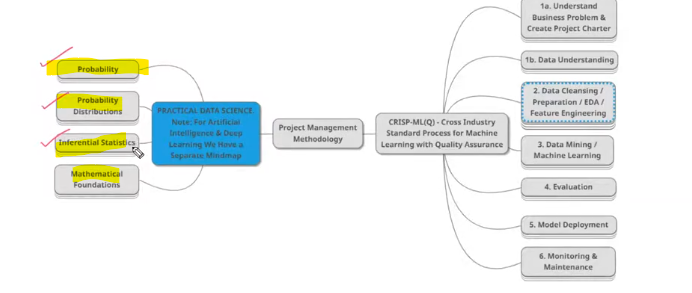
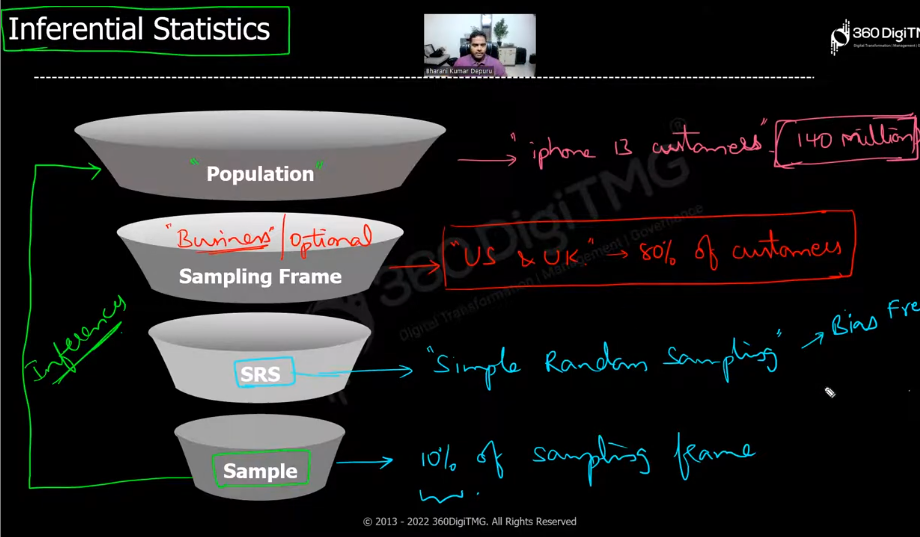
****

**Inferential statistics**

Predict a output from sample subset of large data are called inferential statistics. In the below example we take 10% of data from 80% of population to know, is the 10% of peoples like the I phone features. Actual data is 140 millions, to analyse all this data will takes too much time, so we have some percentage of sample data from it.

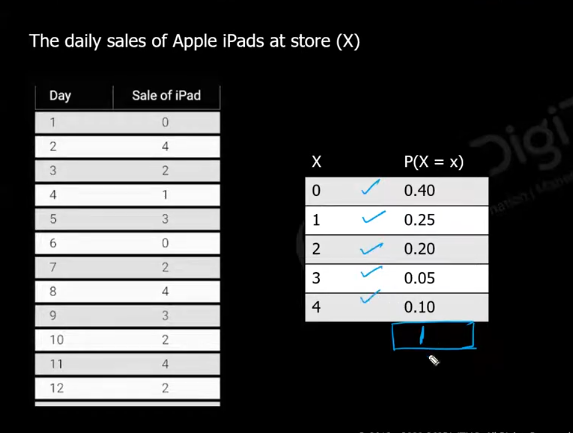
Simple random sampling will be used to take sample data which has no baises.



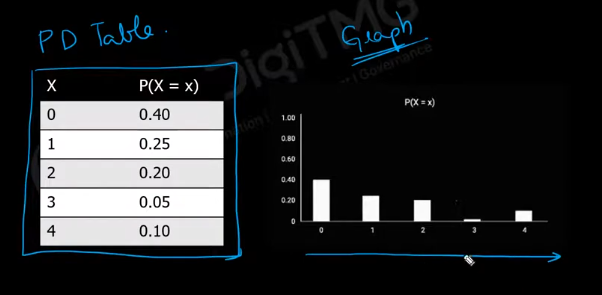
**PROBABILITY DISTRIBUTION**

In the process of probability distribution we have to do normalize to find the maximum range of probability values and that max range will be distribute to all the probabilities to avoid biases, overfitting.

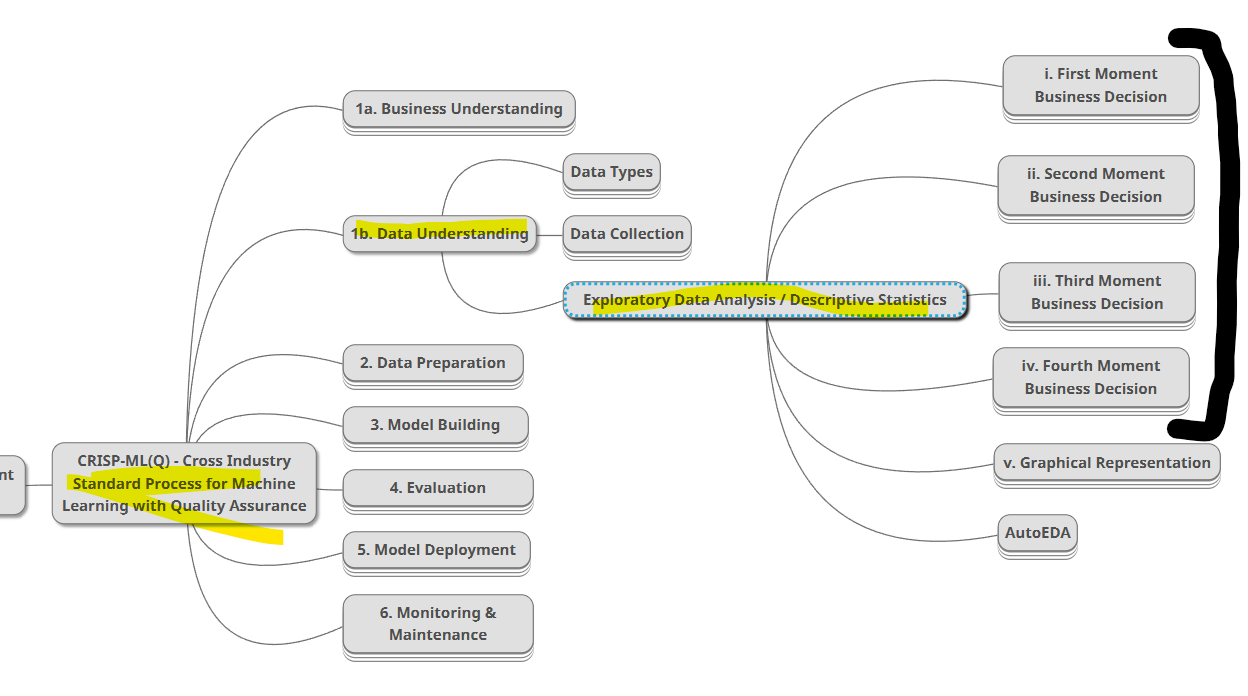
In the below example we find the probability of I phone sale, X- is the no of iphone – p(x) is probability for that no of iphone sale. But summing all those probability we found 1 is the max range that will be distributed to all the probabilities.



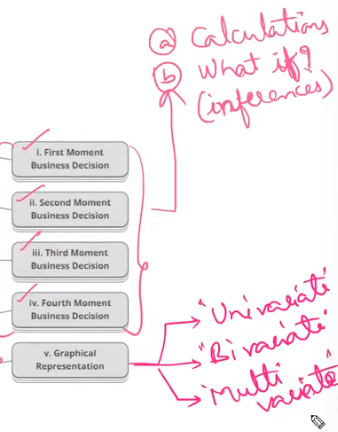
We can represent prob distribution in below tow ways



All this calculations are done for the Exploratory data analysis phase like the below one.



In first four phase of EDA we do some calculation that is nothing to inference statictics, we do calculations with sample and come with a result, and that result will be represented graphically in graphical representation, We have three types of graph representation

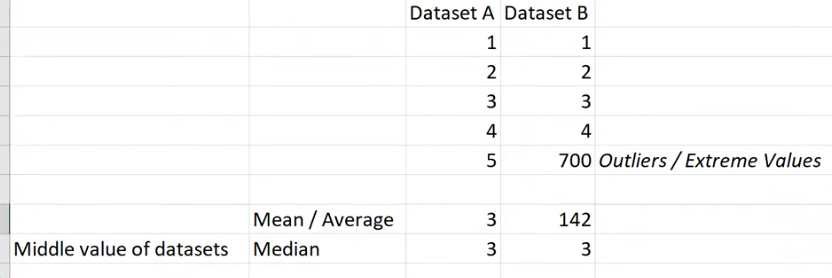


**FIRST MOMENT BUSINESS DECISION**

Average is the first question any one could ask on the starting step of data analysis.

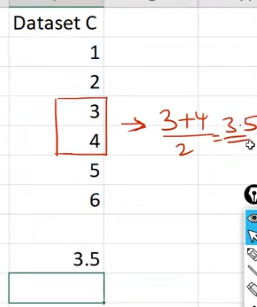
****

Means is influenced by outlier(very big values that doesn’t aligns with all the values), in the below example mean gives the mean of as 3 for dataset A but for dataset B we changed only one single value so the means become 142.

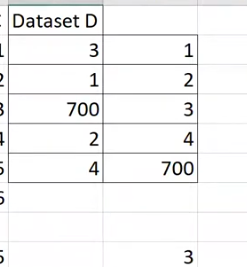


But 142 is not the correct mean here, we have to calculate correct mean value for this outliers data set, this is where **Median** comes to play, it doesn’t affected by outliers

In median it calculates the mean for two middle values like the below one.

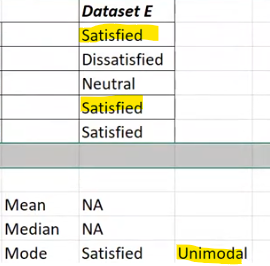


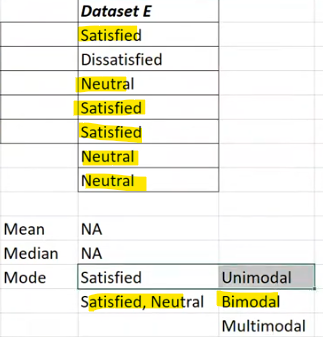
The below kind of data sorting should be applied before going to perform median



Both of this median and mean are only works on numeric data, for categorical data we have mode.

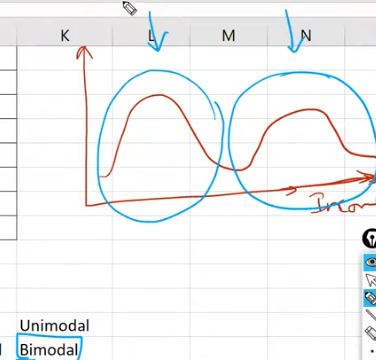
Mode gives the no of occurrences of a data. Three type modes are there





More than two common occurrences are considered as multimode

**Graph rep of bimode**

****