Methods of EDA

Mean

The mean, also known as the average, is a measure of central tendency in a dataset. It is calculated by summing up all the values in the dataset and then dividing the sum by the total number of values. The mean is a commonly used statistic to represent the "center" of a dataset.

Mathematically, the mean (μ) of a dataset with n observations is calculated using the following formula:

$$\mu = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Where:

- μ represents the mean or average.
- $x_1, x_2, \ldots, x_n$ are the individual values in the dataset.
- n is the total number of values in the dataset.

For example, consider the dataset {10, 15, 20, 25, 30}. To find the mean:

$$\mu = \frac{10+15+20+25+30}{5}$$

$$\mu = \frac{100}{5} = 20$$

So, the mean of this dataset is 20.

Mean

- The mean is sensitive to outliers: Extreme values can significantly impact the mean.
- Mean provides a measure of central tendency but may not accurately represent the typical value if outliers are present.
- Other measures such as median and mode are valuable alternatives.
- Important to use alternative measures, especially with skewed or non-normally distributed data.
- Utilizing a combination of mean, median, and mode offers a more comprehensive understanding of the dataset's central tendency.

Median

The median is another measure of central tendency in a dataset. Unlike the mean, which is calculated by summing up all the values and dividing by the total number of values, the median is the middle value of a dataset when it is arranged in ascending or descending order.

To find the median:

- Arrange the data in ascending or descending order.
- If the number of observations (n) is odd, the median is the middle value.
- If the number of observations (n) is even, the median is the average of the two middle values.

For example, consider the dataset {5, 10, 15, 20, 25, 30}. To find the median:

1. Arrange the data in ascending order: {5, 10, 15, 20, 25, 30}.
2. Since there are 6 observations (an even number), the median is the average of the two middle values, which are 15 and 20.
3. Median = (15 + 20) / 2 = 17.5.

So, the median of this dataset is 17.5.

The median is particularly useful when the dataset contains outliers or when the data is skewed, as it gives a better representation of the "middle" of the data compared to the mean.

Mode

The mode is a measure of central tendency in a dataset, representing the value that appears most frequently. In other words, it is the value that occurs with the highest frequency in the dataset.

A dataset can have:

1. **Unimodal distribution:** When there is one mode, meaning one value that occurs most frequently.
2. **Bimodal distribution:** When there are two modes, meaning two values that occur with the same highest frequency.
3. **Multimodal distribution:** When there are more than two modes, meaning multiple values that occur with the same highest frequency.

For example, consider the dataset {2, 3, 3, 5, 5, 5, 7, 8, 8}. In this dataset:

- 5 occurs three times, making it the mode because it appears more frequently than any other value.

In another dataset {2, 3, 3, 5, 5, 6, 6, 8, 8}, there are two modes:

- Both 3 and 5 occur twice, making them the modes because they appear more frequently than any other value.

The mode is often used in conjunction with other measures of central tendency, such as the mean and median, to provide a more complete description of the dataset. It is particularly useful when trying to understand the most common or typical value in a dataset.

Variance

Variance is a measure of dispersion or spread in a dataset, indicating how much the values in the dataset differ from the mean. It quantifies the average squared difference between each data point and the mean of the dataset.

The variance (σ^2) of a dataset with n observations is calculated using the following formula:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Where:

- σ^2 represents the variance.
- x_i represents each individual value in the dataset.
- μ represents the mean of the dataset.
- n is the total number of values in the dataset.

In other words, to calculate the variance:

1. Find the difference between each data point and the mean.
2. Square each of these differences.
3. Sum up all the squared differences.
4. Divide the sum by the total number of values in the dataset.

A larger variance indicates that the data points are more spread out from the mean, while a smaller variance indicates that the data points are closer to the mean. Variance is always non-negative because it involves squaring the differences.

Standard Deviation

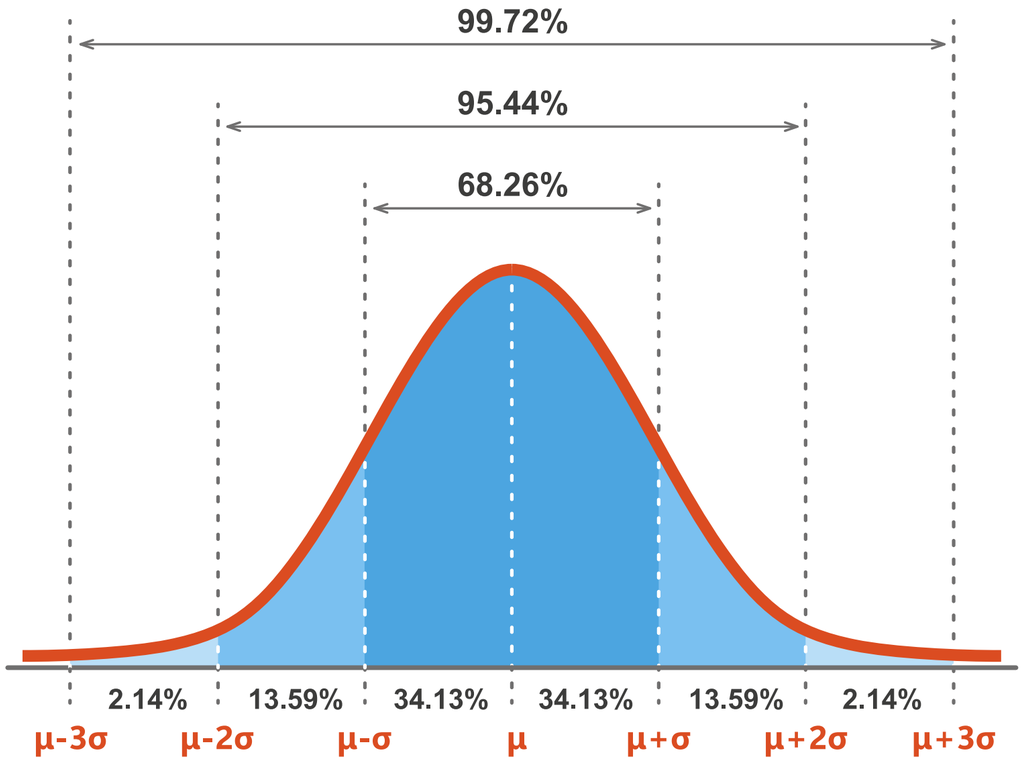
However, the variance is not very intuitive because it is in squared units (e.g., squared meters for area). To address this, the standard deviation is often used, which is the square root of the variance. It is expressed in the same units as the original data, making it more interpretable.

For example, if the variance of a dataset is calculated to be 25, the standard deviation would be 5, since $\sqrt{25} = 5$.

Variance is a fundamental concept in statistics and is used in various fields, including finance, engineering, and natural sciences, to understand the variability or dispersion of data.

Symmetric/Normal Distribution

- Symmetric distributions are balanced or mirror-imaged around the center point (typically the mean).
- Symmetric distributions have similar-shaped and-sized left and right halves.
- The mean, median, and mode of symmetric distributions are usually close.
- The bell-shaped normal distribution, uniform distribution, and triangular distribution are symmetric.

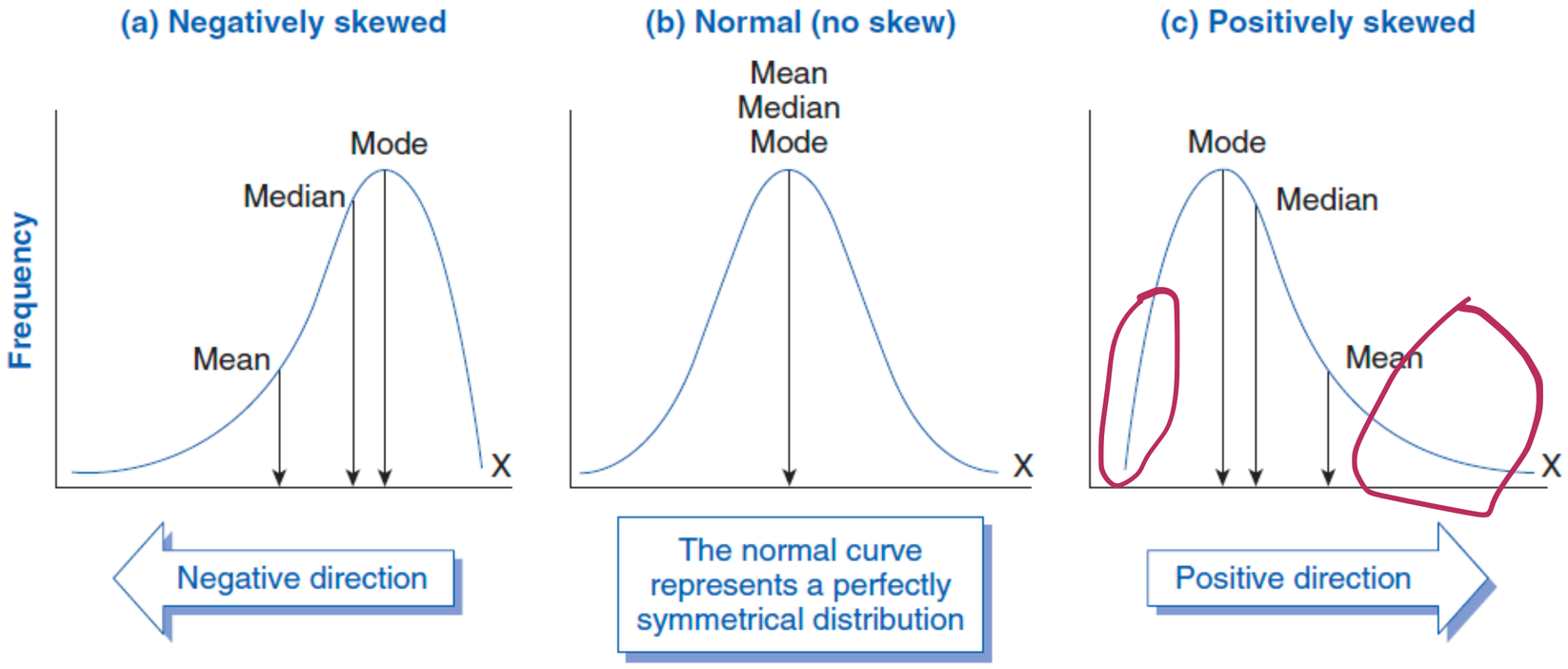


Asymmetric Distribution

- Asymmetric or skewed distributions have an unbalanced form around the center point.
- One tail of an asymmetric distribution may be longer or spread out.

Two forms of asymmetric distributions:

- **Positively skewed:** The distribution's right tail is longer than its left tail. This suggests extremely high values outnumber extremely low values.
- **Negatively skewed:** The distribution's left tail is longer than its right tail. This suggests extremely low values outnumber extremely high values.



Skewness

- Skewness measures the asymmetry of the distribution of data points around the mean.
- A symmetric distribution has a skewness close to zero.
- Positive skewness indicates that the right tail of the distribution is longer or stretched out compared to the left tail, meaning the data is skewed to the right.
- Negative skewness indicates that the left tail of the distribution is longer or stretched out compared to the right tail, meaning the data is skewed to the left.

Mathematically, skewness is often measured using Pearson's moment coefficient of skewness or other similar methods.

Kurtosis

- Kurtosis measures the "peakedness" or "tailedness" of the distribution of data points compared to a normal distribution.
- A normal distribution has a kurtosis of 3 (mesokurtic), which is often subtracted from the calculated kurtosis value, resulting in a measure called excess kurtosis.
- Positive excess kurtosis (greater than 0) indicates a distribution with heavier tails and a sharper peak compared to a normal distribution (leptokurtic).
- Negative excess kurtosis (less than 0) indicates a distribution with lighter tails and a flatter peak compared to a normal distribution (platykurtic).

There are different formulas to compute kurtosis, and excess kurtosis is commonly used in practice.

Percentiles

25

The 25th percentile, also known as the first quartile (Q1), divides the lowest 25% of the data from the highest 75%. It represents the value below which 25% of the observations fall. It is useful for understanding the lower end of the distribution.

50

The 50th percentile, also known as the median, divides the dataset into two equal halves. It represents the middle value of the dataset when arranged in ascending or descending order.

75

The 75th percentile, also known as the third quartile (Q3), divides the lowest 75% of the data from the highest 25%. It represents the value below which 75% of the observations fall. It is useful for understanding the upper end of the distribution.

Maximum

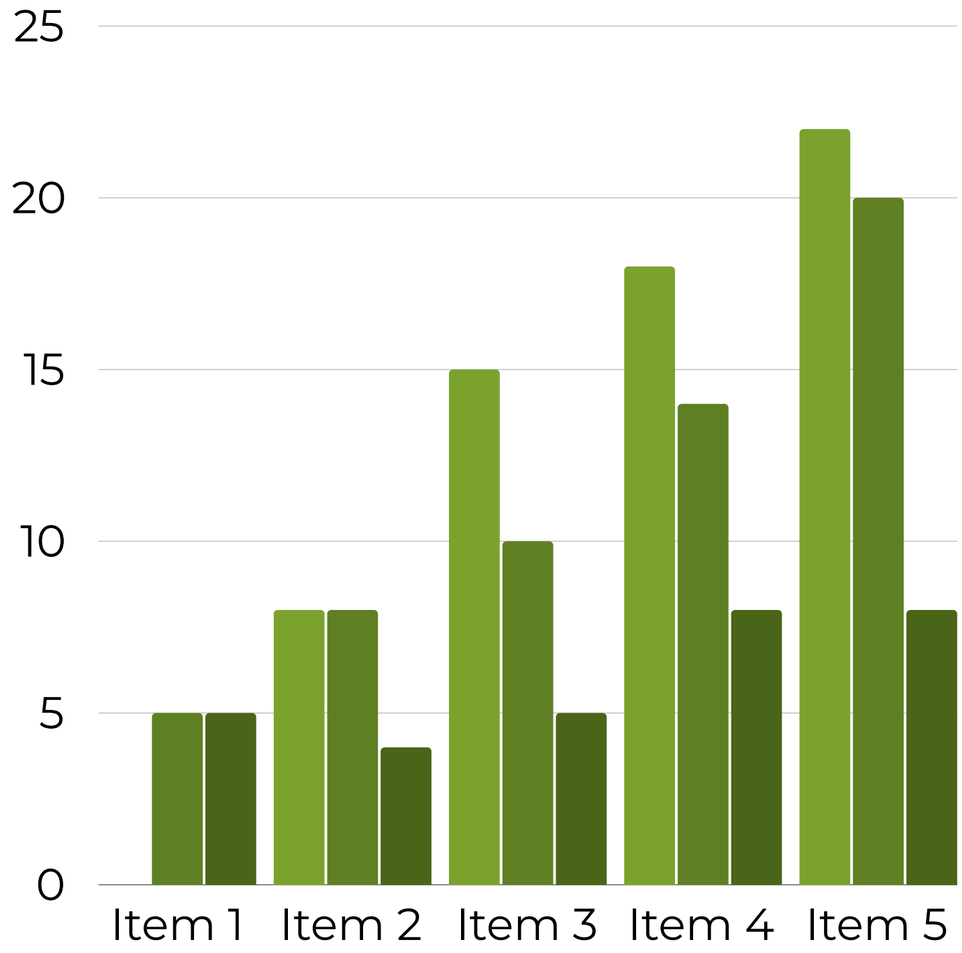
The maximum value is the largest value in a dataset. It represents the highest observation or measurement recorded. It is often used to understand the ceiling or endpoint of the dataset.

Minimum

The minimum value is the smallest value in a dataset. It represents the lowest observation or measurement recorded. It is often used to understand the floor or starting point of the dataset.

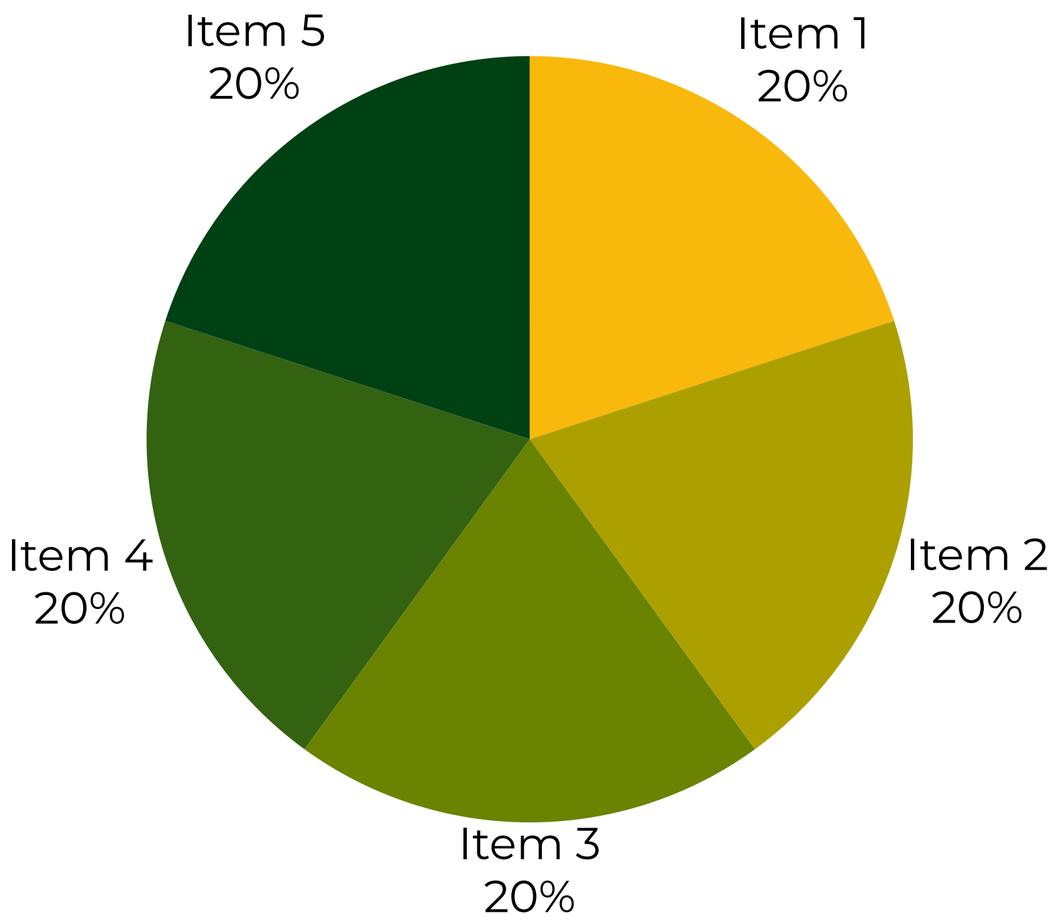
Frequency

- Frequency refers to the number of times a particular value or event occurs in a dataset.
- It provides a count of how many times each value appears.
- Frequencies are often displayed in frequency tables or histograms, where the values are listed along with their corresponding counts.
- For example, in a dataset of exam scores {60, 70, 80, 80, 90, 90, 100}, the frequency of the score 80 is 2 because it appears twice.



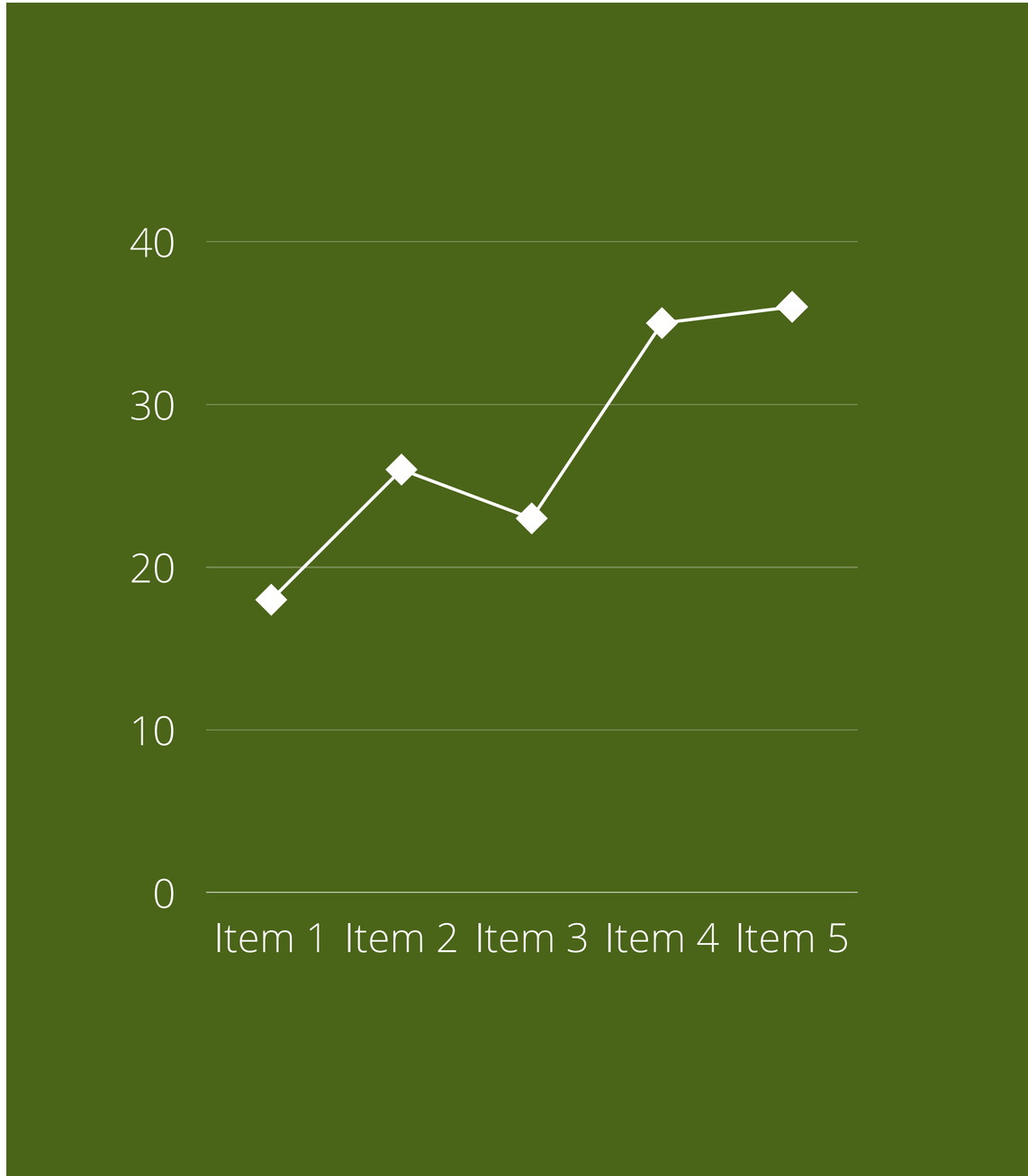
Percentage

- Percentage is a relative measure that expresses a part of the whole as a fraction of 100.
- It is calculated by dividing the frequency of a particular value or event by the total number of observations in the dataset and then multiplying by 100.
- For example, in the same dataset of exam scores {60, 70, 80, 80, 90, 90, 100}, if there are a total of 7 scores, the percentage of the score 80 would be (2/7) * 100 = 28.57%.



Group By

- Data is grouped based on one or more categorical variables. These variables divide the dataset into distinct groups or categories.
- For example, if you have a dataset of student grades with columns for student ID, subject, and grade, you can group the data by subject to analyze the performance of students in each subject.

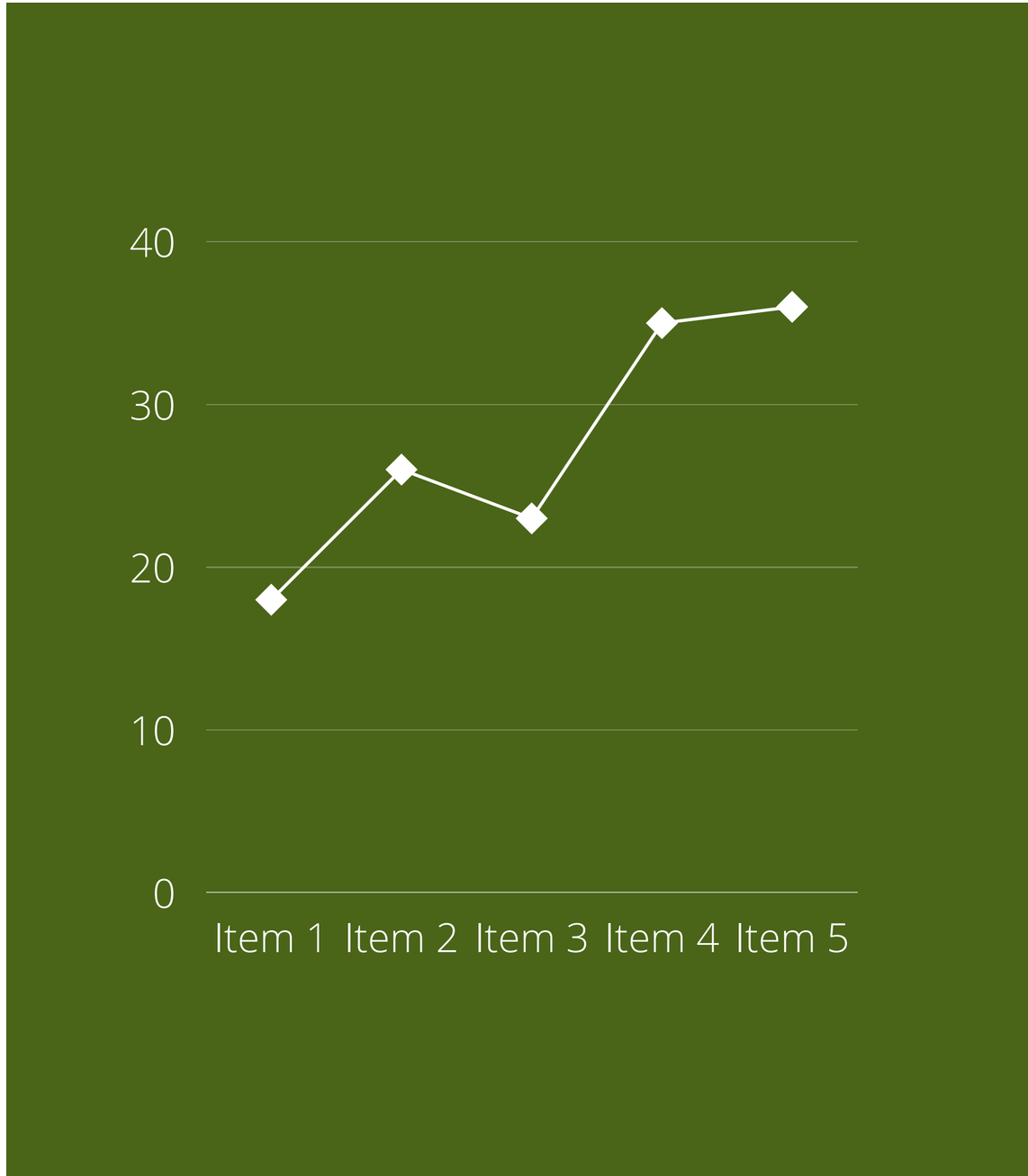


Group By

Subject	Average Grade Point
English	3.96
Mathematics	4.56
History	5.00

Use cases: Group By

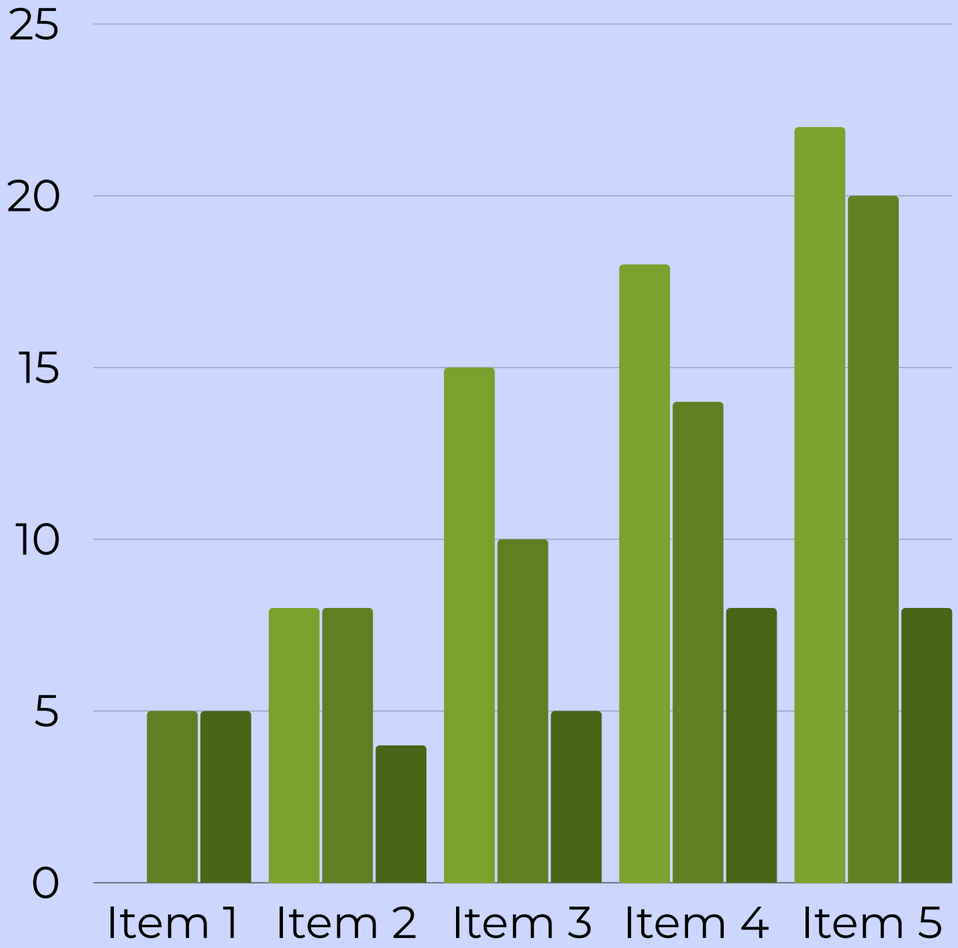
- Understanding patterns and trends in data.
- Comparing the characteristics of different groups.
- Summarizing data for reporting purposes.
- Preparing data for further analysis or modeling.



Cross Tabulation

Cross-tabulation, often known as contingency table analysis, is an approach for analyzing categorical data. Tables show the frequency distribution of the data, making it easy to compare frequencies and find patterns or relationships between variables.

For example, consider a dataset with two categorical variables: "Gender" (male/female) and "Educational Level" (high school, bachelor's, master's).



Cross Table

Education	Male	Female	Total
English	30	50	80
Mathematics	20	30	50
History	10	15	25
Total	60	95	155

Correlation

Correlation analysis is a statistical technique used to measure and quantify the strength and direction of the relationship between two or more variables. It helps determine whether and to what extent changes in one variable are associated with changes in another variable.

Variables:

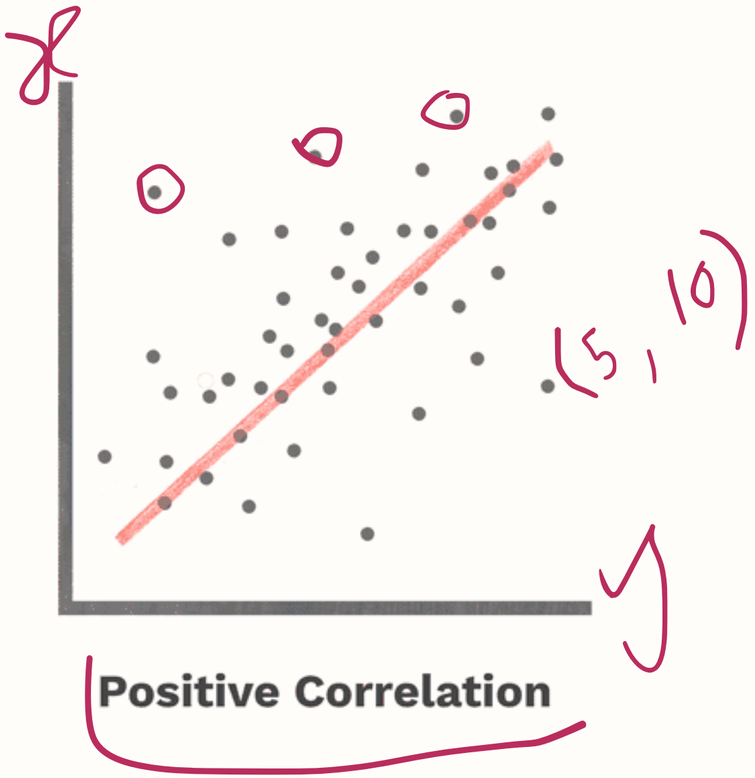
- Correlation analysis requires at least two variables. These variables can be quantitative (numeric) or ordinal (ordered categories).
- For example, in a study examining the relationship between temperature and ice cream sales, temperature would be one variable, and ice cream sales would be the other variable.

Correlation Coefficient:

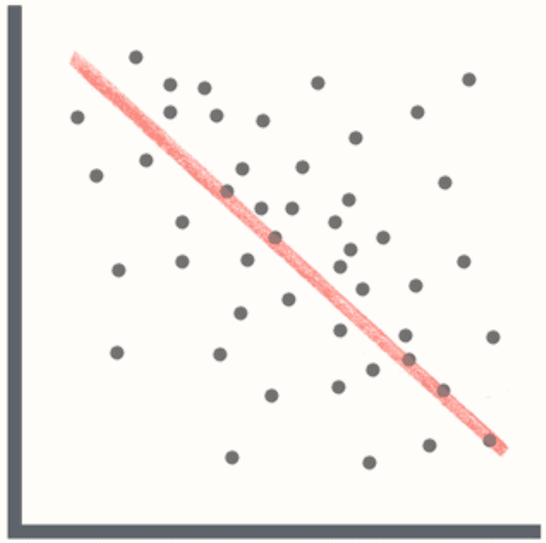
- The correlation coefficient is a numerical measure that summarizes the strength and direction of the linear relationship between two variables.
- The most commonly used correlation coefficient is Pearson's correlation coefficient (r), which ranges from -1 to +1:
    - +1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases in a linear fashion.
    - -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases in a linear fashion.
    - 0 indicates no linear correlation between the variables.
- Other types of correlation coefficients, such as Spearman's rank correlation coefficient and Kendall's tau, can be used for ordinal data or when the relationship is not linear.

- **Interpretation**:
  - The correlation coefficient provides information about the strength and direction of the relationship between the variables:
    - A correlation coefficient close to +1 or -1 indicates a strong linear relationship.
    - A correlation coefficient close to 0 indicates a weak or no linear relationship.
    - The sign of the correlation coefficient (+/-) indicates the direction of the relationship: positive correlation (both variables move in the same direction) or negative correlation (variables move in opposite directions).

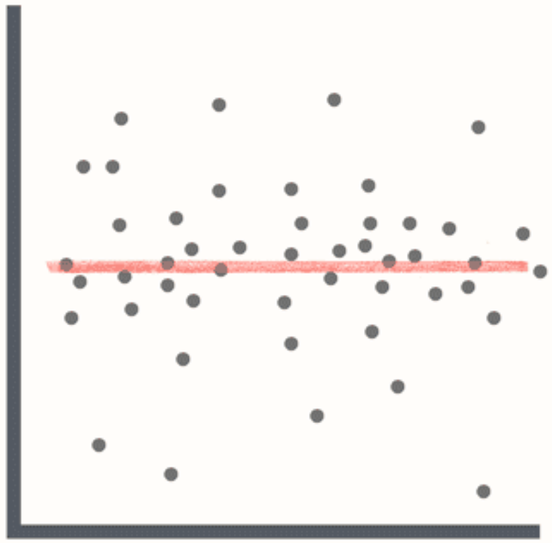
Correlation Coefficient



Positive Correlation



Negative Correlation



No Correlation

THANK YOU