FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Classifiying and Mapping e-Tourism data sets

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Business Informatics

eingereicht von

Mete Sertkan, BSc.
Matrikelnummer 00725297

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Mitwirkung: Mag.rer.nat. Dr.techn. Julia Neidhardt

Wien, 2. Mai 2018

Mete Sertkan	Hannes Werthner

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Classifiying and Mapping e-Tourism data sets

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Mete Sertkan, BSc.
Registration Number 00725297

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Assistance: Mag.rer.nat. Dr.techn. Julia Neidhardt

Vienna, 2nd May, 2018

Mete Sertkan	Hannes Werthner

Erklärung zur Verfassung der Arbeit

Mete Sertkan, BSc.
Franz-Mika-Weg 5/4/18, 1100 Wien, Austria

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 2. Mai 2018

Mete Sertkan

Acknowledgements

I would like to thank my supervisor Prof. Hannes Werthner for the great opportunity to do research on the challenging and interesting topic of this work and for sharing his expertise and experience. Further, my sincere gratitude goes to my co-supervisor Dr. Julia Neidhardt for her extraordinary support and mentoring. I also would like to thank both of them for motivating me for my very first scientific contribution and for giving me the chance to attend and present at a scientific conference.

Further, I would like to thank Rainer Schuster, Jakup Steinbauer, Pavel Radkovic, and Leonhard Seyfang for their support at the initial phase of this work.

I would especially like to thank my mother and brother for giving me the opportunity to study at a university and for their great support.

Finally, I would like to thank my beloved wife for always motivating and encouraging me throughout this work.

Kurzfassung

Heutzutage kann die Online-Recherche vor der Buchung eines Urlaubs als übliche Gewohnheit der Kunden angesehen werden. In diesem Zusammenhang zielen Recommender Systeme darauf ab, die Kunden bei ihrer Suche nach den richtigen Produkten zu unterstützen. Jedoch stehen solche Systeme domänenspezifischen Herausforderungen gegenüber, da Tourismusprodukte typischerweise sehr komplex und mit Emotionen verbunden sind. Um diesen Herausforderungen entgegen zu treten, wurden umfassende Benutzermodelle entwickelt, welche die Präferenzen, die Anforderungen und die Persönlichkeit von Kunden berücksichtigen. Eines dieser Modelle ist das sogenannte Sieben-Faktoren-Modell. In dieser Arbeit werden verschiedene Methoden zur automatisierten Bestimmung der Sieben-Faktoren von Tourismusdestinationen und Hotels untersucht, um Recommender Systeme zu ermöglichen die passendsten Produkte vorzuschlagen. Insbesondere werden explorative Datenanalysen, Clusteranalysen und Regressionsanalysen durchgeführt, um nicht nur die Sieben-Faktoren von Tourismusdestinationen und Hotels zu bestimmen, sondern auch ausschlaggebende Attribute von Tourismusdestinationen und Hotels zu identifizieren. Die Resultate der Clusteranalysen zeigen, dass ähnliche Tourismusdestinationen und auch ähnliche Hotels gruppiert werden können. Die identifizierten Gruppen können mit den Sieben-Faktoren assoziiert werden. Die Ergebnisse der Clusteranalysen ermöglichen es nicht einzelne Faktoren des Sieben-Faktoren-Modells zu bestimmen, aber können für eine direkte Zuordnung verwendet werden. Im Gegensatz zu den Clusteranalysen liefern die Regressionsanalysen einen klaren Beweis dafür, dass die Sieben-Faktoren von Tourismusdestinationen und Hotels unter Berücksichtigung der jeweiligen Attribute bestimmt werden können. Grundsätzlich variiert die Qualität der entwickelten Modelle für verschiedene Faktoren des Sieben-Faktoren-Modells und auch für verschiedene Tourismusprodukte (Destination und Hotels). Der in dieser Arbeit vorgestellte Ansatz kann für neue Datenquellen und auch Produkttypen leicht nachvollzogen werden.

Abstract

Nowadays, researching online before booking a vacation can be seen as a common habit of customers. In this context, Recommender Systems (RSs) are aiming to support the customers to find the right products, but they face domain specific challenges since tourism products are typically very complex and related to emotional experiences. To counteract these challenges, comprehensive user models for capturing the preferences and personality of travelers have been introduced. One of these models is the so-called Seven-Factor Model. This work introduces an automated way for determining the Seven-Factor representation of tourism destinations and hotels to enable a matchmaking for RSs. In particular, exploratory data analyses, cluster analyses, and regression analyses are conducted not only to find a mapping of tourism destinations and hotels onto the Seven-Factors, but also to foster a better understanding of the relationship between destination attributes and the Seven-Factors, and between hotel attributes and the Seven-Factors. The main results show that conceptually meaningful groups of destinations and hotels as well can be identified and associated with the Seven-Factors, but they can only be used for direct allocations rather than for determining each factor of the Seven-Factor Model. Furthermore, the regression analyses provide clear evidence that a tourism destination's Seven-Factor representation and a hotel's Seven-Factor representation can be determined by taking the respective attributes into account. In general, the quality of the developed models varies for different factors of the Seven-Factor Model and also for different tourism products (i.e., destination and hotels). Finally, the introduced approach can easily be followed for new data sources and product types.

Contents

CHAPTER 1

Introduction

1.1 Motivation

The relationship between Internet and Communication Technologies (ICT) and tourism can be described as a symbiosis [WK99]. Thus, the tourism landscape has been strongly affected and shaped by the rapid development of ICT during the last decades. Especially, the emergence of World Wide Web (WWW) led to fundamental changes in the tourism ecosystems, both on supply and demand side. Nowadays, consumers have ubiquitous access to vast amounts of information at a very low cost and a greater control in the information acquisition process compared to traditional media channels (TV or print media). Additionally, they are highly connected, allowing them to exchange experiences and more information among each other. However, increasing cognitive costs to process the amount and variety of information could lead to the problem of information overload. This shows the necessity of new techniques and tools to analyze, categorize and visualize information in a proper way [HGXF06]. On the other side, the Web also allows a massive "informatization" of the whole tourism value chain, resulting in many novel value-generating strategies, to satisfy new consumer needs [WR04].

According to recent study [Med14] people rely on online sources to get inspired where to go or how to travel. The study also shows that 65% of the leisure travelers start researching online before a travel decision and social media, photo, video sites and search engines are listed as top online sources for such a purpose. Particularly, in this early phase of decision making a considerable amount of people has difficulties to explicitly express their preferences and needs [Zin07]. Recommender Systems (RSs) are facilitating this decision-making. In [RRS15] Ricci, Rokach, and Shapira are defining RSs as *"software tools and techniques providing users with suggestions for items a user may wish to utilize"* which are *"primarily directed toward individuals who lack the sufficient personal experience or competence in order to evaluate the potentially overwhelming number of alternative items that a website, for example, may offer"*. Particularly, profiling and personalization

techniques might help in such cases, where preferences and needs are unknown or hard to express. Especially in tourism this is a big challenge, since tourism products are considered as very complex (i.e., they typically combine accommodation, transportation, activities, food, etc.), mostly intangible and highly associated with emotional experiences [WR04]. Consequently, travel and destination decisions are usually not only based upon rational criteria but are rather implicitly given. It has been shown that a legitimate way to counteract this issue are personality based approaches, where preferences and personality are combined and used to build a comprehensive user model, which then can be exploited to recommend an item [NW17].

Taking all this into account, it is clear why sophisticated user models, which enhance understanding and processing of user preferences and needs, and tailored techniques, which reduce the cognitive load people are experiencing, have been and still are challenging issues of research in e-Tourism [WASC$^+$15].

1.2 Problem statement

Neidhardt, Seyfang, Schuster and Werthner [NSSW14, NSSW15] introduced a picture based approach to elicit the preferences of a user and a Seven-Factor Model to capture the respective user's profile within a travel recommender system. The Seven-Factor Model is the result of a factor analysis combining the "Big Five" personality traits [Gol90], representing the long-term behavior, and 17 tourist roles [GY02], representing the short-term behavior. These factors form the basis of a seven-dimensional vector space and are referring to travel behavioral patterns summarized as *Sun & Chill-Out*, *Knowledge & Travel*, *Independence & History*, *Culture & Indulgence*, *Social & Sport*, *Action & Fun*, and *Nature & Recreation*. RSs often tend to suffer from the so-called cold start problem [Bur07], i.e., they require historical user data or knowledge about a user's preferences and needs in order to propose appropriate items to that user. However, for a new user this information is typically missing, which is referred to as "cold start". In such cases preference elicitation can be accomplished explicitly, e.g., by asking the user a number of questions or implicitly, e.g., by observing his or her behavior. In the picture based approach, a user's profile is accurately determined by a simple picture-selection process, where the user has just to select three to seven pictures out of a given picture set. In this way, the well-known cold-start problem and tedious questioners for preference elicitation are avoided. A user's profile comprises a score for each of the factors and thus can be seen as a point in the seven-dimensional vector space. In order to provide recommendations to a user, those items have to be determined that are closest to him or to her. Thus, also the items have to be mapped into the vector space, i.e., represented with respect to the travel behavioral patterns. In order to build up a reasonable recommendation base more than 10,000 tourism products were initially mapped manually by experts. Obviously, this approach does not scale and an automated way (i.e., algorithmic approach) of mapping tourism products onto the Seven-Factors is needed.

1.3 Aim of the work

The presented work aims to introduce an automated way of determining the Seven-Factor representation of tourism products. In contrast to [NSSW14, NSSW15], where Points of Interests (POIs) such as activities, events, restaurants, sights etc. are considered as tourism products, this work will focus on tourism destinations and hotels. Glatzer, Neidhardt, and Werthner [GNW18] introduced a text-mining-based method, where hotels are allocated onto the Seven-Factors. Unlike [GNW18], where hotels are directly allocated to the Seven-Factors, this work aims to determine a score for each factor of the Seven-Factor Model. Similarities among tourism destinations and among hotels will be analyzed in order to identify latent conceptually meaningful groups that can contribute to a better understanding. Furthermore, the relationships between the Seven-Factors and attributes of destinations and hotels respectively will be examined in order to map the destinations and the hotels onto the Seven-Factors.

Considering all this, following research questions (RQ) can be stated:

RQ1 How can (semi)structured, non-textual descriptions of tourism destinations be used to enable an automated mapping of tourism destinations onto the Seven-Factors?

- **RQ1.a** Which tourism destination attributes are relevant (most decisive) for this purpose?
- **RQ1.b** To what extent can the Seven-Factor representation of tourism destinations be determined automatically?
- **RQ1.c** Is there an underlying natural structure of tourism destinations, which might be exploited to determine the Seven-Factor scores of tourism destinations?

RQ2 How can (semi)structured, non-textual descriptions of hotels be used to enable an automated mapping of hotels onto the Seven-Factors?

- **RQ2.a** Which hotel attributes are relevant (most decisive) for this purpose?
- **RQ2.b** To what extent can the Seven-Factor representation of hotels be determined automatically?
- **RQ2.c** Is there an underlying natural structure of hotels, which might be exploited to determine the Seven-Factor scores of hotels?

1.4 Methodological approach

The methodological approach, followed in this work, is based on the Cross Industry Standard Process for Data Mining (CRISP-DM) introduced in [She00, CCK+00]. CRISP-DM is the leading methodology in Data Mining (DM) and Knowledge Discovery (KD) projects and can be considered as a "defacto industry standard" [MMS09, kdn14]. All

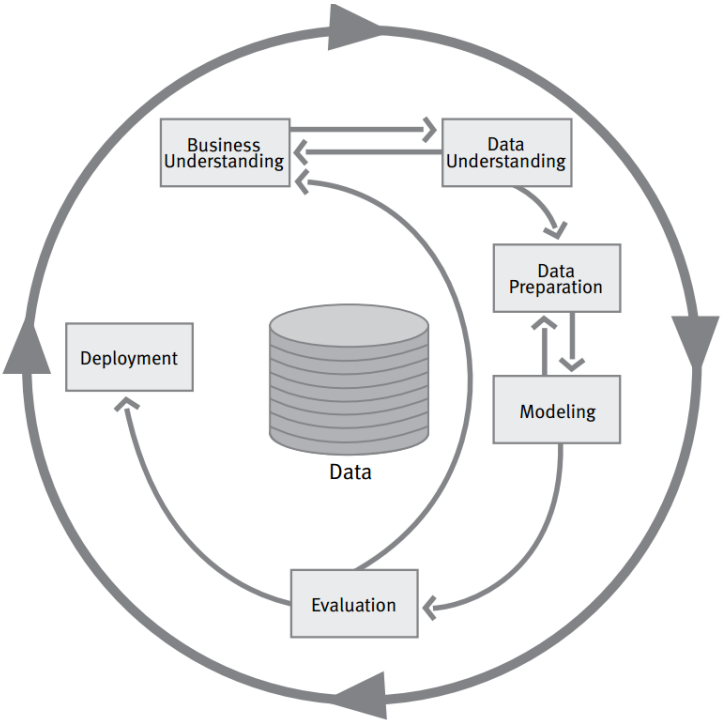


Figure 1.1: Phases of the CRISP-DM Reference Model [CCK+00].

phases of the CRISP-DM reference model are illustrated in Figure 1.1 and briefly described in Table 1.1.

The following describes how this work has applied the CRISP-DM reference model:

Business Understanding. First, the problem was defined, and its relevance was shown. Then, a literature review was conducted to get more knowledge and information about the topic, the related work and the state of the art in the field. In particular, following topics were investigated: the development of different tourist roles; preferences, needs, and personality of travelers; the Seven-Factor Model; the picture-based approach to RSs; data sources in ICT & tourism.

Data Understanding. The data for tourism destinations was provided by Webologen [Gmbc] as SQL-dump and the data for hotels was delivered by GIATA [Gmbb] as an archive of XML-files. Both data sets were transformed into a more convenient tabular format (i.e., CSV). Furthermore, exploratory data analyses were conducted to get more data insights.

Data Preparation. Both data sets were pre-processed in order to feed them into various statistical learning models, mainly realized in Python and the R-Programming Language. First the data was cleansed by deleting unnecessary attributes (textual

data, geo locations etc.), empty columns, and attributes that did not reach a certain frequency threshold. Subsequently, a literature research was conducted in order to find imputation methods for the remaining missing values. Then, the chosen missing value imputation methods were applied and compared.

Modeling. At the beginning of this phase, a literature research was conducted to identify various techniques that could be applied to the given data sets to achieve the stated goals. To identify conceptually meaningful groups clustering methods were implemented using the programming language R and subsequently applied. Furthermore, in order to determine scores for each factor of the Seven-Factor Model different regression models were implemented using the programming language Python and subsequently applied.

Evaluation. The outcomes of the modeling phase were analyzed and discussed thoroughly. First, the resulting models were evaluated and the best performing ones were chosen. Then, the most relevant (decisive) attributes of destinations and hotels were identified and discussed. Furthermore, the results for destinations and hotels were compared together. Also, the outcomes were compared and discussed with the existing literature. Finally, conclusions were drawn and an outline for future work was given.

Deployment. The integration of the outcomes as software as a service or as part of a running project is not the focus of this work. Nevertheless, the publication of this work and thus the sharing of the acquired knowledge can be regarded as deployment in terms of the deployment phase.

1.5 Structure of the work

In Chapter 2 an overview of the related work and the state of the art in the field is presented. In Chapter 3 the provided data for tourism destinations is described and explored. In Chapter 4 clustering and regression analyses with respect to tourism destinations is conducted and evaluated. In Chapter 5 the provided data for hotels is described and explored. In Chapter 6 and regression analyses with respect to hotels is conducted and evaluated. In Chapter 7 methods and outcomes of the previous chapters are analyzed and discussed jointly. Finally, in Chapter 8 conclusions are drawn, limitations are discussed, and an outline for future work is presented.

Phase	Description
Business Understanding	*"This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives."*
Data Understanding	*"The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information."*
Data Preparation	*"The data preparation phase covers all activities needed to construct the final dataset [data that will be fed into the modeling tool(s)] from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools."*
Modeling	*"In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary."*
Evaluation	*"At this stage in the project, you have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached."*
Deployment	*"Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it."*

Table 1.1: Description of all phases of the CRISP-DM reference model [CCK+00].

CHAPTER 2

State of the Art

This chapter provides a brief overview of related work and the state of the art in the field. First, the development of different tourist roles, which are capturing preferences, needs, and the personality of users, is presented. Subsequently, the Seven-Factor Model is introduced and discussed in more detail. Then RSs in the tourism domain are examined and consequently the picture based approach to RSs is introduced and discussed. Finally, different data sources in ICT & tourism research are analyzed and listed.

2.1 Tourist roles

People tend to cluster or classify objects for better understanding, memorization, and communication. For example, even relatively young children are able to divide objects in a picture into classes like buildings, vehicles, people, animals etc. [TSK05]. This is not different in behavioral or motivational studies in the tourism industry. Much research has been conducted in order to identify and categorize tourist roles, describing the relation between a person's travel behavior and his or her preferences, interest, and needs. This has been focus of research since the early seventies, where Cohen [Coh72] introduced following typology of tourist roles: *the organized mass tourist, the individual mass tourist, the explorer,* and *the drifter*). Inspired by Cohen's work and recognizing its limitations, Pearce [P$^+$82] conducted a comprehensive quantitative study on tourist typologies, which resulted in 15 travel related roles (*Tourist, Traveller, Holidaymaker, Jet-setter, Businessman, Migrant, Conservationist, Explorer, Missionary, Overseas student, Anthropologist, Hippie, International athlete, Overseas journalist,* and *Religious pilgrim*). Those roles are describing different travel behavioral and motivational aspects of tourists. Furthermore, Pearce does not differentiate between leisure travelers and those with other intentions like e.g. business or migration, which is crucial due to different underlying motivational influences [MF02].

	Description
Sun Lover	Interested in relaxing and sunbathing in warm places with lots of sun, sand and ocean
Action Seeker	Mostly interested in partying, going to night clubs and meeting people for uncomplicated romantic experiences
Anthropologist	Mostly interested in meeting the local people, trying the food and speaking the language
Archaeologist	Primarily interested in archaeological sites and ruins; enjoys studying history of ancient civilizations
Organized Mass Tourist	Mostly interested in organized vacations, packaged tours, taking pictures/buying lots of souvenirs
Thrill Seeker	Interested in risky, exhilarating activities which provide emotional highs for the participant
Explorer	Prefers adventure travel, exploring out of the way places and enjoys challenge in getting there
jet-setter	Vacations in elite, world class resorts, goes to exclusive night clubs, and socializes with celebrities
Seeker	Seeker of spiritual and/or personal knowledge to better understand self and meaning of life
Independent Mass Tourist I	Visits regular tourist attractions but avoids packaged vacations and organized tours
Independent Mass Tourist II	Plans own destination and hotel reservations and often plays it by ear (spontaneous)
High Class Tourist	Travels first class, stays in the best hotels, goes to shows and enjoys fine dining
Drifter	Drifts from place to place living a hippie-style existence
Escapist I	Enjoys taking it easy away from the stresses and pressures of home environment
Escapist II	Gets away from it all by escaping to peaceful, deserted or out of the way places
Active Sport Tourist	Primary emphasis while on vacation is to remain active engaging in favorite sports
Educational Tourist	Participates in planned study tours and seminars to acquire new skills and knowledge

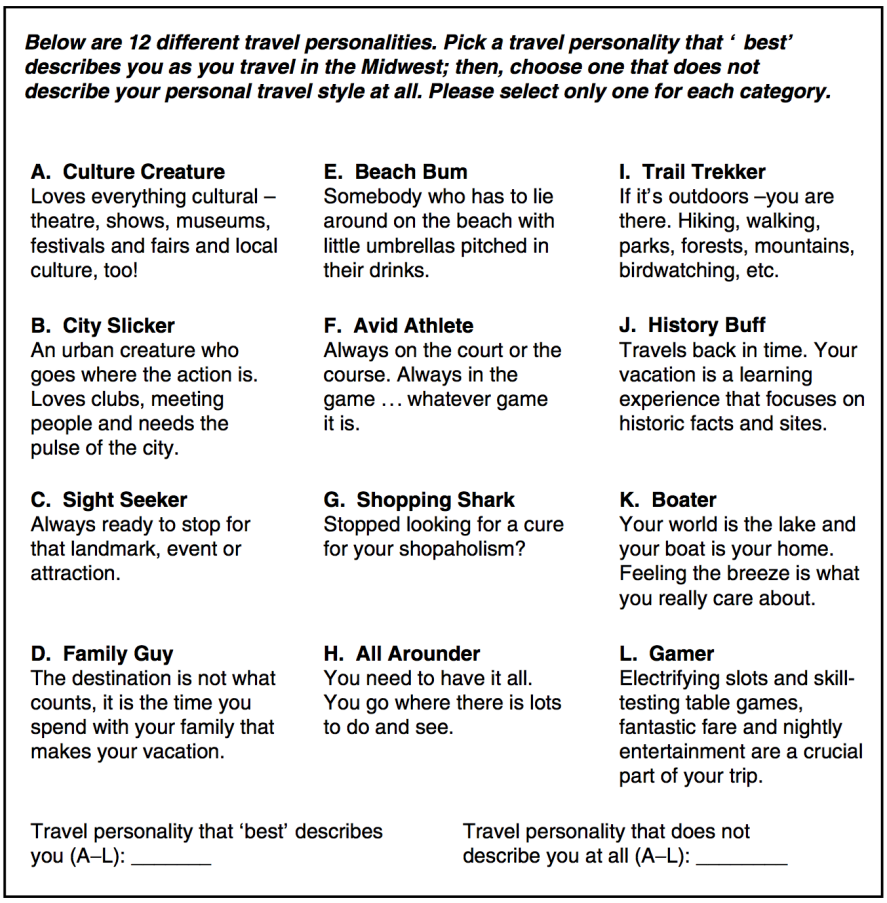
Table 2.1: A Typology of Tourist Roles [GY02].

Based on [Coh72, P$^+$82], but focusing only on leisure travelers Yiannakis and Gibson identified 15 tourist roles. Those roles are able to capture preferences, interest, and motivational indicators of leisure travelers. In a follow-up work Gibson and Yiannakis [GY02] studied the relationship between psychological needs and travel behavioral patterns, which provided significant evidence that they are related and can change over time (short-term behavior). Further, based on the outcomes the original topology of 15 roles is extended to 17 by splitting up the roles *Escapist* to *Escapist 1* plus *Escapist 2* and *Independent Mass Tourist* to *Independent Mass Tourist 1* plus *Independent Mass Tourist 1*. The 17 tourist roles are briefly summarized in Table 2.1. The work of Yiannakis and Gibson [YG92, GY02] delivered a significant contribution, fostered a better understanding of tourist roles, and had a high impact on further studies in this context.

In [GMH$^+$06] Gretzel, Mitsche, Hwang, and Fesenmaier analyzed to which extent pre-defined personality types can contribute to a (destination) recommendation process, by capturing a user's preferences, needs, and in turn, travel behavior. For this purpose, they conducted a questionnaire in order to elicit travel style, psychographic characteristics and actual travel behavior. Travel style was covered by questions addressing importance of certain motivations (e.g., relaxation, excitement, etc.) and importance of certain destination features (e.g., scenery, diversity, good value for money, etc.). Actual travel behavior was determined by questions addressing most recent travel destinations plus activities consumed there. Additionally, the respondents were asked to choose among twelve pre-defined personality types (see Figure 2.1) and among a list of 21 activities.

Figure 2.2 depicts the results of a conducted correspondent analyses assessing the relation between the pre-defined personality types and activities. Note, a correspondence analysis enables a visual exploration of the relationship between variables in a contingency table. A two dimensional solution was proposed, which explained 59,2% of the variance. As one can see, the first dimension captures travel motives ranging from the wish to escape from everyday life to the ambition to gain knowledge. On the other hand, the second dimension reflects the differences between the human made settings (e.g., museums, festivals, etc.) and natural settings (e.g., mountains, lake, beach, etc.). A correspondence between travel personality and respective activities is clearly detectable. For example, the personality type *boater* is very close to the activity *boating* (see Figure 2.2).

Further, a second correspondence analysis was conducted in order to examine the relationship between personality types and destinations (study focus was Northern Indiana region). Surprisingly, no significant evidence of a relationship was found. In order to recommend destinations and to bridge the missing gap a third correspondence analysis was conducted aiming to discover a linkage between activities and destinations. A three-dimensional solution was able to explain 71,6% of the inertia. For the sake of interpretability Gretzel et al. presented in [GMH$^+$06] just a two-dimensional solution (see Figure 2.3), which explained 58.5% of the variance. the first dimension shows the contrast between human made activities and nature-based activities, while the second dimension reflects the differences of travel motives. For example, Chesterton and Angola are located near a variety of nature-based activities, while Merrillville is near a human



Below are 12 different travel personalities. Pick a travel personality that ' best' describes you as you travel in the Midwest; then, choose one that does not describe your personal travel style at all. Please select only one for each category.

A. Culture Creature
Loves everything cultural – theatre, shows, museums, festivals and fairs and local culture, too!

B. City Slicker
An urban creature who goes where the action is. Loves clubs, meeting people and needs the pulse of the city.

C. Sight Seeker
Always ready to stop for that landmark, event or attraction.

D. Family Guy
The destination is not what counts, it is the time you spend with your family that makes your vacation.

E. Beach Bum
Somebody who has to lie around on the beach with little umbrellas pitched in their drinks.

F. Avid Athlete
Always on the court or the course. Always in the game … whatever game it is.

G. Shopping Shark
Stopped looking for a cure for your shopaholism?

H. All Arounder
You need to have it all. You go where there is lots to do and see.

I. Trail Trekker
If it's outdoors –you are there. Hiking, walking, parks, forests, mountains, birdwatching, etc.

J. History Buff
Travels back in time. Your vacation is a learning experience that focuses on historic facts and sites.

K. Boater
Your world is the lake and your boat is your home. Feeling the breeze is what you really care about.

L. Gamer
Electrifying slots and skill-testing table games, fantastic fare and nightly entertainment are a crucial part of your trip.

Travel personality that 'best' describes you (A–L): _____

Travel personality that does not describe you at all (A–L): _____

Figure 2.1: Travel-related personality types [GMH+06].

made activity like museum/concert.

To sum up, Gretzel et al. [GMH+06] demonstrated that tourist roles can be used to recommend touristic activities and, in turn, destinations.

Tourism recommender systems, which are eliciting travel personalities (roles) and recommending items based on that are rare. A really basic example for such a system is the *Airbnb Trip Matcher* [Air17]. It comprises ten predefined travel personality types and each type has just one corresponding destination, which is recommended. All types and corresponding destinations are summarized in Figure 2.4.

Another similar application, which matches a user to pre-defined travel personality types is published by BuzzFeed [Buz15]. The application is just a gimmick rather than a recommender system. The proposed travel types are *The Wanderer, The Expert, The Cautionary Tale, The Minimalist, The Lone Wolf, The Pack Mule, The Cultural Sponge, The Partier,* and *The Repeat Offender.*

In both of the introduced cases, it is not clear how the typologies are defined and if

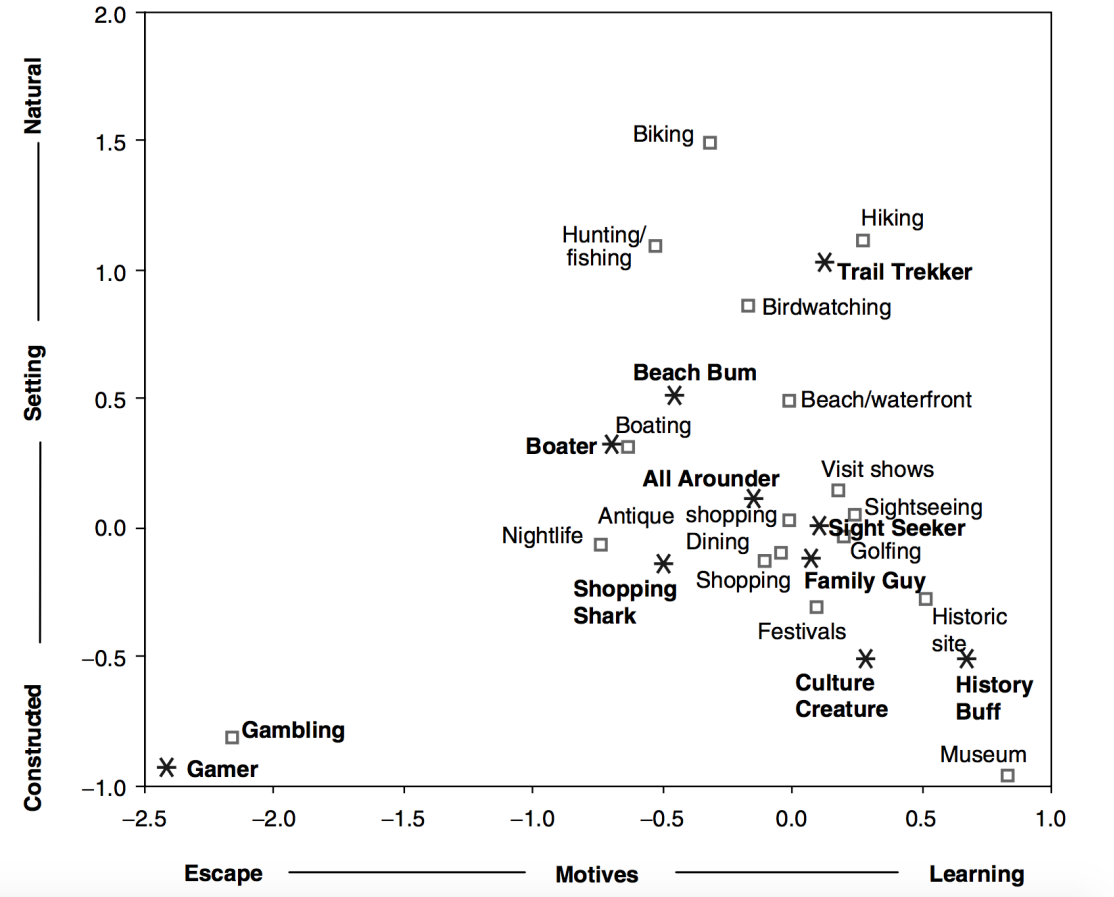


Figure 2.2: Relationship between travel personality and travel activities [GMH+06].

there is an underlying theoretical background. As opposed to this, this thesis is based on the Seven-Factor Model introduced by Neidhardt et al. [NSSW14, NSSW15], which is discussed in more detail in section 2.2.

2.2 Seven-Factor Model

As already mentioned, Gibson and Yiannakis [GY02] introduced a well-established classification framework, distinguishing 17 different tourist roles to capture short-term preferences of tourists, i.e., preferences, which might change depending on the context (e.g., seasonality like summer or winter, special occasions, single or group, etc.). It has also been shown that tourist roles can be related to personality traits. Delić, Neidhardt and Werthner [DNW16] provide significant evidence that there are relations between the well-established "Big-Five" personality traits [Gol90] and the 17 tourist roles [GY02]. In [MDW03] Matthews, Deary, and Whiteman argue that *"large-scale reviews and large single studies offer overwhelming evidence for the stability of personality traits over many*

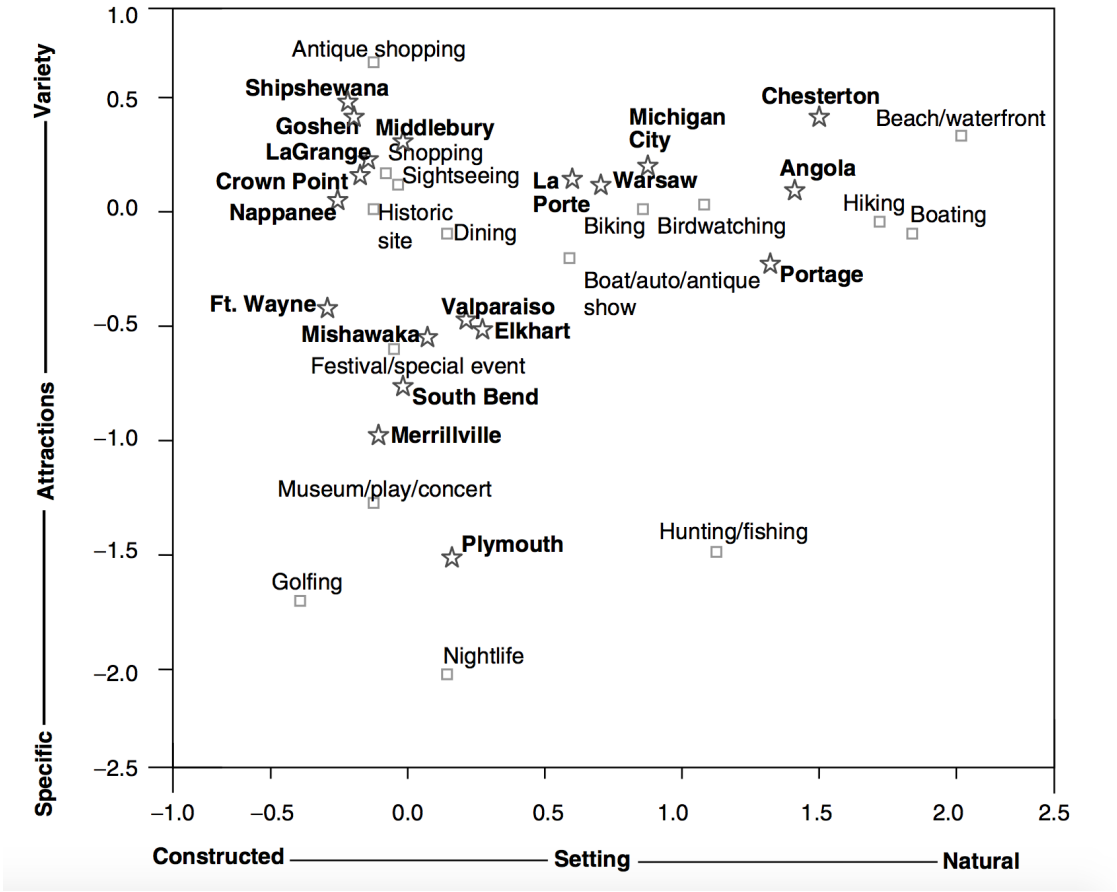


Figure 2.3: Relationship between travel activities and destinations [GMH+06].

years." Thus, they can be considered as long-term preferences of a person [WRS02]. In Table 2.2 the five dimensions of "Big Five" personality traits (also known as the Five-Factor Model) are summarized.

Both, the "Big Five" personality traits and the 17 tourist roles are well-established frameworks and have been subject and bases to many empirical and behavioral studies. Thus, there are existing standardized methods to assess and measure both of them. In [NSSW14, NSSW15] Neidhardt et al. conducted an online and offline survey, with 30 questions addressing the tourist roles and 20 questions addressing the personality traits. About thousand participants completed the questionnaires. Upon the collected data, they conducted a factor analysis in order to reduce the 22 dimensions (the "Big Five" personality traits plus 17 tourist roles) and summarize them in fewer dimensions. The factor analysis resulted in seven independent factors, which are able to capture different travel behavioral patterns. They are summarized in Table 2.3.

These factors are easier to process cognitively as well as computationally compared

Figure 2.4: Airbnb Trip Matcher - Types and Destinations [Air17].

Dimension	Trait facets
Neuroticism	Anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability
Extraversion	Warmth, gregariousness, assertiveness, activity, excitement seeking, positive emotions
Openness	Fantasy, aesthetics, feelings, actions, ideas, values
Agreeableness	Trust, straightforwardness, altruism, compliance, modesty, tender-mindedness
Conscientiousness	Competence, order, dutifulness, achievement striving, self-discipline, deliberation

Table 2.2: Trait facets associated with the five domains of the Five-Factor Model of personality [MDW03].

Factor	Description
Sun & Chill-Out	a neurotic sun lover, who likes warm weather and sun bathing and does not like cold, rainy or crowded places;
Knowledge & Travel	an open minded, educational and well-organized mass tourist, who likes traveling in groups and gaining knowledge, rather than being lazy;
Independence & History	an independent mass tourist, who is searching for the meaning of life, is interested in history and tradition, and likes to travel independently, rather than organized tours and travels;
Culture & Indulgence	an extroverted, culture and history loving high-class tourist, who is also a connoisseur of good food and wine;
Social & Sports	an open minded sportive traveller, who loves to socialize with locals and does not likes areas of intense tourism;
Action & Fun	a jet setting thrill seeker, who loves action, party, and exclusiveness and avoids quiet and peaceful places;
Nature & Recreation	a nature and silence lover, who wants to escape from everyday life and avoids crowded places and large cities.

Table 2.3: Seven-Factor Model [NSSW14, NSSW15].

to the original 22 dimensions. However, dimensionality reduction not only decreased computational and cognitive cost, but also lead to a better understanding and more insights. In [NW17] Neidhardt and Werthner showed that based on different demographic characteristics different user groups can be well distinguished within this model.

2.3 Recommender Systems in Tourism

Travel and tourism have always been major application domains for Web-related services [WK99]. As the amount of information on the Web started to rise, the call for techniques to cope with information overload began to grow. One answer to that are RSs. From the supplier's perspective, RSs are aiming to bring right products to right customers, in order to increase customer experience, satisfaction, trust, and in turn profit. On the other hand, from a consumer's point of view, RSs are aiming to provide suggestions in order to support and simplify various decision-making processes. In case of tourism these decisions might be: Where to go? How to travel? Where to stay? What to do? and much more. Whereas, an item (suggestion) can be a destination, a hotel, an activity, a flight etc.

During the last decades, many recommendation techniques evolved and have been successfully deployed and thoroughly evaluated. Following well-established techniques are the most common ones in the literature [RRS15]:

Content-based. This technique recommends items that are similar to the ones the user liked or bought before. Similarity is measured by comparing features associated with the items. In a more classic way, these techniques are aiming to match features of the user model with features of the item model.

Collaborative Filtering. Here items, which have been of interest to other users with a similar taste like the active user, are recommended. Similarity in taste is modeled based on the similarity of previous rating behavior.

Demographic. Systems based on this technique, are recommending items to users based on their demographic characteristic, such as age, gender, country etc.

Knowledge-Based. These systems are relying on specific domain knowledge about preferences and needs of users and which items (i.e. item features) meet these needs. Here, a utility function asses how good a problem solution (i.e. recommendation) meets a problem definition (a user's preferences and needs). This kind of systems can be seen as case-based. Another sub category of knowledge-based RSs comprises constraint-based systems. In constraint-based systems user requirements are matched with item features by rules, which are defined based on domain-knowledge. Whereas, case-based systems rely on similarity measures (and not rules).

Community-Based. Here, items which have been of interest to a user's friends are recommended. Thus, in this approach the social relationships of users are exploited. Research has shown that people tend to rely on suggestions of people they know more than of unknown people (although they can have same characteristics).

Hybrid Recommender Systems. This approach combines one or more of the techniques mentioned above. Such systems aim is to overcome shortcomings of one methods by combining and exploiting benefits of other methods. For example, one can overcome the known "cold start" problem of collaborative filtering by conducting a content-based approach for new items (where no ratings exist).

Most of the listed techniques rely on user rating behavior and were proposed for products such as movies, music, or books. However, since traveling is costly and time consuming, there are typically less rating data in the tourism domain, which leads to less accurate personalization techniques, compared to other products [NSSW15]. Another challenge for RSs in the tourism domain is, that tourism products are complex (e.g., a bundle of accommodation, transportation, activities etc.), intangible and highly associated with emotional experiences [WK99]. In order to bundle and recommend the right tourism product RSs are relying on content and knowledge [NSSW15]. Also, Burke and Ramezani

[BR11] argue that most appropriate recommendation techniques in the matter of tourism are either knowledge based and/or content-based.

This work aims to find an automated way of determining the Seven-Factor representation of tourism destinations and hotels to enable a matchmaking with user profiles (i.e., Seven-Factor representation of users). The picture-based approach to RSs [NSSW14, NSSW15] uses a gamified and user centric way to elicit the Seven-Factors of a user. Preferences and needs of a user are determined via a simple picture selection process. Users are addressed on an emotional, implicit level and do not have to state their preferences explicitly. This gamified and simple method, which can be considered as content- and knowledge-based approach, counteracts peoples difficulties in explicitly expressing their preferences and needs. As research has shown, such difficulties occur especially in the early phase of travel decision making process [Zin07]. Furthermore, it helps to overcome the so-called cold-start problem [Bur07].

According to Garcia, Sebastia, and Onaindia [GSO11] tourism RSs can be distinguished into two types: one focusing on destination selection the other on activities that can be performed at a certain destination. The presented work can be considered as part of the both groups since it considers tourism destinations and hotels as recommendation items. In contrast to [NSSW14, NSSW15], where the focus lies on Point of Interests (POIs), e.g., activities, events, restaurants, sights. Much research has already been conducted targeting destination recommender systems [FWW06, BMV14], but they are mainly focusing on distinct regions or POIs in a destination. There are few, moreover, that are focussed on personality traits and motifs of a user (see for example [BER14]).

2.4 A picture based approach to recommender systems

A crucial part of the picture based approach to RSs [NSSW14, NSSW15], namely the Seven-Factor Model of travel behavioral patterns, has already been introduced in Section 2.2. A user profile (preferences and needs) is captured through the Seven-Factors and also recommendation items, here POIs, are described via the Seven-Factors. Thus, a recommendation can be done (i.e., items can be ranked) by just calculating the distance between a user and POIs. One can say, that the Seven-Factor Model is spanning a seven-dimensional vector space, where each dimension refers to a travel behavioral pattern. Hence, user profiles and POIs can be seen as a point in this vector space, such that a recommendation can be done by a certain distance measure. Neidhardt et al. are using the Euclidian distance herefore. For a better understanding, Table 2.4 shows the Seven-Factor representation of a user. Furthermore, the Seven-Factor representation of some POIs and the respective Euclidian distance to the user is provided.

In order to determine a user profile (Seven-Factors of a user) accurately, he or she has just to select a three to seven pictures from a given picture set. In the literature, most common approaches to elicit a user's preferences, needs, and personality are critique-based. Thus, a user has to communicate with the systems or fill out questionnaires in order to retrieve his or her personality, preferences and needs. As already stated, many people have

	f1	*f2*	*f3*	*f4*	*f5*	*f6*	*f7*	Distance
User profile	0.14	0.70	0.79	0.88	0.30	0.20	0.12	
Stonehenge	0.09	0.74	0.75	0.89	0.21	0.03	0.17	0.213
Daibutsu	0.06	0.76	0.82	0.82	0.38	0.03	0.07	0.229
Wat Maheyong	0.06	0.65	0.80	0.84	0.35	0.00	0.11	0.231

Table 2.4: User profile and recommended POIs [NSSW15].

difficulties in explicitly expressing their preferences and needs and usually travel decisions (where to go, how to travel etc.) are rather not rationally taken but implicitly given. Thus, by such a simple method of picture selection, the picture based approach avoids tedious communication with the system and addresses also the implicit and emotional level of the decision making.

In order to relate pictures with the Seven-Factors, Neidhardt et al. [NSSW14, NSSW15] asked participants of a workshop to assign pictures (out of 102 travel related pictures) to the Seven-Factors. Additionally, the participants had to find a consensus. In a second study, people (N=105) were asked to select and rank ten pictures out of the 102 travel related pictures by considering their next hypothetical trip. Furthermore, the participants had to fill out the same questionnaires, which were used for the development of the Seven-Factors (see Section 2.2). People tend to select between three and seven pictures. The initial set of 102 travel related pictures is reduced by simply omitting the most and least frequent pictures. This resulted in a more concise set of 63 travel related pictures (capturing the most information). In a third step, 15 travel experts were asked to assign three to seven pictures from the reduced picture set to over 10,000 POIs. Additionally, they had to determine for each POI the affiliation to each travel behavioral factor, simply assigning a value between 0 and 1 for each factor. Through multiple regression analysis (ordinary least squares) the relation between pictures and the Seven-Factors were quantified. This approach resulted in seven equations, each for one of the Seven-Factors. Also, the amount of pictures (minimum three and maximum seven) and their sequence of selection are considered in the resulting model.

$$f_j^u = \sum_{i=1}^{63} b_{ji} x_i^u \tag{2.1}$$

$f_j^u, j = 1, ..., 7$ shows that for user u seven models are fitted, each for one factor. The values $x_i^u, i = 1, ..., 63$ are calculated for each picture and user u. Finally, $b_{ji}, j = 1, ..., 7, i = 1, ..., 63$ are the coefficients to be estimated for each picture and equation. Equations 2.2 shows how pictures, their selected amount and the sequence are actually quantified.

$$x_i = 7\frac{-k + n + 1}{\sum_{j=1}^{n} j} \tag{2.2}$$

$x_i, i = 1, ..., 63$ is the value for the i-th picture if it is chosen and ranked to the k-th place. If a picture is not chosen than this value is 0. Furthermore, $k = 1, \ldots, 7$ denotes the rank of the chosen picture and $n = 1, \ldots, 7$ shows the total amount of chosen pictures. This method outperformed two other suggested approaches, where only dummy variables were used or the amount of chosen pictures was not considered at all.

In [GMH+06] Gretzel et al. pointed out that people can have a variation of travel preferences simultaneously. This crucial finding is considered by the Seven-Factor Model implicitly (i.e., combination of "Big Five" and 17 roles) and explicitly by depicting a user as a mixture of the Seven-Factors (see Equation 2.1 and Table 2.4).

This non-verbal way of eliciting people's preferences and needs through a simple picture selection not only counteracts the mentioned difficulties in explicitly expressing one's preferences and needs, but also gamifies the way of interaction with the system. Krinninger [Kri12] showed in a user response evaluation, that this way of interaction is experienced as interesting, exciting, and inspiring.

2.5 Data sources in Tourism & ICT Research

This section gives a short overview of different data sources, which are commonly accessed in the interdisciplinary area of tourism and ICT research. In order to do so, all papers published in the ENTER 2017 proceedings [SS17] are taken into account. Based on their respective origin the commonly used data sources can be separated into three main groups namely, government-based, non-government/industry-based, and self-acquired.

Government-based. This kind of data are mostly provided by governments thanks to an open data policy. Typically, they can be accessed through APIs or downloaded through governmental online platforms. Another option to access data within this group is by cooperating, for example with tourism ministries, government operated destination marketing organizations (DMO) or convention and visitors bureaus (CVB). Accessed data are usually: arrival data, income, price level, transportation costs, advertising expenditure.

Non-government / industry based. In some cases, researchers have access to data through cooperations with industry partners, for example in one study data from a cellular/mobile provider is used in order to analyze strategic visitor flows. Another common source within this group are (meta) search engines, e.g. using Google Trends in order to improve arrival prediction. Also, data from online traveling agencies (OTA) are on the focus of many researchers, which are accessed either through APIs or have to be "self-scraped". On the other hand, in some studies researches are directly cooperating with hotels (instead of agencies), for example they are using data from property management systems (PMS) in order to analyze the impact of IT-enabled customer experience management on service perceptions and performance. Social media and user generated content (UGC) are getting more and more popular. Platforms like Twitter, Flickr or similar are providing well elaborated APIs to access their data. Another way to get such

data is to extract it with a crawler, for example user reviews from TripAdvisor. Such data is then used for behavioral analysis, network analysis, arrival predictions and much more.

Self-acquired. Most studies based on the previous groups are data driven or based on large amounts of data. Here "self-acquired" refers to the "old, traditional" way of collecting data, namely through questionnaires, interviews or similar. Surprisingly, many studies still rely, and probably will rely, on data collected this way. Such data can be retrieved either online (e.g. online questionnaires, Skype interviews etc.) or offline, and have usually small sample sizes. On the other hand, there are some behavioral studies, which totally rely on big data emerged through emotion tracking, such as electrical activity of the brain (retrieved via Electroencephalography = EEG) or electro dermal activity (EDA). Also in such studies the number of participants (considered people) are usually low compared to social media-based samples for example, but the amount of produced data through this kind of monitoring is vast, such that big data / data driven approaches can be applied.

In this work, two data sets (i.e., data sets for destinations and hotels) have been provided by industrial partners and thus they can be considered as industrial-based. Chapter 3 and Chapter 5 are introducing and discussing both data sets in detail.

CHAPTER 3

Tourism Destinations Data

In this chapter the provided data set for tourism destinations is thoroughly described and analyzed. The data for tourism destinations is provided by Webologen [Gmbc], a German internet and marketing agency, whose focus is tourism and IT-services. In addition to the data set a labeled data sample (manually mapped onto the Seven-Factors) is provided by tourism experts. The upcoming sections are covering following topics: univariate and multivariate analysis, missing values and treatment, feature engineering.

3.1 First Insights

In [B+03] Beirman refers to a tourism destination as *"a country, state, region, city or town which is marketed or markets itself as a place for tourists to visit"*. In this work destinations are defined in a similar way, except that the range is wider, i.e., from a hamlet with a population smaller than 100 to a metropolis with a population larger than one million. The data is provided as a SQL-dump and consists of more than 30,000 destinations all around the world.

Figure 3.1 shows the structure of the tables in the SQL-dump and the relations among them. Destinations are described through 22 geographical attributes and 27 motivational ratings.

Motivational ratings lie in the interval [0,1] and describe the degree of suitability for a particular motif. Following 27 motifs are listed: *nightlife, wellness, shopping, nature & landscape, image & flair, culture, sightseeing, entertainment, mobility, price level, accommodations, gastronomy, beach & swimming, golf, scuba diving, kite & windsurfing, hiking, cycling, horseback riding, winter sports, sports, family, peacefulness, surfing, sailing, gays, mountain biking.* The motivational ratings are determined by the e-Tourism company by considering factors such as infrastructure,

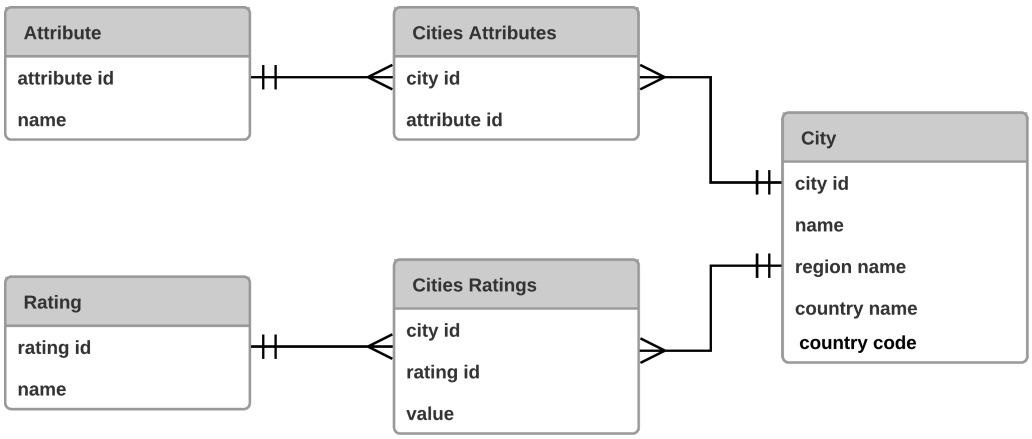


Figure 3.1: ER-Diagram of the Webologen SQL-dump.

climate, user opinions, number of services, image, and marketing. However, not all details are disclosed and thus it is not known how exactly the scores are determined.

Geographical attributes are given in binary format and describe the presence or absence of a particular geographical attribute. Following 22 attributes are listed: *sea, mountain, lake, island, sandy beach, metropolis, forest, river, desert, old town, pebble beach, sand & pebble beach, hill, swamp, volcano, fjord, flat decaying sandy beach, beach promenade, wine-growing, heath, health resort, winter sports resort.*

All possible attributes and ratings are persisted in the tables *Attribute* and *Rating*. On the other side, there are over 30,000 tourism destinations persisted in the table *City*. This table contains an identifier for each destination and textual descriptions to capture destination name, region name, country name, and country code in ISO 3166-1 alpha-2 format, for example AT for Austria. In the table *Cities Attributes* tuples of geographical attributes and tourism destinations are recorded, e.g. (Vienna, *old town*). Similarly, the table *Cities Ratings* persists the motivational ratings of a tourism destination with corresponding (rating) value, e.g. (Vienna, *culture*, 0.99). A major drawback of such a structure is that a tourism destination does not necessarily have an entry for each rating or attribute. Thus, in many cases it is not clear if a destination does in fact not have such attribute or rating, or the data is missing. This ambiguity leads to many "missing values" and in turn to a sparse data set.

Although, detailed descriptions of the motivational ratings and geographical attributes are provided by Webologen, it is still a proprietary solution of a German e-Tourism company. Thus, it is not clear if there is an underlying theoretical background of such rating and attribute structure.

Almost all countries are represented in the database, but the majority (65%) of destinations are located in the USA, Germany, France, Italy, Spain, Great Britain, Austria,

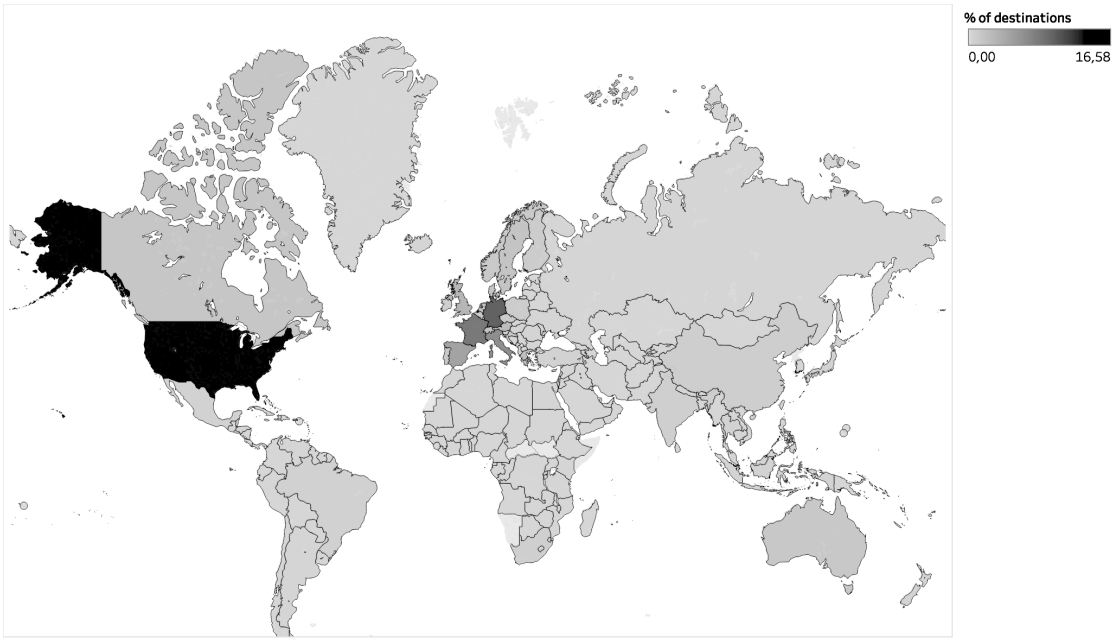


Figure 3.2: Webologen data set - Distribution of tourism destinations over countries.

Greece, Switzerland, and Sweden. This can also be observed in Figure 3.2, where the distribution of tourism destinations over countries is presented as a heat map.

Summary statistics of the distributions of the different motivational ratings for the tourism destinations are listed in Table 3.1. Mean, standard deviation, minimum, median, and maximum are determined by considering only non-missing values. Interesting to see is that motivational ratings actually have a range of minimum 0.01 to maximum 1 and do not start from zero. Thus, tuples of <zero rating, tourism destination> are not persisted in the database, which saves space and transactions costs, but leads to many empty cells in this case. The majority of motivational ratings have a missing value rate greater than 70% and many of them have a missing rates even greater than 90%. Missingness in data will be further analyzed in the upcoming section. Note, that the vast majority of ratings with high missingness (>50%) have an average value greater than 0.5. Hence, one can argue that the Webologen data is biased towards good ratings. Another interesting observation is that the ratings *nature & landscape, hiking, peacefulness, cycling,* and *mountain biking* are not only similar in missingness but also have similar and high average values of 0.61-0.71. Those ratings are not only similar, but they can also be considered as nature and recreation related. Correlation among features are also analyzed and discussed in the upcoming sections.

In Table 3.2 frequencies and missingness of each geographical attribute are listed. The proportion of missing values in geographical attributes is far worse than in motivational ratings. Note, that 66.92% of destinations do not have any geographical attribute at all. Apart from this, most frequent geographical attributes are *island, sea* and *sandy*

	missing(%)	mean	std	min	median	max
pricelevel	0.84	0.57	0.15	0.10	0.58	1.00
gastronomy	14.23	0.35	0.18	0.05	0.27	1.00
sports	14.24	0.34	0.17	0.05	0.27	1.00
accommodations	14.26	0.35	0.17	0.05	0.28	1.00
shopping	14.65	0.52	0.13	0.02	0.52	1.00
nightlife	15.76	0.48	0.14	0.05	0.46	0.99
entertainment	17.19	0.26	0.18	0.01	0.17	0.98
nature_landscape	57.28	0.73	0.16	0.10	0.77	1.00
hiking	58.68	0.70	0.17	0.12	0.71	0.98
peacefulness	60.82	0.72	0.18	0.06	0.76	1.00
cycling	62.68	0.66	0.15	0.09	0.72	0.96
mountainbiking	72.45	0.69	0.20	0.09	0.70	0.99
culture	77.63	0.61	0.16	0.03	0.62	1.00
wintersports	79.49	0.24	0.15	0.01	0.22	0.93
image_flair	85.28	0.79	0.15	0.09	0.80	1.00
mobility	90.36	0.65	0.15	0.13	0.66	1.00
beach_swimming	90.80	0.79	0.19	0.01	0.84	1.00
wellness	90.90	0.56	0.17	0.05	0.57	1.00
family	91.54	0.64	0.20	0.04	0.65	1.00
golf	93.78	0.59	0.20	0.01	0.57	1.00
sightseeing	94.05	0.70	0.20	0.05	0.75	1.00
sailing	95.34	0.55	0.22	0.01	0.54	1.00
diving	95.35	0.55	0.24	0.01	0.55	1.00
horsebackriding	97.47	0.58	0.20	0.01	0.49	1.00
kite_windsurfing	98.27	0.64	0.28	0.01	0.77	1.00
surfing	98.43	0.55	0.30	0.01	0.63	1.00
gays	99.73	0.68	0.31	0.02	0.82	1.00

Table 3.1: Summary statistics of the motivational ratings

	frequency(%)	missingnes(%)
island	14.62	85.38
sea	11.10	88.90
sandy_beach	8.98	91.02
mountains	4.44	95.56
forest	3.39	96.61
volcano	3.24	96.76
hill	3.18	96.82
old_town	2.99	97.01
lake	2.26	97.74
wintersports_resort	2.18	97.82
health_resort	1.88	98.12
river	1.45	98.55
metropolis	1.42	98.58
wine_growing	0.90	99.10
sand_pebblebeach	0.67	99.33
flat_decaying_sandy_beach	0.66	99.34
beach_promenade	0.54	99.46
pebblebeach	0.53	99.47
desert	0.51	99.49
heath	0.05	99.95
fjord	0.04	99.96
swamp	0.03	99.97

Table 3.2: Frequencies and missingness of geographical attributes.

beach, which can be considered as typical attributes of beach resorts. On the other hand, least frequent attributes are *heath, fjord,* and *swamp* with frequencies of just 0.03-0.05%. Also noteworthy is that Webologen is differentiating between five kinds of beach types, namely *sandy beach, pebble beach, sand & pebble beach,* and *flat decaying sandy beach*.

In Section 3.2 and Section 3.3 missing values will be further analyzed and treated and some feature engineering will be conducted.

3.2 Missing Data Analysis

3.2.1 Methods and Concepts

Missing data is a huge topic on its own in the Data Science world. There are many scientific publications and books related to analyzing and/or treating missing data. Before analyzing missing values in the Webologen data set, it is important to understand and differentiate following two concepts in missing data analysis (a) missing data patterns and (b) missing data mechanisms. Often both terms are used interchangeably, but actually

they have very different meanings. In [End10] missing data patterns are characterized as a way to describe the location of "holes" (empty cells) in a data set, but not the reason behind. Also, missing data mechanisms are not giving causal explanations for the missingness in data, but they provide generic mathematical relationship between the data and its missingness.

Enders distinguishes in [End10] six prototypical missing data patterns, which are depicted in Figure 3.3.

Univariate pattern. Missing Values are just appearing in one variable, i.e. isolated to a single feature. This kind of pattern is rare, but may occur in experimental studies.

Unit nonresponse pattern. This is a very common pattern, where there is data for all entries of distinct features (e.g. census data) in a sample, but some surveys people refuse to answer.

Monotone missing data pattern. This is a very typical pattern for longitudinal studies, where for example participants (e.g. patients negatively reacting to some drug) drop out or are excluded from the study.

General missing data pattern. This is probably the most common pattern, where missingness is appearing in an arbitrary fashion. However, this randomness can still have a systematic or dependence in behind.

Planned missing data pattern. Planned missing data approaches are a proper way to reduce the load of participants, but still get much questionnaire items back. Missing data is constructed the way that it can be subsequently imputed.

Latent variable pattern. Origin of such a pattern are latent variable analyses. Although latent variables are per definition unknown and are not necessarily missing data per se, researchers have conducted missing data analysis and methods in order to estimate them.

In contrast to the missing data patterns, the so called missing data mechanism are describing how the probability of missing data is related to the data it self. The most common and widely used classification scheme of such mechanism was introduced by Rubin [Rub76]. According to that scheme, missing data mechanism can be separated into three types, which are discussed in [End10] thoroughly and can be summarized as follows:

Missing at random (MAR). Unfortunately, the name "missing at random" can be misleading, since it does not mean that the missing data is appearing randomly. Actually, missing data are called missing at random, if the probability of missing data on a variable Y is related to one or more other variables in the data set, but not to the values of Y itself.

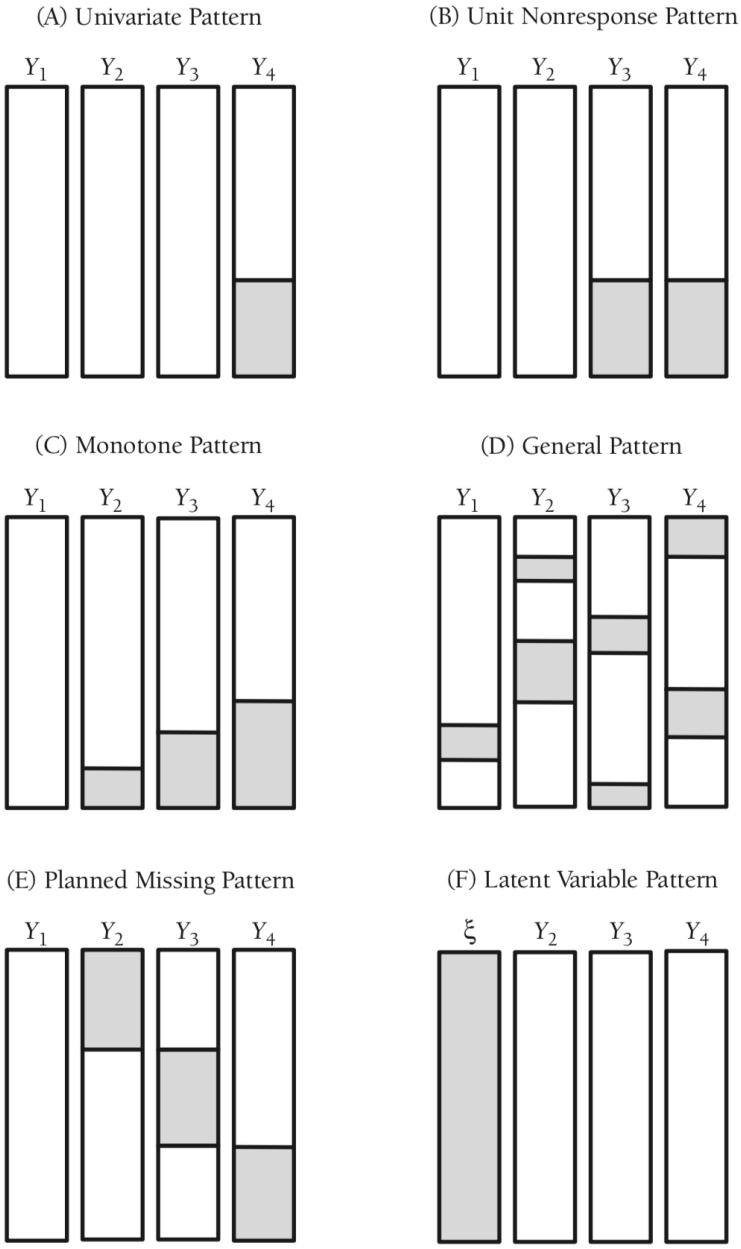


Figure 3.3: Six prototypical missing data patterns. The shaded areas represent the location of the missing values in the data set with four variables [End10].

Missing completely at random (MCAR). Data is called missing completely at random, if missingness is appearing in a pure random way. Hence, the probability of missing data on a variable Y is whether related to other measured variables nor to values of Y itself. Thus, one can say that MCAR is a more restrictive condition than MAR, since it takes missingness as totally unrelated to the data.

Missing not at random (MNAR). In this case the probability of missing data on a variable Y is related to the value of Y itself. Asking somebody for his or her salary in a survey is a good example, where missingness of the variable salary will probably depend on the salary itself.

So far, a brief excerpt of missing data theory is presented, but a more detailed overview and discussion can be found in [End10, Gra12].

3.2.2 Analysis

Considering the summary statistics of the Webologen data set, presented in Table 3.1 and Table 3.2, one can get a first intuition of the extent missingness in data. Next, these first insights will be broadened and enhanced by conducting missing data analysis.

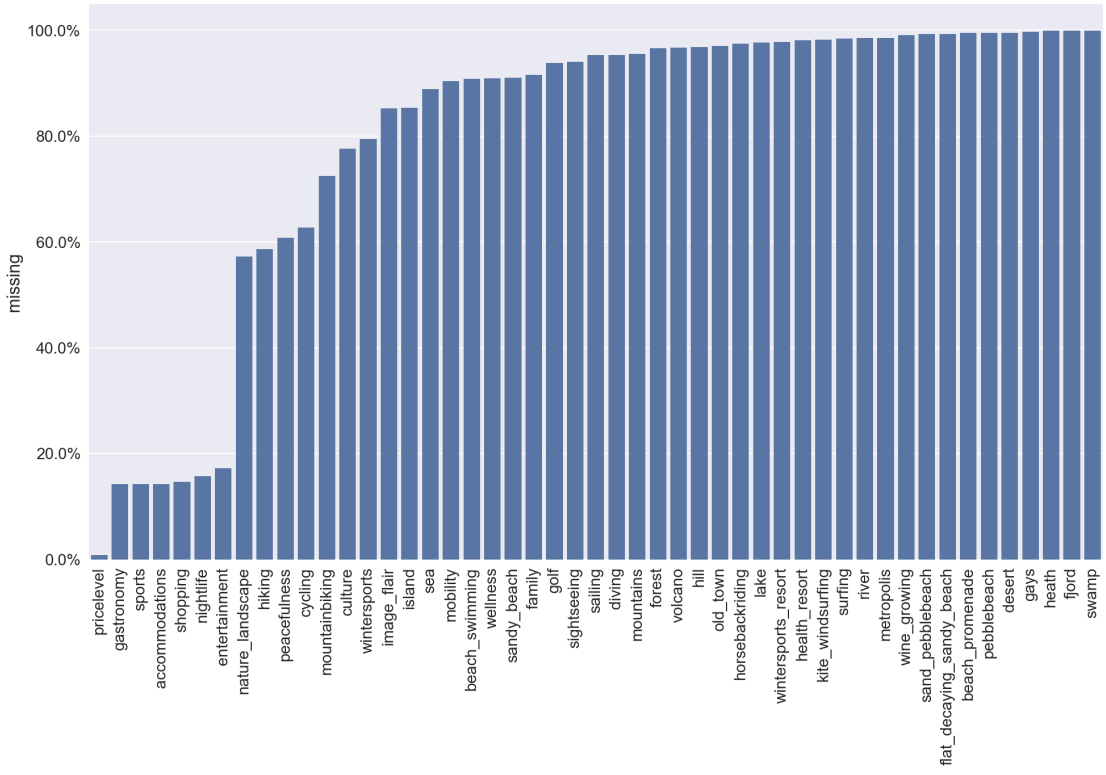


Figure 3.4: An overview of missingness in tourism destination features.

Figure 3.4 shows the extent of missingness in tourism destination features, where most features have a missing rate over 80%. Only motivational rating *price level* is almost complete. Also, motivational ratings *gastronomy, sports, accommodations, shopping, nightlife,* and *entertainment* have relatively low missing value rates in comparison to all other tourism destination features. This might be a sign of different information retrieval approaches, which might have been used by Webologen. Probably, those features were gathered automatically, by aggregating easy to get quantitative measures, like number of bars, hotels, or restaurants. Whereas, all other features might rely more on manual assignments.

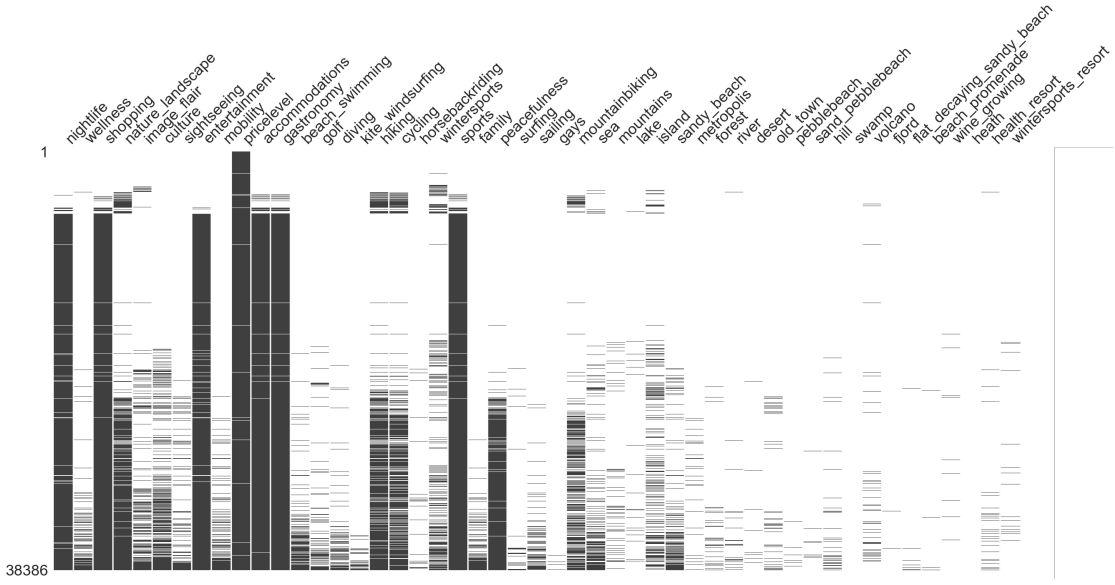


Figure 3.5: Nullity matrix of destination features. Dark shows the presence and white the absence of particular features in each destination.

Figure 3.5 show the nullity matrix of tourism destination features. Each row in the matrix represents a particular destination and each column a feature. A black dash shows the presence of a feature in a destination, whereas the absence is denoted with an empty space. In some destinations only one feature is present and on the other hand, there are some destinations, which are described with 35 features. Destinations are possessing on average ten features.

Referring to the introduced missing data patterns (see Figure 3.3), one can observe here the general pattern. At first glance missingness seems arbitrary, but by examining the nullity matrix in detail, one can observe some similarities and other patterns. The right-hand side is emptier than the left-hand side. This is expected, since on the right-hand side geographical attributes of destinations are shown. Due to the binary nature of geographical attributes and since zero values are not persisted in the Webologen system, such differences are reasonable and expected. Further, one can see that the

most frequent destination features (*nightlife, shopping, accommodations, gastronomy,* and *sports*) also show similar missing data patterns. Also, motivational ratings *nature & landspace, hiking, cycling,* and *mountain biking* have similar missing value patterns. Surprisingly, peacefulness seems to be not so similar to this group as one would expect.

Visually exploring the nullity matrix helps to quickly localize patterns in the missingness. Another more objective way to find relations among the missingness of features is to examine the nullity correlations. The pairwise nullity correlation scales from -1, one feature is always missing if the other one is present, to 1, one feature is always present if the other one is also. In order to encounter and visualize trends going deeper than the pairwise nullity correlations one can hierarchically cluster nullity correlations of features and display the result as a dendrogram. Figure 3.6 show such dendrogram for the nullity correlations of the destination features.

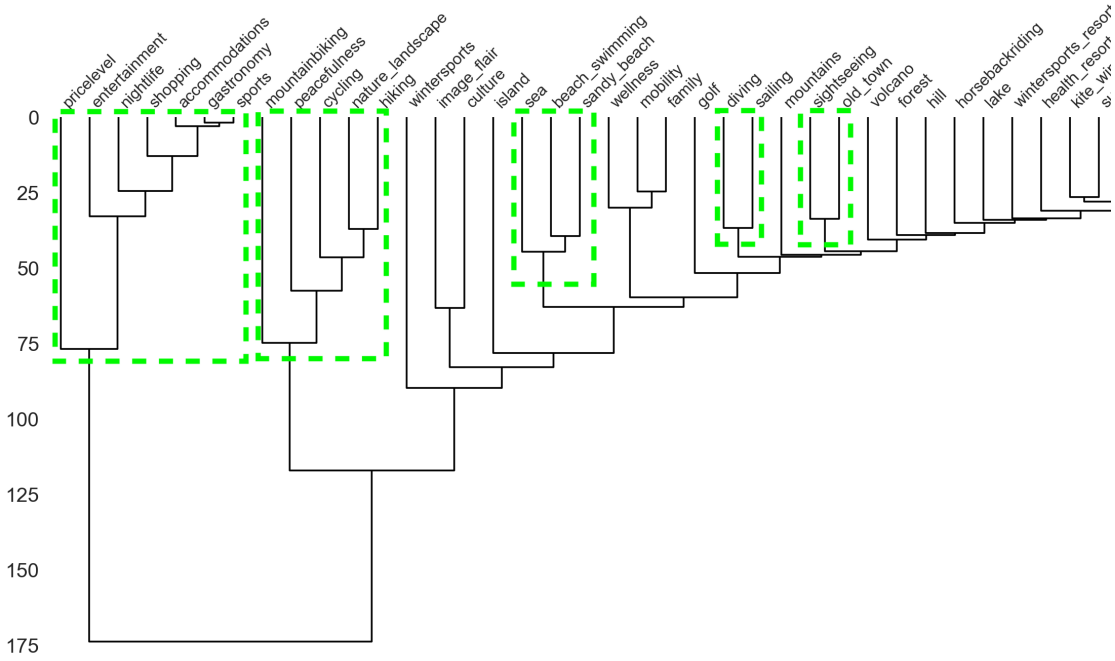


Figure 3.6: Dendrogram of nullity correlation of destination features. Due to space constraints the fully-grown tree is not displayed here, but note that on the right hand side destination features with high missingness (>99%) are grouped.

The Python package *missingno* [Bil16] provides an easy way to generate such dendrogram and the authors are suggesting to interpret it as following: *"To interpret this graph, read it from a top-down perspective. Cluster leaves which linked together at a distance of zero fully predict one another's presence—one variable might always be empty when another is filled, or they might always both be filled or both empty, and so on. The height of the cluster leaf tells you, in absolute terms, how often the records are "mismatched" or*

incorrectly filed – that is, how many values you would have to fill in or drop, if you are so inclined."

As already expected the most frequent motivational ratings *price level, entertainment, nightlife, shopping, accommodations, gastronomy,* and *sports* can be grouped based on their nullity correlation. Also, nature and recreation related ratings *mountain biking, peacefulness, cycling, nature & landscape,* and *hiking* are related in missingness as assumed. Another interesting and plausible grouping is motivational rating *beach & swimming* with geographical attributes *sea* and *sandy beach.* Finally, the relation of missingness in *diving* and *sailing* or *sightseeing* and *old town* are also reasonable. Due to space constraints the fully-grown tree is not displayed here, but note that on the right hand side destination features with high missingness (>99%) are grouped.

Considering the outcomes of the missing data analysis, it is clear that the data is not MCAR, since there are some dependencies between the missingness of some features. Usually, it is hard to differentiate whether missingness can be categorized to MAR or not. The reason why researchers are interested in differentiating if missing data is MAR or MNAR is that most of the missing data methods (analysis or treatment) are assuming the MAR condition.

In [Gra12] Graham argues that one cannot really differentiate between MAR and MNAR, but he also points out that often this not relevant for further analysis. Taking back into account the Webologen data set, one cannot definitely say that missing data is MNAR. There are some indicators for MNAR like the bias towards good motivational ratings (bad ratings or zeroes are missing) and also looking at the geographical attributes in most cases they are missing, because there is no such attribute of the destination (e.g. there is no sea at Vienna, so the sea attribute is missing).

Most machine learning (ML) methods are affected through missing data, where missing values have to be treated before training the model. However, there are some implementation of machine learning methods, which are treating missing values within their training process, like Chen and Guestrin's XGBoost [CG16] a popular, widely used and scalable machine learning method. Since most ML methods need a complete data set and also in order to conduct bivariate analysis, missing data will be treated in the next section.

3.3 Treating Missing Data

3.3.1 Methods and Concepts

Dealing with missing data is a science on its own. There are several conceptually different approaches in order to handle missing values. A commonly known and popular taxonomy of missing data methods was introduced by Little and Rubin [LR14]. They have grouped methods found in the literature into following four categories, which are not mutually exclusive:

Procedures Based on Completely Recorded Units. This method is based on discarding or ignoring units (rows) and/or features (columns) with missing values. It is the most common and default strategy in many statistical tools. Usually, it leads to reasonable and satisfactory results, if there is a low level of missingness. Nevertheless, it may lead to biases, if there are many missing values and / or missingness is considered as MCAR.

Weighting Procedures. *"Randomization inferences from sample survey data without nonresponse commonly weight sampled units by their design weights"* [LR14]. A simplified example for such a procedure is: Considering a population of 50% males and 50% females and a sample of this population with 60% males and 40% females, one can see that females are under-represented in the sample. To adjust such disproportion, one can add weights to each observation, for example 50/60 for males and 50/40 for female participants.

Imputation-Based Procedures. Opposing the already mentioned discarding and ignoring methods, the goal is to retain all observations but still have a complete set in order to do further analysis (e.g. linear regression). In order to do so, missing values are filled in, i.e. imputed. Commonly used imputation methods are *hot deck* imputation, where observations in the sample are used to substitute missing values; *naive* imputation (mean, median etc.), where missing values are substituted with the mean of the observed features for example; and *regression*, where a regression model is build based on the observed features in order to predict the missing values.

Model-Based Procedures. Here, a model is defined for the observed data. Based on that model inferences on likelihood or posterior distributions are made, where parameters are estimated by using maximum likelihood procedures (e.g. variants of Expectations Maximization) for example.

In the upcoming section (Section 3.3.2) a missing data strategy for treating missing values is introduced. It takes into account some of the listed procedures and additionally following imputation methods:

Naive imputation. Usually, naive imputation methods are using the mean or median (depending whether the feature is continuous or categorical) for replacing missing values. In the Webologen case it is known that there is a bias towards good motivational ratings, i.e. most reported ratings have a mean greater than 0.5. A mean imputation would fill in missing values with only good ratings, which is the wrong way to go for the Webologen data. Treating missing motivational ratings as zero is making more sense in this case. However, a naive imputation always leads to a loss in variation of the imputed feature A loss in variation means also a loss in information and predictive power. For example, if all destinations have a value of 80 in motivational rating *family*, one can just say that all destinations in the data set are appropriate for families with children, but one cannot come up with any inference based on good and bad *family* ratings (since it is constant).

KNN imputation. This imputation method can be considered as a variant of hot deck imputation, since observations in the sample are used to substitute the missing values. The K-Nearest-Neighbor algorithm is a proper way to retrieve the closest K neighbors of a tourism destination in the multi-dimensional space, spanned by the destination features. KNN imputation assumes that one can use the values of the K neighbors in order to substitute the empty cells of the considered tourism destination. Before applying KNN imputation the data set is standardized in order to prevent any scaling issues. A K of three or five are the most common values in the literature, where in the case of Webologen K is set to five in order to consider more neighbors and have a more regularized fitting.

SOFT-IMPUTE. Considering the Webologen data set as a large and sparse matrix of size $m \times n$, where m stands for the number of destinations and n for number of features, one can reformulate the challenge to fill in missing values properly as a matrix completion problem. A very popular example of such a problem is the "Netflix" competition [BK07], where the rows of the matrix correspond to viewers and the columns to movies, with values of each cell being a rating from 1 to 5 a particular viewer assigned to a particular movie. There are about 480K viewers and 18K movies, i.e. $8.6 \cdot 10^9$ potential ratings, but only 1.2% of them are actually observed. Thus, the one-million-dollar worth challenge was to complete this sparse matrix, i.e. to predict what viewers would give to movies they have not rated yet. Mazumder, Hastie, and Tibishrani [MHT10] introduced scalable solution for large-scale matrix completion problems, which they named *SOFT-IMPUTE*. The *SOFT-IMPUTE* method shows good training and test error performance while outperforming other well-known, state of the art techniques in timing performance.

3.3.2 Treatment and Evaluation

Considering the taxonomy and the different techniques, introduced in Section 3.3.1, a missing data strategy for the Webologen data is derived and depicted in Figure 3.7.

Initially, features that have in general the same meaning are merged. Such that, *pebble beach, sand & pebble beach, flat decaying sandy beach, beach promenade,* and *sandy beach* are merged to the feature *beach*. Since these features are just binary values the merging is done via a simple OR function. Also, features *winter sports* and *winter sports resorts* are merged to just *winter sports*. Here, merging is done by replacing the binary one of the geographical attribute *winter sports resorts* with the average *winter sports* rating of destinations located in winter sports resorts (i.e., mean(winter sports) if winter sports resort == True). Next, all features, which have been reported as recently introduced and experimental by Webologen, are omitted. Then, destinations, which do not possess a certain amount of features (i.e., a certain threshold), are discarded. Since there are seven destination features (*price level, gastronomy, sports, accommodations, shopping, nightlife,* and *entertainment*) which are mostly non-missing (i.e. in about 90% of the cases), the threshold is set to 10 in order to get at least three more aspects of a tourism destination. These initial steps can be considered as *Procedures Based on Completely*

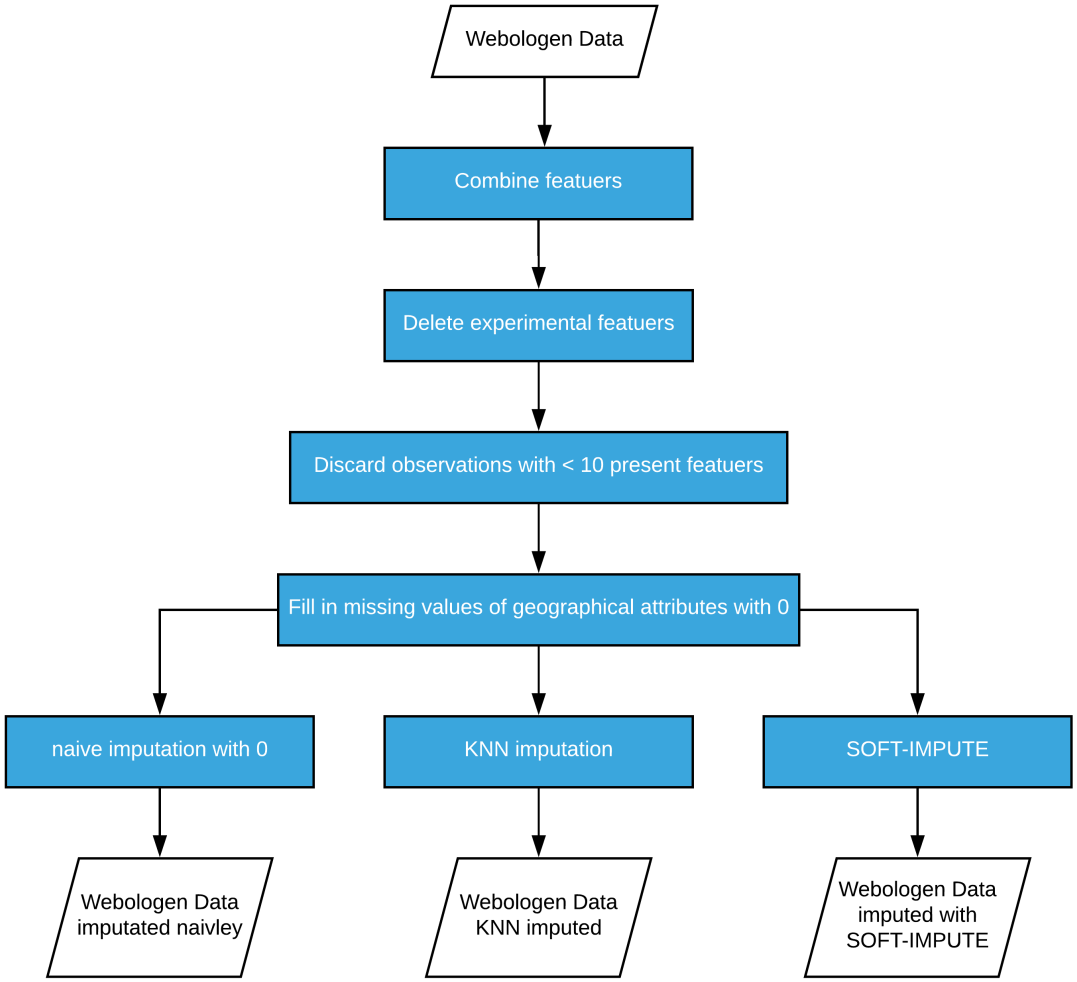


Figure 3.7: Missing Data Strategy for Webologen data set.

Recorded Units. Discarding rows with high missingness leads to a smaller data set of 16950 destinations (44% of the provided data set) but is crucial for further analysis based on complete records.

Geographical attributes are defined by the geographical texture of a tourism destination and are actually representing it. In comparison to motivational ratings they are usually more constraint (through the geography) and specific. For example, the geographical attribute *mountains* represents the presence or absence of a mountains in certain destination. The number of mountains are limited and mountains are not everywhere. Whereas, the motivational rating *family* for example, is dependent on many factors and based on them a destination can be more or less appropriate for travelers with children. Taking all this into account missing values of geographical attributes are imputed naively with zero, i.e. missing values in geographical attributes are indicating the absence of such

geographical characteristics of a tourism destination. In a final step, motivational ratings are imputed via three different methods, namely naive imputation, KNN imputation, and SOFT-IMPUTE. This essential strategy leads to a more concise data set with 16950 destinations and 38 attributes (i.e., 26 motivational ratings and 12 geographical attributes).

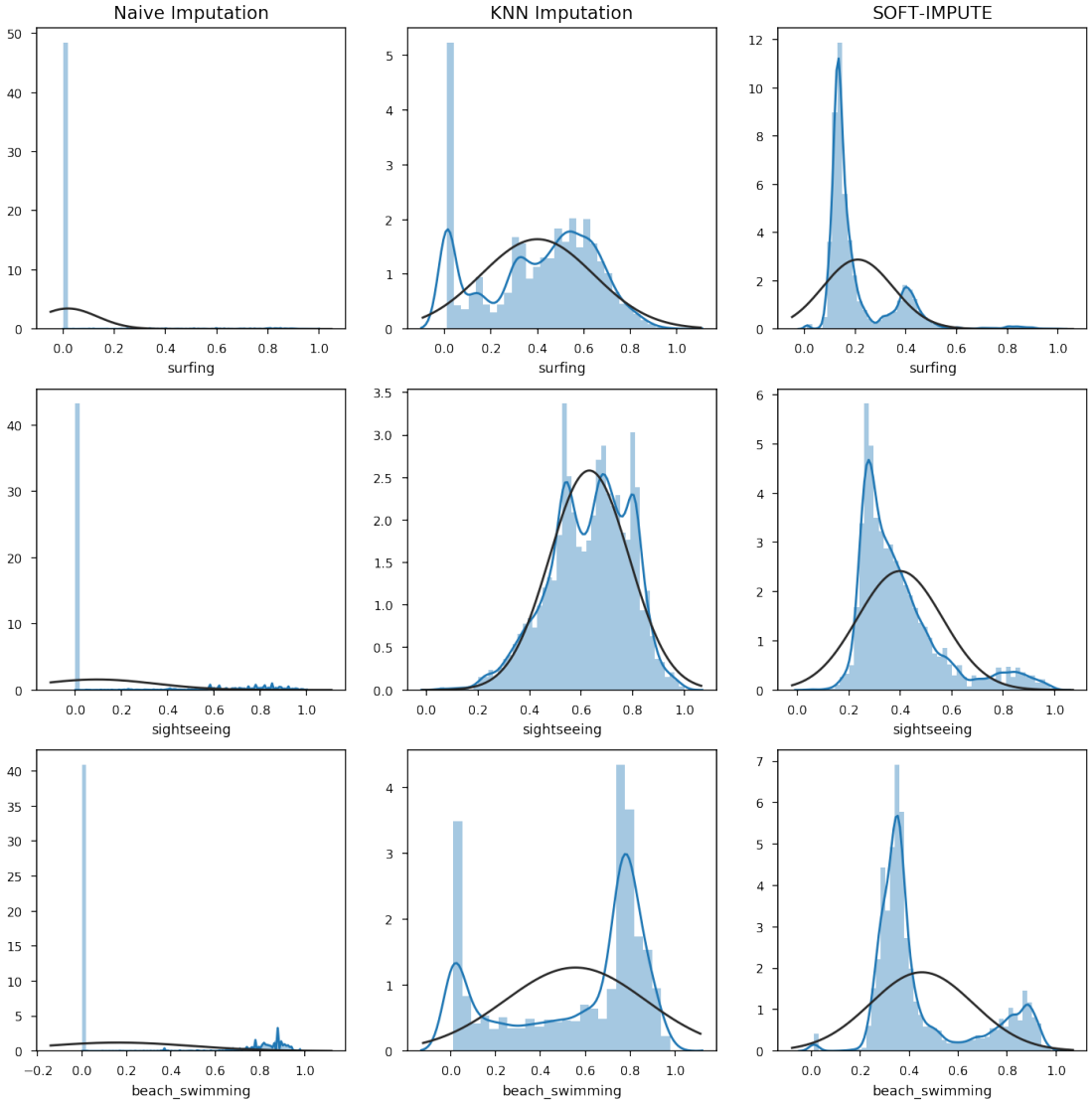


Figure 3.8: Distribution of motivational ratings *surfing, sightseeing,* and *beach & swimming* (low completeness level) in three different imputation strategies. Note, that the black curve shows an estimate for a normal distribution.

In Figure 3.8 probability density functions of motivational ratings *surfing, sightseeing,* and *beach & swimming* in the three different imputations strategies (Naive, KNN, SOFT-

IMPUTE) are depicted. This delivers an overview of how motivational ratings at the lower end of the completeness level are distributed with respect to different imputation strategies. Since there are many missing values, a naive imputation by zero leads to many zero values and thus to a more static representation, but it will also increase the importance (impact) of non-zero values. Interesting to see in the KNN-imputation strategy is that there are no zero-values at all. Simply, because there are no neighbors with zero-values to learn from. Another important observation is, that the KNN-strategy emphasizes the bias towards good ratings. Since, there are no zero-values and the non-zero values have an increased mean (>0.5), learning from nearest neighbors will obviously lead to good ratings. For example, the mean of motivational rating *sightseeing* considering only non-missing values is 0.70 and the mean in the KNN imputed version is 0.63. On the other side, the SOFT-IMPUTE strategy leads to a smoother distribution compared to KNN. Also, it does not emphasize the bias towards good ratings as the KNN strategy. Here motivational rating *sightseeing* scores on average with 0.40.

Figure 3.9 shows the probability density functions of three motivational ratings with midrange completeness level, namely *culture, peacefulness,* and *nature & landscape* (ordered in increasing completeness level), in the proposed imputation strategies. Looking at the naive imputation, this time one can observe not only the peak at zero but also how non-missing (here non-zero) values are actually distributed. Comparing motivational rating *culture* in the KNN imputed version with the SOFT-Imputed version, again one can observe that the KNN-imputation emphasizes the bias toward good ratings, where as SOFT-Impute is not affected. On the other hand, comparing ratings *peacefulness* or *nature & landscape*, where more data is present, in all three imputation strategies (ignoring the zero peak in naive imputation) one can see that the distributions are very similar. This behavior can more clearly be observed in Figure 3.10, where ratings with high completeness level are compared with respect to different imputation strategies.

In Figure 3.10 the probability density functions of motivational ratings *entertainment, gastronomy,* and *price level* with respect to different imputation strategies are illustrated. All three ratings have a high completeness level, i.e. low missingness in data. Since higher completeness means less imputation all three imputation methods show almost the same result. Thus, they just project the real distributions.

Overall, one can say that naive imputation in lower completeness levels leads to a more static rating (loss in variance) and a sparse data set. This can be beneficial for linear models, since they are known to perform well with high dimensional and sparse data sets. On the other hand, more sophisticated imputation methods like KNN and SOFT-IMPUTE are exploiting the given information (present data) and are enriching the variance (information) of the ratings. Whereas, this can be beneficial in tree-based methods, since it can lead to more sensitive and accurate separations. Further, it has been shown that the more data is present the less imputation is needed and the similar the distributions in different imputation methods get. Thus, the performance of imputation methods is more important on the lower end of completeness. KNN imputation is heavily emphasizing the bias towards good ratings and also it can be argued that it loses

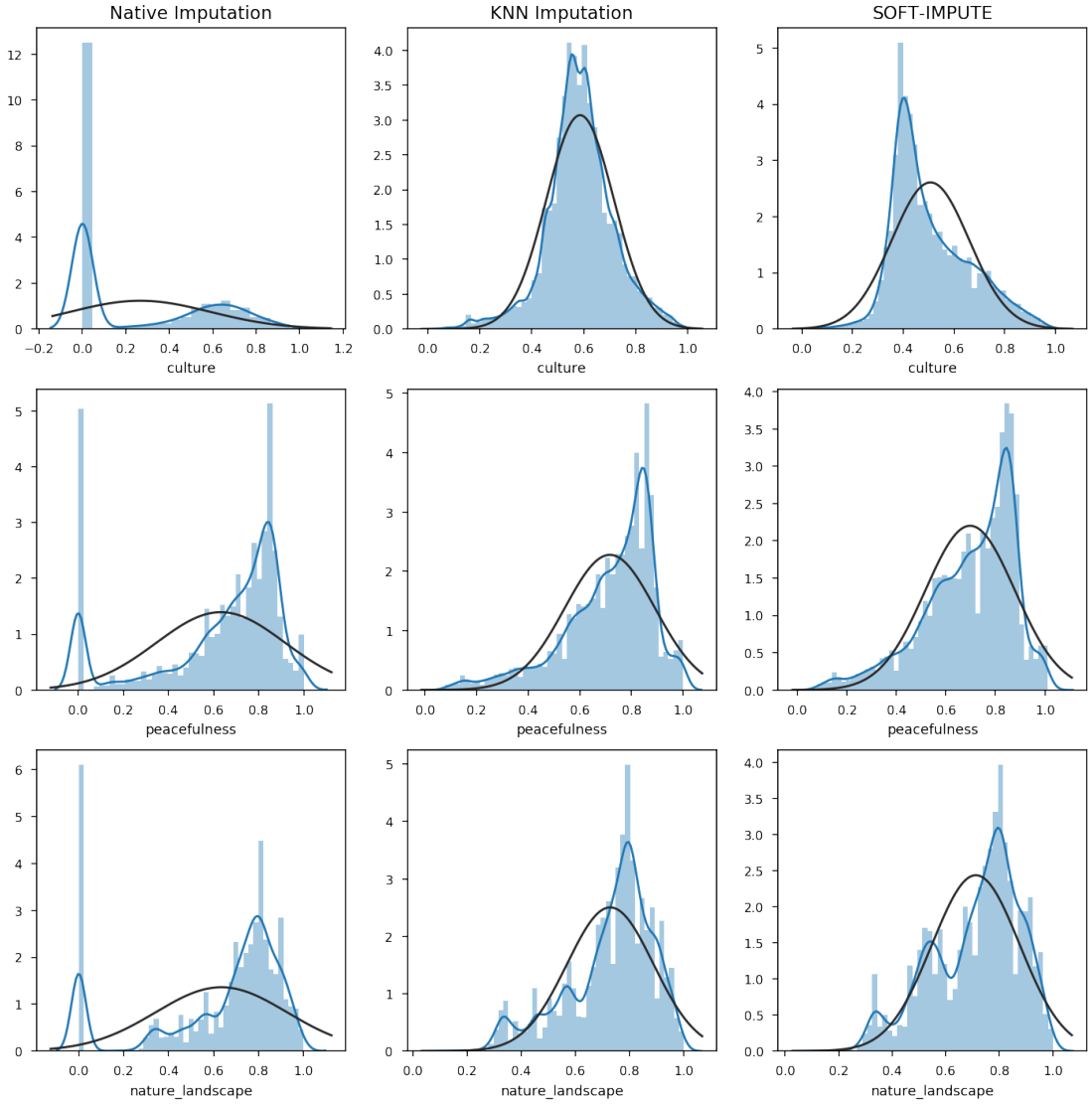


Figure 3.9: Distribution of motivational ratings *culture*, *peacefulness*, and *nature & landscape* (midrange completeness level) in three different imputation strategies. Note, that the black curve shows an estimate for a normal distribution.

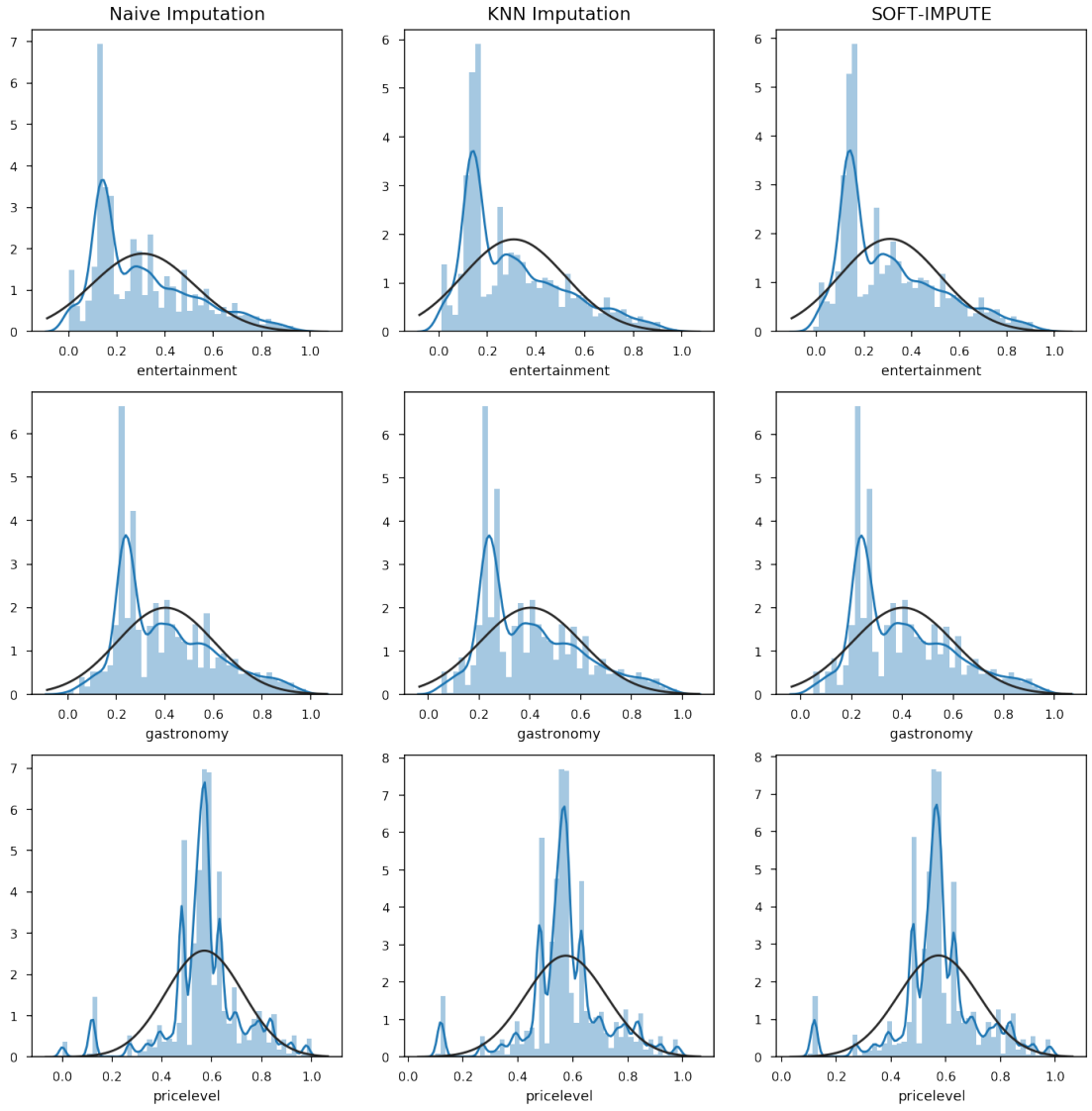


Figure 3.10: Distribution of motivational ratings *entertainment, gastronomy,* and *pricelevel* - (high completeness level) in three different imputation strategies. Note, that the black curve shows an estimate for a normal distribution.

plausibility if the missingnes is too high. For example, looking at the destinations with high *surfing* or *sailing* ratings, one would expect a high probability for the geographical attribute *sea*. But, Table 3.3 shows that this is not the case in the KNN imputed data. Whereas, naive imputation and SOFT-IMPUTE behave as expected. Also noteworthy to mention is that KNN-imputation behaves more or less like a naive means imputation, at least in the analyzed data set and at the lower end of completeness level.

	prob. *sea* - Naive	prob. *sea* - KNN	prob. *sea* - SOFT-IMPUTE
sailing > 0.5	0.83	0.28	0.87
surfing > 0.5	0.98	0.40	0.98

Table 3.3: Probability of geographical attribute sea given motivational ratings sailing and surfing >0.5.

Considering the shown poor performance of KNN imputation, further analysis will only build upon naive imputation and SOFT-IMPUTE.

3.4 Bivariate Analysis of Destination Features

In order to analyse similarities among all destination attributes (i.e., motivational ratings and geographical attributes) a correlation matrix comprising all pairwise Pearson correlation coefficients is calculated. To get a better understanding and overview, the correlation matrix is visualized as a clustered heat map.

Figure 3.11 shows the clustered correlation heat map of the naively imputed data set and following plausible groups are identified (marked by red rectangular):

- Features of the first cluster, namely *peacefulness, nature & landscape, mountain biking, hiking,* and *cycling*, are highly positively correlated and can be interpreted as features of recreational tourism destinations.
- *Mobility, wellness, family, nightlife, entertainment, gastronomy, accommodations,* and *sports* are forming the second cluster. These features are also highly positively correlated and can be interpreted as features of more vibrant destinations compared to the first group. Also, one can clearly differentiate between indicators for family friendliness (*mobility, wellness, family*) and mass tourism (*nightlife, entertainment gastronomy, accommodations,* and *sports*) within this cluster.
- Cluster three groups features related to metropolitan destinations or in other words appropriate destinations for city trips. Members of this group are *shopping, price level, metropolis, image & flair, culture,* and *sight-seeing.* These features are positively correlated. Particularly, *price level* and *shopping* are highly correlated, as expected. Also, *image & flair, culture, sightseeing,* and *old town*, are showing a

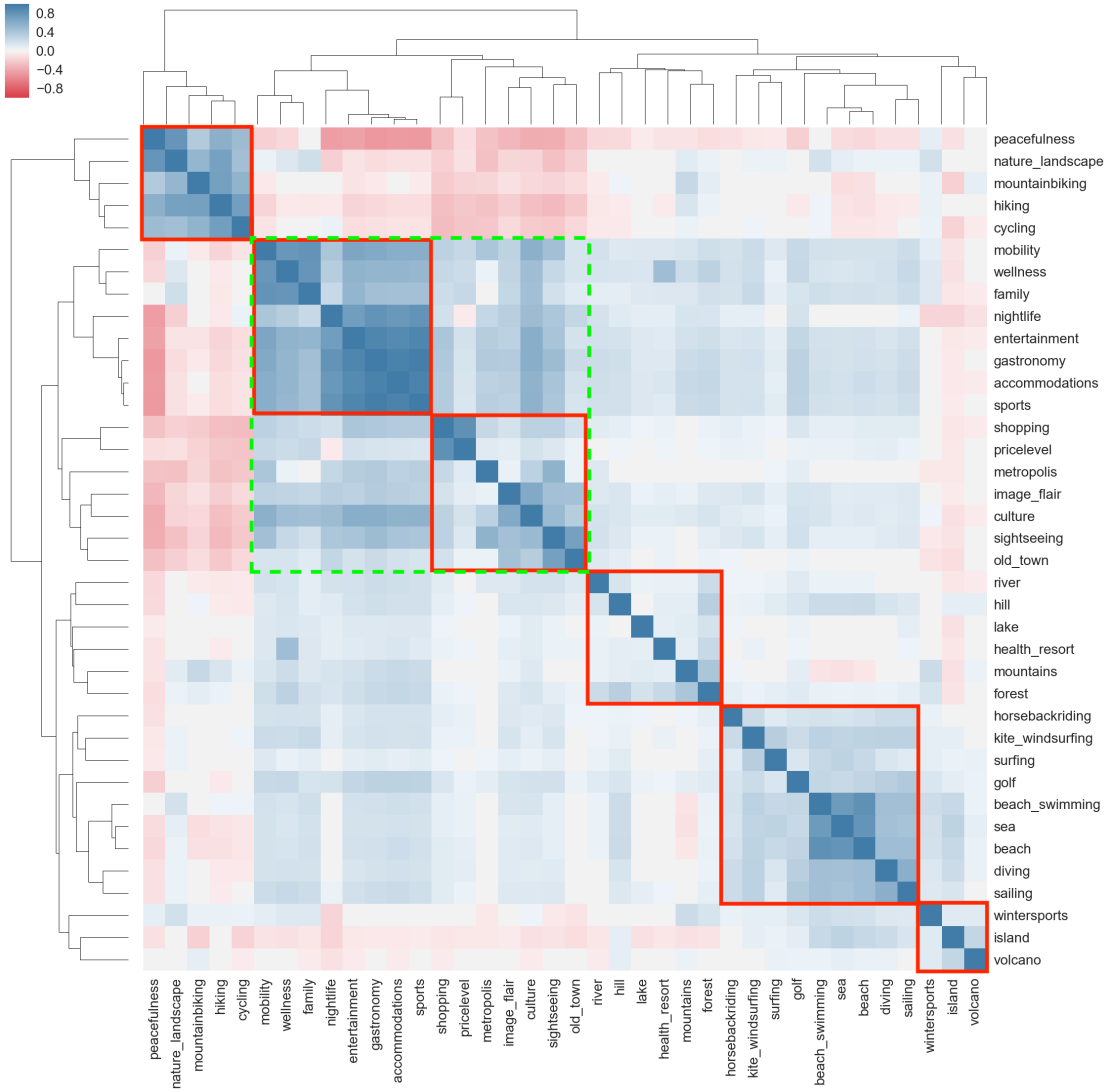


Figure 3.11: Clustered correlation heat map of tourism destination features after naive imputation strategy.

higher correlation coefficient within this cluster. Hence, these are common features of typical mass touristic places.

- In cluster four typical features of destinations at the countryside are represented, namely *river, hill, lake, health resort, mountains,* and *forest.* These features are slightly positively correlated.

- Cluster five consists of *horseback riding, kite- & windsurfing, surfing, golf, beach & swimming, sea, beach, diving,* and *sailing.* As one can see, the majority of features in this cluster are related to water sports and beach vacations.

- Cluster six shows no plausible interpretation and also features within this cluster are just showing a low correlation. Members of this cluster are *winter sports, island,* and *volcano.*

Overall, one can clearly observe a contrast between attributes related to mass tourism (green rectangular) and attributes related to recreational destinations (first group), especially in the case of motivational rating *peacefulness.*

Figure 3.12 shows the clustered correlation heat map of the resulting data set of the SOFT-IMPUTE strategy. Comparing both heat maps (Figure 3.11 with Figure 3.12) one can clearly see that correlation coefficients in the last one are relatively higher. Yet, the resulting groups of features are quite similar (but differently ordered):

- Cluster one is grouping features with affiliation to recreational traveling. Peacefulness, *nature & landscape, mountain biking, hiking, cycling* and *winter sports* are forming this cluster. The feature winter sports is the only difference to cluster one of the previous grouping, but it still makes sense, since destinations appropriate for winter sports are mostly in the nature and tend to show recreational features (except après ski).

- Cluster two consist of *island and volcano* and is corresponding to the last group in the previous clustering. This constellation without *winter sports* makes more sense than the previous one.

- In Cluster three there are typical features of destination at the countryside, namely *river, hill, mountains, forest, lake* and *health resort.* It is the equivalent of cluster four in the previous clustering. Again, there is a slight positive correlation among the features.

- Cluster four is the equivalent of cluster three in the previous clustering and is grouping features related to city trips and metropolitan areas. These features are *shopping, price level, metropolis, and old town.* Whereas, *culture, sightseeing, and image & flair* are not located in this cluster in comparison to the previous clustering.

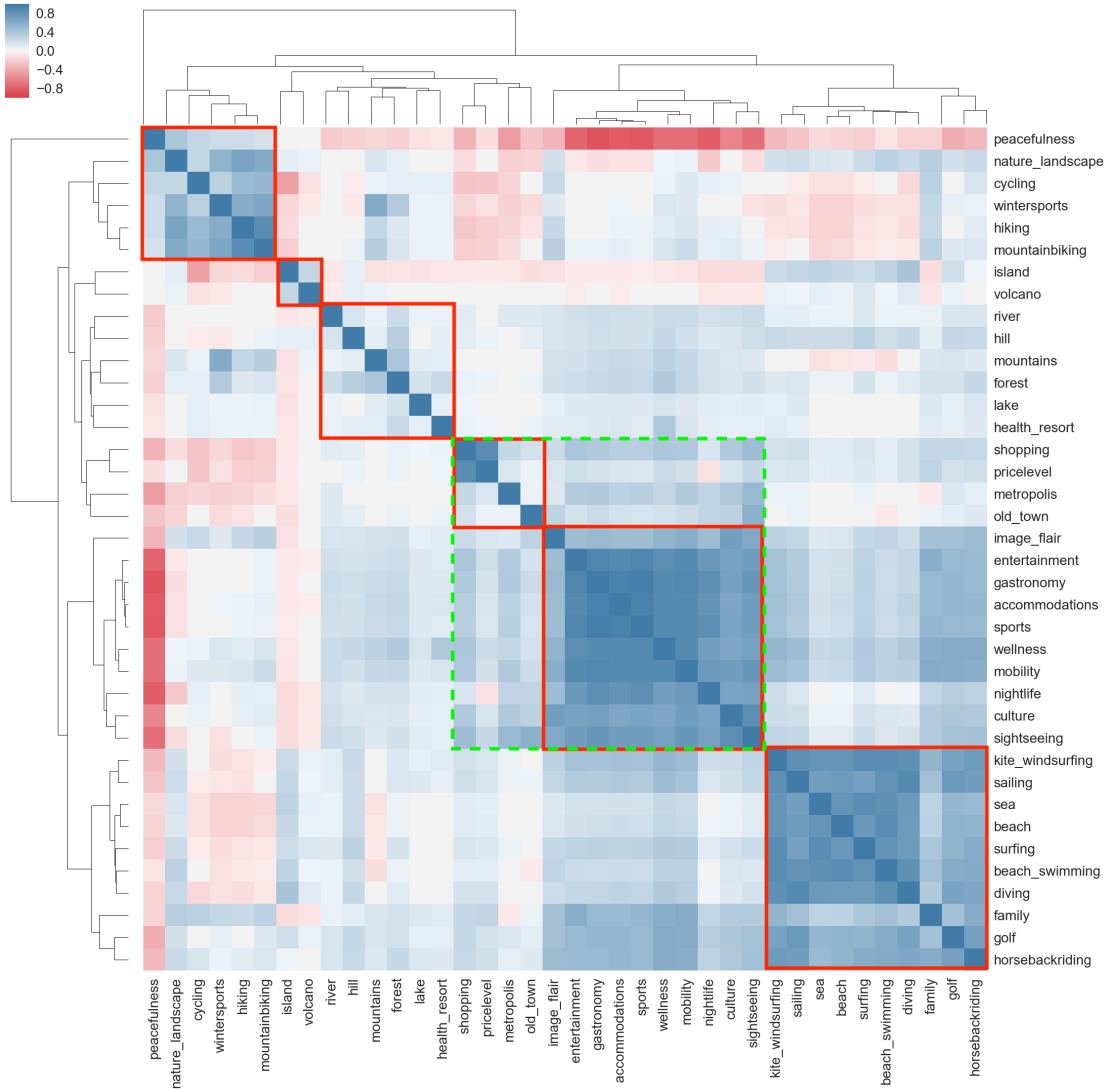


Figure 3.12: Clustered correlation heat map of tourism destination features after SOFT-IMPUTE strategy.

- Cluster five groups features related to mass tourism, such as *image & flair, entertainment, gastronomy, accommodations, sports, wellness, mobility, nightlife, culture,* and *sightseeing*. This cluster is corresponding to cluster two of the previous cluster heat map and is differentiating in features *culture, sightseeing,* and *image & flair*, which can also be seen as features of mass touristic places. Features within this cluster are highly positively correlated.

- Cluster six is the equivalent of cluster five in the naive imputation data set. Here are features mostly related to water sports and beach vacation grouped. Also, one can see that the correlation coefficients are relatively higher compared to the previous results. Features within this cluster are *kite & windsurfing, sailing, sea, beach, surfing, beach & swimming, diving, family, golf, and horseback riding*. The last three features are a bit detached from all other features within this cluster, which can also be seen in the decreased correlation coefficients.

Overall, the contrast between features related to mass tourism and features related to recreational destinations can also be observed here.

3.5 The Data Sample

In addition to the SQL-dump experts of Pixtri [OG], an Austrian e-Tourism company, provided a labeled sample of 561 destinations. In other words, of all destinations, 561 destinations were chosen randomly and mapped manually to the Seven-Factors by experts. These experts were members of an Austrian e-Tourism company using an implementation of the picture based approach [NSSW14, NSSW15]. Thus, they were familiar with both characteristics of tourism destinations and the Seven-Factor Model. For the 561 destinations, three experts assigned first individually a score for each factor using the scale 0 - 0.25 - 0.50 - 0.75 - 1. The higher the score the more suitable, in the expert's opinion, the destination for that specific factor. After the individual mappings, a final mapping was determined in a joint discussion.

Figure 3.13 illustrates the distribution of destinations in a world map. The majority of destinations are located in Germany, USA, France, Greece, Great Britain, Italy, Denmark, Spain, Austria, and Netherlands (62%), which is similar to the distribution in the whole data set.

In Figure 3.14 average motivational ratings in the labeled sample are listed. Again, one can see the same bias toward good ratings in the sample like in the whole data set, i.e. most ratings have a mean higher than 0.5.

Figure 3.15 shows the amount of missing values for each destination feature in the expert sample. Also, with respect to of missingness of data the expert sample shows a similar behavior as the data set. Taking all these into account the labeled sample can be considered as a representative sample of the whole data set.

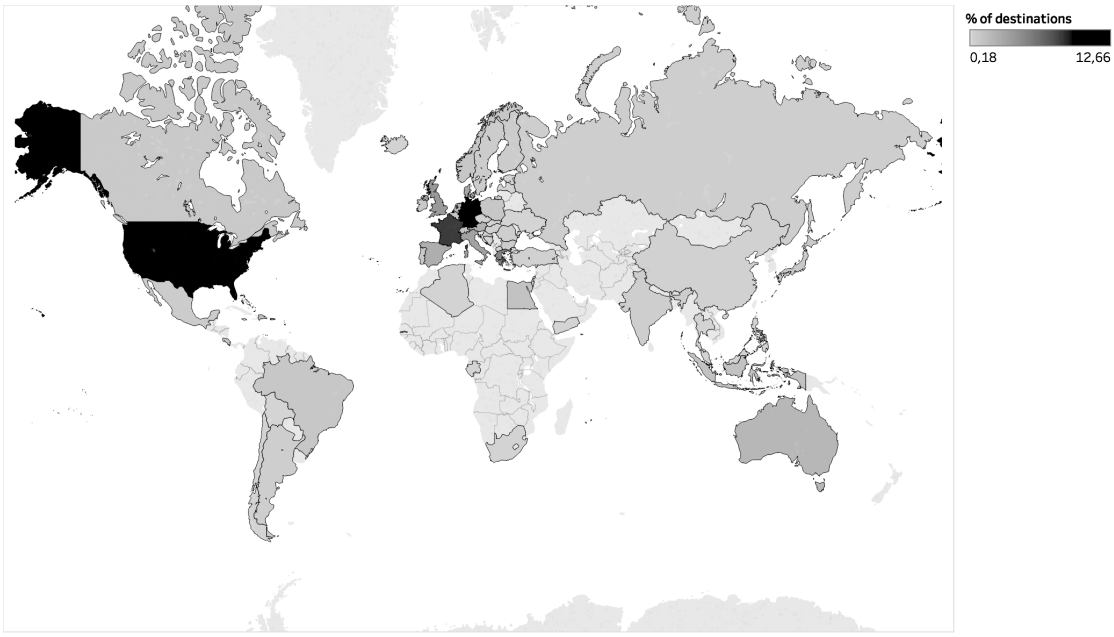


Figure 3.13: Distribution of tourism destinations over countries in the expert data set.

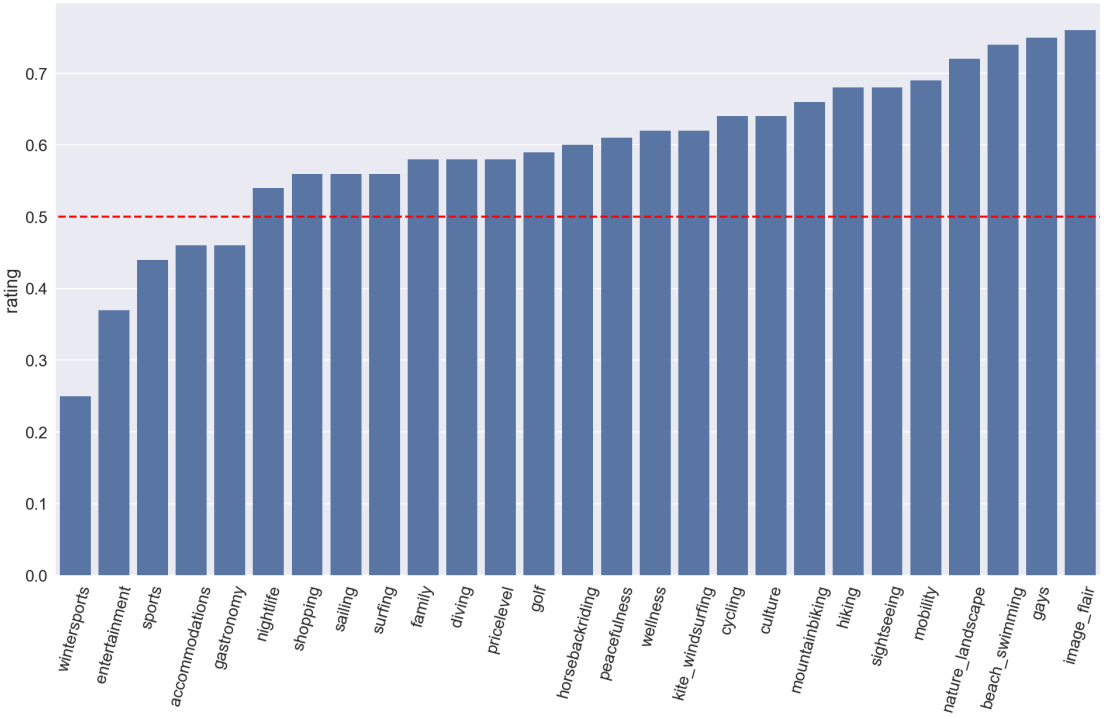


Figure 3.14: Average motivational ratings of tourism destinations in the expert sample.

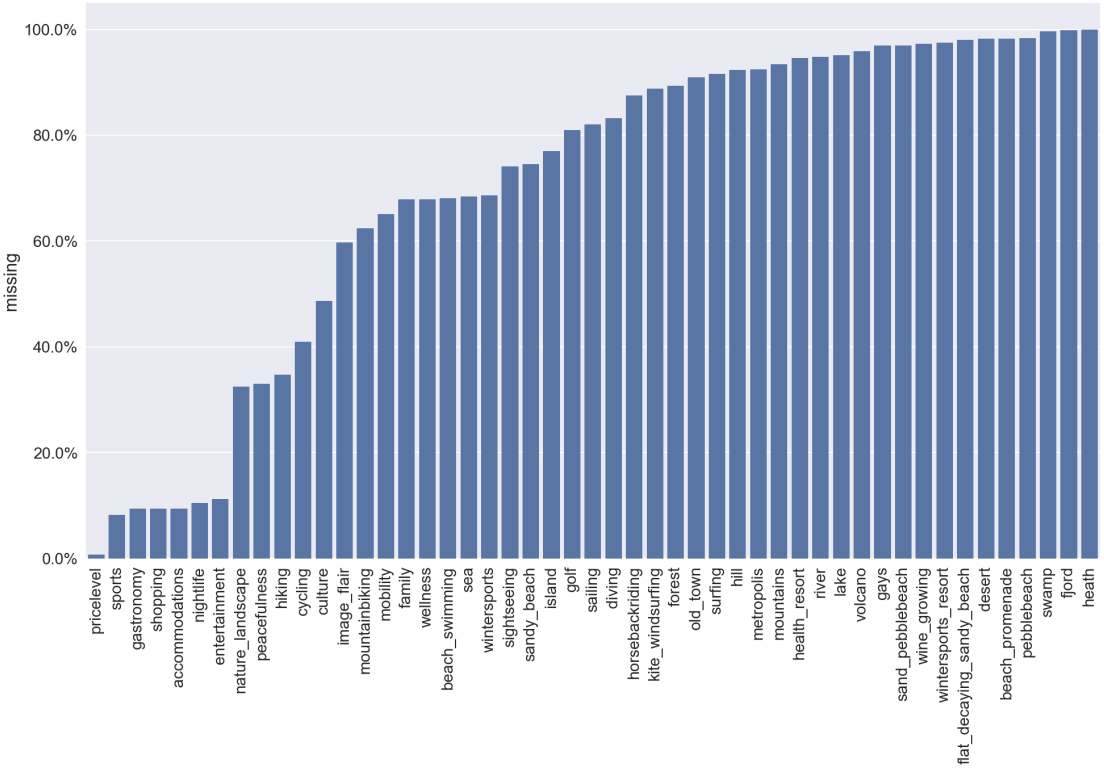


Figure 3.15: Amount of missingness in tourism destination features in the expert sample.

Further analysis is based on complete data, i.e. data with only non-missing values. Due to a shrinked data set, caused by an essential missing value treatment, also the expert sample got smaller in size, namely from N=561 to N=350 (62%).

In Figure 3.16 factor score distributions the Seven-Factors in the expert sample are illustrated. For example, in case of the factor *Sun & Chill-Out* 30% of the destinations scored with 0, 17.1% with 0.25, 15.4% with 0.5, 10.6% with 0.75, and 26.9% with 1. The majority of destinations (56.9%) scored with 0 or 1 in factor *Sun & Chill-Out*. Whereas, the majority of destinations (55.7%) in case of *Knowledge & Travel* scored with either 0 or 0.25 and a few with 1 (10.6%), similar to the distribution in factor *Action & Fun*. Almost half of the destinations have a score in the "lower middles", 0.5 (28.3%) and 0.25 (21.4%), in factor *Culture & Indulgence*. This is similar to the distribution in factor *Nature & Recreation*, where the majority of destinations have scores in the "upper middles" 0.5 (24%) and 0.75 (27.4%). An extreme case of this "upper middles" can be seen in factor *Social & Sports*, where almost all destinations (87.1%) scored with either 0.5 (53.4%) or with 0.75 (33.7%). The only factor where an approximately normal distribution (bell shape) of scores can be observed is *Independence & History*.

Furthermore, the correlation between the Seven-Factors and the features of tourism destinations are examined. In particular, correlation coefficients of each factor of the

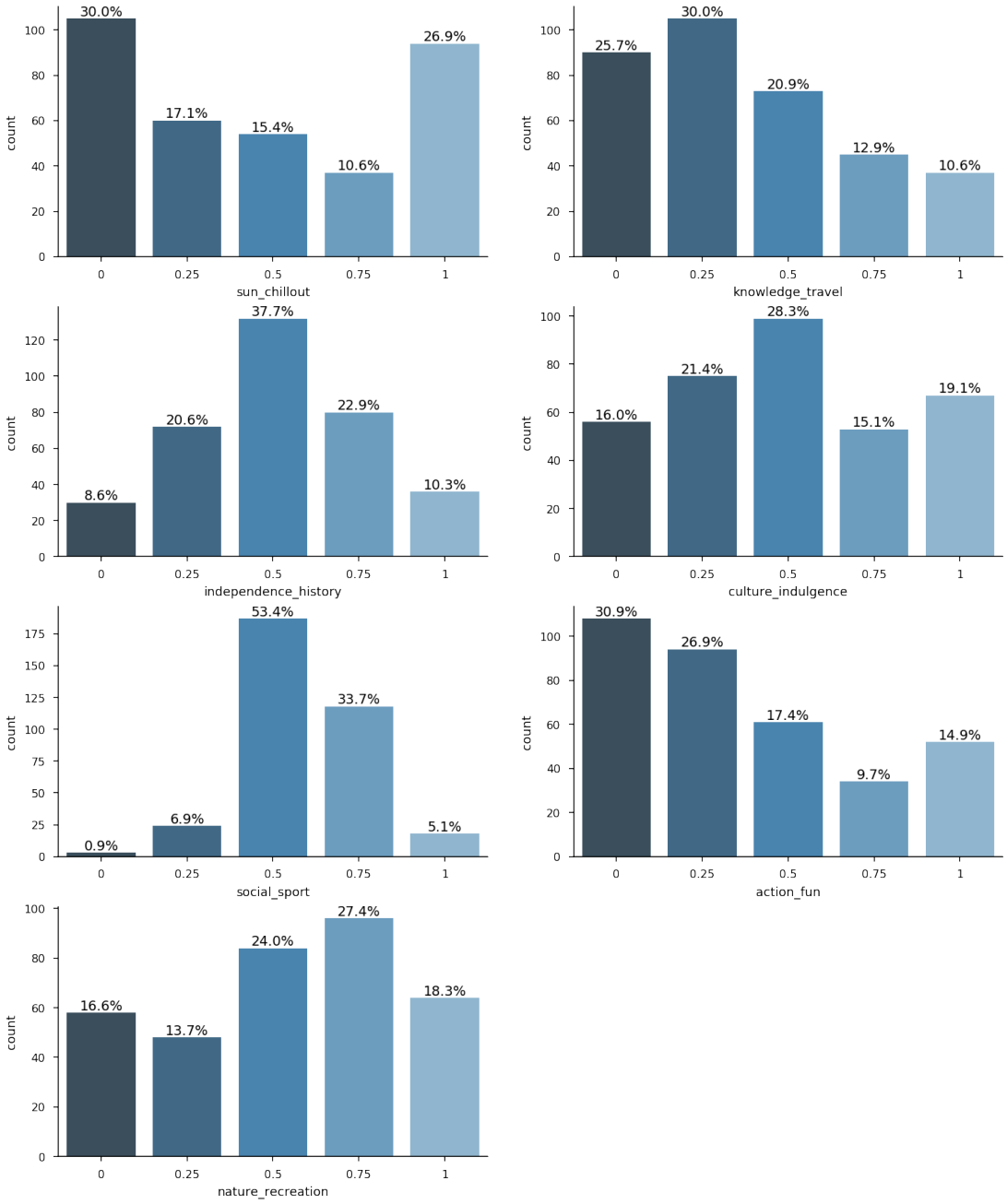
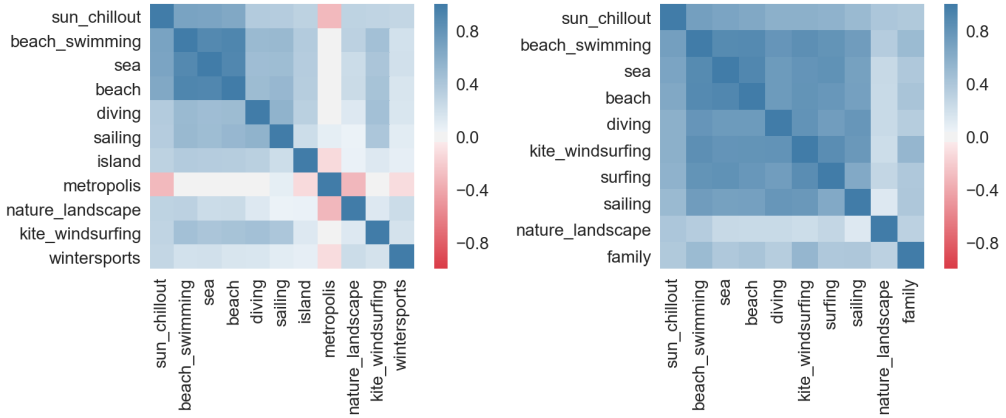


Figure 3.16: Distribution of Seven-Factor scores in the labeled data set of tourism destinations.

Seven-Factor Model and its most correlated destination features are calculated and depicted as heat maps. Note, that correlations are calculated for both, data treated with naive imputation and SOFT-IMPUTE.



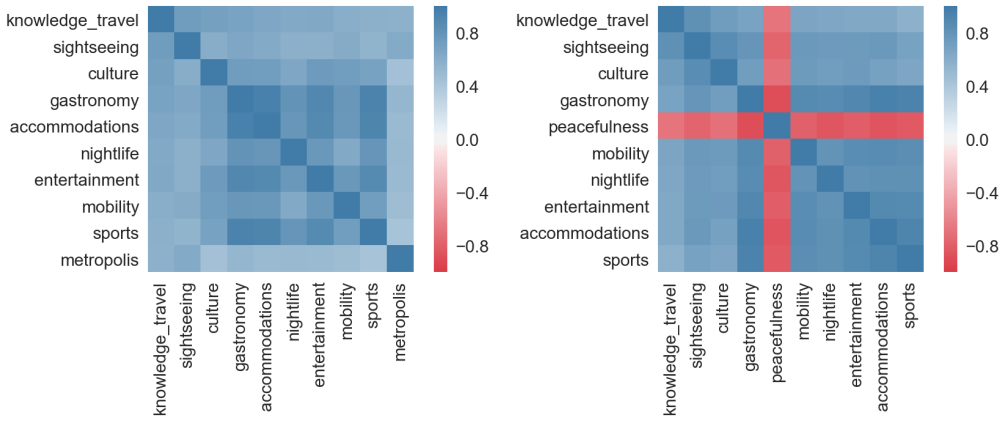
(a) Naive (b) SOFT-IMPUTE

Figure 3.17: Heat map of most correlated destination features of the factor *Sun & Chill-Out* in different imputation strategies. Note, that the first element of the heat map is the factor itself followed by its most correlated features (ordered by the absolute value of the correlation coefficients).

In Figure 3.17 the most correlated features of the factor *Sun & Chill-Out* are illustrated. In both version (3.17a and 3.17b) *beach & swimming, sea* and *beach* are highly correlated with the factor. Also in both, a correlation between the factor and certain water sports, such as *diving, kite- & windsurfing, surfing,* and *sailing*, is observable. Furthermore, the naively imputed sample shows a negative correlation of the geographic attribute *metropolis* with the factor, which can be explained by its contradiction to the chill-out aspect of the factor.

Correlations of the factor *Knowledge & Travel* are shown in Figure 3.18. *Sightseeing* and *culture* are in both, naive (3.18a) and SOFT-IMPUTE (3.18b) versions, the most correlated destination features. These are crucial destination features for an organized mass tourist, who wants to gain knowledge during a trip. Also in both, indicators of mass tourism and urbanization like *gastronomy, accommodations, nightlife, entertainment* etc. are positively correlated with the factor. Note, that only in the SOFT-IMPUTE version *peacefulness* is highly negatively correlated to the factor. Since the motivational rating *peacefulness* is contradiction to mass tourism and urbanization such negative correlation is plausible.

Correlations of destination features with the factors *Independence & History* (Figure 3.19) and *Culture & Indulgence* (Figure 3.20) are similar to the correlations of features with the factor *Knowledge & Travel* (Figure 3.18). Since the factor *Independence & History* is considered as independent, history loving mass tourist and *Culture & Indulgence* as

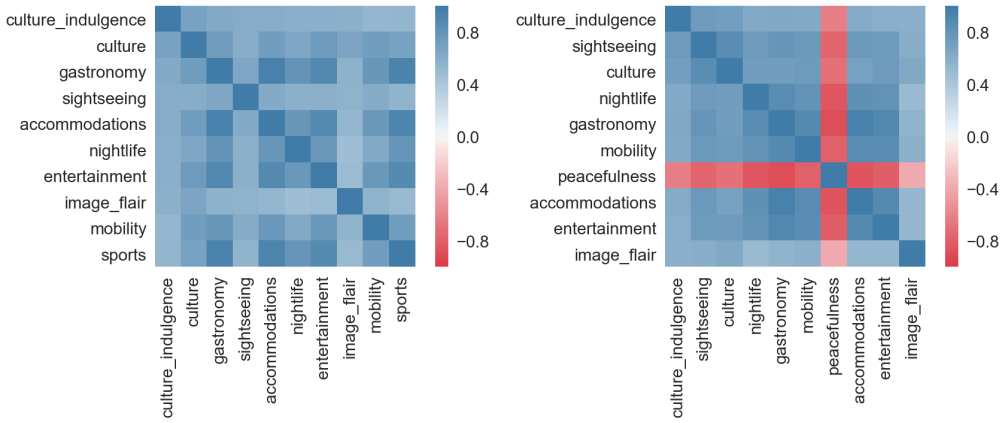


(a) Naive (b) SOFT-IMPUTE

Figure 3.18: Heat map of most correlated destination features of the factor *Knowledge & Travel* in different imputation strategies. Note, that the first element of the heat map is the factor itself, followed by its most correlated features (ordered by the absolute value of the correlation coefficients).

(a) Naive (b) SOFT-IMPUTE

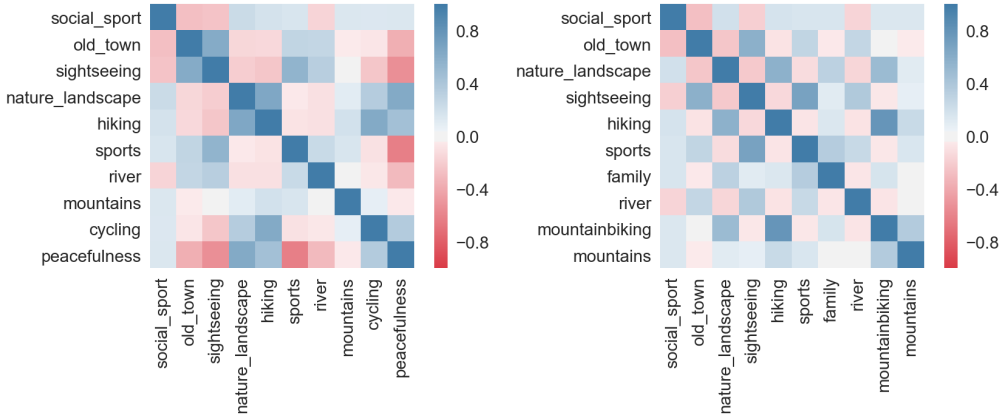
Figure 3.19: Heat map of most correlated destination features of the factor *Independence & History* in different imputation strategies. Note, that the first element of the heat map is the factor itself, followed by its most correlated features (ordered by the absolute value of the correlation coefficients).



(a) Naive (b) SOFT-IMPUTE

Figure 3.20: Heat map of most correlated destination features of the factor *Culture & Indulgence* in different imputation strategies. Note, that the first element of the heat map is the factor itself, followed by its most correlated features (ordered by the absolute value of the correlation coefficients).

culture and history loving gourmet, high positive correlations with *culture, sightseeing, gastronomy* or other mass tourism and urbanization related destination features and a negative correlation with *peacefulness* are plausible.

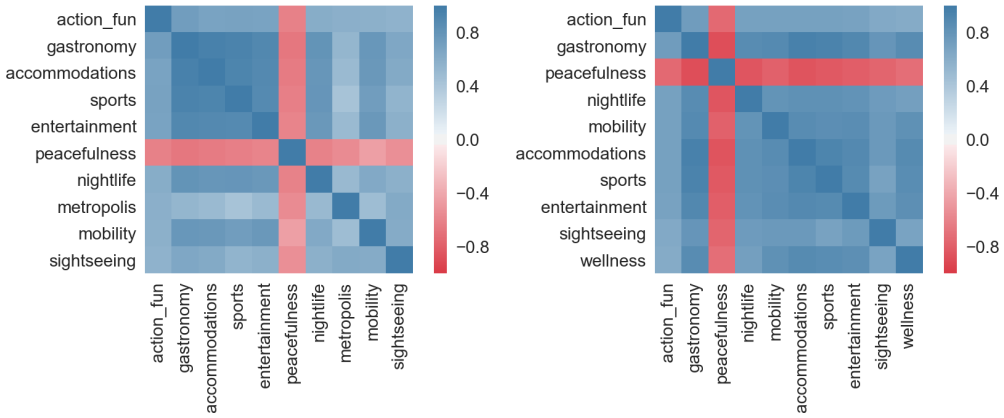


(a) Naive (b) SOFT-IMPUTE

Figure 3.21: Heat map of most correlated destination features of the factor *Social & Sports* in different imputation strategies. Note, that the first element of the heat map is the factor itself, followed by its most correlated features (ordered by the absolute value of the correlation coefficients).

It is known that crowded and touristic places are negatively related to the factor *Social & Sports*. This can also be observed in Figure 3.21, where the factor *Social & Sports* is

negatively correlated with the destination features *sightseeing* and *old town*. The factor is positively correlated with the features *sports, hiking, cycling, mountain biking* and *peacefulness* as expected.



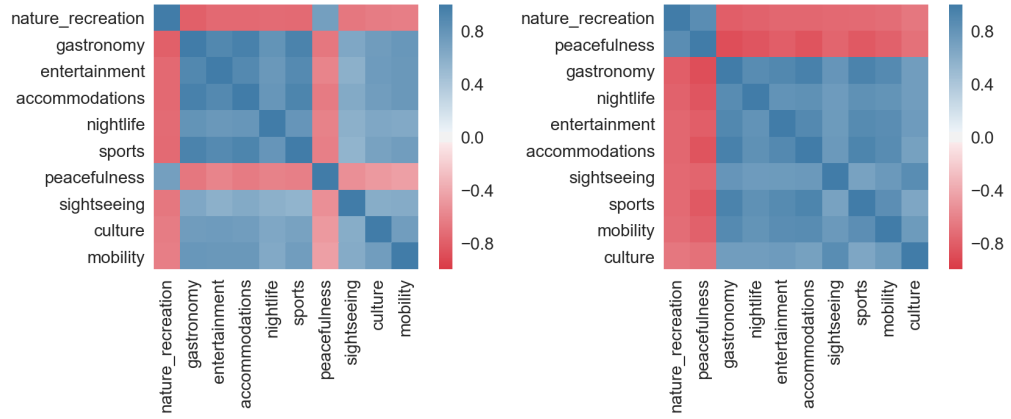
(a) Naive (b) SOFT-IMPUTE

Figure 3.22: Heat map of most correlated destination features of the factor *Action & Fun* in different imputation strategies. Note, that the first element of the heat map is the factor itself, followed by its most correlated features (ordered by the absolute value of the correlation coefficients).

In Figure 3.22 correlations of the factor *Action & Fun* with destination features are depicted. One can immediately see that feature *peacefulness* is negatively correlated with the factor, especially in the SOFT-IMPUTE version. The negative correlation with *peacefulness* is self explaining. Further, the factor *Action & Fun* is positively correlated with indicators of "action" and "fun" like *sports, entertainment, nightlife* and other indicators of vibrant places such as *metropolis, sightseeing, culture, gastronomy, mobility,* and *accommodations*.

Figure 3.23 shows the most correlated destination features of the factor *Nature & Recreation*. Obviously, the motivational rating *peacefulness* is positively correlated with this factor, which is more clear and intense in the SOFT-IMPUTE version. All other listed features are negatively correlated with the factor. These features can be considered as attributes of mass touristic, crowded or highly urbanized places.

Considering the previous bivariate analysis of the whole data set or the correlation analysis of the Seven-Factors here, both SOFT-IMPUTE and naive imputation are leading to overall similar results. More precisely, SOFT-IMPUTE is in some cases better. For example, it emphasizes the contrast between mass touristic places and peaceful places more. Also, it leads to a better ranking in a factor's top correlated features, e.g. the most correlated destination feature of the factor *Nature & Recreation* is *peacefulness*. All in all, SOFT-IMPUTE uses available information in order to intelligently replace missing values, which leads to an overall better result than a simple naive imputation. Hence,



(a) Naive (b) SOFT-IMPUTE

Figure 3.23: Heat map of most correlated destination features of the factor *Nature & Recreation* in different imputation strategies. Note, that the first element of the heat map is the factor itself, followed by its most correlated features (ordered by the absolute value of the correlation coefficients).

further analysis and the model building will only consider the resulting data set of the SOFT-IMPUTE strategy.

Table 7.1 summarizes the most correlated destination features of the Seven-Factors in the SOFT-IMPUTE strategy. Note, that a preceding minus sign indicates a negative correlation. Overall, one can say that the most correlated destination features of the Seven-Factors are reasonable and quite clear. Especially, the factors *Sun & Chill-Out*, *Culture & Indulgence*, and *Action & Fun* are well covered. Only for the factor *Social & Sports* are the correlation coefficients relatively low in comparison with the other factors and furthermore one would expect a higher positive correlation with motivational rating *sports* than observed.

Factor	Most correlated destination features
Sun & Chill-Out	*beach & swimming, sea, beach, diving, kite & windsurfing, surfing, sailing, nature & landscape, family*
Knowledge & Travel	*sightseeing, culture, gastronomy, - peacefulness, mobility, nightlife, entertainment, accommodations, sports*
Independence & History	*sightseeing, culture, mobility, gastronomy, - peacefulness, accommodations, nightlife, entertainment, sports*
Culture & Indulgence	*sightseeing, culture, nightlife, gastronomy, mobility, - peacefulness, accommodations, entertainment, image & flair*
Social & Sports	*- old town, nature & landscape, - sightseeing, hiking, sports, family, - river, mountain biking, mountains*
Action & Fun	*gastronomy, - peacefulness, nightlife, mobility, accommodations, sports, entertainment, sightseeing, wellness*
Nature & Recreation	*peacefulness, - gastronomy, - nightlife, - entertainment, - accommodations, - sightseeing, - sports, - mobility, - culture*

Table 3.4: Most correlated destination features of the Seven-Factors. Note, that a preceding minus sign indicates a negative correlation.

CHAPTER 4

Mapping of Tourism Destinations to the Seven-Factors

In this chapter supervised and unsupervised learning techniques are used in order to understand similarities among destinations, enable an automated mapping onto the Seven-Factors, identify important features, and explain their relationship with the Seven-Factors.

4.1 Cluster Analysis

Identifying conceptually meaningful groups of destinations with shared common characteristics will help to further understand the data and its structure, which may contribute to a more generalized solution. Furthermore, the identified clusters might be addressed by RSs directly. The cluster analysis comprises 16950 destinations (i.e., the data set after pre-processing). Partitional clustering techniques are considered, where most prominent ones are K-means and K-medoids. Since the data comprises binary attributes, using the Euclidean distance and thus centroids (both are essentials of the K-means algorithm) are not meaningful. Therefore, K-medoids is applied. A medoid corresponds per definition to an actual data point, which is considered as the most representative point for the cluster [TSK+06]. Specifically, Partitioning Around Medoids (PAM) [KR90], the most common K-medoids algorithm, is used. Since the data consists of two different data types, i.e., binary (geographical attributes) and continuous (motivational ratings), the Gower distance (appropriate for mixed datatypes) [Gow71] is used as distance metric. In order to find an appropriate number of clusters, the internal evaluation metric silhouette width (i.e., silhouette coefficient) [Rou87] is used for assessment. In the following the used methods and measures are introduced briefly:

PAM. The main objective, when partition objects to clusters is, to separate them in a way that objects grouped in a cluster should be similar as possible, while being as dissimilar as possible to objects of other clusters. The PAM algorithm is separated into two phases, namely build and swap. In the build phase, K most representative objects among all object of the given data set are searched. These objects should represent various characteristics and the structure of the given data and are called medoids. Finally, in the second phase, the swap phase, K clusters are constructed, simply by assigning objects of the data set to the nearest medoid.

Gower Distance. The Gower (dis)similarity metric can be used in case of mixed data types and the approach it follows is rather simple. For each data type in the given data set the most appropriate distance metric is used and scaled to an interval of [0,1]. Finally, dissimilarities among units of the data set are obtained as a weighted sum of dissimilarities of each variable.

Silhouette width. In [Rou87] Rousseeuw introduced silhouettes as *"a graphical aid to the interpretation and validation of cluster analysis"* The silhouette width (= silhouette coefficient), is a measure of cohesion and separation in a cluster. It shows how similar an object is to objects of the same cluster in comparison to objects of other clusters. The silhouette coefficient ranges from -1 to 1, where high values are indicating a good fit of an object to its own cluster and a bad match to other clusters. Low or negative values are indicating that objects are either located in-between clusters or are wrongly assigned. Furthermore, Rousseeuw argues that *"the average silhouette width provides an evaluation of clustering validity, and might be used to select an 'appropriate' number of clusters"* [Rou87]. In [Rou87] Rousseeuw also suggest to use silhouette plots to asses the relative quality of a clustering and describes them as following: *"The entire clustering is displayed by combining the silhouettes into a single plot, allowing an appreciation of the relative quality of the clusters and an overview of the data configuration.*

Figure 4.1 shows the average silhouette width within different cluster sizes. Based on the average silhouette width two, three, four and six cluster solutions are considered, but for the sake of interpretability a six-cluster solution is chosen. Next, the resulting clusters are examined in detail. The number of destinations in each cluster is provided at the beginning of each paragraph.

C1 (N = 1940). The medoid of C1 is Paralia, a small city in Greece. Paralia means in Greek beach and as the name already suggests, the city is located directly on the beach. It is a popular and vibrant seaside resort with many nightlife and shopping opportunities. Interestingly, 93% of the destination in C1 are located at the sea and 92% directly at the beach, whereas globally only about 20% of destination are located at the sea or beach. Also, the rating *beach & swimming* has a high mean value of 0.81. Additionally, ratings *gastronomy, nightlife, sports, accommodations,* and *culture* are showing an increased average value (0.62 - 0.66). To conclude,

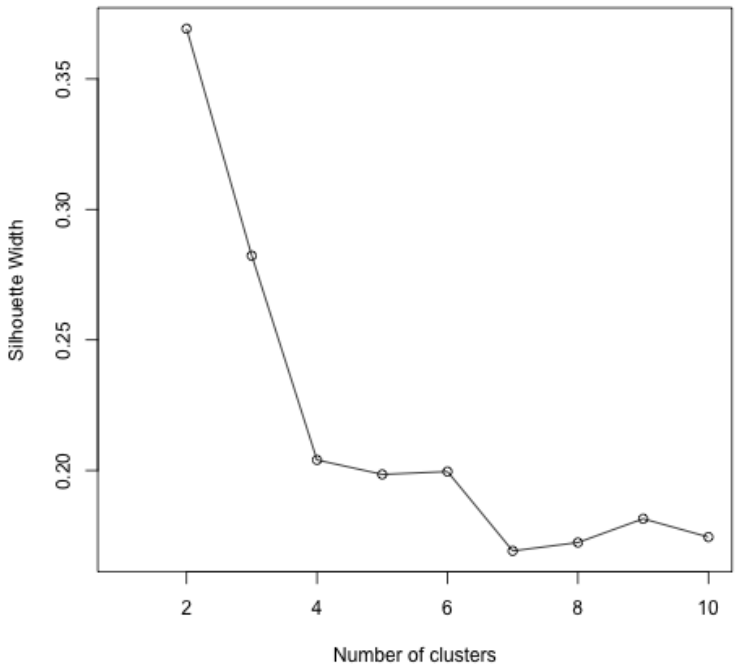


Figure 4.1: Scree plot to determine an appropriate number of clusters.

destinations in C1 are mainly located on the beach, vibrant and lifeful, and also good for various sports.

C2 (N = 2177). The medoid of C2 is Gubbio, a city located on the lowest slope of Mt. Ingino in Italy. Its origins are ancient and reaches to the Bronze Age. Thus, providing many cultural and sightseeing activities. Features *image & flair, hiking, culture, gastronomy, nightlife, mobility, accommodations, sports, sightseeing,* and *entertainment* are showing increased mean values in C2 (0.61 - 0.77). Interestingly, 17% of the destinations in C2 are metropoles, which is about six times more considering the whole data set. Plus, only 1% of the destination in C2 are located at the sea or beach. Hence, destinations in C2 can be considered as mainly vibrant cities or metropoles not located at the beach, offering many nightlife, cultural, sightseeing, gastronomy, and entertainment opportunities.

C3 (N = 1774). The medoid of C3 is Aghios Markos, a small, peaceful village in the nature on the island of Corfu (Greece). 90% of the destination in C3 are located at the sea, 88% at the beach, and 70% on an island. Whereas, in the whole data set only 20% of destinations are located at the sea or beach and 25% on an island. Ratings *beach & swimming, nature & landscape,* and *peacefulness* have an increased

mean value of 0.77-0.79. Furthermore, there is only one metropole in C3. Therefore, destinations in C3 can be seen as small and peaceful townss at seaside, probably on an island, with a few sports opportunities and not much tourists.

C4 (N = 5576). The medoid of C4 is Montbrió del Camp, a small, peaceful village in Catalonia (Spain). The average value of motivational rating *peacefulness* in C4 is 0.81. Also, ratings *nature & landscape, hiking, cycling* and *mountain biking* have an increased mean of 0.62-0.67. Interestingly, none of the 5576 destinations are located on an island or are metropoles. Furthermore, all other features of destinations in C4 are relatively low. Hence, destinations of C4 can be considered as small and peaceful villages, probably in the nature, and more or less good for hiking, cycling, and mountain biking.

C5 (N = 1877). The medoid of C5 is Reynoldston, a small, peaceful village in Wales (Great Britain). Interestingly, all destinations within this cluster are located on an island, only 2% are at the beach, and there is only one metropole. Further, only ratings *peacefulness* (0.76), *nature & landscape* (0.71), and *hiking* (0.64) are showing an increased mean, all other destination features have a relatively low average value. C5 is quite similar to C4, except destinations of C5 are only located on islands, where destinations of C4 are not. Thus, destinations of C5 can be considered as mainly small, peaceful villages, located on an island and in the nature, with some recreational sports offers.

C6 (N = 3606). The medoid of C6 is Irun, a city in Spain at the border to France and on the Atlantic coast. It offers some cultural and sightseeing activities, but also some sports and recreational activities in the nature. Following ratings have an increased average values (0.63 - 0.71) within this cluster: *nature & landscape, peacefulness, image & flair, mountain biking, cycling, nightlife,* and *culture*. In C6, 12% of the destinations are located near a mountain, which is about three times more compared to the whole data set. Only 1% of the destination are considered as metropoles and only 1% are located at the beach or sea. Thus, destination within C6 can be considered as small cities, probably in the nature, with recreational, cultural, and entertainment offers, but none them are dominating.

In summary, it can be said that there is an underlying natural structure of the data. Thus, six conceptually meaningful groups of destinations could be identified. For a better understanding, these groups or clusters can be simplified and summarized as follows: C1 - *vibrant beach resorts*, C2 - *energetic cities*, C3 - *tranquil seaside resorts*, C4 - *peaceful towns*, C5 - *idyllic island villages*, C6 - *ordinary towns*. Also, it is clear that the identified underlying structure is based on following three axes: vibrant/tranquil, land/island, seaside/inland.

The silhouette plot in Figure 4.2 displays the silhouette coefficients of each destination in a cluster in an ordered way. The red dashed line shows the average silhouette width of 0.2, which assisted to find the right cluster size. The silhouette plot enables a visual

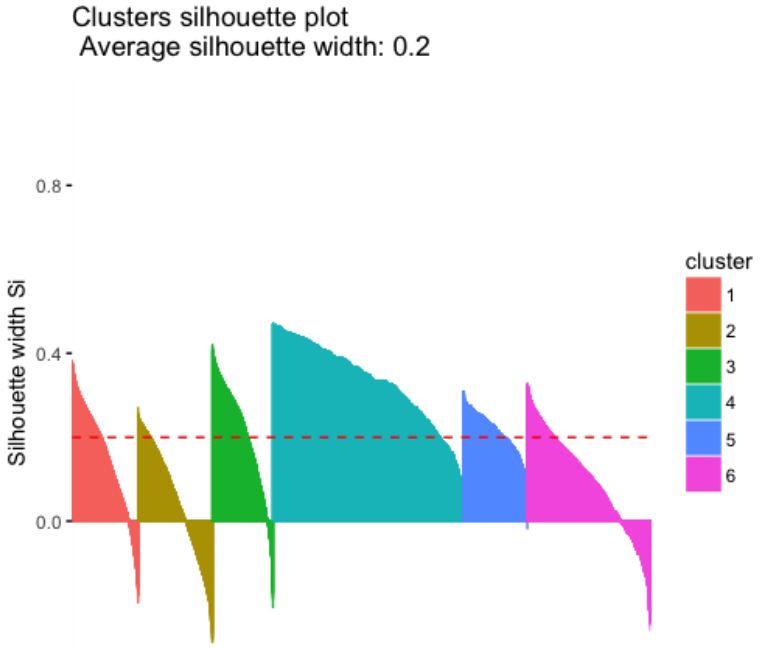


Figure 4.2: Silhouette plot of the 6 cluster solution.

assessment of the relative quality of the developed clustering. A negative silhouette coefficient indicates an incorrect assignment of a destination to a cluster and a very low silhouette coefficient points out that a destination is located in-between two clusters. Hence, almost none of the destinations in C4 and C5 are incorrectly assigned, but some might be located between two clusters. Whereas, in all other clusters there are falsely assigned destinations, especially in C2 and C6.

The conducted cluster analysis considers the complete data set (after missing value treatment). Hence, it also contains the tourism destinations of the expert sample. Therefore, the resulting cluster solution can be further assessed by examining the factor score distribution of each factor of the Seven-Factor Model over the six clusters. In Table 4.1 average factor scores and corresponding standard deviations (SD) in different clusters are listed.

Sun & Chill-Out scores the best in C1 - *vibrant beach resorts* and C3 - *tranquil seaside resorts*, which does not need any further explanation. *Knowledge & Travel* shows only in C2 - *energetic cities*, where many cultural and sightseeing activities are offered, an increased value. *Independence & History* and *Culture & Indulgence* are scoring the best in C2 - *energetic cities* and have increased scores in C1 - *vibrant beach resorts*. Both clusters C1 and C2 are offering cultural, entertainment, and sightseeing activities. *Social & Sports* is scoring the best in C1 - *vibrant beach resorts* with 0.64, but also similarly in all

	F1	F2	C3	F	C4	C5	C6
Sun & Chill-Out	mean	0.71	0.18	0.96	0.22	0.39	0.30
	SD	0.31	0.26	0.11	0.25	0.43	0.33
Knowledge & Travel	mean	0.46	0.65	0.24	0.14	0.18	0.30
	SD	0.31	0.30	0.21	0.21	0.23	0.24
Independence & History	mean	0.58	0.73	0.42	0.31	0.32	0.45
	SD	0.23	0.22	0.21	0.24	0.28	0.24
Culture & Indulgence	mean	0.58	0.76	0.41	0.23	0.24	0.44
	SD	0.30	0.26	0.22	0.26	0.30	0.30
Social & Sports	mean	0.64	0.57	0.50	0.59	0.58	0.58
	SD	0.16	0.19	0.16	0.18	0.14	0.20
Action & Fun	mean	0.59	0.56	0.30	0.05	0.12	0.19
	SD	0.34	0.34	0.19	0.15	0.17	0.20
Nature & Recreation	mean	0.35	0.31	0.76	0.82	0.79	0.72
	SD	0.28	0.30	0.18	0.17	0.22	0.25

Table 4.1: Average factor scores plus standard deviations in different clusters.

other clusters. This is reasonable, since in all clusters different kind of sports are offered. *Action & Fun* is scoring the best in C1 - *vibrant beach resorts* and has a similar score in C2 - *energetic cities*, where both C1 and C2 are the only clusters, which are considered as vibrant and energetic. Finally, *Nature & Recreation* scores well in destinations, which are considered as peaceful and recreational. These destinations are mainly located in C3 - *tranquil seaside resorts*, C4 - *peaceful towns*, C5 - *idyllic island villages*, and C6 - *ordinary towns*.

4.2 Regression Analysis

The aim of the work is not only to project destinations into the seven-dimensional vector space of travel behavioral patterns using their features, but more importantly to understand the relationship between the Seven-Factors and the destination features. In [JWHT13c] it is suggested to choose linear models over more complex ones if inference and interpretability is the goal. Taking this into account, a multiple linear regression model [JWHT13b] with step-wise variable selection [JWHT13a] is applied. All Seven-Factors are considered as independent from each other, since they are obtained from factor analysis. Therefore, they can be treated separately by fitting a model for each travel behavioral pattern, which takes the features of a destination as input and returns the factor score (0 to 1) as output. The regression analysis is considering the expert sample (after missing value treatment), which contains 350 tourism destinations. The expert sample is split into a training and test set in a ratio of 80/20. Model performance is assessed by R^2, the proportion of variance explained, and root mean square error (RMSE), the standard deviation of the residuals / prediction errors. Furthermore, a performance evaluation is

conducted, in order to compare the performance of the linear model against following two more complex models: K-Nearest Neighbors (KNN) Regression and Random Forest Regression. Finally, the outcomes are evaluated by assessing the performance against a baseline and by examining the distribution of predicted factors.

4.2.1 Methods and Measures

Stepwise Regression

Linear regression is a very simple approach for supervised learning, but it is a proper method if inference and interpretability are crucial. For each factor in the Seven-Factors Model a model is fitted. Since there are 38 destination features (after the preprocessing) and always one target variable (a factor) a multiple linear regression model is chosen. Using all available features might lead to an overfitting problem. Overfitting occurs if the constructed models are performing well in the observed data (training data) but poorly out of sample (test data / unseen data). Essentially this means, the models are following errors or noise too closely. Further, if the number of observations is not much larger than the number of features, there can be much variability in the least squares estimate, which might result in overfitting and consequently to poor predictions. Also, it is often the case that not all of the available features can be associated with the target variable and including such leads to unnecessary complexity. Therefore, finding and using essential features will reduce the complexity, work against overfitting, plus lead to a more interpretable model [JWHT13a].

Considering all this, it is clear that a sub set of the tourism destination features hast to be selected for each factor. In [JWHT13a] four well known approaches are discussed and can be summarized as follows:

Best Subset Selection. Here all possible combinations of features are applied and assessed using some criterion/measure. This approach gets computational infeasible if the number of features is large.

Forward Stepwise Selection. This is a more efficient alternative to the best subset selection. This method starts with a model with no features and creates models by adding features one by one, until all features are used. At each step the feature, which adds the most additional improvement to the model, is chosen. Finally, the best performing model among all created models is selected by using some criterion.

Backward Stepwise Selection. This method is similar to the Forward Stepwise Selection, but in contrast it starts with a full model and at each step a model is created by eliminating the least useful feature.

Hybrid Approaches. This approach is a combination of forward and backward stepwise selection. The computational benefits compared to the best subset selection is still given.

In this work a hybrid approach is followed since it mimics the best subset selection more closely. Further, the Bayesian Information Criterion (BIC) [S+78] is used in order to select the best performing model, among all models created during each step. As already mentioned, adding more and more features to a model might lead to overfitting. In order to counter this issue BIC introduces a penalty term for the number of features in the model. BIC is a more conservative criterion compared to other well-known measures used in this context. Thus, it leads to a more concise, but still powerful model.

$\mathbf{R^2}$

Also known as the coefficient of determination, R^2 shows the proportion of variance in the target variable that can be explained through the model. In a multiple linear regression model is equal to $Cor(Y, \hat{Y})^2$, i.e. the squared correlation between the real target values Y and the predicted target $\hat{Y}$. A value near 0 indicates that the model does not explain much of the variability in the target, whereas a number close to 1 shows that the model explains a large proportion of the variance [JWHT13b].

Root-Mean-Square-Error (RMSE)

Gunawardana and Shani are providing in [GS09] an overview of evaluation measures used in different recommendation task. Under the topic "Predicting Ratings" they mention $RMSE$ as the most common and popular measure for evaluation. Taking this into account and considering the fact that predicting a score for each factor of the Seven-Factor Model is similar to predicting ratings, $RMSE$ will also be used to assess and compare the performance of the developed models in this work. $RMSE$ is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n}} \tag{4.1}$$

where y_i represents the true target value of the i^{th} observation, $\hat{y}_i$ the predicted, and n the number of observations. $RMSE$ is measured in the same scale (and units) as the target variable, which facilitates the interpretation and communication of the value. Of course, the smaller the error the better.

Baseline Function

Since the target variable, i.e. the Seven-Factors, are continuous, the proper baseline function is just the simple mean. In other words, the resulting models are compared to the performance of a simple mean function $\hat{y} = mean(Y_{train})$, where $\hat{y}$, the predicted value, is always the mean of the true values in the training set Y_{train}. In Table 4.2 the average values of the Seven-Factors in the training set are listed.

KNN Regression

In [JWHT13b] the KNN regression is described as following: *"Given a value for K and a prediction point x_0, KNN regression first identifies the K training observations that*

	mean
Sun & Chill-Out	0.47
Knowledge & Travel	0.38
Independence & History	0.51
Culture & Indulgence	0.50
Social & Sports	0.58
Action & Fun	0.36
Nature & Recreation	0.56

Table 4.2: Average factor scores of destinations in the training set.

are closest to x_0, represented by N_0. It then estimates $f(x_0)$ using the average of all the training responses in N_0." In other words, the target value of an unseen observation x_0 is predicted by the mean of the K closest neighbors of x_0 in the training set, which can be written as follows:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i \tag{4.2}$$

K can be seen as a kind of regularization parameter. A small K leads to a more "jumpy" solution, since it considers fewer observations and more local phenomena. Whereas, a large K leads to a smoother result, since more observations are considered. This can be observed in Figure 4.3, where a KNN regression on a two-dimensional data set is displayed.

In this work, the optimal K is determined through a cross-validated parameter search. In other words, different values of K are applied and cross-validated and then the best performing K is chosen.

Random Forest Regression

The Random Forest (RF) method is thoroughly explained in [JWHT13d] and will be briefly discussed in the following. The RF method is based on the decision tree approach. Generally, if the relationship between target variable and the features are approximately linear, then a linear method is more likely to outperform decision trees. Whereas, if the relation is more complex and non-linear, then it is more likely that the decision tree will outperform such classical approaches. This is illustrated in Figure 4.4 as a classification problem, where the true decision boundary of the example in the top row is linear. Hence, a linear model (left) is outperforming the decision tree (right). On the other hand, the example in the bottom row has a more complex, non-linear decision boundary. Thus, the decision tree approach (right) is outperforming the linear method (left).

One can see that the decision tree method is performing poorly in some situations. However, by simply aggregating many decision trees such issues can be handled and a

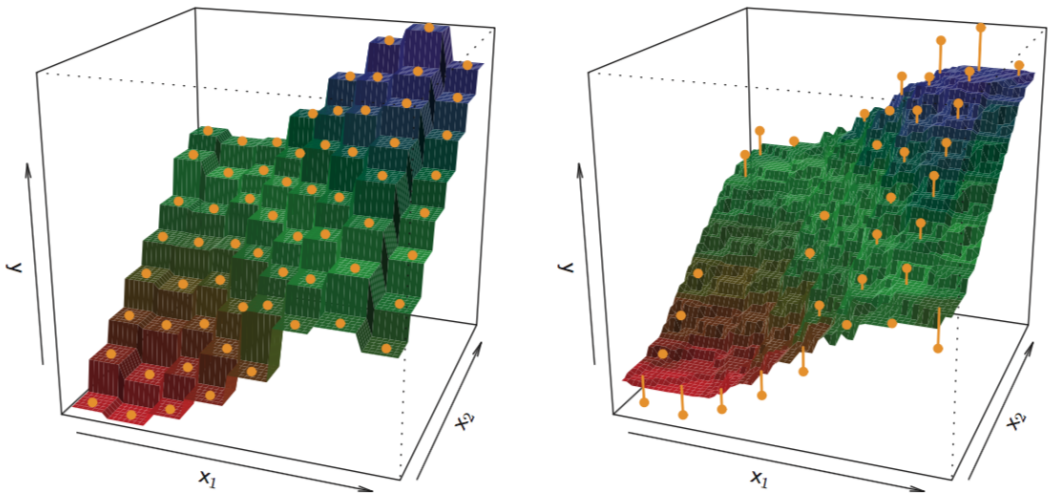


Figure 4.3: Example KNN regression with different K. Left: K = 1. Right: K = 9 [JWHT13b].

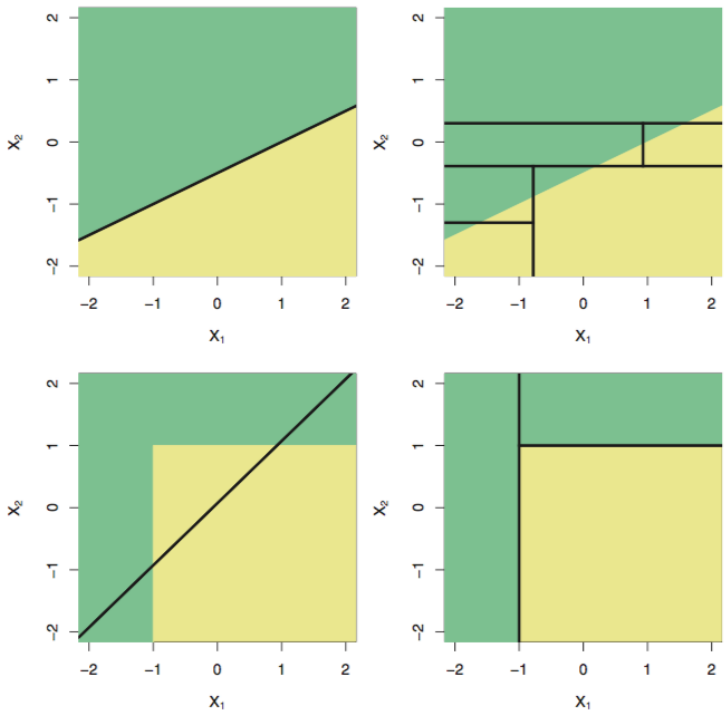


Figure 4.4: Example of two dimensional classification problem. First row: linear true decision boundary. Second row: non-linear true decision boundary [JWHT13d].

better predictive performance can be reached. In the RF method exactly this is done. Additionally, at each split in a tree only a random sample of m features from the full set of p features are chosen. This randomness de-correlates the generated trees in the forest and gives the method its name.

The most important parameters which has to be tuned wisely are the number of trees in the forest and the number of features to consider when looking for the best split. In this work both parameters are determined by a cross validated grid search, where different values are assigned to both parameters and the best performing parameter combination is selected.

4.2.2 Resulting Multiple Linear Regression (MLR) Models

Sun & Chill-Out

	Coefficient	Std. Error	Signif.
(Intercept)	0.41	0.06	***
beach & swimming	0.73	0.11	***
nightlife	-0.76	0.08	***
health resort	0.27	0.05	***
sea	0.23	0.06	***

Table 4.3: Multiple linear regression model for the factor *Sun & Chill-Out*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 4.3 the coefficients of the multiple linear regression model for the factor *Sun & Chill-Out* are listed. The geographical attributes *sea*, *health resort*, and especially the motivational rating *beach & swim* have a significant positive impact on the factor *Sun & Chill-Out*. Those features can be interpreted as indicators for sun and relaxation. On the other side, the motivational rating *nightlife* has a significantly strong, negative impact, which can be associated with crowded places and mass tourism.

Knowledge & Travel

	Coefficient	Std. Error	Signif.
(Intercept)	-0.18	0.04	***
sightseeing	1.02	0.07	***
sea	-0.12	0.02	***
mobility	0.26	0.08	**
winter sports resort	-0.24	0.09	*

Table 4.4: Multiple linear regression model for the factor *Knowledge & Travel*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 4.4 the coefficients of the multiple linear regression model for the factor *Knowledge & Travel* are listed. The motivational rating *sightseeing* has a significant, strongly positive relation with the factor *Knowledge & Travel*. Hence, it is capturing the knowledge part of the factor. Whereas, the motivational rating *mobility* is also positively related to the factor and once can say it captures the travel part. On the other hand, the geographical attributes *sea* and *winter sports resort* are significantly negatively related with the factor. This is reasonable, since in such areas usually the tourism focus does not lie on gaining knowledge.

Independence & History

	Coefficient	Std. Error	Signif.
(Intercept)	0.08	0.06	
culture	0.61	0.1	***
sightseeing	0.39	0.09	***
nature & landscape	-0.17	0.07	*

Table 4.5: Multiple linear regression model for the factor *Independence & History*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 4.5 the coefficients of the multiple linear regression model for the factor *Independence & History* are listed. The motivational ratings culture and *sightseeing* are significantly, positively related to the factor *Independence & History*. Those features can be seen as the main motivation of travelers with interests in history and tradition. Whereas, the motivational rating *nature & landscape* has a significant negative impact on the factor. Since cultural and historical interests are short coming in nature and recreation related destinations, such negative association is reasonable.

Culture & Indulgence

	Coefficient	Std. Error	Signif.
(Intercept)	-0.09	0.08	
sightseeing	0.30	0.12	*
image & flair	0.48	0.11	***
nature & landscape	-0.27	0.09	**
old town	0.17	0.05	***
culture	0.47	0.13	***

Table 4.6: Multiple linear regression model for the factor *Culture & Indulgence*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 4.6 the coefficients of the multiple linear regression model for the factor *Culture & Indulgence* are listed. The motivational ratings *sightseeing*, *culture*, *image & flair*, and

geographical attribute *old town* are significantly, positively related to the factor *Culture & Indulgence*. Those ratings can be interpreted as the main motivation of a culture and history interested high class tourist. On the other side, the motivational rating *nature & landscape* has a significant, negative impact on the factor. Again, this might show that destinations branded with a nature and landscape motif have shortcomings in cultural tourism.

Social & Sports

	Coefficient	Std. Error	Signif.
(Intercept)	0.28	0.09	**
sports	0.85	0.11	***
sightseeing	-0.29	0.06	***
peacefulness	0.27	0.08	**
wellness	-0.31	0.10	**
mountains	0.09	0.03	**

Table 4.7: Multiple linear regression model for factor the *Social & Sports*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 4.7 the coefficients of the multiple linear regression model for the factor *Social & Sports* are listed. The motivational rating *sports* has a strong, significant, positive impact on the factor *Social & Sports*, which is obvious. Also, the motivational rating *peacefulness* and the geographical attribute *mountains* have a significant positive relation to the factor. Since the factor *Social & Sports* factor avoids crowded areas and locations of mass tourism, and prefers more tranquil places, positive associations of both features with the factor are reasonable. On the other hand, the motivational rating *sightseeing* has a significant, negative impact on the factor. It can be seen as an indicator of crowded areas and mass tourism. Surprisingly, the motivational rating *wellness* is significantly, negatively associated with the factor *Social & Sports*. This is caused by an unsound sample, as 55% of the destinations in the expert sample have a larger wellness rating (>0.5) and are located at the beach and 25% are metropoles, which is far less in the whole data set.

Aciton & Fun

In Table 4.8 the coefficients of the multiple linear regression model for the factor *Action & Fun* are listed. The motivational ratings *peacefulness*, *family*, and the geographical attribute *health resort* have a significant, negative impact on the factor. This fits perfectly to the character traits of the factor *Action & Fun*. Whereas, the motivational ratings *nightlife*, *winter sports*, *shopping*, and the geographical attributes *sea* and *metropolis* are significantly positively related to the factor. Those can be interpreted as features of

	Coefficient	Std. Error	Signif.
(Intercept)	0.02	0.15	
peacefulness	-0.53	0.12	***
sea	0.17	0.04	***
metropolis	0.19	0.05	***
wintersports	0.54	0.11	***
shopping	0.42	0.09	***
health resorts	-0.11	0.04	**
nightlife	0.41	0.12	***
family	-0.33	0.09	***
kite & windsurfing	0.10	0.09	***

Table 4.8: Multiple linear regression model for the factor *Action & Fun*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

energetic, vibrant and action loaded places, which are main aspects of destinations for thrill seeking and action loving travelers.

Nature & Recreation

	Coefficient	Std. Error	Signif.
(Intercept)	0.61	0.11	***
peacefulness	0.51	0.09	***
sightseeing	-0.27	0.06	***
hiking	0.35	0.06	***
nightlife	-0.57	0.10	***
health resort	0.09	0.03	**
shopping	-0.22	0.07	**
beach	-0.05	0.02	*

Table 4.9: Multiple linear regression model for the factor *Nature & Recreation*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 4.9 the coefficients of the multiple linear regression model for the factor *Nature & Recreation* are listed. The motivational rating *peacefulness*, *hiking*, and the geographical attribute *health resort* are significantly positively related to the factor *Nature & Recreation*, which is obvious and does not need further explanation. On the other side, the motivational ratings *nightlife*, *sightseeing*, *shopping*, and the geographical attribute *beach* have a significant, negative impact on the factor. Those features can be interpreted as signs of mass tourism and crowded areas. Hence, a negative association on a recreational and escapist traveler is reasonable.

4.2.3 Evaluation

The resulting models are evaluated by assessing both in sample and out of sample performance. In other words, the performance measures are determined using both training set (in sample) and test set (out of sample). Obviously, out of sample performance plays a bigger role, because it delivers an approximation to the question, How the model will perform using unseen data. Still, in sample performance also provides some crucial insights. For example, it might give some hint, whether the developed models are overfitting. As already mentioned, the used performance measures are $RMSE$ and R^2. Furthermore, the resulting linear regression models are compared to an appropriate baseline function f_0 (see Section 4.2.1 and Table 4.2) in order to show, whether the resulting models actually did learn something. Additionally, two more complex and non-linear models (i.e., KNN and RF) are developed, in order to challenge the performance of the simple linear model.

In Table 4.10 the training and test performance of the baseline function (f_0), the multiple linear regression (MLR), the KNN regression (KNN), and the Random Forest regression (RF) are listed. Note, that f_0 is always a constant function. Thus, it does not explain any variance in the factor scores. Therefore, R^2_{train} and R^2_{test} of f_0 is always zero. Training and test performance of the MLR model is close together, which shows that this model is not much overfitting. Whereas, the RF model and especially the KNN model are overfitting the training set, i.e. the training performance is much better than the test performance. For example, an extreme case is the KNN model for factor *Action & Fun*, where R^2_{train} is 1.00 (100% of the variance in the factor is explained) and R^2_{test} is 0.63 and also $RMSE_{train}$ is 0.01 (almost perfect) and $RMSE_{test}$ is 0.21. Although both models are well tuned, the overfitting can be a sign of too few training data, but it also shows a potential for enhancement if more data is used.

Overall, the out-of-sample performance of all three models MLR, KNN, and RF are pretty close. Hence, one can expect that they will perform similar if confronted with unseen data. The overall performance of all three models (MLR, KNN, RF) are always better than the simple mean function f_0, which indicates that the models must have learned something out of the data. The difference is in most cases clear to observe, except in factor *Social & Sports*. Here, the $RMSE_{test}$ of f_0 is 0.19 and MLR, KNN, and RF have a $RMSE_{test}$ of 0.17-0.18. This is caused by an uneven distribution of the expert mapping, where 87% of the destinations have scored with 0.5 or 0.75. Hence, a constant prediction of 0.58, like f_0 does, is performing pretty well, but it also means that there is less information to learn from. On the other hand, the models are performing the best in factor *Nature & Recreation*, where $RMSE_{test}$ is 50% smaller than the baseline. The out of sample performance of the KNN model is always a tick worse than the MLR and RF model, whereas there is almost no difference in the performance of RF and MLR. Thus, discarding the KNN model and choosing the MLR model over RF is reasonable since they are performing similar but the MLR model is much simpler to fit and easier to interpret.

		f_0	MLR	KNN	RF
Sun & Chill-Out	R^2_{train}	0.00	0.68	0.78	0.94
	R^2_{test}	0.00	0.62	0.61	0.64
	$RMSE_{train}$	0.40	0.23	0.19	0.10
	$RMSE_{test}$	0.40	0.25	0.25	0.24
Knowledge & Travel	R^2_{train}	0.00	0.72	0.62	0.85
	R^2_{test}	0.00	0.71	0.64	0.70
	$RMSE_{train}$	0.32	0.17	0.20	0.12
	$RMSE_{test}$	0.33	0.18	0.20	0.18
Independence & History	R^2_{train}	0.00	0.65	0.46	0.71
	R^2_{test}	0.00	0.59	0.58	0.62
	$RMSE_{train}$	0.27	0.17	0.20	0.14
	$RMSE_{test}$	0.28	0.17	0.18	0.17
Culture & Indulgence	R^2_{train}	0.00	0.69	0.99	0.79
	R^2_{test}	0.00	0.61	0.58	0.67
	$RMSE_{train}$	0.33	0.20	0.03	0.15
	$RMSE_{test}$	0.35	0.21	0.22	0.20
Social & Sports	R^2_{train}	0.00	0.28	0.22	0.54
	R^2_{test}	0.00	0.22	0.06	0.16
	$RMSE_{train}$	0.18	0.15	0.16	0.12
	$RMSE_{test}$	0.19	0.17	0.18	0.17
Action & Fun	R^2_{train}	0.00	0.73	1.00	0.88
	R^2_{test}	0.00	0.68	0.63	0.70
	$RMSE_{train}$	0.35	0.18	0.01	0.12
	$RMSE_{test}$	0.36	0.20	0.21	0.19
Nature & Recreation	R^2_{train}	0.00	0.80	1.00	0.92
	R^2_{test}	0.00	0.77	0.69	0.75
	$RMSE_{train}$	0.33	0.15	0.02	0.10
	$RMSE_{test}$	0.34	0.17	0.19	0.17

Table 4.10: Comparison of performance measures of baseline function (f_0), multiple linear regression (MLR), KNN regression (KNN), and Random Forest regression (RF) in test and training set.

In contrast to the previous analysis, where the focus is predictive performance, now the distribution of the predicted factor scores is analyzed. In detail, the factor score distribution of the expert mapping is compared to the distribution behavior of predicted factor scores. In order to do so, the build multiple linear regression model of each factor is fed with the complete data set as input. Then the resulting distribution in factor scores is compared to the one in the expert mapping. This comparison will foster a better understanding of the generalization power of the developed models.

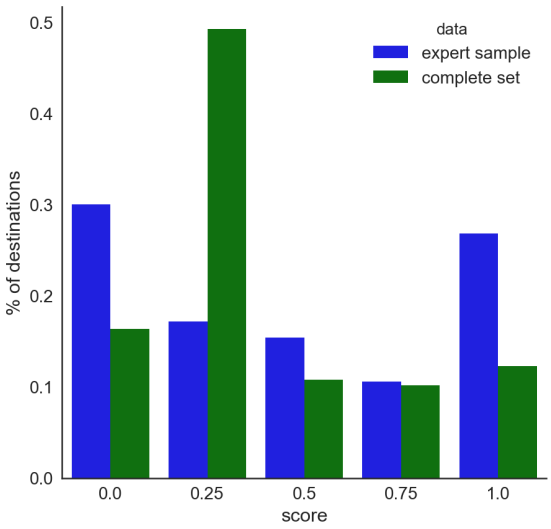


Figure 4.5: Comparison of the factor score distribution in the expert sample versus predicted scores of the complete data set for factor *Sun & Chill-Out*.

In Figure 4.5 the distributions of factor scores for the factor *Sun & Chill-Out* are compared. Here, 49% of the destination in the complete set are scoring with 0.25. This is not observable in the expert sample. The expert sample shows an increased amount of destinations with score 0 (30%) and 1 (27%), whereas 43% of destinations score either with 0.25, 0.5 or 0.75. A similar but damped behavior can be observed in the predicted factor scores of the complete set (setting aside the peak at score 0.25).

Figure 4.6 shows the distributions of factor scores for the factor *Knowledge & Travel*. Taking into account the expert sample, the majority of destinations score either with 0 or 0.25 and with increasing factor score the amount of destinations decays. A similar behavior can be observed in the predicted factor scores of the complete set.

Figure 4.7 compares the distributions of factor scores for the factor *Independence & History*. Considering the predicted factors of the complete set, once again one can see a peak at score 0.25 like previously in *Sun & Chill-Out*. Besides that, the distribution has more or less a normal shape (bell), similar to the factor score distribution of the expert mapping.

The distributions of factor scores for the factor *Culture & Indulgence* are displayed in

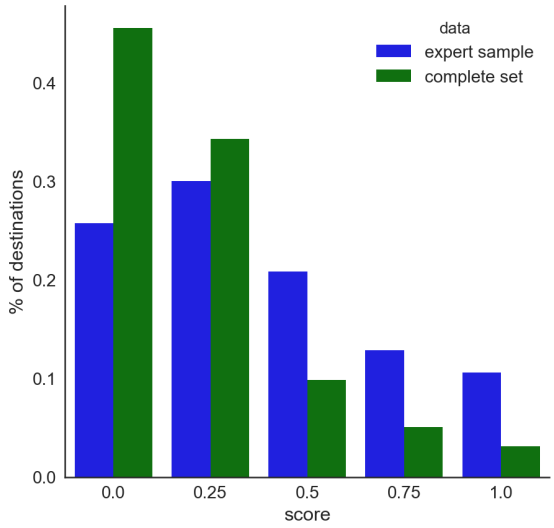


Figure 4.6: Comparison of the factor score distribution in the expert sample versus predicted factor scores of the complete data set for factor *Knowledge & Travel*.

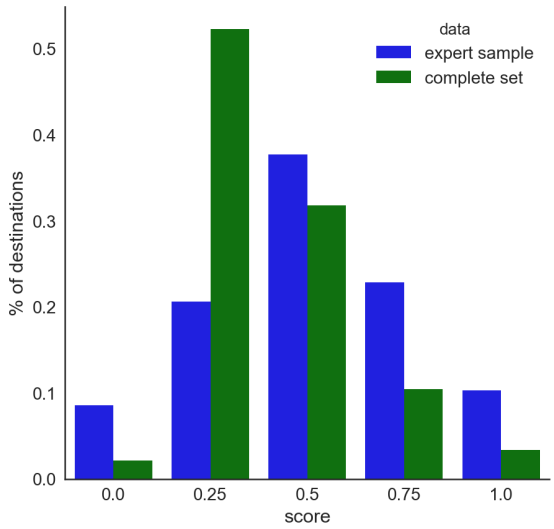


Figure 4.7: Comparison of the factor score distribution in the expert sample versus predicted scores of the complete data set for factor *Independence & History*.

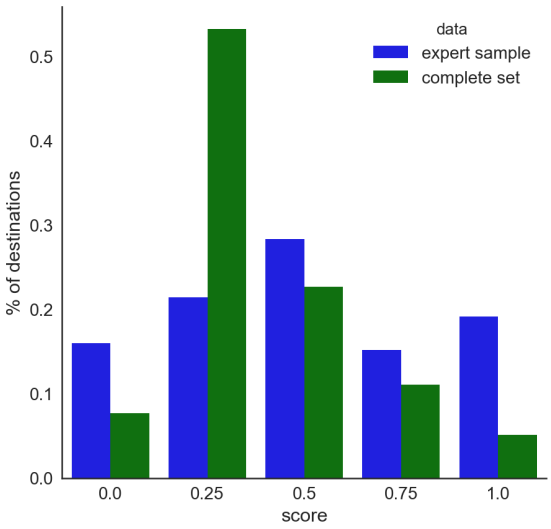


Figure 4.8: Comparison of the factor score distribution in the expert sample versus predicted scores of the complete data set for factor *Culture & Indulgence*.

Figure 4.8. Looking at the predicted factors of the complete set, there is again a peak at score 0.25 (57%), which is not observable in the expert mapping. At score 0.5 and 0.75 the percentage of destinations in the expert sample (28% and 15%) are relatively close to the ones in the complete set (23% and 11%). On the other hand, this is not the case for scores 0 or 1, where the difference is much higher.

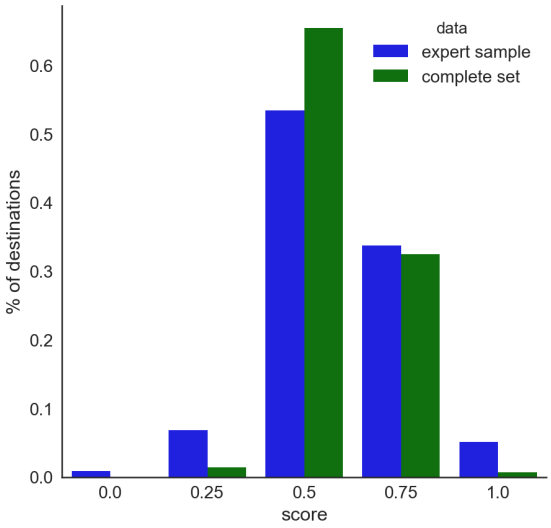


Figure 4.9: Comparison of the factor score distribution in the expert sample versus predicted scores of the complete data set for factor *Social & Sports*.

For the factor *Social & Sports* the factor score distribution in the expert mapping and the complete set are pretty similar, which can be observed in Figure 4.9. The vast majority of destinations score either with 0.5 or 0.75 in both, whereas only few destinations score with 0, 0.25 or 1.

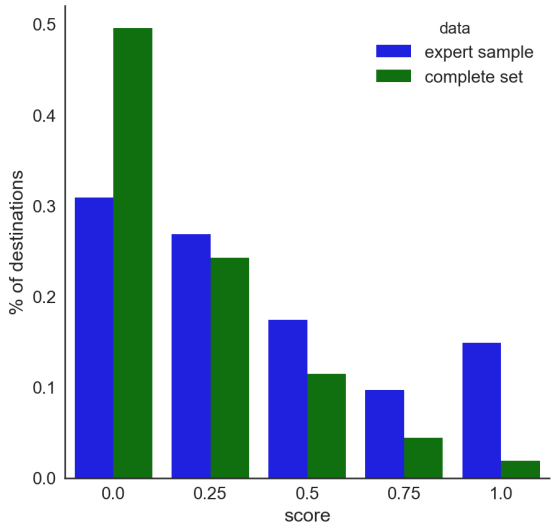


Figure 4.10: Comparison of the factor score distribution in the expert sample versus predicted scores of the complete data set for factor *Action & Fun*.

In Figure 4.10 the distributions of factor scores for the factor *Action & Fun* are displayed. Considering the expert mapping, one can see that the majority of destination score either with 0 or 0.25 and that this amount is decaying the higher the score gets. A similar behavior can be observed by looking at the predicted factors of the complete set.

Finally, Figure 4.11 shows the distributions of factor scores for the factor *Nature & Recreation*. In the predicted scores of the complete set the amount of destinations is increasing with increasing factor scores. This cannot be observed in the expert mapping. Still, in both, the expert sample and the complete set, most of the destinations are scoring with 0.5 or more.

To sum up, there are some differences in the distributions of factor scores between the manually labeled expert sample and the predicted scores of the complete set. But overall, both show similar trends in the distributions. This shows that the build multiple linear regression models are mimicking the experts quite good. Hence, one can expect a sufficient generalization.

In conclusion, one can say that tourism destination features can be used to determine the Seven-Factor representation of a destination. The multiple linear regression outperformed the K-Nearest-Neighbor regression and the Random Forest regression. The multiple linear regression models were able to explain 59 – 77% of the variance in factor scores of the destinations in the test set. Only the model for factor *Social & Sports* showed a

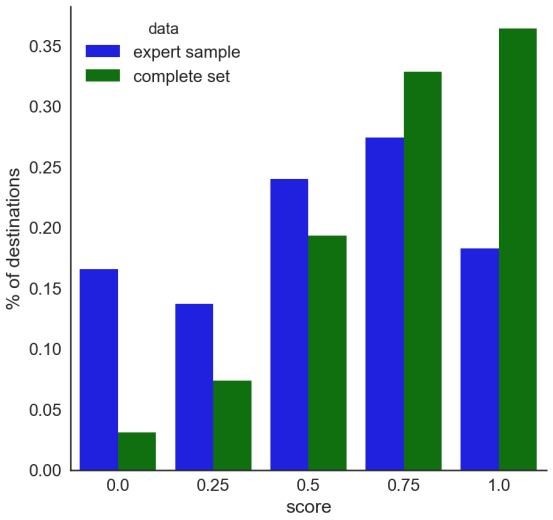


Figure 4.11: Comparison of the factor score distribution in the expert sample versus predicted scores of the complete data set for factor *Nature & Recreation*.

relatively poor performance (in comparison with the models of the other factors), where 22% of the variance could be explained. This is caused by an uneven distribution of factor scores in the expert mapping, where 87% of the destinations scored either with 0.5 or 0.75.

Factor	Destination features
Sun & Chill-Out	*- nightlife, beach & swimming, health resort, sea*
Knowledge & Travel	*sightseeing, mobility, - winter sports resort, - sea*
Independence & History	*culture, sightseeing, - nature & landscape*
Culture & Indulgence	*image & flair, culture, sightseeing, - nature & landscape, old town*
Social & Sports	*sports, - wellness, - sightseeing, peacefulness, mountains*
Action & Fun	*winter sports, - peacefulness, shopping, nightlife, - family, metropolis, sea, - health resort, kite & windsurfing*
Nature & Recreation	*- nightlife, peacefulness, hiking, - sightseeing, - shopping, health resort, - beach*

Table 4.11: Used destination features in the resulting multiple linear regression models.

Nevertheless, all multiple linear regression models were plausible and clearly interpretable. Table 4.11 lists all independent variables (destination features) of the fitted multiple linear regression models of the Seven-Factors. Note, that a minus sign indicates a negative

impact on the corresponding factor. For example, the model of the factor *Sun & Chill-Out* consists of indicators of sun and beach as expected, but there are also indicators off crowdedness, which have a negative impact on the factor. Such model structure is in line with the characteristics of *Sun & Chill-Out*, where crowdedness and mass tourism are negatively associated with the factor.

Futhermore, the resulting multiple linear regression models comprise both motivational ratings and geographical attributes. After the variable selection 15 out of 27 motivational ratings and seven out of 22 geographical attributes in total are used. Motivational ratings *sightseeing, peacefulness, nightlife, culture, nature & landscape,* and *shopping* appear in more than one model. Also, geographical attributes *health resort* and *sea* are used in several models.

CHAPTER 5

Hotels Data

In this chapter the provided data set for hotels is thoroughly described and analyzed. The data for hotels is provided by GIATA [Gmbb], a German e-Tourism company and market leader for tourism content with many internationally known customers like Expedia or TripAdvisor. In addition to the data set two labeled data samples (manually mapped onto the Seven-Factors) are provided by tourism experts. The upcoming sections are covering following topics: univariate and multivariate analysis, missing values and treatment, feature engineering.

5.1 First Insights

Besides tourism destinations, this work is focusing on another kind of tourism products, namely hotels. The data in use is provided as an archive with about one million XML-files deeply structured in folders. These XML-files are called "GIATA Fact Sheets" and each is describing a hotel offered by a certain travel agency like Ruefa or ThomasCook.

```
<?xml version="1.0" encoding="utf-8"?>
<factsheet giataId="75992">
 <factsset>
  <factgroup name="buildinginformation">
   <fact name="buildinginformation:numroomstotal">
    <attributes>
     <attribute name="number" value="278" />
    </attributes>
   </fact>
  </factgroup>
  <factgroup name="category">
   <fact name="category:official">
    <attributes>
```

```
      <attribute name="rating" value="3"/>
     </attributes>
    </fact>
   </factgroup>
   <factgroup name="distance">
   </factgroup>
   <factgroup name="facilities">
   </factgroup>
   <factgroup name="sports">
   </factgroup>
   ...
  </factsset>
</factsheet>
```

Listing 5.1: GIATA Fact Sheet snippet.

In Listing 5.1 a snippet of a GIATA Fact Sheet is given. Every GIATA Fact Sheet corresponds to a distinct hotel, which is identified by the *giataId*. Hotels are described through facts, which in turn are organized by fact groups. In the example snipped just a few fact groups are listed, but actually there are 14, namely *building information, category, distance, entertainment, facilities, location, meals, misc, object information, payment, rooms, spa, sports,* and *type*. Facts are attributes of a hotel with a value assigned. For example, in the given snippet the hotel has 278 rooms in total and is a 3 stars hotel. Next, fact groups, facts, and possible values are investigated further:

Building information. As already the name says, this fact group consists of facts about the hotel building. Facts within this group can only take discrete values and can be separated into two types

- Number of: apartments, bungalows, floors, rooms (total), suits etc.
- Year of: construction and renovation

Category. Category stands for the hotel category in "stars". There are two types of categories. First, there is an official one and second a recommended category from a travel agency or operator. In both cases category takes only discrete values from 1 to 7 with 0.5 stars steps. For example, a hotel can have 3 stars or 3.5 stars.

Distance. Here, distances to relevant touristic places in the near are listed, such as distances to bars/pubs, beach, bus station, city center, forest and much more. Distances are measured in meters.

Entertainment. This group encapsulates entertainment offers of the accommodation like live music, mini disco, childcare, entertainment for children, entertainment for adults and so on. Facts within this group are of type binary, showing the presence of an entertainment offer.

Facilities. Here are general facilities and services of a hotel listed, such as auditorium, babysitter, baggage room, bar, bicycle rental etc. Facts within this group are of type binary, showing the presence of a facility or service.

Location. In contrast to fact group distance, facts in the fact group location are just indicators of how a hotel is situated. Facts within this group are of type binary, showing whether a hotel is situated at the beach, centrally, on the main road, or quietly.

Meals. This group lists different kind of meal packages (e.g. all inclusive, bed and breakfast etc.), special diets (e.g. gluten free) or services (e.g. room service, show cooking etc.). Facts within this group are of type binary, showing the presence of such offer.

Misc. For now, the misc group only contains one fact, namely car rental, a binary value showing if there is a car rental possibility.

Object information. This group contains general information of a hotel, where facts with string values are persisting address, hotel chain name, email, fax, phone, and url of a hotel.

Payment. Also, this group only contains one fact, namely payment, which is categorical and lists different kind of payments methods.

Rooms. This group is similar to the fact group facilities, but it contains possible attributes of a hotel room such as air conditioning, balcony, bathroom, hairdryer, kingsize bed, and much more. Except the fact room size, provided in square meters, all other facts listed within this group are of type binary.

Spa. Here, different kind of wellness offers of a hotel are listed. Facts are of type binary and are showing the presence of a wellness offer like acupuncture, ayurveda, hammam, sauna etc.

Sports. Here, different kind of sport offers of a hotel are grouped. Facts are of type binary and are showing whether a hotel is offering certain kind of sports like aerobics, badminton, beach volleyball, biking, gym, canoe, etc.

Type. Type stands here for hotel type. Facts within this group are for example adults only, airport hotel, apartment hotel, beach hotel, camping ground, casino resort. These facts are of type binary, indicating whether a hotel is of particular type.

Note, that not all possible facts are listed. There are about 300 facts in total and a list of all is provided in [Gmb10].

Since, a hotel can be offered by many travel agencies many times (e.g. winter-, summer-, easter-campaigns) the downloaded data archive of over one million GIATA Fact Sheets includes many duplicates, i.e. the unique identifier *giataId* of a hotel appears in more

than one Fact Sheet. The XML-files are parsed, preprocessed, and transformed into a more convenient and tabular format of a CSV file, where each row corresponds to an offer (Fact Sheet) and each column to a fact. Afterwards, rows are grouped by (merged) *giataId* in order to get only one representation for each hotel. This results in a data set of 143408 distinct hotels. Figure 5.1 shows the distribution of hotels over countries as a heat map. The majority of hotels are located in USA, Italy, Spain, Germany, Greece, and Great Britain (51.74%).

Figure 5.1: Distribution of hotels in the GIATA data set.

There are about 300 distinct facts of hotels, but the provided XML-files do not have an entry for each of them. Only known characteristics of hotels are listed in the GIATA Fact Sheets, i.e. not possessed attributes or not known attributes of hotels do not have an entry in the XML-File. Thus, there is an ambiguity in missing data, which leads to a sparse data set like perviously in the Webologen data set. Missing data will be further analyzed and treated in the upcoming sections.

Furthermore, the size of a hotel can be associated with its total number of rooms. Unfortunately, only for 35202 hotels (25%) the total number of rooms are given. Considering those, one can say that hotels in the GIATA data set are ranging from small apartments with just two rooms to big complexes with more than 1000 rooms. On average hotels have 146 rooms in total. The year of construction is present for 27746 hotels (19%) and the majority of those hotels are built in the 20th century (59%), but there are also few really old ones. Hotel category is ranging from zero stars to seven stars, where most of the hotels have either 3 stars (37%) or 4 stars (23%).

Distance measure are really rare in the used data set and about 87% of the hotels do

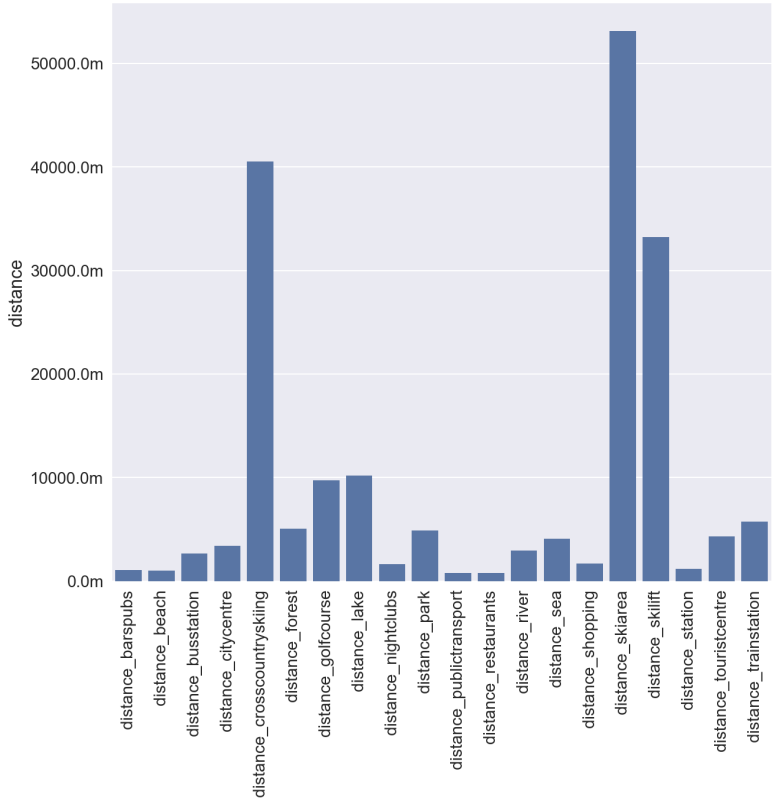


Figure 5.2: Average values of all distance measures in fact group distance.

not have any facts of the fact group distance. Figure 5.2 depicts the mean of various distances considering only the non-missing ones. All distances are below 10km except distances related to skiing namely distance to cross country skiing offers (40km), to skiing area (53km), and to the skiing lift (33km). The majority of distances are even lower or equal to 5km. Based on given distance measures, one can conclude that they are only non-missing if there is something relevant nearby. Hence, facts of fact group distance are binary encoded for further analysis, where one indicates that a particular point of interest is close to the hotel.

Table 5.1 shows the ten most frequent facts of fact group facilities. Most frequent facility facts are Wi-Fi (76%) and internet access (71%), which have in most cases interchangeable meaning. These are followed by common attributes of hotels like carpark (49%), restaurant (43%), reception (41%) and so on. As already mentioned, these are frequencies of facts, which are appearing in the GIATA Fact Sheets. In reality some frequencies are probably higher than in the used data set, for example in the case of reception.

In Table 5.2 the least frequent facility facts are listed. One can immediately see that attributes related with conference/congress hotels are really rare.

	% of hotels
facilities_wifi	76.11
facilities_internetaccess	70.65
facilities_carpark	49.04
facilities_restaurant	43.14
facilities_reception	40.82
facilities_bar	40.49
facilities_outdoorpool	34.41
facilities_elevators	30.70
facilities_safe	28.53
facilities_laundry	28.12

Table 5.1: Ten most frequent facility facts.

	% of hotels
facilities_dvdrental	00.17
facilities_secretarialservice	00.12
facilities_photocopier	00.10
facilities_golfdesk	00.03
facilities_tourdesk	00.02
facilities_overheadprojector	00.02
facilities_flipchart	00.01
facilities_projector	00.01
facilities_translator	$<$ 00.01
facilities_congressfacilities	$<$ 00.01

Table 5.2: Ten least frequent facility facts.

	% of hotels
rooms_bathroom	78.00
rooms_tv	76.62
rooms_internetaccess	70.86
rooms_aircon	67.14
rooms_wifi	66.16
rooms_hairdryer	66.11
rooms_shower	63.85
rooms_phone	63.12
rooms_doublebed	54.71
rooms_safe	47.53

Table 5.3: Ten most frequent room attributes.

Table 5.3 shows the ten most frequent hotel room facts. As in facilities fact group, also here are the top frequent attributes as expected and not surprising like bathroom, TV, internet access, air conditioning, and Wi-Fi.

	% of hotels
rooms_newspaper	0.46
rooms_adapterplug	0.27
rooms_goodnightservice	0.18
rooms_choicetowel	0.14
rooms_welcomegift	0.14
rooms_backgroundmusic	0.13
rooms_phonebathroom	0.01
rooms_choicepillow	0.01
rooms_size	< 0.01
rooms_electricshaver	< 0.01

Table 5.4: Ten least frequent room facts.

Looking at the least frequent hotel room attributes in Table 5.4, one can see that most of them are special and uncommon services and attributes like choice of towel, welcome gift, and choice of pillow. Surprisingly, room size is a rare fact in the GIATA data set.

Figure 5.3 shows the frequencies of spa offers among hotels. Well known spa offers like sauna, whirlpool, or massage are more frequent than special ones like ayurveda, thalasso or acupuncture.

In Figure 5.4 frequencies of sports offers in hotels are depicted. One can see that about 40% of hotels are offering a gym (fitness center), which is the most frequent sports offer. All other kind of sports are offered in about 10% or fewer cases.

Finally, Figure 5.5 shows how many hotels are considered as a particular type of hotel. Note, that a hotel can be assigned to more than one type, for example it can be a family friendly city hotel. Also note, that only 28% of hotels have been assigned to at least one hotel type, i.e. in 72% of the cases hotel type is missing. The most frequent type is city hotel, where 12% of the hotels are explicitly assigned to.

5.2 Missing Data Analysis

Methods and concepts applied or mentioned in this section are already introduced in Sections 3.2.1 and 3.3.1. Overall, the GIATA data set (i.e., the generated table) has about 42 million empty cells and 4.6 million non-empty cells. Thus, the sparsity is about 90%. Following columns do not have an entry at all: *facilities_congressfacilities, facilities_translator, misc_minimumguestage, rooms_electricshaver, type_cyclistshotel, type_villa.* These columns are dropped from the data set. Only 93 columns have a completeness level of greater than 10%, i.e. a missingness lower than 90%. Furthermore, only

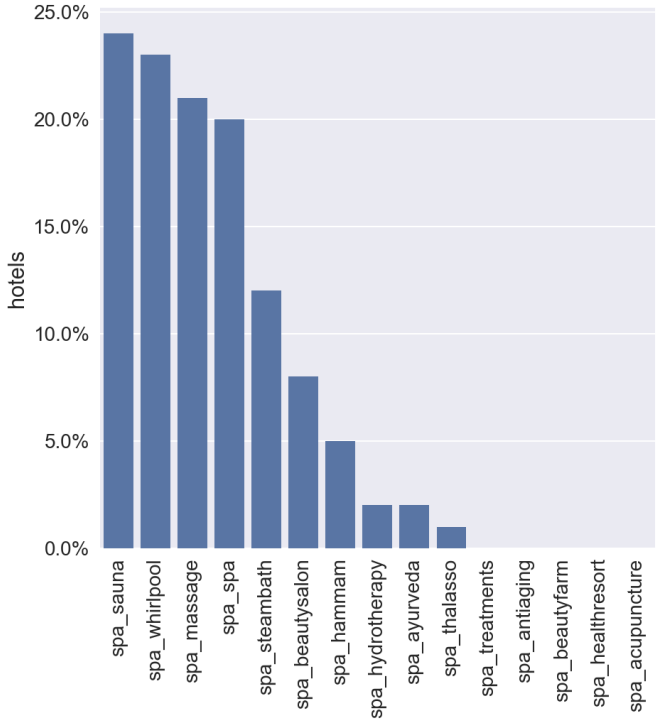


Figure 5.3: Frequencies of different spa offers among hotels.

following 14 columns have a completeness rate greater than 50%: *category_official, category_recommended, facilities_internetaccess, facilities_wifi, meals_breakfast, rooms_aircon, rooms_bathroom, rooms_doublebed, rooms_hairdryer, rooms_internetaccess, rooms_phone, rooms_shower, rooms_tv, rooms_wifi*. Next, missingness within different fact groups are examined in detail.

The total number of rooms, the number of floors in the main building, and the year of construction have a missingness of 75-80% and all other facts in the building information fact group have a much greater missing value rate. Overall, in 74% of the cases there is no building information at all. The official category (in stars) is given in the most cases with a missingness below 10%. Whereas, the recommended category (i.e. suggested category of the offering travel agency) has a missingness of about 50% , which is more than for most facts in this sparse data set. 6% of the hotels do not have any category specification. Facts of fact groups distance and entertainment are really rare and show missing value rates of about 90-100%. For 87% of the hotels there is no distance specification given and also in 82% of the cases there is no entertainment facts at all. Overall, 5% of the hotels do not have any facility specification. Most frequent facility facts are internet access and Wi-Fi (missingness of 24% and 29%) followed by carpark, restaurant, reception, and bar (missingness of 50-60%). Altogether, these facts can be seen as common attributes of hotel facilities, but still they possess higher missingness than one would expect. Also,

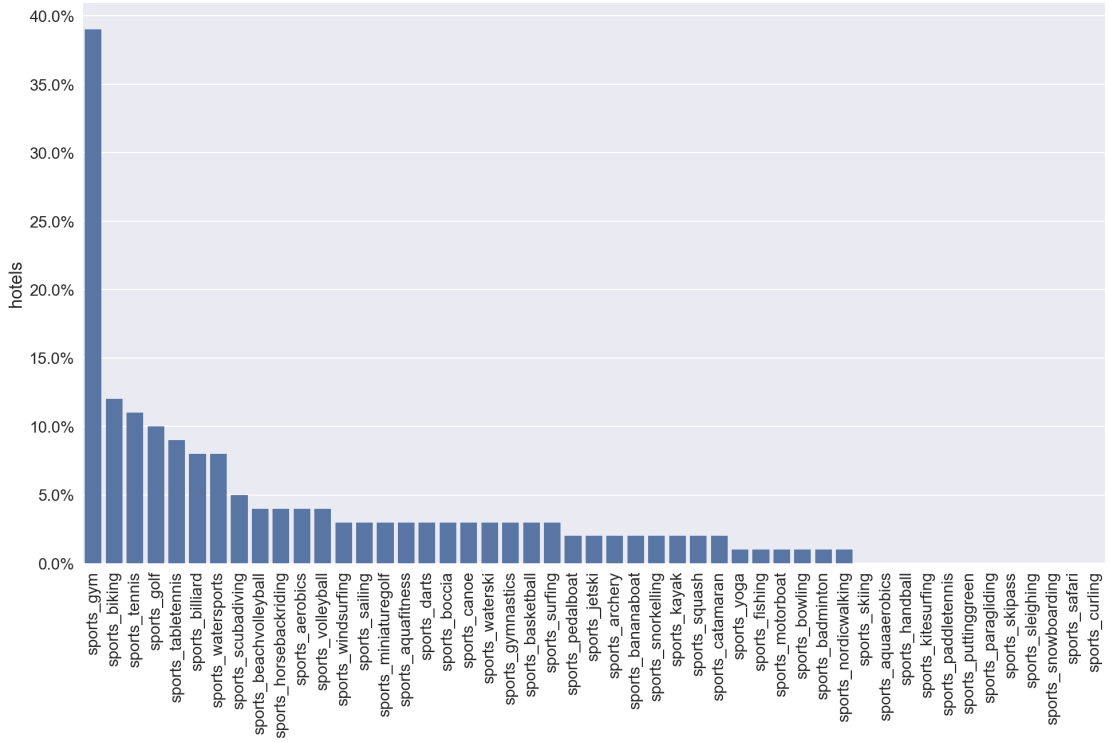


Figure 5.4: Frequencies of different sport offers among hotels.

facts of fact group location have high missing value rates of about 67-95%. Since those facts are binary, missingness is defined as 100% minus the frequency of a fact expressed in percentage. Thus, one can say that 24% of hotels are located centrally, 33% at the beach, 7% on main road, and 5% quietly. In 51% of the cases location facts are missing totally. Furthermore, 25% of hotel do not have any meal information. In 25% of the cases information about breakfast is missing, which means most hotels (75%) have a breakfast offer. Breakfast, lunch, and dinner (the main meals) are the most frequent facts within this factgroup. Only 3% of the hotels do not have any room fact information at all. Further, usual room facts like bathroom, TV, internet access, air conditioning, Wi-Fi, hairdryer, shower, phone, double bed, and safe have lower missing data rates (about 20-50%) than all other facts within the fact group rooms. Frequencies of facts of fact groups spa, sports, and type have already been discussed earlier. Since their missingness is defined as 100% - frequency in percentages, there is no need for repetition. In 58% of the cases there is no information about spa offers at all like in 47% of the cases for sports and 72% of the cases for hotel types.

The sparsity of the data can also be observed in the nullity matrix in Figure 5.6. The nullity matrix helps to visually discover patterns in missingness. Since there are about 300 facts and more than 140K hotels, column names (facts) and row ids (*giataId*) are omitted. The figure should only reveal possible patterns, which can then be analyzed in

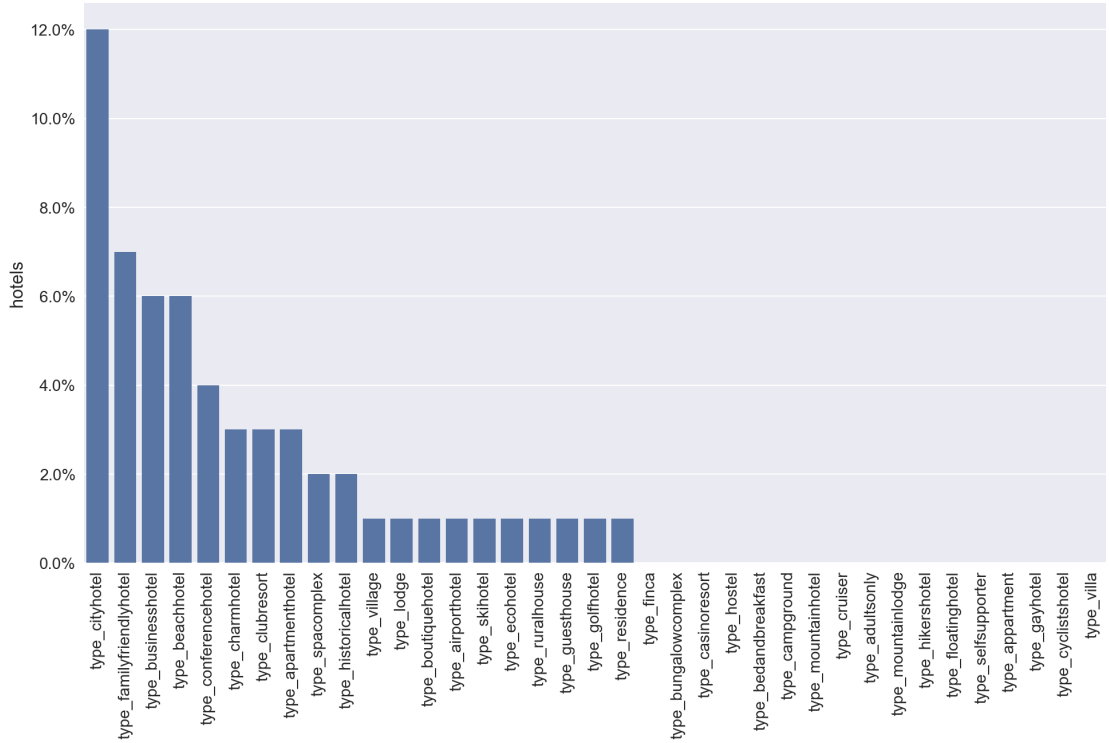


Figure 5.5: Frequencies of different hotel types.

Figure 5.6: Nullity matrix and completeness of hotel facts in the GIATA data set. Dark means presence and white absence.

more detail. Each black dot represents a non-missing fact of a hotel, whereas missing facts are depicted as empty-cells. Hotels are ordered by their completeness level of information. On the right-hand side on can see that completeness is ranging from zero (there is no information at all) to 160 facts. On average hotels have 32 facts and the median lies at 22 facts. Furthermore, one can see that some facts are more complete than others, but an actual pattern is not observable.

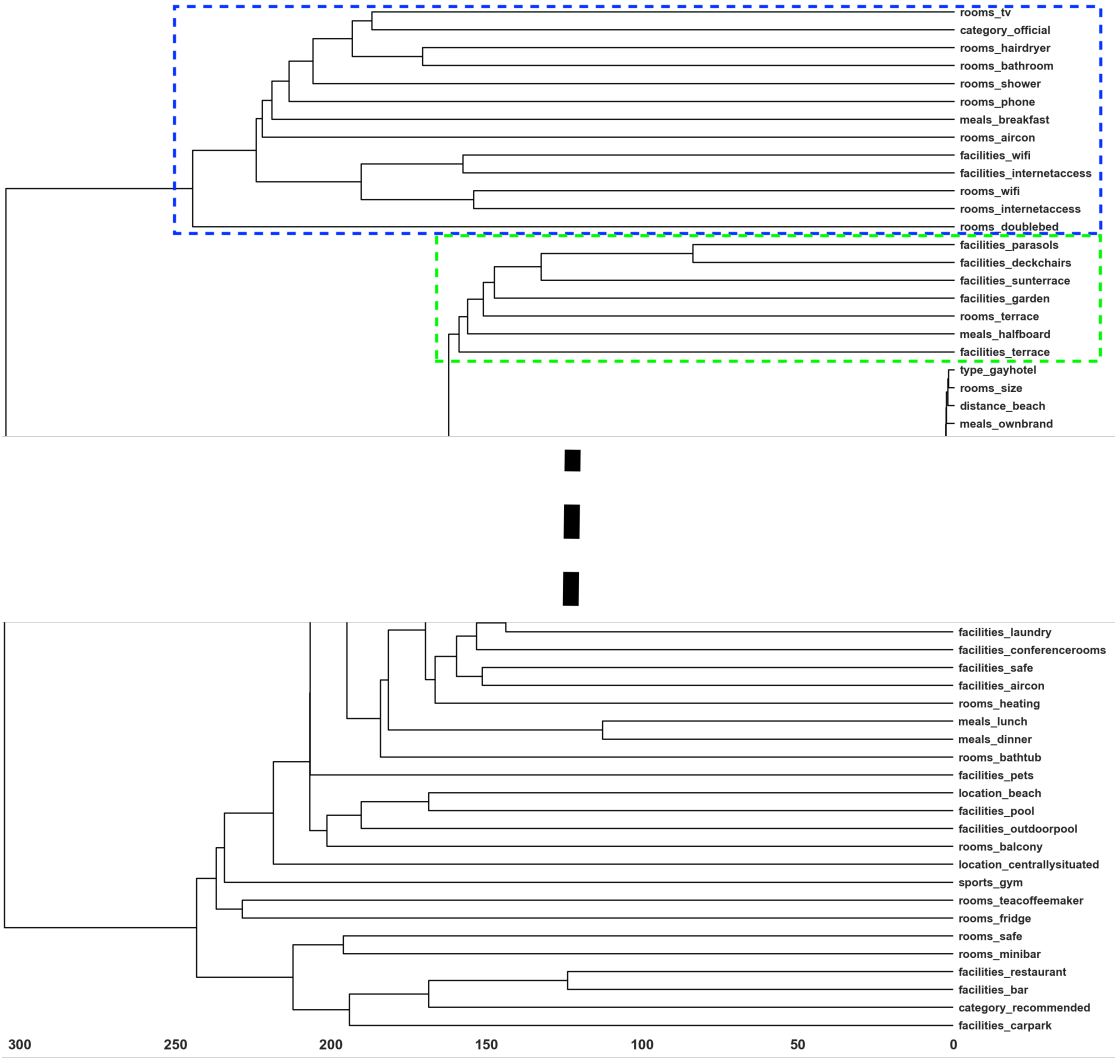


Figure 5.7: Dendrogram of nullity correlation of hotel facts in the GIATA data set.

In order to reveal deeper relations and hidden structures a dendrogram of nullity correlations is developed and displayed in Figure 5.7. The figure shows just parts of the actual dendrogram since lack of space for about 300 facts and the full grown tree. Overall, there are just two meaningful groups identified, which are marked with blue and green dashed rectangles. The first cluster (blue dashed rectangle) is just grouping facts with high

completeness level (low missingness). Whereas, the second highlighted cluster (green dashed rectangle) is interestingly grouping indicators of sun, such as *facilities_parasols, facilities_deckcharis, facilities_sunterrace, facilities_garden, room_terrace,* and *facilities_terrace*. Besides that the facts in the GIATA data set seems to be missing at random (MAR).

5.3 Missing Data Treatment and Feature Engineering

Again, methods and concepts applied or mentioned in this section are already introduced in Sections 3.2.1 and 3.3.1. Previous sections are discussing the shape of the GIATA data, its values and missingness. GIATA Facts are primarily encoded as binary values, showing the presence of a hotel attribute. Only some (distances, building information, and category) have different data types. With a sparsity of about 90%, features in the used data set are overall rare.

To recap, the missing data strategy in the Webologen data set is defined as following:

1. Features with similar meaning are combined
2. Features considered as experimental are deleted
3. Destinations with a completeness rate lower than a threshold are discarded
4. Missing data in geographical attributes are imputed by zero. Since geographical attributes are limited by the geographical nature of a destination and are rather specific features an imputation by zero is reasonable.
5. The rest of the missing data, namely the motivational ratings, are imputed with three different methods (by zero, KNN, and SOFT-IMPUTE).

Compared to the Webologen data set, the GIATA data set is much greater in dimensionality (in both features and observations). Also, it is sparser than the Webologen data set. The means of non-missing binary features (almost all facts) are exactly one, since the absence of a hotel feature is not denoted with zero, but by missingness. Thus, a KNN imputation strategy will only lead to one's since there is no other value to learn from. Matrix completion methods, mainly based on matrix factorization, got really popular thanks to the Netflix competition. Similar to the Netflix competition, where missing user ratings had to be completed, SOFT-IMPUTE is used in the destination data set in order to complete only the missing motivational ratings (geographical attributes have been naively imputed before). SOFT-IMPUTE is not optimized for binary data.

Still, in this work SOFT-IMPUTE was applied and assessed as a first approach for missing value treatment of the GIATA data set. As expected it led to a implausible and unreasonable matrix completion. For example, a lodge/bungalow complex in Slagharen (Netherlands) was predicted as a historical, conference & business hotel in the city.

Another example is that a small spa hotel near a thermal spring in Sinsheim (Germany) was completed as a historical city hotel with a casino in its facilities. Note, unlike the destination data where ratings have been estimated, hard facts are predicted here.

Since KNN imputation and SOFT-IMPUTE are no options, the missing data strategy for destinations cannot be adopted or applied here. Thus, a more suitable missing data strategy is developed and consists of the steps *delete non-relevant features, discard observations, naive imputation* and *feature generation*. These steps are explained in the following and the strategy is displayed in Figure 5.8.

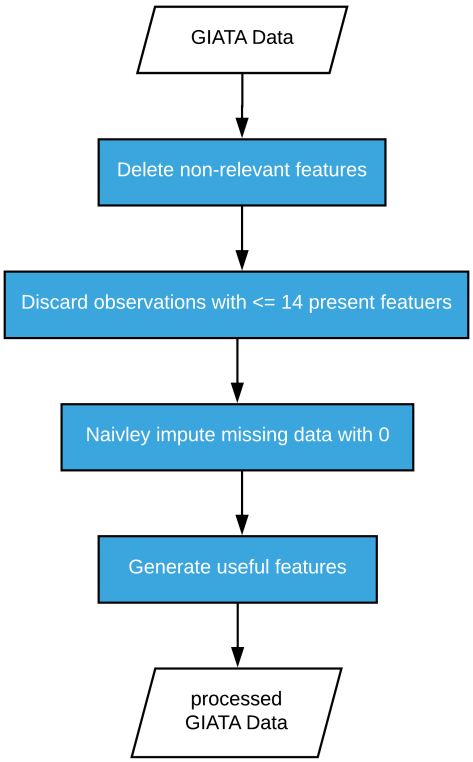


Figure 5.8: Missing data strategy of GIATA data set.

Delete non-relevant features. Relevancy corresponds here to the importance in model building. For example, facts of fact group object information (e.g. phone number, fax, email, address) are not needed to build up a model, but relevant to better identify a hotel. Therefore, fact groups object information, misc, meals, and payment will be ignored during model building.

Discard observations. As already encountered, only 14 hotel features have a missingness lower than 50% and the completeness level of information for each hotel ranges from 0 facts (no information at all) to 160 facts. Thus, in this step hotels with lower or equal to 14 facts are discarded.

Naive Imputation. Missing Values in all fact groups are naively imputed with zero.

Feature Generation. Most machine learning algorithms are not able to consider interactions or meaningful aggregations of features directly. Hence, such interactions or aggregations have to be provided either by brute force (e.g. testing all possible interactions) or with the help of domain knowledge. In order to enhance and support the model building following features are generated:

nightlife_index in an interval of [0,10]. This feature is defined as the sum of all night life indicators of a hotel and is meant to sum up and give a score for night life. Since distances are mostly provided if something is nearby, e.g. if a night club is nearby, they are considered as binary in the equation, i.e. 1 if present else 0.

$$nightlife\_index = distance\_barspubs + distance\_nightclubs + entertainment\_adults + facilities\_bar + facilities\_casino + facilities\_disco + facilities\_nightclub + facilities\_pub + type\_adultsonly + type\_casinoresort$$

nature_index in an interval of [0,18]. This feature is defined as the sum of all facts of a hotel, which are indicators for nature. Again distances are considered as binary.

$$nature\_index = distance\_crosscountryskiing + distance\_forest + distance\_golfcourse + distance\_lake + distance\_park + distance\_skiarea + distance\_skilift + location\_quietlysituated + type\_campground + type\_ecohotel + type\_finca + type\_hikershotel + type\_lodge + type\_mountainhotel + type\_mountainlodge + type\_ruralhouse + tpye\_skihotel + type\_spacomplex$$

family_index in an interval of [0,10]. This feature is summing up all facts, which can be accounted as indicators of family and child friendliness.

$$family\_index = coentertainment\_childcare + entertainment\_children + entertainment\_miniclub + entertainment\_minidisco + facilities\_babysitter + facilities\_childrenspool + facilities\_playground + facilities\_playroom + rooms\_childrensbed + rooms\_videogames$$

sun_index in an interval of [0,12]. This feature is summing up all facts, which are indicators of sun. Again distances are considered as binary.

$$sun\_index = distance\_beach + distance\_sea + facilities\_childrenspool + facilities\_deckchairs + facilities\_outdoorpool + facilities\_parasols + facilities\_pool + facilities\_poolbar + facilities\_sunterrace + facilities\_waterslide + location\_beach + rooms\_seaview$$

spa_count in an interval of [0,15]. This feature is just a simple count of all spa offers, i.e. count of all facts of fact group spa a hotel possesses.

sports_count in an interval of [0,50]. This feature is just a simple count of all sport offers, i.e. count of all facts of fact group sport a hotel possesses.

watersports_count in an interval of [0,17]. This feature is just a simple count of all water sports offers.

$watersports\_count = aquaaerobics + aquafitness + bananaboat + canoe + catamaran + fishing + jetski + kayak + kitesurfing + motorboat + pedalboat + sailing + scubadiving + snorkelling + surfing + waterski + windsurfing$

wintersports_count in an interval of [0,5]. This feature is just a simple count of all winter offers.

$wintersports\_count = nordicwalking + skiing + skipass + snowboarding + sleighing$

recreationalsports_count in an interval of [0,8]. This feature is just a simple count of all recreational sports offers.

$recreationalsports\_count = archery + biking + canoe + fishing + golf + horsebackriding + nordicwarlking + yoga$

actionsports_count in an interval of [0,10]. This feature is just a simple count of all action sports offers. $actionsports\_count = bananaboat + jetski + kitesurfing + motorboat + paragliding + safari + snowboarding + surfing + waterski + windsurfing$

5.4 Bivariate analysis of hotel features

In order to analyze correlations among hotel features and to reveal latent, underlying relations a clustered correlation heat map is generated and displayed in Figure 5.9. Note in addition to the features defined by GIATA the analysis also considers the newly generated features: *nightlife_index, nature_index, family_index, sun_index, spa_count, sports_count, watersports_count, wintersports_count, recreationalsports_count,* and *actionsports_count*.

Unfortunately, there is no space to display all hotel feature labels since the shown figure contains about 300 features. The clustered correlation heat map's intent is to provide an overview rather than to show each pairwise correlation. Hotel features are mostly uncorrelated or slightly positively correlated. Still, there are some meaningful clusters of hotel features, which are highlighted with colored dashed rectangles:

Red cluster. The newly generated features *family_index, sun_index, sports_count, recreationalsports_count,* and *watersports_count* and their corresponding features are within this cluster. Also, *type_clubresort* and *type_beachhotel* are members of this group. Hence, one can conclude that this cluster is grouping features of family friendly and recreational beach resorts with many sports opportunities, especially water sports.

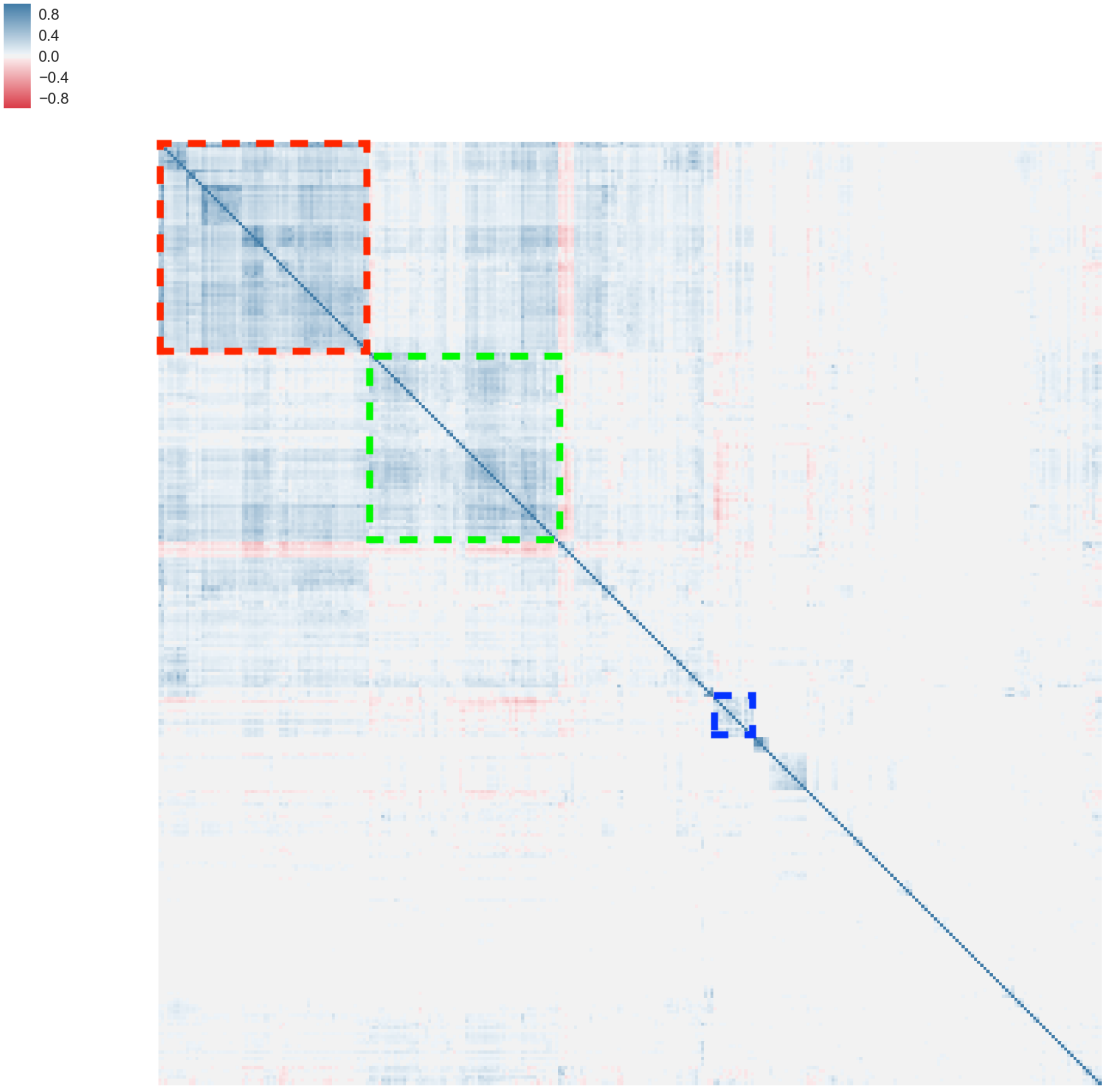


Figure 5.9: Clustered correlation heat map of hotel features.

Green cluster. Hotel types within this cluster are *type_cityhotel, type_businesshotel, type_conferencehotel,* and *type_histroricalhotel*. Also, *location_onmainroad* and *location_centrallysituated* are members of this cluster. One generated feature, namely *nightlife_index*, and some of its corresponding features like *facilities_bar* or *facilities_pub* are also located here. Thus, one can say that this cluster is grouping in contrast to the red cluster features of hotels mainly located in the city.

Blue cluster. This cluster is grouping following features: *buildinginformation_apartments, buildinginformation_numstudios, rooms_fridge, rooms_microwave, rooms_sofabed, rooms_dishwasher, rooms_cooker, rooms_kitchen, rooms_bedroom, rooms_livingroom, rooms_kitchenette, rooms_washingmachine,* and *type_apartmenthotel*. Considering all these features one can say that the blue cluster is grouping attributes of accommodations appropriate for self-supporters. Additionally, a *selfsupporter_index* is added as a new feature, which is summing up all features within this cluster.

5.5 The Data Sample

In addition to the huge archive of XML-files, two Austrian e-Tourism companies have provided labeled samples of hotels. The first sample is again provided by Pixtri [OG]. Experts of Pixtri have mapped 400 randomly chosen hotels to the Seven-Factors, simply by assigning a score for each factor, like previously in the destinations sample. The majority hotels in this sample are located in Italy, Spain, USA, Germany, France, and Greece (51,48%), which is similar to the distribution of the whole data set. Also, this sample is sparse (84%) like the data set itself (90%). Most hotels, similar to the whole data set, are 3-4 stars hotels build in the 20th century and have on average 121 rooms in total. Hence, the sample can be considered as representative.

The second sample is provided by Eurotours [Gmba], another Austrian e-Tourism company. Experts of Eurotours have mapped 620 hotels to the Seven-Factors, by assigning a score for each factor. In contrast to the first sample hotels of this sample were not taken from the whole data set, but from their own product spectrum. Eurotours main focus is Europe, in particular Austria, Germany and Italy. This can also be observed in the provided sample, where the vast majority of hotels are located in Austria, Germany, and Italy (80.57%). Yet, the missingness in data (85%) is similar to the sparsity in the first sample and the whole data set. Also here, most hotels are built in the 20th century, have three to four stars and on average 114 rooms in total. Thus, there can be a bias towards European hotels, but sparsity, size, age, and category are similar to corresponding attributes of hotels in the data set.

Next, the distributions of assigned scores are examined. Figure 5.10 shows how experts of Pixtri assigned scores for each of the Seven-Factors. In all factors, except *Independence & Travel*, the most frequent score is 0. The score distribution in factors *Knowledge & Travel, Culture & Indulgence,* and *Action & Fun* are very similar, where many hotels are scoring with 0 (33-36%) or 0.5 (21-26%). In factor *Sun & Chill-Out* the majority of destinations

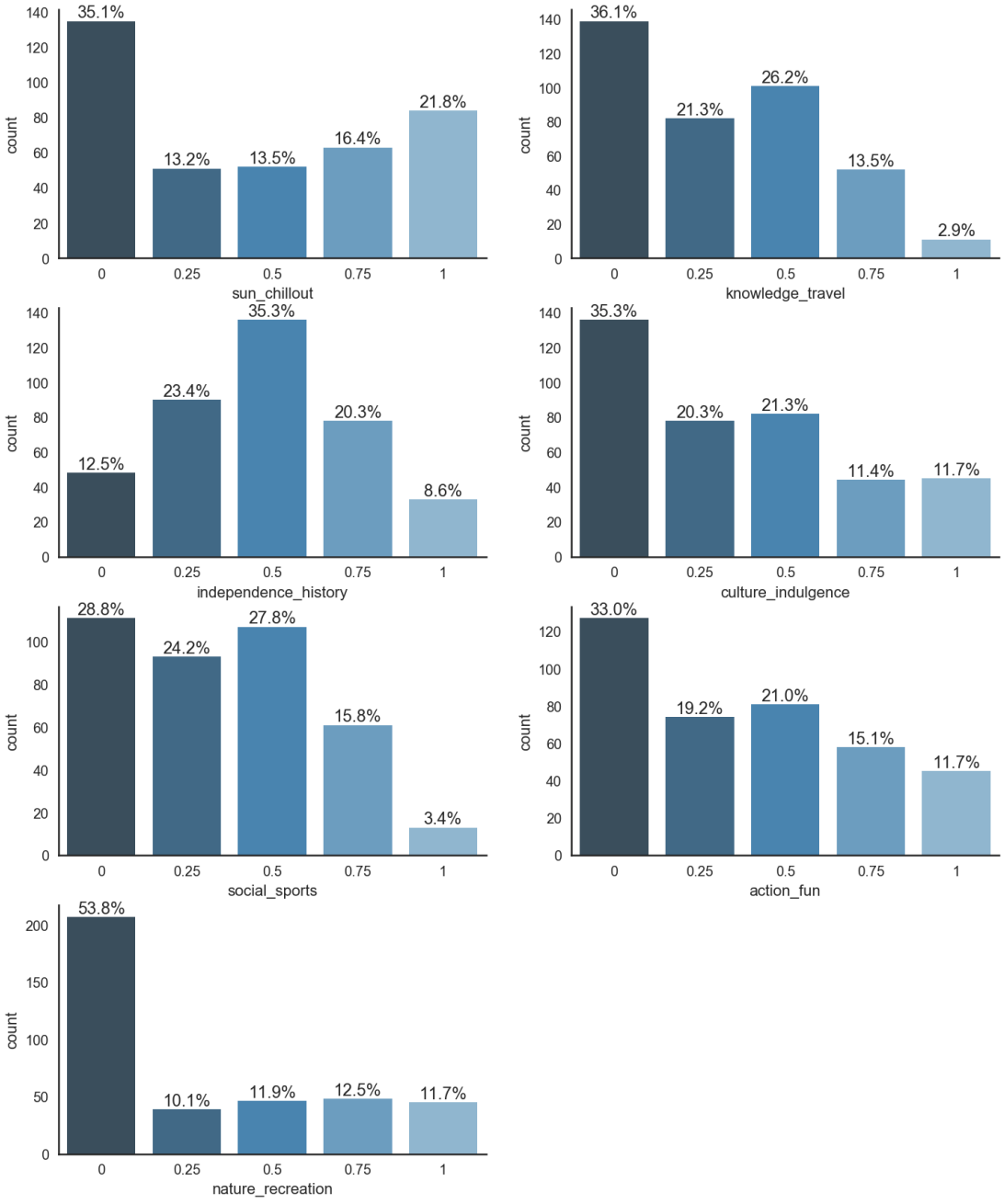


Figure 5.10: Distribution of Seven-Factor scores in the Pixtri sample.

scores with 0 (35%) or 1 (22%). Interestingly, score 1 is given more frequently in factor *Sun & Chill-Out* than in all other factors. Only 3.4% of hotels score with 1 in factor *Social & Sports*. Whereas, the vast majority of hotels in this factor score with 0.5 or lower (81%). Most hotels in factor *Nature & Recreation* score with 0 (54%), where all other scores have frequencies of about 10-12%. *Independence & History* is the only factor, where the distribution of scores has more or less a normal shape.

Similar to Figure 5.10, but for the Eurotours sample, Figure 5.11 shows the distribution of assigned scores in each factor. In factors *Sun & Chill-Out, Knowledge & Travel, Culture & Indulgence, Action & Fun,* and *Nature & Recreation* the most frequent score is 0. Particularly, in factors *Knowledge & Travel* and *Culture & Indulgence* this is more extreme, where 71-82% of hotels scores with 0. *Sun & Chill-Out and Nature & Recreation* have similar score distributions. *Independence & History* is the only factor, where the majority of hotels are scoring with 0.25. Interestingly, the majority of hotels in factor *Social & Sports* score either with 0.25 (22%) or with 0.75 (31%)

Furthermore, the correlations between the Seven-Factors and hotel features are analyzed. In particular, correlation coefficients between each factor and its most correlated hotel features are calculated and displayed as a heat map. The first entry of the heat map is always the factor itself, followed by the most correlated hotel features listed in an ordered way (descending absolute value of correlation coefficient).

In Figure 5.12 the correlations of the factor *Sun & Chill-Out* are depicted. Considering the Pixtri sample (Figure 5.12a), the most correlated feature is the generated feature *sun_index*. As expected, they are positively correlated. All other listed hotel features are related with *sun_index* (indicators of sun) and are also positively correlated to *Sun & Chill-Out*. A similar picture can be observed for the Eurotours sample in Figure 5.12b, where *location_beach* is the most correlated feature. Again, all other features can be considered as indicators of sun. Interestingly, the newly generated feature *watersports_count* and corresponding features are also listed here. Overall, the obvious positive relation between sun, beach, water sports, and *Sun & Chill-Out* can be observed.

Figure 5.13 shows the correlations of the factor *Knowledge & Travel*. In the Pixtri sample (Figure 5.13a) one can see that all listed features are negatively correlated to the factor, except *type_cityhotel*. Almost all negatively correlated hotel features are indicators of sun and beach like *sun_index or location_beach*. This in accordance to the *Knowledge & Travel* description, which states that being lazy and lying at the beach is negatively related to the factor. Surprisingly, the generated feature *family_index* is also negatively related to the factor, which can be a sign of the logistic difficulties of traveling and gaining knowledge with the family. Like in the Pixtri sample, also in the Eurotours sample *room_balcony and rooms_terrace* is negatively correlated to the factor. On the other hand, all other features are positively correlated with *Knowledge & Travel*. Within the positively correlated features are *type_conferencehotel, type_cityhotel, and type_businesshotel*. Obviously, it is easier to gain knowledge in the city, especially in conference and business hotels, which are meant to offer this. Factor *Knowledge & Travel* is also described as organized mass tourist. Thus, a positive correlation to the size of a

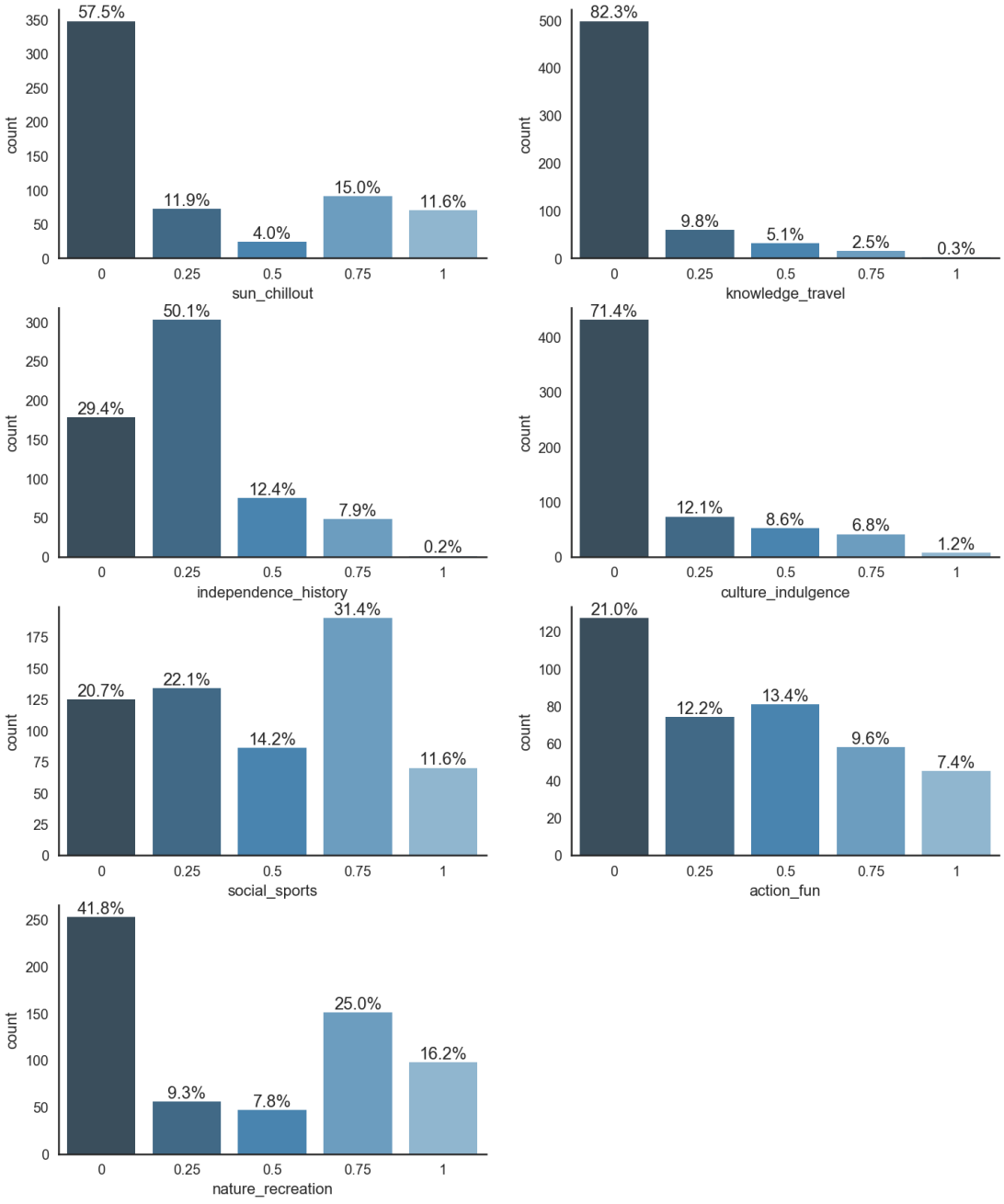
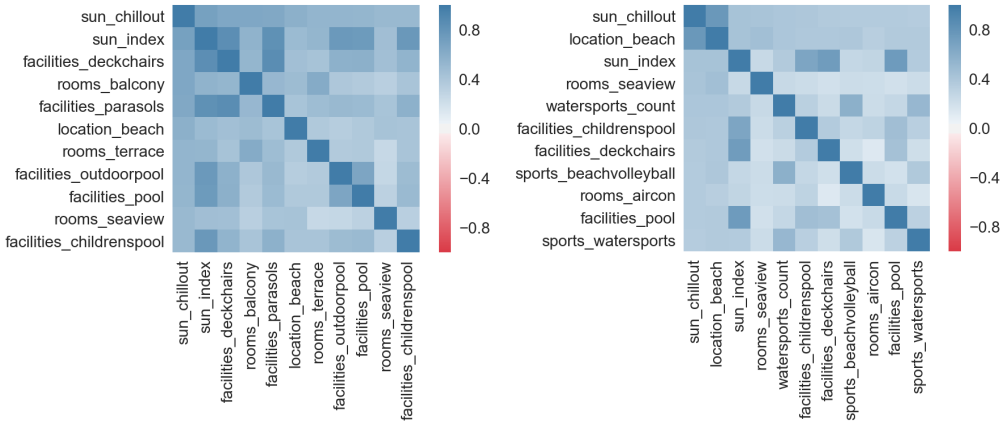
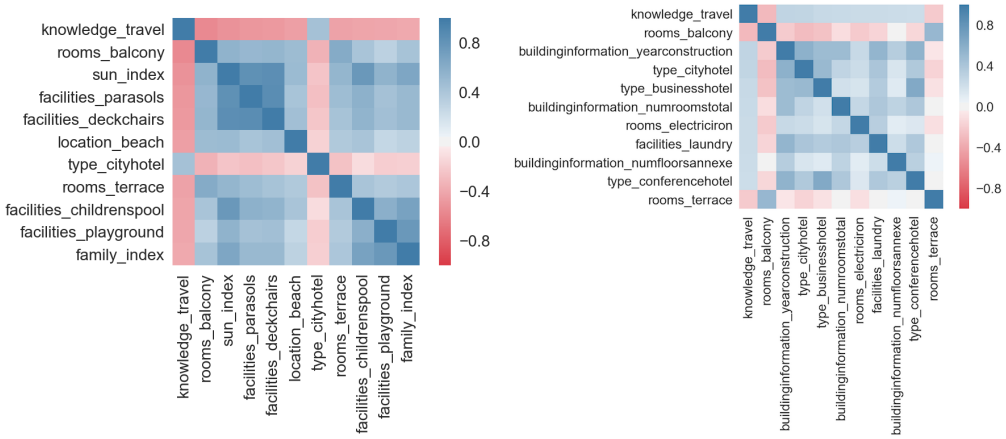


Figure 5.11: Distribution of Seven-Factor scores in the Eurotours sample.



(a) Pixtri (b) Eurotours

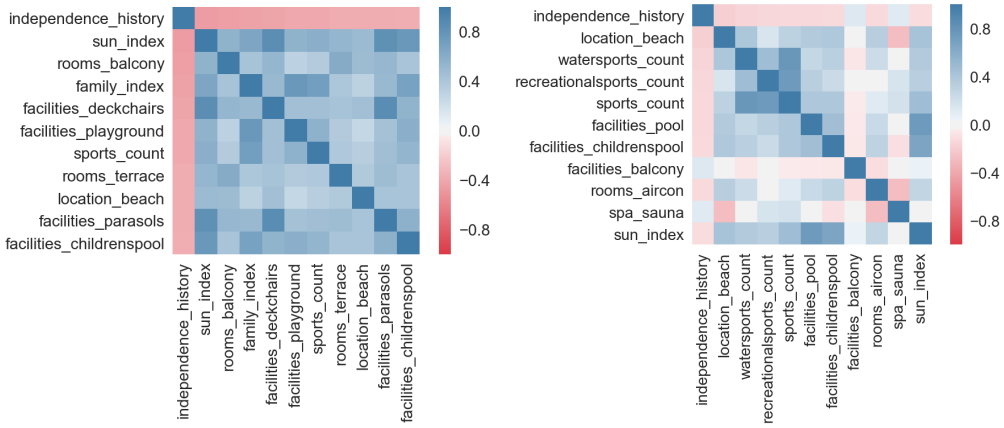
Figure 5.12: Heat map of most correlated hotel features of the factor *Sun & Chill-Out*. Note, that the first element of the heat map is the factor itself, followed by its most correlated features.



(a) Pixtri (b) Eurotours

Figure 5.13: Heat map of most correlated hotel features of the factor *Knowledge & Travel*. Note, that the first element of the heat map is the factor itself, followed by its most correlated features.

hotel (*buildinginformation_numroomstotal, buildinginformation_numroomsannexe*) is reasonable. Overall, the contrast between city hotels and beach resorts is inherently given in the data and their relationship with factor *Knowledge & Travel* is clearly observable.



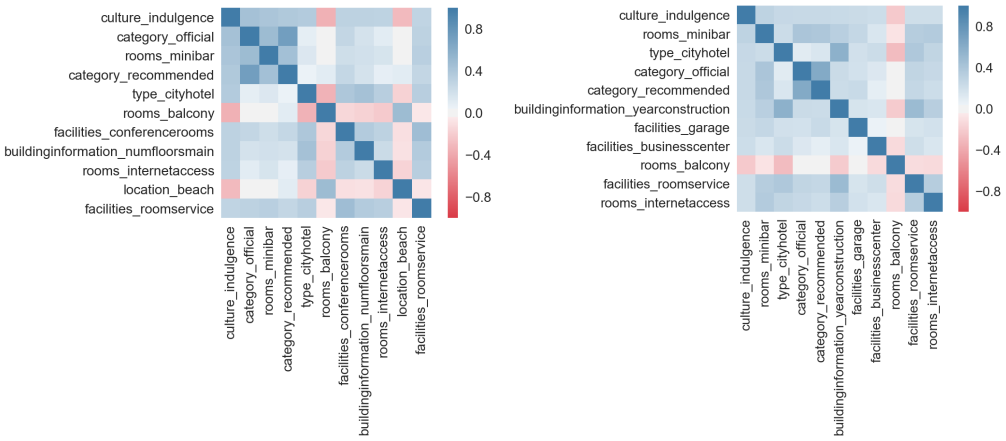
(a) Pixtri (b) Eurotours

Figure 5.14: Heat map of most correlated hotel features of the factor *Independence & History*. Note, that the first element of the heat map is the factor itself, followed by its most correlated features.

Correlations of the factor *Independence & History* are displayed in Figure 5.14. This factor is characterized as a mass tourist, similar to *Knowledge & Travel*, but in contrast as an independent one (not organized). *Independence & History* is positively related with traditions, cultural and historical activities. Thus, negative correlations with hotel features related to sunny beach resorts, observable in both samples (Figure 5.14a and 5.14b), are reasonable since it is more likely to find such activities in the city. Additionally, one can see in the corresponding heat map of the Eurotours sample, that sports, especially recreational- and water sports, are negatively related to *Independence & History*. Again, those kinds of sports are more likely to be found outside of the city. Finally, also this factor is negatively correlated with features *family_index*, if the Pixtri sample is considered. Again, this could be a sign for logistic difficulties of traveling with children.

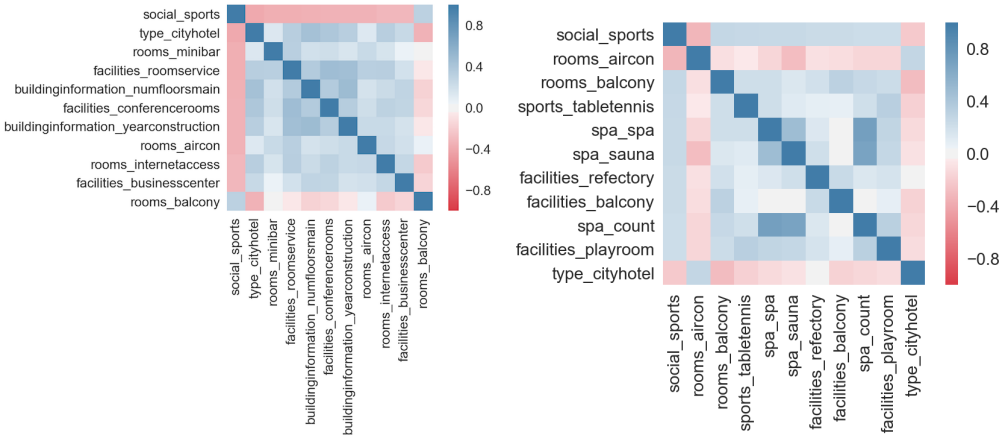
In Figure 5.15 the correlations of the factor *Culture & Indulgence* are shown. This factor is shortly described as culture loving, high class tourist. Thus, a positive correlation with hotel category (in stars) is expected and can be observed in both samples, Pixtri (Figure 5.15a) and Eurotours (Figure 5.15b). Also here, a contrast between features of accommodations in the city and beach resorts is observable, where the first ones are positively correlated with the factor and the second ones negatively.

Figure 5.16 depicts the correlations of the factor *Social & Sports*. Unfortunately, there are almost no plausible correlations in terms of interpretability found. Just a few conclusions can be drawn for both samples. One trait of factor *Social & Sports* is crowd avoidance. Thus, considering the Pixtri sample (Figure 5.16a), negative correlations with features



(a) Pixtri (b) Eurotours

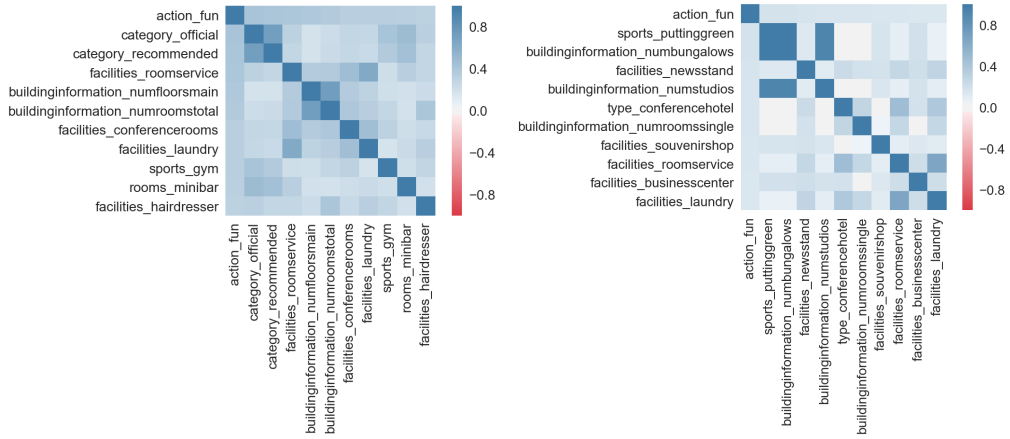
Figure 5.15: Heat map of most correlated hotel features of the factor *Culture & Indulgence*. Note, that the first element of the heat map is the factor itself, followed by its most correlated features.



(a) Pixtri (b) Eurotours

Figure 5.16: Heat map of most correlated hotel features of the factor *Social & Sports*. Note, that the first element of the heat map is the factor itself, followed by its most correlated features.

type_cityhotel, buildinginformation_numfloorsmain, and *facilities_businesscentre*, i.e. possible indicators of big, crowded hotels, are reasonable. On the other hand, considering the Eurotours sample (Figure 5.16b), a positive correlation with *spa_count* (plus corresponding features) is reasonable, since health resorts are mainly located in non-crowded areas.



(a) Pixtri (b) Eurotours

Figure 5.17: Heat map of most correlated hotel features of the factor *Action & Fun*. Note, that the first element of the heat map is the factor itself, followed by its most correlated features.

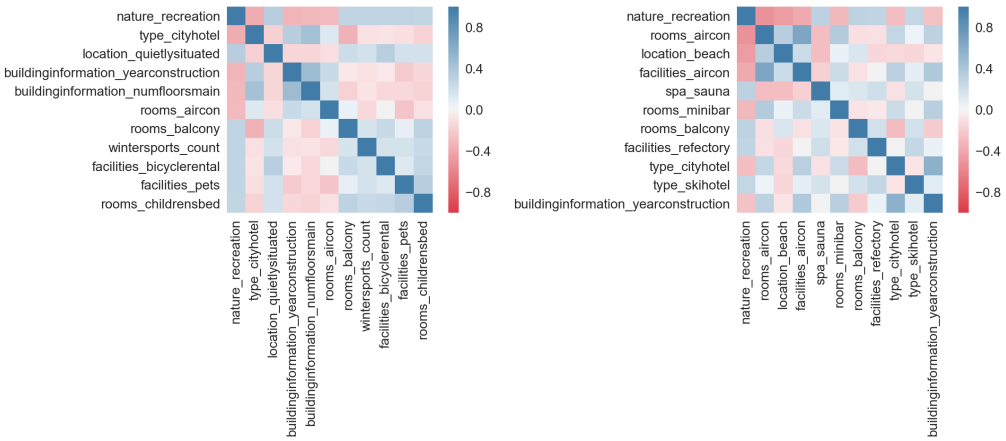
Correlations of the factor *Action & Fun* are displayed in Figure 5.17. This factor is characterized as a high class jet-setter who likes action, fun, and crowded places. Hence, considering the Pixtri sample (Figure 5.17a), positive correlations with hotel category and hotel size indicators (number of main floors or total rooms) are reasonable. In the Eurotours sample (Figure 5.17b) there is a positive correlation with hotel size indicators (number of bungalows, studios, and single rooms) because of the same reason as before in the Pixtri sample. Overall, no further conclusions can be drawn.

Finally, Figure 5.18 displays correlation of the factor *Nature & Recreation*. Considering the Pixtri sample (Figure 5.18a), the most correlated feature with this factor is *type_cityhotel* followed by *location_quietlysituated*, where the first one is negatively correlated and the second positively. Since factor *Nature & Recreation* is characterized as escapist (escaping from everyday life) such correlations are reasonable. Additionally, the number of main floors can be seen as an indicator for the size of a hotel and in turn crowdedness. Hence, a negative correlation with the factor is plausible. Furthermore, *wintersports_count* is positively correlated with *Nature & Recreation*. Since winter sports areas are mainly in the idyllic nature such a positive correlation is also reasonable. Looking at the Eurotours sample (Figure 5.18b), the contrast between crowded areas and quietly situated locations is given by the features *type_cityhotel and type_skihotel*.

Although both expert samples are sampled differently they have many things in common.

Factor	Expert	Most correlated hotel features
Sun & Chill-Out	Pixtri	*sun_index, deckchairs, rooms_balcony, parasols, beach, terrace, outdoorpool, pool, seaview, childrenspool*
	Eurotours	*beach, sun_index, seaview, watersports_count, childrenspool, deckchairs, beachvolleyball, rooms_aircon, pool, watersports*
Knowledge & Travel	Pixtri	*- rooms_balcony, - sun_index, - parasols, - deckchairs, - beach, cityhotel, - terrace, - childrenspool, - playground, - family_index*
	Eurotours	*- rooms_balcony, yearconstruction, cityhotel, businesshotel, numroomstotal, electriciron, laundry, numfloorsanexe, conferencehotel, - terrace*
Independence & History	Pixtri	*- sun_index, - rooms_balcony, - family_index, - deckchairs, - playground, - sports_count, - terrace, - beach, - parasols, - childrenspool*
	Eurotours	*- beach, - watersports_count, - recreationalsports_count, - sports_count, - pool, - childrenspool, facilities_balcony, - rooms_aircon, sauna, - sun_index*
Culture & Indulgence	Pixtri	*category_official, minibar, category_recommended, cityhotel, - rooms_balcony, conferencerooms, numfloorsmain, internetaccess, - beach, roomservice*
	Eurotours	*minibar, cityhotel, category_official, category_recommended, yearconstruction, garage, businesscentre, - rooms_balcony, roomservice, internetaccess*

Table 5.5: Most correlated hotel features of the factors *Sun & Chill-Out*, *Knowledge & Travel*, *Independence & History*, and *Culture & Indulgence*.



(a) Pixtri (b) Eurotours

Figure 5.18: Heat map of most correlated hotel features of the factor *Nature & Recreation*. Note, that the first element of the heat map is the factor itself, followed by its most correlated features.

In Tables 5.5 and 5.6 most correlated hotel features of the Seven-Factors are summarized at one glance. Note, that a preceding minus sign indicates a negative correlation.

Taking into account all listed hotel features in Tables 5.5 and 5.6 one can say that in 41% of the cases the Pixtri sample and the Eurotours sample are delivering exactly the same top correlated features. Note, that there are also similar features, for example: *ski_hotel* and *wintersports_count* are indicators of winter sports or *numberoomstotal* and *numfloorsmain* can be considered for the size of a hotel.

Furthermore, destinations in both expert samples are also similar in hotel size, construction year, category, and missingness. Also, the expert ratings have all in all similar trends. Since both expert samples (Pixtri and Eurotours) behave similar in the conducted analysis and due to the high dimensionality of the given data set (about 300 features) it makes more sense to combine both for further analysis and model building (bigger test and training set) rather than to process them separately.

To sum up, one can say that the correlation analysis delivered the best results for factor *Sun & Chill-Out*. Also, the most correlated features of factors *Culture & Indulgence*, *Action & Fun*, and *Nature & Recreation* are quite reasonable and easy to interpret. Furthermore, the analysis covers only few characteristic aspects of the factors *Knowledge & Travel*, *Independence & History*, and *Nature & Recreation*.

Factor	Expert	Most correlated hotel features
Social & Sports	Pixtri	*- cityhotel, - minibar, - roomservice, - numfloorsmain, - conferencerooms, - yearconstruction, - rooms_aircon, - internetaccess, - businesscentre, rooms_balcony*
	Eurotours	*- rooms_aircon, rooms_balcony, tabletennis, spa, sauna, refectory, facilities_balcony, spa_count, playroom, - cityhotel*
Action & Fun	Pixtri	*category_official, category_recommended, roomservice, numfloorsmain, numroomstotal, conferencerooms, laundry, gym, minibar, hairdresser*
	Eurotours	*puttinggreen, numbungalows, newsstand, numstudios, conferencehotel, numroomssingle, sourvenirshop, roomservice, businesscentre, laundry*
Nature & Recreation	Pixtri	*- cityhotel, quietlysituated, - yearconstruction, - numfloorsmain, - rooms_aircon, rooms_balcony, wintersports_count, bicyclerental, pets, childrensbed*
	Eurotours	*- rooms_aircon, beach, - facilities_aircon, spa, - minibar, rooms_balcony, refectory, - cityhotel, skihotel, - yearconstruction*

Table 5.6: Most correlated hotel features of the factors *Social & Sports*, *Action & Fun*, and *Nature & Recreation*.

CHAPTER 6

Mapping of Hotels to the Seven-Factors

This chapter is the equivalent of Chapter 4, but with respect to hotels. Again, unsupervised and supervised learning methods are used in order to get more insights and to enable an automated mapping onto the Seven-Factors.

6.1 Cluster Analysis

In order to get more insights into the hotel data set and to develop a better understanding of similarities among hotels, based on their features, a cluster analysis is conducted. Also, identifying meaning full groups may contribute to a better generalization in the matter of mapping the hotels onto the Seven-Factors. For the cluster analysis 10,000 randomly chosen destinations (incl. destinations of the expert samples) are used. The same approach, like previously in the destination data set, is followed. Thus, PAM is used as clustering method. Again, the Gower distance is chosen as an appropriate dissimilarity metric, because also the GIATA data set contains mixed datatypes. Finally, the silhouette width is used for determining the right cluster size and as evaluation metric.

Figure 6.1 shows the average silhouette width within different cluster sizes. Based on the silhouette width two, three, and six cluster solutions are considered, but for the sake of interpretability a six-cluster solution is chosen. Next, the resulting clusters are examined in detail. The number of destinations in each cluster is provided at the beginning of each paragraph.

C1 (N=2779). The medoid of C1 is "Holiday Inn Express Fairfax", a small hotel in Fairfax (Virgina, USA). Fairfax is considered as a suburb of Washington, D.C. The hotel has no real eye-catching features and so are the hotels of the whole cluster.

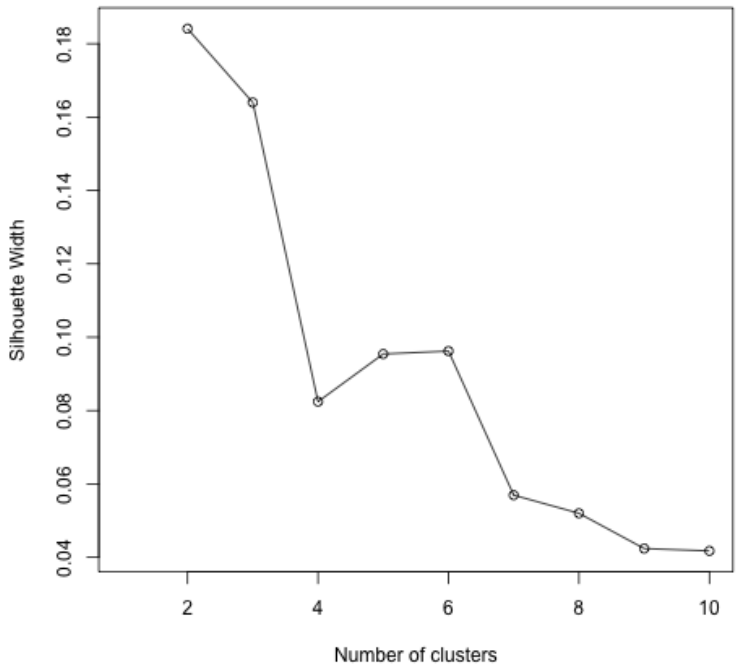


Figure 6.1: Scree plot to determine an appropriate number of clusters.

Only the newly generated feature *selfsupporter_index* has the highest mean value (i.e. 2.1) among all clusters. For none of the hotels within this cluster a distance measure to beach or forest is provided. Further, compared to the other five clusters the mean of the newly created feature *nature_index* is at lowest within this cluster. Thus, hotels of C1 can be considered as ordinary hotels without a special focus, possibly appropriate for self-supporters and backpackers and not located at the beach or nature.

C2 (N=846). The medoid of C2 is "Grünerbaum Hotels" in Bressanone (South Tyrol, Italy). The hotel has 4 buildings, 180 rooms, and a wellness & recreation area of over 5000m^2. Within C2 95% of hotels are offering a sauna, 89% massages, and 88% have a wellness area. The average number of spa offers in C2 is five, which is the highest average compared to the other clusters. Also, the *nature_index* shows the highest mean in C2 considering all other clusters. To conclude, hotels of C2 are most likely located near nature and are appropriate for wellness and recreational tourism.

C3 (N=1442). The medoid of C3 is "Hotel Vanilla", a small (20 rooms) boutique hotel in Fethiye (Turkey). The hotel is located near the beach and in 2km distance to the

city center. It has an outdoor pool, but besides that there is no other eye-catching feature. In C3 81% of the hotels are located at the beach. Hotels in C3 have the 2nd highest average *sun_index* compared to hotels of all other clusters. All other features have nothing extraordinary about them. Thus, hotels of C3 can be considered as small, not crowded beach hotels with not much additional offers.

C4(N=2152). The medoid of C4 is "Park Grand London Heathrow". The hotel is located in Hounslow (Great Britain) and as the name already says it is near the Heathrow airport in West London. One can say that the hotel is located in the suburbs of London. Hotels in C4 have no extraordinary properties, similar to the ones in C5. In general, C4 and C1 are very similar. The only difference is that, in contrast to C1 the *selfsupporter_index* is the lowest here, and in addition much more hotels in C4 are explicitly tagged as city or business hotel. Thus, hotels of C4 can be considered as ordinary hotels without a real special focus, not appropriate for self-supporters or backpackers, but possibly good for business travelers.

C5 (N=2099). The medoid of C5 is "Hotel Santa Clara", which is located in the historical center of the city of Evora, the capital of the Alentejo region in Portugal. The most relevant and important feature of the hotel is its central location. In C5 53% of the hotels are considered as centrally located, which is the highest amount compared to the other clusters. Furthermore, 67% of the hotels are explicitly tagged as city hotel and 30% as business hotel, both values are at its highest in C5 compared to all other clusters. Thus, hotels of C5 are mainly centrally situated city or business hotels, most appropriate for city or business trips.

C6 (N=682). The medoid of C6 is "Olympic Palace Resort Hotel & Convention Center" on the island of Rhodes (Greece). It is a five stars beach resort with 371 rooms, a private beach, in-and outdoor pools, a great wellness area, many sports offers and much more. Such richness of activities, opportunities and features can also be observed throughout the cluster. The average count of sports offers is 14, where five out of these are water sports offers. The average amount of spa offers within C6 is five. Further, the average value of the newly generated *sun_index* is 6.9, the one of *family_index* is 4.9, and *nightlife_index* has a mean of 2.5. Furthermore, 50% of the hotels are explicitly tagged as beach hotel and 25% as club resort and family friendly hotel. In C6 90% of the hotels are located at the beach. Hotels of C6 have on average a higher rating in stars (4 stars on average) than hotels of the other clusters. To conclude, hotels of cluster C6 can be considered as big, family friendly, and high class beach resorts, but also appropriate for nightlife, and with plenty opportunities for wellness and sports.

Generally, it can be said that there is a latent natural structure of the data. Hence, six conceptually meaningful clusters are encountered. For an easier understanding and communication, those clusters can be summarized and simplified as follows: C1 – *basic self-supporter hotel*, C2 – *recreational wellness hotel*, C3 – *lovely beach hotel*, C4 – *simple city hotel*, C5 – *centrally situated city hotel*, C6 – *high class beach resort*.

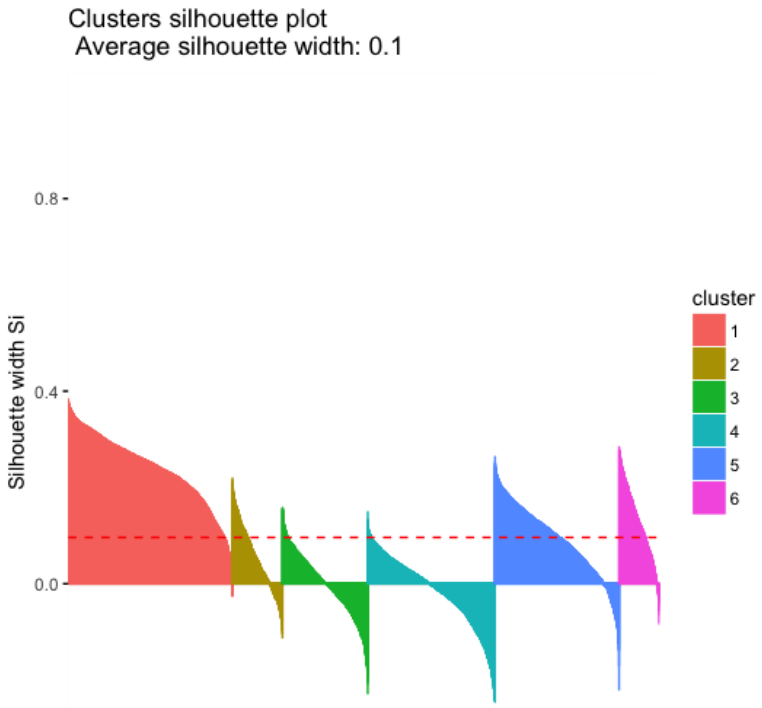


Figure 6.2: Silhouette plot of the 6 cluster solution.

Although some plausible clusters are identified, the resulting clustering has a pretty low average silhouette width, which means that the cohesion of hotels within clusters, respectively the separation of clusters, is weak. This can also be observed in Figure 6.2. Clusters C1 and C6 are relatively robust, but there are many mis assignments (negative silhouette coefficient) in the other clusters, especially in C3 and C4. Hence, the resulting cluster solution should be used carefully.

The conducted cluster analysis considers 10,000 randomly chosen destinations including destinations of the combined expert sample (Pixtri and Eurotours, see Section 5.5). Therefore, the resulting cluster solution can be further assessed by examining the factor score distribution of each factor of the Seven-Factor Model over the six clusters. In Table 6.1 average factor scores and corresponding standard deviations (SD) in different clusters are listed.

The average factor score distribution of the Seven-Factors in different clusters is not clearly interpretable as in the destinations case. Only for the factor *Sun & Chill-Out* a clear interpretation can be made at one glance. The factor *Sun & Chill-Out* scores the best in C6 – *high class beach resort* with 0.81 and in C3 – *lovely beach hotel* with 0.66, which does not need further explanation. Still, all other factors show meaningful, interpretable trends that are not as pronounced as in the factor *Sun & Chill-Out*. Factors

		C1	C2	C3	C4	C5	C6
Sun & Chill-Out	mean	0.30	0.30	0.66	0.18	0.12	0.81
	SD	0.39	0.35	0.35	0.30	0.24	0.25
Knowledge & Travel	mean	0.20	0.06	0.06	0.20	0.44	0.12
	SD	0.22	0.14	0.14	0.28	0.28	0.22
Independence & History	mean	0.48	0.25	0.26	0.35	0.48	0.23
	SD	0.33	0.22	0.23	0.27	0.30	0.23
Culture & Indulgence	mean	0.27	0.16	0.10	0.20	0.46	0.26
	SD	0.33	0.27	0.21	0.29	0.35	0.28
Social & Sports	mean	0.43	0.57	0.45	0.44	0.20	0.34
	SD	0.35	0.31	0.27	0.33	0.26	0.24
Action & Fun	mean	0.17	0.07	0.14	0.13	0.36	0.44
	SD	0.26	0.21	0.25	0.23	0.36	0.35
Nature & Recreation	mean	0.41	0.56	0.30	0.43	0.07	0.25
	SD	0.36	0.36	0.35	0.41	0.19	0.31

Table 6.1: Average factor scores plus standard deviations in different clusters. Note, that both expert samples (Pixtri and Eurotours) are considered.

Knowledge & Travel , *Independence & History*, and *Culture & Indulgence* show their highest mean score in C5 – *centrally situated city hotel*. Since there are more offers for cultural and historical activities, and for gaining knowledge in hotels located in the center of a city, especially in the historical center like the medoid of C5, this observation is plausible. The factor *Social & Sports* has its highest mean score in C2 – *recreational wellness hotels* and it lowest mean score in C5 – *centrally situated city hotel*. Here one can observe clearly that the factor *Social & Sports* avoids crowded areas, which totally fits to its description. Whereas, it is known that the factor *Action & Fun* likes crowded areas, nightlife, thrill, and exclusiveness. All this can also be observed in the factor score distribution, where the factor *Action & Fun* scores the best in C5 - *centrally situated city* and C6 – *high class beach resort*. Finally, the factor *Nature & Recreation* scores the best, as already expected and quite reasonable, in C2 – *recreational wellness hotel*. While it scores the worst in C5 – *centrally situated city hotel*, which is also quite obvious for a nature loving recreational traveler.

6.2 Regression Analysis

For the analyses in this section the combined expert sample (combination of Pixtri and Eurotours expert samples, see Section 5.5) is used. Like previously in Section 4.2 for the destination data, a multiple linear regression analysis is conducted in order to find association between hotel features and the Seven-Factors and to enable an automated mapping onto them. Again, a stepwise variable selection method is used to identify the most decisive features and to avoid an overfitting. The resulting multiple linear regression

models, one for each of the Seven-Factors, are evaluated against a baseline function f_0, which is simply the average score of each factor in the combined expert sample. Model performance is again assed via R^2 and $RMSE$. Also here, the performance of multiple linear regression model (MLR) is compared to the performance of K-Nearest-Neighbor regression (KNN) and Random Forest regression (RF). Finally, the Seven-Factors for 10,000 randomly chosen hotels are predicted and the resulting factor score distributions are compared to the factor score distributions in the expert mapping.

6.2.1 Resulting Models

Sun & Chill-Out

	Coefficient	Std. Error	Signif.
(Intercept)	0.0326	0.0183	
location_beach	0.3318	0.0249	***
sun_index	0.0334	0.0034	***
rooms_seaview	0.1186	0.0306	***
buildinginformation_numroomstotal	-0.0002	0.0001	***
rooms_aircon	0.1154	0.0197	***
type_cityhotel	-0.1098	0.0272	***
facilities_bicyclerental	0.0697	0.0178	***
type_skihotel	-0.1389	0.0379	***
facilities_garage	-0.0501	0.0184	**

Table 6.2: Multiple linear regression model for the factor *Sun & Chill-Out*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 6.2 the coefficients of the multiple linear regression model for the factor *Sun & Chill-Out* are listed. Obviously, sun and beach vacation related hotel features, namely *sun_index, location_beach, rooms_seaview,* and *rooms_aircon*, have a significant and positive impact on the factor *Sun & Chill-Out*. The more rooms a hotel possesses the more visitors it can bear and the less the chill-out factor will get, which can be observed in the negative sign of the feature *buildinginformation_numroomstotal*. Further, *type_skihotel* (cold weather) and *type_cityhotel* (less relaxation) are also significantly negatively related to the factor.

Knowledge & Travel

In Table 6.3 the coefficients of the multiple linear regression model for the factor *Knowledge & Travel* are listed. The features *type_cityhotel* and *tpye_businesshotel*, which indicate that a hotel is located in the city, are significantly positively related with the factor. Whereas, the features *rooms_childrensbed, rooms_bedroom, rooms_balcony, facilities_terrace, facilities_playground,* and *spa_massage* have a significant negative impact on the factor. Those positive and negative relations are reasonable since hotels located in

	Coefficient	Std. Error	Signif.
(Intercept)	0.3116	0.0195	***
rooms_balcony	-0.1210	0.0180	***
type_cityhotel	0.1112	0.0261	***
rooms_childrensbed	-0.0492	0.0159	**
rooms_internetaccess	0.0573	0.0158	***
facilities_playground	-0.0467	0.0166	**
buildinginformation_numroomstotal	0.0002	0.0001	***
spa_massage	-0.0417	0.0157	**
tpye_businesshotel	0.1086	0.0307	***
facilities_terrace	-0.0527	0.0164	**
rooms_bedroom	-0.0553	0.0194	**

Table 6.3: Multiple linear regression model for the factor *Knowledge & Travel*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

the city are more appropriate for gaining knowledge than beach resorts, wellness hotels, or self-supporter accommodations and due to logistic difficulties in traveling with children. Since the factor *Knowledge & Travel* is also characterized as a mass tourist a significant positive relation with feature *buildinginformation_numroomstotal* is not surprising.

Independence & History

	Coefficient	Std. Error	Signif.
(Intercept)	0.49	0.02	***
family_index	-0.02	0.01	**
type_cityhotel	0.08	0.03	**
recreationalsports_count	-0.03	0.01	***
rooms_internetaccess	0.07	0.02	***
tpye_businesshotel	0.07	0.04	***
facilities_pets	-0.07	0.02	***
location_beach	-0.09	0.02	***

Table 6.4: Multiple linear regression model for the factor *Independence & History*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 6.4 the coefficients of the multiple linear regression model for the factor *Independence & History* are listed. The factor *Independence & History* is similar to the factor *Knowledge & Travel*, except the first one is characterized as an independent mass tourist and the last one as an organized mass tourist (tours etc.). Thus, a significant positive impact of features *type_cityhotel* and *tpye_businesshotel* is reasonable since hotels located in the city or centrally are more appropriate for sightseeing and history loving travelers.

On the other side, *location_beach* and *family_index* are negatively related to the factor. Beach resorts usually have another focus than history and sightseeing and the logistic difficulties of travelers with such focus and children have already been mentioned. Hotels with a nature and recreation focus have a relatively higher *recreationalsports_count* and higher possibility of *facilities_pets* (pets are allowed) than the rest of the hotels in the data set. But those hotels are commonly not appropriate for sightseeing and history travelers. Thus, a negative impact of the features *recreationalsports_count* and *facilities_pets* is reasonable.

Culture & Indulgence

	Coefficient	Std. Error	Signif.
(Intercept)	-0.0443	0.0444	
type_cityhotel	0.1587	0.0276	***
category_official	0.0621	0.0152	***
rooms_dvdplayer	0.0205	0.0468	***
rooms_minibar	0.0783	0.0200	***
facilities_pets	-0.0970	0.0179	***
facilities_businesscentre	0.1372	0.0344	***
type_charmhotel	0.1809	0.0457	***
sun_index	-0.0272	0.0037	***
rooms_queensizebed	0.1895	0.0607	**
buildinginformation_numjuniorsuites	0.0003	0.0001	**

Table 6.5: Multiple linear regression model for the factor *Culture & Indulgence*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 6.5 the coefficients of the multiple linear regression model for the factor *Culture & Indulgence* are listed. The factor *Culture & Indulgence* is also characterized as a history and archeology loving high class tourist with gourmet characteristics. Thus, significant positive relations to signs of high class and quality (*category_official, rooms_dvdplayer, rooms_minibar, facilities_businesscentre, buildinginformation_numjuniorsuites, rooms_queensizebed*) and city location (*type_cityhotel*) are reasonable. Also, a significant positive impact of *type_charmhotel* is plausible, since most charm hotels in the data set are high class, repurposed chateaus, villas, and country houses. The feature *facilities_pets* has a significant negative impact on the factor, because of the same reason as as for the factor *Independence & History*. Also, the generated feature *sun_index* is significantly negatively related to the factor, which is reasonable since the higher the *sun_index* the more beach resort characteristics a hotel has and the less is the focus on history and sightseeing.

	Coefficient	Std. Error	Signif.
(Intercept)	0.53	0.05	***
facilities_shuttleservice	-0.13	0.03	***
facilities_pets	0.08	0.02	***
sports_tabletennis	0.09	0.02	***
category_official	-0.06	0.01	***
type_skihotel	0.12	0.04	**
type_cityhotel	-0.10	0.03	**
location_beach	-0.07	0.02	**
facilities_businesscentre	-0.12	0.03	**
spa_spa	0.08	0.02	***
facilities_hairdresser	-0.09	0.03	**

Table 6.6: Multiple linear regression model for the factor *Social & Sports*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

Social & Sports

In Table 6.6 the coefficients of the multiple linear regression model for the factor *Social & Sports* are listed. Hotel features *facilities_shuttleservice, category_official, type_cityhotel, location_beach, facilities_businesscentre,* and *facilities_hairdresser* are typical indicators of high class hotels, beach resorts or city hotels. Since the factor *Social & Sports* is also known for its avoidance of exclusiveness and crowds a significant negative relation with the listed factors is reasonable. On the other side, it is significantly positively related to features *type_skihotel, sports_tabletennis, spa_spa*, which is also plausible since they can be seen as indicators of sports, nature and uncrowdedness.

Aciton & Fun

In Table 6.7 the coefficients of the multiple linear regression model for the factor *Action & Fun* are listed. The hotel type casino resort (*type_casinoresort*) has a significant and highly positive impact on this factor, which is obvious. Also, the number of floors in the main building (*buildinginformation_numfloorsmain*) is significantly positively related to the factor and since the factor *Action & Fun* is known for its passion for crowdedness and partying, this is reasonable. Whereas, the features *nature_index, facilities_bycylerental, facilities_pets,* and *type_village* can be considered as indicators of peacefulness. Hence, their significant negative impact on the factor makes sense. The factor *Action & Fun* is also known for its exclusive taste. Thus, significant, positive relations with indicators of high class and quality, such as *category_official, facilities_businesscentre, facilities_medicalattendance, facilities_roomservice,* and *rooms_minifridge* are reasonable. Finally, the significant positive impact of the feature *sports_catamaran* can be explained by the action and thrill-seeking nature of the factor.

	Coefficient	Std. Error	Signif.
(Intercept)	-0.04	0.04	
buildinginformation_numfloorsmain	0.01	0.01	***
facilities_businesscentre	0.18	0.04	***
facilities_pets	-0.06	0.02	**
category_official	0.06	0.01	***
facilities_bicyclerental	-0.07	0.02	***
facilities_medicalattendance	0.09	0.03	**
nature_index	-0.05	0.01	***
sports_catamaran	0.22	0.06	***
type_village	-0.21	0.06	***
facilities_roomservice	0.07	0.02	**
rooms_minifridge	0.10	0.03	**
type_casinoresort	0.70	0.23	**

Table 6.7: Multiple linear regression model for the factor *Action & Fun*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

Nature & Recreation

	Coefficient	Std. Error	Signif.
(Intercept)	0.44	0.03	***
location_beach	-0.20	0.03	***
type_cityhotel	-0.12	0.03	***
type_skihotel	0.19	0.05	***
facilities_elevators	-0.13	0.02	***
location_quietlysituated	0.08	0.03	**
spa_sauna	0.08	0.02	***
buildinginformation_numapartments	-0.01	0.01	**
facilities_garden	0.06	0.02	**
facilities_fireplace	0.11	0.04	**
facilities_foyer	-0.07	0.03	**
sports_skiing	0.14	0.05	**

Table 6.8: Multiple linear regression model for the factor *Nature & Recreation*. *Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$).*

In Table 6.8 the coefficients of the multiple linear regression model for the factor *Nature & Recreation* are listed. Indicators of nature, peace, and recreation, such as features *location_quietlysituated, type_skihotel, spa_sauna, facilities_garden, facilities_fireplace,* and *sports_skiing*, are significantly positively related to this factor. On the other side, indicators of mass tourism and big city hotels, such as *type_cityhotel, location_beach,*

facilities_elevators, buildinginformation_numapartments, and *facilities_foyer*, are significantly negatively related. Those positive and negative relations with the factor *Nature & Recreation* are pretty obvious and do not need further explanation.

6.2.2 Evaluation

Like previously in the mapping of destinations (see Chapter 4), the evaluation metrics used here are R^2 and $RMSE$. The performances of the resulting multiple linear regression models are compared to an appropriate baseline function (f_0), a KNN regression (KNN), and a Random Forest regression (RF). As baseline the simple mean function is chosen again, i.e. the average factor score for each factor in the training set (see Table 6.9). Finally, the factor scores for 10,000 randomly chosen hotels are determined and the resulting distributions in factor scores are compared to the distributions in the expert mapping.

	mean
Sun & Chill-Out	0.35
Knowledge & Travel	0.17
Independence & History	0.33
Culture & Indulgence	0.22
Social & Sports	0.44
Action & Fun	0.18
Nature & Recreation	0.36

Table 6.9: Average factor scores of hotels in the training set (taken from the combined expert sample).

In Table 6.10 the training and test performance of the baseline function (f_0), the multiple linear regression (MLR), the KNN regression (KNN), and the Random Forest regression (RF) are listed. Since f_0 is a constant function it cannot explain any variance in the target variable, which explains the zero values of R^2_{train} and R^2_{test} in each factor.

Training and test performance of the MLR model are similar, which shows that MLR is overfitting the training data not that much. Whereas, KNN and RF are clearly overfitting the training set. For example, the KNN model for factor *Nature & Recreation* has a R^2_{train} of 1.00 and a $RMSE_{train}$ of 0, but for unseen data this is totally different, namely R^2_{test} is 0.27 and $RMSE_{test}$ is 0.33. Both models, KNN and RF, are thoroughly tuned via cross validated hyper parameter search, but still they are overfitting badly in some cases. This can be caused due to a lack in training data, but it also shows the potential of both models if more data is used.

All three models show a better performance than the simple baseline, i.e. all three models have a smaller $RMSE$ than f_0. Thus, the build models MLR, KNN, and RF are able to explain some variation in the data. As already mentioned, most important is the out of

		f_0	MLR	KNN	RF
Sun & Chill-Out	R^2_{train}	0.00	0.66	0.38	0.87
	R^2_{test}	0.00	0.59	0.23	0.60
	$RMSE_{train}$	0.39	0.23	0.31	0.14
	$RMSE_{test}$	0.38	0.24	0.33	0.24
Knowledge & Travel	R^2_{train}	0.00	0.44	0.27	0.78
	R^2_{test}	0.00	0.35	0.20	0.39
	$RMSE_{train}$	0.26	0.19	0.22	0.12
	$RMSE_{test}$	0.25	0.20	0.22	0.19
Independence & History	R^2_{train}	0.00	0.22	0.14	0.72
	R^2_{test}	0.00	0.13	0.00	0.11
	$RMSE_{train}$	0.27	0.24	0.25	0.14
	$RMSE_{test}$	0.27	0.26	0.27	0.25
Culture & Indulgence	R^2_{train}	0.00	0.44	0.23	0.78
	R^2_{test}	0.00	0.41	0.17	0.43
	$RMSE_{train}$	0.30	0.23	0.27	0.14
	$RMSE_{test}$	0.31	0.24	0.29	0.24
Social & Sports	R^2_{train}	0.00	0.36	1.00	0.77
	R^2_{test}	0.00	0.29	0.16	0.33
	$RMSE_{train}$	0.32	0.26	0.00	0.15
	$RMSE_{test}$	0.32	0.27	0.29	0.26
Action & Fun	R^2_{train}	0.00	0.38	0.22	0.76
	R^2_{test}	0.00	0.35	0.20	0.41
	$RMSE_{train}$	0.30	0.23	0.26	0.14
	$RMSE_{test}$	0.29	0.24	0.26	0.22
Nature & Recreation	R^2_{train}	0.00	0.49	1.00	0.82
	R^2_{test}	0.00	0.42	0.27	0.46
	$RMSE_{train}$	0.38	0.27	0.00	0.16
	$RMSE_{test}$	0.39	0.30	0.33	0.29

Table 6.10: Comparison of performance measures of baseline function (f_0), multiple linear regression (MLR), KNN regression (KNN), and Random Forest regression (RF) in test and training set.

sample performance since it is an approximation of the future performance of the model. Considering the out of sample performance (R^2_{test} and $RMSE_{test}$), the MLR model and RF model are performing similar, but the KNN model is clearly outperformed by them. Again, it makes sense to choose the MLR model over RF and KNN since it performs similar or better, it is simpler to fit, and easier to interpret.

The MLR model is performing the best in factor *Sun & Chill-Out*. Here, 59% of the variation in the unseen target can be explained and $RMSE_{test}$ is the lowest in comparison to the MLR models of the other factors. On the other hand, the MLR model performs the worst in factor *Independence & History*, where the performance is just slightly better than the baseline f_0.

The Seven-Factors for 10,000 randomly chosen hotels are determined, simply by applying the developed MLR models. The resulting factor score distributions are compared to the distributions in the expert mapping.

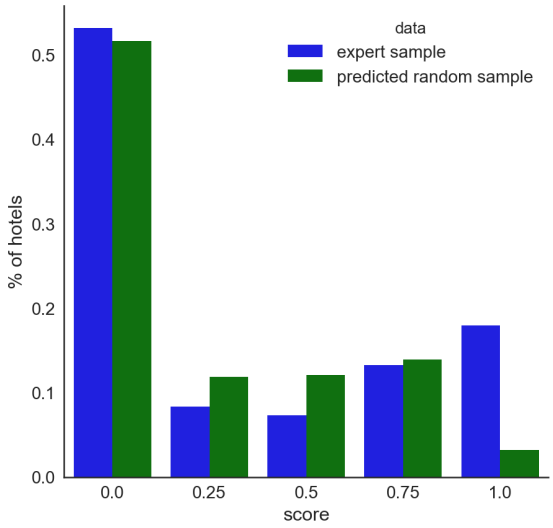


Figure 6.3: Comparison of the factor score distribution of the expert sample versus the predicted scores of 10,000 randomly chosen hotels for factor *Sun & Chill-Out*.

In Figure 6.3 the distributions of factor scores for the factor *Sun & Chill-Out* are compared. The predicted factors of the random sample and the manually labeled factors of the expert sample are overall similar distributed, except the proportion of hotels with score 1 is six times higher in the expert mapping compared to the predicted scores of the random sample.

Figure 6.4 shows the distributions of factor scores for the factor *Knowledge & Travel*. In both samples the vast majority of hotels have scores in the lower end of the scale. Further, in both samples similar proportions of hotels have a score of 0.5 and 1. Besides that, only 1% of the hotels in the random sample have a predicted score of 0.75, whereas experts labeled 7% of the hotels in their sample with 0.75.

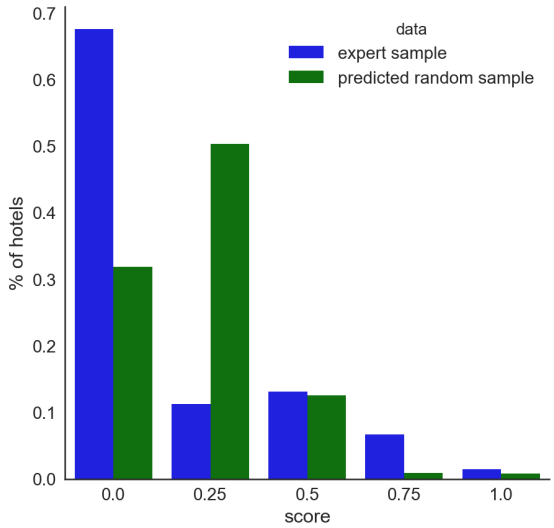


Figure 6.4: Comparison of the factor score distribution of the expert sample versus the predicted scores of 10,000 randomly chosen hotels for factor *Knowledge & Travel*.

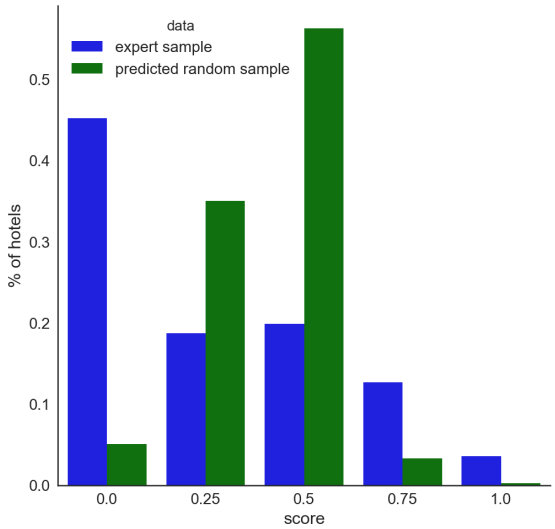


Figure 6.5: Comparison of the factor score distribution of the expert sample versus the predicted scores of 10,000 randomly chosen hotels for factor *Independence & History*.

The distributions of factor scores for factor *Independence & History* are displayed in Figure 6.5. Almost all hotels in the random sample, more precisely 91%, have a predicted score of 0.25 or 0.5. This is not the case in the manually mapped expert sample. Overall, both distributions seem to differ essentially in their behavior, where the distribution of factors scores in the random sample shows more or less a bell shape while the expert sample starts with a bigger amount and decays with increasing score.

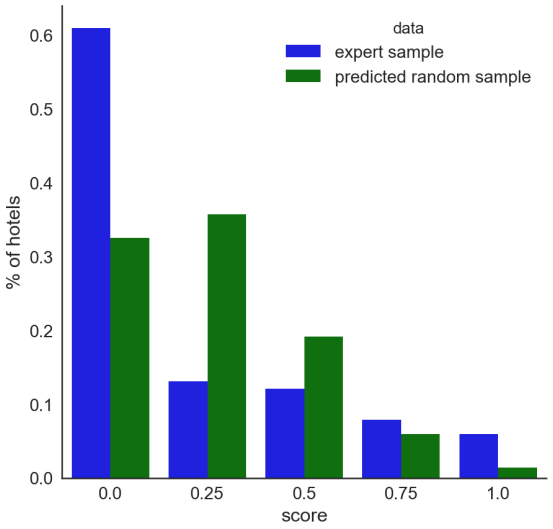


Figure 6.6: Comparison of the factor score distribution of the expert sample versus the predicted scores of 10,000 randomly chosen hotels for factor *Culture & Indulgence*.

In Figure 6.6 the distributions of factor scores for the factor *Culture & Indulgence* are compared. Again, in both samples most of the hotels are scoring with either 0 or 0.25. By comparing the proportion of hotels for each score one cannot identify any similarities. However, their behavior shows roughly similarities, where the majority of hotels score with 0 or 0.25 and the amount is decreasing with increasing score.

Figure 6.7 shows the distributions of factor scores for the factor *Social & Sports*. Here, 77% of the hotels in the random sample have a predicted score of 0.25 or 0.5. Overall there are not much similarities of both distributions, whether in pairwise comparison nor in behavior.

The distributions of factor scores for the factor *Action & Fun* are displayed in Figure 6.8. In both distributions the largest proportion of hotels has a score of 0. Further, both distributions show more or less a similar behavior by starting with a large proportion of hotels with 0 score and getting smaller with increasing score.

Finally, Figure 6.9 shows the distributions of factor scores for the factor *Nature & Recreation*. In both distributions the majority of hotels are located at the lower end of the scale with a score of 0 or 0.25. Besides that, there are no similarities. Setting aside

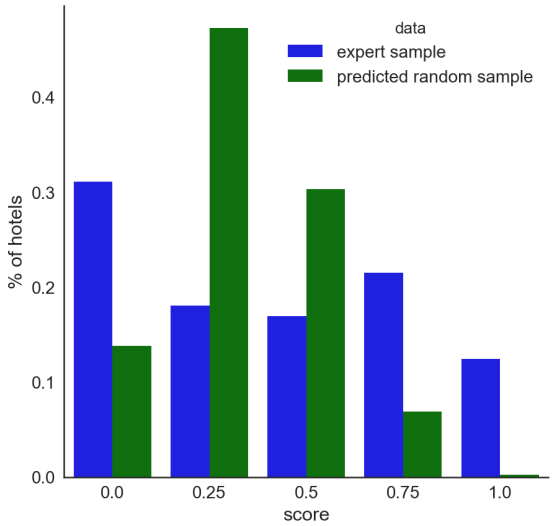


Figure 6.7: Comparison of the factor score distribution of the expert sample versus the predicted scores of 10,000 randomly chosen hotels for factor *Social & Sports*.

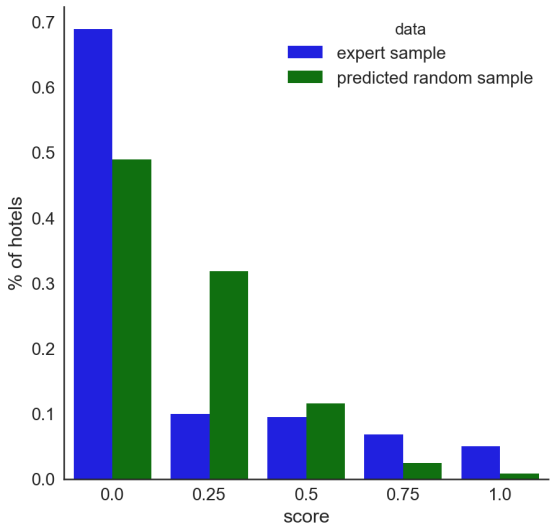


Figure 6.8: Comparison of the factor score distribution of the expert sample versus the predicted scores of 10,000 randomly chosen hotels for factor *Action & Fun*.

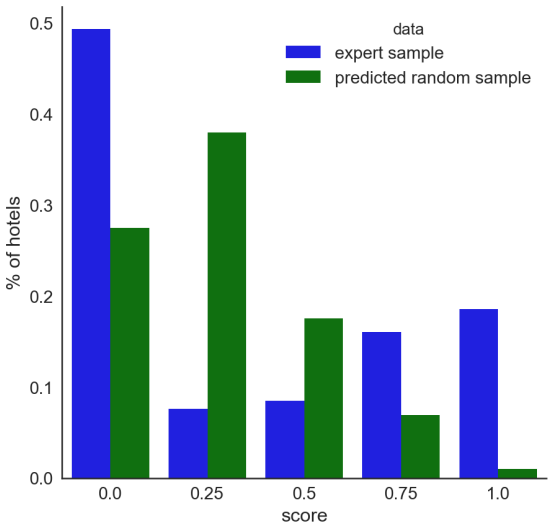


Figure 6.9: Comparison of the factor score distribution of the expert sample versus the predicted scores of 10,000 randomly chosen hotels for factor *Nature & Recreation*.

the proportions of hotels with score 0, the proportions of hotels in the expert mapping are increasing with increasing score while in the random sample they are decreasing.

To sum up, factor *Sun & Chill-Out* shows similar distributions of factor scores in the expert sample and in the predicted factors of the random sample. Further, the factors *Knowledge & Travel*, *Culture & Indulgence*, and *Action & Fun* do not show any similarity in the proportions of hotels for any score in any of the samples, but expert scores and predicted scores are showing roughly similar trends. Finally, no similarities in distributions of the expert mapping and the predicted scores could be identified for the factors *Independence & History*, *Social & Sports*, and *Nature & Recreation*. Considering all this, one can only expect a sufficient generalization of the developed models limited to some factors.

In conclusion, one can say that hotel features can be used to determine the Seven-Factor representation of a hotel. But quality and performance of the developed models highly depend on the factor itself. The model of factor *Sun & Chill-Out* showed very promising results, where 59% of the variance in factor scores in the test set could be explained. On the other hand, the models of *Social & Sports* and especially *Independence & History* showed relatively poor performance, where only 13-29% of the variance could be explained. All other models had midlevel performances, where 35-42% of the variance could be explained.

Table 6.11 lists all independent variables (i.e., used hotel features) of the fitted multiple linear regression models of the Seven-Factors. Note, that a minus sign indicates a negative impact on the corresponding factor. For example, the model of *Sun& Chill-Out* mainly consist of indicators of beach resorts, but there are also indicators of crowdedness

Factor	Hotel features
Sun & Chill-Out	*beach, sun_index, seaview, - numroomstotal, aircon, - cityhotel, bicyclerental, - skihotel, - garage*
Knowledge & Travel	*- balcony, cityhotel, - childrensbed, internetaccess, - playground, numroomstotal, - massage, businesshotel, - terrace, - bedroom*
Independence & History	*- family_index, cityhotel, - recreationalsports_count, internetaccess, businesshotel, - pets, - beach*
Culture & Indulgence	*cityhotel, category_official, dvdplayer, minibar, - pets, businesscentre, charmhotel, - sun_index, queensizebed, numjuniorsuites*
Social & Sports	*- shuttleservice, pets, tabletennis, - category_official, skihotel, - cityhotel, - beach, - businesscentre, spa, - hairdresser*
Action & Fun	*numfloorsmain, businesscentre, - pets, category_official, - bicyclerental, medicalattendance, - nature_index, sports_catamaran, - village, roomservice, minifridge, casinoresort*
Nature & Recreation	*- beach, - cityhotel, skihotel, - elevators, quietlysituated, sauna, - numapartments, garden, fireplace, - foyer, sports_skiing*

Table 6.11: Used hotel features in the resulting multiple regression models.

or cold weather, which have a negative impact. Overall, 47 out of 300 features are actually in use and following 11 features (out of the 47) are appearing in more than one model: *cityhotel, beach, pets, businesscentre, category_official, skihotel, bicyclerental, businesshotel, internetaccess, numroomstotal,* and *sun_index.*

CHAPTER 7

Discussion

This chapter first discusses the outcomes of the exploratory data analysis, the cluster analysis, the model building, and the evaluation with respect to the destination data. Subsequently, an equivalent discussion, but with respect to the hotel data, is made. Finally, results based on both data sets (destination and hotel) are compared.

7.1 Tourism Destinations

The cluster analysis of destinations, based on their similarity, resulted in following six clusters: *vibrant beach resorts, energetic cities, tranquil seaside resorts, peaceful towns, idyllic island villages,* and *ordinary towns*. As already mentioned, one can observe (in the resulting destination clusters) a contrast of *vibrant* to *tranquil*, *land* to *island*, *seaside* to *inland (urban area)*. These contrasts, or so to say axes of separation (cohesion), can also be observed in the clustering of features based on their pairwise correlation. The clustered correlation heat map of tourism destination features (see Figure 3.12) resulted in six clearly separable and interpretable groups of features, which are covering following aspects of tourism destinations: recreational, island, countryside, urban area, mass tourism, and seaside.

Considering the distribution of average factor scores, based on the expert mapping, in each cluster (see Table 4.1) one can say that the factor *Sun & Chill-Out* scores the best in *vibrant beach resorts* and *tranquil seaside resorts*, the factors *Knowledge & Travel*, *Independence & History*, and *Culture & Indulgence* in *energetic cities*, the factor *Social & Sports* has similar scores in each cluster, the factor *Action & Fun* in *vibrant beach resorts* and *energetic cities*, and finally the factor *Nature & Recreation* in *tranquil seaside resorts*, *peaceful towns*, and *idyllic island villages*.

Further, taking into account the previously mentioned axes *vibrant/tranquil, land/island,* and *seaside/inland* one can clearly see that some of the Seven-Factors are near to one

end of the listed axes. The factor *Sun & Chill-Out* is obviously near *seaside* rather than *inland (urban area)*. Factors *Knowledge & Travel*, *Independence & History*, and *Culture & Indulgence* are located at the *vibrant* end rather than *tranquil* and additionally they are near *inland (urban area)* rather than *seaside*. The factor *Social & Sports* is mainly located somehow in the middle of each axis. In other words, there is no underlying grouping, where factor *Social & Sports* is more dominant. The factor *Action & Fun* is obviously more at the *vibrant* side rather than *tranquil*. Finally and clearly, *Nature & Recreation* is far at the *tranquil* end and obviously not at the *vibrant* end.

Overall, one can conclude that there is an clear underlying structure. Some aspects of the discovered latent structure are more or less appealing to some of the Seven-Factors, except for the factor *Social & Sports*, which shows no clear commitment to any cluster. The ambiguity of *Social & Sports* is due to an uneven distribution of factor scores in the expert mapping, where 87% of the destinations scored either with 0.5 or 0.75.

In the last part of Section 3.5 most correlated destination features of each factor of the Seven-Factor Model are examined and discussed. Those correlations are based on the expert mapping. Table 7.1 summarizes the outcomes at one glance. Note, that the features are listed in an descending order based on the absolute value of the correlation coefficients and a minus sign in front of a features indicates a negative correlation.

Factor	Top correlated destination features
Sun & Chill-Out	*beach & swimming, sea, beach, diving, kite & windsurfing, surfing, sailing, nature & landscape, family*
Knowledge & Travel	*sightseeing, culture, gastronomy, - peacefulness, mobility, nightlife, entertainment, accommodations, sports*
Independence & History	*sightseeing, culture, mobility, gastronomy, - peacefulness, accommodations, nightlife, entertainment, sports*
Culture & Indulgence	*sightseeing, culture, nightlife, gastronomy, mobility, - peacefulness, accommodations, entertainment, image & flair*
Social & Sports	*- old town, nature & landscape, - sightseeing, hiking, sports, family, - river, mountain biking, mountains*
Action & Fun	*gastronomy, - peacefulness, nightlife, mobility, accommodations, sports, entertainment, sightseeing, wellness*
Nature & Recreation	*peacefulness, - gastronomy, - nightlife, - entertainment, - accommodations, - sightseeing, - sports, - mobility, - culture*

Table 7.1: Most correlated destination features of the Seven-Factors.

Overall, the listed features in Table 7.1 are highly correlated with the corresponding

factors, except in case of the factor *Social & Sports* (due to an uneven distribution in the expert mapping). Nevertheless, the correlation analysis delivers reasonable results for all factors. Especially, the factors *Sun & Chill-Out*, *Culture & Indulgence*, and *Action & Fun* are well covered.

Furthermore, one can immediately see that factors *Knowledge & Travel*, *Independence & History*, and *Culture & Indulgence* are pretty similar in their correlations with features. All three are correlated with cultural and historical aspects, but also with features indicating mass tourism and urbanization. This observation is in line with the distribution of factors in the clustering, where all three factors are more in the vibrant, urban area side of the axis. Since all three factors have partially similar interests such similarity in clustering and correlation behavior is reasonable. Also, the factor *Action & Fun* is similar to those three, but vibrancy and nightlife are playing a stronger role than historical and cultural aspects. Like in the clustering analysis, the clearest and most reasonable result, can be observed for the factor *Sun & Chill-Out*, where almost all features are related to beach and sea. For the factor *Nature & Recreation* the most correlated feature is *peacefulness*, which is obvious. All other listed features for this factor are indicators of mass tourism and urbanization, which are negatively correlated with the factor. This contrast of recreational features versus urban area and mass tourism features is also clearly observable in the cluster analysis. Surprisingly, also the factor *Social & Sports* shows meaningful, interpretable correlations with destination features. Indicators of crowdedness are negatively correlated, features related to recreation and nature are positively correlated, and of course any feature related to sports is positively correlated. Such correlation behavior is in line with the characteristics of the factor *Social & Sports*.

Factor	Destination feature
Sun & Chill-Out	*- nightlife, beach & swimming, health resort, sea*
Knowledge & Travel	*sightseeing, mobility, - winter sports resort, - sea*
Independence & History	*culture, sightseeing, - nature & landscape*
Culture & Indulgence	*image & flair, culture, sightseeing, - nature & landscape, old town*
Social & Sports	*sports, - wellness, - sightseeing, peacefulness, mountains*
Action & Fun	*winter sports, - peacefulness, shopping, nightlife, - family, metropolis, sea, - health resort, kite & windsurfing*
Nature & Recreation	*- nightlife, peacefulness, hiking, - sightseeing, - shopping, health resort, - beach*

Table 7.2: Used destination features in the resulting multiple linear regression models.

Table 7.2 lists all independent variables (destination features) of the fitted multiple linear regression models of the Seven-Factors. Apart from model of the factor *Social & Sports*, all other models have very promising results, where 59-77% of the variance in factor

scores in the test set is explained. In particular, the models of factors *Action & Fun* and *Nature & Recreation* are describing the respective factor very well and are also showing good generalization performance.

Furthermore, most features used in the regression models are also listed in Table 7.1, where correlated features of the Seven-Factors are shown. Such similarity is expected, but additionally some destination features of the resulting models are covering aspects, which are not touched by the correlation analysis. For example, the model of the factor *Sun & Chill-Out* consists of indicators of sun and beach as expected, but there are also indicators off crowdedness, which have a negative impact on the factor. Such model structure is in line with the characteristics of *Sun & Chill-Out*, where crowdedness and mass tourism are negatively associated with the factor. For the factors *Knowledge & Travel*, *Independence & History*, and *Culture & Indulgence* indicators of history and culture (like in the correlation analysis) are positively associated. Also, like in the correlation results (see Table 7.1) there are some negatively associated features, which might show the lack of cultural or sightseeing offers like winter sports resort. For the factors *Social & Sports*, *Action & Fun*, and *Nature & Recreation* more or less the same aspects of the corresponding factor characteristics are covered as in the correlation analysis (see Table 7.1). For all three factors, regression and correlation analysis resulted in destination features, which are showing the contrast between tranquility and vibrance. Additionally for the factor *Social & Sports* there are also indicators of sports.

All in all, results of the exploratory data analysis, the cluster analysis and the regression analysis are in line. Many independent variables of the regression models are also listed as most correlated features, and those features are also the most decisive variables for the clustering. This can also be seen in the evaluation of the results, where most models show decent performance and sufficient generalization.

7.2 Hotels

Just as for tourism destinations, also the features of the hotels are clustered based on their pairwise correlation among each other. According to the generated clustered correlation heat map (see Figure 5.9) most features are uncorrelated. Only, three meaningful groupings of hotel features are identified. The first one refers to features indicating family friendliness and beach resorts, the second one clusters mainly features related to city hotels, and the last one groups features indicating self-supporter accommodations (kitchen, washing machine etc.). Further, hotels are clustered based on the similarity among each other (see Section 6.1). Six conceptually meaningful groupings are identified namely, *basic self-supporter hotel, recreational wellness hotel, lovely beach hotel, simple city hotel, centrally situated city hotel,* and *high class beach resort.*

Overall, the resulting clustering is not as sharp and clear as in the destinations data, which can also be observed in the distribution of factor scores over the determined clusters. Only, for the factor *Sun & Chill-Out* a clear interpretation can be made, where it scores the best in *high class beach resort* and *lovely beach hotel.*

Based on the expert sample, where experts mapped hotels to the Seven-Factors manually, the most correlated hotel features of each factor of the Seven-Factor Model are identified. Those features are listed in Table 7.3. Note, that the features are listed in an descending order based on the absolute value of the correlation coefficients and a minus sign in front of a features indicates a negative correlation.

Factor	Top correlated hotel features
Sun & Chill-Out	*sun_index, beach, deckchairs, seaview, parasols, watersports_count, pool, childrenspool, balcony*
Knowledge & Travel	*- balcony, - sun_index, - parasols, - deckchairs, cityhotel, businesshotel, numroomstotal, - playground*
Independence & History	*- sun_index, - balcony, - beach, - family_index, - watersports_count, - playground, - deckchairs, - childrenspool, - terrace*
Culture & Indulgence	*category_official, cityhotel, minibar, category_recommended, - balcony, conferencerooms, roomsservice, - beach, numfloorsmain*
Social & Sports	*- cityhotel, - minibar, - roomsservice, - numfloorsmain, tabletennis, spa_count, refectory, - businesscentre, - conferencerooms*
Action & Fun	*category_official, numfloorsmain, numroomstotal, numbungalows, roomsservice, minibar, numroomssingle, gym*
Nature & Recreation	*- cityhotel, - beach, quietlysituated, sauna, - numfloorsmain, - minbar, wintersports_count, bicyclerental, skihotel*

Table 7.3: Most correlated hotel features of the Seven-Factors.

Overall, the most correlated hotel features of the Seven-Factors are not as highly correlated as the most correlated destination features of the Seven-Factors. However, the best result is achieved for the factor *Sun & Chill-Out*, but also correlations of factors *Culture & Indulgence*, *Action & Fun*, and *Nature & Recreation* are clear, reasonable and easy to interpret. Furthermore, the correlation analysis covers only few characteristic aspects of the factors *Knowledge & Travel*, *Independence & History*, and *Nature & Recreation*.

Top correlated features of the factor *Sun & Chill-Out* are all positively correlated with the factor and they can be seen as indicators of beach resorts. On the opposite, the factors *Knowledge & Travel* and *Independence & History* are negatively correlated to obvious features of beach resorts, which probably shows the lack of cultural and historical offers in such hotels. Additionally, one can observe a positive correlation with indicators of city hotels, where the probability of cultural and historical activities is higher. The high-class tourist, *Culture & Indulgence*, is highly positively correlated with the official category in stars, which does not need further explanations. The factor also shows positive

correlations with indicators of city hotels, where as already mentioned the probability of cultural and historical activities is higher. The factor *Action & Fun* is positively correlated with the size of a hotel and with its category in stars. In other words, the bigger the hotel and the more stars it has the higher the score in factor *Action & Fun*, which is totally in line with the jet-setter characteristic of the factor. Finally, factor *Nature & Recreation* is positively correlated with features related to recreation and negatively correlated with the size of a hotel and city hotel/ beach resort indicators.

Factor	Hotel features
Sun & Chill-Out	*beach, sun_index, seaview, - numroomstotal, aircon, - cityhotel, bicyclerental, - skihotel, - garage*
Knowledge & Travel	*- balcony, cityhotel, - childrensbed, internetaccess, - playground, numroomstotal, - massage, businesshotel, - terrace, - bedroom*
Independence & History	*- family_index, cityhotel, - recreationalsports_count, internetaccess, businesshotel, - pets, - beach*
Culture & Indulgence	*cityhotel, category_official, dvdplayer, minibar, - pets, businesscentre, charmhotel, - sun_index, queensizebed, numjuniorsuites*
Social & Sports	*- shuttleservice, pets, tabletennis, - category_official, skihotel, - cityhotel, - beach, - businesscentre, spa, - hairdresser*
Action & Fun	*numfloorsmain, businesscentre, - pets, category_official, - bicyclerental, medicalattendance, - nature_index, sports_catamaran, - village, roomservice, minifridge, casinoresort*
Nature & Recreation	*- beach, - cityhotel, skihotel, - elevators, quietlysituated, sauna, - numapartments, garden, fireplace, - foyer, sports_skiing*

Table 7.4: Used hotel features in the resulting multiple regression models.

Table 7.4 lists all independent variables (hotel features) of the fitted multiple linear regression models of the Seven-Factors. Overall, the model of factor *Sun & Chill-Out* shows the best performance by explaining 59% of the variance in factor scores in the test set. Followed by the moderate performances of the factors *Knowledge & Travel*, *Culture & Indulgence*, *Action & Fun*, and *Nature & Recreation*, where 35-42% of the variance is explained. On the other hand, the models of the factors *Social & Sports* and especially *Independence & History* are performing relatively poor by explaining only 29% and 13% of the variance.

Some features used in the regression models are, as expected, also listed in Table 7.3, where the most correlated features of the Seven-Factors are shown. Additionally, some hotel features of the resulting models are covering aspects, which are not touched by the

correlation analysis. This shows that choosing a stepwise feature selection over a feature selection simply based on correlation is reasonable. For example, the regression model of the factor *Sun & Chill-Out* not only comprises sun and beach related hotel features, but also includes cityhotel and skihotel, which have negative impacts since such types of hotels are less appropriate for the factor. Another example is the factor *Action & Fun* , where not only indicators of size and category of a hotel are part of the regression model, but also nature and recreation related hotel features, which are obviously negatively associated with the factor *Action & Fun*. Also, the regression model of the factor *Nature & Recreation* captures some additional aspects in comparison to the correlation analysis. The model not only shows the positive relation of the factor to recreational aspects of hotels and the negative association with beach and city hotels, but it also reveals the obvious positive relation with indicators of nature (garden, fireplace) and the negative association to the size of a hotel. Features in the models of the factors *Knowledge & Travel*, *Independence & History*, and *Culture & Indulgence* are covering more or less the same aspects as in the correlation analysis (see Table 7.3).

All in all, the results are not as clear and interpretable as for the destinations. The outcomes of the clustering, the correlation, and the regression analysis are in line, very clear and plausible only for factor *Sun & Chill-Out*. This can also be observed in the evaluation of the results, where the multiple linear regression model of factor *Sun & Chill-Out* shows the best performance and generalization potential. The results for all other factors of the Seven-Factor Model are only treating partial aspects of the corresponding factor characteristics.

7.3 Differences in Tourism Products

As already mentioned, this work is based on a picture based approach to recommender systems introduced by Neidhardt et al. in [NSSW14, NSSW15]. Recommendation items in the proposed picture based recommender are defined as point of interests (POIs) and are categorized "as sight (e.g., the Eiffel Tower), activity (e.g., boat ride), restaurant, entertainment (e.g., an opera or musical), shopping, nightlife or tours" [NSSW14, NSSW15]. One can see that recommendation items in the picture based recommender are pretty specific. Hence, it is intuitive to assign POIs to factors of the Seven-Factor Model. For example, "wine tasting" is definitely an activity for *Culture & Indulgence* or "base jumping" can obviously be assigned to *Action & Fun*. This is in line with [GMH+06], where Gretzel et al. are showing that predefined tourist roles can be used in order to recommend touristic activities.

However, this work considers tourism destinations and hotels as recommendation items. These type of tourism products are more generic in comparison to the previously mentioned POIs. Hence, mapping tourism destinations or hotels to the Seven-Factors is not as intuitive as mapping POIs. This can also be observed in the expert samples, where many destination and especially hotels are scoring with 0 in most of the Seven-Factors (see Figures 3.16, 5.10, and 5.11).

In [GMH+06] Gretzel et al. demonstrate that recommending destinations is challenging, but it can be accomplished by determining the touristic activities a destination offers by using predefined tourist roles. Also, Grün et al. highlight in [GNW17] that tourist roles can be used "as a shortcut to propose appropriate tourism objects". Obviously, the more specific product features are the clearer the interpretability of the fitted model gets and also the better the performance is. The used data set for tourism destinations contains motivational ratings and geographical attributes. Motivational ratings can almost always be associated with touristic activities, for example *beach & swim* with "chilling or swimming at the beach". Also geographical attributes like mountains, old town or metropolis can be seen as indicators for some touristic activities or areas where some touristic activities are more likely to be found. Although tourism destinations are more generic than explicit touristic activities, in most cases they can be associated with touristic activities straight forward. One can argue that, such simplicity in identifying associations might be the reason of the promising results based on the destination data.

Furthermore, in [GNW17] Grün et al. aim to "close the gap between users' needs and suppliers' perspectives by matching their respective views" with respect to tourism products. One can argue that the mentioned gap might be smaller or larger depending on the respective tourism product. For example, there are about 300 hotel features in the used data set, which can be seen as 300 points, where the mentioned views of consumers and suppliers might differ. Usually, hotels try to attract as many diverse customer types as possible by offering and listing many features. Thus, as already mentioned they are very generic and hard to allocate to factors of the Seven-Factor Model. Of course, there are exceptions, which are more specific like a lodge in the mountains or a beach resort, but in this case the allocation is also more straightforward. Considering the outcomes of the explorative data analysis and cluster analysis based on the hotel data, one can see that sun and beach related features and clusters are consistently standing out. Further, the evaluation of results has shown that the best performing and only viable model with respect to the hotel data is the model of the factor *Sun & Chill-Out*. In [GNW18] Glatzer et al. are using textual descriptions to allocate hotels onto the Seven-Factors. Also in [GNW18] the best performance is achieved for the factor *Sun & Chill-Out*. Hence, one can argue that the GIATA data set clearly distinguishes between beach hotels and other types and such specificity leads to promising results for the factor *Sun & Chill-Out*.

The overall interpretability and quality of the fitted models based on the hotel data are clearly outperformed by the models based on tourism destinations. Anyway, some specific characteristic aspects of the factors of the Seven-Factor Model are not covered in both data sets, for example: "to gain knowledge", "to search for the meaning of life", "adventures and thrilling activities", "indulgence", "socialization with locals". This is also pointed out by Glatzer et al. in [GNW18], where they argue that such lack might be one reason for the poor performance of some models. In this work, this can especially be observed for the factor *Independence & History*, where none of the used data sets contains features related to spirituality (i.e. "searching for the meaning of life").

CHAPTER 8

Conclusion

8.1 Summary

Primarily, this works aim is to identify and explain associations between destination features and the Seven-Factor Model and between hotel features and the Seven-Factor Model to enable an automated mapping of both tourism products onto the Seven-Factors. To do so, first a literature review was conducted, analyzing predefined tourist roles including the Seven-Factors, RSs in general and then with focus on tourism, a picture based approach to RSs, and common data sources in ICT & tourism research.

Due to a focus on two different tourism products, destinations and hotels, further work was split into two main tasks. The first task (Chapters 3 and 4) was dealing with data acquisition, data preprocessing, missing value analysis and treatment, explorative data analysis, and finally model building and evaluation with focus on tourism destinations. Respectively, the second task (Chapters 5 and 6) was dealing with the same issues, but with focus on hotels.

The data for destinations was provided by Webologen [Gmbc], a German e-Tourism company, as a SQL-dump. Hence, the data was extracted and transformed into a more appropriate form in order to do further analysis, i.e. into a tabular form where columns represent destination features and rows represent distinct destinations.

Missing data analysis showed that the used data set for destinations is relatively sparse. Thus, a missing data strategy was built in order to deal with such sparsity. The strategy for missing value treatment included also a comparison of different data imputation methods, where the SOFT-IMUPTE method was outperforming a naive imputation method and the KNN imputation.

Furthermore, the explorative data analysis showed that destination features can be grouped based on the correlations among each other. Six meaningful groups of features

could be identified such as features of recreational tourism destinations, features of urban areas, features of destinations with intense tourism, features of destination at the countryside, features of destination on an island, features of destinations located at the seaside. Such clear and easy to interpret grouping was the first sign that there might be an underlying natural and meaningful structure among tourism destinations.

In addition to the whole data set experts from Pixtri [OG], an Austrian e-Tourism company, provided an already mapped sample, where scores for all Seven-Factors were assigned to each destination in the sample. Using this expert sample, a correlation analysis was conducted in order to identify the most correlated destination features of the Seven-Factors. The resulting correlations were significant and could be interpreted effortlessly. For example, the factor *Nature & Recreation* had a high positive correlation with *peacefulness* and high negative correlations with features of intense tourism and urban areas.

In the model building part of the destinations data, first a cluster analysis was conducted in order to determine if there is an underlying natural structure. Six conceptually meaningful groups were identified, namely *vibrant beach resorts, energetic cities, tranquil seaside resorts, peaceful towns, idyllic island villages,* and *ordinary towns*. The identified latent structures were in line with the results of the exploratory data analysis. Another interesting observation was that cluster cohesion (separation) was based on following three contrasts: vibrant to tranquil, land to island, seaside to inland (urban areas). Furthermore, the developed clusters were validated against the provided expert mapping, i.e. factor score distributions over the six clusters were examined. Except for the factor *Social & Sports* a clear separation could be made. For example, the factor *Sun & Chill-Out* was scoring the best in clusters *vibrant beach resorts* and *tranquil seaside resorts*, whereas in all other clusters it showed very low scores. Those clusters fostered a better understanding of the similarities among destinations and can be used for more accurate recommendations or can be targeted directly (except for the factor *Social & Sports*).

In contrast to the cluster analysis (an unsupervised learning method) the model building part of the destinations data also includes a supervised learning approach. Since the overall aim is not only to project destinations into the seven-dimensional vector space of travel behavioral patterns, but also to explain which attributes of destinations are more important in this purpose, a multiple linear regression (MLR) analysis with step wise variable selection was conducted. Seven models were established, one for each factor of the Seven-Factor Model. The resulting models are providing strong evidence that there is a significant relation between selected destination features and the Seven-Factors. Furthermore, the developed linear models were challenged by two conceptually different non-linear models, namely random forest regression (RF) and K-Nearest-Neighbor regression (KNN). Additionally, the predictive performances of all three models MLR, RF, and KNN were compared to a baseline function (simple mean of each factor). The evaluation showed that all three models were always better than the baseline function, which indicated that they had learned something out of the data. It has also been demonstrated that the performance of the MLR model is similar to the performance of

the RF model and both are outperforming the KNN model. In the end the MLR model was chosen over the RF model since MLR is simpler to fit and easier to interpret than RF. Overall, all travel behavioral patterns are well described (59-77% of the variance) by the resulting models, except for the factor *Social & Sports*, where only 22% of the variance can be explained. This is caused by an uneven distribution of scores of the factor *Social & Sports* in the expert sample.

The data for hotels was provided as a huge archive of XML-files by GIATA [Gmbb], a German e-Tourism company and the quasi market leader in the matter of tourism content. The XML-files were parsed, preprocessed, and transformed to a more convenient and tabular format, where each row corresponds to a distinct hotel and each column to a hotel feature.

Missing data analysis showed that the sparsity of the hotel data is even worse than the sparsity of the destination data. Unfortunately, the missing data strategy for the destinations data could not be used for the hotels. Thus, an individual missing data strategy for hotel data was developed. In addition to the existing hotel features some generated features were added to the hotel data set in order to support the model building. For example, a counter for all sports offers of a hotel or a counter for all features indicating sun and beach were added as new features into the data set.

Furthermore, the explorative data analysis showed that hotel features are mainly uncorrelated, but some features could be grouped based on the correlation among each other. Only three meaningful groups of features could be identified such as features of family friendly recreational beach resorts, features of centrally situated city hotels, and features of appropriate accommodations for self-supporters. This showed that there might be a weak, latent, natural structure among tourism destinations, especially in comparison to the encountered structure in the destination data.

In addition to the whole data set of hotels experts of Pixtri [OG] and Eurotours [Gmba] (also an Austrian e-Tourism company) provided pre-mapped samples, where scores for all Seven-Factors had been assigned to each hotel in the samples. Using these expert samples, a correlation analysis was conducted in order to identify the most correlated hotel features of the Seven-Factors. The correlations were not so high, clear, and interpretable as in the destinations case, but some interesting observation could be made. For example, the factor *Sun & Chill-Out* was highly positively correlated with typical features of hotels related to beach vacation. Basically, the two expert samples had very similar behavior in the exploratory analysis, so for further analysis merging made more sense than considering them individually.

In the model building part of the hotel data, first a cluster analysis in order to identify latent meaningful natural groupings of hotels was conducted. Six conceptually meaningful groups were found, namely *basic self-supporter hotel, recreational wellness hotel, lovely beach hotel, simple city hotel, centrally situated city hotel,* and *high class beach resort.* Although a plausible cluster solution was found, the cohesion of hotels within the resulting clusters were relatively weak, respectively the separation also. Only the contrast between

seaside and urban areas was clearly observable in the resulting clustering. Furthermore, the developed clusters were validated against the provided expert mapping, i.e. factor score distributions over the six clusters were examined. Only for the factor *Sun & Chill-Out* a clear separation could be made, where it showed a high average score in *lovely beach hotel* (0.66) and *high class beach resort* (0.81) and low scores in all other clusters. Despite the weak cluster result, one could observe some weak but meaningful trends that led to a better understanding. However, except for the factor *Sun & Chill-Out*, the developed clusters cannot be used for the rest of the Seven-Factors.

In the second part of the model building with the hotel data different supervised learning methods were applied and compared. The same approach as in the destinations case was followed. MLR models, RF models, and KNN models were fitted for each factor of the Seven-Factor model and compared with each other and against a baseline function (simple mean of each factor). The evaluation showed that all models were always better than the baseline function except the models of the factor *Independence & History*, where there was almost no difference to the baseline and the models were performing the worst in comparison to the models of all other factors. Besides that, the performance of the MLR models were again similar to the of RF models and both were again outperforming the KNN method. Thus, the MLR method was chosen over the RF method since MLR is simpler to fit and easier to interpret than RF. The resulting models were providing strong evidence that there is a significant relationship between particular hotel features and the Seven-Factors. But only the model for factor *Sun & Chill-Out* showed a viable performance, where 59% of the variance could be explained. Nevertheless, all other models showed plausible associations between hotel features and the Seven-Factor Model and contributed to a better understanding.

8.2 Future work

The main limitations of this work were caused by the used data sources and samples. Poor performances of some fitted models were caused by an unsound sample, i.e. the factor scores of the expert samples were not evenly distributed. Although the thoroughly fitted and fine-tuned RF models were excelling in the training phase of the models their test performance was similar to the of the MLR models. Thus, the RF models were clearly overfitting the training data, which might be a sign of too small samples. Statically sounder and bigger samples will be targeted in future work.

The explorative analysis and also the developed models showed that the used data set were not able to cover all characteristic aspects of the factors of the Seven-Factor Model. Especially, there were no features indicating independence, the passion for knowledge gain, indulgence, or socialization with locals. Other data sources might be able to cover the characteristic aspects of the factors of the Seven-Factor Model better and could be used to enhance the models. However, this aim immediately shows a disadvantage of the followed approach, namely data source dependency. To counter this problem one could build up a comprehensive data model of tourism products. This data model will serve as

an “intermediary” layer between the respective data source and the Seven-Factor Model and can therefore be used to harmonize heterogeneous sources of data (e.g., by mapping different sources of destination data onto this layer).

In [WR04] Werthner and Ricci point out the complexity of tourism products (i.e., they typically combine accommodation, transportation, activities, food, etc.). In other words, tourism offers are in general packages of several tourism products. In this sense it is planned to combine the outcomes of this work in a two step-recommendation process, where first a destination is determined and then a hotel.

List of Figures

List of Tables

Bibliography

[Air17] Airbnb. Airbnb trip matcher. https://press.atairbnb.com/tripmatcher/, June 2017. Online, accessed 12-November-2017.

[B+03] David Beirman et al. Restoring tourism destinations in crisis: A strategic marketing approach. *CAUTHE 2003: Riding the Wave of Tourism and Hospitality Research*, page 1146, 2003.

[BER14] Matthias Braunhofer, Mehdi Elahi, and Francesco Ricci. Usability assessment of a context-aware and personality-based mobile recommender system. In *International conference on electronic commerce and web technologies*, pages 77–88. Springer, 2014.

[Bil16] Aleksey Bilogur. Missingno. https://www.giata.com, October 2016. Online, accessed 13-April-2018.

[BK07] Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

[BMV14] Joan Borràs, Antonio Moreno, and Aida Valls. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, 41(16):7370–7389, 2014.

[BR11] Robin Burke and Maryam Ramezani. Matching recommendation technologies and domains. In *Recommender systems handbook*, pages 367–386. Springer, 2011.

[Bur07] Robin Burke. The adaptive web. chapter hybrid web recommender systems. 2007.

[Buz15] BuzzFeed. What's your travel personality? https://www.buzzfeed.com/jadayounghatchett/what-type-of-traveler-are-you, May 2015. Online, accessed 12-November-2017.

[CCK+00] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. ftp://ftp.software.ibm.com/

software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf, 2000. Online, accessed 28-April-2018.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[Coh72] Erik Cohen. Toward a sociology of international tourism. *Social research*, pages 164–182, 1972.

[DNW16] Amra Delić, Julia Neidhardt, and Hannes Werthner. Are sun lovers nervous? *e-Review of Tourism Research (eRTR)*, ENTER 2016 Special issue, 2016.

[End10] Craig K Enders. *Applied missing data analysis*. Guilford Press, 2010.

[FWW06] Daniel R Fesenmaier, Karl W Wöber, and Hannes Werthner. *Destination recommendation systems: Behavioral foundations and applications*. Cabi, 2006.

[Gmba] Eurotours GmbH. Eurotours. https://www.eurotours.at. Online, accessed 13-April-2018.

[Gmbb] GIATA GmbH. Giata. https://www.giata.com. Online, accessed 13-April-2018.

[Gmbc] Webologen GmbH. Webologen. http://www.webologen.de. Online, accessed 13-April-2018.

[Gmb10] GIATA GmbH. Giata facts. http://www.giata-xml.de/dokumentation/GIATA-Facts/GIATA-Facts-en.html, December 2010. Online, accessed 13-April-2018.

[GMH+06] Ulrike Gretzel, NICOLE Mitsche, Yeong-Hyeon Hwang, Daniel R Fesenmaier, et al. Travel personality testing for destination recommendation systems. *Destination recommendation systems. Behavioural Foundations and Applications. Oxfordshire: CABI*, pages 121–136, 2006.

[GNW17] Christoph Grün, Julia Neidhardt, and Hannes Werthner. Ontology-based matchmaking to provide personalized recommendations for tourists. In *Information and Communication Technologies in Tourism 2017*, pages 3–16. Springer, 2017.

[GNW18] Lisa Glatzer, Julia Neidhardt, and Hannes Werthner. Automated assignment of hotel descriptions to travel behavioural patterns. In *Information and Communication Technologies in Tourism 2018*, pages 409–421. Springer, 2018.

[Gol90] Lewis R Goldberg. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216, 1990.

[Gow71] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

[Gra12] John W Graham. Missing data theory. In *Missing Data*, pages 3–46. Springer, 2012.

[GS09] Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(Dec):2935–2962, 2009.

[GSO11] Inma Garcia, Laura Sebastia, and Eva Onaindia. On the design of individual and group recommender systems for tourism. *Expert systems with applications*, 38(6):7683–7692, 2011.

[GY02] Heather Gibson and Andrew Yiannakis. Tourist roles: Needs and the lifecourse. *Annals of tourism research*, 29(2):358–383, 2002.

[HGXF06] Y Hwang, Ulrike Gretzel, Zheng Xiang, and Daniel R Fesenmaier. Information search for travel decisions. *Destination recommendation systems: Behavioral foundations and applications*, 42(4):357–371, 2006.

[JWHT13a] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear model selection and regularization. In *An Introduction to Statistical Learning*, pages 203–264. Springer, 2013.

[JWHT13b] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear regression. In *An Introduction to Statistical Learning*, pages 59–126. Springer, 2013.

[JWHT13c] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Statistical learning. In *An Introduction to Statistical Learning*, pages 15–57. Springer, 2013.

[JWHT13d] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Tree-based methods. In *An Introduction to Statistical Learning*, pages 303–335. Springer, 2013.

[kdn14] kdnuggets.com. What main methodology are you using for your analytics, data mining, or data science projects? `https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html`, October 2014. Online; accessed 6-November-2017.

[KR90] Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, pages 68–125, 1990.

[Kri12] W Krinninger. Inspiration and information – critical factors to the success of travel platforms. Master's thesis, Vienna University of Economics and Business, November 2012.

[LR14] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[MDW03] Gerald Matthews, Ian J Deary, and Martha C Whiteman. *Personality traits*. Cambridge University Press, 2003.

[Med14] Ipsos MediaCT. The 2014 traveler's road to decision. `https://www.thinkwithgoogle.com/_qs/documents/918/2014-travelers-road-to-decision_research_studies.pdf`, June 2014. Online; accessed 14-July-2017.

[MF02] Robyn McGuiggan and Jo-Ann Foo. Sun and surf or adventure: Who plays what tourist roles? ban australian perspective. *ACR Asia-Pacific Advances*, 2002.

[MHT10] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[MMS09] Óscar Marbán, Gonzalo Mariscal, and Javier Segovia. A data mining & knowledge discovery process model. In *Data Mining and Knowledge Discovery in Real Life Applications*. InTech, 2009.

[NSSW14] Julia Neidhardt, Rainer Schuster, Leonhard Seyfang, and Hannes Werthner. Eliciting the users' unknown preferences. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 309–312, New York, NY, USA, 2014. ACM.

[NSSW15] Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. A picture-based approach to recommender systems. *Information Technology & Tourism*, 15(1):49–69, 2015.

[NW17] Julia Neidhardt and Hannes Werthner. Travellers and their joint characteristics within the seven-factor model. In *Information and Communication Technologies in Tourism 2017*, pages 503–515. Springer, 2017.

[OG] Pixtri OG. Pixtri. `http://www.pixtri.com`. Online, accessed 13-April-2018.

[P+82] PL Pearce et al. The social psychology of tourist behaviour. *The social psychology of tourist behaviour.*, 1982.

[Rou87] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[RRS15] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: introduction and challenges. In *Recommender systems handbook*, pages 1–34. Springer, 2015.

[Rub76] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[S+78] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[She00] Colin Shearer. The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22, 2000.

[SS17] Roland Schegg and Brigitte Stangl. Information and communication technologies in tourism 2017. 2017.

[TSK05] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition).* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[TSK+06] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 8:487–568, 2006.

[WASC+15] Hannes Werthner, Aurkene Alzua-Sorzabal, Lorenzo Cantoni, Astrid Dickinger, Ulrike Gretzel, Dietmar Jannach, Julia Neidhardt, Birgit Pröll, Francesco Ricci, Miriam Scaglione, et al. Future research issues in it and tourism. *Information Technology & Tourism*, 15(1):1–15, 2015.

[WK99] Hannes Werthner and Stefan Klein. *Information technology and tourism: a challenging relationship.* Springer-Verlag Wien, 1999.

[WR04] Hannes Werthner and Francesco Ricci. E-commerce and tourism. *Commun. ACM*, 47(12):101–105, December 2004.

[WRS02] Amy B Woszczynski, Philip L Roth, and Albert H Segars. Exploring the theoretical foundations of playfulness in computer interactions. *Computers in Human Behavior*, 18(4):369–388, 2002.

[YG92] Andrew Yiannakis and Heather Gibson. Roles tourists play. *Annals of tourism Research*, 19(2):287–303, 1992.

[Zin07] Andreas H Zins. Exploring travel information search behavior beyond common frontiers. *Information Technology & Tourism*, 9(3-1):149–164, 2007.