1

THE IMPORTANCE OF DATA VISUALIZATION AND DATA EXPLORATION

OVERVIEW

This chapter introduces you to the basics of the statistical analysis of a dataset. We will look at basic operations for calculating the mean, median, and variance of different datasets and use NumPy and pandas to filter, sort, and shape the datasets to our requirements. The concepts we will cover will serve as a base of knowledge for the upcoming visualization chapters, in which we will work with real-world datasets. By the end of this chapter, you will be able to explain the importance of data visualization and calculate basic statistical values (such as median, mean, and variance), and use NumPy and pandas for data wrangling.

INTRODUCTION

Unlike machines, people are usually not equipped for interpreting a large amount of information from a random set of numbers and messages in each piece of data. Out of all our logical capabilities, we understand things best through the visual processing of information. When data is represented visually, the probability of understanding complex builds and numbers increases.

Python has recently emerged as a programming language that performs well for data analysis. It has applications across data science pipelines that convert data into a usable format (such as pandas), analyzes it (such as NumPy), and extract useful conclusions from the data to represent it in a visually appealing manner (such as Matplotlib or Bokeh). Python provides data visualization libraries that can help you assemble graphical representations efficiently.

In this book, you will learn how to use Python in combination with various libraries, such as **NumPy**, **pandas**, **Matplotlib**, **seaborn**, and **geoplotlib**, to create impactful data visualizations using real-world data. Besides that, you will also learn about the features of different types of charts and compare their advantages and disadvantages. This will help you choose the chart type that's suited to visualizing your data.

Once we understand the basics, we can cover more advanced concepts, such as interactive visualizations and how **Bokeh** can be used to create animated visualizations that tell a story. Upon completing this book, you will be able to perform **data wrangling**, extract relevant information, and visualize your findings descriptively.

INTRODUCTION TO DATA VISUALIZATION

Computers and smartphones store data such as names and numbers in a digital format. **Data representation** refers to the form in which you can store, process, and transmit data.

Representations can narrate a story and convey fundamental discoveries to your audience. Without appropriately modeling your information to use it to make meaningful findings, its value is reduced. Creating representations helps us achieve a more precise, more concise, and more direct perspective of information, making it easier for anyone to understand the data.

Information isn't equivalent to data. Representations are a useful apparatus to derive insights from the data. Thus, representations transform data into useful information.

THE IMPORTANCE OF DATA VISUALIZATION

Instead of just looking at data in the columns of an Excel spreadsheet, we get a better idea of what our data contains by using visualization. For instance, it's easy to see a pattern emerge from the numerical data that's given in the following scatter plot. It shows the correlation between body mass and the maximum longevity of various animals grouped by class. There is a positive correlation between body mass and maximum longevity:

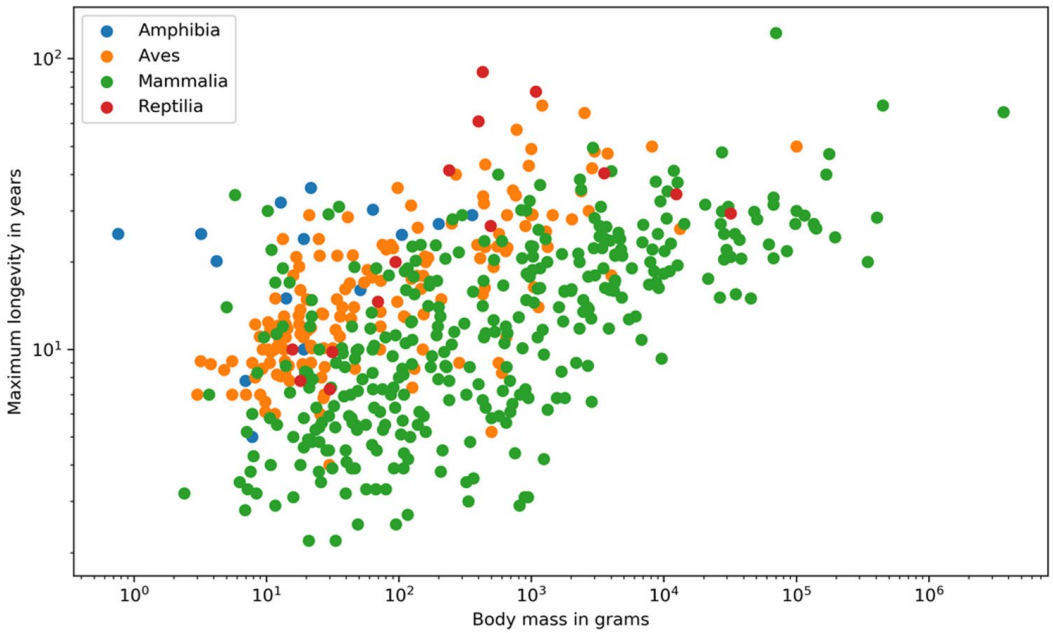

Figure 1.1: A simple example of data visualization

Visualizing data has many advantages, such as the following:

- Complex data can be easily understood.
- A simple visual representation of outliers, target audiences, and futures markets can be created.
- Storytelling can be done using dashboards and animations.
- Data can be explored through interactive visualizations.

DATA WRANGLING

Data wrangling is the process of transforming raw data into a suitable representation for various tasks. It is the discipline of augmenting, cleaning, filtering, standardizing, and enriching data in a way that allows it to be used in a downstream task, which in our case is data visualization.

Look at the following data wrangling process flow diagram to understand how accurate and actionable data can be obtained for business analysts to work on:

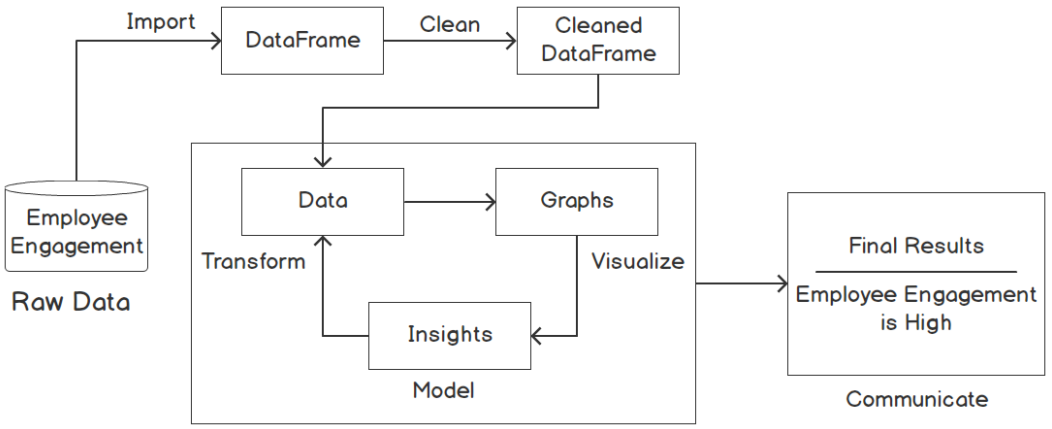

Figure 1.2: Data wrangling process to measure employee engagement

In relation to the preceding figure, the following steps explain the flow of the data wrangling process:

1. First, the Employee Engagement data is in its raw form.
2. Then, the data gets imported as a DataFrame and is later cleaned.
3. The cleaned data is then transformed into graphs, from which findings can be derived.
4. Finally, we analyze this data to communicate the final results.

For example, employee engagement can be measured based on raw data gathered from feedback surveys, employee tenure, exit interviews, one-on-one meetings, and so on. This data is cleaned and made into graphs based on parameters such as referrals, faith in leadership, and scope of promotions. The percentages, that is, information derived from the graphs, help us reach our result, which is to determine the measure of employee engagement.

TOOLS AND LIBRARIES FOR VISUALIZATION

There are several approaches to creating data visualizations. Depending on your requirements, you might want to use a non-coding tool such as **Tableau**, which allows you to get a good feel for your data. Besides Python, which will be used in this book, **MATLAB** and **R** are widely used in data analytics.

However, Python is the most popular language in the industry. Its ease of use and the speed at which you can manipulate and visualize data, combined with the availability of a number of libraries, make Python the best choice for data visualization.

> **NOTE**
>
> MATLAB (https://www.mathworks.com/products/matlab.html), R (https://www.r-project.org), and Tableau (https://www.tableau.com) are not part of this book; we will only cover the relevant tools and libraries for Python.

OVERVIEW OF STATISTICS

Statistics is a combination of the analysis, collection, interpretation, and representation of numerical data. **Probability** is a measure of the likelihood that an event will occur and is quantified as a number between 0 and 1.

A **probability distribution** is a function that provides the probability for every possible event. A probability distribution is frequently used for statistical analysis. The higher the probability, the more likely the event. There are two types of probability distributions, namely discrete and continuous.

A **discrete probability distribution** shows all the values that a random variable can take, together with their probability. The following diagram illustrates an example of a discrete probability distribution. If we have a six-sided die, we can roll each number between 1 and 6. We have six events that can occur based on the number that's rolled. There is an equal probability of rolling any of the numbers, and the individual probability of any of the six events occurring is 1/6:

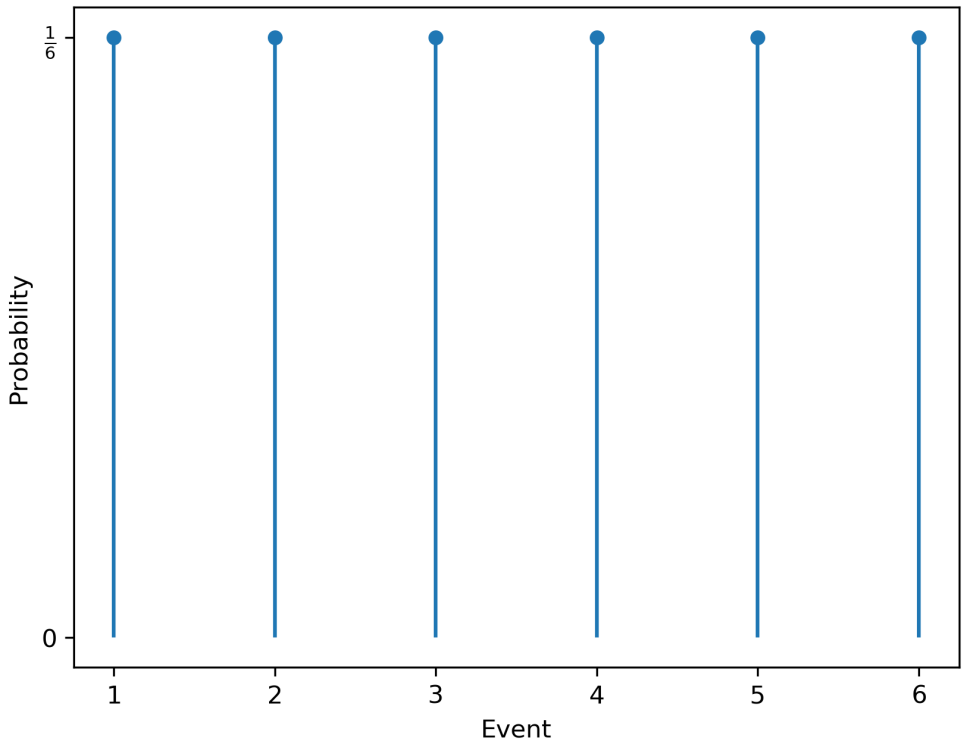

Figure 1.3: Discrete probability distribution for die rolls

A **continuous probability distribution** defines the probabilities of each possible value of a continuous random variable. The following diagram provides an example of a continuous probability distribution. This example illustrates the distribution of the time needed to drive home. In most cases, around 60 minutes is needed, but sometimes, less time is needed because there is no traffic, and sometimes, much more time is needed if there are traffic jams:

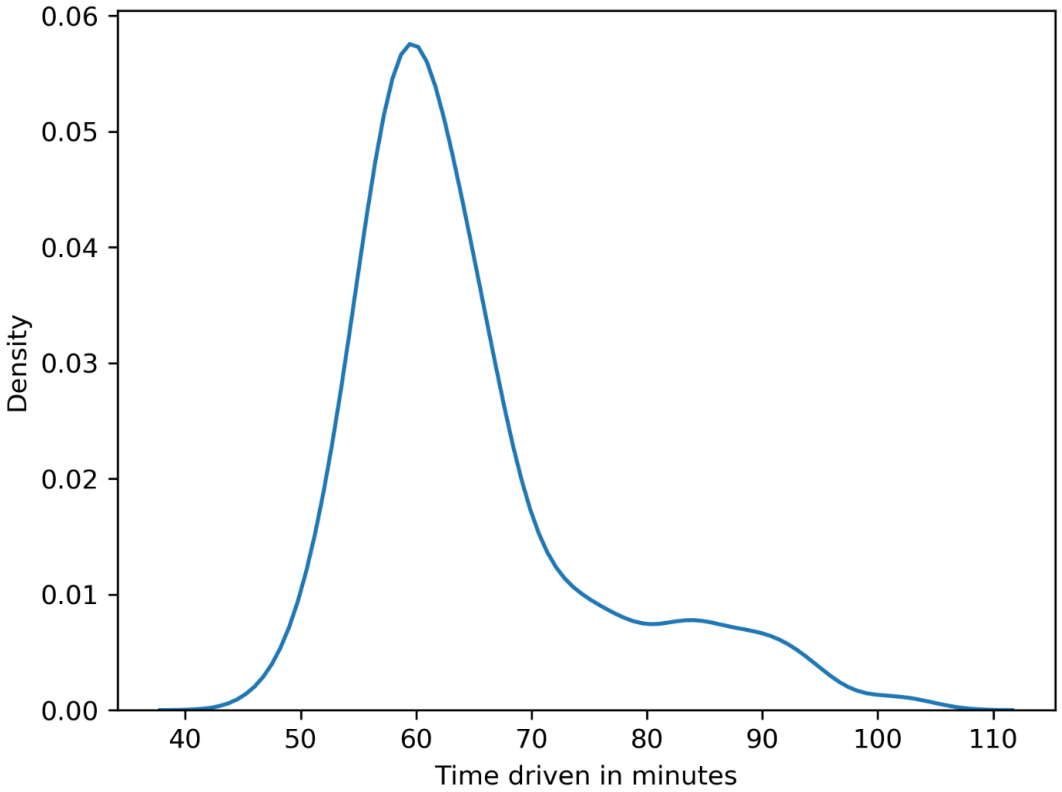

Figure 1.4: Continuous probability distribution for the time taken to reach home

MEASURES OF CENTRAL TENDENCY

Measures of central tendency are often called **averages** and describe central or typical values for a probability distribution. We are going to discuss three kinds of averages in this chapter:

- **Mean**: The arithmetic average is computed by summing up all measurements and dividing the sum by the number of observations. The mean is calculated as follows:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Figure 1.5: Formula for mean

- **Median**: This is the middle value of the ordered dataset. If there is an even number of observations, the median will be the average of the two middle values. The median is less prone to outliers compared to the mean, where outliers are distinct values in data.
- **Mode**: Our last measure of central tendency, the mode is defined as the most frequent value. There may be more than one mode in cases where multiple values are equally frequent.

For example, a die was rolled 10 times, and we got the following numbers: 4, 5, 4, 3, 4, 2, 1, 1, 2, and 1.

The mean is calculated by summing all the events and dividing them by the number of observations: *(4+5+4+3+4+2+1+1+2+1)/10=2.7*.

To calculate the median, the die rolls have to be ordered according to their values. The ordered values are as follows: 1, 1, 1, 2, 2, 3, 4, 4, 4, 5. Since we have an even number of die rolls, we need to take the average of the two middle values. The average of the two middle values is *(2+3)/2=2.5*.

The modes are 1 and 4 since they are the two most frequent events.

MEASURES OF DISPERSION

Dispersion, also called **variability**, is the extent to which a probability distribution is stretched or squeezed.

The different measures of dispersion are as follows:

- **Variance**: The variance is the expected value of the squared deviation from the mean. It describes how far a set of numbers is spread out from their mean. Variance is calculated as follows:

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Figure 1.6: Formula for mean

- **Standard deviation**: This is the square root of the variance.
- **Range**: This is the difference between the largest and smallest values in a dataset.
- **Interquartile range**: Also called the **midspread** or **middle 50%**, this is the difference between the 75th and 25th percentiles, or between the upper and lower quartiles.

CORRELATION

The measures we have discussed so far only considered single variables. In contrast, **correlation** describes the statistical relationship between two variables:

- In a positive correlation, both variables move in the same direction.
- In a negative correlation, the variables move in opposite directions.
- In zero correlation, the variables are not related.

> **NOTE**
>
> One thing you should be aware of is that correlation does not imply causation. Correlation describes the relationship between two or more variables, while causation describes how one event is caused by another. For example, consider a scenario in which ice cream sales are correlated with the number of drowning deaths. But that doesn't mean that ice cream consumption causes drowning. There could be a third variable, say temperature, that may be responsible for this correlation. Higher temperatures may cause an increase in both ice cream sales and more people engaging in swimming, which may be the real reason for the increase in deaths due to drowning.

Example

Consider you want to find a decent apartment to rent that is not too expensive compared to other apartments you've found. The other apartments (all belonging to the same locality) you found on a website are priced as follows: $700, $850, $1,500, and $750 per month. Let's calculate some values statistical measures to help us make a decision:

- The mean is *($700 + $850 + $1,500 + $750) / 4 = $950*.
- The median is *($750 + $850) / 2 = $800*.
- The standard deviation is $\sqrt{\frac{(\$700-\$950)^2+(\$850-\$950)^2+(\$1500-\$950)^2+(\$750-\$950)^2}{4}} = \$322.10$.
- The range is *$1,500 - $700 = $800*.

As an exercise, you can try and calculate the variance as well. However, note that compared with all the above values, the median value ($800) is a better statistical measure in this case since it is less prone to outliers (the rent amount of $1,500). Given that all apartments belong to the same locality, you can clearly see that the apartment costing $1500 is definitely priced much higher as compared with other apartments. A simple statistical analysis helped us to narrow down our choices.

TYPES OF DATA

It is important to understand what kind of data you are dealing with so that you can select both the right statistical measure and the right visualization. We categorize data as categorical/qualitative and numerical/quantitative. Categorical data describes characteristics, for example, the color of an object or a person's gender. We can further divide categorical data into nominal and ordinal data. In contrast to nominal data, ordinal data has an order.

Numerical data can be divided into discrete and continuous data. We speak of discrete data if the data can only have certain values, whereas continuous data can take any value (sometimes limited to a range).

Another aspect to consider is whether the data has a temporal domain – in other words, is it bound to time or does it change over time? If the data is bound to a location, it might be interesting to show the spatial relationship, so you should keep that in mind as well. The following flowchart classifies the various data types:

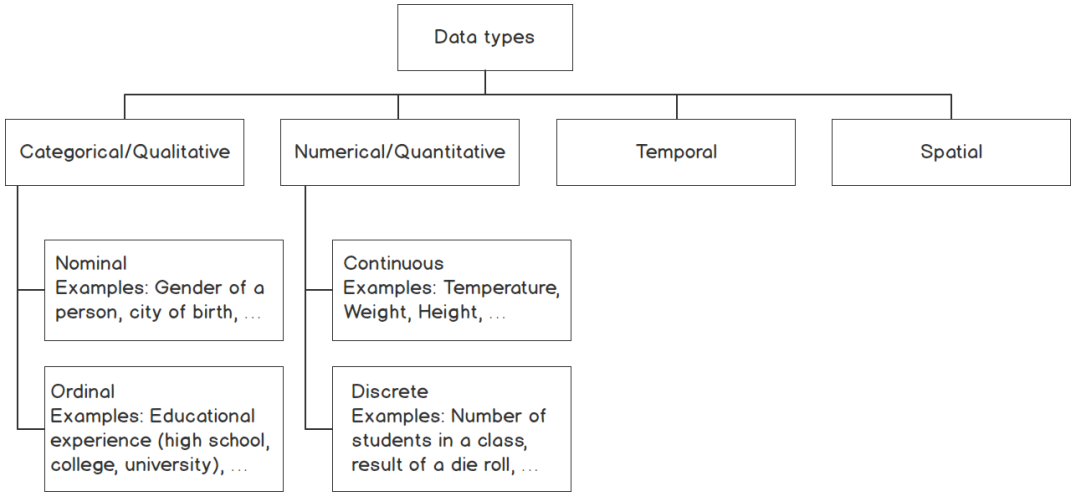

Figure 1.7: Classification of types of data

SUMMARY STATISTICS

In real-world applications, we often encounter enormous datasets. Therefore, **summary statistics** are used to summarize important aspects of data. They are necessary to communicate large amounts of information in a compact and simple way.

We have already covered measures of central tendency and dispersion, which are both summary statistics. It is important to know that measures of central tendency show a center point in a set of data values, whereas measures of dispersion show how much the data varies.

The following table gives an overview of which measure of central tendency is best suited to a particular type of data:

Data type	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Numerical	Mean/Median

Figure 1.8: Best suited measures of central tendency for different types of data

2

ALL YOU NEED TO KNOW ABOUT PLOTS

OVERVIEW

This chapter will teach you the fundamentals of the various types of plots such as line charts, bar charts, bubble plots, radar charts, and so on. For each plot type that we discuss, we will also describe best practices and use cases. The activities presented in this chapter will enable you to apply the knowledge gained. By the end of this chapter, you will be equipped with the important skill of identifying the best plot type for a given dataset and scenario.

INTRODUCTION

In the previous chapter, we learned how to work with new datasets and get familiar with their data and structure. We also got hands-on experience of how to analyze and transform them using different data wrangling techniques such as filtering, sorting, and reshaping. All of these techniques will come in handy when working with further real-world datasets in the coming activities.

In this chapter, we will focus on various visualizations and identify which visualization is best for showing certain information for a given dataset. We will describe every visualization in detail and give practical examples, such as comparing different stocks over time or comparing the ratings for different movies. Starting with comparison plots, which are great for comparing multiple variables over time, we will look at their types (such as line charts, bar charts, and radar charts).

We will then move onto relation plots, which are handy for showing relationships among variables. We will cover scatter plots for showing the relationship between two variables, bubble plots for three variables, correlograms for variable pairs, and finally, heatmaps for visualizing multivariate data.

The chapter will further explain composition plots (used to visualize variables that are part of a whole), as well as pie charts, stacked bar charts, stacked area charts, and Venn diagrams. To give you a deeper insight into the distribution of variables, we will discuss distribution plots, describing histograms, density plots, box plots, and violin plots.

Finally, we will talk about dot maps, connection maps, and choropleth maps, which can be categorized into geoplots. Geoplots are useful for visualizing geospatial data. Let's start with the family of comparison plots, including line charts, bar charts, and radar charts.

> **NOTE**
>
> The data used in this chapter has been provided to demonstrate the different types of plots available to you. In each case, the data itself will be revisited and explained more fully in a later chapter.

COMPARISON PLOTS

Comparison plots include charts that are ideal for comparing multiple variables or variables over time. Line charts are great for visualizing variables over time. For comparison among items, bar charts (also called column charts) are the best way to go. For a certain time period (say, fewer than 10-time points), vertical bar charts can be used as well. Radar charts or spider plots are great for visualizing multiple variables for multiple groups.

LINE CHART

Line charts are used to display quantitative values over a continuous time period and show information as a series. A line chart is ideal for a time series that is connected by straight-line segments.

The value being measured is placed on the y-axis, while the x-axis is the timescale.

USES

- Line charts are great for comparing multiple variables and visualizing trends for both single as well as multiple variables, especially if your dataset has many time periods (more than 10).
- For smaller time periods, vertical bar charts might be the better choice.

The following diagram shows a trend of real estate prices (per million US dollars) across two decades. Line charts are ideal for showing data trends:

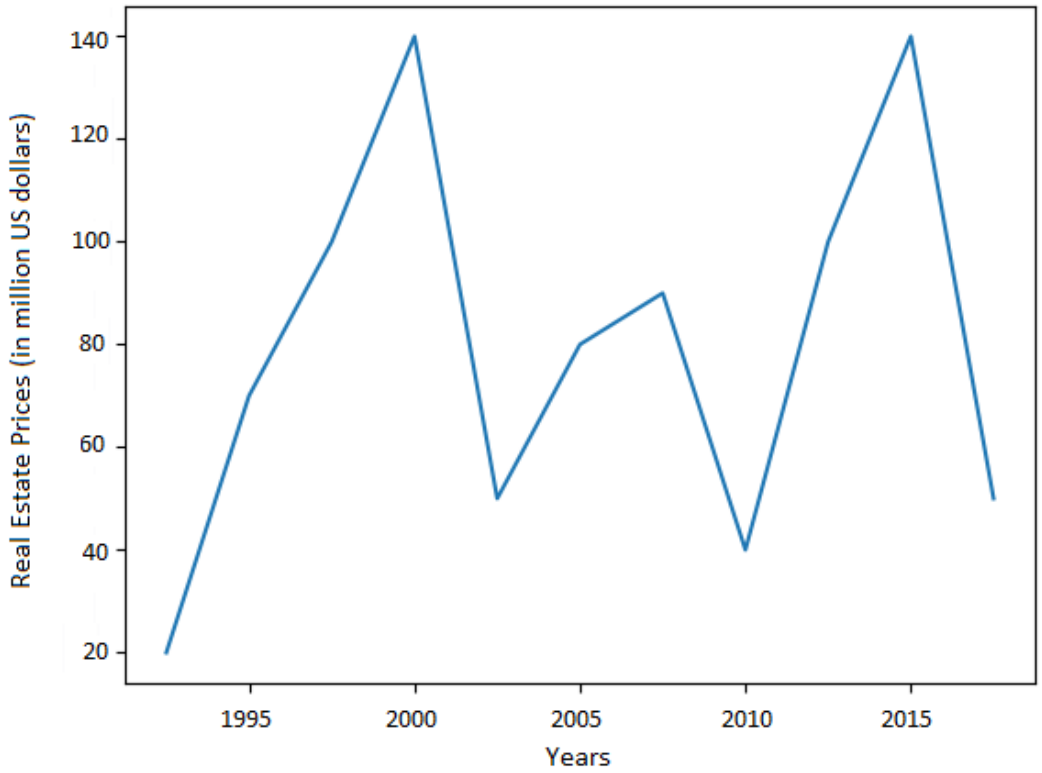

Figure 2.1: Line chart for a single variable

EXAMPLE

The following figure is a multiple-variable line chart that compares the stock-closing prices for Google, Facebook, Apple, Amazon, and Microsoft. A line chart is great for comparing values and visualizing the trend of the stock. As we can see, Amazon shows the highest growth:

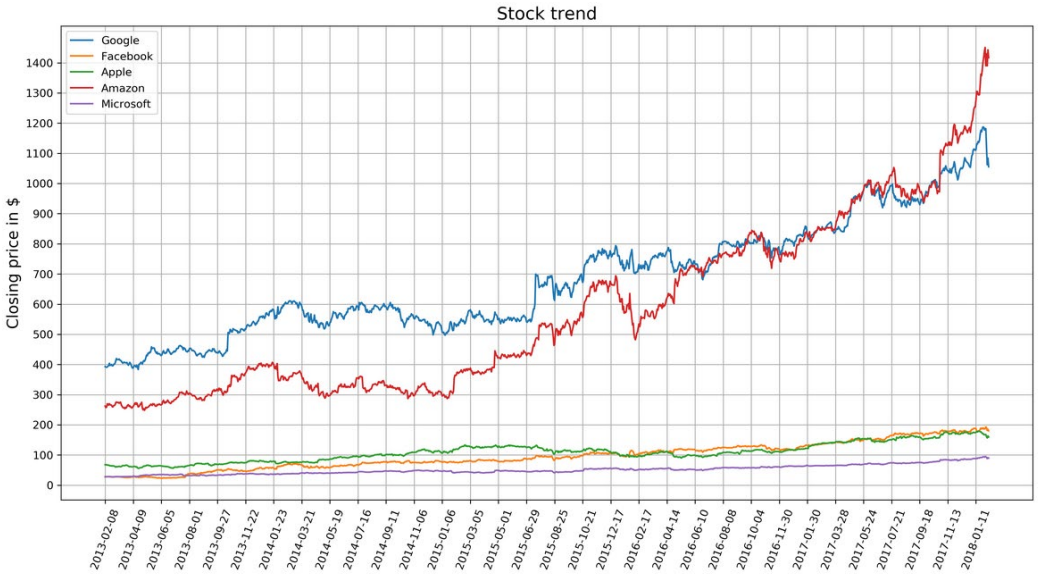

Figure 2.2: Line chart showing stock trends for five companies

DESIGN PRACTICES

- Avoid too many lines per chart.
- Adjust your scale so that the trend is clearly visible.

> **NOTE**
>
> For plots with multiple variables, a legend should be given to describe each variable.

BAR CHART

In a bar chart, the bar length encodes the value. There are two variants of bar charts: vertical bar charts and horizontal bar charts.

USE

While they are both used to compare numerical values across categories, vertical bar charts are sometimes used to show a single variable over time.

DON'TS OF BAR CHARTS

- Don't confuse vertical bar charts with histograms. Bar charts compare different variables or categories, while histograms show the distribution for a single variable. Histograms will be discussed later in this chapter.
- Another common mistake is to use bar charts to show central tendencies among groups or categories. Use box plots or violin plots to show statistical measures or distributions in these cases.

EXAMPLES

The following diagram shows a vertical bar chart. Each bar shows the marks out of 100 that 5 students obtained in a test:

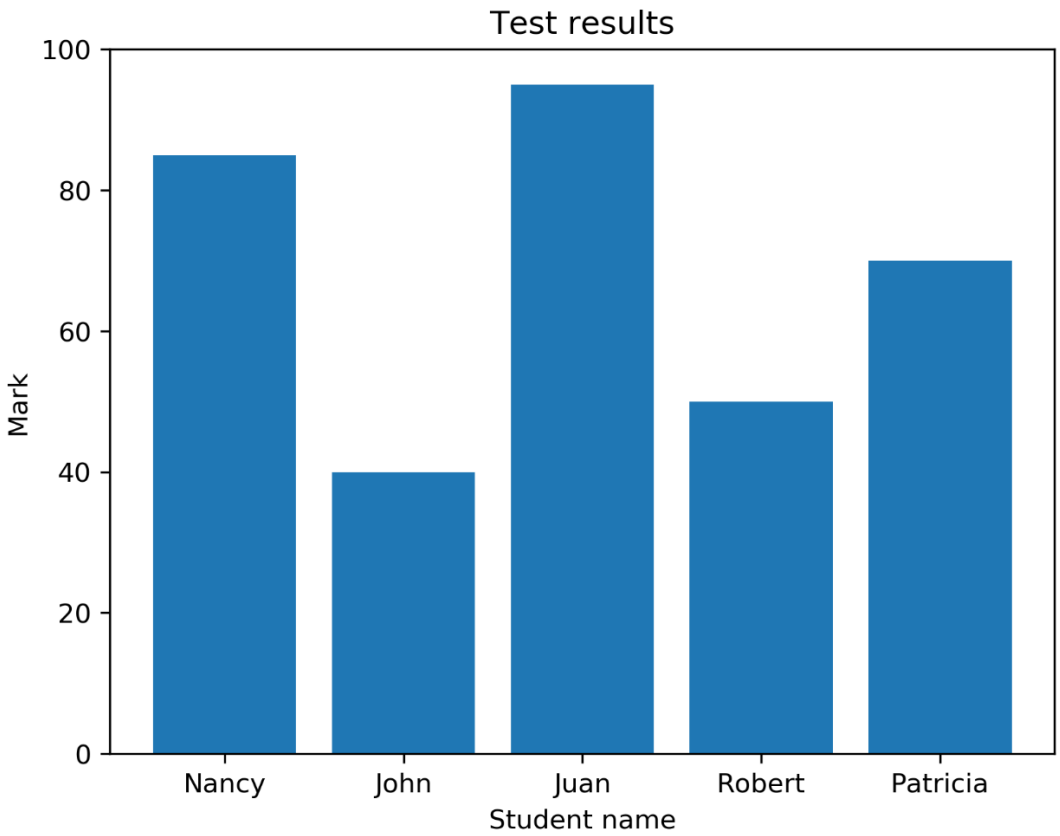

Figure 2.3: Vertical bar chart using student test data

The following diagram shows a horizontal bar chart. Each bar shows the marks out of 100 that 5 students obtained in a test:

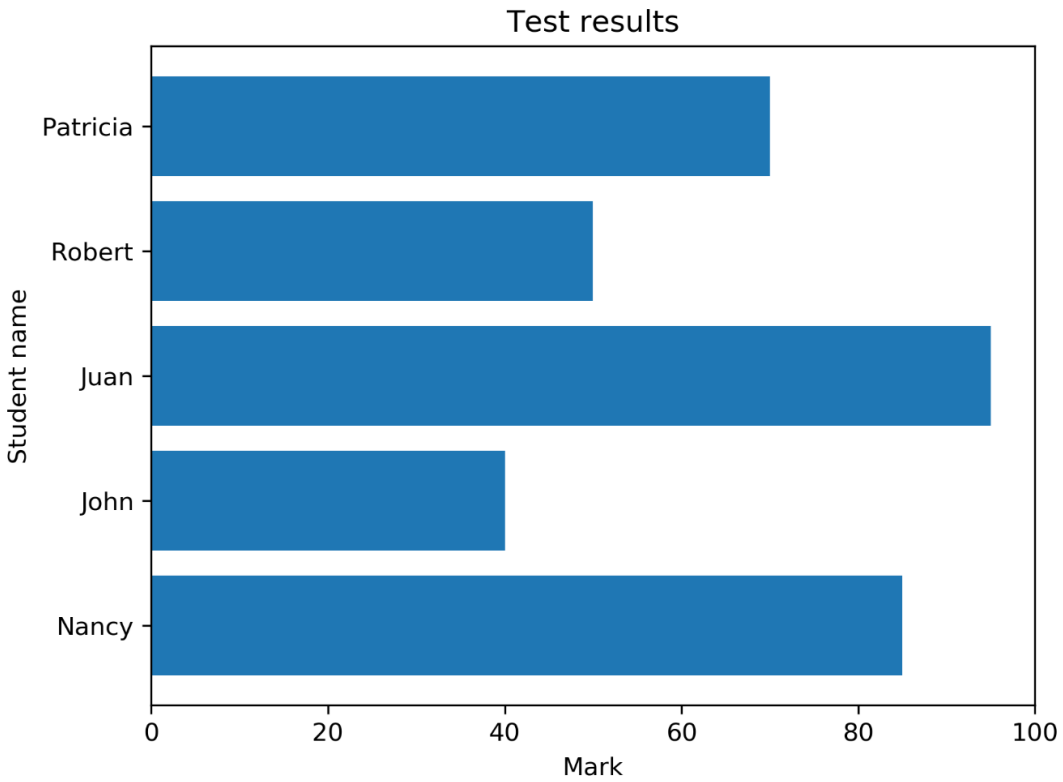

Figure 2.4: Horizontal bar chart using student test data

The following diagram compares movie ratings, giving two different scores. The Tomatometer is the percentage of approved critics who have given a positive review for the movie. The Audience Score is the percentage of users who have given a score of 3.5 or higher out of 5. As we can see, **The Martian** is the only movie with both a high Tomatometer and Audience Score. **The Hobbit: An Unexpected Journey** has a relatively high Audience Score compared to the Tomatometer score, which might be due to a huge fan base:

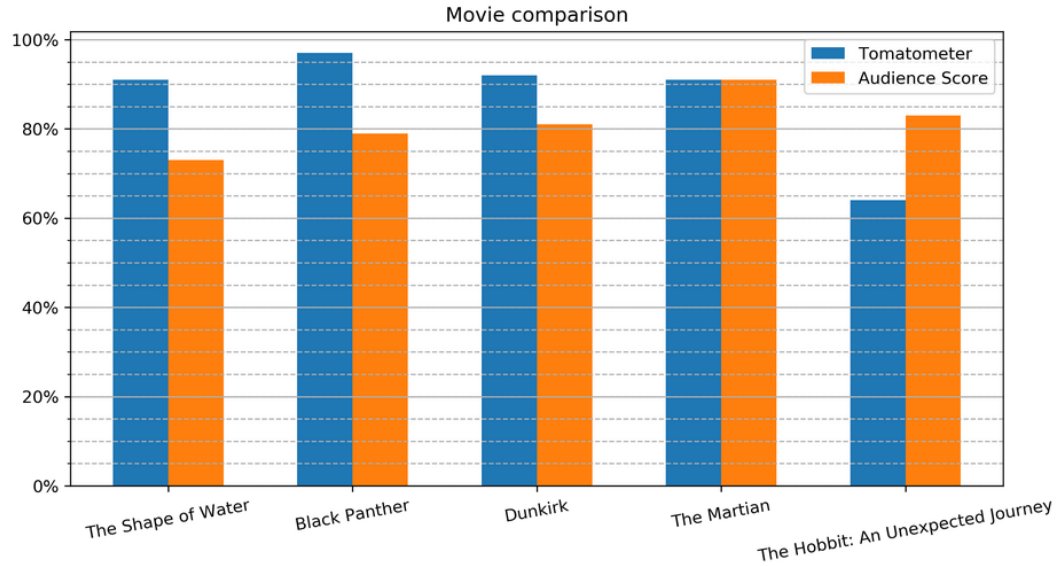

Figure 2.5: Comparative bar chart

DESIGN PRACTICES

- The axis corresponding to the numerical variable should start at zero. Starting with another value might be misleading, as it makes a small value difference look like a big one.
- Use horizontal labels—that is, as long as the number of bars is small, and the chart doesn't look too cluttered.
- The labels can be rotated to different angles if there isn't enough space to present them horizontally. You can see this on the labels of the x-axis of the preceding diagram.

RADAR CHART

Radar charts (also known as **spider** or **web charts**) visualize multiple variables with each variable plotted on its own axis, resulting in a polygon. All axes are arranged radially, starting at the center with equal distances between one another, and have the same scale.

USES

- Radar charts are great for comparing multiple quantitative variables for a single group or multiple groups.
- They are also useful for showing which variables score high or low within a dataset, making them ideal for visualizing performance.

EXAMPLES

The following diagram shows a radar chart for a single variable. This chart displays data about a student scoring marks in different subjects:

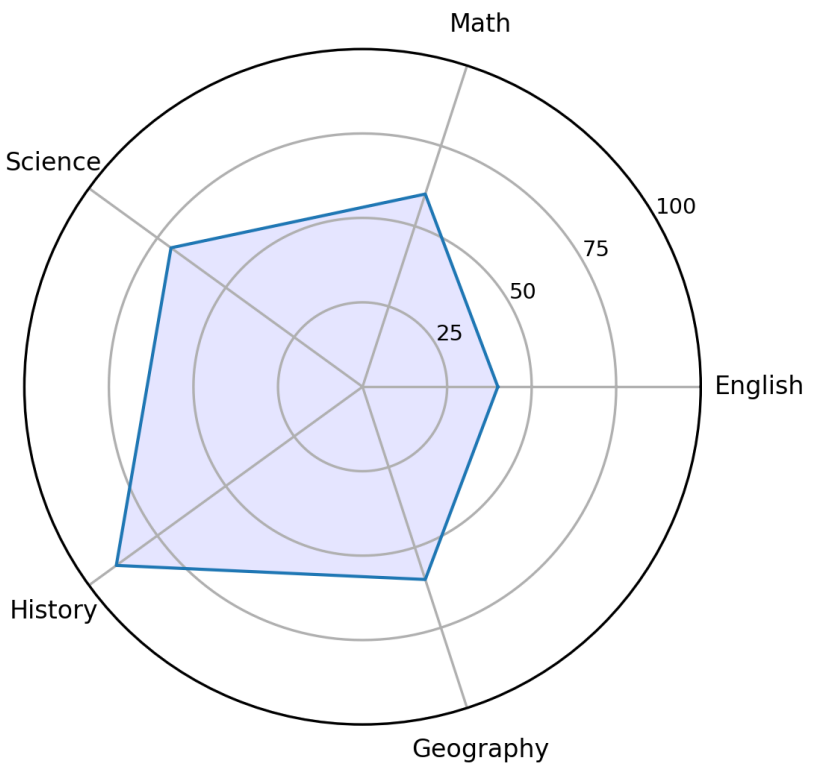

Figure 2.6: Radar chart for one variable (student)

The following diagram shows a radar chart for two variables/groups. Here, the chart explains the marks that were scored by two students in different subjects:

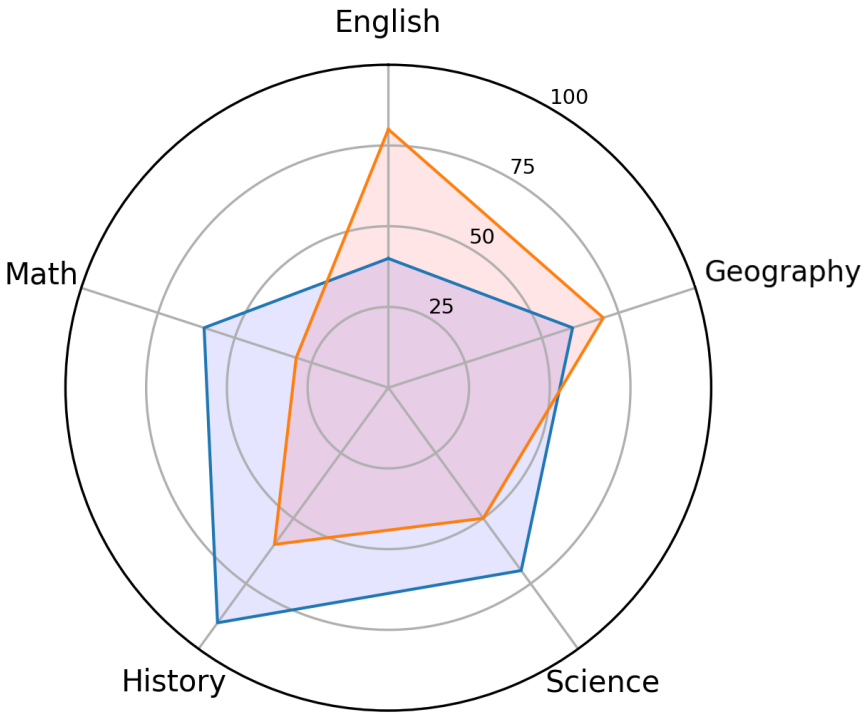

Figure 2.7: Radar chart for two variables (two students)

The following diagram shows a radar chart for multiple variables/groups. Each chart displays data about a student's performance in different subjects:

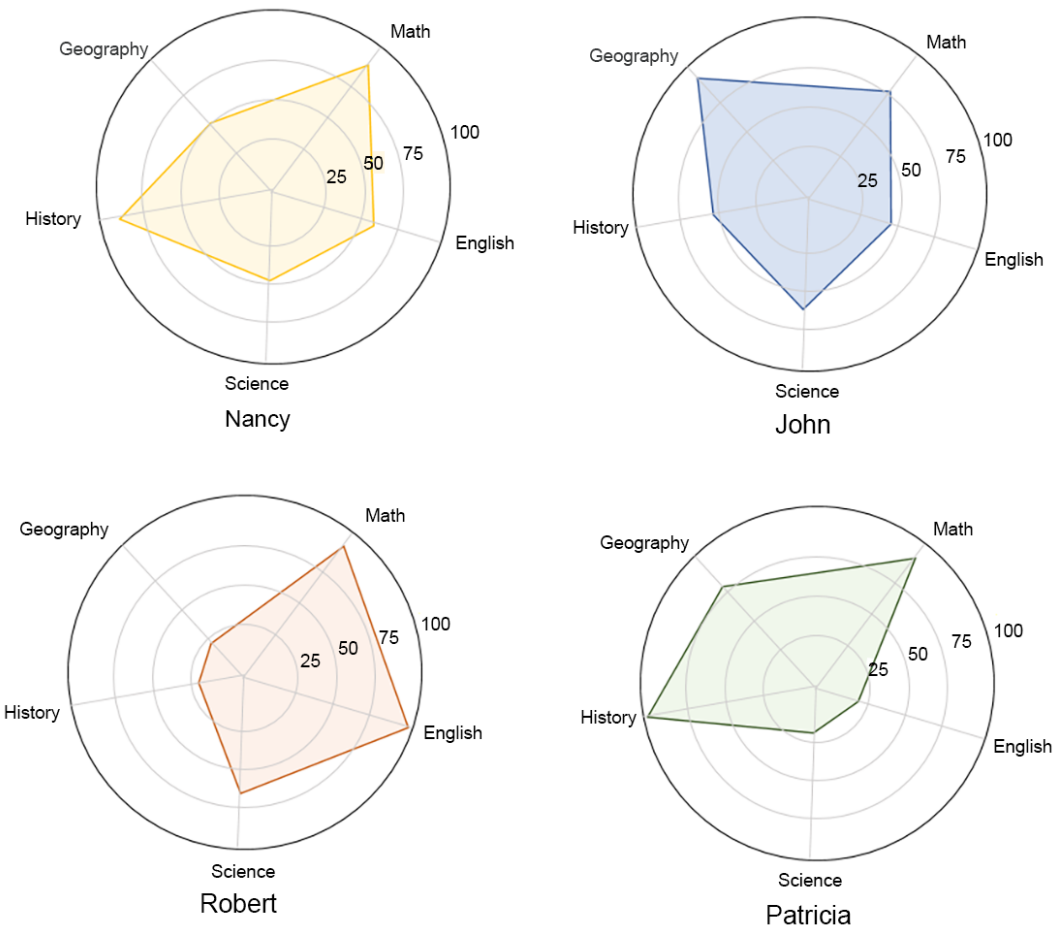

Figure 2.8: Radar chart with faceting for multiple variables (multiple students)

DESIGN PRACTICES

- Try to display 10 factors or fewer on a single radar chart to make it easier to read.
- Use **faceting** (displaying each variable in a separate plot) for multiple variables/groups, as shown in the preceding diagram, in order to maintain clarity.

In the first section, we learned which plots are suitable for comparing items. Line charts are great for comparing something over time, whereas bar charts are for comparing different items. Last but not least, radar charts are best suited for visualizing multiple variables for multiple groups. In the following activity, you can check whether you understood which plot is best for which scenario.

ACTIVITY 2.01: EMPLOYEE SKILL COMPARISON

You are given scores of four employees (Alex, Alice, Chris, and Jennifer) for five attributes: efficiency, quality, commitment, responsible conduct, and cooperation. Your task is to compare the employees and their skills. This activity will foster your skills in choosing the best visualization when it comes to comparing items.

1. Which charts are suitable for this task?
2. You are given the following bar and radar charts. List the advantages and disadvantages of both charts. Which is the better chart for this task in your opinion, and why?

   The following diagram shows a bar chart for the employee skills:

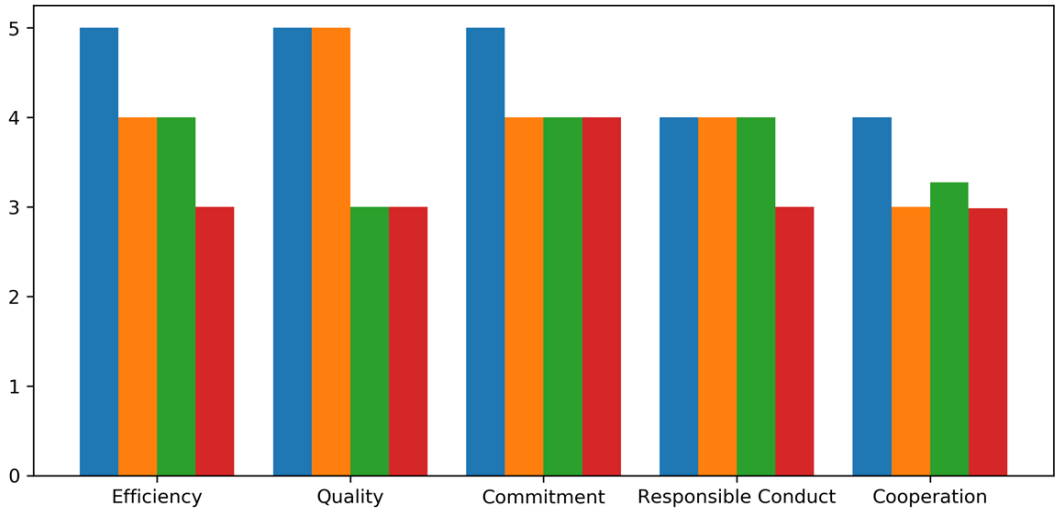

Figure 2.9: Employee skills comparison with a bar chart

The following diagram shows a radar chart for the employee skills:

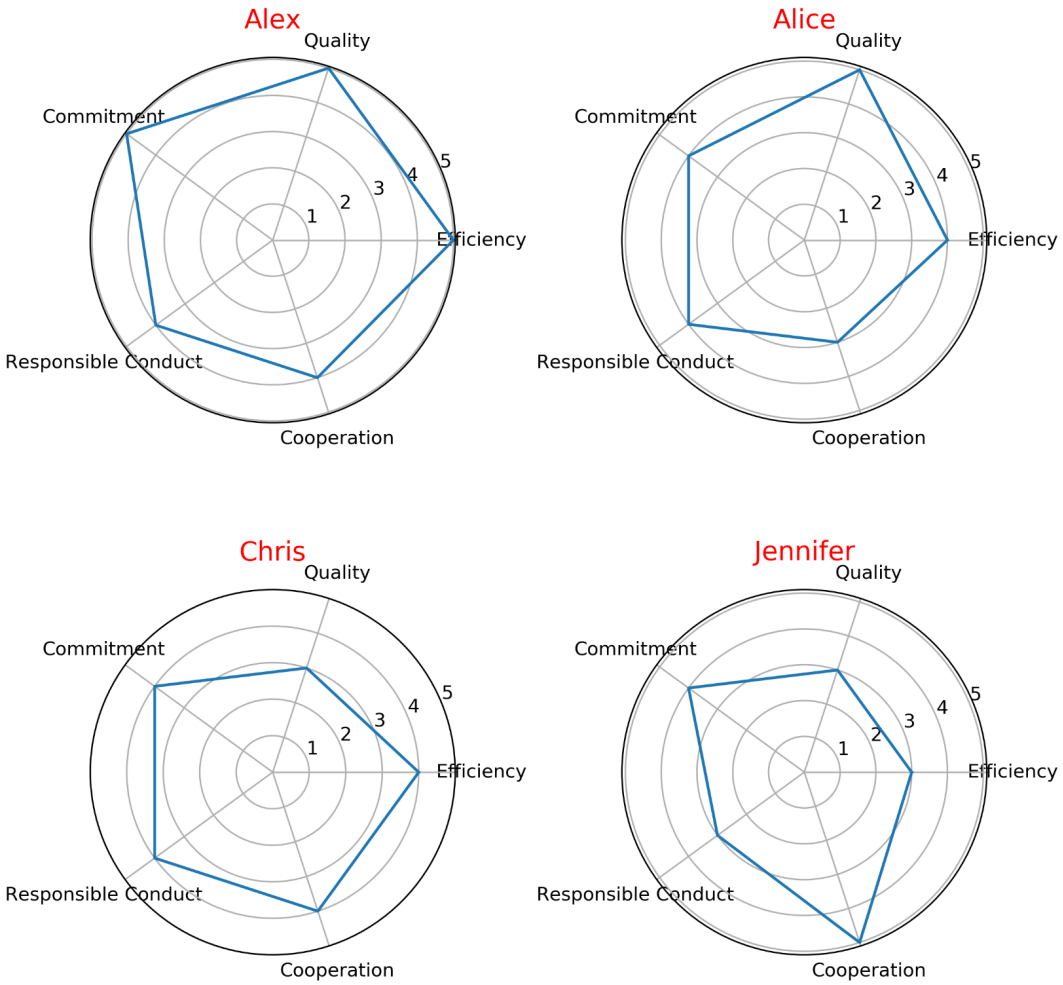

Figure 2.10: Employee skills comparison with a radar chart

3. What could be improved in the respective visualizations?

> **NOTE**
>
> The solution for this activity can be found via this link.

Concluding the activity, you hopefully have a good understanding of deciding which comparison plots are best for the situation. In the next section, we will discuss different relation plots.

RELATION PLOTS

Relation plots are perfectly suited to showing relationships among variables. A scatter plot visualizes the correlation between two variables for one or multiple groups. Bubble plots can be used to show relationships between three variables. The additional third variable is represented by the dot size. Heatmaps are great for revealing patterns or correlations between two qualitative variables. A correlogram is a perfect visualization for showing the correlation among multiple variables.

SCATTER PLOT

Scatter plots show data points for two numerical variables, displaying a variable on both axes.

USES

- You can detect whether a correlation (relationship) exists between two variables.
- They allow you to plot the relationship between multiple groups or categories using different colors.
- A bubble plot, which is a variation of the scatter plot, is an excellent tool for visualizing the correlation of a third variable.

EXAMPLES

The following diagram shows a scatter plot of **height** and **weight** of persons belonging to a single group:

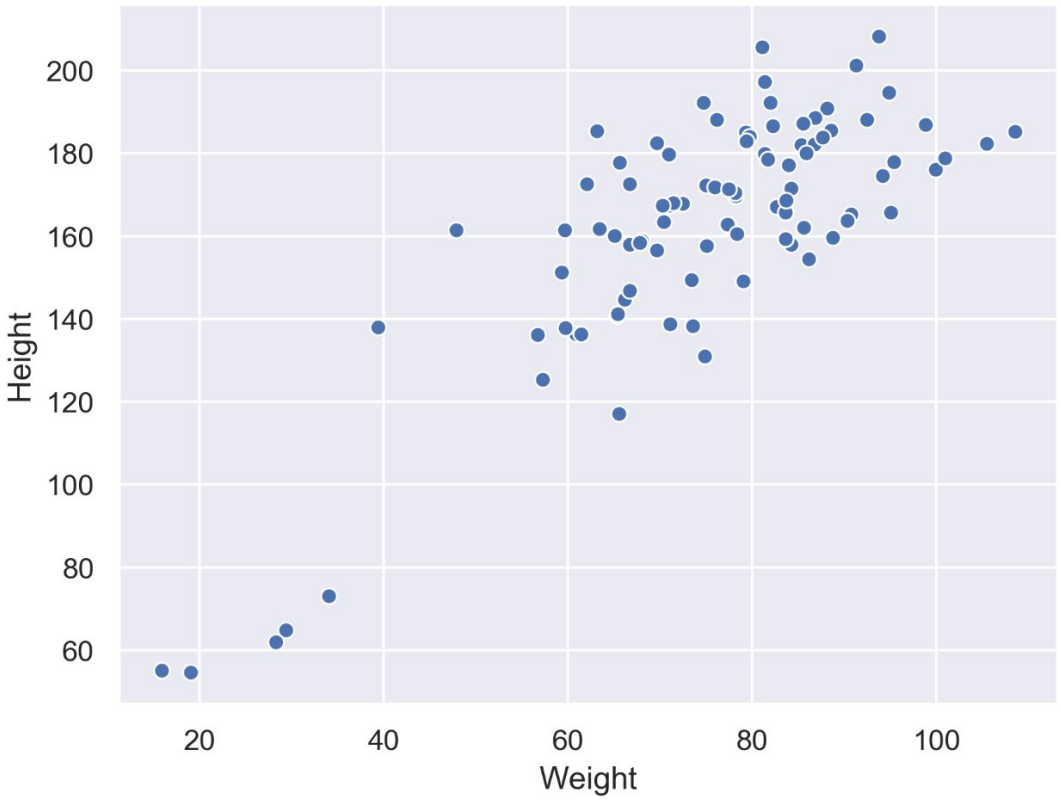

Figure 2.11: Scatter plot with a single group

The following diagram shows the same data as in the previous plot but differentiates between groups. In this case, we have different groups: **A**, **B**, and **C**:

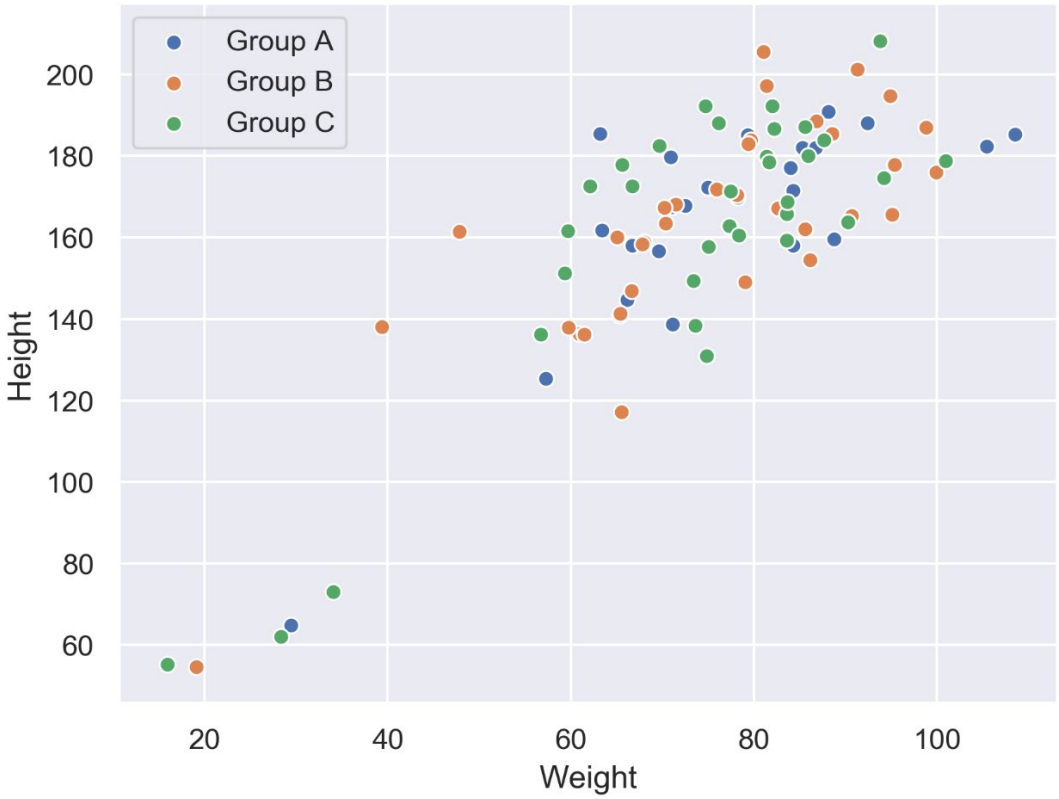

Figure 2.12: Scatter plot with multiple groups

The following diagram shows the correlation between body mass and the maximum longevity for various animals grouped by their classes. There is a positive correlation between body mass and maximum longevity:

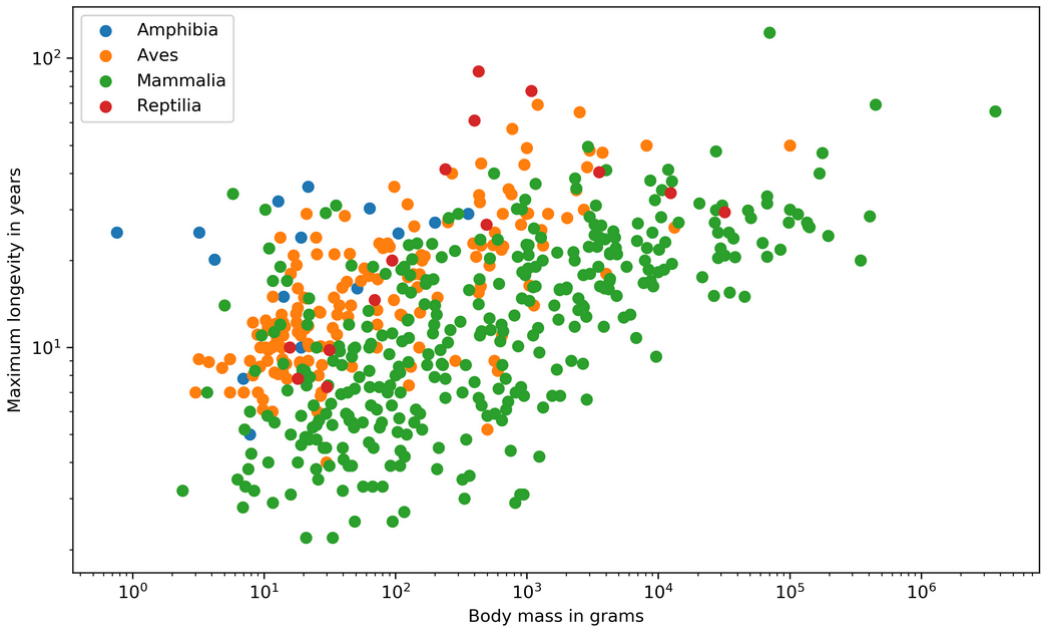

Figure 2.13: Correlation between body mass and maximum longevity for animals

DESIGN PRACTICES

- Start both axes at zero to represent data accurately.
- Use contrasting colors for data points and avoid using symbols for scatter plots with multiple groups or categories.

VARIANTS: SCATTER PLOTS WITH MARGINAL HISTOGRAMS

In addition to the scatter plot, which visualizes the correlation between two numerical variables, you can plot the marginal distribution for each variable in the form of histograms to give better insight into how each variable is distributed.

EXAMPLES

The following diagram shows the correlation between body mass and the maximum longevity for animals in the **`Aves`** class. The marginal histograms are also shown, which helps to get a better insight into both variables:

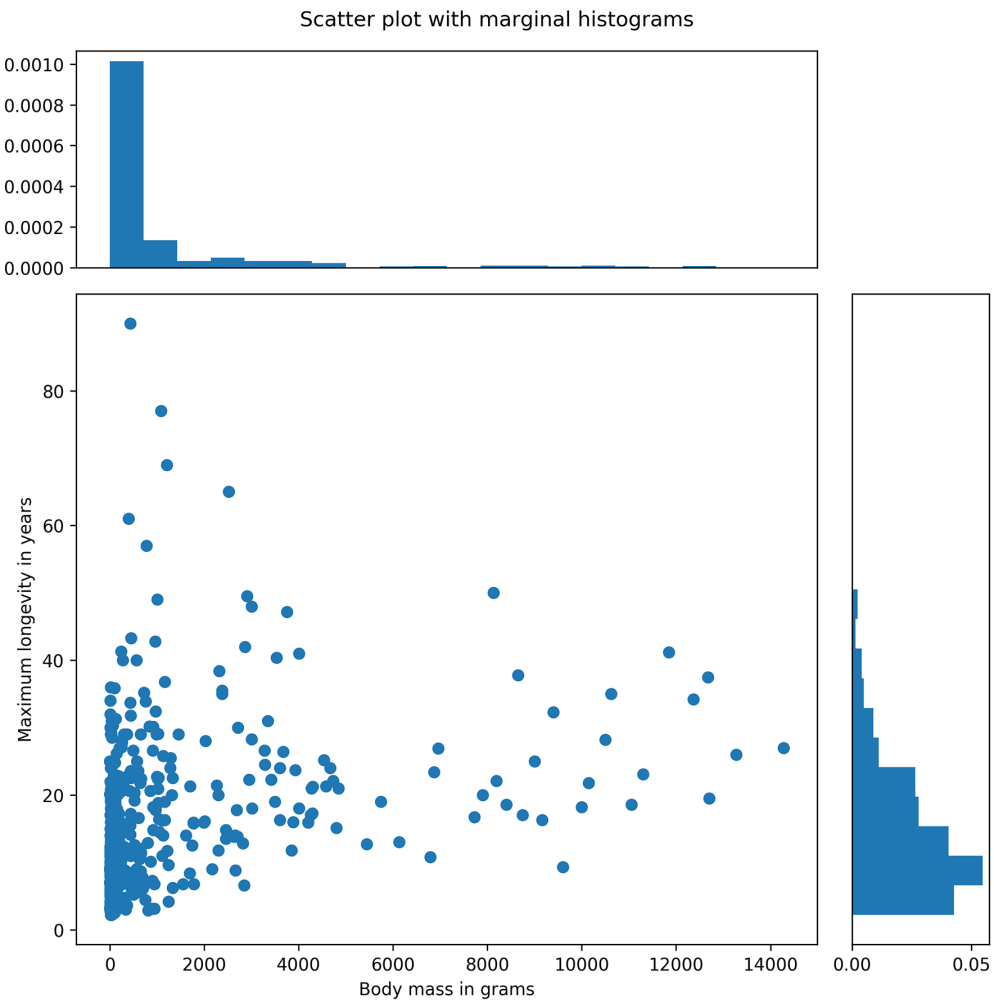

Figure 2.14: Correlation between body mass and maximum longevity of the Aves class with marginal histograms

BUBBLE PLOT

A **bubble plot** extends a scatter plot by introducing a third numerical variable. The value of the variable is represented by the size of the dots. The area of the dots is proportional to the value. A legend is used to link the size of the dot to an actual numerical value.

USE

Bubble plots help to show a correlation between three variables.

EXAMPLE

The following diagram shows a bubble plot that highlights the relationship between heights and age of humans to get the weight of each person, which is represented by the size of the bubble:

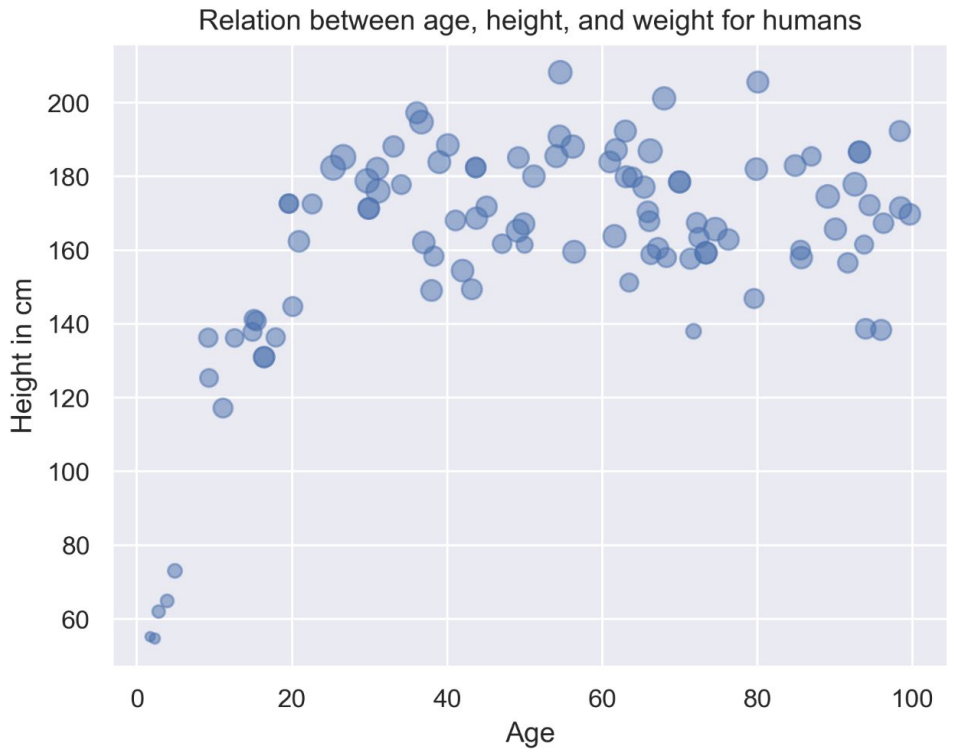

Figure 2.15: Bubble plot showing the relation between height and age of humans

DESIGN PRACTICES

- The design practices for the scatter plot are also applicable to the bubble plot.
- Don't use bubble plots for very large amounts of data, since too many bubbles make the chart difficult to read.

CORRELOGRAM

A **correlogram** is a combination of scatter plots and histograms. Histograms will be discussed in detail later in this chapter. A correlogram or correlation matrix visualizes the relationship between each pair of numerical variables using a scatter plot.

The diagonals of the correlation matrix represent the distribution of each variable in the form of a histogram. You can also plot the relationship between multiple groups or categories using different colors. A correlogram is a great chart for exploratory data analysis to get a feel for your data, especially the correlation between variable pairs.

EXAMPLES

The following diagram shows a correlogram for the height, weight, and age of humans. The diagonal plots show a histogram for each variable. The off-diagonal elements show scatter plots between variable pairs:

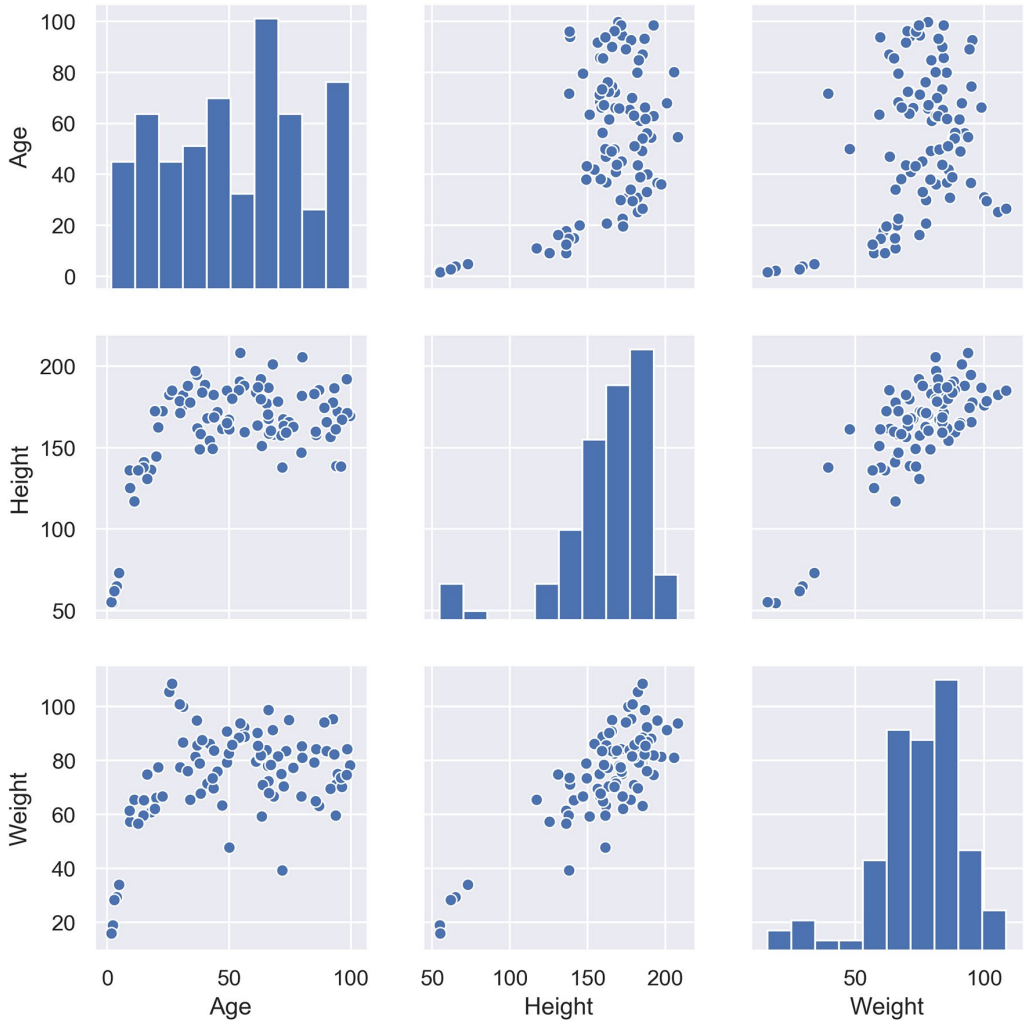

Figure 2.16: Correlogram with a single category

The following diagram shows the correlogram with data samples separated by color into different groups:

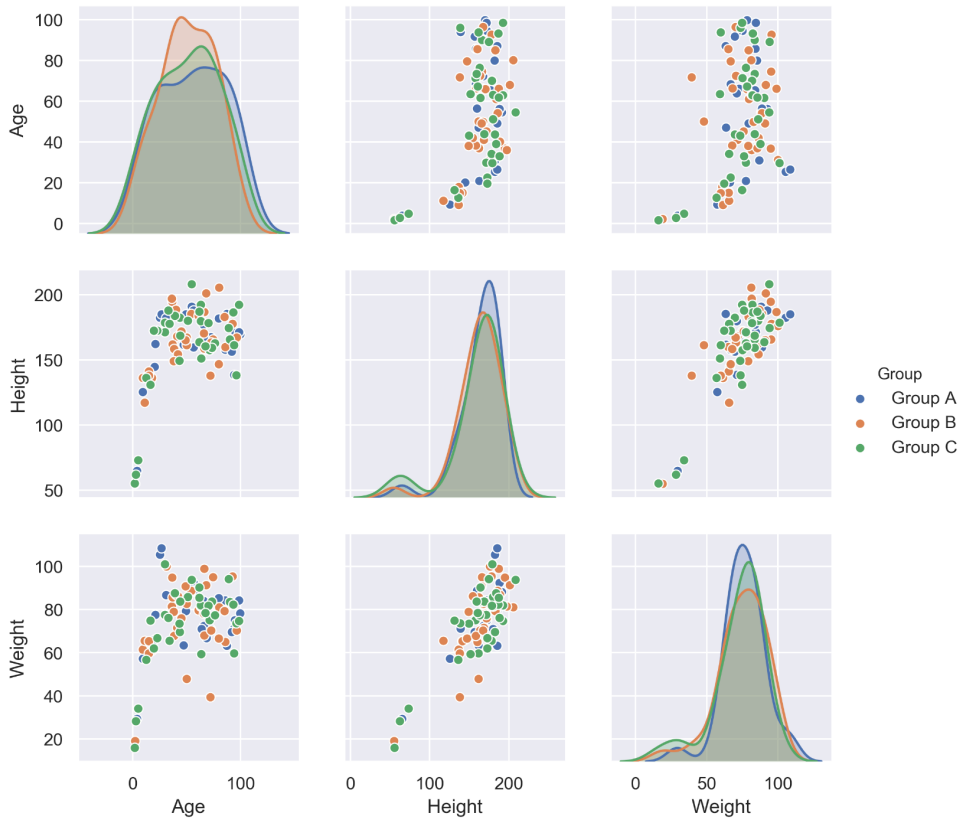

Figure 2.17: Correlogram with multiple categories

DESIGN PRACTICES

- Start both axes at zero to represent data accurately.
- Use contrasting colors for data points and avoid using symbols for scatter plots with multiple groups or categories.

HEATMAP

A **heatmap** is a visualization where values contained in a matrix are represented as colors or color saturation. Heatmaps are great for visualizing multivariate data (data in which analysis is based on more than two variables per observation), where categorical variables are placed in the rows and columns and a numerical or categorical variable is represented as colors or color saturation.

USE

The visualization of multivariate data can be done using heatmaps as they are great for finding patterns in your data.

EXAMPLES

The following diagram shows a heatmap for the most popular products on the electronics category page across various e-commerce websites, where the color shows the number of units sold. In the following diagram, we can analyze that the darker colors represent more units sold, as shown in the key:

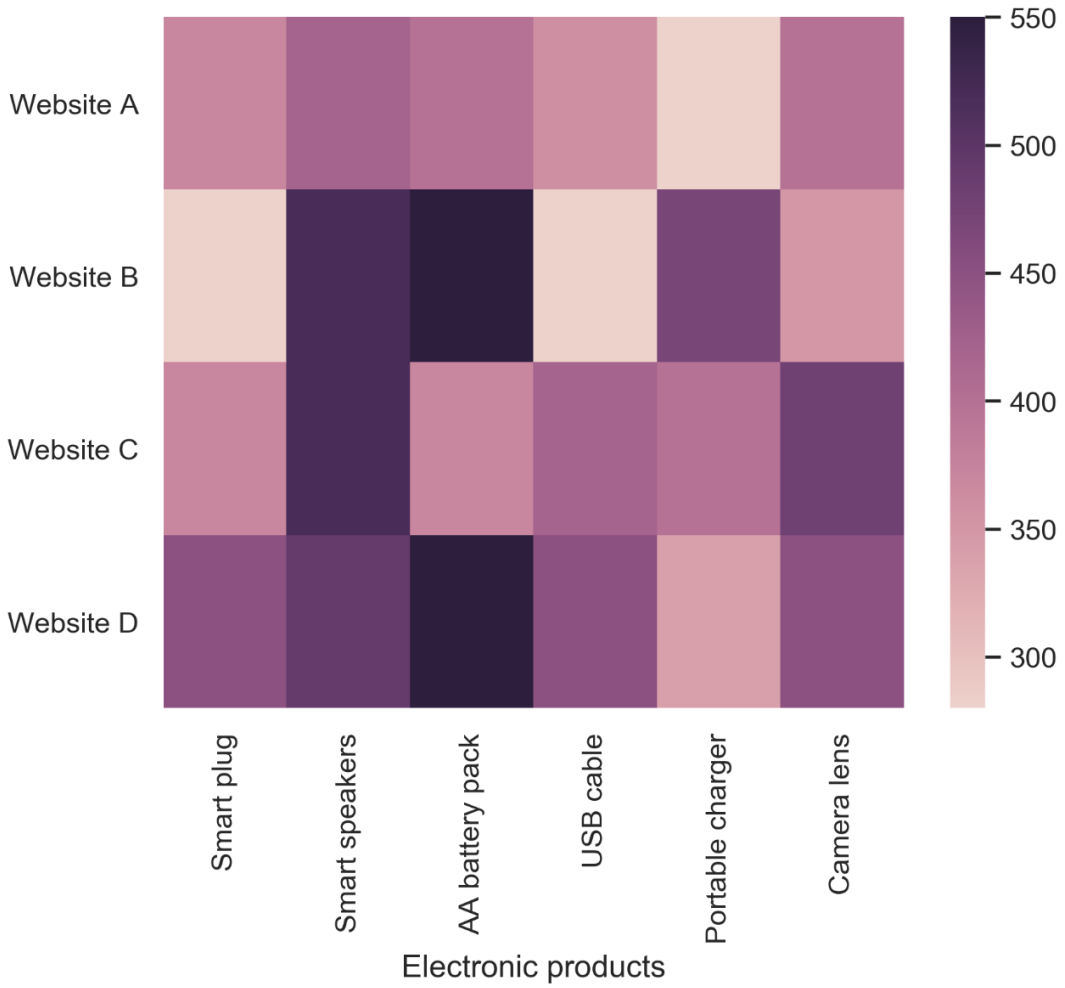

Figure 2.18: Heatmap for popular products in the electronics category

Variants: **Annotated Heatmaps**

Let's see the same example we saw previously in an annotated heatmap, where the color shows the number of units sold:

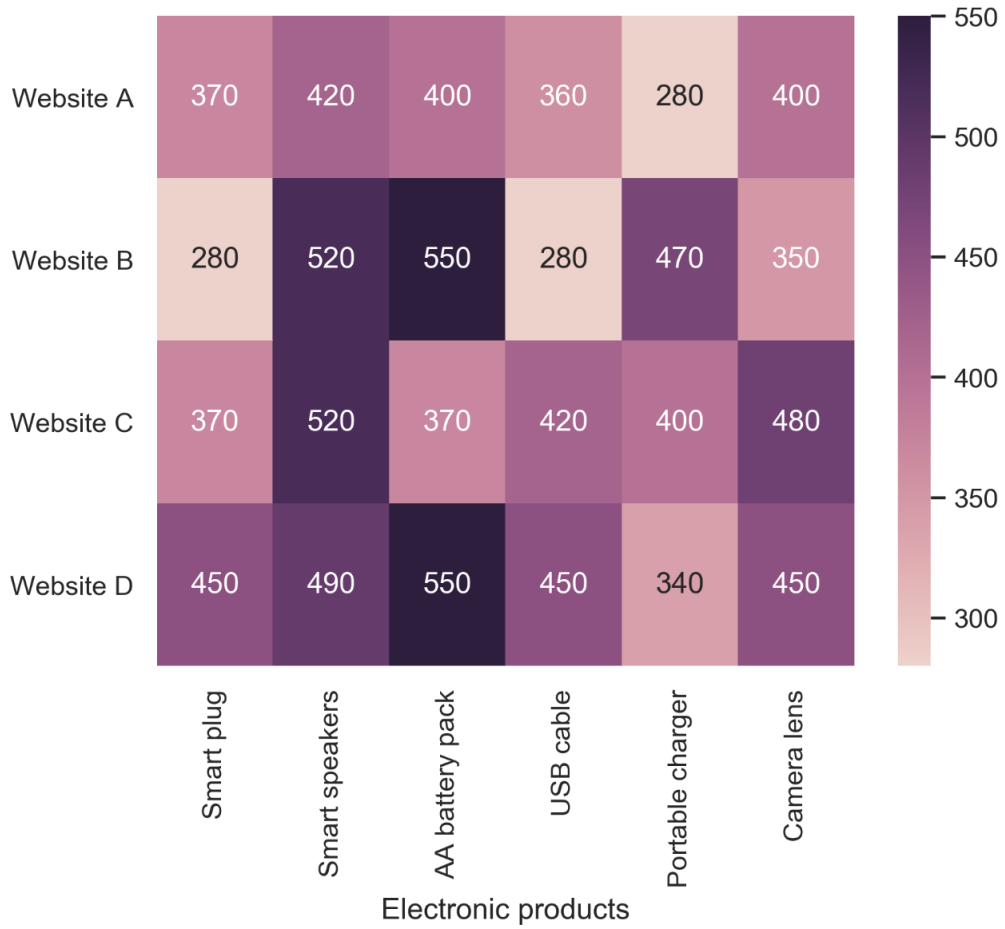

Figure 2.19: Annotated heatmap for popular products in the electronics category

DESIGN PRACTICE

- Select colors and contrasts that will be easily visible to individuals with vision problems so that your plots are more inclusive.

In this section, we introduced various plots for relating a variable to other variables and looked at their uses, and multiple examples for the different relation plots were given. The following activity will give you some practice in working with heatmaps.

ACTIVITY 2.02: ROAD ACCIDENTS OCCURRING OVER TWO DECADES

You are given a diagram that provides information about the road accidents that have occurred over the past two decades during the months of January, April, July, and October. The aim of this activity is to understand how you can use heatmaps to visualize multivariate data.

1. Identify the two years during which the number of road accidents occurring was the least.
2. For the past two decades, identify the month for which accidents showed a marked decrease:

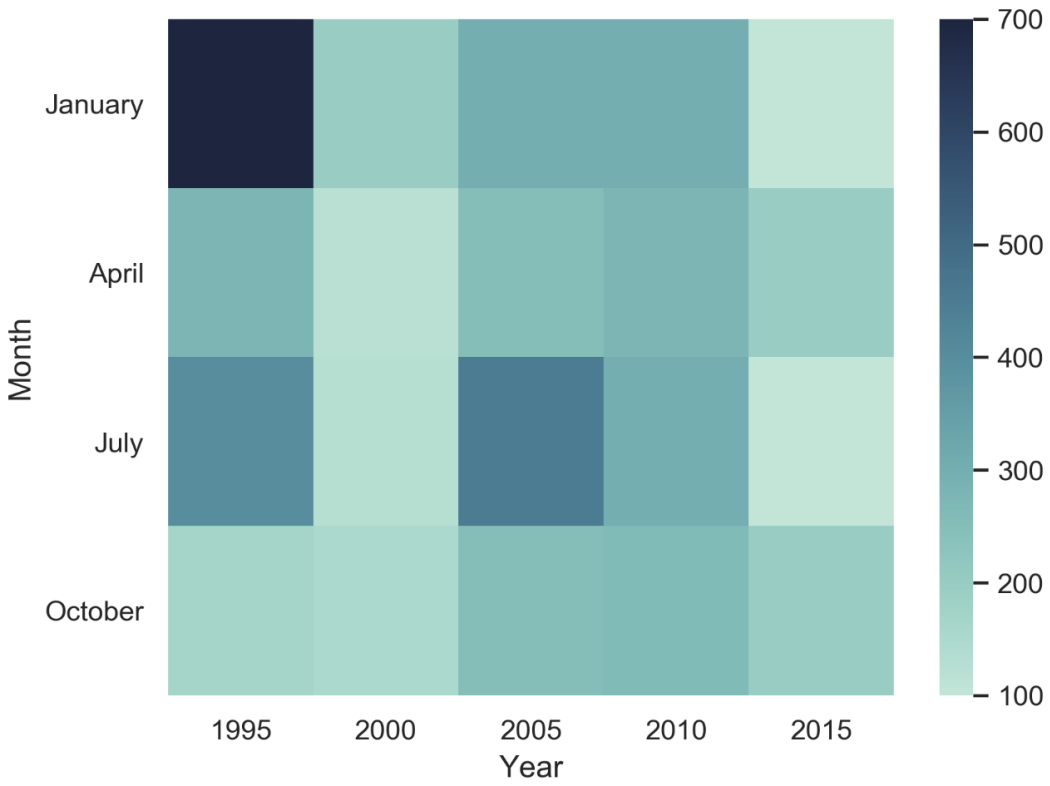

Figure 2.20: Total accidents over 20 years

> **NOTE**
>
> The solution for this activity can be found via this link.

COMPOSITION PLOTS

Composition plots are ideal if you think about something as a part of a whole. For static data, you can use pie charts, stacked bar charts, or Venn diagrams. **Pie charts** or **donut charts** help show proportions and percentages for groups. If you need an additional dimension, stacked bar charts are great. Venn diagrams are the best way to visualize overlapping groups, where each group is represented by a circle. For data that changes over time, you can use either stacked bar charts or stacked area charts.

PIE CHART

Pie charts illustrate numerical proportions by dividing a circle into slices. Each arc length represents a proportion of a category. The full circle equates to 100%. For humans, it is easier to compare bars than arc lengths; therefore, it is recommended to use bar charts or stacked bar charts the majority of the time.

USE

To compare items that are part of a whole.

EXAMPLES

The following diagram shows household water usage around the world:

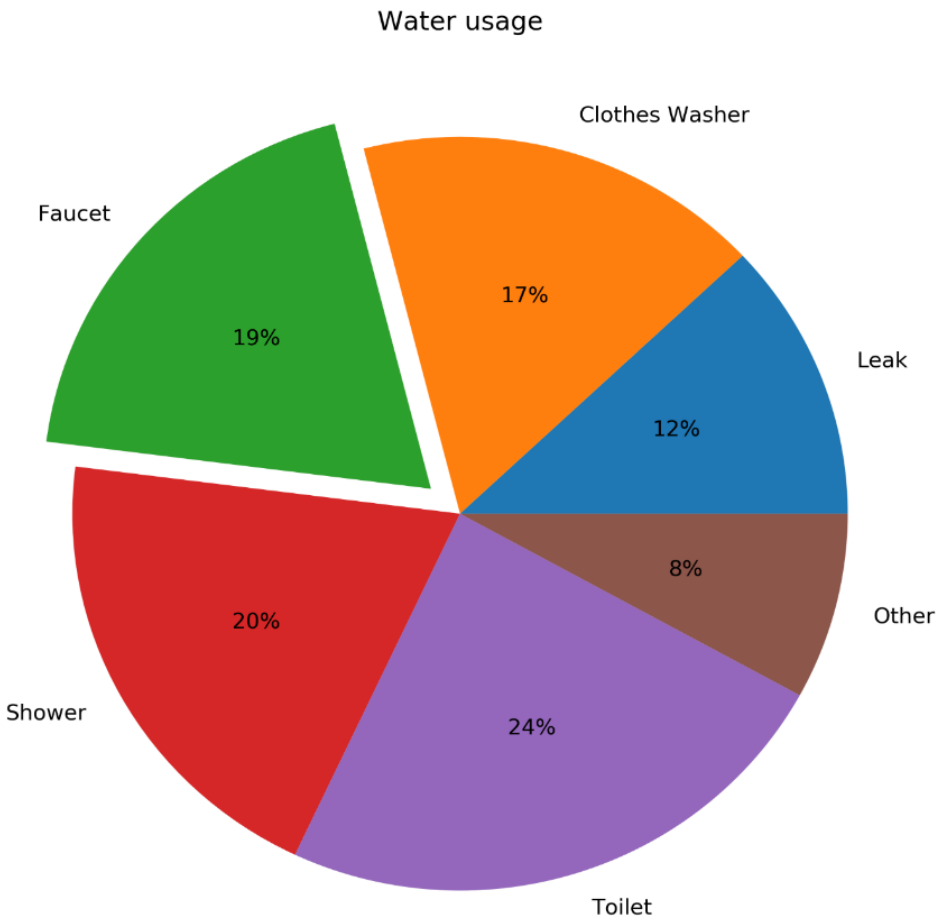

Figure 2.21: Pie chart for global household water usage

DESIGN PRACTICES

- Arrange the slices according to their size in increasing/decreasing order, either in a clockwise or counterclockwise manner.
- Make sure that every slice has a different color.

VARIANTS: DONUT CHART

An alternative to a pie chart is a **donut chart**. In contrast to pie charts, it is easier to compare the size of slices, since the reader focuses more on reading the length of the arcs instead of the area. Donut charts are also more space-efficient because the center is cut out, so it can be used to display information or further divide groups into subgroups.

The following diagram shows a basic donut chart:

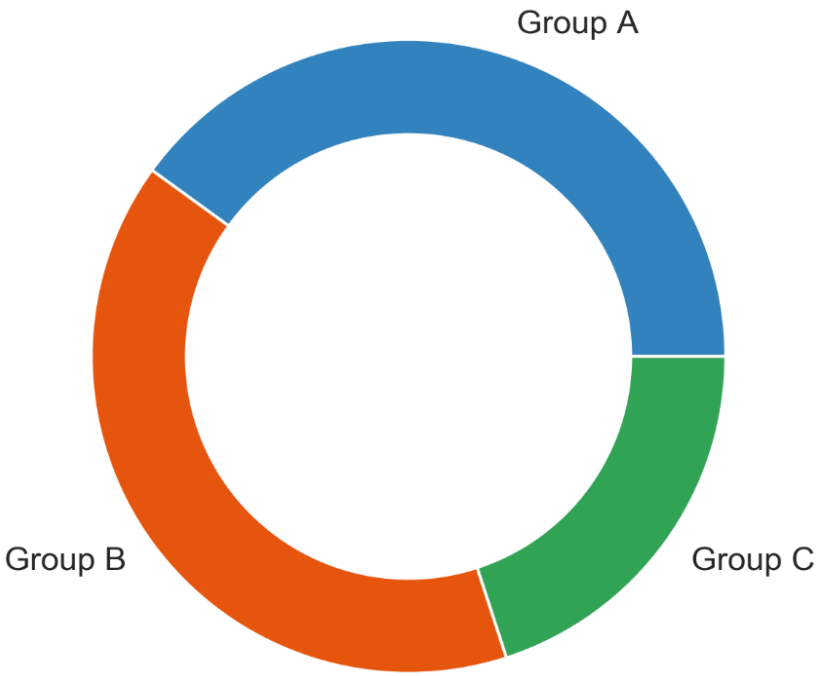

Figure 2.22: Donut chart

The following diagram shows a donut chart with subgroups:

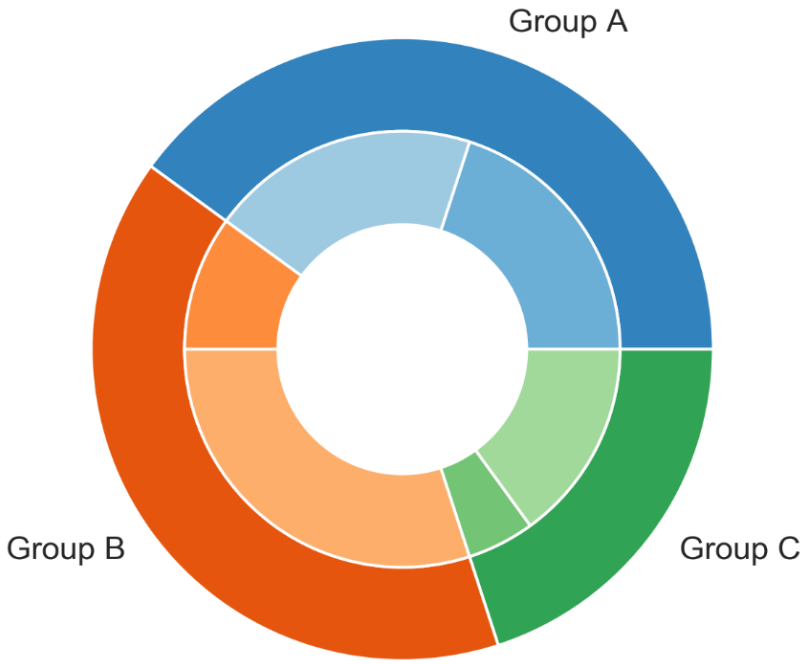

Figure 2.23: Donut chart with subgroups

DESIGN PRACTICE

- Use the same color that's used for the category for the subcategories. Use varying brightness levels for the different subcategories.

STACKED BAR CHART

Stacked bar charts are used to show how a category is divided into subcategories and the proportion of the subcategory in comparison to the overall category. You can either compare total amounts across each bar or show a percentage of each group. The latter is also referred to as a **100% stacked bar chart** and makes it easier to see relative differences between quantities in each group.

USE

- To compare variables that can be divided into sub-variables

EXAMPLES

The following diagram shows a generic stacked bar chart with five groups:

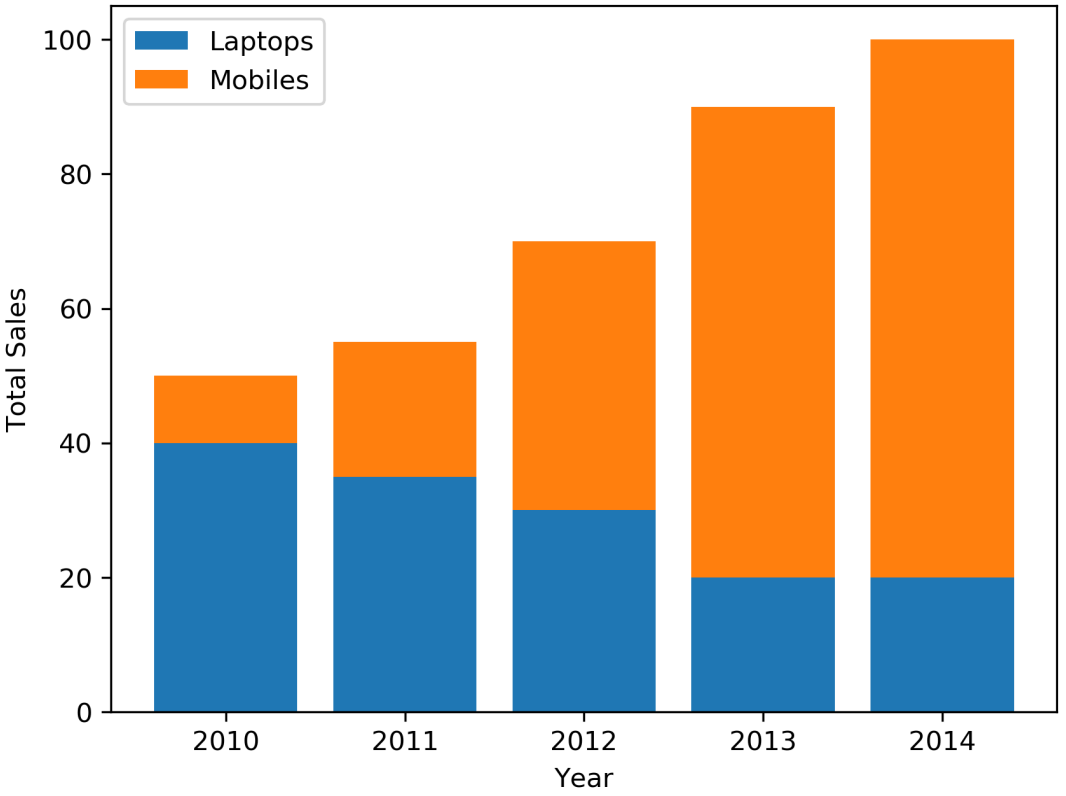

Figure 2.24: Stacked bar chart to show sales of laptops and mobiles

The following diagram shows a 100% stacked bar chart with the same data that was used in the preceding diagram:

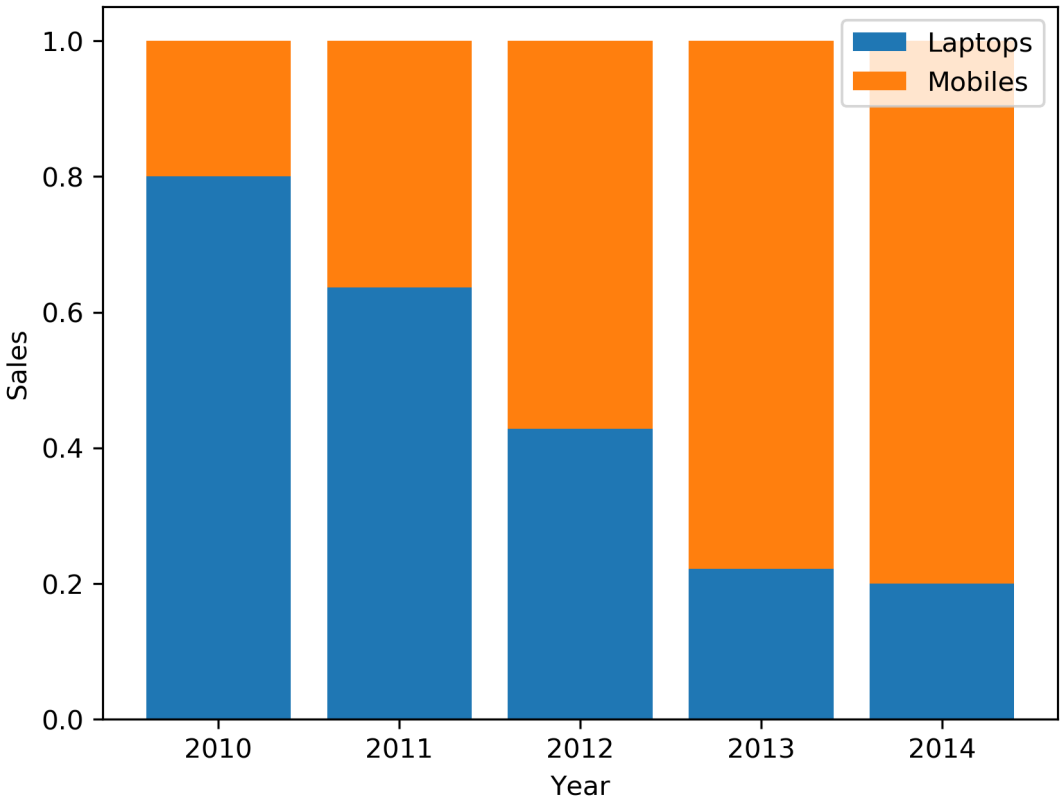

Figure 2.25: 100% stacked bar chart to show sales of laptops, PCs, and mobiles

The following diagram illustrates the daily total sales of a restaurant over several days. The daily total sales of non-smokers are stacked on top of the daily total sales of smokers:

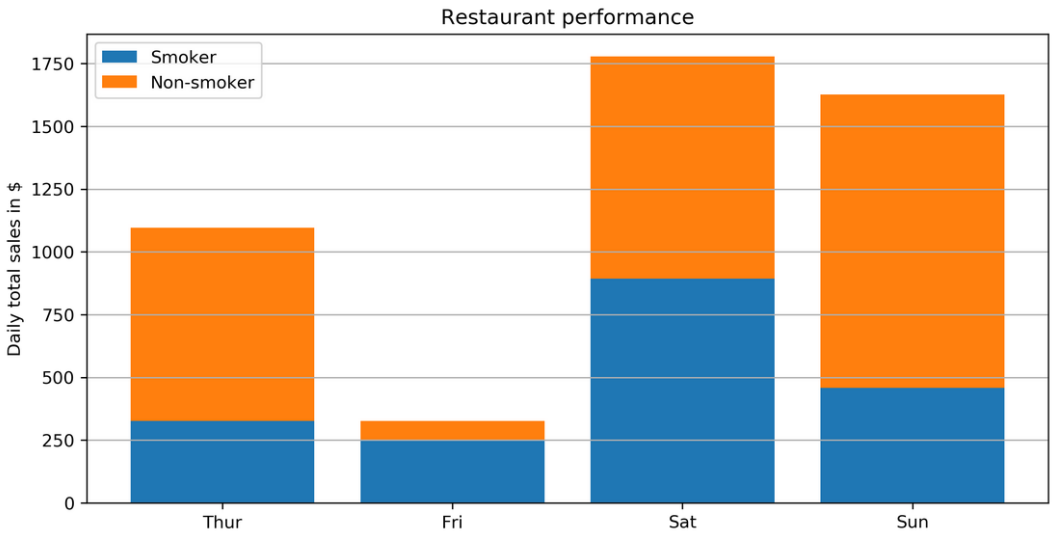

Figure 2.26: Daily total restaurant sales categorized by smokers and non-smokers

DESIGN PRACTICES

- Use contrasting colors for stacked bars.
- Ensure that the bars are adequately spaced to eliminate visual clutter. The ideal space guideline between each bar is half the width of a bar.
- Categorize data alphabetically, sequentially, or by value, to uniformly order it and make things easier for your audience.

STACKED AREA CHART

Stacked area charts show trends for part-of-a-whole relations. The values of several groups are illustrated by stacking individual area charts on top of one another. It helps to analyze both individual and overall trend information.

USE

To show trends for time series that are part of a whole.

EXAMPLES

The following diagram shows a stacked area chart with the net profits of Google, Facebook, Twitter, and Snapchat over a decade:

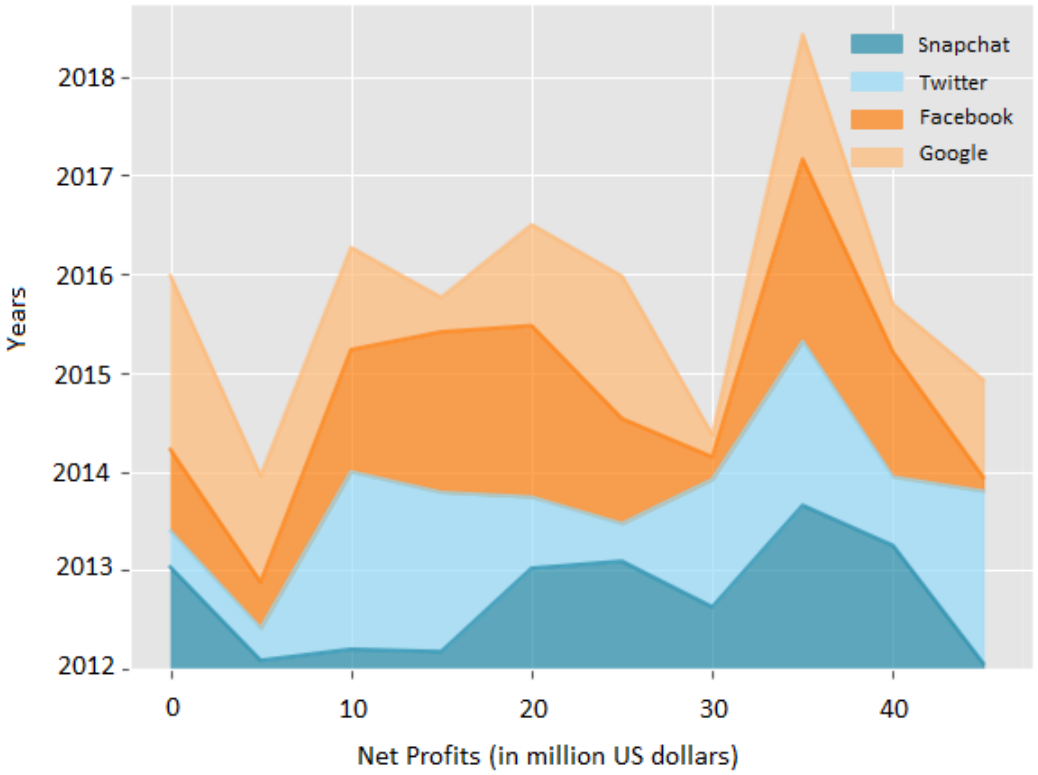

Figure 2.27: Stacked area chart to show net profits of four companies

DESIGN PRACTICE

- Use transparent colors to improve information visibility. This will help you to analyze the overlapping data and you will also be able to see the grid lines.

In this section, we covered various composition plots and we will conclude this section with the following activity.

ACTIVITY 2.03: SMARTPHONE SALES UNITS

You want to compare smartphone sales units for the five biggest smartphone manufacturers over time and see whether there is any trend. In this activity, we also want to look at the advantages and disadvantages of stacked area charts compared to line charts:

1. Looking at the following line chart, analyze the sales of each manufacturer and identify the one whose fourth-quarter performance is exceptional when compared to the third quarter.
2. Analyze the performance of all manufacturers and make a prediction about two companies whose sales units will show a downward and an upward trend:

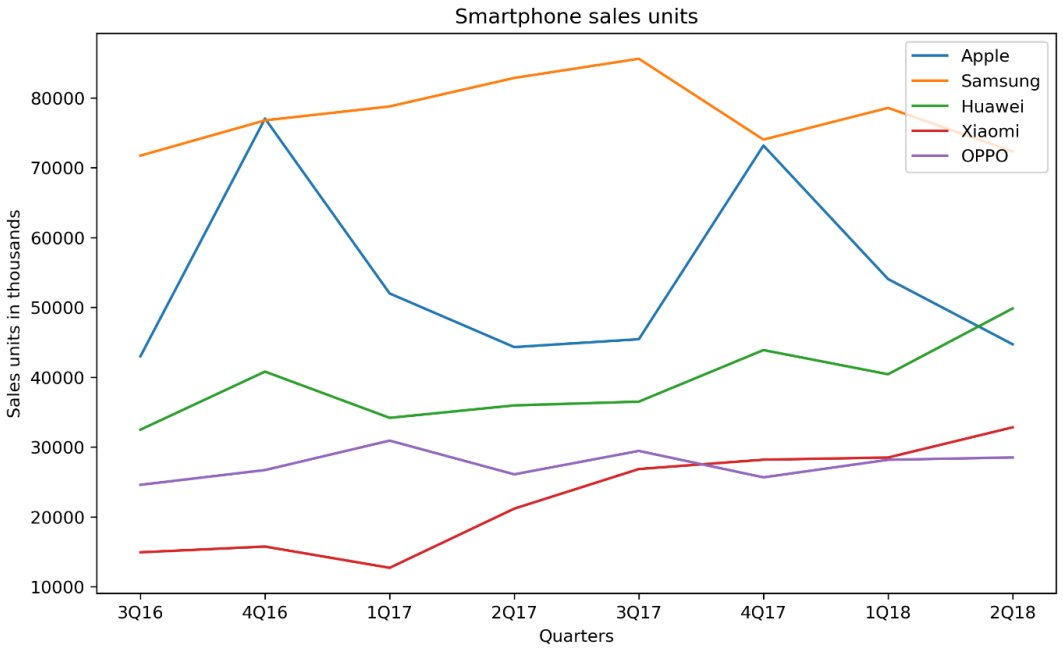

Figure 2.28: Line chart of smartphone sales units

3. What would be the advantages and disadvantages of using a stacked area chart instead of a line chart?

> **NOTE**
>
> The solution for this activity can be found via this link.

VENN DIAGRAM

Venn diagrams, also known as **set diagrams**, show all possible logical relations between a finite collection of different sets. Each set is represented by a circle. The circle size illustrates the importance of a group. The size of overlap represents the intersection between multiple groups.

USE

To show overlaps for different sets.

EXAMPLE

Visualizing the intersection of the following diagram shows a Venn diagram for students in two groups taking the same class in a semester:

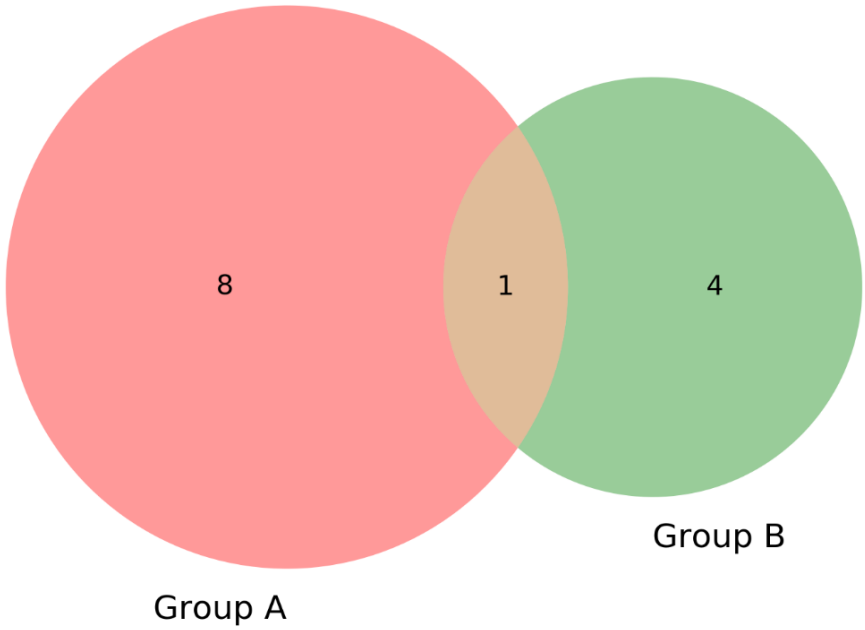

Figure 2.29: Venn diagram showing students taking the same class

From the preceding diagram, we can note that there are eight students in just group A, four students in just group B, and one student in both groups.

DESIGN PRACTICE

- It is not recommended to use Venn diagrams if you have more than three groups. It would become difficult to understand.

Moving on from composition plots, we will cover distribution plots in the following section.

DISTRIBUTION PLOTS

Distribution plots give a deep insight into how your data is distributed. For a single variable, a histogram is effective. For multiple variables, you can either use a box plot or a violin plot. The violin plot visualizes the densities of your variables, whereas the box plot just visualizes the median, the interquartile range, and the range for each variable.

HISTOGRAM

A **histogram** visualizes the distribution of a single numerical variable. Each bar represents the frequency for a certain interval. Histograms help get an estimate of statistical measures. You see where values are concentrated, and you can easily detect outliers. You can either plot a histogram with absolute frequency values or, alternatively, normalize your histogram. If you want to compare distributions of multiple variables, you can use different colors for the bars.

USE

Get insights into the underlying distribution for a dataset.

EXAMPLE

The following diagram shows the distribution of the **Intelligence Quotient** (**IQ**) for a test group. The dashed lines represent the standard deviation each side of the mean (the solid line):

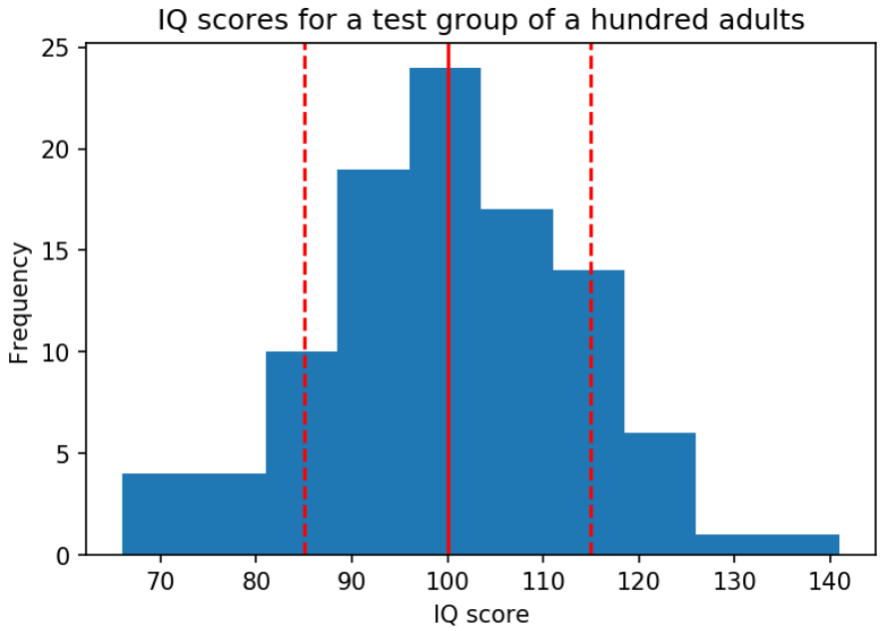

Figure 2.30: Distribution of IQ for a test group of a hundred adults

DESIGN PRACTICE

- Try different numbers of bins (data intervals), since the shape of the histogram can vary significantly.

DENSITY PLOT

A **density plot** shows the distribution of a numerical variable. It is a variation of a histogram that uses **kernel smoothing**, allowing for smoother distributions. One advantage these have over histograms is that density plots are better at determining the distribution shape since the distribution shape for histograms heavily depends on the number of bins (data intervals).

USE

To compare the distribution of several variables by plotting the density on the same axis and using different colors.

EXAMPLE

The following diagram shows a basic density plot:

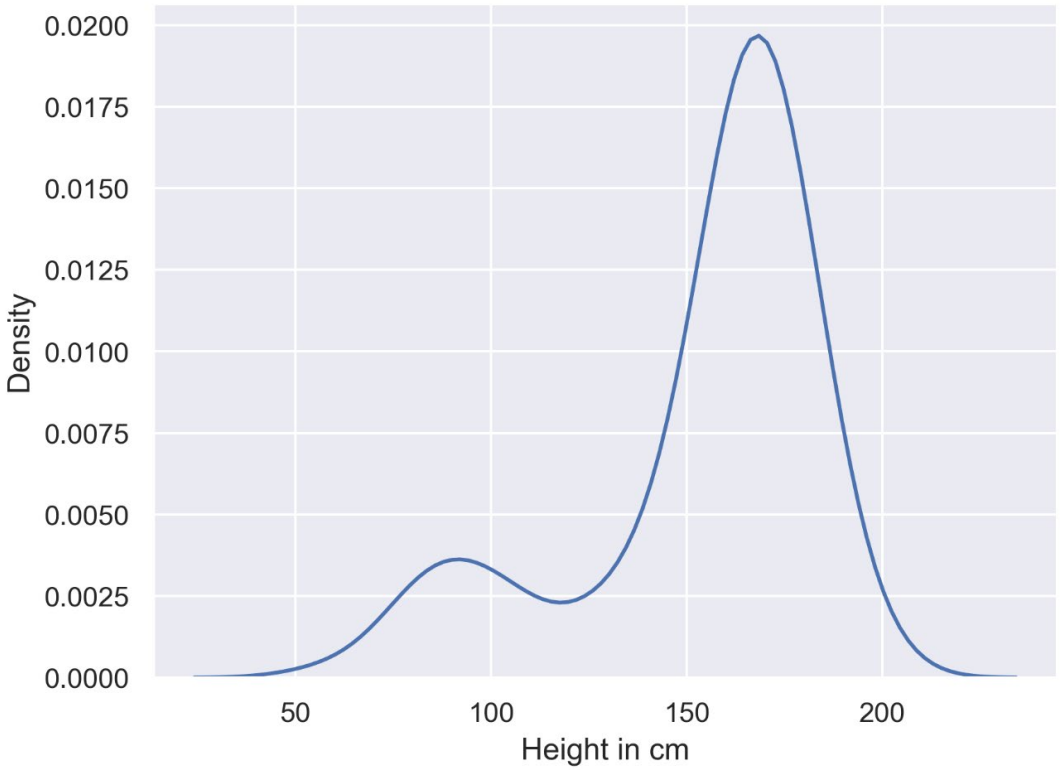

Figure 2.31: Density plot

The following diagram shows a basic multi-density plot:

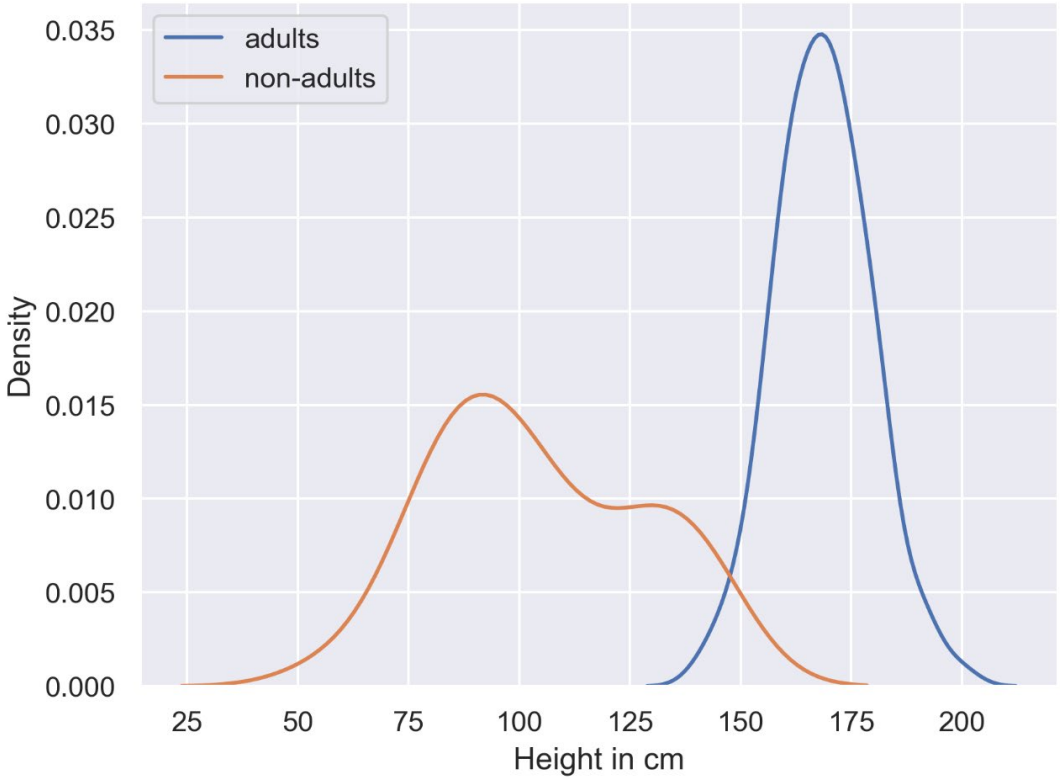

Figure 2.32: Multi-density plot

DESIGN PRACTICE

- Use contrasting colors to plot the density of multiple variables.

BOX PLOT

The **box plot** shows multiple statistical measurements. The box extends from the lower to the upper quartile values of the data, thus allowing us to visualize the interquartile range (IQR). The horizontal line within the box denotes the median. The parallel extending lines from the boxes are called **whiskers**; they indicate the variability outside the lower and upper quartiles. There is also an option to show data **outliers**, usually as circles or diamonds, past the end of the whiskers.

USE

Compare statistical measures for multiple variables or groups.

EXAMPLES

The following diagram shows a basic box plot that shows the height of a group of people:

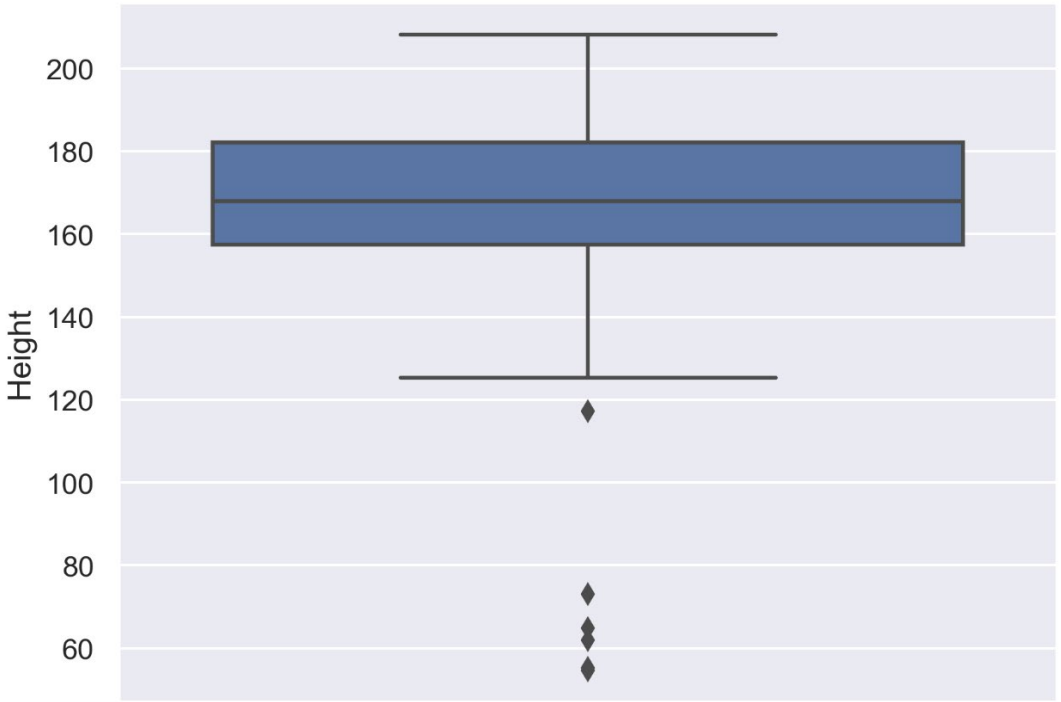

Figure 2.33: Box plot showing a single variable

The following diagram shows a basic box plot for multiple variables. In this case, it shows heights for two different groups – adults and non-adults:

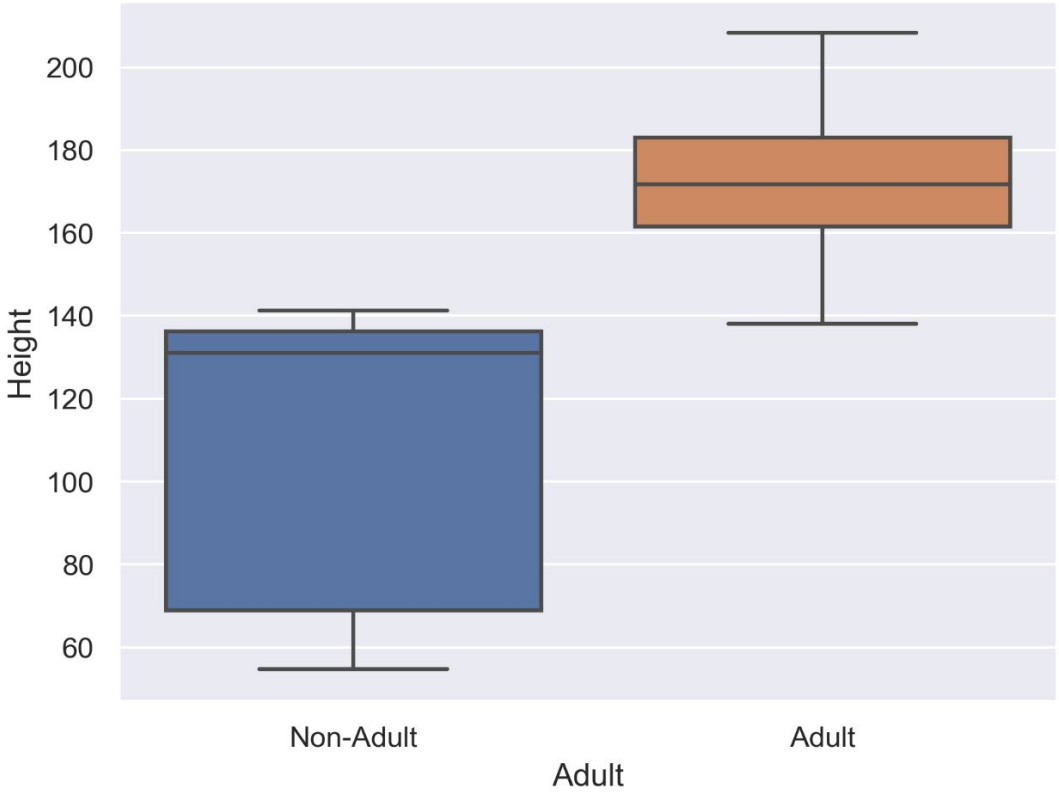

Figure 2.34: Box plot for multiple variables

In the next section, we will learn what the features, uses, and best practices are of the violin plot.

VIOLIN PLOT

Violin plots are a combination of box plots and density plots. Both the statistical measures and the distribution are visualized. The thick black bar in the center represents the interquartile range, while the thin black line corresponds to the whiskers in a box plot. The white dot indicates the median. On both sides of the centerline, the density is visualized.

USE

Compare statistical measures and density for multiple variables or groups.

EXAMPLES

The following diagram shows a violin plot for a single variable and shows how students have performed in **Math**:

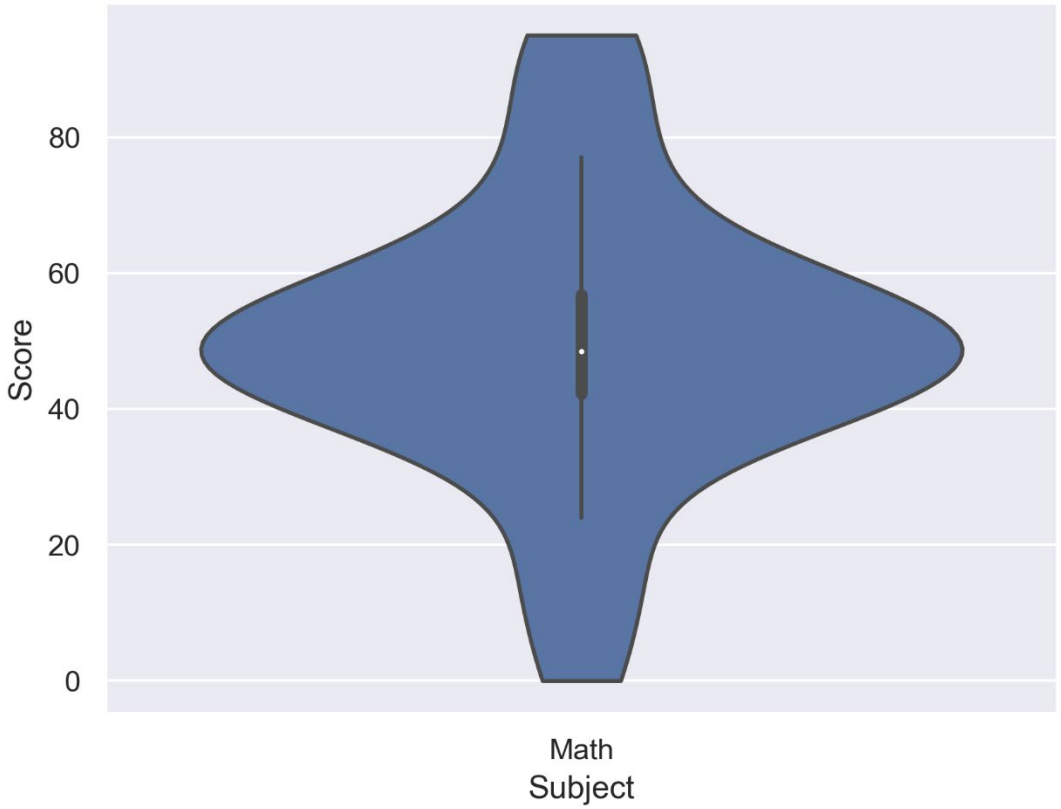

Figure 2.35: Violin plot for a single variable (Math)

From the preceding diagram, we can analyze that most of the students have scored around 40-60 in the **Math** test.

The following diagram shows a violin plot for two variables and shows the performance of students in **English** and **Math**:

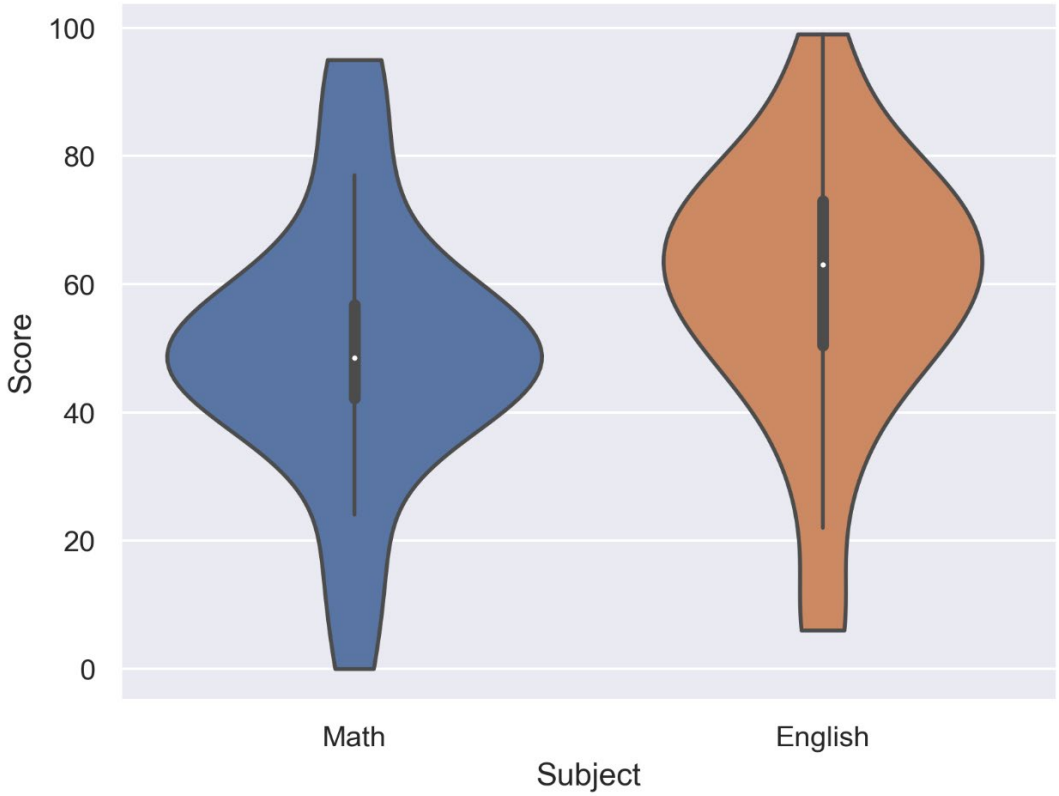

Figure 2.36: Violin plot for multiple variables (English and Math)

From the preceding diagram, we can say that on average, the students have scored more in **English** than in **Math**, but the highest score was secured in **Math**.

The following diagram shows a violin plot for a single variable divided into three groups, and shows the performance of three divisions of students in **English** based on their score:

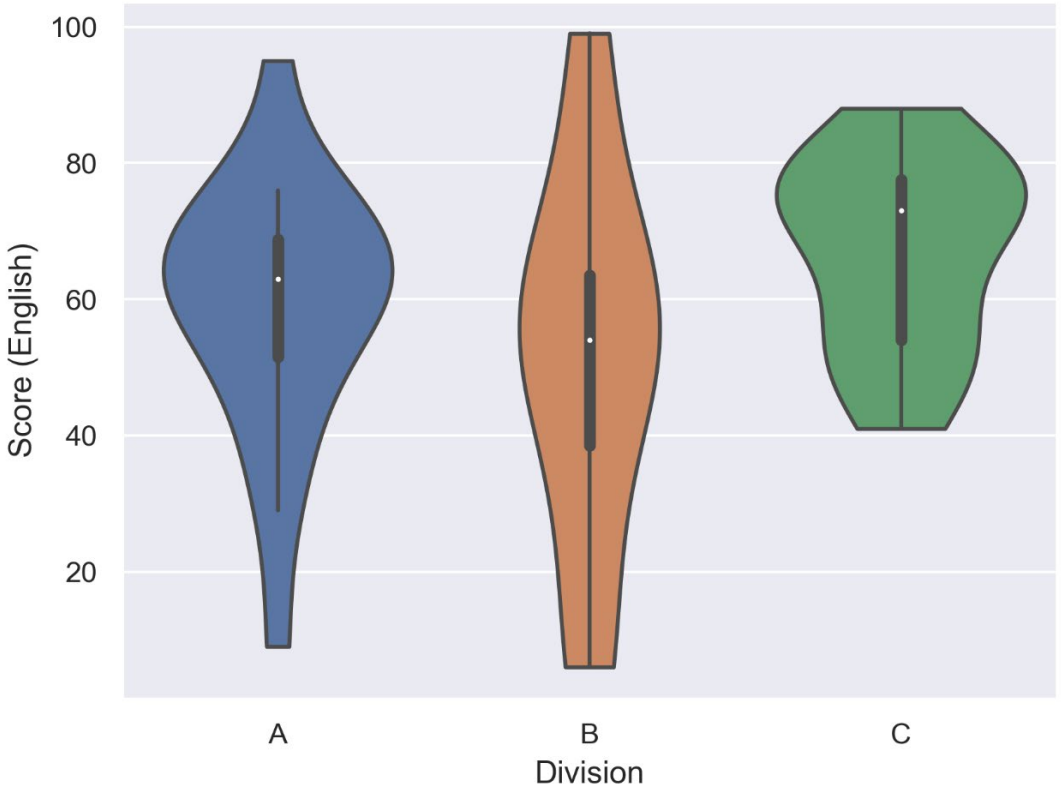

Figure 2.37: Violin plot with multiple categories (three groups of students)

From the preceding diagram, we can note that on average, division C has scored the highest, division B has scored the lowest, and division A is, on average, in between divisions B and C.

DESIGN PRACTICE

- Scale the axes accordingly so that the distribution is clearly visible and not flat.

In this section, distribution plots were introduced. In the following activity, we will have a closer look at histograms.

ACTIVITY 2.04: FREQUENCY OF TRAINS DURING DIFFERENT TIME INTERVALS

You are provided with a histogram that states the number of trains arriving at different time intervals in the afternoon to determine the maximum number of trains arriving in 2-hour time intervals. The goal of this activity is to gain a deeper insight into histograms:

1. Looking at the following histogram, can you identify the interval during which a maximum number of trains arrive?
2. How would the histogram change if in the morning, the same total number of trains arrive as in the afternoon, and if you have the same frequencies for all time intervals?

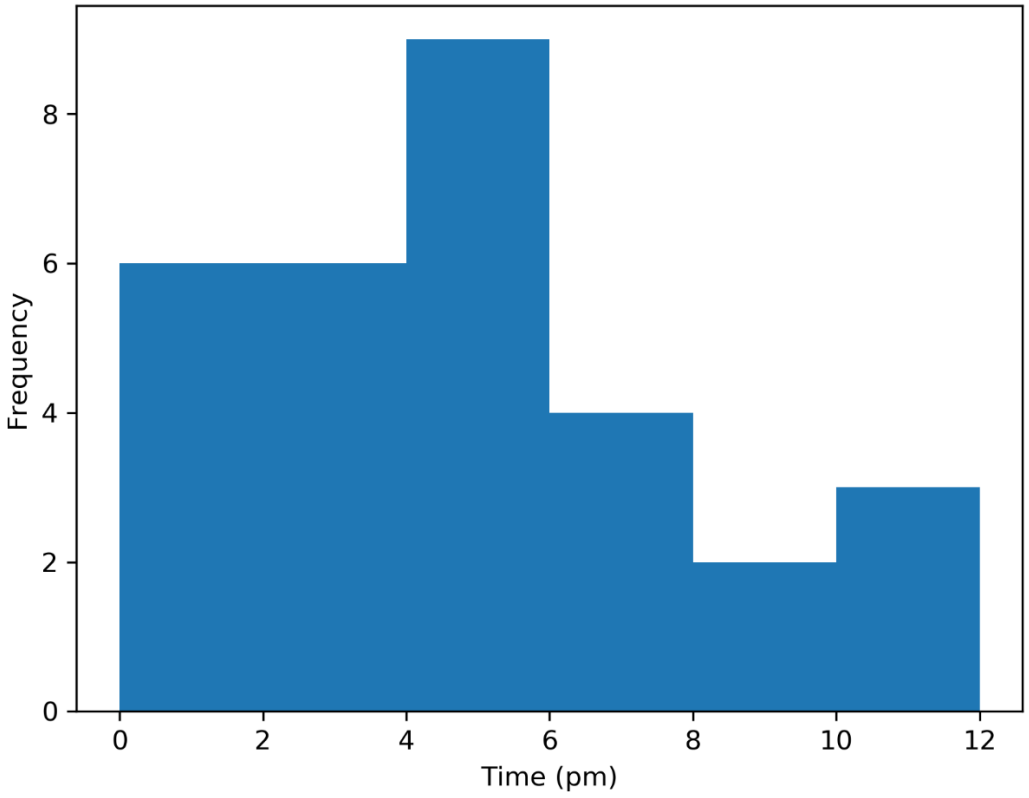

Figure 2.38: Frequency of trains during different time intervals

> **NOTE**
>
> The solution for this activity can be found via this link.

With that activity, we conclude the section about distribution plots and we will introduce geoplots in the next section.

GEOPLOTS

Geological plots are a great way to visualize geospatial data. Choropleth maps can be used to compare quantitative values for different countries, states, and so on. If you want to show connections between different locations, connection maps are the way to go.

DOT MAP

In a **dot map**, each dot represents a certain number of observations. Each dot has the same size and value (the number of observations each dot represents). The dots are not meant to be counted; they are only intended to give an impression of magnitude. The size and value are important factors for the effectiveness and impression of the visualization. You can use different colors or symbols for the dots to show multiple categories or groups.

USE

To visualize geospatial data.

EXAMPLE

The following diagram shows a dot map where each dot represents a certain amount of bus stops throughout the world:

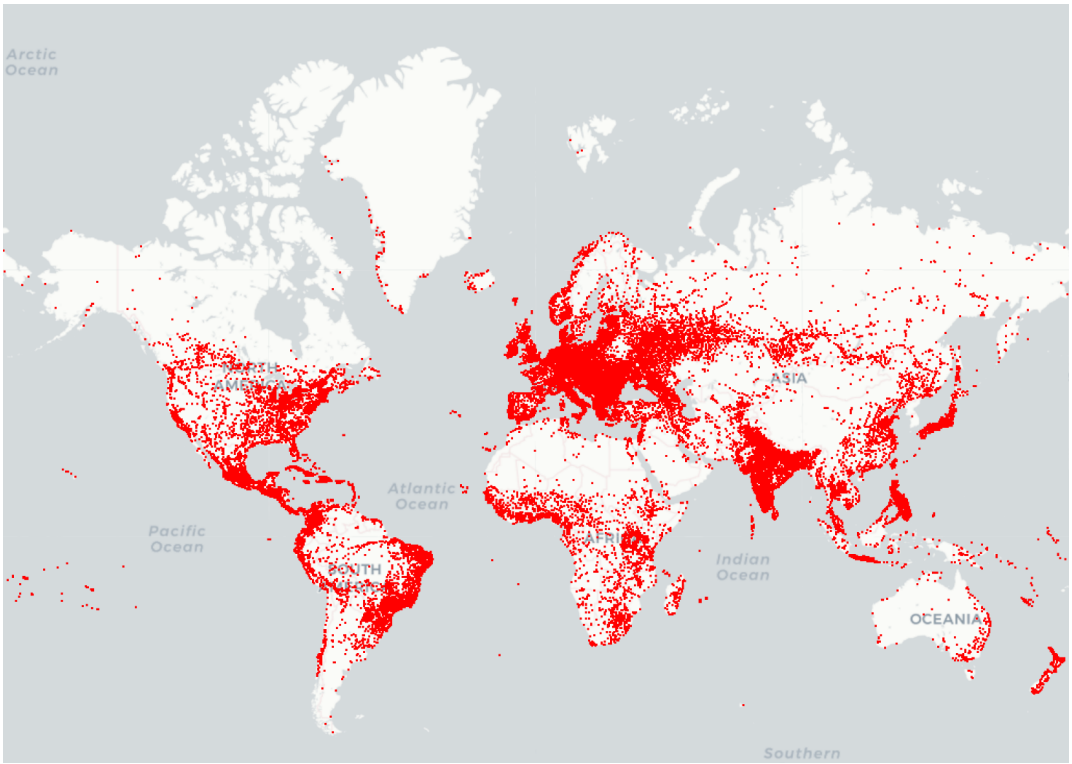

Figure 2.39: Dot map showing bus stops worldwide

DESIGN PRACTICES

- Do not show too many locations. You should still be able to see the map to get a feel for the actual location.
- Choose a dot size and value so that in dense areas, the dots start to blend. The dot map should give a good impression of the underlying spatial distribution.

CHOROPLETH MAP

In a **choropleth map**, each tile is colored to encode a variable. For example, a tile represents a geographic region for counties and countries. Choropleth maps provide a good way to show how a variable varies across a geographic area. One thing to keep in mind for choropleth maps is that the human eye naturally gives more attention to larger areas, so you might want to normalize your data by dividing the map area-wise.

USE

To visualize geospatial data grouped into geological regions—for example, states or countries.

EXAMPLE

The following diagram shows a choropleth map of a weather forecast in the USA:

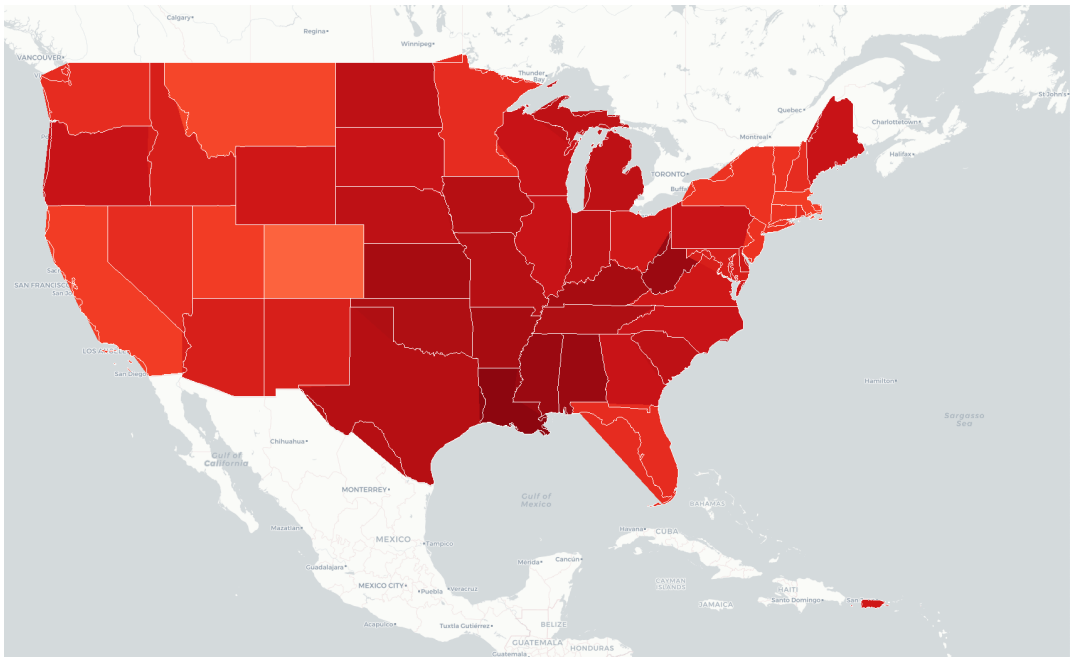

Figure 2.40: Choropleth map showing a weather forecast for the USA

DESIGN PRACTICES

- Use darker colors for higher values, as they are perceived as being higher in magnitude.
- Limit the color gradation, since the human eye is limited in how many colors it can easily distinguish between. Seven color gradations should be enough.

CONNECTION MAP

In a **connection map**, each line represents a certain number of connections between two locations. The link between the locations can be drawn with a straight or rounded line, representing the shortest distance between them.

Each line has the same thickness and value (the number of connections each line represents). The lines are not meant to be counted; they are only intended to give an impression of magnitude. The size and value of a connection line are important factors for the effectiveness and impression of the visualization.

You can use different colors for the lines to show multiple categories or groups, or you can use a colormap to encode the length of the connection.

USE

To visualize connections.

EXAMPLES

The following diagram shows a connection map of flight connections around the world:

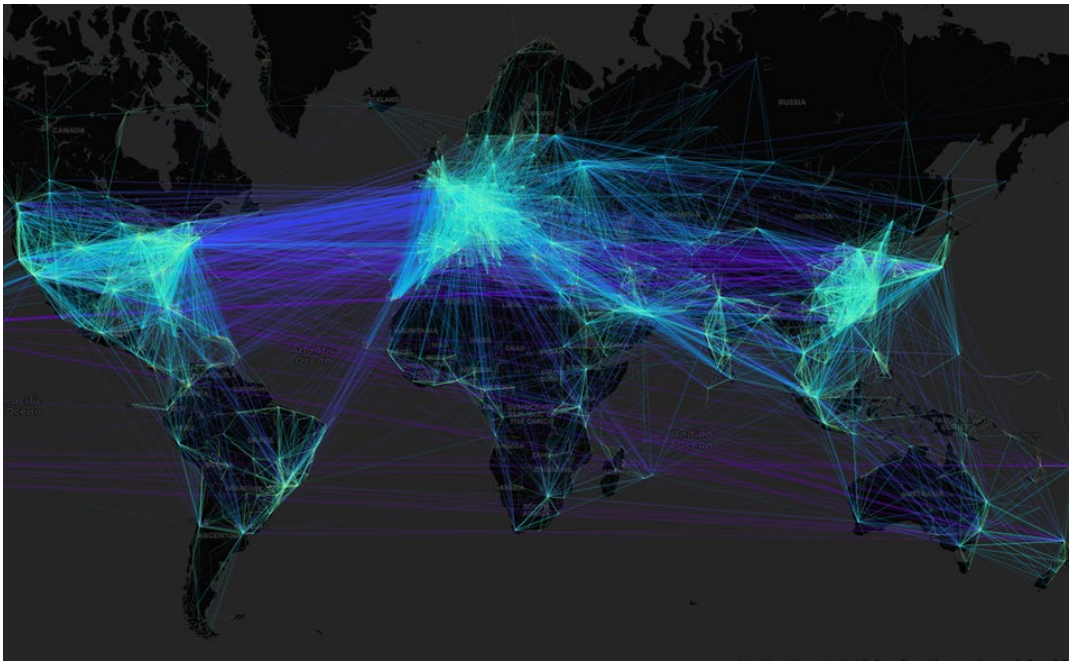

Figure 2.41: Connection map showing flight connections around the world

DESIGN PRACTICES

- Do not show too many connections as it will be difficult for you to analyze the data. You should still see the map to get a feel for the actual locations of the start and end points.
- Choose a line thickness and value so that the lines start to blend in dense areas. The connection map should give a good impression of the underlying spatial distribution.

Geoplots are special plots that are great for visualizing geospatial data. In the following section, we want to briefly talk about what's generally important when it comes to creating good visualizations.

WHAT MAKES A GOOD VISUALIZATION?

There are multiple aspects to what makes a good visualization:

- Most importantly, the visualization should be self-explanatory and visually appealing. To make it self-explanatory, use a legend, descriptive labels for your x-axis and y-axis, and titles.
- A visualization should tell a story and be designed for your audience. Before creating your visualization, think about your target audience; create simple visualizations for a non-specialist audience and more technical detailed visualizations for a specialist audience. Think about a story to tell with your visualization so that your visualization leaves an impression on the audience.

COMMON DESIGN PRACTICES

- Use colors to differentiate variables/subjects rather than symbols, as colors are more perceptible.
- To show additional variables on a 2D plot, use color, shape, and size.
- Keep it simple and don't overload the visualization with too much information.

ACTIVITY 2.05: ANALYZING VISUALIZATIONS

The following visualizations are not ideal as they do not represent data well. Answer the following questions for each visualization. The aim of this activity is to sharpen your skills with regard to choosing the best suitable plot for a scenario.

1. What are the bad aspects of these visualizations?
2. How could we improve the visualizations? Sketch the right visualization for both scenarios.

The first visualization is supposed to illustrate the top 30 YouTube music channels according to their number of subscribers:

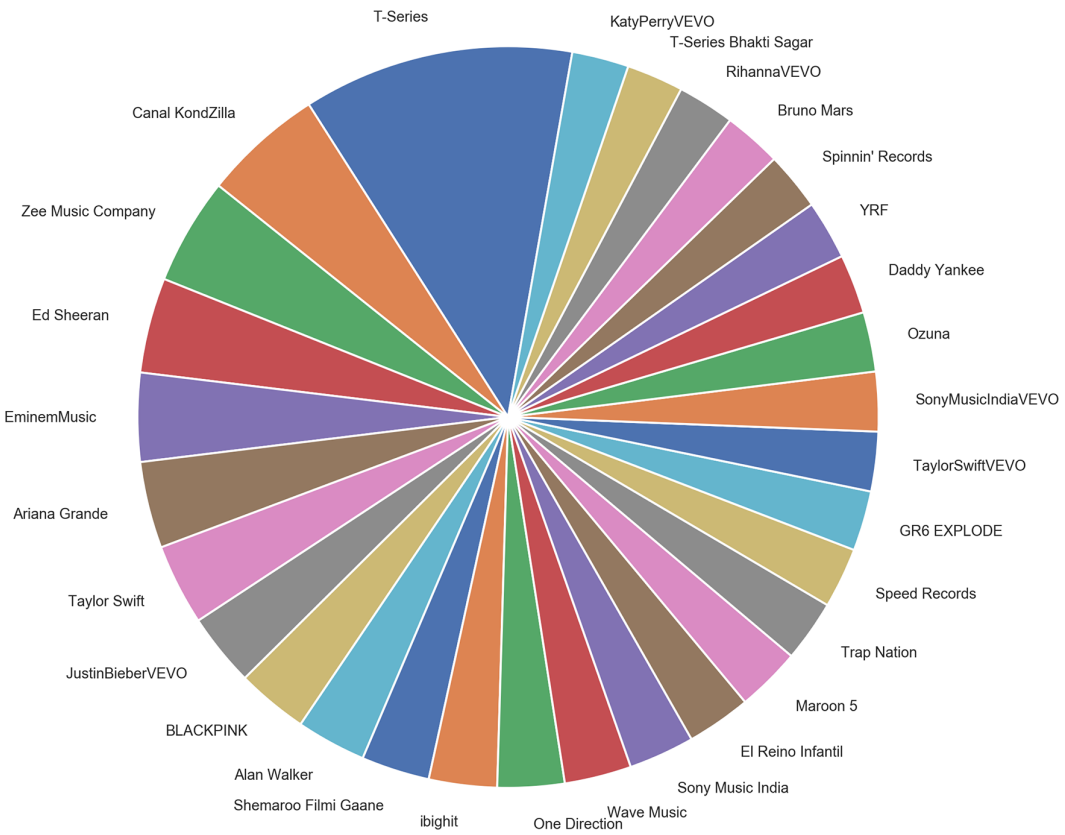

Figure 2.42: Pie chart showing the top 30 YouTube music channels

The second visualization is supposed to illustrate the number of people playing a certain game in a casino over 2 days:

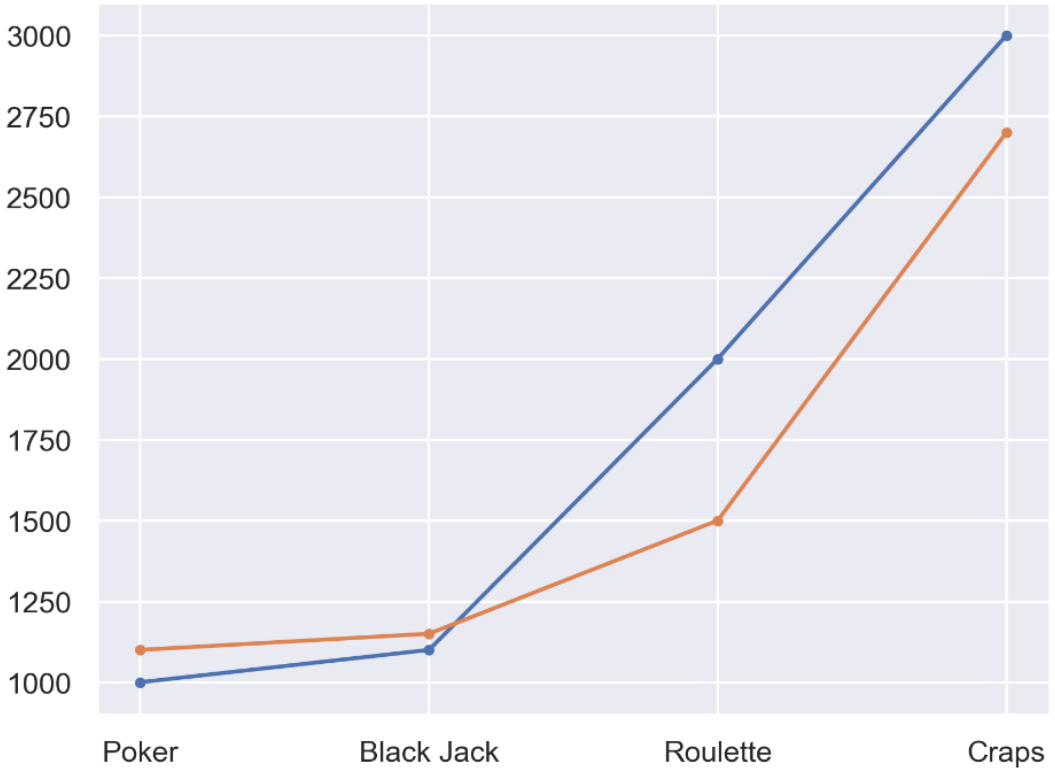

Figure 2.43: Line chart displaying casino data for 2 days

> **NOTE**
>
> The solution for this activity can be found via this link.

ACTIVITY 2.06: CHOOSING A SUITABLE VISUALIZATION

In this activity, we are using a dataset to visualize the median, the interquartile ranges, and the underlying density of populations from different income groups. Following is the link to the dataset that we have used: https://population.un.org/wpp/Download/Files/1_Indicators%20(Standard)/EXCEL_FILES/1_Population/WPP2019_POP_F07_1_POPULATION_BY_AGE_BOTH_SEXES.xlsx. Select the best suitable plot from the following plots.

The following diagram shows the population by different income groups using a density plot:

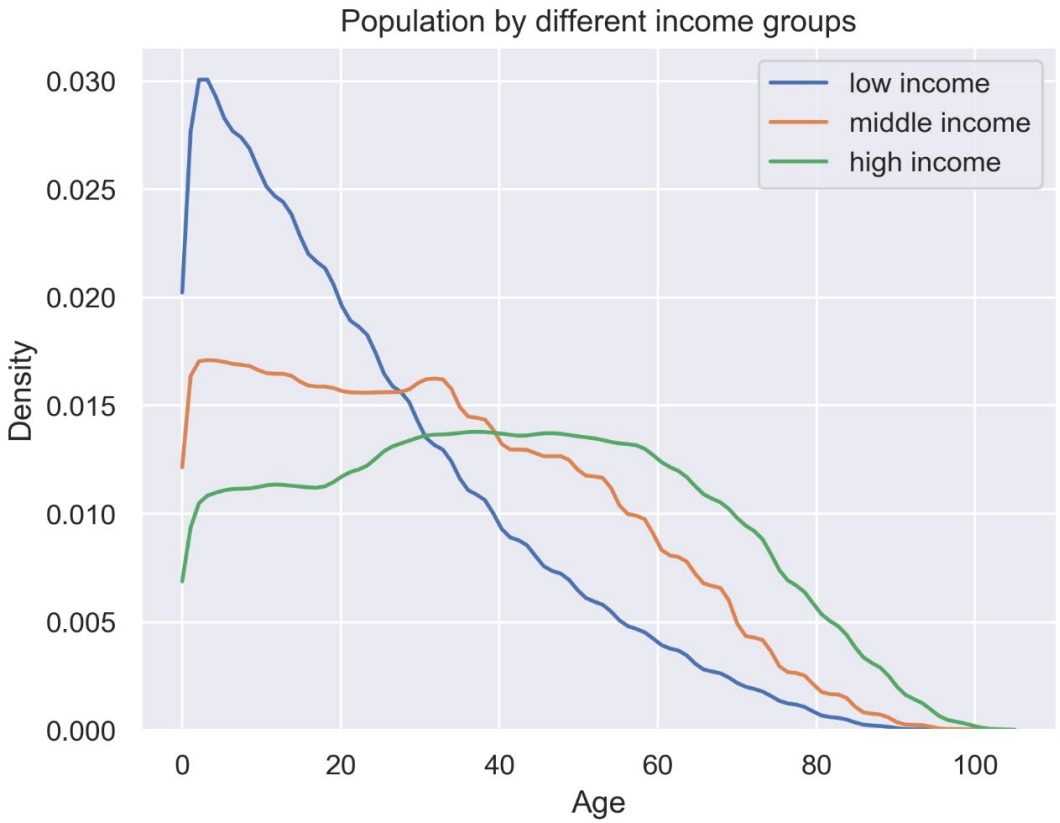

Figure 2.44: Density plot

The following diagram shows the population by different income groups using a box plot:

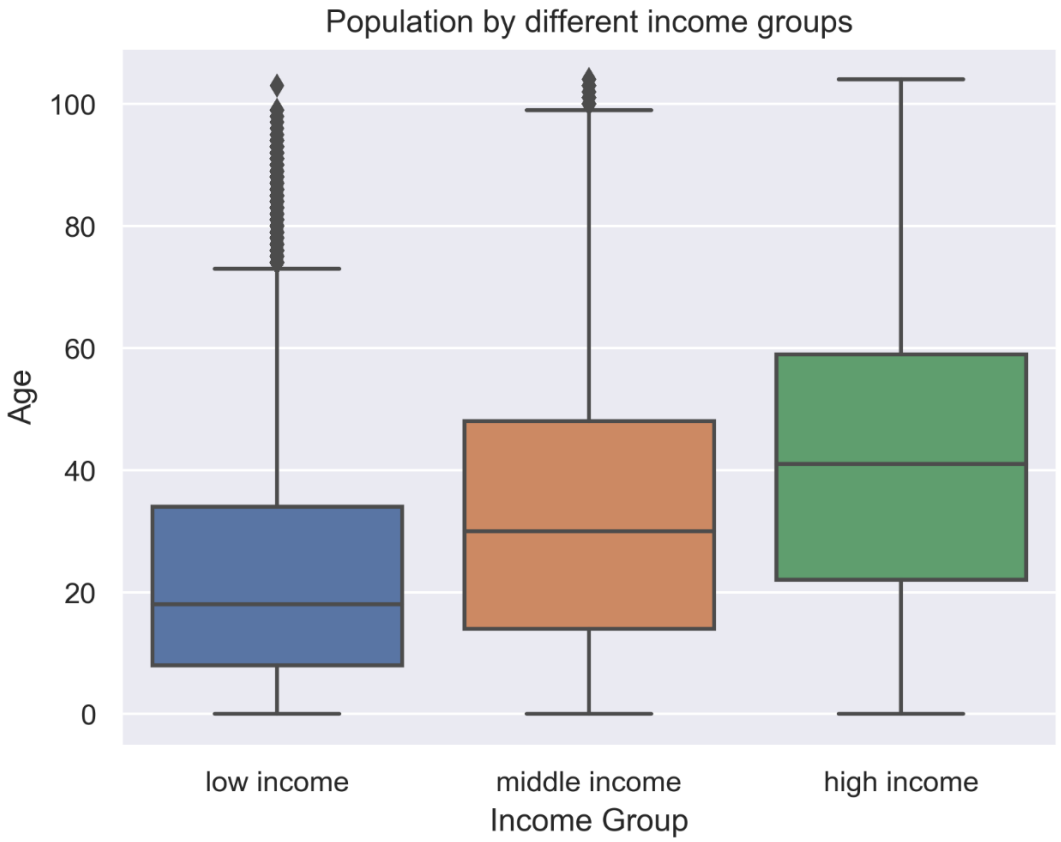

Figure 2.45: Box plot

The following diagram shows the population by different income groups using a violin plot:

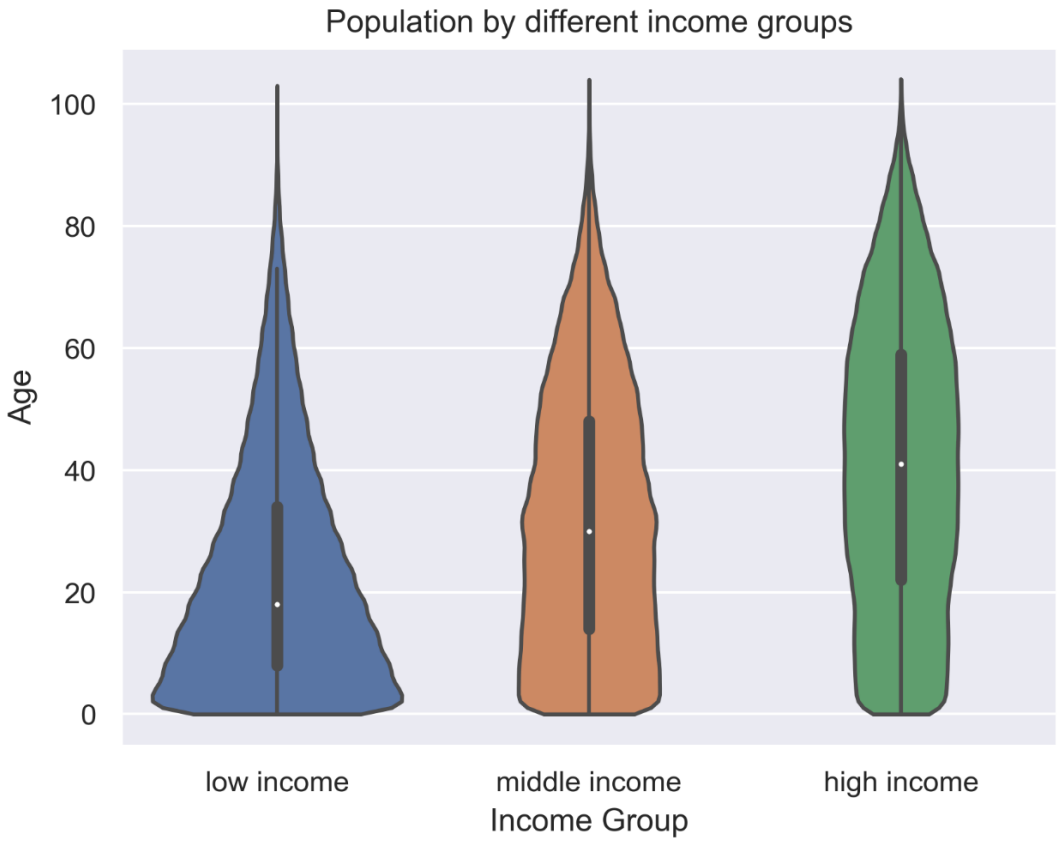

Figure 2.46: Violin plot

> **NOTE**
>
> The solution for this activity can be found via this link.

SUMMARY

This chapter covered the most important visualizations, categorized into comparison, relation, composition, distribution, and geological plots. For each plot, a description, practical examples, and design practices were given. Comparison plots, such as line charts, bar charts, and radar charts, are well suited to comparing multiple variables or variables over time. Relation plots are perfectly suited to show relationships between variables. Scatter plots, bubble plots, which are an extension of scatter plots, correlograms, and heatmaps were considered.

Composition plots are ideal if you need to think about something as part of a whole. We first covered pie charts and continued with stacked bar charts, stacked area charts, and Venn diagrams. For distribution plots that give a deep insight into how your data is distributed, histograms, density plots, box plots, and violin plots were considered. Regarding geospatial data, we discussed dot maps, connection maps, and choropleth maps. Finally, some remarks were provided on what makes a good visualization.

In the next chapter, we will dive into Matplotlib and create our own visualizations. We will start by introducing the basics, followed by talking about how you can add text and annotations to make your visualizations more comprehensible. We will continue creating simple plots and using layouts to include multiple plots within a visualization. At the end of the next chapter, we will explain how you can use Matplotlib to visualize images.