

1 Chapter 2. Theoretical Foundations of Data Visualisation

In the realm of data science, the power of data visualisation is undeniable. The ability to transform raw data into compelling visual representations not only enables us to see patterns and trends but also serves as a bridge for effective communication of empirical findings. To harness this potential to its fullest, it is crucial to lay down a solid theoretical foundation that underpins the art and science of data visualisation.

This chapter, "Theoretical Foundations of Data Visualisation," delves deep into the core principles and concepts that serve as the bedrock of this dynamic field. We seek to understand not only the "how" but also the "why" behind the creation of visualisations that captivate and inform.

2 Chapter 2.1. Introduction to Data Visualisation Theory

In the pursuit of creating effective data visualisations, it is crucial to understand that behind every compelling chart, graph, or map lies a solid theoretical framework. Theoretical underpinnings provide the foundation upon which data visualisation is built, shaping not only how we represent data but also how we perceive, understand, and interpret it.

2.1 2.1.1 Guiding Principles for Data Representation

Within this theoretical framework, we encounter a set of guiding principles that dictate how data should be represented visually. These principles encompass fundamental concepts such as:

- **Accuracy:** Data visualisations should accurately reflect the underlying data, minimising distortion or misinterpretation.
- **Simplicity:** The "less is more" principle applies to data visualisation. Simplified visuals often convey information more effectively than cluttered ones.
- **Clarity:** Visualisations should be clear and understandable to the intended audience, avoiding unnecessary complexity.
- **Relevance:** Information presented should be relevant to the message or question being addressed.
- **Consistency:** Visual elements, such as colour coding and labelling, should be used consistently throughout a visualisation.

2.2 2.1.2 Theoretical Framework and Visual Perception

One of the fundamental aspects of data visualisation theory is an understanding of how the human brain perceives visual information. This knowledge is instrumental in designing visualisations that resonate with viewers. It includes considerations like:

- **Gestalt Principles:** The Gestalt principles of visual perception, including proximity, similarity, and continuity, influence how we group and interpret visual elements in a visualisation.
- **Colour Theory:** The effective use of colour, including colour contrasts and harmonies, can enhance the clarity and impact of a visualisation.
- **Cognitive Load:** Minimising the mental effort required to process information is vital.

3 Chapter 2.2 Visual Perception and Cognition

Understanding the intricacies of how humans interpret visual information is pivotal to the art and science of data visualisation. Thus, we explore human visual perception, along with the application of cognitive psychology principles in data visualisation and highlight the crucial role of pre-attentive attributes in shaping our perception of data.

3.1 2.2.1 Human Visual Perception: Decoding Visual Information

Human visual perception is a remarkable cognitive process that allows us to decode and make sense of the world around us. When applied to data visualisation, it illuminates how viewers interact with and derive meaning from visual representations of data. Key aspects of human visual perception in the context of data visualisation include:

- **Pattern Recognition:** The human brain excels at recognizing patterns, making it adept at identifying trends, outliers, and relationships in data visualisations.
- **Perceptual Grouping:** We tend to group visually similar elements together, a principle known as perceptual grouping. This informs how we interpret clusters of data points and shapes in a visualisation.
- **Hierarchy of Perception:** Certain visual attributes are processed more quickly and efficiently than others. For example, colour is often processed faster than text, influencing the viewer's attention hierarchy.

By harnessing the principles of human visual perception, applying insights from cognitive psychology, and leveraging pre-attentive attributes, data visualisation designers can create visualisations that are not only aesthetically pleasing but also cognitively efficient.

4 Chapter 2.3 Gestalt Principles

Gestalt psychology principles have long been recognised as fundamental to the field of visual perception and design. In this section, we provide an overview of Gestalt principles, delve into their relevance in the context of data visualisation, and explore how these principles can be applied to create more effective and intuitive visualisations.

4.1 2.3.1 Overview of Gestalt Psychology Principles

Gestalt psychology is a school of thought that emphasises how humans perceive and organise visual information. It posits that the mind seeks to create meaningful patterns and wholes from individual visual elements. The Gestalt principles provide a framework for understanding how viewers naturally group and interpret visual stimuli.

4.2 2.3.2 Key Gestalt Principles and Their Relevance

Several key Gestalt principles play a pivotal role in shaping our perception of visual information. These principles include:

- **Proximity:** Elements that are close to each other are perceived as related or belonging to the same group. In data visualisation, proximity can be used to group data points or related information.
- **Similarity:** Elements that share similar visual attributes, such as colour, shape, or size, are perceived as belonging together. Similarity can be harnessed to encode categorical data or highlight relationships.
- **Continuity:** The human mind tends to perceive continuous patterns or lines as a single entity. In data visualisation, continuity can aid in representing trends or connections between data points.
- **Closure:** Viewers tend to mentally complete incomplete shapes or patterns. Closure can be employed to suggest relationships or connections even when not explicitly shown in the visualisation.
- **Symmetry:** Symmetrical elements are often perceived as more balanced and harmonious. Symmetry can be used for aesthetically pleasing and easily understandable visualisations.

4.3 2.3.3 Application of Gestalt Principles in Designing Visualisations

The application of Gestalt principles in designing data visualisations can lead to more intuitive and effective communication of information. Designers can strategically leverage these principles to:

- Group-related data points to enhance clarity and reduce visual clutter.
- Use colour or shape to signify meaningful categories or groupings.
- Create smooth, continuous lines or paths to guide the viewer's gaze through the visualisation.
- Suggest connections or patterns even in complex datasets.

5 Chapter 2.4 Data abstraction and Representation

The transformation of raw data into meaningful representations is a pivotal step in data visualisation. This process, known as data abstraction, involves distilling complex datasets into visual forms that convey insights. In this section, we explore data abstraction, the hierarchies and levels of abstraction in data visualisation, and the critical trade-offs between abstraction and the potential loss of information.

5.1 2.4.1 Data Abstraction: Transforming Raw Data

Data abstraction is the art of simplifying and structuring raw data into formats that are comprehensible and insightful. It is the bridge that transforms numbers, text, and variables into visual elements that convey patterns, trends, and relationships. Effective data abstraction is at the heart of creating informative data visualisations.

5.2 2.4.2 Hierarchies and Levels of Abstraction

In data visualisation, abstraction operates on multiple levels of granularity. Hierarchies of abstraction allow us to represent 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:** As we move up the hierarchy, 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. This level provides a big-picture view.

5.3 2.4.3 Trade-offs Between Abstraction and Information Loss

While abstraction is essential for simplifying complex data, it comes with trade-offs. Increasing levels of abstraction can lead to information loss, where fine-grained details or outliers are overlooked. It is crucial for data visualisation designers to strike a balance:

- **Clarity vs. Detail:** Increasing abstraction can enhance the clarity of a visualisation but may sacrifice detailed information that is important for certain analytical tasks.
- **Generalisation vs. Specificity:** Abstraction can provide a more generalised view of data, making it accessible to a wider audience. However, it may miss specific nuances that experts may require.
- **Context vs. Precision:** High-level abstraction can provide valuable context, but it may lack the precision needed for precise decision-making.

In data visualisation, the art of data abstraction lies in finding the right level of detail that effectively conveys the intended message while minimising the risk of information loss. This balancing act is a critical consideration in the design of informative and meaningful data visualisations.

6 Chapter 2.5 Data Types and Visualisation Techniques

In the world of data visualisation, understanding the nature of your data is key. Data comes in various types, and selecting the appropriate visualisation technique is contingent upon recognising these distinctions. In this section, we categorise data types, and demonstrate how to match each data type with suitable visualisation techniques.

6.1 2.5.1 Categorisation of Data Types

Data types can be broadly categorised into four main types:

- **Nominal data:** nominal data represents categories or labels without any inherent order. Examples include colours, gender categories, and city names.

- **Ordinal data:** 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:** interval data possesses ordered categories with equal intervals between them, but it lacks a true zero point. Temperature is measured in Celsius or Fahrenheit as an example.
- **Ratio data:** ratio data includes ordered categories with equal intervals and a meaningful zero point. Examples are age, income, and weight.

6.2 2.5.2 Matching Data Types with Appropriate Visualisation Techniques

Selecting the right visualisation technique for your data type is pivotal to effective communication. Here are some examples of visualisation techniques matches with corresponding data types:

- **Nominal data:** Techniques such as bar charts and stacked bar charts are effective in displaying categorical information and relative proportions.
- **Ordinal data:** Ordinal data can be visualised using ordered bar charts, dot plots, or stacked bar charts, which maintain the ranking and order of the categories.
- **Interval data:** Interval data benefits from visualisation methods like line charts, histograms, and box plots, which highlight trends and distributions without assuming a true zero point.
- **Ratio data:** Ratio data can be effectively presented using scatter plots, histograms, and line charts, allowing for precise comparisons and measurements due to the presence of a meaningful zero point.

7 Chapter 2.6 Colour Theory in Data Visualisation

Here, we explore 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.

7.1 2.6.1 The Importance of Colour in Conveying Information

Colour is a potent tool for enhancing the understanding and impact of data visualisations. It enables the differentiation of data points, highlights trends, and provides context. Colour can be used to:

- **Encode Categorical Data:** Distinguish between different groups or classes using distinct colours.
- **Represent Quantitative Data:** Use colour intensity or gradients to represent values or magnitudes.
- **Add Context:** Apply colour to background elements, labels, or annotations to provide context and meaning to the visualisation.

7.2 2.6.2 Colour Perception and Colour Encoding in Visualisations

Understanding how colour is perceived by viewers is essential in data visualisation. Key principles include:

- **Colour Discrimination:** Consider that not all viewers may perceive colour in the same way. Ensure your colour choices are accessible to individuals with colour vision deficiencies (colour blindness).
- **Colour Encoding:** Select colour schemes that align with the message you want to convey. For example, warm colours like red and orange often signify caution or heat, while cool colours like blue and green convey calmness or coldness.
- **Colour Combinations:** Pay attention to how colours interact when placed adjacent to each other. Some combinations may create visual vibrations or make text difficult to read.

7.3 2.6.3 Avoiding Misleading Visualisations Due to Colour Choices

Misleading visualisations can result from inappropriate or deceptive use of colour. To avoid this:

- **Consistency:** Maintain consistency in colour usage throughout your visualisation. Use the same colour scheme for similar data categories or elements.
- **Avoid Distortion:** Ensure that colour choices do not exaggerate or distort the data. Overly intense or contrasting colours can lead to misinterpretation.
- **Legend Clarity:** Provide a clear and concise legend to explain the meaning of colours, especially when dealing with complex or unfamiliar colour schemes.
- **Test with Users:** Conduct user testing to ensure that your colour choices effectively convey the intended message and do not confuse or mislead the audience.

In summary, colour theory is a critical component of data visualisation. Thoughtful colour choices can enhance the clarity and impact of visualisations, making data more accessible and understandable. However, it is equally important to use colour judiciously to avoid misleading visualisations and ensure that the message remains accurate and informative.

8 Chapter 2.7 Theoretical Properties of Visualisations

Effective data visualisation is not solely about creating aesthetically pleasing graphics; it's also about adhering to key theoretical properties that optimise the expressiveness, precision, accuracy, and scalability of visual representations. In this section, we delve into these properties, including expressiveness and effectiveness, the data-ink ratio, and the principles of minimal ink, as well as precision, accuracy, and scalability.

8.1 2.7.1 Expressiveness and Effectiveness

- **Expressiveness:** Visualisations should be expressive, meaning they should effectively communicate the intended message or insights within the data. Expressive visualisations capture the richness and complexity of the underlying data, revealing patterns, trends, and relationships.
- **Effectiveness:** An effective visualisation is one that successfully conveys information to its audience. It allows viewers to understand the data, draw meaningful conclusions, and make informed decisions based on the presented information.

8.2 2.7.2 Data-Ink Ratio and the Principle of Minimal Ink

REVISE This **Data-Ink Ratio principle**, introduced by Edward Tufte REFERENCEEEEEEEEEEE, emphasises maximising the ink (or pixels in digital formats) used to represent the actual data while minimising non-essential ink. A higher data-ink ratio results in a cleaner, more efficient visualisation that reduces clutter and enhances comprehension.

The **Principle of Minimal Ink** REFERENCEEEEEEEEEEE builds on the data-ink ratio. This principle advocates for the removal of any visual elements that do not contribute to the viewer's understanding of the data. Eliminating unnecessary ink (e.g., excessive gridlines or decorations) simplifies the visualisation without sacrificing its effectiveness.

8.3 2.7.3 Precision, Accuracy, and Scalability

- **Precision:** Precision in data visualisation refers to the level of detail and granularity in the representation of data. Visualisations should strike a

balance between showing enough detail to support accurate interpretation while avoiding overwhelming viewers with excessive complexity.

- **Accuracy:** Accuracy pertains to how faithfully the visualisation represents the true values in the data. Misleading or distorted visualisations can lead to incorrect conclusions. Therefore, maintaining accuracy is essential.
- **Scalability:** Scalability addresses how well a visualisation adapts to different data sizes or resolutions. Effective visualisations should be scalable, and capable of representing both small and large datasets without sacrificing clarity or performance.

9 Chapter 2.8 Cognitive Load and Visual Complexity

In data visualisation, achieving a balance between complexity and cognitive load is crucial. 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.

9.1 2.8.1 Exploring the Concept of Cognitive Load in Visualisations

In data visualisations, cognitive load plays a significant role in how viewers engage with and understand the presented data. It is essential to strike a balance that ensures the visualisation conveys information effectively without overwhelming or overtaxing the viewer's cognitive capacity.

9.2 2.8.2 Strategies to Reduce Cognitive Load While Maintaining Complexity

- **Visual Hierarchy:** Establish a clear visual hierarchy that guides viewers' attention to the most important elements of the visualisation. Use techniques such as size, colour, and contrast to emphasise key information.
- **Simplify Labels and Text:** Reduce cognitive load by using concise labels and text. Avoid jargon and unnecessary complexity in annotations, ensuring that labels are informative and straightforward.
- **Interactive Features:** Implement interactive elements, such as tooltips and drill-down functionality, to provide additional information when viewers need it, reducing the need for dense, static visualisations.
- **Progressive Disclosure:** Present complex information gradually, allowing viewers to digest it in stages. Start with an overview and provide opportunities for users to explore details as needed.

- **Data Aggregation:** Consider aggregating data when it makes sense. Summarising data can reduce the cognitive load associated with interpreting fine-grained details.

9.3 2.8.3 Information Overload and Simplification Techniques

Information overload occurs when a visualisation overwhelms viewers with excessive data or visual elements, hindering comprehension. To combat information overload, the following simplification techniques can be applied:

- **Filtering:** Allow viewers to filter data based on specific criteria, enabling them to focus on the most relevant information.
- **Data Reduction:** Aggregate or summarise data to present overarching trends or patterns instead of inundating viewers with raw data points.
- **Storyboarding:** Use storytelling techniques to guide viewers through the data in a structured manner, helping them understand the context and relevance of the information presented.
- **Prioritisation:** Identify the most critical information and prioritise its display, relegating less essential data to secondary views or interactions.

10 Chapter 3. Modern Methods of Data Visualisation

11 Chapter 3.1 Introduction to Modern Data Visualization Methods

xxx

12 Chapter 3.2 Scatter Plots and Bubble Charts

Scatter plots and bubble charts are fundamental data visualization techniques that provide valuable insights into the relationships and patterns within datasets. These visualizations are particularly effective for representing data points, making comparisons, and revealing trends. Below, we delve into the characteristics, use cases, and advantages of scatter plots and bubble charts.

12.1 Scatter Plots

A scatter plot is a simple yet powerful visualization that displays individual data points as dots on a two-dimensional plane. Each point represents a unique data entry, with one variable on the horizontal (x-axis) and another on the vertical

(y-axis). Scatter plots are versatile and can accommodate various data types, including nominal, ordinal, interval, and ratio.

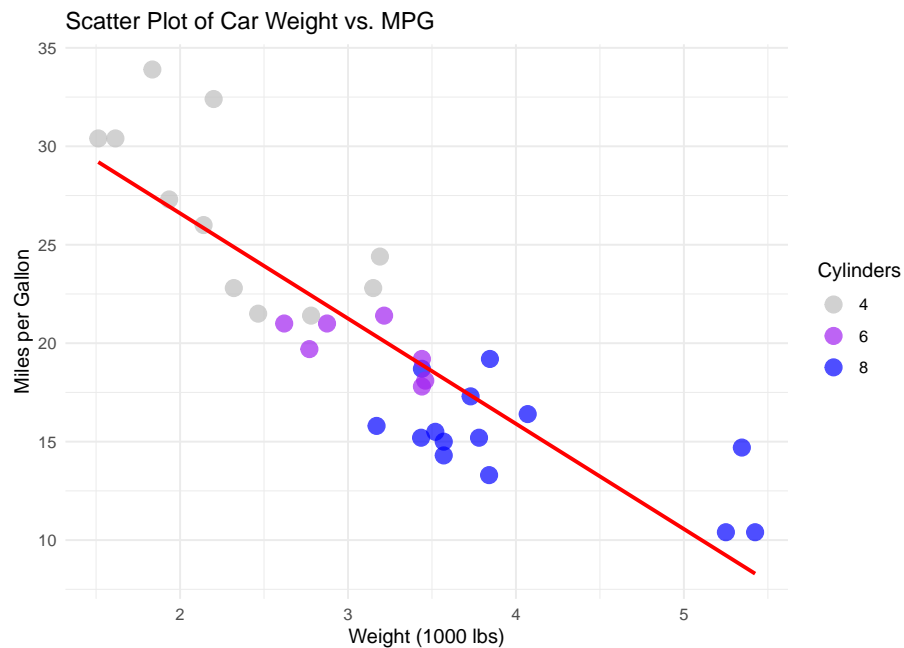
12.1.1 Advantages

- Easy identification of patterns and trends.
- Effective for outlier detection.
- Suitable for both bi-variate and multi-variate analysis.
- Offers a visual representation of the distribution of data points.

12.2 Scatter Plots in Practice

In this example, we'll create a scatter plot that visualizes the relationship between two variables while adding colors, labels, and a trend line for added interest. We'll use the built-in "mtcars" R dataset, which contains information about various car models.

```
## 'geom_smooth()' using formula = 'y ~ x'
```



12.2.1 Analysis

TO BE REVISED The scatter plot displays the relationship between car weight (in thousands of pounds, "wt") and fuel efficiency (miles per gallon, "mpg") for different car models from the "mtcars" dataset. Here are the key characteristics of the plot:

- **Color-Coded Cylinders:** The points are color-coded based on the number of cylinders in the engine (4, 6, and 8 cylinders). This allows for a quick visual differentiation of car types, enhancing the understanding of the data.
- **Point Size and Transparency:** Points are larger in size and have some degree of transparency (alpha) for visual appeal. Larger, more prominent points are easier to see, while transparency helps in the visualization of overlapping points.
- **Linear Regression Line:** A blue linear regression trend line is included. It provides a visual representation of the overall relationship between car weight and fuel efficiency, indicating a negative correlation—cars tend to have lower fuel efficiency as their weight increases.

12.3 Why a Scatter Plot is Appropriate

A scatter plot is an appropriate choice for this dataset and analysis for the following reasons:

1. **Relationship Visualization:** A scatter plot is well-suited for showing the relationship between two continuous variables, such as car weight and fuel efficiency. It allows us to visually assess the pattern and direction of the relationship.
2. **Regression Line:** The inclusion of a regression line provides a clear summary of the overall trend, highlighting the negative correlation between car weight and fuel efficiency. EXPLAIN REGRESSION

12.4 Bubble Charts

A bubble chart is an extension of the scatter plot, where each data point is represented as a bubble or circle. In addition to the x and y axes, bubble charts introduce a third variable that is encoded by the size of the bubbles. This third variable allows for the simultaneous visualization of three data attributes.

12.5 Use Cases

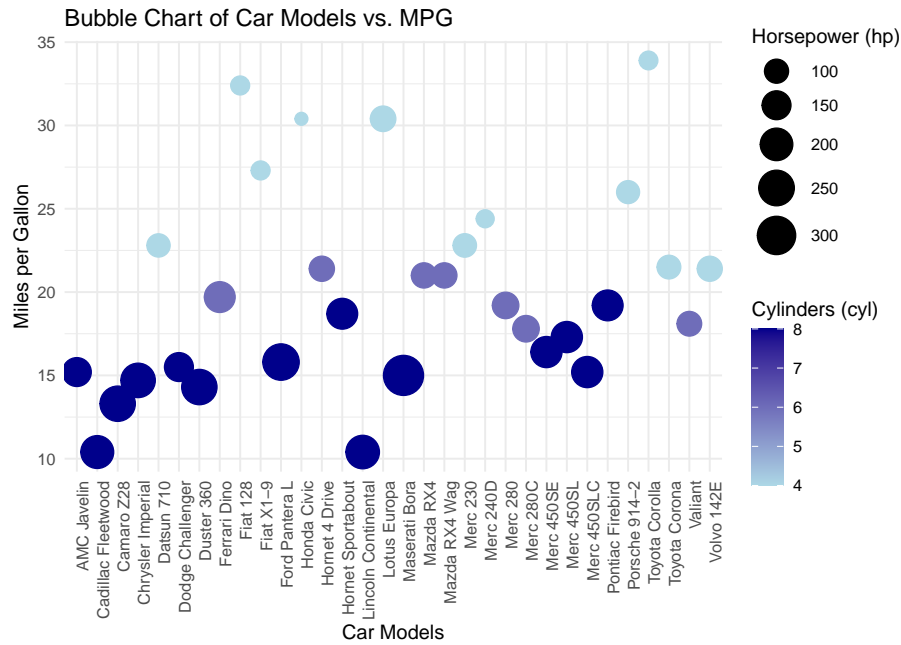
Bubble charts are particularly useful when you need to visualize the relationship between three variables. For example, in finance, a bubble chart can display the market capitalization, price-to-earnings ratio, and stock returns for different

companies simultaneously. In public health, it can represent the population, vaccination coverage, and disease incidence for various regions.

12.5.1 Advantages

- Visualising three data attributes in a single chart.
- Easily conveying the magnitude of the third variable through bubble size.
- Effective for exploring complex relationships and identifying patterns.

12.6 Bubble Charts in Practice



12.7 Key Considerations for Scatter Plots and Bubble Charts

1. **Scaling:** Consider the scaling of the axes and bubble sizes to prevent distortion of the data representation.
2. **Labeling:** Use clear labels, titles, and colour coding to make the visualisation more interpretable and meaning by decreasing the cognitive load.
3. **Interactivity:** Implement interactive features, such as tooltips or zooming, to allow users to explore individual data points in detail.

Chapter 3.3 Bar Charts and Histograms

Bar Charts

A bar chart is a very important method to present data. It organizes information into vertical bars. Bar charts have lots of advantages in data visualization. It can present data categories in a frequency distribution. A bar chart is best for comparing classified data. Especially when the values are close, because the human perception of height is better than other visual elements (such as area, angle, etc.), the use of a bar chart is more appropriate. These bars usually have different lengths, and every length is proportional to the size of the information they present.

R uses the function *barplot()* to create bar charts. R can draw both vertical and Horizontal bars in the bar chart. In the bar chart, each of the bars can be given different colors.

R is a programming language for data analysis and statistical computing, and its advent has made data visualization more straightforward and accessible. Among the various tools available in R, ggplot2 stands out as one of the most renowned and powerful tools for creating data visualizations. It offers a wealth of data visualization capabilities and is celebrated for its versatility and aesthetic appeal. In this chapter, we will focus on how to use ggplot2 to create bar charts for data visualization.

3.3.1 Different Types of Bar Charts

Here is an overview of the different types of bar charts.

Vertical Bar Chart This is the most common bar chart. We use different vertical columns to display and compare the values of different categories in the same dimension, where the X-axis represents the contrasting categories and the Y-axis represents the frequency or count of their categories.

Horizontal Bar Chart This is very similar to a vertical bar chart but rotated 90 degrees. Categories are shown on the y-axis and frequency or count are shown on the x-axis. Horizontal bar charts are especially useful when category names are long or when there are numerous categories.

Multi-set Bar Chart Also known as a grouped bar chart or clustered bar chart. A multi-set bar chart is used to represent and compare different sub-groups within individual categories. This type of chart is useful when you want to show and compare multiple sets of data side-by-side. Multi-set Bar charts can be horizontal or vertical like the other normal bar charts, and the length of each bar represents the frequency or count of their categories.

Stacked bar chart Similar to bar charts, stacked bar charts are often used to compare different classes of values and, within each class of values, are divided into sub-classes, which are often referred to by different colors. Each segment's size is proportional to the frequency or count that it represents from the sub-category. The entire bar's length represents the cumulative total of all the sub-categories. However, it is very easy to get confused when there are too many categories.

3.3.2 Advantages of Bar Charts

1. **Clarity and Simplicity:** Bar charts are structurally simple, making them easy to read and understand, allowing audiences to quickly grasp key information.
2. **Effective Comparison:** They provide a visual representation that makes comparing the size or value of different categories straightforward, especially when comparing a limited number of categorical data.
3. **High Flexibility:** They can be used to represent any type of data, be it continuous or discrete.
4. **Multilevel Representation:** Stacked or grouped bar charts can be used to represent multiple data series.

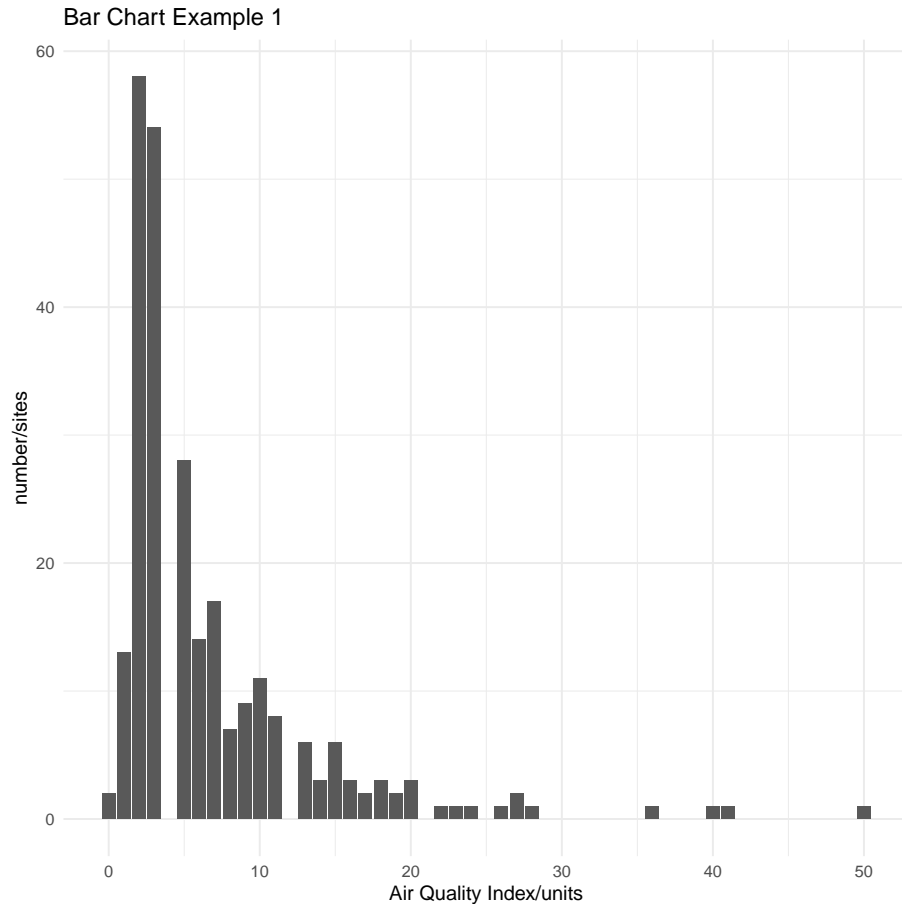
3.3.3 Disadvantages of Bar Charts

1. **Limited Data Representation:** They might not be suitable for representing large datasets as things can get cluttered.
2. **Potential Misinterpretation:** Without a zero baseline, bar charts can be misleading as they might exaggerate differences.
3. **Overcomplexity with Many Categories:** If there are too many bars, it can be challenging to discern information effectively.
4. **Requires Categorical Data:** Bar charts are not ideal for representing trends over continuous data, where line graphs might be more appropriate.

In this section, I will analyze the air quality dataset provided by the United States Environmental Protection Agency. In our dataset, we have data from over 200 locations. The Air Quality Index (AQI) ranges from 0 to 500. A higher AQI indicates increased levels of air pollution, leading to heightened health concerns. This implies that as the AQI rises, there is a greater risk to public health.

```
data <- read.csv("c4_epa_air_quality.csv")
ggplot(data, aes(x = aqi)) +
  geom_bar() +
```

```
labs(title = "Bar Chart Example 1", x =
      "Air Quality Index/units", y = "number/sites") +
theme_minimal()
```



The above code is not optimal. Upon examination, we can see that there is an excessive number of different categories on the x-axis. Consequently, the multitude of vertical bars in the graph can potentially overwhelm and confuse readers. In order to solve the problem, we can use the “cut” function to divide the data into intervals of five units each. For instance, values from 0-5 would constitute one group, 6-10 another, 11-15 would form the next group, and so forth.

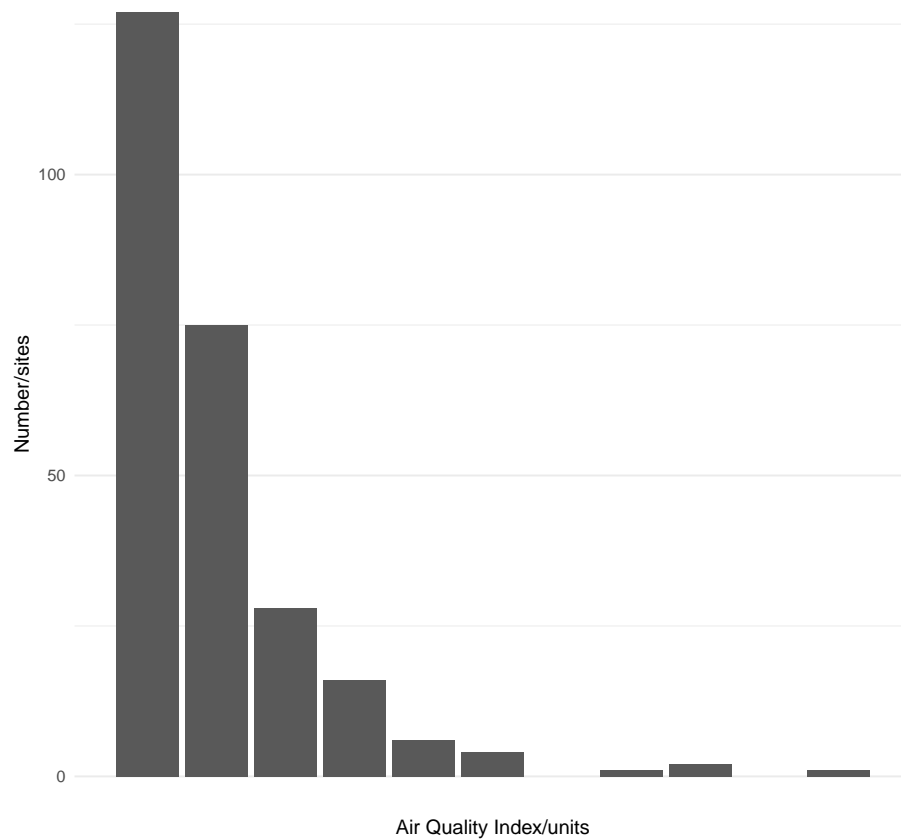
```
data <- read.csv("c4_epa_air_quality.csv") #
# load ggplot2 package
library(ggplot2)
```



```
# Use the cut function to divide the data into groups of five intervals
data$group <- cut(data$aqi, breaks =
                  seq(0, max(data$aqi) + 5, by = 5),
                  right = FALSE, include.lowest = TRUE)
data$group <- as.numeric(data$group)

# Y-axis representing the number of occurrences of the X-axis label in the data
ggplot(data, aes(x = group)) +
  geom_bar() +
  labs(title = "Bar Chart Example 1",
       x = "Air Quality Index/units", y = "Number/sites") +
  theme_minimal() +
  scale_x_discrete(labels = scales::label_number(accuracy = 5))
```

Bar Chart Example 1



We can also add some color to make our plot more attractive, here we can add some color as well. In the code below, we set the color of the bar chart to

blue while specifying the border color as black.

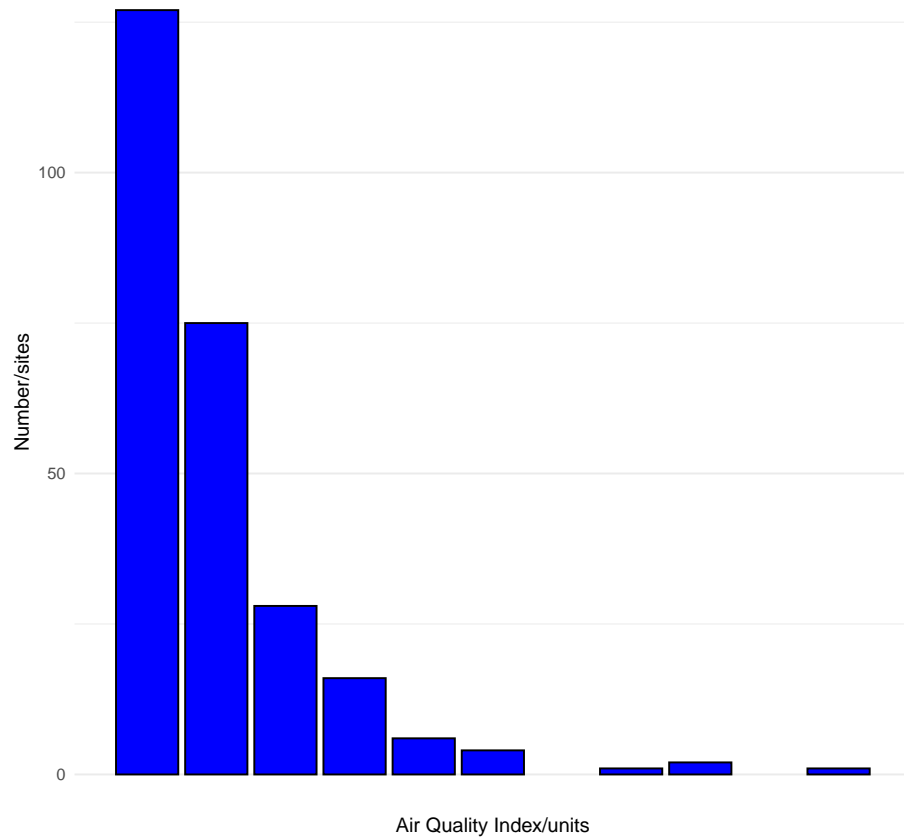
```
# read CSV data set
data <- read.csv("c4_epa_air_quality.csv")

# load ggplot2
library(ggplot2)

# Use the cut function to divide the data into groups of five intervals
data$group <- cut(data$aqi, breaks = seq(0, max(data$aqi) + 5, by = 5),
                  right = FALSE, include.lowest = TRUE)
data$group <- as.numeric(data$group)
# Converts the group column to numeric type

ggplot(data, aes(x = group)) +
  geom_bar(color="black",fill="blue") +
  labs(title = "Bar Chart Example 1", x = "Air Quality Index/units",
       y = "Number/sites") +
  theme_minimal() +
  scale_x_discrete(labels = scales::label_number(accuracy = 5))
```

Bar Chart Example 1



Finally, we add labels to the X-axis to define the range of each categories.

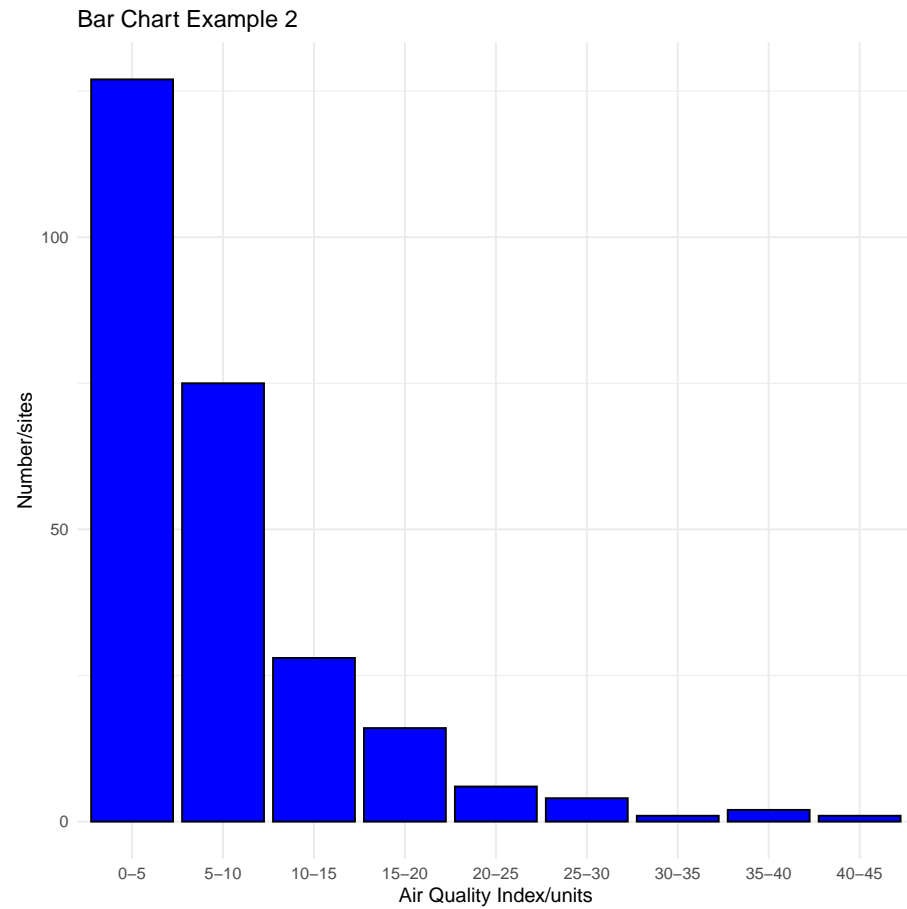
```
# read csv
data <- read.csv("c4_epa_air_quality.csv")

# load ggplot2
library(ggplot2)

# Use the cut function to divide the data into groups of five intervals
breaks_list <- seq(0, max(data$aqi) + 5, by = 5)
data$group <- cut(data$aqi, breaks = breaks_list,
                  right = FALSE, include.lowest = TRUE)

ggplot(data, aes(x = group)) +
  geom_bar(color="black", fill="blue") +
```

```
labs(title = "Bar Chart Example 2", x = "Air Quality Index/units",
     y = "Number/sites") +
theme_minimal() +
scale_x_discrete(labels = paste0
                  (breaks_list[-length(breaks_list)], "-", breaks_list[-1]))
```



13 Chapter 3.4 Heatmaps and Tree Maps

```
# Obtain the Brazil map data
brazil_map <- map_data("world", region = "Brazil")

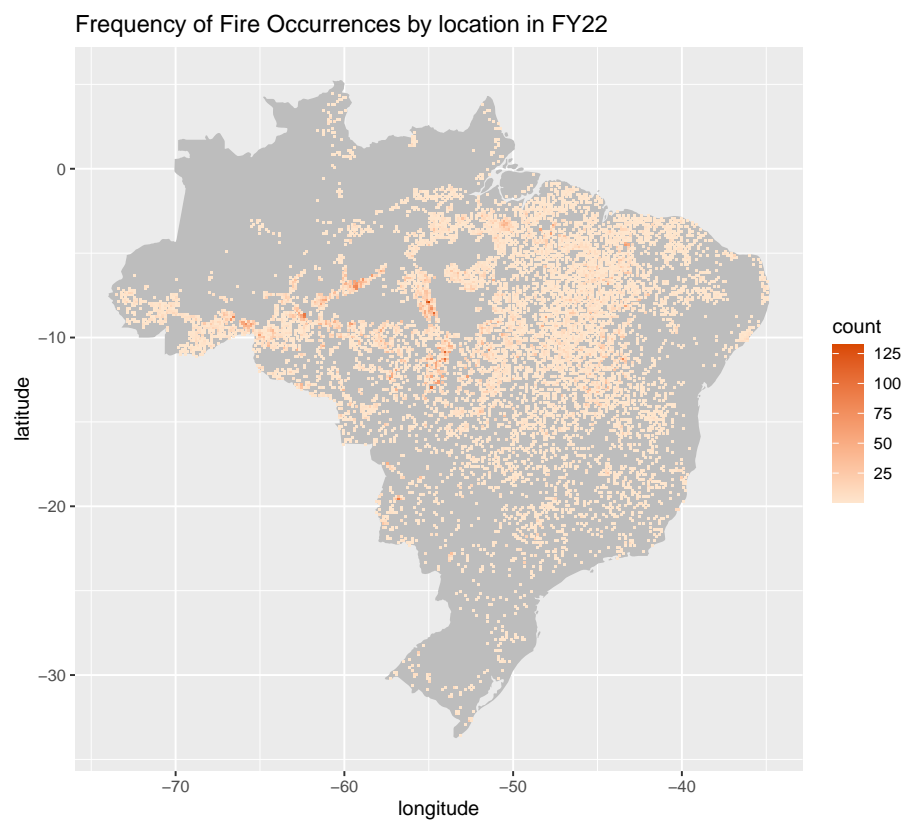
# Create the heatmap of fire occurrences
fire_heatmap <- ggplot(confident_fire_fy22, aes(x = longitude, y = latitude)) +
```

```

geom_polygon(data = brazil_map, aes(x = long, y = lat, group = group),
             fill = "#bdbdbd") +
geom_bin2d(bins = 300) +
scale_fill_gradient(low = "#fee6ce", high = "#d94801") +
coord_fixed(ratio = 1) +
labs(title = "Frequency of Fire Occurrences by location in FY22")

print(fire_heatmap)

```



Colour selection from: <https://colorbrewer2.org/#type=sequential&scheme=Oranges&n=9>

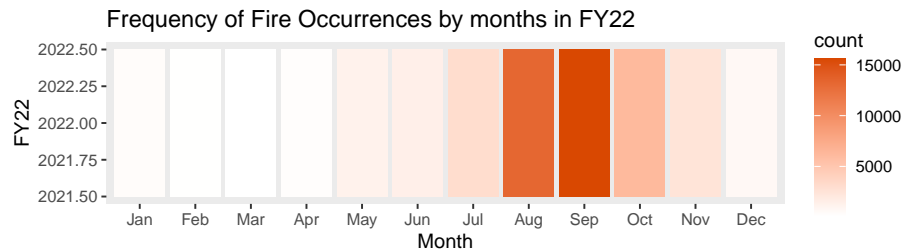
```

heatmap_plot <- ggplot(confident_fire_months_fy22,
                       aes(x = abb_month, y = 2022, fill = count)) +
  geom_tile(width = 0.9, height = 1) + # Create the heatmap tiles

```

```
scale_fill_gradient(low = "white", high = "#d94801") + # Customize the color scale
labs(title = "Frequency of Fire Occurrences by months in FY22", x = "Month", y = "FY22") +
theme(panel.grid = element_blank() )

# Display the heatmap
print(heatmap_plot)
```



From the table, we can clearly see that August and September are the riskiest months in terms of fire hazard, whereas November to July hardly pose any risk at all. It's natural to ask the follow-up question: How does FY22 compare to previous years? Is it valid to claim that August and September are the fire hazard season?

Chapter 3.5 Line Charts and Time Series Visualization

A **Line chart**, often referred to as a line graph or line plot, is a statistical chart composed of a Cartesian coordinate system, some points, and lines. It is commonly used to represent changes in numerical values over continuous time intervals or ordered categories. In a line graph, the x-axis is typically used for continuous time intervals or ordered categories (such as Stage 1, Stage 2, Stage 3). The y-axis is used for quantified data, and if it is negative, it is plotted below the y-axis. Lines are used to connect adjacent data points.

Line graphs are used to analyze trends in things that change over time or ordered categories. If there are multiple sets of data, they are used to analyze the interaction and impact of these data sets over time or ordered categories. The direction of the line represents positive/negative changes, and the slope of the line indicates the degree of change.

In terms of data, a line graph requires a continuous time field or a categorical field and at least one continuous data field.

13.1 Basic Components

- **X-Axis (Horizontal Axis):** Typically represents the independent variable, such as time or date.

- **Y-Axis (Vertical Axis):** Typically represents the dependent variable, like sales numbers, stock prices, or temperatures.
- **Line:** Connects the individual data points. In some line charts, multiple lines can represent different categories or sets of data.

13.2 Suitability for Displaying Trends Over Time:

- **Visual Clarity:** Line charts provide a clear and concise way to view changes over time. When data points are plotted over regular intervals (e.g., days, months, years), it becomes easy to see upward or downward trends.
- **Comparisons:** When you have multiple lines on a single chart, you can easily compare different sets of data. For instance, comparing sales data of two different products over time.
- **Identification of Patterns:** Line charts help in identifying patterns and anomalies. Seasonal patterns, cyclical events, and unexpected spikes or dips become evident.
- **Forecasting:** By viewing historical data trends on a line chart, analysts can make predictions or forecasts for future data points.
- **Simplicity:** They are easy to understand and interpret. Even if someone isn't data-savvy, they can grasp the general trend and major fluctuations from a line chart.
- **Flexibility:** They can be used for both short-term and long-term data. Whether you're looking at stock prices minute-by-minute over a single day or global temperature averages over a century, line charts can effectively represent the data.

13.3 Limitations:

While line charts are excellent for displaying trends over time, they have limitations. They may not be suitable for showing individual data distributions or for data where there's no logical order. eg. too many points, too many lines, too many zeros.

Discuss the importance of time series visualisation in data analysis.

Time series visualization refers to the graphical representation of time-ordered data points. In the world of data analysis, this form of visualization is invaluable for examining patterns, anomalies, and trends in datasets that evolve over time.

Uncovering Trends:

One of the primary advantages of time series visualization is the ease with which it allows analysts to identify long-term upward or downward trends in data. Recognizing these trends can help organizations make informed decisions about future strategies or interventions.

Detection of Seasonality:

Many datasets exhibit patterns that repeat over specific intervals, such as days, months, or years. Time series visualization makes it straightforward to spot such cyclical behaviors, which can be vital for businesses in sectors like retail or agriculture.

Identifying Anomalies:

Graphical representations can quickly highlight data points or periods that deviate significantly from the norm. These anomalies can indicate errors in data collection, or they may reveal significant events that need to be further investigated.

Forecasting and Predictions:

After identifying patterns in historical data, time series visualizations can aid in modeling future data points. Predictive modeling, underpinned by clear visualizations, allows businesses to make proactive decisions.

Facilitating Comparative Analysis:

Time series charts often allow for overlaying multiple data series on a single graph. This capability is useful for comparing different datasets or the same dataset under different conditions, leading to more comprehensive insights.

Conclusion:

Time series visualization is an indispensable tool in the arsenal of data analysts. It condenses large volumes of chronological data into easily interpretable graphics, enabling quick insights, better decision-making, and a deeper understanding of temporal dynamics in datasets. By providing a clear view of data trends, seasonality, and anomalies, time series visualization facilitates more informed and strategic actions in various domains.

Provide best practices for creating clear and informative line charts.

- **Title and Labels:** Every chart should have a descriptive title and axis labels to clearly convey the purpose of the visualization and the data being shown.
- **Use of Colors:** Colors should be chosen to clearly differentiate between different lines or data points but also be consistent with the overall theme or style.
- **Gridlines and Background:** Soft gridlines can help the viewer estimate values. A clean background aids in clarity.

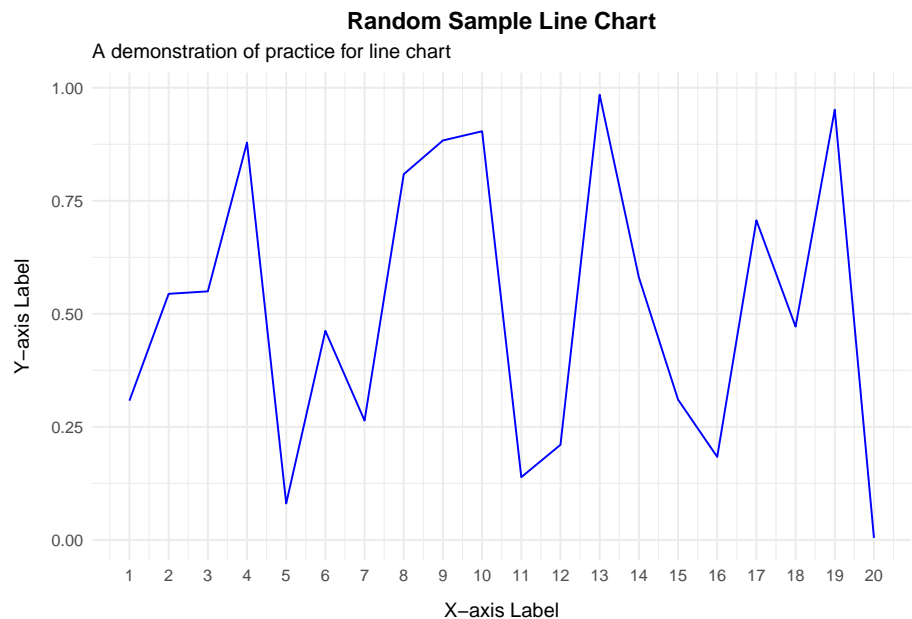
- Line Types and Point Shapes: When multiple lines are on the same chart, use different line types and point shapes to differentiate between them.
- Consistent Scaling: The scale on the y-axis should be consistent so that the viewer isn't misled.
- Annotations: Important points or changes can be annotated directly on the graph.
- Legends: If there are multiple lines or data points with different colors/shapes, a legend should be provided.

Let's apply these practices:

First, we generate 2 series of random data.

```
x <- seq(1, 20)
y <- runif(20)
data <- data.frame(x = x, y = y)
```

Below is a line chart of the random sample:



Source: Randomly generated data

Showcase real-world examples of time series visualisations.

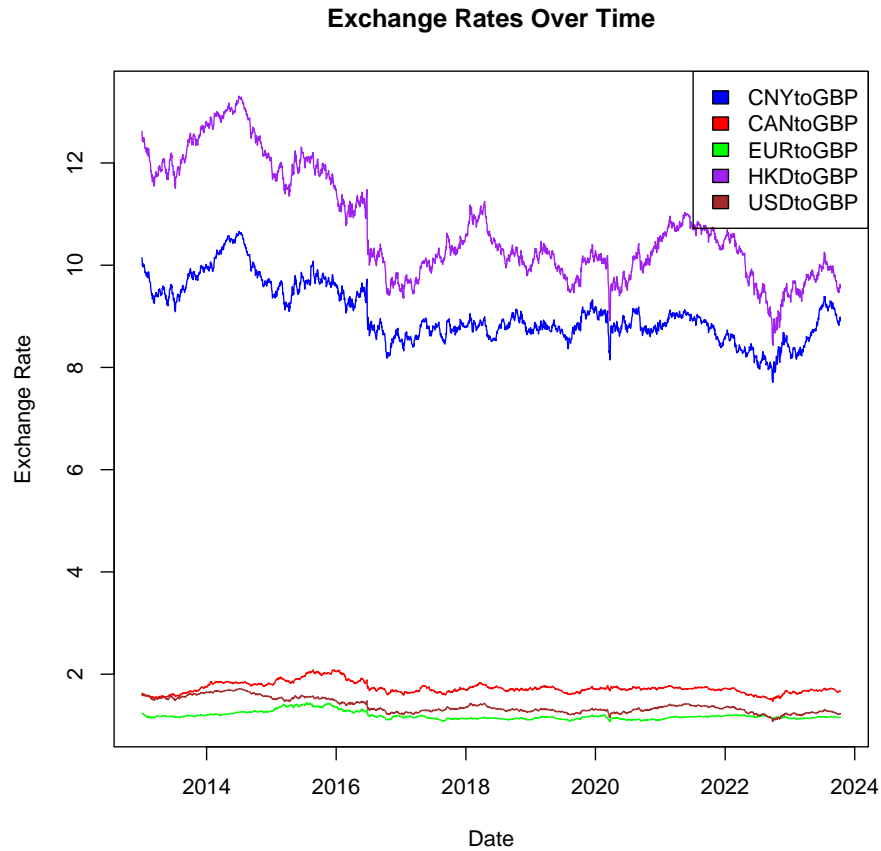
Time series of the daily CNY, CAN, EUR, HKD, USD versus GBP exchange reference rate data published by the European Central Bank over the time period from 01 Jan 2013 to 12 Oct 2023 (without weekends). The exchange rate tells you how many pounds you need to buy/sell 1 CNY, CAN, EUR, HKD, USD.

13.4 The data set has the format as below:

Date	CNYtoGBP	CANtoGBP	EURtoGBP	HKDtoGBP	USDtoGBP
%d-%m-%y	Value	Value	Value	Value	Value

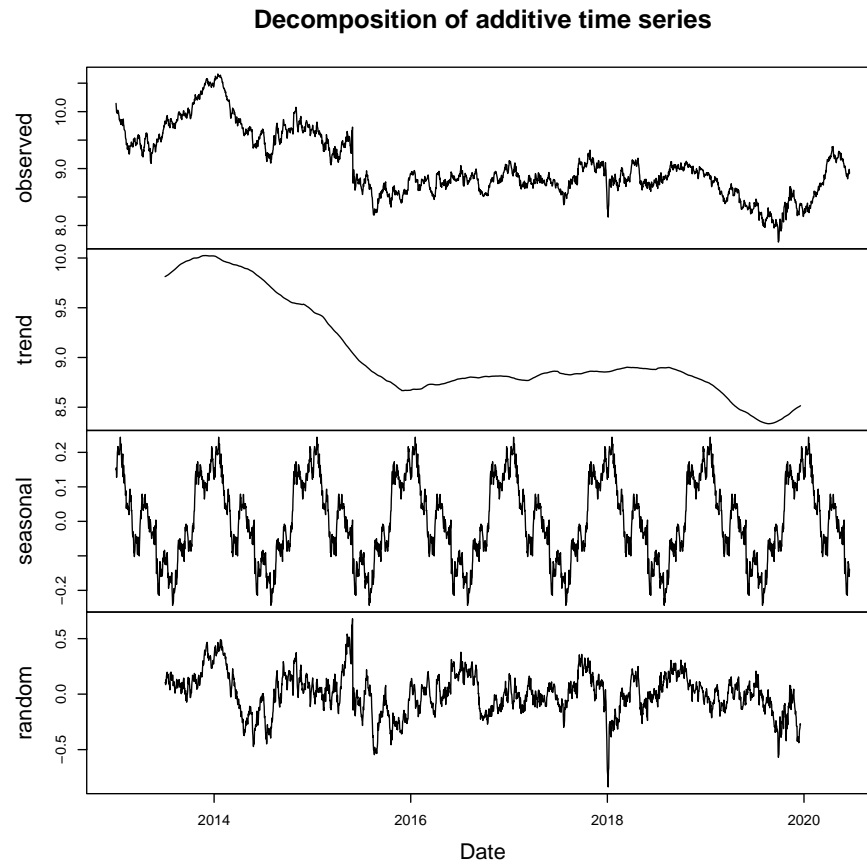
Table 1: Field Information: CNY, CAN, EUR, HKD, USD to GBP

13.5 Multiple time series in one plot:



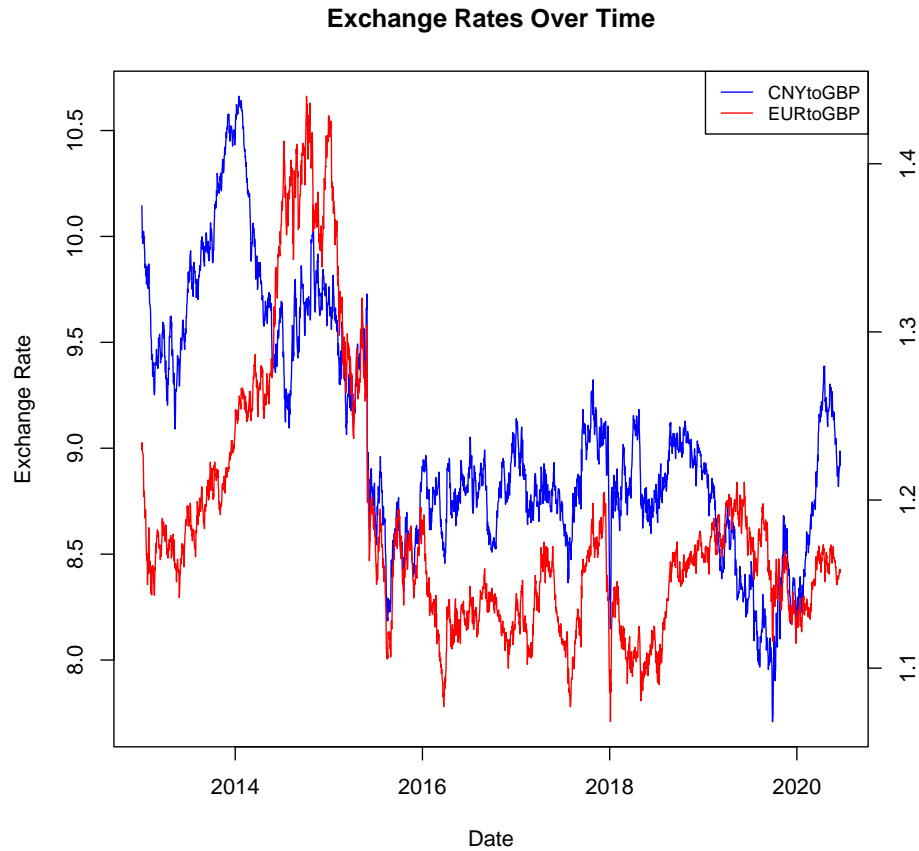
13.6 Decomposition of one time series into trend, seasonal, and random.

One of the primary advantages of time series visualization 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.



13.7 Double y-axis time series plot.

If we want to display two different time series that measure two different quantities at the same time points, we can draw the second series again on the second Y-axis on the right side.



14 Chapter 3.5 Network Graphs and Sankey Diagrams

xxx

15 Chapter 3.6 Geographic Maps and Spatial Data Visualisation

xxx

16 Chapter 3.7 3D and Interactive Visualisations

xxx

17 Chapter 3.8 Advanced Visualisation Techniques

xxx