TU WIEN Informatics

How do we read scatterplots?

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

James Patrick Salazar, BSc
Registration Number 01269132

to the Faculty of Informatics

at the TU Wien

Advisor: O.Univ.Prof. Dipl.-Ing. Dr.techn. A Min Tjoa
Assistance: Univ.Lektorin Dipl.-Ing. Dr.techn. Johanna Schmidt

Vienna, January 10, 2024

James Patrick Salazar	A Min Tjoa

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Declaration of Authorship

James Patrick Salazar, BSc

I hereby declare that I have written this Thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work - including tables, maps and figures - which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.

Vienna, January 10, 2024

James Patrick Salazar

Abstract

People look for patterns, structures, traits, trends, anomalies, and correlations in data. Data visualization helps with this by presenting the data in various formats with various interactions. It can give an qualitative perspective of huge and complex data sets. Additionally, it can provide a data summary, help identify areas of interest, and suggests acceptable parameters for more specialized quantitative research. The scatterplot is arguably the most popular data display method which makes it easier to identify clusters, trends, and correlations. However, they can quickly become too overloaded from the user's perspective when there is a lot of data available. Overplotting is a problem that occurs when multiple observations (points) have the same or strikingly similar values, making it difficult for the user to understand the relationships between the points and variables and producing inaccurate or misleading information in the graph.

In this study, we analyze how the size of data points affect the perception of regression in overloaded scatterplots. Furthermore, we analyze if the education and/or experience in data visualization affects the perception as well. In addition to adhering to the fundamentals of quantitative research by introducing various types, assumptions, techniques, and common mistakes that many researchers make when conducting research studies, this study is dependent on the fundamental and practical issues that should be taken into account when pursuing evaluation studies in information visualization.

Our results show that increasing the dot size does have a positive effect in recognizing the regression in a overplotted scatterplot. Even if the individual dots are not visible anymore the amount of people who see the regression correctly increase. Furthermore, the results show that experience in data visualization does not affect the recognition of regression. Education, however, may affect the recognition of it.

Contents

CHAPTER 1

Introduction

Data visualizations is a very important topic in our research field. Humans scan data for pattern, structure, characteristics, trends, anomalies, and correlations. By displaying the data in multiple forms with diverse interactions, visualization supports this. With the help of Data Visualization, data can be displayed to provide a qualitative overview of large and complex data sets. Furthermore it can summarize data and can assist in identifying regions of interest and appropriate parameters for more focused quantitative analysis. Probably the most often used data mining visualization technique is the scatterplot. Finding clusters, trends, and correlations is made easier. To get further insights from the data, brushing and colored class points are utilized. When too many points overlap or the data is resampled to the point that multiple data points sit at the same (x, y) coordinate, zooming, panning, and jittering can be used to enhance the representation [37].

1.1 Problem statement

There is no denying the effectiveness and popularity of scatterplots as a tool for visual data study [21]. The correct density of data values, however, is difficult to identify when a significant amount of data is used since scatterplots have a high degree of overlap [41]. When there is a lot of data available, scatterplots also seem to soon become overloaded from the user's perspective. When more than one observation (point) has the same or strikingly similar values, this problem is known as *overplotting* which causes data points to overlap to the point where the user finds it difficult to understand the relationships between the points and variables and renders the information of the graph inaccurate or misleading [25]. In Figure 1.1 two plots with the same amount of data points can be seen. One has a smaller dot size. Changing the size is one possible methods to visualize data to be more appealing.

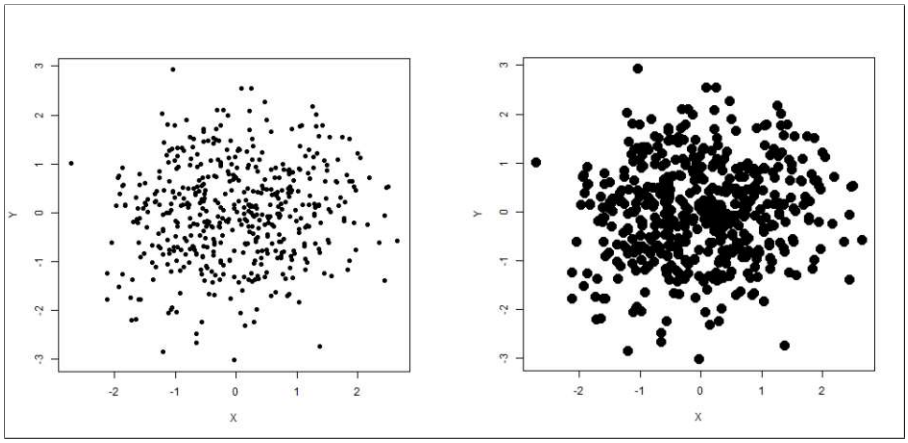


Figure 1.1: Example of overplotting. 500 datapoints are presented in both plots. The dot size is different.

1.2 Goal and expected results

Our hypotheses are based on the knowledge we acquired from literature study, where we discovered that a variety of factors affect how a scatterplot is seen. For instance, the overplotting problem, which arises when a graph contains excessive amounts of data, is one of the reasons listed in Chapter 2.

Our goal is to create a study, that can prove that the size of data points in scatterplots has a significant impact on the perception of the regression by human observers. We expect the participants to misinterpret the correlation (r) when the size of the data points is increased. Furthermore, we want to see if participants with a higher education and experience in data visualization interpret overplotted scatterplots different than participants with lower education and experience in data visualization. We expect the same results from both groups of participants.

1.3 Methodological approach

This study relies on the fundamental and practical issues that should be considered when pursuing evaluation studies in information visualization [26] and also follow the fundamentals of quantitative research by introducing distinct types, assumptions, techniques, and common errors that many researchers make when doing research studies [76]. A study will be conducted to get the numerical data, which will then be interpreted.

We employed a six-step process to plan and conduct the user study:

1. **Literature research**: We reviewed relevant literature about scatterplots, explicitly relating to overplotting and visual clutter visualization (see Section 3.1).
2. **Hypotheses generation**: The basis for conducting the user study is that our hypotheses were generated in favor of a better understanding of overplotting problems in scatterplots (see Section 3.2).
3. **Dataset generation**: We used *R Statistics* [60] to generate random data. Afterward, we generated datasets with different *correlations* for the the experiment (see Section 3.3).
4. **Setup**: The user study was conducted with an online survey tool called *SoSci Survey* [46] in a web-based setting [61]. We set up the study presented for the users with multiple scatterplots created with different parameters for the experiment (see Section 3.4.1). Participants had to complete tasks according to the formulated hypotheses (i.e., judging if the data is positively or negatively correlated) with different scatterplot representations. According to the literature, judgment studies [11] are commonly used for perceptual studies and can provide considerable precision.
5. **Data collection**: For data collection, we asked participants to complete the user study. We collected quantitative data (i.e., accuracy, confidence, and time it took to complete the task) to understand the visual perception of data distribution in scatterplots (see Section 3.5).
6. **Data evaluation**: Using statistical tests (i.e., t-test [24]), we identified differences in participant error rates for data points and parameter settings (see Section 3.6).

1.4 Outline

This work is structured as follows: An initial literature and related work review will be described in Chapter 2. The theoretical groundwork for data visualization is set in this chapter so that the reader may get familiar with the ideas of data visualization, human perception, data literacy, and, in particular, scatterplots. In Chapter 3 we explain our six-step process for the methodological approach. Furthermore, the setup of the user studies is described. In Chapter 4 the results of the collected data is described. Chapter 5 contains a short recap of the main findings. Additionally, current limitations of our work is discussed and plans on future work are explained.

CHAPTER 2

Related Work

"Visualization, as the name implies, is based on exploiting the human visual system as a means of communication" [51, p.6]

- Tamara Munzner, *2014*

In this chapter, we establish the theoretical foundation for data visualization that will allow the reader to become acquainted with the concepts of data visualization, human perception, data literacy, and, in particular, scatterplots.

Section 2.1 provides an introduction to data visualization, beginning with an overview of the rich history of graphical methods and progressing through visualization tools that convert data into visual elements, allowing people to explore and comprehend data. Next, Section 2.2 presents a general introduction that studying human perception is essential in visualization since users view visualizations with their eyes. Finally, in Section 2.3, we introduce scatterplots and their most prominent issue, overplotting; we present techniques that study the problems of visual clutter in scatterplots in two different investigations: cluster and regression analysis.

2.1 Interactive visualization

"Data visualization is part art and part science. The challenge is to get the art right without getting the science wrong, and vice versa" [86, p.1]. Visualization graphically depicts, evaluates, and manipulates data to understand better or more straightforward comprehension [10], and for Wilke [86], a data visualization must first and foremost convey the data accurately. In addition, it must not mislead or distort. At the same time, a data visualization must be aesthetically pleasing since an excellent visual presentation tends to enhance the visualization's message.

Although there have been an increasing number of studies on information visualization in recent years, the field is still growing e.g., Segel & Heer [71] proposed narrative visualization design methodologies, including intriguing and untapped alternatives for journalistic storytelling and instructional media; Hota & Huang [38] demonstrated the effectiveness of self-describing visualizations with two example application implementations: incorporating an embedding filter into the standard rendering process and developing a web reader to automatically and reliably extract provenance information from scientific publications for review and dissemination.

Friendly et al. [29] released their book in 2008. The authors sought to explain the history of data visualization from medieval to modern times, covering the many application areas where data visualization has settled and grown.

Friendly and colleagues stated that the graphical representation of quantitative information has deep roots, tracing back to the history of map production and visual representation, theme mapping, statistics and statistical graphics, medicine, and other areas. Several advancements along the road have contributed to the widespread usage of data visualization today. These include image-drawing and image-reproduction technologies, breakthroughs in mathematics and statistics, and new developments in data gathering, empirical observation, and recording. The authors collected simple analyses examining trends over time in a Milestones project, essentially presented as a chronological list. However, they maintained a relational database (historical objects, references, images) to work with it as data. In Friendly et al.'s Milestones data, we can see the time course of relevant developments in the history of data visualization sorted into eight time periods (see Figure 2.1):

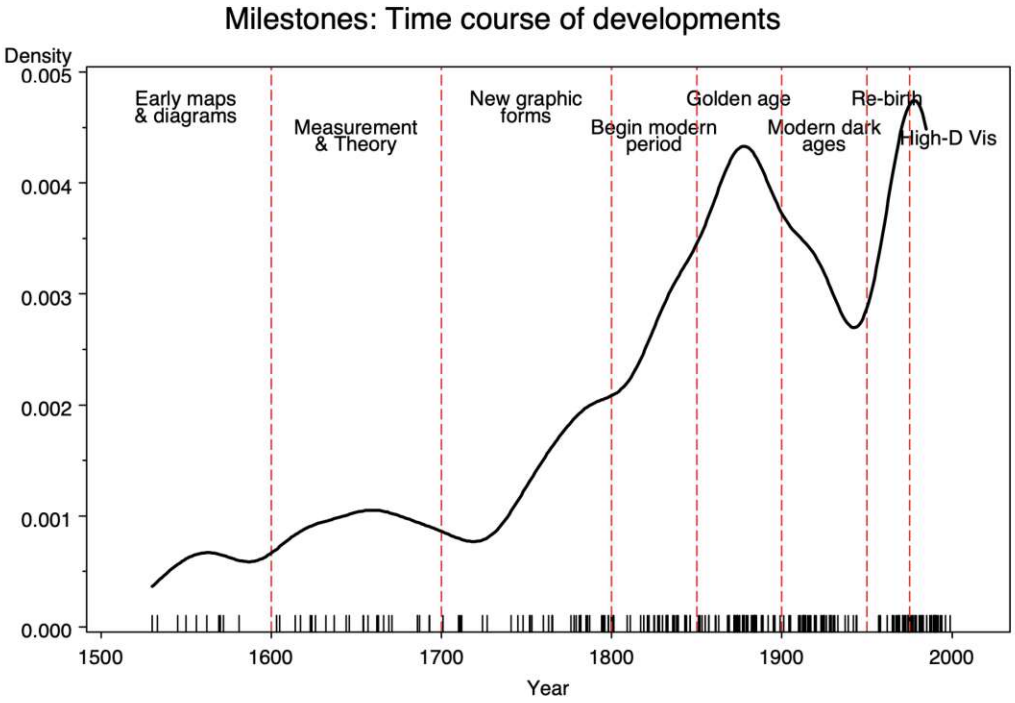


Figure 2.1: The temporal distribution of events considered milestones in the history of data visualization, shown by a rug plot and density estimation. Source: [29, p.3].

1. *Pre-17th century, early maps and diagrams*:
   According to Friendly: "The earliest seeds of visualization arose in geometric diagrams, in tables of the positions of stars and other celestial bodies, and in the making of maps to aid in navigation and exploration" [29, p.27]. The notion of plotting a theoretical function (like a proto bar graph) and the logical link between tabulating numbers and plotting them first surfaced in some work in the 14th century. The earliest picture capture concepts, the recording of mathematical functions in tables, and the first contemporary geographic atlas are also displayed. These are the fundamental steps in data visualization.

2. *1600-1699, measurement and theory*:
   "This century also saw great new growth in theory and the dawn of practical application, theories of errors of measurement and estimation, the birth of probability theory, and the beginnings of demographic statistics and political arithmetic, the study of population, land, taxes, value of goods, etc. for the purpose of understanding the wealth of the state" [29, p.4]. In the 1660s, several European countries began collecting and studying social statistics to educate the state about issues like wealth, population, agricultural land, taxes, and commercial reasons like insurance and annuities based on life tables. At the end of this century, the required pieces for

creating graphical approaches were available. However, perhaps, more importantly, this century might be seen as the birthplace of visual thinking.

3. *1700-1799, new graphic forms*:
   Map-makers began to represent more than simply geographic position on a map in cartography. As a result, new data representations were developed, and thematic mapping of physical quantities grew in popularity. We witness the earliest attempts at the thematic mapping of geology, economic, and medical data around the turn of the century. Several technical advancements also contributed to the development and diffusion of graphic works. Some of these, such as three-color printing, made it easier to reproduce data pictures. However, most of these new graphic forms emerged in low-circulation journals that were unlikely to draw widespread notice, most likely owing to cost.

4. *1800-1850, beginnings of modern graphics*:
   "With the fertilization provided by the previous innovations of design and technique, the first half of the 19th century witnessed explosive growth in statistical graphics and thematic mapping, at a rate which would not be equaled until modern times" [29, p.9]. All contemporary kinds of data display were created in statistical graphics: bar and pie charts, histograms, line graphs, time-series plots, contour plots, and *scatterplots* [8] (see Section 2.3). Mapping moved from single maps to complete atlases in theme cartography, presenting data on various issues (economic, social, moral, medical, physical) and introducing a variety of unique forms of symbolism. Graphical study of natural and physical phenomena (magnetic lines, weather, tides) began to appear routinely in scholarly literature. Graphs began to be acknowledged in certain official circles for economic and governmental planning about the same period, between 1830 and 1850.

5. *1850-1900, the golden age of statistical graphics*:
   By the mid-nineteenth century, all of the requirements for exponential visualization expansion had been developed, "with unparalleled beauty and many innovations in graphics and thematic cartography. So varied were these developments that it is difficult to be comprehensive, but a few themes stand out" [29, p.14]. Many new graphical forms were devised and applied to new fields of investigation, notably in the social domain, as visual representations for interpreting complex data and events became established. The *anti-cyclonic* pattern of winds around low-pressure areas and clockwise rotations around high-pressure zones was possibly the most famous non-statistical graphical finding. Reports and statistical atlases with increasingly diversified visual representations were produced between 1880 and 1900, following each consecutive decennial census.

6. *1900-1950, the modern dark ages*:
   "If the late 1800s were the golden age of statistical graphics and thematic cartography, the early 1900s could be called the modern dark ages of visualization" [29, p.20]. Few graphical advancements, and by the mid-1930s, the late-nineteenth-century

excitement for visualization had been overtaken by the emergence of quantification and formal, generally statistical, models in the social sciences. Numbers, parameter estimates, and standard errors were all exact. Nonetheless, it is reasonable to regard this as a period of essential dormancy, application, and popularization than as one of innovation. Statistical graphics were popular during this period. Furthermore, maybe for the first time, graphical approaches were critical in several new ideas, discoveries, and hypotheses in astronomy, physics, biology, and other disciplines. Graphic innovation was also awaiting new concepts and technology: the creation of current statistical methodology equipment and the arrival of computer capacity and display devices to enable the next wave of advancements in data visualization.

7. *1950–1975, re-birth of data visualization*:
   In the mid-1960s, data visualization awoke from its slumber. Significant crossings and partnerships would commence before the end of this period: computer science research would join forces with advances in data processing, display, and input technologies (pen plotters, graphic terminals, digitizer tablets, the mouse). These advancements created new paradigms, languages, and software packages for expressing statistical ideas and implementing data visualization. As a result, new visualization methods and techniques are exploding. By the end of this century, the first modern Geographic Information System (GIS) and interactive systems for 2D and 3D statistical visualizations had developed. These establish objectives for future development and expansion.

8. *1975–present, high-D, interactive, and dynamic data visualization*:
   "During the last quarter of the 20th-century, data visualization has blossomed into a mature, vibrant, and multi-disciplinary research area, as may be seen in this Handbook, and software tools for a wide range of visualization methods and data types are available for every desktop computer" [29, p.24]. However, it is not easy to present a concise summary of the most recent innovations in data visualization because they are so assorted, have happened rapidly, and span many fields. Many breakthroughs in statistical graphics from the early 1970s to the mid-1980s included static graphs for multidimensional quantitative data, meant to allow the analyst to see relationships in progressively higher dimensions. The introduction of dynamic graphic technologies, which allow immediate and direct modification of graphical objects and related statistical features, has provided enormous potential for current progress in data visualization. These concepts were brought together in the 1990s to create more generic systems for dynamic, interactive visuals integrated with data manipulation and analysis incoherent and expandable computer environments. When all of these elements were combined, they were more potent and impactful than the sum of their parts. Since then, interactive visualizations have taken the limelight as they tend to lead to discovery and do a better job than static data tools [82].

As we have seen, most data visualization breakthroughs emerged from specific, often

practical purposes. The development of visual approaches depended on contemporaneous improvements in technology (big data, digital distribution) [63], [49], data gathering [51], [52], nature and environment [62], and statistical theory [29], [39], [87].

Tufte [80] is one of the most significant scholars in the history of experimental graphics studies. He had a significant effect on the history of data visualization, and it is unquestionably one of today's most important sources for theories in information visualization research. His book examines graphical techniques during the last two centuries since Playfair[1], covering excellent and awful graphics in an attempt to determine the processes that lead to poor graphical displays; also provides a language for discussing graphics and a practical theory of data graphics. When applied to most visual presentations of quantitative information, the approach leads to design revisions and enhancements, proposes why certain graphics may be better than others and develops new forms of graphics.

According to Tufte, the principles for good graphical data displays are [80, p.13]:

- show the data.
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else.
- avoid distorting what the data have to say.
- present many numbers in a small space.
- make large data sets coherent.
- encourage the eye to compare different pieces of data.
- reveal the data at several levels of detail, from a broad overview to the fine structure.
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration.
- be closely integrated with the statistical and verbal descriptions of a data set.

In addition to providing clear principles for good graphical data visualization, Tufte also sought to reaffirm the roots of sign language in practicality while making a quantum leap in its applicability [81]. He aims to reveal a universe of dense, multivariate, and complex information, i.e., an avalanche of data tracked across space, time, and multiple variations using visually transparent means. Moreover, when we examine a map, a photo, a sketch, a graphic, or any other product of deliberate design, why do we need to know the rules and devices that govern good visual communication?

Tufte's work is a pleasant introduction to interpreting the code so that we read what the images and other graphics describe and what they assert, defend, or argue, and the

[1] http://www2.psych.utoronto.ca/users/spence/Spence2004.pdf

extent to which their claims are truthful or reasonable, or practical. For Tufte, "visual displays rich with data are not only an appropriate and proper complement to human capabilities, but also such designs are frequently optimal" [81, p.50], but it is crucial to perform an optimal conversion of data into visual elements.

In his 2019 book, Wilke [86] takes the reader through the critical concepts of presenting data accurately through figures. The author analyzes the aesthetics familiar to all figures: scale, axes, color, and data types to justify which figures are well designed, as well as those that are incorrect or aesthetically ugly, and explains the reasons for judging them good bad, incorrect, or ugly. Wilke builds on the foundation of including information familiar to the reader, such as Cartesian coordinates, to explore the less familiar such as nonlinear scales and curved axes; he also uses easy-to-understand data to demonstrate the comparative strengths and weaknesses of how each figure tells the story of the data.

According to Wilke, best practices for proper data visualization are:

- *The correct chart for the correct data*: The author provides a quick overview of the various graphs and charts commonly used to visualize different data types. For example, Figure 2.2 shows the perfect scatterplots to represent the archetypal visualization when we want to show one quantitative variable concerning another. If we have three quantitative variables, we can assign one to the size of the point, making a bubble chart, a version of the scatterplot.

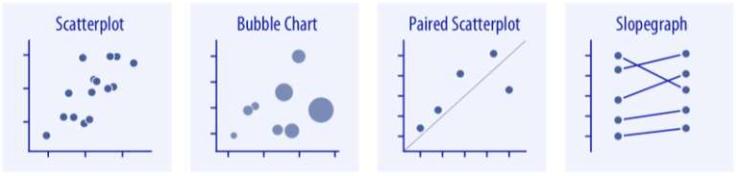


Figure 2.2: x-y relationships plots. Source: [86, p.41].

- *Putting coordinates to good use*: With a few examples, the author tries to keep us from making typical mistakes with something as simple as Cartesian coordinates, where we tend to mislabel a logarithmic scale, apply the wrong transformation, or lack the practice for a polar coordinate plot.
- *Handling superposition*: We commonly face handling overlapping points; this is why the author offers several solutions such as partial transparency and jittering, both handled with careful color selection and shading; binning data into rectangles and hexagons, using contour lines with shading when dealing with high-density points.
- *Data-to-ink ratio management*: Tufte [80] had already warned that optimizing ink to non-data visual elements is very important, so Wilke presents more examples to demonstrate further on this topic, like Figure 2.3, where there is too much ink for

the grids, legend, and frame. However, on the other hand, in Figure 2.4 we have too little, no grids, no frame, which confuses the plot.

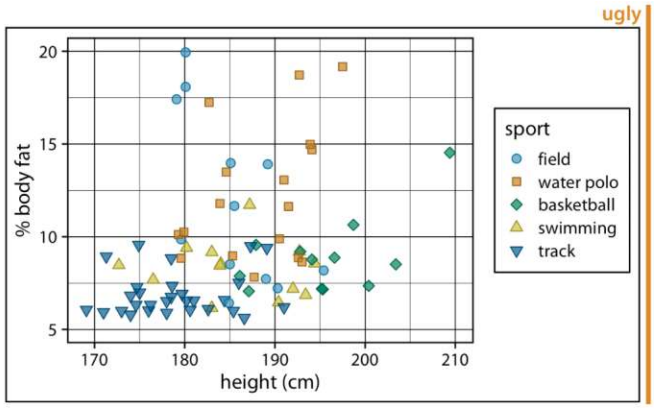


Figure 2.3: A Plot that require a balancing act on data-ink ratio. Source: [86, p.278].

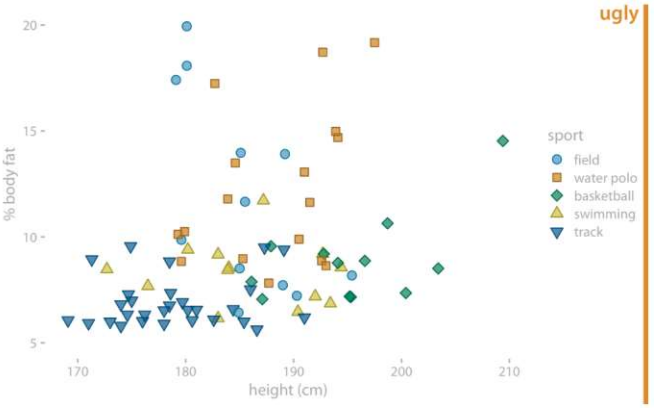


Figure 2.4: A Plot that require a balancing act on data-ink ratio. Source: [86, p.280].

- *3D plotting is not the solution*: The author points out that we are all impressed by 3D, although this is questionable; he presents several examples to determine if a 3D plot is a correct option or conveys the message that the data is carrying. For Wilke, 3D "is unequivocally bad and should be erased from the visual vocabulary of data scientists" [86, p.305].
- *Plots serve to tell a story*: The author borrows from the storytelling patterns of writers to tell a story and thus applies it to data science. According to him, we follow our instincts to frame how we engage with the audience and narrate the challenge and findings. Moreover, many times, we do not do it articulately.
- *The good, the bad, the ugly, and the wrong*: The author helps us identify when a plot lacks proper design, incorrect formatting, or a faulty setup.

Another book that provides the theoretical approach to effective data visualization design is Kirk's [42]; the author relates that we live in a data-rich era. Knowing how to distill data into compelling graphics is a powerful skill. A well-designed data visualization can communicate information and inspire thought and discussion. The accessibility of easy-to-use tools allows almost anyone to become a visualization designer in minutes. However, access to the tools is insufficient because a poorly designed visualization might be disregarded or incorrectly convey incorrect information.

Kirk's book is broken into four parts, which are as follows: Part A, *Fundamentals*, presents Kirk's systematic approach to producing visualizations; it also provides the background necessary to become familiar with data visualization. In Part B, *Hidden Thinking*, the author describes the process and preparation work involved in producing a visualization. Kirk says this includes establishing the who, what, why, and where. This method will work well for designers already immersed in their data. This section also describes data collection for those who have an idea but not the data. Part C, *Developing the Design Solution*, describes the production phase of the visualization design. The author leaves us with the following reflection: "What charts can you actually make and how efficiently can you create them?" [42, p.263]; Kirk also addresses the fundamental concepts of data visualization through visual variables and delves into more artistic concepts such as editorial prominence and functional harmony. Finally, Part D, *Developing Your Capabilities*, contains a single final chapter on visualization literacy. He reviews how to deconstruct and read a visualization from both the reader's and developer's perspective and how to assess and develop skills as a designer.

Kirk recommends three design principles of good data visualization:

1. *Good data visualization is trustworthy*
   Kirk maintains an essential distinction between trust and truth, as the latter is a must. In data visualization, there is rarely a singular view of truth. The half-full glass is also half empty. Both views are true, and it is not easy to choose. In these cases, the author relates that the final solution is potentially composed of many well-informed, well-intentioned, and legitimate choices, no doubt. However, they will equally reflect a subjective perspective. All projects represent the result of a unique path of thought.

2. *Good data visualization is accessible*
   This principle helps to determine the best way to facilitate the viewers' understanding process; for the author, this is the essence of this principle: "a viewer should experience minimum friction between the *act* of understanding (effort) and the *achieving* of understanding (reward)" [42, p.52].

3. *Good data visualization is elegant*
   "Elegant design is about seeking to achieve a visual quality that will attract your audience and sustain that sentiment throughout the experience, far beyond just the initial moments of engagement" [42, p.56]. According to Kirk, any decision

to achieve "elegance" should not undermine the achievement of reliability and accessibility in design; the visual "look" of the work will be the first thing viewers encounter before they experience the consequences of their thinking based on other principles. Therefore, optimizing the perceived appeal of the work will have a significant impact on viewers.

2.1.1 Data transformation into visual components

Munzner, in 2014 stated that:

> "Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively. The design space of possible visualization idioms is huge, and includes the considerations of both how to create and how to interact with visual representations" [51, p.1]

Munzner [51] provides a framework for understanding the fundamental parts of visualization by integrating design decisions with visualization idioms. According to Munzner, the contemporary era is defined by the promise of improved decision-making enabled by more access to data than ever before. Moreover, when humans have well-defined data issues, they can apply purely computational approaches from domains like statistics and machine learning.

Figure 2.5: The three-part analytical framework for a visualization example: why the task is being performed, what data is displayed in the views, and how the visualization idiom is produced in design decisions. Source: [51, p.17].

The Figure 2.5 depicts the author's high-level methodology for examining visualization use in terms of three questions:

1. *What data does the user see?*
   The underlying structure of the four fundamental dataset types is depicted in detail in Figure 2.6. For either the simple flat case or the more sophisticated multidimensional example, cells in tables are indexed by items and attributes.

Tables, networks, fields, and geometries are the four main dataset kinds; additional potential groupings of information include clusters, sets, and lists. These datasets are constructed using combinations of the five data types: items, attributes, connections, locations, and grids. The whole dataset for any of these kinds may be provided instantly in a static file, or it could be dynamic data processed progressively in the form of a stream. An attribute can be categorical or ordered, divided into ordinal and quantitative. The attribute ordering direction might be sequential, divergent, or cyclic.

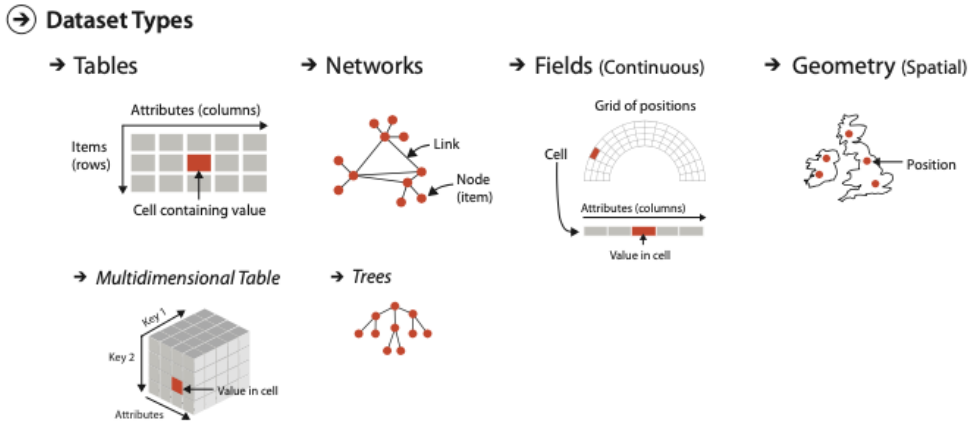


Figure 2.6: The structure of the four fundamental dataset types in detail. Source: [51, p.25].

The list mentioned above of fundamental types provides a starting point for expressing what element of an analysis instance is related to data: the data *abstraction*.

2. *Why does the user want to use a visualization tool?*
Munzner's design in Figure 2.7 is broken down into steps and focuses on why a visualization tool is utilized. At the highest level, the framework differentiates between two possible goals for people who want to analyze data using a visualization tool; the most common use case for visualization is for the user to consume information that has already been generated as data stored in a format amenable to computation. Within that case, the framework distinguishes three additional distinctions: whether the goal is to present something that the user already understands to a third party, to discover something new or analyze information that is not already completely understood, or for users to enjoy a visualization to indulge their casual interests in a topic.

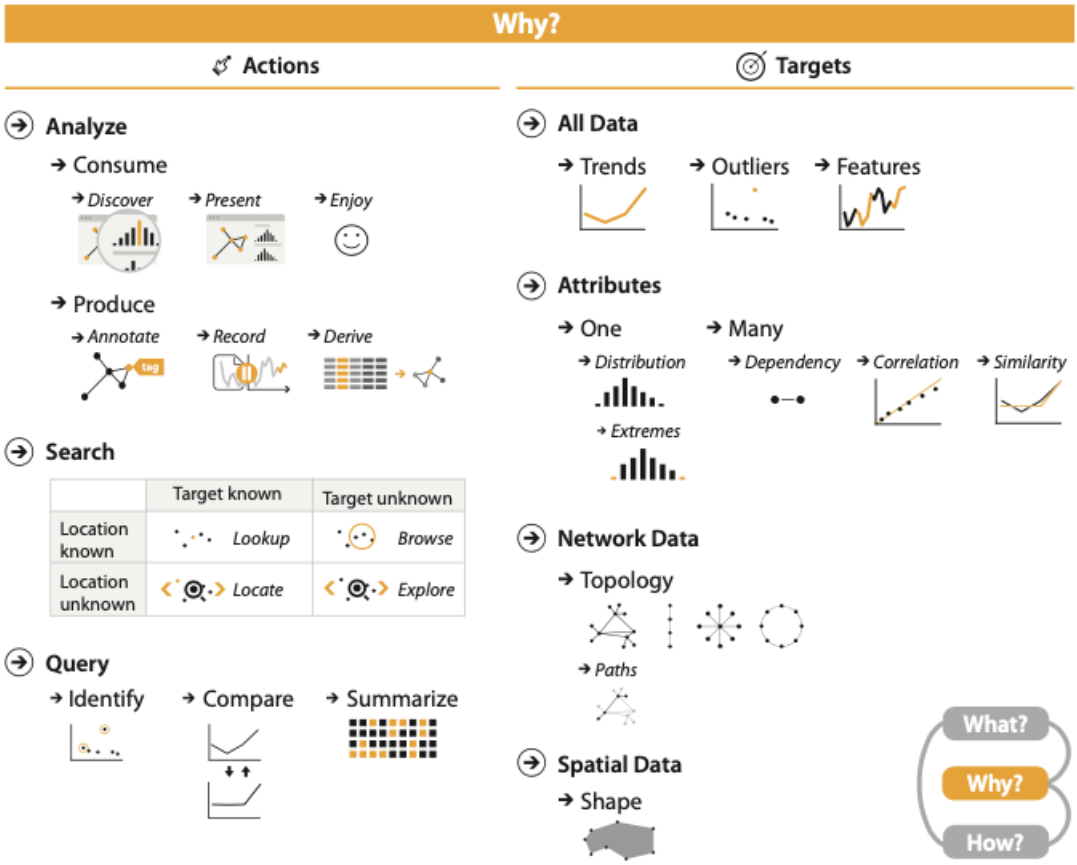


Figure 2.7: *Why* individuals are utilizing visualization in terms of activities and goals. Source: [51, p.42].

At the intermediate level, the search can be classified based on whether the target's identity and location are known or not: both are known with lookup, the target is known, but its location is not for locate, the location is known, but the target is not for browse, and neither the target nor the location is known for exploring.

Queries can have three scopes at the low level: identify a single target, compare many targets, and summarize all targets. The development of these three correlates to a rise in the number of search targets taken into account: one, some, or all. In other words, identify refers to a single target, compare refers to many targets, and summarize refers to the whole collection of possible targets.

3. *In terms of design decisions, how are visual encoding and interaction idioms (a different way to constructing and manipulating visual representations) constructed?* The third component of the analytical example trio is how a visualization idiom may be built from a collection of design options. There are five options for organizing data spatially within the family of encoding data into a view: express values,

separate, sort and align areas, and use provided spatial data. This family also offers information on mapping data using all of the nonspatial visual channels, such as color, size, angle, shape, and many more. Modifying any aspect of the view, picking objects from within the view, and traversing to change the viewpoint are all options for the manipulation family. Finally, the family of ways to facet data across views includes options for juxtaposing and coordinating numerous views, partitioning data between views, and superimposing layers on top of one other.

In addition to these questions that comprise the framework, Munzner presents four degrees of *validation* that are important when beginning the design process. She feels this is significant since most ideas are useless because vast expansive visualization design space is. The author illustrates four-layered design stages in Figure 2.8: domain situation, task and data abstraction, visual encoding and interaction idiom, and algorithm. The task and data abstraction levels handle why and what questions, whereas the idiom level covers how.

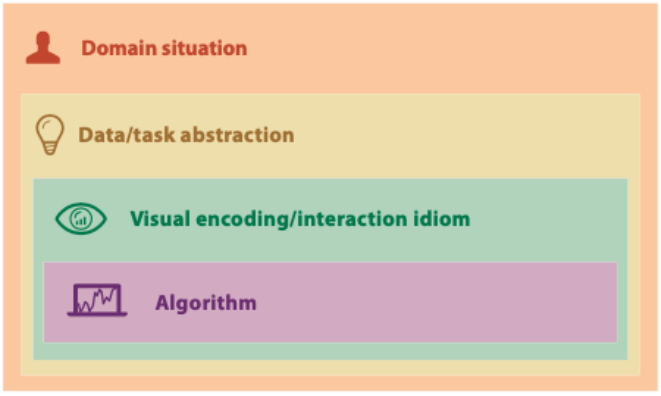


Figure 2.8: The four-layered stages of visualization design. Source: [51, p.68].

For Munzner, the scenario level is at the top; we evaluate the characteristics of a particular application domain for visualization. The next *how* level is the creation of idioms that determine the approach to visual encoding and interaction, which follows the *what–why* abstraction level, a level where we convert those domain-specific issues and data into forms independent of the domain. Finally, the next stage is the creation of algorithms that computationally instantiate those idioms. These levels are nested; one level above is fed into the one below. The difficulty with this layering is that selecting the incorrect block at an upstream level automatically cascades to all downstream levels. The author proves that if we make a poor abstraction decision, even great idiom and algorithm choices will not result in a visualization system that solves the desired goal [51].

Shneiderman [73] also notes that a helpful starting point for advanced graphical user interface design is the *Visual Information-Seeking Mantra*: overview first, zoom and

filter, then details on demand. In an attempt to understand the rich and varied set of information visualizations proposed above, Shneiderman offers in his article a taxonomy of tasks by data type with seven types of data (one-dimensional, two-dimensional, and three-dimensional data, temporal and multidimensional data, and tree and network data); data on which a task is applied at a high level of abstraction.

The seven tasks proposed by Shneiderman are [73, p.337]:

- *Overview*: Gain an overview of the entire collection.
- *Zoom*: Zoom in on items of interest.
- *Filter*: filter out uninteresting items.
- *Details-on-demand*: Select an item or group and get details when needed.
- *Relate*: View relations hips among items.
- *History*: Keep a history of actions to support undo, replay, and progressive refinement.
- *Extract*: Allow extraction of sub-collections and of the query parameters.

Another example of data transformation was introduced by Wilkinson [87], in which the author proposes a system with seven orthogonal classes in his book. The term orthogonal directs that each class has one or more methods (functions) as members, and all tuples in the seven-fold product of these function sets generate valid graphs. Figure 2.9 depicts a data flow diagram with the seven grammar of graphics classes.

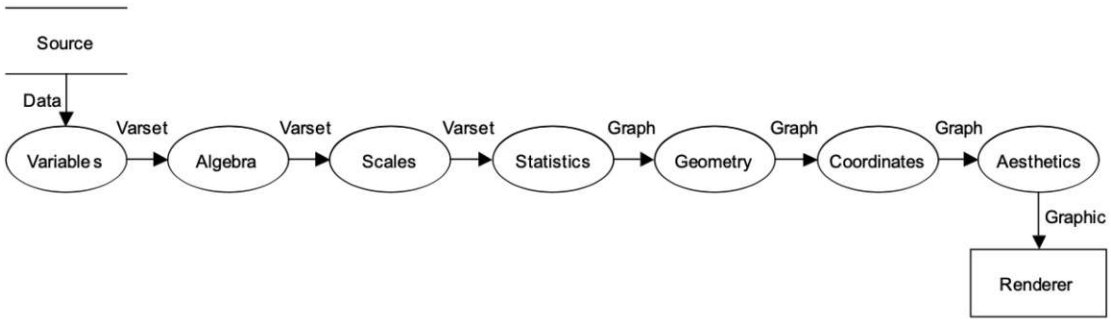


Figure 2.9: The grammar of graphics data flow. Source: [87, p.24].

This *data flow* is a chain that describes the series of mappings required to generate a statistical graph from data collection. The first class (Variables) translates data to a varset object (a set of variables). The following two classes (Algebra and Scales) are varset transformations. Next, Statistics builds a statistical graph from a varset (a statistical

summary). Then, the class Geometry converts a statistical graph into a geometric graph. The following step (Coordinates) embeds a graph in coordinate space. Finally, Aesthetics transforms a graph into a visible or perceptible display known as a graphic. The data flow architecture indicates that the subtasks required to generate a graphic from data must be completed in the sequence provided; if not, it can result in useless graphics.

The author's second grammar of graphics argument is that this approach explains what we do when we create statistical visuals. For Wilkinson, it is more than a taxonomy; it is a computer system built on the underlying mathematics of describing statistical functions of data.

2.2 Human perception

This section discusses everything about human perception and the human visual system (HVS). We will first explain the visualization stages, which show the data visualization process perceived by humans. Afterward, the *Gestalt* theory will be presented. That topic will be about how humans perceive different images like forms or objects and how their presence can change how humans recognize them.

2.2.1 Human Visual System

The human visual system (HVS) is a sophisticated mechanism that is still not fully understood. Furthermore, the HVS's visual qualities are not intuitive [31]. We want to introduce how the visual system is organized and functions to produce visual perception. There have been a lot of discoveries about how our visual system is organized. They stretch from the structural basis of the visual pigments that capture light to the neural basis of higher visual function [79].

On the one hand, there's the human visual system, which includes a flexible pattern finder and an adaptable decision-making process. The computer and the World Wide Web, on the other hand, have the enormous computing power and immense information resources. Interactive visualizations are becoming a more common means of communication between the two. Improving these interfaces can significantly boost the whole system's performance [83].

2.2.2 Visualization

Visualization is a critical topic in human perception since it exploits the visual system to present data naturally, quickly, and language-independent. The amount of brain capacity dedicated to processing visual data vastly outnumbers that dedicated to processing other human senses [40]. Therefore, visualization is an essential tool for understanding and getting insights hidden in data and helps to make decisions[35]. In such large and complex datasets as "sanity checks" by providing evidence that the underlying data are reasonably free of flaws such as missing values or excessive noise that might affect later analysis [18]. The use of visualization is broad and applies in many important fields like architecture,

engineering, and construction (AEC) [9], concept design [19], mathematics [57][50], or even healthcare [13]. It is also used for understanding climate outlooks [32], analysing business data [65], exploring time-dependent data [1] and is a key for industrie 4.0 and industrial internet [56].

In recent years, the use of visualization techniques for exploratory data analysis of complex and multidimensional datasets has increased. Scientists and researchers in various fields rely on that technique to familiarize themselves with their complex data spaces and to generate new insights [55]. The human perceptual and cognitive systems are essential in the process of visualization. Cognitive activities such as forming high-level analysis goals, planning actions, and evaluating results effectively, are required for data exploration [59].

In the early stages of data analysis of a given dataset, scientists are often interested in exploring possible associations that may exist within it. Therefore, they generate and inspect the whole set of scatterplots obtained from all possible pairs of dimensions. However, observing these combinations becomes exceptionally time-consuming, if not impossible, when the number of dimensions grows to tens, hundreds, or even thousands of variables [55].

It is also important to design data visualizations effectively to allow viewers to use their powerful visual system to understand patterns in the data because ineffectively designed visualizations can cause confusion, misunderstanding, or even distrust [28].

Human perception is one of the four primary stages of data visualization. These four stages are combined in several feedback loops, as illustrated in Figure 2.10 [83].

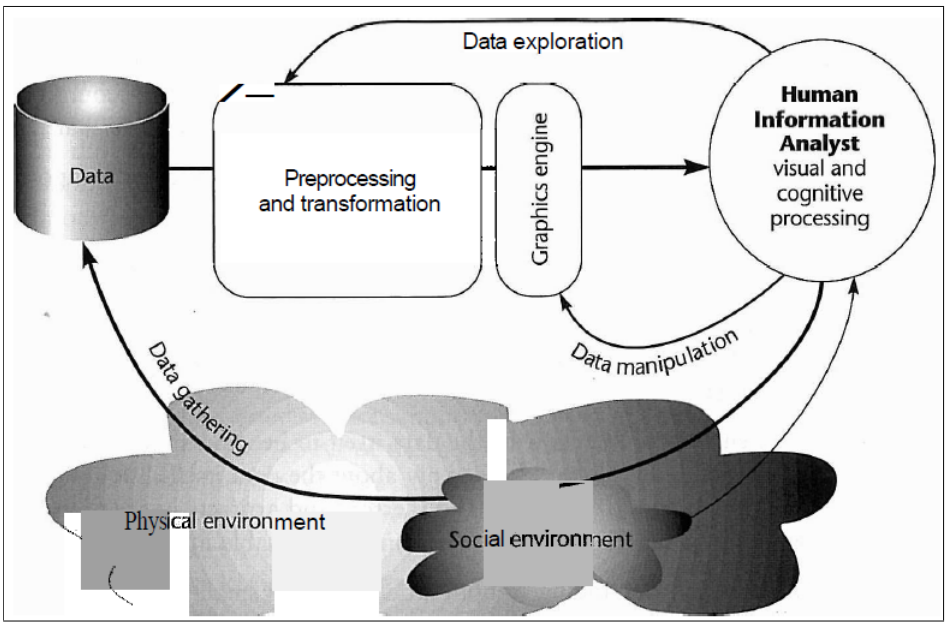


Figure 2.10: A schematic diagram of the visualization process. Source: [83, p.4].

The process of data visualization consists of:

1. The collection and storage of data,
2. The process of transforming the data into something humans can understand,
3. The hardware and the graphics algorithms that are used to process and display an image, and
4. The perceiver (human perceptual and cognitive system).

The most extended feedback loop is the one where data is gathered. Data seekers may choose to collect more data to follow up on an exciting lead. The other loop controls the computational preprocessing that occurs before the visualization. There, the analyst can decide if the data is subjected to an inevitable transformation before visualization and persuade it to give up its meaning. Finally, the visualization process itself could be highly interactive. For example, the scientist may fly to a different vantage point to better understand the emerging structures for a 3D data visualization.

Another example could be that a computer mouse may be used interactively to select the most exciting parameter ranges. The physical and social environments are involved in the data-gathering loop. A physical environment is the source of data, and the social environment determines what is collected and how it is interpreted in different ways [83].

2.2.3 Gestalt theory

The *Gestalt* theory is the well-known perception theory that grew out in the field of psychology but has influenced research from a multitude of disciplines. Those disciplines include for example the medical field [44] or human-computer interaction (HCI) [27]. The theory is based on the following statement: There are wholes, of which their individual elements do not determine the behavior, but where the intrinsic nature of the whole determines the part-processes[85]. According to the *Gestalt* theory, the human brain builds information through the sensory canals, perception, and/or memory. It also says that the perceptive activity is subordinated to an essential factor of *Prägnanz* (good shape). If an object expresses any characteristic in a sufficiently strong way to be obvious, to be imposed, and to be easily evocative, it is *Prägnanz*. The laws of the *Gestalt* theory explain the structural and functional principles of the perceptive field and establish the shape as the constituent element of an image that can be perceived [2]. The Gestalt approach to form perception, developed in Germany in the early twentieth century, is beneficial for comprehending how we perceive groups of objects, or even parts of objects, to form integral wholes [75]. According to Sternberg et al. [75], the *Gestalt* principles of visual perception are as follows:

1. Figure/Background: When perceiving a visual field, characteristics such as size, form, color, and position set the objects /figure apart from the background and seem prominent. Other aspects of the field recede into the background.

2. Proximity: when elements are placed together, they tend to be perceived as a group, even when they are not similar. This law will be very interesting when we analyze the results of our user study about the clusters.

3. Similarity: when elements have similar or equal characteristics, such as color, shape, and texture, they tend to be grouped in sets. Normally, the similarity is not overlapped by proximity.

4. Continuity: when elements look like they build a pattern or a flow in a common direction, they are easier perceived by humans. That means that continuity of direction and continuous ligaments between elements are more accessible to be received than elements that present abrupt modifications in their direction. This law will also play an essential role in the regression analysis of this work.

5. Closure: elements are ordered in a certain way to form an almost closed outline or incomplete shape to become a unity. Human perception realizes complete shapes. Thus, Humans perceive the whole by filling in the missing data.

6. Symmetry: when elements are presented with symmetry regularity and are without textures, they are perceived more efficiently as a whole.

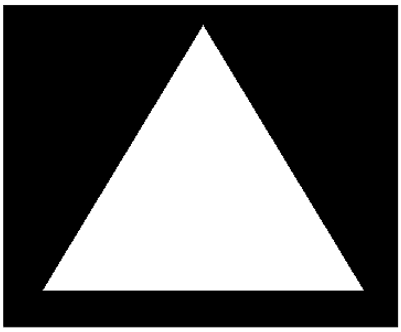


Figure 2.11: Example of the law of Figure/Background. Here, Humans tend to perceive a white triangle on a black background. But it is also possible that it's a black figure with a white background.

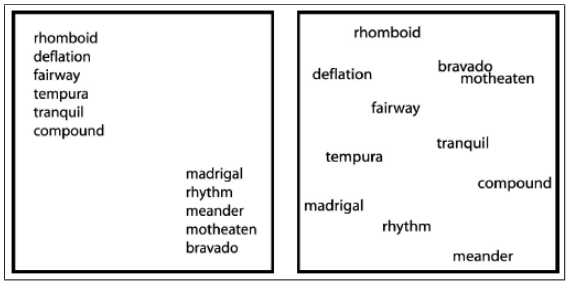


Figure 2.12: Example of the law of proximity. Items located close together seem part of a group (left), while items not close together (right) are not. Source: [34, p.6].

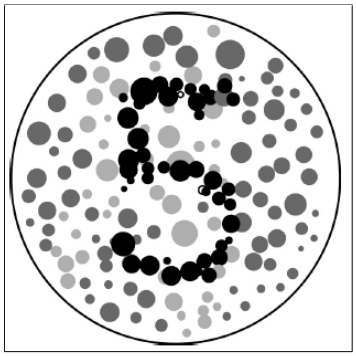


Figure 2.13: Example of the law of similarity. Items that look similar seem to belong together. Source: [34, p.9].

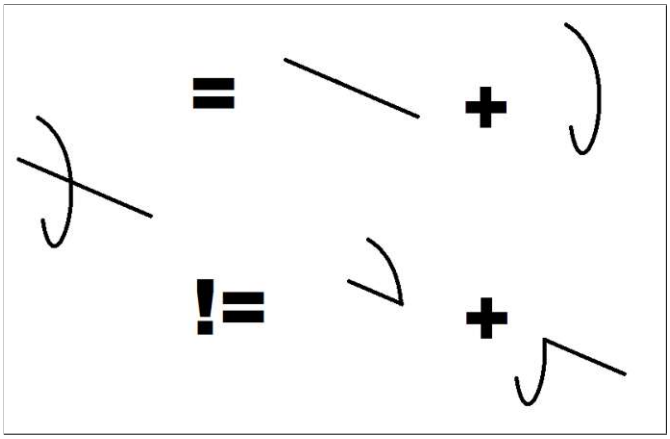


Figure 2.14: Example of the law of continuity. Rather than disrupted or discontinuous forms, we tend to perceive smoothly flowing or continuous ones.

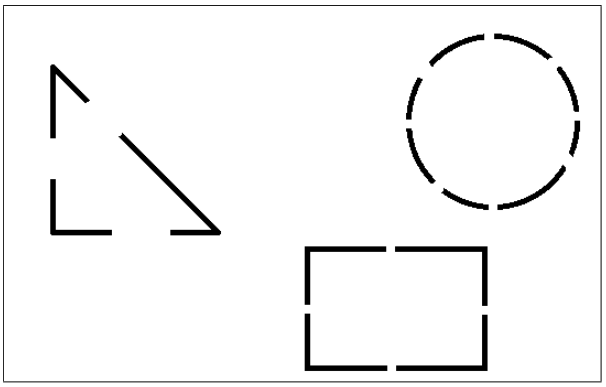


Figure 2.15: Example of the law of closure. Humans tend to close gaps in forms. So in this image, three figures are perceived instead of fifteen different.

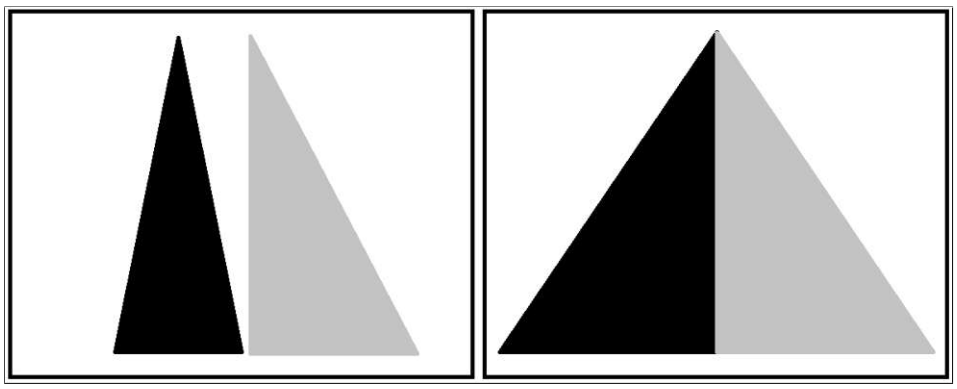


Figure 2.16: Example of the law of symmetry. Humans perceive elements as a whole when presented with symmetry.

2.2.4 Model of visual human perception

All observations related to the human perception system can be applied to obtain the best-produced images for visualization purposes. However, other concepts must be considered to get new information, emphasize others, or intentionally induce the user to perceive some information from the input data. It is necessary to comprehend all phases of the perceptible processing system to do so. Visual perception principles can be used in each of these phases. Much research simplifies human perception system models by discarding some of its phases. A simplified model of an information processing system based on visual human perception is frequently used as the first step in a more detailed investigation. Understanding the involved processes requires a broad examination of the human visual system [2].

Ware [83] describes the human visual information processing in three stages:

1. Parallel processing to extract properties of a low level from the visual scene

2. Pattern perception in the resulting image;

3. Sequential goal-directed processing.

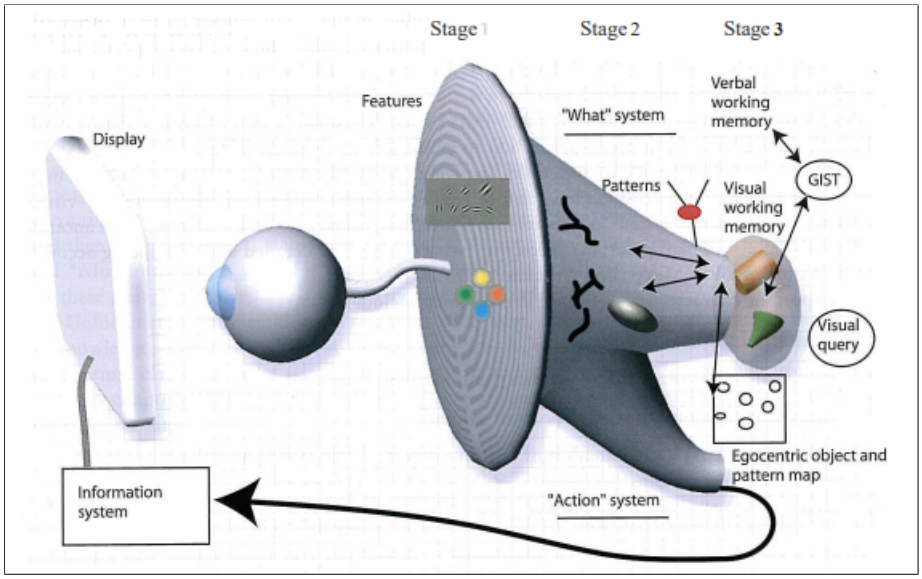


Figure 2.17: Visualization of a three-stage model of human visual information processing. Source: [83, p.21].

Parallel processing

Initially, billions of neurons work in parallel to extract characteristics from parts of the visual image in analysis; for this, dedicated neurons extract specific information from the input data, such as contours, orientation, color, texture, and movement patterns. This phase determines what will be given attention; as a result, the data in this phase is globally transitory [83]. The way the human visual system analyzes images has been the subject of intense research for several years. One of the first and most significant findings was identifying a set of visual properties that the low-level visual system detects precisely and quickly. The pre-attentive was the first to name this property, and it refers to the last moment when our attention was focused [2]. The pre-attentive term can still be used in visualization, and it translates the notion of speed and ease with which humans can identify specific properties from visualized images. Color, shape, movement, and spatial location are the four basic categories of characteristics that are processed in a pre-attentive manner. Furthermore, any change in the pre-attentive characteristics of one element in relation to the others could change the focus of attention within each visual field [36].

Pattern perception

Pattern perception originates with the early work of the *Gestalt* psychologists [7][43]. In this stage, active processes analyze the visual field in regions and simple patterns, such as continuous contours, regions with similar colors, and regions with the same texture [83]. The principle of continuity is very relevant in our user studies. The importance of movement patterns cannot be overstated; however, in Scientific Visualization, the use

of movement as information is often overlooked. Pattern recognition in human visual processing is highly flexible, and it is influenced by the information gathered during the first parallel processing stage. The second stage involves slower processing and long-term memory for object recognition. The attention mechanism is of the top-down and bottom-up types and is visually guided by movements through different paths, emphasizing the prominent aspects [83].

Sequential processing

On a higher level of perception, images are stored in the visual memory due to active attention demands. This memory will aid in responding to visual inquiries. When the human system experiences external visualization, it creates a series of visual searches that will be answered using visual search strategies. At this level, all of the data stored in memory for a set period is used to create patterns, use available ones, and respond to visual searches [2]. Using a road map to find a specific route, for example, the visual inquiry will look for red outlines (which usually represent important roads) between two visual symbols (which represent the desired cities) [83]. So, in terms of visual perception, past experience is another factor to consider; in the case of an association, it is essential for the perception process because we can only comprehend what we are already aware of. Our perception shifts every time we learn something new [2]. As a result, visual perception results from a complex interaction between external information acquired by the visual system and previously acquired internal knowledge [64].

2.3 Scatterplots and the overplotting issue

In this section, we discuss scatterplots and their most apparent difficulty, *overplotting*; we introduced techniques that studied the problems of visual clutter in scatterplots in two separate investigations: cluster and regression analysis.

Cleveland [15] defines scatterplots as an appropriate exploratory tool for offering a first glance at bivariate data to examine how they are distributed throughout the plane, and it encodes the data items using two axes and positions; for him, scatterplots are a potent tool for analyzing and visualizing data [17]. The idiom of scatterplots, according to Munzner, "encodes two quantitative value variables using both the vertical and horizontal spatial position channels, and the mark type is necessarily a point" [51, p.146].

Friendly and Denis [30] mention that scatterplots are a perfect sandbox for early information visualization and perceptual psychology study due to their simplicity and versatility. They are adequate for the abstract tasks of providing overviews and characterizing distributions and specifically for finding outliers and extreme values. These diagrams are also highly effective for the abstract task of judging the correlation between two attributes, and it is possible to examine and demonstrate the relationship between two qualities, clusters of points, and outliers [41]. Variable relationships can be defined in various ways: positive or negative, strong or weak, linear or nonlinear; with this visual

encoding, that task corresponds to the easy perceptual judgment of noticing whether the points form a line along the diagonal. The stronger the correlation, the closer the points fall along a perfect diagonal line; positive correlation is an upward slope, and negative is downward. As we can see, Figure 2.18 shows a highly negatively correlated dataset.

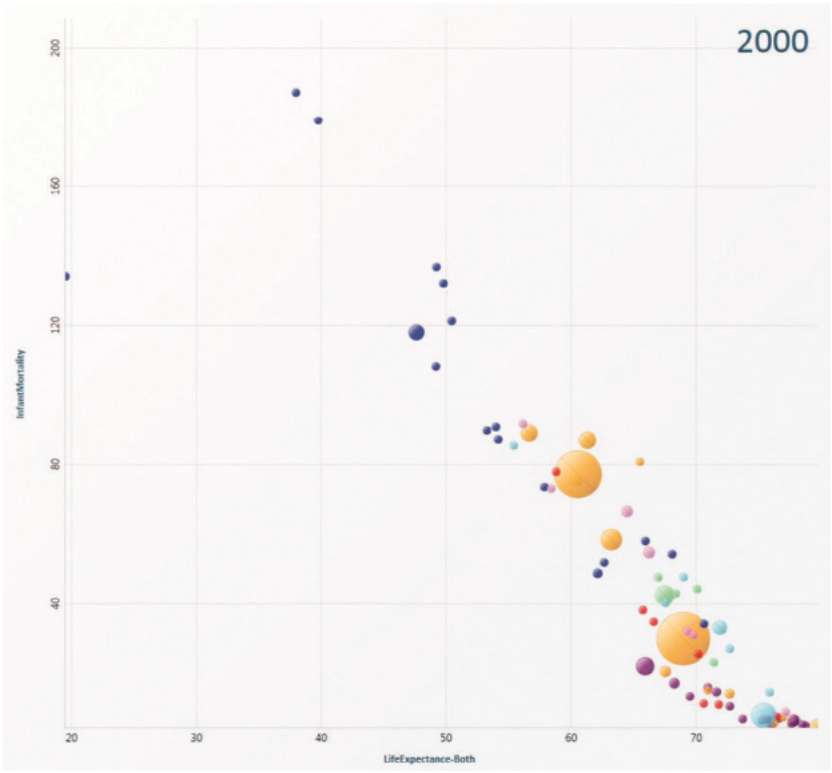


Figure 2.18: Example of a scatterplot. Each point mark represents a nation, with the key quantitative aspects of life expectancy and infant mortality encoded in horizontal and vertical geographical positions. Color is used for qualitative nation attributes, and size is utilized for quantitative population attributes. Source: [51, p.147].

There is no doubt that scatterplots are a powerful and widely used tool for visual data exploration [21]. However, when a large quantity of data is utilized, scatterplots have a high degree of overlap, making the correct density of data values challenging to see [41]. From a user perspective, plots also appear to be quickly overloaded when a large amount of data is available. This issue is called *overplotting* when more than one observation (point) has the same or very similar values, so data points overlap to a degree where the user has trouble seeing relationships between points and variables, making the information of the graph lost or misleading [25].

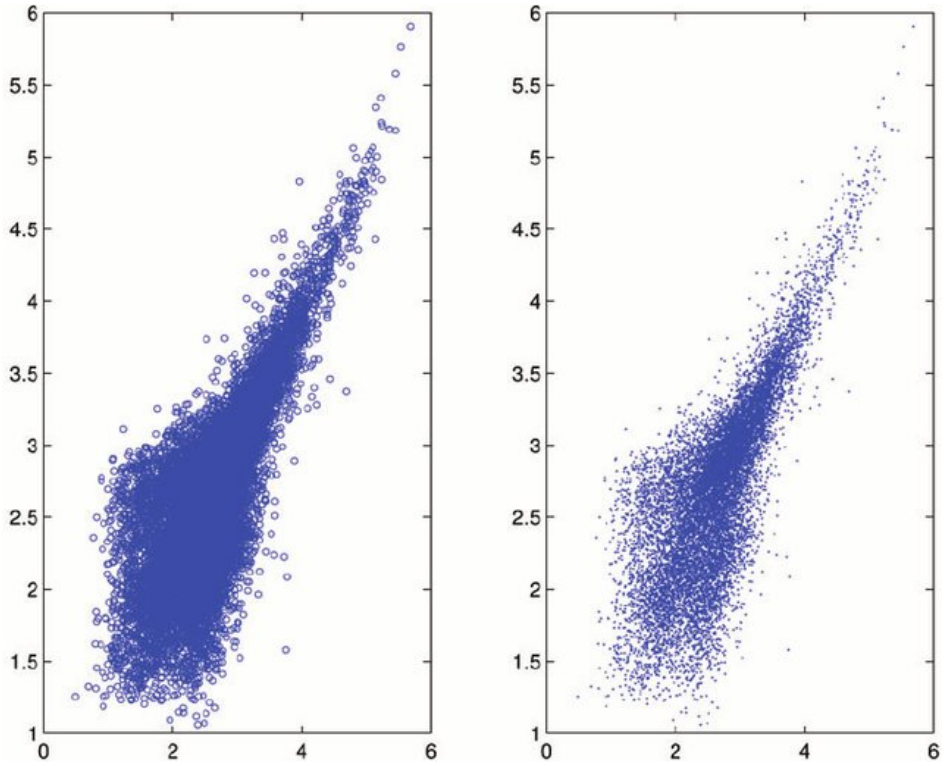


Figure 2.19: Example of overplotting. Left: large points, right: small points. Source: [22, p.624].

Figure 2.19 is an example of overplotting that could be misleading. Both scatterplots have the same data, but the left plot uses more significant symbols than the right one. As a result, the one on the right side, the cluster in the middle, is more visible.

Overplotting has been a longstanding problem in information visualization [25]. Given this problem, where a significant portion of these data values may be clogged, certain studies have tried to analyze visual clutter with different techniques. For instance, Dang et al. [20] proposed a new approach for visualizing and interacting with datasets that maintains density information by *stacking overlapping cases*. Depending on the type of plot, the overlapping data can be points, lines, or other geometric components. Chen et al. [12] developed a visual exploration system that supports visual inspection and quantitative analysis from different perspectives to improve the density contrast of scatterplot data points. The authors presented a new visual abstraction scheme that utilizes a *hierarchical multi-class sampling technique* to show a feature-preserving simplification.

In other investigations, Nguyen et al. [53] revealed that the Multiple-Scatterplot technique is preferred for exploring multivariate data since it is faster and produces higher accuracy. Few and Edge [23] tried to tackle the overplotting problem by changing the shape of the data points; Smart and Szafir [74] measured how the interplay of shape, size, and color encodings influence our ability to distinguish data values along each channel and

measured the symmetry of these effects; Keim et al. [41] proposed the generalized scatterplot approach, which permits an overlap-free depiction of enormous datasets to fit the entire display. The primary concept is for the analyst to optimize the degree of overlap and distortion to provide the most nuanced possible perspective. Other authors [67], [45], [84] tried helping designers make design choices for scatterplot visualizations.

2.3.1 Regression analysis

Regression analysis can answer critical issues that help people and businesses make better decisions [3]. For Arkes, a regression is "an equation that represents how a set of factors explains an outcome and how the outcome moves with each factor" [3, p.14]. Regression analysis is used in different fields like in sports to predict cardiorespiratory status or future outcomes [54], agriculture to predict crop yield [72], and in many more. Some authors presented a scatterplot-based visualization tool for regression error analysis [77].

Even though correlation and regression are closely related, they are distinct concepts. Correlation may be described as the degree of association between two variables, whereas regression expresses the degree of association between two variables. In general, we might argue that studying interdependence leads to the study of correlations, whereas studying dependence leads to regression theory. We are more interested in determining the strength of the linear relationship than in prediction when the x variable is a random covariate to the y variable; thus x and y vary together (continuous variables), and the sample correlation coefficient, rxy (r), is the statistics used for this purpose [4]. The example shown in Figure 2.20 shows different correlation coefficient values such as a perfect positive (a), positive (c and e), perfect negative (d), negative (f) correlation, or no correlation (b).

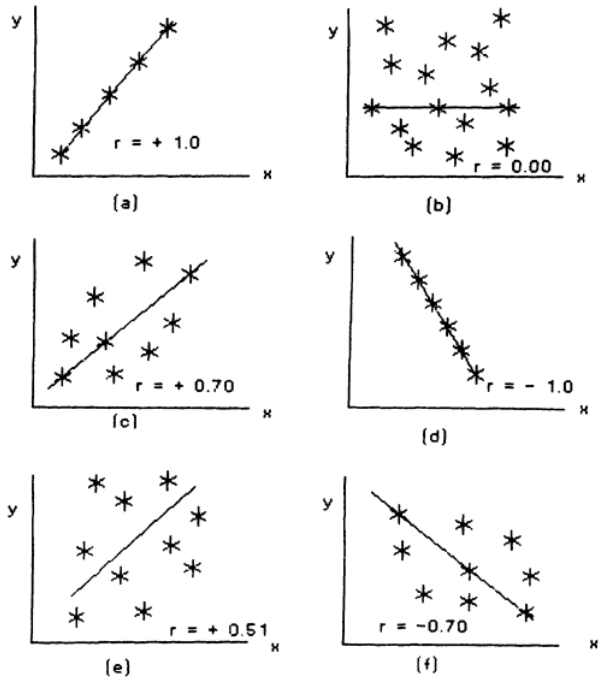


Figure 2.20: Examples of various values of a correlation coefficient r. Each graph shows the correlation indicated by the specific r-value. Source: [78, p.36]

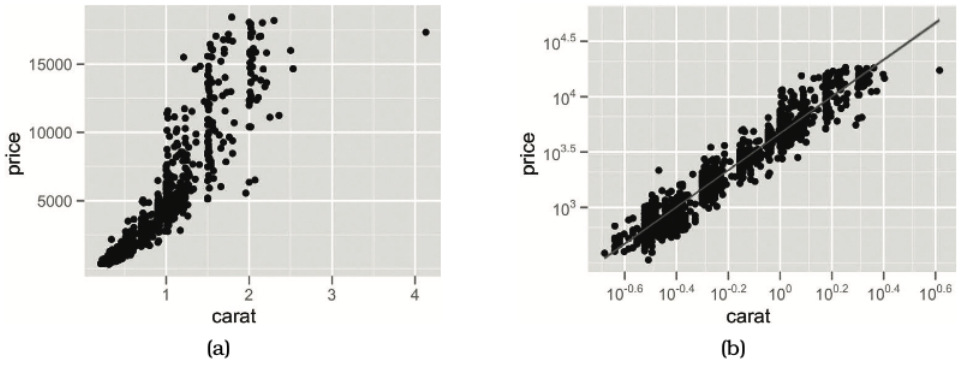


Figure 2.21: Example of a scatterplot being positively correlated. Source: [51, p.148].

Figure 2.21 shows an example of regression; we see a positive correlation between the original diamond price/carat data. However, the regression model does not explain why the variables move together. According to Arkes, we have not established the existence of a causal impact or the size of any causal effect because there are other hypotheses for why the variables move together [3]. When the primary goal of correlation, the calculated regression line, is generated, data is frequently placed on the raw scatterplot of points [51].

Next, we want to discuss studies that tried to discover how human perception can detect regression in a scatterplot. The first experiment compares two different visualization methods in general, while the second analyses how display factors could affect the perception of scatterplots. Finally, the third experiment varied specific parameters in a scatterplot to analyze the bias in the perception of it.

Experiment: Judging correlation

Li et al. [47] identified the most interesting aspects of their task to keep the required experimental effort limited. In their experiment, they identified four independent variables, which are the visualisation method V (scatterplots sc versus paralell coordinate plots (PCPs) pc), the observation duration T (limited display time ld versus unlimited display time ud), sample size n, and population correlation coefficient p. As the dependent variable, they choose the user judgment of correlation U. Their experiment aimed to determine how the independent variable V influences the dependent variable U in the case of different settings of the other three independent variables T, n, and p. Users had only one task, namely judging correlation.

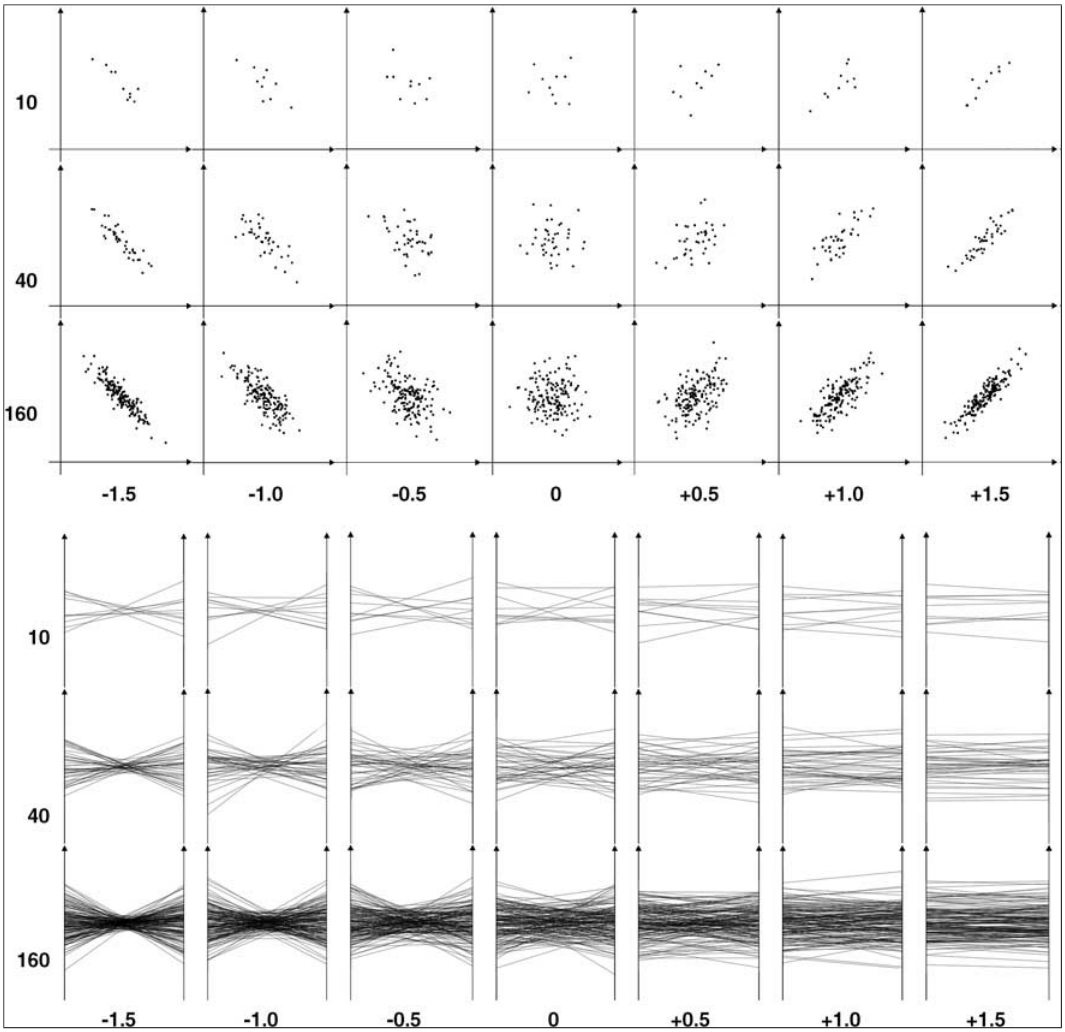


Figure 2.22: Visual stimuli: Scatterplots (top) and PCPs (bottom) with controlled correlations defined by z (in columns) and sample size n (in rows). Source: [47, p.21].

The experiment started with a tutorial in which scatterplots and PCPs were introduced. They also showed the participants how to use them to analyze correlation. Then, the participants were shown characteristic images of both visualization methods for r=-1, r=0, and r=1 on a paper. Afterward, the participants could familiarize themselves with the test environment and test interface in a trial session. Before the formal test session started, trial samples were presented with both time conditions and visualization methods. They used the same images for both time conditions (limited and unlimited) to enable direct comparison. A pilot study found that participants spent quite a long time investigating patterns in the unlimited time condition. Therefore, they first showed the participants the limited time condition to avoid this effect. With that arrangement, they aimed to average out learning effects. After the study, the participants were interviewed

to give their comments. Twenty-five participants took part in the experiment. The age of the participants was between 24 and 45 years old. They were Ph.D. students or faculty from different departments, so they all knew the concept of correlation in statistics. A majority of them were familiar with scatterplots before, but none had used PCPs to analyze correlations. However, all subjects stated that the pre-test tutorial and the trial session gave them enough information to do correlation analysis with the help of PCPs [47].

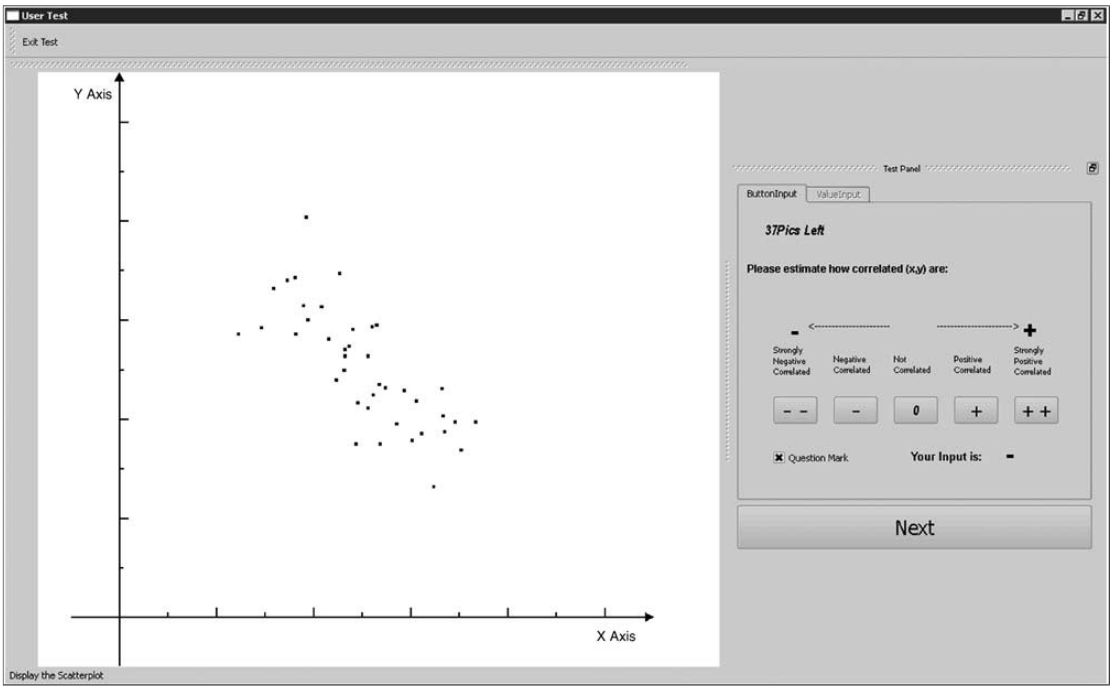


Figure 2.23: The interface of the experimental program. Source: [47, p.23].

Their experiment concluded that for all combinations of sample size n and observation time t, scatterplots allow people to distinguish at least twice as many different correlation levels as PCPs. The authors also stated, that scatterplots are more effective in supporting visual correlation analysis between two variables than PCPs and that for PCPs, the judgment is less accurate. Simultaneously, a diabolo effect is introduced into the perception process of PCPs, causing a bias toward reporting negative correlations. The poor performance of PCPs could be due to unfamiliarity, as none of our subjects had previously used PCPs for correlation analysis. However, we could argue that PCPs' poor performance causes users to be unfamiliar with them. The authors compared and evaluated two different visualization methods using the statistical model of perceived correlation developed in that paper. They plan to investigate whether a similar approach can be applied to other visualization aspects in the future, such as the relationship between cluster detection and icon visual attributes [47].

Experiment: Perception

Cleveland et al. [16] investigated in three experiments how people judge association from scatterplots and how display factors affect their judgments. The participants consisted of students in university courses in statistics, university faculty members in statistics and mathematics, and practicing statisticians in government and industry [16].

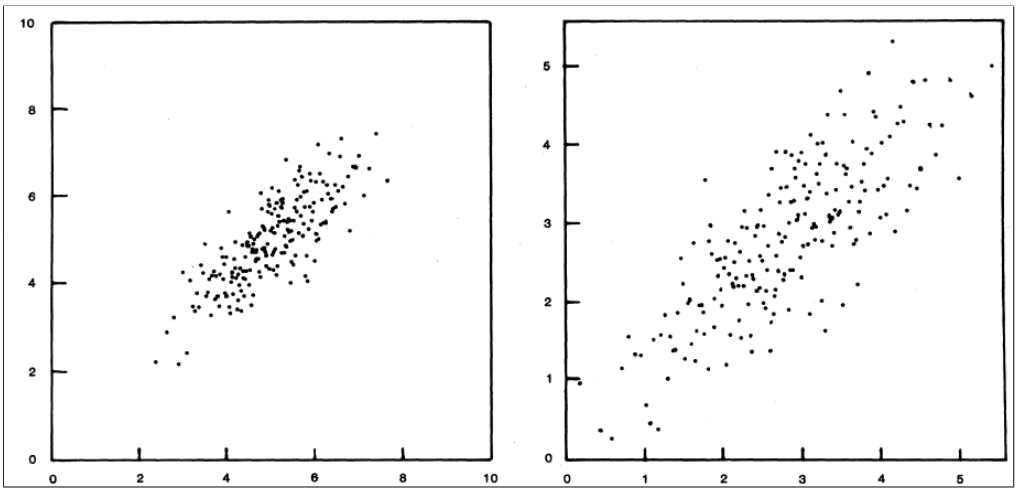


Figure 2.24: Reductions of two scatterplots were used in three types of experiments. The left panel is point-cloud size 2, and the right panel is point-cloud size 4. In both panels w(r) = .4 and r = .8. Source: [16, p.1139].

Nineteen scatterplots, all with 0 or positive correlation coefficients, were shown in the first experiment. For this experiment, the 74 participants were asked to judge linear association on a scale from 0 to 100. 0 would mean that there is no linear association, whereas 100 means a perfect linear association. The figreffig:p2scatter shows two scatterplots from the experiment. In their experiment, they varied two factors: the amount of association and the point-cloud size, but the frame size was kept fixed. They used ten levels of association and each scatterplot had a value of w(r)=1-$\sqrt{1-r^2}$ equal to one of the values 0, .05, .1, .2, . . . , .8. The value w(r) was another numerical measure of linear association that goes from 0 to 1 as r goes from 0 to 1. They choose four different levels for the point-cloud sizes, 1 to 4. Size 1 is the smallest, and size 4 is the largest. There were ten scatterplots with the ten different w(r) values for the point-cloud site 3. There were only three scatterplots for the other sizes with the values of w(r) .1, .4, and .7.

The number of data points and the layout were the same for every scatterplot. Each scatterplot had 200 data points and a square frame with a side equal to 17.3cm. Moreover, they ensured that every plot appeared similar: a linear relation, no peculiar points, and an elliptical appearance.

They used stapled booklets to present the scatterplots. There were written instructions and also samples of scatterplots. For 19 experimental plots, each on a separate page,

the participants were asked to give their own subjective assessment of the amount of linear association rather than to judge the correlation coefficient. They were also asked to work reasonably quickly and not look back or change old answers. Most people could comfortably make a single judgment within 15 seconds.

The judged association increased as the point-cloud site has decreased the scale increase, especially when w(r)=.4. The perceived associations for sizes 1 and 2 were always greater than for sizes 3 and 4. The effect does not appear to extend beyond size 2: The trimmed mean for point-cloud size 2 is only slightly greater than size 1. Sizes 3 and 4 differ by a nontrivial amount only for w(r) = .4 (r = .8).

The authors tested the effect of scale in the second experiment under different conditions. First, they presented the two scatterplots in Figure 2.24 to 109 subjects in three groups of 27, 36, and 46 people using overhead transparency projected onto a screen in the front of the room. Then, on a scale of 0 to 100, they were asked to rate the association of each plot. The value of the 10 percent trimmed mean of [(Judgment for point-cloud size 2) - (judgment for point-cloud size 4)]/100 across subjects was .068 with a standard error of .011 for the second experiment, while the 10 percent trimmed mean of the corresponding values for the subjects in the first experiment was .125 with a standard error of .018.

The third experiment was similar to the second experiment, but they were told, that the correlation coefficients of the two scatterplots were the same. Thirty-two subjects in a single group were shown the scatterplots in Figure 2.24. Their objective was to indicate whether one of the two scatterplots "looked" more highly correlated than the other and, if so, which one. The majority of the participants, namely 66 percent, indicated that the size 2 scatterplot looked more correlated. 13 percent said that the size 4 scatterplot looked more correlated, and 22 percent said they looked the same.

This follows the same pattern as the first experiment, where the corresponding percentages were 81, 18, and 15, and the second experiment, where they were 59, 11, and 30. As a result, the second and third experiments strongly support the first's conclusion: decreasing the point-cloud size by increasing the scales on the horizontal and vertical axes of a scatterplot increases the judged association.

Experiment: Trend judgment

Ciccione et al.[14] conducted an experiment where they tested if human adults can make a quick, intuitive judgment about whether a scatterplot shows an increasing or decreasing trend. The graphs were created using classical linear regression ("ordinary least squares") hypotheses: the values on the ordinate (called yi) were a linear function of the values on the abscissa (called xi) plus independent Gaussian noise ($yi = \alpha xi + \varepsilon i$, where εi are random numbers independently drawn from a normal distribution centered on zero and with standard deviation σ). The slope of the linear trend (α), the number of points (n), and the standard deviation of the noise (σ) were varied orthogonally on the graphs.

Ten people were chosen as participants, of which four were female and six male. The age of the participants was 23.9 ± 1.5. All participants had a normal or corrected vision, no

medical history of epilepsy was right-handed, and did not take psychoactive drugs. The participants were paid 5 euros for their participation in the experiment, which lasted approximately 30 min.

There were 672 scatterplots for every participant to decide if the dataset was increasing or decreasing. Each scatterplot was a graphical representation of a dataset generated at random for each participant using a linear equation plus noise. The number of points (n = 6, 18, 38, or 66), the noise standard deviation (= 0.05, 0.1, 0.15, or 0.2), and the slope of the underlying linear trend (= -0.1875, -0.125, -0.0625, 0, +0.0625, +0.125, or +0.1875) were all varied for a total of 4 x 4 x 7 = 112 combinations. In each of the six experimental blocks, those 112 combinations of parameter values were presented randomly to each participant for a total of 672 trials per participant.

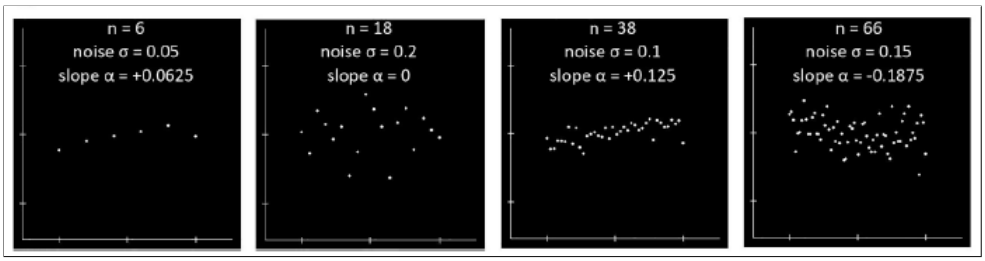


Figure 2.25: Examples of stimuli used in the experiment. Source: [14, p.4].

The first experiment's findings show that participants can quickly extract the linear trend of a scatterplot without any advanced training or prolonged exposures to the stimulus. Participants did not have time to perform complex calculations due to the short presentation time (100 ms) and quick response times (below 900 ms on average). Instead, they had to rely on an intuitive but accurate correlation estimate. Participants' performance remained above chance level even on trials with a prescribed slope of 0, indicating a refined sensitivity to random variations in the graphs [14].

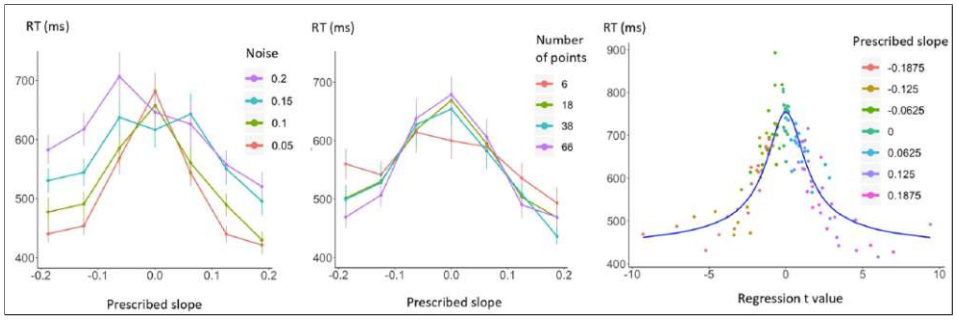


Figure 2.26: Response times in the experiment of Ciccione et al.. Mean response times of the prescribed slope (α) and either the noise (σ, left) or the number of points (n, middle). Source: [14, p.7].

As expected, participants' accuracy was significantly affected by all three parameters of slope, noise, and a number of points, with lower accuracy for shallower slopes, higher noise levels, and smaller datasets. On response times, similar effects were observed, but they were all subsumed by the t value predicted by the Pearson coefficient of correlation. They discovered that participants' decisions followed the prediction of a classic accumulation-of-evidence decision model, where the decision variable was the strength of the t value associated with the Pearson correlation coefficient, by applying Gold and Shadlen's [33] model to their data. This finding suggests that participants gathered evidence on the dataset's trend before giving an answer and that this decision process resembled a statistical regression procedure. Indeed, the performance of participants was better modeled as a function of the t value rather than the prescribed slope. As a result, when detecting a scatterplot's tendency, human adults extract an approximate summary statistic rather than relying solely on the slope of the linear regression. It's worth noting that the average response times did not increase as the number of points n increased. For large values of the slope, response times remained roughly constant or even decreased with n. As a result, participants did not treat the data points in a sequential manner, as would be unavoidable if the data were presented in numbers (such as in a tabular format), but instead processed them in parallel using the graphic presentation.

CHAPTER 3

User Study

> *"Controlled experiments remain the workhorse of evaluation but there is a growing sense that information visualization systems need new methods of evaluation, from longitudinal field studies, insight-based evaluation and other metrics adapted to the perceptual aspects of visualization as well as the exploratory nature of discovery."* [6, p.5]
>
> - BELIV, *2006*

The main goal of this thesis was to find out how people perceive certain information in a scatterplot. To achieve this goal a user study was conducted. In this Chapter, we discuss the methodology of the user study, how the study was set up, and how the data to be used in the study was created. We explain our experiment in-depth and summarize the applied quantitative framework.

3.1 Step 1 - Literature research

As discussed in Chapter 2, data visualization and its presentation are important since they can lead to different results in how humans perceive the information in the displayed data. For example, *overplotting* [25] is still a common problem in information visualization, human observers perceive plots as too overloaded, and it gets challenging to interpret the visualization's information. This issue leads to a loss of information or can even mislead users.

Although many approaches [70], [48], [58] tried to evaluate the effectiveness of clustering techniques, researchers commonly used 2D scatterplots to evaluate the effectiveness of clustering techniques, rather than mathematical frameworks and heuristics [69]. In addition, we have seen that humans are involved in the decision-making process for clustering approaches to aid human discovery [5].

When visualizing regressions, we have seen different techniques [47] [16] [14] being used. It can be seen that regressions between two variables are generally better recognized in scatterplots than in other representations like parallel coordinate plots (PCPs). Different factors influence the perception of regression in visualizations, whether in a scatterplot or any other visualization method [47]. However, some factors (i.e., the education level) are still to be analyzed. There have been studies on how the dot size is influencing the perception of a regression [74]. Our study will provide additional information for understanding how to display the data points correctly so humans perceive the correct information from it.

3.2 Step 2 - Hypotheses generation

This section presents the hypotheses we generated to be evaluated in the user study experiment. We wanted to investigate which parameters affect how human observers perceive regressions in scatterplots. Our hypotheses are derived from the information acquired from the literature research in which we learned that many factors influence the perception of a scatterplot. For example, one of the factors mentioned in Chapter 2 is the overplotting issue in which there is too much data in a graph. With this knowledge, we came up with the following hypotheses:

H1: The size of data points in a scatterplot has a significant effect on the perception of the regression by human observers. With increasing the size of the data points in a scatterplot, the graph gets too overloaded and participants are misinterpreting the correlation (r).

H2: The education and the experience of a participant will not have a significant effect on the perception of the regression of a scatterplot. Participants with higher education and experience in data visualization interpret overplotted scatterplots the same way as participants with lower education and no experience in data visualization.

3.3 Step 3 - Dataset generation

Datasets were needed to create scatterplots that were used for the user study. We decided not to use real-world datasets related to a specific topic but to create datasets with random data. This way it was quickly possible to adjust the dataset parameters that needed to be tested (i.e., dot size, grade of correlation) and create a broad distribution of different data representations. We utilized R [60] to generate the datasets and the scatterplots. The R library ggplot was used to draw the scatterplots. R is a free, open-source statistical analysis software based on the S programming language. It provides various statistical tools (e.g., linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering) and tools for graphical representations, making the tool suitable for our approach.

We had to find the relevant and important parameters for our study to test our hypothesis as defined in Section 3.2. For this approach, we took a closer look at scatterplots and what parameters would be interesting to investigate. As a result, we were able to narrow the possibilities down to four parameters (see Figure 3.1) which are:

1. Number of Samples (n),
2. Size (s) of the dots in the scatterplots, and
3. Shape (sh) of the dots in the scatterplots,
4. Correlation (r) - which can be negative or positive.

Number of Samples (n)	Size (s)	Shape (sh)	Correlation (r)
500	1	●	0,1
	2	▲	0,2
	3	■	0,3
			0,4
			0,5
			-0,1
			-0,2
			-0,3
			-0,4
			-0,5

Figure 3.1: Possible parameters for the regression experiment

To be able to compare the plots with each other, the attributes of the scatterplots only differ by one attribute. Starting with one scatterplot, the following one should have almost the exact attributes to see if the change of one attribute has a significant effect. Therefore, only one attribute should be changed at once to make them comparable. That means, for example: If scatterplot *A* is assigned attributes n = 20, r = 0.1, s = 1, and sh = round, then scatterplot *B* should have the same attributes, but one of the attributes should be varied. For every possibility of an attribute, the total amount of scatterplots multiplies. That means, if we take two attributes with e.g. two possibilities to vary each, there would be four scatterplots. Adding another attribute with five variations led to 20 scatterplots.

Since we want to test our hypothesis H1 that increasing the dot size will have a negative effect on the human observer, the parameter Size (s) may have one of three different values: Size 1, Size 2, and Size 3 (see Figure 3.2).

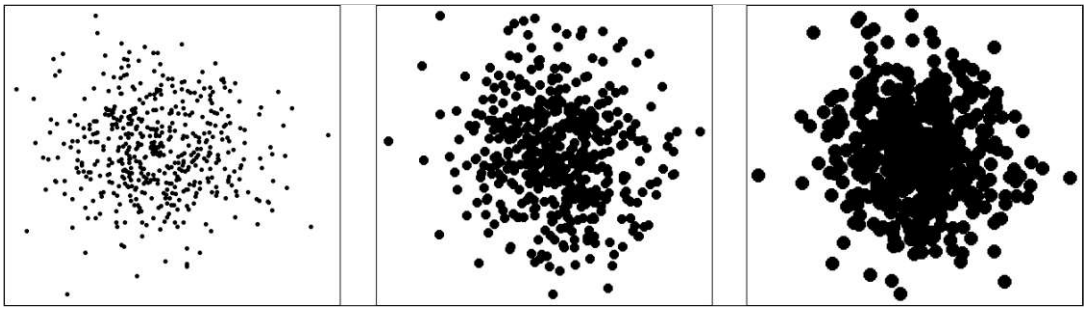


Figure 3.2: Sizes of the dots in the survey in comparison to each other. Left: Size 1; Middle; Size 2; Right: Size 3

We decided to take a fixed number of samples (n) n = 500. The parameter Correlation (r) was varied from 0.10 to 0.50 for the positive- and from -0.10 to -0.50 for the negatively correlated plots in steps of 0.10. We also added a random deviation of up to 0.05 in each direction to minimize the risk of creating a pattern that could be recognized by the participants during the study. We generated 30 different scatterplots in total (see Figure 3.3).

ID	Samples	Pos/Neg	Size	Correlation base
Q1	500	Neg	1	0,1
Q2	500	Neg	1	0,2
Q3	500	Neg	1	0,3
Q4	500	Neg	1	0,4
Q5	500	Neg	1	0,5
Q6	500	Neg	2	0,1
Q7	500	Neg	2	0,2
Q8	500	Neg	2	0,3
Q9	500	Neg	2	0,4
Q10	500	Neg	2	0,5
Q11	500	Neg	3	0,1
Q12	500	Neg	3	0,2
Q13	500	Neg	3	0,3
Q14	500	Neg	3	0,4
Q15	500	Neg	3	0,5
Q16	500	Pos	1	0,1
Q17	500	Pos	1	0,2
Q18	500	Pos	1	0,3
Q19	500	Pos	1	0,4
Q20	500	Pos	1	0,5
Q21	500	Pos	2	0,1
Q22	500	Pos	2	0,2
Q23	500	Pos	2	0,3
Q24	500	Pos	2	0,4
Q25	500	Pos	2	0,5
Q26	500	Pos	3	0,1
Q27	500	Pos	3	0,2
Q28	500	Pos	3	0,3
Q29	500	Pos	3	0,4
Q30	500	Pos	3	0,5

Figure 3.3: Chosen parameters for survey with all possible combinations

We generated the scatterplots in R with the help of the command *mvrnorm*[1] from the *MASS*[2] package. This command produces one or more samples from the specified multivariate normal distribution.

In Listing 3.1 the source code for generating the positively correlated scatterplots in R is shown. In addition, an adapted code was used for creating the negatively correlated

[1]mvrnorm: Minor revision of mvrnorm (from MASS) to facilitate replication: https://www.rdocumentation.org/packages/rockchalk/versions/1.8.152/topics/mvrnorm

[2]MASS package version 7.3-57: https://cran.r-project.org/web/packages/MASS/index.html

scatterplots as well:

```
#defining start-parameters
plots <- 15
samples = 500
rmin <- 0.05 //0.10 - 0.05 as a minimum correlation (r)
rmax <- 0.15 //0.10 + 0.05 as a maximum correlation (r)
size <- 1

i <- 0
while (i < plots) {
    #generating a random value between rmin and rmax
    r <- runif(1, rmin, rmax)
    library('MASS') //needed for mvrnorm-function
    data = mvrnorm( n=samples,
                    mu=c(0, 0),
                    Sigma=matrix(c(1, r, r, 1), nrow=2),
                    empirical=TRUE )
    X = data[, 1]
    Y = data[, 2]

    #saving as a image-file
    png(paste("v2", i, "neg", size, samples, r, ".png", sep="-"))
    plot(data, cex = size, pch = shape, xlab="X", ylab="Y")
    dev.off()

    #increasing rmin and rmax for the next iteration
    rmin = rmin + 0.1
    rmax = rmax + 0.1

    #after every 5th iteration, the size is increased...
    #...and the rmin and rmax are reset to 0.10 +- 0.05
    if((i%%5)==0){
        size = size+1
        rmin = 0.05
        rmax = 0.15
    }
}
```

Listing 3.1: Generation of positively correlated scatterplots.

3.4 Step 4 - Setup user study system

3.4.1 SoSci Survey

The user study was conducted in a web-based manner. As an essential requirement, the collected data needed to be stored safely and anonymously. Therefore, we employed the web-based survey tool. We needed a platform to present our questionnaires, record responses for our experiment, and share them with the participants. Furthermore, the data we acquired had to be stored safely and anonymously. So, we decided to use the web-based survey tool SoSci Survey [46]. This german professional tool allowed us to implement and distribute the survey online. SoSci Survey is free when used for academic survey initiatives. It is also mentioned that the data is protected following the General

Data Protection Regulation (GDPR). Therefore, the survey projects can be designed individually to meet different study requirements.

In our surveys, we designed and showed our questionaries with the possibility for the participants to choose among pre-defined answers. The scatterplots used during the survey were pre-rendered using R and uploaded to the survey tool as images. Furthermore, all the data, e.g., answers and time values, were safely stored. SoSci Survey allows exporting the collected data in XML (Microsoft Excel File) or CSV format.

The final surveys could be accessed through a link provided by the system, which we used to distribute among participants. Among other channels, we also used social media platforms (i.e., Facebook and WhatsApp), where we asked people to share the survey with others, e.g., from their personal or work environments.

Project structure

The projects in SoSci Survey for the experiment were composed of four parts:

1. Introduction page,
2. Task page,
3. Personal information page, and
4. Appreciation page.

On page 1 (introduction) an example plot was given for each positive, negative, and no correlation. A red regression line was added to emphasize a positive or negative correlation in these example plots. It was also noted that this line was only for demonstration purposes and that it will not be visible in the subsequent questions.

Page 2 (task) started after the introduction. Here participants had to judge the correlation in 30 different scatterplots. The scatterplots were pre-rendered according to the descriptions in Section 3.3. During data generation, we made sure that only one parameter (i.e., size, shape, correlation) differed from one scatterplot to another. If we presented the scatterplots in the order they were generated, it would be possible to only judge to correlation according to the pattern instead of their perception. In order to prevent this from happening, the questions were randomly shuffled within SoSci Survey in a way that the pattern of regression levels was not recognizable.

In each question, the participant had the option to answer if they saw a positive, negative correlation or if they did not see any correlation in the displayed data. The participants also had to choose on a scale how confident they were about their chosen option for each answer. To avoid a neutral answer, we decided to use a confidence scale. An Example for page 2 (task) can be seen in Figure 3.6.

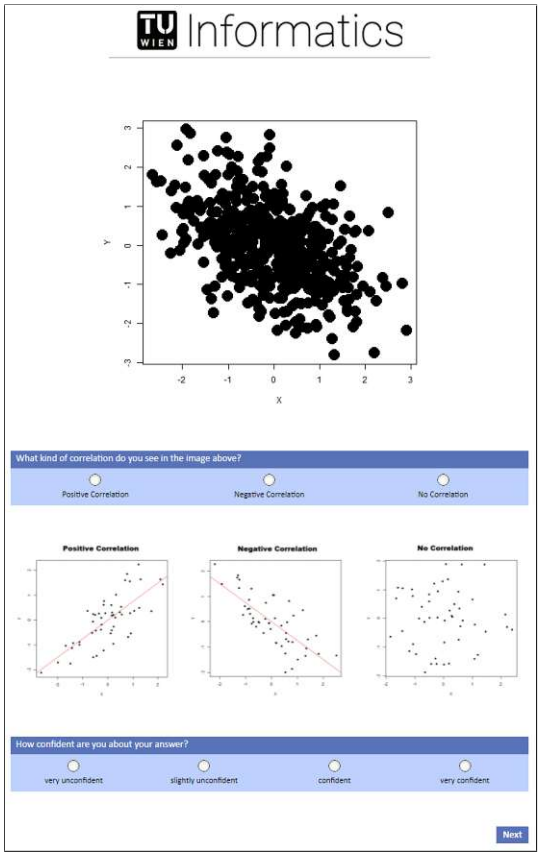


Figure 3.6: Example of page 2 (task)

Page 3 (personal information) gathered additional information from the participants to analyze if they could be relevant to the perception of the scatterplots. Therefore, at the end of the survey, we asked the participants for some information about their age, education, experience in data visualization, and if they had issues with their vision.

For the age, we offered the selection groups of:

- Under 18 years old
- 18-24 years old
- 25-34 years old
- 35-44 years old
- 45-54 years old
- 65-74 years old
- 75 years or older

Education could also influence the perception or interpretation of scatterplots. Because of that, we asked the participants about their highest completed degree or school level. Since the study started in Austria, we chose the joint Austrian degrees:

- No schooling completed
- Grund-/Hauptschulabschluss
- Gymnasium (Matura)
- Abgeschlossene Ausbildung
- Fachhochschulabschluss
- Bachelor's degree
- Master's degree
- Professional degree
- Doctorate

To test the effect of the participant who worked with data visualizations, users had to choose between the following:

- I have no experience in data visualization
- I used to read about data visualization
- I work and create data visualization

The very last question was about their vision in general. The participant had to choose between the following:

- I do not have any issues with my vision
- I am colorblind
- I have a corrected vision

Figure 3.7 shows the final design of page 3 (personal information) using the SoSci Survey tool.

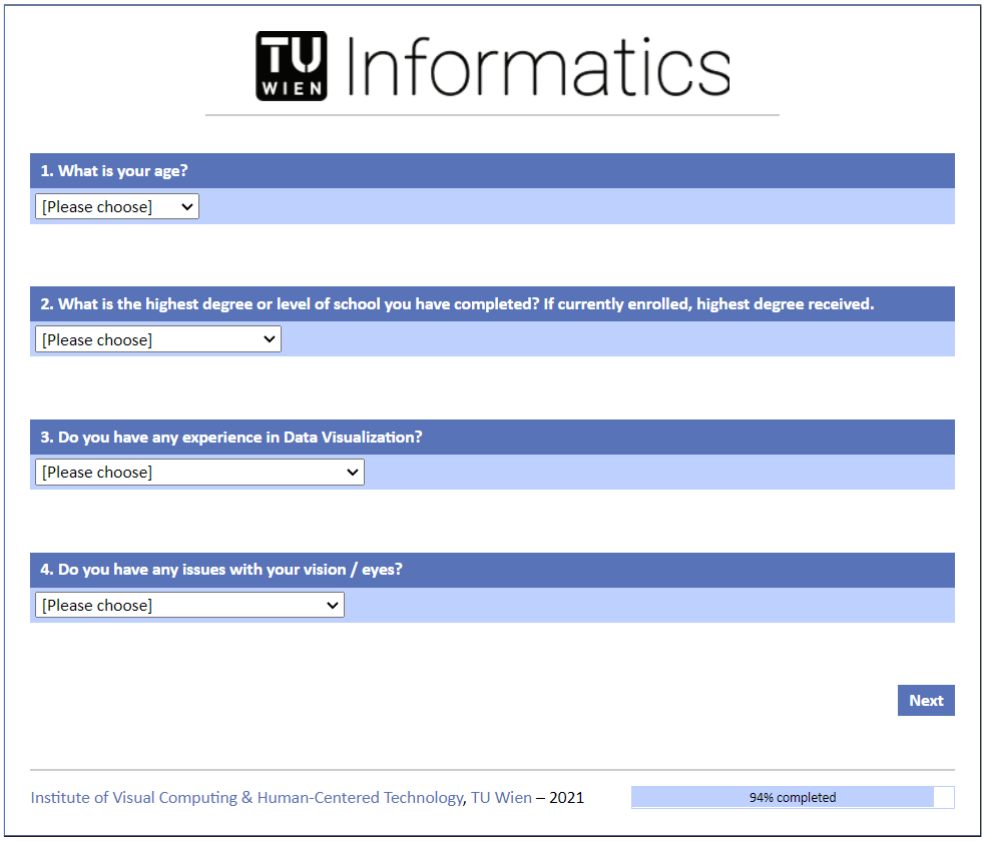


Figure 3.7: Page 3 (personal information) built with the SoSci Survey tool.

Page 4 (appreciation) was the final page we displayed in our survey, where we thanked the participants for their answers.

The study was composed of 33 pages in total; one for the introduction page, 30 task pages (in randomized order), one page for the personal information, and at the end, one for the appreciation page.

3.5 Step 5 - Data collection

The survey were kept online for two months. The collected data was stored in the SoSci Survey database from where it could be downloaded. This *raw data* contained detailed information about the answers chosen during the survey but did not provide any analysis.

The survey for the experiment was started 112 times, of which only 89 cases were valid (see Figure 3.8). In 23 cases the questionnaire was started but not completed, meaning they closed the survey before finishing it. The invalid cases were removed from the dataset that was analyzed.

Questionnaire	clicks	Data sets finished	interviews total	valid cases (download)	
Correlation-Survey v2 cs2	455	89	112	112	
Overall	**455**	**89**	**112**	**112**	

Figure 3.8: Questionnaire response rates for the experiment

3.6 Step 6 - Data evaluation

In this final step, we evaluated the collected data we downloaded from the SoSci survey database. Then, we processed the raw data to present and interpret the results. The detailed data evaluation process and interpretation of the results will be discussed in the following Chapter 4.

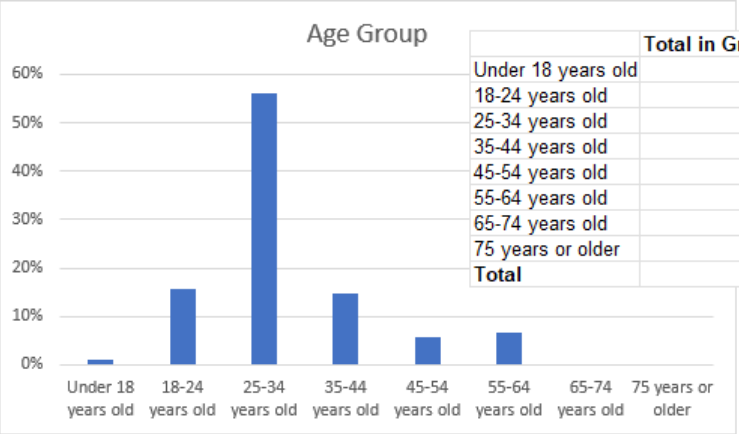
CHAPTER 4

Results

As explained in Chapter 3.2, we investigated human perception of regressions in scatterplots. In this chapter we show the findings of the conducted experiment. We analyze the participant parameters and the general findings. We also grouped the results to see which parameters have specific effects on human perception.

4.1 Participants

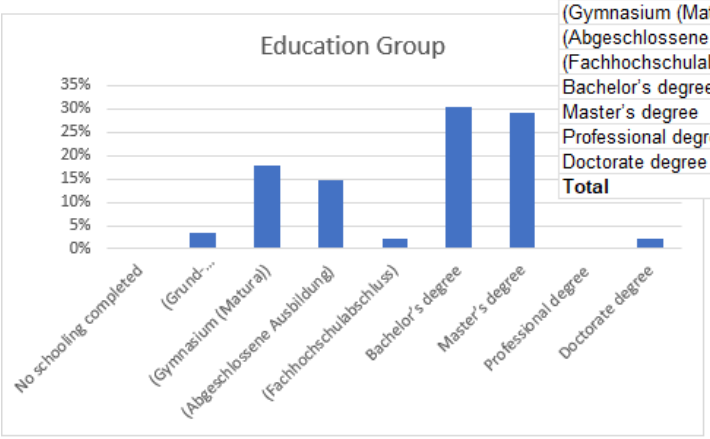
In total, 89 participants completed the survey where we asked them to judge regression, i.e., correlation, in scatterplots. The majority (50 people, or 56%) fell into the age group of 25 to 34 years. The other two main representative age groups were people between 18 and 24 years old (14 people, or 16%) and people between 35 and 44 years old (13 people, or 15%). The only age groups that were not represented at all in the study were people from 65 to 74 years old and people older than 75. A detailed representation of the age distribution can be seen in Figure 4.1.



	Total in Group	Percentage
Under 18 years old	1	1%
18-24 years old	14	16%
25-34 years old	50	56%
35-44 years old	13	15%
45-54 years old	5	6%
55-64 years old	6	7%
65-74 years old	0	0%
75 years or older	0	0%
Total	**89**	

Figure 4.1: Overview of the age distribution of the participants. The majority of 56% were between 25 and 34 years old.

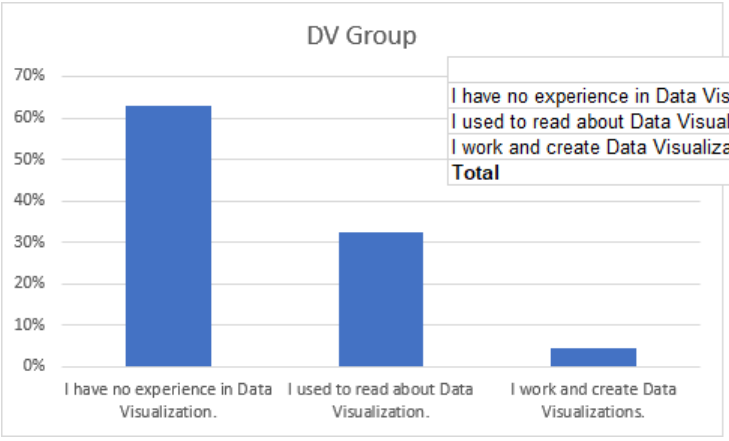
Regarding education, it can be said that every participant had at least any form of education since nobody choose the option "No schooling completed". About one third of participants had a bachelor's degree (27 people, or 30%). The next main representative degree was the master's degree with 26 people, or 29%. People with a *Matura* or with a completed training (*Abgeschlossene Ausbildung*) made up the other two main representative education groups. None of the participants had a professional degree. A detailed representation of the education distribution can be seen in Figure 4.2.



	Total in Group	Percentage
No schooling completed	0	0%
(Grund-/Hauptschulabschluss)	3	3%
(Gymnasium (Matura))	16	18%
(Abgeschlossene Ausbildung)	13	15%
(Fachhochschulabschluss)	2	2%
Bachelor's degree	27	30%
Master's degree	26	29%
Professional degree	0	0%
Doctorate degree	2	2%
Total	**89**	

Figure 4.2: Overview of the education distribution of the participants. The Bachelor's and Master's degree made up the majority with only one person less having a Master's degree than persons having a Bachelor's degree.

When it comes to the topic of having experience in data visualization, only 4 people work and create data visualizations. The majoriy of almost two third (56 people, or 63%) said, that they have no experience in working with and creating data visualization. One third, however, used to read about data visualisations. A detailed representation of this distribution can be seen in Figure 4.3.

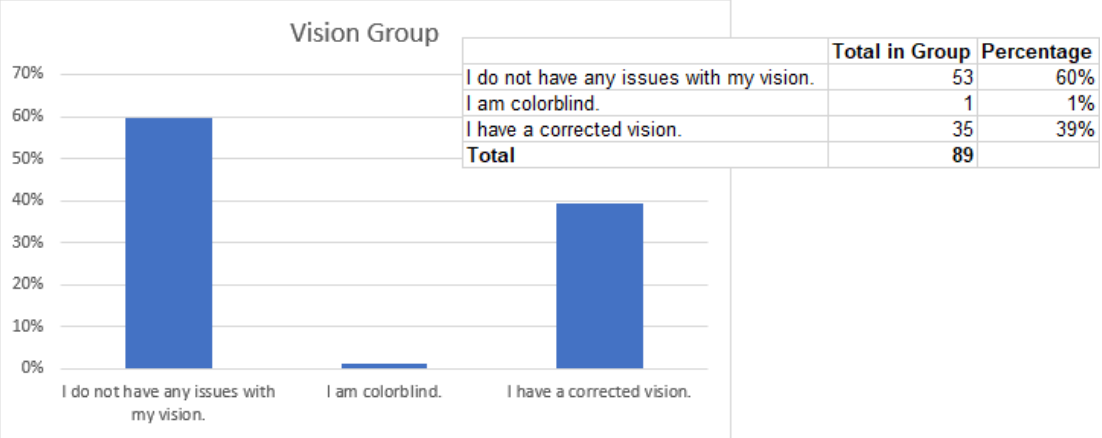


	Total in Group	Percentage
I have no experience in Data Visualization.	56	63%
I used to read about Data Visualization.	29	33%
I work and create Data Visualizations.	4	4%
Total	**89**	

Figure 4.3: Overview of the distribution of having experience in data visualization. Two thirds of the participants had no experience in it.

Only one person stated to be colorblind. 53 persons, or 60%, did not have any issues

with their vision and 35 persons, or 39%, had a corrected vision. We were made aware that some people were uncertain what to choose in this question. Some people said that they have a corrected vision and having no issues any more. Also some did not know, if wearing glasses would mean that they have a corrected vision. This information was provided to us after the experiment, meaning that the data could be inaccurate because of the high uncertainty. In Figure 4.4 the data für the distribution of participants regarding their vision is provided.



	Total in Group	Percentage
I do not have any issues with my vision.	53	60%
I am colorblind.	1	1%
I have a corrected vision.	35	39%
Total	**89**	

Figure 4.4: Overview of the distribution of participants having problems with their vision. Over one third had a corrected vision.

4.2 Findings

In this section we describe our findings we gained from analyzing the study data. We used statistical methods to analyze the quantitative user study data [68]. To see if parameter distributions were statistically significantly different, we used a two-tailed t-Test [66]. The generated p-value shows if two compared datasets are significantly different. A p-value less or equal than 0.05 indicates that the presented parameter distributions are statistically significantly different to each other.

All plots for this experiment had the same sample size of 500. The parameters that differed from plot to plot were the dot size and the correlation coefficient r.

The question Q1-Q5 all had the same dot size of 1 and were negatively correlated. Only the degree of the correlation differed between the questions. For further analysis, we group these questions (plot) into one group, called Group 1 (G1). The five plots can be seen in Figure 4.5.

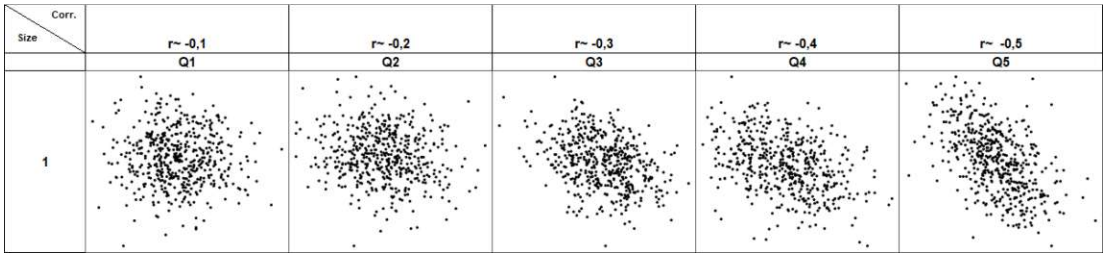


Figure 4.5: Group 1: Plots used for the questions Q1-Q5. Each plot has the dot size 1 and a negative correlation coefficient r.

The majority of participants could see a negative correlation (i.e., chose the option "negative" for their answer) for the question Q3, Q4 and Q5. For the questions Q1 and Q2 the majority of participants did not see any correlation in the data. It is worth mentioning, though, that 25% of the participants saw a negative correlation in Q2. Only 3% of the participants noticed the negative correlation in Q1. In Figure 4.6 we visualized the answers for G1 in percentage.

		r~ -0,1	r~ -0,2	r~ -0,3	r~ -0,4	r~ -0,5
Size		Q1	Q2	Q3	Q4	Q5
1	Pos	9%	4%	6%	4%	6%
	Neg	3%	25%	75%	87%	92%
	No Corr	88%	71%	19%	9%	2%

Figure 4.6: Answers for Q1-Q5 visualized in percentage. Participants start to recognize a correlation with a negative correlation coefficient r~ -0.3

The settings for the parameters of the next five questions Q6-Q10 were the same as for the question of G1, except that the dot size was bigger, namely size 2. The correlation coefficient r was again negative between r~ -0.1 and r~ -0.5. For further analysis, we group these questions (plots) into one group, called Group 2 (G2). These five plots can be seen in Figure 4.7.

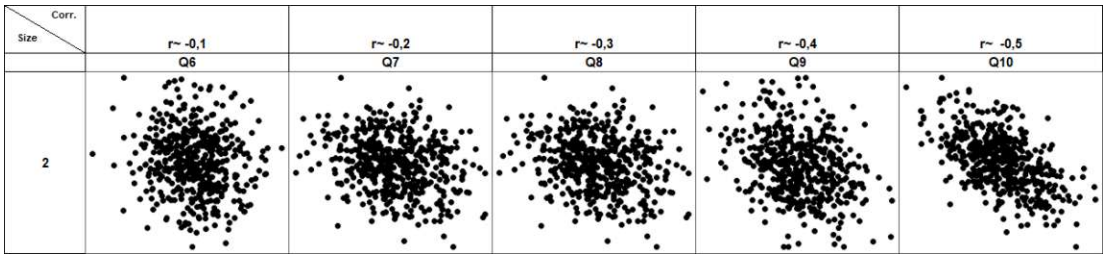


Figure 4.7: Group 2: Plots used for the questions Q6-Q10. Each plot has the dot size 2 and a negative correlation coefficient r.

We compared this group G2 with group G1 and noticed that for Q7, 39% of the participants were able to tell that there was a negative correlation whereas in G1-Q2, only 25% chose the correct answer. As mentioned in the beginning of this chapter, the sample size for all plots were the same (500). The only difference of Q2 and Q7 in this regard is the dot size. We then checked the p-value to see if the change of the dot size was significant meaning it has an influence on the recognition of the correlation. As seen in Table 4.1 only for Q2 and Q7 the p-value was below the 0.05 mark (0.03). This means, that in this setting (r~ -0.2) the change of the dot size was significant. For the other settings, changing the dot size did not have an effect on the recognition of the negative correlation. Figure 4.8 represents the answers for G2 in percentage.

		r~ -0,1	r~ -0,2	r~ -0,3	r~ -0,4	r~ -0,5
Size		Q6	Q7	Q8	Q9	Q10
2	Pos	9%	4%	9%	6%	7%
	Neg	8%	39%	73%	90%	90%
	No Corr	83%	56%	18%	4%	3%

Figure 4.8: Answers for Q6-Q10 visualized in percentage. Similar to the plots of Group 1, participants started to recognize a correlation with a correlation coefficient r~ -0.3. The increased dot size had a statistically significant influence on the recognition of the correlation at least for r~ -0.2.

	p-value	significant?
Q1 compared with Q6	0,508061912	no
Q2 compared with Q7	0,031784024	yes
Q3 compared with Q8	0,4172a98433	no
Q4 compared with Q9	0,320052349	no
Q5 compared with Q10	1	no

Table 4.1: Comparison of G1 with G2.

The next questions Q11-Q15 have been grouped together into group 3 (G3). For these plots the dot size was increased to 3 while the correlation coefficient r again was negative between r~ -0.1 and r~ -0.5. The plots for G3 can be seen in Figure 4.9.

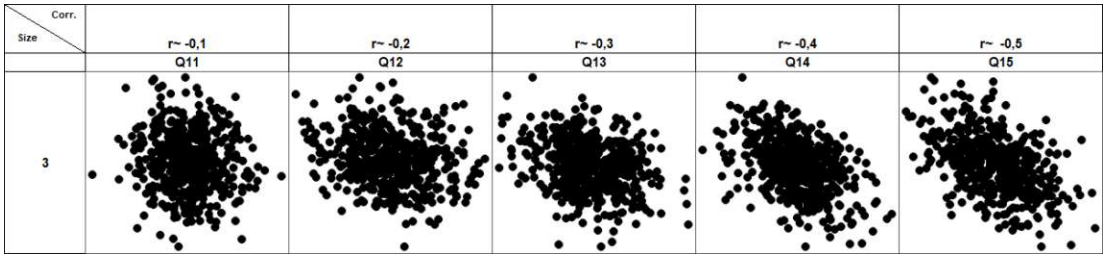


Figure 4.9: Group 3: Plots used for the questions Q11-Q15. Each plot has the dot size 3 and a negative correlation coefficient r.

As can see in Figure 4.10 that in this group, the amount of correct answers for the plot with the correlation coefficient r~ -0.2 increased to 44%. The significance test, however, shows that increasing the dot size from 2 to 3 has no significant effect on the perception of the participants. Table 4.2 shows that all p-values were over the 0.05 mark.

		r~ -0,1	r~ -0,2	r~ -0,3	r~ -0,4	r~ -0,5
Size		Q11	Q12	Q13	Q14	Q15
3	Pos	6%	8%	7%	3%	6%
	Neg	6%	44%	67%	90%	90%
	No Corr	89%	48%	26%	7%	4%

Figure 4.10: Answers for Q11-Q15 visualized in percentage. Changing the dot size from 2 to 3 did not have a significant effect on the perception.

	p-value	significant?
Q6 compared with Q11	0,145103627	no
Q7 compared with Q12	0,067605758	no
Q8 compared with Q13	0,094846566	no
Q9 compared with Q14	0,25044658	no
Q10 compared with Q15	0,56664024	no

Table 4.2: Comparison of G2 with G3.

We then compared the answers of G3 with the answers of G1 to see if the change of the dot size from size 1 to size 3 was significant. The change of the dot size was only significant for the plot with the correlation coefficient r~ -0.2. The values of the t-test can be seen in Table 4.3.

	p-value	significant?
Q1 compared with Q11	0,508061912	no
Q2 compared with Q12	0,00043016	yes
Q3 compared with Q13	0,372144822	no
Q4 compared with Q14	0,8099234	no
Q5 compared with Q15	0,619808566	no

Table 4.3: Comparison of G1 with G3.

The next five plots (Q16-Q20) had the dot size of 1. The correlation coefficient for this plots were r~ 0.1, meaning they were positively correlated. They have been grouped together into group 4 (G4). The plots used for G4 can be seen in Figure 4.11.

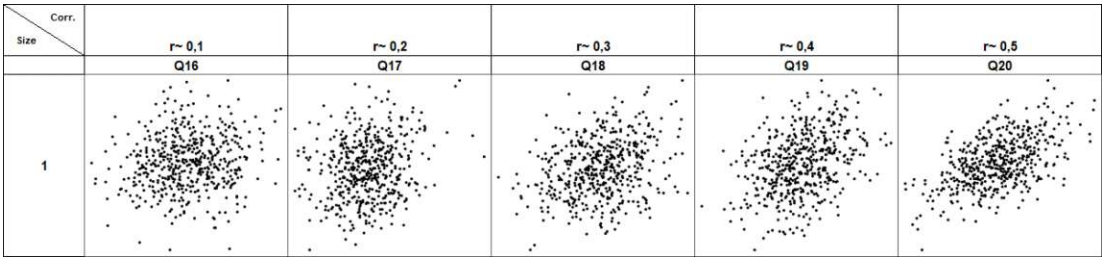


Figure 4.11: Group 4: Plots used for the questions Q16-Q20. Each plot has the dot size 1 and a positive correlation coefficient r.

The majority of participants started to recognize a positive correlation (i.e., select the option "positive correlation" as answer) with a correlation coefficient of greater than r~0.3. Unlike in the group G1, there were significantly more people choosing the correct answer where the correlation was r~0.2. In Q2 only 25% of the participants saw the correct answer whereas in Q17 40% choose the correct answer 4.12.

		r~ 0,1	r~ 0,2	r~ 0,3	r~ 0,4	r~ 0,5
Size		Q16	Q17	Q18	Q19	Q20
1	Pos	10%	40%	63%	88%	98%
	Neg	6%	6%	6%	3%	2%
	No Corr	84%	54%	31%	9%	0%

Figure 4.12: Answers for Q16-Q20 visualized in percentage. Participants start to recognize a positive correlation with r~ 0.2.

For the plots for Q21-Q25 the dot size was incresed to 2. All other parameters stayed the same as for the previous group G4. We summarize these five plots into group 5 (G5). The plots used for this group can be seen in Figure 4.13.

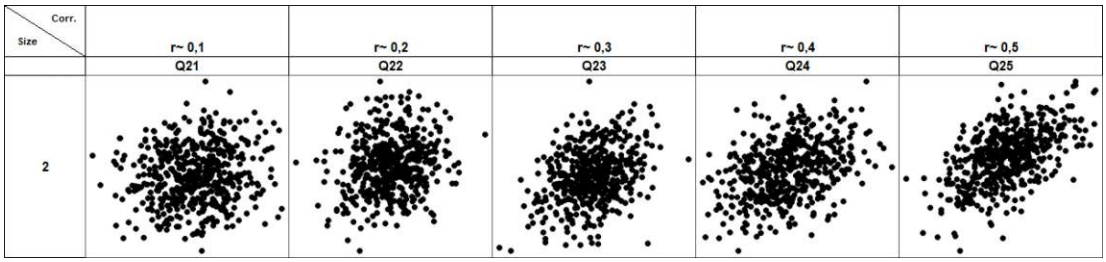


Figure 4.13: Group 5: Plots used for the questions Q21-Q25. The plots have the same parameters as G4 except the increased dot size of 2.

In the groups before, the number of correct answers increased with the degree of the correlation. This was not true for group G5. In this group, the answers of the Q21 was recognized as a positively correlated plot by 30% of the participants. Q21 is the plot with the weakest correlation in this group (r~0.1). Nevertheless, a high percentage of participants answered it correctly. Q22 had a higher correlation of r~0.2 but only 19% of the participants recognized it as a positively correlated plot. The correct answer for Q23, Q24 and Q25 seemed to be clearly visible for the participants as the had a high amount of correct answers (79%, 90% and 92% respectively). In Figure 4.14 the percentage of the choosen answers for this group can be seen.

		r~ 0,1	r~ 0,2	r~ 0,3	r~ 0,4	r~ 0,5
Size		Q21	Q22	Q23	Q24	Q25
2	Pos	30%	19%	79%	90%	94%
	Neg	3%	6%	4%	4%	6%
	No Corr	66%	75%	17%	6%	0%

Figure 4.14: Answers for Q21-Q25 visualized in percentage. The percentage of correct answers did not steadily increase with the correlation.

The comparison of group G5 with G4 is very interesting since there are significant differences between them. Changing the dot size from 1 to 2 seems to affect the perception of the plots with the correlation coefficient r~0.1, r~0.2 , and r~0.3. For Q19 or Q20 and their counterparts, there were no significant changes in the answers. The outcomes of the t-test for this comparison can be seen in Table 4.4.

	p-value	significant?
Q16 compared with Q21	0,000065394	yes
Q17 compared with Q22	0,000636704	yes
Q18 compared with Q23	0,008934382	yes
Q19 compared with Q24	0,44887823	no
Q20 compared with Q25	0,259149675	no

Table 4.4: Comparison of G4 with G5.

The last five plots (Q26-Q30) were grouped into group 6 (G6). The plots of this group can be seen in Figure 4.15. They had the same attributes of group G5 but the dot size was increased to 3.

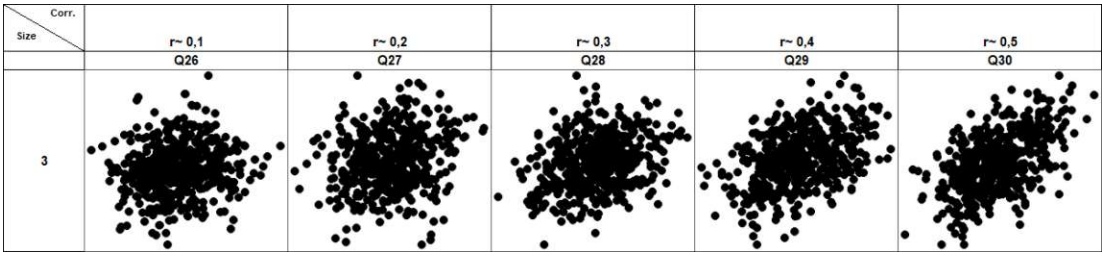


Figure 4.15: Group 6: Plots used for the questions Q26-Q30. Those plots had a dot size of 3 and were positive correlated.

In this group, the plot with the weakest degree of correlation (Q26, r~0.1) was regconized correctly by 21% of the participants. 52% saw the correcct correlation in Q27. Q28 was answered correctly by 74% of the participants while Q29 and Q30 were recognized correctly by 92% and 96%. In Figure 4.16 the answers of the participants can be seen.

		r~ 0,1	r~ 0,2	r~ 0,3	r~ 0,4	r~ 0,5
Size		Q26	Q27	Q28	Q29	Q30
3	Pos	21%	52%	74%	92%	96%
	Neg	3%	2%	3%	2%	2%
	No Corr	75%	46%	22%	6%	2%

Figure 4.16: Answers for Q26-Q30 visualized in percentage. Particiants start to recognize the correlation with r~ 0.2.

Unlike in the groups before (except in group G5), the trend of choosing the correct answer does not start at a correlation index of r~0.3, but at a correlation index of r~0.2. When we took a closer look at the result of this particular question Q27, we saw that out of 46 participants (who chose the correct option), a majority of 26 participants were confident about their answer. In Table 4.5 the outcomes of the significance test can be seen.

	p-value	significant?
Q21 compared with Q26	0,08079397	no
Q22 compared with Q27	0,00000029432	yes
Q23 compared with Q28	0,281085634	no
Q24 compared with Q29	0,740932722	no
Q25 compared with Q30	0,707720287	no

Table 4.5: Comparison of G5 with G6.

Comparing G4 with G6 showed that the change of the dot size from 1 to 3 was only significant for the plots with a correlation with r~ 0.1. For the other correlations, it was not significant.

	p-value	significant?
Q16 compared with Q26	0,043513507	yes
Q17 compared with Q27	0,164972499	no
Q18 compared with Q28	0,063014899	no
Q19 compared with Q29	0,29933297	no
Q20 compared with Q30	0,25044658	no

Table 4.6: Comparison of G4 with G6.

After analyzing the results of all the answers from all the participants, we recognized that there was a trend where participants started to correctly identify a correlation. We can see that the participants started to recognize a positive correlation when the scatterplot had a correlation of r~0.3, and a negative correlation with a correlation of r~-0.3, respectively. It is worth mentioning that in this study participants started to choose the correct answer at a correlation of r~0.2 on the positive plots and with a dot size of 3.

If we take a look at the answers of the plot where the correlation was r~0.1, namely Q16, Q21, and Q26, we can see that most of the participants did not see any correlation. The same applies for the plots with a correlation of r~-0.1 (Q1, Q6, and Q11). Regarding the negatively correlated plots with the mentioned correlation only 3% correct answers were chosen for dot size 1 (Q1). For dot size 2, 8% correct answers were made and 6% for dot size 3 (Q11). For the positively correlated plots with the same degree of correlation r (r~0.1) there were 10% correct answers for dot size 1, 30% correct answers for dot size 2, and 21% correct answers for dot size 3.

Next we divided the answers into correct and incorrect and compared the distribution of the age, education-level, experience in DV and the visual limitations. We counted those answers as correct when the answer for the plots with a (negative) correlation coefficient r~-0.3, r~-0.4 and r~-0.5 has been answered with the "Negative Correlation"-Option in

the survey. Also correct were the answers for the plots with a (positive) correlation coefficient r~0.3, r~0.4, r~0.5 if they were answered with the "Positive Correlation"-option. The answers for the plots with a correlation of r~-0.1, r~-0.2 and r~0.1, r~-0.2 were not included in this analysis due to the weak grade of the correlation. We weighted the number of the remaining answers since there was an uneven distribution for each category. An example for the uneven distribition would be, that there were 50 participants were between 25 and 34 years old but only 14 participants between 18 and 24 years old. The exact distribution of the participants attributes was explained in chapter 4.1.

All correct answers were then further split up into two parts: answers for the negatively correlated plots and answers for the positively correlated plots. Regarding age, participants between 25 and 34 years old had the largest share of correct answers. This applies to both the negatively and positively correlated plots (46% for negative and 58% for positive correlations). In Figure 4.17 the weighted correct answers by age can be seen.

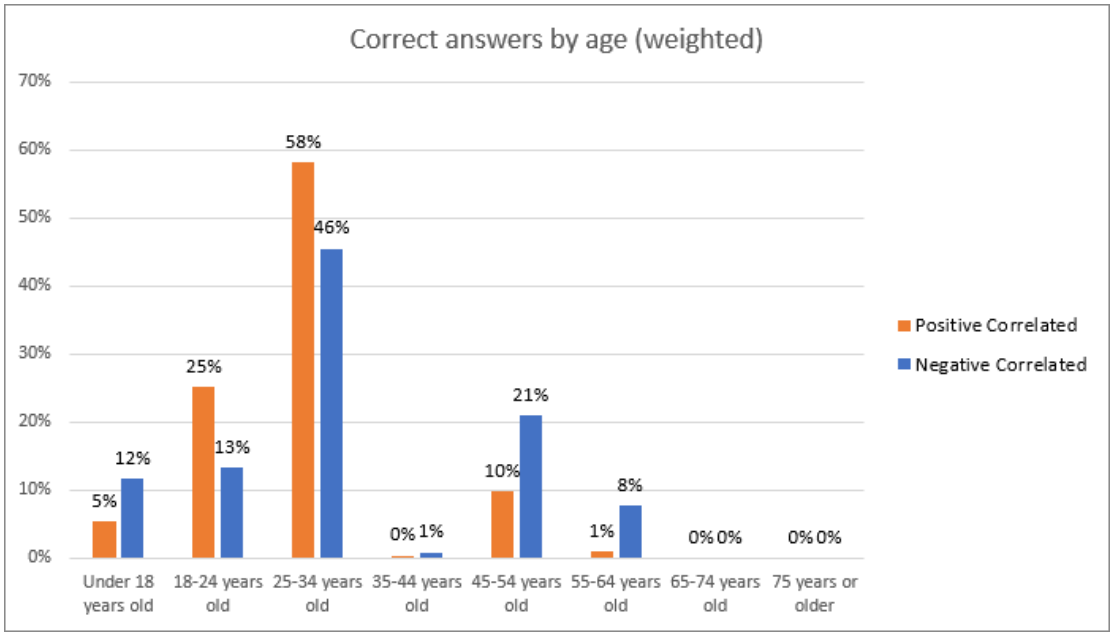


Figure 4.17: Overview of the distribution of the participant age groups with respect to correct answers.

Taking a look at the distribution of the participants' education we noticed that people with a bachelor's degree had the largest share of correct answers of 45%. However, it has to be mentioned that this is only the case for positively correlated plots. For the negatively correlated plots, people with a master's degree represent the majority of 26%. This is only 1% more than the participants with a bachelor's degree (25%) and only 5% more than people with a Matura-degree (21%). The weighted correct answers by education can be seen in Figure 4.18.

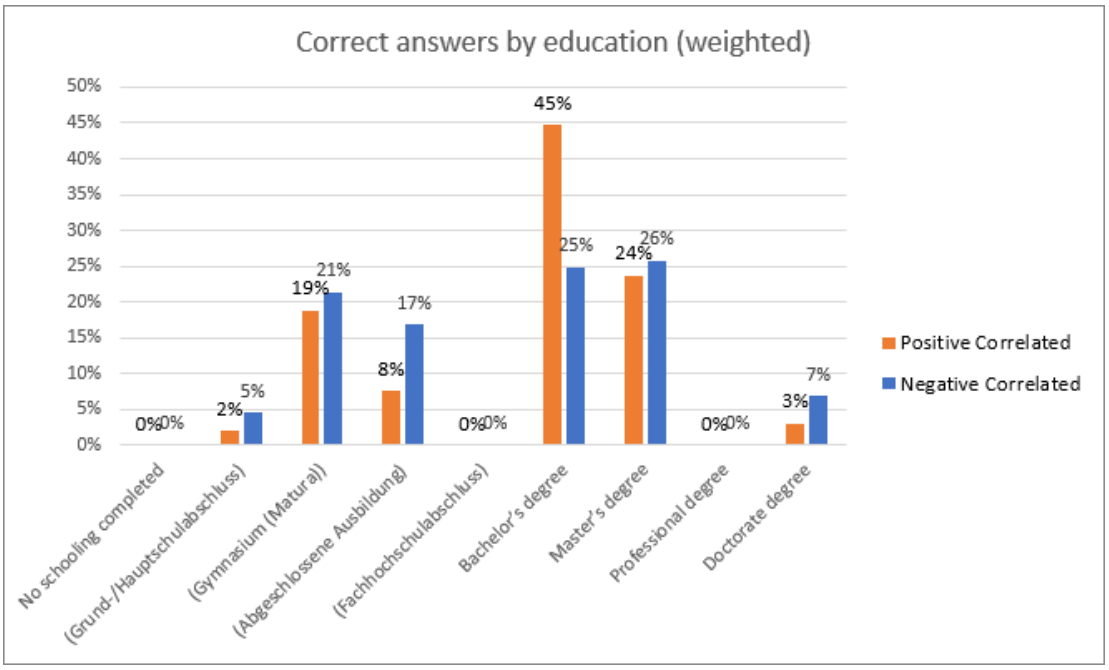


Figure 4.18: Overview of the distribution of the participants' education with respect to correct answers.

When it comes to the correct answers regarding the experience in DV, people with no experience had the largest share of correct answers. For the positively correlated plots the share was 66% and for the nagatively correlated plots it was 74%. Only 2% (for positively correlated plots) and 4% (for negative correlated plots) were correctly chosen by participants who work and create DVs. The detailed barchart regarding the correct answers by experience can be seen in Figure 4.19.

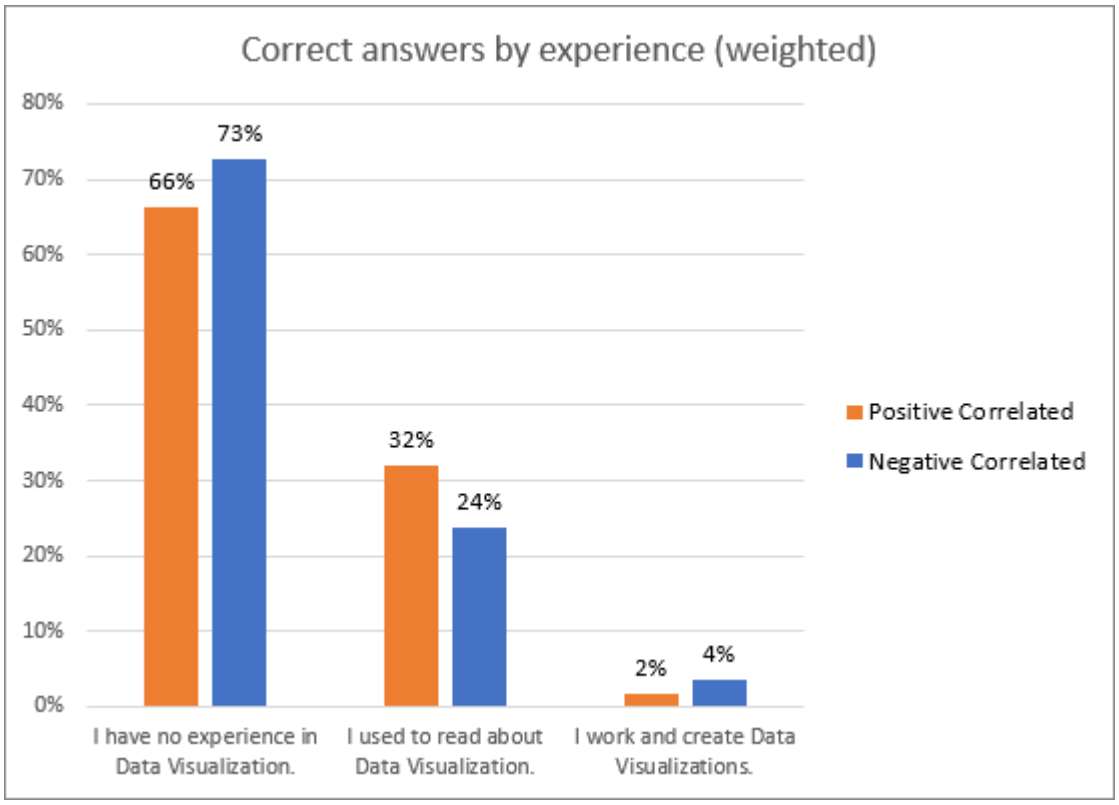


Figure 4.19: Overview of the distribution of the participants' experience in data visualization with the correct answers.

Participants who have no issues with their vision represent the majority of this group with 64% for the positive correlated plots and 52% for the negative correlated plots. We had no colorblind people who had correct answers. This doesn't necessarily mean anything since there was only one colorblind person participating in the experiment. People with a corrected vision were represented by 36% for the positive correlated plots and 48% for the negative correlated plots. As mentioned before, this data has a high uncertainty since the participants were not sure if wearing glasses means that they have a corrected vision. Nevertheless, the representation of the weighted correct answers by vision can be seen in Figure 4.20.

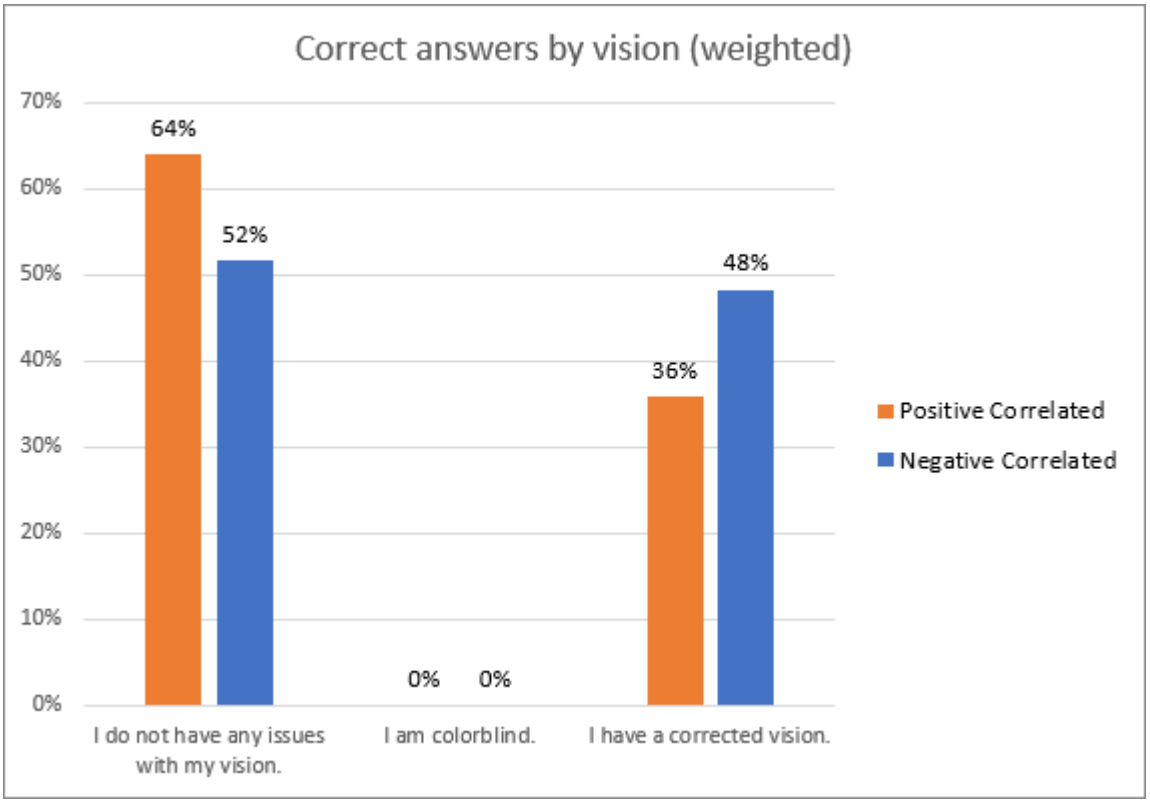


Figure 4.20: Overview of the distribution of the participants having problems with their vision with the correct answers.

CHAPTER 5

Conclusion

The results of this work are summarized in this chapter. Additionally, this thesis' limitations and future research are discussed.

5.1 Summary and main findings

This thesis focuses on the research of the perception of scatterplots. Our main task in our work was to prove that the visualisation of a scatterplot can influence on how human perceive them. In order to fulfill that task, we did a literature research followed by a creation of a user study. This study helped us to develop additional knowledge on what parameters are influencing the perception. Details on how we developed the survey are presented in Chapter 3. The results of the mentioned survey are presented in Chapter 4 in more detail.

In our experiment we analyzed if the change of the dot size affects the perception of regression of a scatterplot. The results showed that changing the dot size does have an effect on the perceiver's recognition for specific correlations. This means, the change of the size was significant for negatively correlated plots with a correlation coefficient r~-0.2 when changing the dot size from 1 to 2. For positively correlated plots, it was significant for r~0.1, r~0.2, and r~0.3 when the dot size was increased from size 1 to size 2. It was also significant when changing the dot size from 2 to 3 for the positively correlated plots with r~0.2. Plots with a stronger correlation, namely r~-0.4 and r~-0.5 or r~0.4 and r~0.5, were not affected by the change. This could be due to the fact, that a stronger correlation is easier to interpret by the viewer.

Regarding education, results show that for the positive correlated plots people with a Bachelor's degree had by far the biggest share of correct answers, namely 45%. The second biggest share were People with a Master's degree which were 24%. Whereas for negatively correlated plots, participants with a Master's degree were the majority of

26% and people with a Bachelor's degree were accounted for 25%. It seems that the type of education affects the number of visualizations people already had contact with in their life, and which increases their confidence in interpreting them. In terms of having experience, the biggest share of correct answers were people without any experience in Data Visualization. Participants who work and create Data Visualization, had the smallest share of correct answers. This applies in both cases for negatively as well for positively correlated scatterplots. This is a very interesting result which would need further exploration in the future. We assume that people with a lot of experience in data visualization may spend less time looking at the graphs, and, therefore, more easily give incorrect answers.

Our work showed, that people only start to recognize correlations from a certain degree. This leads to our suggestion for visualization designers to add a regression line when visualizing plots with a correlation coefficient between r~-0.3 and r~0.3. Plots with a higher or respectively lower degree of correlation are recognized by at least over 60% of the observers.

Regarding the dot size, visualization designers should take into account, that increasing it can significantly affect the recognition of a regression in a scatterplot positively. Even if the individual dots are not visible anymore the amount of people who see the regression correctly increase.

5.2 Limitations and future work

One drawback we have to mention is the fact that we chose to do a web-based study (to reach more participants), which means that the study was not done under controlled settings. The participants were chosen by a distribution of the link for the survey via social media platforms (i.e., Facebook and WhatsApp), where the people were also asked to share the survey with others, e.g., from their personal or work environments. The distribution of participants was very unequal among each and every category, namely 'Age', 'Education', 'Experience', and 'Vision'. One example would the the category regarding the age. 56% were 25-34 years old. We did not evaluate the results for vision, because during the study we discovered that some participants did not know what "corrected vision" means.

Future work will focus on improving the study setting. This might be achieved by either including more participants and/or by refining the survey answers. More participants could bring more reliable and generalizeable data. The surveys could be refined by explaining or defining of the possible answers in more detail. This would address the problem with the question regarding the vision of the participant.

List of Figures

List of Tables

Bibliography

[1] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visualizing time-oriented data—a systematic view. *Computers & Graphics*, 31(3):401–409, 2007.

[2] D. S. Alexandre and J. M. R. Tavares. Introduction of human perception in visualization. *International Journal of Imaging*, 4, 2010.

[3] J. Arkes. *Regression Analysis: A Practical Introduction*. Routledge, 01 2019.

[4] A. G. Asuero, A. Sayago, and A. Gonzalez. The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1):41–59, 2006.

[5] M. Aupetit, M. Sedlmair, M. M. Abbas, A. Baggag, and H. Bensmail. Toward perception-based evaluation of clustering techniques for visual analytics. In *Proceedings of the 2019 IEEE Visualization Conference (VIS)*, pages 141–145, 2019.

[6] AVI (Conference) Corporate Author. *BELIV '06: Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, New York, NY, USA, 2006. Association for Computing Machinery.

[7] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[8] J. Bertin, W. Berg, H. Wainer, and U. of Wisconsin Press. *Semiology of Graphics*. University of Wisconsin Press, 1983.

[9] D. Bouchlaghem, H. Shang, J. Whyte, and A. Ganah. Visualisation in architecture, engineering and construction (aec). *Automation in construction*, 14(3):287–295, 2005.

[10] K. Brodlie, L. Carpenter, R. Earnshaw, J. Gallop, R. Hubbold, A. Mumford, C. Osland, and P. Quarendon. *Scientific Visualization: Techniques and Applications*. Springer Berlin Heidelberg, 2012.

[11] S. Carpendale. *Evaluating Information Visualizations*, pages 19–45. Springer Berlin Heidelberg, 2008.

[12] H. Chen, W. Chen, H. Mei, Z. Liu, K. Zhou, W. Chen, W. Gu, and K.-L. Ma. Visual abstraction and exploration of multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1683–1692, 2014.

[13] A. Chinnaswamy, A. Papa, L. Dezi, and A. Mattiacci. Big data visualisation, geographic information systems and decision making in healthcare management. *Management Decision*, 2019.

[14] L. Ciccione and S. Dehaene. Can humans perform mental regression on a graph? accuracy and bias in the perception of scatterplots. *Cognitive Psychology*, 128:101406, 2021.

[15] W. S. Cleveland. *Visualizing Data*. Hobart Press, 1993.

[16] W. S. Cleveland, P. Diaconis, and R. McGill. Variables on scatterplots look more highly correlated when the scales are increased. *Science*, 216(4550):1138–1141, 1982.

[17] W. S. Cleveland and R. McGill. The many faces of a scatterplot. *Journal of the American statistical association*, 79(388):807–822, 1984.

[18] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger. Looks good to me: Visualizations as sanity checks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):830–839, 2018.

[19] D. W. Dahl, A. Chattopadhyay, and G. J. Gorn. The importance of visualisation in concept design. *Design Studies*, 22(1):5–26, 2001.

[20] T. N. Dang, L. Wilkinson, and A. Anand. Stacking graphic elements to avoid overplotting. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1044–1052, 2010.

[21] M. E. Doherty, R. B. Anderson, A. M. Angott, and D. S. Klopfer. The perception of scatterplots. *Perception & Psychophysics*, 69(7):1261–1272, 2007.

[22] P. Eilers and J. Goeman. Enhancing scatterplots with smoothed densities. *Bioinformatics (Oxford, England)*, 20:623–8, 04 2004.

[23] S. Few and P. Edge. Solutions to the problem of over-plotting in graphs. *Visual Business Intelligence Newsletter*, 2008.

[24] A. Field and G. Hole. *How to Design and Report Experiments*. SAGE Publications, 01 2003.

[25] P. Flom. Scatterplots: Basics, enhancements, problems and solutions. In *Scatterplots: Basics, enhancements, problems and solutions*, 01 2014.

[26] C. Forsell. A guide to scientific evaluation in information visualization. *2010 14th International Conference Information Visualisation*, pages 162–169, 2010.

[27] R. Fraher and J. Boyd-Brent. Gestalt theory, engagement and interaction. In *Proceedings of CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3211–3216. Association for Computing Machinery, 2010.

[28] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman. The science of visual data communication: What works. *Psychological Science in the Public Interest*, 22(3):110–161, 2021.

[29] M. Friendly, C.-h. Chen, W. K. Härdle, and A. Unwin. *A Brief History of Data Visualization*, pages 15–56. Springer Berlin Heidelberg, 01 2008.

[30] M. Friendly and D. Denis. The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41:103–30, 02 2005.

[31] X. Gao, W. Lu, D. Tao, and X. Li. Image quality assessment and human visual system. In *Visual Communications and Image Processing 2010*, volume 7744, pages 316–325. SPIE, 2010.

[32] M. D. Gerst, M. A. Kenney, A. E. Baer, A. Speciale, J. F. Wolfinger, J. Gottschalck, S. Handel, M. Rosencrans, and D. Dewitt. Using visualization science to improve expert and public understanding of probabilistic temperature and precipitation outlooks. *Weather, Climate, and Society*, 12(1):117–133, 2020.

[33] J. I. Gold and M. N. Shadlen. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36(2):299–308, 2002.

[34] L. Graham. Gestalt theory in interactive media design. *Journal of Humanities & Social Sciences*, 2(1), 2008.

[35] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. Lineup: Visual analysis of multi-attribute rankings. *IEEE transactions on visualization and computer graphics*, 19(12):2277–2286, 2013.

[36] C. Healey, V. Interrante, and P. Rheingans. Fundamental issues of visual perception for effective image generation. In *SIGGRAPH*, volume 99, pages 1–42, 1999.

[37] K. Healy. *Data Visualization: A Practical Introduction*. Princeton University Press, 2018.

[38] A. Hota and J. Huang. Embedding meta information into visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 26(11):3189–3203, 2020.

[39] D. Huff and I. Geis. *How to Lie with Statistics*. W. W. Norton, 2010.

[40] T. A. Keahey et al. Using visualization to understand big data. *IBM Business Analytics Advanced Visualisation*, 16, 2013.

[41] D. Keim, M. Hao, U. Dayal, H. Janetzko, and P. Bak. Generalized scatter plots. *Information Visualization*, 9:301–311, 12 2010.

[42] A. Kirk. *Data Visualisation: A Handbook for Data Driven Design*. SAGE Publications, 2016.

[43] K. Koffka. *Principles of Gestalt Psychology (Nueva York, Harcourt, Brace &Co)*. Taylor & Francis, 1935.

[44] N. A. Koontz and R. B. Gunderman. Gestalt theory: implications for radiology education. *American Journal of Roentgenology*, 190(5):1156–1160, 2008.

[45] M. Kozak. Improved scatterplot design. *IEEE Computer Graphics and Applications*, 30(6):3–7, 2010.

[46] Leiner, Dominik. *SoSci Survey (Version 3.1.06)*. Munich, Germany, 2019.

[47] J. Li, J.-B. Martens, and J. J. Van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, 9(1):13–30, 2010.

[48] H. Liao, Y. Wu, L. Chen, and W. Chen. Cluster-based visual abstraction for multivariate scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 24(9):2531–2545, 2018.

[49] M. Lima. *Visual Complexity: Mapping Patterns of Information*. Princeton Architectural Press, 2013.

[50] A. Makina. The role of visualisation in developing critical thinking in mathematics. *Perspectives in Education*, 28(1):24–33, 2010.

[51] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2014.

[52] S. Murray. *Interactive Data Visualization for the Web: An Introduction to Designing with D3*. O'Reilly, 2017.

[53] Q. V. Nguyen, N. Miller, D. Arness, W. Huang, M. L. Huang, and S. Simoff. Evaluation on interactive visualization data with scatterplots. *Visual Informatics*, 4(4):1–10, 2020.

[54] P. B. Palmer and D. G. O'Connell. Regression analysis for prediction: understanding the process. *Cardiopulmonary physical therapy journal*, 20(3):23, 2009.

[55] A. V. Pandey, J. Krause, C. Felix, J. Boy, and E. Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3659–3669, 2016.

[56] J. Posada, C. Toro, I. Barandiaran, D. Oyarzun, D. Stricker, R. De Amicis, E. B. Pinto, P. Eisert, J. Döllner, and I. Vallarino. Visual computing as a key enabling technology for industrie 4.0 and industrial internet. *IEEE computer graphics and applications*, 35(2):26–40, 2015.

[57] N. C. Presmeg. Visualisation in high school mathematics. *For the learning of mathematics*, 6(3):42–46, 1986.

[58] G. J. Quadri and P. Rosen. Modeling the influence of visual density on cluster perception in scatterplots using topology. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1829–1839, 2021.

[59] G. J. Quadri and P. Rosen. A survey of perception-based visualization studies by task. *IEEE Transactions on Visualization and Computer Graphics*, abs/2107.07477(01):1–1, jul 2021.

[60] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[61] U.-D. Reips. *The methodology of Internet-based experiments*, pages 373–390. Oxford University Press, 2007.

[62] S. Rendgen and J. Wiedemann. *Understanding the World*. Taschen, 2014.

[63] S. Rendgen and J. Wiedemann. *History of Information Graphics*. Taschen, 2019.

[64] R. A. Rensink. Internal vs. external information in visual perception. In *Proceedings of the 2nd international symposium on Smart graphics*, pages 63–70, 2002.

[65] R. C. Roberts and R. S. Laramee. Visualising business data: A survey. *Information*, 9(11):285, 2018.

[66] A. Ross and V. L. Willson. *Basic and advanced statistical tests: Writing results sections and creating tables and figures*. Springer, 2018.

[67] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412, 2018.

[68] J. Sauro and J. R. Lewis. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.

[69] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2634–2643, 2013.

[70] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum (Proc. EuroVis)*, 31(3):1335–1344, 2012.

[71] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, 2010.

[72] V. Sellam and E. Poovammal. Prediction of crop yield using regression analysis. *Indian Journal of Science and Technology*, 9(38):1–5, 2016.

[73] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier, 2003.

[74] S. Smart and D. A. Szafir. *Measuring the Separability of Shape, Size, and Color in Scatterplots*, page 1–14. Association for Computing Machinery, 2019.

[75] R. J. Sternberg, K. Sternberg, and J. Mio. *Cognitive psychology*. Cengage Learning Press, 2012.

[76] S. Sukamolson. Fundamentals of quantitative research. *Language Institute Chulalongkorn University*, 1(3):1–20, 2007.

[77] C. Suzuki, T. Itoh, K. Umezu, and Y. Motohashi. A scatterplot-based visualization tool for regression analysis. In *Proceedings of the 20th International Conference Information Visualisation (IV)*, pages 75–80. IEEE, 2016.

[78] R. Taylor. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39, 1990.

[79] M. J. Tovée. *An introduction to the visual system*. Cambridge University Press, 1996.

[80] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2001.

[81] E. Tufte and G. Press. *Envisioning Information*. Graphics Press, 1990.

[82] L. Wang, G. Wang, and C. A. Alexander. Big data and visualization: Methods, challenges and technology progress. *Digital Technologies*, 1(1):33–38, 2015.

[83] C. Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2019.

[84] Y. Wei, H. Mei, Y. Zhao, S. Zhou, B. Lin, H. Jiang, and W. Chen. Evaluating perceptual bias during geometric scaling of scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):321–331, 2020.

[85] M. Wertheimer. A source book of gestalt psychology. *Psyccritiques*, 13(8), 1968.

[86] C. Wilke. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media, 2019.

[87] L. Wilkinson. *The Grammar of Graphics (Statistics and Computing)*. Springer-Verlag, Berlin, Heidelberg, 2005.