Module 4

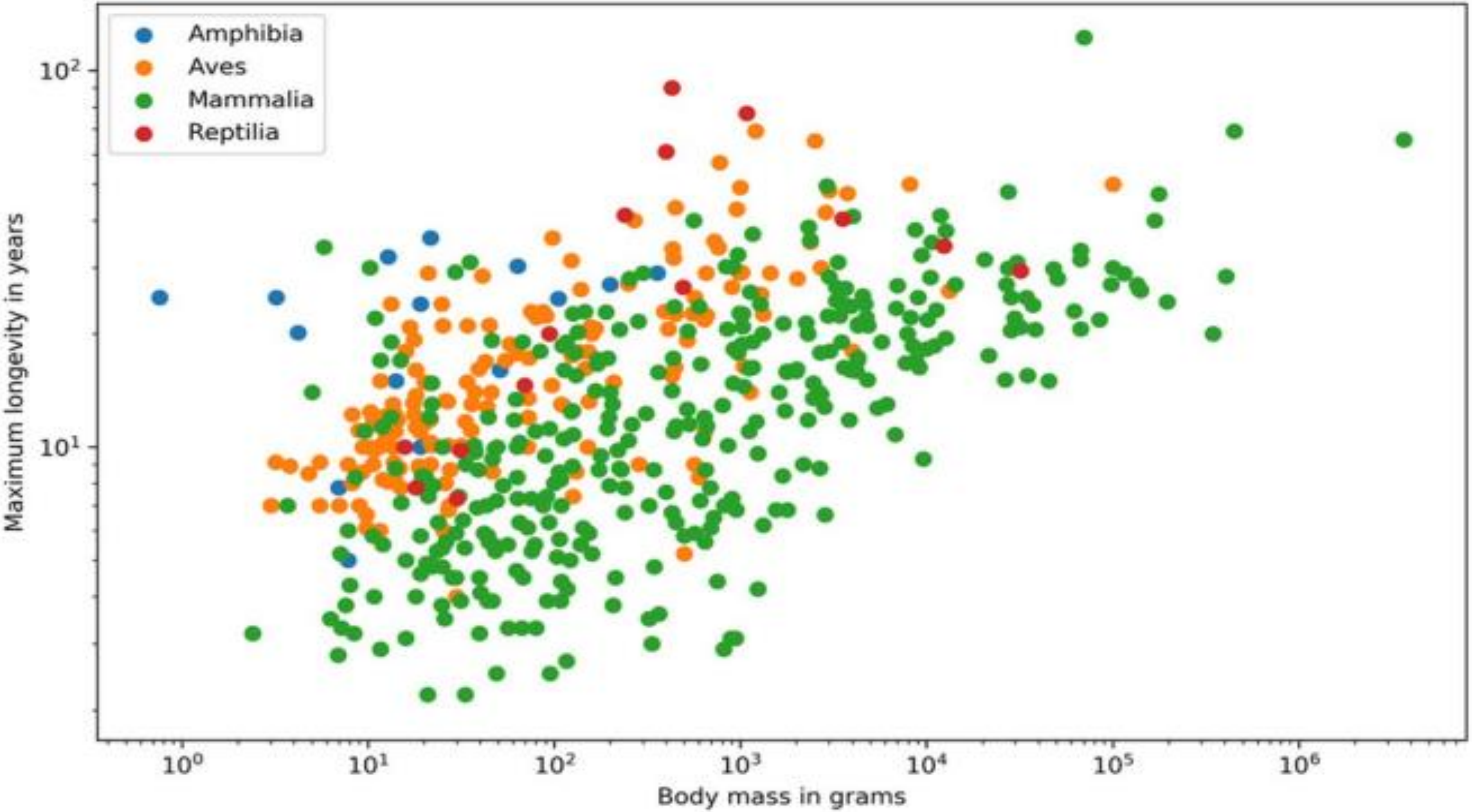
Data Visualization

Data Visualization and Data Exploration

- **Introduction**: Data Visualization, Importance of Data Visualization, Data Wrangling, Tools and Libraries for Visualization
- **Comparison Plots**: Line Chart, Bar Chart and Radar Chart
- **Relation Plots**: Scatter Plot, Bubble Plot , Correlogram and Heatmap;
- **Composition Plots**: Pie Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram;
- **Distribution Plots**: Histogram, Density Plot, Box Plot, Violin Plot;
- **Geo Plots**: Dot Map, Choropleth Map, Connection Map; What Makes a Good Visualization?

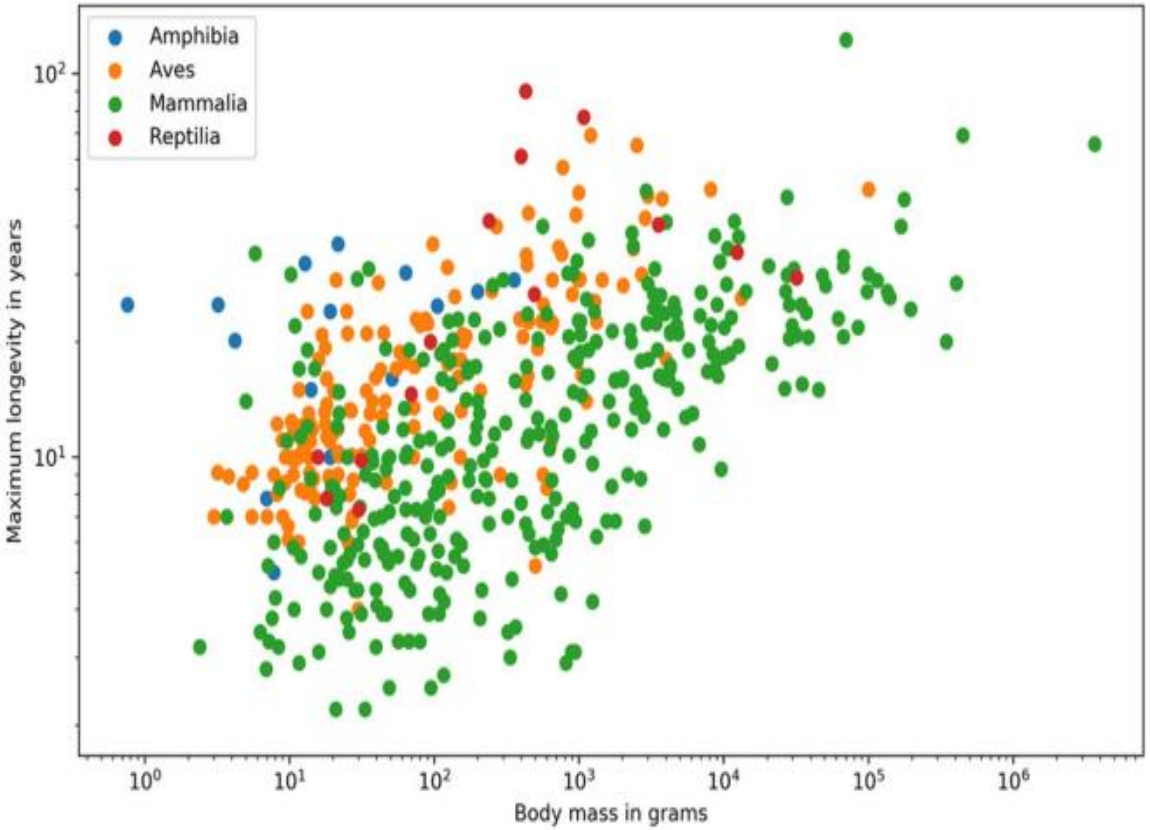
Introduction

- Humans are better at **understanding information visually** rather than through raw data.
- Python has become a key language for data analysis, featuring libraries such as **Pandas** for data manipulation, **NumPy** for analysis, and **Matplotlib** and **Bokeh** for visualization.
- While computers and smartphones **store data in digital formats**, data representation is about how this ***data is stored, processed, and transmitted***.
- Effective data representation **can tell a story and convey key findings**, enhancing the value of the data by making it more understandable.
- Representations help transform raw data into ***meaningful information, providing clearer and more concise insights***. This transformation is essential because information derived from data is what holds true value.


Amphibia
Aves
Mammalia
Reptilia
Maximum longevity in years
10^2
10^1
Body mass in grams
10^0
10^1
10^2
10^3
10^4
10^5
10^6

Importance of Data Visualization

- Instead of just looking at **data in Excel columns, using visualization** helps us understand the data better.
- For example, a scatter plot can show patterns, like the relationship between **body mass and lifespan** in different animal classes, where we can see a **positive correlation**.



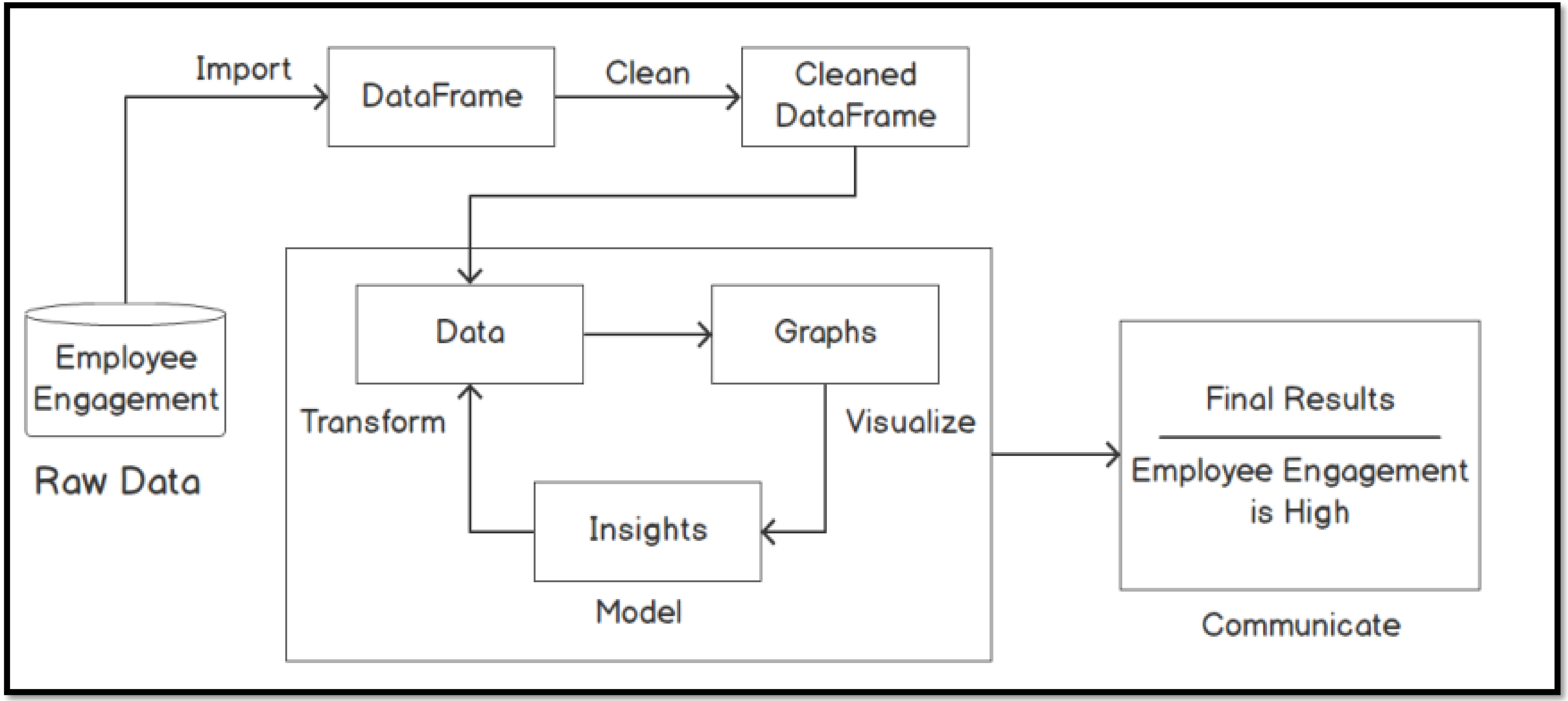
Advantages of Visualizing Data

- 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 converting raw data into a usable format for different tasks.
- It involves enhancing(augmenting), cleaning, filtering, standardizing, and enriching data to make it suitable for use, particularly for data visualization.

Data wrangling process to measure employee engagement



In relation to the preceding figure, the data wrangling process involves the following steps:

1. The **raw Employee Engagement data** is collected.
2. The data is imported into a **DataFrame and cleaned**.
3. The **cleaned data is then transformed** into graphs to derive findings.
4. Finally, the **data is analyzed** to communicate the final results.

Employee engagement can be measured using raw data from:

- **Feedback surveys**
- **Employee tenure**
- **Exit interviews**
- **One-on-one meetings**

This data is cleaned and transformed into graphs based on parameters such as:

- **Referrals**
- **Trust in leadership**
- **Promotion opportunities**

Tools and Libraries for Visualization

- Non-coding tools like **Tableau** offer a user-friendly way to explore data.
- **Python, MATLAB, and R** are widely used in data analytics.
- **Python** is the most popular language for data visualization in industry.
- **Python's** ease of use and speed in data manipulation and visualization, along with its **extensive library support, make it ideal for data visualization**.

Important Links:

- **Python** (https://www.python.org/)
- **MATLAB** (https://www.mathworks.com/products/matlab.html),
- **R** (https:// www.r-project.org),
- **Tableau** (https://www.tableau.com)

Overview of Statistics

Statistics vs Probability

- **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.**

Probability Distribution

- A function that assigns probabilities to every possible event.
- It categorizes into discrete (e.g., rolling a die) and continuous (e.g., driving time distribution).

Fig: Discrete probability distribution for die rolls

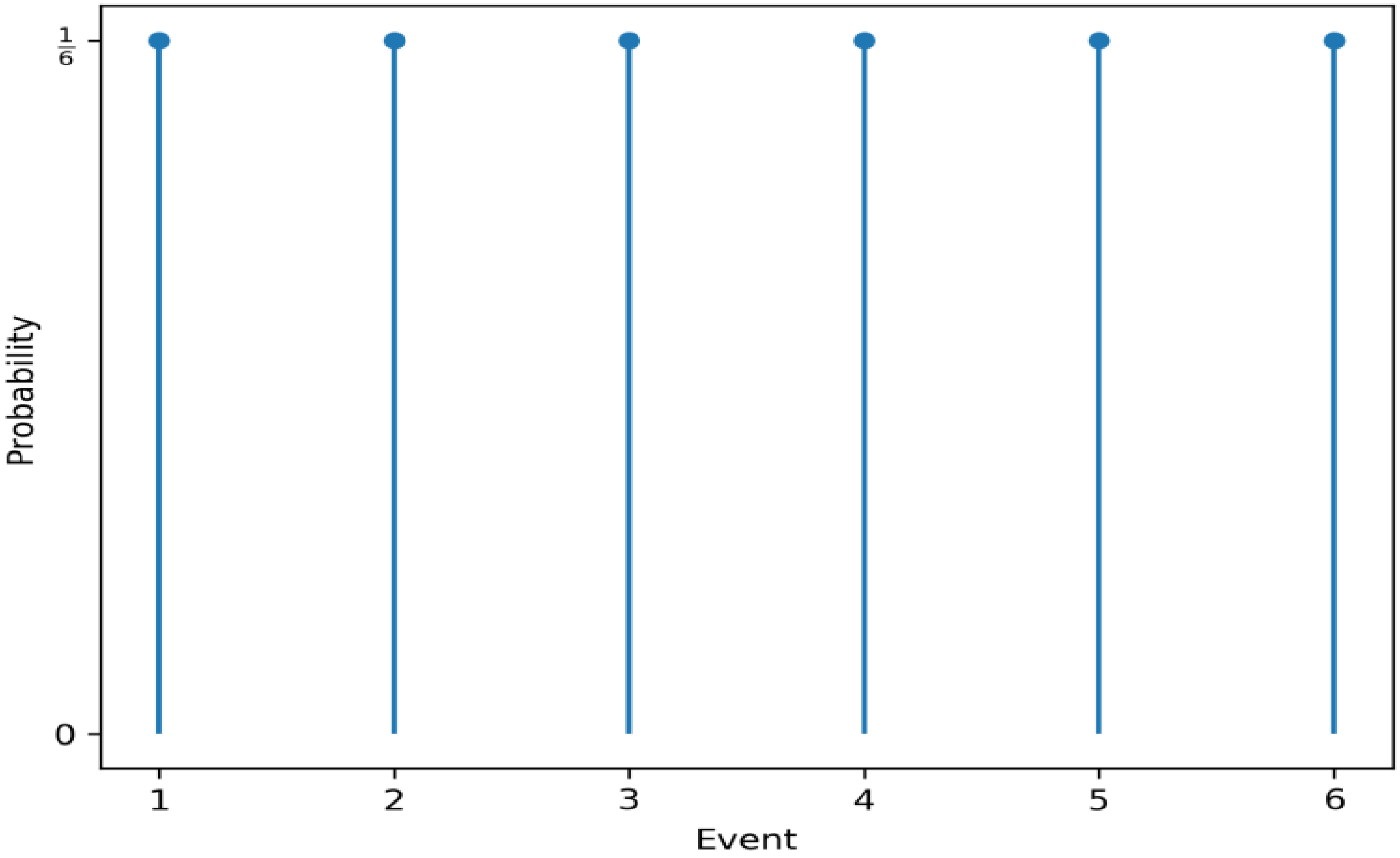
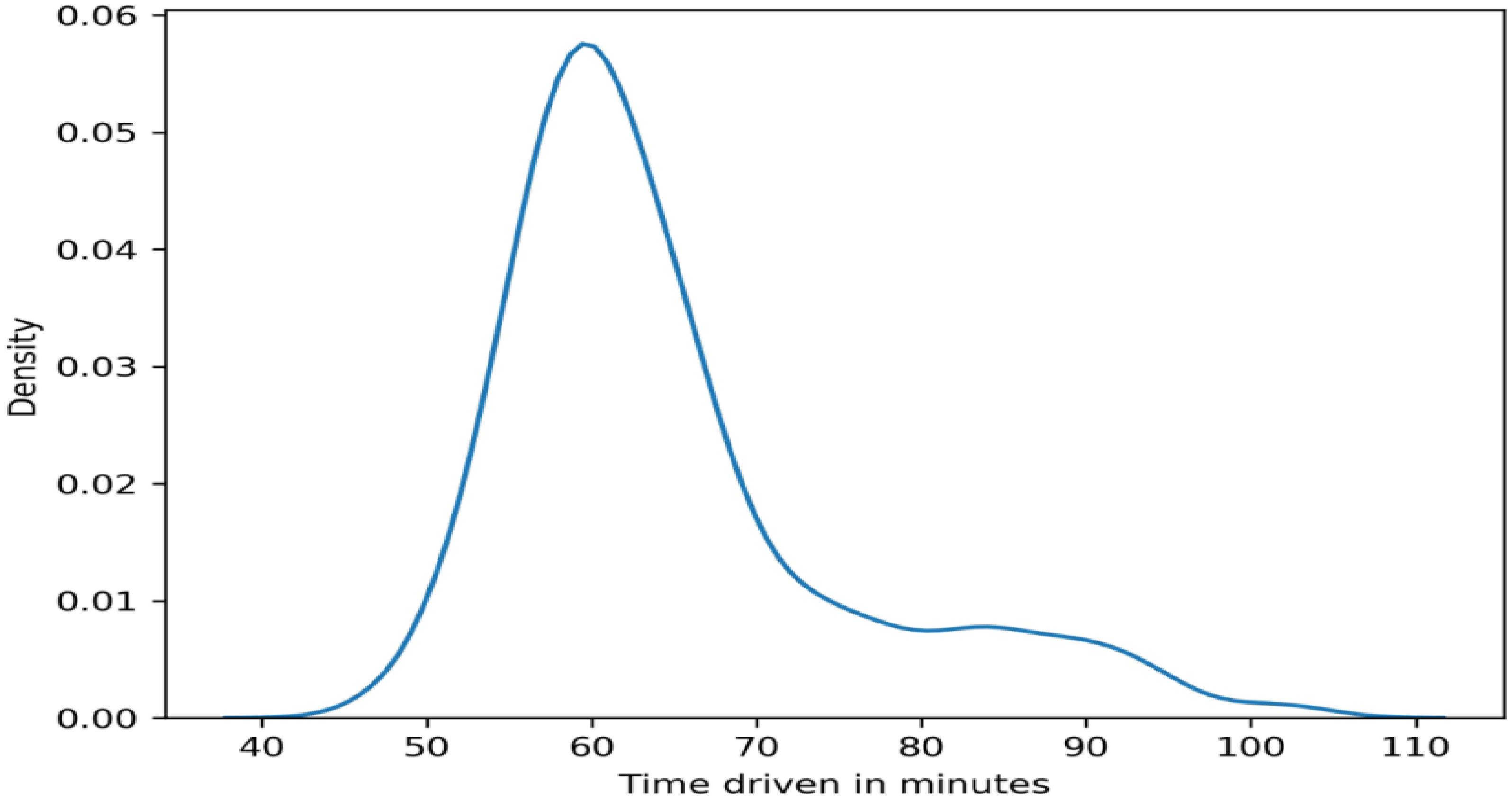


Fig: Continuous probability distribution for the time taken to reach home



Measures of Central Tendency:

- **Mean**: Average obtained by summing values and dividing by the number of observations.

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

- **Median**: Middle value in ordered data; less sensitive to outliers.

- **Mode**: Most frequent value; can have multiple modes if frequencies are equal.

Example : For instance, after rolling a die 10 times, the outcomes were: 4, 5, 4, 3, 4, 2, 1, 1, 2, and 1.

- **Mean**: Calculated by summing all outcomes and dividing by the number of rolls: (4 + 5 + 4 + 3 + 4 + 2 + 1 + 1 + 2 + 1) / 10 = 2.7.
- **Median**: Arranging the outcomes in ascending order: 1, 1, 1, 2, 2, 3, 4, 4, 4, 5. Since there are an even number of rolls, the median is the average of the two middle values: (2 + 3) / 2 = 2.5.
- **Modes**: The most frequent outcomes are 1 and 4, making them the modes.

Measures of Dispersion:

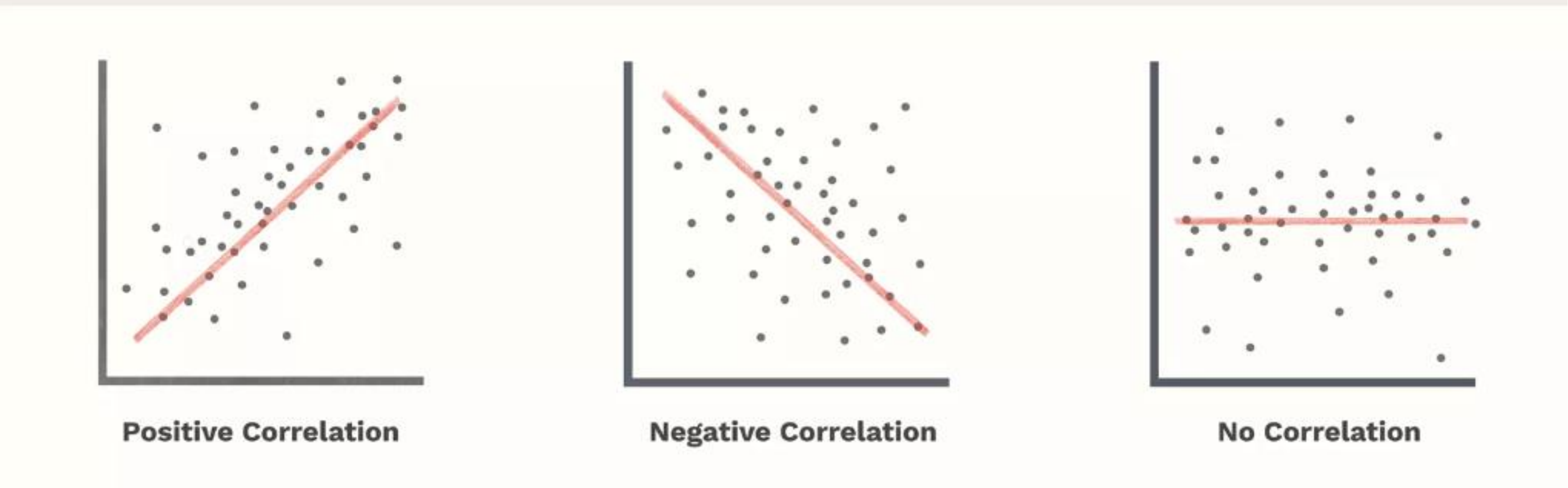
- **Variance**: Average of squared deviations from the mean; indicates spread.

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

- **Standard Deviation**: Square root of variance.
- **Range**: Difference between the largest and smallest values.
- **Interquartile Range**: Difference between the upper and lower quartiles. 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: Describes the relationship between two variables:

- **Positive Correlation**: Variables move in the same direction.
- **Negative Correlation**: Variables move in opposite directions.
- **Zero Correlation**: No relationship between variables

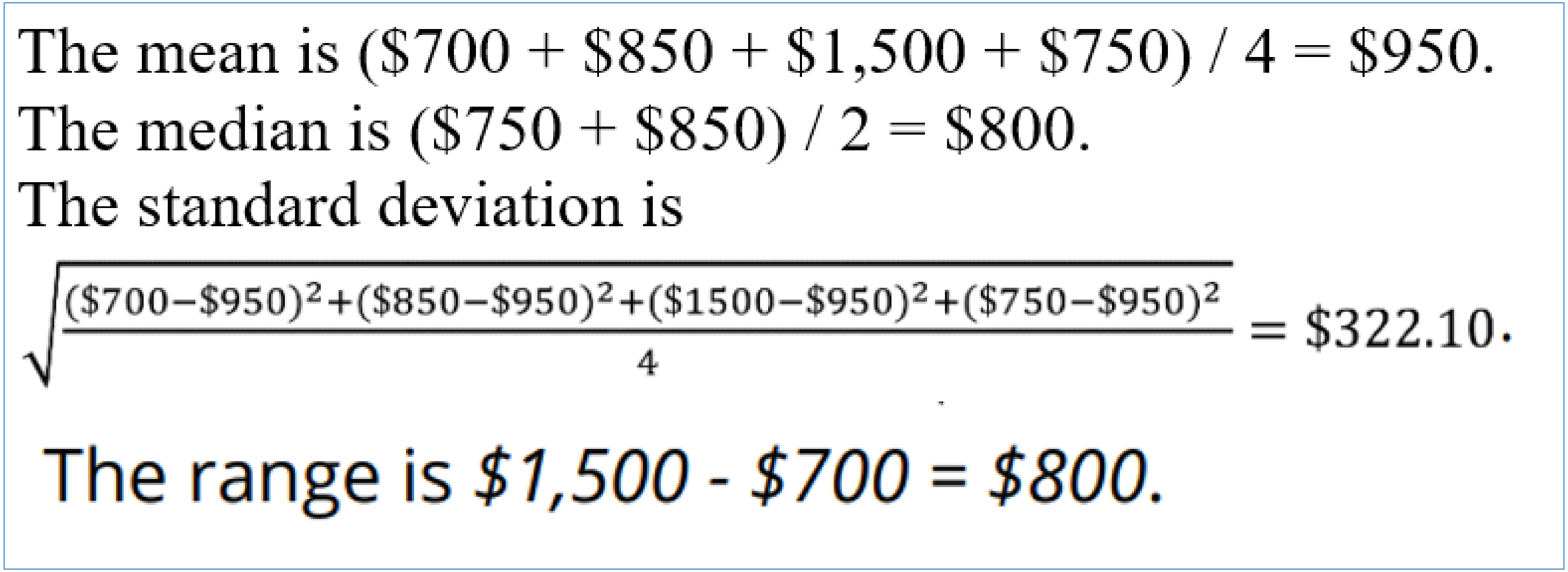


Correlation vs. Causation:

- Correlation indicates a relationship between variables, while causation explains how one event causes another.
- **Example:** Ice cream sales and drowning deaths may show a correlation, but it doesn't mean one causes the other.
- **Third Variable:** Factors like temperature could influence both ice cream sales and swimming, which may actually explain the increase in drowning deaths.

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.
- Calculate the following statistical measures to help us make a decision:
  - Mean:
  - Median:
  - Standard Deviation:
  - Range :



The mean is (\$700 + \$850 + \$1,500 + \$750) / 4 = \$950.

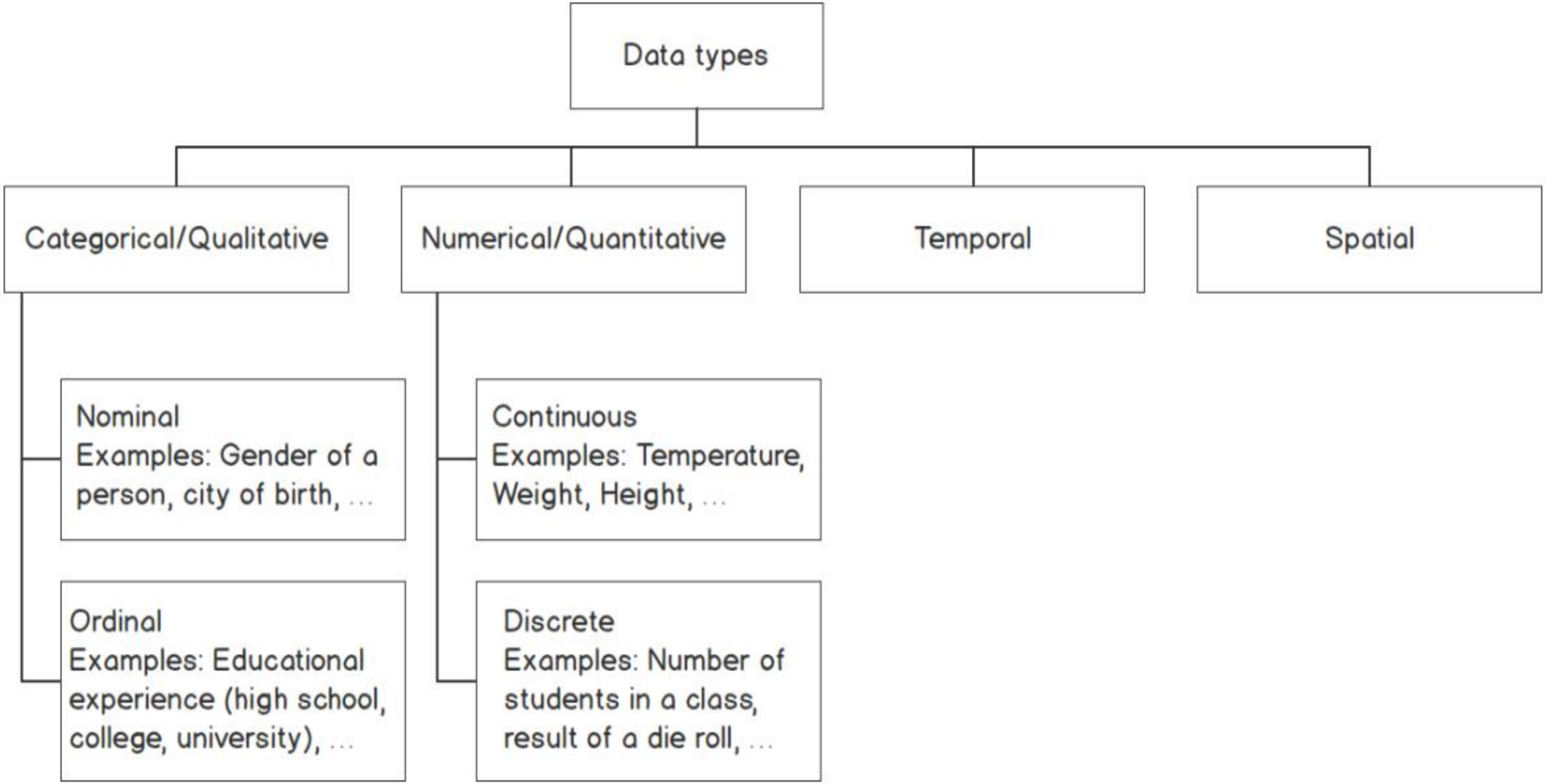
The median is (\$750 + \$850) / 2 = \$800.

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*.

Types of Data



Data types

Categorical/Qualitative

Numerical/Quantitative

Temporal

Spatial

Nominal
Examples: Gender of a person, city of birth, …

Ordinal
Examples: Educational experience (high school, college, university), …

Continuous
Examples: Temperature, Weight, Height, …

Discrete
Examples: Number of students in a class, result of a die roll, …

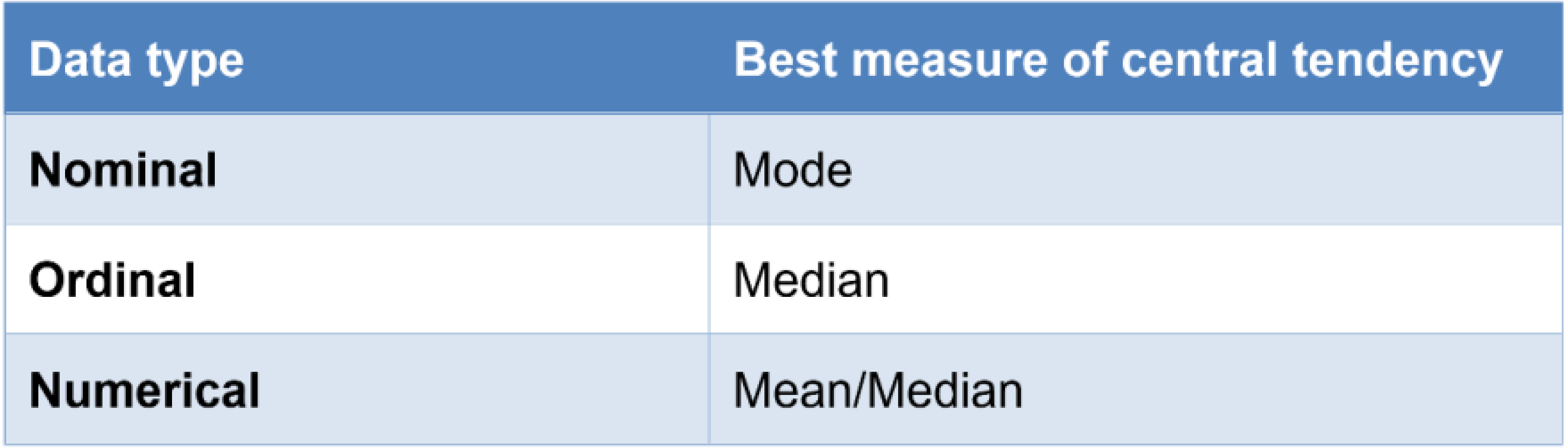
Summary Statistics

Statistical summary refers to a set of **descriptive statistics** that provide an **overview of the characteristics** of a dataset. These statistics typically include measures such as:

- **Central Tendency:**
  - **Mean** (average)
  - **Median** (middle value)
  - **Mode** (most frequent value)
- **Dispersion or Variability**:
  - **Range** (difference between the maximum and minimum values)
  - **Variance** (average of the squared differences from the mean)
  - **Standard deviation** (square root of the variance, measures how spread out the values are)
- **Shape of the Distribution**:
  - **Skewness** (measures the asymmetry of the distribution)
  - **Kurtosis** (measures the "tailedness" of the distribution)
- **Other Measures**:
  - **Count** (number of observations)
  - **Percentiles** (values below which a given percentage of observations fall)

Summary Statistics

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

Comparison Plots

- **Comparison plots** are charts used to compare multiple variables or variables over time.
- **Line charts** are ideal for visualizing variables over time,
- **Bar charts (or column charts)** are best for comparing items.
- **Vertical bar charts** are suitable for fewer than **10 time points**.
- **Radar charts**, also known as spider plots, are effective for visualizing **multiple variables across multiple groups**.

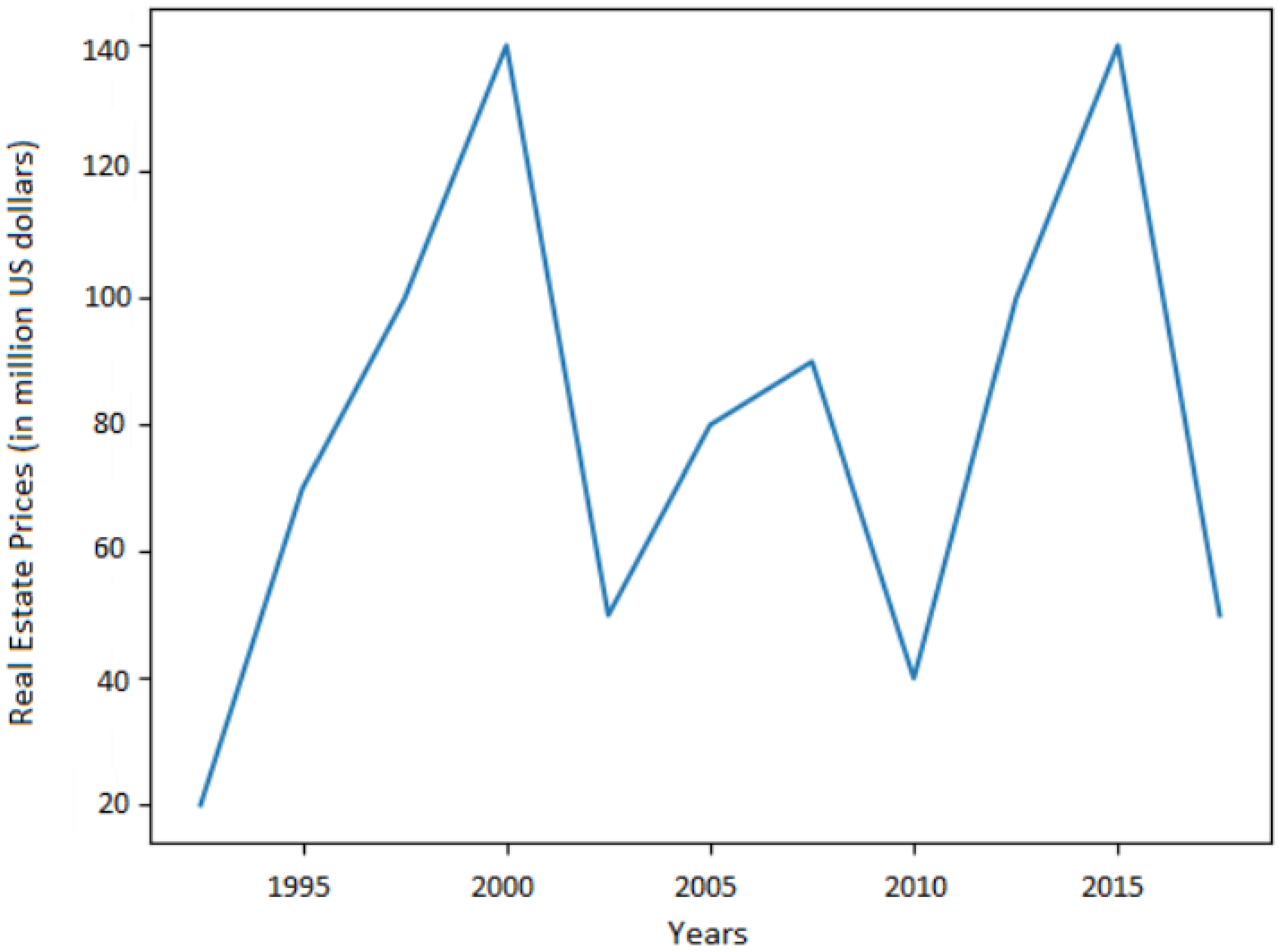
Line Chart

- 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

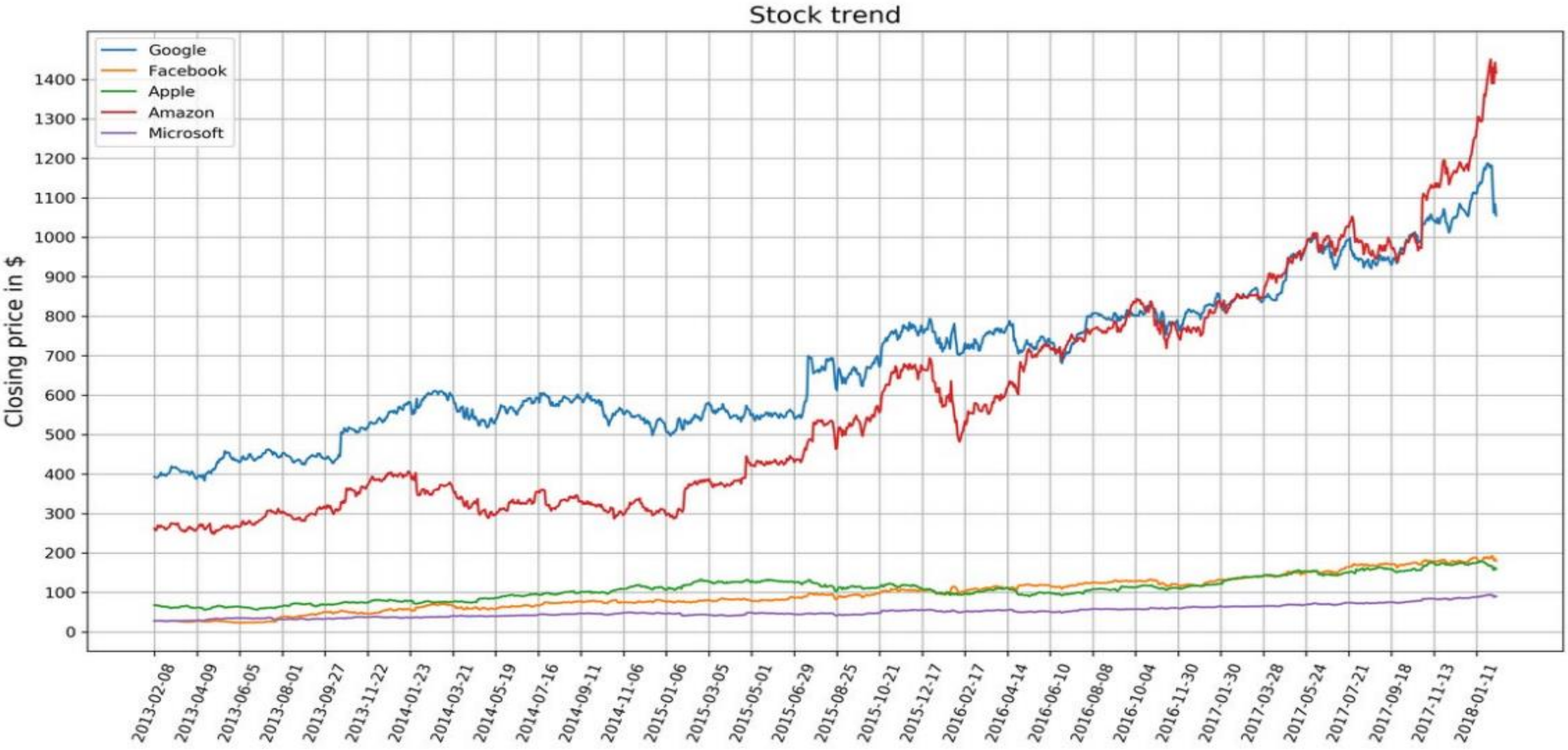
Uses

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

Example: Line chart for a single variable



Example2: The following figure is a multiple-variable line chart that compares the stock-closing prices for Google, Facebook, Apple, Amazon, and Microsoft.



Design Practices

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

Bar Chart

- In a bar chart, the bar length encodes the value.
- There are two variants of bar charts:
  1. **vertical bar charts**
  2. **horizontal bar charts.**

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**.
- 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

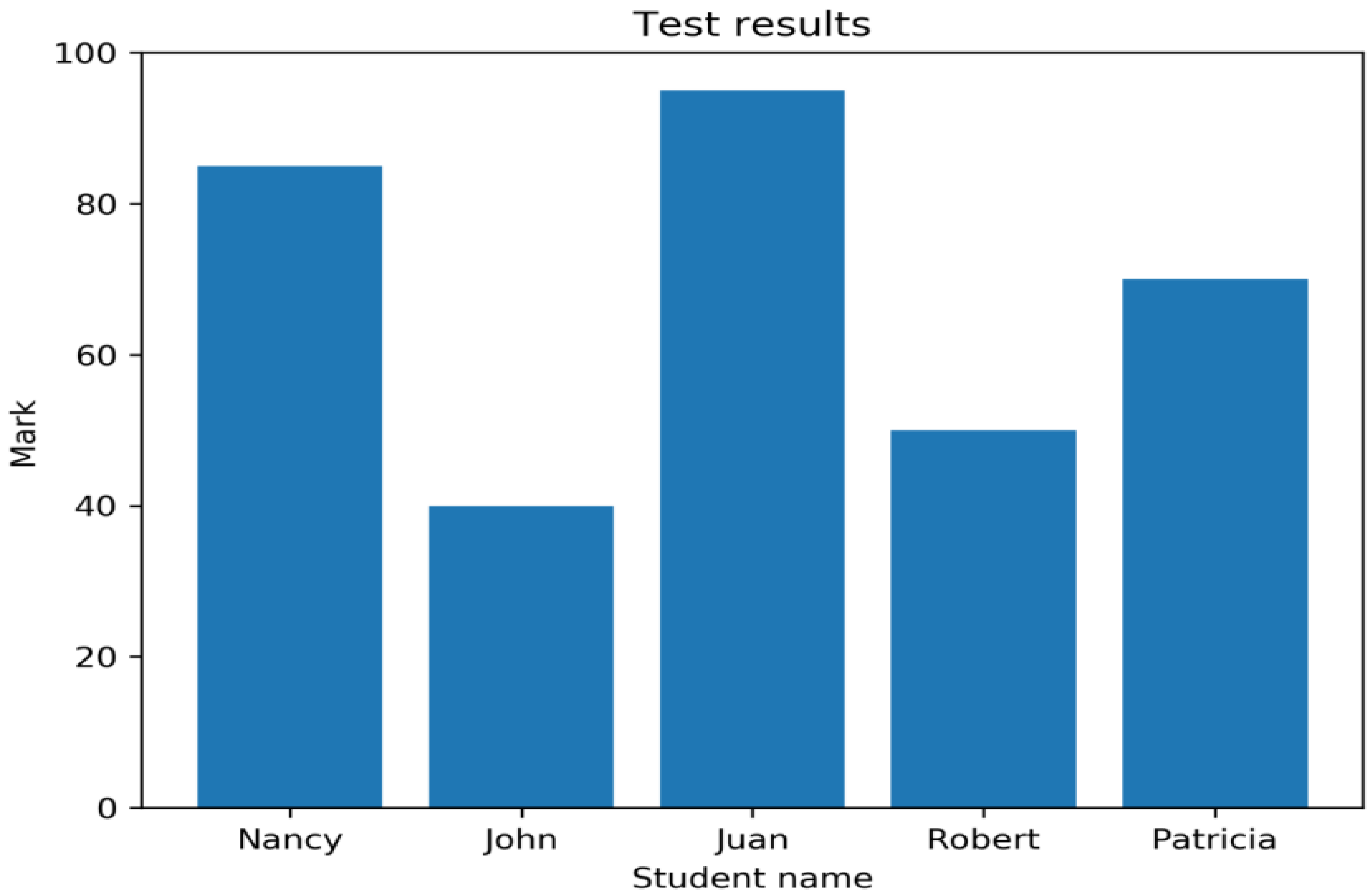


Figure 2.3: Vertical bar chart using student test data

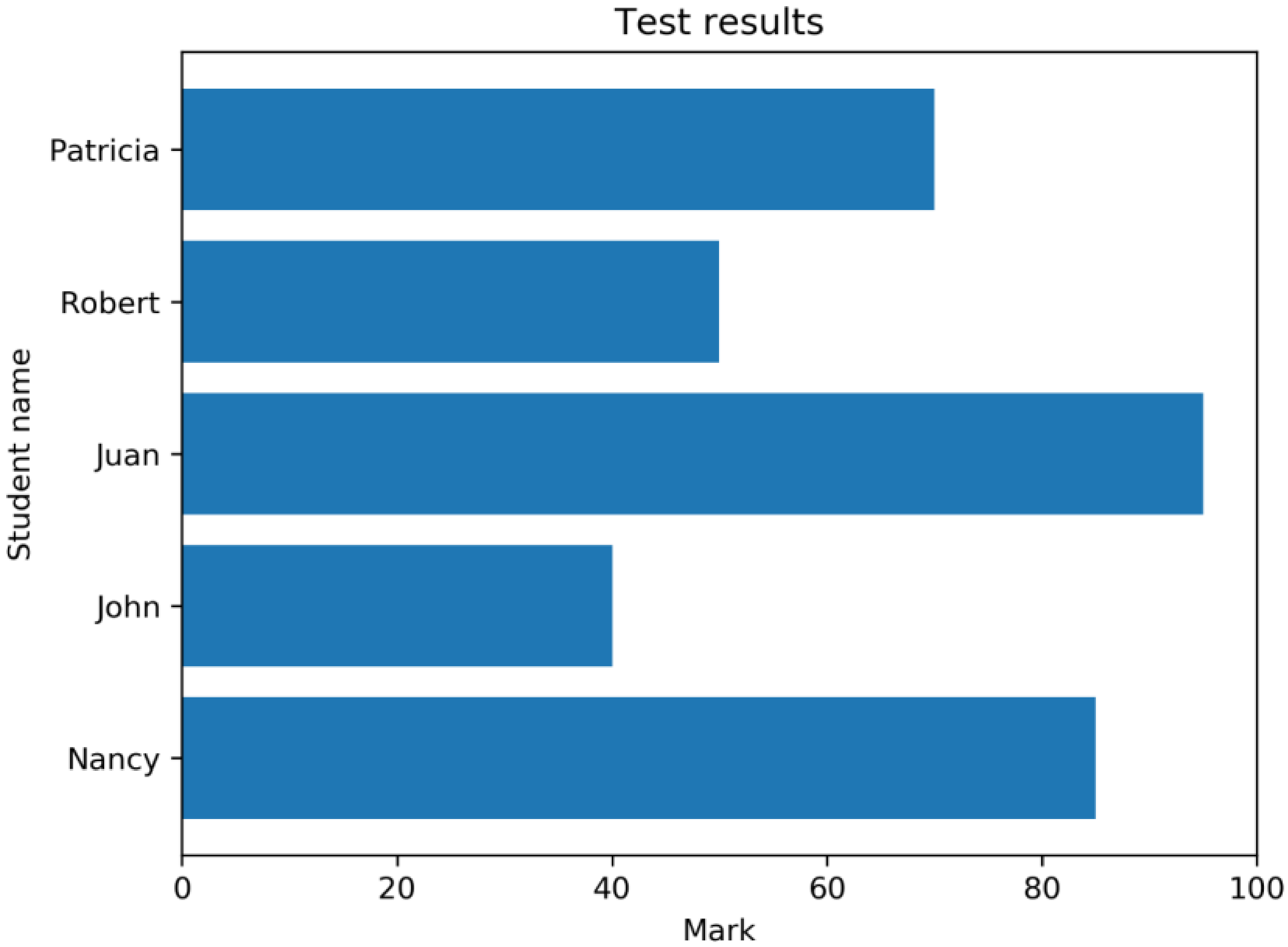


Figure 2.4: Horizontal bar chart using student test data

Example: The following diagram compares movie ratings, giving two different scores.

1. The **Tomatometer** is the percentage of approved critics who have given a positive review for the movie.
2. The **Audience Score** is the percentage of users who have given a score of **3.5** or higher out of **5**.

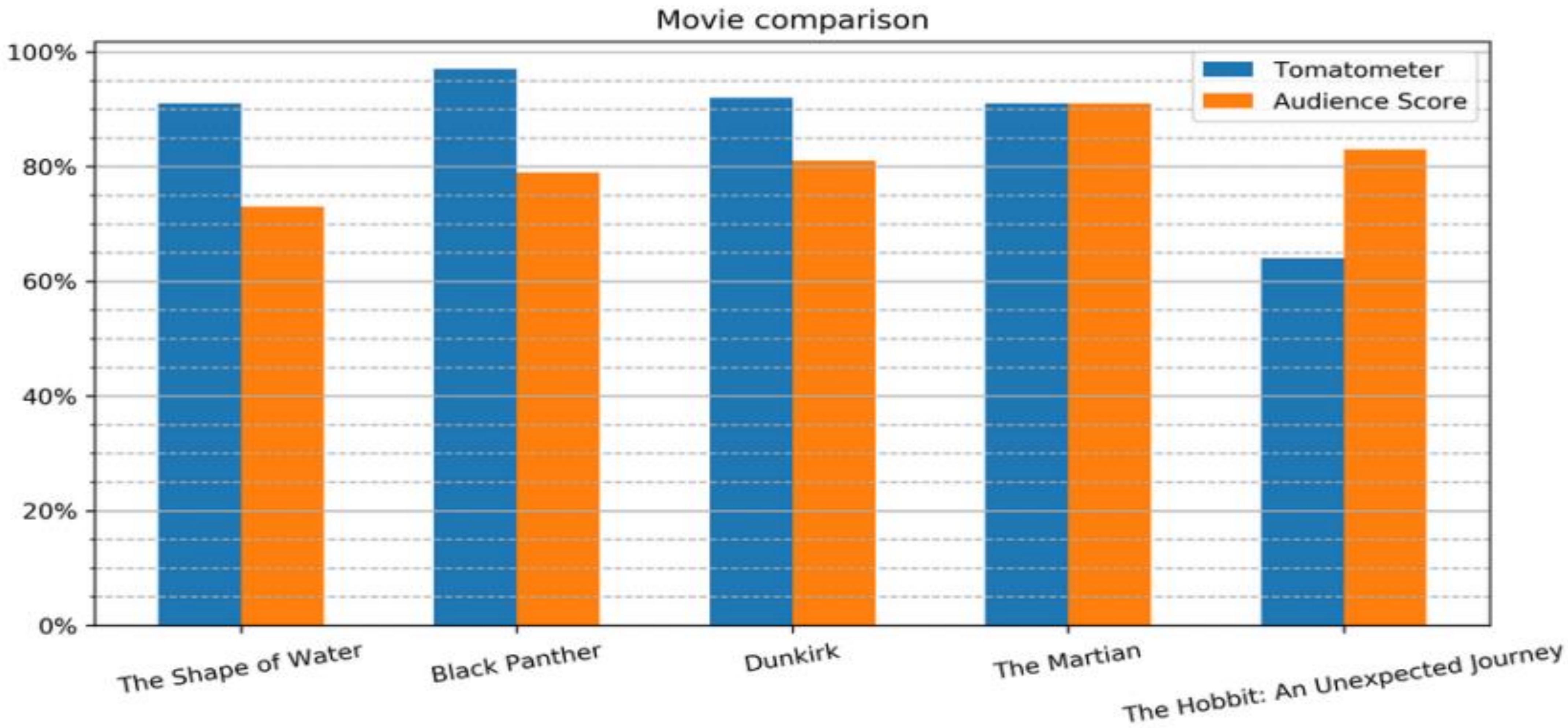
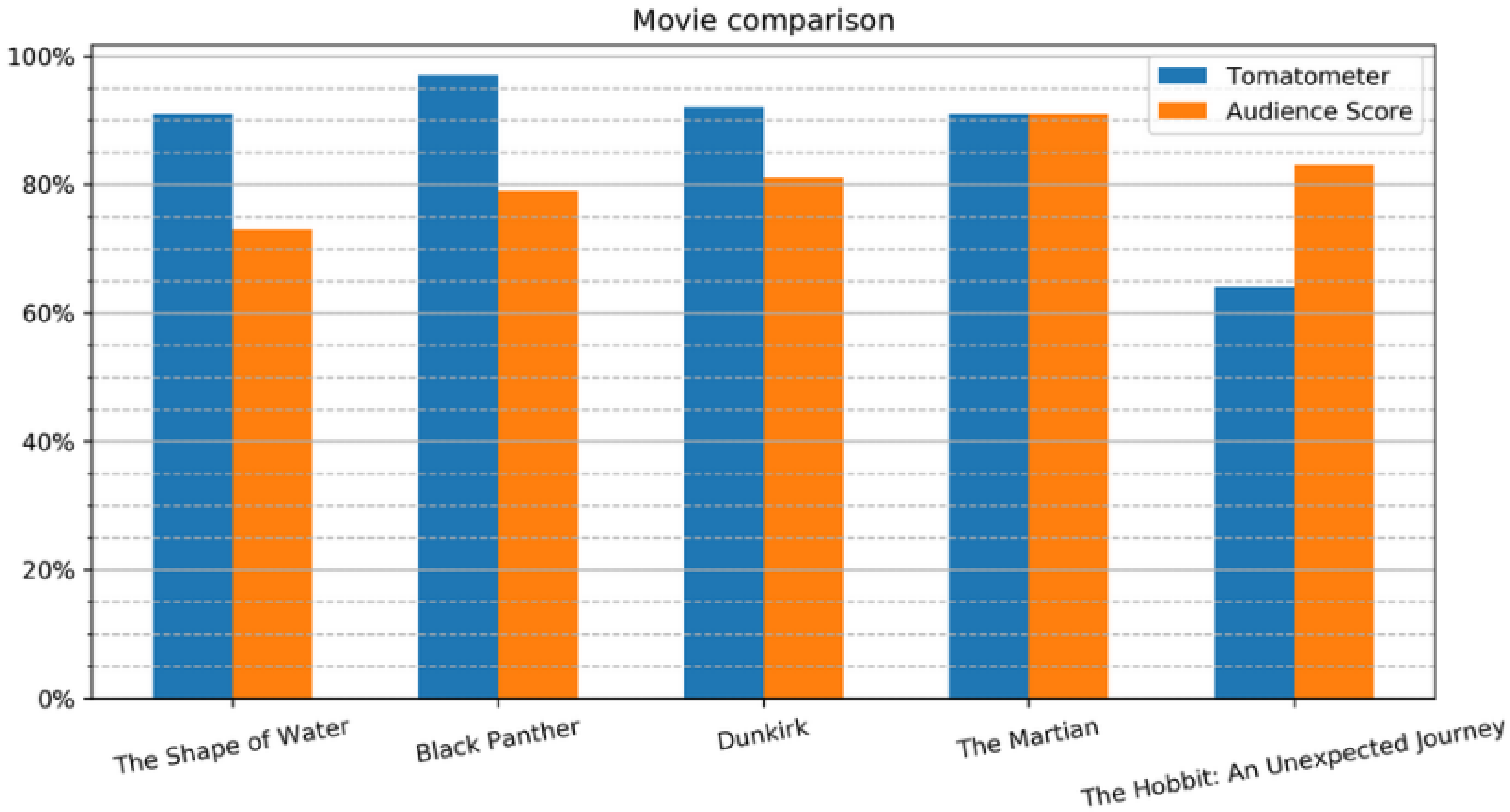


Figure 2.5: Comparative bar chart



Inference:

- **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**

Design Practices

- The axis corresponding to the **numerical variable should start** at zero.
- **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.

Types of Plots

1. **Comparison Plots**: Line Chart, Bar Chart and Radar Chart
2. **Relation Plots**: Scatter Plot, Bubble Plot , Correlogram and Heatmap;
3. **Composition Plots**: Pie Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram;
4. **Distribution Plots**: Histogram, Density Plot, Box Plot, Violin Plot;
5. **Geo Plots**: Dot Map, Choropleth Map, Connection Map; What Makes a Good Visualization?

Radar Charts

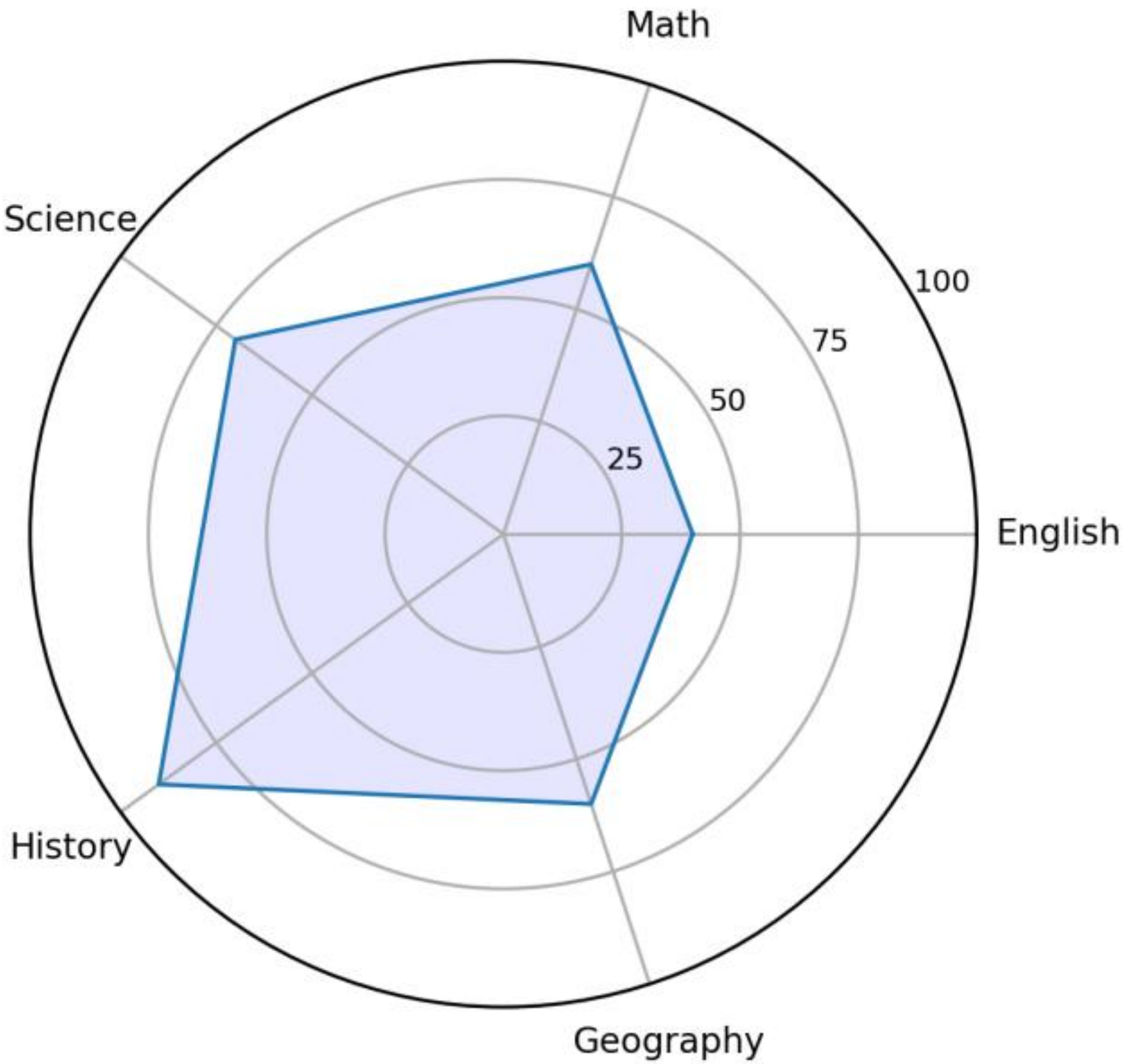
- **Radar charts** (also known as **spider or web charts**) visualize multiple variables with each variable plotted on its own axis, resulting in a polygon.
- All axes are arranged radially, starting at the **center** with equal distances between **one another, and have the same scale**.

Uses

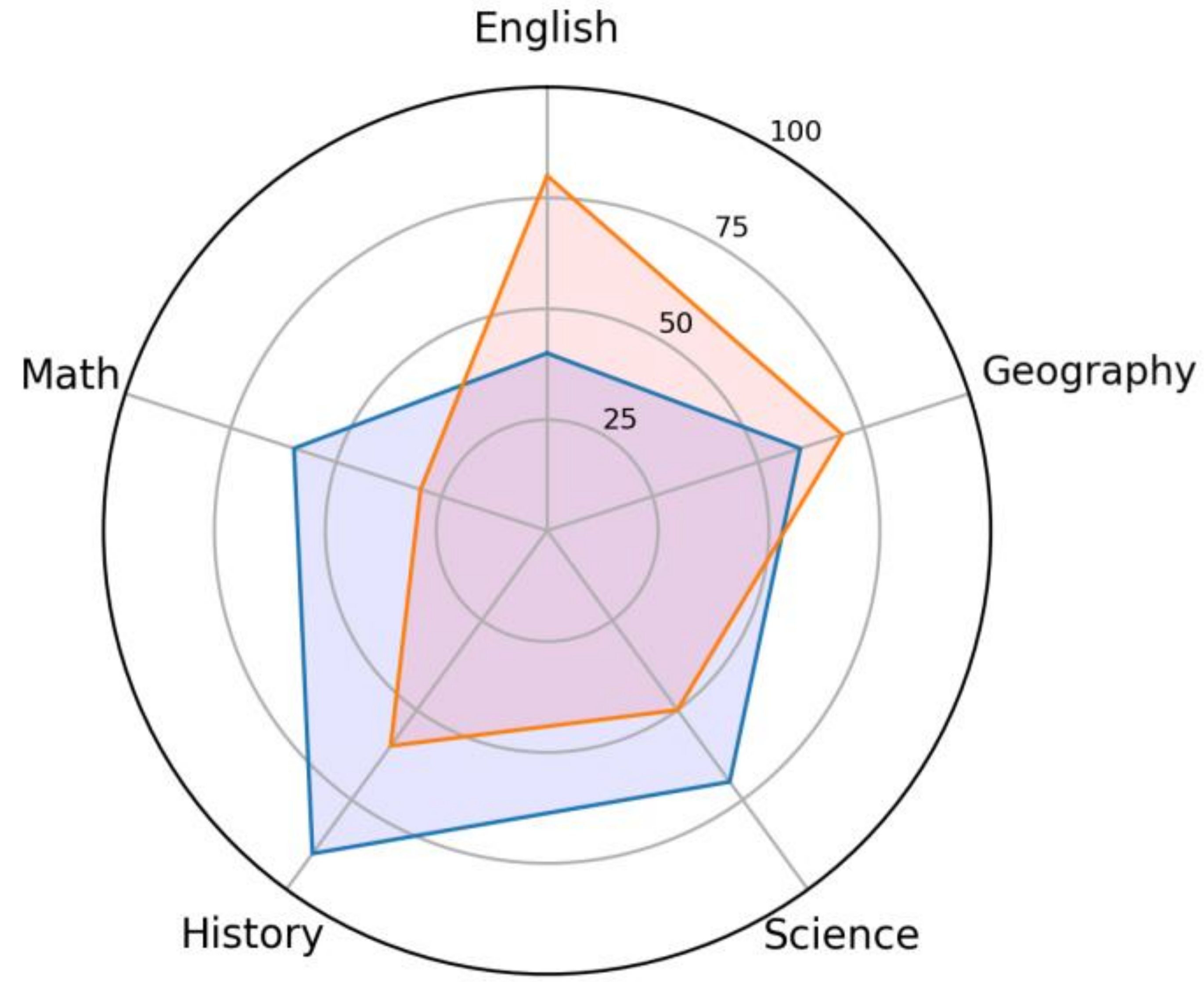
- **Radar charts** are great for comparing multiple quantitative variables for a **single group or multiple groups**.
- They are also useful for showing which variables score high or low within a **dataset, making them ideal for visualizing performance**.

Examples

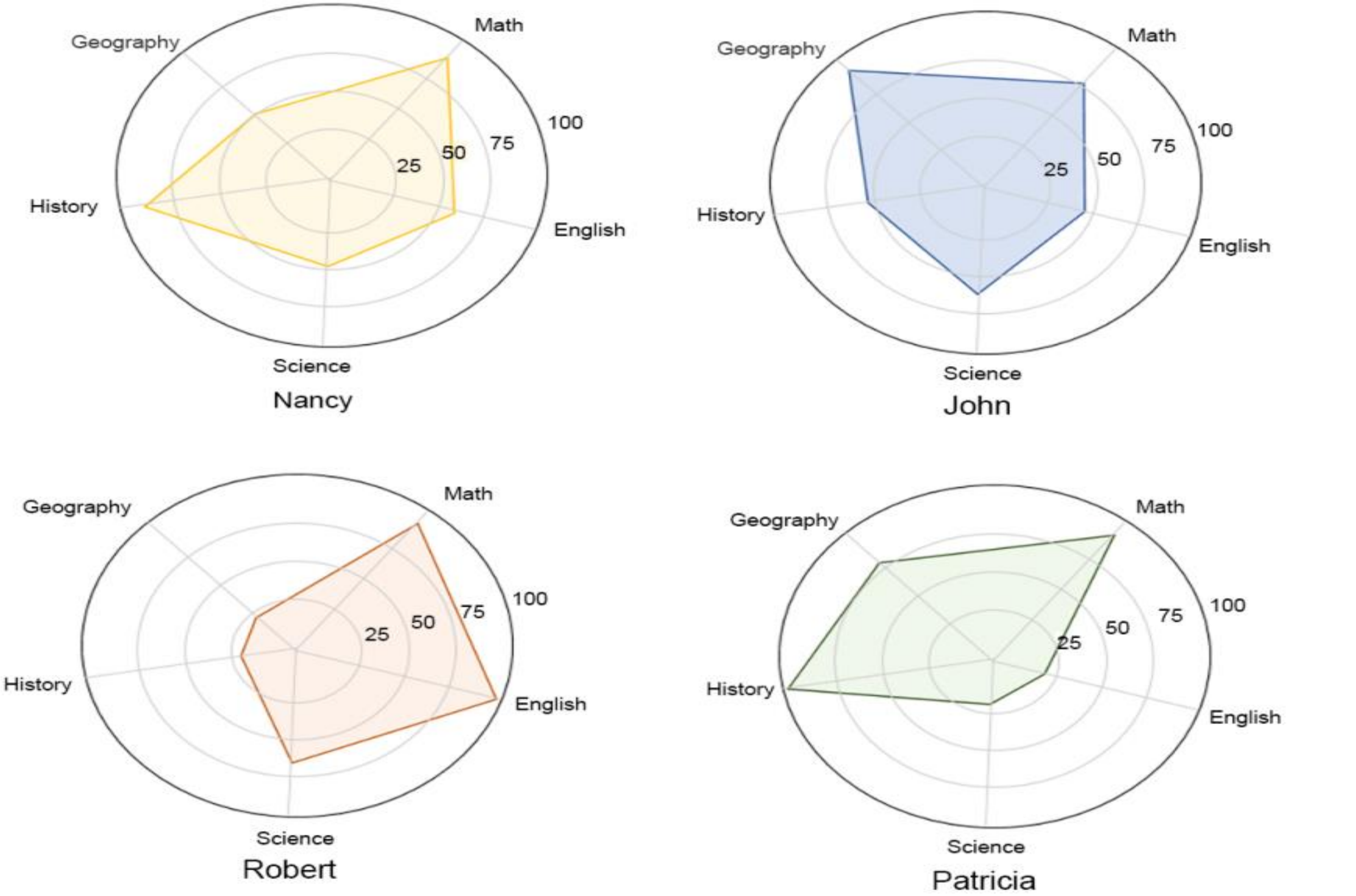
Radar chart for **one variable (student)**



Radar chart for **two variables (two students)**



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.

Summary

- **Line charts** are great for comparing **something over time**, whereas **bar charts** are for **comparing different items**.
- **Radar charts** are best suited for visualizing multiple variables for multiple groups.

Activity : 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.

Activity : Employee Skill Comparison

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**.
3. **Which is the better chart for this task in your opinion, and why?**
4. **What could be improved in the respective visualizations?**

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

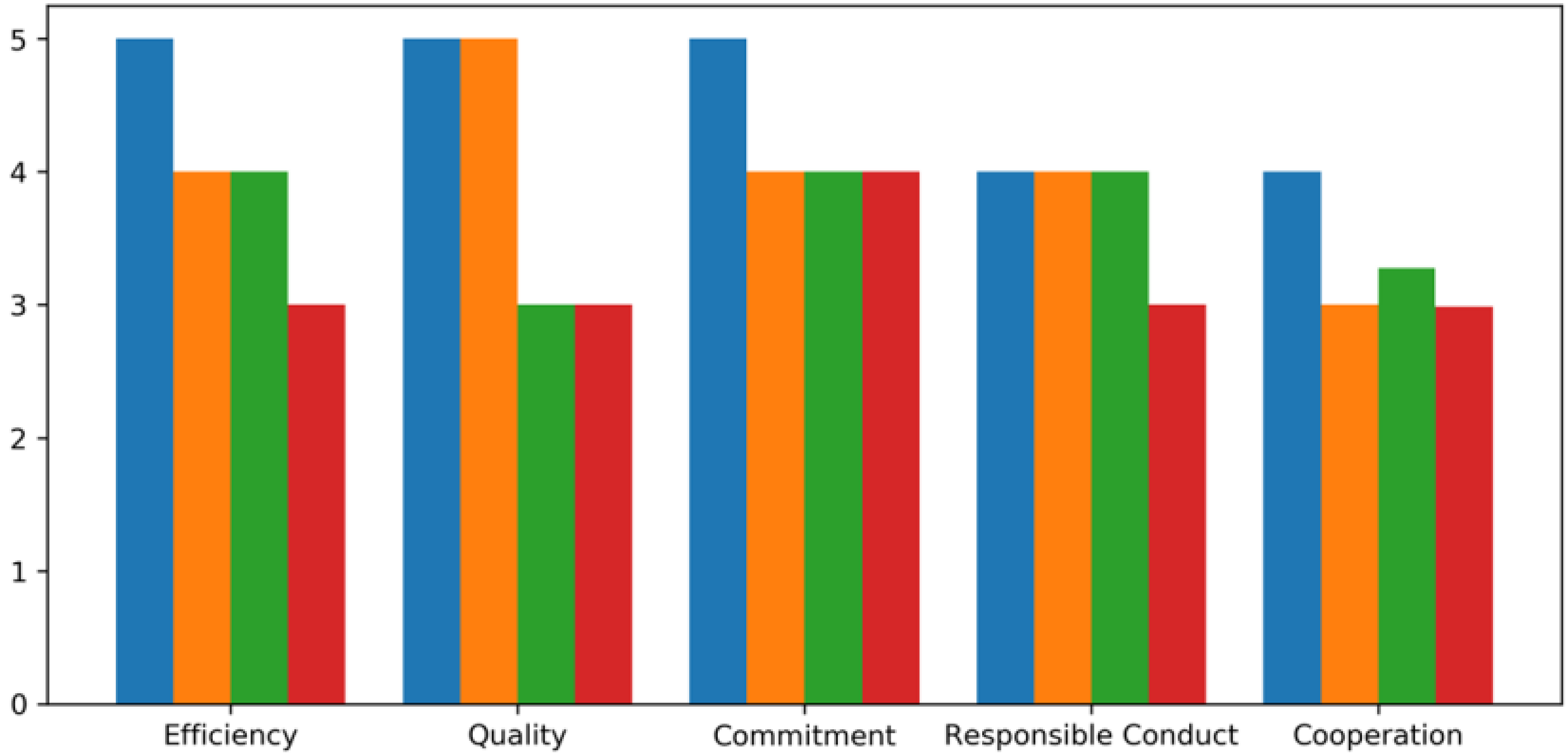
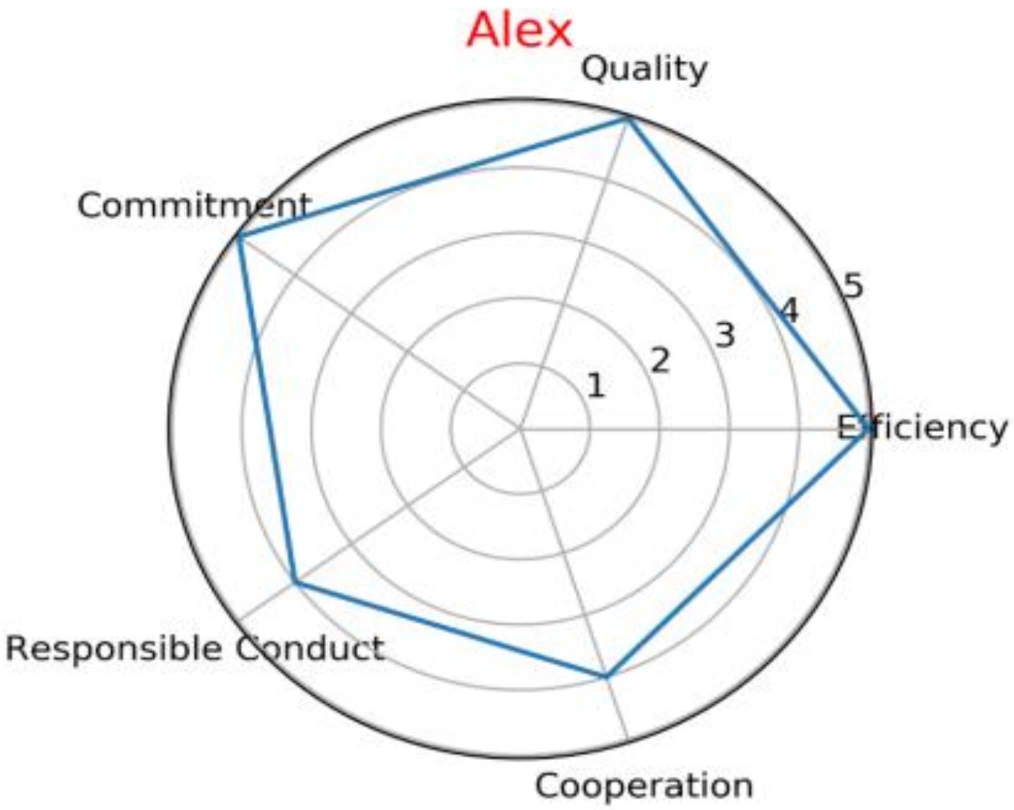
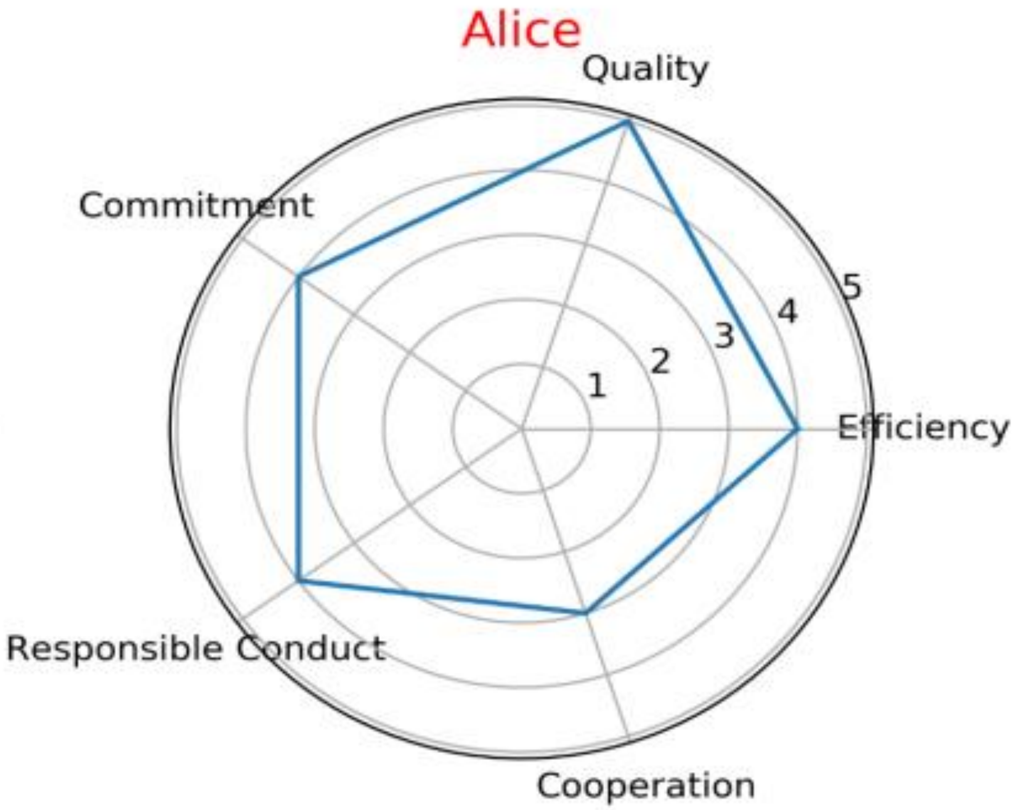


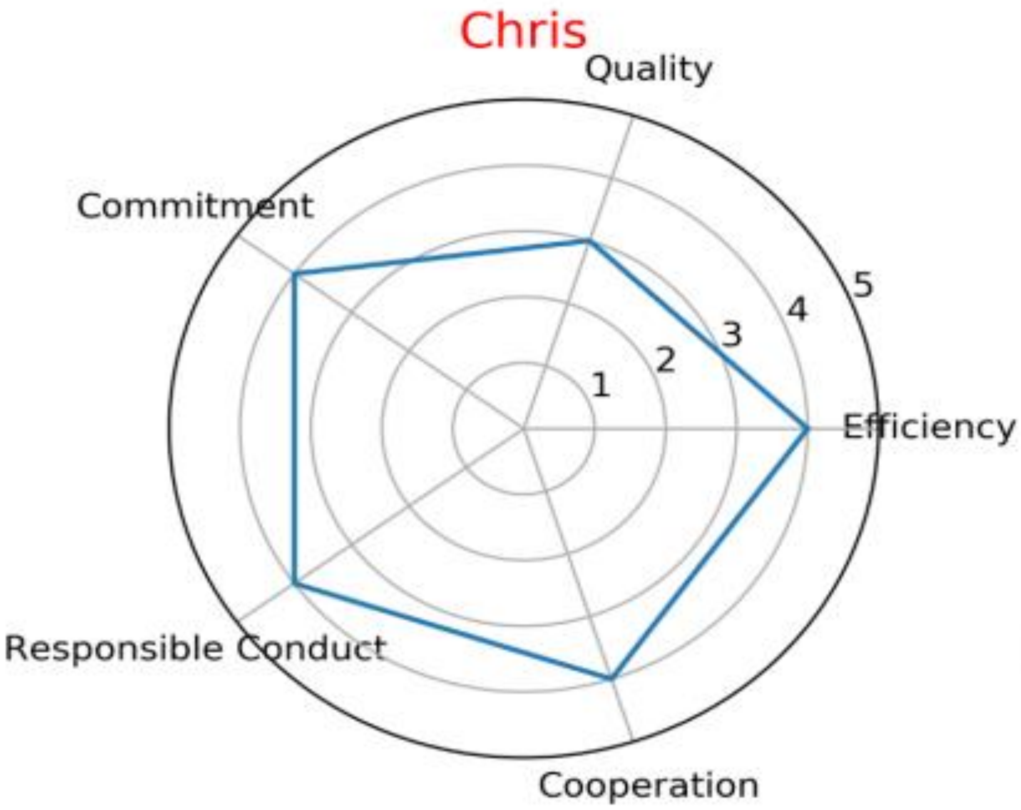
Figure 2.9: Employee skills comparison with a bar chart

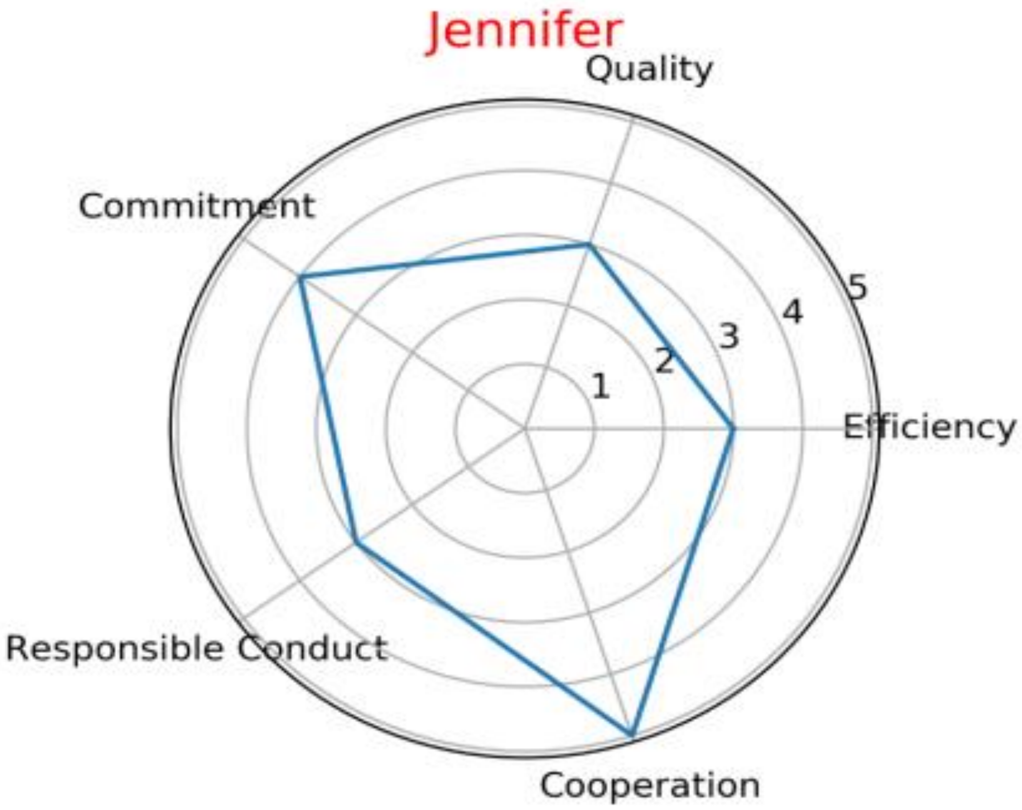
Employee skills comparison with a radar chart





Employee skills comparison with a radar chart





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**.

Example1: The following diagram shows a scatter plot of height and weight of persons belonging to a single group:

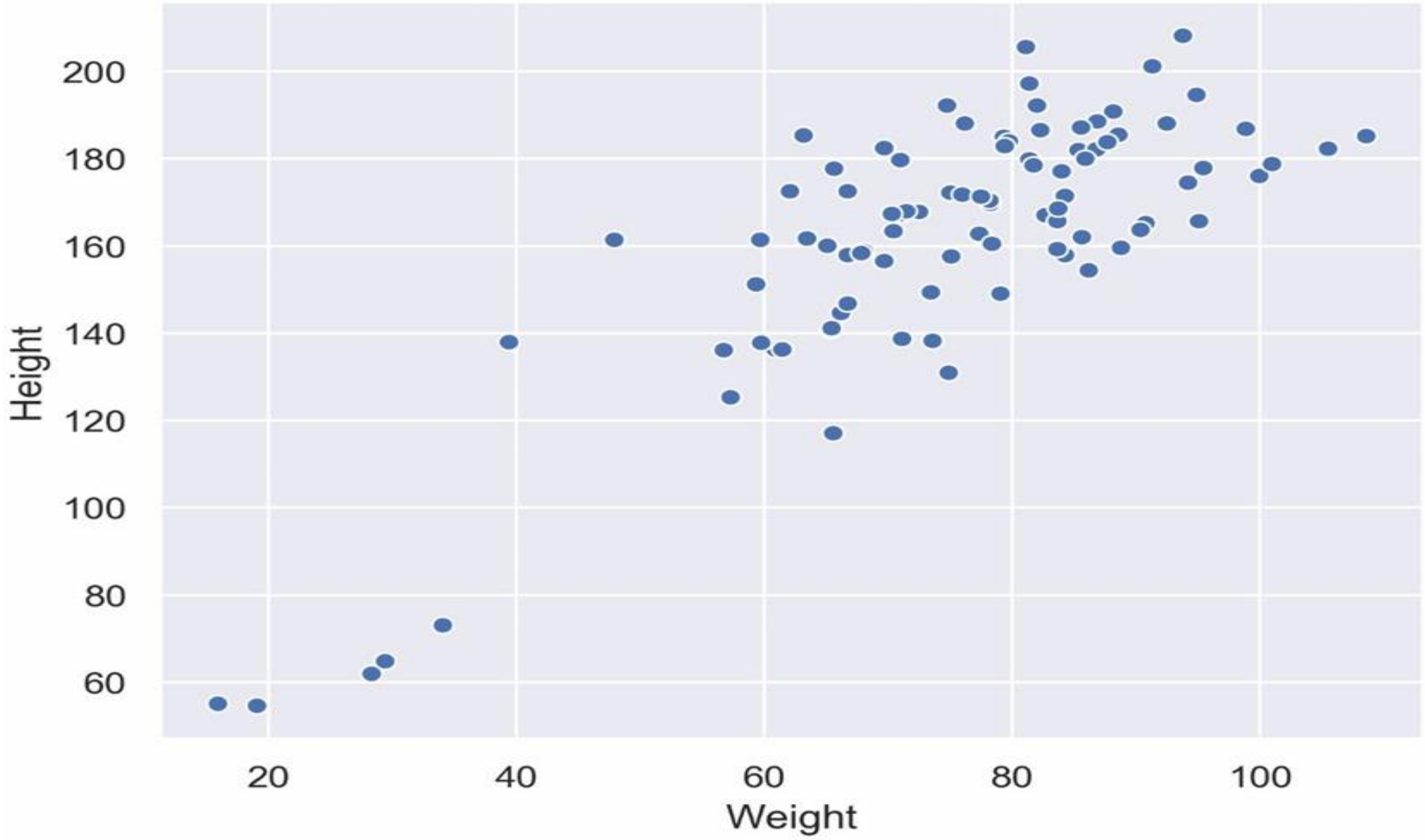


Figure 2.11: Scatter plot with a single group

Example2: 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:

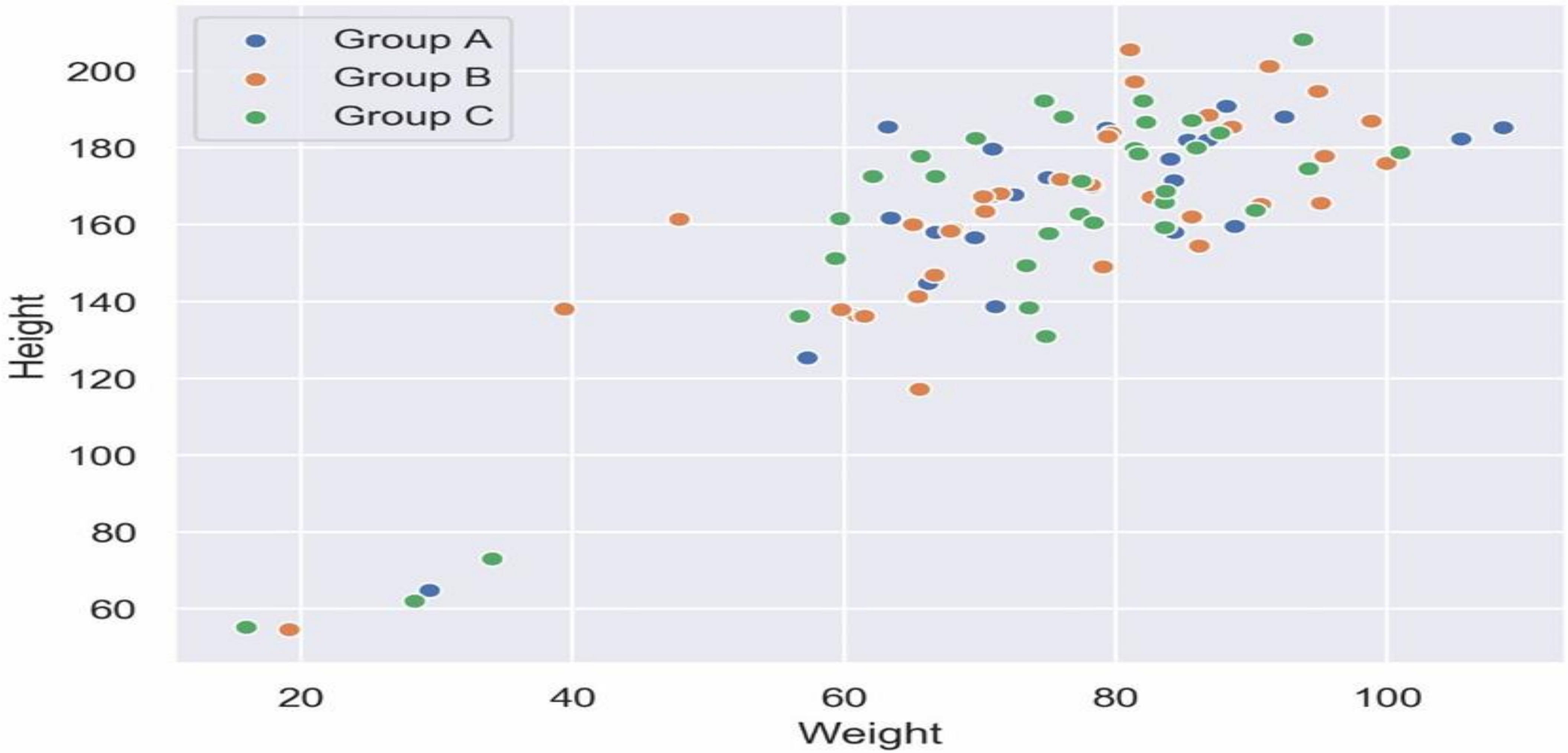


Figure 2.12: Scatter plot with multiple groups

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*:

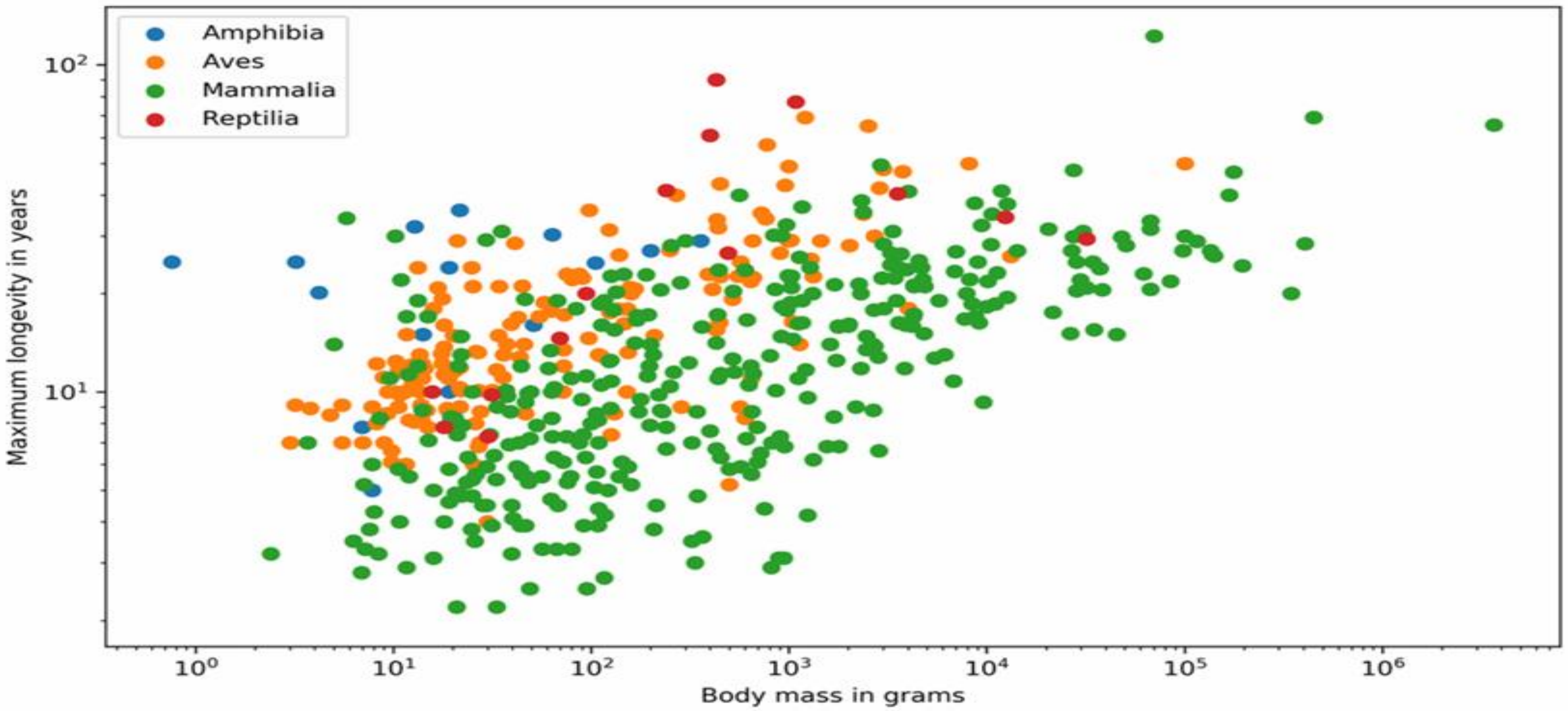


Figure 2.13: Correlation between body mass and maximum longevity for animals

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

- One can plot the marginal distribution for each variable in the form of histograms to **give better insight into how each variable is distributed**.
- **Example:** 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:

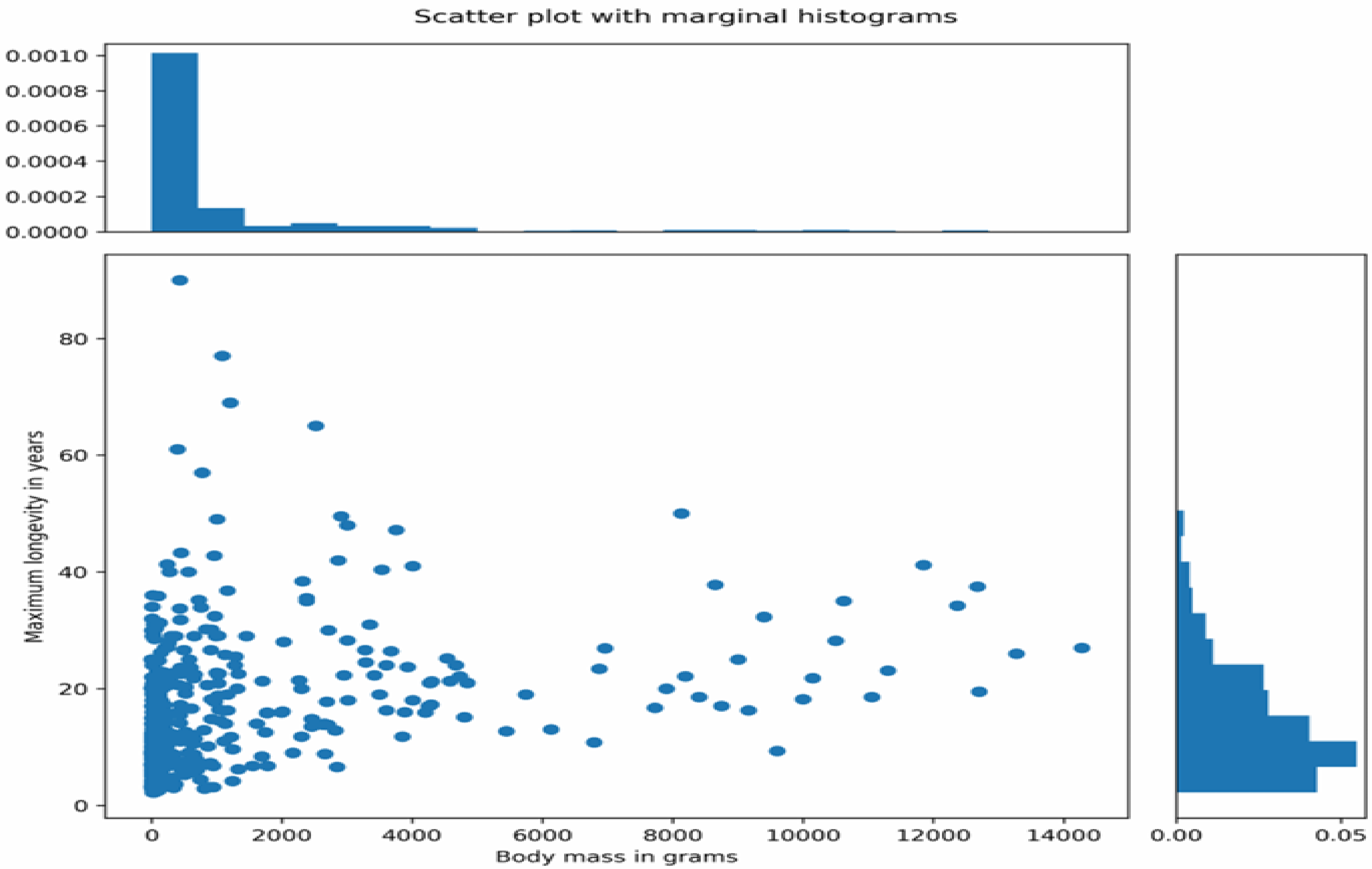
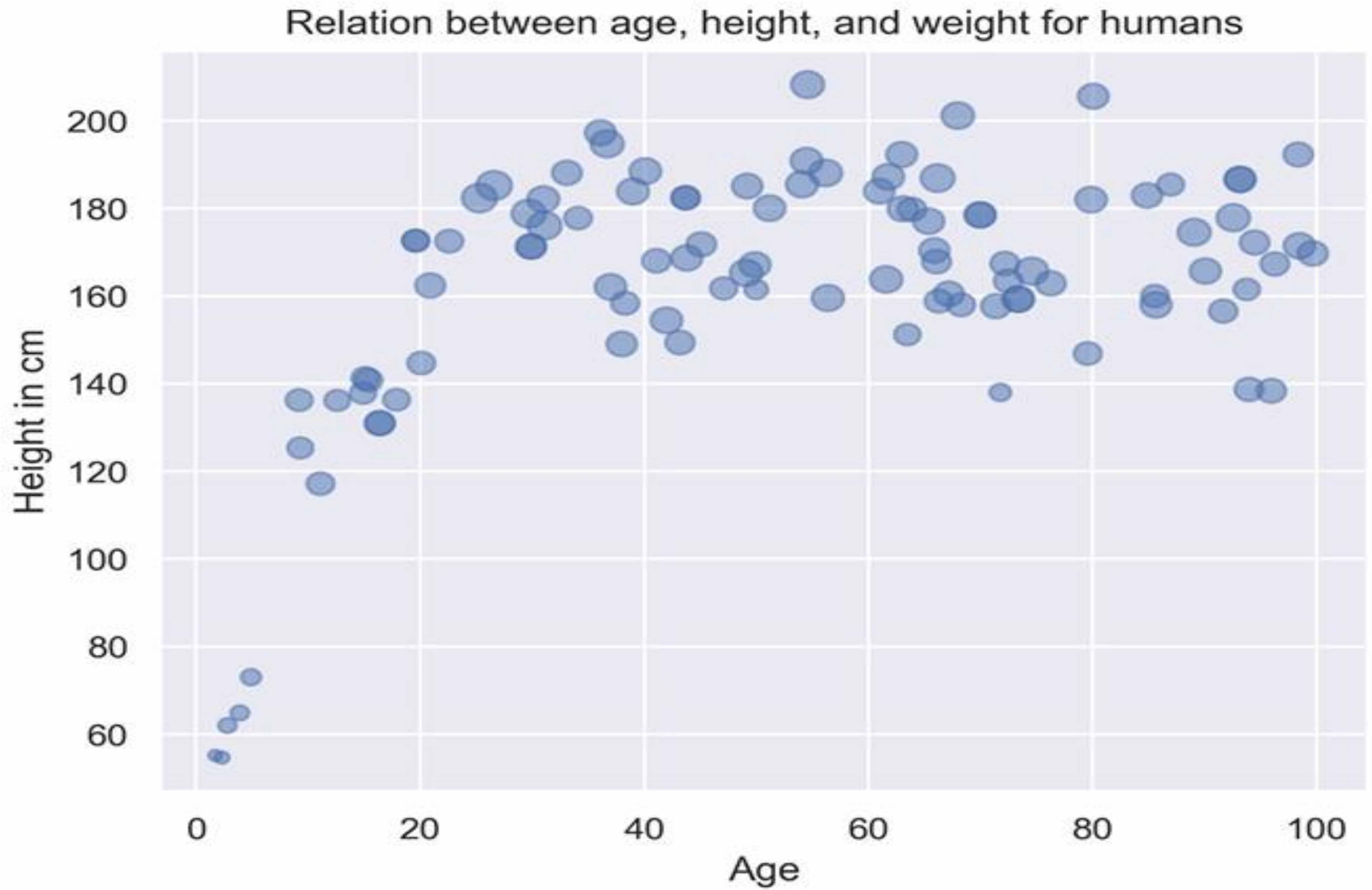
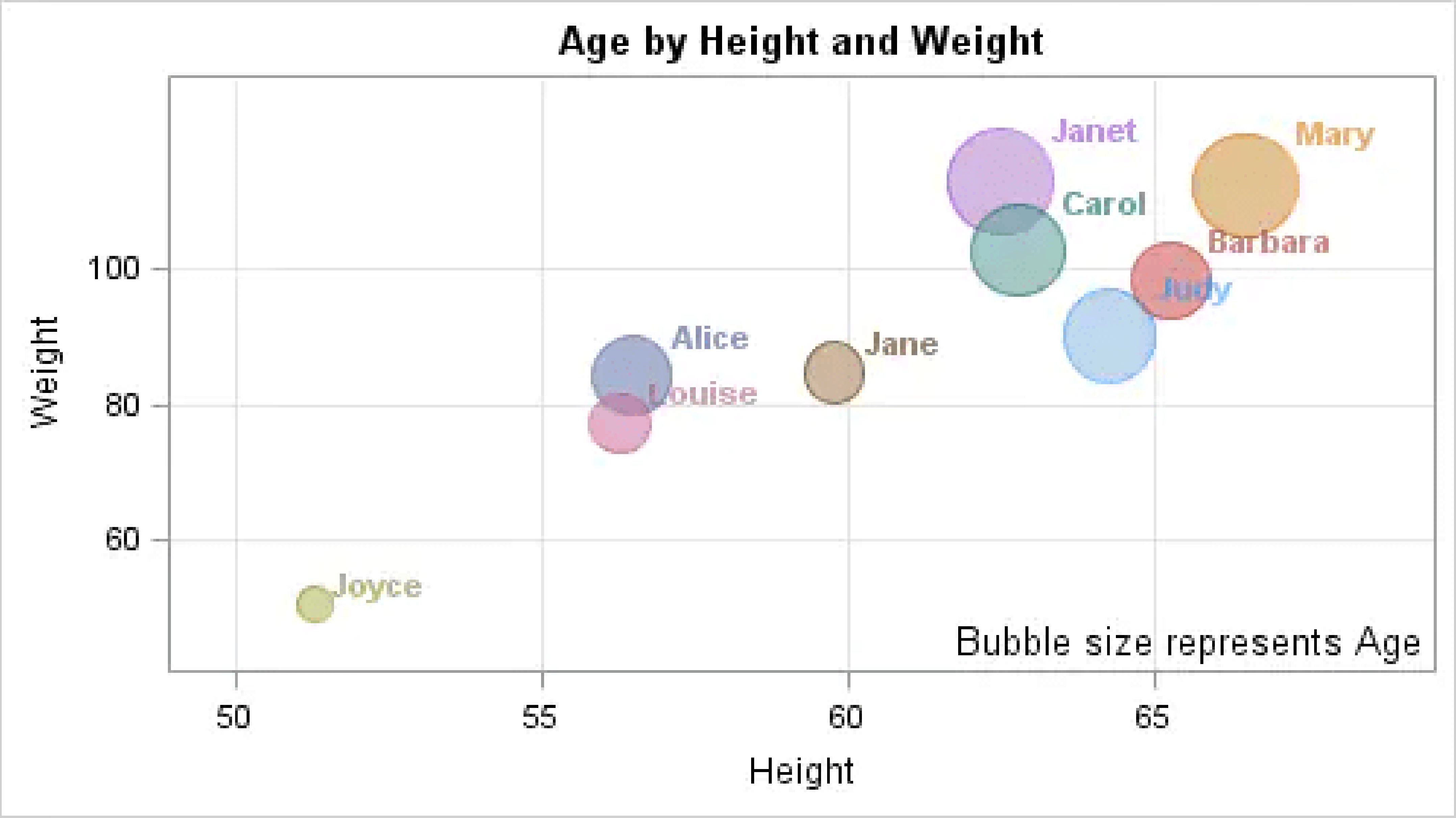


Figure 2.14: Correlation between body mass and maximum longevity of the Aves class with marginal histograms

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


Relation between age, height, and weight for humans
Height in cm
200
180
160
140
120
100
80
60
0
20
40
60
80
100
Age


Age by Height and Weight
Janet
Mary
Carol
Barbara
Judy
Alice
Jane
Louise
Joyce
100
80
60
Weight
Bubble size represents Age
50
55
60
65
Height

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.
- 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/linegraph.
- One can also plot the relationship between **multiple groups or categories** using different colors.

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.

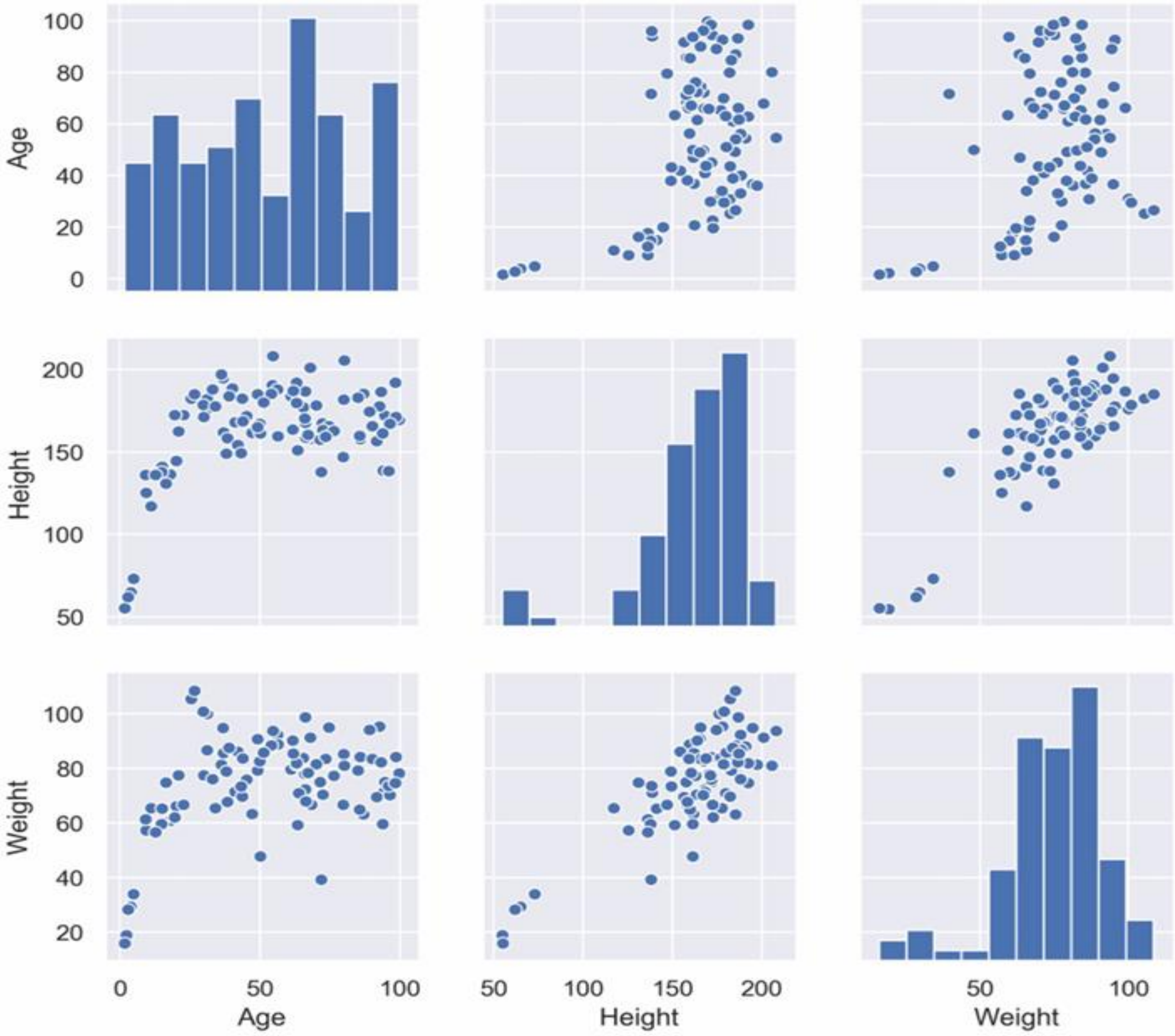


Figure 2.16: Correlogram with a single category

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

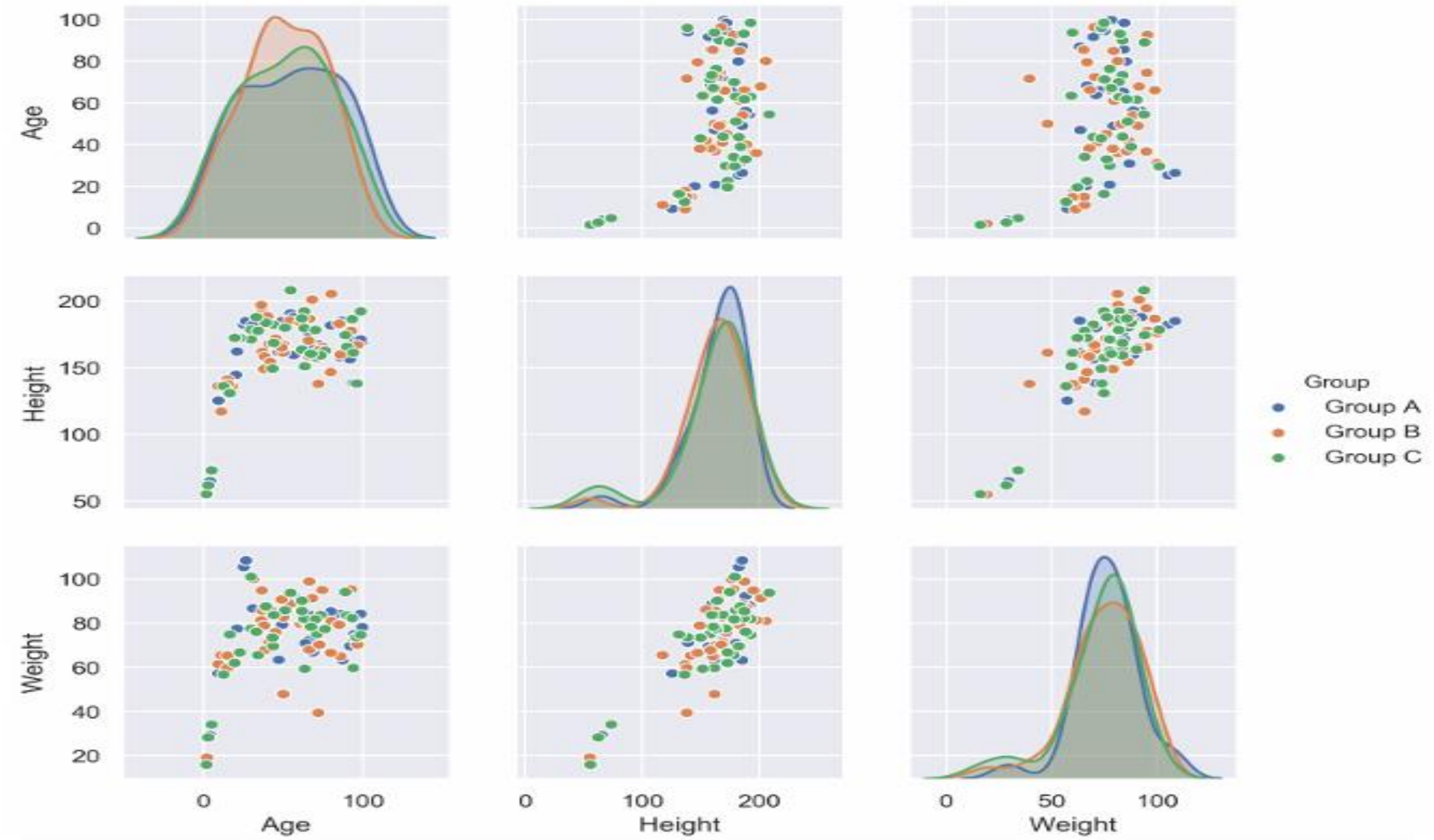


Figure 2.17: Correlogram with multiple categories

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.

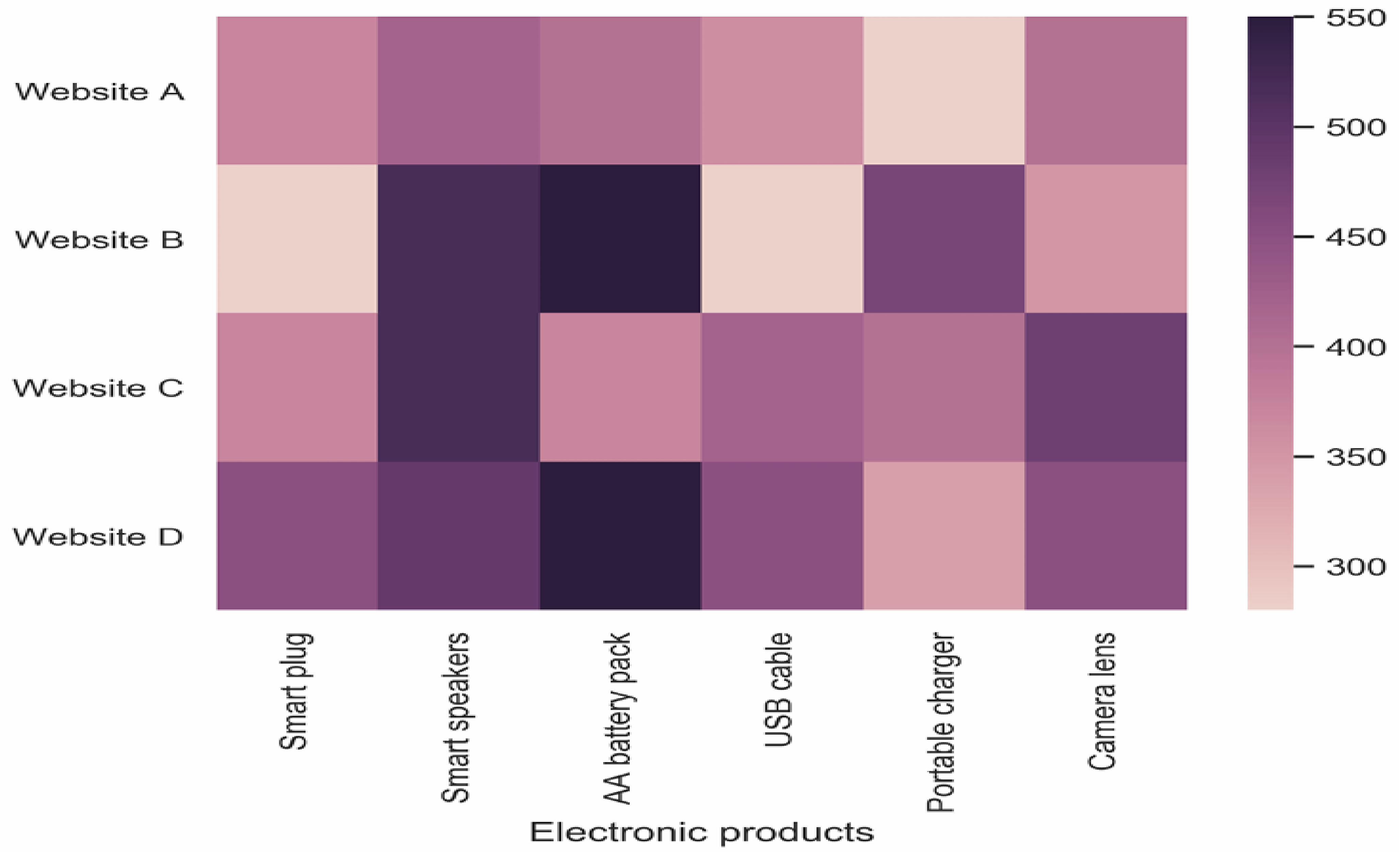
Heatmap

- 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:

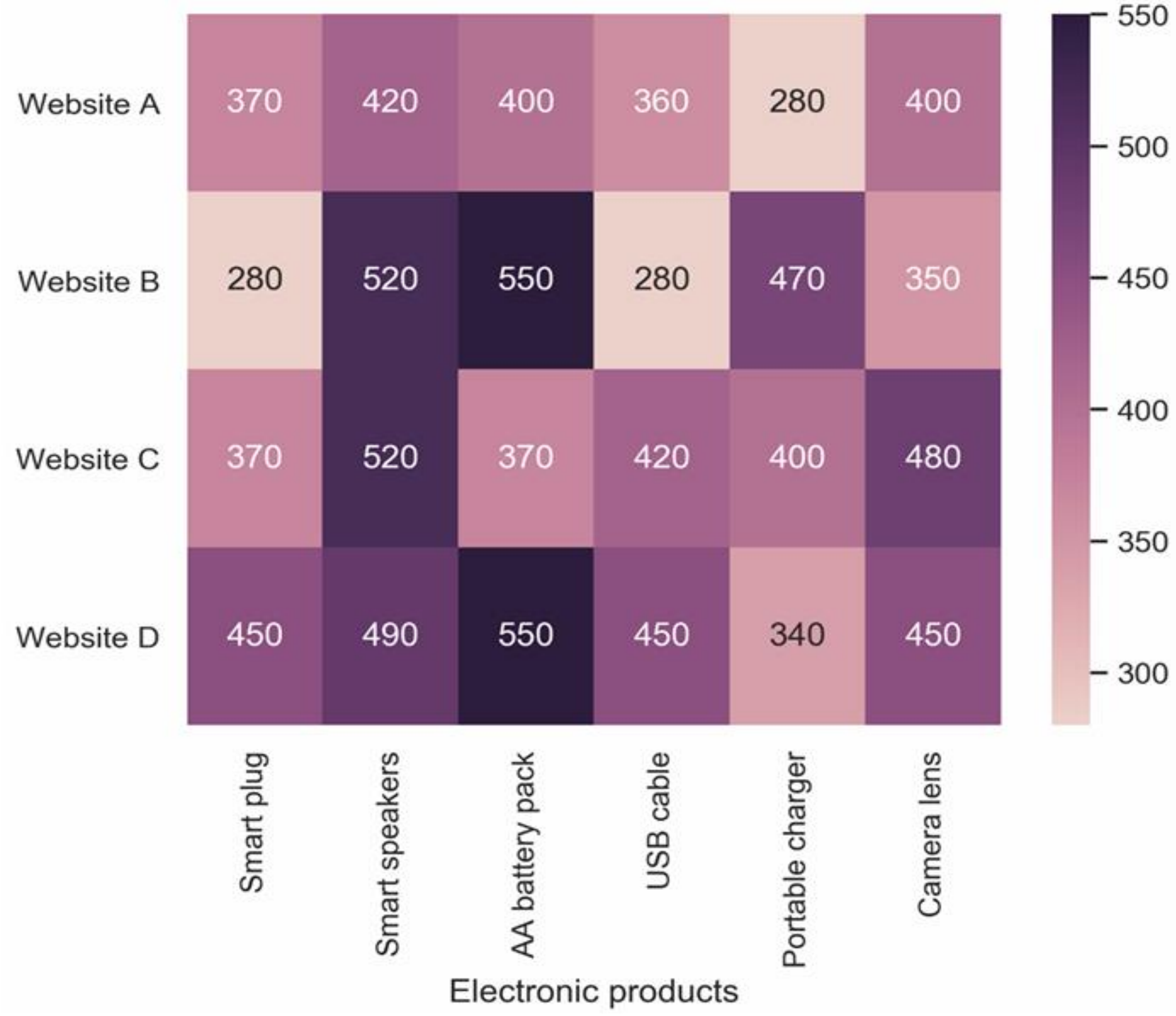
Figure : Heatmap for popular products in the electronics category



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 : 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.

Activity 2: 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:

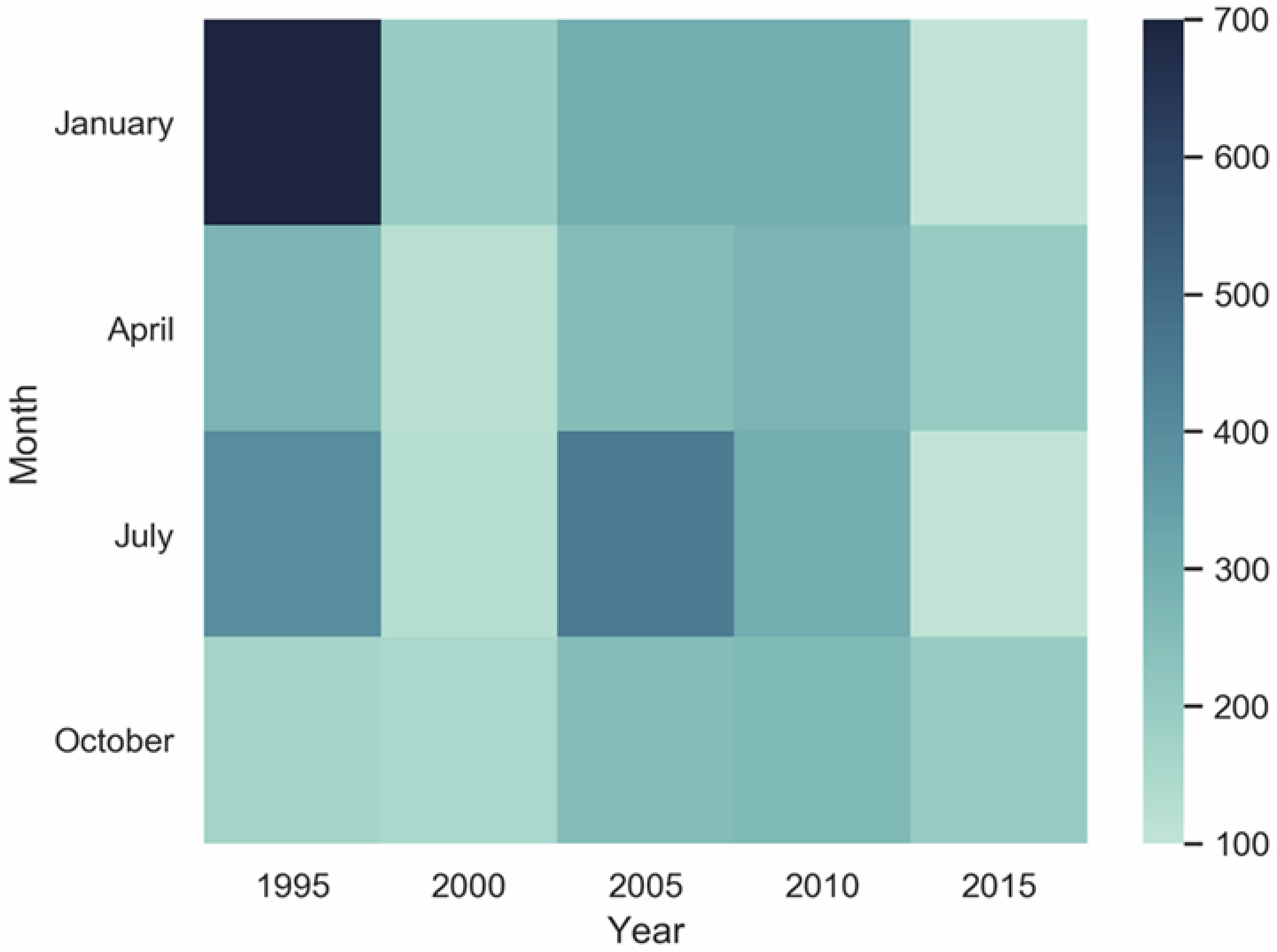


Figure 2.20: Total accidents over 20 years

Solution:

1. **Suggested response**: If we look at Figure, we can see that the years 2000 and 2015 have the lightest colored squares overall. These are the two years that have the lowest accident rates.
2. **Suggested response**: If we look at the trend for each month, that is, January, April, July, and October for the past two decades, we can see a decreasing trend in the number of accidents taking place in **January.**

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.
    - **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%.
- Use : To compare items that are part of a whole.

Example: The following diagram shows household water usage around the world

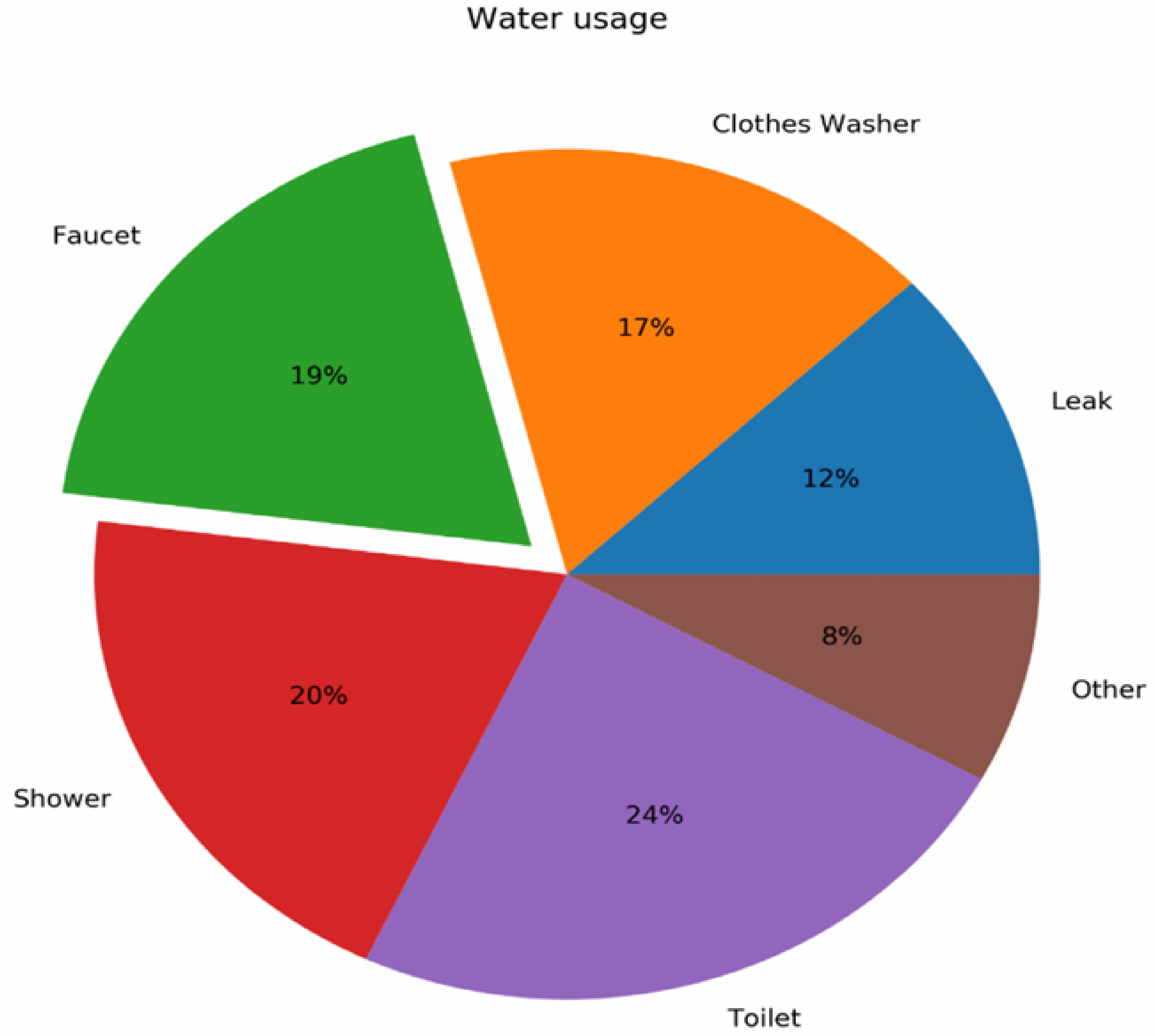


Figure 2.21: Pie chart for global household water usage

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**:

Example: The following diagram shows a basic donut chart:

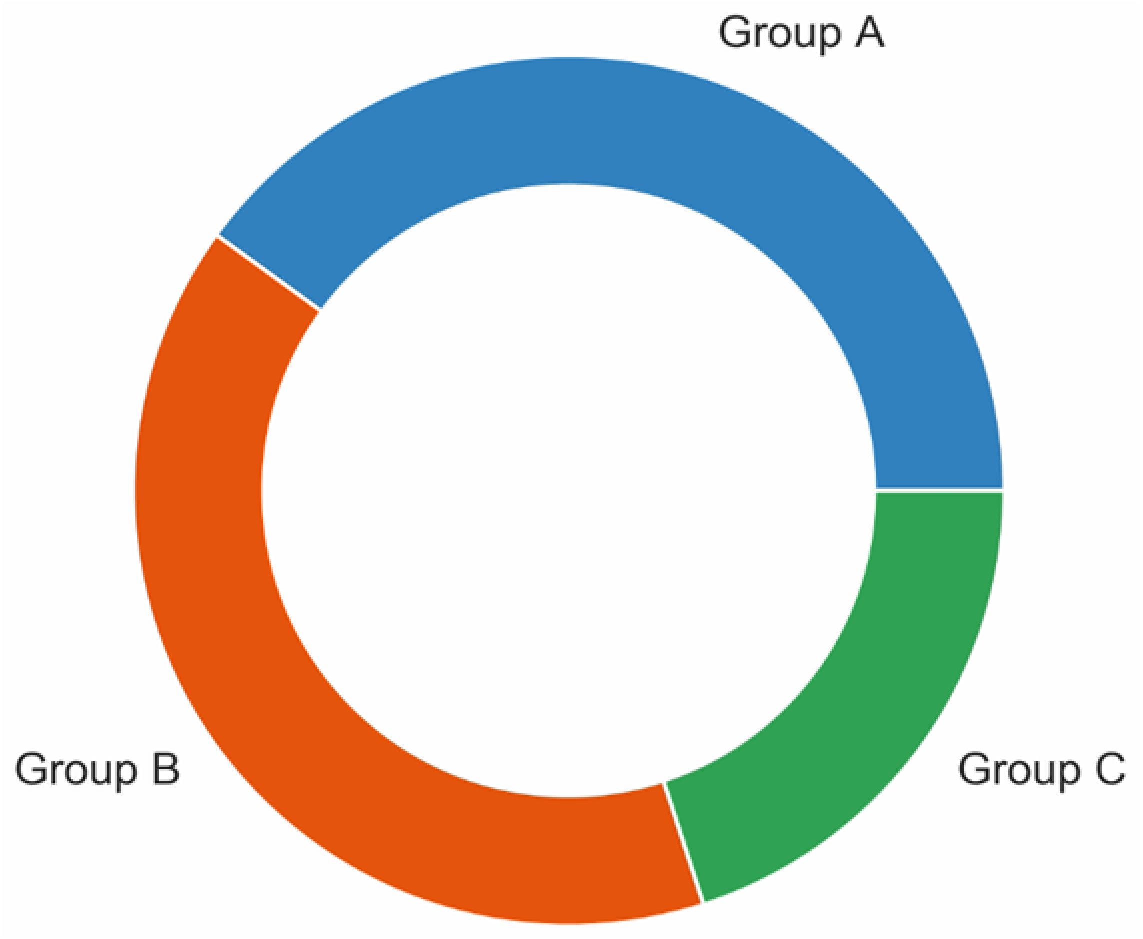
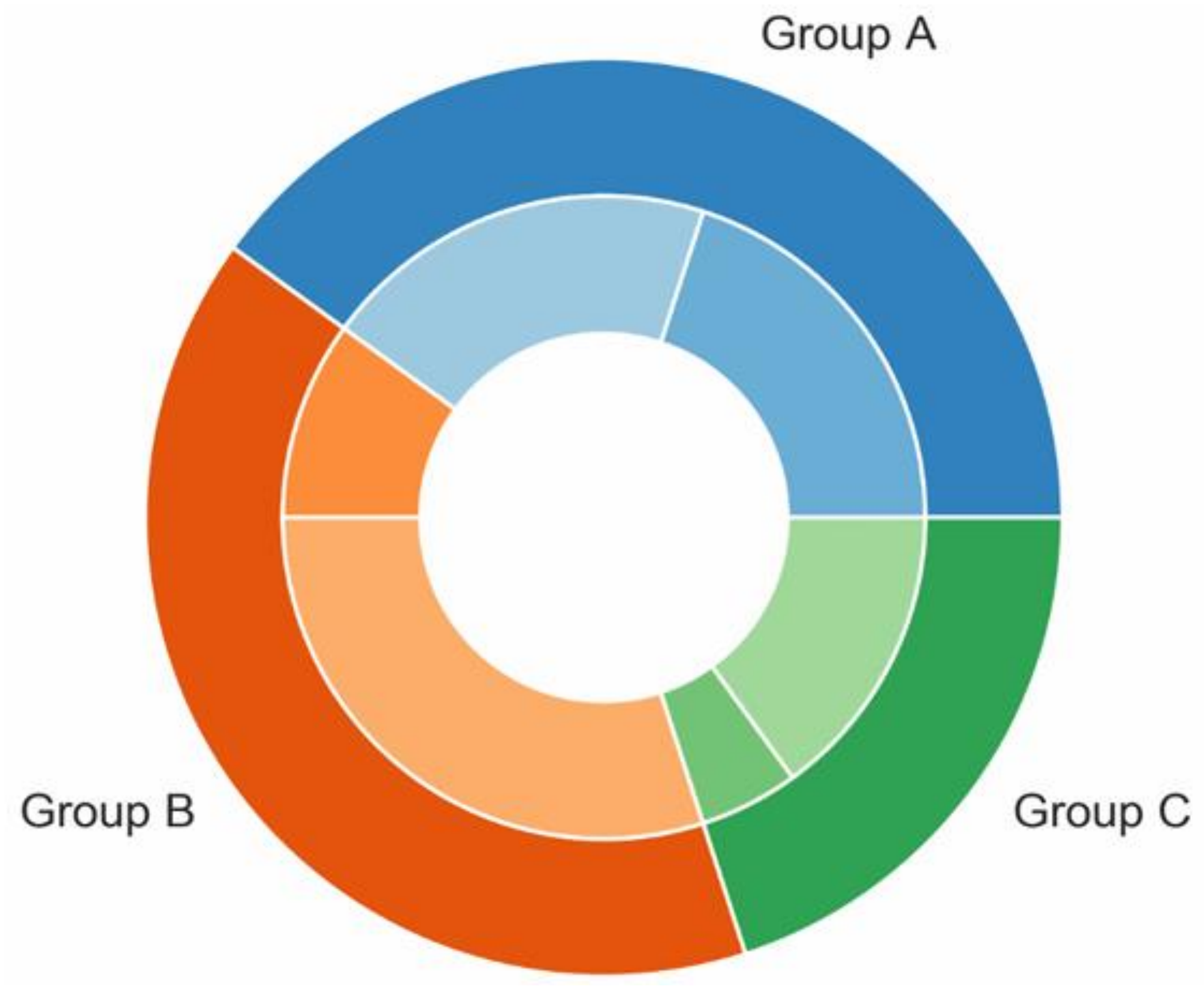


Figure 2.22: Donut chart

Example: **The following diagram shows a 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

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

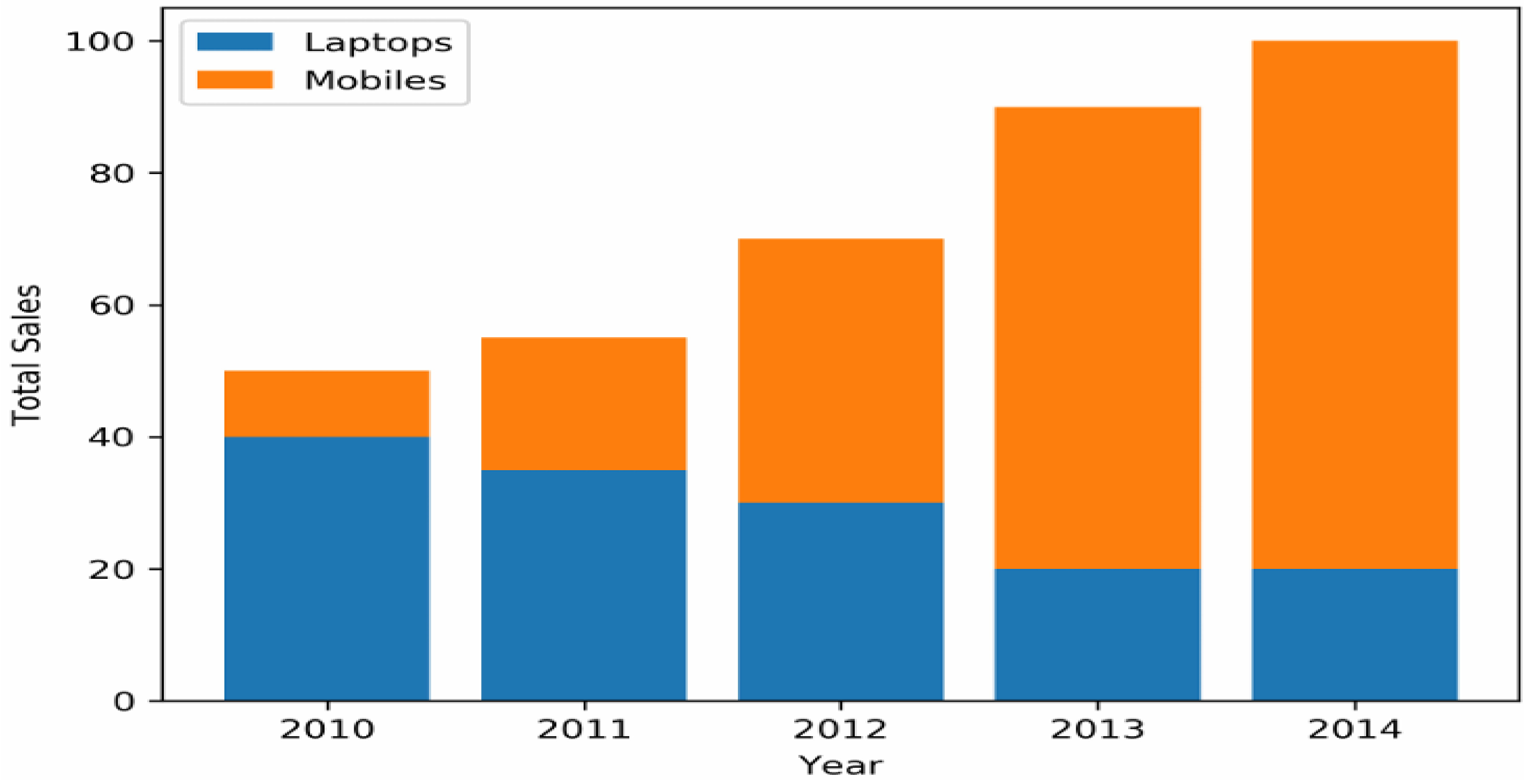


Figure 2.24: Stacked bar chart to show sales of laptops and mobiles

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

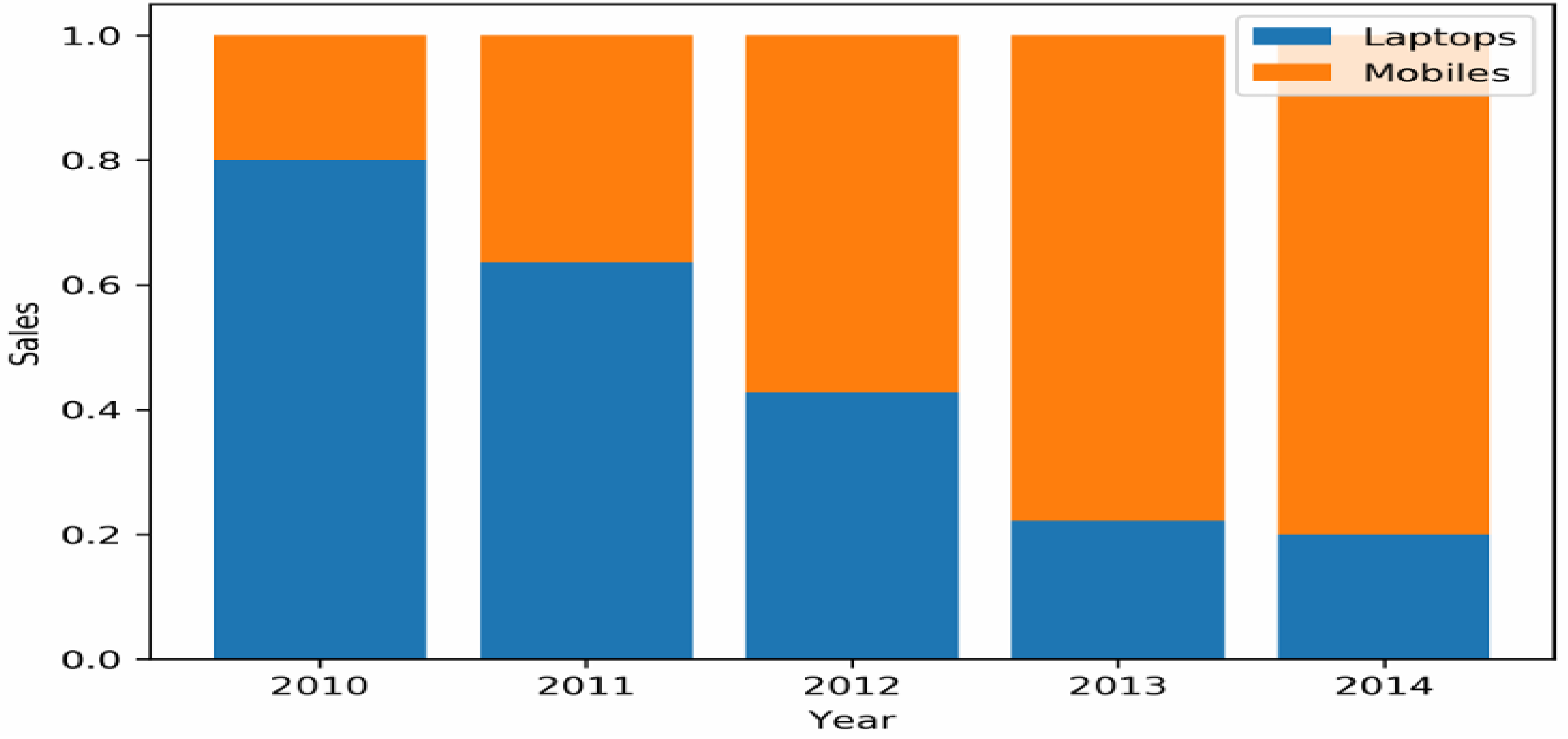


Figure 2.25: 100% stacked bar chart to show sales of laptops, PCs, and mobiles

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

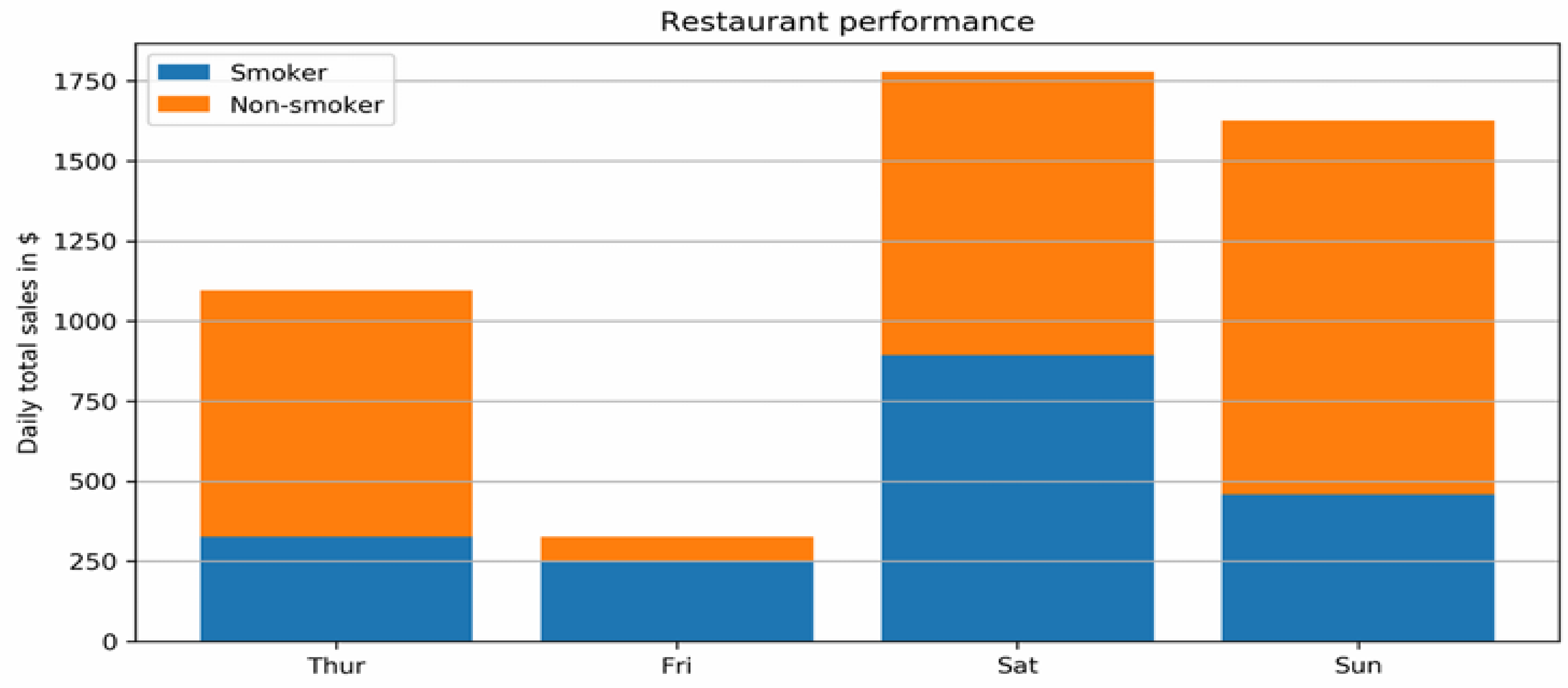


Figure 2.26: Daily total restaurant sales categorized by smokers and non-smokers

Design Practices

1. **Use contrasting colors for stacked bars.**
2. **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.**
3. **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**

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

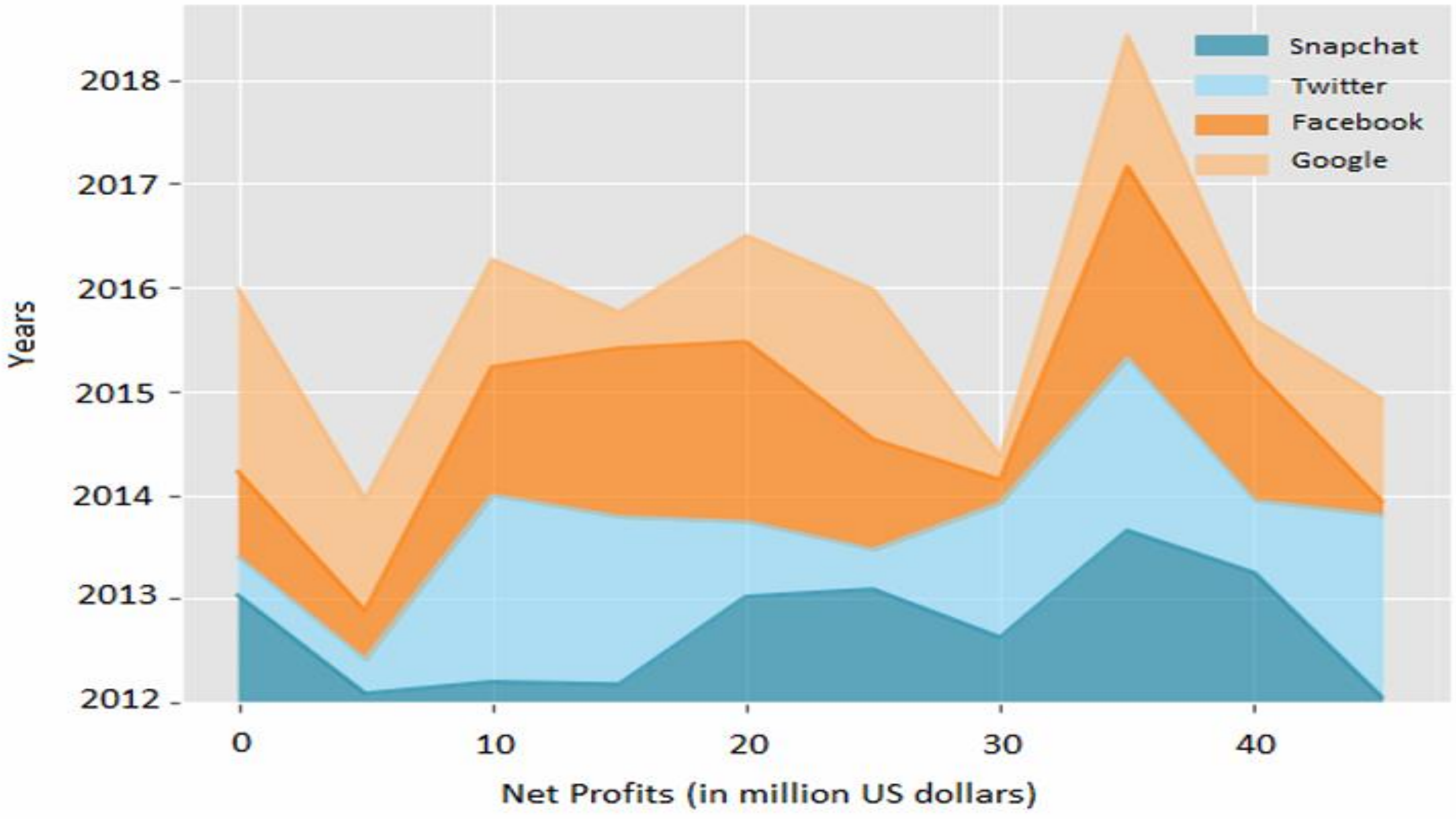


Figure 2.27: Stacked area chart to show net profits of four companies

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.

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

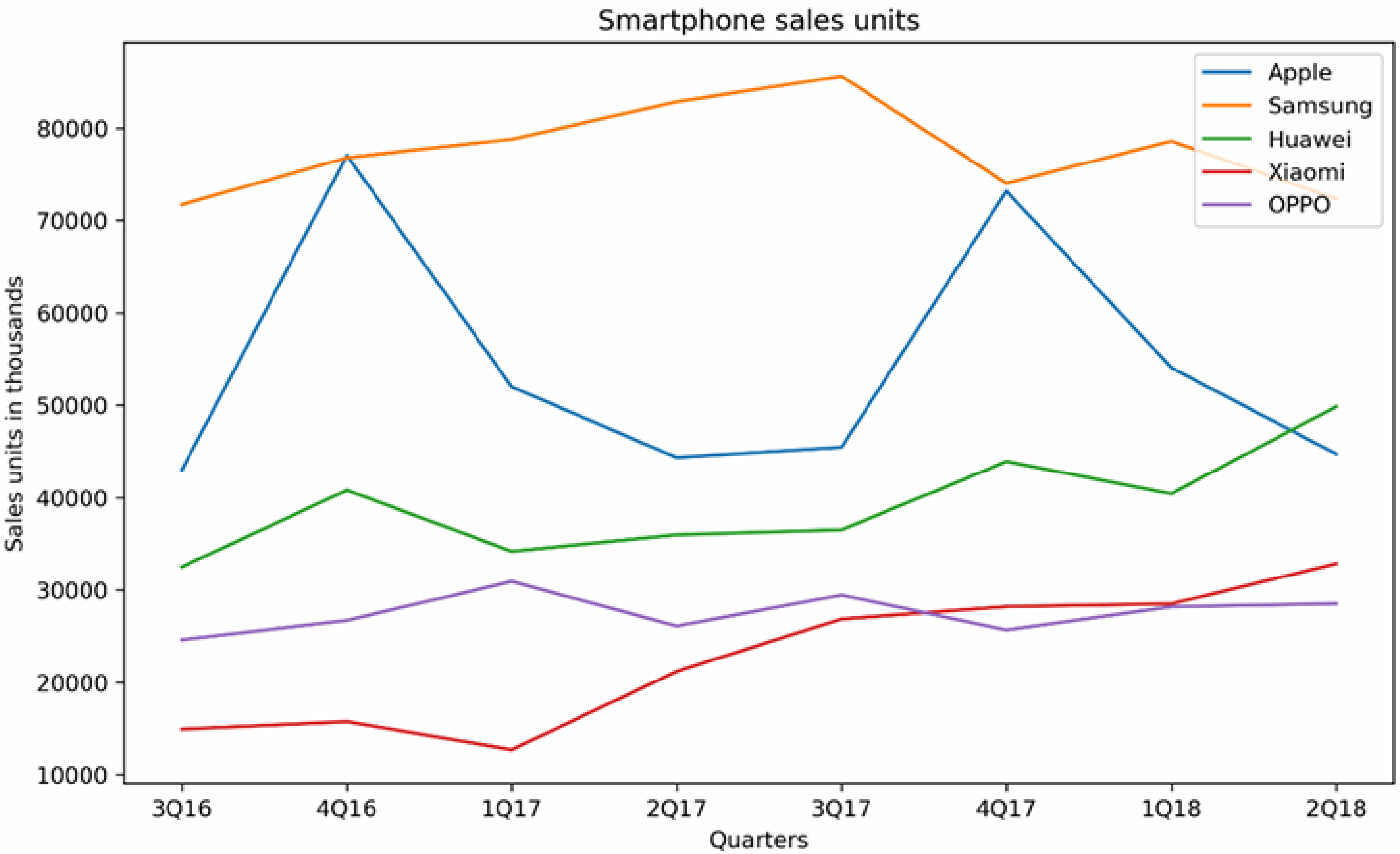


Figure 2.28: Line chart of smartphone sales units

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:

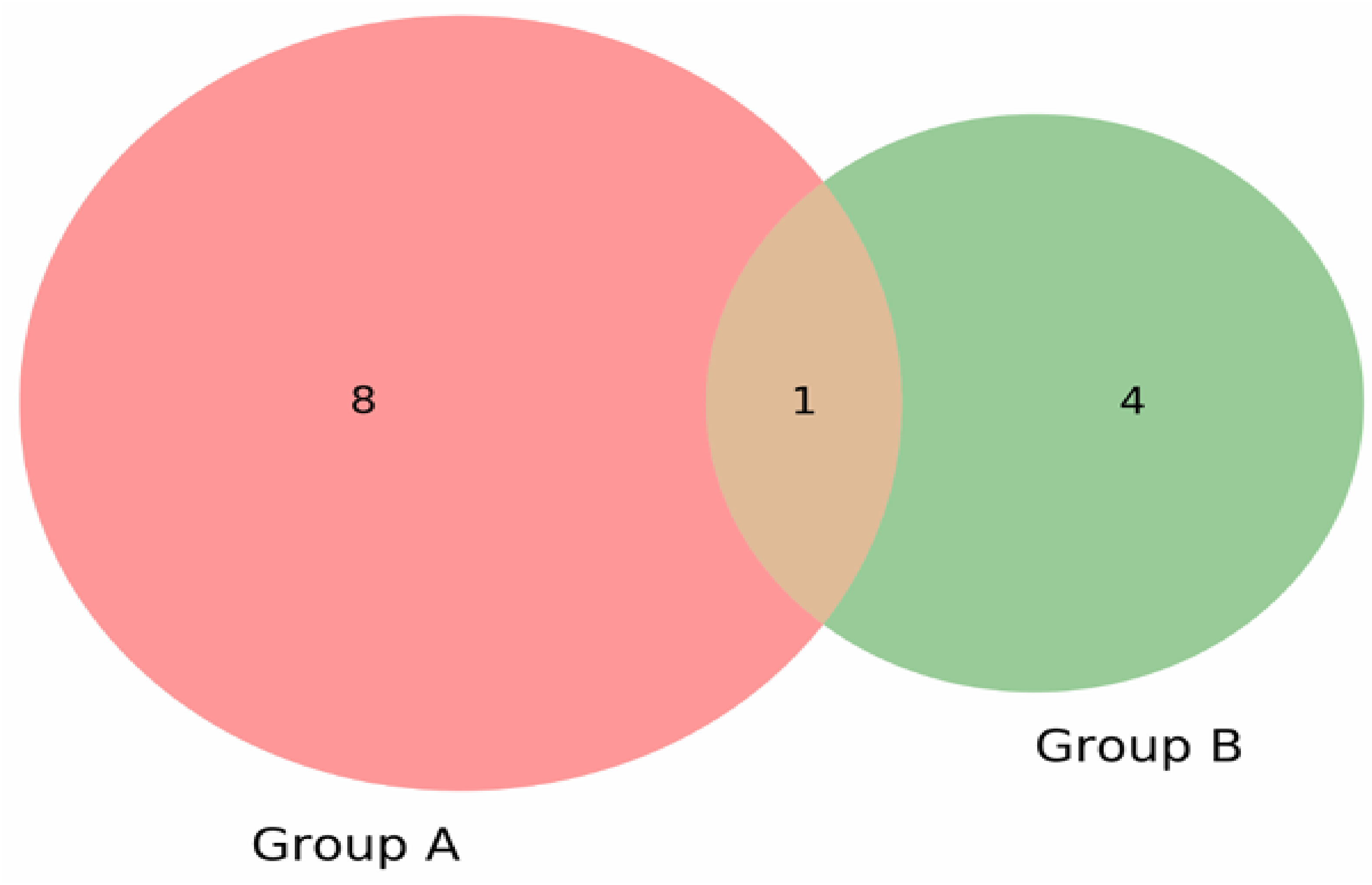


Figure 2.29: Venn diagram showing students taking the same class

Design Practice

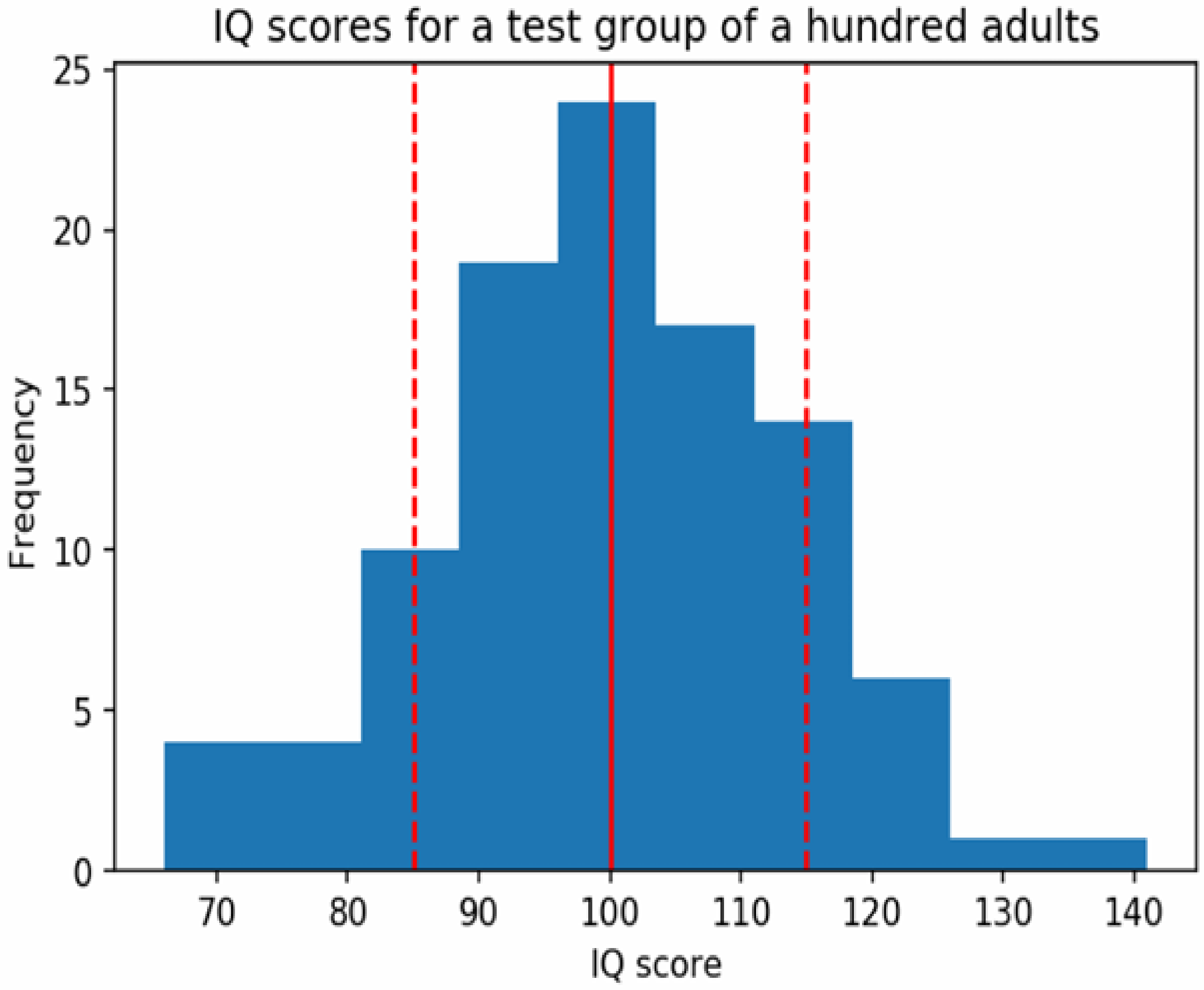
- It is not recommended to use Venn diagrams if you have more than three groups. It would become difficult to understand.

Distribution Plots

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

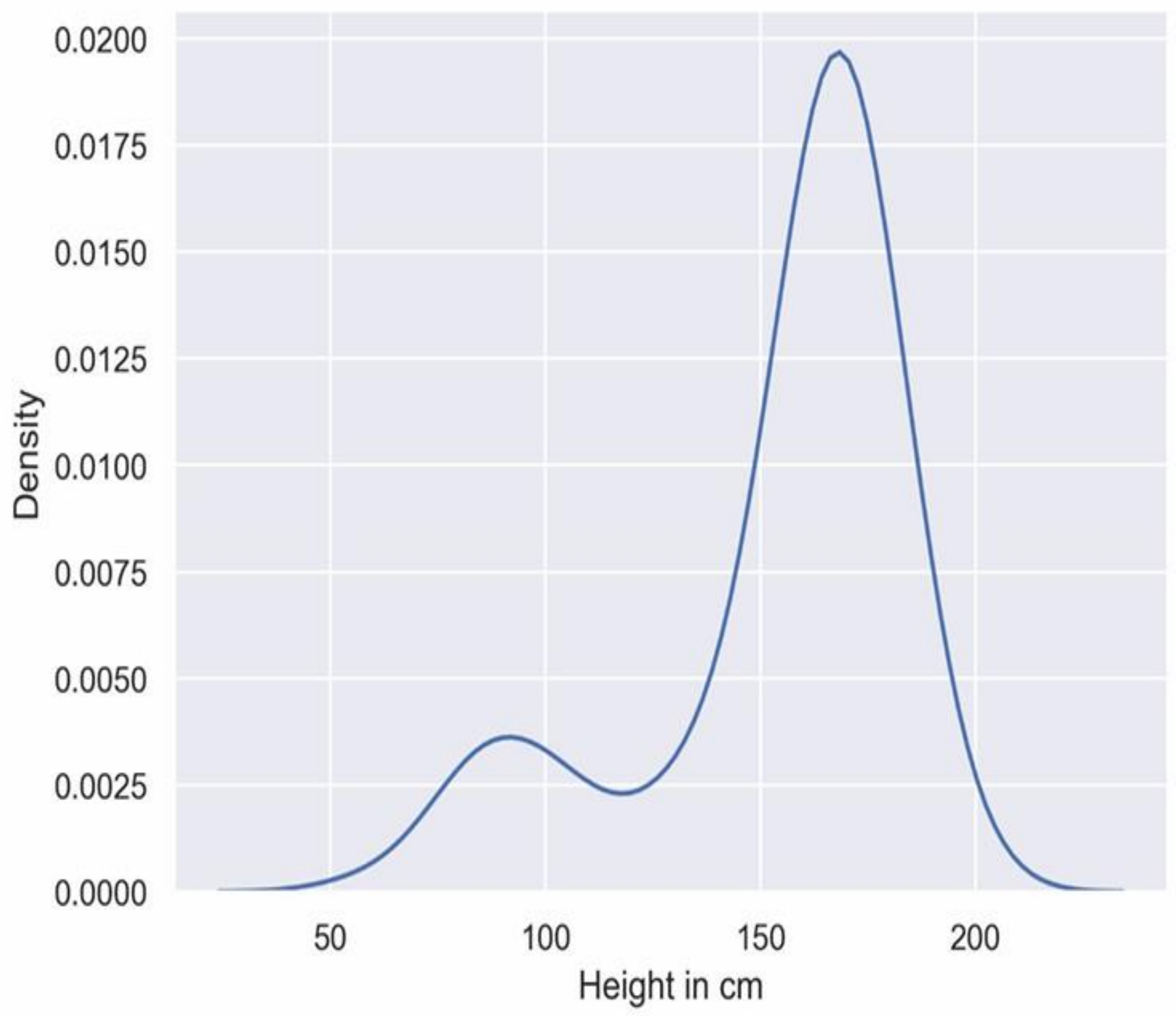
Histogram

- A histogram visualizes the distribution of a single numerical variable. **Each bar** represents the frequency for a certain interval.
- **Use:** Get insights into the underlying distribution for a dataset.
- **Design Practice** : Try **different numbers** of bins (data intervals), since the shape of the histogram can vary significantly



Density Pot

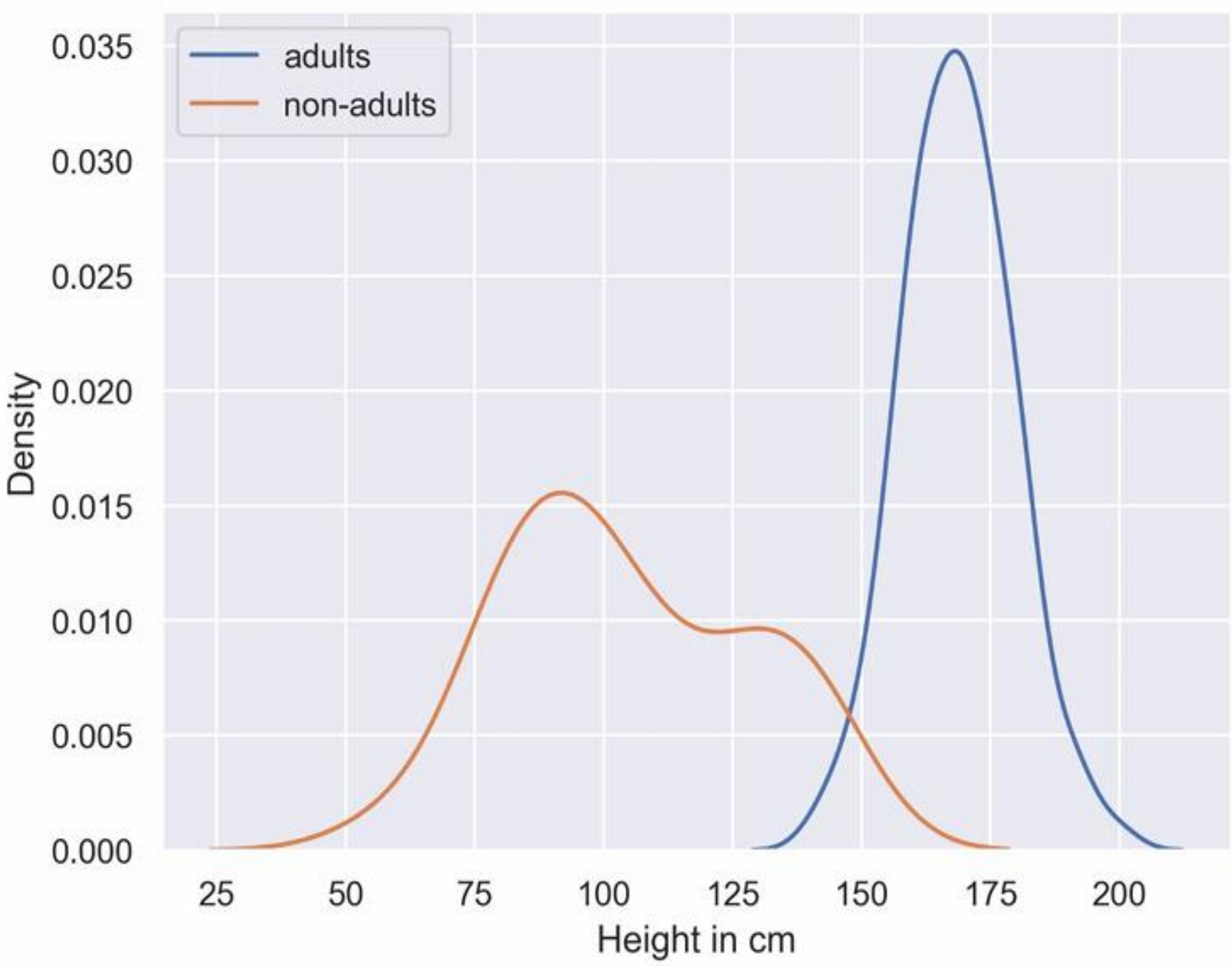
- A density plot shows the distribution of a numerical variable. It is a variation of a histogram that allowing for smoother distributions
- Use :To compare the distribution of several variables .



Density Pot

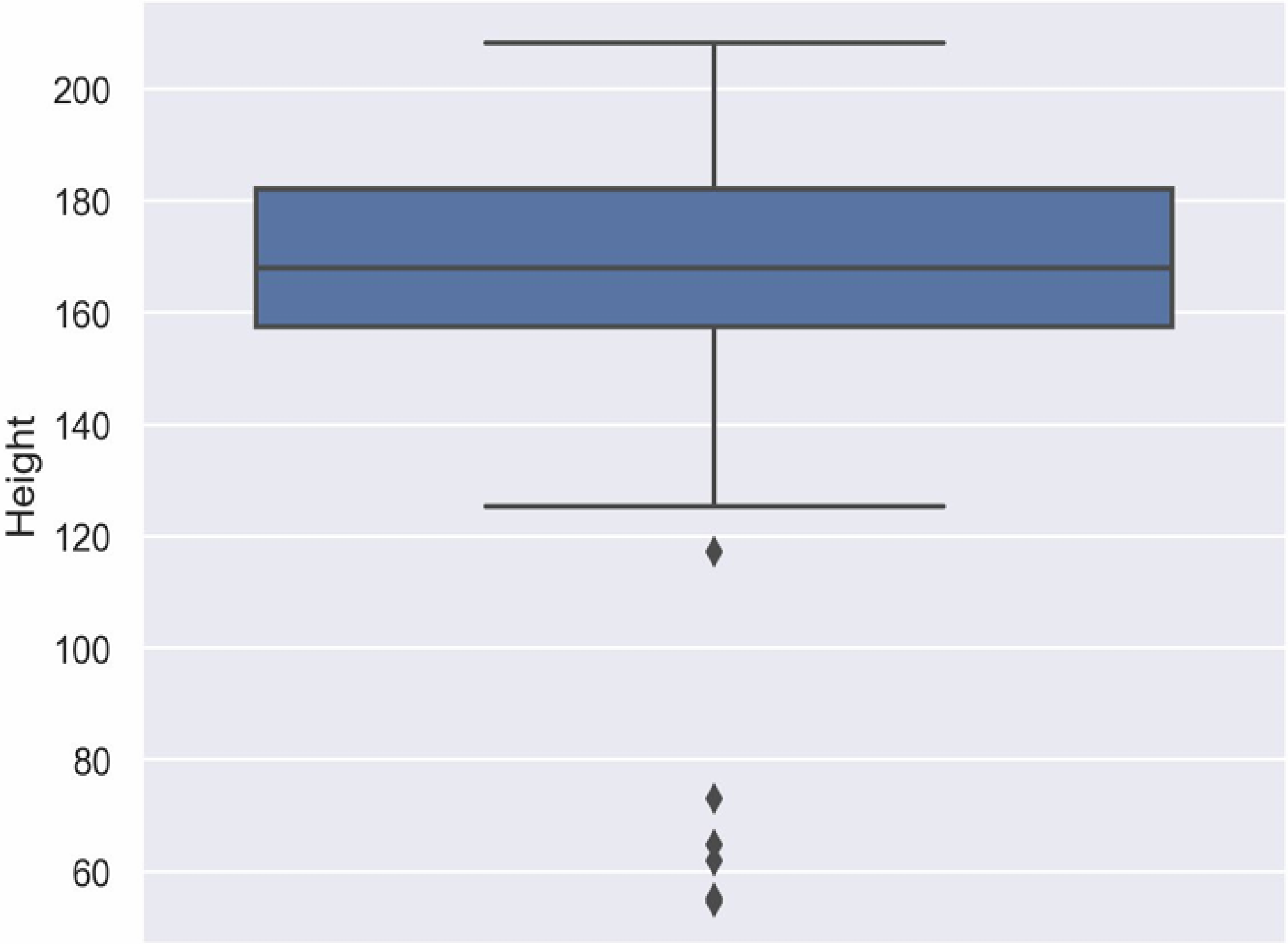
- **Design Practice** : Use **contrasting colors** to plot the density of multiple variables.

The following diagram shows a basic multi-density plot:

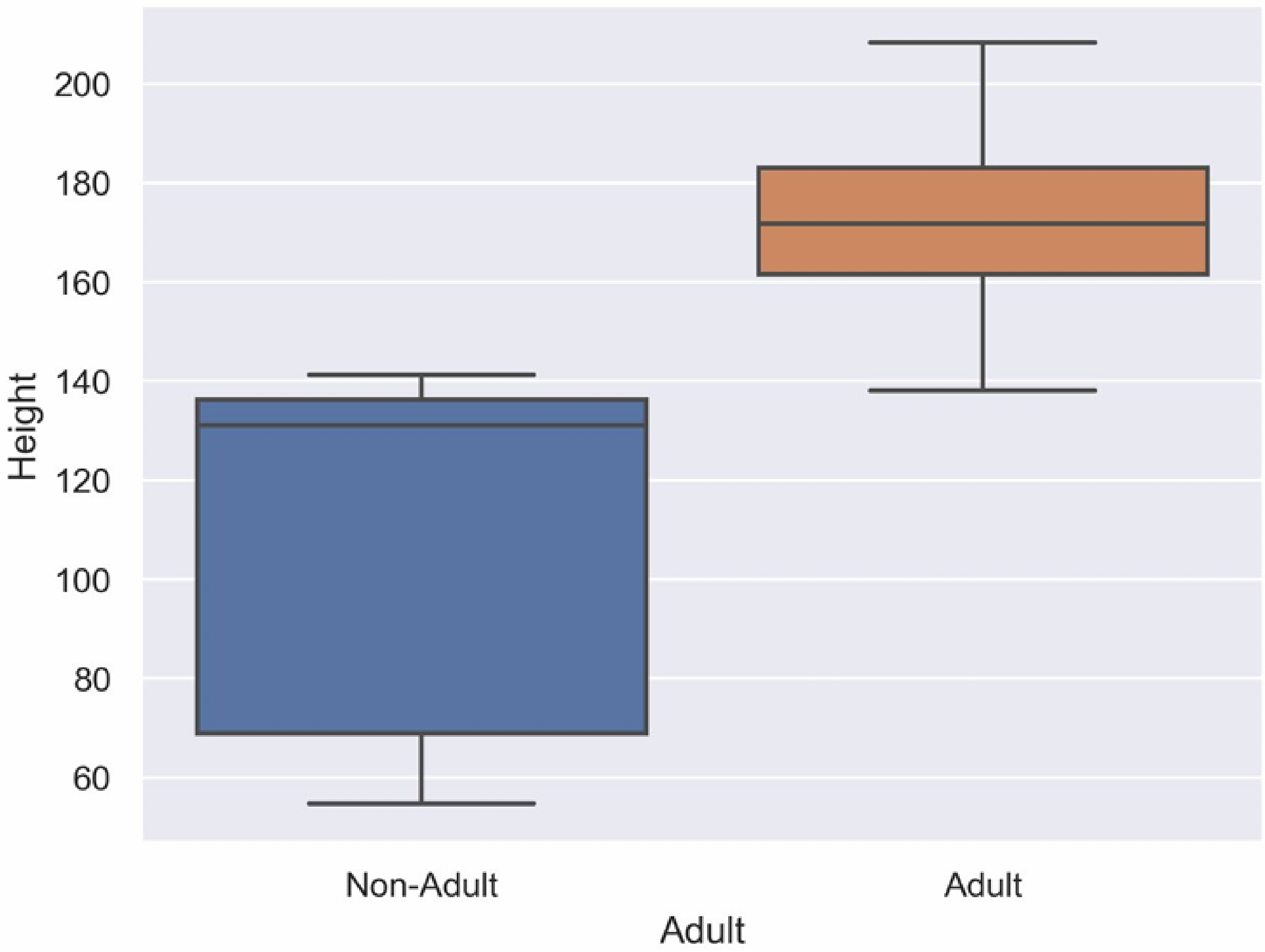


Box Plot

- 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 parallel extending lines from the boxes are called **whiskers**; they indicate the variability outside the lower and upper quartiles.
- **Outliers** are represented as circles or diamonds

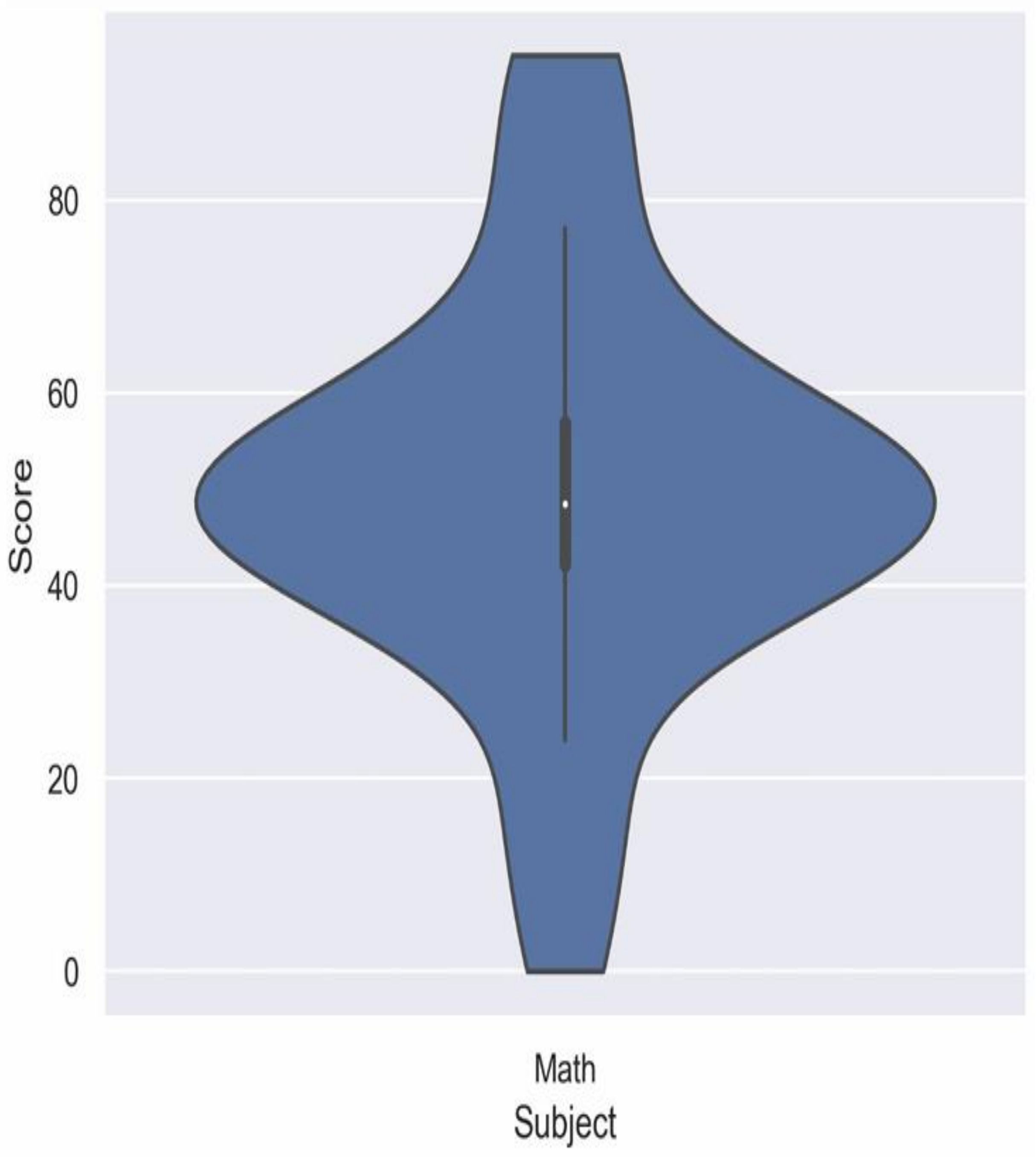


Box plot for multiple variables

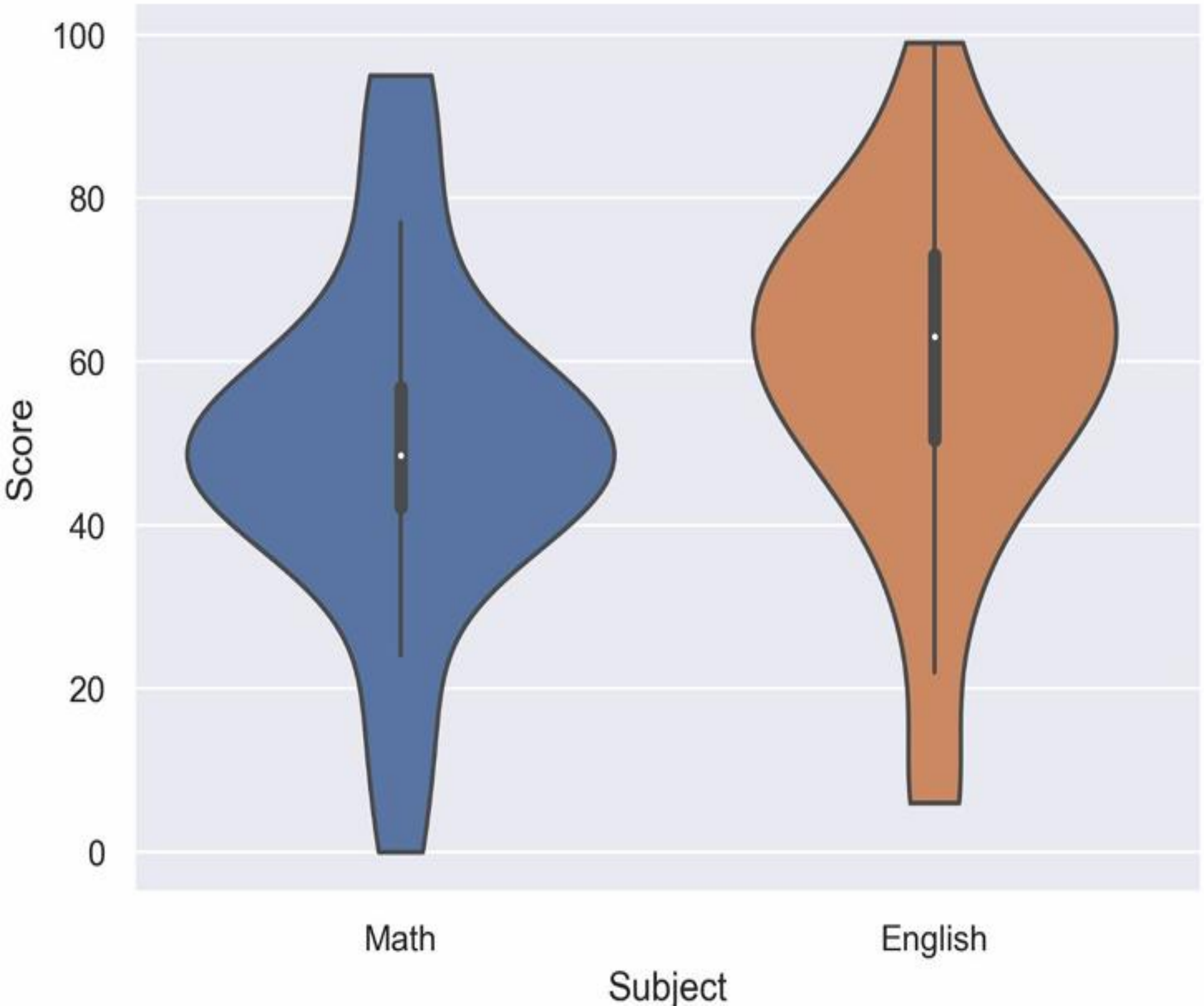


Violin Plot

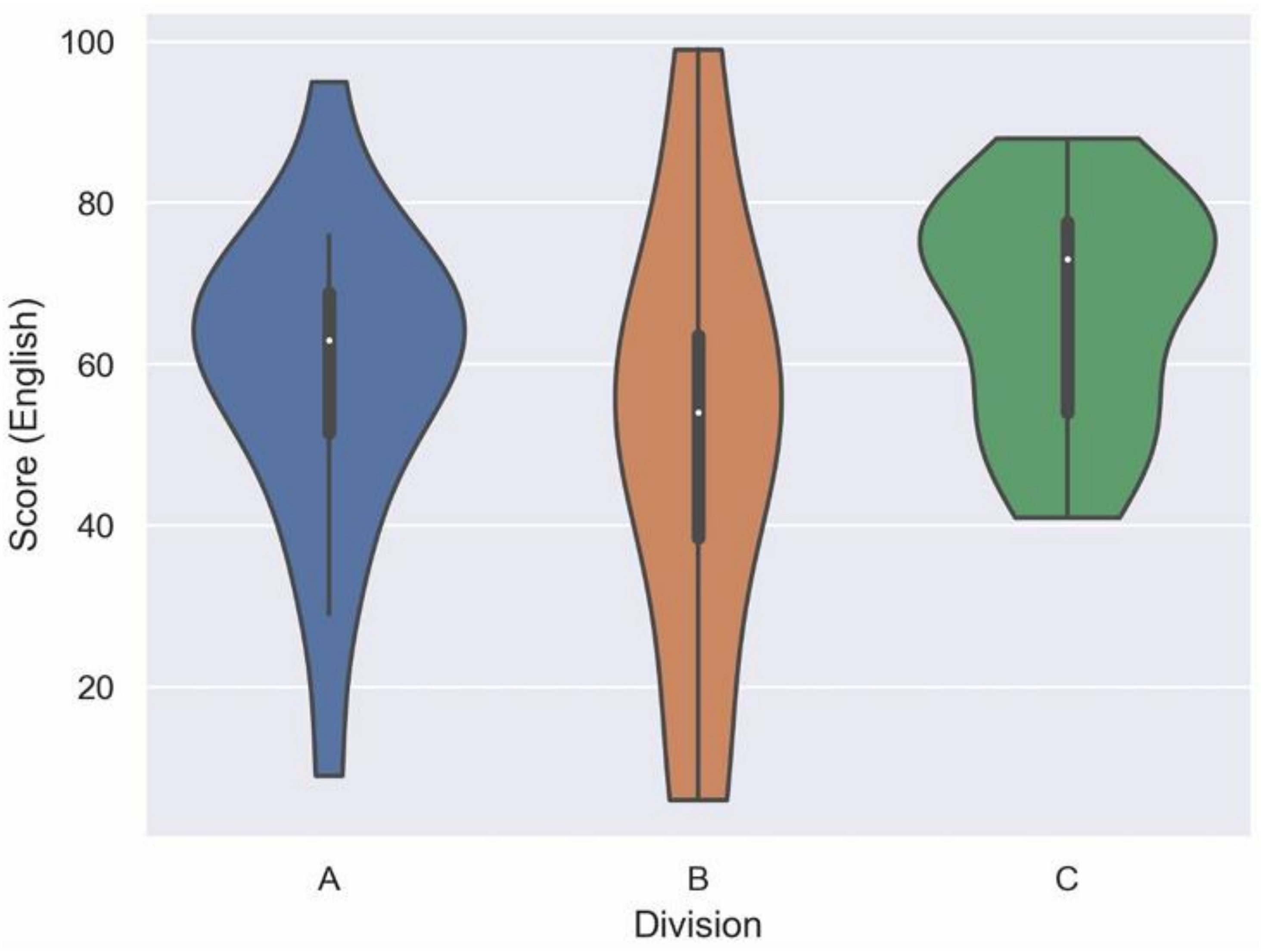
- Violin plots are a combination of box plots and density plots.
- Both the statistical measures and the distribution are visualized.
- The **thick black bar** in the center represents the **interquartile range**,
- while **the thin black line** corresponds to the **whiskers in a box plot**.
- The **white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



Violin plot for multiple variables (English and Math)



Violin plot with multiple categories (three groups of students)

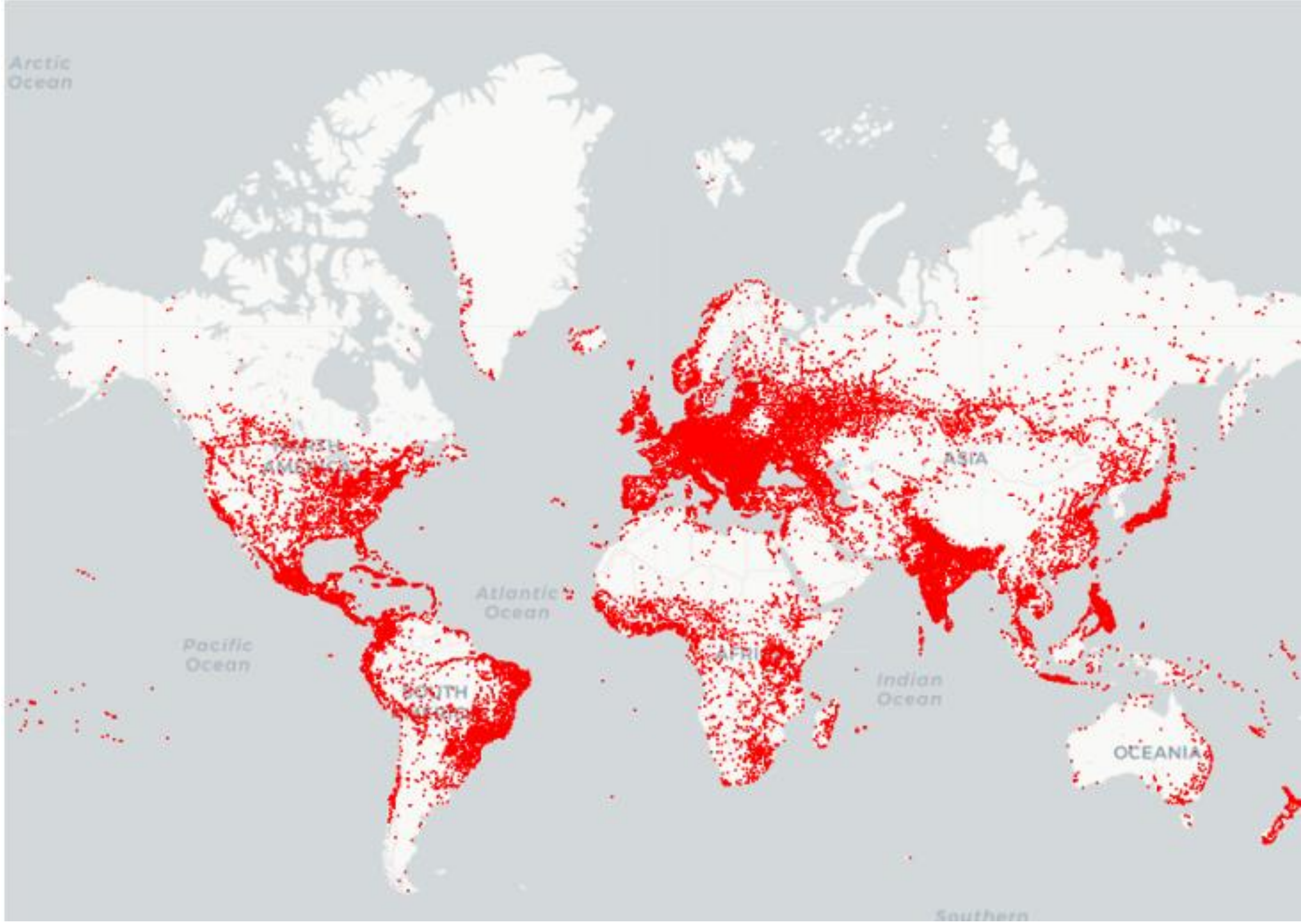


Geoplots

- 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.

Dot Map

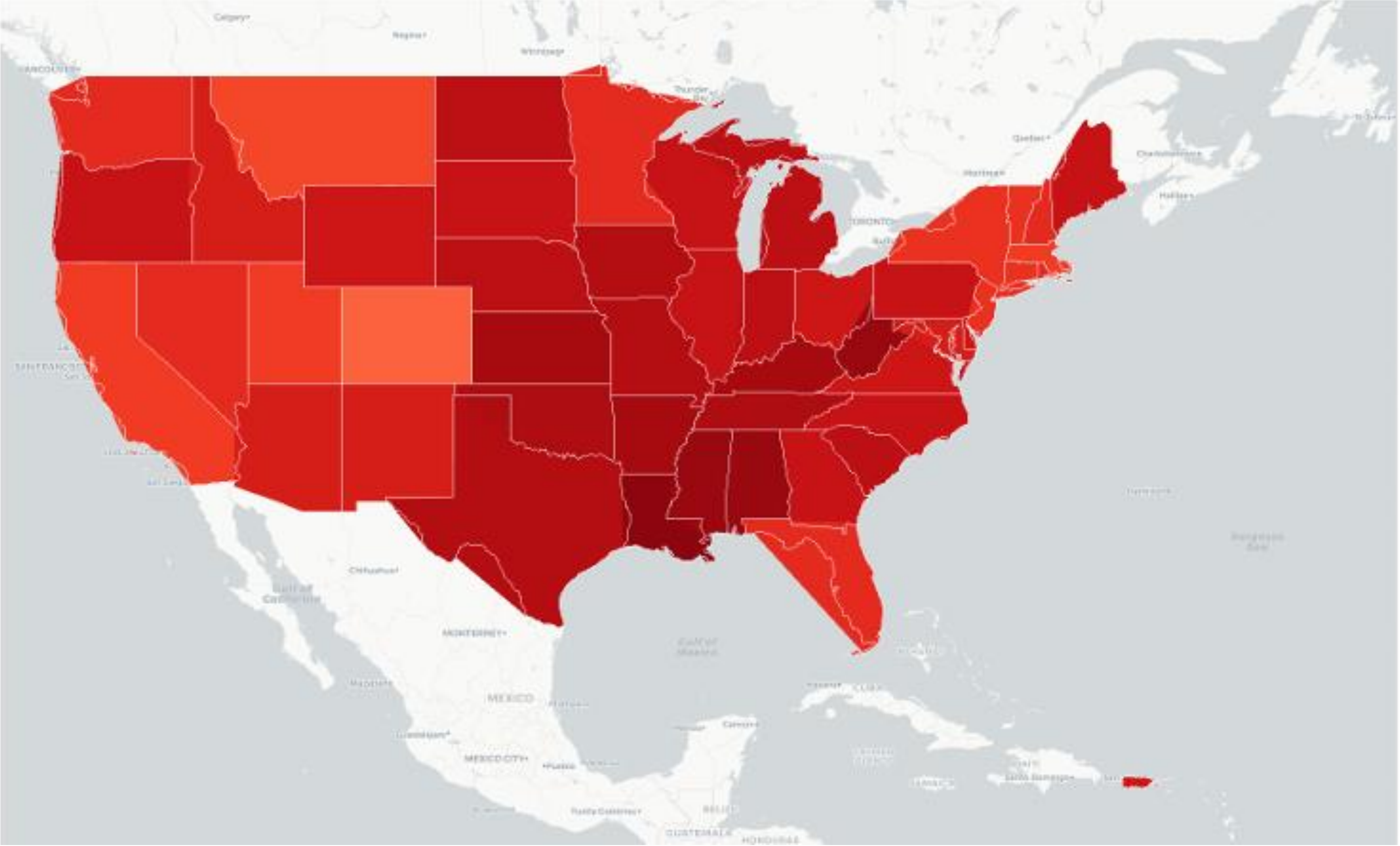
- In a dot map, each dot represents a certain number of observations. Each dot has the same size and value (the number of observations each dot represents).



Dot map showing bus stops worldwide

Choropleth Map

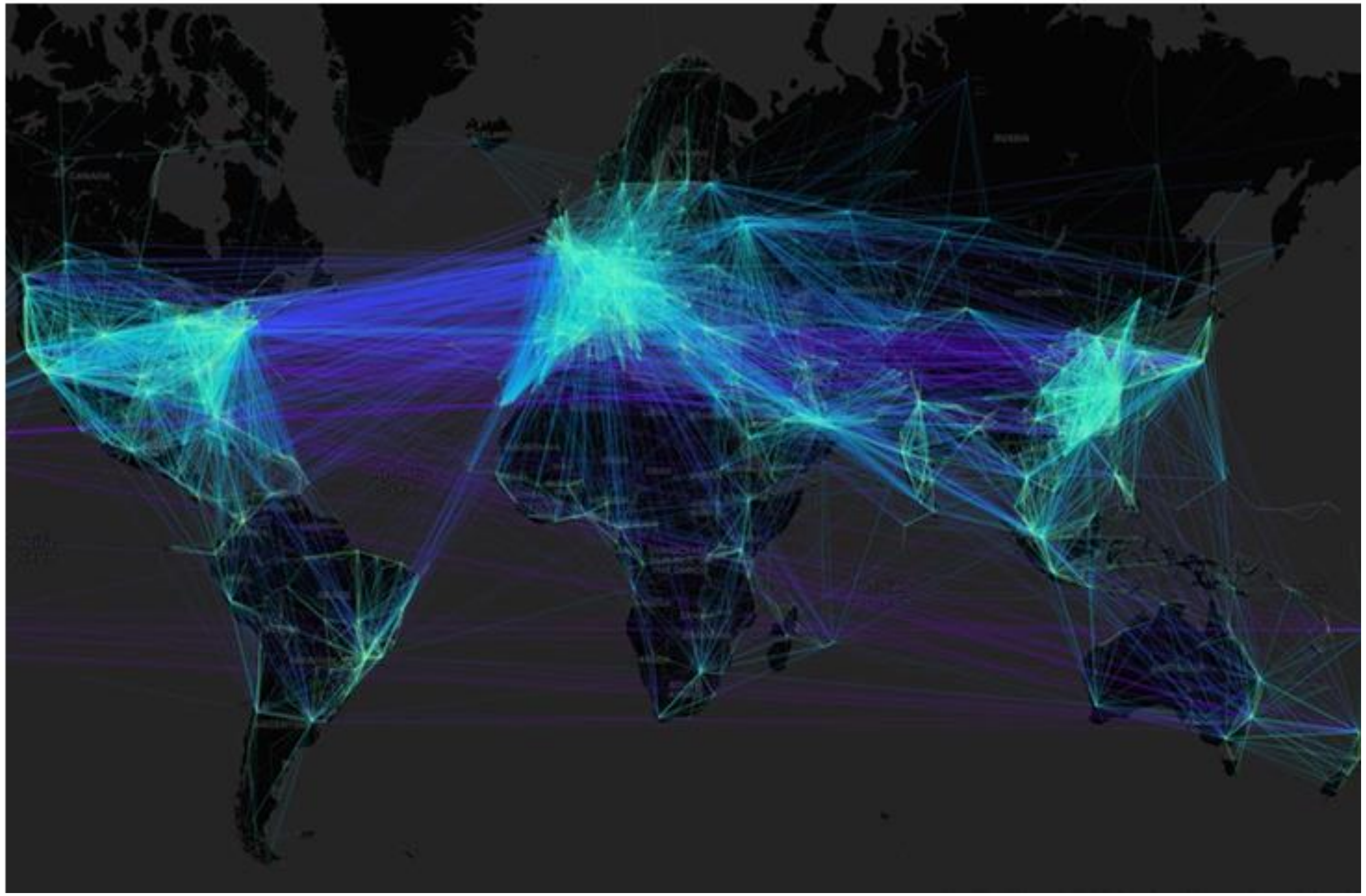
- In a choropleth map, **each tile is colored to encode a variable**.
- For example, a tile represents a **geographic region and countries.**
- Choropleth maps provide a good way to show how a variable varies across a geographic area.



Choropleth map showing a weather forecast for the USA

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



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.
- Choose a line thickness and value so that the lines start to **blend** in dense areas.

What Makes a Good Visualization?

- Most importantly, the visualization should be **self-explanatory** and visually appealing.
  - To make it s**elf-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**.

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**?
3. Sketch the right visualization for both scenarios.

Figure 2: Pie chart showing the top 30 YouTube music channels

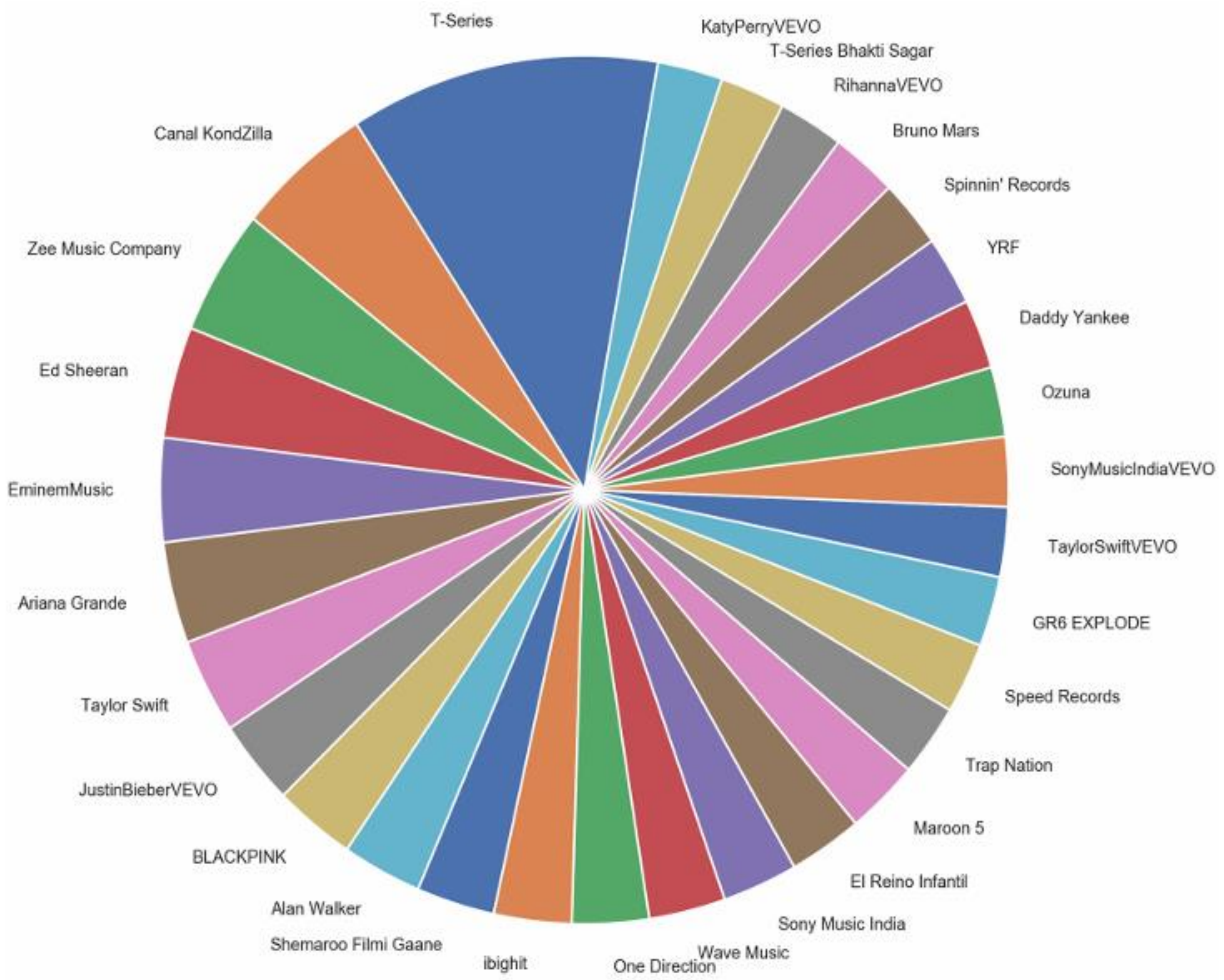
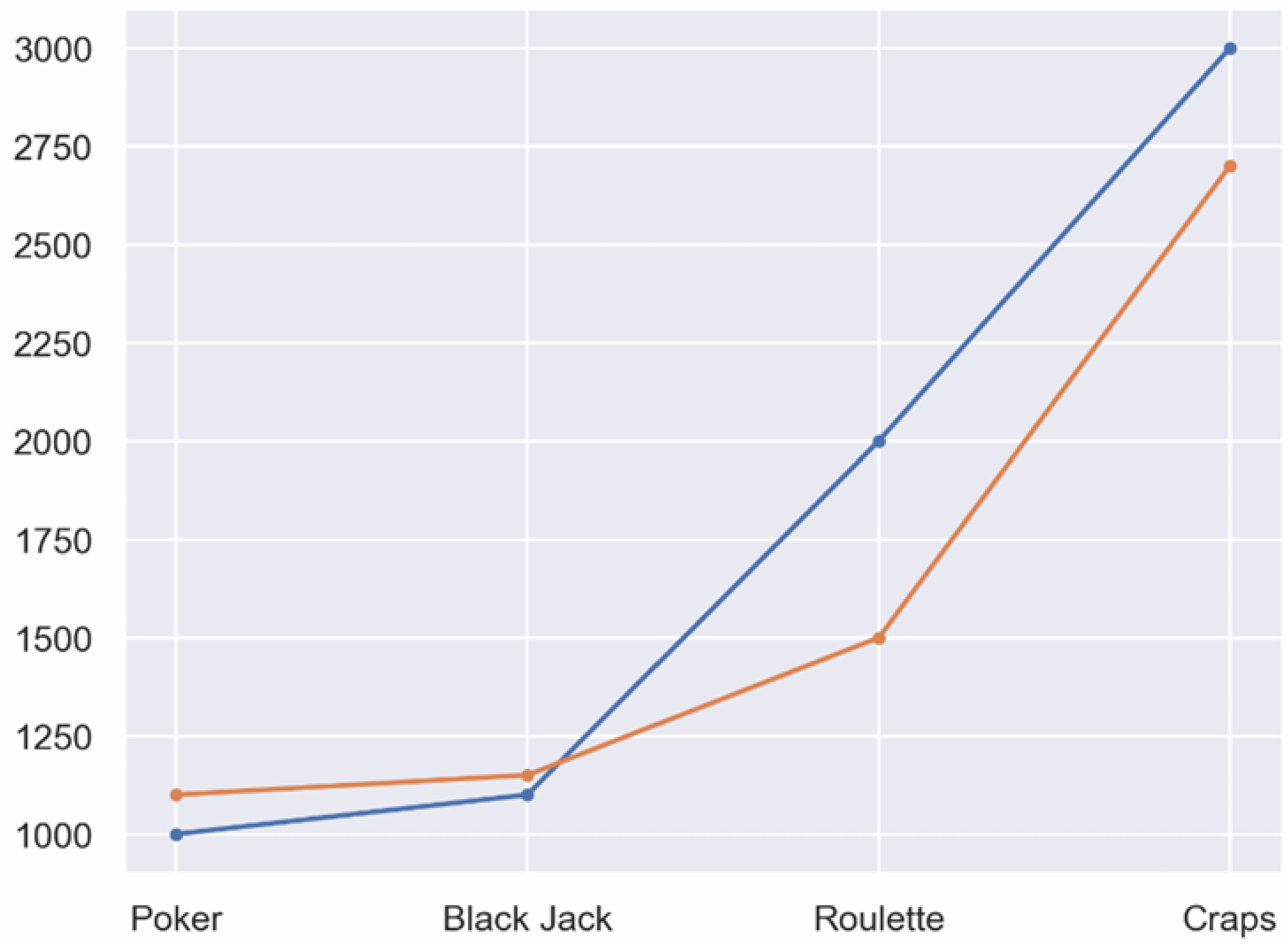
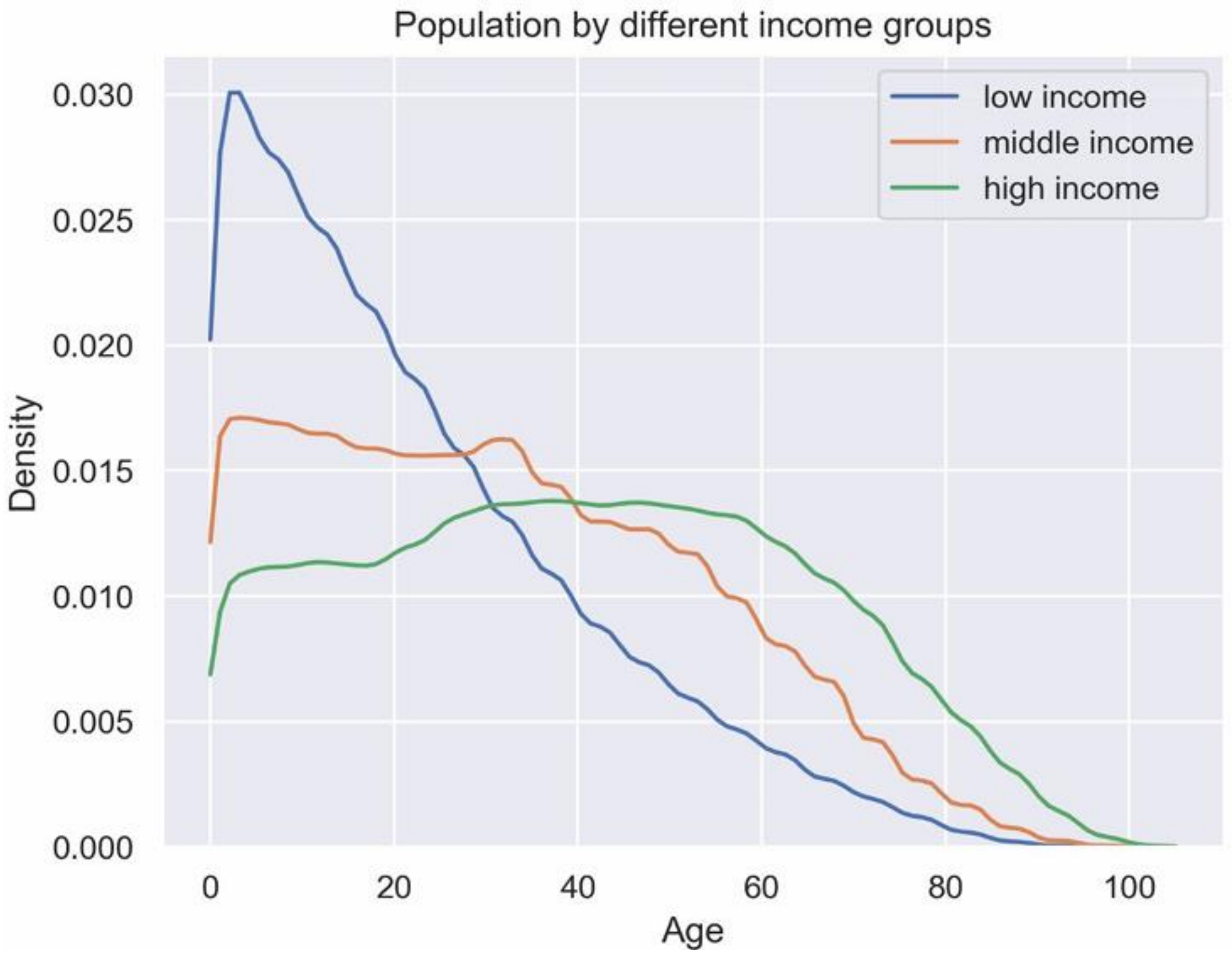


Figure : Line chart displaying casino data for 2 days



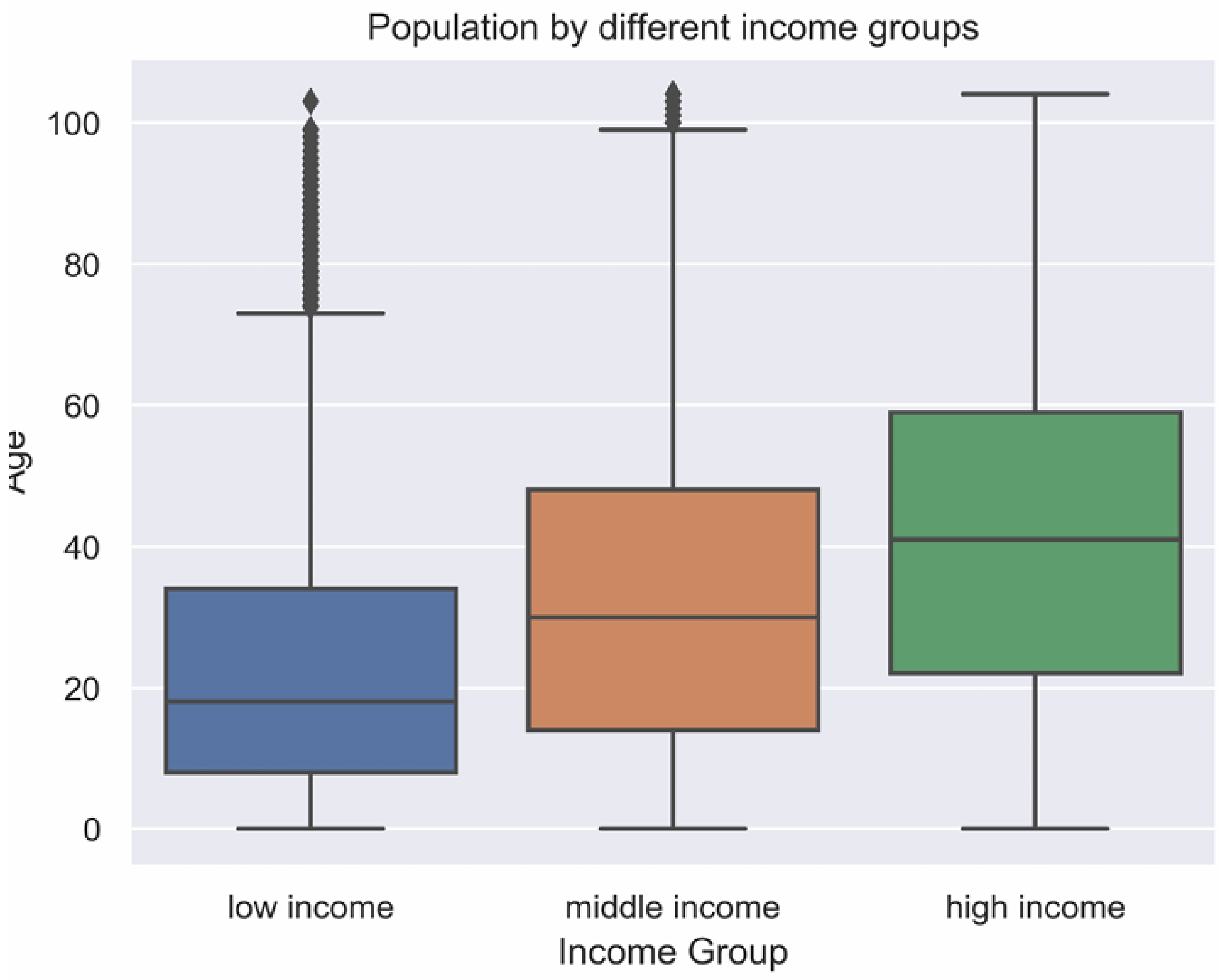
Activity : Choosing a Suitable Visualization

- 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.
- Select the best suitable plot from the following plots.



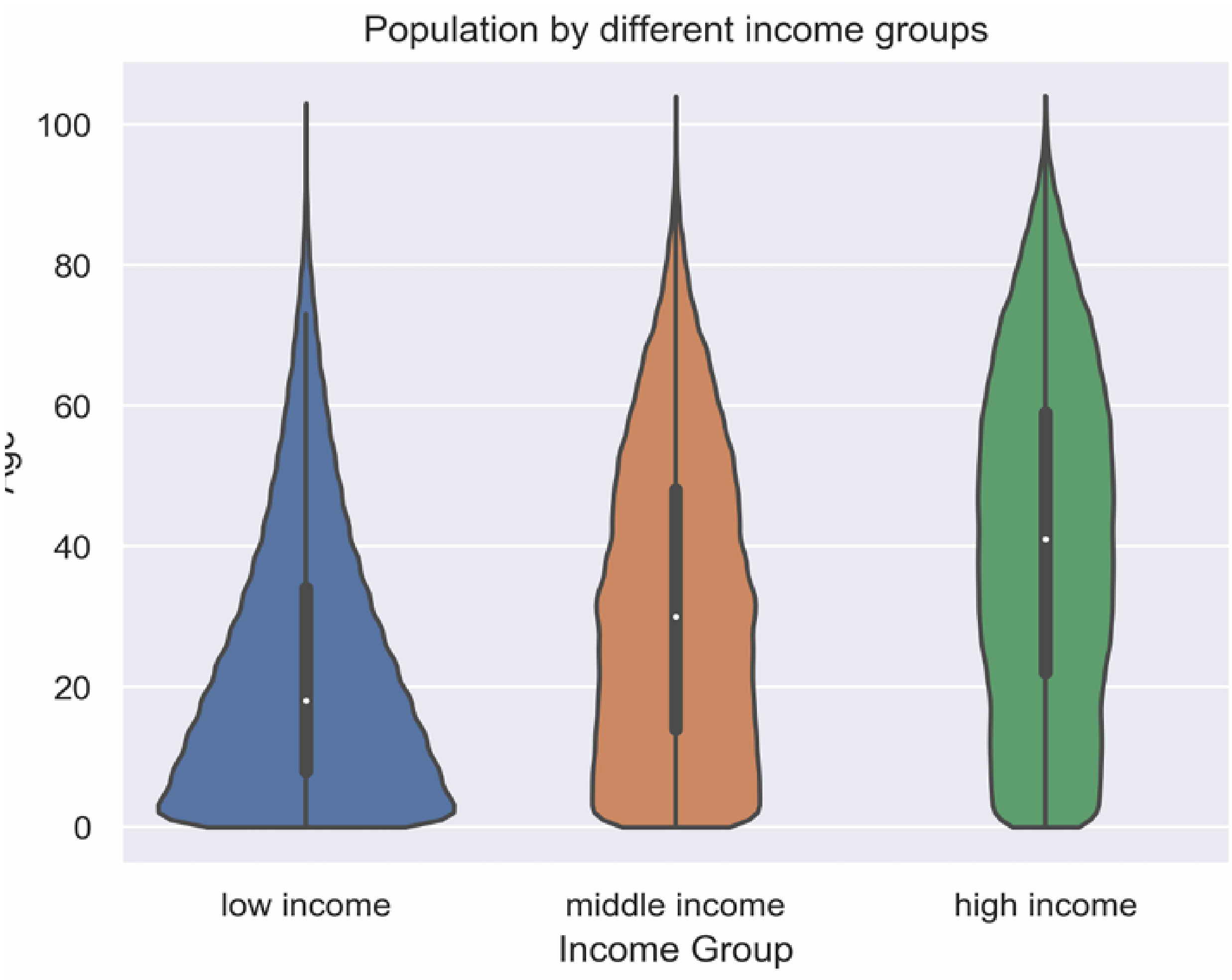
Activity : Choosing a Suitable Visualization

- 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.
- Select the best suitable plot from the following plots.



Activity : 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.
- Select the best suitable plot from the following plots.



End of Module 4

This chapter covered the most important visualizations, categorized into

1. comparison,
2. relation,
3. composition,
4. distribution, and
5. geological plots.

Will be Continued