Informatics

# Linienwahrnehmung in Parallelen Koordinaten unter verschiedenen Seitenverhältnissen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieur**

im Rahmen des Studiums

**Media and Human-Centered Computing**

eingereicht von

**Leon Thierry Meka, Bsc**
Matrikelnummer 12045662

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller
Mitwirkung: Univ.Lektorin Dipl.-Ing.in Dr.in techn. Johanna Schmidt

Wien, 19. September 2024

Leon Thierry Meka    Eduard Gröller

# Line Perception in Parallel Coordinates under different Aspect Ratios

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieur**

in

**Media and Human-Centered Computing**

by

**Leon Thierry Meka, Bsc**
Registration Number 12045662

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller
Assistance: Univ.Lektorin Dipl.-Ing.in Dr.in techn. Johanna Schmidt

Vienna, September 19, 2024

Leon Thierry Meka

Eduard Gröller

# Erklärung zur Verfassung der Arbeit

Leon Thierry Meka, Bsc

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

Wien, 19. September 2024

Leon Thierry Meka

# Kurzfassung

In dieser Arbeit werden die Auswirkungen verschiedener Seitenverhältnisse auf die Wahrnehmung von Winkelverhältnissen und Linien in Parallelen Koordinaten untersucht. Parallele Koordinaten sind eine Visualisierungstechnik zur Darstellung multivariater Daten, bei der jede zu untersuchende Variable als eine parallele Achse dargestellt wird und Datenpunkte durch Linien verbunden werden, die sich durch diese Achsen ziehen. Diese Methode ermöglicht die gleichzeitige Visualisierung von mehr als zwei Variablen und die Interpretation von Korrelationsmustern innerhalb eines Datensatzes.

Die Zuverlässigkeit und Genauigkeit der Interpretation kann jedoch durch das Seitenverhältnis des Plots erheblich beeinflusst werden. Das Ziel dieser Arbeit ist es daher, den Einfluss von Variationen des Seitenverhältnisses auf die Genauigkeit und das Vertrauen der Benutzer bei der Wahrnehmung von Korrelationen in Parallelen Koordinaten zu untersuchen.

Der methodische Ansatz umfasst drei Komponenten: die Entwicklung eines webbasierten Visualisierungswerkzeugs, eine statistische Analyse der Linien- und Winkelparameter, sowie eine empirische Nutzerstudie. Das Visualisierungswerkzeug ermöglicht die Darstellung Paralleler Koordinaten in verschiedenen Seitenverhältnissen und die Analyse geometrischer Eigenschaften der Linien im Plot. Die statistische Analyse zeigt, dass das Seitenverhältnis signifikante Korrelationen mit den minimalen und maximalen Linienwinkeln in parallelen Koordinaten aufweist, was die visuelle Wahrnehmung und Interpretation der Daten beeinflusst. Diese Ergebnisse werden durch eine webbasierte Benutzerstudie bestätigt, die darüber hinaus aufzeigt, dass bestimmte Seitenverhältnisse präzisere und zuverlässigere Korrelationsschätzungen liefern. Die Ergebnisse verdeutlichen den bewussten Einsatz von flexiblen Seitenverhältnissen, um Verzerrungen zu minimieren und die Zuverlässigkeit der visuellen Interpretation in Parallelen Koordinaten zu gewährleisten.

# Abstract

This thesis investigates the impact of different aspect ratios on the perception of angles and lines in parallel coordinates. Parallel coordinates are a visualization technique for representing multivariate data where each variable is drawn as a parallel axis, and data points are connected by lines across these axes. This method allows for the simultaneous visualization of more than two variables and enables the interpretation of correlation patterns within a given dataset.

However, the reliability and accuracy of this interpretation can be significantly influenced by the aspect ratio of the plot. This thesis aims to explore how variations in aspect ratios affect the accuracy and confidence of users in perceiving correlations within parallel coordinates.

The methodological approach comprises three components: the development of a web-based visualization tool, a statistical analysis of line and angle parameters, and an empirical user study. The visualization tool enables users to display parallel coordinates in various aspect ratios and analyze the geometric properties of the lines in the plot. The statistical analysis reveals that aspect ratios significantly correlate with the minimum and maximum angles in parallel coordinates, which in turn affects the visual perception and interpretation of the data. These findings are validated through a web-based user study, demonstrating that specific aspect ratios lead to more accurate and reliable correlation estimates. The results underscore considerate usage of flexible aspect ratios to minimize distortion and ensure the reliability of visual data interpretation in parallel coordinates.

# Contents

# CHAPTER 1

# Introduction

Parallel coordinates are a powerful visualization technique for representing datasets with multiple variables. By arranging axes in parallel rather than perpendicularly, as in traditional Cartesian coordinates [1], this method allows the visualization of more than two variables simultaneously. As visualized in Figure 1.1, each axis corresponds to a variable, and data points are displayed as lines intersecting each axis at points corresponding to their respective values [1].

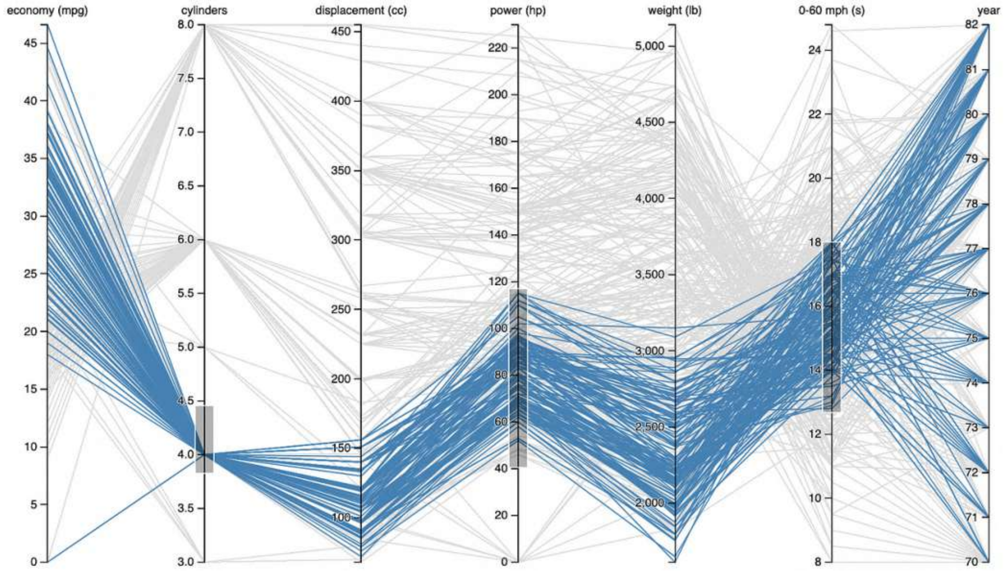


Figure 1.1: Example of a parallel coordinates plot, visualizing various characteristics of automobiles. The highlighted subset emphasizes cars with four cylinders, medium to low horsepower, and medium acceleration rate. Image taken from Ilievski [2].

This parallel axes arrangement can be used for interpretation of data patterns between variables, which are otherwise challenging to gauge in other visualization techniques. Therefore, one of the most valuable features of parallel coordinates is the ability to identify correlation patterns in variable pairs. This works by analyzing the angles and intersections of lines between axes, which tend to form interpretable patterns. These visual cues can reveal relationships such as how closely related two variables are and whether the relationship is direct or inverse. Since correlations are manifested through these patterns, parallel coordinates are especially useful for visual correlation estimation.

Parallel coordinates are widely adopted in various visualization libraries and applications due to their ability to visually display complex datasets. Although traditionally used in fields like demographics or finance, they have also found recent usage in fields like machine learning, where understanding the complex relationships between many hyperparameters is valuable for model fine-tuning [3].

Parallel coordinates remain a promising visualization technique for gathering information on datasets that might be difficult to uncover using other methods. As a result, they continue to be a popular choice for researchers and data analysts.

## 1.1 Motivation and Problem Statement

The reliability of parallel coordinates can significantly vary with their representation – notably with the aspect ratio. The aspect ratio defines the proportion between the width and height of a visualization and can have noticeable effects on the visual dynamics of these plots [4].

Previous research has shown, aspect ratios' influences on angles between lines and axes in visualizations, misleading viewers about the true nature of data. For example, altering the aspect ratio in line charts can significantly affect slope perception, leading viewers to either exaggerate or understate trends [5][6].

Despite its significance, the aspect ratio is often unintentionally altered in responsive user interfaces. As users adjust application window sizes or interact with multi-view dashboards, the aspect ratios of parallel coordinates may shift unpredictably. While beneficial for layout customization, this flexibility can unintentionally distort the critical visual cues for accurate data interpretation (see Figure 1.2).

While it is well-established that aspect ratio affects the perception of slopes and angles in traditional visualizations, its specific impact on the perception of angular parameters in parallel coordinates remains unexplored. This gap in research motivated our study, as we aimed to investigate how different aspect ratios influence the accuracy and reliability of interpreting correlation patterns in parallel coordinates.

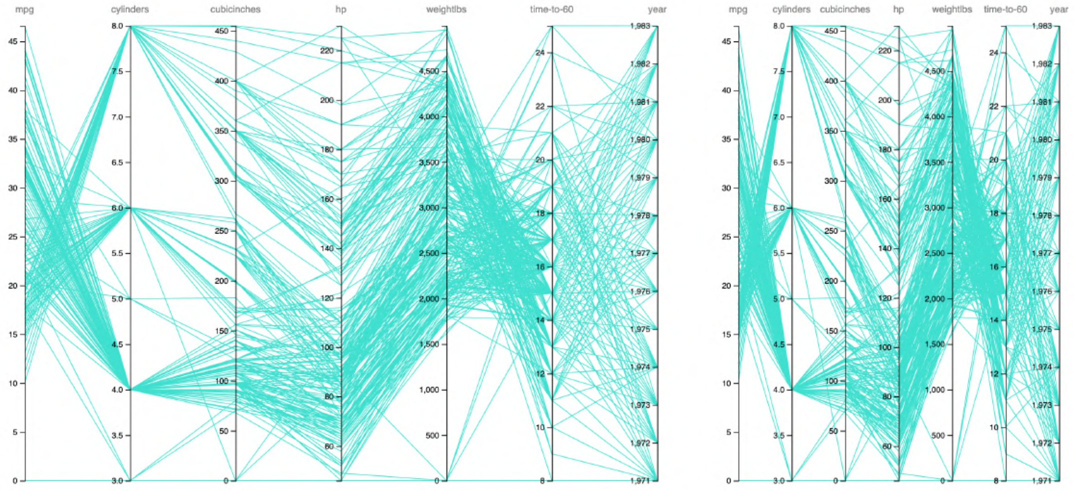


Figure 1.2: The same parallel coordinates plot with different aspect ratios: the left image shows a wide aspect ratio (4:3), while the right one is narrower (3:4). Note how the chosen aspect ratio has noticeable effects on the perception of line angles within the plot.

## 1.2 Research Question

Our research sought to close this gap by answering the following research question:

**RQ:** *How does aspect ratio influence the perception of correlation in parallel coordinates?*

To thoroughly investigate this topic, we focused on exploring how varying aspect ratios affect the user's ability to accurately judge correlations between variables. Given that angles and lines are key indicators of relationships in parallel coordinates, we argued that changes in aspect ratio might distort these visual indicators, leading to incorrect interpretations. Moreover, this study sought to bridge the gap between existing knowledge on the effects of aspect ratios in traditional visualization techniques and their effects on parallel coordinates.

## 1.3 Goal and Expected Results

To answer the proposed research question, the goal of this thesis was to provide statistical and empirical evidence for understanding how different aspect ratios influence correlation perception in parallel coordinates. To achieve this, the thesis focused on generating the following results:

**Visualization Tool:** A web-based application capable of adjusting parallel coordinates across a range of aspect ratios, equipped with features for analyzing and capturing underlying geometric properties such as line angles and correlation.

**Statistical Analysis:** A statistical analysis investigating the relations between aspect ratio variations and the geometric properties of parallel coordinates. This includes identifying specific properties that correlate with the aspect ratio in order to validate our initial hypothesis.

**Quantitative Analysis:** The results of a web-based study providing a quantitative analysis of human perception of relationships between variables under different aspect ratios. This analysis may empirically answer our research question and suggest which aspect ratios were most effective for accurate visual correlation estimation.

## 1.4 Methodological Approach

We developed a methodological approach aimed at contributing to a better understanding and improved utilization of parallel coordinates. We began with an extensive review of related literature on correlation analysis and data visualization, with a particular focus on parallel coordinates and the influence of aspect ratios. This phase was intended to establish fundamental concepts and gather findings from previous studies related to the field.

Following this research, we formulated hypotheses based on the results from our literature review. These hypotheses focused on how aspect ratios might affect users' perception of correlations within the plots.

The next step involved developing a web-based visualization tool. This tool allowed users to interactively adjust and analyze parallel coordinates with varying aspect ratios, serving as the primary means for visual experimentation and data collection in subsequent stages. Further, we conducted a statistical analysis using data collected from the web application. This analysis focused on how different geometric properties correlate with the aspect ratio. Based on this preliminary analysis, first findings regarding the influence of aspect ratio on the geometric properties of parallel coordinates were gathered.

The next phase involved the generation of datasets designed to evaluate our hypotheses. Once the datasets were ready, the experimental setup was configured, ensuring all necessary tools and systems were in place for conducting the user study. This included integrating the web application with data collection mechanisms to track user interactions and responses.

A user study was conducted to collect empirical data. Participants used the developed web-based visualization tool to interact with parallel coordinates, and their responses regarding the perceived correlations and self-confidence were captured. The collected data was evaluated and based on these results, conclusions were drawn to either accept or reject the formulated hypotheses.

CHAPTER 2

# Fundamentals and Related Work

The following chapter covers the fundamental concepts and related work necessary to understand correlation analysis, data visualization, parallel coordinates, and the effects of aspect ratios in data visualization. Following, will take a closer look at correlation analysis and explore its different manifestations, measures like the Pearson's correlation coefficient, and important distinctions between correlation and causation. We also look at methods for assessing statistical significance and potential errors in hypothesis testing.

Next, we move to data visualization, outlining common and novel taxonomies for their categorization and describe the data visualization pipeline. Further, we examine techniques for visualizing different types of datasets, highlighting their key distinctions and individual use cases. We then focus on how parallel coordinates represent complex datasets, which interaction techniques they offer and illustrate their relevance by exploring selected case studies. Finally, we explain the role of aspect ratios in graphical representations and discuss proposed solutions for counteracting its effects.

## 2.1 Correlation Analysis

Correlation analysis is a discipline in descriptive statistics that aims to understand how variables relate to each other. More specifically, it denotes the magnitude and nature of the relationship between two or more variables [7]. Understanding these relationships involves several key concepts.

### 2.1.1 Variance

Variance measures the dispersion of points in relation to an underlying mean value [7]. It describes how much the values in a dataset deviate from this mean value. For a given sample of $n$ observations $x_1, x_2, \ldots, x_n$, the variance $\mathrm{Var}(X)$ is calculated as:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

where $\overline{x}$ is the sample mean for variable X:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

As shown in Figure 2.1, high variance is present in datasets where data points are largly scattered away from the center. On the other hand, lower variance suggests that data points are closer to the center [7] [8].

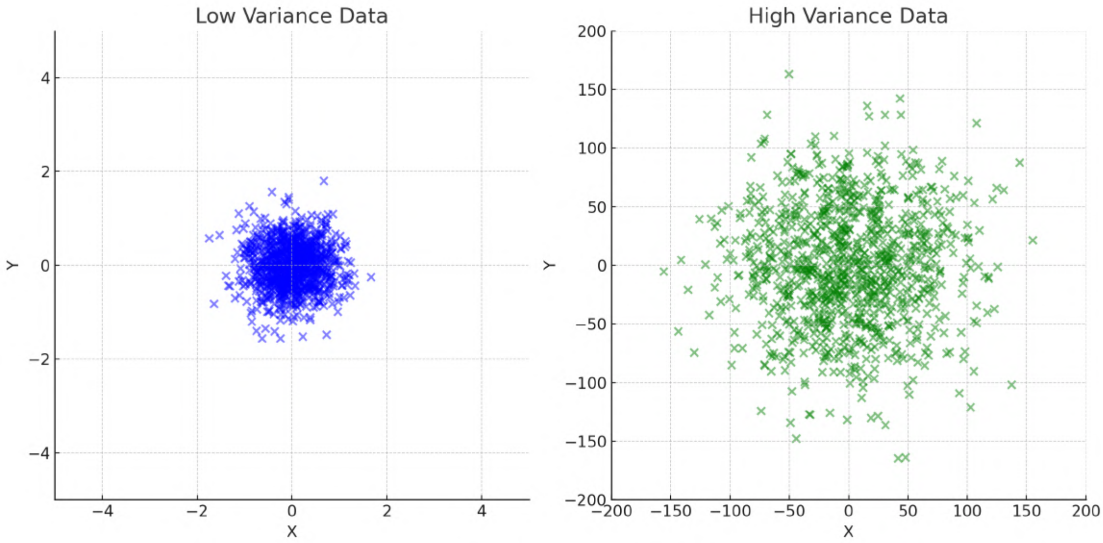


Figure 2.1: Scatterplots of datasets with low (left) and high variance (right). The horizontal axis represents variable $X$, and the vertical axis represents variable $Y$.

Variance is a fairly important metric for understanding the degree of variability within a dataset and is often used in scientific research as for evaluating experimental results.

### 2.1.2 Covariance

Unlike variance, covariance describes the linear relationship between two variables. Figure 2.2 highlights the characteristics of covariance within two different datasets. Essentially, covariance describes the extent of concurrent or opposite trends within two given variables [7] and can be expressed by:

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{N - 1}$$

Where $x_i$ and $y_i$ are individual data points from two variables $X$ and $Y$, $\overline{x}$ and $\overline{y}$ are their respective means, and $N$ is the number of data points.

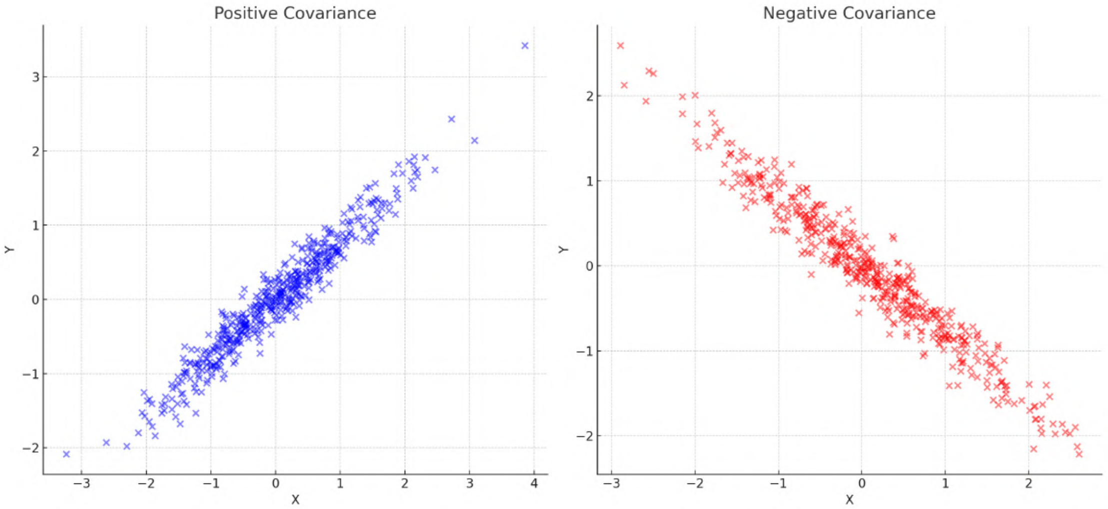


Figure 2.2: Scatterplots depicting datasets with strong positive and negative covariance. The left plot illustrates a positive covariance, where $X$ and $Y$ increase together, while the right plot displays negative covariance, where $Y$ decreases as $X$ increases.

Covariance describes a positive or negative value, which expresses the nature of the linearity between variables $X$ and $Y$.

### 2.1.3 Types of Relationships

Closer examining the covariance's sign gives us a more nuanced look into the different manifestations of relationships. As visualized in Figure 2.3, researchers generally divide relationships into three categories.

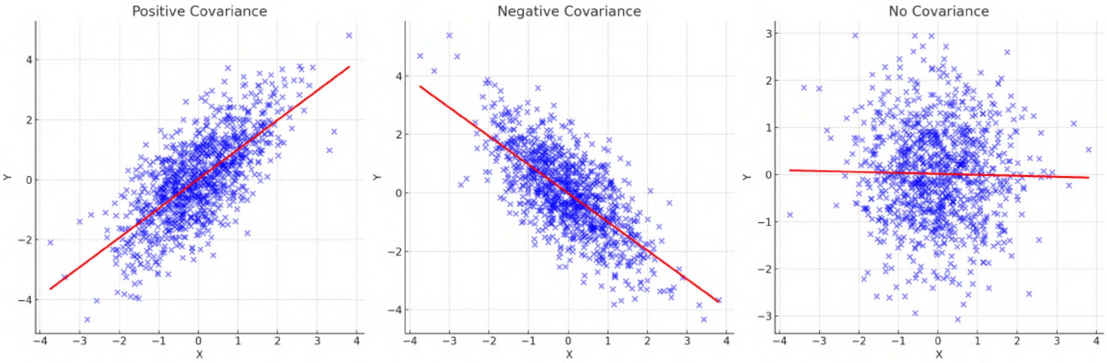


Figure 2.3: Scatterplots illustrating three types of linear relationships: positive covariance (left), negative covariance (center), and no covariance (right). The red lines indicate the trend direction for each relationship, highlighting how $X$ and $Y$ vary together (positive or negative covariance) or show no consistent trend (no covariance).

### Positive Relationship

A positive relationship exists if an increase in variable $X$ leads to an increase in variable $Y$. This means that the variables move in the same direction. A typical example in this context is the relationship between hours studied $(X)$ and exam results $(Y)$ [9]. Generally, as the number of hours studied increases, the exam scores also increase, reflecting a positive relationship. In terms of covariance, this means $\sigma_{xy} > 0$.

### Negative Relationship

Negative relationships, where $\sigma_{xy} < 0$, are present if an increase in variable $X$ leads to a decrease in variable $Y$, indicating that the variables move in opposite directions. In the exam case, a negative relationship would be represented by the number of hours spent playing video games and exam grades. As the number of hours spent on distractions increases, the examination scores tend to decrease, describing a negative relationship.

### No Relationship

Whenever covariance nears 0 ($\sigma_{xy} \approx 0$), it can be assumed that no underlying relationship is present. In this case, changes in variable $X$ do not predictably affect variable $Y$. In other words, the variables do not exhibit any consistent linear trend. To illustrate this, one could argue that there is no real relationship between the number of hours studied and the amount of rainfall in a city.

## 2.1.4 Correlation

Since covariance can manifest as arbitrarily high positive or negative values, researchers often aim for a normalized measurement to make further calculations easier [7]. To do this, the covariance is divided by the product of the standard deviations:

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

This value, also known as the *Pearson product-moment correlation coefficient* [7], results in a normalized expression between -1 and 1. The Pearson correlation coefficient essentially attempts to determine the degree of linear relationship between two variables. As previously illustrated in 2.3, it indicates how accurately the association between the two variables can be expressed using a linear equation in the form of $y = ax + b$ [10].

In practical terms, a strong linear relationship is represented by $\rho_{xy}$ being very close to either 1 or -1. In contrast, values close to 0 imply a weaker or non-existent linear relationship.

### 2.1.5 Correlation Coefficients

Similar to the Pearson correlation coefficient, researchers can employ several other coefficients to describe the relationship between variables within a dataset. These coefficients can be broadly classified based on the criteria they pose on a given dataset, as well as the scale level of the underlying variables. In literature, they are commonly referred to as *parametric* and *non-parametric* [11][7].

**Parametric correlation coefficients** are used if the variables are measured on an interval or ratio scale [7]. These coefficients assume specific characteristics about the data distribution, such as linearity, normality, and homoscedasticity[1]. They are sensitive to deviations from these assumptions and may not necessarily provide accurate results if these assumptions are violated. The Pearson correlation coefficient is usually the most commonly used parametric measure [7].

**Non-parametric correlation coefficients**, on the other hand, are suitable for variables measured on an ordinal scale or if the data does not meet the assumptions required for parametric methods. These coefficients do not make strict assumptions about the underlying data distribution, making them more robust for non-normally distributed data. Examples of non-parametric measures include Spearman's rank correlation and Kendall's tau [7].

### 2.1.6 Significance

Statisticians conduct tests to reliably deem a correlation significant enough to be important. Testing for significance is vital since researchers must collect statistical evidence to make sense of data. When conducting a significance test, researchers are essentially saying that the observed correlation in a given sample is unlikely to have occurred randomly [7].

[1] Homogenity of variance, i.e., the distribution of the $y_i$ values must have the same variance for all groups of $x_i$ [12].

When testing for significance in correlation analysis, standardized $t$- or $z$-tests are used. For both, the population is defined as a baseline of comparison. Note that the method employed for significance testing is dependent on the underlying population correlation coefficient $\rho$.

### t-Test for $\rho = 0$

The t-test [7] effectively answers the question of whether a given sample correlation, e.g., the Pearson correlation between two variables, significantly differs from zero (given it is known that $\rho = 0$), which indirectly tests whether there is any correlation in the population at all. This can be calculated as follows:

$$t_{N-2} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

where $r$ is the sample correlation coefficient, and $N$ is the number of paired observations. The hypotheses for the t-test are:

- Null hypothesis ($H_0$): The sample correlation $r$ is not significantly different from zero.
- Alternative hypothesis ($H_1$): The sample correlation $r$ is significantly different from zero.

The test statistic $|t_{N-2}|$ is compared to the critical value $t_{\alpha/2,N-2}$ from the t-distribution with $N-2$ degrees of freedom at a chosen significance level $\alpha$. If $|t_{N-2}| > t_{\alpha/2,N-2}$, $H_0$ is rejected, concluding that there is a statistically significant correlation in the population. If not, $H_0$ is not rejected, meaning that the evidence is insufficient to conclude that the correlation is different from zero.

### z-Test for $\rho \neq 0$

Alternatively, if it is known that $\rho \neq 0$, researchers encounter either right-skewed ($\rho > 0$) or left-skewed ($\rho < 0$) sample score distributions. To handle this skewness and to test the significance of the correlation, Fisher's Z-transformation is typically applied, followed by a z-test [7].

1. Fisher's $Z$-transformation:

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

where $r$ is the sample correlation coefficient.

2. Mean of the Fisher's $Z$-transformed values:

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

where $\rho$ is the population correlation coefficient.

3. Standard error of $Z$:

$$\sigma_z = \frac{1}{\sqrt{N-3}}$$

where $N$ is the number of paired observations.

4. $z$-score:

$$z = \frac{Z - \mu_z}{\sigma_z}$$

After applying the transformation and calculating $z$, researchers set up the following hypotheses:

- Null hypothesis ($H_0$): The sample correlation $r$ is not significantly different from the population correlation $\rho$.
- Alternative hypothesis ($H_1$): The sample correlation $r$ is significantly different from the population correlation $\rho$.

Similar to the t-test, the calculated $z$-score is compared to the critical values from the standard normal z-distribution. If $|z| > z_{\alpha/2}$, $H_0$ is rejected, concluding that the correlation is statistically significant. Otherwise, $H_0$ is not rejected, suggesting that the sample correlation differs from the population correlation.

### 2.1.7 Error Types

When conducting hypothesis tests, researchers aim to reduce two primary types of errors: *Type I* and *Type II* errors [7]. A Type I error, also known as a false positive, occurs when $H_0$ is rejected if it is actually true. This error leads to the incorrect conclusion that there is a significant effect or correlation when there is none.

The previously selected significance level $\alpha$ controls the probability of making a Type I error. This expression indicates that there could be a $\alpha = 5\%$ chance of rejecting $H_0$ when it is true.

On the other hand, a Type II error (false negative) occurs when $H_0$ is not rejected despite being actually false. This error leads to the incorrect conclusion that there is no significant effect or correlation when there is one. The probability of making a Type II error is denoted by $\beta$. In addition to that, the statistical power of the test, $1 - \beta$, represents the likelihood of correctly detecting an effect if it truly exists.

In decision-making regarding which error to optimize for, it is generally recommended to consider the consequences of Type I and Type II errors in the context of the specific problem or domain [13]. If the consequences of a false positive are more severe, it would make sense to set a lower significance level to minimize the probability of this error. Conversely, if a false negative poses a greater risk, efforts should be made to increase the power of the test. This can be done e.g., by increasing the sample size or reducing the $\alpha$ level.

### 2.1.8 Correlation and Causation

Another critical distinction in statistical analysis is understanding the difference between correlation and causation. A common fallacy of correlation and data analysis is the assumption of an inherent causal relationship between variables that can be deduced by the correlation alone [7][14].

An often-cited [15][16][17] example in discussions of correlation versus causation is the relationship between chocolate consumption and the number of Nobel Prize winners in various countries [18]. Messerli [19] found a significant positive correlation between a country's per capita chocolate consumption and the number of Nobel winners it produces.

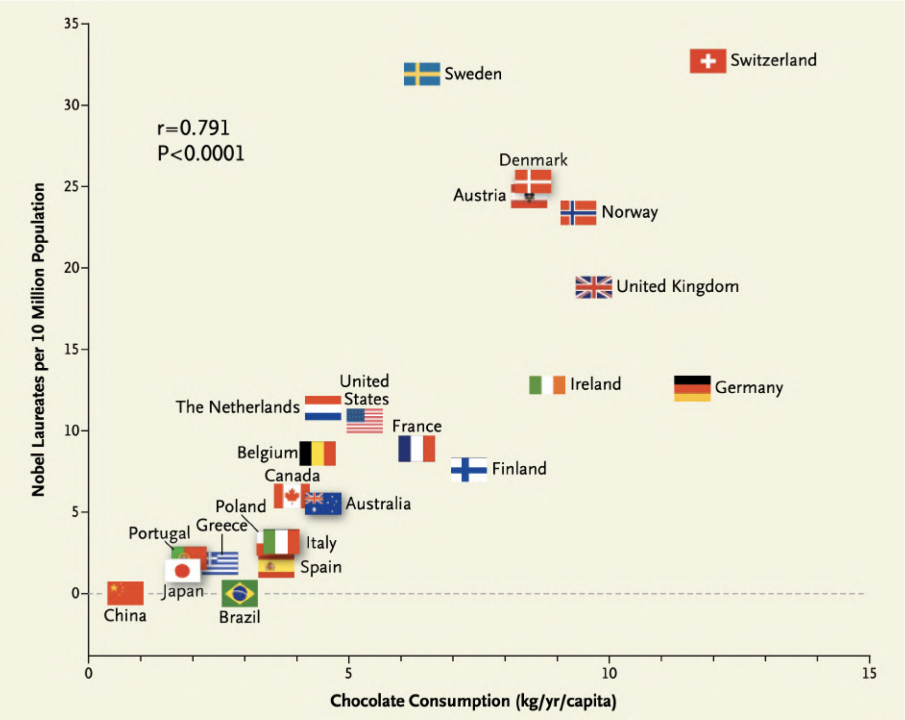


Figure 2.4: Correlation between chocolate consumption and Nobel Prize winners per 10 million population. Image taken from Messerli [19].

The study's reported correlation results suggest that countries with higher chocolate consumption tend to have more Nobel Prize winners. As illustrated in Figure 2.4, the author pointed at a linear relationship between the two variables. The peak was marked by Messerli's home country [20], Switzerland, known for a considerable number of Nobel laureates and exceptionally high chocolate consumption [19], further underlining this trend.

However, these assumptions were solely drawn from correlation patterns and are not backed by further investigation. The author even pointed out that "[...] a correlation between X and Y does not prove causation but indicates that either X influences Y, Y influences X, or X and Y are influenced by a common underlying mechanism" [19].

In this particular study, Messerli proposed some potential hypotheses to explain his observations; however, controlled experiments or longitudinal studies are required to move from correlation to causation effectively. Ultimately, this study, perfectly illustrates the issues of inferring causation from correlation.

### 2.1.9 Spurious and Confounding

In this context, understanding *spurious associations* and *confounding variables* as contextual factors is also important when interpreting correlations [14]. These phenomena can significantly distort our understanding of the relationship between variables, leading to false conclusions if not carefully considered [21]. The interconnected relationship between confounder, spurious associations and the variables involved can be seen in Figure 2.5.

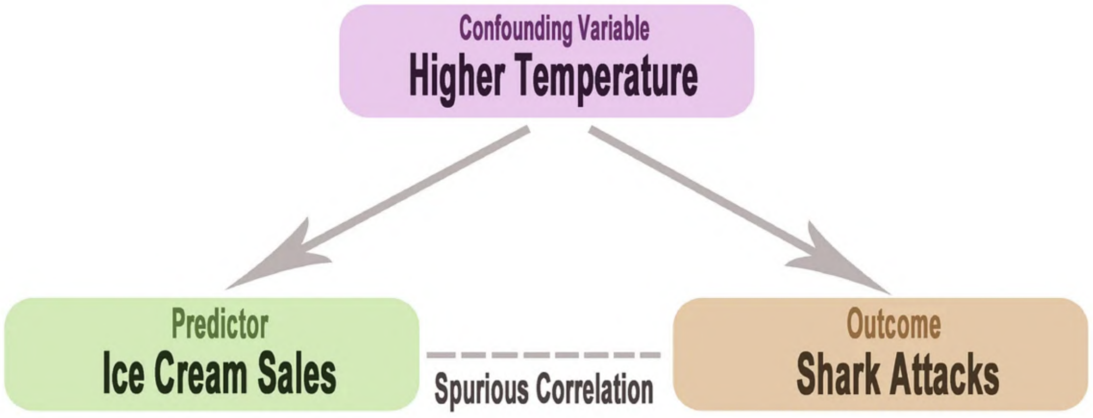


Figure 2.5: Illustration depicting the relationship between the confounding variable (high temperature) and spurious correlations with variable A (ice cream sales) and B (shark attacks). Image taken from Das [22].

**Spurious Association:** Spurious associations, or correlations, refer to a perceived high relationship between variables that occurs by chance or due to the influence of an external factor, leading to misleading conclusions about the actual relationship. The findings of the proposed study by Messerli [19] illustrate such a phenomenon.

**Confounding Variable:** A confounding variable is an external factor that influences both variables under study, creating a spurious association. For instance, a positive correlation between physical fitness and academic performance might be influenced by socioeconomic status, a confounding variable that affects both fitness opportunities and educational resources [23].

The interplay between confounding and observed variables, especially in bivariate correlation analysis, is never fully transparent [22]. However, considering additional variables when interpreting correlation patterns drawn from observed data, can already help reduce false conclusions. Confounding variables, in particular, need careful consideration as they can mask genuine relationships between the variables under study or suggest a relationship where none exists.

To effectively address these phenomena, researchers can employ restrictions or randomization to isolate the effects of the variables of interest [21]. If additional variables or data are available, running a multivariable regression analysis can further help in reasoning about underlying relationships [22].

## 2.2 Data Visualization

Data visualization is used as a means of visually presenting insights gathered from data. By translating data into graphical representations, characteristics in the data sets can be visualized that may not be apparent from raw numbers alone. [1].

To understand and develop effective visualization techniques, it is crucial to categorize them systematically. Tory et al. [24] claim this is due to two factors; First, taxonomies help user classify visualizations in order to guide them in effectively describing, using and thinking about them. Secondly, they advance the field as a whole by providing a meaningful way of organization, which serves as a foundation for research and discussion. For this purpose, various categorization schemes have been proposed over the years, each focusing on different aspects.

### 2.2.1 Categorizations

One common approach is categorizing visualizations by application area [25]. Especially, *scientific* and *information visualization* are coined terms in this field (see Figure 2.6).

Scientific visualizations typically focus on illustrating physical or mathematical phenomena with particular consideration for an accurate representation. It often involves data that has spatial or physical properties, such as medical imaging [26] or fluid dynamics [27]. In contrast, information visualization primarily deals with non-spatial data, such as a large corpus of text or numerical data, where the inherent goal is to make the underlying information more accessible, understandable, and usable.

Another framework for categorizing visualizations was presented by Shneiderman [28]. His taxonomy introduced the "Visual Information-Seeking Mantra"; with overview first, zoom and filter, then details on demand, as a guiding principle for designing graphical user interfaces. This taxonomy classifies visualizations based on seven data types (one-dimensional, two-dimensional, three-dimensional, temporal, multi-dimensional, tree, and network data) and seven tasks (overview, zoom, filter, details-on-demand, relate, history, and extract). By integrating these data types and tasks, the framework provides a more granular approach on categorization of the different types of data visualizations.

There are also categorization methods that solely focus on the interaction techniques used in visualizations. The taxonomy by Yi et al. [29] identified several fundamental interaction techniques that are essential for an effective visualization: select, explore, reconfigure, encode, abstract/elaborate, filter, and connect. Each technique addresses a different aspect of how users engage with visual data. Notably, this categorization heavily emphasizes the importance of user intent and interaction in the visualization process.

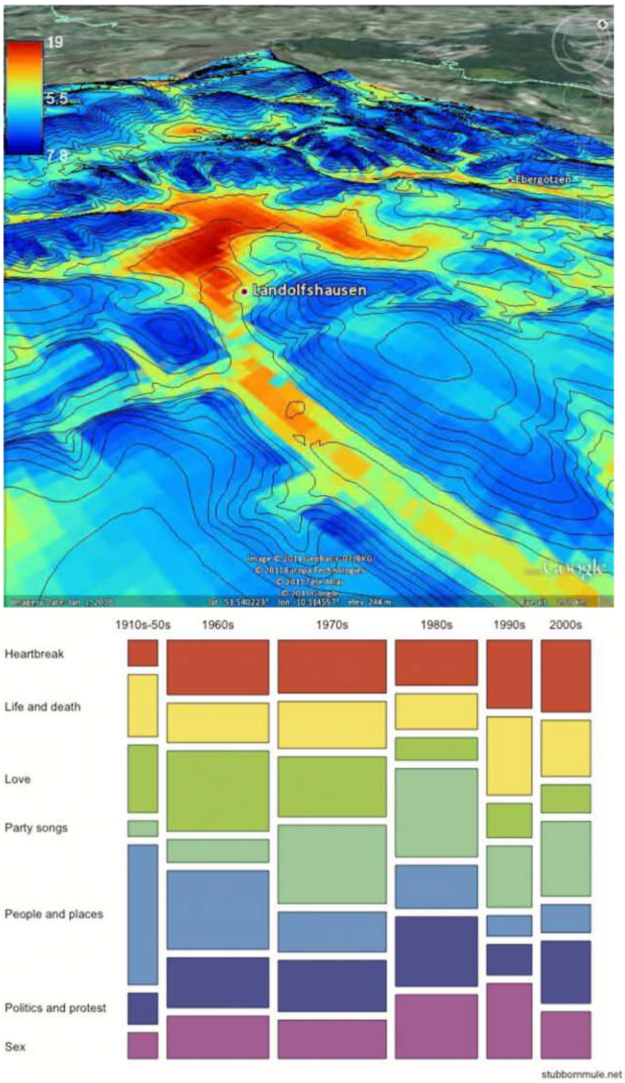


Figure 2.6: Examples of scientific and information visualizations. The left image represents a scientific visualization showing a topographical map with color-coded areas displaying geophysical measurements. The right image is an example of information visualization, depicting a treemap that categorizes music themes across decades. Images taken from Hengl et al. [30] and Chen [31].

### 2.2.2 Novel Taxonomies

Ongoing discussions in visualization research question the prevalence and necessity of existing classifications in visualization. A notable example was the panel discussion on the established distinctions within the field, hosted by Rhyne in 2003 [25]. The discussion included several key researchers from different areas. It was aimed at debating over maintaining separate categories for scientific and information visualization, questioning whether such distinctions are really needed [25].

Building on this debate, Tory and Möller proposed a novel taxonomy for classifying visualization techniques "[...] based on the design models of algorithms rather than the data itself" [24]. This model-based taxonomy emphasizes the human aspect of visualization, considering both the assumptions made by designers and the conceptual models held by users. The authors argue that common categories like *scientific* and *information visualization* are often ambiguous and overlapping, hindering the understanding and development of hybrid visualization techniques [25]. Instead, they classify design models as either discrete or continuous and analyze to the use of display attributes such as transparencies or color.

Tory and Möller's taxonomy is meant to offer a flexible framework that can accommodate a broader range of visualization techniques as a result of challenging the existing categorization approaches.

### 2.2.3 Data Visualization Pipeline

A central aspect of data visualization is the data visualization pipeline [32]. This pipeline proposes a step-wise process of transforming raw data into visual representations. It involves stages that process data, convert them into visual formats, and render them in ways that users can easily interpret (see Figure 2.7).

The term *Data Visualization Pipeline* and its conceptual framework were introduced and popularized by Card et al. [32]. This framework has since become a foundational concept in data visualization, guiding the design and implementation of effective data visualization applications. In the following, we discuss the different stages of the pipeline and outline their influence on the final representation.

#### Data Transformation

The initial stage in the pipeline is *data transformation*, which involves converting data into a structured and organized format suitable for visualization. Generally, this step is meant to both reduce and enhance the information contained in raw data [32].

According to Card et al. [32], data transformations can be classified into four types. First, he mentions transformations that can derive new values from existing ones through mathematical operations, i.e., calculating means or sums. These operations are mostly used to enhance the information contained in raw values.

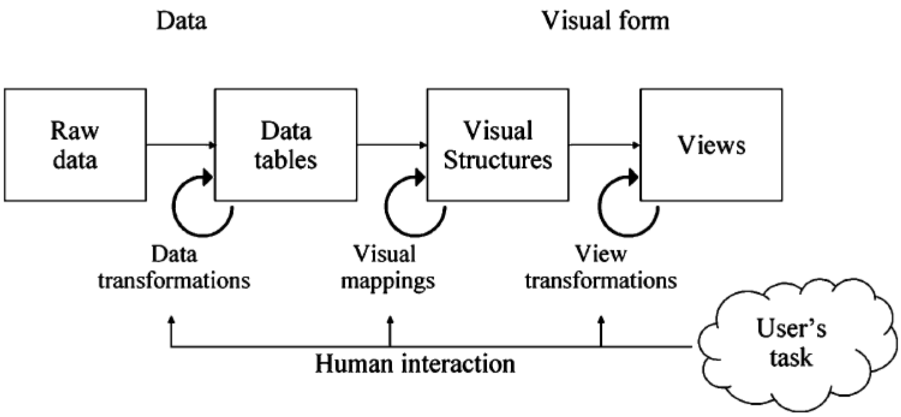


Figure 2.7: Illustration of the data visualization pipeline. Key stages include: transforming raw data into structured tables, mapping these tables into abstract visual structures, and finally generating views for user interaction. Human interaction influences the pipeline at various stages. Image taken from Card et al. [32].

Data transformations can also modify the structure of the data itself, such as reorganizing it into tables, which can be used to compare or classify. This step is more steered toward changing the representation of data and allowing for a structured look on them.

Third, the created structures can be used to generate new derived values. This can mean extracting aggregations or ranges from the organized data structures. These transformations consume and convert the structured organization of data to generate higher-level information that is not visible in the raw data or derived values alone.

Finally, transformations can derive new structures from existing one, which in turn provide a different perspective or highlight different relationships in the data. Card et al. give an example of this, where they take a table of ranges and transform it into a binary table. In this case, existing ranges are used to define criteria for binarization, where each cell in the binary table represents whether a specific data point falls in a given range or category.

In essence, all these transformations try to ensure that the data is in an "idiosyncratic format" [32]. This means they are optimized for subsequent stages of the visualization pipeline.

### Visual Mapping

*Visual mapping* is the core of the visualization process. In this stage, the derived structures and values within data tables are mapped to visual glyphs using a specific function $F$ [33]. This function takes the structured data as input and produces a visual representation as output. The goal is to create a visual form that users can easily interpret.

A well-designed visual mapping function must be computable, meaning it can be executed algorithmically. It also has to be invertible, allowing users to reconstruct data from the visual representation. Additionally, the function or its inverse must be understandable to users and should aim to minimize the cognitive load required to interpret the visual representation [33].

According to Card et al. [32], the visual mapping process involves two sub-steps: mapping data entities to visual *glyphs*[2] and assigning data attribute values to the visual properties of these glyphs (e.g., size, color, and shape) [32]. Since this step builds on several aspects of human perception, concepts like visual structures and Gestalt principles are crucial for making these visualization effective.

### View Transformation

After the visual forms are created, they are embedded into views. Views display these visual forms on the screen and enable various transformations such as zooming, panning, and rotating to adjust the perspective [33].

View transformation involves affine view transformations, which allow the user to focus on different parts of the data. Multiple views might be created to show different aspects of the data, and synchronization ensures that interactions in one view are reflected in other related views. The goal is to present the visual forms in an accessible and interpretable way, focusing on better understandability and providing some degree of exploration.

### Interaction

*Interaction* is a crucial part of the visualization pipeline, enabling users to manipulate the visualization and generate meaning through exploration. Common interaction techniques include selecting, filtering, linking, and rearranging or remapping [33].

**Selecting** involves marking specific data entities or subsets to view detailed information or perform further analysis.

**Filtering** reduces the quantity of data in the display, allowing users to focus on the information of interest.

**Dynamic queries** enable users to adjust query parameters and see filtered results in real-time.

**Linking** relates information between multiple views to show users how selections in one view correspond to data in another one.

[2] In a more general sense, Ward [34] defines glyphs as a "graphical entity" that encodes additional information found in the data with the help of visual cues.

**Rearranging and remapping** provides tools for changing the visual mapping or choosing different mappings to explore various aspects of the data.

The underlying reasoning for the effectiveness of interactivity can generally be attributed to a concept referred to as *Human in the Loop* [35]. This approach underlines the vital role of human judgment and expertise in the analytical process. Human in the Loop systems involve continuous feedback between the user and the system, where users iteratively refine data analysis and visualization based on their evolving understanding of the data at hand.

Although having its roots in fields such as artificial intelligence and machine learning, where human oversight is essential to guide a system's performance [35], Heer and Shneidermann specifically pronounced interactivity and exploration as a central aspect of supporting the sense-making process within visualization [36].

### 2.2.4 Data Visualization Techniques

Different types of data require different visualization techniques to appropriately convey the underlying information. Understanding the nature of the data therefore heavily guides the selection of appropriate visualization methods [1].

There are several ways to categorize visualization techniques. The most common approaches are distinctions by:

- **Problem:** This approach categorizes visualizations based on the specific problem they aim to solve, such as communication, exploration, or confirmation.
- **Data Type:** Visualizations can be categorized based on the type of data they handle, such as categorical, numerical, or time-series data.
- **Dimensions:** This refers to the number of dimensions (univariate, bivariate, multivariate) represented in the data visualization.
- **Data structure:** Visualizations can be classified based on the structure of the data (linear, hierarchical, circular)
- **Interaction:** This category separates visualizations based on the level of interactivity they offer, such as static, interactive, or dynamic visualizations.

In the following sections, we will further examine data visualization techniques classified by *dimension* and how each of these techniques highlight different characteristics in the data.

### Univariate Data

Univariate data consists of observations on a single variable, and its visualization usually aims to understand the data's distribution, central tendency, and dispersion [1]. Central tendency describes the center or typical value of the data. The most common measures include the (weighted) mean, median, and mode [7]. Although each of those measures focuses on highlighting different subtleties in a given dataset, all of them help summarize aspects of the data with a single representative value.

As mentioned earlier, dispersion (variance) quantifies how much the data values deviate from the central tendency. Typically, certain variability in data is highlighted using measures like ranges, percentiles, and standard deviations [37].

The distribution of univariate data shows how variable values are spread, which in larger populations is generally assumed to be normally distributed. However, this must not always be the case. Visualizations in this regard often then focus on illustrating data points' relative or cumulative frequencies. As shown in Figure 2.8, various techniques are employed to visualize univariate data effectively [38]:

**Histograms** are graphical representations of data distributions, where the data is divided into bins of equal width, and the height of each bin reflects the frequency of data points within it. This technique helps understand the shape of the data distribution, whether it is skewed, uniform, or normal [37].

**Bar charts** are another popular method for displaying univariate data, especially if dealing with categorical data. Each category is represented by a bar, with the length of the bar corresponding to the frequency of observations in that category.

**Box plots,** also known as box-and-whisker plots, summarize the data distribution through five main statistics: the minimum, first quartile, second quartile (median), third quartile, and maximum [39].

### Bivariate Data

Bivariate data involves observations on exactly two variables, and its analysis aims to uncover relationships between them. As outlined in the previous sections on correlation analysis, the primary goal is to identify the nature and strength of associations. There are various visualization techniques that are used to visualize bivariate data [38][7], as illustrated in Figure 2.9.

**Scatterplots** are widely used to display the relationship between two variables. Each point on the scatterplot represents a single observation, with its position determined by the values of the two variables $X$ and $Y$. Scatterplots can reveal patterns such as linear or non-linear relationships and are particularly suited for visualizing clusters or outliers [37].

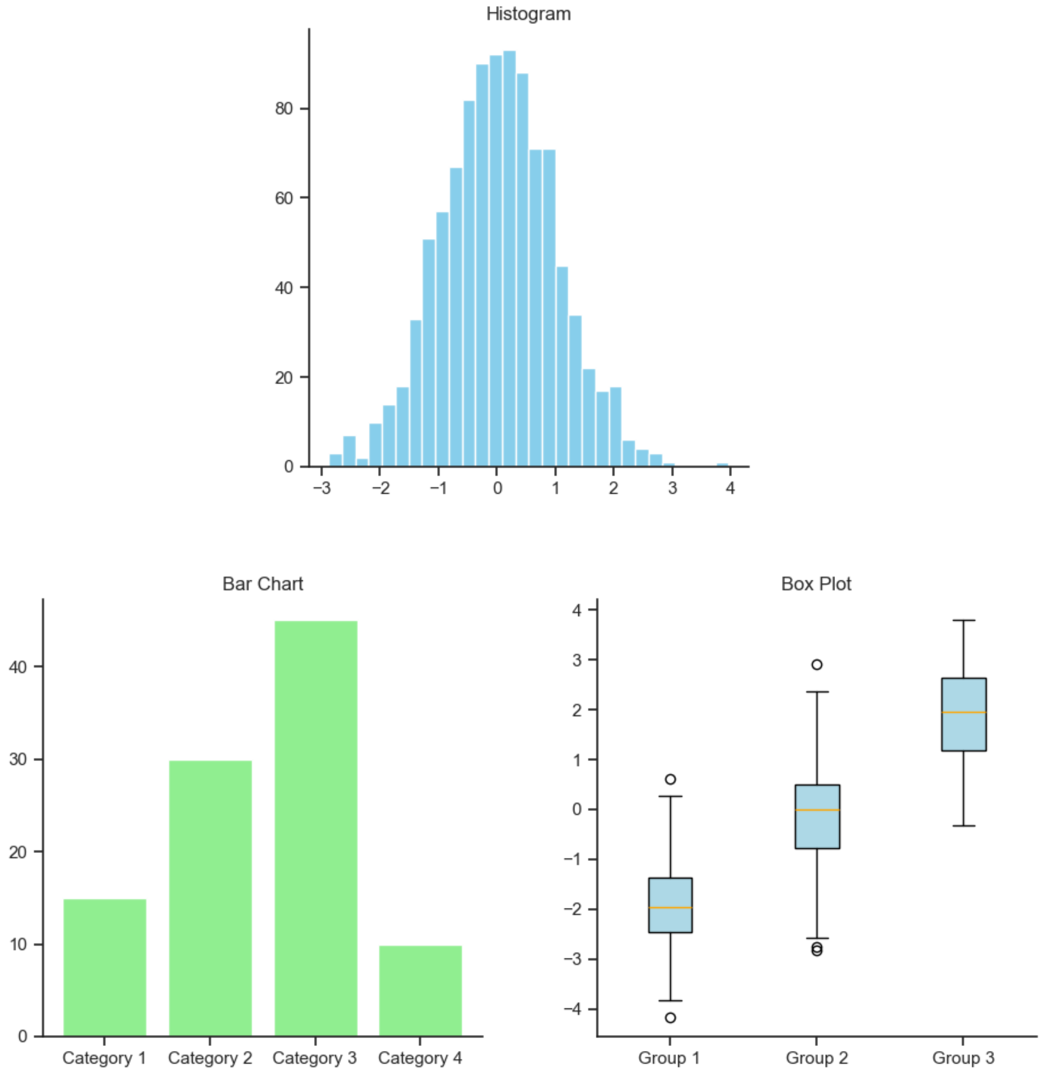


Figure 2.8: Data visualization techniques for univariate datasets using three common methods: histograms, bar charts, and box plots. The histogram (top) displays the frequency distribution of continuous data, revealing the shape and spread of the data. The bar chart (bottom left) is used for categorical data, where each bar represents the frequency of a given category. The box plot (bottom right) summarizes the data's central tendency and variability.

**Line graphs** are ideal for visualizing the relationship between two variables if one of them is a time variable [38]. This technique is commonly used to display trends over time, with data points connected by lines to show the time-series progression of values.

**Heatmaps** use a spectrum of color to represent the values of two variables within a matrix. This technique is effective for dealing with large amounts of data, as it provides a clear visual representation of areas with higher or lower concentrations of values [37].

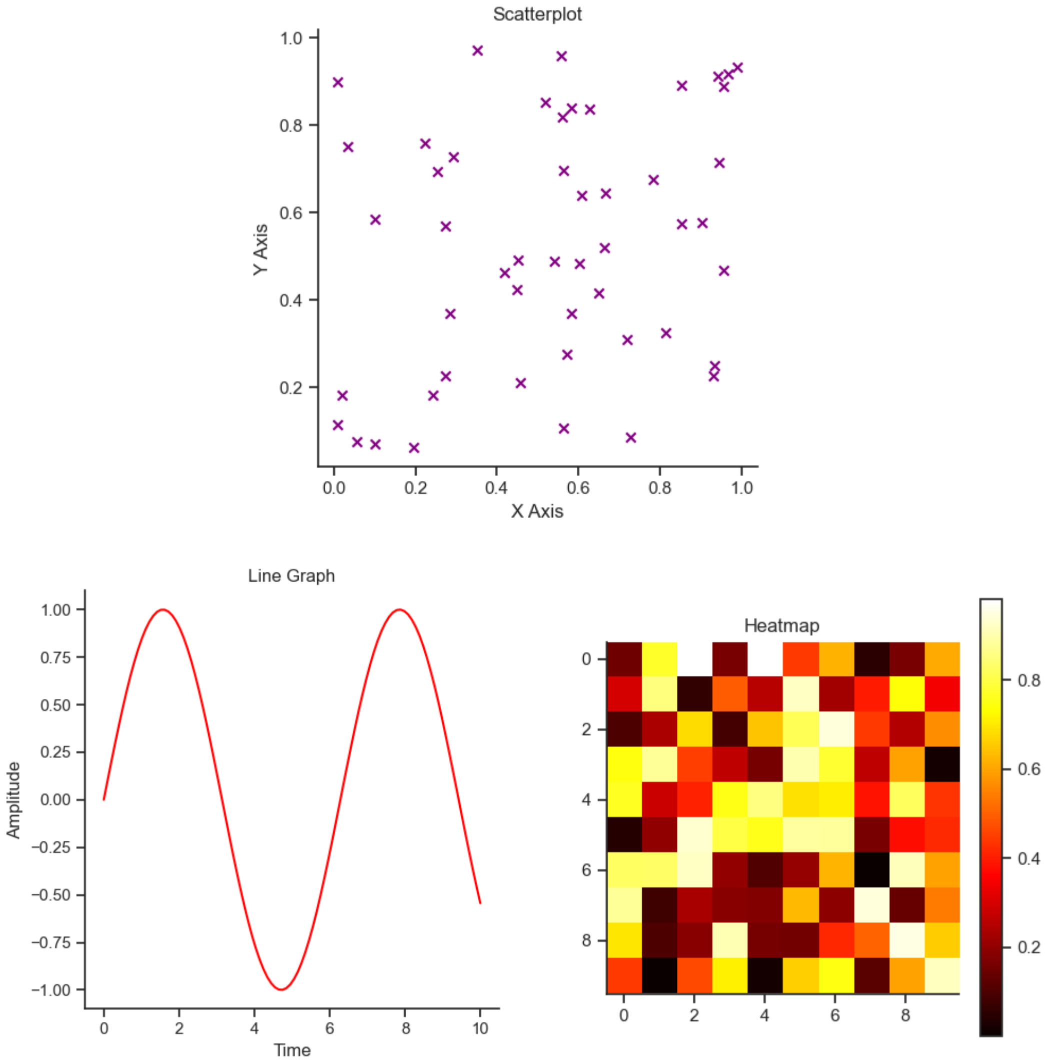


Figure 2.9: Examples of scatterplots, line graphs, and heatmaps as key data visualization techniques for bivariate data. The scatterplot (top) visualizes the relationship between two variables. The line graph (bottom left) plots time-series data, showing trends over time by connecting data points with lines. The heatmap (bottom right) presents values of two variables across a grid, effectively displaying color-encoded patterns.

### Multivariate Data

Multivariate data involves more than two variables, and visualizing such data requires techniques that can display complex relationships and interactions among them. If working with multivariate data, the goal is generally to uncover details that are not visible by examining each variable in isolation.

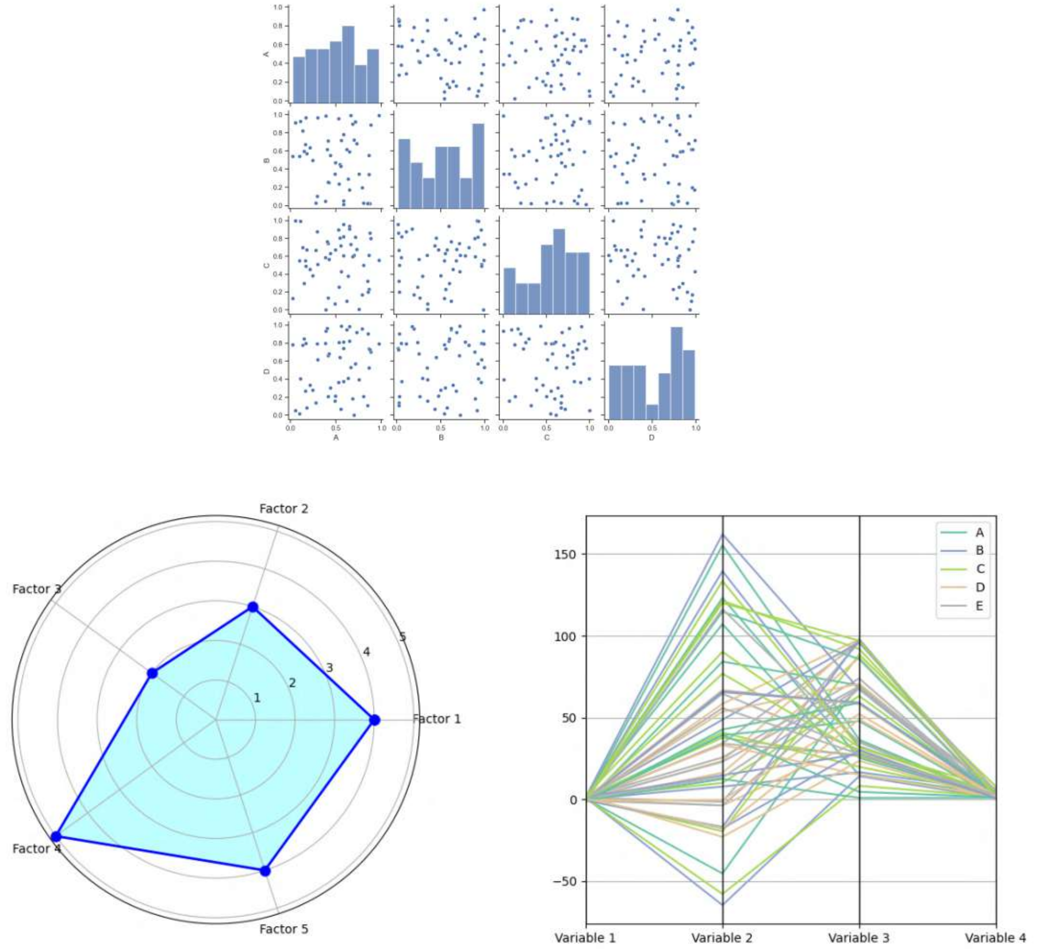


Figure 2.10: Examples for multivariate datasets employing scatterplot matrices, radar charts and parallel coordinates. The scatterplot matrix (top) displays pairwise scatterplots for multiple variables. The radar chart (bottom left) visualizes data across multiple axes. The parallel coordinates plot (bottom right) shows relationships between multiple variables across different axes.

**Scatterplot matrices,** also known as pair plots, display multiple scatterplots in a grid format, where each plot shows the relationship between a distinct pair of variables. This method helps identify pairwise relationships and potential correlations among several variables simultaneously. This type of visualization can quickly become overwhelming if increasing the number of plots to compare [37].

**Radar charts,** stardinates [40], or star plots represent multivariate data on a circular layout with each variable corresponding to a *spoke*, meaning rays being projected from the center [41]. The length of each spoke reflects the variable's value and the points are connected to form a polygon. Radar charts are usually used for comparing the profiles of different observations across multiple variables.

**Parallel coordinates,** are a method of visualizing high-dimensional data by plotting each variable on a separate parallel axis. Each observation represents a polyline crossing each axis at the corresponding value [1]. This becomes important for visually gauging correlations and patterns across multiple variables [1]. Parallel coordinates are the central visualization technique explored in this thesis and are further explained in Section 2.3.

### Hierarchies and Structures

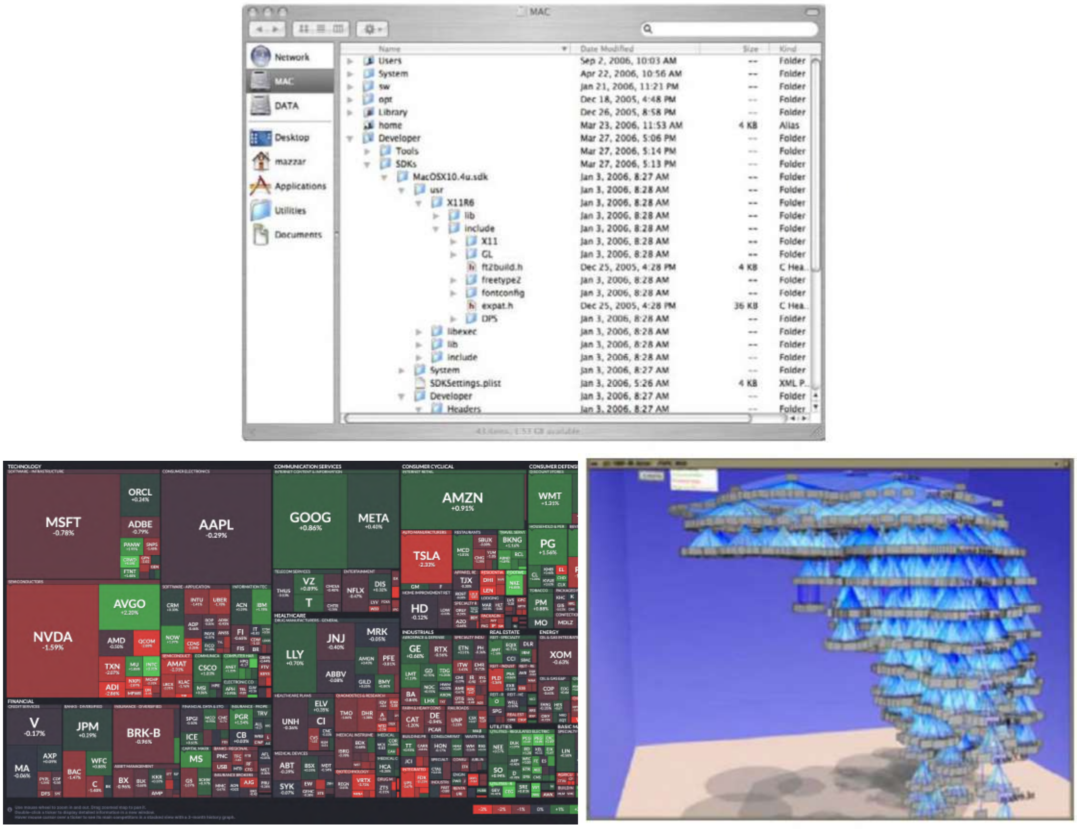
Hierarchical data consists of elements organized in parent-child structures. These can be visualized in various forms, as illustrated in Figure 2.11. They generally aim to represent various levels of relationships from the root to the leaves. This organization reflects nested relationships where higher-level elements encompass lower-level ones.

**File systems** are a typical example of hierarchical data, where files and folders are organized in a tree-like structure. Visualizing file systems helps users navigate and manage data by showing the hierarchical relationships and organization of files. This metaphor is very common mental model and can be found across all nearly all modern operating systems.

**Cone trees** are 3D representations of hierarchical data, where the hierarchy is visualized in a cone shape, with the root at the top and leaves at the base. Cone trees provide an intuitive way to explore large hierarchical structures by allowing users to rotate and zoom into different parts of the hierarchy.

**Botanical trees** represent hierarchical data using a tree metaphor, where branches represent the hierarchy and leaves represent the individual elements. This visualization method effectively shows the overall structure and relationships within the hierarchy.

**Treemaps** are space-filling visualizations that display hierarchical data as nested rectangles. Each rectangle represents an element in the hierarchy, with its size proportional to a specified attribute, such as file size or value. Treemaps are suitable for comparing the relative sizes of elements within the hierarchy and for identifying patterns and outliers.



| Name | Date Modified | Size | Kind |
|---|---|---|---|
| Users | Sep 2, 2006, 10:03 AM | -- | Folder |
| System | Apr 22, 2006, 10:56 AM | -- | Folder |
| sw | Jan 21, 2006, 11:21 PM | -- | Folder |
| opt | Dec 18, 2005, 4:48 PM | -- | Folder |
| Library | Dec 26, 2005, 8:58 PM | -- | Folder |
| home | Mar 23, 2006, 11:53 AM | 4 KB | Alias |
| Developer | Mar 27, 2006, 5:06 PM | -- | Folder |
| Tools | Mar 27, 2006, 5:14 PM | -- | Folder |
| SDKs | Mar 27, 2006, 5:13 PM | -- | Folder |
| MacOSX10.4u.sdk | Jan 3, 2006, 8:27 AM | -- | Folder |
| usr | Jan 3, 2006, 8:28 AM | -- | Folder |
| X11R6 | Jan 3, 2006, 8:28 AM | -- | Folder |
| lib | Jan 3, 2006, 8:28 AM | -- | Folder |
| include | Jan 3, 2006, 8:28 AM | -- | Folder |
| X11 | Jan 3, 2006, 8:28 AM | -- | Folder |
| GL | Jan 3, 2006, 8:28 AM | -- | Folder |
| ft2build.h | Dec 25, 2005, 4:28 PM | 4 KB | C Hea... |
| freetype2 | Jan 3, 2006, 8:28 AM | -- | Folder |
| fontconfig | Jan 3, 2006, 8:28 AM | -- | Folder |
| expat.h | Dec 25, 2005, 4:28 PM | 36 KB | C Hea... |
| DPS | Jan 3, 2006, 8:28 AM | -- | Folder |
| libexec | Jan 3, 2006, 8:28 AM | -- | Folder |
| lib | Jan 3, 2006, 8:28 AM | -- | Folder |
| include | Jan 3, 2006, 8:28 AM | -- | Folder |
| System | Jan 3, 2006, 8:27 AM | -- | Folder |
| SDKSettings.plist | Jan 3, 2006, 5:26 AM | 4 KB | XML P... |
| Developer | Jan 3, 2006, 8:27 AM | -- | Folder |
| Headers | Jan 3, 2006, 8:27 AM | -- | Folder |

Figure 2.11: Hierarchical data structures represented with three different techniques. The top image depicts a traditional file system hierarchy, showcasing the nested folder structure in a directory tree. The bottom left image displays a treemap that visualizes stocks in the S&P 500 index with rectangle sizes based on market capitalization. On the bottom right, the image displays a cone tree, a three-dimensional visualization method used to present complex hierarchical data with a focus on depth and layering. Images taken from Mazza [1] and Finwiz [42].

### Networks and Graphs

Network data often consists of entities (nodes) and the relationships (edges) between them. Visualizing network data helps understand the structure and dynamics of the relationships within them. The visualization possibilities for this technique are mainly rooted in network and graph theory and are often used for modeling complex social phenomena [43] or networks [44]. As showcased in Figure 2.12, these techniques usually employ different visual techniques like annotations or expansion to further contextualize information within them.

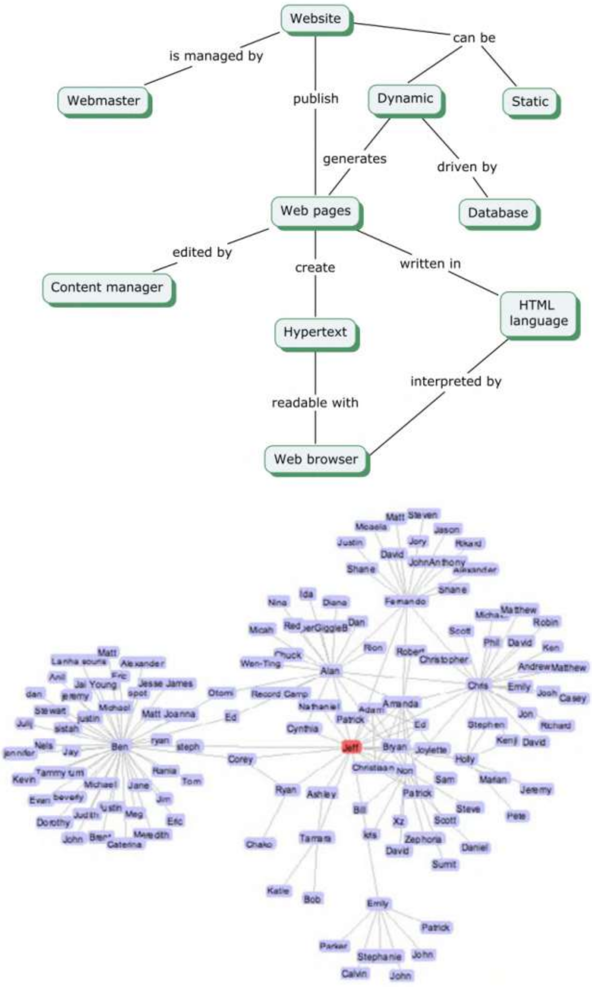


Figure 2.12: Examples of data visualization techniques for network data. A traditional node-link diagram is depicted at the top and shows entities as annotated nodes and their relationships as edges. The bottom image illustrates a force-directed tree layout in a social network, with nodes radiating out from central roots. Images taken from Mazza [1].

Adjacency matrices are another commonly used method for representing graphs (see 2.13). In an adjacency matrix, the rows and columns represent nodes, and the matrix entries indicate whether there is an edge between corresponding nodes [45]. This matrix-based approach results in a more compact, tabular view, which can be beneficial for visualizing dense graphs [46]. Alternatively, adjacency lists are also frequently used to represent network relationships. An adjacency list stores the same information as a matrix but in a more space-efficient way by keeping track of each node and the nodes it is directly connected to. This type of visualization is well suited for sparse graphs.

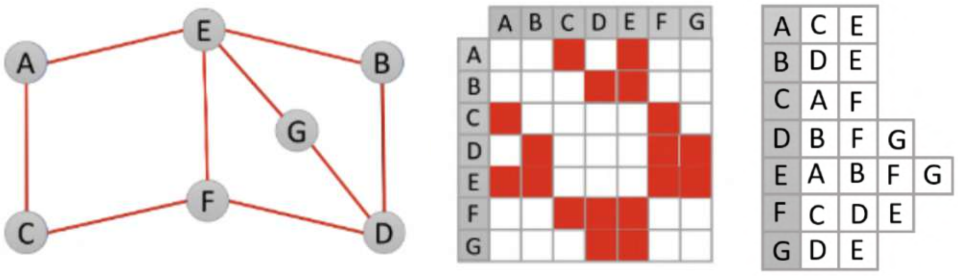


Figure 2.13: Depiction of a relational mapping from a graph (left) to an adjacency matrix (center) and adjacency list (right). Images taken from Kale et al. [45].

**Knowledge graphs** focus on reflecting semantic relationships. They organize information by connecting nodes, which represent entities or concepts, through labeled edges that define the relationships between them [47]. These graphs are particularly valuable for organizing and representing complex relationships and interdependencies derived from various sources, such as text data or databases. Figure 2.14, for example, shows a knowledge graph depicting a fraction of the interconnected page network on Wikipedia.

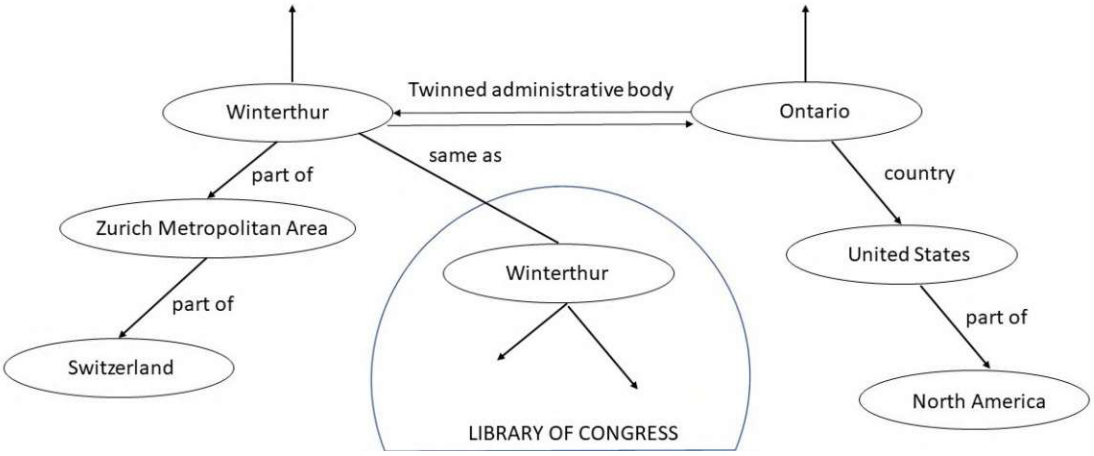


Figure 2.14: An example of a knowledge graph as generated from the Wikipedia page on the city of Winterthur in Switzerland. Images taken from Chaudhri et al. [47].

**Concept maps and mind maps** serve as visual tools for organizing information, but compared to knowledge graphs, differ in scope and structure. Concept maps highlight relationships between ideas, where the connections are also labeled to explain how different concepts relate to one another. However, unlike knowledge graphs, concept maps are generally used for educational or ideation purposes, showing how broader concepts are connected [48]. Mind maps, on the other hand, start with a central idea and branch into related subtopics, using lines and images to organize information hierarchically.

### 2.2.5 Geographic Representations

Geographic representations visualize spatial data by mapping the relationships between data and locations [49]. Geographic Information Systems (GIS) can overlay various data layers on a map or other geographic structures to provide enhancing information based on geographic proximity (see Figure 2.15).

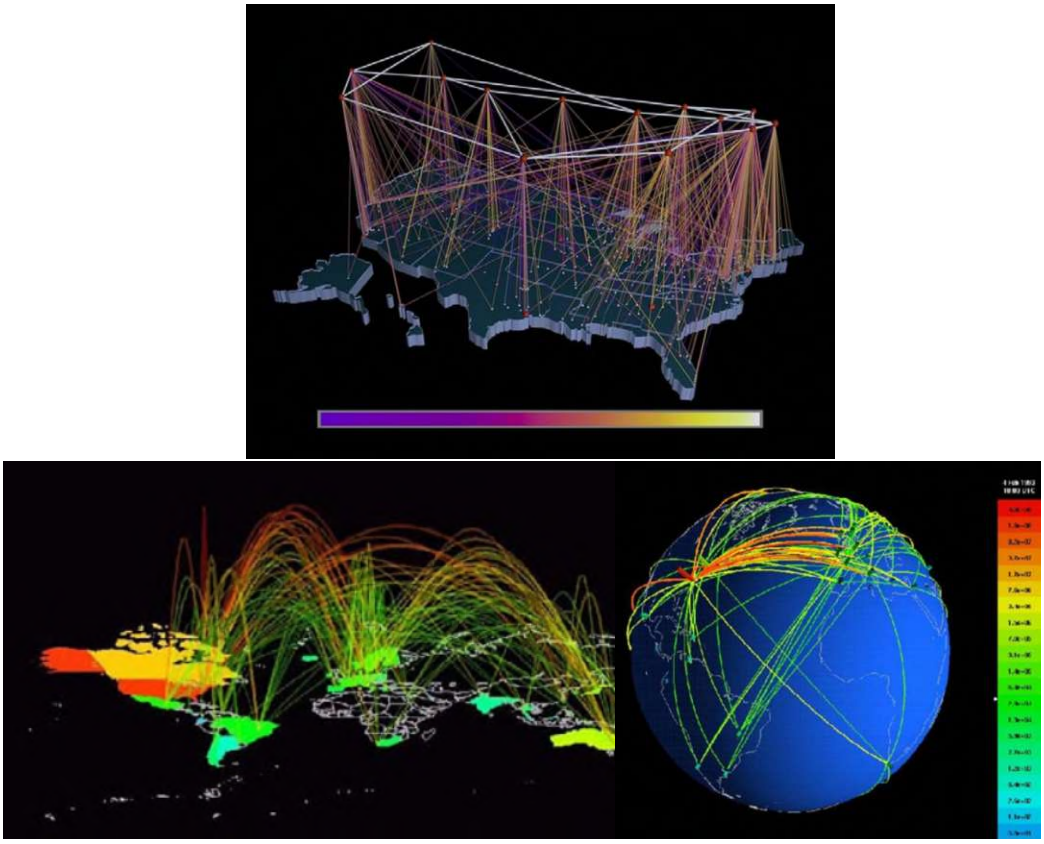


Figure 2.15: Data visualization techniques for GIS data. The top image depicts a network of connections across the US, highlighting the relationships between different locations through a 3D representation. The bottom left and right images both show visualizations of global internet network traffic using color gradients. Images taken from Mazza [1].

**3D graphs** allow for the visualization of network data in three dimensions, providing a more immersive perspective on network structures. This technique is suitable for understanding complex networks with many interconnected nodes and edges, as it reduces overlap and enhances the clarity of relationships.

## 2.3 Parallel Coordinates

Parallel coordinates provide a unique perspective on representing complex datasets on a two-dimensional plane (see Figure 2.16). Unlike traditional Cartesian coordinates [37], which struggle to represent more than three dimensions effectively, parallel coordinates can handle numerous dimensions simultaneously, making them ideal for multivariate data analysis [1].

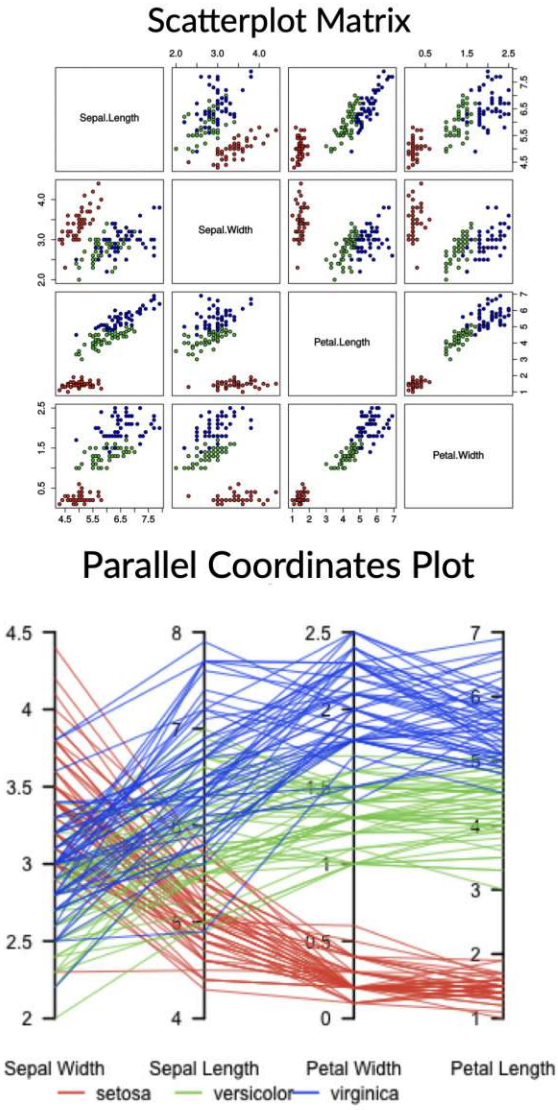


Figure 2.16: Comparative visualization of a multi-dimensional dataset using scatterplot matrices (top) and parallel coordinates (bottom). Note that both plots visualize the same dataset, but provide two distinctly different ways of exploration. Images taken from Pezzotti [50].

### 2.3.1 Mathematical Framework

First conceptualized by Inselberg in the 1980s [1][51], the theoretical foundations of parallel coordinates date back to the work of the several key figures, including d'Ocagne [52], Gannet [53] and Guerry [54]. They laid the groundwork for this visualization technique by introducing the concept of data representation using a system of coordinates that allowed the simultaneous display of variables on parallel axes.

From a mathematical standpoint, when talking about parallel coordinates, we consider a dataset with $n$ observations and $m$ variables. Each observation can be represented as a vector in an $m$-dimensional space [51]:

$$\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m}) \quad \text{for} \quad i = 1, 2, \dots, n$$

### 2.3.2 Geometric Construction

Each vertical axis corresponds to one of the $m$ variables, arranged in parallel and equidistant from each other. The position of the axes is critical for the interpretation of the data, as adjacent axes can only easily reveal relationships between themselves [55]. To ensure comparability across variables with different units or scales, each observation $x_j$ is further normalized to a standard scale [56], typically to [0, 1], using *Min-Max Scaling* [57]:

$$x'_{i,j} = \frac{x_{i,j} - \min(x_j)}{\max(x_j) - \min(x_j)}$$

where $\min(x_j)$ and $\max(x_j)$ are the minimum and maximum values of the $j$-th variable, respectively. For each observation $x_i$, a polyline is drawn so that it intersects the vertical axis $j$ at the heigh corresponding to the normalized value $x'_{i,j}$. The ployline then connects the points $(j, x'_{i,j})$ for $j = 1, 2, \dots, m$.

### 2.3.3 Correlation Analysis

Parallel coordinates are especially useful for visually assessing correlations between variables. The interpretation of these plots is typically limited to adjacent axes. Only correlations between consecutive variables are directly observable, as the polylines connecting these axes show whether two variables move in a similar or opposite direction.

As illustrated in 2.17, the nature and strength of correlations can be inferred by observing the patterns manifested between adjacent axes:

**Positive Correlation:** If two variables are positively correlated, the polylines between the corresponding axes tend to be parallel and move in the same direction. This means that as one variable's value increases, the other variable's value also increases.

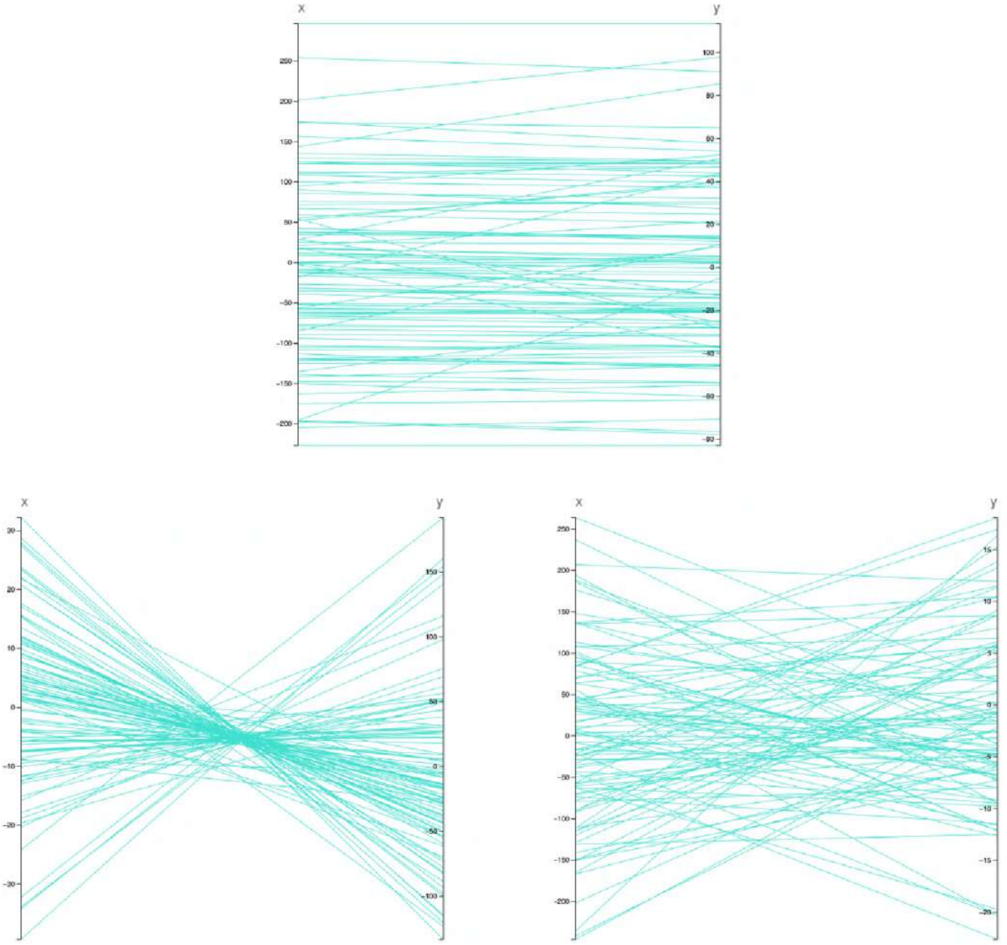


Figure 2.17: Correlation patterns manifested in parallel coordinates. The top plot shows a strong positive correlation ($r \approx 0.96$), where lines between the axes are mostly parallel and move in the same direction. The bottom left plot represents a strong negative correlation ($r \approx -0.96$), in which lines cross, indicating an inverse relationship between the variables. The bottom right plot shows no correlation ($r \approx -0.02$) with lines not following a clear pattern.

**Negative Correlation:** A negative correlation is indicated by distinct lines that cross each other and run in opposite directions. This pattern suggests an inverse relationship.

**Missing correlation:** If there is no apparent correlation between two variables, the lines between the corresponding axes appear more random, with no consistent pattern of parallelism or crossing. This randomness suggests that changes in one variable do not systematically relate to changes in the other one.

### 2.3.4 Clusters and Outliers

Clusters of observations appear as groups of polylines that follow similar paths across the axes. These clusters indicate subsets of the data with similar profiles [55]. In real-world use cases, detecting clusters can help identify segments in the data that share common characteristics, such as demographic groups or customer segments.

On the other hand, outliers are identifiable as polylines that diverge significantly from the majority. These outliers represent observations that noticeably differ from the average data pattern, which can also be of particular interest, depending on the use case.

### 2.3.5 Visual Clutter

Plotting an arbitrary number of observations in a single visualization also comes with challenges. Traditional parallel coordinates face issues with over-plotting, where too many overlapping data points obscure important details in the plot, making it overall difficult to read. This visual clutter can significantly reduce the effectiveness of the visualization, requiring specialized techniques to manage the complexity [58].

Over time, researchers proposed several ways of counteracting this issue. Early optimizations for this problem especially focused on the use of curved lines instead of straight lines for visually seperating parts of the data in order to improve readability.

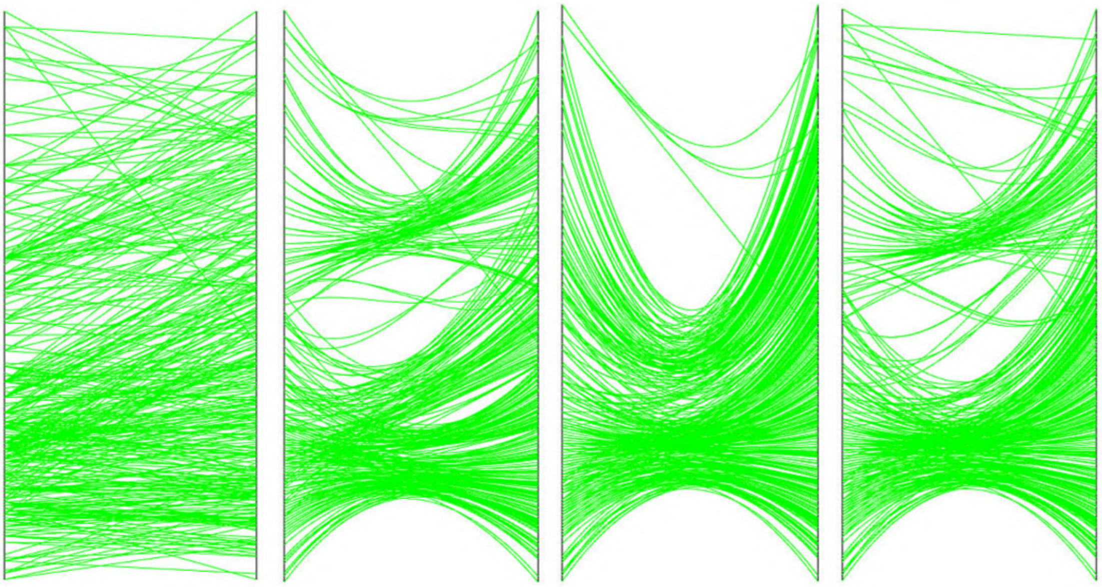


Figure 2.18: Depiction of different clustering weights applied to parallel coordinates. The axes represent different variables, with each line connecting corresponding data points across the variables. The varying curvature of the lines indicates the degree of clustering applied. Image taken from Zhou et al. [59].

Zhou et al. [59] proposed a novel visual clustering algorithm based on energy minimization to reduce clutter by deforming and bundling lines as clusters. Specifically, the method optimizes the arrangement of curved lines through minimizing their curvature and maximizing the parallelism of adjacent edges, leading to more organized visual clusters. This geometric bundling is performed during plotting, allowing users to control the clustering level dynamically through weight parameters in the energy function (see Figure 2.18).

Furthermore, the authors incorporated color and opacity enhancements to highlight different aspects of the clustered data. By assigning color and opacity according to the local density of polylines, users can distinguish clusters more effectively. The authors demonstrated the effectiveness of their approach through experiments on several representative datasets, showing how visual clutter is reduced.

Johansson et al. [60] proposed a solution for over-plotting by using high-precision textures to represent clusters, preserving visualization clarity by adjusting intensity ranges based on data overlap. A key part of their approach is the use of transfer functions, which map data values to opacity. By adjusting these functions, different aspects of the data can be highlighted.

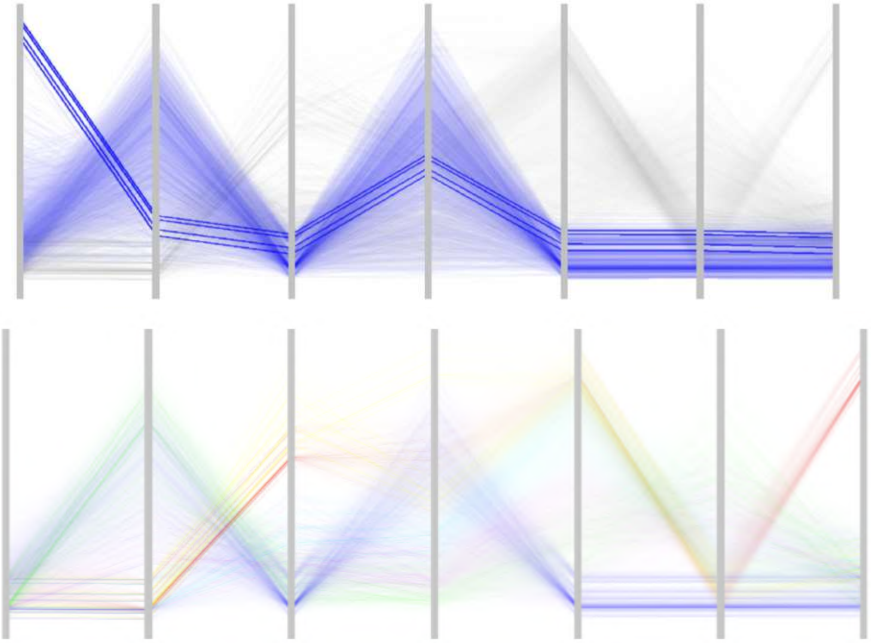


Figure 2.19: Managing over-plotting using different transfer functions. The first image uses a square root function to bring out outliers in areas where data points are tightly clustered. The second image employs a linear function, which simplifies the visualization and makes the overall data structure more visible by reducing the intensity in more populated regions. Images taken from Johansson et al. [60].

For example, a linear transfer function helps provide a clear overview by reducing the visual complexity in dense areas, while a square root transfer function emphasizes outliers, making them stand out even in regions with many data points. They also provide the flexibility for users to create custom transfer functions, allowing more control over what features of the data are emphasized. The effectiveness of this method is illustrated in Figure 2.19, which shows how two different transfer functions are applied to enhance the visibility of structures in the clustered data.

### 2.3.6 Interaction

Interaction techniques play a crucial role in enhancing the utility of parallel coordinates, enabling viewers to explore and analyze multivariate data more effectively [36]. The following techniques are commonly implemented with this specific type of visualization [61][58][55]:

**Brushing** is one of the most widely used interaction techniques. It allows users to select and highlight specific ranges of data on one or more axes, making it easier to focus on particular subsets of the data.

**Axis Reordering** is another important interaction technique. The order of axes in a parallel coordinates plot can significantly influence the visibility of patterns and relationships in the data. By allowing users to interactively reorder axes, they can explore different perspectives and uncover details or structures that may not be visible with a static axis order.

**Axis Inversion** involves flipping the direction of an axis, which can help in identifying patterns, such as negative correlations that may not be easily visible in the default orientation. This technique can be fairly useful for analyzing datasets with both positive and negative correlations.

**Density Adjustments** increasingly enhance the readability of parallel coordinates, especially if dealing with large datasets. By using varying degrees of semi-transparent lines or density-based rendering, users can better understand the distribution and concentration of data points, reducing visual clutter and highlighting prominent patterns.

**Coordinated Multiple Views** allow for the integration of parallel coordinates with other types of visualizations. An approach, known as brushing-and-linking, enables users to select data points in one view and see the corresponding points highlighted in other views, providing a better understanding of the data across multiple dimensions.

### 2.3.7 Case Studies

In this section, we focus on practical applications of parallel coordinates through detailed case studies. These studies highlight the versatility and effectiveness of parallel coordinates in various domains, as well as showcasing their strengths and weaknesses compared to other visualization methods.

Lanzenberger et al. [40] presented a detailed comparative study of two information visualization techniques: stardinates and parallel coordinates. Their research aimed to assess the effectiveness of both methods in visualizing complex datasets (see Figure 2.20). The authors conducted an empirical study with 22 participants to evaluate the techniques based on several criteria: time taken for interpretation, accuracy of information extracted, and subjective feedback from the users. The study involved two distinct visualization tasks.

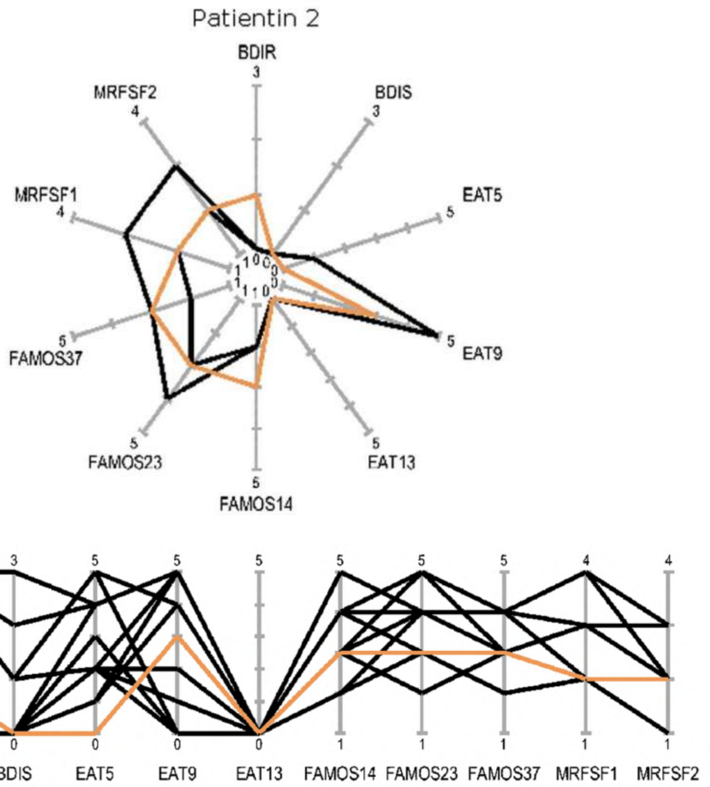


Figure 2.20: Comparison of the two techniques used for visual estimation tasks. The top image shows psychotherapeutic data using stardinates, highlighting patient ID 2 in orange. The bottom image presents the same data using parallel coordinates. Images taken from Lanzenberger et al. [40].

In the first task, participants were asked to identify whether there had been a collision between aircrafts, which acted as a basic test of their ability to interpret data. This scenario was selected for its simplicity, allowing participants to become familiar with both visualization techniques without needing extensive background knowledge.

The second task, which involved psychotherapeutic data, was more complex. Here, participants needed to identify patterns and changes in patients' conditions over time, a more challenging task given the participants' lack of prior familiarity with this specific dataset, increasing the difficulty of the analysis.

To guide their evaluation, the researchers focused on several research questions, including how easily users could find information at first glance, their ability to identify critical details, and the extent to which the visualization techniques supported them in their decision.

The results of the study revealed that stardinates were more effective for the in-depth interpretation of highly structured data, leading participants to report on more detailed observations. Participants using stardinates were better at spotting nuanced changes and relationships in the psychotherapeutic data. On the other hand, parallel coordinates excelled in offering an immediate understanding of the data, enabling viewers to quickly extract information at a glance. These findings suggest, that parallel coordinates are especially useful if the goal is to gain a broad overview of complex data without diving into finer details.

Bok et al. [62] introduced the parallel histogram plot, a novel visualization technique designed to enhance traditional parallel coordinates (see Figure 2.21). Parallel histogram plots integrate histograms with parallel coordinates. They developed this technique to provide an enhanced overview of data – without the scalability and cluttering issues that usual parallel coordinates suffer from. Each histogram displays a color-coded data ranking in relation to a selected attribute, making it easier to examine relationships between attributes even when displayed far apart.

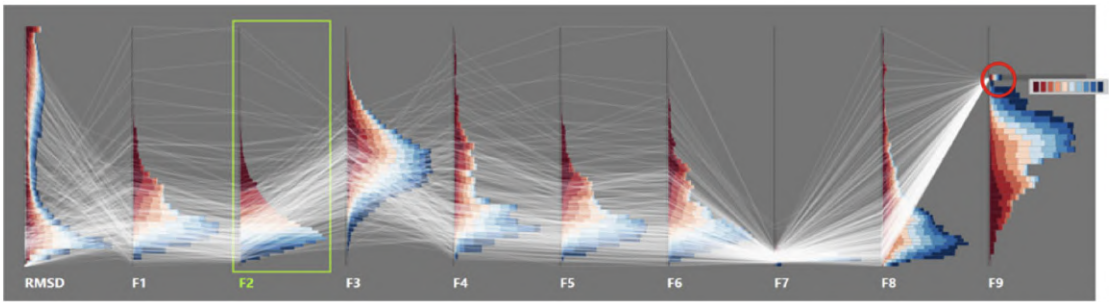


Figure 2.21: Example of a parallel histogram plot, visualizing histograms with parallel coordinates to overlay correlations across multiple attributes. Correlations are inferred by comparing color patterns across axes: similar colors indicate positive correlations while contrasting colors suggest negative correlations. Images taken from Bok et al. [62].

The authors presented real-world applications of this technique and a controlled user study to evaluate its performance in estimating correlations between attributes. The authors demonstrated that viewers using the histogram overlays consistently performed high in correlation estimation tasks, highlighting its effectiveness in addressing parallel coordinates' limitations.

Li et al. [63] conducted a comparison between scatterplots and parallel coordinates, as shown in Figure 2.22, to evaluate their effectiveness in visual correlation assessment. Participants were tasked with estimating correlations while varying the visualization method, sample size, and display time.

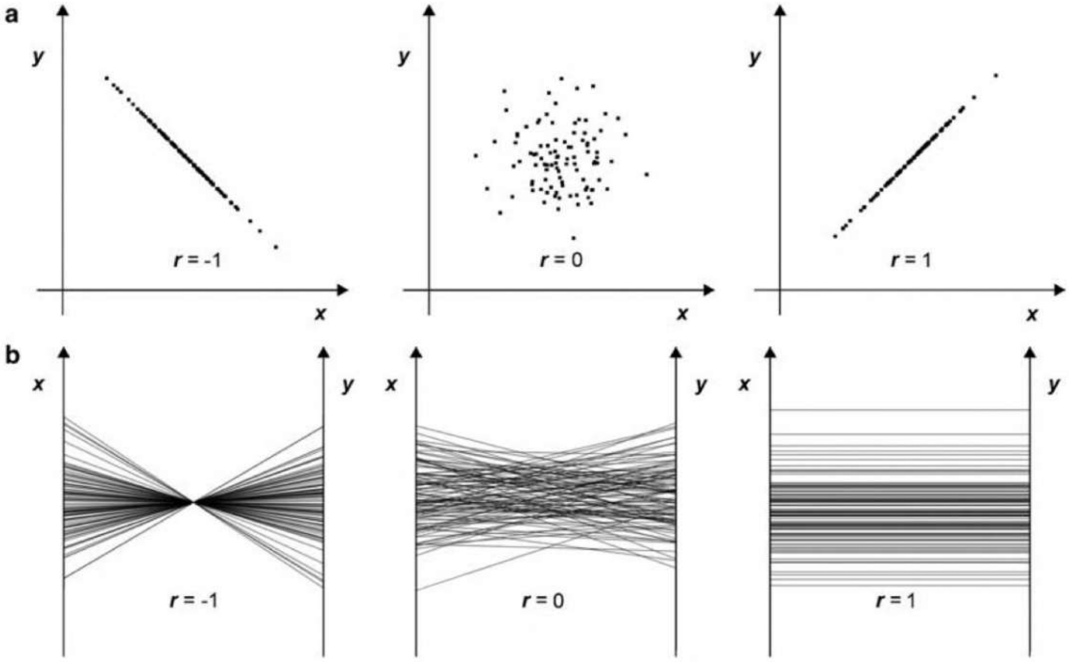


Figure 2.22: Comparison of correlation patterns in scatterplots and parallel coordinates. Scatterplots (top) visualize correlations through linear trends. Parallel coordinates (bottom) indicate correlations by patterns in line sets between adjacent axes. Image taken from Li et al. [63].

The results showed that users were more effective at identifying different levels of correlation using scatterplots compared to parallel coordinates Scatterplots provided higher accuracy and consistency, allowing users to make more reliably estimate correlations across different sample sizes and visualization durations. In contrast, parallel coordinates often led to an underestimation of correlation, particularly for weaker correlations, and users exhibited greater variability in their judgments. This case study was pivotal because it systematically evaluated the effectiveness of parallel coordinates against scatterplots, providing an understanding of their strengths and limitations in visual correlation analyses.

## 2.4 Aspect Ratio

The aspect ratio, defined as the ratio of the width to the height of a visual representation [4], has a strong impact on how information is perceived and interpreted. As shown in Figure 2.23, aspect ratio can introduce various phenomena that affect the readability, interpretability, and overall effectiveness of visual data presentations.

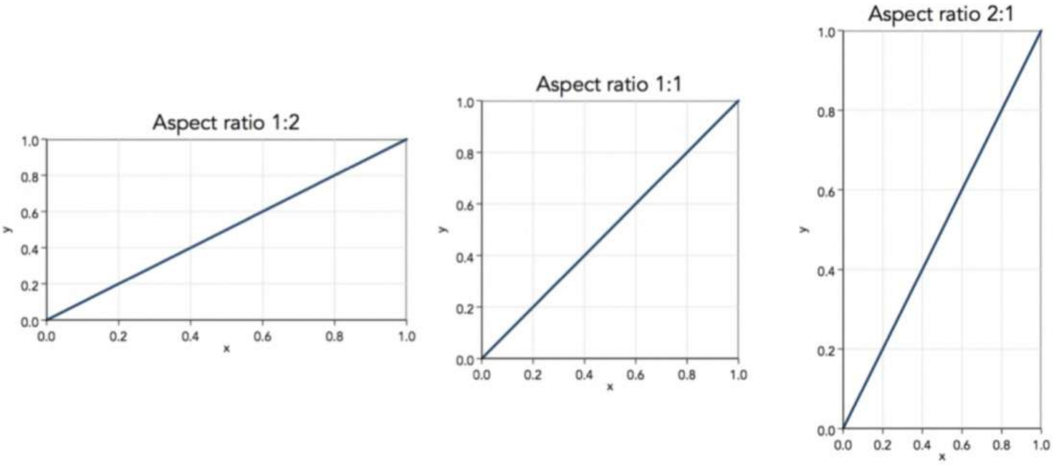


Figure 2.23: Visualization of a linear slope under three different aspect ratios: 1:2, 1:1, and 2:1. In the 1:2 aspect ratio (left), the slope appears to have a gentle incline, whereas, in the 2:1 aspect ratio (right), the same slope appears much steeper. Image taken from Christodoulou [4].

The aspect ratio of a plot can significantly influence our perception of data. In line charts, altering the aspect ratio by manipulating the axes can change visual cues within the plot. Extending vertical axis limits creates a taller aspect ratio with steeper slopes, while shrinking them results in a wider aspect ratio with shallower slopes [5].

Aspect ratios also influence the perception of trends and patterns in data. An optimal aspect ratio can enhance trend detection, making it easier to see patterns. Conversely, a suboptimal aspect ratio can obscure important patterns, making it difficult to discern them. The distortion caused by inappropriate aspect ratios can lead to misinterpretations of the data, as noted in studies examining the effects of visual presentation on data interpretation [6].

### 2.4.1 Case Studies

Optimizing the aspect ratio is crucial for effective data visualization. Various methodologies have been developed to select and optimize aspect ratios based on use-case specific criteria.

Earlier work in this field surrounds the concept of *Banking to 45°* as introduced by Cleveland et al. [5]. *Banking to 45°* changes the aspect ratio of a given chart so that the average absolute line orientation is 45 degrees, thereby maximizing the discriminability between line segments. According to Cleveland et al., this change enhances the perception of trends, making it easier for viewers to interpret.

Heer et al. [64] proposed two extensions to Cleveland's technique. They introduced optimizations designed to further improve the perception of line segments and additionally developed a Multi-Scale Banking method based on spectral analysis and Cleveland's *Banking to 45 degrees* [5].

Their enhanced method analyzes a given chart at different frequencies in order to find underlying data trends. In detail, it applies a discrete Fourier transformation to the data to find noticeable periodic trends and compute aspect ratios that optimize the display of these trends. From this, they generate charts for each of those derived frequencies, effectively revealing patterns for different scales. The Multi-Scale Banking technique automates the identification of scales of interest, reducing the need for a manual, iterative adjustment.

Heer et al. [64] illustrated their approach with a few use cases. One use case leverages $CO_2$ measurements from the *Mauna Loa* observatory, in which Multi-Scale Banking automatically identified aspect ratios that best display trends at different scales. As illustrated in Figure 2.24, a broader aspect ratio made yearly oscillations more visible in one plot, while another aspect ratio highlighted a long-term accelerating increase in $CO_2$ levels.

Fink et al. [65] explored the impact of aspect ratio on the readability of scatterplots. The authors argued that selecting an appropriate aspect ratio was crucial because it directly affected the accuracy with which users interpreted the underlying data. They developed a method to compute aspect ratios based on optimizing the plot using Delaunay triangulation based on six specific geometric criteria: maximizing the minimum triangle angle, minimizing total edge length, maximizing triangle inradius, minimizing squared angles, maximizing triangle compactness, and minimizing triangle *uncompactness*. In this context, minimizing uncompactness just meant ensuring that the points are distributed in a way that looks more natural and less distorted, avoiding shapes that are too stretched or compressed.

Their approach relied on the idea that the geometric properties of a scatterplot could guide the selection of an effective aspect ratio. To validate this idea, the authors conducted an empirical study involving 64 participants, who selected aspect ratios for 18 different scatterplots. The study compared the participants' choices with the results produced by their optimization criteria.

The authors concluded that minimizing the total edge length and minimizing the uncompactness of the triangles in their approach resulted in aspect ratios that closely matched those selected by human participants, demonstrating the effectiveness of these two measures.

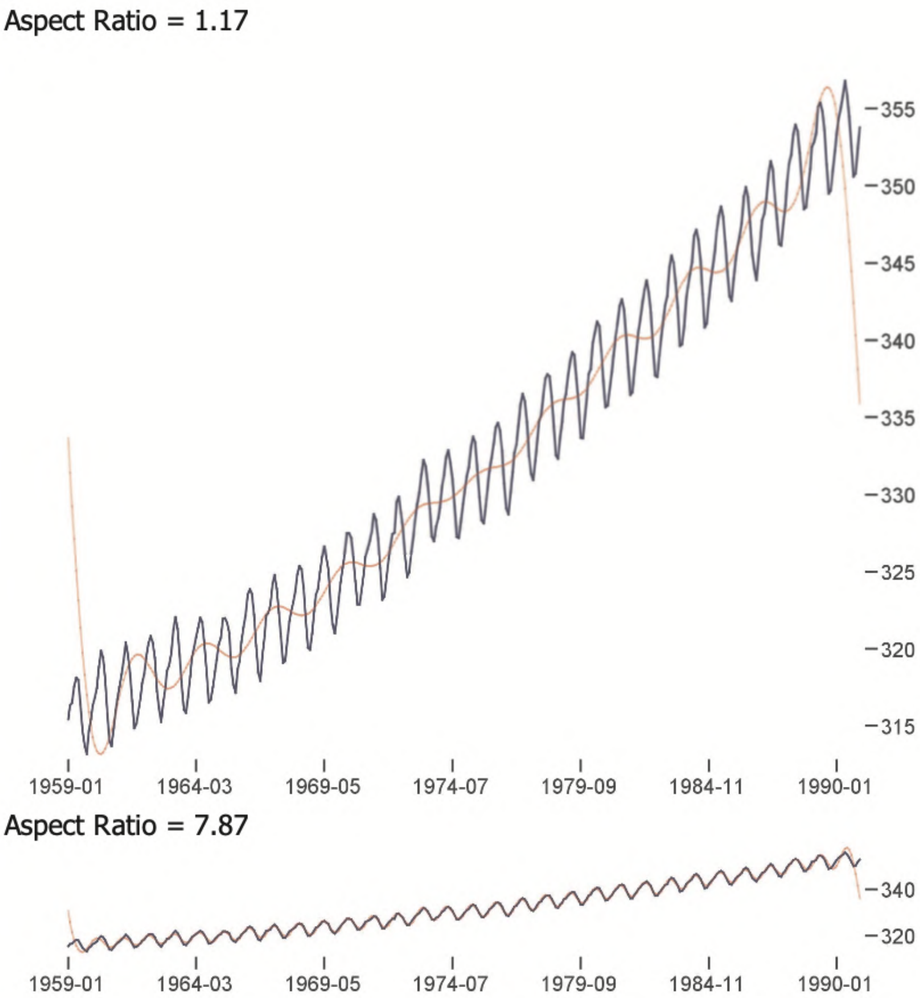
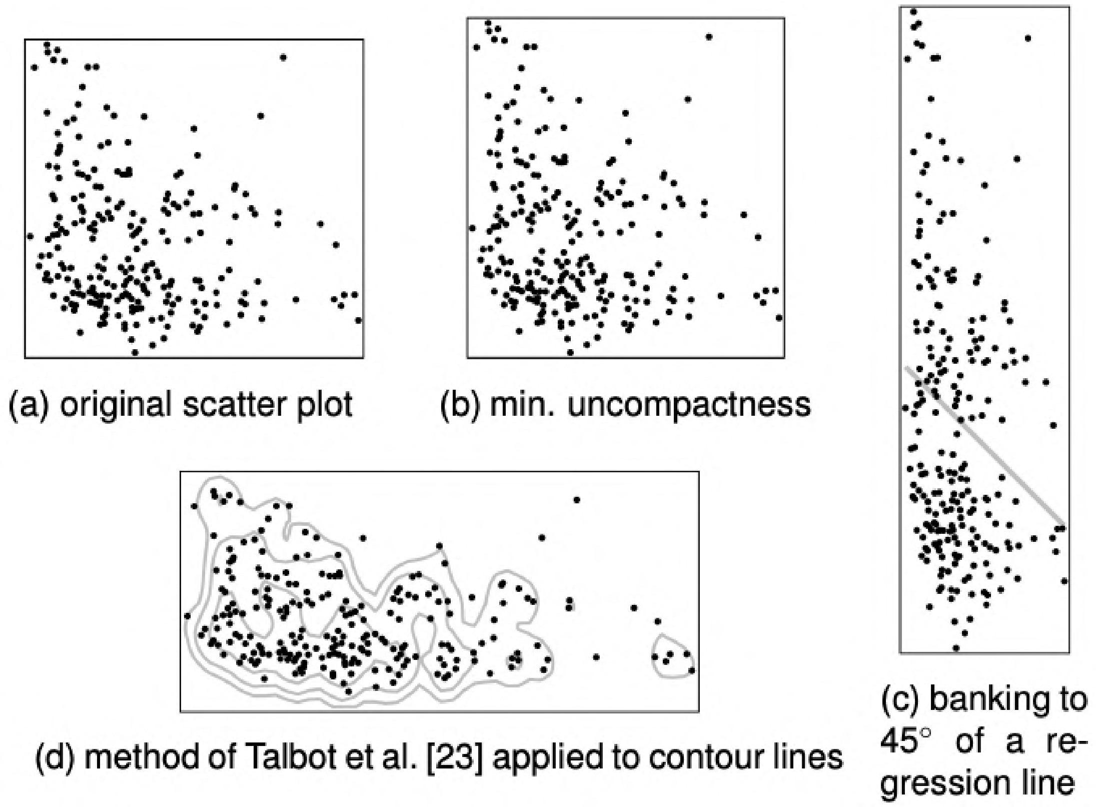


Figure 2.24: Two charts with monthly atmospheric $CO_2$ measurements, highlighting both yearly oscillations (top) and long-term accelerating increase (bottom) through Multi-Scale Banking. Image taken from Heer et al. [64].

In addition, the authors compared their optimization approach to existing techniques, including Cleveland's *Banking to 45°* (see Figure 2.25). Their comparative analysis showed that each method had its strengths and weaknesses, depending on the characteristics of the data.

Cleveland's [5] *Banking to 45°* performed well if the data exhibited strong linear trends, making the slopes easier to interpret by adjusting the aspect ratio to align the median slope with 45°. However, this method did not perform as well on scatterplots with distinct clusters, as the focus on slope alignment sometimes distorted the appearance of clusters, making them harder to recognize.



(a) original scatter plot

(b) min. uncompactness

(c) banking to 45° of a regression line

(d) method of Talbot et al. [23] applied to contour lines

Figure 2.25: Comparison of three different aspect ratio selection methods for scatterplots. This image demonstrates the visual impact of the chosen optimization techniques on data interpretation. Image taken from Fink et al. [65].

Compared to the other approaches, Fink et al.'s [65] proposed method provided a balanced solution, performing well in both cluster recognition and trend detection. This balance made their method more robust across a wider variety of scatterplot types, as it did not overly emphasize trends or density, but instead maintained the geometric properties of the scatterplot.

Recently, research has increasingly focused on directly inferring the optimal aspect ratio from the visualization image itself. Talbot et al. [66] laid the groundwork for an image-based approach on aspect ratio selection using an isoline-based method, which focuses on visualizing optimal density fields.

Building on this, Wang et al. [67] introduced an enhanced image-based method that simplifies the process by directly utilizing density fields. By bypassing the isoline extraction, their approach significantly reduces computational overhead, providing a more efficient solution compared to previous methods (see Figure 2.26).

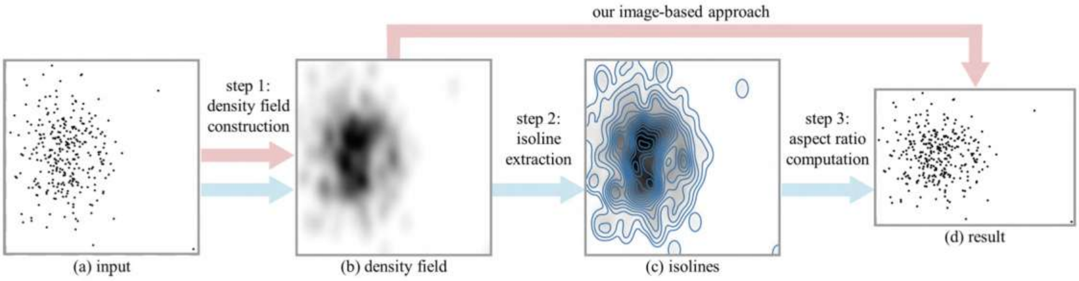


Figure 2.26: Comparison of the proposed image-based approach and traditional workflows for selecting aspect ratios in visualizations. Note that the proposed method bypasses the need for intermediate isoline representations, and instead directly works with density fields. Image taken from Wang et al. [67].

The authors compared their approach to existing methods and user studies. Their results indicate, that their method better reflects the actual distribution of the data and produces a clearer picture of underlying patterns. Additionally, their approach could also be applied to any non-negative and unnormalized use cases, extending its applicability.

# CHAPTER 3

# Methodology

The foundation of this thesis was built on the central research question regarding the influence of aspect ratios on the perception of parallel coordinates. While extensive research exists on the general effects of aspect ratios in traditional data visualization contexts [64][67][65], there remained a notable gap in understanding how these effects manifest in this specific type of visualization technique. To address this gap, our research focused on exploring the visual and cognitive effects of varying aspect ratios on users' ability to gauge correlation patterns in these plots. In detail, we aimed to answer the following question:

**RQ:** *How does aspect ratio influence the perception of correlation in parallel coordinates?*

In this chapter, we detail the research methodology employed to answer this question. We begin by explaining the rationale behind adopting an exploratory approach and further introduce the central hypotheses that guided our investigation. Following this, we present the development of a visualization tool that enables interactive exploration of parallel coordinates. We discuss the requirements, key features, and implementational choices.

Next, we describe the statistical analysis procedures, including the selected datasets and metrics we explored. We outline the hypothesis testing process and the expected impact of this pre-analysis. Lastly, we outline the design of the user study, covering the proposed objectives, participant recruitment, preparation, procedure, timeframe, and subsequent analysis.

## 3.1 Rationale

The decision to adopt an exploratory approach for this study was driven by the complexity of the visual and cognitive processes involved in interpreting parallel coordinates. Traditional visualizations, such as line charts, have been extensively studied in terms of how their aspect ratios influence perception [5]. However, parallel coordinates differ in their structure and representation of multidimensional data, making it unclear whether existing knowledge directly applies to them.

Given the lack of prior research on aspect ratio effects specifically in parallel coordinates, an exploratory methodology was deemed necessary to understand these dynamics in more detail. Therefore, we first conducted a statistical pre-analysis to calculate key metrics, such as correlations between variables and geometric properties (e.g., angles between data points), across various aspect ratios. This analysis helped identify which metrics were most influenced by aspect ratio changes, providing preliminary insights into how these variations affect the interpretability of parallel coordinates.

Following the statistical analysis, we conducted a user study to further investigate how users perceive correlation patterns under different aspect ratios. The combination of these two methods — statistical analysis and user evaluation — enabled us to understand how aspect ratio manipulations impact correlation estimation tasks in parallel coordinates. By exploring the impact from both a theoretical and empirical standpoint, we ensured that the study was more robust and comprehensive.

## 3.2 Hypotheses

Our methodological approach was structured around two main hypotheses, designed to be tested through statistical analysis and empirical evaluation:

**H1:** *Aspect ratio has a significant impact on the interpretability of parallel coordinates.*

**H2:** *Aspect ratios that minimize visual distortion result in improved accuracy and confidence in interpreting parallel coordinates.*

Our first hypothesis (**H1**) argued that aspect ratio has a noticeable effect on the interpretability of parallel coordinates. This hypothesis was based on the premise that a plot's visual layout and proportion can significantly influence how users perceive and interpret the data.

Studies have shown that data representation can affect users' performance in graph reading tasks [68][5]. Aspect ratios unoptimized for the problem domain may distort the visual representation, making it harder to interpret. Applied to parallel coordinates, these effects may lead to issues in accurately interpreting the relationships between variable pairs.

Our second hypothesis (**H2**) suggested that aspect ratios, which inherently have less distortion, lead to better accuracy and higher confidence in data interpretation. This hypothesis was supported by research indicating that minimal distortion allows for a more accurate visual representation of data, enhancing users' ability to discern patterns and trends. Cleveland's concept [5] of **Banking to 45°**, Heer et al.'s [64] *Multi-Scale Banking* method, and Wang et al.'s [67] image-based approach to aspect ratio selection all highlight the importance of optimal aspect ratios in improving visual accuracy and interpretability.

Although the mentioned methods aimed to optimize the aspect ratio based on intrinsic properties or specific metrics, our assumption for this hypothesis was that maintaining a balanced aspect ratio of 1:1 ensures that the data's natural proportions are preserved, leading to higher visual accuracy and interpretability.

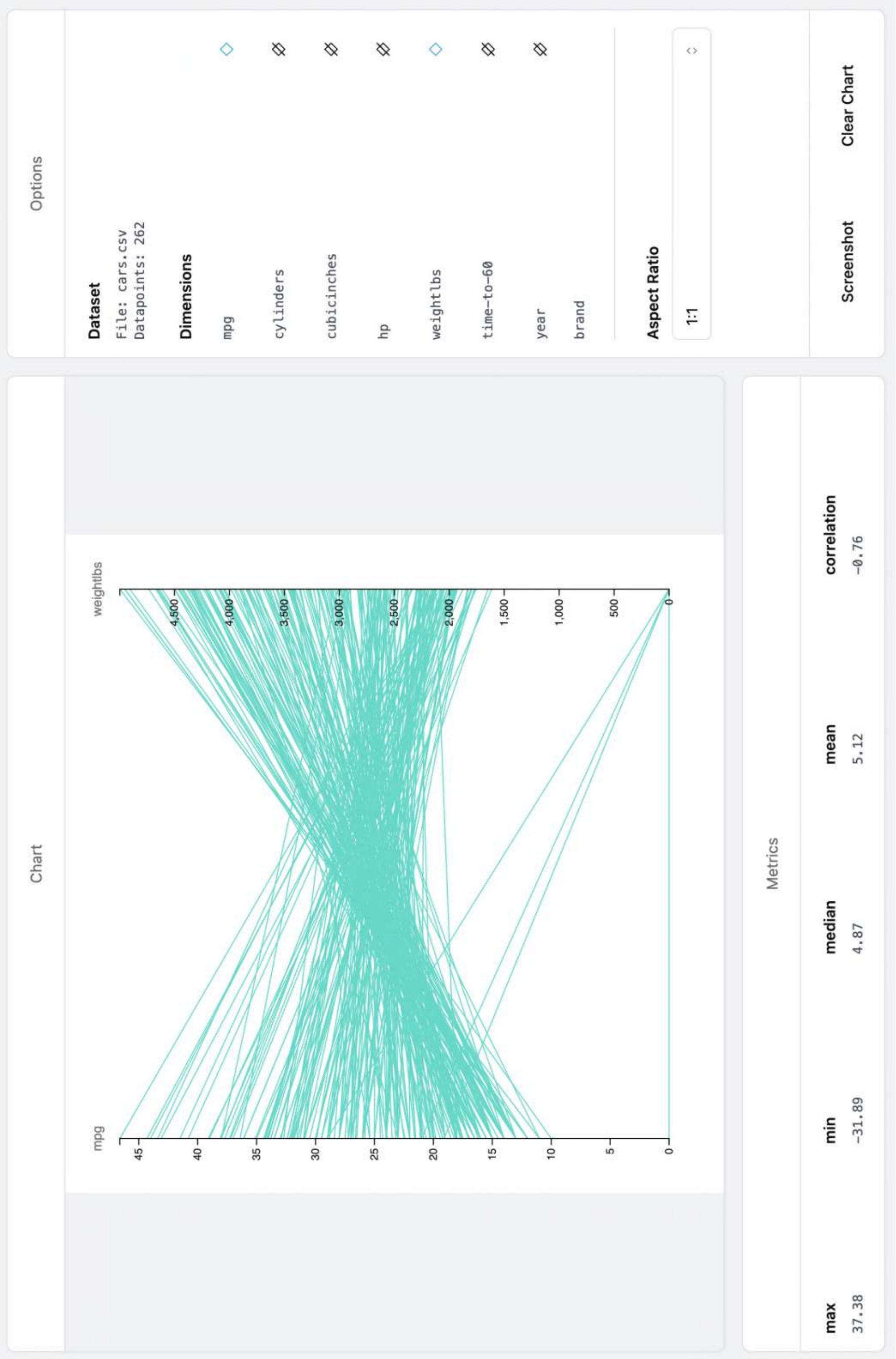
## 3.3 Visualization Tool

Addressing the research questions of this thesis involved creating a dedicated visualization tool. The "Editor", the tool developed for this study, was crucial for interactively exploring parallel coordinates and extracting key metrics (see Figure 3.1).

### 3.3.1 Requirements and Features

The primary objective of the Editor was to load, visualize, and interact with multivariate datasets using parallel coordinates. The Editor allowed to visually explore and experiment with different datasets and aspect ratios. To achieve this, the Editor had to include several features:

- **Interactive Data Loading and Visualization:** The Editor had to enable users to load multivariate datasets in standard formats such as CSV or JSON. Once loaded, the data had to be visualized interactively, enabling users to explore the relationships between variables.
- **Aspect Ratio Adjustment:** The Editor needed to allow users to adjust the aspect ratio of the parallel coordinates dynamically. This functionality was crucial for studying how different aspect ratios impact the perception of data correlations.
- **Data Interaction and Highlighting:** The Editor needed to enable users to interact with the plots by highlighting specific data points or variable axes and to dynamically hide or show variable pairs.
- **Metric Extraction:** The Editor needed to be capable of calculating and displaying metrics for the visualized data. These metrics included aspect ratio, maximum angle, minimum angle, median angle, mean angle, and correlation. These metrics were essential for the subsequent statistical analysis, providing quantitative measures of the geometric properties of the plots.



**Dataset**
File: cars.csv
Datapoints: 262

**Dimensions**

**Aspect Ratio**: 1:1

| max | min | median | mean | correlation |
|---|---|---|---|---|
| 37.38 | −31.89 | 4.87 | 5.12 | −0.76 |

Figure 3.1: Screenshot of the Editor showing a dataset, correlation and angle metrics, and an interactive sidebar for toggling variables and adjusting aspect ratio selection.

### 3.3.2 Implementation Details

The Editor was implemented as a web-based application accessible from a browser. It aimed to be a dynamic and responsive web interface and was developed using a combination of React, Next.js, and D3.js.

#### React

React is a JavaScript library for building web-based user interfaces, especially suited for single-page applications handling complex state and data flow using its component-based architecture [69]. This architecture allows developers to subdivide the UI into reusable components, enabling efficient data handling and management. Each component can maintain its state, simplifying the development and debugging of complex interfaces by isolating functionality and making tracking and managing changes easier.

React's virtual DOM[1] is an important feature that optimizes rendering performance. The virtual DOM is a minimal in-memory representation of the actual DOM, allowing React to efficiently compute the necessary changes needed to update the real DOM. When state changes, React first updates the virtual DOM and then compares it to the actual DOM to identify the changes required to perform [70]. This process, known as reconciliation, noticeably enhances performance by minimizing the effort of direct manipulations to the real DOM, which can be slow and resource-intensive.

The virtual DOM's optimization is particularly beneficial for applications with dynamic and interactive elements that require frequent updates. By reducing the performance overhead associated with DOM manipulations, React ensures responsive user interfaces even in scenarios involving large datasets and complex visualizations. This performance efficiency makes React a good choice for building the Editor's interactive components.

#### Next.js

Next.js is a robust framework built on top of React, designed to enhance the development process of React applications by offering a set of features. One of its core strengths is its file-based routing system, which simplifies navigation within applications [71]. This routing system supports layouts, nested routing, loading states, and error handling, allowing developers to create well-structured and efficient navigation by specifying routes via files and folders.

Next.js also provides extensive support for various styling methods, including CSS modules, pre-processors, and CSS-in-JS, giving developers the flexibility to choose their preferred approach to styling. For Editor's styling, we are relying on *Tailwind CSS* [72] in combination with the minimal component library *shadcn/ui* [73], which are both fully integrated with Next.js.

[1] The Document Object Model (DOM) is an interface for web documents that represents the structure of a document as a hierarchical tree of nodes. It provides a standardized way for accessing and manipulating the elements of a webpage using JavaScript.

Furthermore, Next.js offers enhanced support for TypeScript, including a custom TypeScript plugin and type checker. It also comes with a robust development environment, offering features like hot module replacement, which allows developers to see changes in real-time without a full page reload.

### D3.js

D3.js [74], or Data-Driven Documents, is a JavaScript library known for its ability to produce complex and dynamic data visualizations. It binds data to DOM elements and then applies data-driven transformations to these elements [75]. This makes D3.js exceptionally powerful for creating more complex visualizations that are interactive to user input.

For the Editor, D3.js manages and renders parallel coordinates. Its interactive nature allows users to manipulate these plots, such as by highlighting specific data points, filtering dimensions (i.e., the variables or axes), and dynamically adjusting the visual representation based on user interactions.

D3.js's extensive API exposes helpful interfaces and functions around plotting in general, which are essential for building data visualizations. In combination with React, it enables the creation of reusable and efficient visualization components that can react to data manipulation in real-time.

## 3.3.3 Calculations

D3.js additionally comes with helper functions for data handling and calculation. These functions are particularly useful if working with data in the native D3.js format, as they simplify complex mathematical operations. The following algorithms were used to calculate line angles, correlations, and key metrics.

### Line Angles

Calculating the angular relationships between lines defined by pairs of data points was necessary for analyzing the geometric relationships in a visualizations. The line angle became particularly important in later stages when we aimed to statistically analyze the data.

For determining the minimum, maximum, median, and mean angles for each variable pair we will first need to calculate the individual line angles. The function displayed in Listing 3.1 demonstrates how to calculate the angle between the vertical axis and a line established by two points:

```
const calculateAngle = (point1: Point, point2: Point) => {
    const deltaY = point2.coords[1] - point1.coords[1];
    const deltaX = point2.coords[0] - point1.coords[0];
    const angleRadians = Math.atan2(deltaX, deltaY);

    return angleRadians * (180 / Math.PI) - 90;
};
```

Listing 3.1: Function to calculate the angle between the vertical axis and a line formed by two points based on their coordinates.

### Pearson's Correlation

D3.js's mean and sum functions were leveraged to calculate the correlation coefficient between two variable pairs as detailed in Listing 3.2:

```
const calculateCorrelation = (pointsA: number[], pointsB: number[]) => {
  const meanA = d3.mean(pointsA);
  const meanB = d3.mean(pointsB);

  const numerator = d3.sum(
    pointsA.map((pointA, i) => (pointA - meanA) * (pointsB[i] - meanB)),
  );

  const denominator = Math.sqrt(
    d3.sum(pointsA.map((pointA) => Math.pow(pointA - meanA, 2))) *
      d3.sum(pointsB.map((pointB) => Math.pow(pointB - meanB, 2))),
  );

  return numerator / denominator;
};
```

Listing 3.2: Function to calculate the Pearson correlation coefficient between two sets of data points.

### Key Metrics

Calculating the key metrics between variable pairs was essential for understanding the dataset's general properties in the data points. The function in Listing 3.3 computes various angle metrics, including the maximum, minimum, median, mean angles, and the Pearson correlation coefficient between two selected variables.

```
const calculateMetrics = (
    dataset: Dataset,
    selectedDimensions: string[],
    xScale: d3.ScalePoint<string>,
    yScale: Record<string, d3.ScaleLinear<number, number>>,
    height: number,
): Metrics | null => {
    // Ensure exactly two dimensions are selected for comparison
    if (selectedDimensions.length !== 2) return null;

    // Extract the data points for each dimension
    const pointsA: number[] = dataset.map(
    (data) => +data[selectedDimensions[0]],
    );
    const pointsB: number[] = dataset.map(
    (data) => +data[selectedDimensions[1]],
    );

    // Calculate angles based on the selected dimensions
    const angles = dataset.map((data) => {
        const points = selectedDimensions.map((dim: string) => ({
          coords: [
            xScale(dim),
            isNaN(+data[dim]) ? height : yScale[dim](+data[dim]),
          ],
        })) as Point[];

        return calcualteAngle(points[0], points[1]!);
    });

    // Compute statistical metrics from the angles
    const max = d3.max(angles);
    const min = d3.min(angles);
    const median = d3.median(angles);
    const mean = d3.mean(angles);
    const correlation = calculateCorrelation(pointsA, pointsB);

    return { max, min, median, mean, correlation };
};
```

Listing 3.3: Function to calculate various metrics from a dataset based on selected dimensions.

## 3.4 Statistical Analysis

For the statistical analysis, our primary objective was to calculate key metrics for each sample in a dataset, including the correlation between each pair of variables and the minimum angle, maximum angle, median angle, and mean angle between data points. These calculations were performed for several aspect ratios and exported into a spreadsheet.

Specifically, we sought to identify which of these metrics most strongly correlate with the chosen aspect ratio. By doing so, we were directly testing Hypothesis H1, which claimed that aspect ratio has a significant impact on the interpretability of parallel coordinates.

If our analysis revealed that certain metrics are strongly influenced by changes in aspect ratio, it would support H1 by demonstrating that the aspect ratio significantly affects the geometric properties of the plots. This impact on the geometric properties would be crucial, as it influences how easily and accurately users can interpret the relationships depicted in the plots. Such findings would confirm the hypothesis that the aspect ratio is a key factor in the interpretability of parallel coordinates.

### 3.4.1 Key Metrics

The line angle metrics — minimum, maximum, median, and mean — were chosen because we deemed them essential indicators for how aspect ratio influences the visual interpretation of data. The median and mean angles offered a more general overview of the data's typical angular distribution, helping us identify potential systematic biases introduced by different aspect ratios. The minimum and maximum angles helped us understand the extremes in the data's visual representation, showing how the steepest upward and steepest downward angles between lines might distort perception.

This focus on angles was particularly important with regard to research on graphical perception, which suggests that humans are not equally adept at interpreting all visual encodings. Foundational studies by Cleveland and McGill [76] demonstrated that while people are generally proficient at perceiving certain visual elements, such as positions along a common scale, they are less accurate when interpreting others, particularly angles, areas, and volumes.

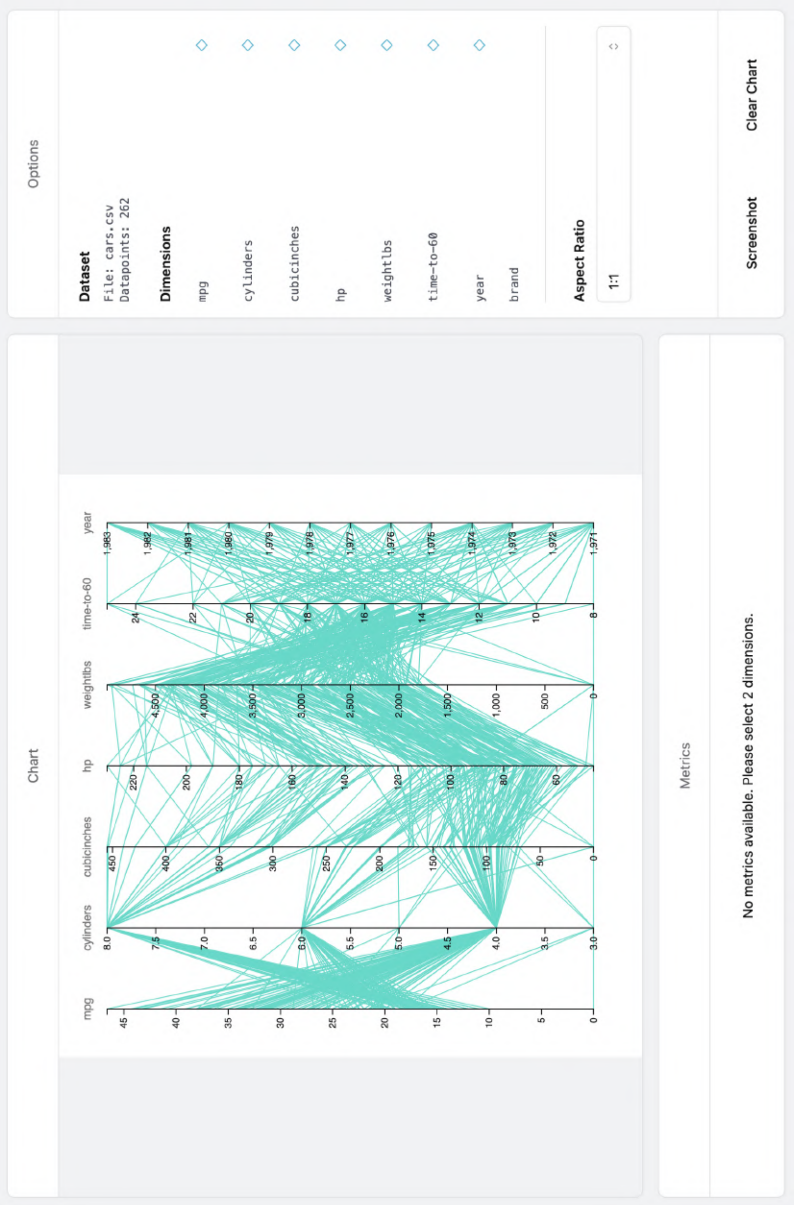
Their research suggests that perceptual accuracy decreases if users only have to rely on angular cues, which are inherently more challenging to judge precisely. This implies that when angles in a visualization become increasingly extreme — either highly acute or obtuse — these perceptual difficulties are likely to be intensified, potentially leading to misinterpretations of the data relationships being represented.

### 3.4.2 Datasets

We selected three diverse datasets for the initial statistical analysis, each with multiple variables and varying numbers of data points. These datasets spanned different domains, providing a broad basis for our analysis.

### Cars Dataset

The cars dataset [77] offered detailed information on various car models, including attributes like engine size, horsepower, weight, and miles per gallon (see Figure 3.2). With over 400 data points and eight performance-related variables, it is a commonly used dataset for prediction tasks.



Options

Dataset
File: cars.csv
Datapoints: 262

Dimensions
mpg
cylinders
cubicinches
hp
weightlbs
time-to-60
year
brand

Aspect Ratio
1:1

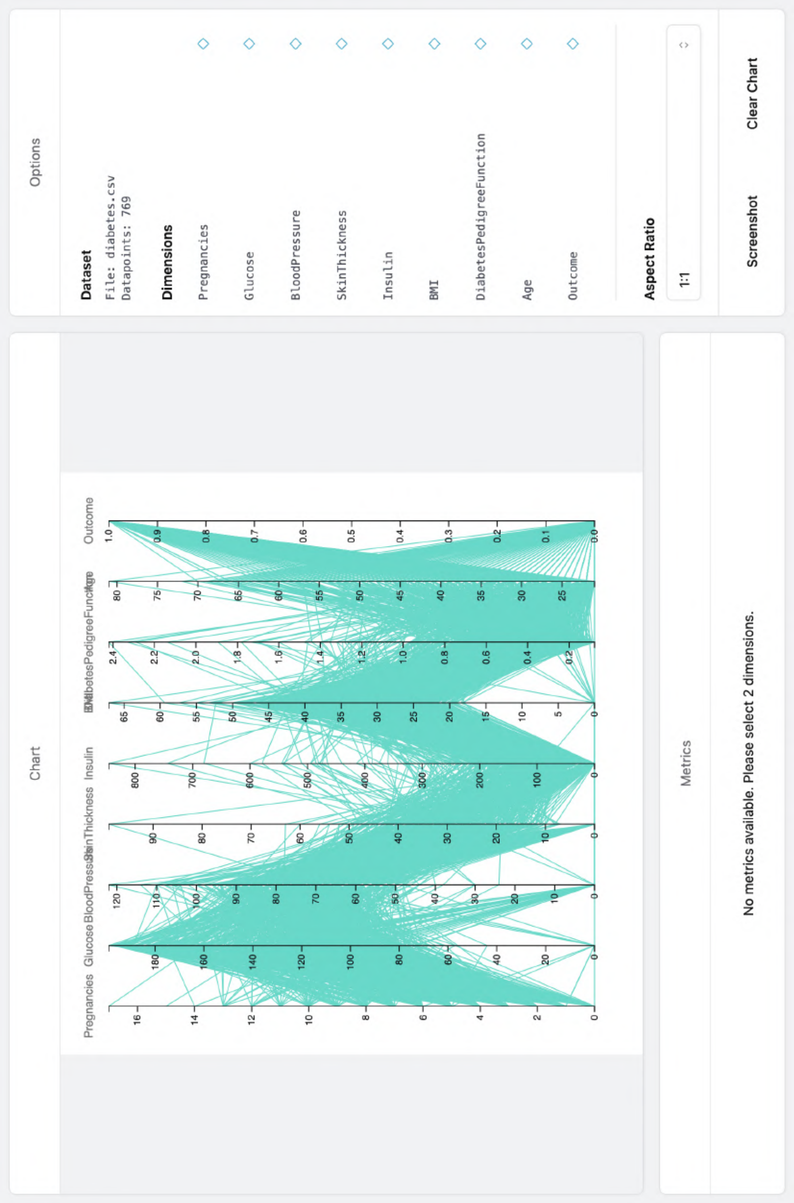
Metrics

No metrics available. Please select 2 dimensions.

Figure 3.2: Cars dataset visualized in the Editor. This dataset has the smallest number of data points among the used datasets, which is underlined by a relatively low amount of visual clutter.

### Diabetes Dataset

Another selected dataset involved medical information for patients, such as age, BMI, blood pressure, and various blood test results [78]. As shown in Figure 3.3, the dataset contained around 770 data points with nine variables and was generally used for examining the relationships between health markers and the presence of diabetes.



No metrics available. Please select 2 dimensions.

**Dataset**
File: diabetes.csv
Datapoints: 769

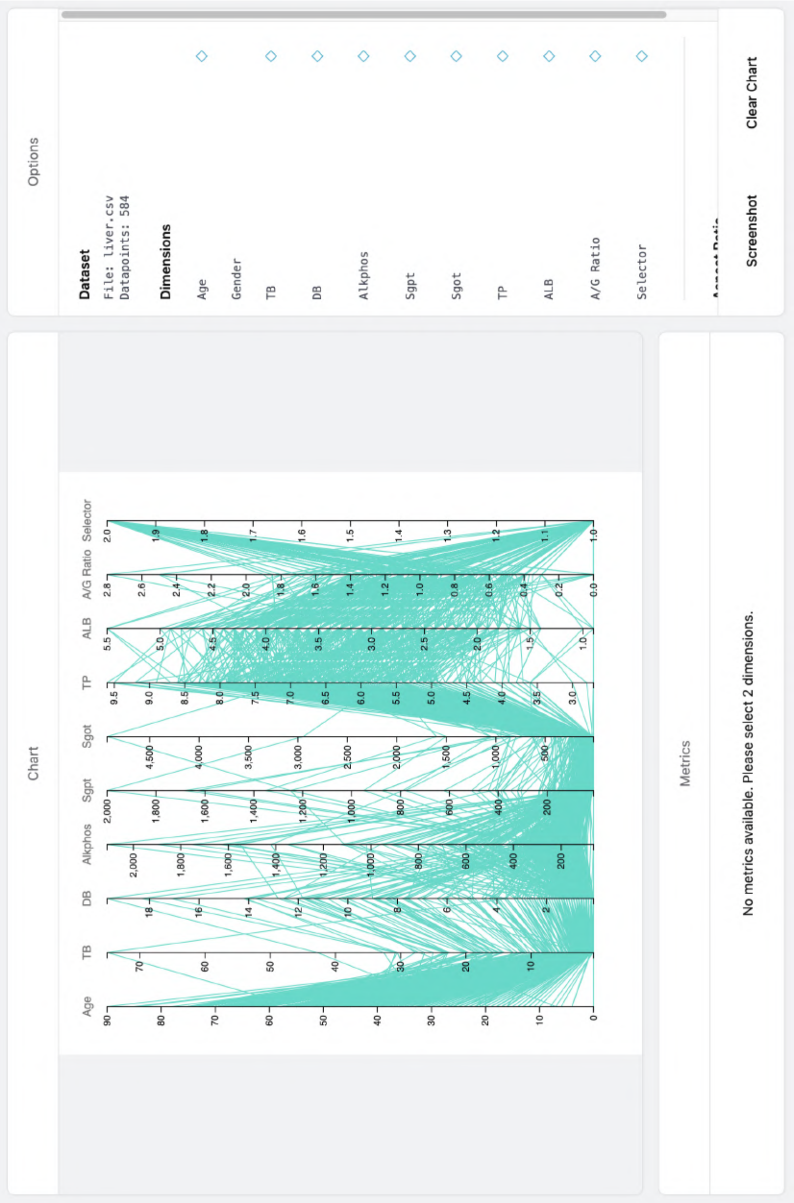
**Dimensions**
Pregnancies ◇
Glucose ◇
BloodPressure ◇
SkinThickness ◇
Insulin ◇
BMI ◇
DiabetesPedigreeFunction ◇
Age ◇
Outcome ◇

**Aspect Ratio**
1:1

Figure 3.3: The Editor displaying the diabetes dataset. Note how this dataset is much denser compared to the cars dataset.

### Liver Disorders Dataset

The third dataset [79] comprised medical test results related to liver disorders, focusing on metrics like albumin or bilirubin levels and enzyme activities (see Figure 3.4). With over 580 data points and 11 variables, this dataset helped explore correlations between biochemical indicators and liver disorders.



Options

Dataset
File: liver.csv
Datapoints: 584

Dimensions
Age ◇
Gender
TB ◇
DB ◇
Alkphos ◇
Sgpt ◇
Sgot ◇
TP ◇
ALB ◇
A/G Ratio ◇
Selector ◇

Screenshot
Clear Chart

Chart

Metrics

No metrics available. Please select 2 dimensions.

Figure 3.4: Parallel coordinates plot of the liver disorder dataset displayed in the Editor. A medium-sized dataset with densely packed lines.

### 3.4.3 Significance Testing

We created a correlation matrix for each dataset to quantify the relationships between aspect ratio and underlying angular metrics. As correlation values range between -1 and 1, we considered correlations with an absolute value above 0.3 ($r \geq |0.33|$) as relevant.

Additionally, we performed significance tests for each correlation coefficient to determine if the observed correlations were statistically meaningful. For this, we run a t-test using an alpha level of 0.05 with the following hypotheses:

- **Null Hypothesis ($H_0$)**: There is no significant correlation between the metric pair.
- **Alternative Hypothesis ($H_1$)**: There is a significant correlation between the metric pair.

$p$-values less than 0.05 were considered statistically significant, indicating that the correlation was unlikely to be due to random chance.

### 3.4.4 Expected Results

By performing initial hypothesis testing through statistical means, we could ensure that subsequent experiments are not solely grounded in assumptions but supported by empirical data. This approach allowed us to establish a quantitative baseline, which was helpful in validating the results obtained in the later stages of the research.

## 3.5 User Study

We aimed to extend our findings from the statistical analysis with quantitative results acquired by a dedicated user study. For this purpose, we enhanced the Editor web application to support various functionalities necessary for conducting the study.

### 3.5.1 Technical Aspects

In developing the infrastructure to support our user study, we focused on creating a robust system that could efficiently manage and process user data. This section describes the underlying technical aspects of the implementation, detailing the technologies used to extend the Editor.

Key aspects of our technical implementation include the use of a PostgreSQL database for reliable data storage, Prisma ORM for streamlined database interactions, and React's Zustand for efficient state management in the application. Additionally, we implemented specific API routes to facilitate the dynamic functionality required for the study, such as randomizing image selection and securely handling user submissions.

### Database: PostgreSQL

We utilized a PostgreSQL database to store user study entries. The database schema included tables for user demographics and responses capturing all data necessary for later analysis.

### Prisma

Prisma served as an ORM[2] client to interact with the PostgreSQL database. It simplified database access by providing an intuitive API, ensuring type safety, and automating query generation [80]. Prisma's integrated migration tools were used to manage database schema changes efficiently, supporting more rapid development and iteration. Within the Prisma schema (see Listing 3.4), we outlined the following models for storing study related data in our PostreSQL database:

```
// Model representing a participant in the study
model Participant {
  id String @id @default(nanoid())
  createdAt DateTime @default(now())

  age        String
  education  String
  occupation String
  experience String

  submission   Submission?
}

// Model representing a submission made by a participant
model Submission {
  id String @id @default(nanoid())
  createdAt DateTime @default(now())

  participant   Participant @relation(
    fields: [participantId], references: [id], onDelete: Cascade
  )
  participantId String @unique

  answers       Answer[]
  Answer model
}
```

[2] Object Relational Mapper (ORM) is an additional layer of abstraction between a database's and application's system design. It allows for query generation using an object-oriented interface, providing a more accessible way of writing queries compared to raw SQL [80].

```
// Model representing an answer submitted in response to a submission
model Answer {
  id Int @id @default(autoincrement())
  createdAt DateTime @default(now())

  correlation String
  confidence  String

  submission   Submission @relation(
    fields: [submissionId], references: [id], onDelete: Cascade
  )
  submissionId String

  image String
}
```

Listing 3.4: Database schema for all user study related entries modeled with Prisma's proprietary syntax.

### Zustand

To manage the state of the user study, including demographics and user answers, we employed the *Zustand* [81] library. Zustand is a small, fast, and scalable state-management library for React applications. It allowed us to maintain all states related to the user study, making it easy to handle user interactions and data flow in the application without the complexity of more heavyweight state management solutions.

### API Routes

Since we had decided to build the application with Next.js, adding backend functionalities was straightforward. The following server-side API routes were introduced for the user study:

- **/api/images**: This endpoint returned 25 random images from a pre-generated image pool, effectively randomizing each test run.
- **/api/submit**: This endpoint handled the submission of user results. It stored user responses, including their correlation judgments and confidence levels, ensuring that all data required for subsequent analysis was securely saved.
- **/api/results**: With this endpoint, the server returned all submissions with participant information and converted the individual entries into an Excel sheet.

### 3.5.2 Objectives

The primary objective of this study was to understand whether users can accurately identify correlations between pairs of variables in parallel coordinates and how various randomly chosen aspect ratios affect their judgments. Additionally, we aim to measure participants' confidence levels in their correlation judgments to gauge the subjective certainty influenced by aspect ratio variations.

### 3.5.3 Participants

The study was targeted toward participants knowledgeable about visualizations or interpreting parallel coordinates. We aimed for a statistically sound sample size, considering the variability in data visualization literacy, with a minimum of 30 participants.

### 3.5.4 Preparation

Randomization was a crucial aspect of this experimental design, aimed at minimizing biases and ensuring that the results are generalizable. By randomly generating images for our study, we ensured a diverse and unbiased data collection.

In total, we generated **135 images** for the study. These images were broken down as follows:

- **5 different aspect ratios** (16:9, 4:3, 1:1, 3:4, 9:16)
- **3 different positive correlations** ($\approx 0.6, 0.8, 1.0$)
- **3 different negative correlations** ($\approx -0.6, -0.8, -1.0$)
- **3 different non-correlations** ($\approx 0$)

We focused on the following aspect ratios for the parallel coordinates to be created:

- **16:9**: Common for widescreen displays and presentations.
- **4:3**: Traditional display ratio.
- **1:1**: Square aspect ratio for equal visual space.
- **3:4**: Taller than wide, the inverse of 4:3.
- **9:16**: Vertical display, generally less suitable for parallel coordinates.

These aspect ratios were selected to cover a broad range of standard formats, ensuring that our study captured how different proportions affect user interpretation in various contexts. Essentially, the selected ratios aimed to replicate common scenarios users might encounter in responsive real-world applications. In the plots (see Figure 3.5), variables of interest were highlighted to ensure they were visible and interpretable by the participants.

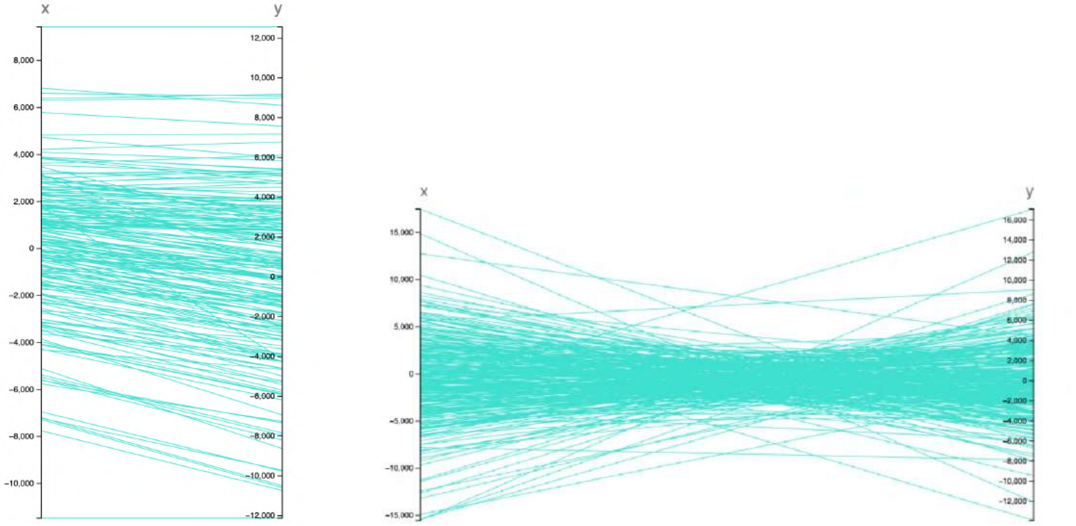


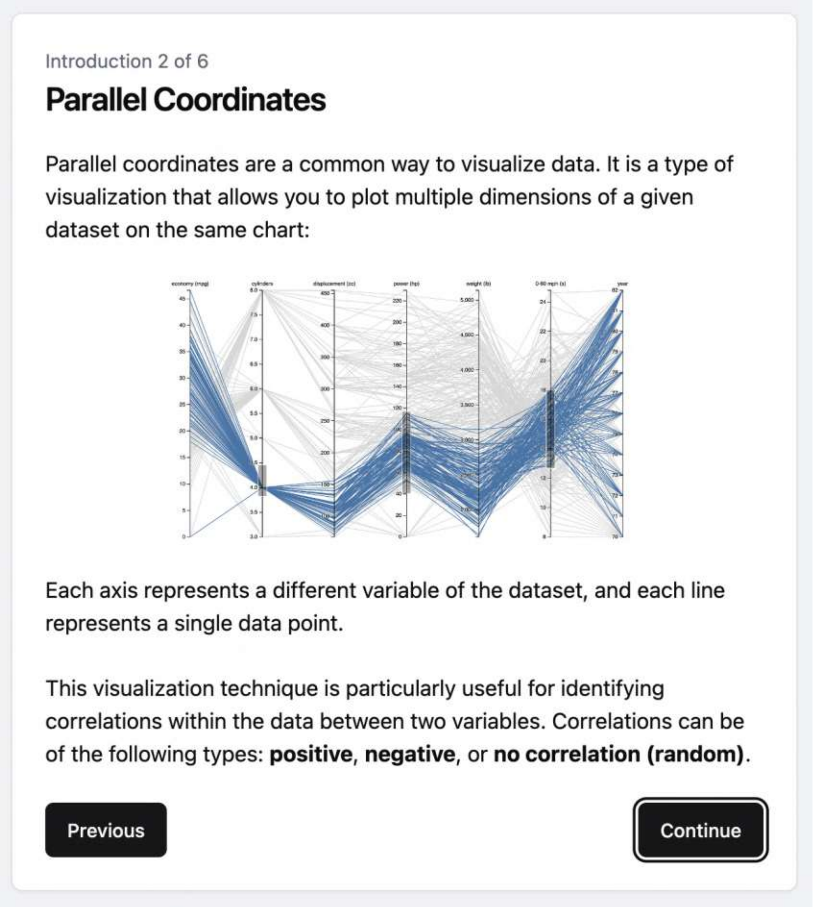
Figure 3.5: Two different plots with varying aspect ratios (9:16 & 16:9) from the pool of generated images.

We ensured that the area displaying the images remained constant. Additionally, we incorporated variations in sample size, noise, and outliers to make the datasets more realistic.

The images included only a distinct variable pair labeled with $x$ and $y$. Although displaying only a single variable pair in parallel coordinates isn't entirely realistic, we chose to do this in order to isolate the specific effects of aspect ratios. By focusing on a single variable pair, we could better control the experiment's conditions, ensuring that participants' performances were directly related to aspect ratio rather than due to complexity induced by additional variables.

### 3.5.5 Procedure

Participants received an introduction to the study's topic, outlining key concepts in interpreting parallel coordinates without revealing the specific focus on aspect ratios (see Figure 3.6). Following this, they underwent a training phase where they were shown examples of plots with clear positive, negative, and non-correlations. This familiarization step ensured that participants understood the task at hand.



Introduction 2 of 6

Parallel Coordinates

Parallel coordinates are a common way to visualize data. It is a type of visualization that allows you to plot multiple dimensions of a given dataset on the same chart:

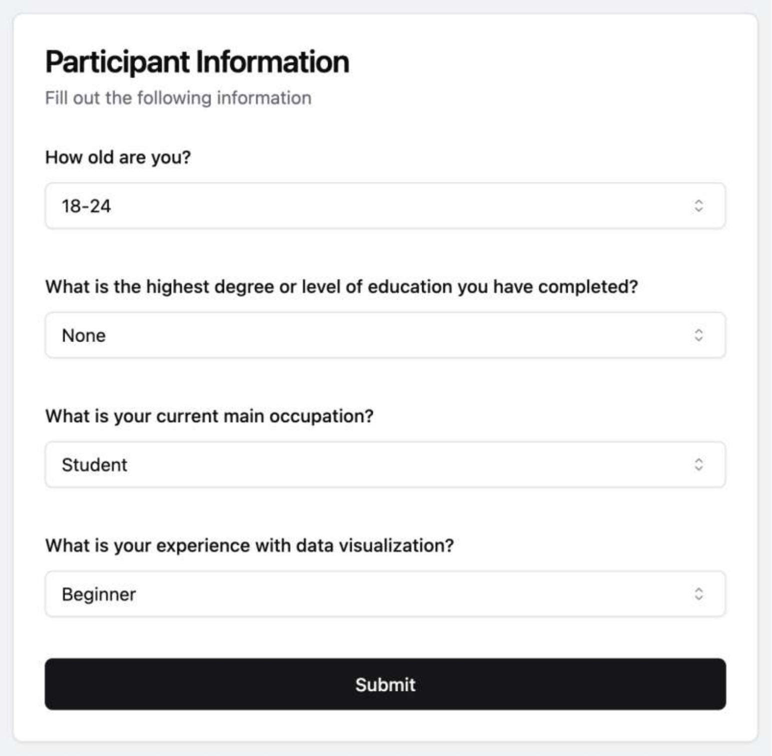
Each axis represents a different variable of the dataset, and each line represents a single data point.

This visualization technique is particularly useful for identifying correlations within the data between two variables. Correlations can be of the following types: **positive**, **negative**, or **no correlation (random)**.

Figure 3.6: Screenshot of the briefing phase in which basic principles on parallel coordinates and visual correlation estimation were introduced.

Before the central part of the study, participants answered questions related to their demographics. As shown in Figure 3.7, this included experience with data visualization, current occupation, higher education, and age.

Participants were then presented with a series of randomly selected plots in the main task (see Figure 3.8). For each plot, they were asked to categorize the correlation between a specified variable pair as positive, negative, or non-existent. They also rated their confidence in their categorization on a textually encoded 4-point Likert scale, where 1 indicated high confidence and 4 represented low confidence. This process helped gather objective accuracy and subjective confidence data across different aspect ratios.



## Participant Information

Fill out the following information

How old are you?

18-24

What is the highest degree or level of education you have completed?

None

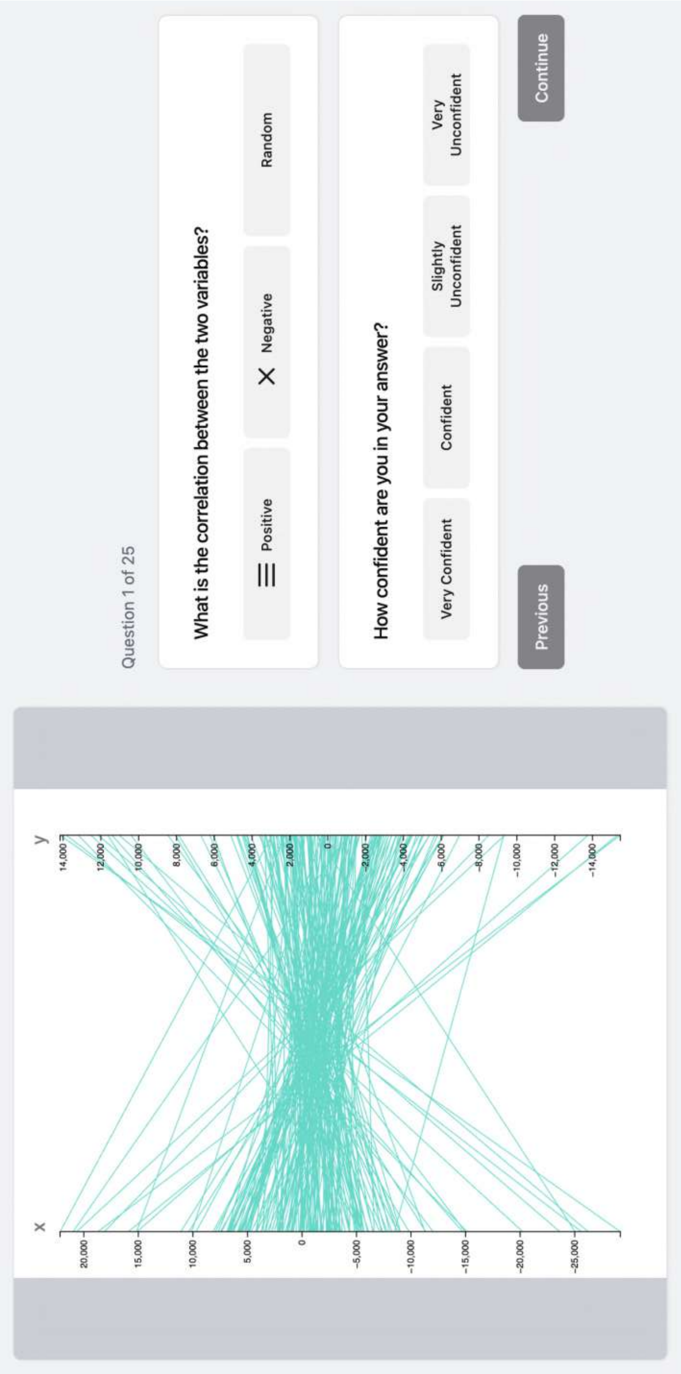
What is your current main occupation?

Student

What is your experience with data visualization?

Beginner

Figure 3.7: Screenshot of the initial questionnaire, which prompted users to specify information on demographic data (e.g., higher education, occupation, experience with data visualization).



Question 1 of 25

What is the correlation between the two variables?

How confident are you in your answer?

Figure 3.8: Screenshot of the main task during the user study. Participants were exposed to a random selection of pre-generated images, asked to give a visual correlation estimation, and report on their confidence.

### 3.5.6 Timeframe

The user study was conducted over six weeks. This timeframe allowed for thorough participant recruitment, ensuring a diverse and representative sample. Additionally, the extended duration provided a buffer for addressing potential technical issues (e.g., deployment and unwanted response caching) during the study, ensuring the integrity and reliability of the collected data.

### 3.5.7 Analysis

We analyzed the user study data in several steps. Initially, we calculated the accuracy of correlation judgments against known correlations in the datasets, comparing perceived correlations with the ground truth. We then analyzed confidence ratings about judgment accuracy and aspect ratios, comparing average confidence ratings for correct and incorrect judgments.

Subsequently, we assessed how different aspect ratios of the parallel coordinates affected judgment accuracy and confidence. This involved a detailed comparison of accuracy and confidence ratings across the various aspect ratios. Lastly, we clustered the results into demographic groups to understand if certain demographic aspects additionally influence both accuracy or confidence ratings. To determine if the differences in accuracy and confidence across different aspect ratios were statistically significant, we performed a single-factor analysis of variance and a post-hoc Scheffé test.

#### Analysis of Variance

Analysis of Variance (ANOVA) is a statistical method used to compare means across more then two groups, generally giving an idea about whether there are statistically significant differences in their means [82][83]. ANOVA is aimed to work with data having three or more groups. The rationale behind ANOVA is to assess the general impact of one or more factors by comparing the means of different samples.

More generally, ANOVA operates under the null hypothesis $H_0$ that all group means are equal, while claiming the alternative hypothesis $H_1$ that at least one group mean is different. The test involves calculating the F-statistic; the ratio of variance estimates is defined as:

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}}$$

The steps in ANOVA include calculating the group means and overall mean, computing the sum of squares (Between-Group Sum of Squares (SSB) and Within-Group Sum of

Squares (SSW)), determining the mean squares (MSB and MSW), and finally calculating the $F$-statistic [83]:

$$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$MSB = \frac{SSB}{k-1}$$

$$MSW = \frac{SSW}{N-k}$$

$$F = \frac{MSB}{MSW}$$

Where $\bar{x}_i$ is the mean of the $i$-th group, $\bar{x}$ is the overall mean, $n_i$ is the number of observations in the $i$-th group, $k$ is the total number of groups, $N$ is the total number of observations, and $x_{ij}$ is the $j$-th observation in the $i$-th group.

If the group means $\bar{x}_i$ are very different from each other compared to the variability within the groups, the $F$-value will be rather large, indicating that at least one group mean is significantly different from the others [83].

This F-statistic is then used in two ways:

1. **P-value ($p$)**: The $p$-value is determined by looking up the calculated $F$-statistic in the F-distribution, which depends on the degrees of freedom. It indicates the probability that the observed differences occurred by chance – a smaller $p$-value provides stronger evidence against the null hypothesis. If the $p$-value is below the significance level (e.g., $\alpha = 0.05$), we reject the null hypothesis, indicating significant differences between the group means.

2. **Critical F-value ($F_{crit}$)**: $F_{\text{crit}}$ is obtained from the F-distribution table using the significance level $\alpha$ and the degrees of freedom ($k$). If the calculated $F$-value exceeds $F_{\text{crit}}$, we reject the null hypothesis, also indicating significant differences. If it does not exceed $F_{\text{crit}}$, we fail to reject the null hypothesis, suggesting no significant differences.

### Scheffé's Test

Following the ANOVA test, we employed a post-hoc Scheffé test to pinpoint specific pairs of aspect ratios that show significant differences. Scheffé's test is a conservative post-hoc analysis that controls the Type I error rate when making multiple comparisons [84].

The conservativeness of Scheffé's test arises from the fact that it adjusts the critical value used to assess significance based on the number of comparisons being made [84]. This adjustment increases the threshold required to declare a result as significant, thereby reducing the likelihood of falsely identifying a difference as statistically significant if no such difference exists. This makes Scheffé's test particularly useful in situations where all possible comparisons are of interest, as it maintains the overall error rate across the set of comparisons, unlike multiple t-tests, which increase the cumulative risk of Type I errors with each additional comparison [7][85].

We calculated the Scheffé statistic for each pair of group means, comparing the differences with the Scheffé critical value. The formula for the Scheffé statistic $F_s$ is given by:

$$F_s = \frac{(\bar{x}_i - \bar{x}_j)^2}{MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

where $\bar{x}_i$ and $\bar{x}_j$ are the means of groups $i$ and $j$ and $MSW$ is the within-group mean square error from the ANOVA test. $n_i$ and $n_j$ are the number of observations in groups $i$ and $j$.

Further, we needed to calculate the critical Scheffé value $S$, which was obtained by taking the degrees of freedom (between-groups) df and multiplying it with the critical F-value $F_{\text{crit}}$ from the ANOVA test.

$$S = \text{df} \times F_{\text{crit}}$$

To determine whether the difference between the group means is significant, we compared the calculated $F_s$-value with the critical Scheffé value $S$:

- If $F_s$ is greater than the critical Scheffé value $S$, the difference between the group means was considered statistically significant.
- If $F_s$ is less than or equal to $S$, the difference was considered not statistically significant.

CHAPTER 4

# Results

This chapter covers the results of our evaluations. It is divided into two main parts: statistical analysis and empirical user study.

We begin with the description of the results of the statistical analysis, examining the results of evaluating three diverse datasets to identify correlations between aspect ratio and angular metrics in the plots. This section includes the results of correlation analyses and significance tests, highlighting the relationship between aspect ratio and geometric properties.

Following the statistical analysis, we present the outcomes of our web-based user study. This study empirically evaluated how different aspect ratios affect users' accuracy and confidence. We describe the significance of accuracy and confidence ratings, using ANOVA and Scheffé post-hoc tests to determine the statistical significance of the observed differences.

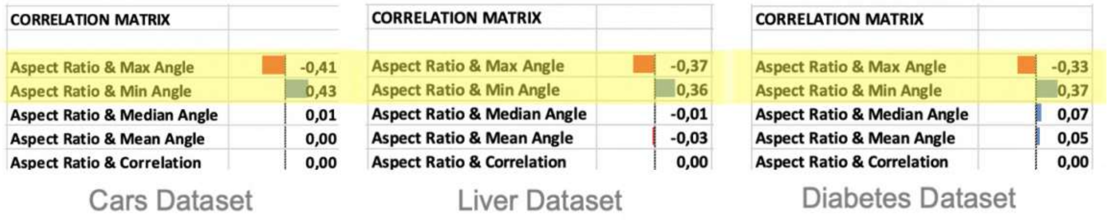
Further, we discuss the implications of our findings regarding the impact of aspect ratios on the perception of correlations in parallel coordinates. We interpret the results from both the statistical analysis and the user study, reflect on our hypotheses, and consider the practical implications of our findings.

## 4.1 Statistical Analysis

The statistical analysis aimed to quantitatively identify which metrics most strongly correlate with changing aspect ratios across all datasets. By examining the three selected datasets and evaluating various angular metrics, we determined the magnitude and significance of these correlations.

### 4.1.1 Correlation Matrix

As outlined in Section 3.4, the statistical analysis involved examining three datasets across three different aspect ratios (16:9, 4:3, 1:1). Each dataset contained between 400 and 770 data points. We extracted metrics for each variable pair, focusing on the maximum, minimum, median, and mean angles. We then calculated the Pearson correlation between each metric pair and established a correlation matrix for each dataset.



| CORRELATION MATRIX | |
|---|---|
| Aspect Ratio & Max Angle | -0,41 |
| Aspect Ratio & Min Angle | 0,43 |
| Aspect Ratio & Median Angle | 0,01 |
| Aspect Ratio & Mean Angle | 0,00 |
| Aspect Ratio & Correlation | 0,00 |

Cars Dataset

| CORRELATION MATRIX | |
|---|---|
| Aspect Ratio & Max Angle | -0,37 |
| Aspect Ratio & Min Angle | 0,36 |
| Aspect Ratio & Median Angle | -0,01 |
| Aspect Ratio & Mean Angle | -0,03 |
| Aspect Ratio & Correlation | 0,00 |

Liver Dataset

| CORRELATION MATRIX | |
|---|---|
| Aspect Ratio & Max Angle | -0,33 |
| Aspect Ratio & Min Angle | 0,37 |
| Aspect Ratio & Median Angle | 0,07 |
| Aspect Ratio & Mean Angle | 0,05 |
| Aspect Ratio & Correlation | 0,00 |

Diabetes Dataset

Figure 4.1: Results of the statistical analysis on three selected datasets (cars, liver, and diabetes). The matrices show the Pearson correlation coefficients $r$ for all metric pairs. Significant correlation pairs ($r \geq |0.33|$) were highlighted in yellow.

.

As highlighted in Figure 4.1, the generated correlation matrices revealed notable correlations of $r \geq |0.33|$ between *aspect ratio* and both *maximum* and *minimum angle*. This indicates that changing the aspect ratio noticeably affects the angles between lines, especially at the extremes. These alterations affect the visual representation and perception of the data, making it arguably easier or harder to identify patterns and correlations.

### 4.1.2 Significance

To determine the significance of the correlations, we performed a t-test on our findings. The results across all datasets showed that the correlations between aspect ratio and minimum/maximum angle were significant at a 0.05 confidence interval. This confirmed that the observed correlations were not due to random chance and that aspect ratio significantly alters the angles in parallel coordinates.

### 4.1.3 Geometric Impact

In retrospect, we could have calculated the angle metrics from the changed aspect ratio directly. When the aspect ratio changes, it effectively stretches or compresses the visual elements along one axis relative to the other. This geometric transformation directly influences the angular metrics, leading to the observed correlations. This, in turn, reinforces our findings, demonstrating that the observed correlation is actually due to an underlying functional relationship driven by the geometric effects of aspect ratio changes.

## 4.2 User Study

We further conducted a user study to investigate how aspect ratio influences users' accuracy and confidence in identifying correlations. This study aimed to validate the statistical results through empirical evaluation, focusing on user perception.

### 4.2.1 Demographics

The web-based user study involved 57 participants who were familiar with data visualizations and mostly had experience in interpreting parallel coordinates. As outlined in Figure 4.2, participants were primarily between the ages of 25 and 34, which accounted for 65% of the group. Smaller segments included those aged 35 to 44 and 45 to 54, with the youngest and oldest age groups being the least represented.

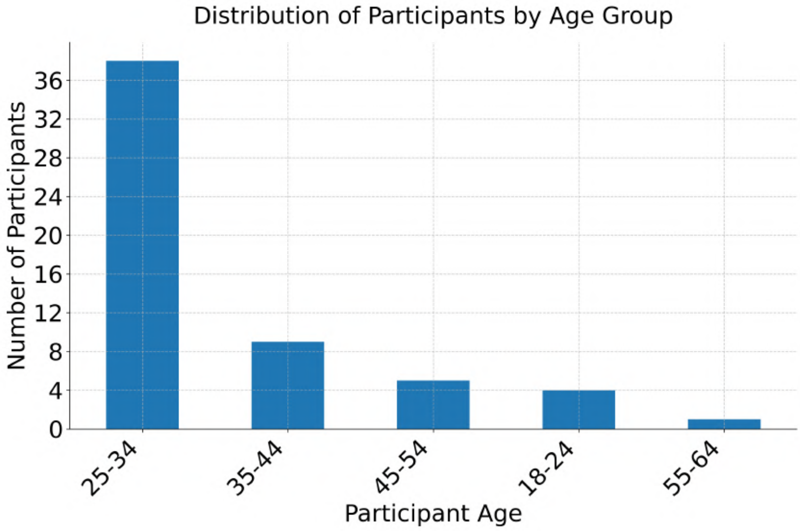


Figure 4.2: Age distribution of participants in the study. The majority of participants belong to the 25-34 age group, representing a significant portion of the sample.

The participants' experience levels varied (see Figure 4.3), with a significant portion (21 participants) being advanced users. Nearly equal parts of participants identified themselves as intermediates or beginners, indicating a balanced mix of expertise in data visualization.

Higher educational backgrounds were predominantly at the graduate level, with most participants holding a Master's degree (see Figure 4.4). Bachelor's degree holders formed

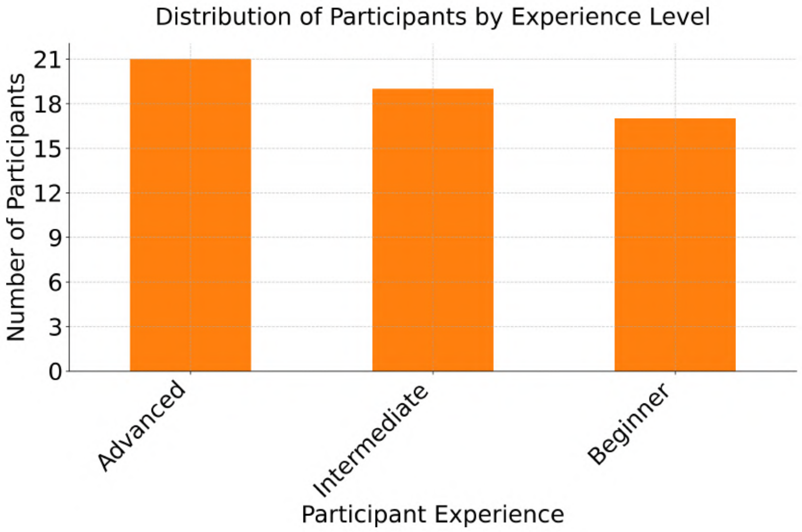


Figure 4.3: Level of experience with data visualization among participants. The distribution across advanced, intermediate, and beginner levels was fairly balanced.

the second-largest group, while fewer had achieved a PhD or higher. Only one participant did not report having a higher educational qualification. This hints at an educational bias in our study, as the majority of participants had advanced educational backgrounds.

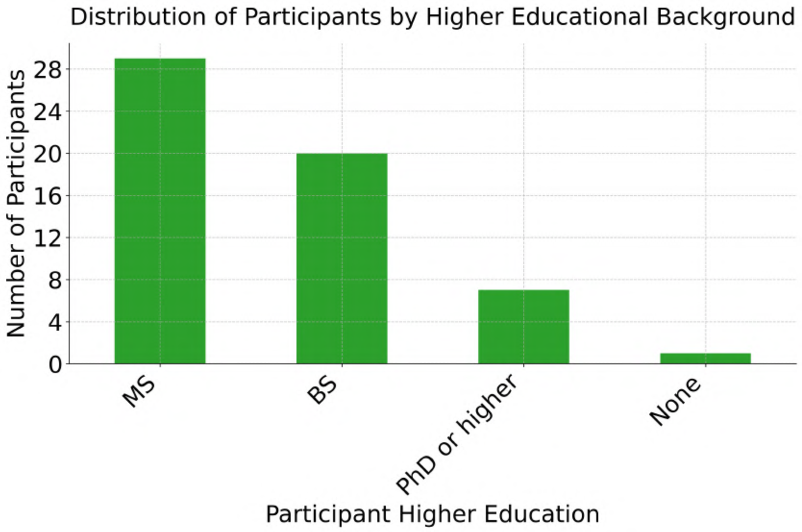


Figure 4.4: Higher educational background of the participants. The majority of participants had an MS degree, followed by those with a BS degree. Fewer participants had a PhD or no higher education.

The participants' occupational distribution, visualized in Figure 4.5, included 25 professionals, 15 academics, 15 students, and 2 individuals in other occupations.

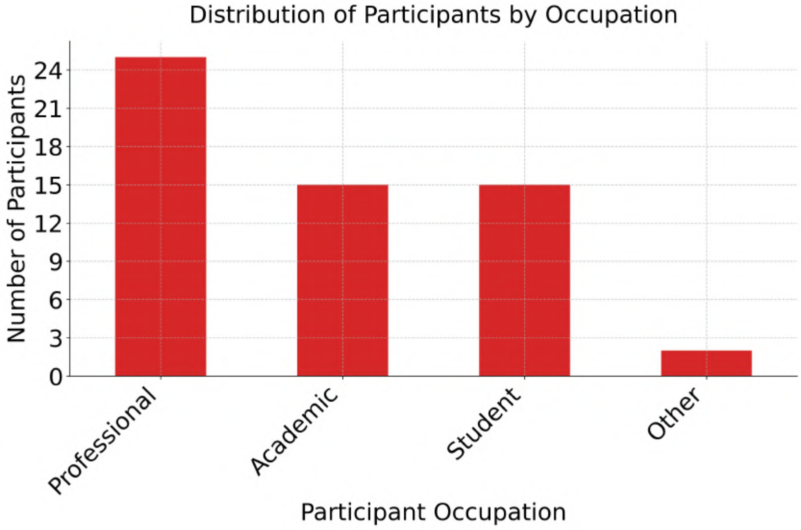


Figure 4.5: Participants' current occupations categorized by profession. The largest group of participants were professionals, followed by academics and students, while a small portion of participants fell into the "Other" category.

### 4.2.2 Mean Analysis

The accuracy of participants' correlation judgments was determined by comparing their answers to the known correlations in the datasets. The same was done for the confidence ratings. The insights presented in the next sections are based on the mean values calculated for all participants. In Section 4.2.3, we will take a closer look at the data, exploring our findings beyond just central tendencies.

#### Accuracy

The overall accuracy across all aspect ratios was 0.71. This level of accuracy suggests a reasonably high degree of proficiency among the participants in interpreting the visual data.

Taking a look at Figure 4.6, the 1:1 aspect ratio showed the highest accuracy at 0.77, suggesting that the most "balanced" visual representation aids in better data interpretation. The wider and taller aspect ratios, 16:9 and 9:16, resulted in lower accuracy at 0.67 and 0.66, respectively, indicating potential distortion in data perception. The aspect ratios 4:3 and 3:4 had accuracies of 0.71 and 0.73, demonstrating moderate performance.

Regarding accuracy per experience level, detailed in Figure 4.7, advanced users had the highest accuracy at 0.75, while beginners and intermediate users had accuracies of 0.69 and 0.68. This suggests that, in this study, more experienced users better interpreted correlations in parallel coordinates.

Accuracies per occupation showed that academics had the highest accuracy at 0.80,

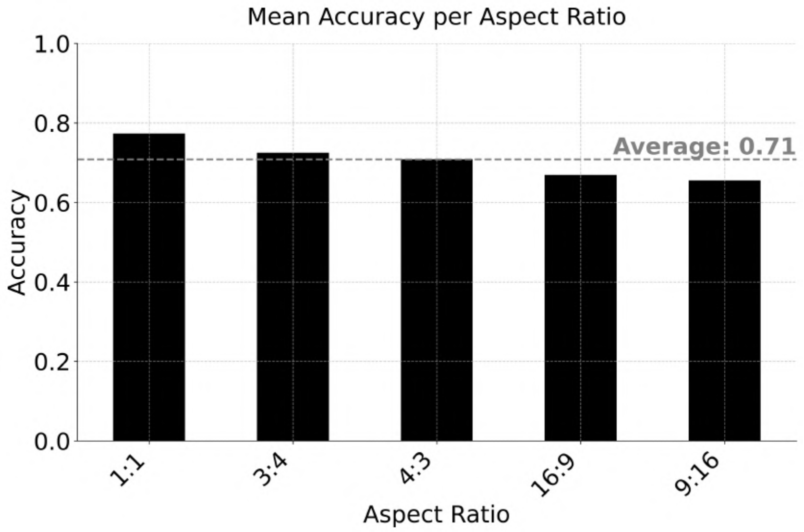


Figure 4.6: Mean accuracy of participants ranked by aspect ratio. The 1:1 aspect ratio achieved the highest accuracy, while the accuracy steadily decreases for other aspect ratios.

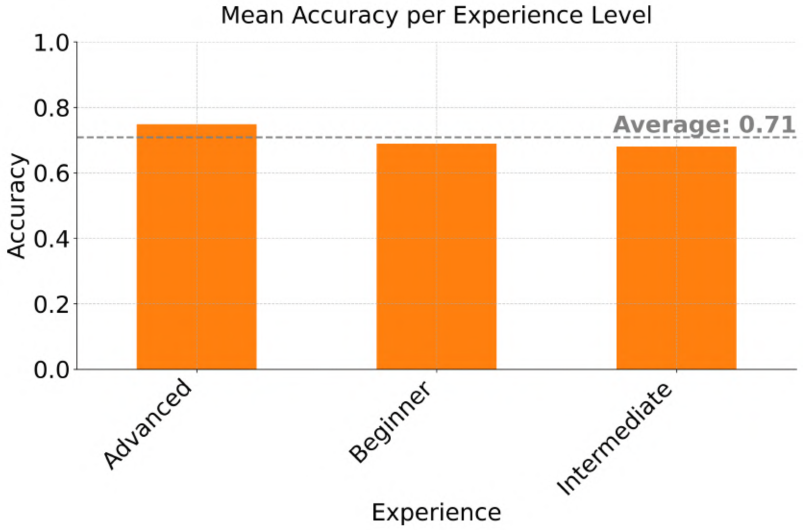


Figure 4.7: Participant accuracy grouped by experience level in data visualization. Results indicate a similar performance across all experience levels, with advanced users showing a marginally higher accuracy compared to beginners and intermediates.

followed by professionals at 0.68 and students at 0.67. This indicates that participants with academic backgrounds, arguably more exposed to analysis and visualization tasks, generally perform slightly better (see Figure 4.8).

Figure 4.9 shows the accuracy per education level. Participants with a PhD or higher

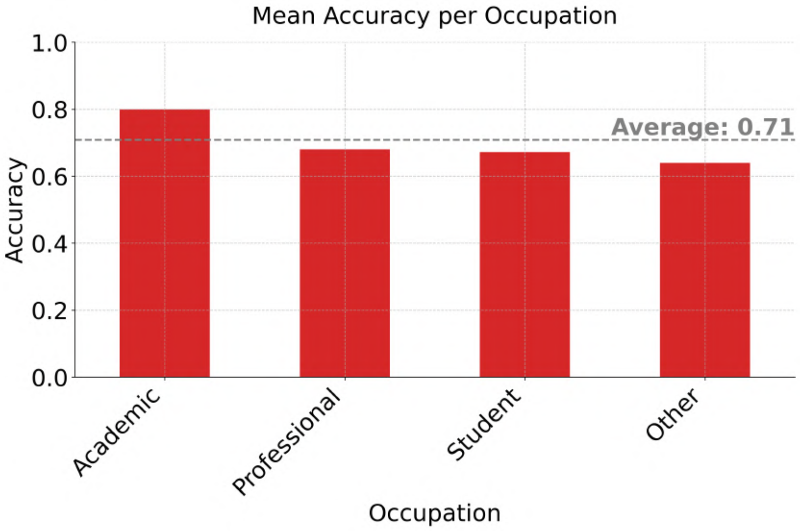


Figure 4.8: Participant accuracy based on occupation. Academics exhibited the highest accuracy, followed by professionals and students.

had the highest accuracy at 0.74, followed by participants with a Master's degree at 0.73, a Bachelor's degree at 0.69, and no higher education at 0.20. This trend suggests that, on average, participants with higher education levels show better accuracy in interpreting parallel coordinates.

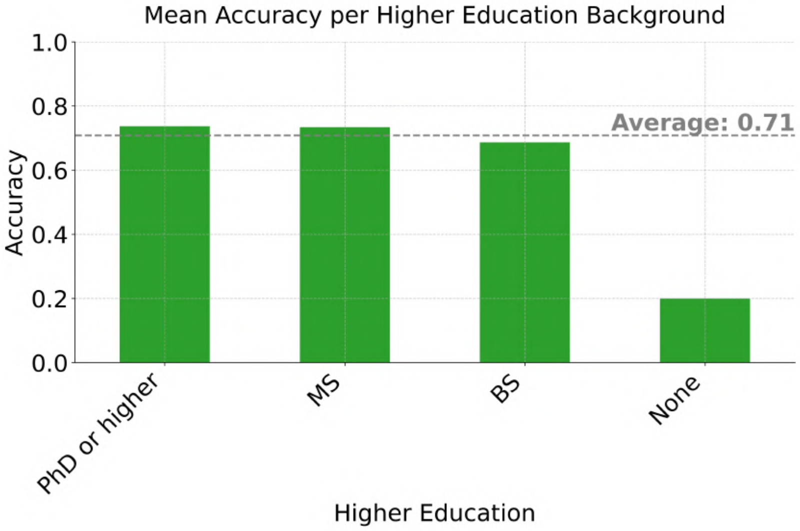


Figure 4.9: Participant accuracy according to educational background. Individuals with a PhD or MS degree performed similarly and achieved the highest accuracy, while those with a BS degree performed slightly lower.

Among the groups we invited to participate in this study, we included members of VRVis as well as fellow student colleagues. Consequently, many of these participants had prior experience and training in interpreting complex visual data, which likely contributed to higher accuracy rates.

### Confidence

In addition to accuracy, participants rated their confidence in their correlation judgments. This self-reported measure provided insight into how participants felt about their own assessments. Confidence was originally rated on a 4-point Likert scale. Ratings of 1 and 2 were categorized as confident, while ratings of 3 and 4 were grouped as not confident. For consistency, we also normalized the results to a value between 0 and 1.

The average confidence rating was 0.69, indicating that while participants generally trusted their interpretations, they were still somewhat uncertain about their selection. For the confidence per aspect ratio (see Figure 4.10), the 16:9 aspect ratio had the highest ranking at 0.72, followed by 4:3 at 0.70, 9:16 at 0.69, 1:1 at 0.67, and 3:4 at 0.66. This suggests that participants felt slightly more confident with specific aspect ratios despite their actual accuracy.

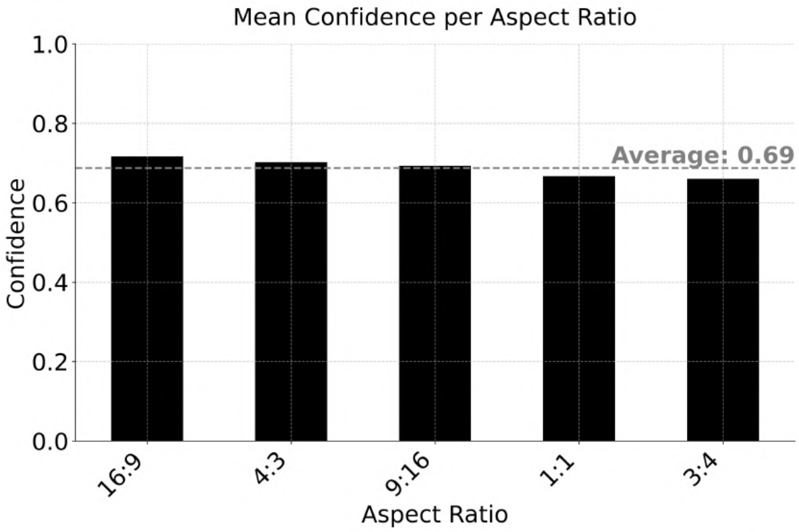


Figure 4.10: Participant confidence levels compared across different aspect ratios. The 16:9 aspect ratio led in confidence, while confidence ratings on the other aspect ratios were gradually decreasing.

As seen in Figure 4.11, with confidence per experience level, advanced users reported the highest confidence at 0.74, while intermediate and beginner users had confidence levels of 0.67 and 0.64. Thus, more experienced users performed slightly better and felt more confident in their judgments.

For confidence per occupation (see Figure 4.12), it is important to mention that the highest accuracy was observed in the "Others" category, but this is not representative as

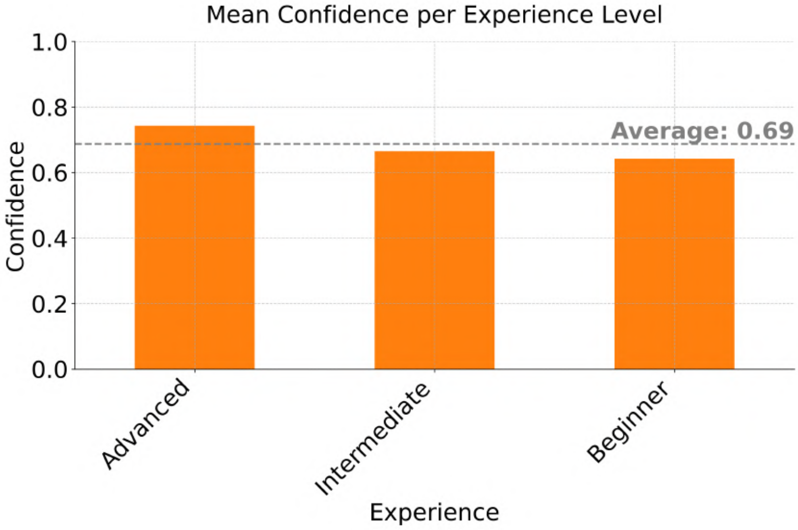


Figure 4.11: Confidence levels of participants based on their experience with data visualization. Advanced users reported the highest confidence, as opposed to slightly lower confidence levels for the intermediate and beginner participants.

it consisted of only one person. Among the more representative groups, professionals had the highest confidence at 0.71, followed by academics at 0.68 and students at 0.66.

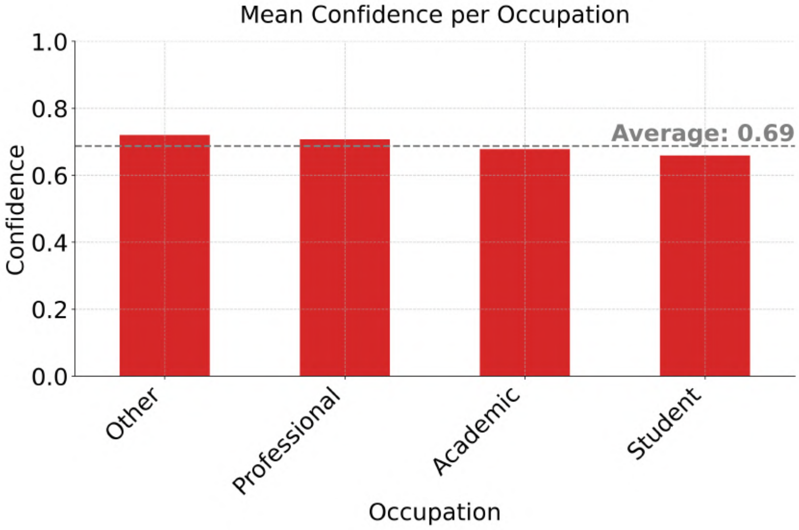


Figure 4.12: Participants' confidence categorized by occupation. Confidence was mostly consistent across occupations, with only minor differences.

As Figure 4.13 indicates, participants with a PhD or higher reported the highest confidence at 0.78, followed by those with a Master's degree at 0.70. Participants with either a Bachelor's degree or no university education both ranked similarly at 0.64. Trendwise, higher education levels seemed to rank slightly higher in confidence with respect to interpreting parallel coordinates.

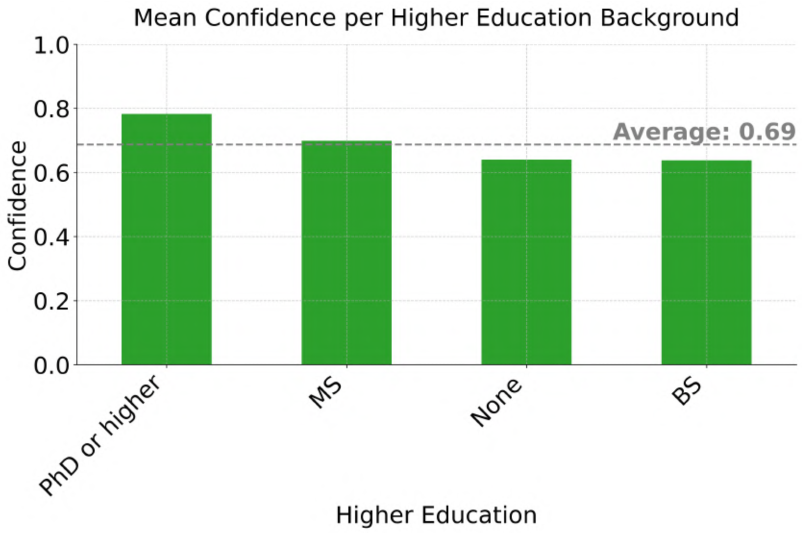


Figure 4.13: Confidence ratings of participants categorized by educational background. Individuals with a PhD or higher expressed the greatest confidence, while participants with Master's, Bachelor's, or no degree demonstrated somewhat lower levels of confidence.

### 4.2.3 Significance Analysis

We also aimed to determine whether the differences in accuracy and confidence across various aspect ratios were statistically significant. For this, we performed a single-factor ANOVA test. The significance level was set to 0.05 for all statistical tests. If significance was found, we further identified which specific pairs of aspect ratios had significant differences using a Scheffé post-hoc test.

### 4.2.4 Accuracy

As detailed in Section 3.5.7, the ANOVA test was employed to analyze whether the mean accuracy across different aspect ratios in our study varied significantly. We needed this since insights from mean values alone generally do not provide a complete understanding of the differences between groups. The ANOVA test allowed us to assess whether any observed differences in accuracy are statistically significant, rather than just being due to random variation.

#### ANOVA Test

The input for the ANOVA consisted of accuracy measurements collected from participants as they interacted with visualizations under different aspect ratios. The output of the ANOVA includes summary statistics as shown in 4.1, as well as the ANOVA table with the F-statistic $F$ and $p$-values (see Table 4.2), which help determine whether there are statistically significant differences between the group means by comparing the between-group variance to within-group variances.

In the summary table 4.1, accuracy scores refer to the proportion of correct interpretations made by participants when observing the visualizations and the calculated summary statistics are bases on the respective data points for each aspect ratio group.

| Groups | Count | Sum | Average | Variance |
|---|---|---|---|---|
| 16:9 | 275 | 184 | 0.6691 | 0.2222 |
| 4:3 | 292 | 207 | 0.7089 | 0.2071 |
| 1:1 | 300 | 232 | 0.7733 | 0.1759 |
| 3:4 | 291 | 211 | 0.7251 | 0.2000 |
| 9:16 | 267 | 175 | 0.6554 | 0.2267 |

Table 4.1: Summary statistics for accuracy per aspect ratio. This table provides the count, sum, average, and variance of accuracy scores for each aspect ratio group.

| Source of Variation | SS | df | MS | $F$ | $p$ | $F_{crit}$ |
|---|---|---|---|---|---|---|
| Between Groups | 2.5199 | 4 | 0.6300 | **3.0632** | **0.0159** | **2.3782** |
| Within Groups | 292.0373 | 1420 | 0.2057 | | | |
| Total | 294.5572 | 1424 | | | | |

Table 4.2: ANOVA single factor analysis for accuracy per aspect ratio. This table contains the results of a one-way ANOVA test, comparing accuracy across different aspect ratios.

The ANOVA table generates various statistics related to the analysis of variance and reads as follows:

- **SS (SSB/SSW)**: The total variability in the data.
- **df (Degrees of Freedom)**: The number of independent values in the analysis.
- **MS (MSW/MSB)**: The average of the sum of squares.
- ***F***: The ratio of the variance between the groups and within the groups. The higher the $F$-statistic, the greater the disparity between group means.
- ***p***: The probability that the observed results occurred by chance. A $p$ below 0.05 typically suggests statistical significance.
- **$F_{crit}$**: The critical value that the $F$-statistic must exceed to reject the null hypothesis at the 0.05 significance level.

For accuracy, the ANOVA results show a $p$-value of 0.015, which is below the alpha level of 0.05. This allows us to reject the null hypothesis that the mean accuracy across all five aspect ratios is the same and instead accept the alternative hypothesis that there are significant differences in accuracy across the different aspect ratios. Comparing the $F$-statistic of 3.0632 to the critical value $F_{crit} = 2.3782$ further supports the presence of significant differences between the accuracy levels across the different aspect ratio groups.

### Scheffé Post-Hoc Test

Following the ANOVA test, a Scheffé post-hoc test was conducted. The Scheffé test was particularly useful for making multiple comparisons between group means while controlling for Type I errors (see Section 3.5.7). The input for the Scheffé test consists of the same accuracy data used in the ANOVA, and its output highlights which specific pairs of aspect ratios show significant differences in accuracy.

| **Comparison** | $\boldsymbol{F_s}$ |
|---|---:|
| 16:9 vs 4:3 | 1.0915 |
| 16:9 vs 1:1 | 7.5810 |
| 16:9 vs 3:4 | 2.1555 |
| 16:9 vs 9:16 | 0.1229 |
| 4:3 vs 1:1 | 2.9867 |
| 4:3 vs 3:4 | 0.1856 |
| 4:3 vs 9:16 | 1.9391 |
| 1:1 vs 3:4 | 1.6720 |
| 1:1 vs 9:16 | **9.5487** |
| 3:4 vs 9:16 | 3.2849 |

Table 4.3: Scheffé Test results for accuracy per aspect ratio. This table summarizes the pairwise comparisons of variances in accuracy results between aspect ratios using the Scheffé test. The F-statistic ($F_s$) for each comparison is calculated to determine the significance of differences between groups. Statistically significant differences in accuracies between aspect ratios are marked by a bold $F_s$ value.

The Table 4.3 displays the results for the Scheffé post-hoc test. It contains the pairs of aspect ratios being compares and the computed Scheffé F-statistic for each comparison. A higher $\boldsymbol{F_s}$-value indicates a greater difference between the groups. To be considered significant, this value must exceed the critical Scheffé value $S$ of **9.5128** ($S = df \times F_{crit}$).

The results indicate that comparing the 1:1 and 16:9 aspect ratios revealed a significant difference, with the 1:1 aspect ratio leading to notably higher accuracy, as evidenced by an $F_s$-value of 9.5487, which exceeds the critical Scheffé value. This suggests that the balanced 1:1 aspect ratio provides a clearer and more proportional view of the data, minimizing distortions that can mislead users.

### 4.2.5 Confidence

#### ANOVA Test

As shown in Tables 4.4 and 4.5, the ANOVA results for confidence show a $p$-value of 0.55, which is well above the threshold of 0.05. This leads us to retain the null hypothesis that the mean confidence across all five aspect ratios is the same. The $F$-value of 0.7570, which is below the $F_{crit}$-value of 2.3782, also supports the conclusion that there are no significant differences in confidence across the different aspect ratios.

| Groups | Count | Sum | Average | Variance |
|---|---|---|---|---|
| 16:9 | 275 | 197 | 0.7164 | 0.2039 |
| 4:3 | 292 | 205 | 0.7021 | 0.2099 |
| 1:1 | 300 | 200 | 0.6667 | 0.2230 |
| 3:4 | 291 | 192 | 0.6598 | 0.2252 |
| 9:16 | 267 | 185 | 0.6929 | 0.2136 |

Table 4.4: Confidence summary per aspect ratio. The averages and variances are relatively consistent across the aspect ratios, with 16:9 having a slightly higher average and 3:4 displaying the highest variance.

| Source of Variation | SS | df | MS | $F$ | $p$ | $F_{crit}$ |
|---|---|---|---|---|---|---|
| Between Groups | 0.6520 | 4 | 0.1630 | **0.7570** | **0.5534** | **2.3782** |
| Within Groups | 305.7579 | 1420 | 0.2153 | | | |
| Total | 306.4098 | 1424 | | | | |

Table 4.5: ANOVA results comparing confidence across aspect ratios. The test shows an $F$-value of 0.7570 and a $p$-value of 0.5534, which suggests no statistically significant difference in confidence between the aspect ratio groups.

This shows that the aspect ratio of parallel coordinates does not significantly influence users' confidence in identifying correlations. Regardless of the aspect ratio, participants' confidence levels remain relatively consistent, suggesting that while aspect ratio affects accuracy, it does not strongly impact how confident users feel about their interpretations.

#### Scheffé Post-Hoc Test

Since the ANOVA test did not show significant differences in confidence across aspect ratios, the Scheffé post-hoc test was unnecessary. In this case, the lack of a significant ANOVA result means we can conclude there are no statistically notable differences in confidence between any pairs of aspect ratios.

### 4.2.6 Correlation Type Differences

In addition to our previous analyses, we conducted a further investigation into whether the type of correlation displayed (positive or negative) affects users' accuracy and confidence when interpreting parallel coordinates. Specifically, we performed t-tests to compare the accuracy and confidence for user judgments on both positive and negative correlations.

#### Accuracy

The t-tests showed a statistically significant difference in accuracy between positive and negative correlations, with a p-value of $p = 1.79581 \times 10^{-5}$, which is well below the 0.05 threshold. As indicated by the results, participants were more accurate when interpreting negative correlations. The mean accuracy for negative correlations was higher (0.8151) compared to positive correlations (0.6975), suggesting that users found it easier to detect more distinguishable patterns in negative correlation visualizations.

#### Confidence

Similarly, the confidence ratings for positive and negative correlations differed significantly, with a p-value of $p = 0.0279$. However, in contrast to accuracy, participants reported higher confidence in their judgments for positive correlations (0.7878) compared to negative correlations (0.7276).

## 4.3 Discussion

In the following section, we discuss the results of our findings regarding the impact of aspect ratios on the perception of correlation in parallel coordinates. We interpret the results from both the statistical analysis and the user study, reflect on our hypotheses, and consider the practical implications of our findings. The interpretation of our results is divided into three subsections: accuracy, confidence, and reflections on our hypotheses.

### 4.3.1 Accuracy

Our user study found that the aspect ratio of parallel coordinates significantly influences the accuracy with which users interpret correlations. The 1:1 aspect ratio (see Figure 4.14), which provides a balanced visual representation, leads to the highest accuracy rates. This suggests that maintaining a proportional view minimizes visual distortions, allowing users to correctly identify correlation patterns in the plots.

In contrast, wider or taller aspect ratios such as 16:9 and 9:16 (see Figure 4.15) result in lower accuracy, implying visual misrepresentation and users' difficulty in interpreting skewed data presentations. Additionally, intermediate aspect ratios, 4:3 and 3:4 (see Figure 4.16), induce moderate accuracies.

The user study results also align with the findings in the statistical analysis. A 16:9 aspect ratio, which is much wider than tall, tends to stretch the lines in the plot horizontally,

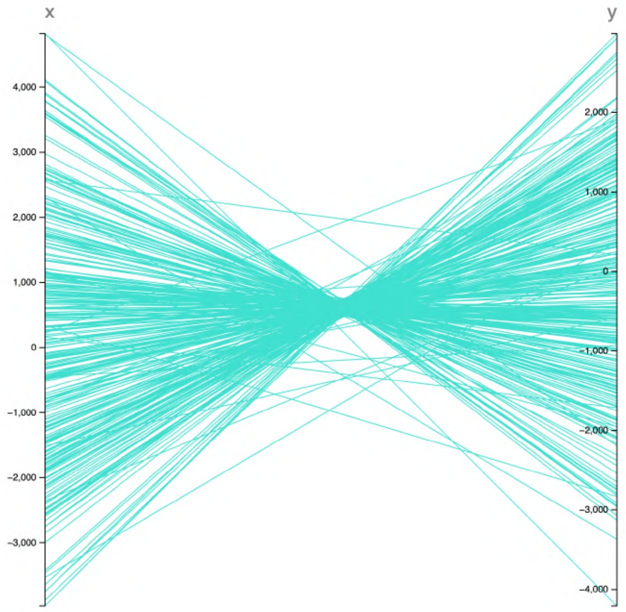


Figure 4.14: Parallel coordinates with an aspect ratio of 1:1, displaying a correlation of $r \approx -0.96$. This aspect ratio was identified as the most effective for visual correlation estimation in our user study.

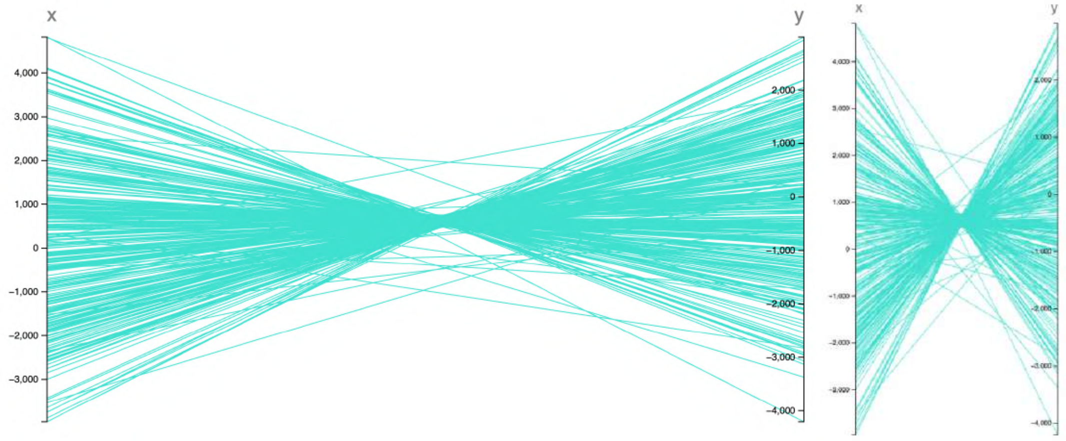


Figure 4.15: Parallel coordinates ($r \approx -0.96$) with aspect ratios of 16:9 (left) and 9:16 (right). According to our user study, these aspect ratios resulted in the least accurate visual correlation estimations.

making it harder for users to perceive the actual angles between data points. This horizontal stretching can cause users to overestimate or underestimate the strength of

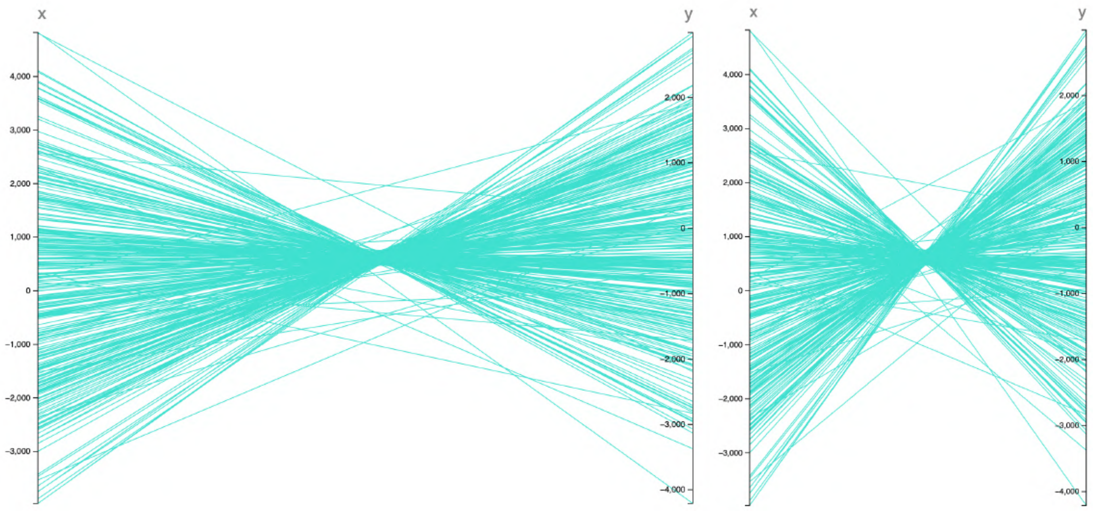


Figure 4.16: The 4:3 (left) and 3:4 (right) aspect ratio plots, with $r \approx -0.96$, introduced moderate distortion. However, these distortions were less severe than those observed with 16:9 and 9:16 ratios.

correlations, leading to inaccurate conclusions. Similarly, the 9:16 aspect ratio, much taller than wide, compresses the lines horizontally, causing similar interpretative challenges.

The data also suggest that the moderate aspect ratios (4:3 and 3:4) provide a compromise between the "extreme" ratios and the balanced 1:1 ratio. While these ratios do introduce some level of distortion, it is less pronounced than with the 16:9 or 9:16 ratios. This means that users can still interpret the data reasonably, though likely less effectively than with the 1:1 ratio.

### 4.3.2 Confidence

Interestingly, while accuracy varies with aspect ratio, user confidence does not significantly differ across the various aspect ratios. Participants reported similar confidence levels regardless of whether the aspect ratio was balanced or not. This indicates a potential disconnect between perceived and actual accuracy. Users might feel equally confident in their judgments, even if the aspect ratio compromises their accuracy. This finding suggests that users' subjective confidence may not be a reliable indicator of their performance in interpreting parallel coordinates with varying aspect ratios.

Moreover, the consistent confidence levels across aspect ratios suggest that users are not aware of the distortive effects of different aspect ratios. This could stem from a general assumption that visualizations are accurate representations of data without recognizing the subtleties introduced by design choices such as aspect ratio. In this regard, if users had been aware of the study's focus, their accuracy and confidence ratings could have differed, as they might have paid more attention to the potential distortions caused by the different aspect ratios.

### 4.3.3 Reflection

Our study was structured around two main hypotheses, each addressing how aspect ratios influence the interpretation of parallel coordinates. In this section, we want to reflect on these hypotheses in the light of our findings.

Hypothesis **H1** proposed that aspect ratios have a noticeable effect on the interpretability of parallel coordinates. Our results support this hypothesis, as we observe significant variations in accuracy across different aspect ratios. In this study, the 1:1 aspect ratio, in particular, was the most effective for accurate data interpretation. This finding aligns with previous research that suggests maintaining balanced proportions can minimize visual distortions and enhance data interpretability.

Hypothesis **H2** proposed that aspect ratios with less distortion generally lead to better accuracy and higher confidence in data interpretation. This hypothesis is partially supported. While the 1:1 aspect ratio improved accuracy, it did not significantly impact user confidence. This partial support potentially indicates that while users can interpret data more accurately with optimal aspect ratios, they may not recognize the improved accuracy, most likely due to a lack of awareness about the effects of aspect ratios as a whole.

### 4.3.4 Practical Implications

Our study's findings have several practical implications for designing and using parallel coordinates in data visualization. Understanding these implications can help practitioners make more informed decisions that improve the effectiveness and accuracy of visual data interpretations.

Firstly, our study demonstrates that – when viewing variable pairs – balanced aspect ratios, such as 1:1, can help enhance the accuracy of data interpretation by minimizing visual distortions and helping users to identify correlations in the data. However, we can only confirm the significance for the 1:1 aspect ratio, which suggests that further research is needed to conclusively determine the optimal range of aspect ratios for accurate data interpretation. Designers should be cautious when using aspect ratios with high degrees of distortion, as these influence data perception and lead to misinterpretation, potentially resulting in false conclusions.

Secondly, given the disconnect between accuracy and confidence, it is essential to educate users about the potential distortions caused by different aspect ratios. Practitioners and decision-makers should know the importance of aspect ratios and other design elements in influencing perception. By emphasizing these factors, they can develop a more critical eye when working with visual data.

Moreover, instead of allowing fully responsive interfaces that dynamically change the viewport, users might benefit from locked views where the aspect ratio remains constant, but only the scale of the entire plot changes. This could provide a consistent alternative to responsive views. Locking the aspect ratio could ensure that the visual proportions of the data remain unchanged, preventing misinterpretations arising from stretching or compressing the visualization.

More generally, by locking the aspect ratio and adjusting only the scale, users can maintain a consistent frame of reference while viewing data in a fixed aspect ratio that the designer has determined to be optimal.

CHAPTER 5

# Conclusion

This thesis investigated the impact of aspect ratios on the perception of correlations within parallel coordinates. We developed a visualization tool to enable the interactive exploration of datasets, allowing for the dynamic adjustment of aspect ratios. Through a combination of a statistical analysis and an empirical user study, we explored how different aspect ratios influence the users' ability to interpret multivariate data.

Our study revealed that aspect ratios significantly affects the accuracy in data interpretation, with the 1:1 aspect ratio proving to be the most effective one. For visual correlation estimation tasks isolated to two variables, our results suggest that a balanced aspect ratio helps users to give more accurate estimations. Our findings align with existing literature on data visualization, which emphasizes the importance of generally maintaining proportionality.

However, results also indicate that user confidence did not vary significantly across different aspect ratios, revealing a disconnect between perceived and actual performance. This suggests that users are often unaware of how aspect ratios influence their interpretive accuracy. This unawareness could result from a general assumption that all visualizations are equally effective, regardless of the aspect ratio, highlighting a critical area for education.

In practical terms, these results imply that designers of data visualization tools should prioritize balanced aspect ratios to improve the accuracy of data interpretation. Additionally, we suggest that users may benefit more from interfaces with fixed, predefined aspect ratios for data interpretation tasks.

## 5.1 Limitations

While providing valuable insights into the impact of aspect ratios, our study's design has some limitations. Firstly, the scope of our research was constrained by the specific aspect ratios we investigated. We focused only on a limited set of aspect ratios for the statistical analysis (16:9, 4:3, 1:1) and the user study (16:9, 4:3, 1:1, 3:4, 9:16). While generally representative, this may not cover all possible variations that could affect data interpretation.

Another area for improvement is the reliance on self-reported confidence levels, which may not accurately reflect the participants' true confidence or understanding. Although we used these self-reports to gauge subjective confidence, there could be discrepancies between reported confidence and actual comprehension.

Lastly, we did not integrate the impact of sample size or data density in this study. Higher data densities may introduce visual clutter and make it more difficult to interpret correlations accurately, this variable was not considered in our analysis.

## 5.2 Future Work

Building on our study's findings, some areas for future research arise. To address the limitations of our work, further research should explore a wider range of aspect ratios beyond those we examined. Investigating more varied aspect ratios could provide a more nuanced understanding of how different visual proportions affect data interpretation.

Additionally, research should incorporate more objective quantitative measures of user confidence and understanding, such as task completion times. In this context, collecting qualitative data, which provides a more detailed picture of how participants felt during the study, may also be valuable. These additional data points could complement self-reported confidence levels and provide a better assessment of how aspect ratios impact performance.

As mentioned in the limitations, further research should also aim to integrate the influence of sample size and its effects. A similar approach to our study, incorporating datasets with a broader range of observation counts, could be a good starting point.

# List of Figures

# List of Tables

# Listings

# Bibliography

[1] R. Mazza, *Introduction to Information Visualization*. Springer Science & Business Media, 2009.

[2] V. Ilievski, "The Importance of Interactive Data Visualization," https://medium.com/@ilievski.vladimir/the-importance-of-interactive-data-visualization-5e125cb04ce3, 2020, [Accessed 19-09-2024].

[3] H. Fung, "Data Storytelling: Making Sense of Complex, Multi-Dimensional Data with Parallel Coordinates Plots," https://medium.com/@hokifung_/data-storytelling-making-sense-of-complex-multi-dimensional-data-with-parallel-coordinates-plots-e73ddb1eb7f4, 2022, [Accessed 19-09-2024].

[4] D. Christodoulou, "Aspect Ratio," https://graphworkflow.com/enhancement/aspect/, 2019, [Accessed 19-09-2024].

[5] W. S. Cleveland, "A Model for Studying Display Methods of Statistical Graphics," *Journal of Computational and Graphical Statistics*, vol. 2, no. 4, pp. 323–343, 1993.

[6] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini, "How Deceptive are Deceptive Visualizations? An Empirical Analysis of Common Distortion Techniques," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1469–1478.

[7] R. Leonhart, *Lehrbuch Statistik: Einstieg und Vertiefung*. Hogrefe AG, 2022.

[8] Kmshilpamurali, "Understanding Variation: An Introduction to Measures of Variability," https://medium.com/@kmshilpamurali/understanding-variation-an-introduction-to-measures-of-variability-d8c174239ec2, 2024, [Accessed 19-09-2024].

[9] R. Webb, "12.1.2: Hypothesis Test for a Correlation," https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Mostly_Harmless_Statistics_(Webb)/12%3A_Correlation_and_Regression/12.01%3A_Correlation/12.1.02%3A_Hypothesis_Test_for_a_Correlation, 2023, [Accessed 19-09-2024].

[10] A. Demeusy, "Pearson correlation: Methodology, Limitations & Alternatives - Part 1: Methodology," https://medium.com/@anthony.demeusy/pearson-correlation-

methodology-limitations-alternatives-part-1-methodology-42abe8f1ba90, 2023, [Accessed 19-09-2024].

[11] M. Moran, “Selecting Between Parametric and Non-Parametric Analyses,” https://www.statisticssolutions.com/selecting-between-parametric-and-non-parametric-analyses/, 2024, [Accessed 19-09-2024].

[12] Alexander, “Homoscedasticity / Homogeneity of Variance/ Assumption of Equal Variance,” https://www.statisticshowto.com/homoscedasticity/, 2024, [Accessed 19-09-2024].

[13] Statistics How To, “The Difference Between Type I and Type II Errors in Statistics,” https://www.thoughtco.com/difference-between-type-i-and-type-ii-errors-3126414, 2024, [Accessed 19-09-2024].

[14] T. Vigen, “Spurious Correlations,” https://tylervigen.com/spurious-correlations, 2015, [Accessed 19-09-2024].

[15] A. L. Prinz, “Chocolate Consumption and Noble Laureates,” *Social Sciences & Humanities Open*, vol. 2, no. 1, p. 100082, 2020.

[16] M. Grandjean, “About Nobel Laureates and Chocolate, Correlations and Unreliable Data,” https://www.martingrandjean.ch/nobel-chocolate-correlation/, 2015, [Accessed 19-09-2024].

[17] F. B. Ortega, “The Intriguing Association Among Chocolate Consumption, Country’s Cconomy and Nobel Laureates,” *Clinical Nutrition*, vol. 32, no. 5, pp. 874–875, 2013.

[18] F. Dablander, *An Introduction to Causal Inference.* PsyArXiv, 2020.

[19] F. H. Messerli, “Chocolate consumption, Cognitive function, and Nobel Laureates,” *New England Journal of Medicine*, vol. 367, no. 16, pp. 1562–1564, 2012.

[20] ——, “Franz H. Messerli, MD: A Conversation With the Editor. Interview by William Clifford Roberts.” *American Journal of Cardiology*, vol. 96, no. 1, pp. 154–165, 2005.

[21] K. Van Stralen, F. Dekker, C. Zoccali, and K. Jager, “Confounding,” *Nephron Clinical Practice*, vol. 116, no. 2, pp. c143–c147, 2010.

[22] V. Das, “Confounding Variable and Spurious Correlation: Key Challenge in making Causal Inference,” https://vivdas.medium.com/confounding-variable-and-spurious-correlation-key-challenge-in-making-causal-inference-4e33d8ba60c2, 2020, [Accessed 19-09-2024].

[23] A. D. Bohr, D. D. Brown, K. R. Laurson, P. J. K. Smith, and R. W. Bass, “Relationship Between Socioeconomic Status and Physical Fitness in Junior High School Students,” *Journal of School Health*, vol. 83, no. 8, p. 542–547, 2013.

[24] M. Tory and T. Möller, “Rethinking Visualization: A High-Level Taxonomy,” in *IEEE Symposium on Information Visualization*, 2004, pp. 151–158.

[25] T.-M. Rhyne, M. Tory, T. Munzner, M. O. Ward, C. R. Johnson, and D. H. Laidlaw, “Information and Scientific Visualization: Separate but Equal or Happy Together at Last,” in *IEEE Visualization*. Seattle, 2003, pp. 611–614.

[26] L. Zhou, M. Fan, C. Hansen, C. R. Johnson, and D. Weiskopf, “A Review of Three-dimensional Medical Image Visualization,” *Health Data Science*, vol. 2022, p. 9840519, 2022.

[27] H. Hagen, A. Ebert, R. H. van Lengen, and G. Scheuermann, “Scientific Visualization: Methods and Applications,” in *Proceedings of the 19th Spring Conference on Computer Graphics*, 2003, pp. 23–33.

[28] B. Shneiderman, “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations,” in *The Craft of Information Visualization*. Elsevier, 2003, pp. 364–371.

[29] J. S. Yi, Y. Ah Kang, J. Stasko, and J. A. Jacko, “Toward a Deeper Understanding of the Role of Interaction in Information Visualization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[30] T. Hengl, P. Roudier, D. Beaudette, and E. Pebesma, “Plotkml: Scientific Visualization of Spatio-Temporal Data,” *Journal of Statistical Software*, vol. 63, pp. 1–25, 2015.

[31] H. M. Chen, “Information Visualization Principles, Techniques, and Software,” *Library Technology Reports*, vol. 53, no. 3, pp. 8–16, 2017.

[32] S. K. Card, J. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.

[33] C. North, “Information Visualization,” Center for Human-Computer Interaction, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA, Tech. Rep., 2005.

[34] M. O. Ward, “A Taxonomy of Glyph Placement Strategies for Multidimensional Data Visualization,” *Information Visualization*, vol. 1, no. 3-4, pp. 194–210, 2002.

[35] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, “Human-In-The-Loop Machine Learning: A State of the Art,” *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2023.

[36] J. Heer and B. Shneiderman, “Interactive Dynamics for Visual Analysis: A Taxonomy of Tools That Support the Fluent and Flexible Use of Visualizations,” *Queue*, vol. 10, no. 2, pp. 30–55, 2012.

[37] C. O. Wilke, “Fundamentals of Data Visualization,” https://clauswilke.com/dataviz/visualizing-associations.html, 2019, [Accessed 19-09-2024].

[38] Ibrahimkamran, “Understanding Univariate, Bivariate, and Multivariate Plots: A Visual Guide to Data Analysis,” https://medium.com/@ibrahimkamran490/understanding-univariate-bivariate-and-multivariate-plots-a-visual-guide-to-data-analysis-a4ebb5692819, 2024, [Accessed 19-09-2024].

[39] S. McLeod, “Box Plot Explained: Interpretation, Examples, & Comparison,” https://www.simplypsychology.org/boxplots, 2023, [Accessed 19-09-2024].

[40] M. Lanzenberger, S. Miksch, and M. Pohl, “Exploring Highly Structured Data: A Comparative Study of Stardinates and Parallel Coordinates,” in *Proceedings of the 9th International Conference on Information Visualisation, IV ’05*, 2005, pp. 312–320.

[41] D. Kaczynski, L. Wood, and A. Harding, “Using Radar Charts with Qualitative Evaluation: Techniques to Assess Change in Blended Learning,” *Active Learning in Higher Education*, vol. 9, no. 1, pp. 23–41, 2008.

[42] Finwiz, “S&P 500 Map,” https://finviz.com/map.ashx, 2024, [Accessed 19-09-2024].

[43] F. Harary and R. Z. Norman, “Graph Theory as a Mathematical Model in Social Science,” *University of Michigan, Institute for Social Research Ann Arbor*, 1953.

[44] S. A. Hale, “Multilinguals and Wikipedia Editing,” in *Proceedings of the 2014 ACM conference on Web science*, 2014, pp. 99–108.

[45] B. Kale, M. Sun, and M. E. Papka, “The State of the Art in Visualizing Dynamic Multivariate Networks,” *Computer Graphics Forum*, vol. 42, no. 3, pp. 471–490, 2023.

[46] F. Busato and N. Bombieri, *Graph Algorithms on GPUs.* Elsevier, 2017, p. 163–198.

[47] V. K. Chaudhri, C. Baru, N. Chittar, X. L. Dong, M. Genesereth, J. Hendler, A. Kalyanpur, D. B. Lenat, J. Sequeda, D. Vrandečić, and K. Wang, “Knowledge Graphs: Introduction, History, and Perspectives,” *AI Magazine*, vol. 43, no. 1, p. 17–29, 2022.

[48] A. J. C. Joseph D. Novak, “The Theory Underlying Concept Maps and How to Construct and Use Them,” Florida Institute for Human and Machine Cognition (IHMC), Tech. Rep., 2008.

[49] N. Andrienko and G. Andrienko, *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach.* Springer Science & Business Media, 2006.

[50] N. Pezzotti, “Dimensionality-Reduction Algorithms for Progressive Visual Analytics,” Ph.D. dissertation, Delft University of Technology, 2019.

[51] A. Inselberg, "The Plane with Parallel Coordinates," *The Visual Computer*, vol. 1, pp. 69–91, 1985.

[52] M. d'Ocagne, "Étude de Deux Systèmes Simples de Coordonnées Tangentielles Dans le Plan: Coordonnées Parallèles et Coordonnées Axiales," *Nouvelles Annales de Mathématiques: Journal des Candidats aux Écoles Polytechnique et Normale*, vol. 4, pp. 110–130, 1885.

[53] H. Gannett and F. W. Hewes, "General Summary, Showing the Rank of States, by Ratios, 1880," https://archive.org/details/dr_general-summary-showing-the-rank-of-states-by-ratios-1880-based-on-the-4521152, 1883, [Accessed 19-09-2024].

[54] A.-M. Guerry, *Essai Sur la Statistique Morale de la France*. Crochard, 1833.

[55] J. Heinrich and D. Weiskopf, "Parallel Coordinates for Multidimensional Data Visualization: Basic Concepts," *Computing in Science & Engineering*, vol. 17, no. 3, pp. 70–76, 2015.

[56] Y. Holtz and C. Healy, "Parallel Coordinates Plot," https://www.data-to-viz.com/graph/parallel.html, 2018, [Accessed 19-09-2024].

[57] P. V. Singh, "All About Min-Max Scaling," https://medium.com/@poojaviveksingh/all-about-min-max-scaling-c7da4e0044c5, 2023, [Accessed 19-09-2024].

[58] J. Heinrich and D. Weiskopf, "State of the Art of Parallel Coordinates," in *Eurographics 2013 - State of the Art Reports*, M. Sbert and L. Szirmay-Kalos, Eds. The Eurographics Association, 2013.

[59] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen, "Visual Clustering in Parallel Coordinates," in *Computer Graphics Forum*, vol. 27, no. 3, 2008, pp. 1047–1054.

[60] J. Johansson, P. Ljung, M. Jern, and M. Cooper, "Revealing Structure within Clustered Parallel Coordinates Displays," in *Proceedings of the IEEE Symposium on Information Visualization. INFOVIS '05.* IEEE, 2005, pp. 125–132.

[61] H. Siirtola and K.-J. Räihä, "Interacting with Parallel Coordinates," *Interacting with Computers*, vol. 18, no. 6, pp. 1278–1309, 2006.

[62] J. Bok, B. Kim, and J. Seo, "Augmenting Parallel Coordinates Plots With Color-Coded Stacked Histograms," *IEEE Transactions on Visualization & Computer Graphics*, vol. 28, no. 07, pp. 2563–2576, 2022.

[63] J. Li, J.-B. Martens, and J. J. Van Wijk, "Judging Correlation from Scatterplots and Parallel Coordinate Plots," *Information Visualization*, vol. 9, no. 1, pp. 13–30, 2010.

[64] J. Heer and M. Agrawala, "Multi-Scale Banking to 45 Degrees," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 701–708, 2006.

[65] M. Fink, J.-H. Haunert, J. Spoerhase, and A. Wolff, “Selecting the Aspect Ratio of a Scatter Plot Based on its Delaunay Triangulation,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2326–2335, 2013.

[66] J. Talbot, J. Gerth, and P. Hanrahan, “Arc Length-Based Aspect Ratio Selection,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2276–2282, 2011.

[67] Y. Wang, Z. Wang, C.-W. Fu, H. Schmauder, O. Deussen, and D. Weiskopf, “Image-Based Aspect Ratio Selection,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 840–849, 2018.

[68] J. Talbot, J. Gerth, and P. Hanrahan, “An Empirical Model of Slope Ratio Comparisons,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2613–2620, 2012.

[69] M. T. Thomas, *React in Action.* Simon and Schuster, 2018.

[70] Meta Platforms, Inc., “Reconciliation – React,” https://legacy.reactjs.org/docs/reconciliation.html, 2024, [Accessed 19-09-2024].

[71] S. Srivastava, H. Shukla, N. Landge, A. Srivastava, and D. Jindal, “A Comprehensive Review of Next.js Technology: Advancements, Features, and Applications,” *SSRN Electronic Journal*, 2024.

[72] Tailwind Labs Inc., “Install Tailwind CSS with Next.js - Tailwind CSS,” https://tailwindcss.com/docs/guides/nextjs, 2024, [Accessed 19-09-2024].

[73] shadcn, “Next.js,” https://ui.shadcn.com/docs/installation/next, 2024, [Accessed 19-09-2024].

[74] M. Bostock and Observable, Inc., “D3 by Observable | The JavaScript Library for Bespoke Data Visualization,” https://d3js.org, 2024, [Accessed 19-09-2024].

[75] E. Meeks, *D3.js in Action: Data Visualization with JavaScript.* Simon and Schuster, 2017.

[76] W. S. Cleveland and R. McGill, “Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods,” *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984.

[77] K. Abineshkumar, “Cars Data,” https://www.kaggle.com/datasets/abineshkumark/carsdata, 2017, [Accessed 19-09-2024].

[78] A. D. Khare, “Diabetes Dataset,” https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset, 2022, [Accessed 19-09-2024].

[79] F. Mehrparvar, “Liver Disorders,” https://www.kaggle.com/datasets/fatemehmehrparvar/liver-disorders, 2024, [Accessed 19-09-2024].

[80] Prisma Data Inc., “Is Prisma ORM an ORM? | What is an ORM? | Prisma Documentation,” https://www.prisma.io/docs/orm/overview/prisma-in-your-stack/is-prisma-an-orm, 2024, [Accessed 19-09-2024].

[81] pmndrs, “GitHub - pmndrs/zustand: Bear Necessities for State Management in React,” https://github.com/pmndrs/zustand, 2019, [Accessed 19-09-2024].

[82] R. Henson, “Analysis of Variance (ANOVA),” *Brain Mapping: an Encyclopedic Reference. Elsevier*, pp. 477–481, 2015.

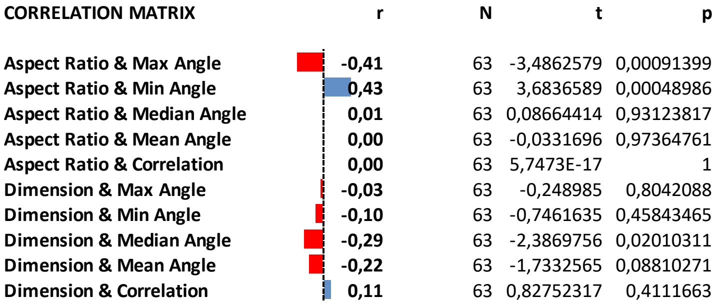
[83] L. Sullivan, “Hypothesis Testing-Analysis of Variance (ANOVA),” https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistesting-anova/bs704_hypothesistesting-anova_print.html, 2016, [Accessed 19-09-2024].

[84] S. Midway, M. Robertson, S. Flinn, and M. Kaller, “Comparing Multiple Comparisons: Practical Guidance for Choosing the Best Multiple Comparisons Test,” *PeerJ*, vol. 8, p. 10387, 2020.

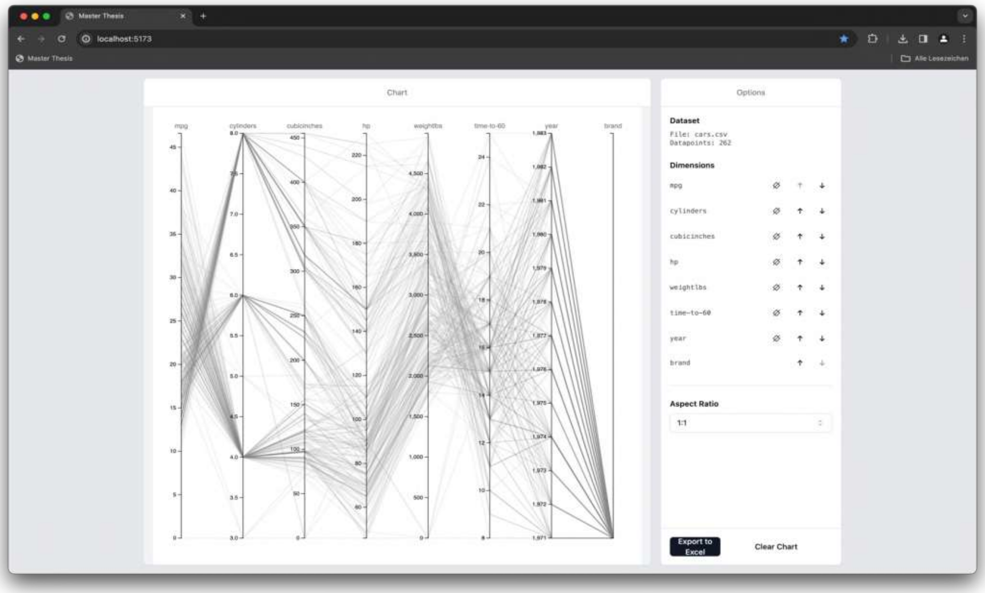
[85] D. C. Howell, *Statistical Methods for Psychology.* PWS-Kent Publishing Co, 1992.

# Appendices

# Statistical Analysis Results (Cars Dataset)



| CORRELATION MATRIX | r | N | t | p |
|---|---|---|---|---|
| Aspect Ratio & Max Angle | -0,41 | 63 | -3,4862579 | 0,00091399 |
| Aspect Ratio & Min Angle | 0,43 | 63 | 3,6836589 | 0,00048986 |
| Aspect Ratio & Median Angle | 0,01 | 63 | 0,08664414 | 0,93123817 |
| Aspect Ratio & Mean Angle | 0,00 | 63 | -0,0331696 | 0,97364761 |
| Aspect Ratio & Correlation | 0,00 | 63 | 5,7473E-17 | 1 |
| Dimension & Max Angle | -0,03 | 63 | -0,248985 | 0,8042088 |
| Dimension & Min Angle | -0,10 | 63 | -0,7461635 | 0,45843465 |
| Dimension & Median Angle | -0,29 | 63 | -2,3869756 | 0,02010311 |
| Dimension & Mean Angle | -0,22 | 63 | -1,7332565 | 0,08810271 |
| Dimension & Correlation | 0,11 | 63 | 0,82752317 | 0,4111663 |



**Dataset**

```
File: cars.csv
Datapoints: 262
```

**Dimensions**

mpg

cylinders

cubicinches

hp

weightlbs

time-to-60

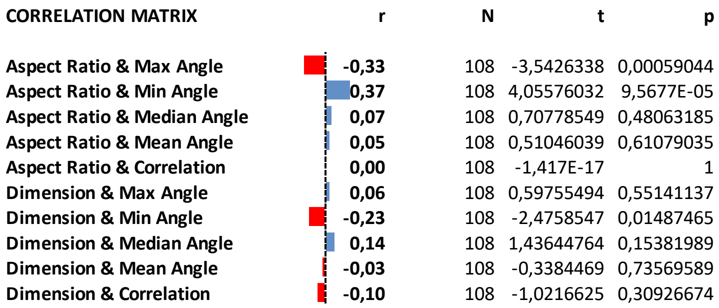
year

brand

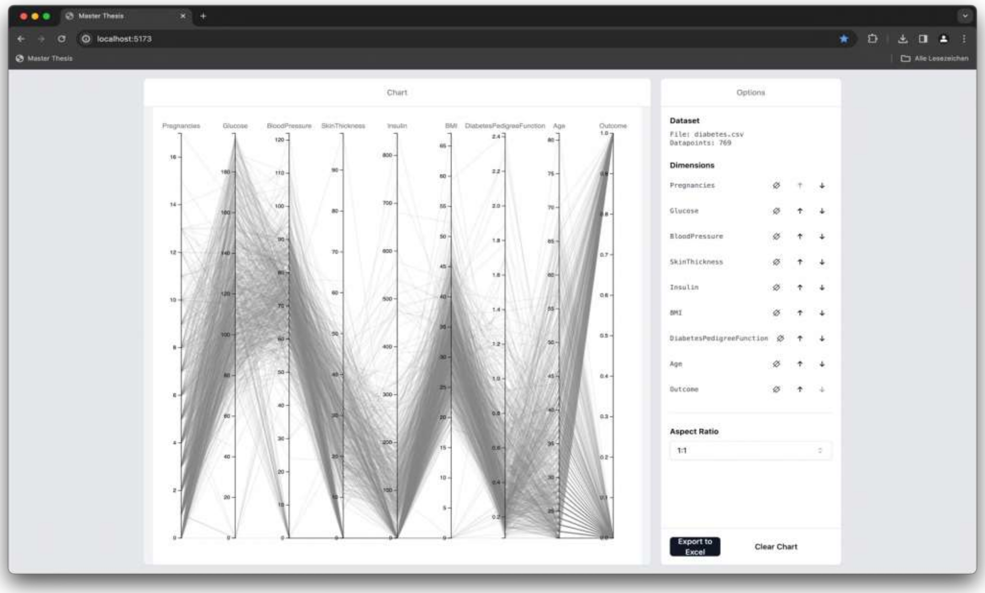
**Aspect Ratio**

1:1

# Statistical Analysis Results (Diabetes Dataset)



| CORRELATION MATRIX | r | N | t | p |
|---|---|---|---|---|
| Aspect Ratio & Max Angle | -0,33 | 108 | -3,5426338 | 0,00059044 |
| Aspect Ratio & Min Angle | 0,37 | 108 | 4,05576032 | 9,5677E-05 |
| Aspect Ratio & Median Angle | 0,07 | 108 | 0,70778549 | 0,48063185 |
| Aspect Ratio & Mean Angle | 0,05 | 108 | 0,51046039 | 0,61079035 |
| Aspect Ratio & Correlation | 0,00 | 108 | -1,417E-17 | 1 |
| Dimension & Max Angle | 0,06 | 108 | 0,59755494 | 0,55141137 |
| Dimension & Min Angle | -0,23 | 108 | -2,4758547 | 0,01487465 |
| Dimension & Median Angle | 0,14 | 108 | 1,43644764 | 0,15381989 |
| Dimension & Mean Angle | -0,03 | 108 | -0,3384469 | 0,73569589 |
| Dimension & Correlation | -0,10 | 108 | -1,0216625 | 0,30926674 |



Dataset

File: diabetes.csv
Datapoints: 769

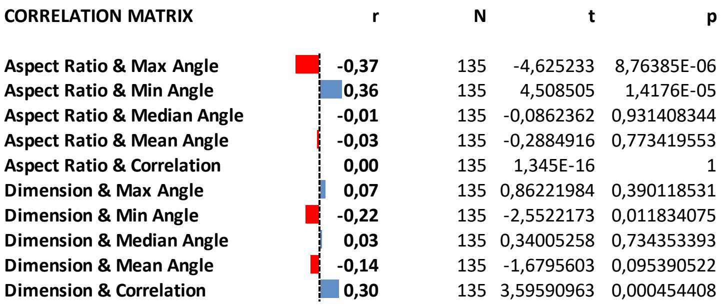
Dimensions

Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome

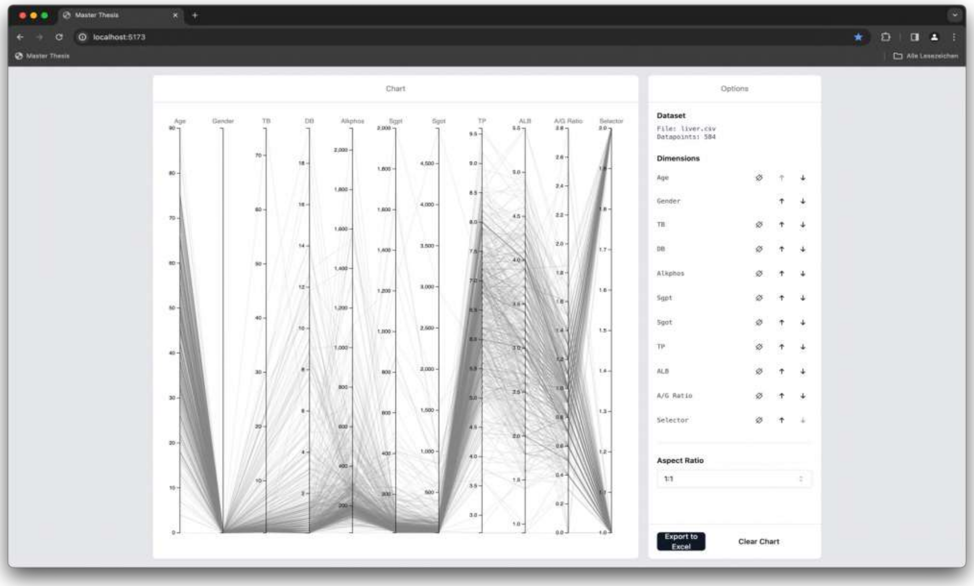
Aspect Ratio

1:1

# Statistical Analysis Results (Liver Dataset)



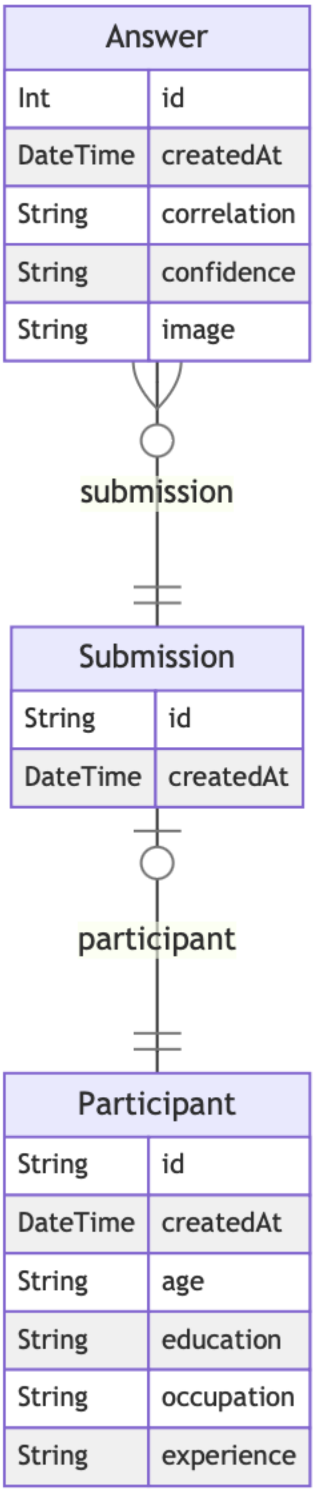
| CORRELATION MATRIX | r | N | t | p |
|---|---|---|---|---|
| Aspect Ratio & Max Angle | -0,37 | 135 | -4,625233 | 8,76385E-06 |
| Aspect Ratio & Min Angle | 0,36 | 135 | 4,508505 | 1,4176E-05 |
| Aspect Ratio & Median Angle | -0,01 | 135 | -0,0862362 | 0,931408344 |
| Aspect Ratio & Mean Angle | -0,03 | 135 | -0,2884916 | 0,773419553 |
| Aspect Ratio & Correlation | 0,00 | 135 | 1,345E-16 | 1 |
| Dimension & Max Angle | 0,07 | 135 | 0,86221984 | 0,390118531 |
| Dimension & Min Angle | -0,22 | 135 | -2,5522173 | 0,011834075 |
| Dimension & Median Angle | 0,03 | 135 | 0,34005258 | 0,734353393 |
| Dimension & Mean Angle | -0,14 | 135 | -1,6795603 | 0,095390522 |
| Dimension & Correlation | 0,30 | 135 | 3,59590963 | 0,000454408 |



Dataset
File: liver.csv
Datapoints: 584

Aspect Ratio
1:1

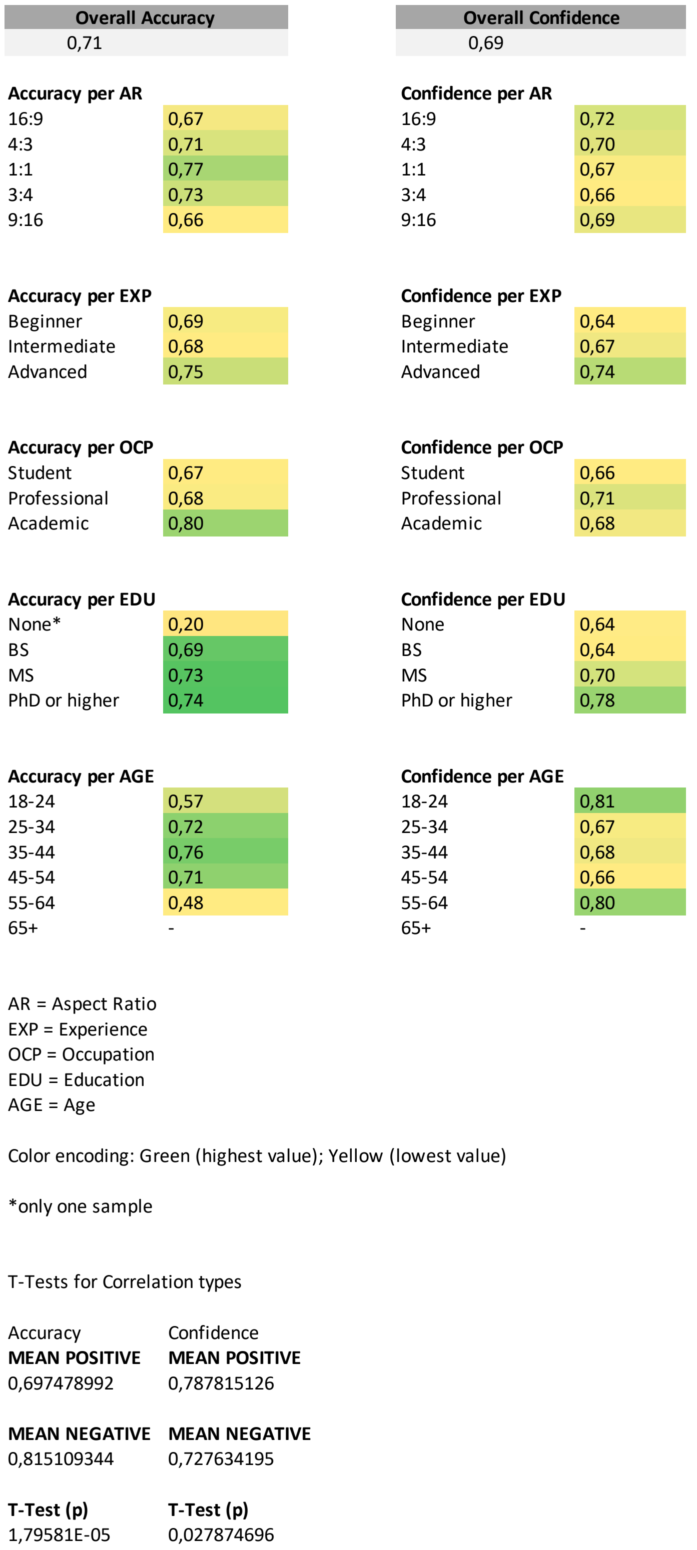
## Prisma ERD (User Study)



# User Study Participants

| Submission ID | Participant Occupation | Participant Higher Education | Participant Experience | Participant Age |
|---|---|---|---|---|
| vW3diM2TXzC2DqGgmMruF | Student | MS | Intermediate | 18-24 |
| k7VU7GOTSpubLt_xyC2p_ | Academic | MS | Intermediate | 25-34 |
| StKDApFyBpBIY_4e7WeNa | Professional | MS | Advanced | 35-44 |
| gSZrWhSgzxZNixJ2pIWEE | Professional | MS | Intermediate | 25-34 |
| nBEUULFfrrAxzEYqh09TO | Professional | MS | Beginner | 25-34 |
| tSDlqZvVWK8LmnWWqNUSv | Professional | MS | Advanced | 35-44 |
| gSThpIgRtEo3vAX5YqTx_ | Academic | MS | Intermediate | 45-54 |
| HS1Lf3ICZ9L2l_h1vS-3n | Academic | MS | Intermediate | 25-34 |
| 7lCFp9tnM6UX_RfrETy9G | Other | PhD or higher | Advanced | 35-44 |
| iz56PMP0h0ogogI7XXIKd | Student | BS | Beginner | 25-34 |
| mBhKdE5SrVoUGstMNOEVc | Professional | PhD or higher | Intermediate | 55-64 |
| p7KfCpMG82N30IcrVIY-5 | Academic | MS | Advanced | 35-44 |
| RjCjiC4-rPSbmg8kUXouE | Professional | MS | Beginner | 25-34 |
| szHnQWByGa2bne80a6QhW | Professional | PhD or higher | Advanced | 35-44 |
| d5tOfoa92gPBlbHz7SnQx | Academic | PhD or higher | Advanced | 18-24 |
| netRBaSsl_FVye0P6tEfD | Professional | MS | Beginner | 25-34 |
| TASD5_5aUGTLzYWkZifQ4 | Student | MS | Intermediate | 35-44 |
| h6fyVxVCymZBGTDUdoF1L | Professional | MS | Intermediate | 45-54 |
| zin1JcgkrUJlDA1YGvFPA | Student | BS | Intermediate | 25-34 |
| 7dgv38DZ-vGPXDzW7QgLd | Professional | MS | Beginner | 25-34 |
| l1S1AstXxK698khZXhLxz | Student | BS | Intermediate | 25-34 |
| 7AGs_9e8Ab6maSqgpn-xp | Student | BS | Beginner | 25-34 |
| yg8O0IXY9JCH6gpmPG3RJ | Student | BS | Intermediate | 25-34 |
| ZRHi7IdFmPNLSY7K48NLN | Student | BS | Beginner | 25-34 |
| BbFMxgzjfGf6HqKGMN8WK | Professional | MS | Beginner | 35-44 |
| OmcKjy-OOCr-MCSp9a4Hr | Student | MS | Advanced | 25-34 |
| PuN2bv4Ad3f8gv1XJpj5c | Professional | BS | Beginner | 25-34 |
| TpEO0DrbtqcIs3MA5ebQp | Professional | BS | Intermediate | 25-34 |
| TBoiVcUHBQDHHD5ny-FjW | Academic | MS | Advanced | 25-34 |
| prxqktCiEqAmNK4QMqrZa | Academic | PhD or higher | Advanced | 25-34 |
| oESe26shyXjn-pel7oJhl | Academic | MS | Intermediate | 25-34 |
| ZH3bwOj9e74NRVhvXELdF | Academic | MS | Advanced | 25-34 |
| iVlI83PB3BCbIbbjwqElU | Student | None | Intermediate | 18-24 |
| Ow6-ZukebMu-Fe6wUp87m | Professional | MS | Beginner | 25-34 |
| FmNzl-tGE-3tRUPp6E4NB | Professional | BS | Advanced | 45-54 |
| x0OsO-J1DRKQpiKnCwmhn | Academic | PhD or higher | Advanced | 25-34 |
| 8_TwFrwIQCCLf_hio7vcy | Academic | MS | Advanced | 25-34 |
| i3uIJxhv8CNMNP-ekBreJ | Professional | MS | Beginner | 25-34 |
| nhw8yydLRamFdmrm-rIPr | Other | BS | Beginner | 25-34 |
| 57CQrODrygx7IEqieVxXh | Academic | BS | Beginner | 25-34 |
| y9cjTe2gVQzHkHM3Eo32s | Student | BS | Beginner | 25-34 |
| YEqXJo9U6PF4fYp6zT3zy | Professional | BS | Beginner | 25-34 |
| 0clgyr3-DAkjYRKdmtLs_ | Professional | BS | Intermediate | 18-24 |
| V31MZmlayzjgYOapT8bV6 | Professional | BS | Beginner | 25-34 |
| sXtV5x5bat1Afeup1g9QA | Professional | MS | Advanced | 35-44 |
| l2XwUpdbw_bx4lwvda0rS | Professional | PhD or higher | Advanced | 45-54 |
| lbXoWVuZmiEMPsO3imG7z | Professional | BS | Beginner | 25-34 |
| 6P3qWUe0eyi4b85WTLsV5 | Student | BS | Intermediate | 25-34 |
| HS-wF_dpKftwkiNTCw-Gq | Professional | MS | Advanced | 25-34 |
| BjWProM49Hb5_yWKFs-Ql | Student | MS | Advanced | 35-44 |
| 1Ike47jfP-evCqddxrva4 | Academic | MS | Advanced | 25-34 |
| wBtfHECMKs59ZytWfJmYb | Professional | BS | Intermediate | 45-54 |
| HwYR7GJUCNdtYM7fVRJLR | Academic | MS | Advanced | 25-34 |
| ASEbpdK2LzNKnJeSY0Rrv | Student | BS | Intermediate | 25-34 |
| GkCYcnwOUWPAIDjmaFKLq | Professional | MS | Advanced | 25-34 |
| d7MvgGVN91-zbiCe8dqFT | Student | BS | Intermediate | 25-34 |
| 06vyKIzc0WBViFBD8qxo3 | Academic | MS | Advanced | 25-34 |

# User Study Results (Means)



| Overall Accuracy | | | Overall Confidence | |
|---|---|---|---|---|
| 0,71 | | | 0,69 | |

| Accuracy per AR | | | Confidence per AR | |
|---|---|---|---|---|
| 16:9 | 0,67 | | 16:9 | 0,72 |
| 4:3 | 0,71 | | 4:3 | 0,70 |
| 1:1 | 0,77 | | 1:1 | 0,67 |
| 3:4 | 0,73 | | 3:4 | 0,66 |
| 9:16 | 0,66 | | 9:16 | 0,69 |

| Accuracy per EXP | | | Confidence per EXP | |
|---|---|---|---|---|
| Beginner | 0,69 | | Beginner | 0,64 |
| Intermediate | 0,68 | | Intermediate | 0,67 |
| Advanced | 0,75 | | Advanced | 0,74 |

| Accuracy per OCP | | | Confidence per OCP | |
|---|---|---|---|---|
| Student | 0,67 | | Student | 0,66 |
| Professional | 0,68 | | Professional | 0,71 |
| Academic | 0,80 | | Academic | 0,68 |

| Accuracy per EDU | | | Confidence per EDU | |
|---|---|---|---|---|
| None* | 0,20 | | None | 0,64 |
| BS | 0,69 | | BS | 0,64 |
| MS | 0,73 | | MS | 0,70 |
| PhD or higher | 0,74 | | PhD or higher | 0,78 |

| Accuracy per AGE | | | Confidence per AGE | |
|---|---|---|---|---|
| 18-24 | 0,57 | | 18-24 | 0,81 |
| 25-34 | 0,72 | | 25-34 | 0,67 |
| 35-44 | 0,76 | | 35-44 | 0,68 |
| 45-54 | 0,71 | | 45-54 | 0,66 |
| 55-64 | 0,48 | | 55-64 | 0,80 |
| 65+ | - | | 65+ | - |

AR = Aspect Ratio
EXP = Experience
OCP = Occupation
EDU = Education
AGE = Age

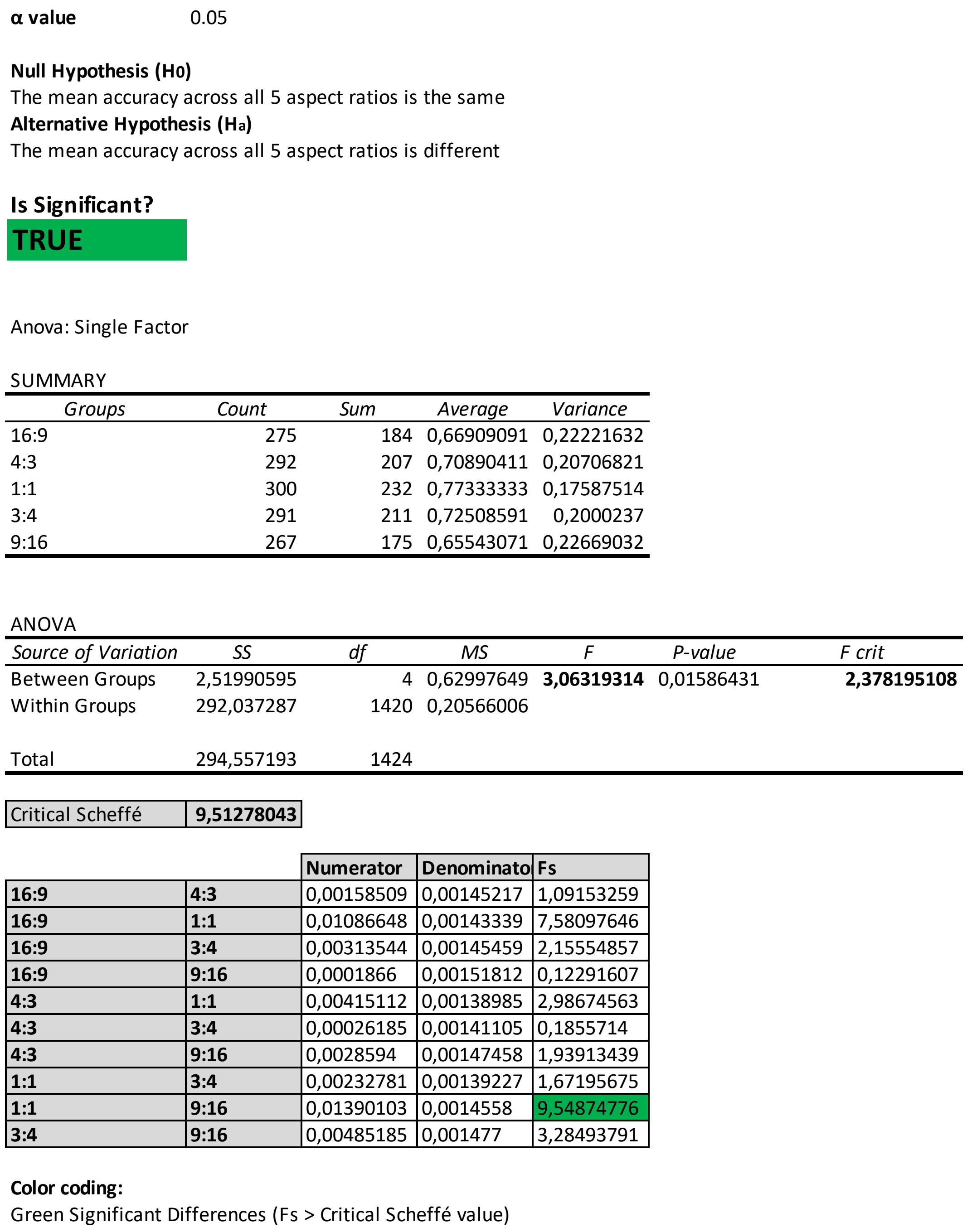
Color encoding: Green (highest value); Yellow (lowest value)

*only one sample

T-Tests for Correlation types

| Accuracy | Confidence |
|---|---|
| **MEAN POSITIVE** | **MEAN POSITIVE** |
| 0,697478992 | 0,787815126 |
| | |
| **MEAN NEGATIVE** | **MEAN NEGATIVE** |
| 0,815109344 | 0,727634195 |
| | |
| **T-Test (p)** | **T-Test (p)** |
| 1,79581E-05 | 0,027874696 |

# User Study Results (ANOVA Accuracy)



**α value** 0.05

**Null Hypothesis (H0)**
The mean accuracy across all 5 aspect ratios is the same
**Alternative Hypothesis (Ha)**
The mean accuracy across all 5 aspect ratios is different

**Is Significant?**
**TRUE**

Anova: Single Factor

SUMMARY

| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
|---|---|---|---|---|
| 16:9 | 275 | 184 | 0,66909091 | 0,22221632 |
| 4:3 | 292 | 207 | 0,70890411 | 0,20706821 |
| 1:1 | 300 | 232 | 0,77333333 | 0,17587514 |
| 3:4 | 291 | 211 | 0,72508591 | 0,2000237 |
| 9:16 | 267 | 175 | 0,65543071 | 0,22669032 |

ANOVA

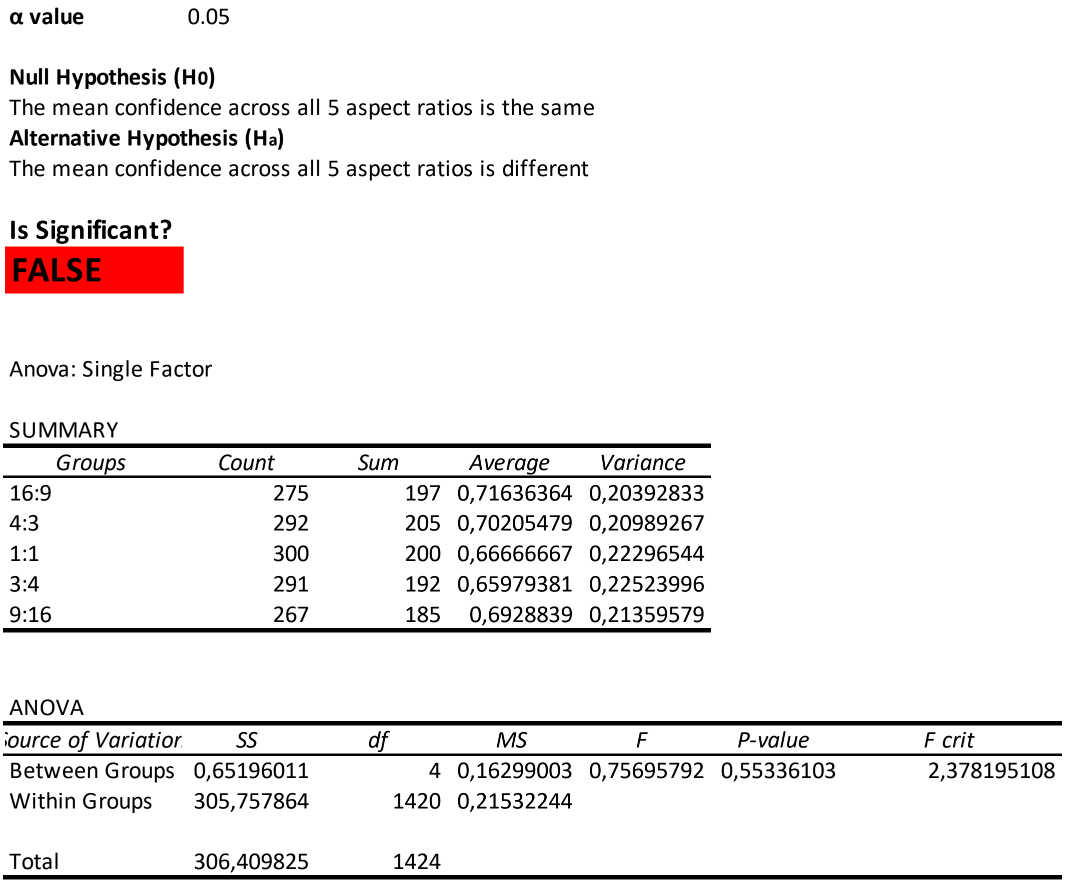
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
|---|---|---|---|---|---|---|
| Between Groups | 2,51990595 | 4 | 0,62997649 | **3,06319314** | 0,01586431 | **2,378195108** |
| Within Groups | 292,037287 | 1420 | 0,20566006 | | | |
| | | | | | | |
| Total | 294,557193 | 1424 | | | | |

| Critical Scheffé | **9,51278043** |
|---|---|

| | | **Numerator** | **Denominato** | **Fs** |
|---|---|---|---|---|
| **16:9** | **4:3** | 0,00158509 | 0,00145217 | 1,09153259 |
| **16:9** | **1:1** | 0,01086648 | 0,00143339 | 7,58097646 |
| **16:9** | **3:4** | 0,00313544 | 0,00145459 | 2,15554857 |
| **16:9** | **9:16** | 0,0001866 | 0,00151812 | 0,12291607 |
| **4:3** | **1:1** | 0,00415112 | 0,00138985 | 2,98674563 |
| **4:3** | **3:4** | 0,00026185 | 0,00141105 | 0,1855714 |
| **4:3** | **9:16** | 0,0028594 | 0,00147458 | 1,93913439 |
| **1:1** | **3:4** | 0,00232781 | 0,00139227 | 1,67195675 |
| **1:1** | **9:16** | 0,01390103 | 0,0014558 | 9,54874776 |
| **3:4** | **9:16** | 0,00485185 | 0,001477 | 3,28493791 |

**Color coding:**
Green Significant Differences (Fs > Critical Scheffé value)

# User Study Results (ANOVA Confidence)



**α value** 0.05

**Null Hypothesis ($H_0$)**
The mean confidence across all 5 aspect ratios is the same
**Alternative Hypothesis ($H_a$)**
The mean confidence across all 5 aspect ratios is different

**Is Significant?**

**FALSE**

Anova: Single Factor

SUMMARY

| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
|---|---|---|---|---|
| 16:9 | 275 | 197 | 0,71636364 | 0,20392833 |
| 4:3 | 292 | 205 | 0,70205479 | 0,20989267 |
| 1:1 | 300 | 200 | 0,66666667 | 0,22296544 |
| 3:4 | 291 | 192 | 0,65979381 | 0,22523996 |
| 9:16 | 267 | 185 | 0,6928839 | 0,21359579 |

ANOVA

| *ource of Variatior* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
|---|---|---|---|---|---|---|
| Between Groups | 0,65196011 | 4 | 0,16299003 | 0,75695792 | 0,55336103 | 2,378195108 |
| Within Groups | 305,757864 | 1420 | 0,21532244 | | | |
| | | | | | | |
| Total | 306,409825 | 1424 | | | | |