DATA SCIENCE, DASHBOARDS, AND THE WAY IT WORKS WITH STATISTICS

by

Denise Renee Bradford

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Statistics

Under the Supervision of Susan R. VanderPlas, Ph.D

Lincoln, Nebraska

Month, Year

DATA SCIENCE, DASHBOARDS, AND THE WAY IT WORKS WITH STATISTICS

Denise Renee Bradford, Ph.D.

University of Nebraska, Year

Adviser: Susan R. VanderPlas, Ph.D

Here is my abstract. *(350 word limit)*

COPYRIGHT

© Year, Denise Renee Bradford

DEDICATION

Dedicated to...

ACKNOWLEDGMENTS

Thank you to all my people!

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Statisticians use graphs in almost every stage of their work: we create charts when we get new data, to explore what we have and identify potential problems and opportunities. We fit models based on relationships between variables which are often identified visually. We identify problems with those models based on residual plots and other visual diagnostics. When our modeling work has been completed, we present our results to interested parties using visual displays, because non-statisticians often find it easier to understand data and models through an intuitive visual medium rather than through the mathematical formulae which underlie the statistical work.

Given the wide range of uses for graphs and visual data displays in statistical modeling, it is unsurprising that some graphs are more useful for specific applications such as exploratory analysis, and are unsuitable for other applications, such as presentation to an outside group. In addition, we know that not all visual displays have equal perceptual value (Aspillaga, 1996). The best graphics are designed to account for both the features of the dataset and the features of the intended audience. Some design constraints stem from limitations of the human perceptual system and are common to most potential consumers of the visualization: the sine illusion affects anyone with binocular depth perception, and color recommendations are built around the specific characteristics of the human retina (VanderPlas & Hofmann, 2015b). Other design constraints are due to the audience's experience level: are they used to working with data? Do they understand specialized techniques such as principal component analysis to the point where a plot of factor loadings might be a useful visual display? When we create visualizations for public consumption we have to consider both perceptual factors and the target audience's domain knowledge. In this introduction, we explore previous research related to the construction of interactive and static visual displays for different audiences and consider the implications of this research when designing interactive data displays such as dashboards.

Most research in statistical graphics has been done on static graphics; usually, research also strips away all but the most essential contextual information. As a result, it can be hard to generalize this research to practical applications, where the contextual information surrounding the data is critical and the chart does not just exist in a vacuum.

In the "real world", however, conventions and familiarity often win out over best practice validated by perceptual experiments.

For example, in sports, many coaches desire printable diagrams containing all necessary and valuable information on a single page. As data in sports becomes more prominent, extensive, and collected, this information must be refined.

Thus, in addition to the experimental evidence, we must consider the human element: how to introduce new graphical concepts to stakeholders, and the considerations involved in encouraging stakeholders to adopt these improved graphics. Let us first consider the audience characteristics that affect the selection of graphics. Then, we will engage with considerations based on the data to be displayed. Finally, we will consider the interactions between the audience and the data: how graphics are tested, amended, and hopefully eventually adopted into common use.

1.0.1 The increase of dashboards in day-to-day

Dashboards have seen a significant increase in day-to-day usage as a potent data visualization and decision-making tool. The proliferation of dashboards can be attributed to several factors, including the growing availability of data from various sources and the increasing need for organizations to extract actionable insights. Few found that the widespread use of dashboards is attributable to their capacity to present key performance indicators and relevant metrics in a visually appealing and easily digestible format, Few (2006a). Moreover, technological advancements and the development of user-friendly dashboard platforms have facilitated the creation and effective utilization of dashboards by individuals from diverse fields. Dashboards have revolutionized data analysis and presentation, allowing users to gain valuable insights and make data-driven decisions more effectively.

1.0.2 Misleading/Terrible Graphics

1.1 Audience Considerations

Several factors, including perception, attention, and expertise, can influence our desire and ability to read and engage with data visualization.

Perception and attention are crucial cognitive processes that allow us to interpret and make sense of data visualizations. Perception refers to the manner in which we interpret and organize sensory information from our environment, whereas attention refers to the capacity to selectively focus on particular aspects of this information.

Our ability to perceive and pay attention to pertinent features, such as patterns, trends, and relationships, is essential for comprehending data visualizations. This is especially true when working with unfamiliar or complex data sets, as our ability to focus on pertinent information becomes more difficult.

In addition to perception and focus, domain-specific knowledge is essential for understanding and interacting with data visualizations. Expertise in a particular field can enable individuals to better interpret and comprehend the significance of the presented data, as well as identify potential biases or errors in the visualization.

In conclusion, the ability to perceive and interact with data visualizations requires a combination of perceptual and attentional processes, as well as domain-specific knowledge, in order to interpret and comprehend the presented information.

The term "data visualization" dates back to the 2nd century A.D. drawings and other visual representations were used to investigate the world and record historical events in ancient societies. Throughout human history, data visualization has significantly contributed to invention and discovery Crapo, Waisel, Wallace, & Willemain (2000). The introduction of computer technology dramatically changed the visual representation of data. Using computer-based graphical data visualization, data analysts have become faster and more precise. Data visualization has become an integral component of research in numerous disciplines, such as algorithms, human perception, animation, computer vision, etc. The origin of data visualization being a sub-category, it is regarded as "the science of the visual representation of data," "Data Visualization: a successful design process" defined data visualization as "the representation and presentation of data that exploits our visual perception abilities to amplify cognition," Kirk (2012). This suggests that data visualization involves the exploitation of human visual perception in addition to the presentation of data. Assigning meaning to visualization is not a statistical or computational step but a cognitive one. Each step in the data analysis process is part of a more extensive mental process.

1.1.1 Perception

Human perception is an essential component of data visualization. Colin Ware suggests that perception can significantly enhance both the content and quantity of displayed information, Ware (2012). Perception refers to the organization, interpretation, and conscious experience of sensory data. Perception is also defined as "the process of recognizing (being aware of), organizing (gathering and storing), and interpreting (binding to knowledge) sensory information," Ward, Grinstein, & Keim (2010). Ward et al. explain the notion of perception as the following:

> Simply put, perception is the process by which we interpret the world around us, forming a mental representation of the environment. This representation is not isomorphic to the world, but it's subject to many correspondence differences and errors. The brain

> makes assumptions about the world to overcome the inherent ambiguity in all sensory data, and in response to the task at hand.

Human visual perception is a highly complex and subjective process, and the efficacy of a visualization in communicating objective understanding depends on a vast array of subtle factors, Reuter, Tukey, Maloney, Pani, & Smith (1990). Furthermore, certain situations present unique challenges and lead to systematic errors; can these provide insight into how the brain solves the problem of which objects are represented by which images in general Gregory (1968). Additional to the general process of human visual perception process and efficacy of a visualization in communicating, short-term memory is crucial to the effectiveness of statistical graphics. Research indicates that our short-term memory can only store a limited amount of information at any given time. Therefore, designers must present data in a manner that is simple to comprehend and remember.

Untrained analysts can and do "analyze" data with only their natural mental abilities - The mind performs its data analysis-like process to create detailed understandings of reality from bits of sensory input. In a later chapter, we will show how a utilizing parallel coordinate plot is one method for achieving a simple design to comprehend. These plots enable viewers to compare multiple variables concurrently, thereby reducing cognitive load and making it easier to identify patterns and trends.

Examining the Gestalt principles, which describe how our brain organizes and interprets sensory information to form coherent patterns and objects, is one way to gain a deeper understanding of how perception functions.

1.1.2 Gestalt Principles

Our visual interpretation of the world is a major factor in why humans perceive the world and its objects as organized, regular, and simple shapes, schemas, figures, or forms. The theory of Gestalt has philosophical and psychological roots that date back to the late 1800s. Gestalt therapy is founded on the notion that overall perception is contingent upon the interaction of numerous factors. These principles include proximity, similarity, closure, continuity, symmetry, and figure-ground relationships formally outlined in "Principles of Gestalt Psychology", Koffka (2013).

Gestalt principles include:

- Proximity: Objects close to each other are perceived as related or grouped.
- Similarity: Objects that are similar in some way (e.g., shape, color, size) are perceived as related or grouped.
- Continuation: The human eye follows lines and patterns, so designers can use this principle to guide the viewer's gaze through a display.
- Closure: The human brain tends to complete incomplete figures or patterns, so designers can use this principle to create the illusion of missing information.
- Figure-Ground: The human brain separates the foreground (figure) from the background (ground), so designers can use this principle to create visual hierarchy and emphasis.
- Contrast: The human eye is drawn to high-contrast areas, so designers can use this principle to create emphasis and hierarchy.

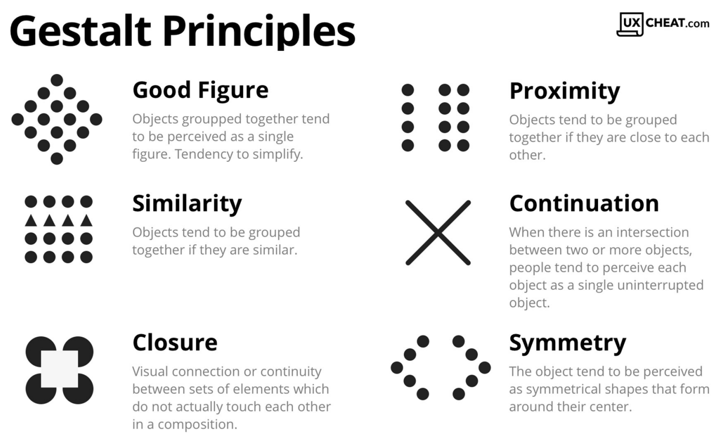


Figure 1.1: Gestalt Principles with Examples

- Symmetry and Balance: The human eye finds symmetry and balance visually pleasing, so designers can use this principle to create a sense of harmony and order.

These principles are based on cognitive psychology and understanding how the human brain processes visual information. By applying these principles to dashboard design, designers can create visual arrangements that make it easier for viewers to understand the relationships between data elements. For example, proximity can be used to group related elements together, while symmetry can be used to create balance and harmony in the overall layout of the dashboard. At its most basic, the entire form is perceived (or emerges to our visual pathways) as opposed to its component parts.

All of this suggests that our brain frequently perceives things differently than what is actually present. If you are familiar with optical illusions or the famous “gorilla in the crowd” experiment, you are aware that we do not always process everything in our visual field. There is simply too much information for our brains to process, and if we tried to interpret it all, we would be rendered paralyzed. We organize the world according to Gestalt principles and pre-attentive attributes so that it is familiar, makes sense, and is easy to process. These principles guide how people perceive and make sense of the world around them, and they play a critical role in designing effective visual displays, such as dashboards. Understanding the Gestalt principles can also cast light on how information is processed and stored in short-term memory and attention's critical role in this process.

1.1.3 Attention and Memory

The Gestalt principles describe how humans perceive and organize visual information into meaningful patterns and structures to help the brain to effectively process and organize incoming visual information, making it easier to attend to and remember. It's worth exploring the efficacy of short-term memory when developing effective dashboards.

Short-term memory (STM), also known as working memory, is the stage of temporary storage and processing where the majority of memory retention effort is expended. According to Alan Baddeley's Working memory: Theories, models, and controversies, STM is a limited-capacity system prone to interference and decay, A. Baddeley (2012). Selective attention is essential for the maintenance of STM because it allows us to filter out irrelevant information and concentrate on what is essential, Cowan (2001).

Visual aids such as charts and diagrams can improve short-term memory by allowing us to encode and retain information more effectively, according to research, Alvarez & Cavanagh (2004). Consequently, utilizing visual aids such as charts can be advantageous for enhancing our short-term memory.

Furthermore, annotations can also help aid short-term memory. By adding annotations, such as notes or highlights, to information we are trying to remember, we can improve our recall of the information later on, Alvarez & Cavanagh (2004).

Feature Integration Theory (FIT)

According to the Feature Integration Theory (FIT), STM is composed of two stages: pre-attentive processing and focused attention A. Treisman (1998). Parallel and independently, the brain processes the physical characteristics of an object, such as its color, shape, and orientation, during pre-attentive processing. However, focused attention is required to bind these features into a coherent object representation in STM. STM can be improved through various strategies, such as rehearsal, chunking, and elaboration Oberauer (2009). For example, by repeating a phone number several times or breaking it down into chunks of two or three digits, we can increase the likelihood of it being stored in STM. Similarly, by elaborating on the information we want to remember, such as creating mental associations or visual images, we can enhance its retention in STM Bui & Myerson (2014).

STM is a dynamic and malleable cognitive system that is crucial to our daily lives. Understanding the mechanisms underlying STM and how to improve it can have significant implications for learning, memory, and the treatment of memory disorders. By analyzing the relationship between attention and working memory, we can gain insight into how we construct meaning from the information in our environment.

1.1.4 Constructing Meaning

Gestalt psychology suggests that humans actively construct meaning by organizing information into patterns and wholes Wertheimer (1938). Both top-down and bottom-up processing are involved in the process of meaning construction. Bottom-up processing entails analyzing sensory data from the environment and constructing perceptions based on this data. Top-down processing is the influence of prior knowledge, expectations, and context on the perception and interpretation of incoming sensory data.

Together, top-down and bottom-up processing facilitate the encoding and retrieval of information in the context of short-term memory. Selective attention, the ability to focus on relevant information while ignoring irrelevant information, is an example of top-down processing that aids in the encoding and retrieval of information in short-term memory Cowan (2010).
According to the feature integration theory, the perception of objects involves both the bottom-up analysis of individual features and the top-down processing of higher-level features in order to form a complete perception A. M. Treisman & Gelade (1980).

The Gestalt principles of perception emphasize the significance of bottom-up and top-down processing in constructing meaning from sensory data. Both types of processing are involved in encoding and retrieving information, which has significant implications for understanding how short-term memory works.

1.1.5 Expertise

However, creating effective graphics is not a simple task, and proficiency in this area is required to create high-quality visualizations. This essay will discuss the contributions to the field of graphics made by research on psychological processes, automaticity, readily available information, and practice effects.

Cognitive Processes - the way we think about and approach a task.

As we become more proficient in a particular skill, we develop more complex and efficient mental models or schemata, a heuristic technique to encode and retrieve memories, the majority of typical situations do not require much strenuous processing. These mental models help us to organize information in a meaningful way, and to quickly identify and solve problems related to the task. This process is known as cognitive restructuring and is facilitated by developing domain-specific knowledge Ericsson & Lehmann (1996). For example, a basketball coach is able to quickly recognize patterns and positions on a court that are common in basketball, which allows them to make decisions more quickly and accurately than a novice coach or player.

Automaticity - the ability to perform a task without conscious effort or attention

As our proficiency in a task increases, our performance becomes more automatic, thereby freeing up cognitive resources for other tasks. The development of procedural knowledge, which is the ability to perform a series of steps or actions in a particular order, facilitates this process Fitts & Posner (1967). For example, a well-trained quarterback can throw a ball without looking at the wide receiver because their throwing movements have become automatic.

Information Readily Available - the way we process information related to a task

As our proficiency increases, we can recognize and retrieve pertinent information more rapidly and precisely than a novice. This is made possible by the creation of domain-specific knowledge structures that allow us to retrieve pertinent information from memory quickly Chase & Simon (1973). For example, a medical expert can quickly identify signs and diagnose a patient using their knowledge of disease symptoms and risk factors.

Practice Effects - extensive practice and experience

Practice effects are the performance enhancements that result from repeated practice. These gains are frequently most significant at the outset of practice, but gradually diminish as the individual approaches their performance ceiling Anderson (1982). The development of procedural knowledge and automaticity, which allow for more efficient and accurate task performance, facilitates the effects of the practice.

The contributions of research on psychological processes, automaticity, readily available information, and practice effects to the field of graphics have significant implications. Expertise is required to create high-quality graphics, which requires a thorough understanding of design principles and the capacity to work quickly and efficiently. The use of automatic processing and domain-specific knowledge can aid designers in processing and deciding on design elements efficiently and quickly. The creation of standardized design templates, workflows, and other tools can aid in enhancing the efficiency and effectiveness of the design process. Design skills can be improved through deliberate guidelines, and educators must focus on developing skills that can be applied in a variety of contexts.

As the significance of graphics in numerous fields continues to rise, the demand for specialists in this area will only intensify. By comprehending the contributions of research on psychological processes, automaticity, readily accessible information, and practice effects, designers, educators, and trainers can develop more effective approaches to graphics design and education. This can help ensure that the graphics used to convey complex information are clear, concise, and effective, making it easier for individuals to comprehend and interpret the required information.

1.1.6 Engagement with the data

The goal of data analysis is to extract meaningful insights, patterns, and knowledge from data. The process of data analysis involves collecting, cleaning, transforming, and modeling data, followed by the use of statistical and machine learning methods to uncover patterns and relationships within the data. The end goal of data analysis is to support decision making and provide a basis for informed action. Data analysis can help organizations to better understand their customers, market trends, and operational performance. Additionally, data analysis can support scientific research by helping researchers to test hypotheses, develop theories, and gain a deeper understanding of complex phenomena. Ultimately, data analysis aims to turn data into actionable insights and information that can inform and improve decision-making.

1.2 Data Considerations

John Tukey was the first to organize the collection and methods associated with philosophy into Exploratory Data Analysis (EDA). Previous research by Tukey focused on graphics as a tool for exploratory analysis. In “Exploratory

Data Analysis," Tukey wrote that graphics and charts often display data with more enhanced understanding than a table, Tukey & Wilk (1966). Tukey outlines detailed the types of different graphics and in which situations to utilize these graphics. He was a strong advocate for the importance of EDA as a crucial first step in the data analysis process and emphasized the need for visualization and interactive techniques to understand patterns and relationships in data.

Tukey's Principles of EDA have become a cornerstone in the field of statistics and have been adopted by data professionals in various industries. Tukey's principles have enabled data professionals to understand complex data sets better and make more informed decisions by emphasizing the importance of visual exploration, data characterization, and model critique. In this way, Tukey's Principles have revolutionized our data analysis approach and become the foundational framework for EDA.

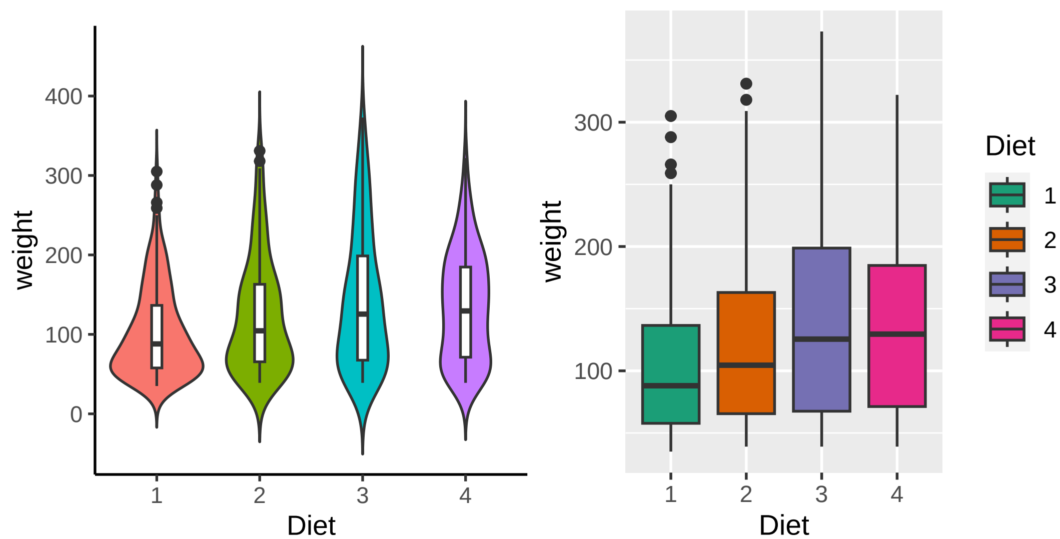
Tukey's Principles in EDA:

1. Graphical exploration, looking for patterns or displaying fit, the method demonstrates things about data that a single numeric metric does not understand. This has been useful in graphing the data before you develop summary statistics.

2. Describing the general patterns of the data. This step should be insensitive to outliers. In general, think about the types of resistant measures (i.e., median or mean). This step is making sure to determine data patterns.

3. The natural scale/state that the data are at their best. This will be the step at which the scale of data can be helpful for analysis. The reexpressing data to a new scale by taking the square root or logarithmic scale.

4. The mostly known parts of EDA but is done in the way of accessing fit of the data. This is taught in every statistics 101 class. The growth of machine learning and prediction methods have now used residuals more in the toolbox to assessing the best prediction models.



Data visualizations are an integral part of the EDA process, enabling analysts to discern patterns and relationships in the data that would otherwise be difficult to discern from tabular data alone. Through data visualization, analysts can quickly identify trends, outliers, and other patterns that may be missed through numerical analysis alone. Moreover, visualizations facilitate the communication of findings to non-technical stakeholders, allowing them to comprehend complex data sets more efficiently. Through visualizations, analysts can also identify potential issues or biases in the data, resulting in better decisions and models. Thus, visualizations play a crucial role in the EDA process by enabling analysts to more effectively explore, comprehend, and communicate data-derived insights. During the initial EDA stage, an analyst may find that a variable or a covariate is directly related to the dependent variable when looking at a correlation heatmap or a scatterplot. The basic understanding can be formalized to visualize the discovery process.

The field of graphical communication, which is directly related to EDA, semiology, and their use in touch, has been a valuable tool and extension of the EDA thoughts that Tukey expressed. One of the fundamental principles of semiology is the relationship between signifier and signified, in which a visual element (the signifier) represents a particular meaning or concept (the signified), (**barthes1972?**). Another essential concept in semiology is using syntax and semantics to convey meaning in graphic communication effectively. This includes both the syntax and semantics of a graphic's visual elements, Monmonier (1985).

Using color to represent data on maps is an example of successful graphical communication utilizing semiology. By using different colors to represent different data points, viewers can comprehend patterns and relationships in the data quickly and easily. Jacques Bertin writes in "Semiology of Graphics" that color can be used to "emphasize a point, distinguish one category from another, or establish a relationship between two points", Monmonier (1985). In addition, Bertin explains that the use of color can help overcome language barriers, making it easier for the audience to comprehend the presented information.

The application of semiology in graphical communication is not devoid of obstacles. One difficulty is the possibility of misinterpretation, in which viewers may assign a different meaning to a visual element than was intended, Monmonier (1985). Another concern is the possibility of cultural differences in interpretation, in which a visual element may have a different meaning in one culture versus another, Norman (2013).

Despite these obstacles, semiology in graphical communication remains an indispensable tool for effectively conveying information. By understanding semiology principles and syntax and semantics' role in graphical communication, designers can create compelling visual representations that convey information clearly and concisely.

By utilizing visual elements such as bars and lines to represent data, graphs can make complex information more understandable to viewers. For instance, a line graph can be used to illustrate the change in the value of a stock over time, making it easier for investors to identify trends and patterns. Leland Wilkinson writes in his book "The Grammar of Graphics" that "graphical methods are not only superior to other forms of communication, but also superior to numerical or verbal methods for certain types of data and reasoning," Wilkinson (2012).

It proposes that any statistical graphic can be broken down into a set of essential components, or "grammar," that can be combined in different ways to create a wide range of visualizations, following a layered approach to describe and construct visualizations or graphics in a structured manner.

The components of the grammar of graphics include:

- Data: The raw data being visualized represents a set of observations or values.

- Aesthetic Mappings: The mapping of data variables to visual properties such as position, color, shape, and size.
- Scales: The mapping of data values to visual values, such as mapping a numerical value to a bar height.
- Geometries: The basic shapes representing the data, such as points, lines, bars, and histograms.
- Facets: The plot division into multiple subplots, each representing a different subset of the data.

For example, a bar chart can be created by mapping a categorical variable to the x-axis, mapping a numerical variable to bar heights, and using rectangular bars as the geometry. Moreover, mapping two numerical variables can create a scatter plot to the x and y positions and use points as the geometry. Finally, the “Grammar of Graphics” provides a systematic way of thinking about visualizations, making it easier to choose the appropriate visual representation for a given dataset.

Michael Friendly, a leading expert in data visualization, has utilized the principles of the grammar of graphics to develop innovative teaching methods that make complex data visualization concepts more accessible to a broader audience. Friendly has investigated the origins and development of graphic techniques, tracing their evolution from antiquity to the present. He used SAS with hands-on experiments to present categorical data analysis visually, Friendly (2014). By emphasizing the role of graphical methods in scientific

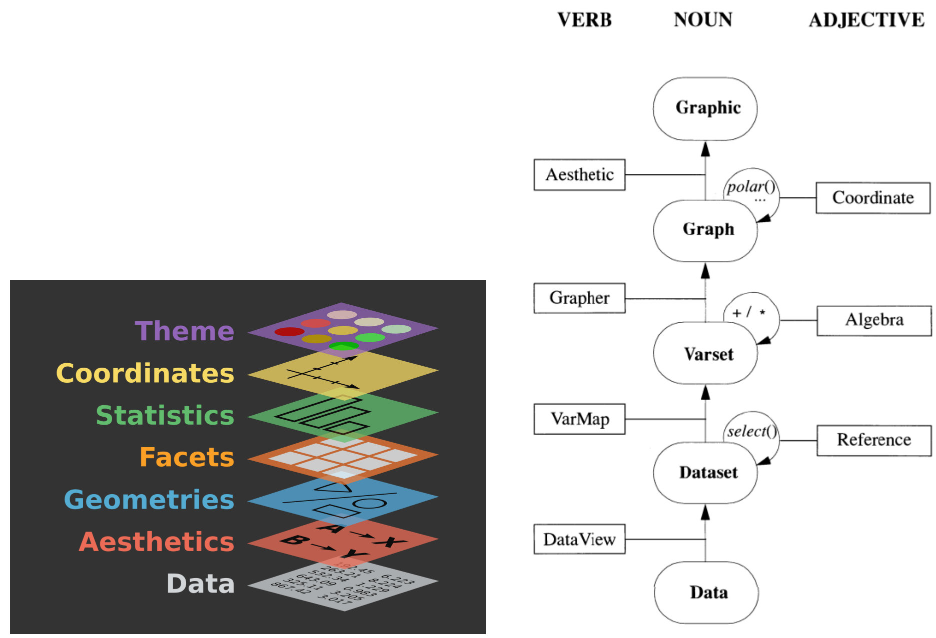


Figure 1.2: Grammar of Graphics Diagram of Wickham and Wilkinson's work

discovery, Friendly has helped promote his use in various disciplines, from the natural sciences to the social sciences and beyond. In his book "Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization," Friendly provides a comprehensive overview of the key milestones in the evolution of statistical graphics, including the contributions of pioneers like William Playfair, Charles Minard, and John Tukey, Friendly & Denis (2001).

As discussed regarding semiology, Tukey's Exploratory Data Analysis (EDA), and the introduction of the Grammar of Graphics, we should be mindful that a well-constructed graphic can be misleading out of context. Compelling graphics can be a powerful tool for communicating complex information, making numerical accuracy, engagement, correct decision-making, and accurate predictions crucial.

In the context of data visualization, numerical accuracy refers to the precision and correctness of the numerical data displayed in graphics. Accurate graphics can assist users in comprehending complex numerical data and making more informed decisions, Cardoso, Leite, & Aquino (2016).

Engagement in the context of data visualization is the extent to which viewers are drawn to and interested in the displayed data. Engaging graphics can encourage users to interact with and explore data further, resulting in a more thorough comprehension of the data.

Correct decision making refers to the capacity of data visualization to enable users to make informed and precise decisions based on the presented information. Clear, accurate, and well-designed graphics can help users recognize patterns and insights, resulting in more effective decision making.

Correct predictions refers to the capability of graphics to accurately forecast or predict future events or outcomes. For accurate predictions, data visualization must include trustworthy data, sound statistical models, and efficient visualization techniques.

"Graphical Tests for Power Comparison of Competing Designs" by Hofmann et al. presents a graphical method for comparing the power of two or more competing designs in an experimental study (**hofmann2012?**). The article demonstrates that the graphical method is a useful tool for comparing the effectiveness of various experimental designs. It enables researchers to visualize and compare the effectiveness of different designs in an intuitive and straightforward manner.

Two methods for measuring the particle size distribution in a chemical process were compared in one study. The study evaluated both designs under various operating conditions and compared their power using a graphical method. The results demonstrated that one design was more effective at detecting differences in operating conditions than another.

Static Visualization is commonly used in the communication phase of data science workflows, and data scientists sometimes use them as part of the analysis. John Tukey's EDA methods are currently known and well-vetted in the field. However, Satyanarayan et al. addressed this by introducing a high-level grammar of graphics called "Vega-Lite," which presents a set of standardized linguistic rules for producing interactive information visualizations using a concise JSON format for data to be represented by the grammar Satyanarayan, Moritz, Wongsuphasawat, & Heer (2016). Vega-Lite has been directly implemented in R via the `ggvis` package using the same - albeit slightly lower-level.

Understanding cognitive load is crucial for designing compelling data visualizations, as it influences how users perceive, process, and remember the data presented in the visualization. When designing visuals, it is essential to consider the cognitive load they may place on the viewer. Cognitive load is the amount of mental effort required to process information, and minimizing it can enhance a graphic's effectiveness. In addition, displaying as much raw data as possible while minimizing cognitive load can improve the graphic's clarity and precision. Here are some general guidelines for making better graphics with works from Few ((**few2012?**)), Tufte ((**tufte1983?**)), and Cairo (Cairo (2016)):

1. Keep it simple - Avoid overwhelming the viewer with too much information at once by employing a clear and concise design with minimal distractions.

2. Use visual hierarchy - Utilize size, color, contrast, and placement to highlight important information and direct the viewer's focus.

3. Choose appropriate charts - Choose the chart type that best illustrates the data and facilitates comprehension.

4. Label clearly - Use labels that are clear and concise for axes, legends, and other essential information to avoid confusion.

5. Use data-to-ink ratio - Focus on the data by minimizing the amount of non-data ink, such as decorative elements or excessive grid lines.

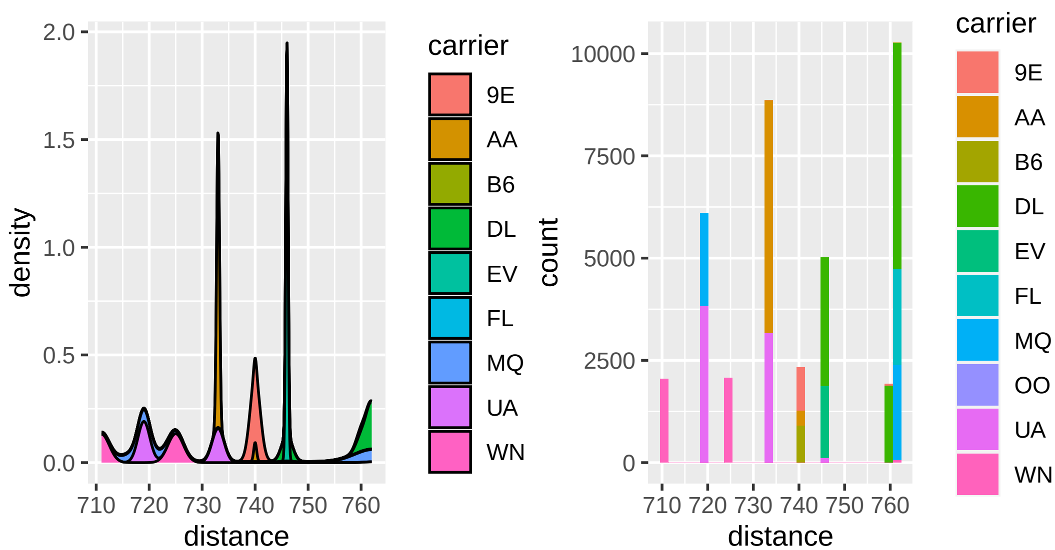
6. Avoid distortion - Use appropriate scaling and avoid distortions to ensure that the graphics accurately represent the data.

7. Provide context - Add context to assist the viewer in comprehending the significance of the data and its relevance to the topic.

```
## Warning: Groups with fewer than two data points have been
## dropped.
```

```
## Warning: Removed 1 rows containing missing values
## (`position_stack()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with
## `binwidth`.
```



On the other hand, Interactive graphics provide a more dynamic and engaging way to explore and analyze complex data sets than traditional static visualizations. By allowing users to manipulate and explore data in real-time, interactive graphics can reveal hidden patterns and relationships that may be difficult to discern in static visualizations, making them a valuable tool for data analysis.

1.2.0.1 Interactive Graphics

As previously mentioned, theories behind visual representation include - graphical comprehension (Cleveland & McGill (1984)), preattentive processing (Ware (2012)), gestalt theory (Few (2009)), and graphical excellence (Edward R. Tufte (2001)). Interactive graphics offer a number of advantages when analyzing complex data sets, and technological progress has played a crucial role in making these tools more accessible and widely adopted, especially in interactive graphics. Compared to static visualizations, interactive graphics offer a more engaging and dynamic way to explore and analyze complex data sets. To gain a deeper understanding of the data, users are able to modify parameters, zoom in on specific regions, and rapidly explore various variables. In addition, interactive graphics offer a more intuitive method of communicating findings to non-technical users, making them a valuable asset for data-driven decision making. Interactive graphics are excellent for EDA; they are designed for exploring rather than presenting information (and more) and can be obtained by directly querying the graphic, Unwin, Volinsky, & Winkler (2003). Overall, interactive graphics are a potent data analysis tool, allowing analysts to gain a deeper understanding of complex data sets and to make more informed decisions.

The area of interactive graphics is still very much a work in progress despite existing as a field of research since the late 1960s. Developments are driven partly by new technology, such as `d3` (Bostock, Ogievetsky, & Heer, 2011). Visualizations are more than just a picture. They are now a tool that facilitates analytic activity through different modes of interaction (Yi, Kang, Stasko, & Jacko, 2007). Visualization is context-free, as it can mean different things to different people depending on the situation (Parsons & Sedig, 2014).

In recent years, interactive graphics that enable users to manipulate and explore visualizations in real-time have grown in popularity and align with the van Wijk Simple Visualization Model framework. The van Wijk Simple Visualization Model is a diagrammatic representation that provides a simple and effective way to understand and visualize the flow of information and data through a system. van Wiij's simple visualization model shows how insights are generated as the human participates in a feedback loop between reading and interacting with visualization, Van Wijk (2005). It is a commonly used

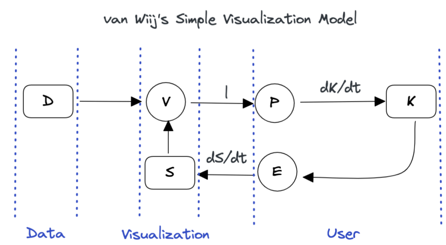


Figure 1.3: van Wiij Simple Visualization Model

tool in EDA, the initial step in the data analysis. The van Wijk model can represent data flow from data sources, through intermediate processing stages, to the final visualization of results. This model is also context-free, allowing for the focus to be on the feedback loops between visualization and the user. The model helps to identify the various steps involved in the visualization process, from the collection and processing of data to the presentation of results. By doing so, it supports the design of more effective and user-friendly visualizations, which can enhance the overall user experience.

While the van Wijk Simple Visualization Model provides a valuable framework for designing compelling visualizations, it should not be used in isolation. Instead, effective data visualization requires a thorough understanding of human perception, cognition, and Human-Computer Interaction (HCI) principles. Later in this review, we will explore the fundamental principles of Human-Computer Interaction and how they can be applied to the design of compelling interactive graphics.

Based on the above best practices on the concept of cognitive load in graphics, the theory of manipulation of visualizations provides a set of guidelines and best practices for designing interactive graphics that minimize cognitive load and facilitate practical data analysis.

As interactive visualizations play a more significant role in information systems, designers must know what tasks, visual representations, and interaction techniques are available and how they work to facilitate analytical reasoning. They must decide on the most effective visual representation without being able to estimate every user's ability to read and interpret the visualization.

An influential framework developed by Vessey et al refers to the degree to which a person's cognitive abilities match the cognitive demands of a task, Vessey & Galletta (1991), in the field of cognitive load research has been improved by Heer and Bostock, providing examined how the complexity of interactive visualizations influences users' cognitive load, Heer & Bostock (2010). The authors discovered that more intricate visualizations tend to increase cognitive load, especially for users with lower visual literacy. They recommended that designers consider the cognitive load of interactive visualizations and strive to reduce complexity whenever possible.

While the direct relationship to traditional mobile devices, such as phone screens, will not be discussed, it is noteworthy to account for the impact of larger screen mobile devices, such as iPads and other portable devices, concerning a user's cognitive abilities. Eissele et al investigated the cognitive load of mobile interactive visualizations, Eissele, Weiskopf, & Ertl (2009). The authors discovered that the limited input options and small screen size of mobile devices can increase cognitive load in interactive visualizations. To reduce cognitive load on mobile devices, they suggested that designers employ suitable visual encoding techniques and simplify interactions.

Lastly, Toyama et al. investigated the impact of interactive elements on cognitive load in visualizations, Toyama, Sonntag, Orlosky, & Kiyokawa

(2015). Based on the nature of the task and the user's familiarity with the interactive features, the authors discovered that interactive features can both increase and decrease cognitive load. They suggested that designers should evaluate the cognitive load caused by the interactive elements they include in visualizations

These studies suggest that designers should carefully consider the cognitive load implications of interactive graphics and strive, whenever possible, to reduce complexity. Simplifying interactions and implementing suitable visual encoding techniques can reduce cognitive load, especially on mobile devices. In addition, designers should evaluate the cognitive load caused by the interactive features they include in visualizations.

1.3 Audience-Data Interactions

Effective design of interactive graphics necessitates a comprehensive knowledge of Human-Computer Interaction (HCI) and User Experience (UX) design principles to ensure that the visualizations are engaging, informative, intuitive, and user-friendly.

Designing an effective interactive dashboard involves much more than simply selecting a set of visualizations and arranging them on a page. It requires a comprehensive knowledge of Human-Computer Interaction (HCI) and User Experience (UX) design principles and the ability to apply these principles to data analysis and visualization. Before designing an interactive dashboard, it is crucial to have a thorough understanding of the underlying HCI and UX frameworks and the specific needs and preferences of the intended audience. By understanding the principles of HCI and UX design, it is possible to create

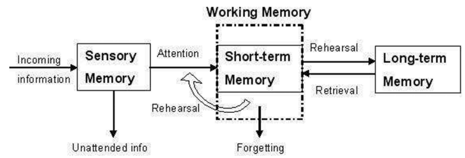


Figure 1.4: Information Processing Model created by Atkinson and Shiffin

interactive dashboards that facilitate effective data analysis and offer a seamless and engaging user experience. This can result in greater engagement with the data, more profound insights, and better-informed decisions.

1.3.1 Human-Computer Interaction (HCI)

Human-Computer Interaction (HCI) refers to the study of the interaction between humans and computers. It encompasses the design, evaluation, and implementation of computer systems that are intuitive, efficient, and effective to use.

In 1968, the model proposed by Atkinson and Shiffrin posited a sequential flow of information through memory systems, emphasizing the processes of encoding, storage, and retrieval Atkinson & Shiffrin (1968). Unless actively practiced, short-term memory has a limited capacity and a short duration of retention. If information is deemed significant or sufficiently rehearsed, it can be encoded and transferred to long-term memory, which has an almost unlimited capacity and long-term storage.

Baddeley expanded our understanding of working memory by emphasizing its active processing nature, expanding upon the model of Atkinson and Shiffrin. The influential model developed by Baddeley, known as the working memory model, proposed a more complex structure with multiple components, Baddeley Alan D. (1976).

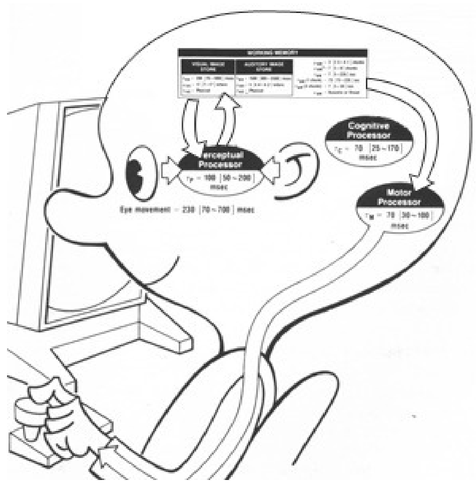


Figure 1.5: GOMS Model created by Card Moran and Newell

Together, Atkinson and Shiffrin's model of information processing and Baddeley's model of working memory provided a comprehensive framework for studying memory and cognition. They contributed to the understanding of how information is processed, encoded, stored, and retrieved in human memory systems, laying the groundwork for subsequent research and theories in cognitive psychology.

In the 1970s and 1980s, HCI researchers began to concentrate on developing user interface design theories and methodologies. The "GOMS" model, created by Card, Moran, and Newell in 1983, was an influential framework that provided a way to model user behavior when interacting with computer systems, Card, Moran, & Newell (1983). During this time, the field of usability engineering emerged, which involved designing and testing interfaces to ensure that they were user-friendly and efficient.

The 1990s witnessed a significant rise in the popularity and accessibility of personal computers, as well as the emergence of the World Wide Web as the dominant information-sharing platform. This resulted in an emphasis on designing interfaces for web-based applications and research into the usability of websites. In 1995, Don Norman published his influential book "The Design of Everyday Things," emphasizing the importance of designing intuitive and user-friendly products and interfaces (revisited and revised in 2013, Norman (2013)). His influential book examines the principles of user-centered design and the critical role that psychology plays in shaping user experiences. It emphasizes the significance of user-friendly and intuitive designs in everyday objects.

In the 2000s, HCI research expanded to include new technologies such as mobile devices, touchscreens, and virtual and augmented reality, O'Brien & Toms (2008). There was also an increased focus on designing interfaces for diverse user populations, including people with disabilities and older adults.

HCI research evolves and adapts to new technologies and user requirements. The use of artificial intelligence and machine learning in interface design, designing interfaces for wearable devices, and designing interfaces for social and collaborative systems are some current research areas.

1.3.2 User Experience (UX)

The history of UX can be traced back to the early days of human-computer interaction (HCI) research, but it has evolved to become a distinct discipline in its own right. User Experience (UX) refers to the overall interaction between a user and a product or service. It includes the interface, functionality, content, and aesthetics.

The origins of user experience research can be traced back to the 1970s and 1980s. At this time, researchers were primarily focused on developing efficient and effective user interface design theories and methodologies. As computer use became more prevalent in the 1990s, the emphasis shifted to designing interfaces that were not only functional but also enjoyable to use. In his influential 1995 book, “The Design of Everyday Things,” Donald Norman argued that good design should center on the needs and behaviors of the user. Jesse James Garrett coined the term “user experience design” in 2000 to describe the process of designing intuitive, efficient, and enjoyable user interfaces, Garrett (2003). The rise of mobile devices in the late 2000s and early 2010s presented UX designers with new challenges and opportunities. As users began interacting with devices featuring smaller screens and touch-based interfaces, designers were forced to reevaluate traditional interface design principles and adopt new interaction paradigms, Hassenzahl (2010).

Alan Cooper’s “The Inmates Are Running the Asylum” - Cooper discusses the significance of user-centered design and the challenges and opportunities in the software design field in this book. He advocates placing the needs and objectives of users at the forefront of the design process, Cooper (1999). “About Face: The Essentials of Interaction Design” - This comprehensive guide covers interaction design principles, methods, and best practices. It provides designers with useful insights and real-world examples to assist them in developing effective and engaging user experiences, Cooper, Reimann, Cronin, & Noessel (2014).

Dan Saffer’s “Designing for Interaction: Creating Smart Applications and Clever Devices” delves into the history and evolution of interaction design and offers practical advice for creating meaningful user experiences. The book examines a variety of design strategies and techniques for designing interactive applications and devices, Saffer (2007).

“A Project Guide to UX Design: For User Experience Designers in the Field or in the Making” by Russ Unger and Carolyn Chandler provides step-by-step approaches, techniques, and examples for the entire UX design process. It is intended for both seasoned designers and newcomers to the field, Chandler & Unger (2012).

“Information Architecture: For the Web and Beyond” by Louis Rosenfeld and Peter Morville focuses on the optimal organization of information for online and offline user experiences. It investigates the principles and methods of information architecture as well as its influence on user-centered design, Morville, Rosenfeld, & Arango (2015).

Overall, understanding the design principles of UX and HCI and the importance of testing static graphics can significantly inform the design of interactive dashboards, as we will explore in more detail later in this review.

1.3.3 UX and Cognitive Load

Since 1980s, cognitive load theory (CLT) of Human-Computer Interaction (HCI) was developed by cognitive psychologist John Sweller. The primary focus of Sweller’s research was on learning and the cognitive processes underlying the acquisition of new knowledge and skills.

The evolution of CLT was influenced by a number of fundamental concepts and theories. Information processing theory, which posits that individuals process information sequentially using sensory memory, working memory, and long-term memory, was one of the major influences. Sweller’s research extended this theory by highlighting the limitations of working memory and its role in cognitive load, Sweller & Chandler (1994). In addition, CLT distinguished between intrinsic, extraneous, and relevant cognitive load. The intrinsic load of a task is its inherent complexity, which cannot be altered by the interface design. On the other hand, the extraneous load is the cognitive load imposed by the design of the learning environment or interface, which can be reduced through practical methods. Finally, germane load refers to the cognitive effort required for learning and acquiring knowledge, (**hollender20210?**).

The original purpose of CLT was to improve instructional design and learning outcomes by minimizing irrelevant cognitive load and optimizing relevant cognitive load. However, its principles and concepts have been applied to the field of HCI, recognizing the significance of cognitive load in the design of interactive systems and user interfaces.

The cognitive load theory of HCI illuminates human-computer interaction cognitive processes. By comprehending the cognitive load imposed by interfaces and learning environments, designers can create more efficient and user-friendly systems that maximize users' mental resources. CLT has continued to evolve and inform interface design over time, improving HCI's usability, user experience, and cognitive support.

Both CLT and HCI were based on the same theories of cognition. Both emphasized the reduction of unnecessary cognitive load. With the concept of germane cognitive load, CLT incorporated principles to promote germane learning processes, which may result in a rise in cognitive load.

In relation to Cognitive Load Theory (CLT), the split-attention principle addresses the negative effects of dividing a learner's attention between multiple sources of information during learning tasks. The split-attention principle focuses specifically on reducing extraneous cognitive load, which refers to cognitive processing that does not contribute directly to learning, (**nielsen1994?**).

Because the goal of instructional design should be to present information in an integrated and coherent manner to reduce the effect of divided attention. This can be accomplished by spatially aligning relevant text and visuals, such as by placing labels near the corresponding elements in a diagram or by employing captions to provide explanations directly alongside visual representations. The split-attention principle optimizes cognitive resources for learning tasks by reducing the necessity for learners to mentally integrate distinct sources of information.

1.3.4 Infomation Search Enviroment

In the article, "Web Designers and web users: Influence of the ergonomic quality of the web site on the information search," written by Aline Chevalier and Maud Kicka outline investigate the difficulties of web designers or compare their respective activities, Chevalier & Kicka (2006). The two experimental studies presented in this article compare the strategies developed by professional web designers and (novice versus experienced) web users while searching for information on websites with varying ergonomic qualities. The results of the two experiments, that web designers are incapable of predicting the strategies of novice users and do not behave like novices. Therefore, techniques for assisting web designers in developing a user-centered activity are required, and a few of these techniques are suggested at the conclusion of this article.

In the article, "Predicting graph reading performance: a cognitive ap-

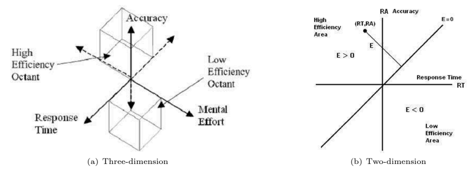


Figure 1.6: Multi-dimensional visualization efficiency (a) modified from Tuovinen et al. [2004]; (b) modified from Kalyuga et al. [1999]

proach, by Huang, Hong, and Eades outlines that performance and preference metrics are frequently employed in evaluating visualization techniques, Huang, Hong, & Eades (2006). To introduce a cognitive method for assessing the effectiveness and efficiency of visualization. Huang et al. propose a model of user performance, mental effort, and cognitive load (memory demand) and incorporate additional mental effort and visualization efficiency measures into our analysis.

Huang et al. build on the understanding that Richard Atkinson and Richard Shiffrin, wrote the information processing model of memory.

Measuring Visualization Efficiency can be calculated using the method proposed in 2004 to measure the relative effectiveness of various instructional conditions, Tuovinen & Paas (2004). This method relates mental effort to performance measures within the context of visualization efficiency, which is defined as the ratio of cognitive cost (in this study, mental effort and response time) to response accuracy

The efficiency (E) is calculated by converting mental effort (ME), response time (RT) and accuracy (RA) into z-scores, and combining them using the following formula:

$$E = \frac{RA - ME - RT}{\sqrt{3}}$$

where E of zero means that cognitive cost and performance accuracy are balanced.

1.3.5 Testing static graphics

While understanding HCI and UX design principles helps create compelling interactive graphics, it is equally important to rigorously test and evaluate the usability and efficacy of both interactive and static graphics to ensure they achieve their intended purpose. Kosara highlighted the significance of user testing in identifying potential problems in static visualizations, Kosara (2007). User testing aids in comprehending the users' behavior and perceptions of the visualization, thereby facilitating the identification of issues that designers may have overlooked. User testing also identifies the graphic's confusing or ineffective visual aspects, allowing designers to refine the design and increase its effectiveness.

The evolution of testing static graphics has progressed through several chronological moments, each building on the knowledge and techniques of the past, starting with the thoughts of Cleveland and McGill to Hofmann, who explored the statistical testing of categorical data.

Firstly, Cleveland and McGill investigated the effect of various bar chart designs on the accuracy with which viewers perform simple perceptual tasks Cleveland & McGill (1986). People perform significantly worse on stacked bar charts than on simple bar charts, and bar charts with filled bars are generally superior to those with empty ones.

"The cognitive style of PowerPoint: Pitching out corrupts within" examined the effect of static graphics, such as bar charts, line charts, and pie charts, on the comprehension and decision-making of viewers, Edward R. Tufte (2008). The results indicated that the bar chart was most effective at accurately communicating information, while the pie chart was least effective. In addition, the researchers discovered that the type of graphic employed influenced the viewers' perceptions and decision-making, highlighting the significance of selecting the most appropriate graphics when conveying information.

"Using color to code different types of uncertainty in visualization," investigated the effect of color on the communication effectiveness of static graphics (**ware2010?**). Researchers discovered that the use of color in static graphics enhanced viewers' comprehension and memory retention of the presented information. Specifically, it was determined that the use of color to highlight particular data points or trends in a graph or chart was particularly effective.

"Communicating Uncertainty: A Look at Graphical Representations of Error Bars, Box Plots, and Confidence Intervals" examined the efficacy of static graphics in communicating data uncertainty and variability (**wainer2020?**). Researchers discovered that static graphics, such as boxplots and error bars, effectively communicated the range and variability of data. Furthermore, the use of static graphics enhanced viewers' comprehension and decision-making in uncertain situations.

Overall, these above studies indicate that static graphics effectively convey information to viewers. Nonetheless, their effectiveness is influenced by the type of graphic exercised, the use of color, and their capacity to express uncertainty and variability in data. It is essential to note that the effectiveness of static graphics can also be affected by audience knowledge and experience, the complexity of the presented information, and the context in which the graphics are used.

1.3.6 Testing interactive graphics

The significance of testing and evaluating static graphics to ensure their effectiveness and user-friendliness naturally segues into the similar, yet distinct challenge of evaluating interactive graphics

The design space of interactive visualizations, including the use of multiple views and coordinated various perspectives, is discussed in a paper by Heer and Bostock Heer & Shneiderman (2012). The authors argue that interactive visualizations can enhance comprehension and decision-making by enabling users to investigate complex data and the relationships between multiple variables. It was discovered that interactive graphics are beneficial for investigating data that is difficult to comprehend with static graphics, such as time series data or data with multiple aggregation levels.

Compared to static graphics, interactive graphics can improve users' ability to identify patterns and outliers in data, according to a second study, (**wu2017?**). This is because interactive graphics enable users to zero in on particular aspects of the data and adjust the visualization in real-time to reveal hidden patterns.

It has also been discovered that interactive graphics increase user engagement with data and improve their decision-making abilities. Moreover, interactive graphics can help users identify potential biases or errors in their data analysis by allowing them to rapidly explore various visualizations and test their hypotheses.

In addition to their benefits for data exploration and analysis, interactive graphics have been found to be more engaging and enjoyable for users Eppler & Bresciani (2013). This can increase the motivation of users to explore and analyze data, resulting in better decision-making.

In summary, the studies highlighted indicate that interactive graphics can be an effective data visualization and analysis tool, especially for complex and multidimensional data. They can improve users' comprehension, engagement, and decision-making skills, and are ideal for examining data that is difficult to comprehend with static graphics.

1.4 Dashboard Design

Given that the intended audience has limitations, there are design constraints around the data, and the ability of the audience to successfully use the graphical displays of the data, what can we take from this body of research that applies to more complicated sets of graphics? How do we maintain user attention, desire to explore, and accurately communicate the data through the medium of an interactive data dashboard? Solutions to these questions can start with a dashboard.

A dashboard is a visual display of the essential information needed to achieve one or more objectives, consolidated and arranged on a single screen so the data can be monitored at a glance (Few, 2006a). Dashboards have particular characteristics:

- Achieve specific objectives

- Fits on a single computer screen
- Information can be displayed in multiple mediums (web browser or mobile device)
- Can be used to monitor information at a high level

Dashboards can present various statistical data, such as financial performance, website traffic, or customer engagement metrics. They allow users to quickly and easily understand complex data sets using visual elements such as charts, graphs, and tables to display the information. Additionally, statistics can be used to analyze data presented on a dashboard, providing insights into trends and patterns that can inform decision-making.

While a dashboard can be handy, it may be worth describing that a poorly designed dashboard will not be used. A dashboard should be concise, clear, and intuitive when displaying components in combination with a customized list of requirements of users.

Much of the work done within statistical research and dashboard design involves collaboration with other researchers and users. While this may be the best for the growth of the discipline, one will find that working with collaborators with non-STEM backgrounds. Dashboards can help understand and support many data types in essential business objectives. There are many different ways to label and utilize dashboards in different kinds.

Dashboards are cognitive tools that should be used to improve understanding of data, which should help people visually find relationships, trends, patterns, and outliers. Most importantly, dashboards should leverage people's visual cognitive capabilities.

Cowan suggested that the average person can only hold two to six pieces of information in their attention, Cowan (2001). People can develop detailed understandings of reality, which is infinitely complex.

It is important to consider how users construct mental models of the presented data and information. Researchers have introduced cognitive structures consisting of mental models and their relationships (Rumelhart & Ortony (1976); Carley & Palmquist (1992); (**jonassen1996?**)). In cognitive psychology, a schema is a mental model representing a person's general knowledge or expectations regarding a particular concept or situation. Schemas are organized into semantic networks based on their relationships to other schemas (Wertheimer (1938); Rumelhart & Ortony (1976)). This arrangement helps the brain process its experiences instead of storing every sensory observation; the brain only needs to maintain its schemas, which are good summaries of all previous observations. Some "memories" may even be complete recreations built with a schema (Bartlett & Remembering (1932); Klein, Phillips, Rall, & Peluso (2007))

Wixon introduced the concept of "contextual design" as a systems development method in which the researcher partners with the user at the user's place of work to "develop a shared understanding" of the user's activities, and they define contextual inquiry as the first part of the broader process, Cowan (2001). Contextual inquiry is the data collection step of the field research element of the contextual design method, and it emphasizes four essential principles:

1. The context of the activity being performed by the user

2. The partnership between the researcher and the participant
3. The spoken verification that the investigator's interpretation of the activity matches the user's
4. The focus of the study is central to the approach taken by the interviewer

Kandal conducted what might be considered a contextual interview study in that they analyzed data scientists' self-reported work processes, Kandel, Paepcke, Hellerstein, & Heer (2012). They proposes three main archetypes that data scientists may be classed into the following:

- Hackers: who build processes chaining together multiple programming languages of different types (analytical, scripting, and database languages) and use visualization in various environments.
- Scripters perform most of their analysis in an analytical environment (e.g., R or Python) and execute the most complex statistical modeling of the types but do not perform statements in their Extract, Transform, and Load (ETL).
- Application Users: who performed most or all of their work in an application such as Excel or SPSS and, like scripters, relied on others (namely, their organizations' IT departments) for ETL.

Combining two compelling graphics does not necessarily result in a successful visualization. In certain instances, suboptimal combinations can result in confusion, misinterpretation, and the failure to convey the intended message. Combining two charts with distinct scales or units is an example of suboptimal graphic design, which can result in misinterpretation and flawed comparisons. For example, if a bar chart displaying the number of sales is combined with a line chart showing revenue, meaningful comparisons between the two metrics can be challenging. According to a study conducted by Cleveland and McGill, people frequently make inaccurate judgments when comparing graphs with different scales, Cleveland & McGill (1984).

In addition, combining two difficult-to-compare graphics with redundant visual cues or unnecessary embellishments such as colors, 3D effects, or patterns can increase cognitive load and reduce the dashboard's effectiveness. Although adding extra elements to a chart or graph may be tempting, doing so can detract from the primary message and make it more difficult for the audience to focus on the essential information. Tufte discovered that adding unnecessary visual cues to a graph decreases its effectiveness because viewers are more likely to focus on the embellishments rather than the data, Edward R. Tufte (1985).

For instance, if a scatter plot and a bar chart are combined, the resulting visualization may be difficult to interpret due to the two graphics types' incompatibility. This issue was highlighted by Hullman et al., who discovered that viewers had difficulty interpreting a visualization that incorporated a scatter plot and a line chart, Hullman, Adar, & Shah (2011).

1.5 Dissertation Map

This dissertation will be constructed as follows.

Chapter 1 thoroughly reviews the literature regarding graphical and human-computer interaction/UI-UX methods. Chapter 2 will explore the process of designing dashboards for public use through parallel coordinate plots as a central component of data exploration to make decisions. Chapter 3 focuses on graphical methods for multidimensional categorical variables and visualization methods have for growth. We conclude with a Shiny application that facilitates a better understanding of the possible forms a parallel coordinate plots in exploratory data analysis can take by accommodating a through examination through variables and structural changes to the parallel coordinate plot with a click of a mouse. Chapter 4 further explores multidimensional categorical data visualizations and develops an approach to using parallel coordinate plots to assess predictive model. We identify visual indicators for parameters in different models and extend the connection between parallel coordinate plots of binary tables and odds ratios to include logistic regression models with categorical variables.

1.6 Conclusion

The lack of integration between statistical interactive graphics and Gestalt principles is a research gap in this field. Gestalt principles provide a framework for understanding how visual elements are perceived and organized in the human brain, whereas statistical interactive graphics enable the manipulation and exploration of complex data sets. This integration can result in more effective and efficient visualization tools that take human perception and cognition into account.

In addition, there is a lack of research on the application of dashboards in visual communication. Dashboards are becoming increasingly popular to present data and information in a concise and easily accessible format. Still, there is a shortage of research on designing and employing dashboards to communicate complex data effectively.

This research seeks to bridge the gap between statistical interactive graphics, Gestalt principles, and dashboards so that users can more effectively explore complex data. This can be accomplished by designing interactive visualizations that adhere to Gestalt principles, enabling users to recognize patterns and trends in the data quickly and easily. Incorporating statistical interactive graphics into dashboards can also provide users with a more comprehensive view of the data, enabling them to make more effective data-driven decisions.

Chapter 2

Chapter Paper on Rural Shrink Smart Manuscript submitted to Journal of Data Science Special Issue

2.1 Abstract

Many small and rural places are shrinking. Interactive dashboards are the most common use cases for data visualization and context for exploratory data tools. In our paper, we will explore the specific scope of how dashboards are used in small and rural area to empower novice analysts to make data-driven decisions. Our framework will suggest a number of research directions to better support small and rural places from shrinking using an interactive dashboard design, implementation and use for the every day analyst.

2.2 Introduction

As the amount of data has increased in nearly every facet of life, the need to make sense of that data in an approachable, accessible form has become ever more important. As a result, many companies and organizations use interactive dashboards to present these data in a more useful and visually appealing form (Sarikaya, Correll, Bartram, Tory, & Fisher, 2019).

In many cases, dashboards support viewers' information processing, helping to make sense of complex data, navigate through a dataset, and supporting decision making based on the data.

Dashboards are often used, as with the car display of the same name, to provide summary information about many separate attributes of a common entity. One glance at a car's dashboard will tell you the speed, RPM, engine temperature, amount of gas in the tank; more importantly, however, the goal is not for the user to remember all of these characteristics, but to assess whether any of these quantities is outside of the expected range. Similarly, interactive dashboards for data are often used to display many different attributes and performance metrics which are of importance for stakeholders.

In this paper, we discuss the process of designing a dashboard to present publicly available government data to stakeholders in small Iowa towns to facilitate decision making and objective comparison with other similarly-situated towns.

Some communities continue to thrive as they lose population because they adapt, maintaining quality of life and community services for residents while investing in the future. This process, *smart shrinkage*, is important for rural areas who have experienced shrinking populations for decades. As small rural towns do not have access to data scientists or even the ability to easily leverage data collected locally to support decisions, our research team will provide communities with data about services in small town Iowa in order to assist with developing strategies to improve quality of life for their residents amid shrinking populations (Rural Shrink Smart Team, 2022). We hope to allow towns to explore their own data and compare to other similar towns, centering decision-making on data in the context of small-town Iowa life.

2.3 Data Description

The Smart and Connected Community (SCC) dashboard data are primarily assembled from data.iowa.gov (State of Iowa, 2020), with some additional datasets assembled from federal and private sources. Most of these data sets are collected at a town/city or county spatial resolution, requiring us to carefully join data to ensure that these differences are respected while collating relevant information at the city level. In addition to the more commonly available statistics derived from e.g. the census and American Community Survey, data.iowa.gov contains several unique data sets, including local liquor sales, school building locations, town budgets and expenditures, hospital beds, Medicaid reimbursements, and other details that may provide information about local quality of life.

Data available on Iowa's data portal were augmented in some cases with higher-quality data sets in cases where the Iowa data were out of date or insufficiently accurate. Data collected from ELSI (National Center for Education Statistics, 2020) from https://nces.ed.gov were used to show the distance to any private or public school. The National Center for Education Statistics (NCES) is the primary federal entity for collecting and analyzing data related to education (Zarecor, Peters, & Hamideh, 2021).

Data collected from the Index of Relative Rurality (IRR) (USDA - ERS, 2020a) were used in the SCC dashboard to help classify the towns. The Index of Relative Rurality (IRR) is a continuous, threshold-free, and unit-free measure of rurality. It is an alternative to the traditional discrete threshold-based classifications.The IRR ranges between 0 (low level of rurality, i.e., urban) and 1 (most rural). Four steps are involved in its design:

1. Identifying the dimensions of rurality: population size, density, remoteness, and built-up area.

2. Selecting measurable variables to adequately represent each dimension:

    - Size: logarithm of population size
    - Density: logarithm of population density.
    - Remoteness: network distance.
    - Built-up area: urban area (as defined by the US Census Bureau) as a percentage of total land area.

3. Re-scaling the variables onto bounded scales that range from 0 to 1.

4. Selecting a link function: unweighted average of the four re-scaled variable.

Data collected from Rural Urban Commuting Area Codes (USDA - ERS, 2020b) were used to help identify towns with commuting behaviors in our rural areas. The rural-urban commuting area (RUCA) codes classify U.S. census tracts using measures of population density, urbanization, and daily commuting. This data is on a zip code-level that will help identify those communities that commute to more urban areas. The most recent RUCA codes are based on data from the 2010 decennial census and the 2006-10 American Community Survey. The classification contains two levels. Whole numbers (1-10) delineate metropolitan, micropolitan, small town, and rural commuting areas based on the size and direction of the primary (largest) commuting flows.

One of the interesting features of this assembled data set is that missing data can be missing for multiple reasons: not all state data is complete, but data about certain services may also be missing because towns do not offer that service. Thus, in addition to the usual challenges of working with real-world data that is “messy” in a variety of ways, we also have to contend with missing data that is missing due to the size of the community or the lack of services. This makes both visualization and statistical analysis more complicated (and more interesting).

2.4 Dashboard Design Considerations

One problem we identified early in the process of assessing smart-shrinkage strategies in small towns is that these towns do not have the resources to make data-driven decisions. Typically, small towns in Iowa are managed by at most a few part-time employees or volunteers. In some cases, essential management functions of the town are paid, but the municipalities we are interested in do not have sufficient funding to hire professionals to gather and analyze data.

As part of a wider project investigating the strategies towns use to maintain quality of life amid shrinking population, our research team provides communities with data about their own town, but also comparable towns across the state which may have a different approach to city services. In combination with other engagement strategies that are more qualitative, we hope to use this interactive dashboard approach to assist small Iowa cities with generalizing and developing strategies to improve or maintain quality of life amid shrinking populations.

One factor at the forefront of our visualization design is the importance of

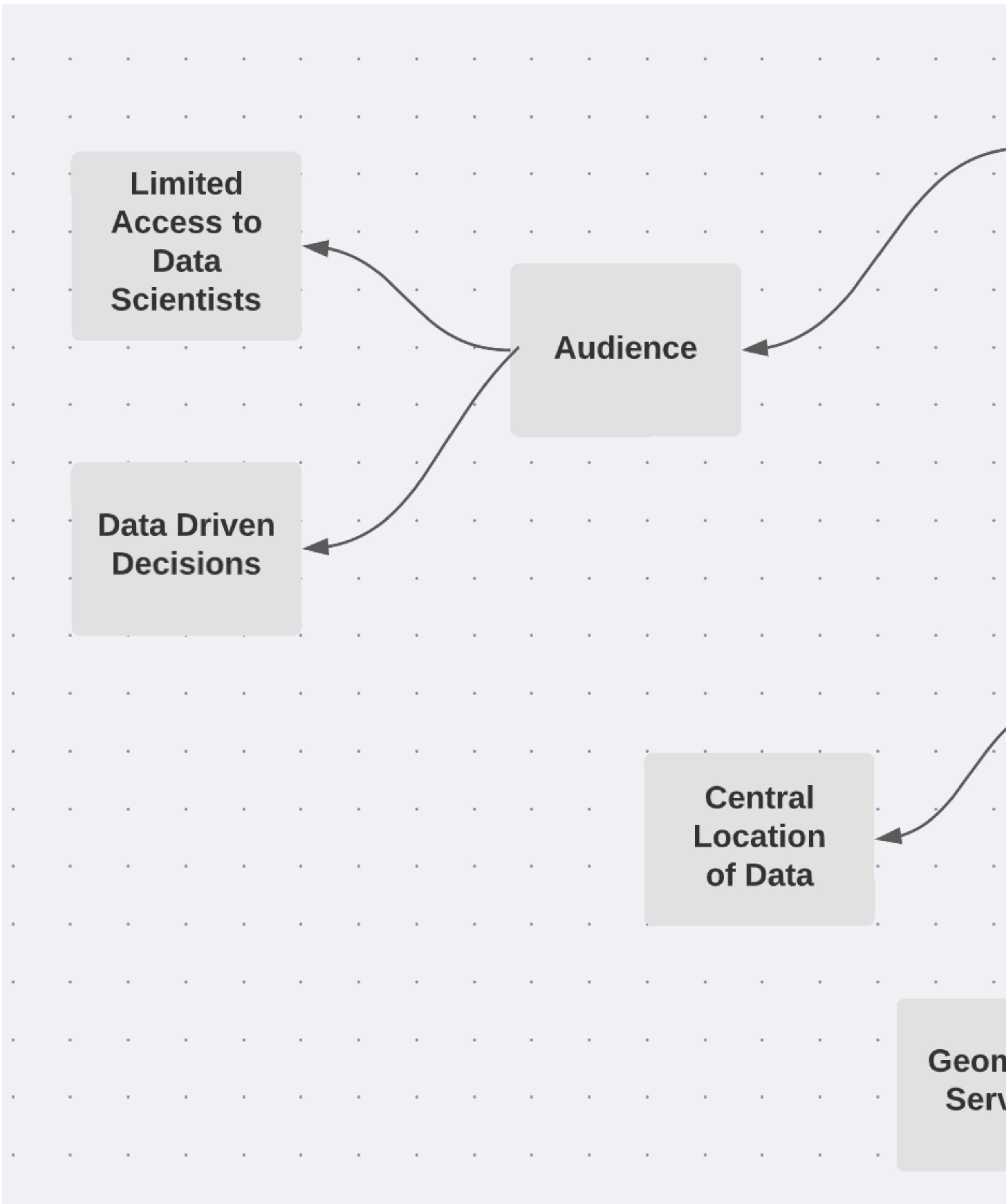


Figure 2.1: Diagram of considerations for our dashboard design process.

reducing the cognitive demands on viewers: we have assembled an incredible amount of data, and it is easy for even statisticians who deal with much larger datasets to get lost in the details of this data. At the same time, we want to invite viewers to engage with the data - to imagine, to draw comparisons, to generalize across towns, and to integrate outside information into the conclusions drawn based on the data we present. This invitation to engage with the data is similar to the approach advocated in Guided Discovery Learning, a framework leverages hints, feedback, and other helpful information to guide users in interactive exploration (DeDonno, 2016).

We expect that users will be interested in “sets” of variables from the wider dataset, which we assembled based on quality of life factors in the Iowa Small Town Poll (Peters, 2019). For instance, users might be interested in medical and social services available to residents, such as a local primary care clinic, nursing homes which are within driving distance, and the distance to the nearest emergency room; these factors might be explored separately from variables describing the services provided directly by city government, such as parks and recreation expenditures, snow removal services, and the distance to the closest fire station.

As a consequence of this massively multivariate structure, we very quickly focused on the use of parallel coordinate plots; other alternatives, such as tours (Wickham, Cook, Hofmann, & Buja, 2011), require much more sustained attention to interactive plots as well as a deeper understanding of projections in multidimensional space which we cannot assume our users will have. Introduced in the 1880s (d’Ocagne, 1885), parallel coordinate or parallel set plots feature a series of vertical axes representing different variables arranged horizontally, with lines connecting each observation. When representing categorical data, parallel set plots may show "blocks" of data instead of individual lines, and are useful for representing conditional relationships between adjacent variables (Bendix, Kosara, & Hauser, 2005); modifications of this design, such as common-angle plots (Hofmann & Vendettuoli, 2013), address the issues which arise due to line-width illusions VanderPlas & Hofmann (2015a). Parallel coordinate plots have been generalized to allow for continuous data and additional summaries beyond individual data points, such as densities (Heinrich & Weiskopf, 2009). In this paper, we use the `ggpcp` package, which leverages the grammar-of-graphics framework introduced in Wickham (2016), allowing us to use not only parallel coordinate plots, but also to overlay other statistical summaries, such as boxplots or violin plots, to provide additional context about the marginal distributions of each variable in addition to allowing for exploration of the multivariate space.

We also anticipate that users will be interested in comparing their town to other, similar towns. We will discuss the different ways that this comparison strategy was implemented in each dashboard in the next section, which describes the evolution of the dashboard over time and accounting for feedback from users and other researchers on the wider project.

One final component of this project is that our dashboard is part of a wider effort to work with towns to understand the different strategies used to maintain resident quality of life amid shrinking populations. Thus, while the town leaders are our primary audience, we also are creating this applet for use in parallel with a team of other researchers: sociologists, economists, city planning specialists, and artists. These researchers opinions and feedback about the dashboard are also useful and important, as they regularly work with town leaders in different capacities and have an understanding of what factors are most important to them and what types of questions these leaders may have when faced with data and unfamiliar statistical visualizations.

Throughout the design process, we will assess our visualizations to determine which strategies for user interface and interactive graphics design are most useful to empower town leaders to make discoveries in publicly available data assembled with a focus on items that impact rural quality of life.

2.5 Guiding Design Principles

Research on dashboard creation and interactive visualization tends to be very task-specific and hard to apply to more generalized settings. That is, it is relatively easy to create a dashboard that works for a particular task, but it is hard to generalize from that process what will work for the next dashboard. With this in mind, we set out to clearly document our intentions at each stage of the design and evaluation process, with the goal of gathering some useful information about general dashboard design from the process of creating this specific dashboard.

Thus, our initial set of dashboard design principles is as follows:

- The town leaders are the focus audience; thus, the town itself should be the central focus of the app.
- We should facilitate comparisons with other towns in order to allow the user to explore other potential solutions to offering services that enhance resident quality of life.

- We will present the user with peer comparisons in order to widen the scope of exploration beyond the initial set of obvious peers in the local region.
- We will implement feedback mechanisms that allow us to provide more detailed data and respond to feature requests to improve the dashboard design over time.

As with many dashboards, this project is under continuous development; while it makes for an unsatisfactory conclusion, we do not have a "final" dashboard design because the application will continue to evolve. However, we have some useful insights into the process of creating an application designed to invite users to explore a large and complex dataset that we believe to be a useful contribution to work in this area.

2.6 Dashboard Design Process

2.6.1 Dashboard Components

In this section, we discuss the philosophy behind the basic "building blocks" of the dashboard. This philosophy is present in all of the iterations of the dashboard that we present in this discussion, and we will evaluate the overall philosophy's effectiveness in the conclusion.

The large set of publicly available data (primarily from `data.iowa.gov`) we have assembled is useful, but we must be careful with how we present this data because it would be easy to overwhelm the user with small details that mask the bigger picture. We select a small subset of towns (out of the 999 towns in Iowa) and a small subset of variables of interest to start with, and then allow the user to increase the complexity of the display in accordance with their interest. This avoids some of the pitfalls of dashboard design that can easily lead to user overload (Few, 2006b).

Our primary objective is to provide users with a town-centric approach: their town is at the center of our application, and comparisons to other, similar towns are secondary. As a result, the next component of the dashboard is intended to provide a brief overview of the information we have about a specific town of interest. This design is based on research into visualization sensemaking (Lee et al., 2016), in that we allow users to explore outward from the familiar to the unknown. The map visuals were built using Open Source Routing Machine (OSRM) route functions (Luxen & Vetter, 2011) in R (R Core Team, 2022) to amplify the accuracy of the distances from necessary services in town-centric point. OSRM allows for finding the "As the Crow Flies" distance and time on the road for our vital services map, since OSRM technology is similar to Google maps.

When faced with the next component, a parallel coordinate plot (PCP), a novice user will be able to determine two basic components: Visual Object (textual objects and non-textual objects) and Frame (frame of content and frame of visual encoding).

Taken together, the app is a single page; the initial "solid ground" which the user explores from consists of maps showing the route from the center of town to necessary services, including the fire department, schools, post offices, and hospitals. In version 2, as shown in Figure 2.3, the map portion is condensed, and more space is given to value boxes that show vital statistics about the town's QoL and financial metrics. This relatively straightforward display is followed by a parallel coordinate plot that allows the user to see similar towns along dimensions such as economic indicators or population size.

2.6.2 Initial Draft

The initial design sketch and implementation are shown in Figure 2.2.

Users' towns are at the center of our application, and comparisons to other, similar towns are secondary. As it can be extremely difficult to predict which towns are optimal for comparison purposes (similar may involve population, region, economic indicators, sports rivalries, and any number of other variables), we allow users to modify a set of suggested comparison towns to indicate other towns of interest.

We implemented some suggested town comparisons using unsupervised clustering methods to help our towns make decisions that are informed in comparison to similar towns, for budget size, population size and location. We initially focused on determining the next five to ten similar towns, based on distances to services. This feature became an important diagnostic for our data quality, as it became clear that towns which were grouped with big cities but which did not have a large population were so grouped because of missing data. Unfortunately, this clustering feature was not as useful to the application users, as they came to the dashboard with a pre-existing set of towns to compare to; our suggested comparisons were in the way.

The initial dashboard design featured several responsive maps showing the distance to the nearest hospital, fire department, post office, and school. These maps were ineffective for several reasons:

- Town residents already know this information (though it was useful for

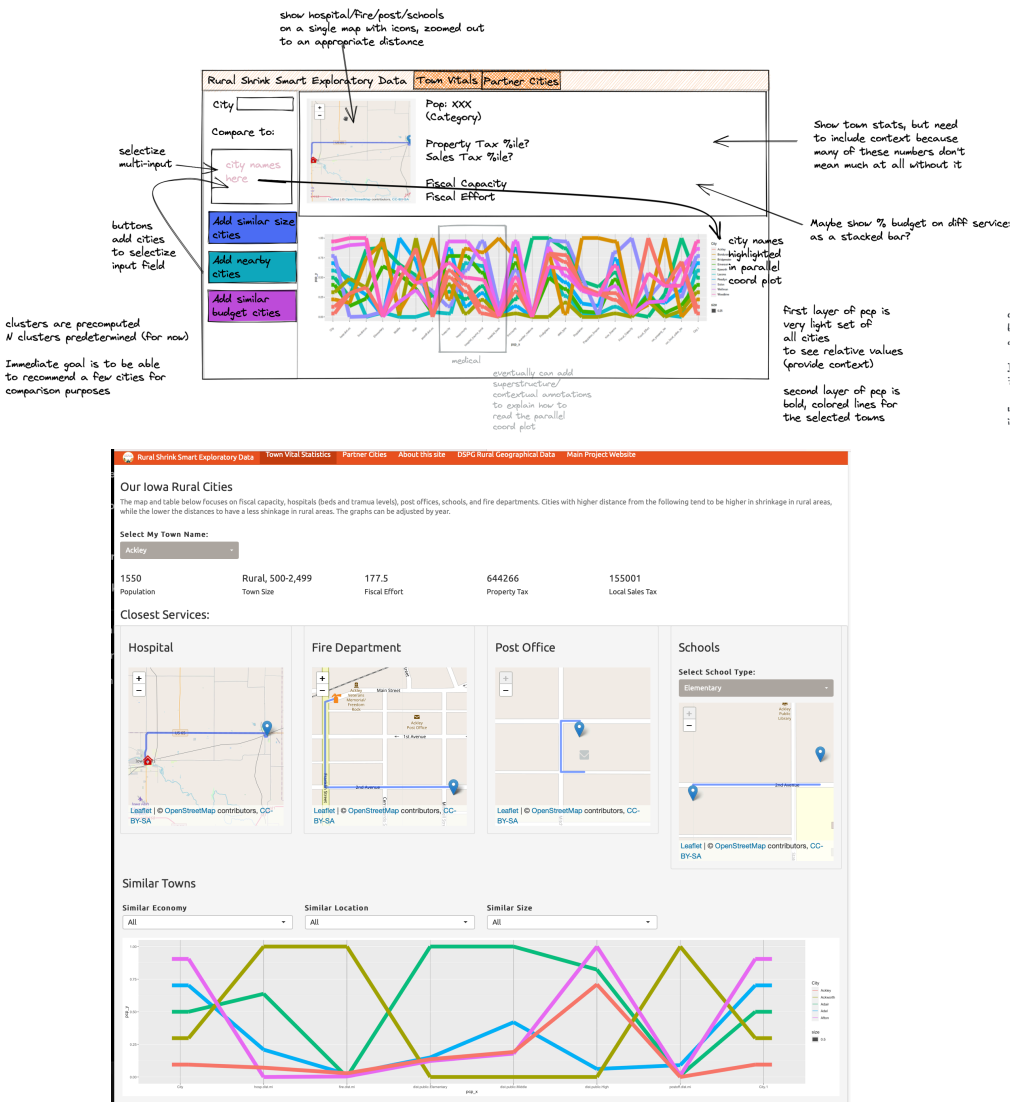


Figure 2.2: Initial dashboard design sketch (top) and implementation (bottom).

us as the dashboard designers, because we aren't nearly as familiar with the 900+ small towns in Iowa)

- We computed distance from services relative to the center of town - coordinates provided in the data from data.iowa.gov. Generally speaking, the post office is at the center of town and the fire department is usually very close to the center of town; these two maps were useless. The school and hospital maps were less useless, but still did not provide particularly useful information to people already familiar with the town.
- It became clear that it might be more useful to show the comparison towns on a map (relative to the town of interest) so that users could compare geographical ratings for unfamiliar data to familiar data.

In addition, we received feedback on the parallel coordinate plot at the bottom of the app which was surprising: the viewers (in this case, other researchers on the team) were not as intimidated by the parallel coordinate plot as we had expected. They did need some explanation of how to read the plot, and these hints need to be included in the dashboard, but they grasped the fundamental idea of the plot very quickly.

Our conclusion, based on this initial dashboard draft, was that we needed to restructure the application. Our attempt to show familiar information first to "build up" to the more unfamiliar structure of a parallel coordinate plot was not effective; there was too much clutter and not enough new information to draw users in.

Figure 2.3: A second iteration of the sketched design (top) and the implementation (bottom).

2.6.3 Redesign

In the initial design, we included a map for each vital service, this initially created a lag for the users' experience. As a result, we cached map directions from OSRM for each service in our database, which drastically reduced the response time for the user. Our initial design did not naturally focus the user's eye on the most important parts of the dashboard; the redesign allowed for a cleaner flow from the top to the bottom.

In addition to the timing due to the map loading slowly, we added the vital statistics at the county level to allow for a more robust understanding of the town and it's surroundings. The rurality index provided a better classification and the USDA sources allowed for the town to understand the impact of the closest major city due to commuting for work and shopping at larger stores not available within the town.

We also modified the parallel coordinate plots in several ways:

- Our x-axis had a large number of variables that we as researchers believed to be the most strongly associated with quality of life. However, there were still too many variables for users to successfully parse. We reduced the number of variables, focusing on variables that had the highest data quality, and we grouped these variables by quality of life factor (Peters, 2019).
- Originally, parallel coordinate bands were scaled based on the selected comparison towns. This had the effect of truncating the range of variables and over-emphasizing differences between selected towns relative to the overall range of each variable over all towns in our data set. We

chose to show all towns in the data set in a very light α grey color to provide some information about the overall range of each variable. Unfortunately, even with the low-α value, this increased the visual complexity of the plot and confused users. Future iterations will likely make use of another aesthetic, such as boxplots or violin plots, to show the range of values for all towns, and then use lines only for towns that are selected by the user. This should strike a balance between visual complexity and representing the data accurately.

- We noticed that users did not make use of our suggested comparison towns, and so we removed that option in favor of allowing users to enter their own comparison towns directly. Users already had pre-determined towns they wanted to compare to, and our suggestions were just in the way.

While not all of these modifications were well received in our second round of user testing, the changes did incrementally move the dashboard display towards our goal of allowing users to explore the data and engage with it. We continued to be surprised with how well users reacted to the parallel coordinate plots, which we initially thought might be too abstract for users unfamiliar with multivariate data displays, but the ability to compare towns across multiple dimensions and examine the similarities and differences between their approaches to different services seemed to be intuitive for users once they understood that each vertical axis was a different variable.

2.7 Discussion

Our dashboard design philosophy worked primarily to promote a town-centric approach application with comparisons to other similar towns being secondary. This approach created a way for the user to see their town information at the top of the page and to explore the PCP after reviewing their own town's essential statistics. The PCP in the lower part of the dashboard allowed for the user to see the plot and adjust to the fact that they could add more towns to the plot, providing an opportunity to explore the wider dataset from a base of familiar knowledge.

While we initially framed the design around guided discovery learning, the approach did not seem to suffice for our user base; instead, we found that users were more drawn to the unfamiliar from the start. We will likely leverage this in future iterations by using visual forms such as flower plots to draw the users in; even though these plots are not ideal for numerical display of data, the visual novelty and aesthetic appeal will provide some motivation to continue exploring and thinking about the data.

One factor that we have briefly considered and have seen hints of in our user feedback is that towns may not want to be compared negatively with other towns. While users have very definite ideas about which towns they would like to compare to, we can always mask the town names and move back to comparisons based on town size and other factors (for instance, whether or not a town is the county seat is a factor that is important outside of population). Using this approach, we would label each town as “Town 1”, “Town 2”, and so on, which would eliminate some of the fears about negative comparisons, but would also remove some of the novelty of the data dashboard for our users and would prevent users from drawing on their own outside knowledge about each of the comparison towns.

We also recognize that we need to leverage the expertise of others in our research team: we are working with artists, researchers in architecture, economists, and sociologists; these researchers provide outside knowledge that we do not have and may be able to help us create insightful use-cases to showcase the app and teach towns how to use it. We can also leverage the app to connect users with our research team, providing additional value to those who use the applet and facilitating development of strategies for maintaining quality of life amid shrinking populations.

2.8 Future Work

One avenue we will explore in future iterations of the dashboard is to incorporate other dashboards generated by different groups within this project. This will create a wider field of information to explore: for instance, some of the additional work will focus on the 99 towns featured in the Iowa Small Town Poll; this will allow us to showcase survey-based measures of quality of life alongside the more objective measurements assembled in the dataset discussed in this paper. While at least one tab of this omni-dashboard will still focus on wider EDA and discovery, we hope to incorporate other information as well to provide a more well-rounded data display encompassing most of the facets of this complex project.

We are also mindful of a distinction between "eye candy" and purpose-driven data visualization. While we have typically focused on the latter, there is certainly a place in our dashboard for the former as well. "Eye candy" visualization is intended to draw the viewer in and motivate them to explore; while these visualizations may not be particularly effective at communicating quantitative information, if they motivate the user to engage with the rest of the dashboard, they still serve a purpose. It is with this mindset that we intend to explore the use of flower plots - the artistic opportunities combined with the display of quantitative information (even in a form that isn't optimal for quantitative comparisons) may be useful to engage viewers before transitioning to more useful data visualizations intended to provide accurate quantitative comparisons.

EDA can be a difficult for a variety of groups of people, novice users and experienced researchers. One of the more difficult components of this project has been clearly articulating the purposes of EDA to a diverse group of researchers unfamiliar with the concept. One of the most useful parts of this dashboard iteration process has been as an aid to data discovery: that is, the dashboard motivated us to find additional data sources and incorporate them into the project. Having conversations with other researchers about the EDA process helped to facilitate these conversations, as each discussion seemed to uncover additional data sources that someone remembered after looking at the dashboard. While this facet of the dashboard process may be difficult to study formally, it would be an interesting avenue for investigation.

2.9 Conclusions

In this paper, we have documented the process of designing a dashboard for exploration and visualization of a large and complex data set assembled from many different sources. Our primary audience was leaders of small towns in Iowa, with a secondary audience of researchers in fields other than statistics collaborating on this project with us. Through the process of revising our dashboard, we found that the idea of guided discovery learning as implemented in our first version did not work as well as we had anticipated. It was more important to focus on allowing users to explore their questions about the dataset by facilitating user-driven comparisons and exploration, rather than attempting to anticipate user desires by providing comparison towns. In addition, we found that it would be more effective to draw users in with novel visual displays, as these seemed to attract more interest than providing known facts and an opportunity to explore outwards from an initial area of familiarity.

While it is hard to apply the findings from one fairly specific visualization project more widely, there is a lack of resources in this area that provide both design philosophies and actual analysis of user feedback in a qualitative sense. We have attempted to address this dearth of information by providing the design strategies, user feedback, and our planned and executed modifications, in the hopes that others facing the daunting challenge of designing a dashboard for EDA may learn something from our experiences.

Chapter 3

Chapter 2 Stuff

3.1 Many useful visualizations don't have easy interactivity

Chapter 4

Tables, Graphics, References, and Labels

4.1 Dashboard EDA

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

```
library(knitr)
library(palmerpenguins)
library(tidyverse)
library(nycflights13)
data(flights)

library(ggpcp)
library(ggplot2)
library(dplyr)
data(nasa)

library(scales)
library(datasets)
data("ChickWeight")
library(formatR)
```

In Chapter 4:

Appendix B

The Second Appendix, for Fun

Colophon

This document is set in EB Garamond, Source Code Pro and Lato. The body text is set at 11pt with *lmr*.

It was written in R Markdown and LaTeX, and rendered into PDF using huskydown and bookdown.

This document was typeset using the XeTeX typesetting system, and the University of Washington Thesis class class created by Jim Fox. Under the hood, the University of Washington Thesis LaTeX template is used to ensure that documents conform precisely to submission standards. Other elements of the document formatting source code have been taken from the Latex, Knitr, and RMarkdown templates for UC Berkeley's graduate thesis, and Dissertate: a LaTeX dissertation template to support the production and typesetting of a PhD dissertation at Harvard, Princeton, and NYU

The source files for this thesis, along with all the data files, have been organised into an R package, xxx, which is available at `https://github.com/xxx/xxx`. A hard copy of the thesis can be found in the University of Washington library.

This version of the thesis was generated on 2023-06-09 19:28:51. The repository is currently at this commit:

The computational environment that was used to generate this version is as follows:

```
## - Session info -------------------------------------------
##  setting  value
##  version  R version 4.2.2 (2022-10-31)
##  os       macOS Big Sur ... 10.16
##  system   x86_64, darwin17.0
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       America/New_York
##  date     2023-06-09
##  pandoc   2.19.2 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/ (
##
## - Packages -----------------------------------------------
##  package       * version date (UTC) lib source
##  assertthat      0.2.1   2019-03-21 [1] CRAN (R 4.2.0)
##  bookdown        0.33    2023-03-06 [1] CRAN (R 4.2.0)
##  cachem          1.0.7   2023-02-24 [1] CRAN (R 4.2.0)
##  callr           3.7.3   2022-11-02 [1] CRAN (R 4.2.0)
##  cli             3.6.1   2023-03-23 [1] CRAN (R 4.2.0)
##  colorspace      2.1-0   2023-01-23 [1] CRAN (R 4.2.0)
##  crayon          1.5.2   2022-09-29 [1] CRAN (R 4.2.0)
##  devtools        2.4.5   2022-10-11 [1] CRAN (R 4.2.0)
##  digest          0.6.31  2022-12-11 [1] CRAN (R 4.2.0)
##  dplyr         * 1.1.2   2023-04-20 [1] CRAN (R 4.2.0)
##  ellipsis        0.3.2   2021-04-29 [1] CRAN (R 4.2.0)
##  evaluate        0.21    2023-05-05 [1] CRAN (R 4.2.0)
##  fansi           1.0.4   2023-01-22 [1] CRAN (R 4.2.0)
```

```
## farver          2.1.1   2022-07-06 [1] CRAN (R 4.2.0)
## fastmap         1.1.1   2023-02-24 [1] CRAN (R 4.2.0)
## forcats       * 1.0.0   2023-01-29 [1] CRAN (R 4.2.0)
## formatR       * 1.14    2023-01-17 [1] CRAN (R 4.2.0)
## fs              1.6.2   2023-04-25 [1] CRAN (R 4.2.0)
## generics        0.1.3   2022-07-05 [1] CRAN (R 4.2.0)
## ggpcp         * 0.2.0   2022-11-28 [1] CRAN (R 4.2.0)
## ggplot2       * 3.4.2   2023-04-03 [1] CRAN (R 4.2.0)
## glue            1.6.2   2022-02-24 [1] CRAN (R 4.2.0)
## gtable          0.3.3   2023-03-21 [1] CRAN (R 4.2.0)
## hms             1.1.3   2023-03-21 [1] CRAN (R 4.2.0)
## htmltools       0.5.4   2022-12-07 [1] CRAN (R 4.2.0)
## htmlwidgets     1.6.2   2023-03-17 [1] CRAN (R 4.2.0)
## httpuv          1.6.9   2023-02-14 [1] CRAN (R 4.2.0)
## knitr         * 1.42    2023-01-25 [1] CRAN (R 4.2.0)
## labeling        0.4.2   2020-10-20 [1] CRAN (R 4.2.0)
## later           1.3.0   2021-08-18 [1] CRAN (R 4.2.0)
## lifecycle       1.0.3   2022-10-07 [1] CRAN (R 4.2.0)
## lubridate     * 1.9.2   2023-02-10 [1] CRAN (R 4.2.0)
## magrittr        2.0.3   2022-03-30 [1] CRAN (R 4.2.0)
## memoise         2.0.1   2021-11-26 [1] CRAN (R 4.2.0)
## mime            0.12    2021-09-28 [1] CRAN (R 4.2.0)
## miniUI          0.1.1.1 2018-05-18 [1] CRAN (R 4.2.0)
## munsell         0.5.0   2018-06-12 [1] CRAN (R 4.2.0)
## nycflights13  * 1.0.2   2021-04-12 [1] CRAN (R 4.2.0)
## palmerpenguins * 0.1.1  2022-08-15 [1] CRAN (R 4.2.0)
## pillar          1.9.0   2023-03-22 [1] CRAN (R 4.2.2)
## pkgbuild        1.4.0   2022-11-27 [1] CRAN (R 4.2.0)
## pkgconfig       2.0.3   2019-09-22 [1] CRAN (R 4.2.0)
## pkgload         1.3.2   2022-11-16 [1] CRAN (R 4.2.0)
## prettyunits     1.1.1   2020-01-24 [1] CRAN (R 4.2.0)
```

```
## processx       3.8.1   2023-04-18 [1] CRAN (R 4.2.0)
## profvis        0.3.7   2020-11-02 [1] CRAN (R 4.2.0)
## promises       1.2.0.1 2021-02-11 [1] CRAN (R 4.2.0)
## ps             1.7.5   2023-04-18 [1] CRAN (R 4.2.0)
## purrr        * 1.0.1   2023-01-10 [1] CRAN (R 4.2.0)
## R6             2.5.1   2021-08-19 [1] CRAN (R 4.2.0)
## RColorBrewer   1.1-3   2022-04-03 [1] CRAN (R 4.2.0)
## Rcpp           1.0.10  2023-01-22 [1] CRAN (R 4.2.0)
## readr        * 2.1.4   2023-02-10 [1] CRAN (R 4.2.0)
## remotes        2.4.2   2021-11-30 [1] CRAN (R 4.2.0)
## rlang          1.1.1   2023-04-28 [1] CRAN (R 4.2.0)
## rmarkdown      2.20    2023-01-19 [1] CRAN (R 4.2.2)
## rstudioapi     0.14    2022-08-22 [1] CRAN (R 4.2.0)
## scales       * 1.2.1   2022-08-20 [1] CRAN (R 4.2.0)
## sessioninfo    1.2.2   2021-12-06 [1] CRAN (R 4.2.0)
## shiny          1.7.4   2022-12-15 [1] CRAN (R 4.2.0)
## stringi        1.7.12  2023-01-11 [1] CRAN (R 4.2.0)
## stringr      * 1.5.0   2022-12-02 [1] CRAN (R 4.2.0)
## tibble       * 3.2.1   2023-03-20 [1] CRAN (R 4.2.0)
## tidyr        * 1.3.0   2023-01-24 [1] CRAN (R 4.2.0)
## tidyselect     1.2.0   2022-10-10 [1] CRAN (R 4.2.0)
## tidyverse    * 2.0.0   2023-02-22 [1] CRAN (R 4.2.0)
## timechange     0.2.0   2023-01-11 [1] CRAN (R 4.2.0)
## tzdb           0.3.0   2022-03-28 [1] CRAN (R 4.2.0)
## urlchecker     1.0.1   2021-11-30 [1] CRAN (R 4.2.0)
## usethis        2.1.6   2022-05-25 [1] CRAN (R 4.2.0)
## utf8           1.2.3   2023-01-31 [1] CRAN (R 4.2.0)
## vctrs          0.6.2   2023-04-19 [1] CRAN (R 4.2.0)
## withr          2.5.0   2022-03-03 [1] CRAN (R 4.2.0)
## xfun           0.37    2023-01-31 [1] CRAN (R 4.2.0)
## xtable         1.8-4   2019-04-21 [1] CRAN (R 4.2.0)
```

```
##  yaml             2.3.7   2023-01-23 [1] CRAN (R 4.2.0)
##
##  [1] /Library/Frameworks/R.framework/Versions/4.2/Resources/library
##
## ----------------------------------------------------------
```

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369.

Aspillaga, M. (1996). Perceptual foundations in the design of visual displays. *Computers in Human Behavior*, *12*(4), 587–600.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (Vol. 2, pp. 89–195). Elsevier.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29.

Baddeley, Alan D. (1976). *The psychology of memory / alan d. baddeley* (pp. xvii, 430 p. :). Book, Basic Books New York.

Bartlett, F. A., & Remembering, A. (1932). A study in experimental and social psychology. New York: Cambridge University Press.

Bendix, F., Kosara, R., & Hauser, H. (2005). Parallel Sets: Visual Analysis of Categorical Data. In *2005 IEEE symposium on information visualization (INFOVIS'05)* (pp. 133–140). IEEE. http://doi.org/10.1109/INFOVIS.2005.27

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309.

Bui, D. C., & Myerson, J. (2014). The role of working memory abilities in lecture note-taking. *Learning and Individual Differences*, *33*, 12–22.

Cairo, A. (2016). *The truthful art: Data, charts, and maps for communication*. New Riders.

Card, S., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale [etc.]: Lawrence Erlbaum.

Cardoso, R. L., Leite, R. O., & Aquino, A. C. B. de. (2016). A graph is worth a thousand words: How overconfidence and graphical disclosure of numerical information influence financial analysts accuracy on decision making. *PloS One*, *11*(8), e0160443.

Carley, K., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. *Social Forces*, *70*(3), 601–636.

Chandler, C., & Unger, R. (2012). *A project guide to UX design: For user experience designers in the field or in the making*. New Riders.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Chevalier, A., & Kicka, M. (2006). Web designers and web users: Influence of the ergonomic quality of the web site on the information search. *International Journal of Human-Computer Studies*, *64*(10), 1031–1048.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*(387), 531–554.

Cleveland, W. S., & McGill, R. (1986). An experiment in graphical perception. *International Journal of Man-Machine Studies*, *25*(5), 491–500.

Cooper, A. (1999). *The inmates are running the asylum*. Springer.

Cooper, A., Reimann, R., Cronin, D., & Noessel, C. (2014). *About face: The essentials of interaction design*. John Wiley & Sons.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57.

Crapo, A., Waisel, L., Wallace, W., & Willemain, T. (2000). Visualization and the process of modeling: A cognitive-theoretic view. In *Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 218–226). http://doi.org/10.1145/347090.347129 d'Ocagne, M. (1885). Coordonnées parallèles et axiales : Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles. *Gauthier-Villars*, 112. Retrieved from https://archive.org/details/coordonnesparal00ocaggoog/page/n10

Day, R. H., & Stecher, E. J. (1991). Sine of an illusion. *Perception*, *20*, 49–55.

DeDonno, M. A. (2016). The influence of IQ on pure discovery and guided discovery learning of a complex real-world task. *Learning and Individual Differences*, *49*, 11–16.

Eissele, M., Weiskopf, D., & Ertl, T. (2009). Interactive context-aware visualization for mobile devices. In *International symposium on smart graphics.*

Eppler, M. J., & Bresciani, S. (2013). Visualization in management: From communication to collaboration. A response to zhang. *Journal of Visual Languages & Computing*, *24*(2), 146–149.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*(1), 273–305.

Few, S. (2006a). *Information dashboard design: The effective visual communication of data.* Newton, MA: O'Reilly Media, Inc.

Few, S. (2006b). *Information dashboard design: The effective visual communication of data.* Newton, MA: O'Reilly Media, Inc.

Few, S. (2009). *Now you see it: Simple visualization techniques for quantitative analysis*. Analytics Press.

Fitts, P. M., & Posner, M. I. (1967). Human performance.

Friendly, M. (2014). Comment on "the generalized pairs plot." *Journal of Computational and Graphical Statistics*, *23*(1), 290–291.

Friendly, M., & Denis, D. J. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. *URL Http://Www. Datavis. Ca/Milestones*, *32*, 13.

Garrett, J. J. (2003). The elements of user experience: User-centered design for the web.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *171*(1024), 279–296.

Hassenzahl, M. (2010). Experience design: Technology for all the right reasons. *Synthesis Lectures on Human-Centered Informatics*, *3*(1), 1–95.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 203–212).

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Commun. ACM*, *55*(4), 45–54. http://doi.org/10.1145/2133806.2133821

Heinrich, J., & Weiskopf, D. (2009). Continuous Parallel Coordinates. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1531–1538. http://doi.org/10.1109/TVCG.2009.131

Hofmann, H., & Vendettuoli, M. (2013). Common Angle Plots as Perception-True Visualizations of Categorical Associations. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2297–2305. http://doi.org/10.1109/TVCG.2013.140

Huang, W., Hong, S.-H., & Eades, P. (2006). Predicting graph reading performance: A cognitive approach. In *ACM international conference proceeding series* (Vol. 164, pp. 207–216).

Hullman, J., Adar, E., & Shah, P. (2011). Benefitting infovis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2213–2222.

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2917–2926.

Kirk, A. (2012). *Data visualization: A successful design process*. Packt publishing LTD.

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data–frame theory of sensemaking. In *Expertise out of context* (pp. 118–160). Psychology Press.

Koffka, K. (2013). *Principles of gestalt psychology* (Vol. 44). routledge.

Kosara, R. (2007). Visualization criticism-the missing link between information visualization and art. In *2007 11th international conference information visualization (IV'07)* (pp. 631–636). IEEE.

Lee, S., Kim, S.-H., Hung, Y.-H., Lam, H., Kang, Y., & Yi, J. S. (2016). How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE*, *22*, 499–508.

Luxen, D., & Vetter, C. (2011). Real-time routing with OpenStreetMap data. In *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 513–516). New York, NY, USA: ACM. http://doi.org/10.1145/2093973.2094062

Monmonier, M. (1985). *Annals of the Association of American Geographers*, *75*(4), 605–609. Retrieved from http://www.jstor.org/stable/2563117

Morville, P., Rosenfeld, L. B., & Arango, J. (2015). Information architecture for the web and beyond. *(No Title)*.

National Center for Education Statistics. (2020). National center for education statistics. https://nces.ed.gov/.

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955.

Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation*, *51*, 45–100.

Parsons, P., & Sedig, K. (2014). Adjustable properties of visual representations: Improving the quality of human-information interaction. *Journal of the Association for Information Science and Technology*, *65*(3), 455–482.

Peters, D. J. (2019). Community Resiliency in Declining Small Towns: Impact of Population Loss on Quality of Life over 20 Years. *Rural Sociology*, *84*(4), 635–668. http://doi.org/10.1111/ruso.12261

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reuter, L. H., Tukey, P., Maloney, L. T., Pani, J. R., & Smith, S. (1990). Human perception and visualization. In *Proceedings of the 1st conference on visualization'90* (pp. 401–406).

Rumelhart, D. E., & Ortony, A. (1976). The representation of knowledge in memory. In *Center for human information processing*.

Rural Shrink Smart Team. (2022). Rural shrink smart. https://ruralshrinksmart.org/.

Saffer, D. (2007). Designing for interaction: Creating smart applications and clever devices. New riders. AIGE design press.

Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2019). What do we talk about when we talk about dashboards? *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 682–692. http://doi.org/10.1109/TVCG.2018.2864903

Satyanarayan, A., Moritz, D., Wongsuphasawat, K., & Heer, J. (2016). Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 341–350.

State of Iowa. (2020). Iowa data portal. `https://data.iowa.gov`.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, *12*(3), 185–233.

Toyama, T., Sonntag, D., Orlosky, J., & Kiyokawa, K. (2015). Attention engagement and cognitive state analysis for augmented reality text display functions. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 322–332).

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1373), 1295–1306.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Tufte, Edward R. (1985). The visual display of quantitative information. *The Journal for Healthcare Quality (JHQ)*, *7*(3), 15.

Tufte, Edward R. (2001). *The visual display of quantitative information (2nd edition)*. USA: Graphics Press.

Tufte, Edward R. (2008). The cognitive style of PowerPoint: Pitching out corrupts within. *Education Review*.

Tukey, J. W., & Wilk, M. B. (1966). Data analysis and statistics: An expository overview. In *Proceedings of the november 7-10, 1966, fall joint computer conference* (pp. 695–709).

Tuovinen, J. E., & Paas, F. (2004). Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional Science*, 133–152.

Unwin, A., Volinsky, C., & Winkler, S. (2003). Parallel coordinates for exploratory modelling analysis. *Computational Statistics & Data Analysis*, *43*, 553–564. http://doi.org/10.1016/S0167-9473(02)00292-X

USDA - ERS. (2020a). Rural classifications. `https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/`.

USDA - ERS. (2020b). Rural-urban commuting area codes. `https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/`.

Van Wijk, J. J. (2005). The value of visualization. In *VIS 05. IEEE visualization, 2005.* (pp. 79–86). IEEE.

VanderPlas, S., & Hofmann, H. (2015a). Signs of the sine illusion–why we need to care. *Journal of Computational and Graphical Statistics*, *24*(4), 1170–1190. http://doi.org/10.1080/10618600.2014.951547

VanderPlas, S., & Hofmann, H. (2015b). Signs of the sine illusion—why we need to care. *Journal of Computational and Graphical Statistics*, *24*(4), 1170–1190.

Vessey, I., & Galletta, D. (1991). Cognitive fit: An empirical study of information acquisition. *Information Systems Research*, *2*(1), 63–84.

Ward, M., Grinstein, G., & Keim, D. (2010). *Interactive data visualization: Foundations, techniques, and applications* (p. 81). http://doi.org/10.1201/b10683

Ware, C. (2012). *Information visualization: Perception for design*. Morgan Kaufmann.

Wertheimer, M. (1938). Laws of organization in perceptual forms.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed.). Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., Cook, D., Hofmann, H., & Buja, A. (2011). tourr: An R Package for Exploring Multivariate Data with Projections. *Journal of Statistical Software, Articles*, *40*(2), 1–18. http://doi.org/10.18637/jss.v040.i02

Wilkinson, L. (2012). *The grammar of graphics*. Springer.

Yi, J. S., Kang, Y. ah, Stasko, J., & Jacko, J. A. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1224–1231.

Zarecor, K. E., Peters, D. J., & Hamideh, S. (2021). Rural smart shrinkage and perceptions of quality of life in the american midwest. In *Handbook of quality of life and sustainability* (pp. 395–415). Springer.