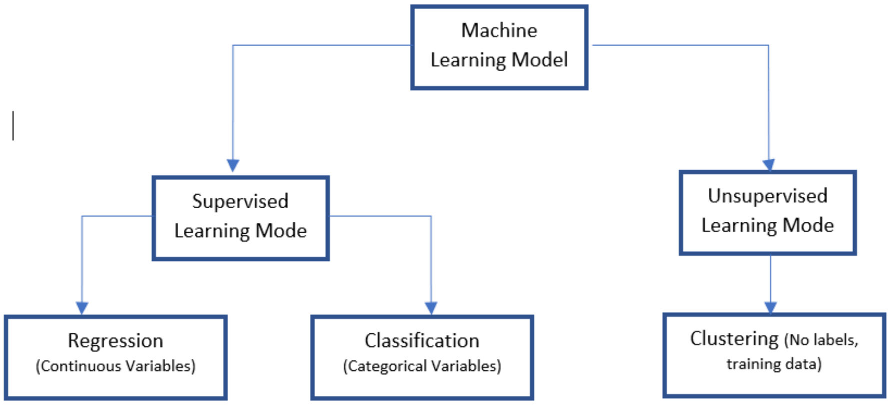
1. Explain the linear regression algorithm in detail.

One of the Machine Learning models is based on concept of Regression which basically means predicting continuous variable. Linear regression falls under supervised models.


Machine Learning Model
Supervised Learning Mode
Unsupervised Learning Mode
Regression (Continuous Variables)
Classification (Categorical Variables)
Clustering (No labels, training data)

Linear regression algorithm establishes linear relationship between dependent and independent variables, if any. Linear regression algorithm helps in predicting the values of dependent variable based on change in value of independent variable.

Equation for predicting dependent variable is

$$\hat{Y} \ = \ \beta_0 + \beta_1 X_1 \ldots\ldots + \beta_n X_n$$

Here β_0 is the intercept, β_1 is slope and X_i is independent variable.

A best fit line is drawn and β_0 and β_1 are determined for the best fit line.

Method used for determining the Best fit line is Ordinary Least Squared Method which essentially calculates the Residual Sum of Squares. Residual is nothing but error i.e True value – Predicted value

If we minimize RSS, we will be able to determine Best fit line and hence β_0 and β_1 as well.RSS is an absolute quantity and is dependent on unit of variable and hence there is a need to define Total Sum of Square which is relative.

Goodness of fit is indicated by R2 which is calculated based on RSS and TSS

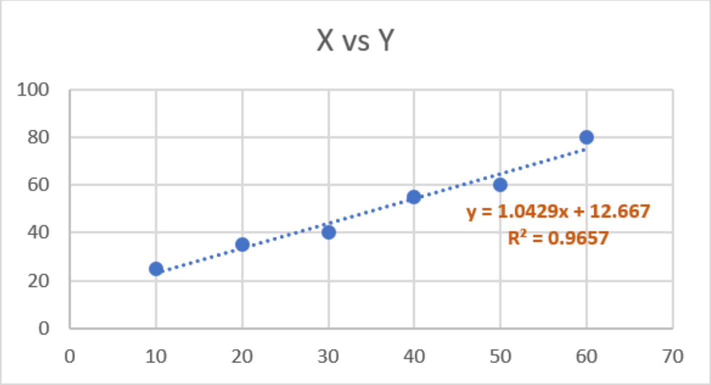
$$R^2 = 1 \ - \ \frac{RSS}{TSS}$$

For the below sample data best-fit line is drawn in excel which demonstrate slope and intercept.

Data:

X	Y
10	25
20	35
30	40
40	55
50	60
60	80

Representation using best-fit line:



R^2 is goodness of fit and higher the value better it explains the variation in data points.

2. What are the assumptions of linear regression regarding residuals?

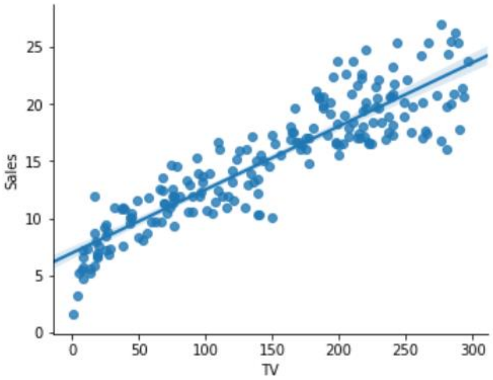
There are 4 key assumptions in Linear regression

1) X and Y have linear relationship
2) Error terms are normally distributed (have same standard deviation with mean around 0)
3) Homoscedasticity - Error terms have constant variance
4) Error terms are independent of each other.

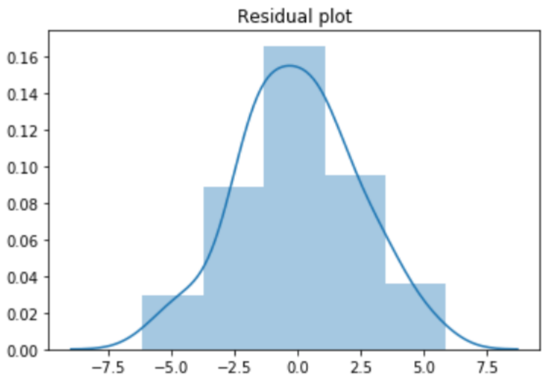
Note – Only assumption# 2,3 and 4 are regarding error terms (residual) as asked in questions. We are discussing everything here.

1) Linear relationship between X and Y – We cannot build Linear Regression model if we do not have linear relationship with between X and Y. One good indicator is to plot scatter plot or reg plot between variables.

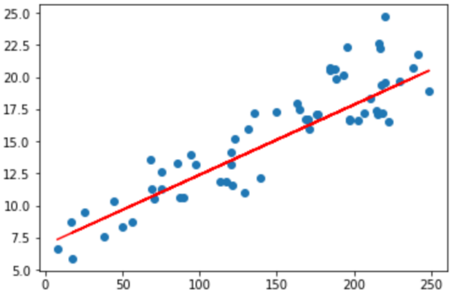
   If this assumption does not hold true then predictions would not be correct when we test model.



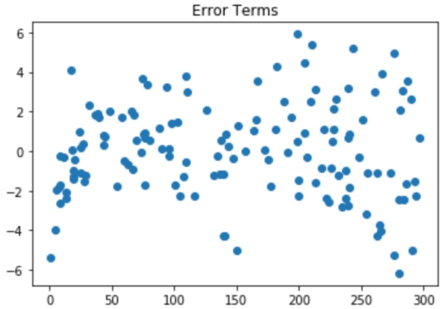
2) Error terms are normally distributed (have same standard deviation with mean around 0) – This is key for making inferences. If error terms are not normally distributed, then we will not be able to say with confidence that model is reliable. We can plot residuals to check the distribution.



3) Homoscedasticity - Error terms have constant variance. As discussed in point#2, the error terms are normally distributed. Homoscedasticity means that standard variance is constant across independent variables. Ordinary least squares (OLS) regression assumes that all residuals have a constant variance (**homoscedasticity**).



4) Error terms are independent of each other – Plotting error terms can indicate any visible pattern indicating that error terms are not independent.



3. What is the coefficient of correlation and the coefficient of determination?

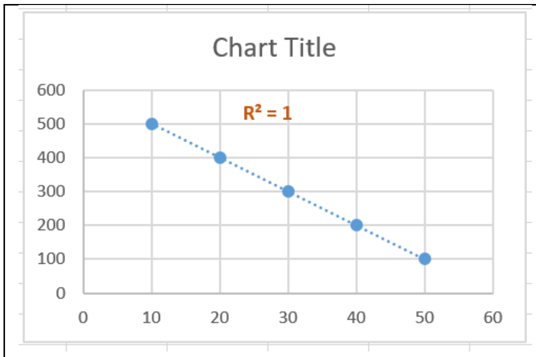
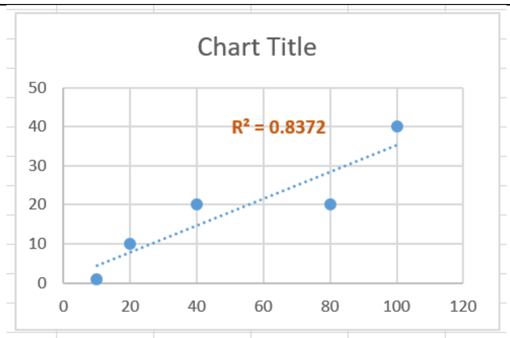
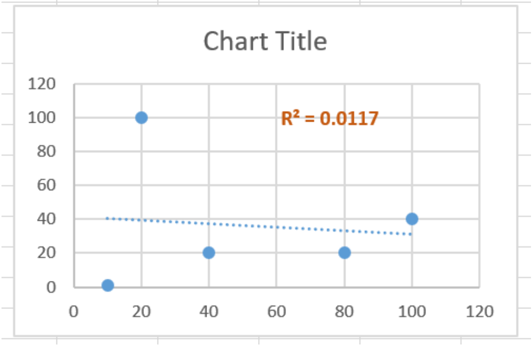
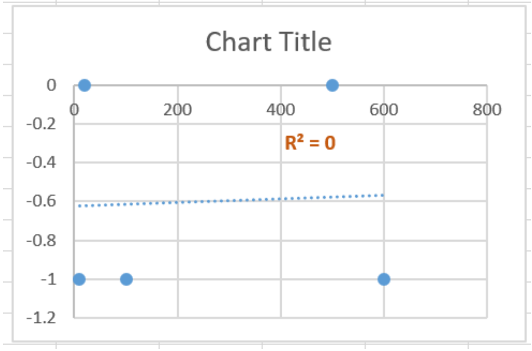
Coefficient of correlation is statistical measure to indicate relationship (positive or negative) between 2 quantitative variables. Coefficient of correlation ranges from -1 to +1. -1 indicates strong negative correlation whereas +1 indicates strong positive correlation.

4 common types of correlations are:

- Pearson's = Also called as Person's R is most common correlation which is used to measure linearity between variables. This is used in Linear models.
- Kendall = Measures ordinal relationship between 2 sets using nonparametric test
- Spearman = This is mostly used when relationship between variables is not linear.

- Cohen = This determines the effect size. .10 to .29 is small, .30 to .49 is medium association, and .50 and above is large association. Effect size is nothing but magnitude of phenomenon.

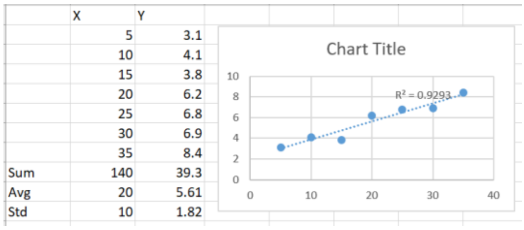
Coefficient of determination is nothing but R2 (R-squared) that we discussed above.
Value of R2 ranges from 0 to +1 with +1 indicating all variance in data is explained by model. Higher value of R2 is preferred since this indicates that greater proportion of variable is explained by model. This is also referred to as strength of linear regression.

 R2 = 1 means that all variance is explained by model	 R2 = 0.8372 indicates good population is explained by model
 R2 = 0.0117 indicates very less variance explained by model.	 R2 = 0 means no variance can be explained

4. Explain the Anscombe’s quartet in detail.

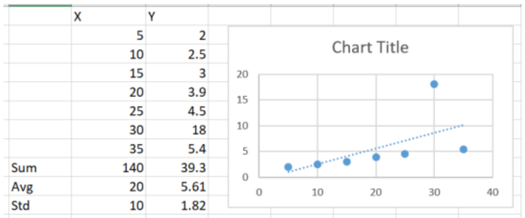
Anscombe’s quartet is aspect which basically indicates that descriptive statistics does not accurately and/or completely depict the distribution of data. It encourages to use graphs along with descriptive statistics to get true picture of data distribution. 4 sets of data with same descriptive statistics when plotted show completely different distribution.

If we just analyze descriptive statistics of below 4 sets, they do not correctly indicate the different distribution of data. They all have same total, average and standard deviation. However, when plotted shows different data distribution.



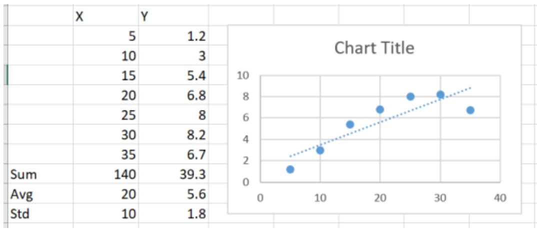
	X	Y
	5	3.1
	10	4.1
	15	3.8
	20	6.2
	25	6.8
	30	6.9
	35	8.4
Sum	140	39.3
Avg	20	5.61
Std	10	1.82

This is a true representation of linear relationship



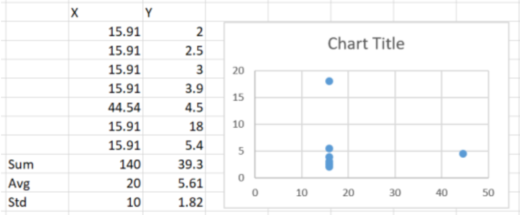
	X	Y
	5	2
	10	2.5
	15	3
	20	3.9
	25	4.5
	30	18
	35	5.4
Sum	140	39.3
Avg	20	5.61
Std	10	1.82

Outlier is evident from the plot. Apart from outlier it has tight linear relationship.



	X	Y
	5	1.2
	10	3
	15	5.4
	20	6.8
	25	8
	30	8.2
	35	6.7
Sum	140	39.3
Avg	20	5.6
Std	10	1.8

Data is not distributed normally, and linear model cannot be developed for this



	X	Y
	15.91	2
	15.91	2.5
	15.91	3
	15.91	3.9
	44.54	4.5
	15.91	18
	15.91	5.4
Sum	140	39.3
Avg	20	5.61
Std	10	1.82

X=c, X is constant except one outlier.

5. What is Pearson's R?

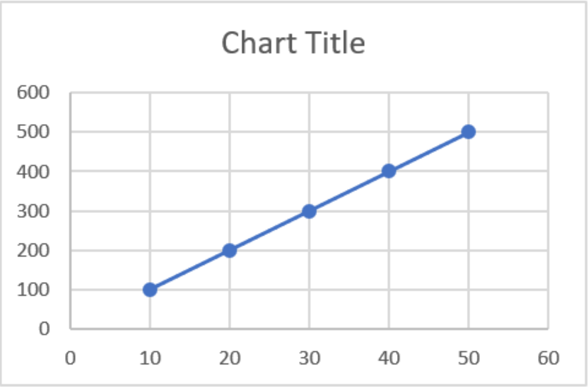
Pearson R's is nothing but Person's correlation coefficient. This measures linear correlation between two variables. The value of coefficient ranges from -1 to +1.

Strong Positive Correlation (+1):

Data:

X	Y
10	100
20	200
30	300
40	400
50	500
Correlation	1

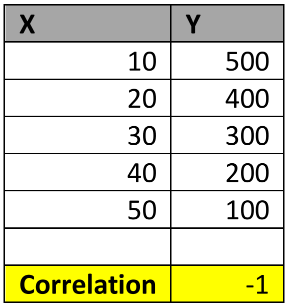
Representation of line:



Positive correlation indicates increase in 1 variable with increase in value of another variable. In this case Y increases with increase in value of X.

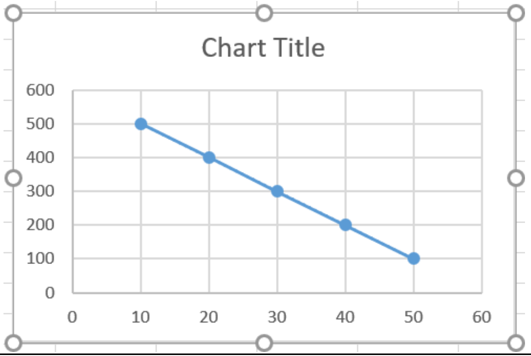
Strong Negative Correlation (+1):

Data:



X	Y
10	500
20	400
30	300
40	200
50	100
Correlation	-1

Representation of line:

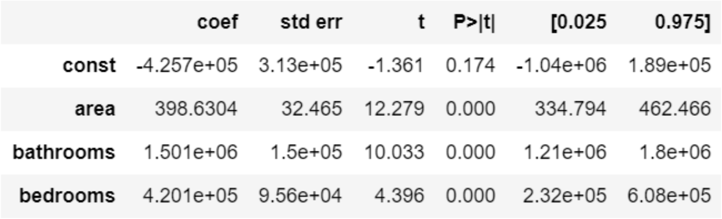


Negative correlation indicates decrease in 1 variable with increase in value of another variable. In this case Y decreases with increase in value of X.

6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

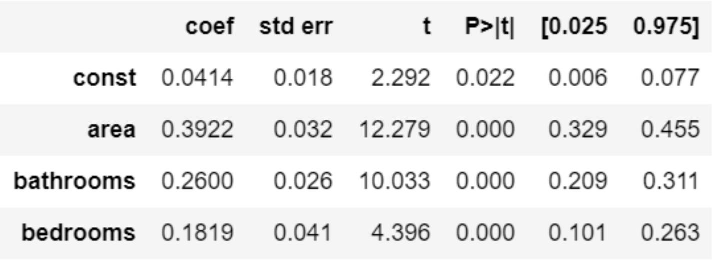
Scaling - Scaling bring the data in same range. This is very important when your quantitative data is on different scale. For an example price of house and interest rate of loan. Scaling improves the performance in terms of gradient descent and interpretation of the model. Apart from coefficients no other statistical parameters are changed.

Coefficients without normalizing:



	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.257e+05	3.13e+05	-1.361	0.174	-1.04e+06	1.89e+05
area	398.6304	32.465	12.279	0.000	334.794	462.466
bathrooms	1.501e+06	1.5e+05	10.033	0.000	1.21e+06	1.8e+06
bedrooms	4.201e+05	9.56e+04	4.396	0.000	2.32e+05	6.08e+05

Coefficients after normalizing:



	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0414	0.018	2.292	0.022	0.006	0.077
area	0.3922	0.032	12.279	0.000	0.329	0.455
bathrooms	0.2600	0.026	10.033	0.000	0.209	0.311
bedrooms	0.1819	0.041	4.396	0.000	0.101	0.263

There are 2 scaling techniques as mentioned below.

Normalized Scaling – Also known as Min-Max scaling where smallest value/min is 0 and largest value/Max is 1. All other values lie between 0 and 1.

Standardized Scaling – Here the mean is 0 and max is standard deviation. All other values lie with standard deviation.

Below is illustration of scaling on data.

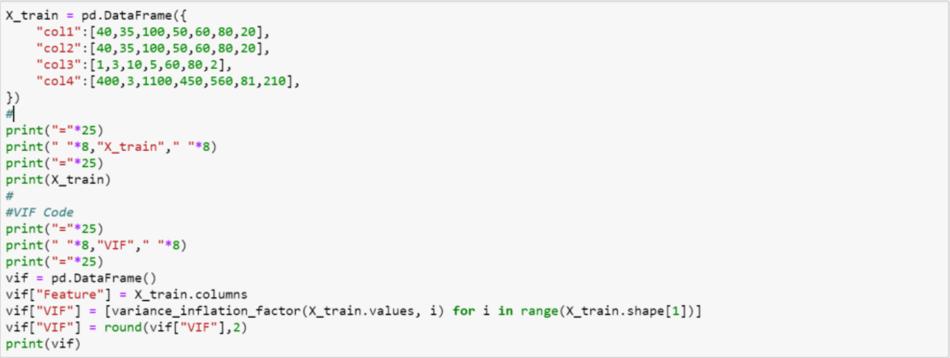
Variable	Normalized Scaling	Standardized Scaling
40	0.25	-0.588348405
35	0.1875	-0.784464541
100	1	1.765045216
50	0.375	-0.196116135
60	0.5	0.196116135
80	0.75	0.980580676
20	0	-1.372812946

Output prediction is not impacted by scaling technique used.

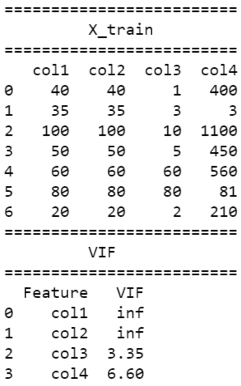
7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF indicates correlation between independent variables and is a strong indicator of multicollinearity in data. Higher the VIF greater is the possibility of that independent variable being explained by other variable(s). This helps in explaining relationship of independent variables. A smaller value preferably (less than 10) is recommended during multiple linear regression.
If an independent variable is completely explained by other variable(s) then its value is infinity.

Example1:

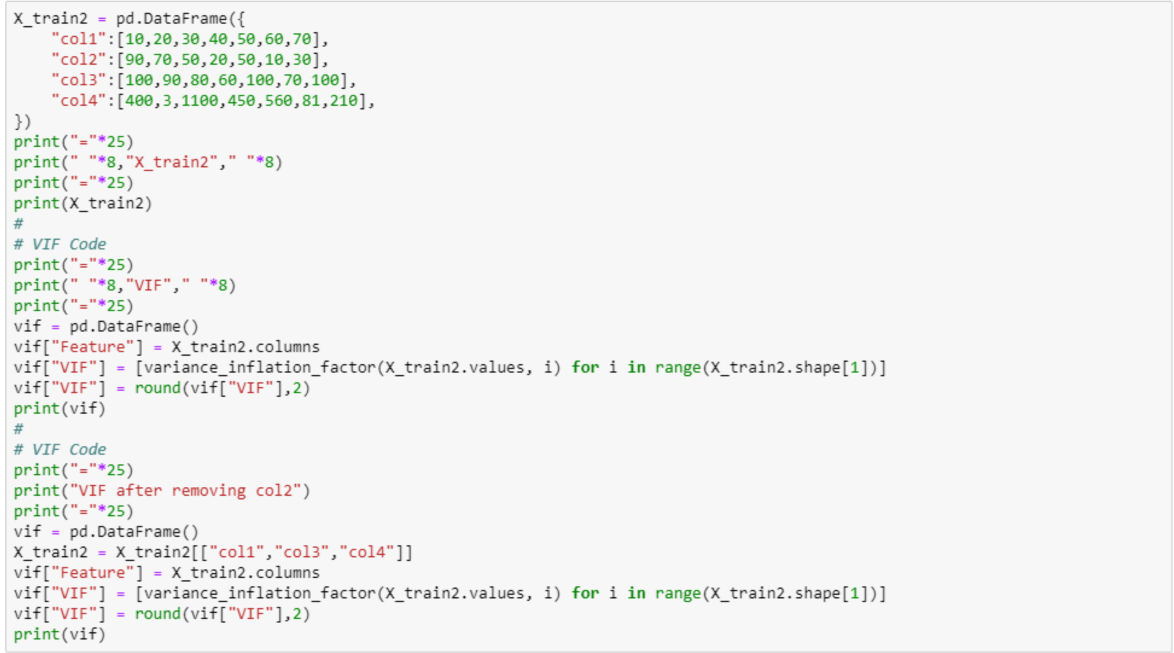


```
X_train = pd.DataFrame({
    "col1":[40,35,100,50,60,80,20],
    "col2":[40,35,100,50,60,80,20],
    "col3":[1,3,10,5,60,80,2],
    "col4":[400,3,1100,450,560,81,210],
})
#
print("="*25)
print(" "*8,"X_train"," "*8)
print("="*25)
print(X_train)
#
#VIF Code
print("="*25)
print(" "*8,"VIF"," "*8)
print("="*25)
vif = pd.DataFrame()
vif["Feature"] = X_train.columns
vif["VIF"] = [variance_inflation_factor(X_train.values, i) for i in range(X_train.shape[1])]
vif["VIF"] = round(vif["VIF"],2)
print(vif)
```

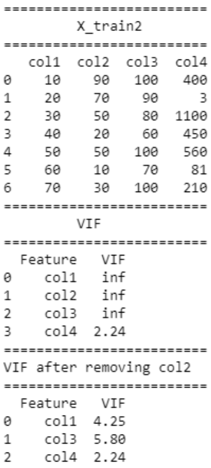


```
=========================
         X_train
=========================
   col1  col2  col3  col4
0    40    40     1   400
1    35    35     3     3
2   100   100    10  1100
3    50    50     5   450
4    60    60    60   560
5    80    80    80    81
6    20    20     2   210
=========================
         VIF
=========================
  Feature   VIF
0    col1   inf
1    col2   inf
2    col3  3.35
3    col4  6.60
```

Example2:



```python
X_train2 = pd.DataFrame({
    "col1":[10,20,30,40,50,60,70],
    "col2":[90,70,50,20,50,10,30],
    "col3":[100,90,80,60,100,70,100],
    "col4":[400,3,1100,450,560,81,210],
})
print("="*25)
print(" "*8,"X_train2"," "*8)
print("="*25)
print(X_train2)
#
# VIF Code
print("="*25)
print(" "*8,"VIF"," "*8)
print("="*25)
vif = pd.DataFrame()
vif["Feature"] = X_train2.columns
vif["VIF"] = [variance_inflation_factor(X_train2.values, i) for i in range(X_train2.shape[1])]
vif["VIF"] = round(vif["VIF"],2)
print(vif)
#
# VIF Code
print("="*25)
print("VIF after removing col2")
print("="*25)
vif = pd.DataFrame()
X_train2 = X_train2[["col1","col3","col4"]]
vif["Feature"] = X_train2.columns
vif["VIF"] = [variance_inflation_factor(X_train2.values, i) for i in range(X_train2.shape[1])]
vif["VIF"] = round(vif["VIF"],2)
print(vif)
```



```
=========================
         X_train2
=========================
   col1  col2  col3  col4
0    10    90   100   400
1    20    70    90     3
2    30    50    80  1100
3    40    20    60   450
4    50    50   100   560
5    60    10    70    81
6    70    30   100   210
=========================
         VIF
=========================
  Feature   VIF
0    col1   inf
1    col2   inf
2    col3   inf
3    col4  2.24
=========================
VIF after removing col2
=========================
  Feature   VIF
0    col1  4.25
1    col3  5.80
2    col4  2.24
```

Summarizing the examples

- In example1 col1 and col2 are exactly same and hence 1 col can explain another column. This results in infinity VIF.
- In example2 col1+col2 = col3 and hence they completely explain one another resulting in VIF a s infinity.
- In example2, after we dropped col2 there was no correlation and hence VIF values were changed.

Formula for VIF is $\frac{1}{1-R^2}$. So, if there is high correlation then R2 will be 1. This is result in VIF value of infinity since (1/0 will be infinity).

8. What is the Gauss-Markov theorem?

Gauss-Markov theorem states that under Gauss-Markov assumptions, OLS estimators will be BLUE. This explains why we select $\beta_0 \ and \ \beta_1$ such that RSS is minimum.

BLUE stands for Best Linear Unbiased Estimator.

Gauss-Markov assumptions:

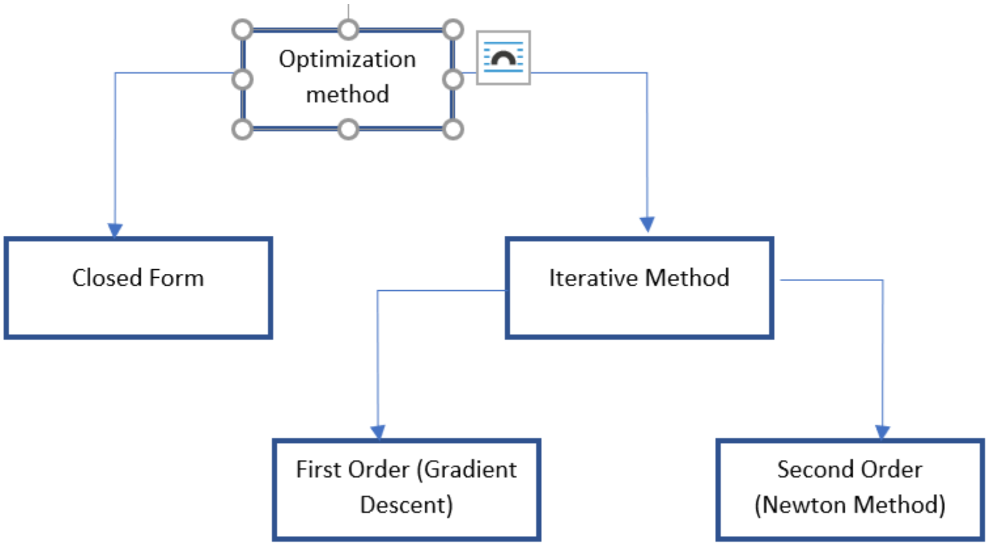
a. Linearity -> Regression model is linear in model and therefore can be expressed as
$$\hat{Y} \ = \ \beta_0 + \beta_1 X_1 \ldots\ldots + \beta_n X_n$$
b. Exogeneity -> An exogenous variable is one that isn’t affected by any other variables in the model
c. Xi values are non-stochastics i.e. non-random. Independent variables are non-stochastics whereas dependent variables are stochastics.
d. Homoscedastic – Same variance across X

Given all these assumptions, OLS estimators are BLUE i.e.

a. Estimator is linear is it is linear function of Y. OLS estimators are linear estimator.
b. Unbiased estimator is more efficient than another unbiased estimator if it has small variance.
c. We say an estimator is BLUE (Best Linear Unbiased Estimator) if it is
    i. Linear
    ii. Unbiased
    iii. More efficient than another linear unbiased estimator.

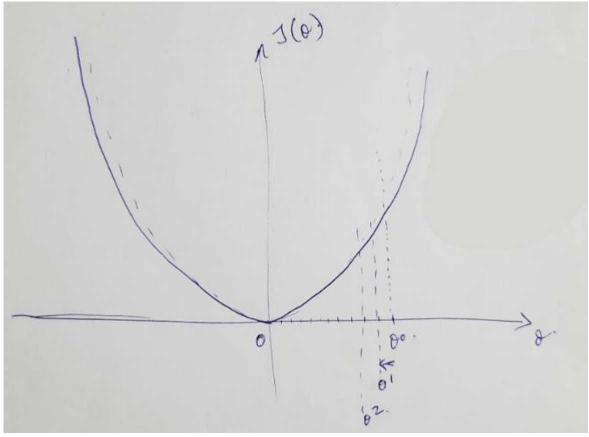
9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization technique to reduce cost function (RSS) to reach optimal solution. Gradient descent falls under first order iterative solution.

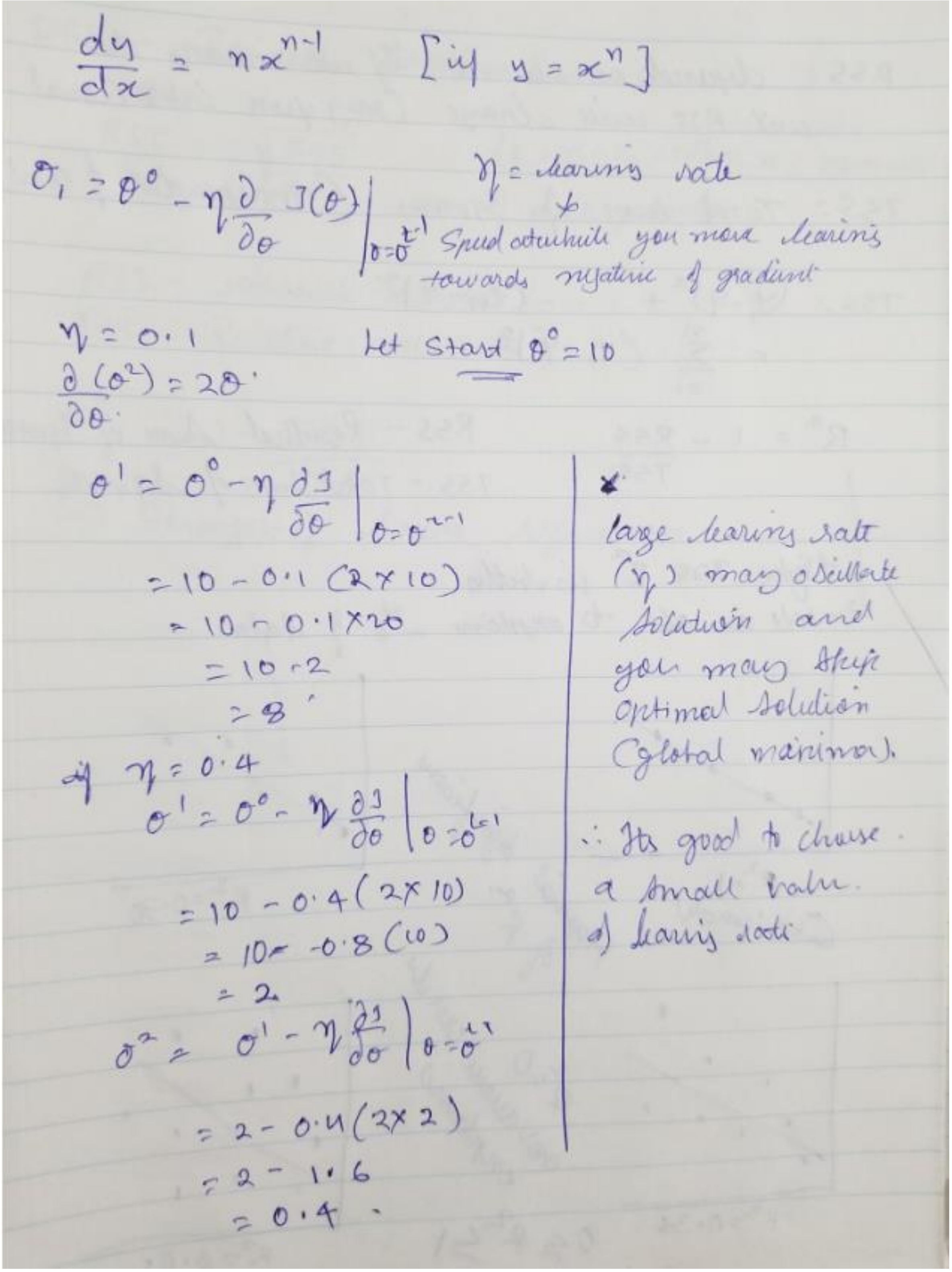


Gradient descent is an iterative method where algorithm starts at a given point and based on learning rate it slowly moves toward optimal value. Rate at which it moves is learning rate.

Small value of learning rate will result in slower convergence whereas Large value of learning rate may oscillate the solution and might result in skipping the optimal solution (i.e. global minima)



How learning rate impacts convergence is depicted below.



$$\frac{dy}{dx} = nx^{n-1} \quad [\text{if } y = x^n]$$

$$\theta_1 = \theta^0 - \eta \frac{\partial}{\partial \theta} J(\theta) \Big|_{\theta=\theta^{t-1}}$$

η = learning rate
&
Speed atwhich you move learning towards negative of gradient

$\eta = 0.1$ Let start $\theta^0 = 10$

$$\frac{\partial (\theta^2)}{\partial \theta} = 2\theta$$

$$\theta^1 = \theta^0 - \eta \frac{\partial J}{\partial \theta} \Big|_{\theta=\theta^{t-1}}$$

$$= 10 - 0.1 (2 \times 10)$$

$$= 10 - 0.1 \times 20$$

$$= 10 - 2$$

$$= 8$$

if $\eta = 0.4$

$$\theta^1 = \theta^0 - \eta \frac{\partial J}{\partial \theta} \Big|_{\theta=\theta^{t-1}}$$

$$= 10 - 0.4 (2 \times 10)$$

$$= 10 - 0.8 (10)$$

$$= 2$$

$$\theta^2 = \theta^1 - \eta \frac{\partial J}{\partial \theta} \Big|_{\theta=\theta^{t-1}}$$

$$= 2 - 0.4 (2 \times 2)$$

$$= 2 - 1.6$$

$$= 0.4$$

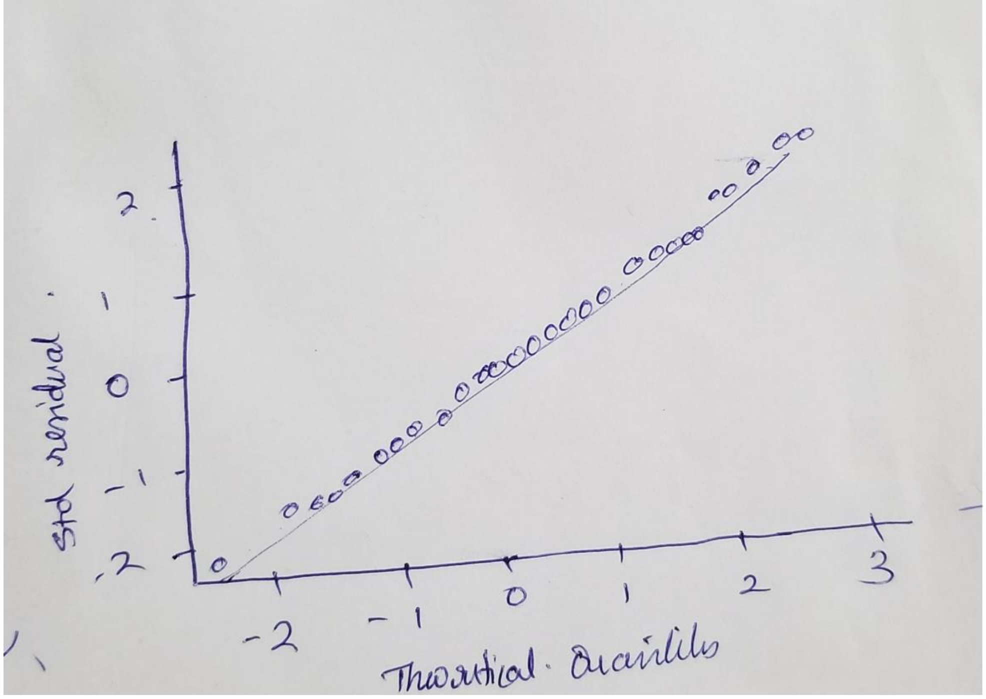
*
large learning rate (η) may oscillate solution and you may skip optimal solution (global minima).

∴ Its good to choose a small value of learning rate

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

QQ plot stands for quantile-quantile plots. QQ-plot is built by plotting 2 set of quantiles against each other. QQ- Plot helps in determining if sample data came from normally distributed population. If both type of distribution are same, then the QQ plot will be a straight line.

QQ confirms that data stands true on assumption on normality. It plots the quantile of data to that of quantile of normal distribution. For normally distributed data, it follows approximately straight line.


Std residual
Theoretical Quantiles