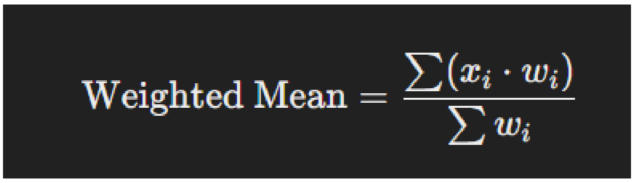
Comprehensive list of statistical topics for **Descriptive Analytics** in the context of **business analysis**:

1. Measures of Central Tendency

- **Purpose**: Summarize the centre or typical value of a dataset.
- **Topics**:
  - Mean (Arithmetic Average)
  - Median
  - Mode (Most Frequent Value)
  - Weighted Mean

A **weighted mean** (also known as a **weighted average**) is a measure of central tendency that takes into account the relative importance, or weight, of each data point in a set. Unlike the simple mean, where all values are treated equally, in the weighted mean, some values contribute more to the average than others based on their weight.

Formula:



$$\text{Weighted Mean} = \frac{\sum(x_i \cdot w_i)}{\sum w_i}$$

Where:

- Xi is each value in the dataset.
- Wi is the weight of the corresponding value.
- ∑ represents the summation over all values.

Steps to Calculate Weighted Mean:

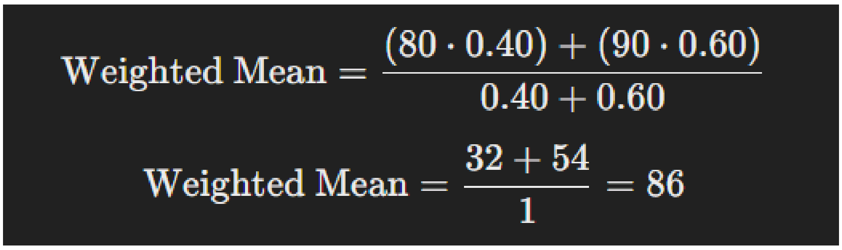
1. Multiply each data value by its respective weight.
2. Add up all the weighted values.
3. Add up all the weights.
4. Divide the sum of the weighted values by the sum of the weights.

Example:

Suppose you are calculating the average grade for a class where the midterm exam has a weight of 40% and the final exam has a weight of 60%. If a student scored 80 on the midterm and 90 on the final exam, the weighted mean of their grades would be:

- Midterm grade (x1) = 80, Weight (w1) = 0.40
- Final grade (x2) = 90, Weight (w2) = 0.60

Now, apply the formula:



$$\text{Weighted Mean} = \frac{(80 \cdot 0.40) + (90 \cdot 0.60)}{0.40 + 0.60}$$

$$\text{Weighted Mean} = \frac{32 + 54}{1} = 86$$

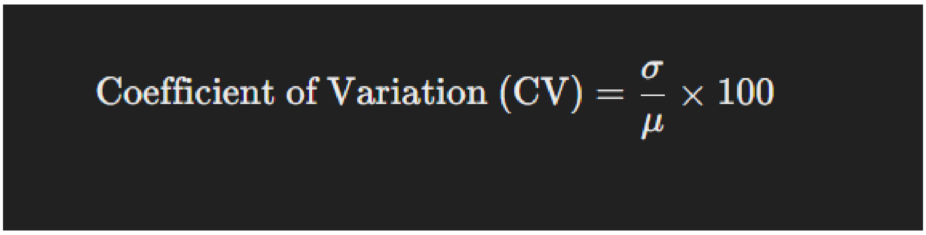
So, the student's weighted average score is **86**.

Why Use Weighted Mean?

- It gives more importance to some values, which is useful when certain data points are more significant than others (e.g., exam scores with different weights).

2. Measures of Dispersion

- **Purpose**: Understand the spread or variability in the data.
- **Topics**:
  - Range (Max - Min)
  - Interquartile Range (IQR)
  - Variance
  - Standard Deviation
  - Coefficient of Variation (CV)
- **Coefficient of Variation (CV)**
- The **Coefficient of Variation (CV)** is a statistical measure of the relative dispersion or spread of a data set. It is the ratio of the **standard deviation** to the **mean** of the dataset. It is expressed as a percentage and helps compare the variability of data from different distributions, especially when the data have different units or scales.
- **Formula:**

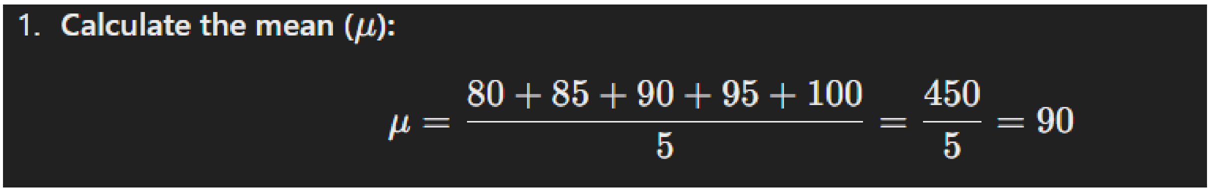


$$\text{Coefficient of Variation (CV)} = \frac{\sigma}{\mu} \times 100$$

- Where:
- σ\sigma is the **standard deviation** of the data.
- μ\mu is the **mean** of the data.
- The result is multiplied by 100 to express CV as a percentage.
- **Interpretation of CV:**
- **Low CV**: A low CV indicates that the data points are relatively close to the mean, showing less variation.
- **High CV**: A high CV indicates greater variability or spread in relation to the mean.

- The Coefficient of Variation is particularly useful when comparing the variability of datasets that have different units of measurement or different means. A higher CV indicates higher relative variability.
- **Steps to Calculate CV:**
- **Find the mean** (μ\mu) of the data set.
- **Calculate the standard deviation** (σ\sigma) of the data set.
- **Divide the standard deviation by the mean** and multiply by 100 to get the CV as a percentage.

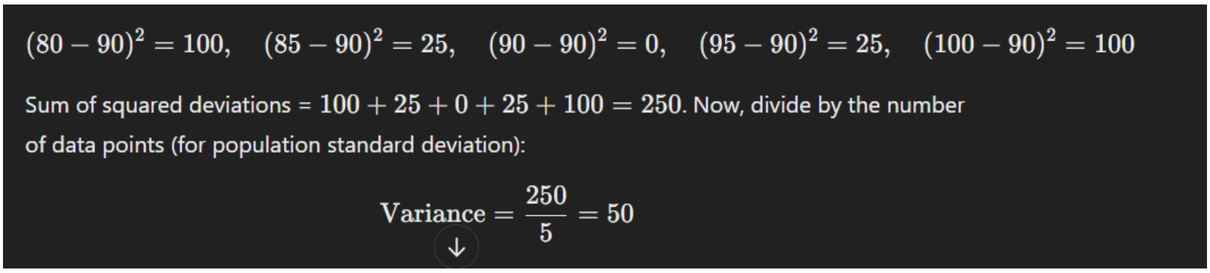
- **Example:**
- Suppose we have two datasets representing the test scores of two classes. We want to compare the variability in scores between these two classes using the Coefficient of Variation.
- **Class A (Scores: 80, 85, 90, 95, 100)**
- **Calculate the mean (μ\mu):**



1. Calculate the mean (μ):

$$\mu = \frac{80 + 85 + 90 + 95 + 100}{5} = \frac{450}{5} = 90$$

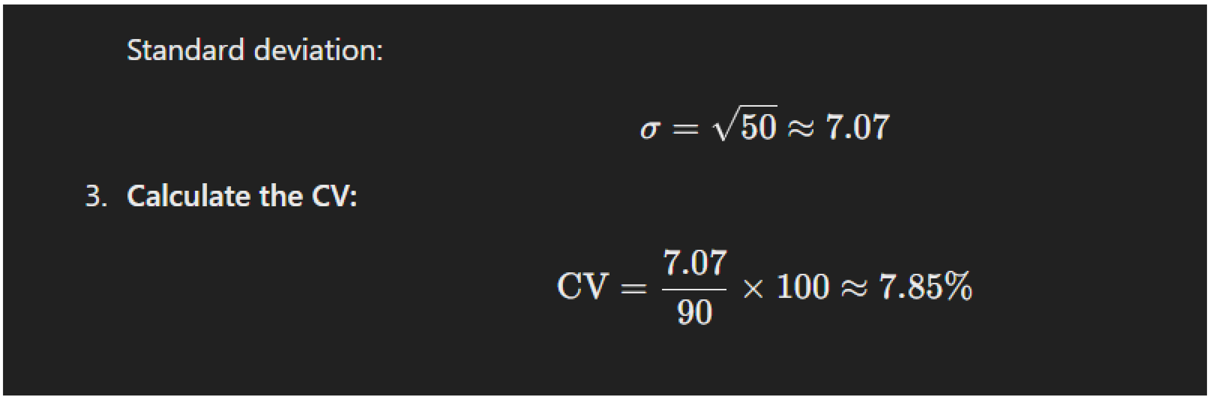
- **Calculate the standard deviation (σ\sigma):** First, find the squared deviations from the mean:



$$(80 - 90)^2 = 100, \quad (85 - 90)^2 = 25, \quad (90 - 90)^2 = 0, \quad (95 - 90)^2 = 25, \quad (100 - 90)^2 = 100$$

Sum of squared deviations = $100 + 25 + 0 + 25 + 100 = 250$. Now, divide by the number of data points (for population standard deviation):

$$\text{Variance} = \frac{250}{5} = 50$$



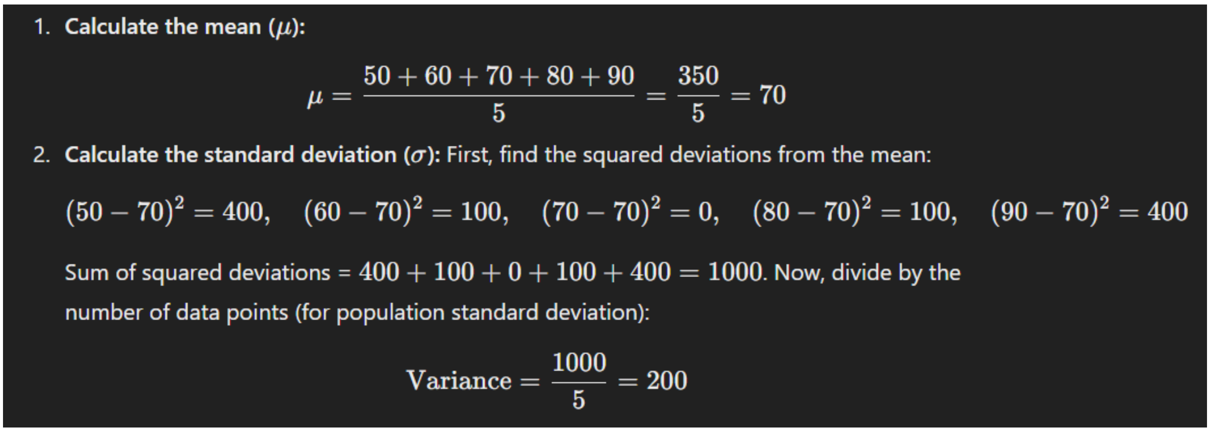
Standard deviation:

$$\sigma = \sqrt{50} \approx 7.07$$

3. **Calculate the CV:**

$$\text{CV} = \frac{7.07}{90} \times 100 \approx 7.85\%$$

- **Class B (Scores: 50, 60, 70, 80, 90)**



1. **Calculate the mean (μ):**

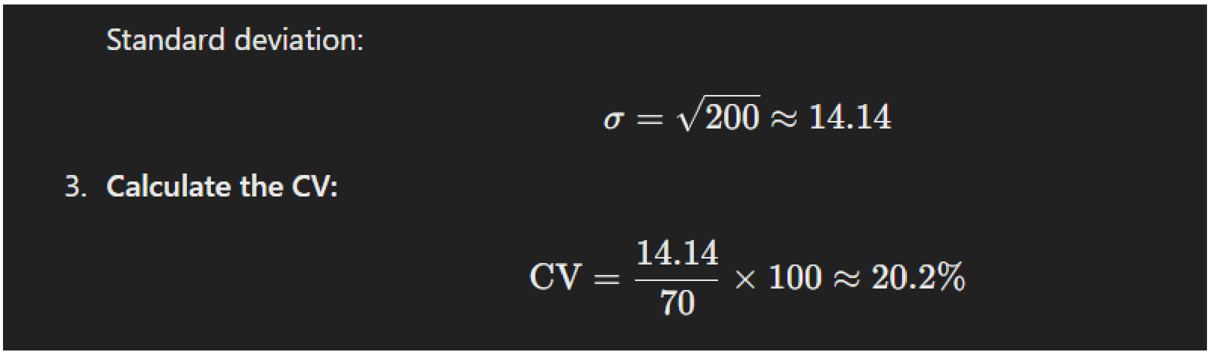
$$\mu = \frac{50 + 60 + 70 + 80 + 90}{5} = \frac{350}{5} = 70$$

2. **Calculate the standard deviation (σ):** First, find the squared deviations from the mean:

$$(50 - 70)^2 = 400, \quad (60 - 70)^2 = 100, \quad (70 - 70)^2 = 0, \quad (80 - 70)^2 = 100, \quad (90 - 70)^2 = 400$$

Sum of squared deviations = $400 + 100 + 0 + 100 + 400 = 1000$. Now, divide by the number of data points (for population standard deviation):

$$\text{Variance} = \frac{1000}{5} = 200$$



Standard deviation:

$$\sigma = \sqrt{200} \approx 14.14$$

3. **Calculate the CV:**

$$\text{CV} = \frac{14.14}{70} \times 100 \approx 20.2\%$$

- **Interpretation:**
- **Class A**: The Coefficient of Variation (CV) is **7.85%**, meaning the scores have relatively low variability compared to the mean (the scores are tightly clustered around 90).

- **Class B**: The Coefficient of Variation (CV) is **20.2%**, meaning the scores have much higher variability relative to the mean (the scores spread more widely around 70).
- From this, we can conclude that **Class B** has greater relative variability in scores compared to **Class A**, even though both datasets contain five data points. The CV allows for a direct comparison of variability across datasets with different means.

---

3. Distribution Analysis

- **Purpose**: Examine the shape and spread of the dataset.
- **Topics**:
  - Skewness (Symmetry of Distribution)
  - Kurtosis (Peakedness or Flatness)
  - Frequency Distribution

Kurtosis (Peakedness or Flatness)

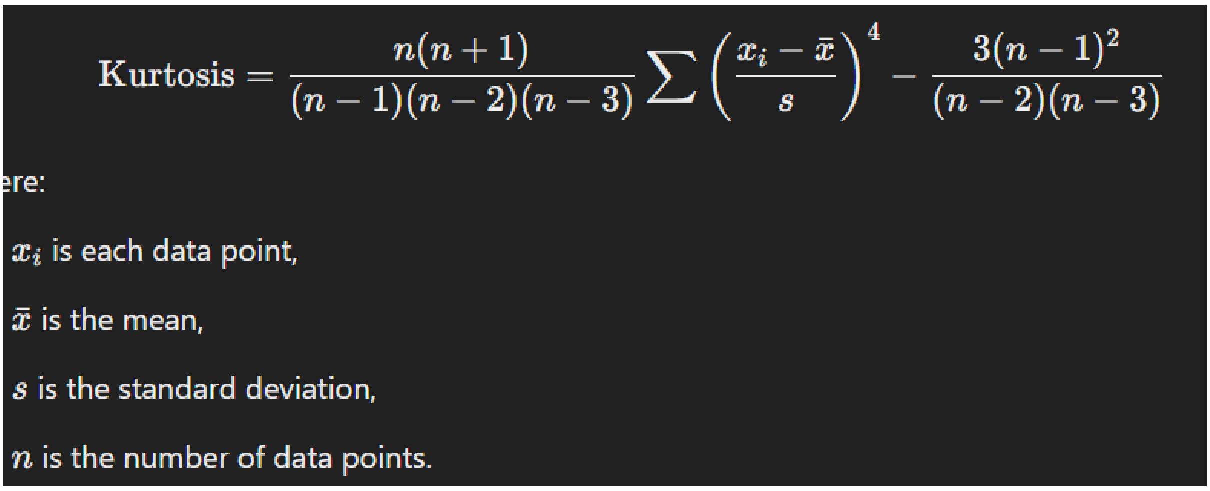
Kurtosis is a statistical measure that describes the shape of a data distribution, particularly the "tailedness" or the extent to which the data are concentrated around the mean. In simpler terms, it tells us how "peaked" or "flat" a distribution is relative to a normal distribution.

There are three main types of kurtosis:

1. **Leptokurtic** (Peaked): The distribution has a higher peak and heavier tails than a normal distribution. The data is more concentrated around the mean, with more extreme values.
2. **Mesokurtic** (Normal): The distribution has a shape similar to a normal distribution, with moderate tails and a moderate peak.
3. **Platykurtic** (Flat): The distribution has a lower peak and lighter tails than a normal distribution, meaning data points are more spread out.

Formula for Kurtosis:

The formula for excess kurtosis is given by:



$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

ere:

x_i is each data point,

$\bar{x}$ is the mean,

s is the standard deviation,

n is the number of data points.

For simplicity, when the kurtosis value is **greater than 0**, the distribution is **leptokurtic** (more peaked). If the kurtosis is **0**, the distribution is **mesokurtic** (similar to the normal distribution). A **negative kurtosis** indicates a **platykurtic** distribution (flatter).

Example of Kurtosis:

Consider the following two distributions:

1. **Leptokurtic Distribution**: A distribution of exam scores where most students scored either very high or very low, with few average scores.
2. **Platykurtic Distribution**: A distribution of scores where most students' scores are spread out more evenly across the scale, with fewer extreme high or low scores.

In the leptokurtic case, you would expect a sharp peak around the mean and more extreme scores (higher or lower). In the platykurtic case, the data would be more evenly spread, with a lower peak.

➢ Frequency Distribution

A **frequency distribution** is a way to organize data to show how often each value or range of values (called "bins") occurs in a dataset. It is a helpful tool to summarize large datasets and observe patterns, such as the most common values and the spread of data.

The frequency distribution can be represented in:

1. **Table format**: A simple table with values and their corresponding frequencies.

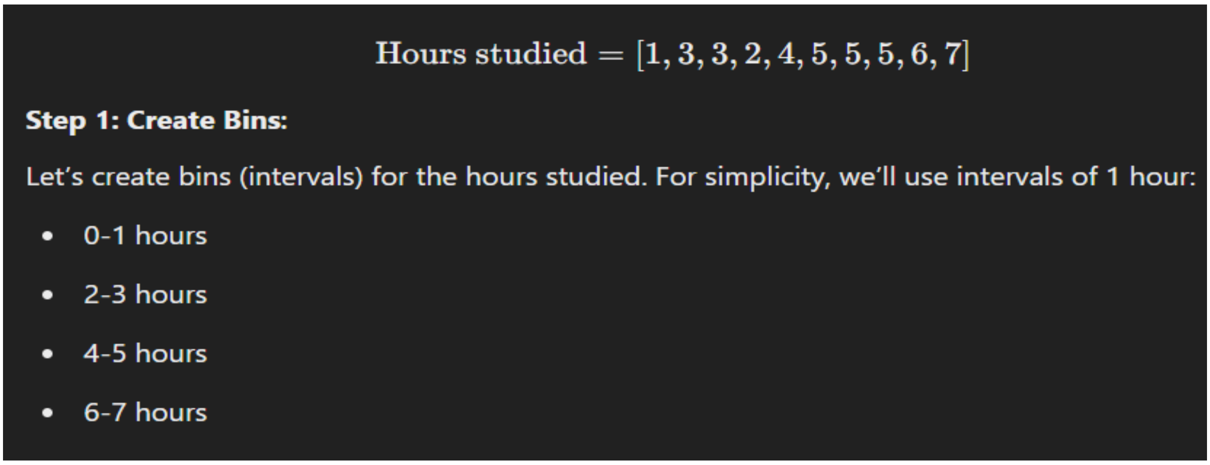
2. **Histogram**: A graphical representation of the frequency distribution.

Steps to Create a Frequency Distribution:

1. **Sort the data**: Arrange the data in ascending order.
2. **Create intervals or bins**: Divide the range of values into equal intervals (for continuous data).
3. **Count the frequency**: Count how many data points fall into each interval.
4. **Summarize**: Create a table or graph that summarizes the frequencies.

Example of Frequency Distribution:

Suppose we have the following dataset representing the number of hours five students studied for an exam:



$$\text{Hours studied} = [1, 3, 3, 2, 4, 5, 5, 5, 6, 7]$$

Step 1: Create Bins:

Let's create bins (intervals) for the hours studied. For simplicity, we'll use intervals of 1 hour:

- 0-1 hours
- 2-3 hours
- 4-5 hours
- 6-7 hours

Step 2: Count Frequencies:

Now, count how many values fall into each bin:

- **0-1 hours**: 1 value (1 hour)
- **2-3 hours**: 3 values (2, 3, 3 hours)
- **4-5 hours**: 4 values (4, 5, 5, 5 hours)
- **6-7 hours**: 2 values (6, 7 hours)

Step 3: Create the Frequency Distribution Table:

Hours Studied	Frequency
0-1 hours	1
2-3 hours	3
4-5 hours	4
6-7 hours	2

This table shows how many students studied within each range of hours.

Step 4: Plot the Data (Optional):

We can also create a histogram to visualize this frequency distribution, where the x-axis represents the hours studied (or the bins), and the y-axis represents the frequency (the number of students).

Why Use Frequency Distribution?

1. **Summarization**: It condenses large datasets into a more understandable form.
2. **Visualization**: Helps in visualizing patterns or trends in the data.
3. **Comparison**: Allows comparison of frequencies across different groups or datasets.
4. **Identifying Outliers**: Helps identify data points that fall outside the expected range or frequency.

In this example, we see that most students studied between 2-5 hours, and fewer students studied for longer periods.

Summary:

- **Kurtosis** tells us how "peaked" or "flat" a distribution is compared to a normal distribution. It helps understand the shape of the data's distribution, such as whether it has heavy tails (leptokurtic) or lighter tails (platykurtic).
- A **frequency distribution** helps summarize data by showing how often values occur in certain intervals or bins, providing insights into the distribution, trends, and patterns within the data.

- Histograms

4. Data Summarization Techniques

- **Purpose**: Aggregate data for easier interpretation.
- **Topics**:
    - Count, Sum, and Averages
    - Percentages and Proportions
    - Cumulative Sums (Running Totals)

Cumulative Sums (Running Totals) - Explanation

A **cumulative sum** (or running total) is the running total of a sequence of numbers. It is calculated by adding each successive number to the total of all the previous numbers in the sequence. This means that each value in the cumulative sum is the sum of all the previous values, including the current value.

Example:

Let's say you have a list of numbers representing the sales of a product over several days:

Day	Sales ($)
1	100
2	150
3	200
4	120

To calculate the cumulative sum for these sales:

1. **Day 1**: The cumulative sum is simply the sales on Day 1:
    - Cumulative sum = 100
2. **Day 2**: Add the sales of Day 2 to the cumulative sum of Day 1:
    - Cumulative sum = 100 (Day 1) + 150 (Day 2) = 250
3. **Day 3**: Add the sales of Day 3 to the cumulative sum of Day 2:
    - Cumulative sum = 250 (Day 2) + 200 (Day 3) = 450

4. **Day 4**: Add the sales of Day 4 to the cumulative sum of Day 3:
   - Cumulative sum = 450 (Day 3) + 120 (Day 4) = 570

So, the cumulative sum for each day would look like this:

Day	Sales ($)	Cumulative Sum ($)
1	100	100
2	150	250
3	200	450
4	120	570

Key Points:

- The **cumulative sum** grows as you keep adding the next number in the sequence.
- It helps to visualize or track how a total accumulates over time.
- It's commonly used in tracking sales, expenses, performance, or any scenario where running totals are needed.

   - Grouped Summaries (e.g., by region, product, or segment)

---

5. Measures of Relationship

- **Purpose**: Explore relationships between two or more variables.
- **Topics**:
  - Correlation (e.g., Pearson, Spearman)
  - Covariance
  - Cross-tabulation (Contingency Tables)

Cross-tabulation (Contingency Tables) - Explanation

A **cross-tabulation** (also known as a **contingency table**) is a data analysis technique used to examine the relationship between two or more categorical variables. It displays the frequency distribution of variables, showing how the categories of one variable intersect with the categories of another variable.

In simpler terms, it is a way to display the joint distribution of two categorical variables in a table format. The rows represent one variable's categories, the columns represent another variable's categories, and the cells in the table show the frequency or count of observations that fall into the corresponding categories.

Example:

Let's say we have a survey of 100 people, asking them about their preferred type of drink (coffee or tea) and whether they prefer hot or cold drinks. We want to create a contingency table to analyze the relationship between the type of drink and the temperature preference.

Survey Data:

Respondent	Type of Drink	Temperature Preference
1	Coffee	Hot
2	Tea	Cold
3	Coffee	Hot
4	Tea	Hot
5	Tea	Cold
6	Coffee	Cold

Respondent	Type of Drink	Temperature Preference
7	Coffee	Hot
8	Tea	Hot

Step 1: Create a Contingency Table

Now, we create a table where the rows represent **Type of Drink** and the columns represent **Temperature Preference**.

Type of Drink \ Temperature Preference	Hot	Cold	Total
Coffee	3	1	4
Tea	3	2	5
Total	6	3	9

Step 2: Analyse the Table

- The table shows that **3 people prefer hot coffee**, while **1 person prefers cold coffee**.
- **3 people prefer hot tea**, and **2 people prefer cold tea**.
- This can help identify patterns or trends, such as whether certain drink types are more associated with hot or cold preferences.

Key Points:

- **Rows**: Represent one categorical variable (e.g., Type of Drink).
- **Columns**: Represent the other categorical variable (e.g., Temperature Preference).
- **Cells**: Show the frequency or count of observations that fall into the corresponding combination of categories.

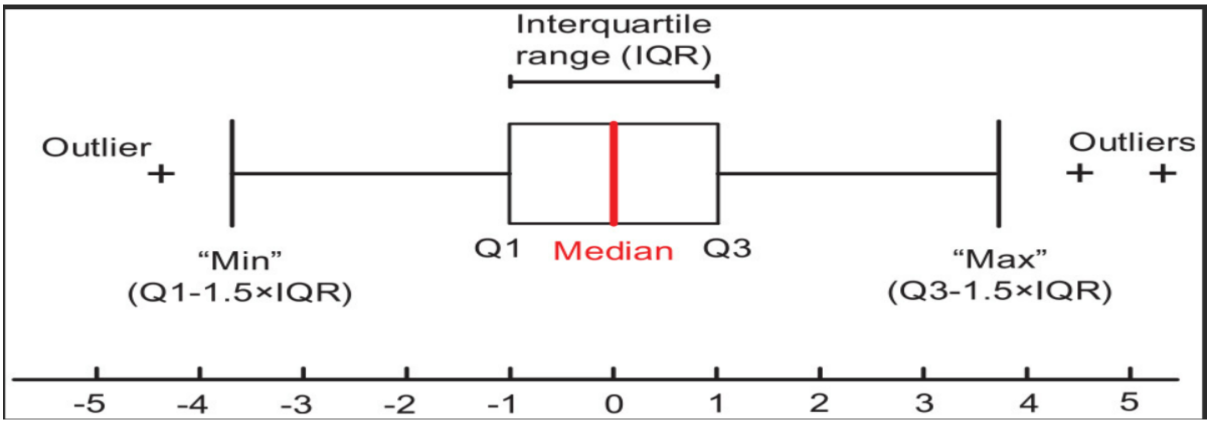
- **Marginal Totals**: The totals for rows and columns give insights into the overall distribution of each variable.

Uses of Cross-Tabulation:

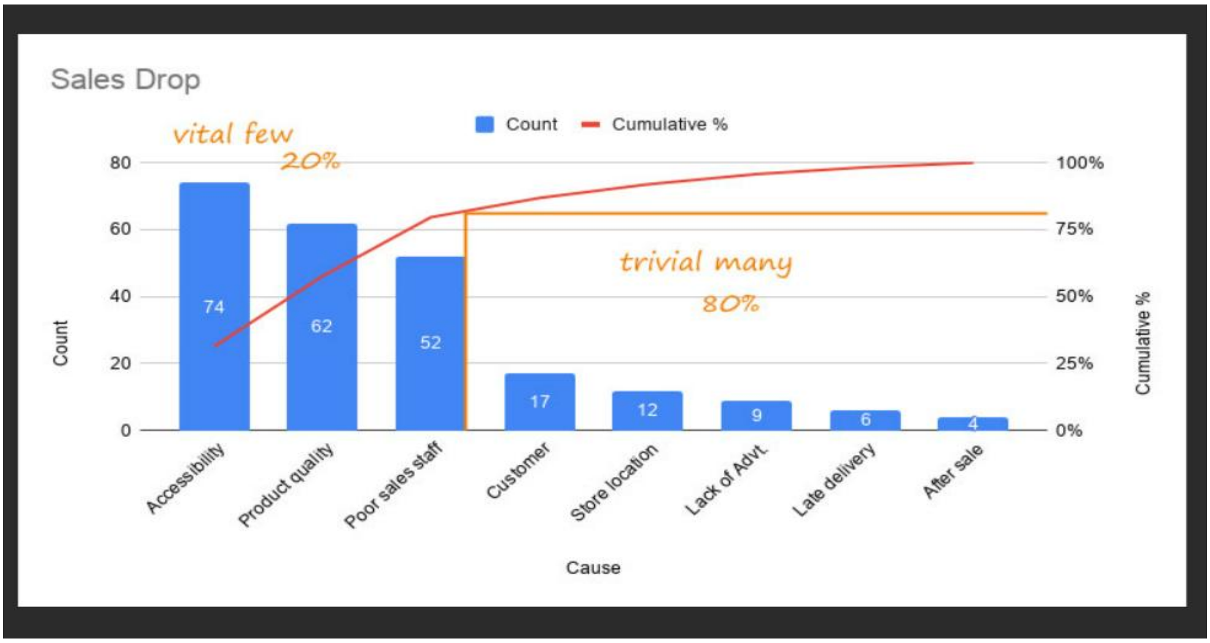
- It helps to understand the relationship between two categorical variables.
- It is commonly used in market research, social sciences, and business analytics.
- It provides a simple way to summarize data and identify trends or associations between variables.

6. Visual Representation of Data

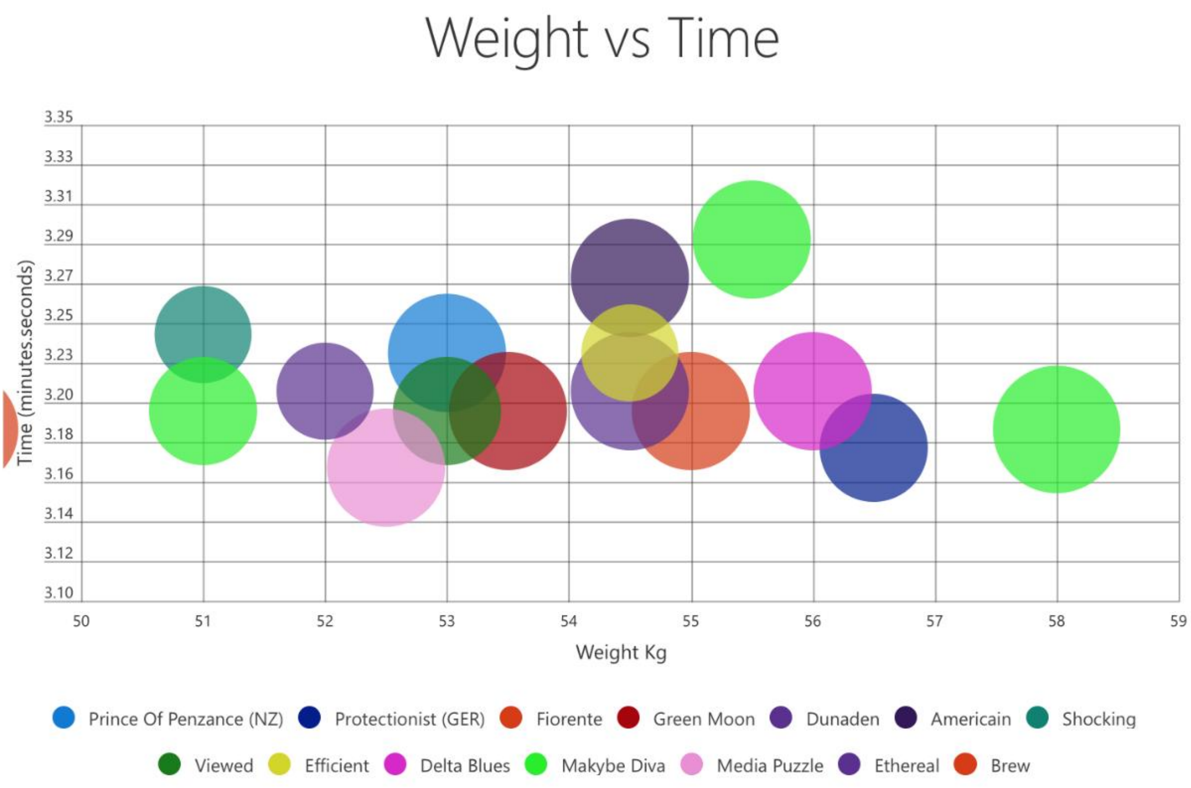
- **Purpose**: Communicate data insights effectively.
- **Topics**:
  - Charts and Graphs:
    - Bar Chart
    - Line Chart
    - Pie Chart
    - Scatter Plot
    - Box Plot



  - Heatmaps (e.g., Correlation Heatmap)
  - Pareto Charts (80/20 Rule Visualization)



- Bubble Charts



See in the chart above how the weight of the last 15 years' worth of winners has influenced the final times. The size of each winner's dot is dictated by their age. Roll your mouse over each dot to learn more about the winner.

---

7. Percentiles and Quartiles

- **Purpose**: Divide the data into equal parts to understand distributions.
- **Topics**:
  - Percentiles (e.g., 25th, 50th, 75th)
  - Quartiles (Q1, Q2/Median, Q3)
  - Outlier Detection (e.g., using IQR)
