**Statistics**

**Statistics: Way to get Information from Data**

* Collecting the Data
* Analyzing the Data
* Interpreting the Data
* Presenting the Data (Stack & Unstack)
* Organizing the Data(Segregating Data)

**Why learn Statistics:**

* Make it more structured
* Draw Insights
* Describe the Data
* East to Understand
* Summarizing Data
* Numbers can misleading but chart never misleading

**2 Types of Statistics:**

* Descriptive Statistics
  + Presenting /Organizing/Summarizing Data
* Inferential Statistics
  + Drawing conclusions on data observed in sample

**Basic Statistical Analysis Include:**

* Measure of Central Values
* Measure the spread around central values
* Overall distribution shape

Descriptive analysis of the input and output attribute values helps understand the current state and behavior of the process.

**Variables**:

* Qualitative (Categorical)

Nominal (Emp ID, Gender, Religion) Does not have order

Ordinal (Ratings, Rankings) – Does have order

* Quantitative(Numerical)

Discrete (Countable – 1, 2, 3...)

Continuous (Measurable - 1.1, 1.2, 1.3…)

Age can comes under 25 years(Categorical), 25(Numerical)

**Characteristics of Frequency Distribution:**

**Modality**

* Unimodal
* Bimodal

**Symmetry**

* Symmetric(+ve)
* Asymmetric(-ve)

Based on tail can predict skewness whether positive or negative

**Central Tendency:**

* Mean – Average of Data point
* Median – Central value of Data Points
* Mode – Most frequency occurring values

Time to prepare the pizza





ABC pizzeria is much more consistent than XYZ pizza in terms of time to prepare

a. If you want to be sure about the amount of time you will need to get your lunch you will go to ABC Pizza

b. Going to XYZ Pizza will make one less sure about the time required to gather the lunch

c. However, if you are a risk taker, you may go to XYZ Pizza even when under time pressure as there is a chance of getting the lunch within 4.2 minutes. Saving a full 2+ minutes compared to ABC!

**Points to note:**

* In Symmetric – Mean/Median/Mode can be same
* Always Outliners better to go with Median
* If table with missing values, impute with Mean, Median and Mode
* Mode not for Continuous values, it’s for segregated values
* Mode not for separate values
* Imputation of mode, only for categorical values

**Variability**: (Spread of the Data)

* Range (Spread of Data Range)
* Measure of Dispersion:
  + Standard Distribution (Deviation from Mean)
  + Variance

**Percentile**:

A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

**Boxplot**:

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median.

Contains below 5 ranges to determine the Quartile portion.

* Minimum
* Q1(25%)
* Q2(25%) - Median Value
* Q3(25%)
* Maximum

**Quartile**:

A quartile is a type of quantile which divides the number of data points into four more or less equal parts, or quarters. The first quartile (Q1) is defined as the middle number between the smallest number and the median of the data set.

Divided into 1/4 parts:

Inter – Quartile Range = Q3-Q1