Data Visualisation: Theory and Practice

by Yujie Chu, Pia Fullaondo, Qinqing Li, Jacko Zhou

The School of Mathematics
Undergraduate Mathematics Project

Supervised by
Dr. Miguel de Carvalho

Own Work Declaration

We declare that this paper was composed by ourselves, that the work contained herein is our own except where explicitly stated otherwise, and that this work has not been submitted for any other degree or professional qualification except as specified.

Abstract

This project aims to provide the theory of graphical statistics methods and their visual implementations using R, utilising open-source datasets from different sectors. The authors recognise that graphical statistics is an active area of research, therefore textbook methods and state-of-the-art methods will be covered. Chapters 3 to 5 concentrate on standard implementations of graphical statistics methods for univariate, bivariate, and multivariate data, respectively. Chapter 6 focuses on functional boxplots and Q–Q boxplots, which were published in the *Journal of Computational and Graphical Statistics* in 2011 and 2022, respectively.

Contents

1 Introduction

This paper is motivated by the enduring significance and power of data visualisation throughout history. This chapter contextualises the importance of the field by showcasing some of its influential applications, as well as its potentially dangerous misuses. Drawing from BBC Ideas [23], we analyse case examples to underscore both the power and pitfalls of data visualisations. Additionally, essential R packages including *ggplot2* are introduced, along with the datasets used throughout paper.

Literature and Research
To appreciate the significance and impact of data visualisation, it is imperative to recognise its prominence within various method-specific publications. Notably, the *Journal of Computational and Graphical Statistics* stands out as a significant source of ongoing research in the realm of computational and graphical methods within statistics. This prestigious journal publishes ongoing research on the latest techniques in computational and graphical methods in statistics, encompassing data analysis and numerical graphical displays. Additionally, the *Journal of Statistical Software* publishes peer-reviewed articles about statistical software, together with the source code. Collectively, these publications reveal the vast potential for current and future research in this ever-evolving domain, particularly as our digital landscape and data sphere rapidly expand.

1.1 Historical Background and Misuses of Data Visualisation

This section analyses two case studies showcasing the effectiveness and influence of data visualisation, followed by two additional case studies illustrating instances of problematic data visualisation misuse. The primary aim is to emphasise the importance of data visualisation while stressing the need for careful attention to detail, especially considering the potential for malicious exploitation of visualised data.

Motivations for Using Data Visualisations — Case 1

Florence Nightingale was not only a social reformer and the founder of modern nursing but also a pioneering statistician. It was her application of data visualisation during the Crimean War that transformed the field of healthcare and pushed for social reform.

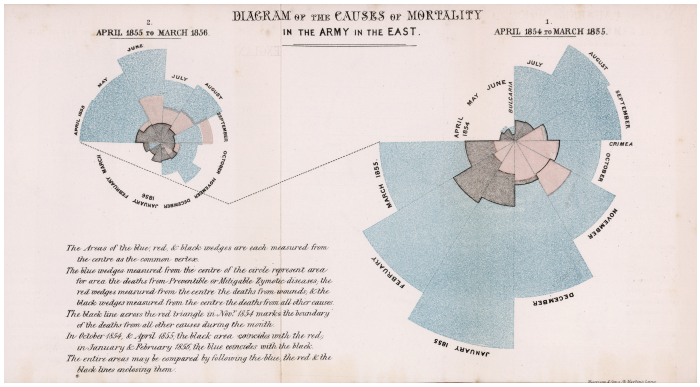

Figure 1: "Diagram of the causes of mortality in the army in the East" (1858) by Florence Nightingale

During the Crimean War, Nightingale recognised that unsanitary hospital conditions were claiming more lives than the battlefield itself. With the help of William Farr, Nightingale created the coxcomb aimed to illustrate the toll of preventable mortality on soldiers, as shown in Figure 1 [20]. The coxcomb, resembling an unconventional pie chart, partitioned mortality by causes. Blue indicates preventable deaths, red indicates deaths by wounds, and black indicates other causes. The blue areas outweighed the red and black sections combined, highlighting the disproportionate impact of unsanitary hospital conditions on the mortality rate.

Nightingale leveraged the compelling visualisations in her advocacy efforts, presenting them to members of Parliament and government officials who otherwise were unlikely to read or understand statistical reports. She successfully persuaded Queen Victoria, head of the British Army at the time, to allocate funding for the improvement of military hospitals.

Motivations for Using Data Visualisations — Case 2
Sometimes, one glance is enough to convey a powerful idea. Edward Hawkins, a British climate scientist and Professor of climate science at the University of Reading, is renowned for his exceptional data visualisations of climate change.

In 2018, Edward Hawkins was invited to deliver a lecture on climate change in Wales to an audience with diverse backgrounds. It was important to effectively convey the growing urgency surrounding global warming. To achieve this, he created a chart that used only colours, without any words, titles, or legends, as shown in Figure 2 [15]. This seemingly simple, yet remarkably powerful chart visually illustrated the Earth's warming trend since 1850.

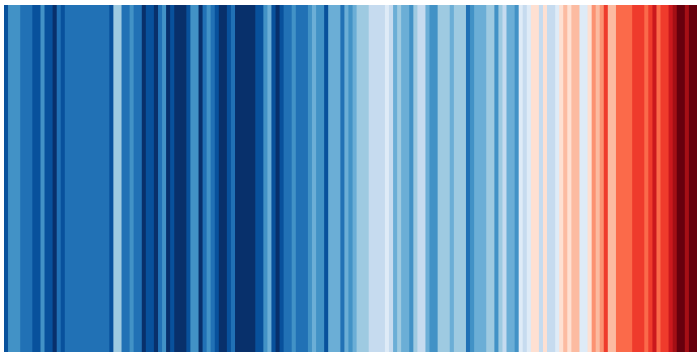

Figure 2: "Latest global stripes" (1850-2020) by Edward Hawkins

Known as the "warming stripes," this chart cleverly employs blues to indicate cooler-than-average years and reds to signify hotter-than-average years. Its influence reached far and wide, gracing the front pages of major media outlets and featured in news broadcasts worldwide. It became a symbol in climate change demonstrations. Arguably, it stands as one of the most iconic graphics in modern times.

Misuses of Data Visualisation — Case 1
Having observed the remarkable effectiveness of data visualisations, the significance of employing them correctly becomes apparent. The improper use of data visualisations holds the potential to significantly influence the public in misleading ways, resulting in undesired consequences.

In fact, inappropriate data visualisation can conceal trends rather than reveal them. Figure 3 [6] illustrates an instance of this issue. On the left-hand side of the figure, an inappropriate scale is used — the y-scale ranges from 0 to 30 million dollars. In this way, the fluctuations in payroll spending are obscured. Conversely, on the right-hand side, a significant increase of over 500,000 dollars in just two months can be observed due to the use of a different scale. This revelation is substantial; considering inflation, 500,000 dollars in 1937 is worth well over 10 million dollars today [1].

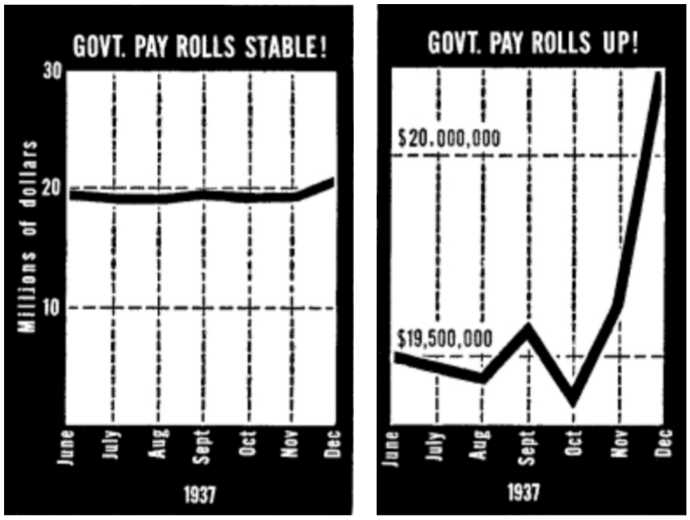

Figure 3: Incomplete Data Analysis by Miguel de Carvalho, "How to lie with statistics" [17]

Thus, the scale used in graphs serves as a critical tool, enabling the clear representation of data. However, it also holds the potential to mislead if not employed appropriately. The manipulation of scales can distort the interpretation of data, leading to misrepresentations of reality.

Misuses of Data Visualisation — Case 2
One striking example of data visualisation misuse is found in the Kallikak Family tree — one of the most prominent eugenic narratives of the 20th century.

The visualisation shown in Figure 4 [4], was created by the psychologist Henry Goddard and presented in his 1912 book, "The Kallikak Family: A Study in the Heredity of Feeble-Mindedness." Goddard's narrative centered around Martin Kallikak, a soldier who, in addition to his marriage to a respected citizen, had a one-night stand with a "feeble-minded" maid. Goddard believed that intellectual disabilities were inherited traits. In Goddard's account, the legitimate family was successful, while the children of the "feeble-minded" maid were labeled as "the lowest types of human beings." However, research has since revealed that the entire story was fictitious, as there was no record of the maid's existence [30].

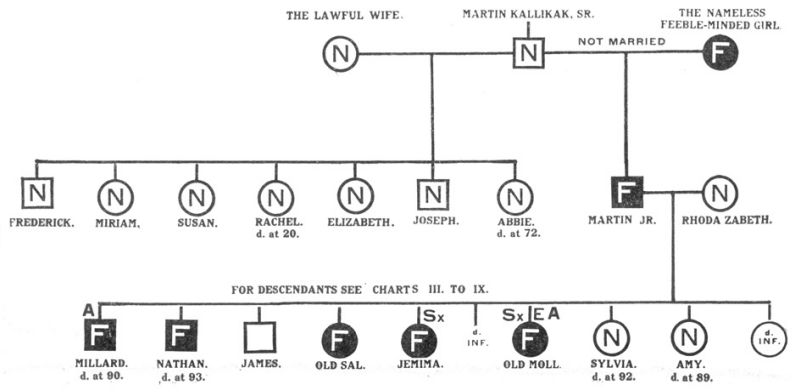

Figure 4: The Kallikak Family tree (1912) by Henry Goddard

Regrettably, the Kallikak family tree became a central element in the eugenics movement. Figure 4 was featured in the 1935 Nazi propaganda film "Das Erbe" (The Inheritance), which was used to promote public acceptance of Nazi eugenics laws. This propaganda laid the groundwork for the forced sterilisation of approximately 400,000 people under Nazi eugenics policies [27].

1.2 Computing and Data Visualisation

Over the years, numerous tools have been developed to facilitate the generation of data visualisation. The programme R stands out as a powerful tool for statistical analysis and graphic creation, offering a wide array of functionalities to cater to various data visualisation needs. We will first introduce the commonly used R package *ggplot2*. Then, other key packages used to produce clear and compelling graphs in R will be presented. In this project, R package names such as *ggplot2* are italicised, while R commands like `faceting` are displayed in a teletype font.

The *ggplot2* Package

The *ggplot2* package is the main tool in R used to create plots representing datasets of various natures. The foundation of *ggplot2* lies in Leland Wilkinson's *The Grammar of Graphics*, allowing the sequential construction of the individual elements composing a graph [39]. These individual elements are combined to create a unified graphical representation.

This R package, *ggplot2*, is notable for its robustness and flexibility, enabling the creation of customised graphics rather than adhering strictly to pre-set options. Despite the initial learning curve, *ggplot2* is crafted to be user-friendly, offering sensible defaults and an iterative method for constructing plots. Its focus lies in uncovering the underlying message within the data. Furthermore, *ggplot2* is structured to support layered and annotated graphics that enhance data analysis, especially for beginners and those new to data exploration.

Key *ggplot2* Commands

As shown in subsequent sections, it is possible to produce visualisations rather effortlessly by using internal data from R, or even external data, together with *ggplot2*. When using this package, certain key tools are used repeatedly to create visual plots from datasets. Some of these integrals tools for

creating sophisticated visualisations are presented subsequently [38]:

Firstly, the `geom` function refers to "geometric objects". It is pivotal took in *ggplot2* used to specify the type of geometric objects or shapes to be drawn on a plot. Additionally, the `scales` function enables the adjustment of various mapping details, including colour choices, label formatting, legend arrangement, and more.

Furthermore, the `coord` function provides the axes and gridlines which aid in structuring and interpreting the graph. Various coordinate systems such as cartesian, polar, and map projections are available. Moreover, `faceting` is a robust feature enabling the partitioning of a single plot into multiple plots based on factors present in the dataset. This functionality proves particularly beneficial for exploring and presenting data with multiple groups or categories.

The `theme` function holds significance in customising the non-data elements of plots. The theme system within *ggplot2* allows precise adjustment of aesthetic aspects such as fonts, labels, legends, and background colours. This tool is essential for enhancing plot readability and creating visually captivating graphics tailored to specific audiences or publication requirements.

key Additional Packages

The R language, on top of the *ggplot2*, has a wide array of packages facilitating diverse functions. Throughout this project, various packages will be used, each enhancing the analysis and presentation of data.

The *tidyverse* is collection of R packages designed for data science and manipulation. It includes several packages that work seamlessly together, providing a consistent approach to data processing, visualisation, and analysis.

Some of the key packages included in the collection are widely used packages such as *ggplot2* and *dplyr*. Particularly, the *dplyr* package plays a pivotal role in data manipulation and transformation. It enables the filtering, sorting, summarising, and transforming of datasets, thereby streamlining data analysis and management.

Moreover, recent research developments in graphical methods, as discussed in the *Journal of Computational and Graphical Statistics* such as t-SNE and Q–Q boxplots. The *Rtsne* and *qqboxplot* packages have allowed us to implement these methods in the report.

1.3 Datasets

This section provides an overview of the different datasets analysed and visualised in the subsequent chapters.

mtcars: The dataset was extracted from the 1974 Motor Trend US magazine. It includes the fuel consumption and 10 aspects of automobile design for 32 automobiles (1973–74 models). (built-in R)

iris: The dataset was collected by Edgar Anderson. This dataset was famously used by British statistician Ronald Fisher to demonstrate linear discriminant analysis in 1936. It encompasses 5 flower characteristics for 3 types of iris, each with 50 samples. (built-in R)

gapminder: The dataset was provided by the Gapminder Foundation. It encapsulates key demographic statistics like life expectancy, GDP per capita, and population across various countries spanning multiple years, illustrating global development trends.

palmerpenguins: This dataset contains biometric measurements from three distinct penguin species inhabiting the Palmer Archipelago in Antarctica, namely the Adélie, Chinstrap, and Gentoo penguins. (built-in R)

ToothGrowth: The dataset measures the tooth growth of 60 guinea pigs. Each animal was either administered with vitamin C or orange juice. (built-in R)

Fire in Brazil: Open-source fire observation data is provided by NASA. The analysis focuses on Brazil (2013-2022). Note that the dataset contains the variable "confidence", ranging from 0% to 100%. The dataset was filtered with a confidence level of $\geq 95\%$ to ensure an accurate account of fire occurrences [7].

Exchange Rate: The exchange rate data from the Bank of England provides daily spot exchange rates against the pound Sterling from 2005 to the present. This report examines the daily spot rates of the Canadian Dollar, Euro, and US Dollar against the pound Sterling.

Metabolic Syndrome Data: The oxygen saturation in blood, for a healthy population of 80 women and a diseased population of 35 women. These data were collected through a survey conducted in North-West Spain. The dataset is accessible via DATAstudio, a collection of datasets from research articles by Miguel de Carvalho.

Taipei Housing: The historical real estate valuation from Sindian Dist., New Taipei City, Taiwan, available at UC Irvine Machine Learning Repository.

covid_test_scores: The synthetic dataset for simulating COVID-19 Diagnostic Test Scores consists outcomes of two different diagnostic tests for COVID-19, referred to as Score_Test1 and Score_Test2. The scores are continuous variables that represent the likelihood of a positive diagnosis, with higher scores indicating a higher probability of infection. The dataset includes a Condition column, which indicates the true condition of each simulated individual, with 1 representing a positive COVID-19 case and 0 representing a negative case.

2 Theoretical Foundations of Data Visualisation

This chapter delves into the core principles and concepts that serve as the base of the field of data visualisation. We seek to understand not only the "how" but also the "why" behind the creation of visualisations that captivate and inform.

2.1 Introduction to Data Visualisation Theory

Creating effective data visualisations requires familiarity with the robust theoretical framework underlying every chart, graph, or plot. These theoretical underpinnings not only form the basis of data visualisation, but also shape the way in which we represent, perceive, understand, and interpret data.

Guiding Principles for Data Visualisation
The theoretical framework of data visualisation involves guiding principles dictating the visual representation of data. These principles include accuracy, emphasising faithful reflection of underlying data to reduce distortion or misinterpretation; simplicity, advocating for streamlined visuals to convey information effectively; clarity, ensuring visuals are easily understood without unnecessary complexity; relevance, presenting information pertinent to the message or question addressed; and consistency, maintaining uniform use of visual elements like colour coding and labeling throughout a visualisation [14].

Theoretical Framework and Visual Perception
Furthermore, understanding how the human brain processes visual information is a fundamental aspect of data visualisation theory. This knowledge plays a crucial role in designing visualisations that effectively connect with viewers. It encompasses several key considerations which will be studied subsequently: the Gestalt Principles, concerned with how visual elements are grouped and interpreted; Colour Theory, involving the strategic use of colour contrasts and harmonies to improve clarity; and the management of Cognitive Load, which emphasises the importance of reducing mental effort needed to process information.

2.2 Data Types and Visualisation Techniques

Understanding the nature of the data is a key prerequisite before delving into discussions about data representation. Data comes in various types, and selecting the appropriate visualisation technique hinges on recognising these distinctions. In this section, the key data types are categorised and matched with their most suitable visualisation techniques.

2.2.1 Categorisation of Data Types

Data can be broadly categorised into four main types [39]:

Nominal data: represents categories or labels without any inherent order. Examples include colours and gender categories.

Ordinal data: implies a meaningful order or ranking among categories but lacks equal intervals between them. Examples include survey responses (eg. "very satisfied", "satisfied", "neutral", "dissatisfied", "very dissatisfied").

Interval data: possesses ordered categories with equal intervals between them, but it lacks a true zero point. An example is temperature, measured in Celsius or Fahrenheit.

Ratio data: includes ordered categories with equal intervals and a meaningful zero point. Examples are age and income.

Matching Data Types with Appropriate Visualisation Techniques
Various data types demand specific visualisation methods for optimal representation. For nominal data, bar charts and stacked bar charts are effective in displaying categorical information and relative proportions. Ordinal data benefits from ordered bar charts, dot plots, or even stacked bar charts, maintaining the ranking and order of categories. Moreover, interval data is best visualised using line charts, histograms, and box plots, showcasing trends and distributions without assuming a true zero point. Finally, ratio data is effectively represented through scatter plots, histograms, and line charts, enabling precise comparisons and measurements due to the presence of a meaningful zero point [16]. An example of each of these is represented in Figure 5, where artificial data from a class of 4th-year university students is visualised.

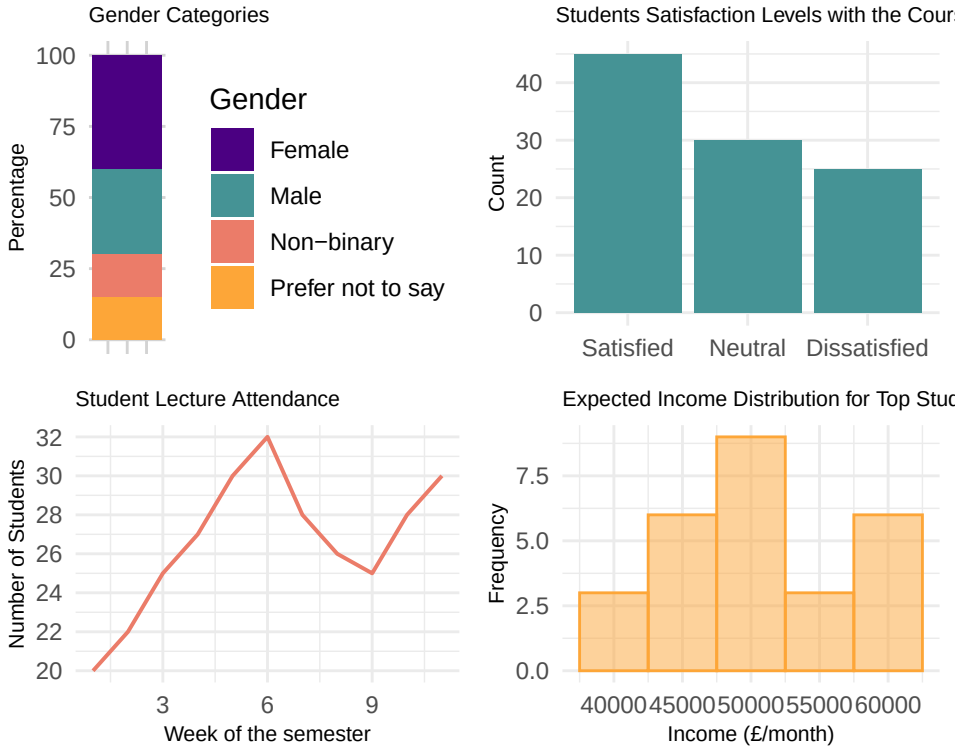

Figure 5: Different types of visualisation techniques according to the data type

2.3 Data Abstraction and Representation

The transformation of complex, raw data into simplified and comprehensible formats by focusing on its essential characteristics and hiding unnecessary details is a pivotal step in data representation [37]. This process, known as data abstraction, involves distilling complex datasets into visual forms

that convey insights. In this section, we explore the hierarchies and levels of abstraction in data visualisation, and the critical trade-offs between abstraction and the potential loss of information.

2.3.1 Hierarchies and Levels of Abstraction

Hierarchies of abstraction enable the representation data at varying levels of detail:

1. **Low-Level Abstraction**: At the lowest level, raw data is preserved in its most detailed form. This might include individual data points, measurements, or unprocessed text.
2. **Mid-Level Abstraction**: At the mid-level, data is grouped or aggregated to provide a broader overview. For example, hourly data points may be aggregated into daily or weekly averages.
3. **High-Level Abstraction**: At the highest level, data is represented in a condensed and abstracted form, often as summary statistics or key insights. Thus, this level provides a "big-picture view" of the data.

These different levels of abstraction are represented in Figure 6, where the mtcars dataset is represented. The first visualisation is a scatter plot that provides detailed information about the relationship between car weight and miles per gallon, with points coloured by the number of cylinders. The second is a bar plot presenting aggregated information about the average miles per gallon for different numbers of cylinders. Finally, the third is an abstract visualisation using a box-and-whisker plot to provide a high-level summary of the distribution of miles per gallon for different numbers of cylinders.

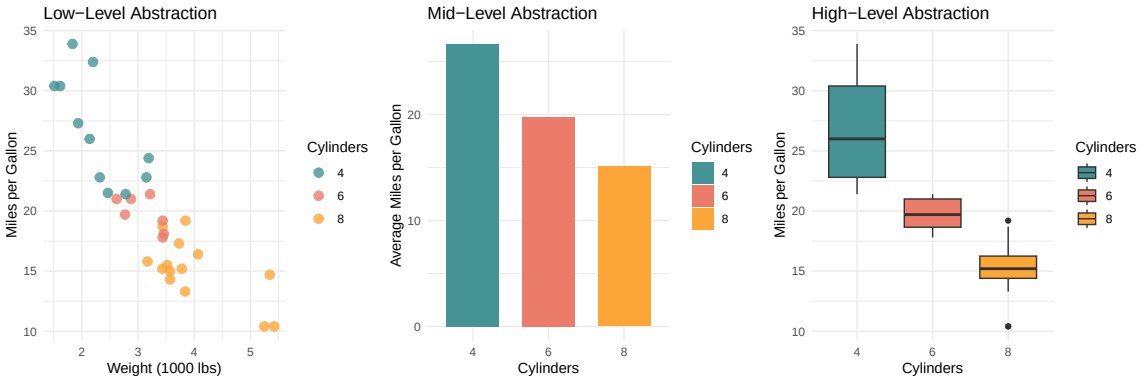

Figure 6: Mtcars dataset visualised on three different levels of abstraction

Trade-offs Between Abstraction and Information Loss
While abstraction simplifies complex data, it presents trade-offs. Creators of data visualisations must find a balance between clarity and detail, as well as between generalisation and specificity. Abstraction often increases clarity, but may sacrifice crucial details of the data necessary for certain analytical tasks. It also offers a more generalised view accessible to a wider audience, but may however overlook specific nuances essential for experts.

In data visualisation, the art of data abstraction lies in finding the right level of detail to effectively convey the intended message while minimising the risk of information loss. This delicate balance is a crucial consideration in the design of informative and meaningful data visualisations.

2.4 Visual Perception and Cognition

In this section, human visual perception is explored, along with the application of cognitive psychology principles to data visualisation.

Human Visual Perception: Decoding Visual Information
Visual perception profoundly influences our understanding of the world. When applied to data visualisation, it shapes the way that individuals engage with and derive meaning from visual data representations.

The most significant aspects of human visual perception within the realm of data visualisation include, firstly, pattern recognition, enabling the identification of trends, outliers, and relationships in data representations. Additionally, perceptual grouping, which causes visually similar elements to be grouped together, and thus, influences the interpretation of data clusters and shapes. Moreover, the hierarchy of perception dictates that certain visual attributes are processed more swiftly and effectively than others, such as colour being processed faster than text, influencing the viewer's attention hierarchy [5, 37].

The Gestalt Principles
Furthermore, the Gestalt Principles play an important role in the realm of visual perception and thus, data visualisation design [26]. Key Gestalt principles include proximity, which groups related elements; similarity, that links similar attributes; continuity, aiding trend representation; closure, for implying connections; and symmetry, for balance and aesthetics in visualisations [33].

Thus, note how by harnessing the principles of human visual perception and applying insights from cognitive psychology, designers of data visualisations can create visualisations that are not only aesthetically pleasing, but which are also cognitively efficient.

2.5 Colour Theory in Data Visualisation

In this section, the significance of colour in data visualisation, the principles of colour perception and encoding, and the importance of avoiding misleading visualisations through thoughtful colour choices are explored.

The Importance of Colour in Conveying Information
Colour significantly enhances the impact and comprehension of data visualisations. Some of the multiple purposes that colour serves in data visualisation include emphasising trends, distinguishing data points, and offering contextual information. It is often used to encode categorical data, differentiating between various groups with distinct colours, and to represent quantitative data by using colour intensity or gradients to portray values or magnitude [16].

Colour Perception and Colour Encoding in Visualisations
Understanding colour perception is crucial in the field of data visualisation. Key principles involve considering colour discrimination, ensuring accessibility for individuals with colour vision deficiencies. Figure 7 illustrates how impactful this is by comparing what is be perceived by "normal" observer, and what is perceived an observed with a colour vision deficiency when a specific colour palette is used. Furthermore, careful selection of colour schemes aligned with the intended message is essential — for instance, using warm colours like red and orange to indicate caution or warmth,

and cool colours like blue and green to convey calmness or coldness. Additionally, attention should be paid to how colours interact when combined; certain combinations might create visual vibrations, optical illusions, or impact text legibility.

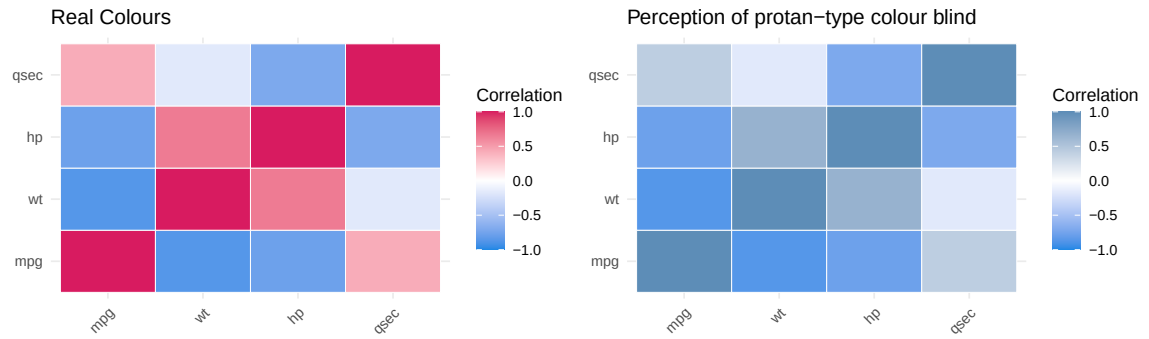

Figure 7: Colour perception of a heatmap for by a colour-blind person

Avoiding Misleading Visualisations Due to Colour Choices
Misleading visualisations often stem from inappropriate or deceptive use of colour. To avoid this, maintaining consistency in colour usage throughout the visualisation is essential. Furthermore, employing a uniform colour scheme for similar data categories or elements helps establish coherence and understanding. Finally, it's crucial to avoid colour choices that distort or exaggerate the data. Overly intense or contrasting colours might mislead interpretations, emphasising the necessity for judicious colour selection.

2.6 Cognitive Load and Visual Complexity

In data visualisation, achieving a balance between complexity and cognitive load is essential. Cognitive load significantly influences how viewers engage with and comprehend presented data, and finding a balance is crucial to effectively convey information without overwhelming the viewer's cognitive capacity [34]. This section explores the concept of cognitive load in visualisations, strategies to reduce cognitive load while maintaining complexity, and techniques to combat information overload through simplification.

2.6.1 Strategies to Reduce Cognitive Load While Maintaining Complexity

Several strategies can be employed to mitigate cognitive load while preserving complexity. Firstly, establishing a clear visual hierarchy using size, colour, and contrast helps direct attention to crucial elements. This strategy is used to simplify the graph on the left side of Figure 8, to the one on the right side. Secondly, simplifying labels and text by avoiding unnecessary complexity ensures information is clear and easily digestible.

Furthermore, employing interactive features like tooltips and drill-down functionality can assist in providing additional information when required or desired by the observer, while reducing the density of static visualisations.

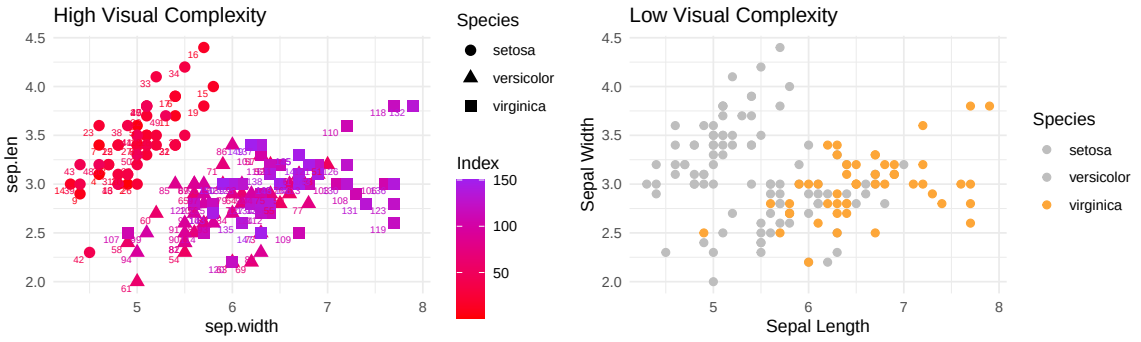

Figure 8: High versus low cognitive load demand through the reduction of visual complexity

2.6.2 Information Overload and Simplification Techniques

Addressing information overload in visualisations requires the strategic application of simplification techniques. The filtering technique enables focused data selection, while data reduction aggregates information to highlight overarching trends. Furthermore, storyboarding structures data presentation, aiding in contextual comprehension, and prioritisation ensures critical information is prominently displayed, elevating the visualisation's clarity and impact. These strategies collectively combat overwhelming data or excessive visual elements, enhancing comprehension and the effective communication of insights to viewers.

Chapter Overview and Further Reading References

This chapter delves into the theoretical underpinnings of data visualisation, drawing upon established concepts elucidated in academic literature. Notably, the first chapters in Robert Grant's comprehensive work, *Data Visualization: Charts, Maps, and Interactive Graphics* (2018), offers in-depth exploration of these concepts.

3 Univariate Data Visualisation Methods

This chapter initiates the discussion regarding data visualisation methods. Particularly, it introduces techniques for graphically representing for datasets of the simplest nature — that is, single-variable datasets. It focuses on histograms, bar charts, the kernel density estimate, time series analysis, and ROC Curves.

3.1 Histograms

Histograms are essential tools in univariate data visualisation. They provide a concise yet comprehensive overview of the frequency or density distribution of a single variable. This chapter delves into the construction of two types of histograms, as well as an example of their application in practice.

3.1.1 Classic Frequency Histograms

A classic histogram is a visual representation of the distribution of a (numerical) dataset. It is composed by a series of contiguous rectangles, also known as "bins", each representing a range of data values. The height of each of these corresponds to the frequency or count of data points within that range, depending on the type of histogram. On a cartesian coordinate system, typically, the x-axis shows the endpoints of each bin, and the y-axis represents count or frequency. Hence, histograms facilitate, among other thing, the visualisation of the distribution, spread, and central tendency of datasets.

Mathematical Notation
Suppose the data points within a datset are partitioned into k non-overlapping bins, denoted as $B_1, \ldots, B_k$ such that $B_i = [t_i, t_{i+1})$ for $i = 1, \ldots, k$. Here, t_i and t_{i+1} represent the lower and upper boundaries of the i-th bin, respectively.

Then, the classical frequency histogram of the data can be represented as a set of k bars, where the height of the i-th bar corresponds to the frequency or count of data points falling within the interval B_i.

Theory of Bin Number and Bin Width
The classical frequency histogram can be fully described by two main factors: the bin width b, and the bin origin t_0. However, in order for the bin counts to be comparable, the bins should all have the same width. Furthermore, it is essential to correctly choose the number of bins, since these have a huge impact on how the data is displayed and interpreted. Too few bins may hide the information in a dataset, and too many bins can cause a lot of noise in a dataset.

Sturges' Rule, established by Herbert Sturges in 1926, offers a systematic approach to determining the optimal number of bins constructing a frequency histogram [28]. This rule is based on the concept of normality, with the binomial distribution, $B(n, p = 0.5)$, serving as a model for an optimally constructed histogram. It links the binomial distribution with normally distributed data, providing a basis for selecting the number of bins to achieve a histogram resembling a normal density curve.

It suggests setting the number of bins, denoted as k, it can be expressed as:

$$k = 1 + \log_2 n,$$

where n represents the sample size. This formula stems from constructing a frequency histogram with k bins, each of width 1 and centered on the points $i = 0, 1, \ldots, k-1$. The bin count of the i-th bin is chosen to be the binomial coefficient $\binom{k-1}{i}$. As k increases, this ideal frequency histogram assumes the shape of a normal density with mean $(k-1)/2$ and variance $(k-1)/4$. The total sample size is:

$$n = \sum_{i=0}^{k-1} \binom{k-1}{i} = (1+1)^{k-1} = 2^{k-1},$$

by the binomial expansion. Hence, Sturges' rule follows immediately. [28].

Note that Sturges' rule functions as a number-of-bins rule rather than a bin-width rule. Assuming all bins are of equal width, to find the bin-width, Sturges' rule is implemented by partitioning the sample range of the data into the recommended number bins [28].

3.1.2 Density Histograms

A frequency histogram differs from a density histogram by its normalisation, which integrates to 1. That is, let $B_k = [t_k, t_{k+1})$ represents the k-th bin, as defined previously. If $t_{k+1} - t_k = b$ for all k, the histogram has a fixed bin width of b. In a frequency histogram, blocks of height 1 and width b are stacked in the appropriate bins, resulting in an integral equal to nb. Conversely, a density histogram uses blocks of height $1/(nb)$ to ensure each block has an area of $1/n$ [28]. In this way, the height of each bin represents the probability distribution of the data, such that the total area of the histogram equals 1.

Theory of Density Histograms
Let v_i denote the bin count of the i-th bin, representing the number of sample points falling in bin B_i. The density histogram is mathematically defined as:

$$\hat{f}(x) = \frac{v_k}{nb} = \frac{1}{nb} \sum_{i=1}^{n} I_{[t_i, t_{i+1})}(x_i) \quad \text{for } x \in B_i.$$

With this definition, it is straightforward to verify that $\hat{f}(x) \geq 0$ and that $\int \hat{f}(x)\, dx = 1$, ensuring that $\hat{f}(x)$ is a proper density function.

The density histogram's bin counts v_i can be considered as binomial random variables, where each bin count v_i follows a binomial distribution $B(n, p_i)$. The probability p_i for each bin is the integral of the density function $f(t)$ over the bin interval B_i [28]:

$$p_i = \int_{B_i} f(t)\, dt.$$

Consider the Mean Squared Error (MSE) of the estimator $\hat{f}(x)$ for $x \in B_k$. In this case, we have $\mathrm{E}[v_k] = np_k$ and $\mathrm{Var}[v_k] = np_k(1-p_k)$. Consequently, we can derive expressions for the variance and bias of $\hat{f}(x)$ as follows:

$$\mathrm{Var}\hat{f}(x) = \frac{Var v_k}{(nb)^2} = \frac{p_k(1-p_k)}{nb^2},$$

$$\mathrm{Bias}\hat{f}(x) = \mathrm{E}\hat{f}(x) - f(x) = \frac{1}{nb}\mathrm{E}v_k - f(x) = \frac{p_k}{b} - f(x).$$

The Mean Squared Error (MSE) of an estimator $\hat{f}(x)$ can be bounded using Lipschitz continuity and the Mean Value Theorem (MVT). Assuming $f(x)$ is Lipschitz continuous over a bin B_k, with a Lipschitz constant γ_k, the MSE at x can be expressed as the sum of the variance and the square of the bias. This is denoted by the inequality

$$\operatorname{MSE} \hat{f}(x) \leq \frac{f(\xi_k)}{nb} + \gamma_k^2 b^2,$$

where ξ_k is a point in the bin B_k, n is the number of observations, and b is the bin width, reflecting the combined effects of variability and estimation error.

3.1.3 Histogram in Practice

As the number of sample means used to plot a histogram increases, one can usually observe the shape of the histogram gradually resembling that of the normal distribution. Consider the histogram found on the left side of Figure 9

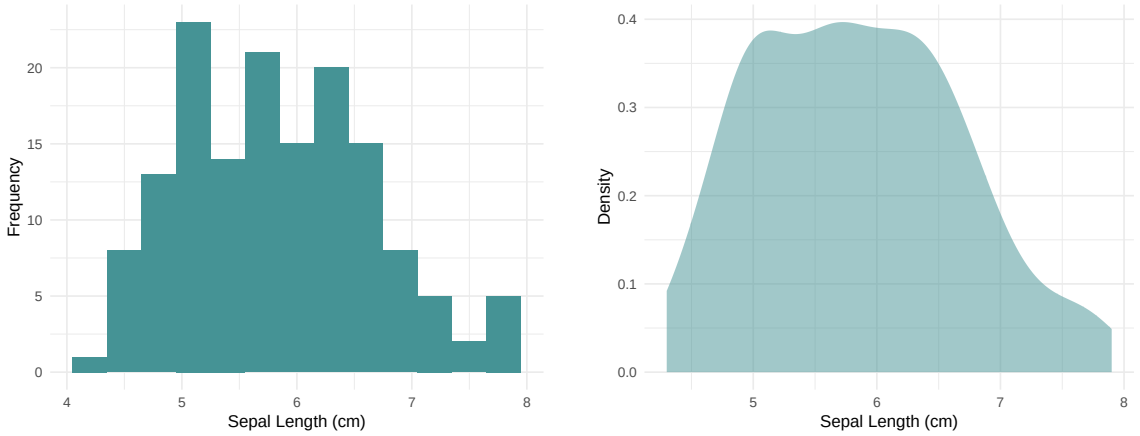

Figure 9: Histogram and Kernel Density Estimation of Sepal Length in Iris Dataset

The frequency histogram displayed in Figure 9 exhibits variable bin lengths, with a constant bin width set to 0.3. The shape of the histogram is slightly shifted to the left, deviating from the symmetric bell shape of a normal distribution. This deviation may indicate some level of asymmetry or non-normality in the Sepal Length data.

3.2 Kernel Density Estimation

Kernel Density Estimation(KDE) evolves from the concept of histogram, offering a method for estimating the the probability density function of a dataset.

Kernel Density Estimation (KDE) is an very useful tool in statistics. Instead of discrete histograms, it helps create a smooth curve from values in a dataset. KDE is used to infer the distribution of a population based on a limited sample. Thus, the result of the kernel density estimation is an estimate of the sample's probability density function. Based on this estimated probability density function, one can ascertain certain characteristics of the data distribution, such as the regions where

data is concentrated.

The KDE algorithm takes a parameter called bandwidth, h, that affects how smooth the resulting curve is. Changing the bandwidth changes the shape of the kernel: a lower bandwidth implies only points very close to the chosen position are given any weight, which leads to the estimate looking squiggly. In contrast, a higher bandwidth means having a shallow kernel, where distant points can contribute. Thus, leading to a smoother curve.

3.2.1 Theory of Kernel Density Estimation

Let us begin with the formula for the kernel density estimator:

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

In KDE, the variable x represents the point of density estimation. The total number of data points, denoted as n, constitutes the sample size, and all these points are used for estimating the dataset's probability density function. The bandwidth (h) is a crucial parameter that controls the smoothness of the estimation, with smaller bandwidths leading to closer fits to data and larger ones smoothing out details. The kernel function (K), such as the Gaussian or triangular kernel, assigns mass around each data point, influencing the density estimate, while X_i represents the individual sample points that collectively contribute to the estimated density function, guided by the kernel function and the bandwidth[36]. By introducing the re-scaling notation $K_h(u) = h^{-1}K(\frac{u}{h})$, we can also write the formula in a more compact way:

$$\hat{f}(x;h) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i).$$

Here, the kernel function $K(u)$ is a normalized non-negative function that satisfies $\int K(u)du = 1$. The mean squared error, $\mathrm{MSE}\hat{f}(x;h)$, can be used to quantify the difference for a given x between the "true" density function $f(x)$ and its estimator $\hat{f}(x)$, and with some simple transformation, it can be presented as follows:

$$\mathrm{MSE}\hat{f}(x;h) = E[(\hat{f}(x) - f(x))^2] = [\mathrm{bias}\hat{f}(x)]^2 + \mathrm{var}\hat{f}(x).$$

We can see that in order to compute $\mathrm{MSE}\hat{f}(x;h)$, we will require expression for the mean and variance value of $\hat{f}(x;h)$. This can be derived from the equation of KDE, and by using the convolution notation, we can get the value of the bias. The bias is the difference between a smoothing of f and f itself. Using similar calculations, we can get the variance, and the MSE by combining the variance and bias. So the MSE is expressed as:

$$\mathrm{MSE}\hat{f}(x;h) = n^{-1}(K_h^2 * f)(x) - (K_h * f)^2(x) + (K_h * f)(x) - f(x)^2.$$

Now, move our attention to the integration of MSE over all x, and it gives a global measure of conformity of $\hat{f}(x)$ with $f(x)$, called the mean integrated quare error, MISE, and it is one of measures used to estimate the smoothing parameter:

$$\begin{aligned}\mathrm{MISE}(\hat{f}) &= \int \mathrm{MSE}\hat{f}(x;h)\,dx \\ &= n^{-1}\int (K_h^2 * f)(x) - (K_h * f)^2(x)\,dx + \int ((K_h * f)(x) - f(x))^2\,dx.\end{aligned}$$

In Figure 9, the kernel density estimation of the dataset is presented. The distribution of sepal lengths is depicted by the kernel density curve. The peaks observed in the curve correspond to the primary concentration trends of sepal length within the data. A unimodal curve signifies a concentration of sepal lengths for most irises in that specific region. Conversely, a bimodal or multimodal curve suggests the existence of multiple concentration areas.

3.3 Bar Charts

Bar charts are a crucial tool in data presentation, arranging data into vertical or horizontal bars. The varying lengths of these bars directly correspond to the magnitude of the information they represent. Bar charts excel in comparing classified data, especially when values are closely aligned. This stems from the nature of human perception, as our visual acuity for height surpasses that of other visual elements like area or angle.

Bar charts represent a versatile tool for data visualisation, frequently employed to compare distinct categories. The vertical bar chart, commonly recognised, exhibits categories along the x-axis and their frequencies or counts along the y-axis. Horizontal bar charts, rotated 90 degrees, prove beneficial for extended category names or numerous categories, displaying categories on the y-axis and frequencies on the x-axis. Multi-set or grouped bar charts facilitate side-by-side comparisons of subgroups within categories, available in both vertical and horizontal orientations. Stacked bar charts illustrate classes of values subdivided into sub-classes, often differentiated by colour, where each segment's size signifies its frequency or count, and the total bar length reflects the cumulative total.

3.3.1 Bar Charts in Practice

The two bar charts on in Figure 10 illustrate the impact of varying vitamin dosages on tooth growth, further categorised by supplement type. On the x-axis, three distinct levels of vitamin dosage are presented, while the y-axis indicates the average length of tooth for each dosage.

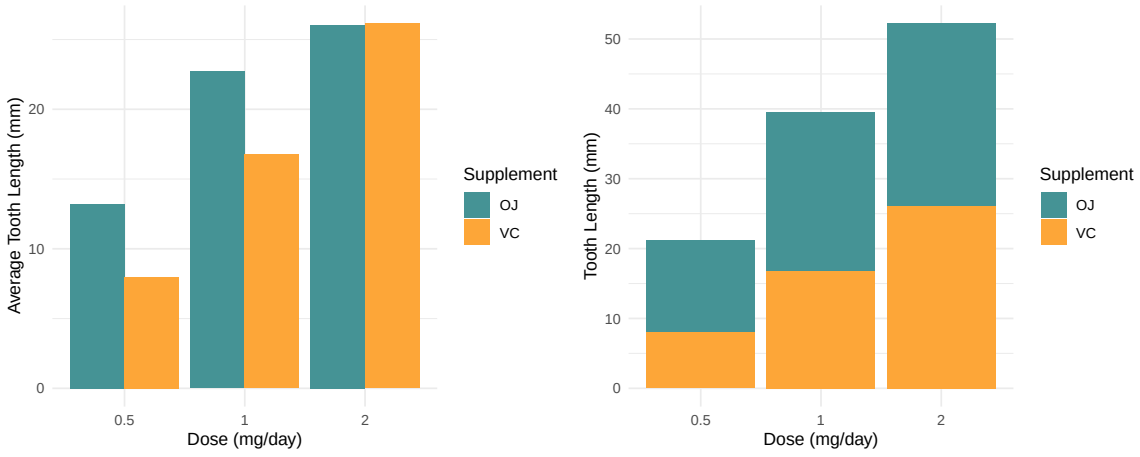

Figure 10: Visualising tooth growth by dosage: A comparison of grouped and stacked bar chart techniques

From the grouped bar, chart found on the left of Figure 10, due to the side-by-side positioning of the bars, it is easy to note that tooth growth varies not only with the dosage but also with the supplement type. In contrast, the stacked bar graph on the right of Figure 10 facilitates the understanding of the combined effects of the two supplements at each dosage level. However, compared to the latter, it becomes more challenging to differentiate the individual contribution of each supplement.

3.4 Line Charts and Time Series

Previous subsections explored histograms, kernel density estimation, and bar charts, which are univariate methods used for analysing independent variables in isolation, revealing patterns and intrinsic properties within data. Now, the focus shifts to univariate methods of dependent variables, specifically line charts, which are crucial for visualising relationships between variables to understanding their interactions and dependencies over time.

3.4.1 Theory of Line Charts

Line charts are fundamental tools in data visualisation, particularly useful for displaying time series data. A line chart represents n data points $\{(x_i, y_i)\}_{1 \leq i \leq n}$ on a Cartesian coordinate system, with the x-axis often denoting time intervals or ordered categories and the y-axis representing the measured values.

In a line chart, consecutive data points are typically connected by straight lines. The line segment between two points (x_i, y_i) and (x_{i+1}, y_{i+1}) can be described by the equation of a line in the slope-intercept form: $y = mx + b$, where m is the slope and b is the y-intercept. Also, a series of linear interpolations between pairs of data points could be used. These interpolations assume that the change between two points is uniform or linear. This linear approach is mathematically represented as:

$$y = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \times (x - x_i), \quad x_i \leq x \leq x_{i+1}.$$

This equation highlights that for any point x between x_i and x_{i+1}, the corresponding value of y on the line chart is determined by a linear relation. This method effectively "fills the gaps" between actual observed data points and provides a continuous view of the data.

3.4.2 Time Series Analysis

Time series visualisation is particularly suited to line charts because they effectively display changes and trends over time, allowing for easy visualisation of relationships between time points. A time series [3] is a collection of observations $\{x_t\}$, where $t = 0, \ldots, n$ denotes the time point at which the observation is recorded. An index set T_0 which collects all the time points when observations are available. For instance, $T_0 = \{0, \ldots, n\}$ for $n \in \mathbb{N}$. By plotting these data points over time, line charts help in identifying long-term trends, seasonal patterns, and anomalies.

A time series can be viewed as a realisation of a stochastic process. A stochastic process [3] $X = (X_t)_{t \in T_0}$ is a collection of random variables X_t, where t denotes the time index and T_0 the index set. For a fixed event $\omega \in \Omega$ we obtain the realisation of the stochastic process (sometimes also called a sample path) which is given by $x_t = X_t(\omega)$, $t \in T_0$.

A time series $\{X_t\}$ is a moving average process [3] of order q(MA(q)) if it can be expressed as:

$$x_t = \omega_t + \theta_1\omega_{t-1} + \cdots + \theta_q\omega_{t-q},$$

where $\{\omega_t\} \sim \mathrm{N}(0, \sigma^2)$ are i.i.d. white noise events and $\theta_1, \ldots, \theta_q$ are real valued coefficients.

3.4.3 Case Example: Visualisation of Exchange Rates

Here, plot daily and 49-day moving average exchange rates of exchange rate data in one figure.

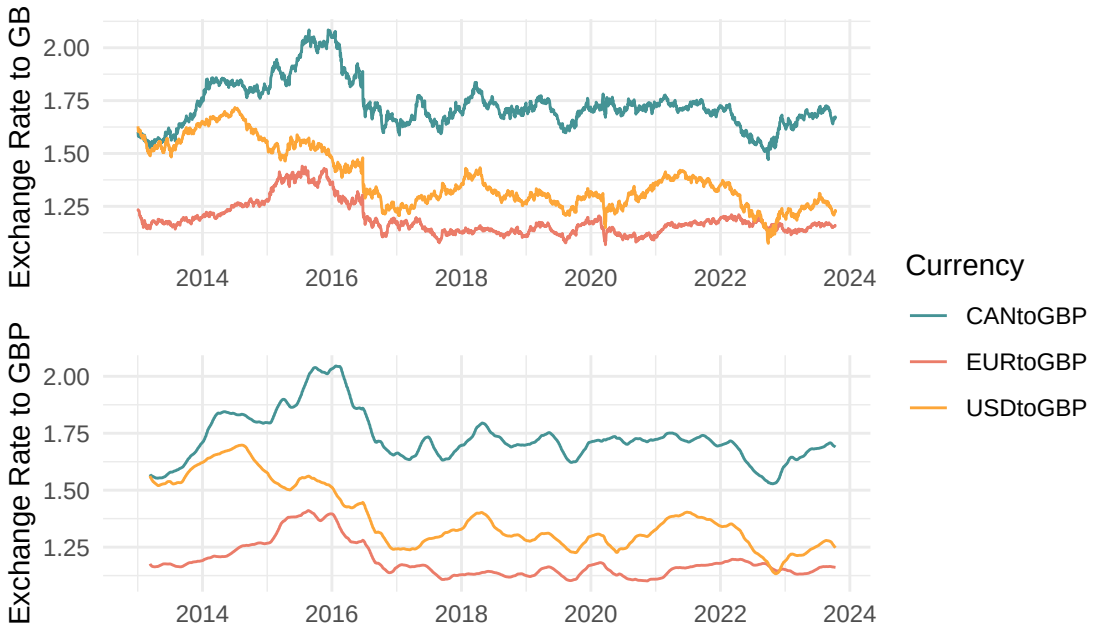

Figure 11: Daily (top) and 49-day moving average (bottom) exchange rates of CAN, EUR, USD to GBP

The first plot in Figure 11 presents a comparative visualisation of daily exchange rates for CAD, EUR, and USD against GBP, offering an overview of their trends and relative performance. This plot enables the identification of overall trends and periods of volatility for each currency pair, allowing for an assessment of their stability and strength relative to GBP. Notice that, there was a simultaneous and significant drop in all three currencies relative to the GBP. The concurrent nature of these declines across diverse currency pairs suggests that the driving factor is a depreciation of the GBP, rather than independent appreciations of the USD, EUR, and CAN. This notable depreciation of the GBP occurred around 2016, which coincides with the commencement of the BREXIT process. Usually, the long term trend attracts financial analyst most. Therefore, filtering fluctuations and anomalies is important. Then, a 49-day moving average MA(49) was employed, shown in the second plot of Figure 11, to elucidate long-term trends while mitigating short-term fluctuations. This is the smoothing method of simple average of past observations, expressed as:

$$\mathrm{MA}_{49}(t) = \frac{1}{49} \sum_{k=t-48}^{t} x_k,$$

where x_k denotes the exchange rate on day k.

This method allows a clearer view of overarching trends in currency movements against the GBP. The overlay of these moving averages on the daily exchange rates in visualisations provides both a clear comparative and a quantitative perspective.

Autocorrelation Analysis of CNY to GBP Exchange Rate
Autocorrelation describes the correlation of a time series with its own previous and proceeding values. The autocorrelation function (ACF) measures the linear predictability of the series at lag h, which is the time t with its values at a previous time $t - h$. The mathematical formulation of ACF of time series will be given in the following paragraph.

Suppose we have a time series with observations denoted by $\{x_1, \ldots, x_n\}$. Then the sample mean is given by $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$. And the sample autocovariance function [3] at lag h in days of our time series is

$$\hat{\gamma}(h) := \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \bar{x})(x_t - \bar{x}), \quad -n < h < n.$$

Hence, the sample autocorrelation function [3] at lag h in days is given by

$$\hat{\rho}(h) := \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{n-|h|}(x_{t+|h|} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}, \quad -n < h < n.$$

The value of $\hat{\rho}(h)$ lies between -1 and +1. A value close to +1 indicates a strong positive correlation, while a value close to -1 indicates a strong negative correlation. A value near 0 suggests little to no linear correlation. A slow decay in the ACF plot indicates a strong relationship between past and present values, while spikes at specific lags may suggest seasonality. In addition, the bounds in an ACF plot help to determine whether the observed autocorrelations are significant or merely due to random fluctuation. The theoretical bounds of the ACF plot are typically set at $\pm 1.96/\sqrt{n}$, where n is the length of the time series. This $\pm 1.96/\sqrt{n}$ formula derives from the assumption of a normal distribution of the autocorrelation coefficient under the null hypothesis of no autocorrelation; And 1.96 corresponds to the 95% confidence interval of a standard normal distribution. This means that if the ACF of a certain lag falls outside these bounds, the correlation at that lag is statistically significant at the 5% level, suggesting that the series exhibits autocorrelation at that lag.

Next, Figure 12 visualises the ACF for the CNY to GBP exchange rate to understand its time-dependent structure better.

```
# PLOT OF THE AUTOCORRELATION FUNCTION
acf_data <- acf(MyData$CNYtoGBP, plot = FALSE)
plot(acf_data, main = "", xlab = "Lag h", ylab = "ACF")
```

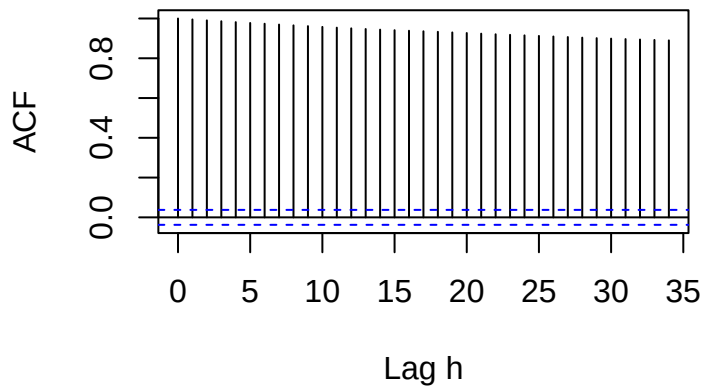

Figure 12: Autocorrelation Function at lag h in days of CNY to GBP Exchange Rate

From Figure 12, the ACF plot for the CNY to GBP exchange rate series reveals that the ACF starts near 1 and decreases gradually. This pattern suggests a strong persistence in the time series, indicating that past values have a significant influence on future values. In time series analysis, such a slow decay in the ACF is indicative of a non-stationary series, where the mean, variance, and autocorrelation structure do not remain constant over time. Also, since the ACF for a wide range of lag falls outside these bounds, the series exhibits strong autocorrelation at those lags. This means past values of the series have a significant influence on future values.

This persistent autocorrelation suggests that short-term movements in the CNY to GBP exchange rate are heavily influenced by its recent history. Such a characteristic is crucial for forecasting models, as it implies that recent historical data can be a powerful predictor of near-future trends. Models like ARIMA (Autoregressive Integrated Moving Average), which are well-suited for data with high autocorrelation, may be particularly effective in this context.

Decomposition of Time Series
One of the primary advantages of time series visualisation is the ease with which it allows analysts to identify long-term upward or downward trends in data and patterns that repeat over specific intervals. By decomposing the time series, it would be easy to see those features. Non-stationary time series data, X_t, can often be described as a combination of several distinct components: Trend component t_t: The underlying progression in the series, Seasonal component s_t: Periodic fluctuations due to seasonal factor, Residual r_t: The irregular or error component. Hence, the decomposition of a time series can be described in two main models:

Additive Model [3]: In the additive model, the components are added together:

$$X_t = t_t + s_t + r_t.$$

Multiplicative Model [3]: In the multiplicative model, the components are multiplied together:

$$X_t = t_t \times s_t \times r_t \quad \text{or} \quad \log(X_t) = \log(t_t) + \log(s_t) + \log(r_t).$$

In practice, the choice between the additive and multiplicative models often depends on the nature of the time series. If the magnitude of the seasonal fluctuations or the variation around the trend does not vary with the level of the time series, then an additive model is appropriate. If the magnitude of the seasonal fluctuations or the variation around the trend increases or decreases as the time series level changes, then a multiplicative model may be more suitable.

R function `decompose()` is able to decompose the time series by additive model or multiplicative model. And as shown in Figure 13 is the demonstration of decomposition.

```
# PLOT OF DECOMPOSITION OF ADDICTIVE TIME SERIES MODEL
ggplot(decomposed_df, aes(x = time, y = value)) +
  geom_line() +
  facet_wrap(~component, scales = "free_y", ncol = 1) +
  labs(x = "Date",y = "Exchange Rate to GBP") +
  theme_minimal()
```

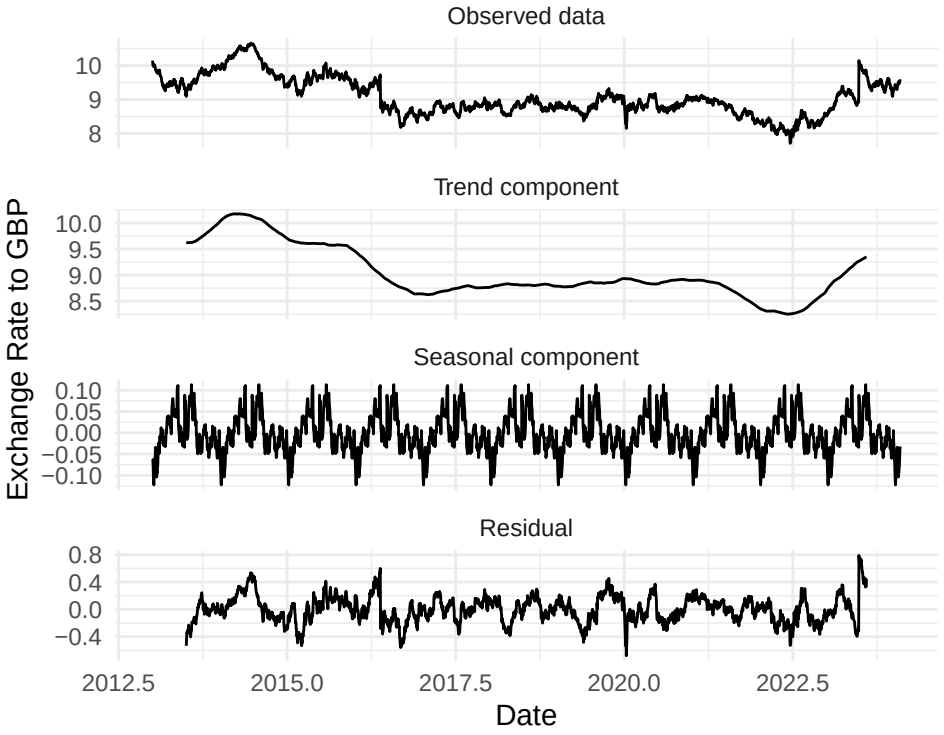

Figure 13: Decomposition of addictive time series model of CNY to GBP exchange rates

From Figure 13, the CNY to GBP exchange rate time series was decomposed into its fundamental components: trend, seasonality, and residual noise by additive model. This additive model, represented mathematically as $X_t = t_t + s_t + r_t$.

As illustrated in Figure 13, the trend component t_t of CNY to GBP exchange rate exhibits a distinct pattern over time: initially, it shows a gradual decrease and reached crest in 2022, followed by a sudden increase afterwards. It coincides with the tax reduction policy issued by UK government in 2022, which leads to a depreciation of GBP. This trend is pivotal for understanding the broader economic relationship between these currencies.

Moreover, the seasonal component s_t of the decomposition highlights cyclical fluctuations, indicative of recurrent patterns within the year. These could be attributed to seasonal economic activities, policy changes, or other cyclical factors influencing the currency market. The clear demarcation of these cyclical trends in the seasonal component helps in isolating such effects from the overarching trend.

Lastly, the residual component r_t encompasses the random, unexplained variations after accounting for the trend and seasonal factors. Analysing these residuals is crucial for understanding the unpredictability in the exchange rate and can be pivotal in risk management and forecasting.

3.5 ROC Curve

The Receiver Operating Characteristic (ROC) analysis is a technique for assessing the performance of classification models as its discrimination threshold is varied. Unlike univariate methods in previous subsections that are constructed from a single dataset representing one variable, ROC analysis divides a single dataset into two samples based on the actual condition of each observation, but focuses on evaluating the performance of a model on one variable, typically the predicted probability that a given observation belongs to a positive class.

3.5.1 Theory of ROC curve

Suppose we have N observations $\{y_1, \ldots, y_N\}$, the ROC curve can be used to access the model performance of binary predictions, such as diseased and non-diseased or default and non-default. For binary predictions, we define the True positive (TP) as the sum of instances where model's prediction $\hat{y}_n$ is positive when real observation y_n is positive. Similarly, we can also define false positive (FP), false negative (FN), true negative (TN) [9] and they can be expressed mathematically as follows:

$$\mathrm{TP} = \sum_{n=1}^{N} \mathbb{I}(y_n = 1)\mathbb{I}(\hat{y}_n = 1), \quad \mathrm{FP} = \sum_{n=1}^{N} \mathbb{I}(y_n = 0)\mathbb{I}(\hat{y}_n = 1),$$

$$\mathrm{FN} = \sum_{n=1}^{N} \mathbb{I}(y_n = 1)\mathbb{I}(\hat{y}_n = 0), \quad \mathrm{TN} = \sum_{n=1}^{N} \mathbb{I}(y_n = 0)\mathbb{I}(\hat{y}_n = 0).$$

Let Y be a binary variable indicating the true class of an instance, with $Y = 1$ for positive instances and $Y = 0$ for negative instances. Let X be a continuous variable representing the predicted score or probability of an instance being classified as positive by the classifier, with $X \geq c$ for positive and $X < c$ for negative given a threshold c.

ROC curve plots True Positive Rate (TPR is the proportion of positive instances correctly identified) and False Positive Rate (FPR is the proportion of negative instances incorrectly identified as positive) [11], which are defined as follows:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN}.$$

Alternatively, the TPR and the FPR can be defined using conditional probabilities as follows [22]:

$$\text{TPR}(c) = \mathbb{P}[X \geq c | Y = 1], \quad \text{FPR}(c) = \mathbb{P}[X \geq c | Y = 0].$$

The ROC curve is then the set of points $(\text{FPR}(c), \text{TPR}(c))$ for all possible values of threshold c. This curve plots the trade-off between sensitivity (or TPR) and specificity (1 - FPR) across different thresholds. Let $t = \text{FPR}(c)$, then $c = \text{FPR}^{-1}(t)$ and $\text{TPR}(c) = \text{TPR}(\text{FPR}^{-1}(t))$. Hence, let $\text{ROC}(t) = \text{TPR}(\text{FPR}^{-1}(t))$, the ROC curve can be represented as a parametric function of t (classification threshold corresponding to $t = \text{FPR}(c)$) [22] , with t varying from 0 to 1:

$$\text{ROC} = \{(t, \text{ROC}(t)) \mid t \in [0, 1]\}.$$

The Area Under the Curve (AUC) [22] provides a single measure of model performance. This definition represents the integral of $\text{ROC}(t)$ over the interval from 0 to 1. It can be defined as:

$$\text{AUC} = \int_0^1 \text{ROC}(t) dt.$$

The larger the AUC, the better the classifier.

3.5.2 ROC analysis in COVID-19 test

The COVID-19 pandemic underscores the need for accurate diagnostic tests to differentiate between positive and negative cases [12]. An effective mean for evaluating accuracy of various diagnostic tests is the ROC analysis.

An ideal test minimises FPR, avoiding unnecessary treatments or quarantine. The ROC curve visualises the trade-off between TPR and FPR at various thresholds. A curve arching towards the upper left indicates high sensitivity and low FPR — the hallmarks of a reliable test.

The Figure 14 plots the ROC curve of simulated COVID-19 test data.

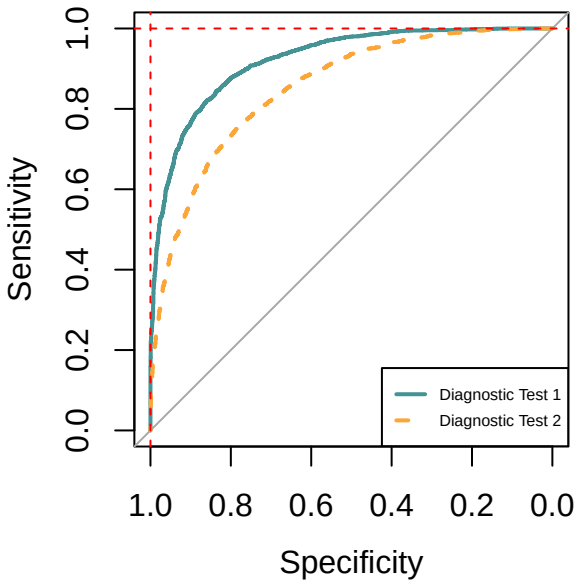

Figure 14: ROC Curves for Two Diagnostic Tests

From the plot 14, the Diagnostic Test 1 has higher AUC, which is closer to the top left corner. This test is likely to be more accurate in diagnosing COVID-19. It suggests that the test has a higher combined sensitivity and specificity, meaning it can identify positive cases more correctly and has fewer false alarms. Also, the Diagnostic Test 2 has a lower AUC than Test 1, This test is less accurate but still better than random guessing. It may miss more true cases (lower sensitivity) or incorrectly identify healthy individuals as having COVID-19 (higher false positive rate).

Chapter Overview and Further Reading References

In this chapter, we have delved into the theory behind histograms, density estimators, line charts, time series, and the ROC curve. These concepts are well-established in statistical literature. For further insights into histograms and kernel smoothing, David W. Scott's book *Multivariate Density Estimation: Theory and Practice, and Visualization* (1992) provides comprehensive details of the theory behind these concepts. Additionally, *Introduction to Time Series and Forecasting* (2016) by Peter J. Brockwell and Richard A. Davis offers in-depth exploration of line charts and time series theory. Lastly, Margaret Sullivan Pepe's *The Statistical Evaluation of Medical Tests for Clarification and Prediction* (2003), particularly Chapters 4 and 5, provides thorough coverage of the ROC curve theory.

4 Bivariate Data Visualisation Methods

This chapter transitions from the study of univariate data visualisations to the exploration of bivariate data. Fundamental bivariate data visualisation methods, including heatmaps, scatter plots, and bubble charts, are introduced, scrutinised, and modelled throughout this chapter. In order to facilitate a more profound analysis of bivariate data, the chapter further delves into the examination of linear regression, as well as LOESS regression. Notably, within the context of scatter plots, an intriguing deviation is observed with the incorporation of animated data visualisations.

4.1 Heatmaps

The heatmap is a data visualisation technique that uses colour coding to represent different intensities. It can be represented as an $m \times n$ matrix $\mathbf{M}$, with m observations for variable 1 and n observations for variable 2:

$$\mathbf{M} = \begin{bmatrix} \mathrm{M}_{11} & \mathrm{M}_{12} & \dots & \mathrm{M}_{1n} \\ \mathrm{M}_{21} & \mathrm{M}_{22} & \dots & \mathrm{M}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{M}_{m1} & \mathrm{M}_{m2} & \dots & \mathrm{M}_{mn} \end{bmatrix},$$

where each entry M_{mn} represent an observation.

In this illustrative example, heatmaps are used to visualise fire occurrences in Brazil. These heatmaps provide a spatially coherent representation, highlighting regions at high risk and seasonal patterns. The data-driven insights could empower policymakers to make informed decisions regarding preventive measures and firefighting strategies.

In Figure 15, it can be observed that significantly higher fire counts are found in certain locations. The presence of two strips with high frequencies of fires are highly unusual. The vertical trend corresponds to the location of BR-230 (Trans-Amazonian Highway) passing through the city of Apuí, State of Amazonas. The horizontal trend corresponds to BR-163 (Brazil highway) passing through Três Pinheiros in Novo Progresso, State of Pará. The western coastal area with a high frequency of fire occurrence corresponds to regions in close proximity to the cities of Vista Alegre do Abunã and Rio Branco. Research has indicated that 95 % of active fires and the most intense ones (FRP $>$ 500 megawatts) occurred at the edges in forests [8].

The seasonal pattern of fire frequency is shown in Figure 16. Observe that more fire occur in the months of August to October compared to the rest of the year.

```
# HEATMAP PLOT
heatmap_plot <- ggplot(pivot_table,
                       aes(x = factor(abb_month, levels = custom_order),
                           y = as.character(year), fill = count)) +
  geom_tile() +
  scale_fill_gradient(low = "#fff7ec", high = "#d7301f") +
  labs(x = " ", y = " ") +
  theme_minimal() +
  theme(axis.text = element_text(size = 9))

print(heatmap_plot)
```

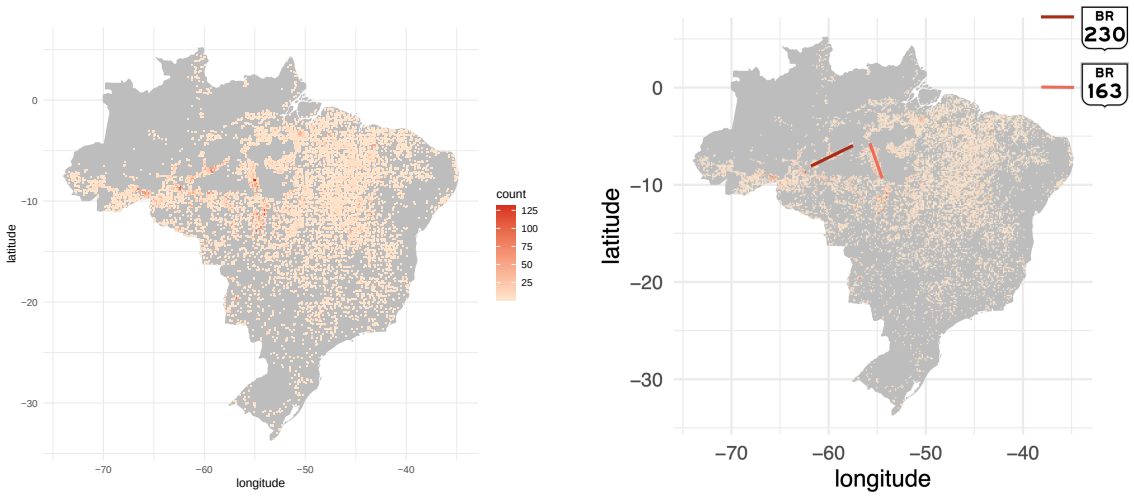

Figure 15: Frequency of fire in Brazil (2022), two strips of high frequencies of fires are highly unusual

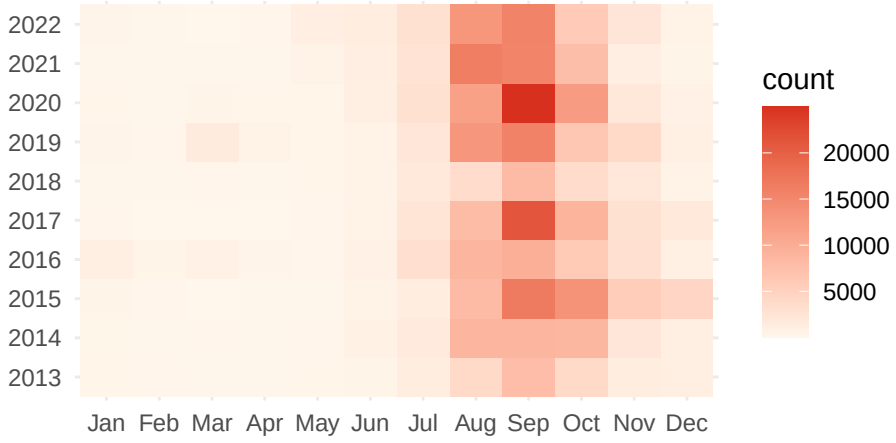

Figure 16: Frequency of fire in Brazil (2013-2022)

4.2 Scatter Plots

A scatter plot is a graphical representation of a set of data points in a two-dimensional coordinate system. Each data point is represented by a dot, and the position of the dot is determined by the values of two variables in the system. Some examples of this type of visualisation can be found in Figures 6, 8, and 17.

In general, the Y-axis denotes the response, or dependent, variable and the X-axis denotes the

explanatory, or independent, variable. Each observation of a dataset is mapped to a dot in the 2-dimentional space. Let (x_i, y_i) represent the coordinates of the i-th data point on the scatter plot produced by mapping a set of data. The scatter plot can be mathematically described as a set of points: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where n is the number of observations in the set.

4.2.1 Scatter Plots in Practice

The *ggplot2* package in R offers a versatile built-in function for generating scatter plots called `geom_point()`. This function works by mapping variables in a dataset to aesthetic attributes of points in a scatter plot, such as position (x and y coordinates), size, color, shape, and transparency.

Illustrations of plots generated in this manner are depicted in Figure 17. The scatter plots in this figure visually depict relationships between variables in the mtcars dataset.

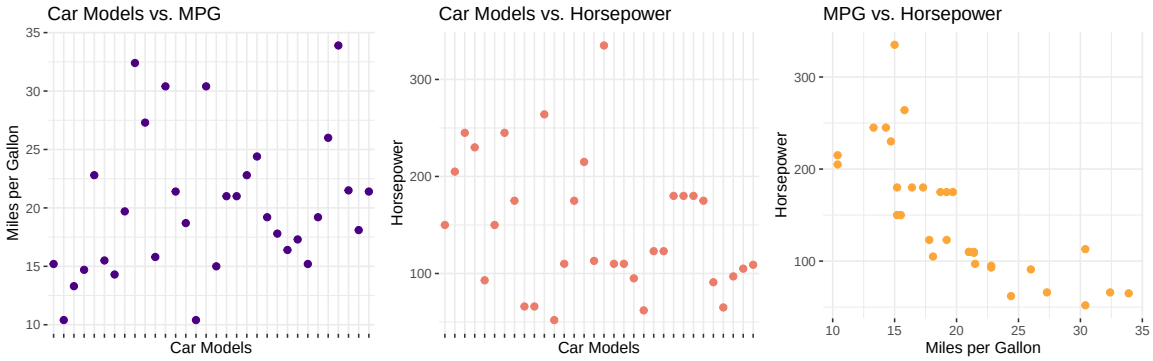

Figure 17: Partial scatter plot deconstruction of the mtcars dataset

The first plot in Figure 17 illustrates the relationship between car models and miles per gallon (MPG), showing potential differences in fuel efficiency among different models. Similarly, the second plot explores the association between car models and horsepower. Finally, the third plot examines the correlation between MPG and horsepower.

As expected, discerning a pattern or clear relationship between variables proves challenging in the first and second scatter plots. This challenge arises from the fact that the variable depicted along the X-axis in these plots is the car model, which lacks any relevant ordering that would relate on the technical details of the cars. Conversely, in the third scatter plot, a negative correlation between horsepower and miles per gallon is evident. This relation highlights potential trade-offs between fuel efficiency and engine performance.

4.2.2 Animated Scatter Plots

While static scatter plots are effective in depicting relationships between two variables, animated scatter plot visualisations take this a step further by introducing, for example, a temporal dimension to the data visualisation. Unlike static plots, animated scatter plots enable the depiction of changes in relationships, clusters, or outliers over time. In R, the *gganimate* package, building on the

foundation of *ggplot*2 package, facilitates the creation of animated plots, including animated scatter plots.

4.2.3 *gganimate* in Practice

Figure 18 represents the gapminder dataset in a static scatter plot demonstrating the relationship between gross domestic product (GDP) per capita and life expectancy across various countries. GDP per capita is depicted on a logarithmic scale along the X-axis to accommodate a wide range of values, while life expectancy is represented on the Y-axis. Moreover, the colour each of the dots indicate the country they represent data of.

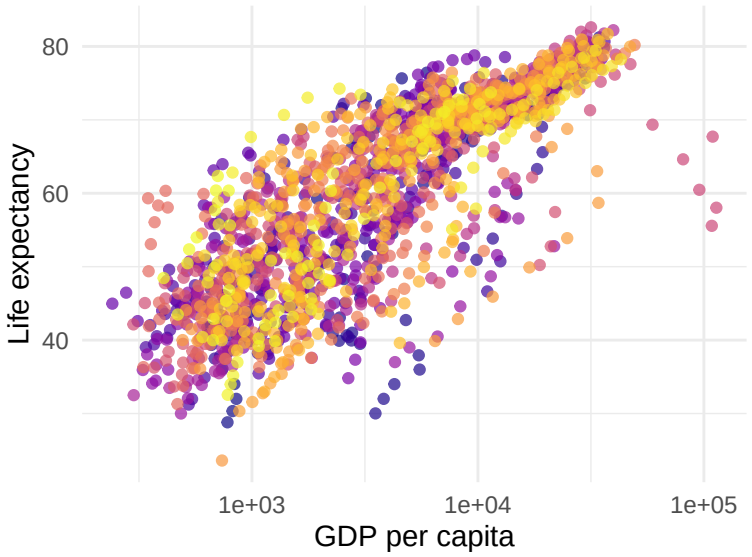

Figure 18: Static scatter plot deconstruction of the gapminder dataset

In the static scatter plot displayed in Figure 18, there appears to be a positive correlation between GDP per capita and life expectancy: as GDP per capita increases, the life expectancy behaves in a similar manner. However, the spread of points and variability in life expectancy at similar GDP levels make it difficult to draw precise conclusions. Furthermore, the complexity and clutter in the plot demand an extremely high cognitive load, making it difficult for the viewer to understand the details of the data visualised.

However, one can refer to the animated version of the scatter plot to solve many of these issues and discover details in the data that the static plot is unable to illustrate. While here the code chunk and screenshots are provided, the set animated plots of the gapminder dataset, produced using *gganimate*, can be found in the GitHub code folder.

```
# CREATE ANIMATED GGPLOT
gapplot<-ggplot(gapminder, aes(gdpPercap, lifeExp, colour = country)) +
  geom_point(alpha = 0.7, show.legend = FALSE) +
  scale_colour_manual(values = country_colors) +
```

```
scale_x_log10() +
facet_wrap(~continent) +
theme_minimal()

# gganimate specific code
labs(title = 'Year: {frame_time}', x = 'GDP per capita', y = 'life expectancy') +
transition_time(year) +
ease_aes('linear')
```

Figure 19: Screenshots of animated scatterplot at different time stamps

The animated scatter plots provide a dynamic visualisation of the relationship between GDP per capita and life expectancy in each country over time. The animation is displayed across 5 separate animated scatter plot, each grouping the data from countries from a specific continent to reduce clutter and cognitive load. This animation helps in understanding temporal trends and patterns in the data that may not be apparent in static plots.

The conclusions drawn from the animated plots might differ from those of the static plot. For instance, while the static plot may suggest a general positive correlation between GDP per capita and life expectancy, the animated plots reveals nuances such as the pace of improvement in life expectancy over time, variations in the relationship between different countries and continents, and the impact of historical events or economic changes on life expectancy trends. In general, the animated plots offer a richer and more dynamic exploration of the data, leading to potentially deeper insights and understanding.

4.3 Bubble Charts

4.3.1 Theory of Bubble Charts

Bubble charts are a captivating data visualisation tool that extends beyond the typical two-dimensional scatter plot by introducing an extra dimension. They represent data points as bubbles or circles on a two-dimensional plane, where the size of each bubble encodes a third variable.

The construction of bubble charts, is parallel to that of a scatter plot, but involves the scaling the data values to determine the size of each bubble. While it isn't always the case, the size of these is typically proportional to the variable it represents. The choice of scaling method depends on the data distribution and the message the chart aims to convey.

Generally, the formula for calculating the bubble radius (R) involves applying the scaling function

$$R = kV,$$

where R represents the size of the bubble, V the value of the variable being represented, and k a scaling factor to control the bubble size. Selecting an appropriate scaling factor (k) is critical for maintaining the proportionality between the bubble size and the variable being represented.

Hence, bubble chart represents data points in three dimensions: x-coordinate, y-coordinate, and bubble size. Each data point is denoted by a triplet of values (x_i, y_i, C_i), where x_i and y_i represent the coordinates, and C_i is the equation of the circle for the i-th observation. In the context of a bubble chart, the radius R_i of each bubble is expressed as $k \cdot V_i$, allowing us to formulate the equation for each circle as:

$$(x - x_i)^2 + (y - y_i)^2 = (kV_i)^2,$$

Here, x_i and y_i denote the coordinates of the circle's center.

Since each bubble B_i can be defined mathematically by the triplet (x_i, y_i, C_i), akin to the scatter plot, the bubble plot can be described as a set of "bubbles" $\{B_1, \ldots, B_n\}$, where n represents the number of observations in the set.

4.3.2 Bubble Charts in Practice

The bubble chart shown in Figure 20, as the scatter plots above, visualises data from the mtcars dataset. The single graph deplics the relationship between car models and their fuel efficiency (mpg) while using the size of the bubbles to represent the car's horsepower (hp) and even colour-coding the bubbles based on the number of cylinders (cyl).

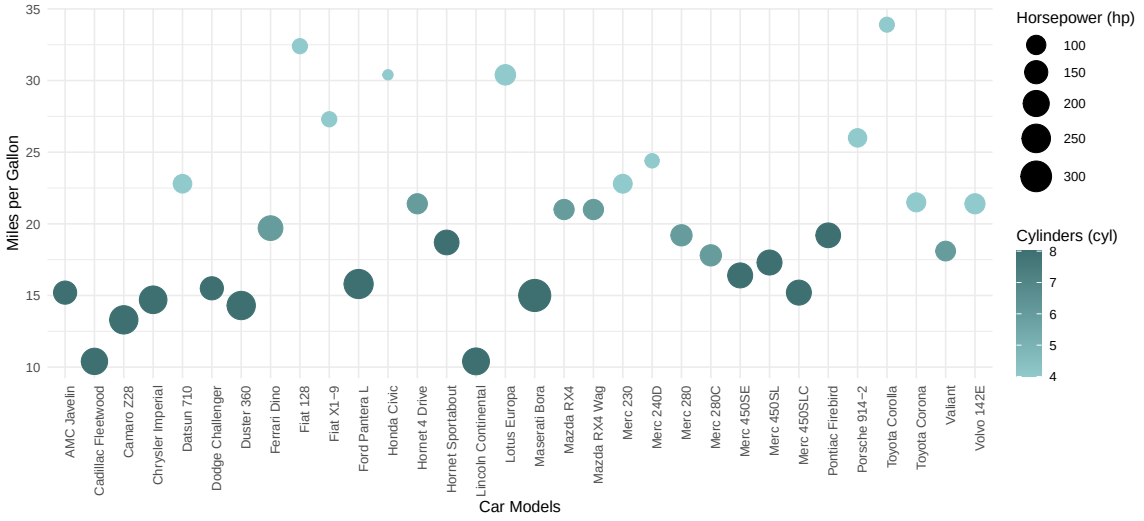

Figure 20: Bubble plot illustrating trends in car types and performance

In Figure 20 a strong correlation between the cars' number of cylinder and their horsepower can be quickly identified. This is due to the fact that as size of the dots (representing horsepower) increases, their colour (representing the number of cylinders) simultaneously becomes darker. Furthermore, a correlation between the miles per gallon consumed by the different types of cars and their number of cylinders and horsepower is easily identifiable in Figure 20. This is shown by the fact that, in the same way that the darker dots cluster towards the bottom of the graph and become lighter as they reach the top of the graph, those with a large diameter also cluster near the lower part and decrease in size as they ascend.

The efficiency of bubble charts is highlighted by the fact that capturing just some of the information provided by this single bubble chat - that is, the information regarding the car models, the miles covered per gallon, and the horsepower of each of these, requires three different scatter plots. These are displayed in Figure 17. The contrast between the simplicity and readability of Figure 20, and the density and complexity of Figure 17 emphasises the advantages of bubble charts in representing datasets with a higher number of variables.

4.4 Simple Linear Regression

Regression models are statistical tools that provide functions to estimate the relationship between the response variable and one or more explanatory variables. Regression analysis is widely adopted by data scientists, who use large datasets to build predictive models for trend forecasting. The following paragraphs will introduce simple linear regression models and demonstrate their usage using the mtcars dataset.

4.4.1 Theory of Simple Linear Regression

Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ denote n explanatory variables and let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ denote n corresponding response variables.

In a simple linear model, it is assumed that the response variables $Y_1, \ldots, Y_n$ are uncorrelated with a common variance σ^2, and their expectations are given by $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$. The expectations generated by β_0 and β_1 given x_i can be expressed as:

$$\mathrm{E}(\mathbf{Y}|\mathbf{x}) = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta_0 + \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \beta_1 = \mathbf{1}_n \beta_0 + \mathbf{x}\beta_1, \tag{4.1}$$

where $\mathbf{1}_n$ is an n-vector of 1's.

Given design matrix $\mathbf{X}$ where $\mathbf{x}_i = (1, x_i)$ and $\beta = (\beta_0, \beta_1)^T$, then $E(Y_i|x_i) = \mathbf{x}_i\beta$. These assumptions can be equivalently written in the vector form:

$$\mathrm{E}(\mathbf{Y}|\mathbf{x}) = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\mathbf{Y}|\mathbf{x}) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}_n. \tag{4.2}$$

4.4.2 Theory of Least Squares Estimation

The residual sum of squares (RSS) is a measure of the goodness of fit in a regression model, where residuals are the differences between the response variables y_i and responses generated by the regression model $\mathrm{E}(\mathbf{Y}_i|\mathbf{x})$. In least squares estimation, the goal is to find values of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ to minimise the RSS, denoted by Q:

$$\mathrm{Q} = \sum_{i=1}^{n} [y_i - \mathrm{E}(Y_i|\mathbf{x})]^2 = [\mathbf{y} - \mathrm{E}(\mathbf{Y}|\mathbf{x})]^T [\mathbf{y} - \mathrm{E}(\mathbf{Y}|\mathbf{x})] = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{4.3}$$

where $\mathbf{y}$ is n-vector of response variables and $\mathbf{X}$ is the $n \times 2$ design matrix. The partial derivative of Q with respect to vector $\boldsymbol{\beta}$ is:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = 2(\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^T\mathbf{y}), \tag{4.4}$$

Equating $\frac{\partial Q}{\partial \boldsymbol{\beta}} = \mathbf{0}$, the vector $\hat{\boldsymbol{\beta}}$, the least squares estimate of $\boldsymbol{\beta}$, can be written as:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{4.5}$$

The least squares estimate of $\boldsymbol{\beta}$ is given by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{4.6}$$

The least squares estimate $\hat{\boldsymbol{\beta}}$ is an unbiased estimator:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathrm{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

4.4.3 Case Example: 1970s Automobiles

In this section, we will study the performance of 1970s automobiles using the mtcars dataset, employing the method of linear regression. Performance is measured in Miles per Gallon (mpg); the higher the mileage, the more efficient the automobile. We will start with the visualisation of a simple linear regression model, followed by the discussion of linear regression models and the model selection method.

In the preliminary stages of data exploration, calculating the correlation matrix is a crucial step before engaging in regression modeling, as shown in Figure 21. In real-world scenarios, variables are often correlated, and entirely independent relationships are seldom encountered. Therefore, analysing pairwise correlations becomes essential. This helps in understanding multicollinearity issues within the model. Multicollinearity occurs when one covariate within the model can be accurately predicted from another covariate. When this happens, the coefficient estimates of the model can change unpredictably due to minor changes in the data.

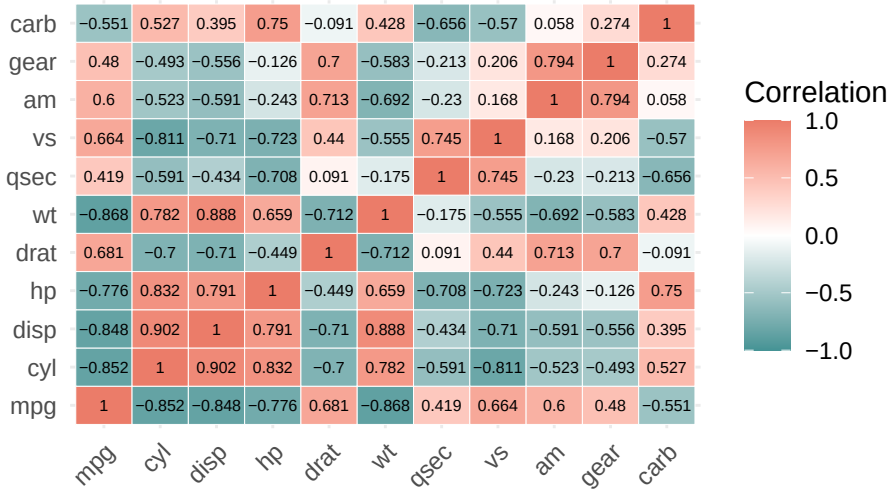

Figure 21: Correlation matrix of all variables in mtcars dataset

For the simple linear regression model, the response variable is Miles per Gallon (mpg), and we select weight (wt) as the explanatory variable. Note that mpg and wt are highly correlated, with a correlation coefficient of -0.868. This suggests that wt may have strong predictive power for mpg. Use the R function `lm()` to calculate the linear regression model, with the summary displayed below.

Observe that the t-test yields a p-value of 1.29×10^{-10}, which is less than 0.001. This indicates that the variable wt holds high statistical significance in this model. For the fitted model, the slope is $\beta_1 = -5.3445$, meaning that for every increase of 1000 lbs, the car efficiency decreases by 5 miles per gallon. The RSS is 3.046. The simple linear regression line is displayed in Figure 22.

```
# SUMMARY FOR SIMPLE LINEAR REGRESSION, WITH RESPONSE VARIABLE MILES PER GALLON
Modelwt <- lm(formula = mpg ~ wt, data = mtcars)
summary(Modelwt)

##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

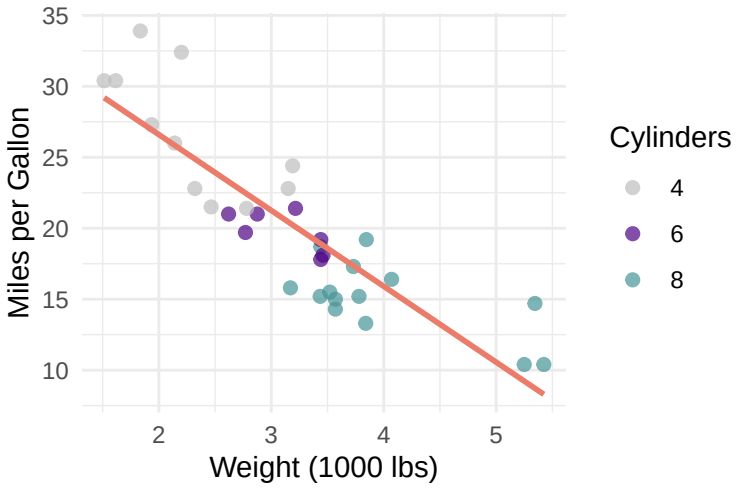

Figure 22: Scatter plot of car weights vs MPG

4.5 LOESS Regression

Locally weighted scatterplot smoothing (LOWESS) is a local regression method for a response variable Y on a single predictor X. LOWESS was proposed by William S. Cleveland in 1979, it is the special case to higher dimention locally estimated scatterplot smoothing (LOESS) regression method. They are non-parametric regression methods that fits a linear regression model for each neighbourhood of data. Unlike GLM, the LOESS model does not provide a global function to fit the data; rather, it fits neighborhoods of the data.

4.5.1 Theory of LOESS Regression

Suppose we have n ordered observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with predictors x_i and response variables y_i. Assume a model of the form

$$y_i = g(x_i) + \varepsilon_i,$$

where g is an unknown smooth function and ε_i is i.i.d.Gaussian error terms with mean 0 and variance σ^2.

Let $\Delta_i(x) = |x - x_i|$ represent the horizontal distance between x and x_i. Let $\Delta_{(i)}(x)$ denote the ordered distances to x, arranged from smallest to largest.

For each point (x, y) consider the neighbourhood to fit a line segment around it. The size of the neighbourhood indicates the number of data points used to fit the line segments and is determined by the smoothing parameter α. For $\alpha \leq 1$, each neighbourhood consists of $\frac{n}{\alpha}$ number of points [24]. Each point (x_i, y_i) in the neighbourhood of point (x, y) is assigned with a weight, "importance" of the data point to the line segment [31]. The weight of (x_i, y_i) is denoted as $w_i(x)$, which is defined as:

$$w_i(x) = T(\Delta_i(x); \Delta_{(q)}(x)). \tag{4.7}$$

Tricube weight function $T(u,t)$ is defined as:

$$T(u;t) = \begin{cases} (1-(u/t)^3)^3 & \text{for } 0 \le u < t \\ 0 & \text{for } u \ge t \end{cases}, \tag{4.8}$$

where u is the horizontal distance between neighbourhood point of interest (x_i, y_i) to point (x, y) and t is the maximum distance threshold. For points in the neighbourhood with horizontal distance exceeding threshold t, they are assigned with weight 0 by the tricube weight function.

For $\alpha > 1$, for each point (x, y) the neighbourhood include all points. The maximum distance threshold is assumed to be $\alpha^{\frac{1}{p}}$, for p explanatory variables. The weight is given by:

$$w_i(x) = T(\Delta_i(x); \Delta_{(n)}(x))\alpha. \tag{4.9}$$

The line segment is fitted by weighted least square, and this is the preliminary estimation $\hat{g}(x)$ for every point (x, y).

Both LOWESS and LOESS penalise the outliers in the preliminary estimation by assigning points further away from $\hat{g}(x)$ with less weight than before. The robustness weight (the adjusted weight) is given by:

$$r_i = B(\hat{\varepsilon}_i, 6\text{m}),$$

where bisquare weight function is defined as:

$$B(u;b) = \begin{cases} (1-(u/b)^2)^2 & \text{for } 0 \le |u| < b \\ 0 & \text{for } |u| \ge b \end{cases}, \tag{4.10}$$

and the median absolute residual m is defined as $\text{m} = \text{median}(|\hat{\varepsilon}_i|)$, the residual $\hat{\varepsilon}_i$ is defined as $\hat{\varepsilon}_i = y_i - \hat{y}(x_i)$.

Note the process of weight adjustment using previous estimation is iterated several times until a smooth curve is obtained.

The updated model estimate $\hat{g}(x)$, is computed using the local fitting method, but with the neighborhood weights $w_i(x)$ replaced by $r_i w_i(x)$ [10].

4.5.2 Case Example: Estate price in Taipei

Property valuation can be modelled as a regression problem, in this analysis, we delve into the pricing of properties based on their age. Using the Taipei Housing dataset, we investigate the relationship between the age of houses (measured in years) and their prices (measured in New Taiwan dollars per unit area). Linear regression model and the LOESS regression model are used, as shown in Figure 23.

The linear model has a residual standard error of 13.32, while the LOESS model has a residual standard error of 12.16. Cross-validation is a method to compare model goodness of fit. LOESS, with its ability to capture local variations may outperform linear regression, which minimises the least squares estimate $\hat{\beta}$.

```r
# LINEAR REGRESSION VS. LOESS MODEL
ggplot(data = estate, aes(x = house_age, y = price_per_area)) +
  geom_point(size =0.5) +
  theme_minimal()+
  scale_x_continuous(labels = scales::number_format(scale = 1)) +
  scale_y_continuous(labels = scales::number_format(scale = 1)) +
  geom_smooth(method = "loess", se = FALSE, color = "#EB7C69", span = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "#459395", formula = y ~ x) +
  labs(title = "House Age vs. House Price",
       x = "house age (year)",
       y = "price (unit area)") +
  theme_minimal()
```

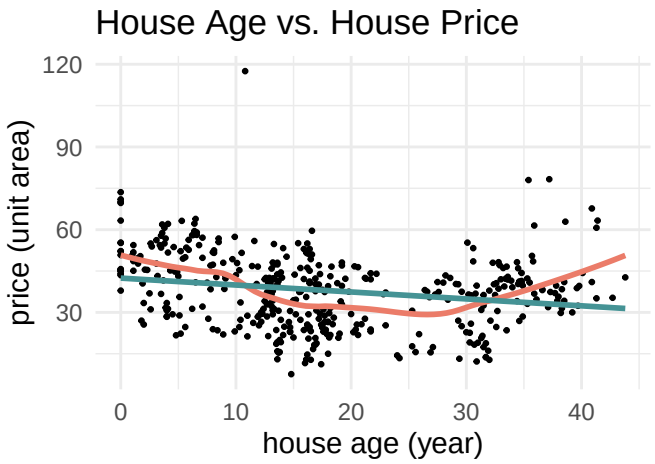

Figure 23: Estate valuation using house age, linear regression and smooth regression methods

Chapter Overview and Further Reading References

This chapter delves into the theory of scatter plots, and bubble charts, as well as simple linear regression and LOESS regression. These concepts are firmly established the existing statistical literature. Further insights into scatter plots and bubble charts can be found in Leland Wilkinson's *The Grammar of Graphics* (2005), while comprehensive details on simple linear regression theory are available in *Introduction to Linear Regression Analysis* (2021) by Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. Additionally, *Local Regression Models* (1992) by William S. Cleveland, Eric Grosse, and William M. Shyu offers an in-depth exploration of LOESS regression theory.

5 Visualising Beyond Two Dimensions

This chapter extends beyond the realm of two dimensional data visualisations. Within this chapter, advanced techniques are explored to unravel complex relationships involving multiple variables [37]. The Principal Component Analysis (PCA) is introduced as a powerful method for dimensionality reduction, enabling a concise representation of high-dimensional datasets. Multiple innovative approaches facilitating the visualisation of high-dimensional data in lower-dimensional biplots and t-distributed Stochastic Neighbor Embedding (t-SNE) are studied.

5.1 Principal Component Analysis (PCA)

5.1.1 Theory of PCA

Principal Component Analysis (PCA) is one of the most widely-used dimensionality reduction techniques. The basic idea is to use the dependencies between the variables to project the data into a low-dimensional subspace, without losing too much information.

The feature matrix $\mathbf{X}$ is an $N \times D$ real-valued matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ & \vdots & & \\ x_{N,1} & x_{N,2} & \cdots & x_{N,D} \end{pmatrix},$$

with N observations and D variables (also known as features).

Define a linear and orthogonal mapping $\mathbf{W} : \mathbf{x} \to \mathbf{z}$, acting as a 'transformer' between observation $\mathbf{x}$ and its low-dimensional projection $\mathbf{z}$. The matrix $\mathbf{W}$ is a $D \times L$ orthogonal and normalised matrix, where its column vectors are represented as $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L]$. Then, for all $i \neq j$, $\mathbf{w}_i \mathbf{w}_j = 0$ and $\mathbf{w}_i^T \mathbf{w}_i = 1$. This matrix allows for the projection of data from its original high-dimensional space to the low-dimensional subspace using $\mathbf{z} = \mathbf{W}^T \mathbf{x}$, known as encoding. To reconstruct $\mathbf{x}$ from $\mathbf{z}$, $\hat{\mathbf{x}} = \mathbf{W}\mathbf{z}$ will give the high-dimensional approximation of the observation, this process is known as decoding.

The optimal solution of PCA is obtained by minimising the loss function, which represents the reconstruction error subject to constraint $\mathbf{W}$. The loss function is defined as follows:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}_n - \hat{\mathbf{x}}||_2^2. \tag{5.1}$$

5.1.2 Derivation of PCA

In general, the vector $\mathbf{w}_i$ is given by the equation $\mathbf{\Sigma}\mathbf{w}_i = \lambda \mathbf{w}_i$, where $\mathbf{\Sigma}$ is the covariance matrix and λ is the i^{th} largest eigenvalue of $\mathbf{\Sigma}$, for $i = 1, \ldots, L$.

Define the covariance matrix $\mathbf{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$. Without loss of generality, let's assume that all observations in matrix $\mathbf{X}$ are centred around 0 mean, then $\bar{\mathbf{x}} = 0$. The covariance matrix can be written in a simple form:

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T = \frac{1}{N}\mathbf{X}\mathbf{X}^T. \tag{5.2}$$

The optimal encoding $\mathbf{Z}$ ensures minimum information loss when data is projected to a lower-dimensional subspace. To determine this optimal encoding, we minimise the reconstruction error with respect to $\mathbf{w}$ and $\mathbf{z}$.

$$\begin{aligned}\mathcal{L}(\mathbf{w},\mathbf{z}) &= \frac{1}{N}\sum||\mathbf{x}_n - \mathbf{W}\mathbf{z}_n||^2 \\ &= \frac{1}{N}\sum(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)^T(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n) \\ &= \frac{1}{N}\sum\mathbf{x}_n^T\mathbf{x}_n - 2\mathbf{x}_n^T\mathbf{W}\mathbf{z}_n + \mathbf{z}_n^T\mathbf{W}^T\mathbf{W}\mathbf{z}_n.\end{aligned} \tag{5.3}$$

To determine the optimal encoding $\mathbf{z}_n$ for the n^{th} observation, we set the derivative with respect to $\mathbf{z}_n$ equal to 0:

$$\frac{\partial\mathcal{L}(\mathbf{w},\mathbf{z})}{\partial\mathbf{z}_n} = \frac{1}{N}(-2\mathbf{x}_n^T\mathbf{W} + 2\mathbf{z}_n) = \mathbf{0}.$$

Thus we find that $\mathbf{z}_n = \mathbf{W}^T\mathbf{x}_n$ yields the optimal encoding for the n^{th} observation.

Next, we aim to find the matrix $\mathbf{W}$ which gives the optimal encoding. We start by expanding and simplifying the loss function $\mathcal{L}(\mathbf{W})$ from equation 5.3:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{N}\sum_{n=1}^{W}\mathbf{x}_n^T\mathbf{x}_n - \sum_{n=1}^{L}\mathbf{W}_l^T(\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T)\mathbf{W}_l. \tag{5.4}$$

WLOG, our data is assumed to be centred around 0 mean (by 5.2), the loss function can be represented as:

$$\mathcal{L}(\mathbf{W}) = \text{constant} - \sum_{l=1}^{L}\mathbf{W}_l^T\mathbf{\Sigma}\mathbf{W}_l. \tag{5.5}$$

Since $\mathbf{W}$ is a normalised matrix, it has the property $\mathbf{W}_i^T\mathbf{W}_i = 1$ for all $i = 1,\ldots,N$. Therefore, when computing the partial derivative of $\mathcal{L}(\mathbf{W})$ with respect to vector $\mathbf{w}_1$, we can apply the Lagrange multipliers to $\mathcal{L}(\mathbf{w}_1)$ and calculate the derivative of $\hat{\mathcal{L}}(\mathbf{w}_1)$ instead:

$$\hat{\mathcal{L}}(\mathbf{w}_1) = -\mathbf{w}_1^T\mathbf{\Sigma}\mathbf{w}_1 + \lambda(\mathbf{w}_1^T\mathbf{w}_1 - 1), \tag{5.6}$$

Equating $\frac{\partial\hat{\mathcal{L}}(\mathbf{w}_1))}{\partial\mathbf{w}_1} = \mathbf{0}$, the weight vector $\mathbf{w}_1$ which minimise the loss function can be expressed as [29]:

$$\frac{\partial\hat{\mathcal{L}}(\mathbf{w}_1))}{\partial\mathbf{w}_1} = 2\mathbf{\Sigma}\mathbf{w}_1 + 2\lambda\mathbf{w}_1 = \mathbf{0}. \tag{5.7}$$

Therefore $\mathbf{\Sigma}\mathbf{w}_1 = \lambda\mathbf{w}_1$ that is, $\mathbf{w}_1$ is the eigenvector of $\mathbf{\Sigma}$ with eigenvalue λ.

Premultiply by $\mathbf{w}_1^T$, $\mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1 = \lambda\mathbf{w}_1^T\mathbf{w}_1 = \lambda$. To minimise the loss $\mathcal{L}(\mathbf{w}_1)$ we maximise $\mathbf{w}_1^T\boldsymbol{\Sigma}\mathbf{w}_1$. Thus the largest eigenvalue λ would minimise the loss $\mathcal{L}(\mathbf{w}_1)$.

Adopting the same methodology, the second largest eigenvalue λ_2 would minimise the loss $\mathcal{L}(\mathbf{w}_2)$.

Therefore, order the eigenvectors by their corresponding eigenvalues $\{\mathbf{w}_1, \ldots, \mathbf{w}_L\}$. These form the columns of the weight matrix $\mathbf{W} \in \mathbb{R}^{D\times L}$ [9].

5.2 Biplots

A biplot is a graphical representation that succinctly captures the relationships between variables and observations in a reduced-dimensional space defined by the principal components. This visualisation method facilitates the exploration of complex multivariate data by providing a simultaneous display of both samples (observations) and variables in a single plot. They are commonly understood to be a generalisation of the simple two-variable scatterplot.

At their core, biplots offer a graphical representation of a standardised data matrix, whose rows n are the samples, and whose columns p are the variables, by projecting these onto a two-dimensional plane. To do this, mathematical computations derived from techniques like Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are required.

In recent years, there have been substantial developments in biplots [13]. The technique has expanded far beyond what is now referred to as the "classical biplots" or "PCA biplots". The concept now finds broad application in conjunction with various other techniques of multivariate analysis. Nevertheless, this section studies the foundations of this visualisation technique, and thus, the emphasis is on biplots within the framework of PCA.

5.2.1 Construction of PCA Biplots

Singular Value Decomposition
Principal component biplots are based on singular value decomposition of the $(n \times p)$ data matrix $\mathbf{X}$ of n observations on p variables with rank r. The singular value decomposition of the standardised matrix $\mathbf{X}$ is found to be the following:

$$\mathbf{X} = \mathbf{UBS}^T,$$

where $\mathbf{U}$ is an $(n \times r)$ orthogonal matrix, and $\mathbf{S}$ is an $(p \times r)$ orthogonal matrix. The columns of $\mathbf{U}$ and $\mathbf{S}$ are the left and right singular vectors. $\mathbf{B}$ is a $(r \times r)$ diagonal matrix with entries $b_1, \ldots, b_r$ such that $b_1 \geq \ldots b_r \geq 0$ [19]. These entries the singular values discussed later on.

In order to find this singular value decomposition, firstly we require the computation of the orthogonal diagonalisation

$$\mathbf{X}^T\mathbf{X} = \mathbf{PDP}^T,$$

where $\mathbf{D}$ is the diagonal matrix of eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{P}$ is an orthogonal matrix composed of the normalised eigenvectors of $\mathbf{X}^T\mathbf{X}$.

Now, $\mathbf{B}$ and $\mathbf{S}$ are computed. Matrix $\mathbf{S}$ is equivalent to the matrix $\mathbf{P}$ previously computed, and matrix $\mathbf{B}$ is a $(n \times p)$ diagonal matrix, where the diagonal entries are the "singular values" with any additional rows and columns of zero to make the dimensions of $\mathbf{B}$ match those of $\mathbf{X}$. The singular values referenced are square root of the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Hence matrix $\mathbf{B}$ is of the form

$$\mathbf{B} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Finally, the $\mathbf{U}$ orthogonal matrix is calculated. The i-th column of $\mathbf{U}$ is given by

$$\mathbf{u}_i = \frac{1}{\sigma_1}\mathbf{X}\mathbf{s}_i,$$

where σ_i are the singular values, and $\mathbf{s}_i$ are the columns of the matrix $\mathbf{S}$ [2].

PCA Biplot Construction

Now, the heuristic argument motivating PCA biplots is the following. Define $\mathbf{B}^\alpha$, for $0 \leq \alpha \leq 1$, as a diagonal matrix who's diagonal elements are $b_1^\alpha, \ldots, b_r^\alpha$. The definition for $\mathbf{B}^{1-\alpha}$ is similar [18]. Then, with the decomposition of matrix $\mathbf{X}$ found through SVD, the data can be further rewritten as

$$\mathbf{X} = \mathbf{U}\mathbf{B}^\alpha\mathbf{B}^{1-\alpha}\mathbf{S}^T = \mathbf{U}\mathbf{B}\mathbf{S}^T = \mathbf{G}\mathbf{H}^T,$$

where

$$\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_n] = \mathbf{U}\mathbf{B}^\alpha, \quad \mathbf{H}^T = [\mathbf{h}_1, \ldots, \mathbf{h}_p]^T = \mathbf{B}^{1-\alpha}\mathbf{S}^T.$$

Hence, the ith observation of the jth variable can be rewritten as

$$x_{ij} = \mathbf{g}_i^T\mathbf{h}_j.$$

Both the $\mathbf{g}_i$ and $\mathbf{h}_j$ have r elements, and hence, if $\mathbf{X}$ has rank 2, all could be plotted as points in two-dimensional space [18]. Note that for the more general case, where $r > 2$, x_{ij} can be written as

$$x_{ij} = \sum_{k=1}^{r} u_{ik}b_k s_{jk}$$

which is often well approximated by

$$_m\tilde{x}_{ij} = \sum_{k=1}^{m} u_{ik}b_k s_{jk}, \quad \text{with } m < r.$$

But this can be written

$$_m\tilde{x}_{ij} = \sum_{k=1}^{m} \mathbf{g}_{ik}\mathbf{h}_{jk} = \mathbf{g}_i^{*T}\mathbf{h}_j^*,$$

where $\mathbf{g}_i^*, \mathbf{h}_j^*$ contain the first m elements of $\mathbf{g}_i$ and $\mathbf{h}_j$ respectively. Hence, suggesting that if instead of using the $\mathbf{g}_i$ and the $\mathbf{h}_j$, we can just use their first elements (say, 2), respectively denoted by $\mathbf{g}_i^*$ and $\mathbf{h}_j^*$, to get

$$x_{ij} \approx \mathbf{g}_i^{*T}\mathbf{h}_j^*.$$

Thus, $\mathbf{g}_i$ and $\mathbf{h}_j$ should provide a reasonable two-dimensional approximation of the n observations and the p variables [19].

5.2.2 Case Example: Vintage cars attributes

This section presents a comparative analysis of PCA biplots generated through two distinct methodologies: manual mathematical construction and use of R's in-built functions. The focus lies on elucidating the mathematical intricacies, exploring relations, and highlighting differences between the two approaches. The mtcars dataset will serve as a benchmark.

Methodologies Discussion
Traditionally, biplots are constructed manually through mathematical operations based on SVD discussed previously. However, with computational tools like R, in-built functions such as `prcomp()` and `biplot()` streamline the process, removing tidious mathematical intricacies, which becomes particularly interesting when working with larger datasets.

The manual construction, which results in the biplot shown in Figure 24, begins with data standardisation followed by SVD decomposition to obtain matrices representing row and column contributions to principal components. These matrices are then utilised to construct the biplot. Conversely, the in-built function approach employs R's `prcomp()` function for PCA and directly generates the biplot using `biplot()`, without the need for explicit mathematical operations. The product of this method is shown in Figure 24.

```
#MANUAL CONSTRUCTION OF A BIPLOT IN R
# Standardise the data
scaled_data <- scale(mtcars)

# Set the alpha parameter for SVD
alpha <- 0

# Preprocess the data using SVD
preprocess <- scaled_data - matrix(colMeans(scaled_data), nrow = nrow(scaled_data), ncol = ncol(scaled_data), byrow = TRUE)
svd_mtcars <- svd(preprocess)
U <- svd_mtcars$u
L <- diag(svd_mtcars$d)
A <- svd_mtcars$v

# Calculate G and H matrices
G <- U %*% (L%^%(alpha))
H <- t((L%^% (1-alpha)) %*% t(A))

# Extract the first two columns for biplot
G2 <- G[, 1:2]
H2 <- H[, 1:2]

# Set up a side-by-side layout
par(mfrow = c(1, 2))

# Create a biplot-like plot for manual construction with title
plot(G2, xlim = c(-0.25, 0.4), xlab = "PC1", ylab = "PC2", main = "Manual Mathematical Construction")

par(new = TRUE)
plot(H2, xlim = c(-4.5, 8), ylim = c(-4.5, 8), col = "#EB7C69",
     xaxt = "n", yaxt = "n", xlab = "", ylab = "")
for (i in 1:ncol(scaled_data)) {
  lines(c(0, H2[i, 1]), c(0, H2[i, 2]), col = "#EB7C69", lwd = 1.5)
}

axis(3)
axis(4)

#GENERATION OF A BIPLOT WITH R'S BUILT-IN FUNCTIONS
# Standardise the data
scaled_data <- scale(mtcars)

# Perform Principal Component Analysis (PCA)
```

```r
pca_result <- prcomp(scaled_data, scale. = TRUE)

# Create a biplot with title
biplot(pca_result, cex = 0.7, main = "R's Built-in Function")
```

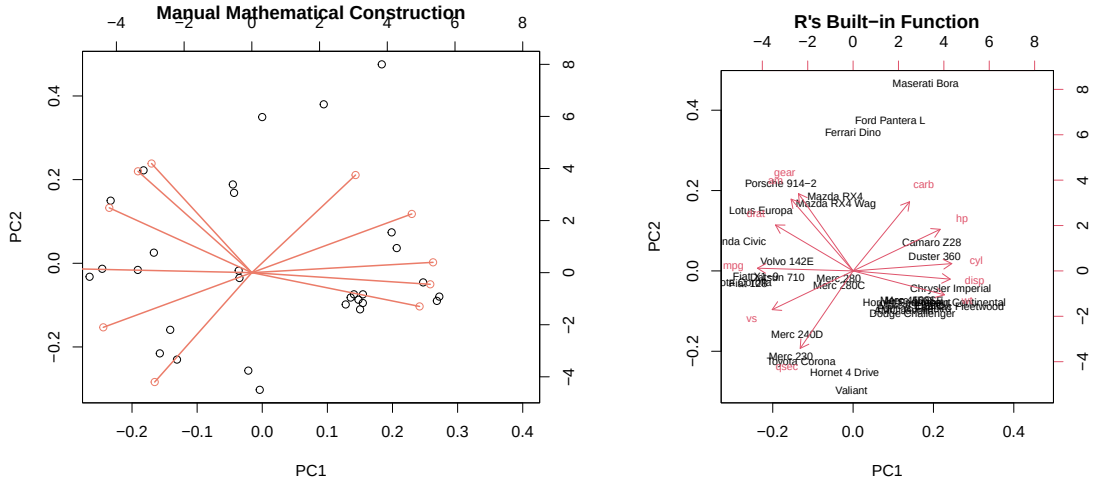

Figure 24: PCA biplots of mtcars dataset

Comparative analysis of these two plots reveals several key findings. The manual construction approach offers meticulous control over the biplot's customisation, allowing for fine-tuning of parameters and appearance. Conversely, the in-built function approach provides a user-friendly ideal for practitioners less inclined towards mathematical intricacies the visualisation. However, despite differences in implementation, both methodologies produce visually comparable biplots, showcasing identical patterns and relationships within the data.

5.3 Principal Curves

Principal curves stand as a vital tool in exploratory data analysis, providing a nonlinear analogue to PCA for capturing the underlying structure of multidimensional data. While PCA focuses on finding linear projections that maximise variance, principal curves seek to identify smooth, nonlinear trajectories that capture the most significant variations in the data. These curves are essentially one-dimensional paths embedded within the multidimensional space of the data points.

Hence, if the points don't fall in a linear subspace, that is, they fall in a non-linear subspace such as a curve in 2-dimension, PCA is insufficient. An example of this is the scatter plot on the left side of Figure 25. In this case, a straight line can't be placed in the middle of the plot, however, one could fit a curve that passes through the centre of the points.

5.3.1 Construction of Principal Curves

As with PCA, take $(n \times p)$ data matrix $\mathbf{X}$, with row vectors $\mathbf{x}_i$. The goal is to extract a lower-dimensional set of "factors" or "scores" $\mathbf{y} = (y_1, \ldots, y_n)^T$. These satisfy the following model:

$$\mathbf{x}_i = f(y_i) + \epsilon_i,$$

where ϵ_i are errors with mean 0 and variance $\sigma^2\mathbf{I}_p$ and $f : \mathbb{R}^1 \to \mathbb{R}^p$ is a continuous function called a curve [19].

For curve $\mathbf{f}$, the projection index $y_f(\mathbf{x})$ maps observation $\mathbf{x} \in \mathbb{R}^p$ to the point of $\mathbf{f}$ that is closest, returning the score (index). If there are several such points, it returns the largest (this is arbitrary choice to ensure it is a well defined function). Hence, the projection index is [19]:

$$y_f(\mathbf{x}) = \sup_y\{y : ||\mathbf{x} - \mathbf{f}(y)|| = \inf_\mu ||\mathbf{x} - \mathbf{f}(\mu)||\}.$$

As seen in a similar form in PCA, principal curves estimates the following least-squares objective function

$$\min_{\mathbf{f}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{f}(y_{\mathbf{f}}(\mathbf{x}_i))||^2.$$

To construct the principal curve, an algorithm put to work. This iterates between finding a curve $\mathbf{f}$ and then projecting the points onto that. More explicitly, the algorithms Principal curves!algorithmhas the following form [19]:

1. Initialise: let iteration counter $h = 1$ and set $\mathbf{y}^{(0)} = \mathbf{Xu}$, where $\mathbf{u}$ is the first PC vector.
2. Smoothing step: With $\mathbf{y}^{(h-1)}$ fixed, estimate $\hat{\mathbf{x}}_i^{(h)}$ with a smoother.
3. Projection step: With $\hat{\mathbf{X}}^{(h)}$ fixed, use the projection index to update the scores $\mathbf{y}(h)$ so that they have unit speed.
4. Loop: Increment h and return to step 1 while the change in the objective function above some threshold.

5.3.2 Principal Curves in Practice

In Figure 25, we see an example of how a principal curve is fitted to the observations of an artificial dataset. The points in the plots are not grouped in a linear subspace, but rather follow a similar shape to that of cubic function.

Using the built-in `principal-curve()` function, a principal curve can be fitted to the dataset. The principal curve traverses through the densest regions of the data distribution, capturing the underlying nonlinear structure. The panel on the left of Figure 25 depicts the principal curve overlaid on the dataset, providing insights into the central tendency and inherent curvature of the data. Additionally, whiskers are drawn to visualise the perpendicular distances between data points and the principal curve.

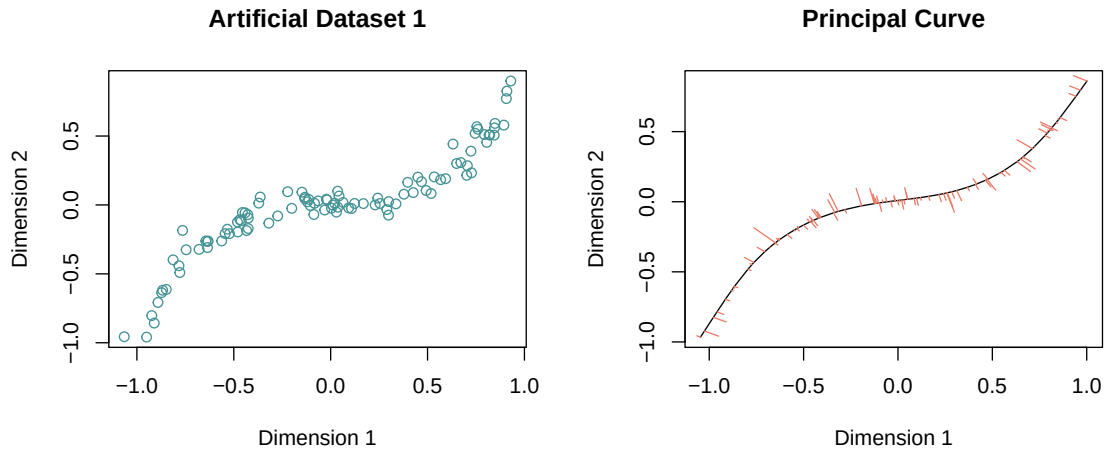

Figure 25: Principal curve fitting illustration on artificial dataset

The principal curve effectively captures the nonlinear relationship between the dimensions of the artificial dataset. By flexibly adapting to the curvature of the data distribution, the principal curve reveals intricate patterns and trends that may not be captured by linear methods like PCA. The visualisation aids in understanding the underlying structure of the dataset.

5.4 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Since PCA is a linear method and may not perform well with non-linear data structures, often missing complex non-linear relationships. Although effective in capturing the global structure of data, PCA focuses on maintaining large pairwise distances to maximise variance and overlooks important local patterns and structures.

Unlike PCA, t-SNE [35] is a nonlinear technique for embedding high-dimensional data for visualisation in a low-dimensional space of two or three dimensions, focusing on preserving the small pairwise similarities between data points thus retain the local structure of the dataset in a lower-dimensional space. It starts by converting high-dimensional Euclidean distances into probabilities that reflect the similarity between points, then maps these points to a lower-dimensional space in a way that tries to preserve these similarities.

5.4.1 Theory of t-SNE

For a given high-dimensional dataset $\{x_1, ..., x_n\}$, the t-SNE algorithm quantifies the similarity between pairs of data points, x_i and x_j, using conditional probabilities $p_{j|i}$. These probabilities are calculated under the assumption that if x_i were to choose its neighbours, it would do so in proportion to their probability density under a Gaussian distribution centered at x_i. The conditional probability $p_{j|i}$ [35] for $i \neq j$ is mathematically expressed as:

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i^2)},$$

where σ_i denotes the variance of the Gaussian that is centered on data point x_i. Also, set similarities of the same pair wise points $p_{i|i} = 0$ and $\sum_j p_{j|i} = 1$ for all i.

To create a symmetrical joint probability distribution in the high-dimensional space, t-SNE averages the conditional probabilities $p_{j|i}$ and $p_{i|j}$ [35], resulting in:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n},$$

where n is a number of data points. Also, $p_{ii} = 0$ and $\sum_{i,j} p_{ij} = 1$.

In the low-dimensional space (usually 2D or 3D), we have the corresponding dataset $\{y_1, ..., y_n\}$. Since t-SNE aims to find mapped points y_i and y_j in a low-dimensional space that reflects the similarities p_{ij} as well as possible. In the low-dimensional space, t-SNE computes similarities q_{ij} using a similar formula but with a Student t-distribution (with one degree of freedom, equivalent to the Cauchy distribution) which has heavier tails than a Gaussian distribution to compute the joint probabilities [35] of the mapped points:

$$q_{ij} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}}.$$

Similarly, set $q_{ii} = 0$.

The goal of t-SNE is to minimize the discrepancy between the probabilities p_{ij} in the high-dimensional space and q_{ij} in the low-dimensional space. Therefore, the objective is to minimise the Kullback-Leibler (KL) divergence [35] between the joint probability distributions P and Q:

$$C = \mathrm{KL}(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

where the C means the cost function that we wish to minimize, and the Kullback-Leibler (KL) divergence is a measure of how probability distribution P differs from probability distribution Q. In addition, the lower the KL divergence value, the closer the two distributions are. A KL divergence of zero means that these two distributions are the same.

To continue with the minimisation, compute the gradient of symmetric SNE [35] as follows:

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1}.$$

Once the gradient of the cost function with respect to the positions of the points in the low-dimensional space, $\frac{\partial C}{\partial y_i}$, is computed, t-SNE employs gradient descent to minimizes the Kullback-Leibler (KL) divergence between the high-dimensional probability distribution P and the low-dimensional probability distribution Q. Gradient descent is an iterative optimisation algorithm used to minimize the cost function by updating the parameters in the opposite direction of the gradient of the cost function. In the case of t-SNE, the parameters are the positions of the points in the low-dimensional space. The update rule [35] at each iteration t for a point y_i is given by:

$$y_i^{(t+1)} = y_i^{(t)} - \eta \frac{\partial C}{\partial y_i},$$

where η is the learning rate, a positive scalar determining the step size.

The process is repeated for a number of iterations or until the change in the cost function between iterations is below a predetermined threshold, indicating convergence. The gradient descent process in t-SNE is crucial for effectively reducing the dimensionality of high-dimensional data while preserving the local structure of the data. By iteratively updating the positions of the points in the low-dimensional space, t-SNE minimizes the KL divergence, resulting in a meaningful representation of the data in lower dimensions.

5.4.2 Case Example: Classifiication of penguins

In Figure 26 we observe a t-SNE visualisation of the penguin dataset. In this case, using the t-SNE algorithm helps us to reduce the multidimensional numerical data of the penguin dataset to a 2-dimensional space. We also use scatter plots and elliptical areas to visualise the distribution of different penguin species. The data points are colour-coded to distinguish between three penguin species: Adelie in red, Chinstrap in green, and Gentoo in blue. This colour coding aids in the visual differentiation of the species clusters.

```
# PLOT OF T-SNE OF PENGUINS DATASET
penguins %>%
  select(where(is.numeric),-year,species) %>%
  mutate(ID=row_number()) %>%
  column_to_rownames("ID") %>%
  na.omit()-> df

tSNE_fit<-df %>% select(-species) %>% scale() %>% Rtsne()
tSNE_fit$Y %>%
  as.data.frame() %>%
  rename(tSNE1 = "V1",tSNE2 = "V2") %>%
  mutate(Species = df$species) -> tSNE.plot

ggplot()+
  geom_point(data = tSNE.plot, aes(x = tSNE1,y = tSNE2,color = Species)) +
  stat_ellipse(data = tSNE.plot,
               geom = "polygon",
               aes(x = tSNE1,y = tSNE2,
                   group = Species,
                   fill = Species),
               alpha = 0.5,
               lty = "dashed",
               color = "black",
               key_glyph = "blank") +
  theme_bw()
```

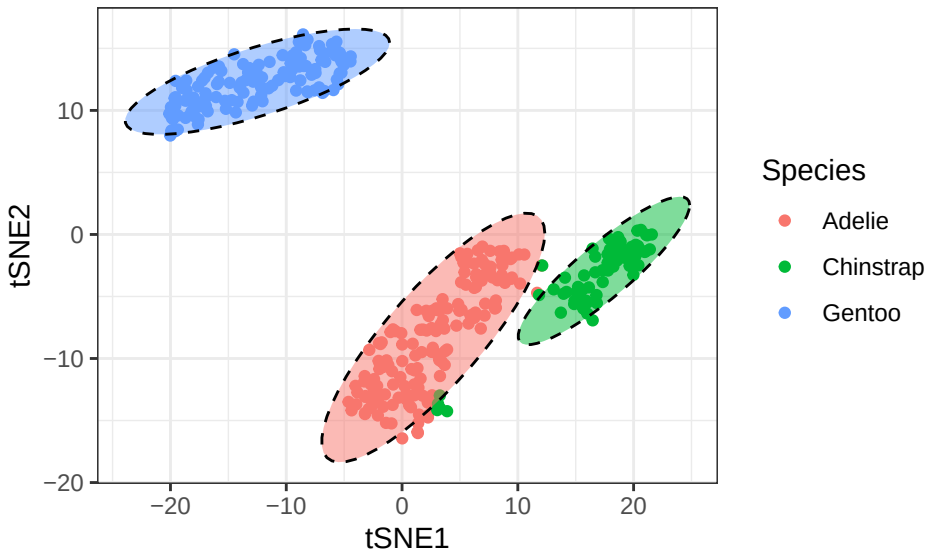

Figure 26: t-SNE of Penguins Dataset

The axes are labeled $tSNE1$ and $tSNE2$, corresponding to the dimensions reduced by the t-SNE algorithm. Dashed ellipses overlaid on the scatter plot demarcate the clusters of each species, providing a visual guide to the density and separation of the species within the transformed feature space. The plot effectively uses the t-SNE technique to illustrate the grouping of species, highlighting the algorithm's utility in discerning inherent data patterns in a lower-dimensional representation.

While t-SNE is an exceptionally potent tool for data visualisation, it does come with its own set of constraints. Firstly, it has a significant memory footprint and can be time-consuming to run, which may pose challenges when dealing with large datasets.

Secondly, t-SNE is tailored specifically for visualisation, which constrains the embedding space to two or three dimensions. This limitation means that t-SNE is optimised for human interpretability rather than for capturing higher-dimensional relationships.

Additionally, it requires experimentation with different initialisation to mitigate the risk of local suboptimal solutions affecting the results. Therefore, it's crucial to carefully consider and select the most suitable method based on the specific requirements and nature of the data at hand.

Chapter Overview and Further Reading References

Chapter 5 explores the theory PCA, and its applications constructing biplots and principal curves, as well as the theory of t-SNE. These concepts date back to 1901 when PCA was first introduced by Karl Pearson [21]. Thus, they are well established in existing literature. Substantial detail on PCA, biplots and principal curves is available in *Principal Component Analysis* (2003) by Ian T. Jolliffe. Transitioning from PCA, *Visualising Data using t-SNE* (2008) by Laurens van der Maaten and Geoffrey Hinton is a seminal paper exploring the theory of t-SNE.

6 State-Of-The-Art Modern Approaches

In this chapter, we will introduce some state-of-the-art graphical statistical methods published in recent research. In Chapter 6.1, we will introduce the foundational visualisation methods box plots and Q–Q plots. Building on this, we will discuss the theory and implementation of functional boxplots and Q–Q Box plots, as published in recent articles in the *Journal of Computational and Graphical Statistics*.

6.1 Introduction

6.1.1 Box Plots in Practice

Box plots, also known as box-and-whisker plots, were introduced by John W. Tukey in 1977. Since then, they have become essential in exploratory data analysis, offering a summary of a dataset's central tendency, skewness, and outliers. These plots typically showcase the median, quartiles, and range, making them valuable for comparing and discerning patterns across different variables.

An example of box plots using the iris dataset is displayed in Figure 27. Specifically, the right graph showcases a violin plot — which is a recent advancement in box plot techniques. The violin plot combines a box plot with a kernel density function, providing a mirrored density plot displayed in the same way as a box plot.

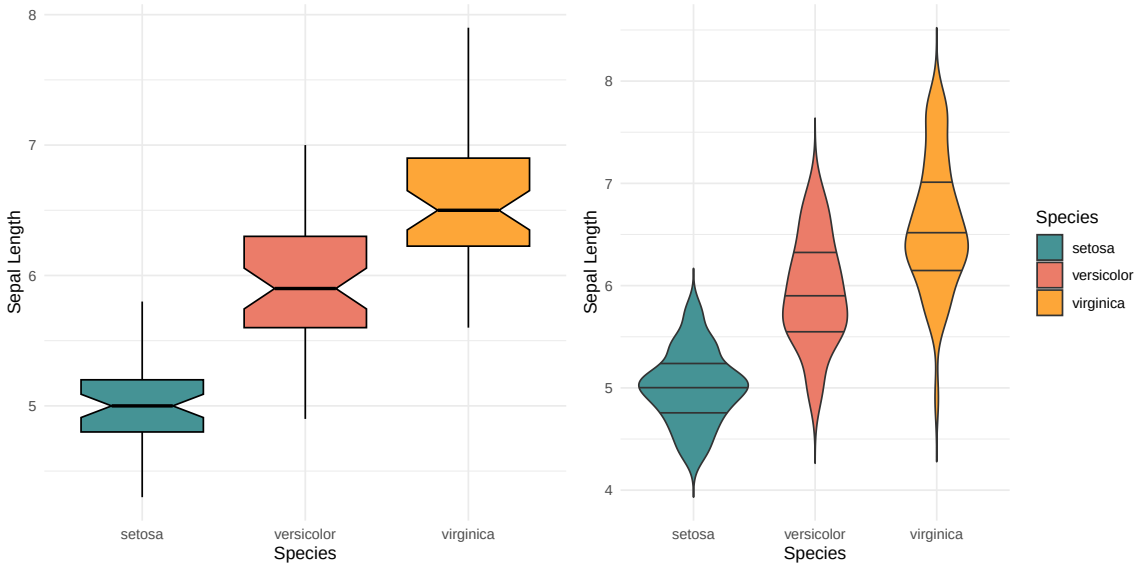

Figure 27: Comparison of Violin and Box Plots of Sepal Length by Species

6.1.2 Q–Q Plots in Practice

Q–Q plots are constructed by plotting the quantiles of two datasets against each other. Typically, one dataset serves as the basis for theoretical quantiles, often assumed to follow a particular probability distribution, while the other dataset comprises observed values. This comparative depiction

enables a direct assessment of how closely the observed data align with the theoretical distribution, facilitating the identification of deviations from expected patterns.

In this section, side-by-side QQ plots are displayed offering a visual comparison of the distributional properties of two variables from the mtcars and the iris datset. These plots serve as a means to assess normality assumptions and identify potential deviations from expected distributions. Through this visual exploration, we aim to gain insights into the underlying structure of the data and guide further investigation.

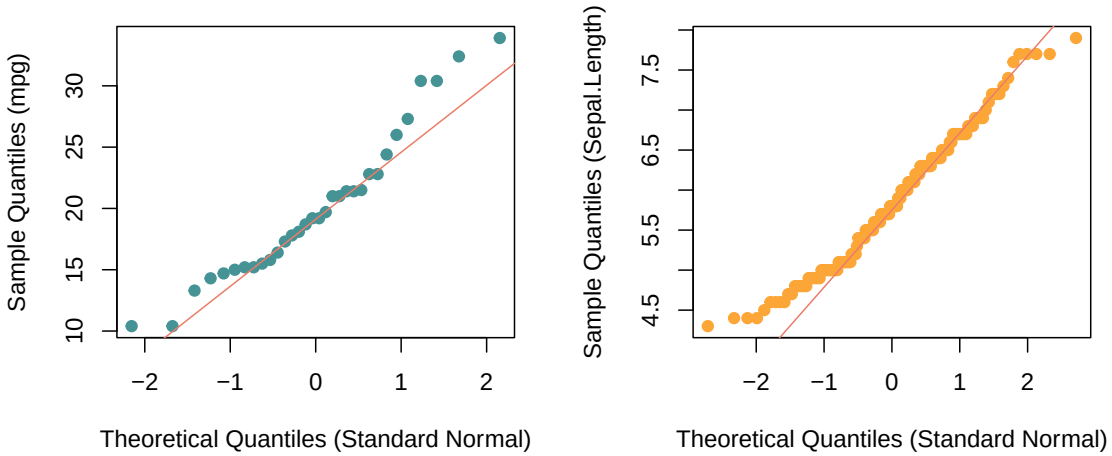

Figure 28: Q–Q plot of mpg in mtcars dataset

In the first plot in Figure 28, which examines the mpg variable in the mtcars dataset, deviations, particularly in the tails of the distribution, are noticeable. These deviations suggest potential departures from normality, possibly indicating the presence of outliers or non-normal behaviour in the data.

The second plot examines the distribution of the Sepal.Length variable in the iris dataset. Here, the points exhibit a slight departure from the theoretical quantiles line. The deviation is less pronounced compared to the previous plot. The distribution of sepal lengths within the iris dataset appears to be approximately normal.

In general, the Q–Q plots provide a succinct and insightful representation of the distributional properties of the variables under examination. They offer a visual means to assess normality assumptions and identify potential outliers or deviations from expected distributions. As such, Q–Q plots serve as a valuable tool in exploratory data analysis and model diagnostics, aiding researchers in understanding the underlying structure of their data and guiding subsequent analyses.

6.2 Functional Boxplot

Building upon the traditional box plot, the functional boxplot is an innovative extension that provides a comprehensive visual representation of functional data. The i^{th} observation of the population is represented as a real function $y_i(t)$, where $i = 1, \ldots, n, t \in \mathcal{I}$, where $\mathcal{I}$ is an interval on the real line.

This section aims to unravel the methodology behind functional boxplots, elucidating their construction, interpretation, and application in real-world datasets, which enables a deeper understanding of the patterns and anomalies inherent in functional data.

6.2.1 Theory of Funtional Boxplot

The construction of a box plot relies on data ordering, where you simply arrange univariate observations from smallest to largest. However, with multivariate data, this process becomes significantly more complex, necessitating advanced methodologies to establish a meaningful order. Various versions of data depth have been developed to assess the centrality (or depth) of an observation or its deviation (outlyingness), where observations have a population distribution of density P.

Consider a set of observations $\{y_1(t), \ldots, y_n(t)\}$ defined over an interval $\mathcal{I} \in \mathbb{R}$. The graph of a function $y(t)$ is defined as $G(y) = (t, y(t)) : t \in I$. Applying the concept of range from univariate data to two dimensions, we derive a band in $\mathbb{R}^2$ bounded (also known as delimited) by curves $y_{i_1}, \ldots, y_{i_k}$ for time (or states) $1, \ldots, k$. Band is defined as $B(y_{i_1}, \ldots, y_{i_k}) = \{(t, x(t)) : t \in \mathcal{I}, \min_{r=1,\ldots,k} y_{i_r}(t) \leq x(t) \leq \max_{r=1,\ldots,k} y_{i_r}(t)\}$. Let J denote the number of curves that determines the band, where J is a fixed integer $2 \leq J \leq n$.

For a stochastic process $Y(t)$ generating the observations $y_1(t), \ldots, y_n(t)$, the population version of the band depth for a curve $y(t)$ with respect to the probability measure P is defined as:

$$BD_J(y, P) = \sum_{j=2}^{J} BD^{(j)}(y, P) = \sum_{j=2}^{J} P\{G(y) \subset B(Y_1, \ldots, Y_j)\}, \tag{6.1}$$

where $B(Y_1, \ldots, Y_j)$ is a band delimited by j random curves [32]. Here, $BD^{(j)}(y, P)$ calculates the probability that the graph of $y(t)$ lies entirely within a band formed by any selection of j functions. These bands are delineated by the minimum and maximum values at each point t among the selected j functions. This formulation of BD facilitates a nuanced understanding of curve centrality and variability within functional data, laying the groundwork for the construction of functional boxplots. Then we can define the sample version of $BD(j)(y, P)$, which is the fraction of the bands determined by j different sample curves containing the whole graph of the curve $y(t)$ [32]:

$$BD_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} I\{G(y) \subseteq B(y_{i_1}, \ldots, y_{i_j})\}, \tag{6.2}$$

where $I\{\cdot\}$ denotes the indicator function. Hence, by computing the fraction of the bands containing the curve $y(t)$, the bigger the value of band depth, the more central position the curve has. Then, the sample band depth of a curve $y(t)$ is [32]

$$BD_{n,J}(y) = \sum_{j=2}^{J} BD_n^{(j)}(y). \tag{6.3}$$

A sample median function is a curve from the sample with largest depth value, defined by [32]

$$\text{argmax}_{y\in y_1,\ldots,y_n} BD_{n,J}(y). \tag{6.4}$$

In a traditional boxplot, the box represents the central 50% of the data. In functional box plot, estimate the central region of data using a band defined by the α proportion (where $0 < \alpha < 1$) of the deepest curves in the sample. Specifically, the central region Funtional boxplot!central regionfor 50% of the sample, denoted as $C_{0.5}$, is calculated using the formula [32]:

$$C_{0.5} = \{(t, y(t)) : \min_{r=1,\ldots,\lceil n/2\rceil} y[r](t) \leq y(t) \leq \max_{r=1,\ldots,\lceil n/2\rceil} y[r](t)\}, \tag{6.5}$$

where $\lceil n/2\rceil$ is the smallest integer not less than half of the sample size. This 50% central region functions similarly to the inter-quartile range (IQR) in a boxplot, providing a measure of the spread of the central 50% of the curves. This method is robust, less affected by outliers or extreme values, and offers a clearer visualisation of the distribution's central spread. Within this central region, the curve that represents the median, denoted as $y_{[1]}(t)$, signifies the most centrally located curve with the highest band depth, serving as a robust measure of central tendency.

6.2.2 Case Example: Visualisation of metabolic syndromes

The functional boxplot shown in Figure 29 illustrates the metabolism rates between healthy and diseased populations. The outer blue line highlights a wider range for the diseased group, accompanied by a broader grey band. The range indicates greater variability and the band indicates lower oxygen saturation for most patients in the blood of the diseased group.

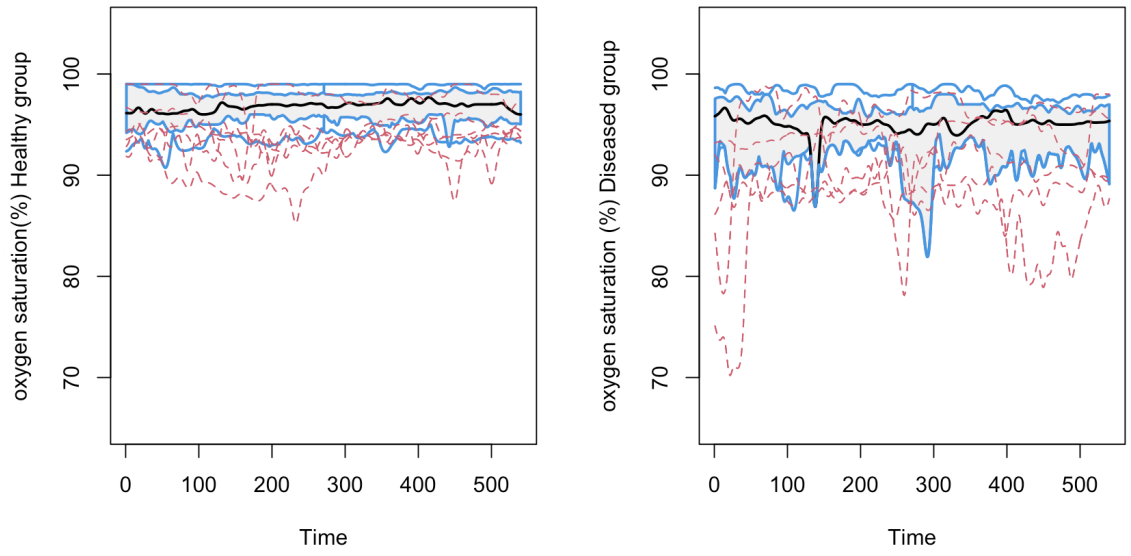

Figure 29: Left graph demostrates curves of arterial oxygen saturation (%) for 80 women without metabolic syndrome and right graph demostrates curves of arterial oxygen saturation (%) for 35 women with metabolic syndrome. Outliers from the population are highlighted in red.

6.3 Q–Q Boxplots

In this section, we will introduce the Q–Q boxplot, which combines the advantages from the boxplot and the Q–Q plot. The Q–Q boxplot integrates the box plot's ability to summarize data through its quartiles and the Q–Q plot's capability to compare data distributions to a theoretical distribution (often the normal distribution).

Although Q–Q plots and box plots are very effiective in data visualisation, they have there limitations as well. Q–Q plot is less effective than box plot in highlighting summary quantitles, and it also suffers the readability when we place multiple samples together. Box plot usually has unreliable tail information when $n < 100$, therefor may not suitable for some research questions that need more accurate estimates of extreme quantiles.

6.3.1 Construction of Q–Q Boxplots

The construction of a Q–Q boxplot involves a three-step process [25]. The first setp is to construct the box, which is similar to the standard boxplot and represents the interquartile range of the data. The second step is to draw the whiskers and confidence bands. The third step is to draw the outside values. The first and third steps are as same as in box plots.

The unique aspect of the Q–Q boxplot comes in the second step. In ordor to draw the whiskers and confidence bands, we have to standardise both the comparison distribution and the reference distribution. This standardisation process involves matching the quantiles from the reference distribution to those in the comparison distribution, either by subsampling or interpolation. Once this is done, the deviations of the comparison distribution from the reference distribution are calculated relative to each data point. Alongside this, the point-wise confidence bands are also determined, which provides a visual indication of how closely the data follows the expected theoretical distribution. To integrate the relative deviations and confidence bands into the Q–Q boxplot, we include both the x and y coordinates in the plot. The y-axis coordinates correspond to the quantile values from the comparison dataset, while the x-axis coordinates represent the deviations relative to the reference distribution.

To calculate the x-coordinate for a Q–Q boxplot, we set a scale for the plot and determin the maximum x-coordinate, $x_{\max} = (\frac{b_w}{2})$ where b_w is the box width. Let dev_{max} be the maximum of the absolute values of the whisker deviations and confidence band values. The value of a relative deviaition d is computed by either taking the difference between the standardised values from the comparison distribution and those from the reference distribution, or by using a value from the confidence band associated with the reference distribution. We can then set:

$$x^* = \frac{d \cdot x_{\max}}{\text{dev}_{\max}}.$$

The x coordinate is then $x_b + x*$, where x_b marks the center of the boxplot.

6.3.2 Case Example: Iris dataset using Q–Q boxplot

The resulting Q–Q boxplot shows the distribution of sepal length values within each iris species, allowing comparison whether they come from a normal distribution. Outliers are also indicated in the plot.

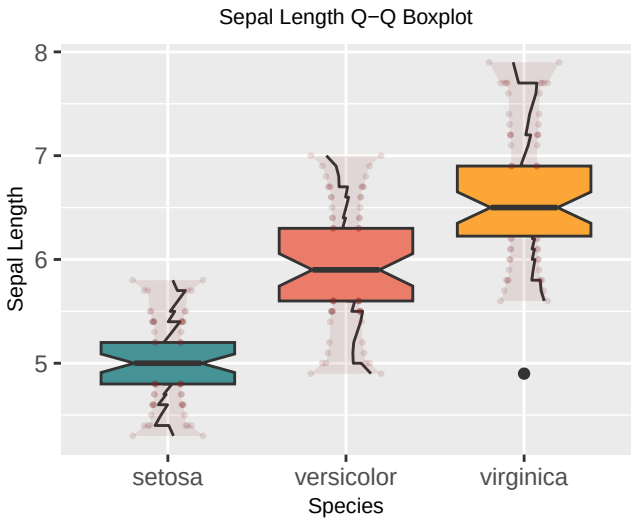

Figure 30: Q–Q boxplot of iris dataset

From the above Figure 30, we can find out easily that the middle of our Q–Q boxplot is same as in the boxplot as we seen in Figure 27. In addition, there are more details in tails behavior simultaneously. We can see that the right-tail deviations and left-tail deviations all lie inside of the confidence bands calculated with the bootstrap method, which means the sample distribution of our dataset is consistent with the theoretical distribution being compared to, that is a normal distribution.

Chapter Overview and Further Reading References

This concluding chapter provides an overview of traditional boxplots and Q-Q plots, followed by an examination of modern data visualisation techniques such as functional boxplots and Q-Q boxplots. Despite being relatively new methods, they have garnered attention in the literature. For further understanding of functional boxplots, readers can refer to Ying Sun and Marc G. Genton's article *Functional Boxplots*. Similarly, additional insights into Q-Q boxplots can be found in Jordan Rodu and Karen Kafadar's article *The Q-Q Boxplot*, both of which are published in the *Journal of Computational and Graphical Statistics*.

Subject Index

References

[1] $500,000 in 1937 is how much today? https://www.calculateme.com/inflation/500000-dollars/from-1937/to-now, n.d. Accessed 9 Nov. 2023.

[2] Kirk Baker. Singular value decomposition tutorial. *The Ohio State University*, 24:22, 2005.

[3] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer, 2016. Accessed 13 Nov. 2023.

[4] K. Dakwa. The kallikak family – historical influences, current controversies, teaching resources. https://intelltheory.com/intelli/the-kallikak-family/, 2001. Accessed 9 Nov. 2023.

[5] Mehdi Dastani. The role of visual perception in data visualization. *Journal of Visual Languages & Computing*, 13(6):601–622, 2002.

[6] M. de Carvalho. Incomplete data analysis. University of Edinburgh, Lecture 2, 9-10, 2023.

[7] Earth Science Data Systems. Firms frequently asked questions — earthdata. https://www.earthdata.nasa.gov/faq/firms-faq#ed-confidence, 2023. Accessed 12 Nov. 2023.

[8] Celso Silva et. al. Deforestation-induced fragmentation increases forest fire occurrence in central brazilian amazonia. *Forest*, 9(6):305, 2018.

[9] Sara Wade et. al. Machine learning in python. http://www.drps.ed.ac.uk/19-20/dpt/cxmath11205.htm, 2024. Accessed 15 Feb. 2024.

[10] William S. Cleveland et. al. *Local regression models. Chapter 8 of Statistical Models in S*. Wadsworth Brooks/Cole, 1992.

[11] Tom Fawcett. Introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 06 2006.

[12] JP García, JC Ferreira, and CM Patino. Receiver operating characteristic analysis: an ally in the pandemic. *J Bras Pneumol*, 47(2):e20210139, Apr 30 2021.

[13] John C Gower and David J Hand. *Biplots*, volume 54. CRC Press, 1995.

[14] Robert Grant. *Data visualization: Charts, maps, and interactive graphics*. Crc Press, 2018.

[15] E. Hawkins. Warming stripes — climate lab book. https://www.climate-lab-book.ac.uk/warming-stripes/, n.d. Accessed 9 Nov. 2023.

[16] Kieran Healy. *Data visualization: a practical introduction*. Princeton University Press, 2018.

[17] Darrell Huff. *How to Lie with Statistics*. W. W. Norton Company, 1954.

[18] M. J. Blanchet, de Carvalho. Applied statistics. Ecole Polytechnique Federale De Lausanne, Lecture 6, 2011.

[19] Ian T Jolliffe. Principal component analysis. *Technometrics*, 45(3):276, 2003.

[20] F. Nightingale. Diagram of the causes of mortality in the army in the east. https://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~327826~90096398:Diagram-of-the-Causes-of-Mortality-;JSESSIONID=802dd785-32fc-4b43-a53d-72ed57e15cc8?qvq=q%3Aauthor%3D%22Nightingale%2C%20Florence%22%3Blc%3ARUMSEY%7E8%7E1&mi=1&trs=10, 1859. Accessed 20 Oct. 2023.

[21] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[22] Margaret Sullivan Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, volume 28. Oxford University Press, United Kingdom, 2003. Print.

[23] A. Proctor. Five charts that changed the world - bbc ideas. `https://www.bbc.co.uk/ideas/videos/five-charts-that-changed-the-world/p0fb69c1`, 2023. Accessed 20 Oct. 2023.

[24] B. D. Ripley. loess: Local polynomial regression fitting. `https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/loess`, 1998. Accessed 20 Jan. 2024.

[25] Jordan Rodu and Karen Kafadar. The q–q boxplot. *Journal of Computational and Graphical Statistics*, 31(1):26–39, 2022.

[26] Muhammad Hafiz Wan Rosli and Andres Cabrera. Gestalt principles in multimodal data representation. *IEEE computer graphics and applications*, 35(2):80–87, 2015.

[27] A. Rutherford. Where science meets fiction: the dark history of eugenics. `https://www.theguardian.com/science/2022/jun/19/where-science-meets-fiction-the-dark-history-of-eugenics`, 2022. Accessed 9 Nov. 2023.

[28] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[29] Cosma Shalizi. Undergraduate advanced data analysis, chp.18.1.2. `https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf`, 2012. Accessed 18 Feb. 2024.

[30] J.D. Smith and M.L. Wehmeyer. Who was deborah kallikak? *Intellectual and Developmental Disabilities*, 50(2):169–178, 2012.

[31] Josh Starmer. Lowess and loess, clearly explained!!! `https://www.youtube.com/watch?v=Vf7oJ6z2LCc`, 2018. Accessed 3 Jan. 2024.

[32] Ying Sun and Marc G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011.

[33] Dejan Todorovic. Gestalt principles. *Scholarpedia*, 3(12):5345, 2008.

[34] Edward R Tufte. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 2001.

[35] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[36] Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994.

[37] Matthew O Ward, Georges Grinstein, and Daniel Keim. *Interactive data visualization: foundations, techniques, and applications*. CRC press, 2010.

[38] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[39] Leland Wilkinson. *The grammar of graphics*. Springer, 2012.