Chapter 2

All You Need to Know About Plots

Learning Objectives

By the end of this chapter, you will be able to:

- Identify the best plot type for a given dataset and scenario
- Explain the design practices of certain plots
- Design outstanding, tangible visualizations

In this chapter, we will learn the basics of different types of plots.

Introduction

In this chapter, we will focus on various visualizations and identify which visualization is best to show 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. Relation plots are handy to show 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.

Composition plots, which are used to visualize variables that are part of a whole, as well as pie charts, stacked bar charts, stacked area charts, and Venn diagrams are going to be explained. To get a deeper insight into the distribution of variables, distribution plots are used. As a part of distribution plots, histograms, density plots, box plots, and violin plots will be covered. Finally, we will talk about dot maps, connection maps, and choropleth maps, which can be categorized into geo plots. Geo plots are useful for visualizing geospatial data.

Comparison Plots

Comparison plots include charts that are well-suited for comparing multiple variables or variables over time. For a comparison among items, bar charts (also called column charts) are the best way to go. Line charts are great for visualizing variables over time. For a certain time period (say, less than ten 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, which is connected by straight-line segments.

The value 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 (roughly more than ten).
- For smaller time periods, vertical bar charts might be the better choice.

The following diagram shows a trend of real-estate prices (in million US dollars) for two decades. Line charts are well-suited for showing data trends:

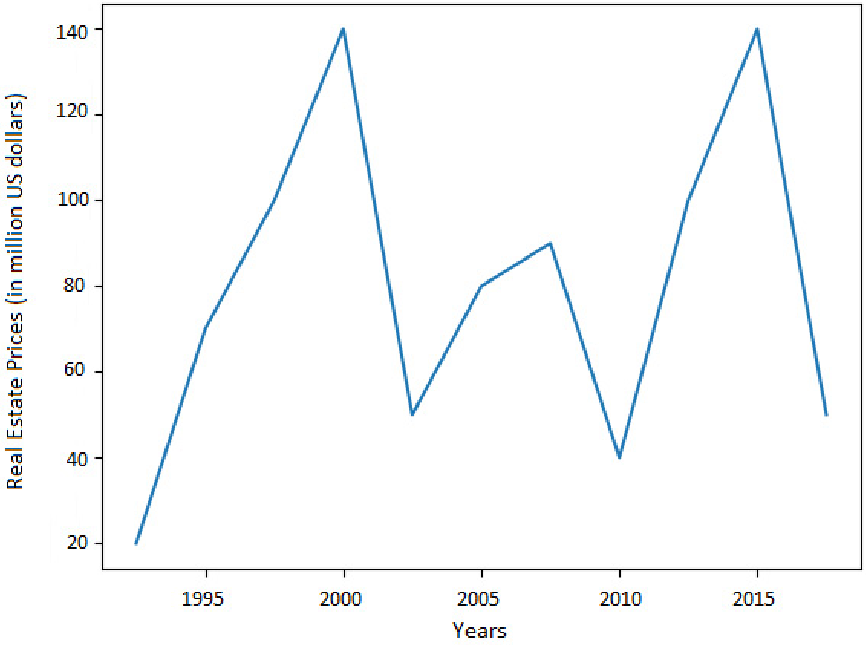

Figure 2.1: Line chart for a single variable

Example:

The following diagram 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:

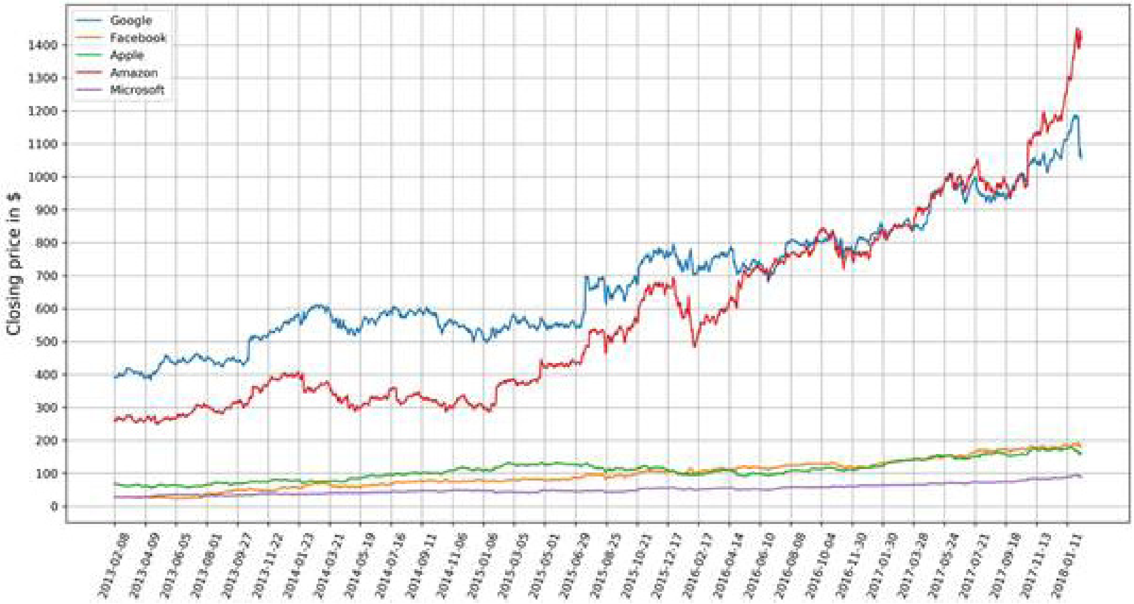

Figure 2.2: Line chart showing stock trends for the five companies

Design practices:

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

Note

Design practices for plots with multiple variables. A legend should be available to describe each variable.

Bar Chart

The bar length encodes the value. There are two variants of bar charts: vertical bar charts and horizontal bar charts.

Uses:

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

The do's and the 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 five students obtained in a test:

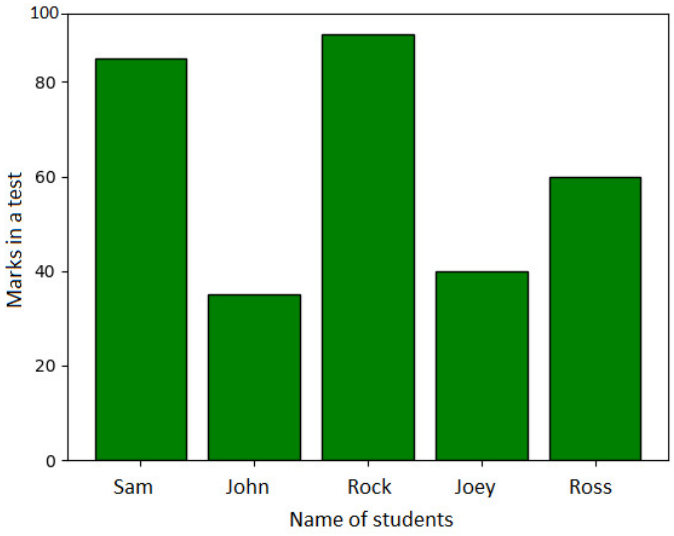

Figure 2.3: Vertical bar chart using student test data

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

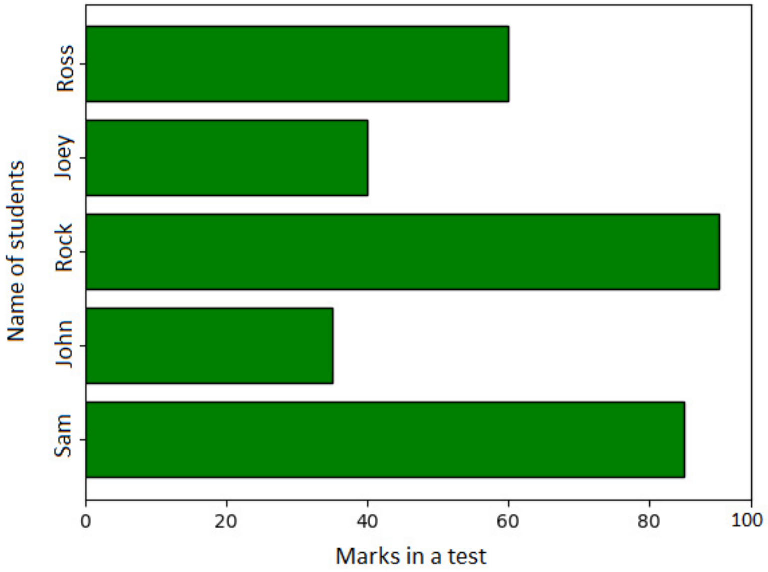

Figure 2.4: Horizontal bar chart using student test data

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 score 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:

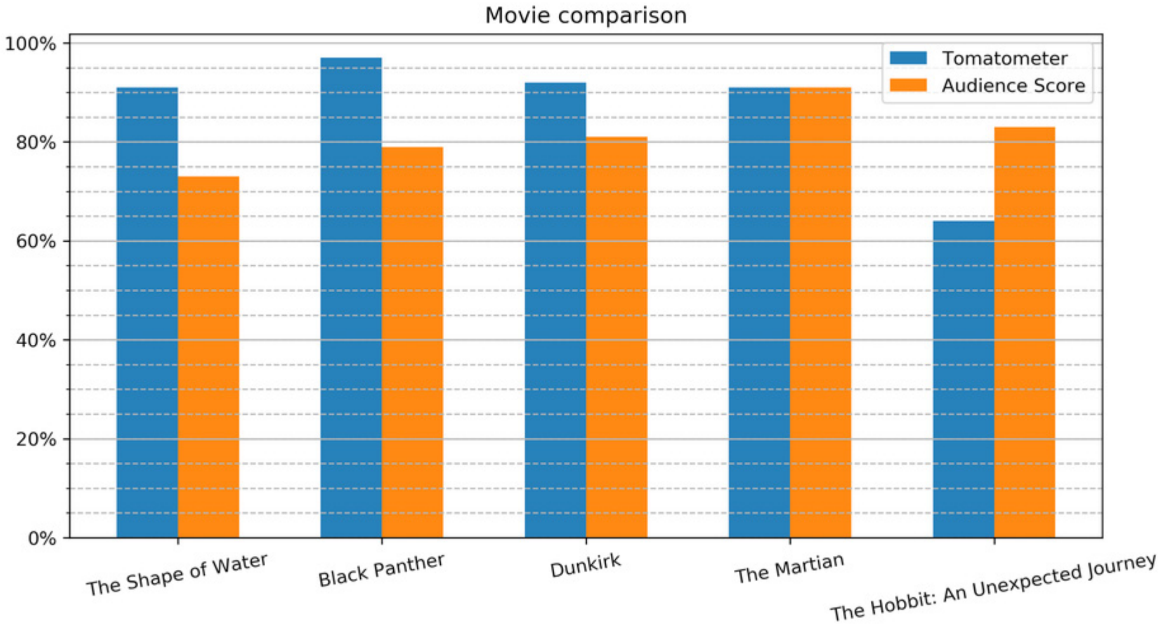

Figure 2.5: Comparative bar chart

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.

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 to show which variables score high or low within a dataset, making them ideal to visualize 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:

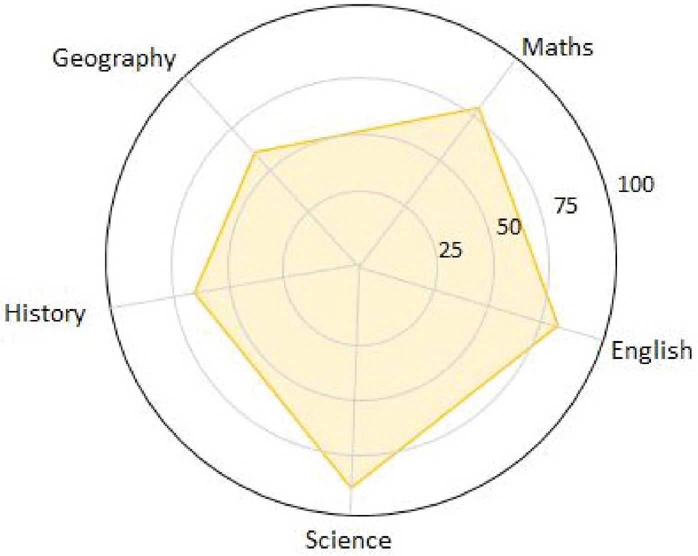

Figure 2.6: Radar chart for one variable (student)

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:

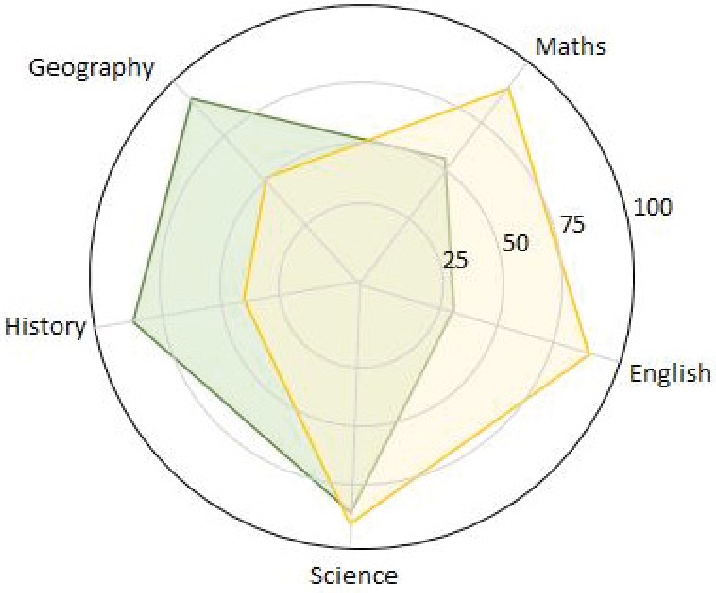

Figure 2.7: Radar chart for two variables (two students)

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

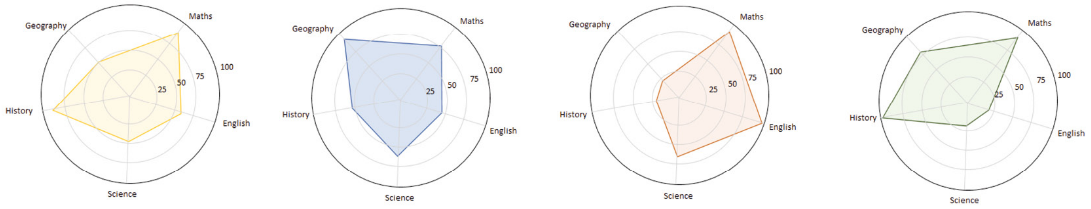

Figure 2.8: Radar chart with faceting for multiple variables (multiple subjects)

Design practices:

- Try to display ten factors or fewer on one radar chart to make it easier to read.
- Use **faceting** for multiple variables/groups, as shown in the preceding diagram, to maintain clarity.

Activity 7: Employee Skill Comparison

You are given scores of four employees (A, B, C, and D) for five attributes: Efficiency, Quality, Commitment, Responsible Conduct, and Cooperation. Your task is to compare the employees and their skills:

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

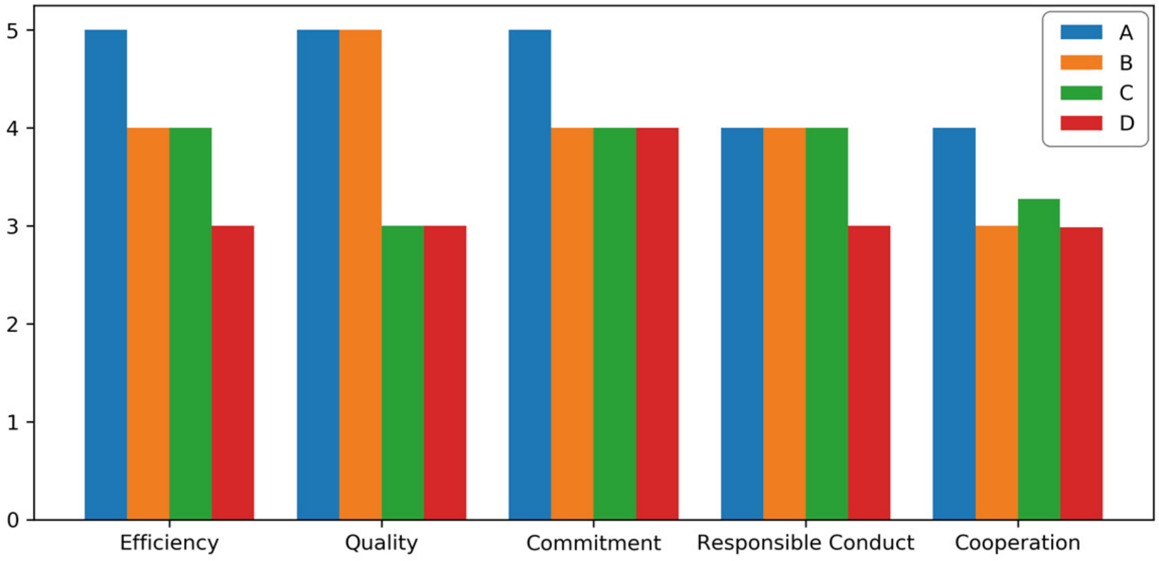

Figure 2.9: Employee skills comparison with a bar chart

The following figure shows a radar chart for employee skills:

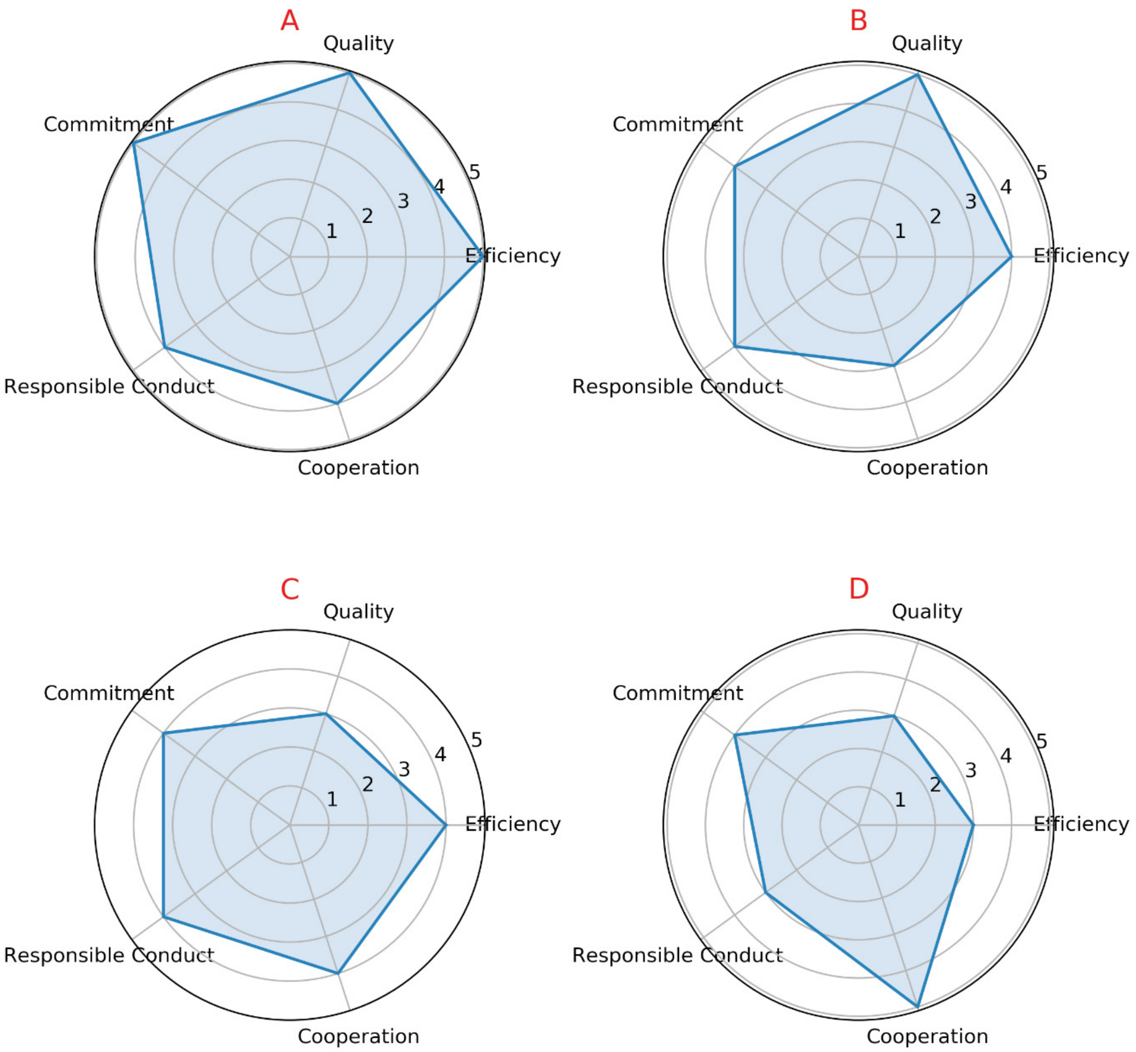

Figure 2.10: Employee skills comparison with a radar chart

3. What could be improved in the respective visualizations?

Note:

The solution for this activity can be found on page 275.

Relation Plots

Relation plots are perfectly suited to show 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 correlating between two qualitative variables. A correlogram is a perfect visualization to show 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 for 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:

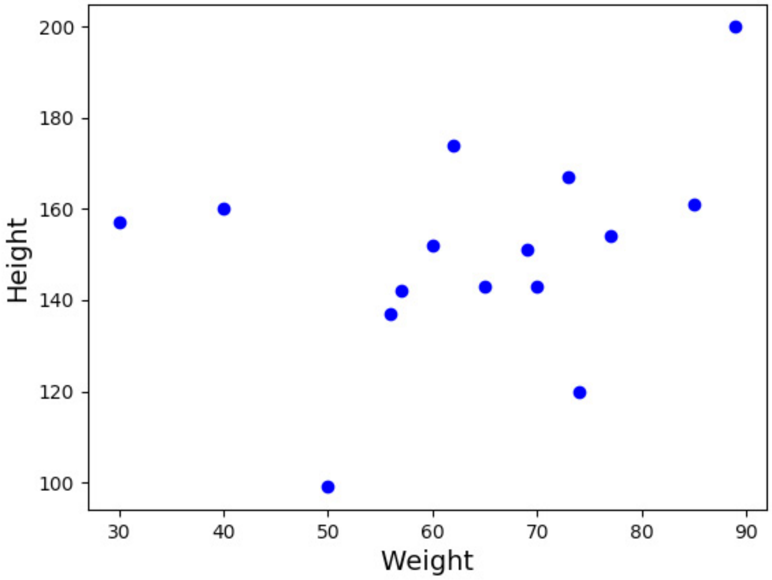

Figure 2.11: Scatter plot with a single variable (one group)

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**:

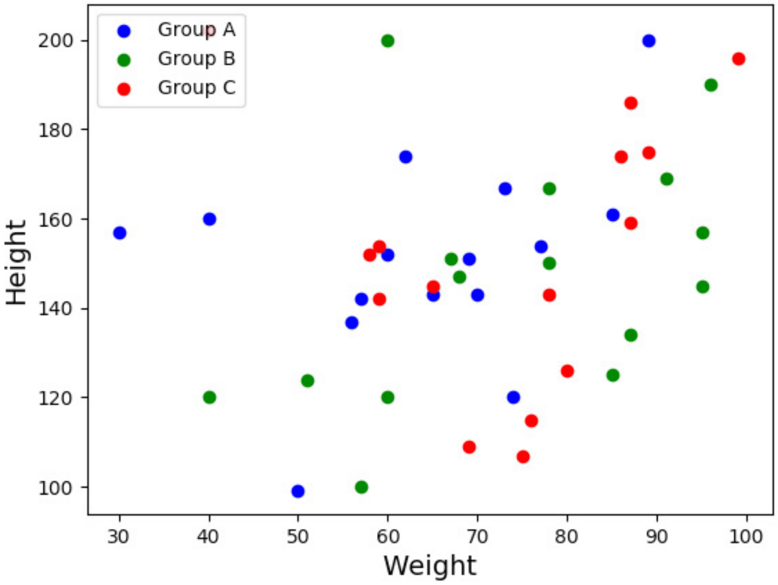

Figure 2.12: Scatter plot with multiple variables (three groups)

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

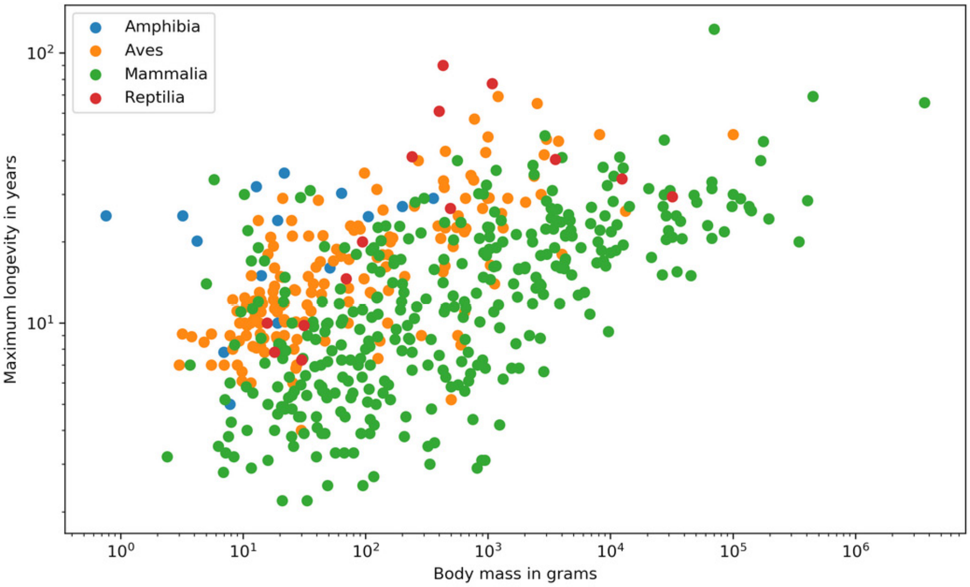

Figure 2.13: Correlation between body mass and maximum longevity for animals

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 the 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:

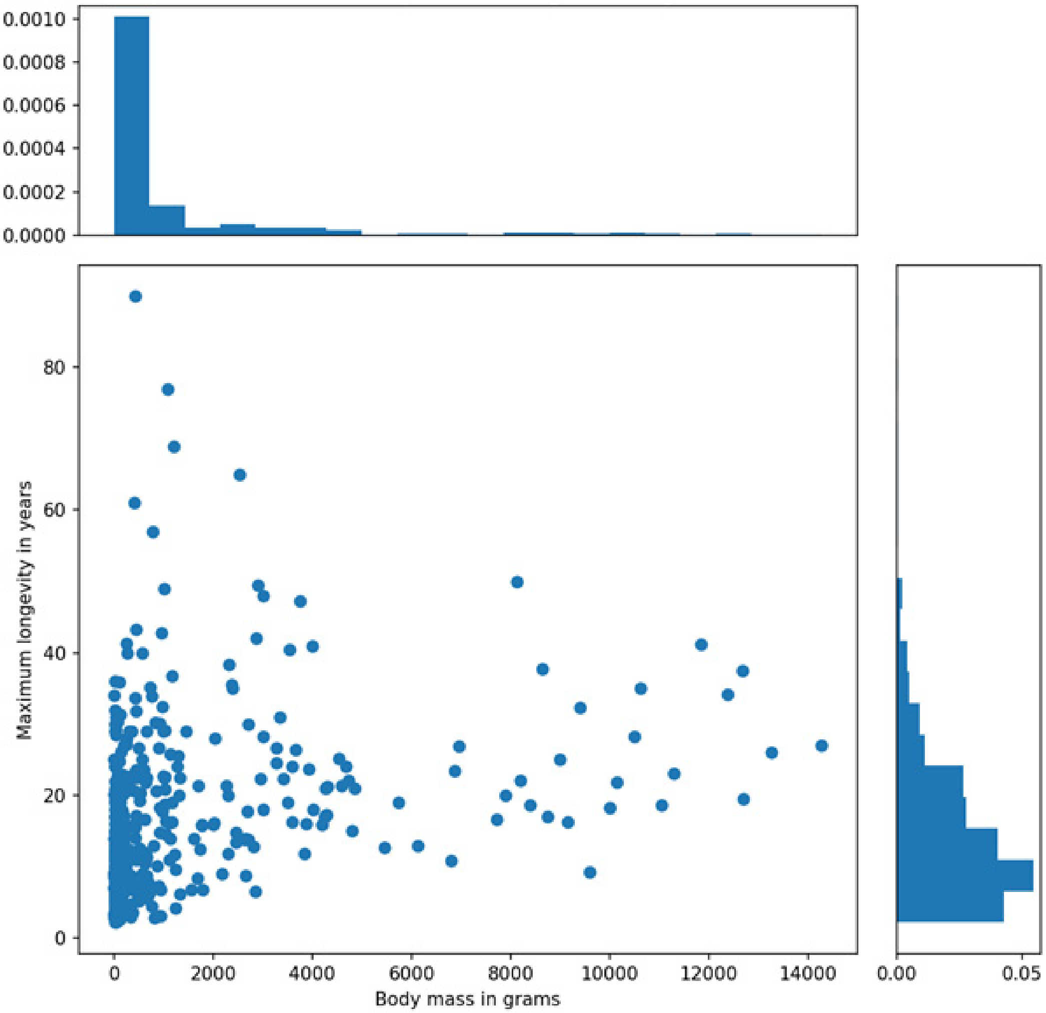

Figure 2.14: Correlation between body mass and maximum longevity of the Aves class with marginal histograms

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.

Uses:

- 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:

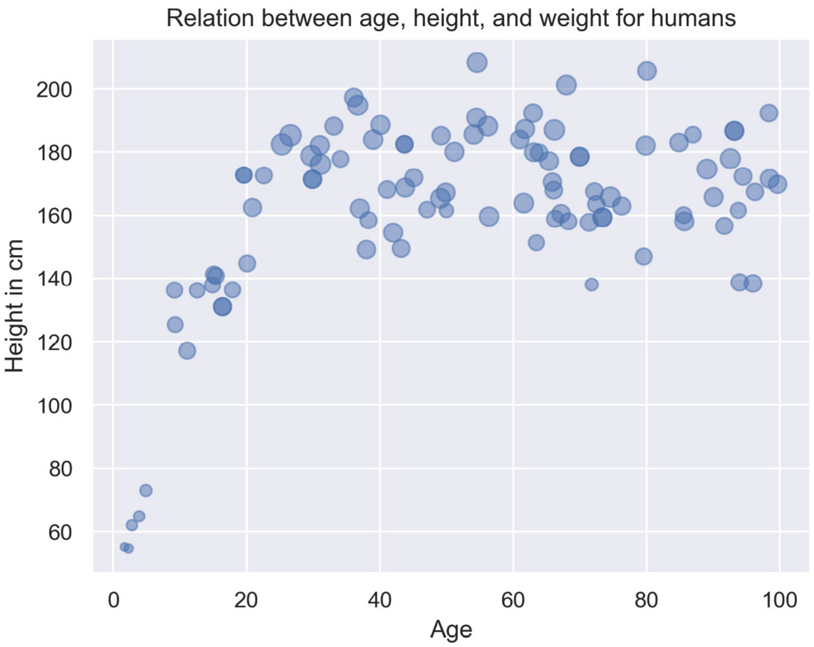

Figure 2.15: Bubble plot showing relation between height and age of humans

Design practices:

- Design practices for the scatter plot are also applicable to the bubble plot.
- Don't use it for very large amounts of data, since too many bubbles make the chart hard 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 for multiple groups or categories using different colors. A correlogram is a great chart for exploratory data analysis to get a feeling for your data, especially the correlation between variable pairs.

Examples:

The following diagram shows a correlogram for 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:

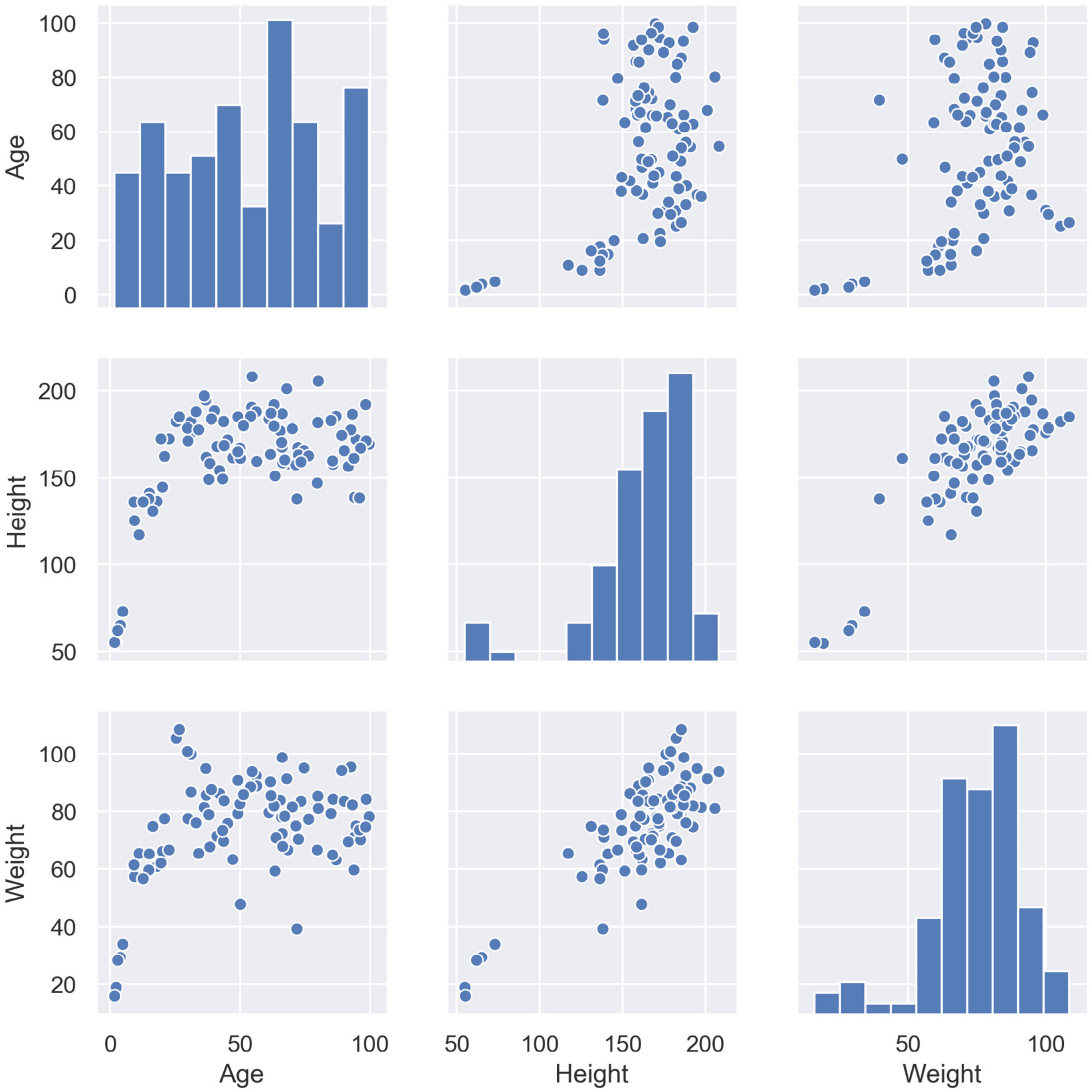

Figure 2.16: Correlogram with single category

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

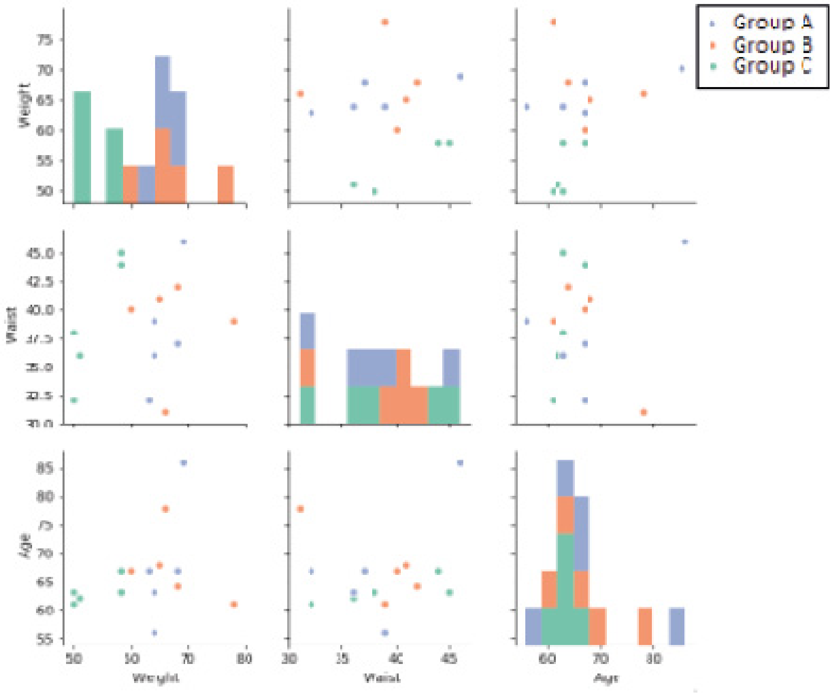

Figure 2.17: Correlogram with multiple categories

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, where categorical variables are placed in the rows and columns and a numerical or categorical variable is represented as colors or color saturation.

Uses:

- Visualization of multivariate data. 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:

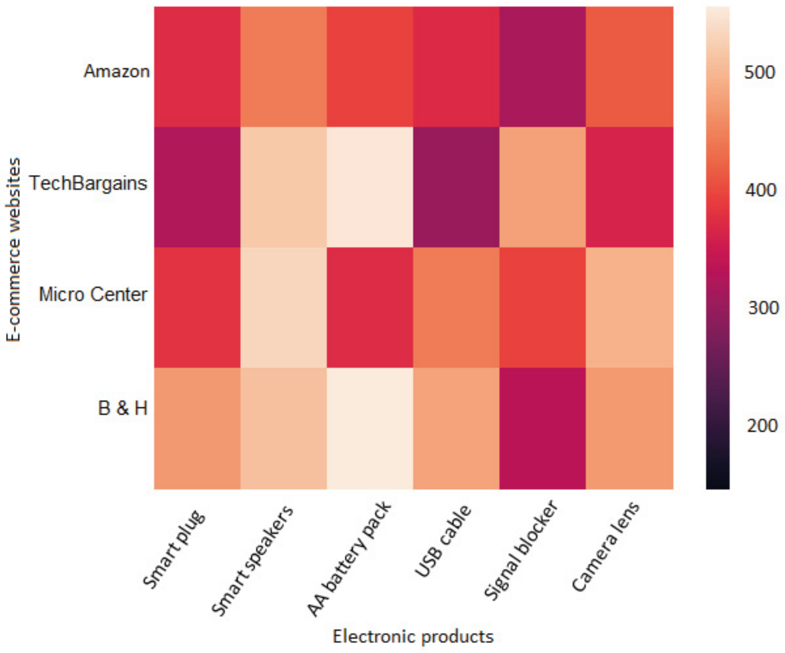

Figure 2.18: Heatmap for popular products in the Electronics category

Variants: annotated heatmaps

Let's see the same example the we saw previously in an annotated heatmap:

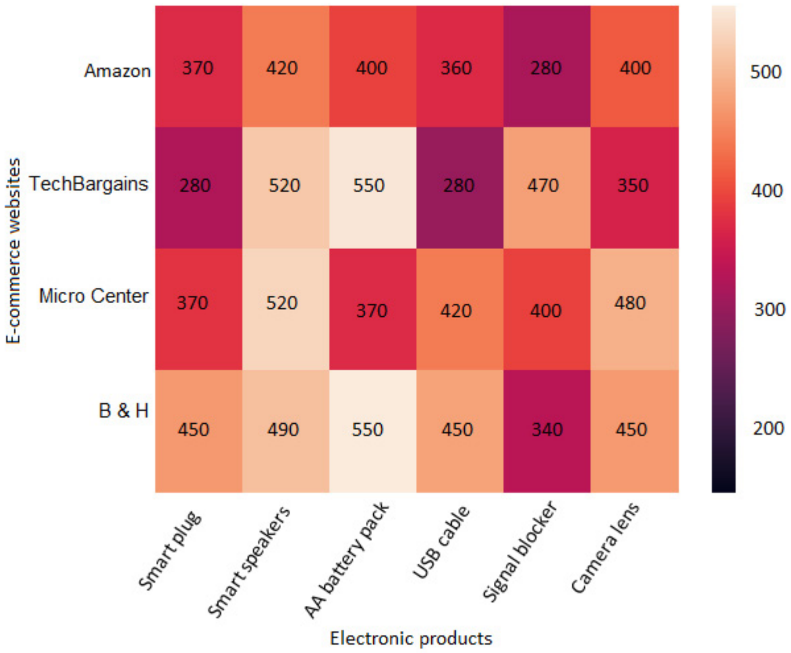

Figure 2.19: Annotated heatmap for popular products in the Electronics category

Activity 8: Road Accidents Occurring over Two Decades

You are given a diagram that gives information about the road accidents that have occurred over the past two decades during the months of January, April, July, and October:

1. Identify the year during which the number of road accidents occurred were the least.
2. For the past two decades, identify the month for which accidents show a marked decrease:

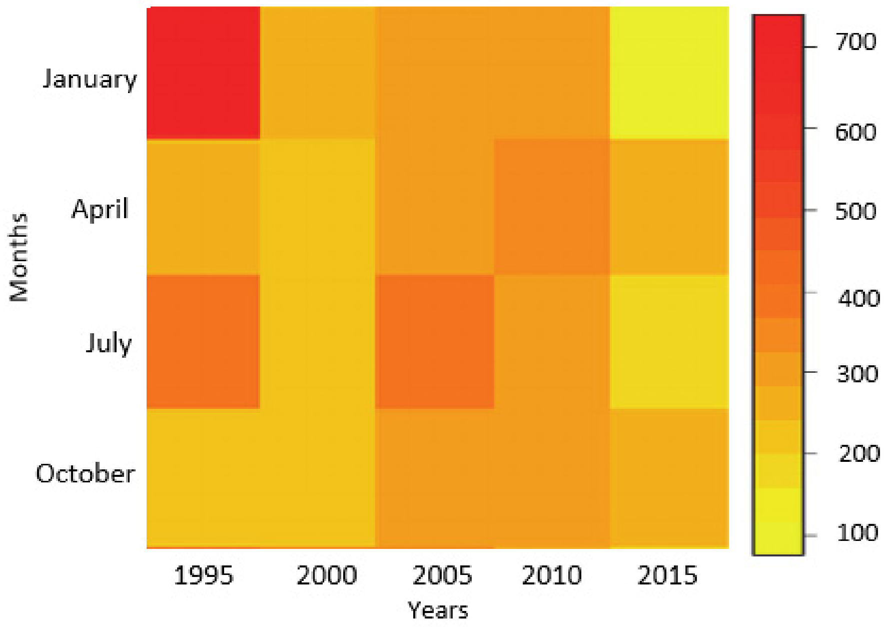

Figure 2.20: Total accidents over 20 years

Note:

The solution for this activity can be found on page 275.

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 proportion by dividing a circle into slices. Each arc length represents a proportion of a category. The full circle equals 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 most of the time.

Uses:

- Compare items that are part of a whole.

Examples:

The following diagram shows a pie chart that shows different fielding positions of the cricket ground, such as long on, long off, third man, and fine leg:

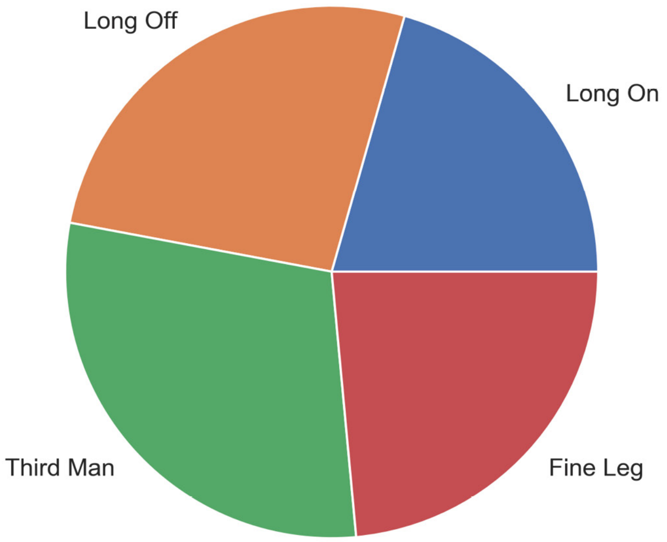

Figure 2.21: Pie chart showing fielding positions in a cricket ground

The following diagram shows water usage around the world:

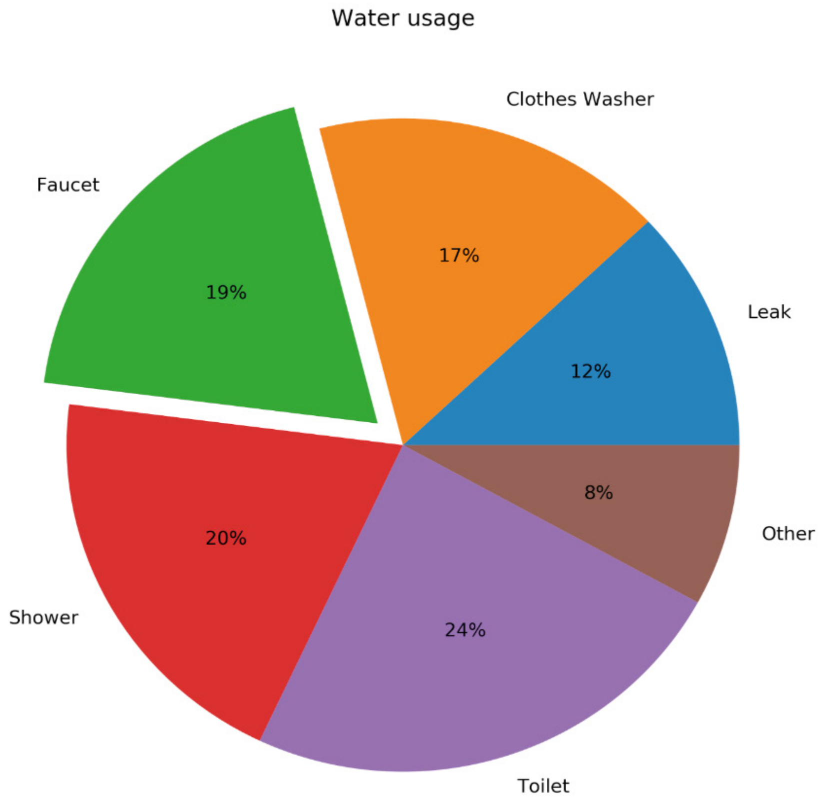

Figure 2.22: Pie chart for global water usage

Design practices:

- Arrange the slices according to their size in increasing/decreasing order, either in a clockwise or anticlockwise 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 sub-groups.

The following figure shows a basic donut chart:

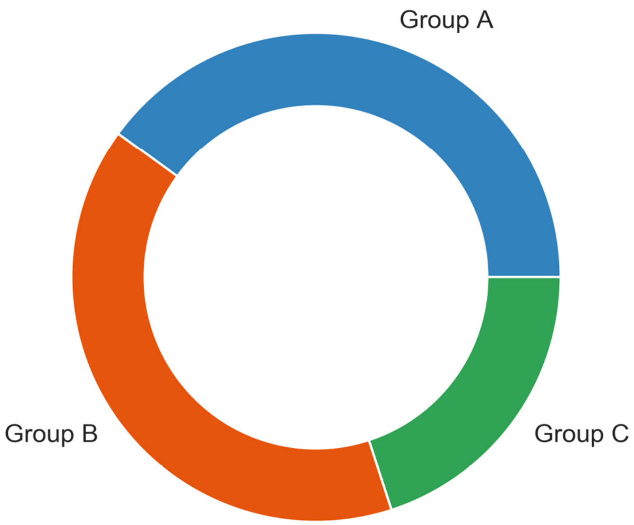

Figure 2.23: Donut chart

The following figure shows a donut chart with subgroups:

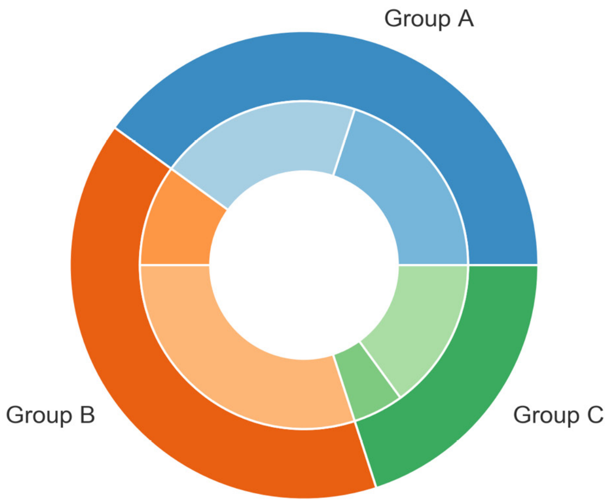

Figure 2.24: Donut chart with subgroups

Design practices:

- 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 sub-categories and the proportion of the sub-category, 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 referenced as a **100% stacked bar chart** and makes it easier to see relative differences between quantities in each group.

Uses:

- Compare variables that can be divided into sub-variables.

Examples:

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

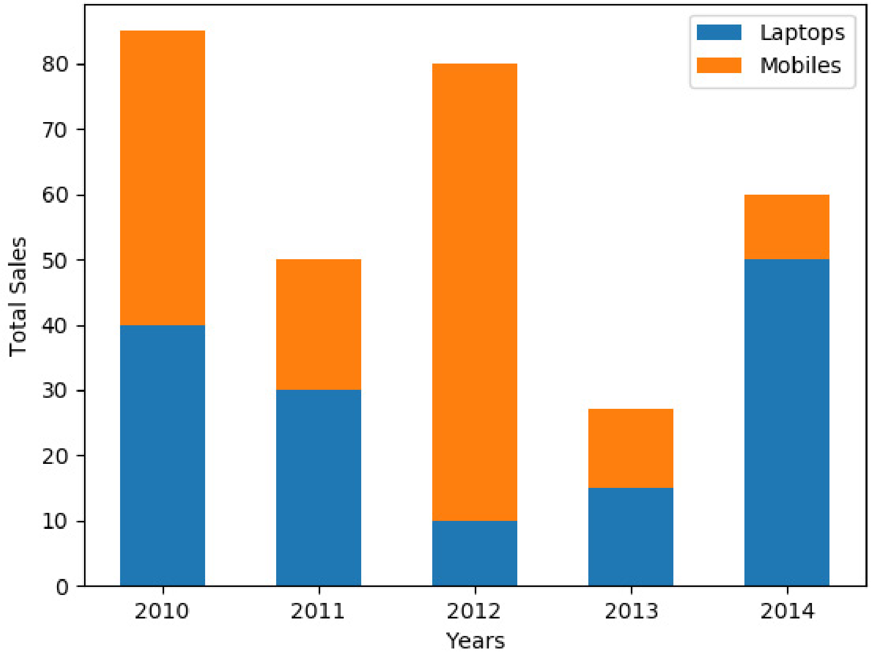

Figure 2.25: Stacked bar chart to show sales of laptops and mobiles

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

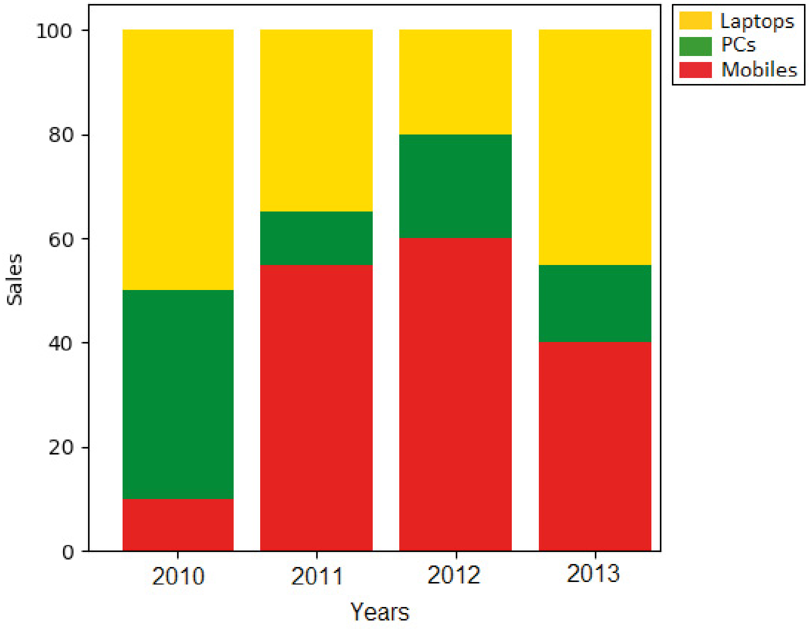

Figure 2.26: 100% stacked bar chart to show sales of laptops, PCs, and mobiles

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:

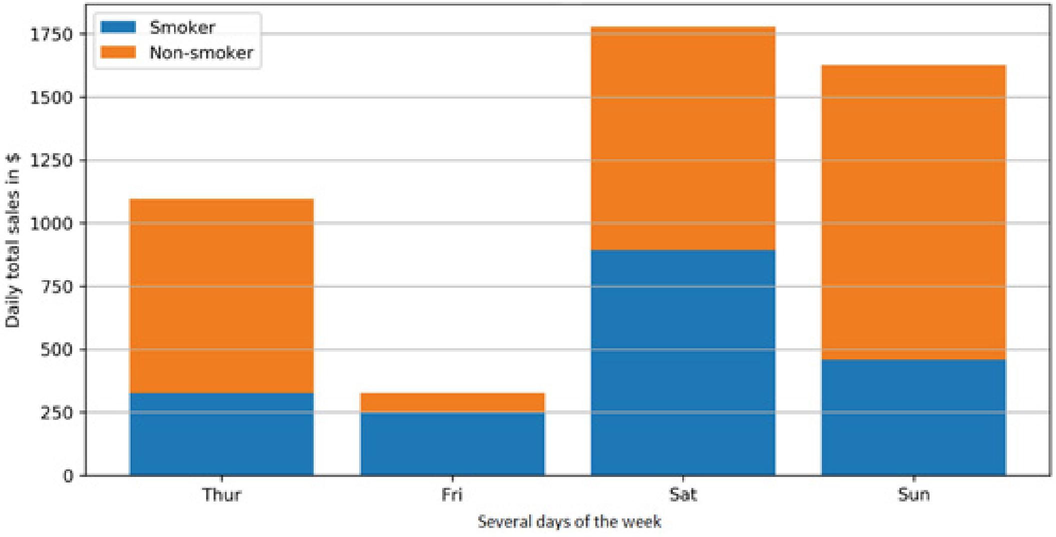

Figure 2.27: Daily total sales of restaurant categorized by smokers and non-smokers

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 order it uniformly 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 on top of one another. It helps to analyze both individual and overall trend information.

Uses:

- 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 companies like Google, Facebook, Twitter, and Snapchat over a decade:

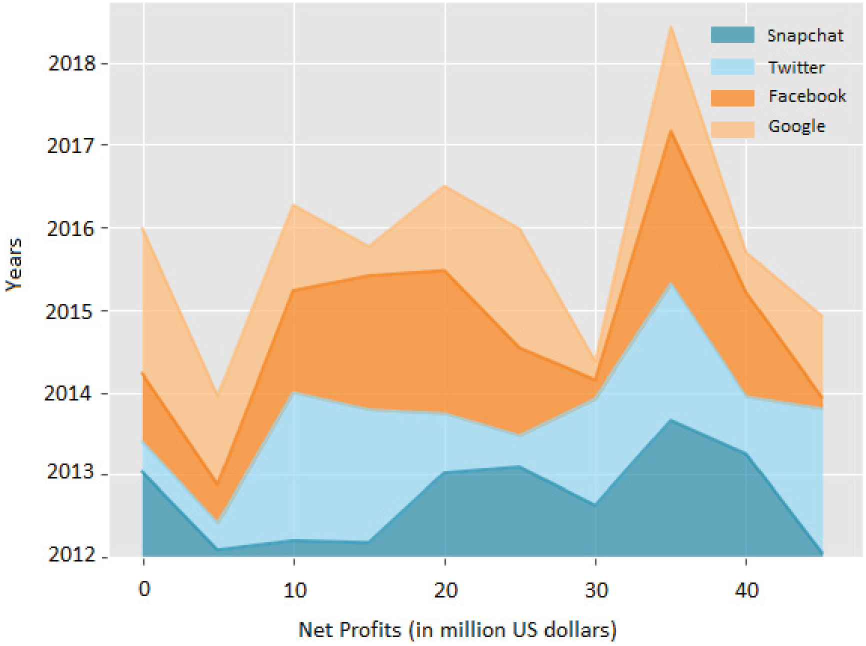

Figure 2.28: Stacked area chart to show net profits of four companies

Design practices:

- Using transparent colors might improve information visibility.

Activity 9: 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:

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

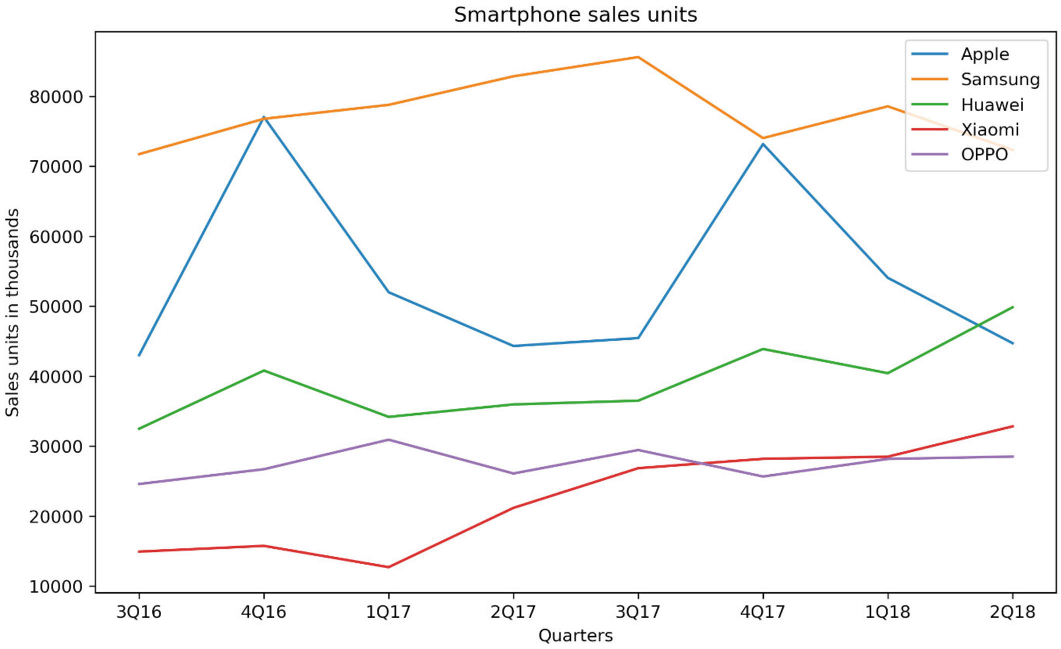

Figure 2.29: Line chart of smartphone sales units

Note:

The solution for this activity can be found on page 275.

Venn Diagram

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

Uses:

- 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:

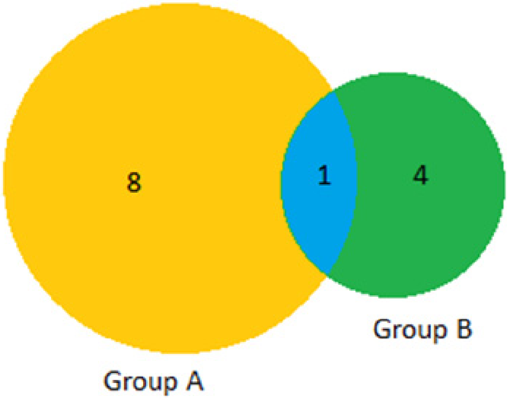

Figure 2.30: Venn diagram to show students taking the same class

Design practices:

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

Distribution Plots

Distribution plots give a deep insight into how your data is distributed. For a single variable, a histogram is well-suited. 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.

Uses:

- 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 solid line indicates the mean and the dashed lines indicate the standard deviation:

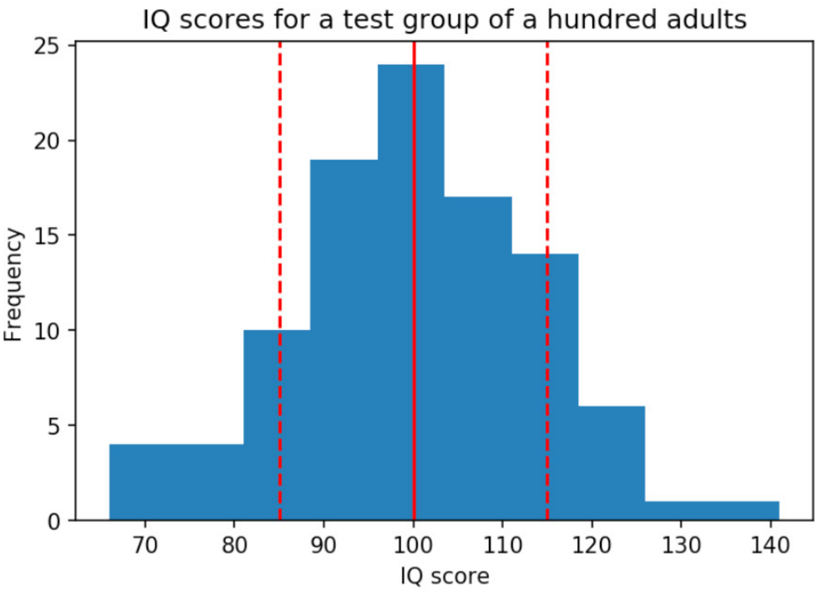

Figure 2.31: Distribution of Intelligence Quotient (IQ) for a test group of a hundred adults

Design practices:

- Try different numbers of bins, 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. An advantage they 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).

Uses:

- You can 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:

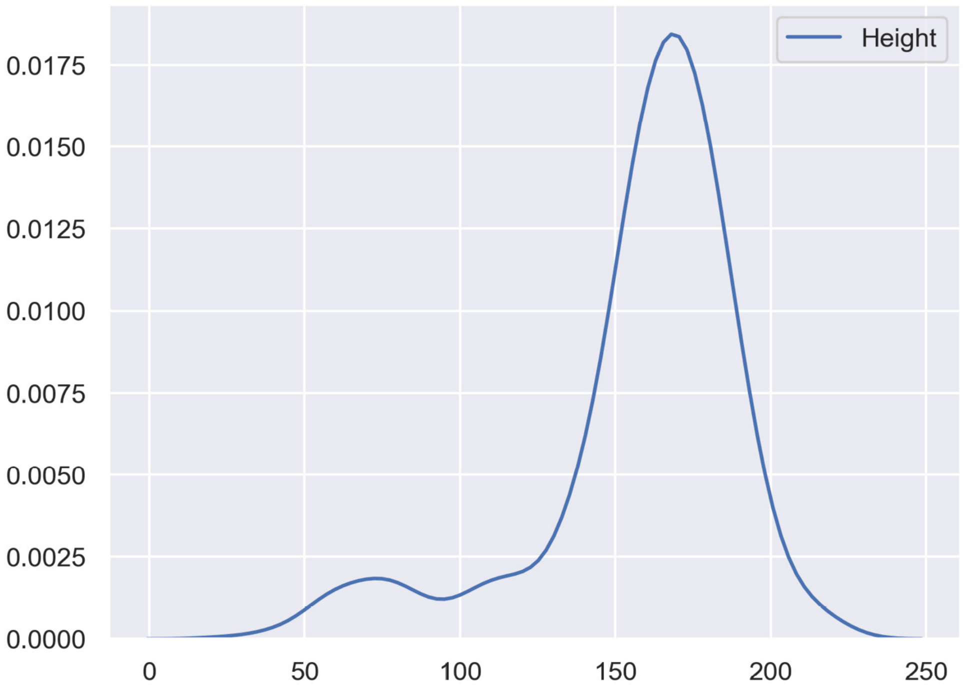

Figure 2.32: Density plot

The following diagram shows a basic multi-density plot:

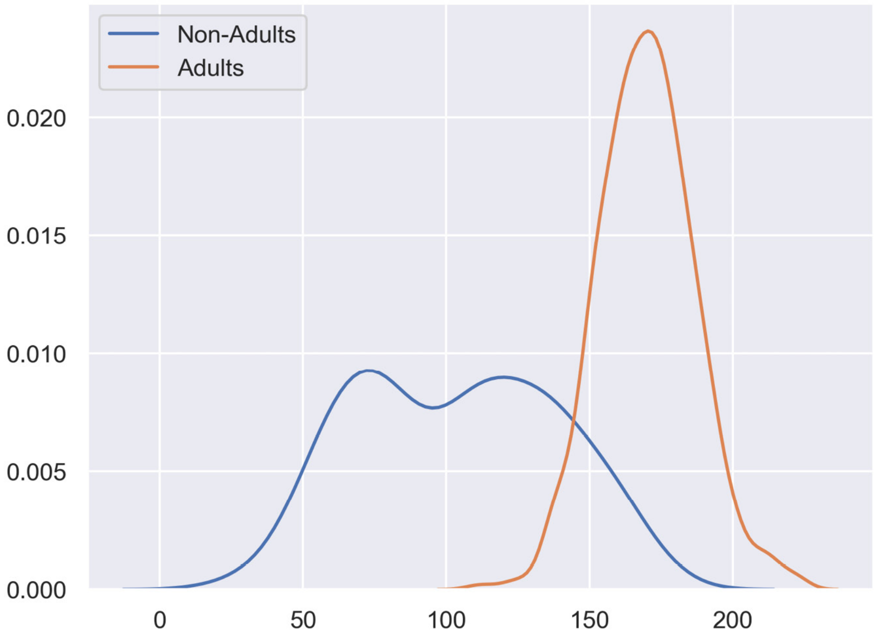

Figure 2.33: Multi-density plot

Design practices:

- Use contrasting colors for plotting 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. The horizontal line within the box denotes the median. The **whiskers** extending from the box show the range of the data. It is also an option to show data **outliers**, usually as circles or diamonds, past the end of the whiskers.

Uses:

- If you want to compare statistical measures for multiple variables or groups, you can simply plot multiple boxes next to one another.

Examples:

The following diagram shows a basic box plot:

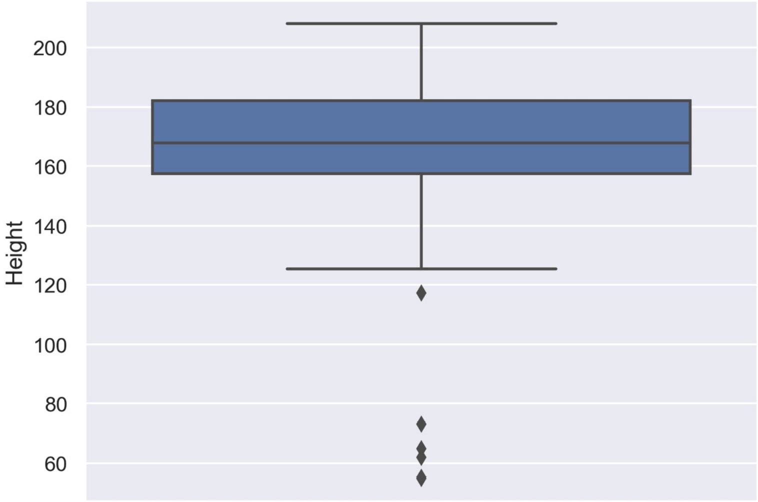

Figure 2.34: Box plot showing single variable

The following diagram shows a basic box plot for multiple variables:

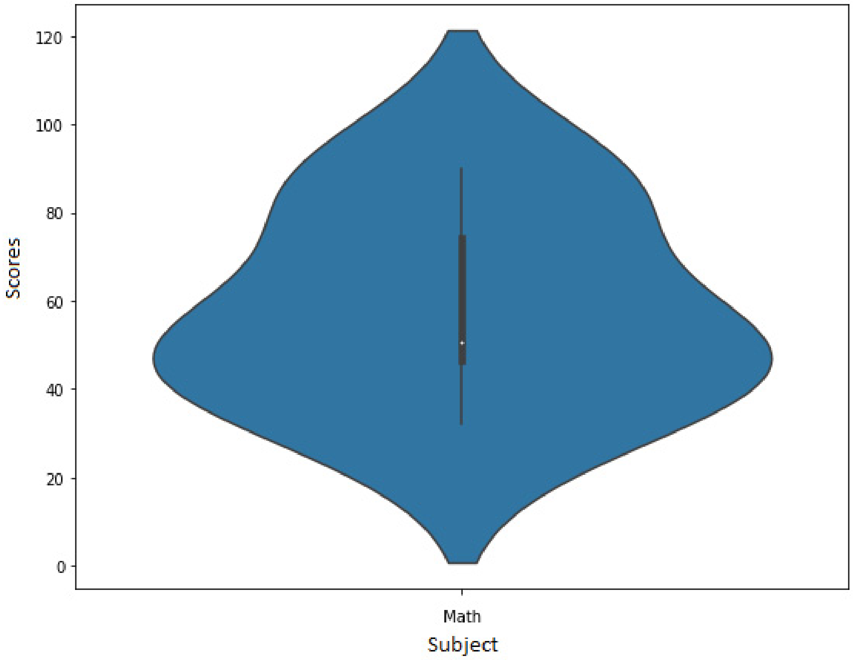

Figure 2.35: Box plot for multiple variables

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, the thin black line shows the 95% confidence interval, and the white dot shows the median. On both sides of the center-line, the density is visualized.

Uses:

- If you want to compare statistical measures for multiple variables or groups, you can simply plot multiple violins next to one another.

Examples:

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

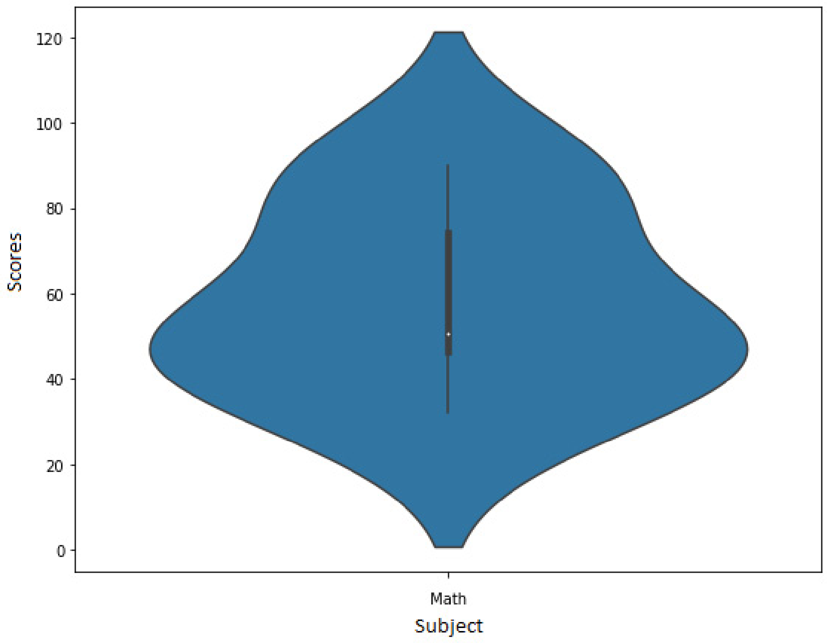

Figure 2.36: Violin plot for a single variable (math)

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

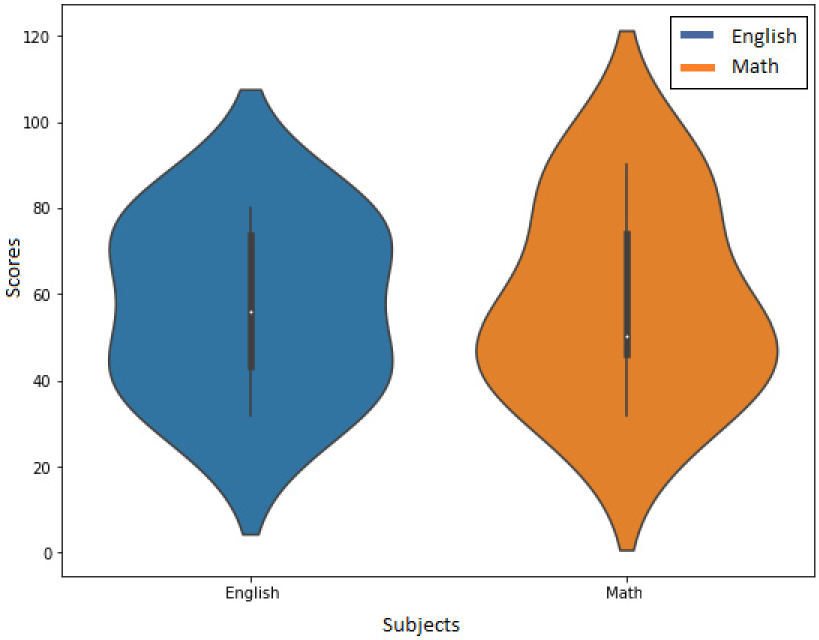

Figure 2.37: Violin plot for multiple variables (English and math)

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**:

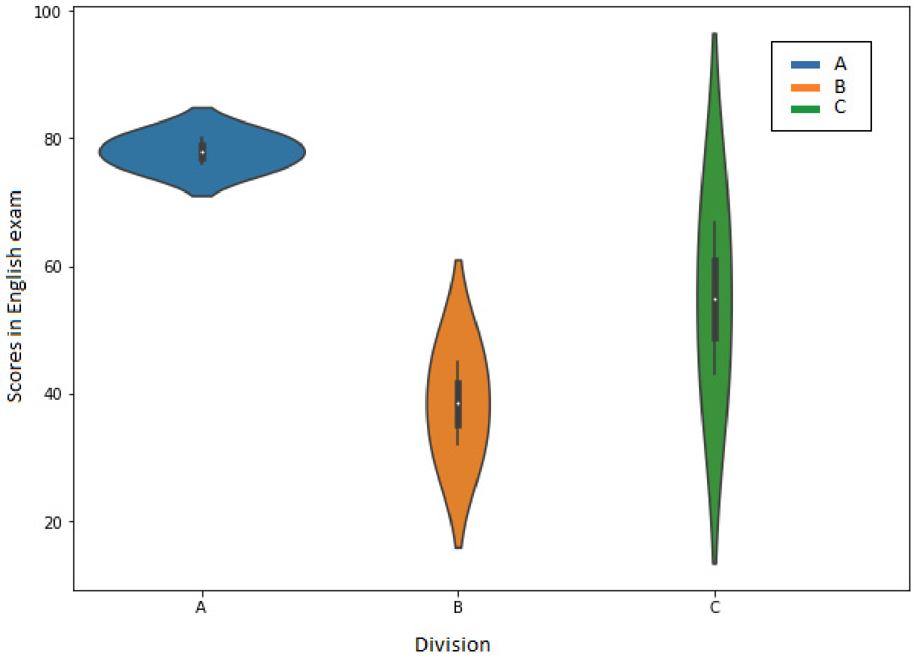

Figure 2.38: Violin plot with multiple categories (three groups of students)

Design practices:

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

Activity 10: Frequency of Trains during Different Time Intervals

You are provided with a histogram that states the total number of trains arriving at different time intervals:

1. By looking at the following graph, can you identify the interval during which the most number of trains arrive?
2. How would the histogram change if the number of trains arriving between 4 and 6 pm were to be increased by 50?

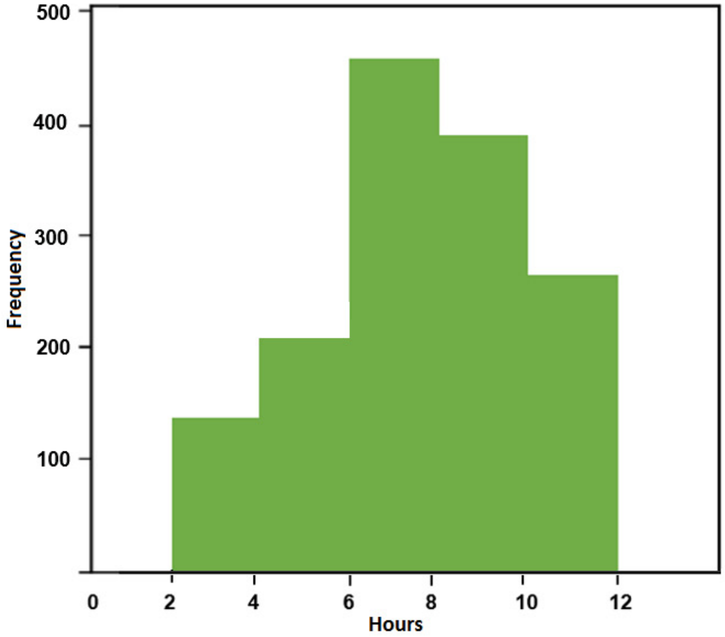

Figure 2.39: Frequency of trains during different time intervals

Note:

The solution for this activity can be found on page 276.

Geo Plots

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.

Uses:

- For the visualization of geospatial data

Example:

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

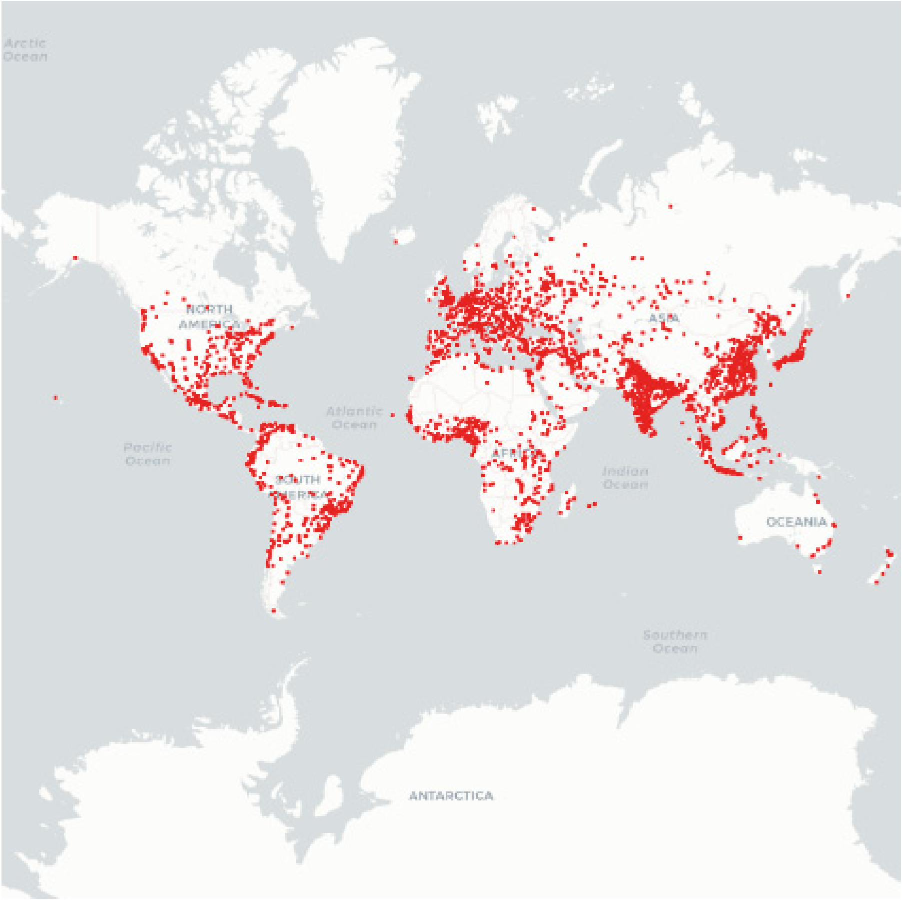

Figure 2.40: Dot map showing bus stops worldwide

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. A tile represents a geographic region for, for example, 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 prominence to larger areas, so you might want to normalize your data by dividing the map area-wise.

Uses:

- For the visualization of 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:

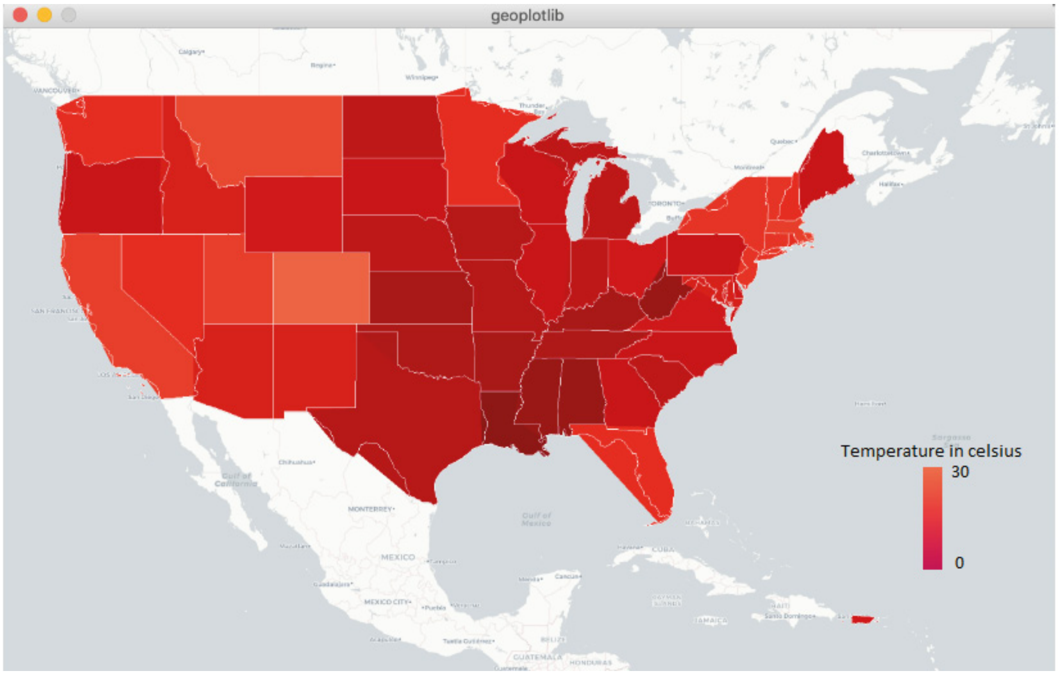

Figure 2.41: Choropleth map showing weather forecast of USA

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 to 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 (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.

Uses:

- For the visualization of connections

Examples:

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

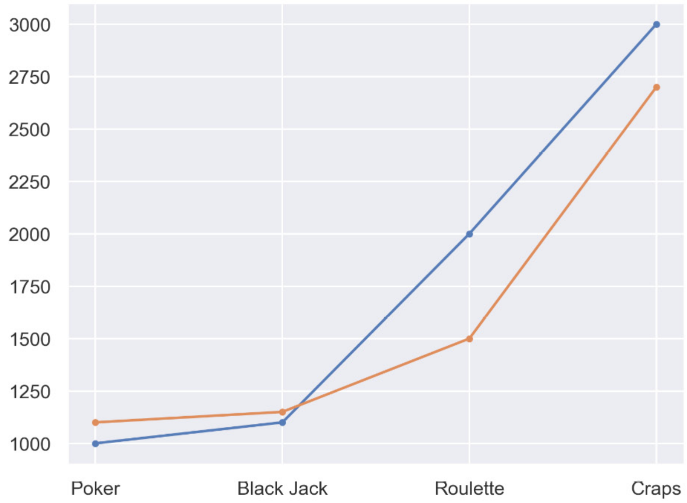

Figure 2.42: Connection map showing Flight connections around the world

Design practices:

- Do not show too many connections. You should still see the map to get a feel of 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.

What Makes a Good Visualization?

There are multiple aspects to what makes a good visualization:

- Most importantly, a 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:

- Colors are more perceptible than symbols.
- 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 11: Identifying the Ideal Visualization

The following visualizations are not ideal as they do not represent data well. Answer the following questions for each visualization:

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 YouTubers according to the number of subscribers:

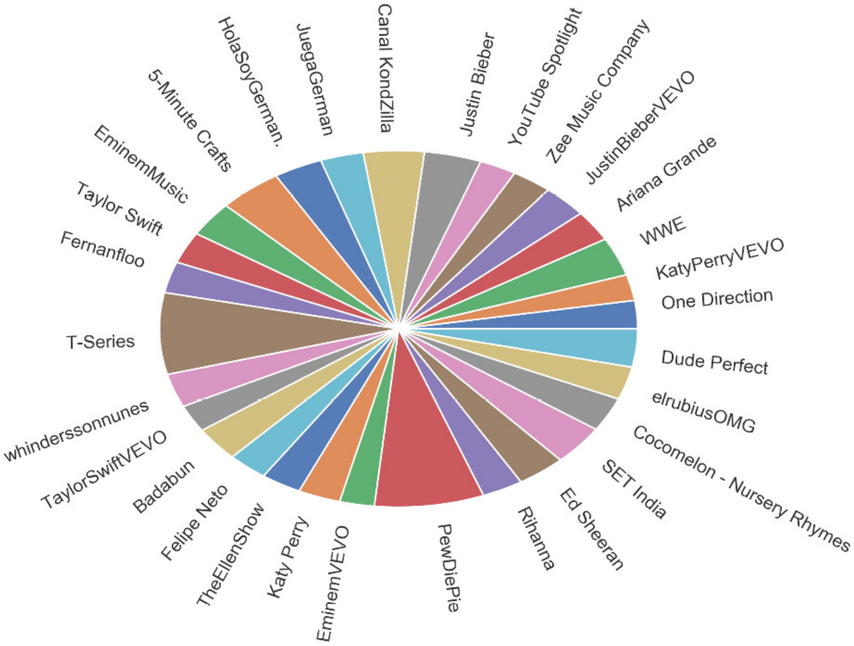

Figure 2.43: Pie chart showing Top 30 YouTubers

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

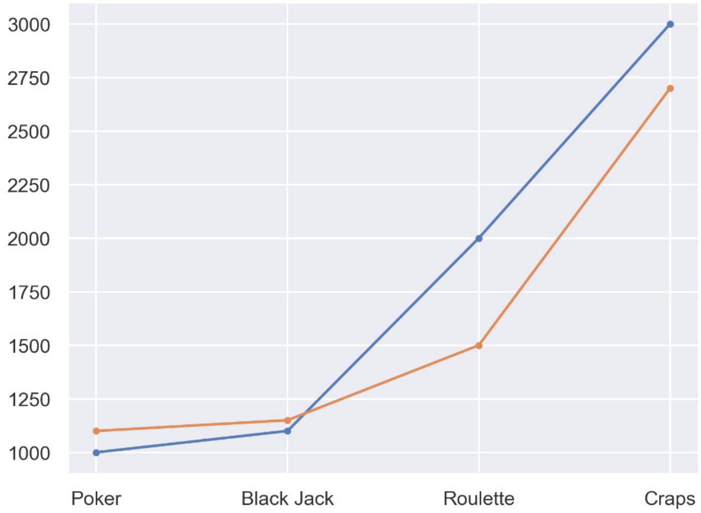

Figure 2.44: Line chart displaying casino data for two days

Note:

The solution for this activity can be found on page 277.

Summary

In this chapter, the most important visualizations were discussed. The visualizations were 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 for 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 think about something as a 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 given on what makes a good visualization. In the next chapter, we will dive into Matplotlib and create our own visualizations. We will cover all the plots that we have discussed in this chapter.