**Statistics**

Statistics is the branch of mathematics that deals with collecting, analyzing, interpreting, presenting and organizing data. Statistics is like a tool that helps us see patterns, trends and relationships in the world around us. Statistics helps us make decisions based on data.

Types of Statistics:-

1. Descriptive statistics
2. Inferential statistics

**Descriptive statistics** – helps us simplify and organize big chunks of data. This makes large amounts of data easier to understand.

It helps us to understand the data in more depth. Descriptive statistics provides insights for what has happened in the past before attempting to explain why it happened or predicting what will happen in the future.

It commonly uses a measure of central tendencies such as mean, median & mode. And measures of variability or dispersion such as standard deviation or variance, the min and max values of the variables, kurtosis & skewness of the data to give us an idea about the distribution of our data.

Histogram – to understand the characteristics of data

Boxplot, Scatterplot – to understand relationships between the two variables

Two statistical methods for analysis while doing exploratory data analysis:-

1. Univariate Analysis
2. Bivariate Analysis

Univariate Analysis – provides statistics of a single variable using measure of central tendency and measure of variability.

Ex:- df.describe()

A screenshot of a computer

Description automatically generated

#summary statistics for categorical features

For categorical features, df.describe(include=[‘O’])

A screenshot of a computer

Description automatically generated

Summary statistics for all features, df.describe(include=’all’)

For unapplicable values it returns NaN

A screenshot of a computer

Description automatically generated

To get unique count of each value in categorical feature, df[‘Gender’].value\_counts()

Bivariate Analysis – examines the relationship between two or more variables. Visualization is often used together with qualitative analysis as a more intuitive way of presenting the result. It helps us determine whether variables are correlated or not. It provides insights into the problem and helps to develop ideas or hypotheses for potential quantitative research.

Ex:- df.groupby()

**Inferential statistics** – It uses smaller data to conclude a larger group. It helps us predict and draw conclusions about a population.

Ex:- testing a hypothesis and concluding a population from sampled data. However, the sample might not provide a perfect estimation of the population all the time. To measure this uncertainty, we can create a confidence interval.

A confidence interval is a range of values that is likely to contain a population parameter with a certain level of confidence. Ex:- we might produce a 95% confidence interval of [61.5, 64.5] which says that we are 95% confident that the average weight of students from a certain school is between 61.5kg and 64.5kg.

Basics of Statistics:-

1. Population mean
2. Sample mean
3. Sample/population std dev
4. Sample/population variance
5. Class Interval
6. Frequency
7. Range

Types of Data:-

1. Qualitative data – data is descriptive
2. Quantitative data – numerical information

Types of Quantitative data:-

1. Discrete data – it has a particular fixed value & can be counted
2. Continuous data – it is not fixed but has a range of data and can be measured

Measure of Central Tendency:-

1. Mean

A number equation with numbers and symbols

Description automatically generated with medium confidence

1. Median

A white background with black text

Description automatically generated

1. Mode

Measure of Dispersion:-

1. Range
2. Variance
3. Standard Deviation
4. Interquartile Range(IQR)
5. Quartiles
6. Mean Absolute deviation
7. Coefficient of Va