Data Science & Machine Learning

* **Data Science**: It is a process of extracting knowledge & insights from data using scientific methods.

Scientific Methods:
(1) Machine Learning
(2) Deep Learning
(3) NLP
(4) Statistics
(5) Data visualization.

Data can be in:
(i) Structured format
(ii) Unstructured format
(iii) Semistructured format.

* **Machine Learning**: Type of artificial intelligence which allows software applications to become more accurate to predict the outcomes using historical data.
e.g: House Price, Stock Price, Sentiment Analysis, weather prediction, Speech recognition, self driving cars, chatbots.

* **Types of Machine Learning:**
(1) Supervised ML - Data is labelled
(2) Unsupervised ML - Data is unlabelled (Cricket Data)
(3) Reinforcement ML - Reward based learning

Supervised ML

(A) Classification (Contains Discrete values)
- Binary classification
  - Yes / No
  - True / False
  - Approved / Declined
  - Good / Bad
- Multiclass Classification
  - High / Medium / Low

(B) Regression (Target column contains continuous values)
e.g
- House Price
- Cibil Score
- Share Price
- Temp. control Systems

* **Data Science Project Steps:**
(1) Problem Statement -
(2) Data Gathering (JSON, csv, excel, images, videos, pdf, database etc) -
(3) Exploratory Data Analysis (EDA) - (we use pandas, numpy, matplotlib, seaborn).
Imp. Part
- If more than 50% NaN values, drop that column
- If same value in one column, drop that column
- If different data in one column, drop it.
- Data should have some pattern, then it is helpful for analysis.
- If we have three different types of data, then we have to make separate column for each value.
- Convert text or string to numeric data using encoders (~~mean, median~~)
- If we have NaN value, fill it using (mean or median).
- If we want to find ?, use valuecount fn & then replace.

(4) Feature ~~Selection~~ Engineering - Scaling, Binning, Handling missing values.

(5) Feature Selection - (Feature means column)
- Here, we have to select required columns to train model.
- Here, we have to separate data

(6) Model Building / Model Training -
- This is easy part
- We use linear regression, logistic regression, KNN, Decision Tree, Random Forest classification

(7) Model Evaluation -
- Find MSE, classification report, confusion matrix
- without evaluating a model, we can not deploy it.

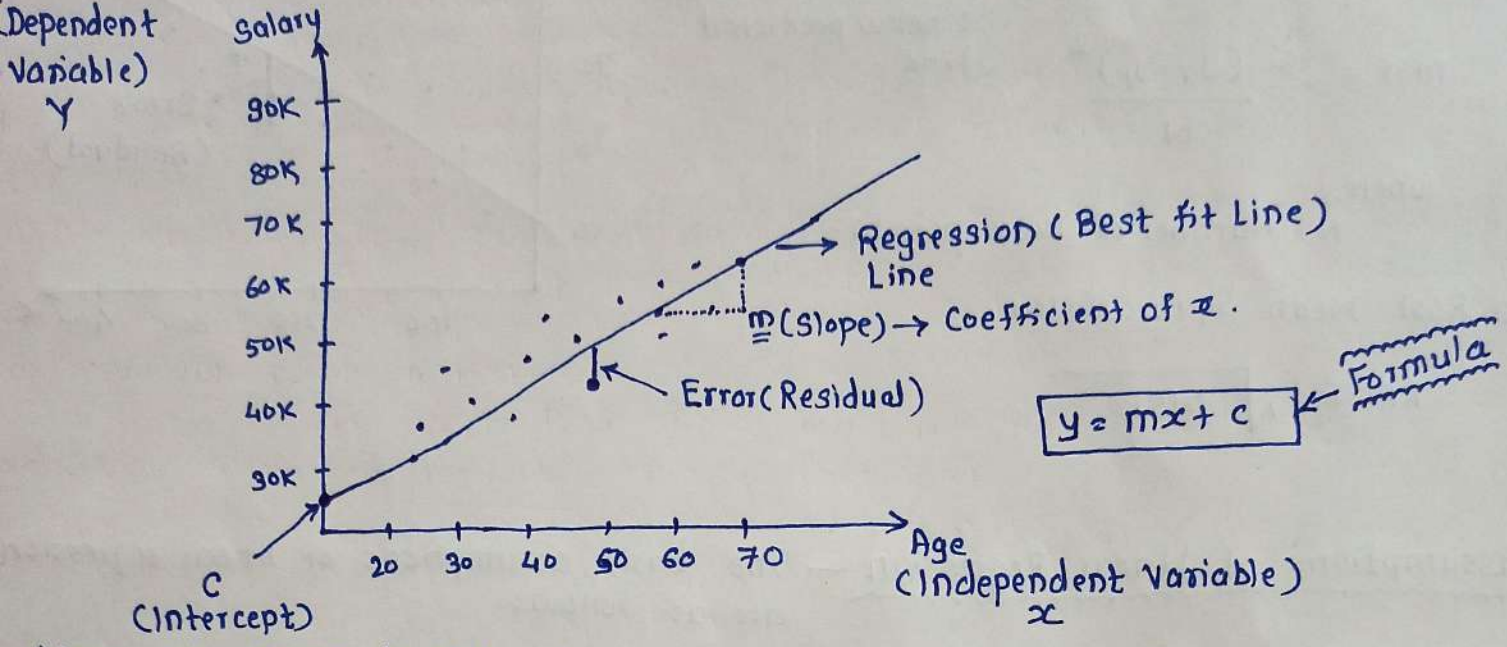
(8) Project Deployment -
- Project can be deployed using AWS, GCP, Azure, Herokku)
- We can deploy 5 free projects on Herokku.
- All cloud platforms are used to deploy project.

◉ Important points in EDA -

(a) If we are not satisfied in step 7, again go to step no. 03 - EDA.
(b) For KNN algorithm, ~~scle~~ scaling is important.
(c) We can use normalization / standardization / log for scaling.
(d) In decision tree & Random forest algorithm, no need to use scaling.
(e) Target column should be clean.
(f) If values in column are continues, use mean/median for replacing missing values.
(g) If we have different data values, to replace it we use mode (from scify import stats). (Mode is used for categorization)

Linear Regression — (Regression = Continuos).

Defn: It is a predictive model used to find linear relationship between dependent variable and one or more independent variables.
- It is supervised machine learning model,
- It finds best fit line.



Linear regression is of two types : (a) Simple Linear Regression
(b) Multiple Linear Regression

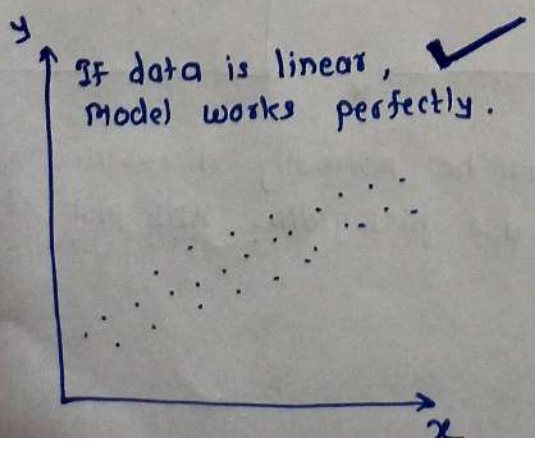
(a) **Simple Linear Regression** → Here only one independent variable is present & model has to find linear relation between dependent & independent variables.

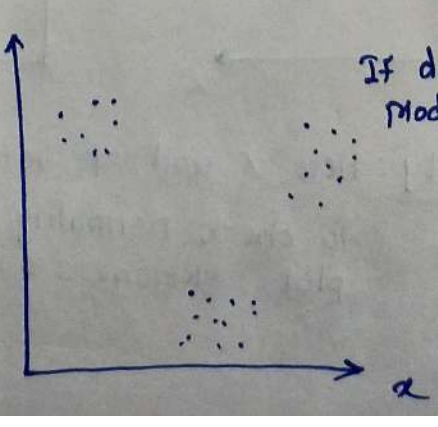
formula : $y = mx + c$ Task : Find y

where,
m = Coefficient
To find m, formula is $m = \frac{y_2 - y_1}{x_2 - x_1}$
c = Intercept

(b) **Multiple Linear Regression** → There are more than one independent variables for model to find relationship.

$$y = m_1x_1 + m_2x_2 + m_3x_3 + \ldots\ldots + m_nx_n + C$$





Mathematical Approach :

(i) Residual / Error = $y_a - y_p$

(ii) Sum of Residuals / Errors = $\sum (y_a - y_p)^2$

↳ Using square because points may be above & below predicted line.

(iii) Mean Square Error

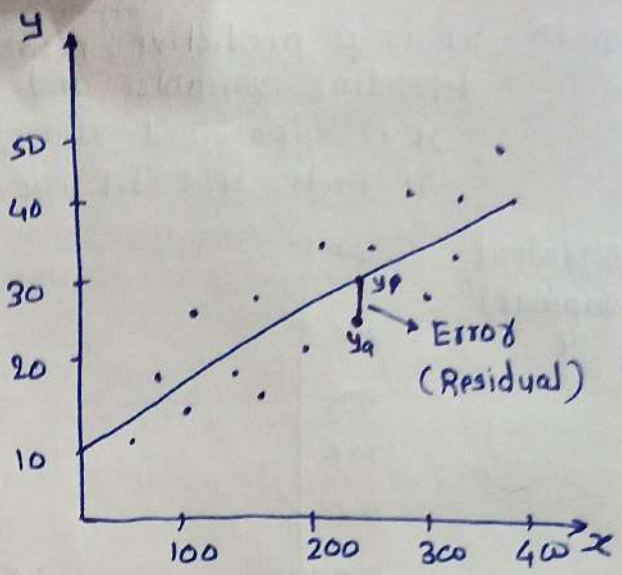
$$MSE = \sum_{0}^{n} \frac{(y_a - y_p)^2}{N}$$

where,

N = Number of Data points.

(iv) Root Mean Square Error

$$RMSE = \sqrt{MSE}$$



* **Assumptions of Linear Regression** — The basic assumptions of linear regression are as follows.

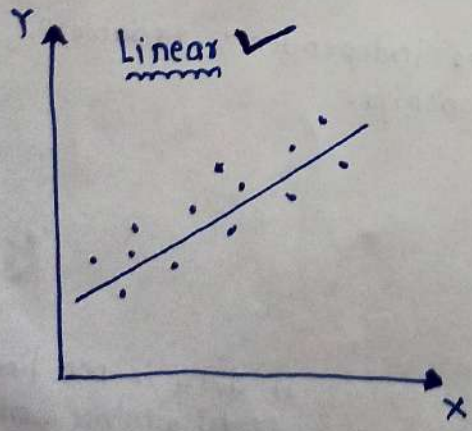
(a) Linearity - (Use Scatter Plot)

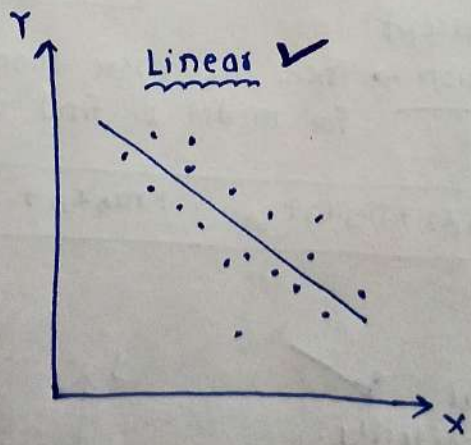
(b) Normality - (Use Histogram, KDE plot, Q-Q plot, Skewness & Kurtosis)

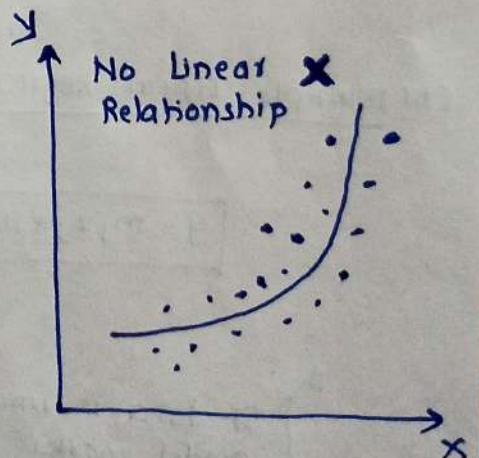
(c) Homoscedasticity - (Use residual plot)

(d) No Multicollinearity / Independence. - (Use correlation matrix or VIF score)

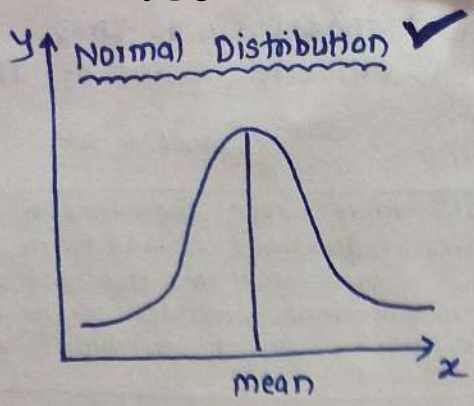
(a) **Linearity** : There should be (must) linear relationship between dependent variable & independent variable. This assumption can be checked by plotting a scatter plot between both variables. (X and Y)

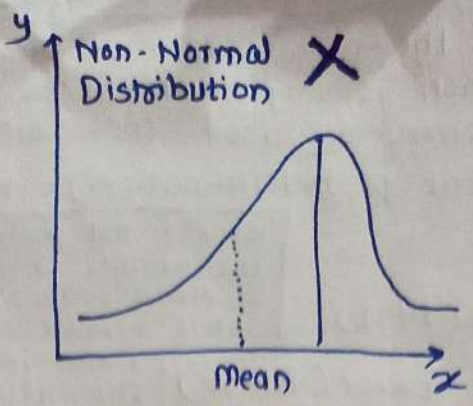


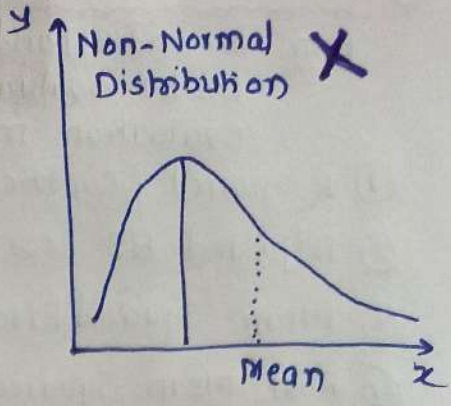




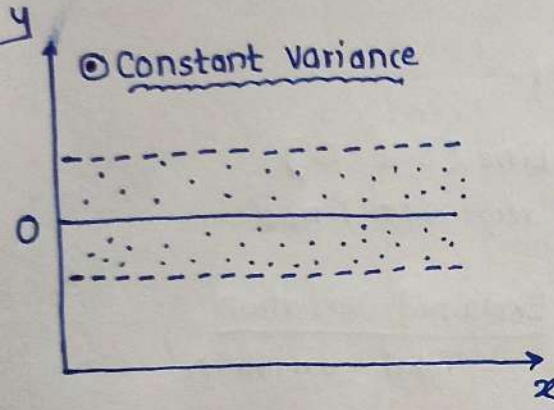
(b) **Normality** : Here X and Y variables must be normally distributed. To check normality, we can use histograms, KDE plots, Q-Q plots, skewness & kurtosis

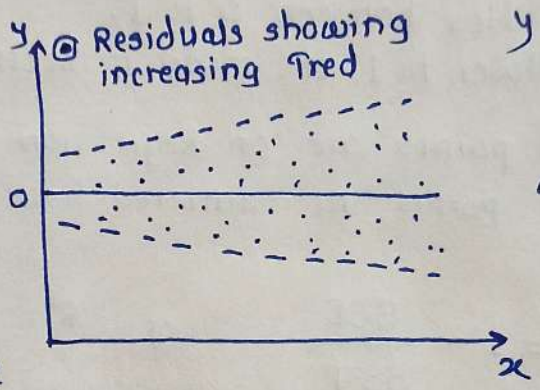


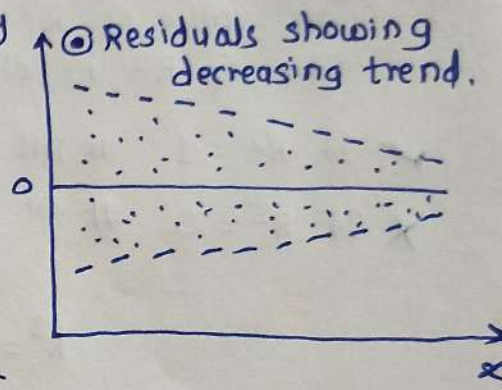




(c) **Homoscedasticity**: The variance of the error should be constant i.e. the spread of residuals should be constant for all values of x. This assumption can be checked by plotting residual plot. If the assumption is violeted, then the points will form a funnel shape otherwise they will be constant.







(d) **No Multicollinearity: (Independence)**

The variables should be independent of each other i.e. there should not be correlation between the independent variables. To check the assumption, we can use correlation matrix or VIF score (Variance Inflation Factor). If the VIF score is greater than 5 then the variables are highly correlated.

How to deal with the violation of any of the Assumption →

* **Impact of Violation** → It may lead to the decrease in the accuracy of the model. Therefore, predictions will not be accurate & errors will be high.

(a) **Violation of Normality Assumption** →

Solution → To treat this problem, we can transform the variables to the normal distribution using various transformation functions such as log transformation, reciprocal or Box-Cox Transformation.

(b) **Violation of multicollinearity Assumption** →

Solution → (i) Remove some of the highly correlated independent variables.
(ii) Deriving a new feature by adding them or performing some mathematical operations.

⊙ Evaluation Metrics for Regression Analysis →

Aim: To understand the performance of Regression model. To do this model evaluation is very much necessary. To evaluate, some of the evaluation metrics are used. These are —

1. R squared (Coefficient of Determination) – R^2
2. Adjusted R2 score
3. Mean Squared Error (MSE)
4. Root Mean Squared Error (RMSE)

Drawback of R^2

The R^2 is not perfect. Its value never decreases no matter the number of variables (redundent) we add to model. It either remain same or increases. This does not make sense because some independent variables might not be useful in determining target variable. To avoid this, adjusted R^2 works.

① R Squared or Coefficient of Determination (R^2) →

- This is the most commonly used metric for model evaluation in regression analysis. It is defined as Ratio of the variation to the total variation.
- The value of R^2 lies between 0 to 1.
- The of R^2 is closer to 1, the model is better.

✔ If $R^2 = 1$ # All points are on Regression Line (SSE = 0)

✗ If $R^2 = 0$ # All points are scattered from regression line.

$$R^2 = 1 - \frac{SSE}{SST} \quad \text{or} \quad R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

Formula → $$R^2 = \frac{SST - SSE}{SST}$$

where,

(a) $SST = \sum (Y_a - \bar{Y})^2$ ← Total Error

(b) $SSE = \sum (Y_a - Y_p)^2$ ↑ Sum of Squared Errors.

Keynote: (i) $R^2 = 1$ ← Best Score

(ii) $R^2 = 0$ ← Worst Score

(iii) $R^2 = -ve$ ← We will not use this (Improve value? → Not Possible) If R^2 is is negative, we have to move or check adjusted R^2.

② Adjusted R Squared (Adjusted R^2) → (Best evaluation Metrics) ✔

- It is improvement to R squared.
- The drawback with R^2 is as number of features increase, the value of R^2 also increase, which gives illusion of good model.
- So Adjusted R^2 solves drawback of R^2.
- It only considers that features which are important for the model & in results, it shows real improvement of the model.

Adjusted R^2 is alway lower than R^2 ✔

- It is used to find goodness of the Best Fit Line.
- It will increase only for good predictors.

Formula :→ $$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

where,
R^2 = Sample R^2 score
p = Number of predictors
N = Total Sample size

③ Mean Squared Error (MSE) →
- It is mean of the squared difference of actual vs. predicted values.

Formula :→ $$MSE = \sum_{0}^{n} \frac{(y_a - y_p)^2}{N}$$

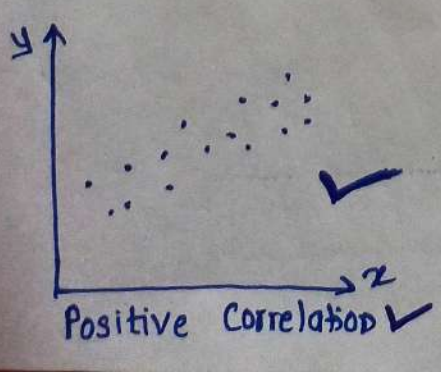
④ Root Mean Squared Error (RMSE) →
- It is the root of MSE.
- It penalizes the large errors whereas MSE doesn't.

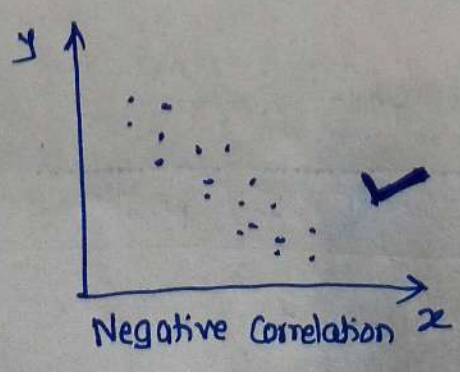
● Coefficient of Correlation → (R)
- These are used to meagure how strong relationship is between two variables.
- The most popular correlation coefficient is Pearson's Correlation Coefficient.

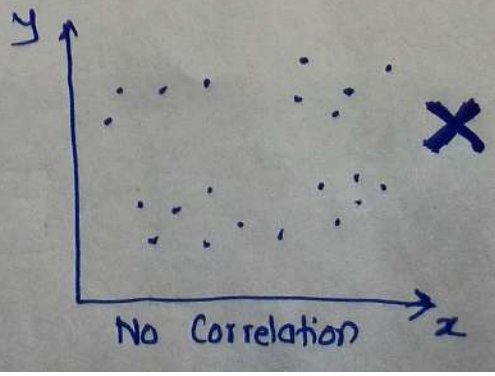
Formula: $$\text{Coefficient of Correlation (R)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where,
R = Pearson's Correlation Coefficient
x_i = x variable samples
y_i = y variable samples
$\bar{x}$ = mean of values in x variables
$\bar{y}$ = mean of values in y variables

* Value of R →
- Range = '−1' to '+1'
- value 0 (zero) specifies that there is no relation between two variables.

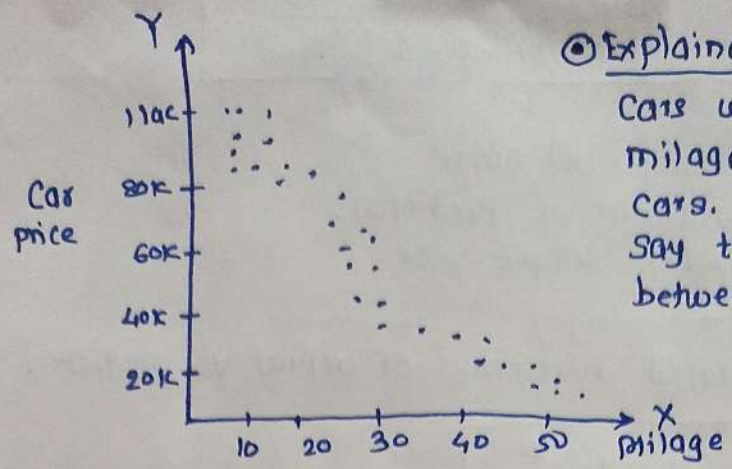






sathe.manoj@gmail.com

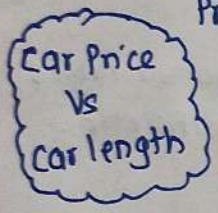
Example of Negative Correlation →

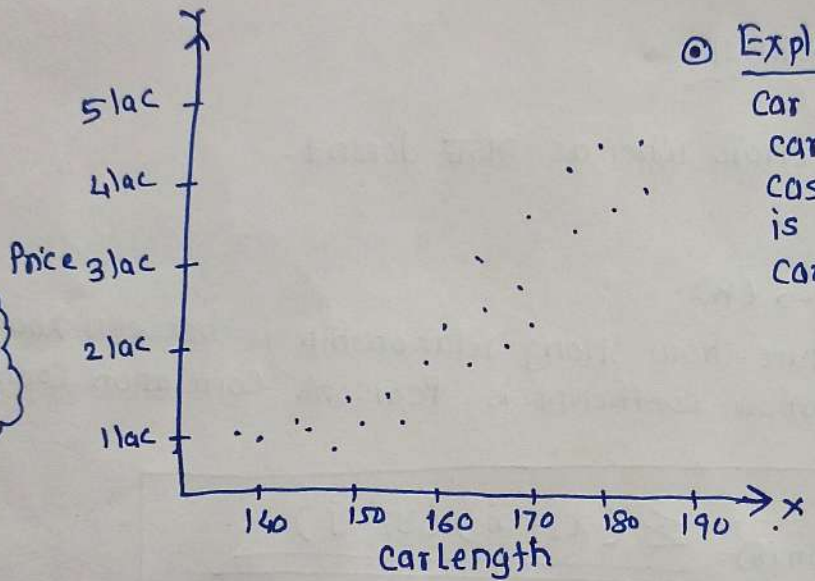


⊙ Explaination →

Cars with high prices have very low milage as compared to low range of cars. Hence, in this case we can say that there is negative correlation between car price & milage.

Example of Positive Correlation →





⊙ Explaination →

Car with high price have high car length. Hence, in this case we can say that there is positive correlation between car price & car length.

To find correlation, we use df.corr()

Gradient Descent Algorithm →

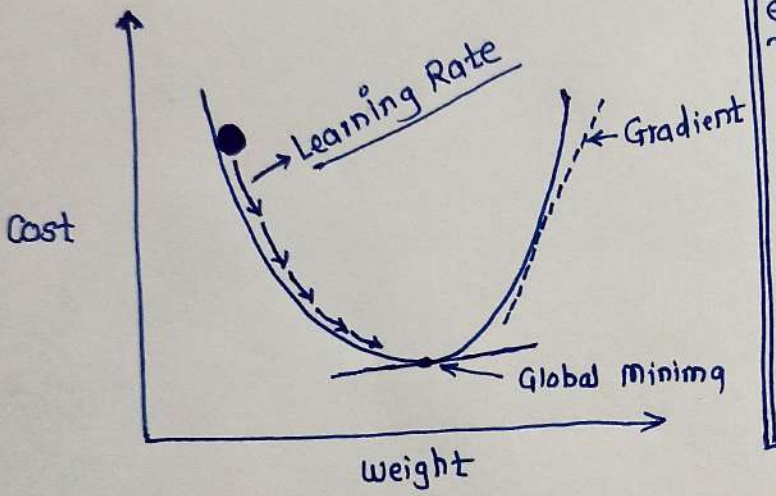
Many popular machine algorithms depend on optimization techniques such as linear regression, k-nearest etc. In this part, we will go through optimization technique called Gradient Descent.

- It is used to reduce cost function. (MSE)
- It is used to find best m & c values or Best Fit Line (BFL).
- Gradient Descent will try infinite m & c values.
- It use partial derivate.

$$MSE = \sum_{0}^{n} \frac{(y_a - y_p)^2}{N}$$

["A gradient measures how much the output of a function changes if you change the inputs a little bit" ✓]

* Explaination with Graph →



* Explaination with Practical Example →

e.g: Imagine a valley and a person with no sense of ~~direction~~ direction who wants to get to the bottom of the valley. He goes down the slope & takes large steps when the slope is steep & small steps when the slope is less steep. He decides his next position based on his current position & stops when he gets to the bottom of the valley which was his goal.

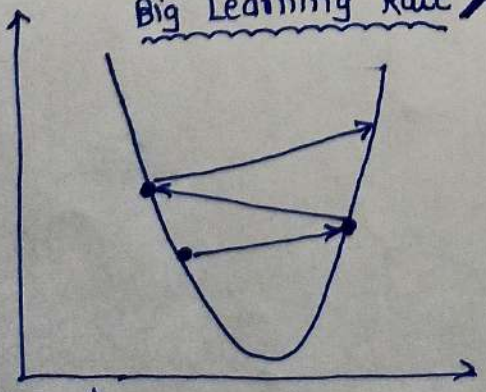
* What is Gradient →

Defn: In machine learning, a gradient is a derivative of a function that has more than one input variable.

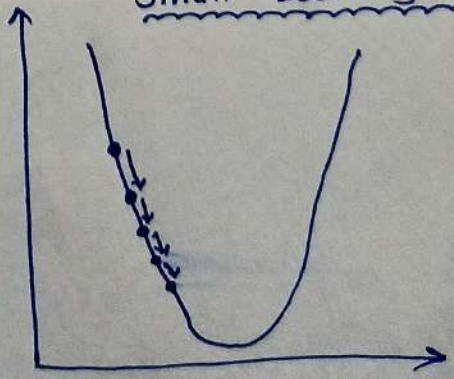
* Importance of Learning Rate →

For gradient descent to reach the local minimum we must set the learning rate to an appropriate value which is neither too low nor too high. This is important because if the steps it takes are too big, it may not reach the global minima.

Big Learning Rate ✗



Small Learning Rate ✓



So, the learning rate should never be too high or too low for this reason. You can check if you are learning rate is doing well by plotting on graph.