# 

# 

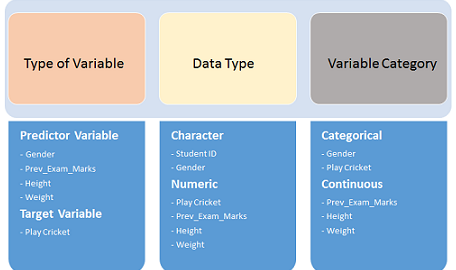
# EDA Data Analysis

**Data Analysis:**

Identify predictor (Input) and Target variable(Output)

Divide variables into type of variable,datatype,variable category

<https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_2.png>



EDA i.e. Exploratory Data Analysis is the process of studying data by leveraging various statistical and visualization techniques

1. Univariate Analysis
2. Bivariate Analysis
3. Multivirate Analysis

**Univariate Analysis** : Impact of one variable in dataset….Study of single variable at one point of time

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous. Let’s look at these methods and statistical measures for categorical and continuous variables individually

Ex: marks of students

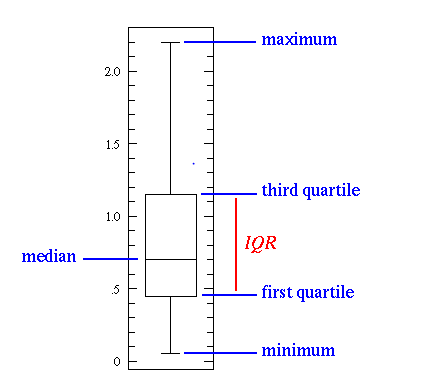
Marks is a variable for student dataset

* There are two types of variables :

1. Categorical ----Strings
2. Numerical (Continous)------Numbers

**Plots on Categorical Variables : Bar chart & Pie Chart--** In case of Categorical variable,we need to measure count of category

**Plots on Numerical Variables: Box Plot & Histogram ---** In case of continuous variables, we need to understand the central tendency and spread of the variable

****

**Histogram**:

Variance= represents data dispersion

Standard deviation : sqrt(Variance)

Co-efficient deviation : Standard Deviation/mean

Skewness : represents symmetry of graph ,if it is Inclined right means +ve symmetry

Kurtosis: Whether data is dispersed peaked or flatted

Example:

Random =[2,4,6,8,10]

Variance = (sum (random)^2/N-1

Mean = sum/num

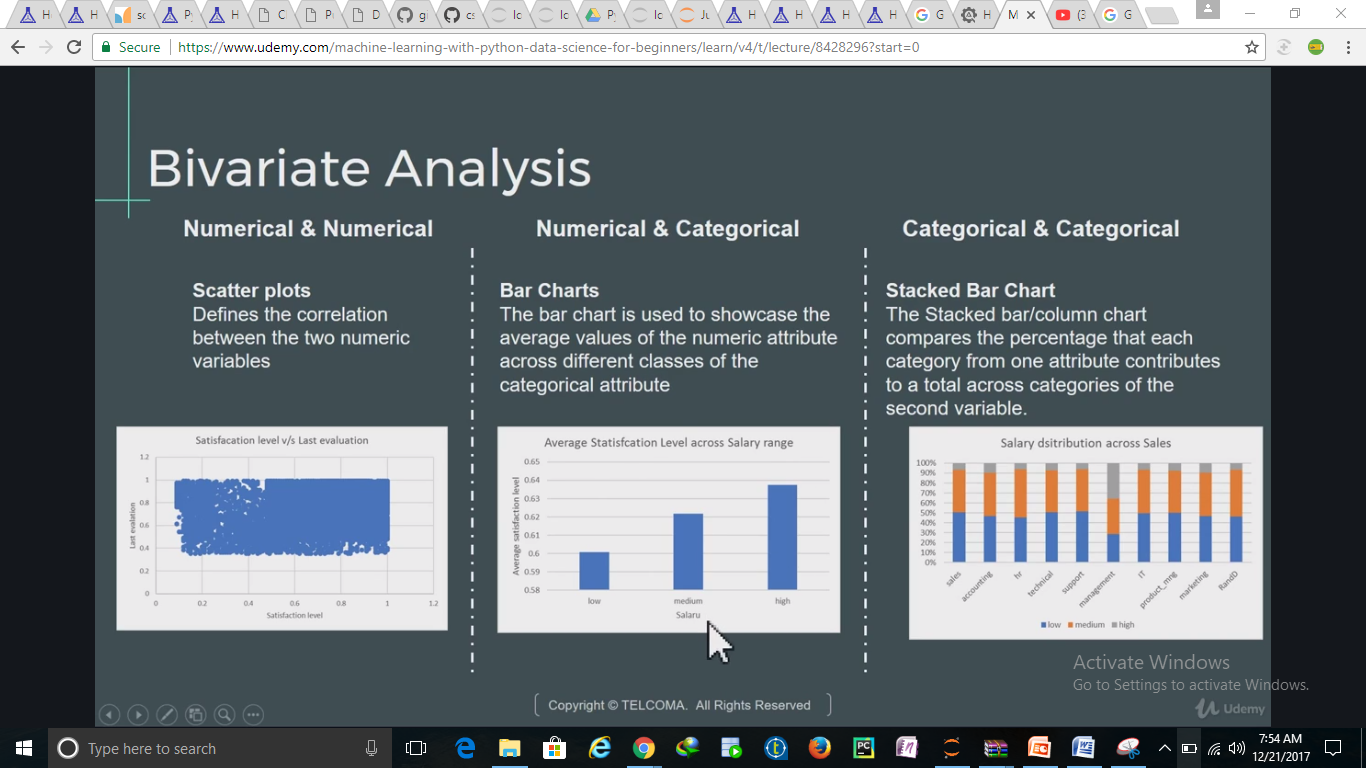
S.D =sqrt(Variance)

Coefficient = S.D /mean

**Bivariate Analysis**:

Analyzing two variables at one point of time.

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables.



## Missing Values:

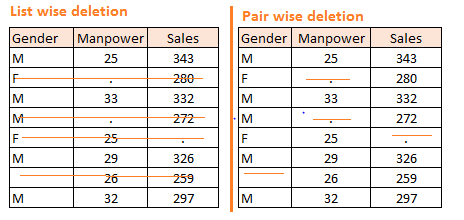
**Methods to treat Missing Values**:

1)Deletion :

There are two types of deletion

1)List deletion ---this method reduces the power of model because it reduces the Sample Size

2)Pairwise deletion--- One of the disadvantage of this method, it uses different sample size f for different variables



**Impute Missing Values:(**Replace Missing Values)

There are many options we could consider when replacing a missing value, for example:

* A constant value that has meaning within the domain, such as 0, distinct from all other values.
* A value from another randomly selected record.
* A mean, median or mode value for the column.----df.fillna(df.mean(),inplace=True)
* A value estimated by another predictive model

Sklearn has imputer preprossing class to fill missing values

# Outlier:

Outlier is an observation that appear faraway and diverge from regular pattern in a Sample.

# Feature Engineering:

Adding More value to data

**Linear Regression Analysis:**

1. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X¹ is constant, regardless of the value of X¹. An additive relationship suggests that the effect of X¹ on Y is independent of other variables.
2. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
3. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
4. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.
5. The error terms must be normally distributed.

**Variance:**

