Module4

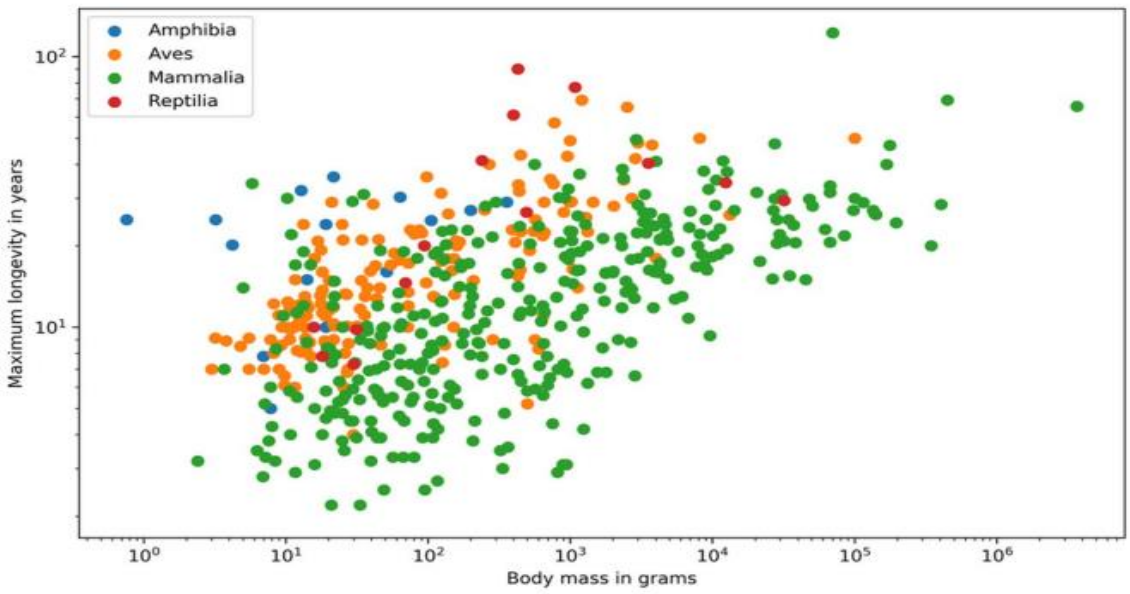
4.1 Introduction:

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.

4.1.1 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 (Fig below) can show patterns, like the relationship between body mass and lifespan in different animal classes, where we can see a positive correlation.



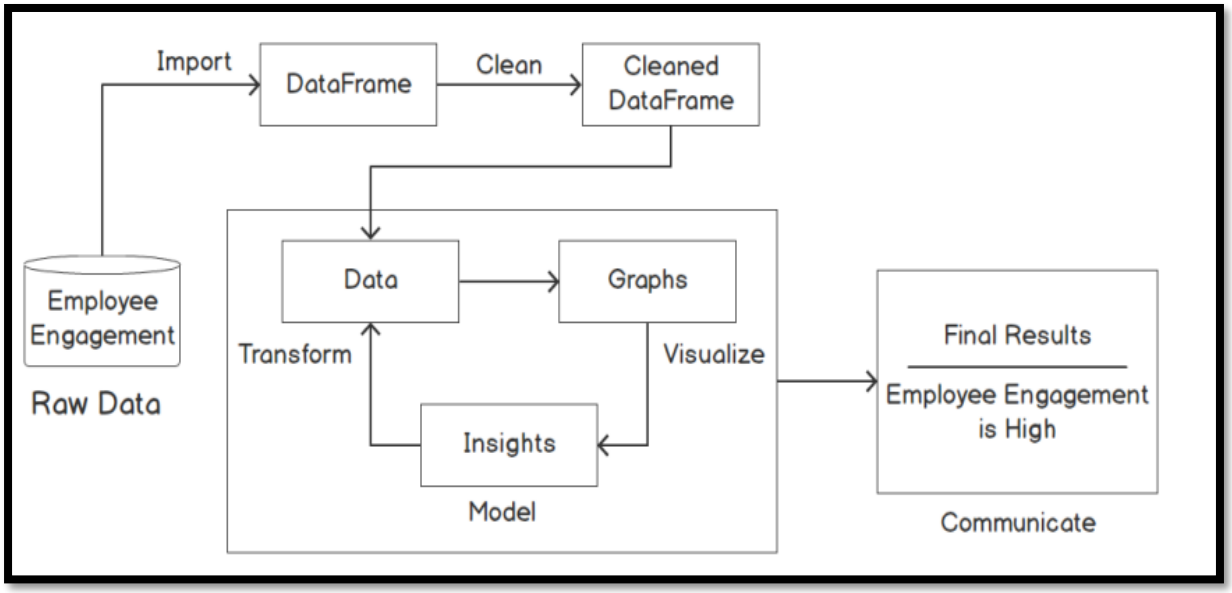
4.1.2 Advantages of Visualizing Data:

- 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.

4.1.3 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.

The flow diagram below shows the steps in the data wrangling process (***to measure employee engagement***), demonstrating how to obtain accurate and actionable data for business analysts.



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.

For example, employee engagement can be measured using raw data from **feedback surveys, employee tenure, exit interviews, and one-on-one meetings**. This data is cleaned and then transformed into graphs based on parameters like **referrals, trust in leadership, and promotion opportunities**. The percentages and insights derived from these graphs help us determine the level of employee engagement.

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 Sources:

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

4.1.4 Overview of Statistics

Statistics: Involves analysing, collecting, interpreting, and representing numerical data. Probability quantifies the likelihood of an event, ranging from 0 to 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

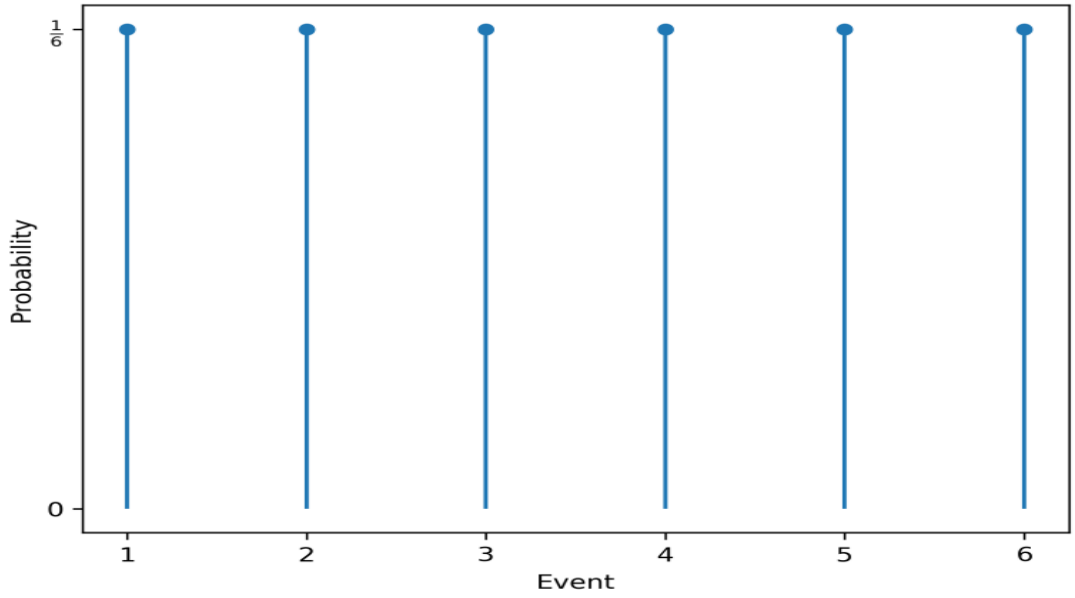
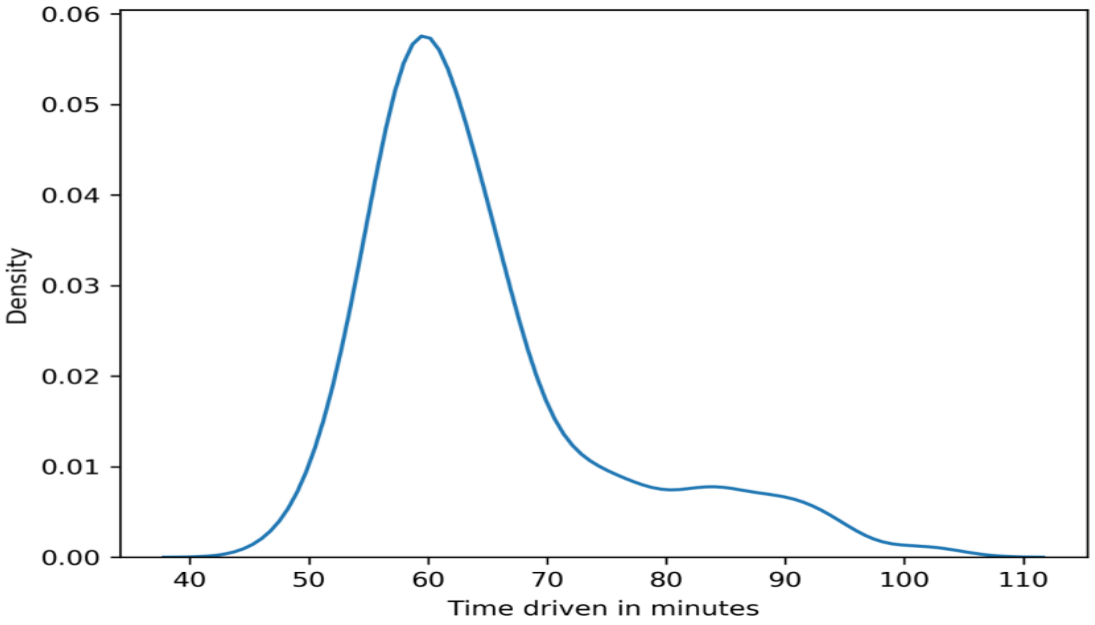


Fig: Continuous probability distribution for the time taken to reach home



4.1.4.1 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.

4.1.4.2 Measures of Dispersion:

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

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

2. **Standard Deviation**: Square root of variance.

3. **Range**: Difference between the largest and smallest values.

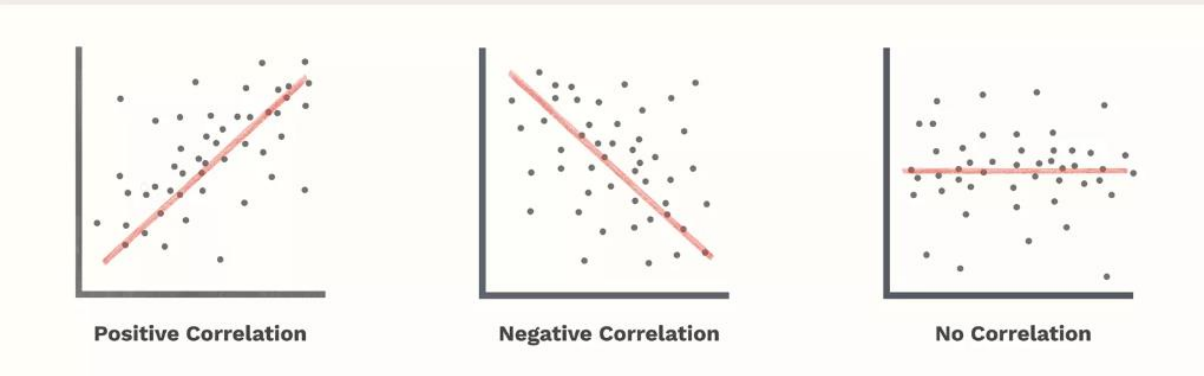
4. **Interquartile Range**: Difference between the upper and lower quartiles. Also called the midspread or middle 50%, this is the difference between the 75th and 25th percentiles, or between the upper and lower quartiles

4.1.4.3 Correlation: Describes the relationship between two variables:

1. **Positive Correlation**: Variables move in the same direction.

2. **Negative Correlation**: Variables move in opposite directions.

3. **Zero Correlation**: No relationship between variables.



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 calculate the variance. However, in this case, the median value ($800) is a more reliable statistical measure because it is less affected by outliers (such as the $1,500 rent). All apartments being in the same locality, it's evident that the $1,500 apartment is significantly higher priced than the others. This simple statistical analysis has effectively helped narrow down our options.

4.1.4.4 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.

4.2 Comparison Plots

Comparison plots are charts used to compare multiple variables or variables over time. Line charts are ideal for visualizing variables over time, while 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)**.
- 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:

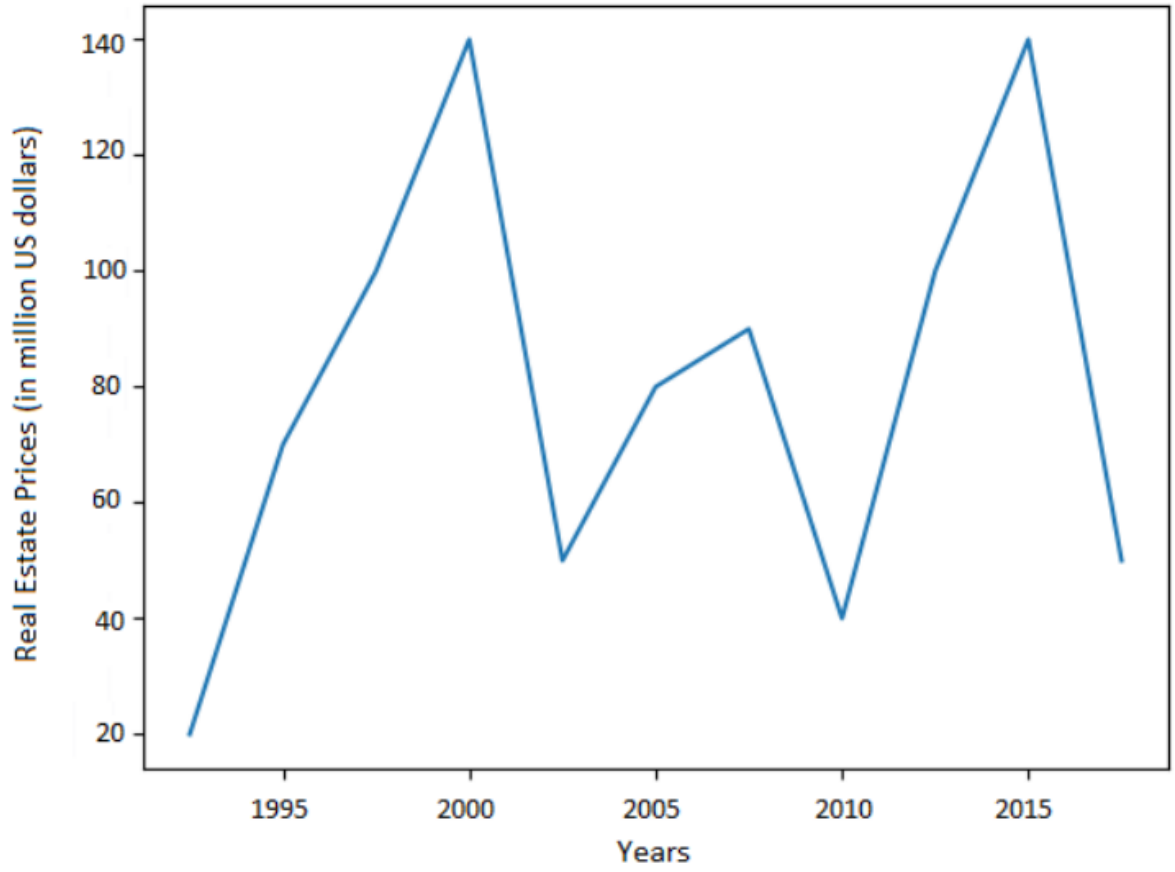


Figure 2.1: Line chart for a single variable

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:

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 Charts

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

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

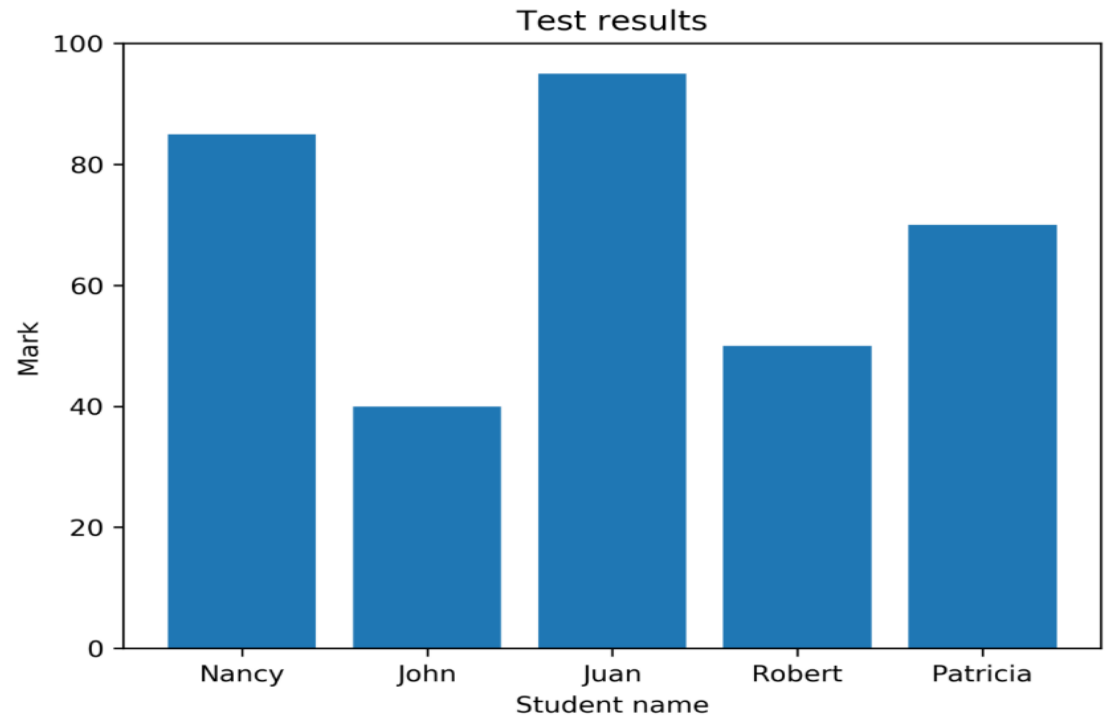


Figure 2.3: Vertical bar chart using student test data

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

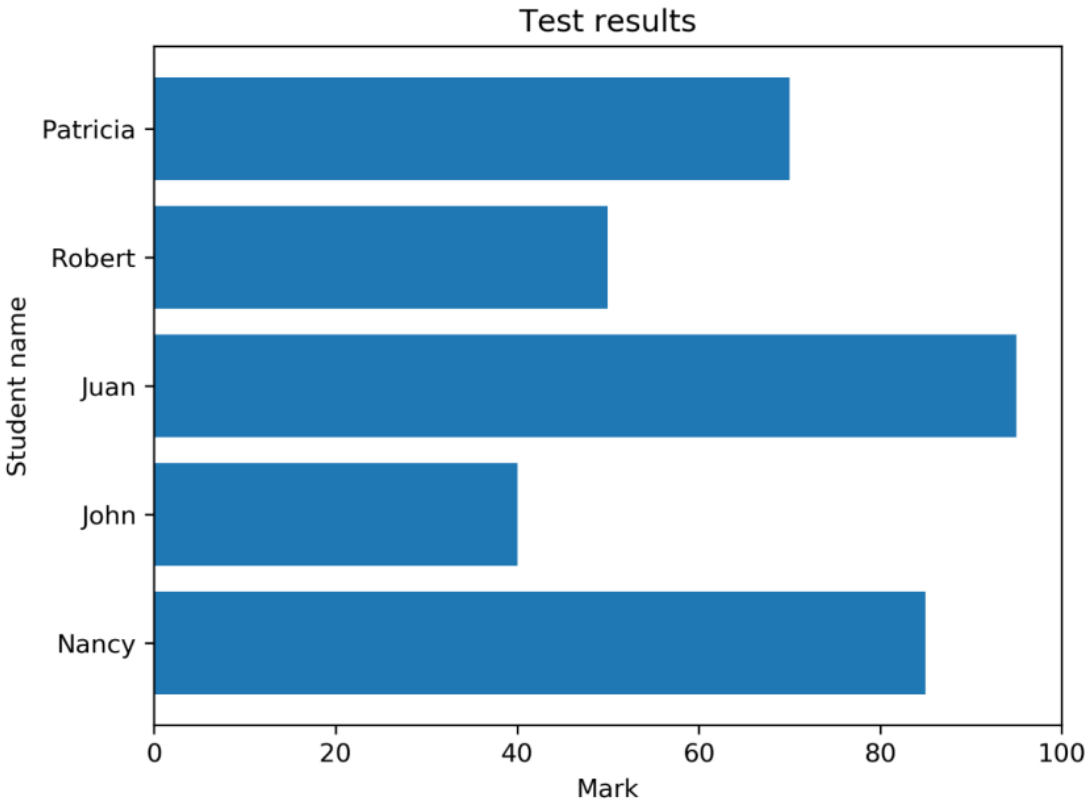


Figure 2.4: Horizontal bar chart using student test data

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.

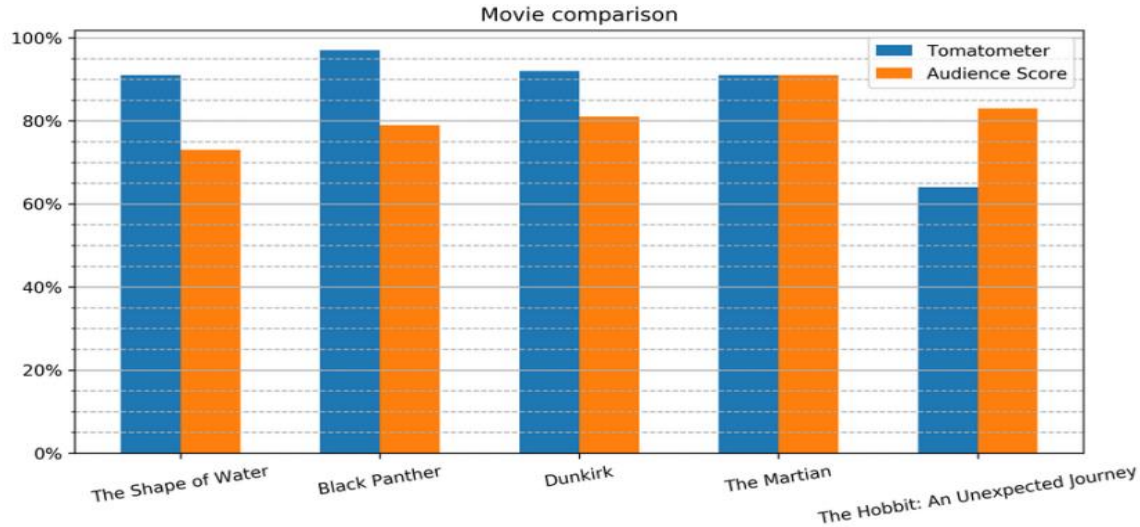


Figure 2.5: Comparative bar chart

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

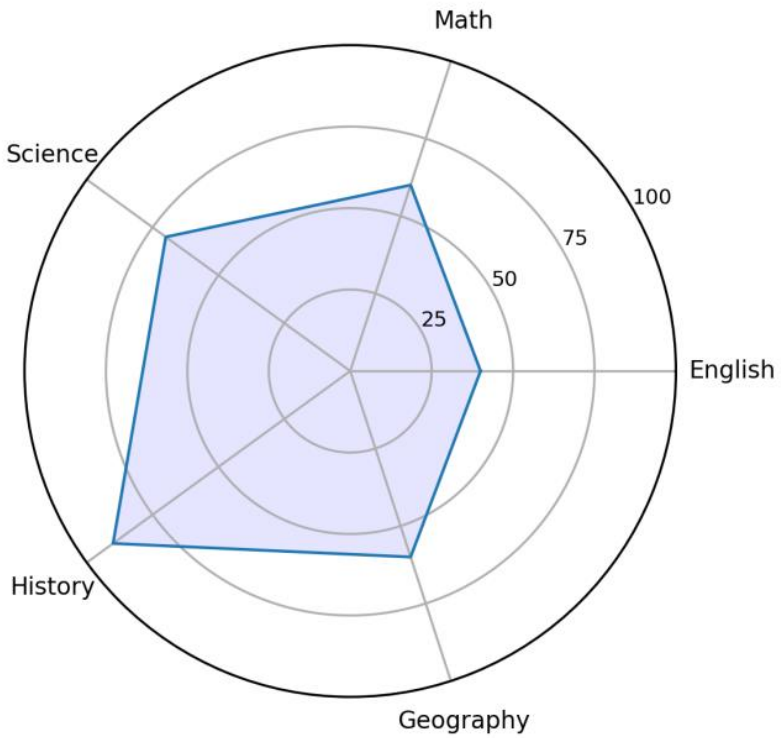
Radar charts (also known as **spider or web charts**) visualize multiple variables with each variable plotted on its own axis, resulting in a polygon.
All axes are arranged radially, starting at the **center** with equal distances between **one another, and have the same scale**.

Uses

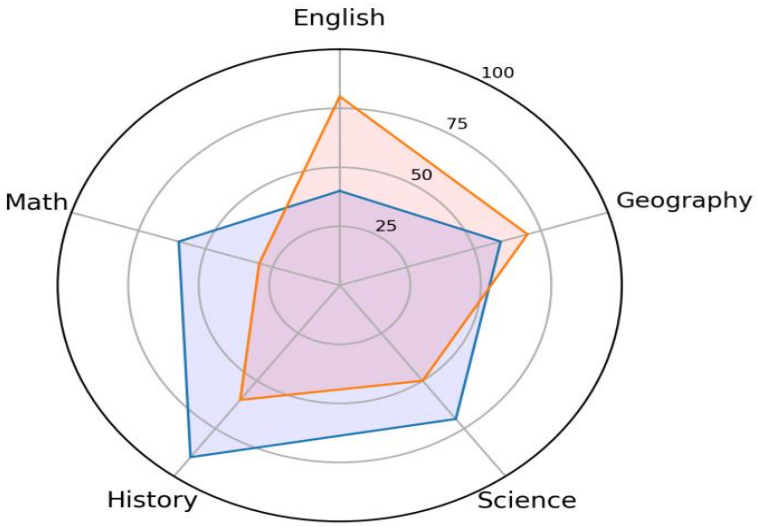
• **Radar charts** are great for comparing multiple quantitative variables for a single group or multiple groups.
• They are also useful for showing which variables score high or low within a **dataset, making them ideal for visualizing performance**.

Examples

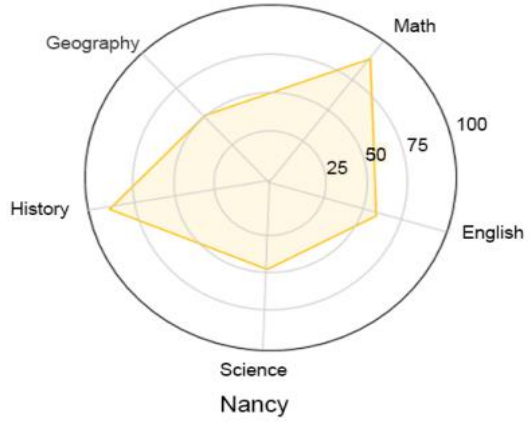
The following diagram shows a radar chart for a single variable. This chart displays data about a student scoring marks in different subjects:

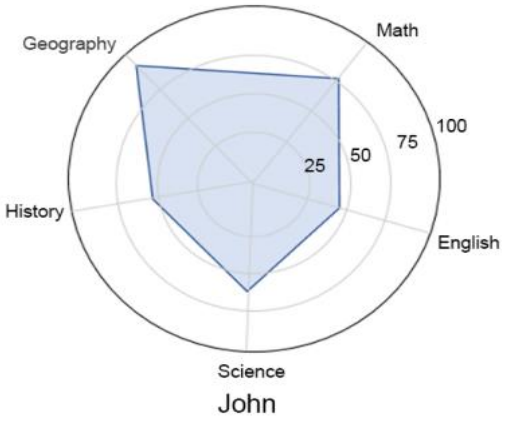


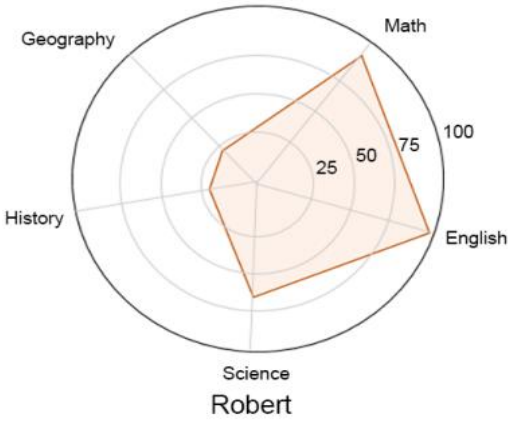
The following diagram shows a radar chart for two variables/groups. Here, the chart explains the marks that were scored by two students in different subjects:

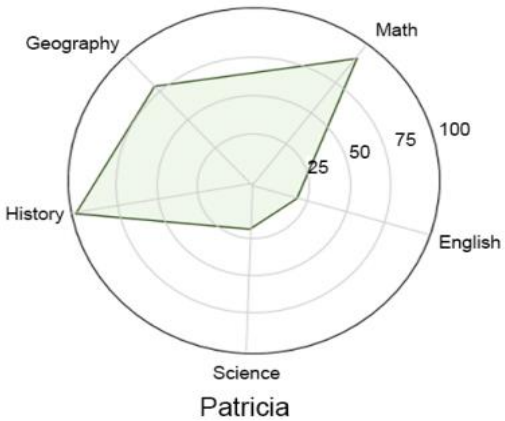


The following diagram shows a radar chart for multiple variables/groups. Each chart displays data about a student's performance in different subjects:









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.

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.

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:

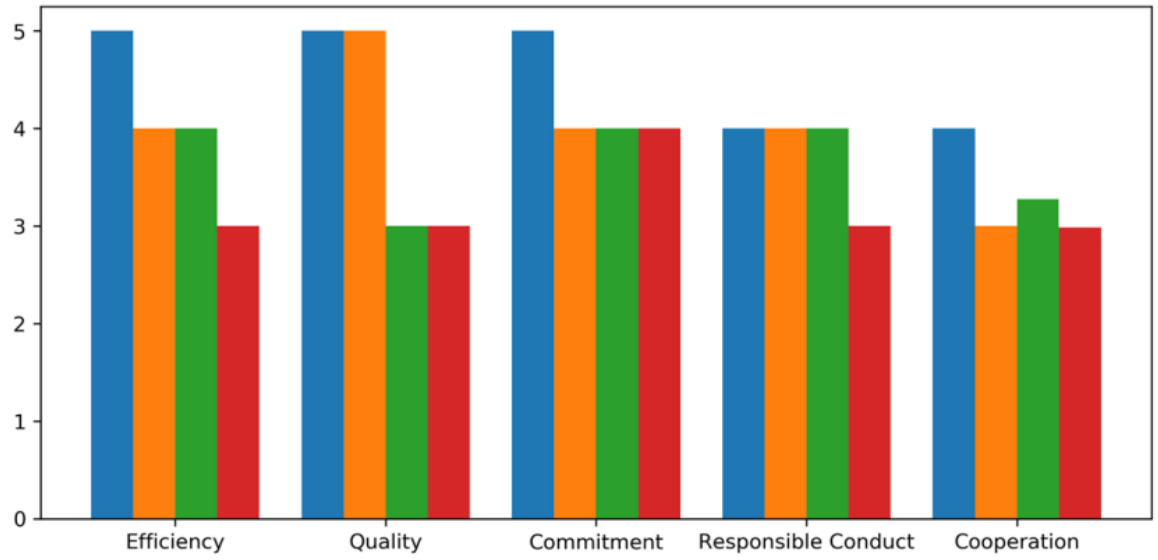


Figure 2.9: Employee skills comparison with a bar chart

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