* N- 1 is the degree of freedom – which is how many variable which can be independent to the result
* Central Limit Theorem :

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed

* Alpha – Significance level (Risk factor)
* (1 – Alpha) - Confidence level (how much confident you are 95% or 99%, default is 95%)
* For a Random Variable, If SD of population is known and sample is greater than 30 then use Z test
* For a Random Variable, If SD of population is unknown or sample is less than 30 then use T test
* One sample T-test – to check whether sample belongs to a population or not
* Two sample T-test – to check whether both the sample belongs to a population or not
* Paired T- test – to check the mean when two samples are dependent
* ANOVA test - When there is more than 2 samples we analyze the variance of the samples
* F- Stat – Variance between groups / Variance Within groups
* For two tailed T- test, p value is 0.025
* For one sided T- test, p values in 0.05
* If P value > than confident interval then failed to reject the H0
* If P value < than confident interval then reject the H0
* Covariance – How two variables vary together
* +ive Covariance – Both variables increases/decreases together
* -ive Covariance – Both variables increases/decreases in the opposite way (inverse relationship)

**Linear Regression:**

* Linear regression can be used only on Continuous variable and will not work on Categorical variable
* R Square increases when the error decreases
* R Square maximum value will be 1 when the error is 0
* R square minimum value will be 0
* Adjusted R square increases only when the variable is significant in the model. We need to look at the Adjusted R square
* If there is no relationship between the dependent and independent variable then the coefficient is 0
* If the p-value for a variable is less than your significance level, your sample data provide enough evidence to reject the null hypothesis for the entire population. Your data favor the hypothesis that there is a non-zero correlation. Changes in the independent variable are associated with changes in the response at the population level. This variable is statistically significant and probably a worthwhile addition to your regression model
* If the p-value that is greater than the significance level indicates that, there is insufficient evidence in your sample to conclude that a non-zero correlation exists. Variable should be omitted.
* For linear regression, error should be independent and normally distributed [qqnorm()]
* If the data is not linear make it linear by transformation like log, square root
* There should not be correlation between independent variable. So, R square value should be low if we repeat linear regression by having that independent variable as dependent
* Correlation gives the strength of the relationship and it will vary between -1 and +1
* While prediction, y axis will be the dependent variable
* Variance inflation factor (1 to 2 – Very good,2 to 5 – Good, 5 to 10 –Moderate, >10 – Correlation is there)
* There should not be any heteroscedasticity for a linear regression between error and dependent variable[if the variability of the random disturbance is different across elements of the vector]. It should be constant and not linear

**Classification:**

* Classification will be used for Categorical variable
* Linear and nonlinear classification
* Linear – Logistic regression, Support vector machine
* Non Linear – Decision trees, Random forest, K nearest neighbor, Naïve bayes
* **Linear - Logistic regression**
  + - Maximum Likelihood estimate
    - Null deviance – Total errors when we don’t take independent variables
    - Residual deviance – Total errors when we take independent variables
    - AIkake Information Criterion to compare two models – Lower the better
    - Convert continuous variable into categorical variable by using threshold logic
    - True positive rate – How many times model is correctly able to identify +ive
    - Accuracy or Sensitivity – True +ive
    - Specificity – True –ive
    - Truth table or confusion matrix
* **Linear - Support Vector Machine**
  + - Create a plain as the decision boundary
    - Find the margin(gap) between the points of different classes
    - One of the fastest classification technique
    - It is a black box and not known what exactly happens between variables
    - Kernel – type of separation made (linear,radial,polynomial)
    - Dependent variable should be factor
* Non-Linear – Naïve bayes
  + - Assume that the variables are independent of each other
    - Dependent variable should be factor
* Non-Linear – Decision Tree
  + - How a Node is selected? Entropy (when there is randomness in data)

**Correlation**:

Relationship between two variable

1. - No relationship

0 to 1 – Positive linear relationship

0 to -1 – Negative linear relationship

**Generalized linear model (GLM)**

Generalized linear model (GLM) is a generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution like Gaussian distribution.

**Cluster**:

Kmeans clustering – Need to know the no of clustering before hand