Statistics

Mean : The mean is the sum of all values in a dataset divided by the total number of values.

$$\text{Mean}(\mu) = \frac{\sum x_i}{N}$$

Key Characteristics:

Sensitive to extreme values (outliers).
Works well for datasets with a symmetric distribution.

$$\text{Mean} = \frac{4+8+6+5+3}{5} = \frac{26}{5} = 5.2$$

Median: The median is the middle value of a dataset when it is ordered from smallest to largest.

If the dataset has an odd number of values, the median is the middle value.
If the dataset has an even number of values, the median is the average of the two middle values.

Odd dataset: $[3, 5, 6, 8, 10]$

- Median: 6 (middle value).

Even dataset: $[2, 4, 6, 8]$

- Median: $\frac{4+6}{2} = 5$.

Key Characteristics:

Less sensitive to outliers than the mean.
A better measure of central tendency for skewed data.

Mode: The mode is the value(s) that appear most frequently in a dataset.

Example: Dataset: $[1, 2, 2, 3, 3, 3, 4, 5]$

- Mode: 3 (appears 3 times, more than any other value).

Key Characteristics:

Useful for categorical data (e.g., most common color or category).
Can provide insights about the distribution of data.

Variance: Measures the spread of data points around the mean.

$$\text{Variance}(\sigma^2) = \frac{\sum (x_i - \mu)^2}{N}$$

Key Characteristics:

A high variance indicates that the data points are widely spread,
while a low variance shows that they are closer to the mean.

Standard Deviation: Provides a measure of spread in the same unit as the data, making it more interpretable.

$$\text{Standard Deviation}(\sigma) = \sqrt{\text{Variance}}$$

Correlation: Correlation quantifies the relationship between two variables and helps identify how one variable changes with respect to another.

$r = 1$: Perfect positive correlation (as one variable increases, the other also increases).

$r = -1$: Perfect negative correlation (as one variable increases, the other decreases).

$r = 0$: No correlation (variables are unrelated).

$$r = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum (x_i - \mu_x)^2 \sum (y_i - \mu_y)^2}}$$

- x_i, y_i: Values of variables x and y.
- μ_x, μ_y: Mean of variables x and y.

Data Cleaning

Handling missing values:

1) Removing rows or columns with excessive missing data.
2) Imputing values using the mean, median, mode, or predictive techniques.

Outlier Detection:

1) Removing them if they result from errors.
2) Transforming or capping them if they are valid but extreme.

Correcting Inconsistent Formats:

1) Standardize inconsistent data formats (e.g., date formats or text casing).
2) Ensure uniformity in units of measurement (e.g., all temperatures in Celsius or Fahrenheit).