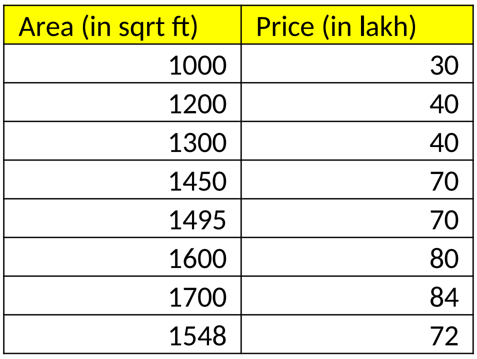
Question 1: Explain the linear regression algorithm in detail ?

Answer 1: Linear Regression: *It is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).*
So, to understand dependent and independent variable let take one example in which we have to predict the price of the house on the basis of the "Area of the house". See below image



Area (in sqrt ft)	Price (in lakh)
1000	30
1200	40
1300	40
1450	70
1495	70
1600	80
1700	84
1548	72

So, from the above image we can see that **price of the house is dependent on the Area of the house** so, **Price is a Dependent variable.**
Area is not dependent on any variable so **Area is an Independent variable.**
There are two type of Linear Regression:-
1. Simple Linear Regression: In this, only one independent variable and one dependent variable.
2. Multiple Linear Regression: In this, more than one independent variable and one dependent variable.
Equation of the Linear regression is:-
$Y = c + mX$

(Equation of Simple Linear Regression)
where c is intercept (value of y when X = 0) and m is slope of the line which can be measure by ratio of unit change in X and unit change Y.
What Linear Regression do?
Linear Regression algorithm find the best fit line (find the optimal value of c and m) through which we can predict the value of dependent variable by the given value of independent variables.

Best fit line:- A line which passes near or through the data points in such a way that the difference between **actual value** and the **predicted value** is minimum i.e. error should be minimum.
error = (y_actual – y_predicted)

Where **actual value** (Price in our case) of independent variable is given in the data while the **predicted value** (Price in our case) of independent variable is predicted by the Linear Model. Which means sum of(y_actual – y_predicted)^2 should be minimum. This is known as RSS-Residual sum of square. This is also called the Cost function of the Linear Model.

Cost Function = RSS = sum of (y_actual – y_predicted)^2

Linear Regression uses **Gradient Descent** algorithm to find best fit line.

Let's see below image to understand the concept-:

In our case, equation will be

Price = c + m * Area

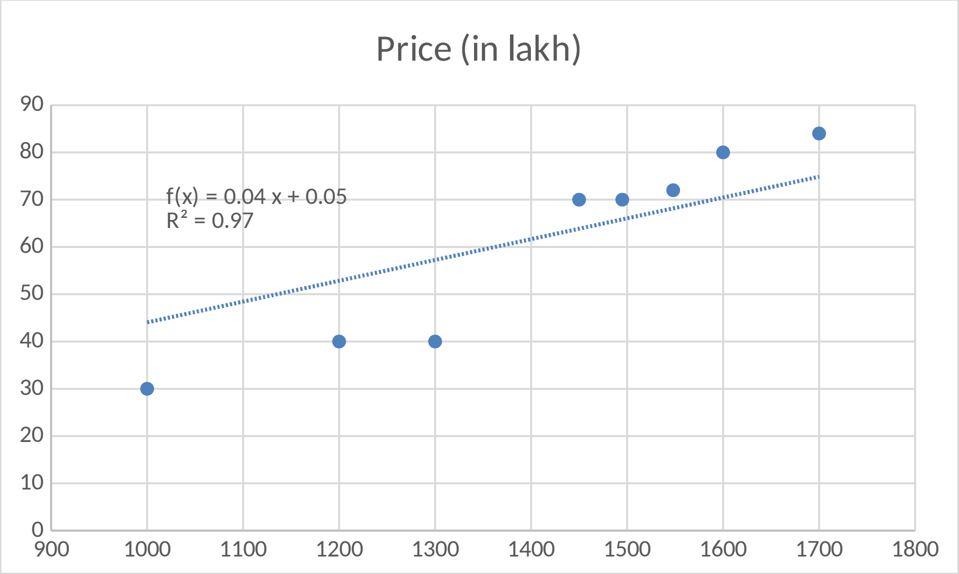
So Linear Regression algorithm finds the best fit line through which we can predict the value of Price from given of Area of the house.

In our case best fit is :

Price = 0.044 * Area + 0.05

Here optimal value of c is 0.044 and optimal value of m is 0.05.

RSS value of this line is minimum.



How to measure the strength of Linear Regression Model?

Strength of the linear model can measure by R-square and by assumptions of linear model.

R-square-: It is measure how much percentage of variance of dependent variable explained by the independent variable.

R-square = **(1 – RSS/TSS)** where RSS (Residual sum of square) and TSS (Total sum of square).

Higher the percentage means model is good.

In the case of Multiple Linear Regression there are other measure as well like F-statistic Adjusted R-square.

Assumptions:

1. Linear relationship between X and y.
   As we can see from above image that there is Linear relationship between price and area. Basically this measure we should use to check is Linear Regression Model can be create for below data.
2. Normal distribution of error terms.
3. Independence of error terms.
4. Constant variance of error terms.

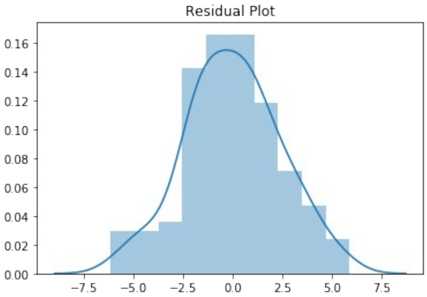
Question 2: What are the assumptions of linear regression regarding residuals?

Answer 2: There are 4 assumptions of linear regression regarding residuals

1. Linear relationship between X and y.

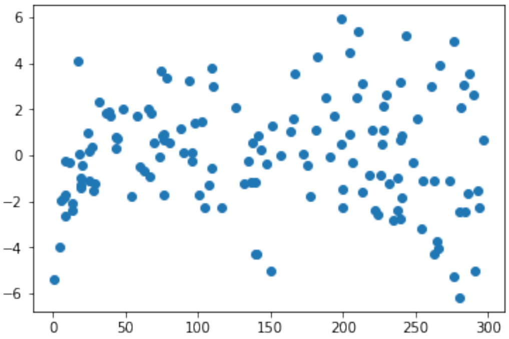
   There should at-least one independent variable which should be linearly related to dependent variable. Basically this assumptions we should use to check whether we can use the linear regression model or not. Means if there is no other independent which not linearly related to independent variable.

2. Normal distribution of error terms: Mean of the error term should be zero.



3. Independence of error terms: Error term should not be dependent.

4. Constant variance of error terms.: As we can see from the below image variance of the error term almost constant.



Question 3: What is the coefficient of correlation and the coefficient of determination?

Answer 3: Coefficient of correlation: *It is a statistical measure of the degree to which changes to the value of one continuous variable predict change to the value of another continuous variable.*

There are mainly two type of correlations which mostly used-:

1. Pearson's product moment correlation coefficient (r)
2. Spearman's rank correlation coefficient (rs)

Mostly used correlation coefficient is Pearson’s product moment correlation coefficient.

Formula for Pearson Coefficient

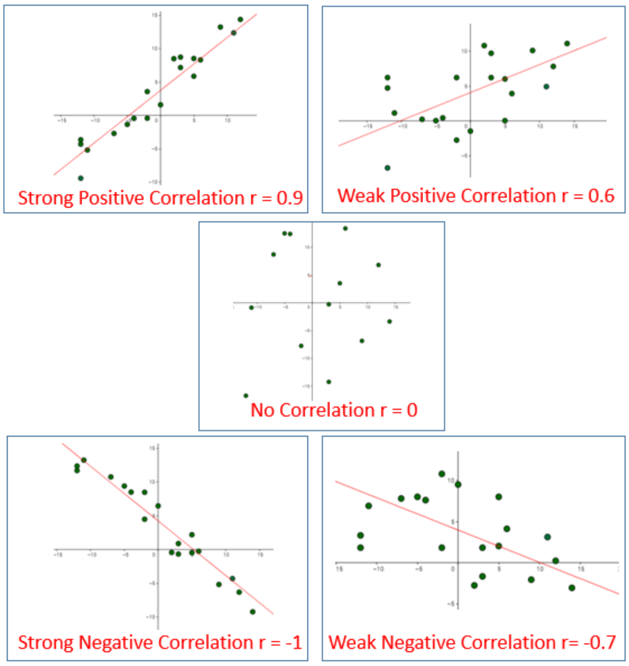
$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable
$\overline{Y}$ = mean of Y variable

And X and Y are continuous variables.
Correlation coefficient can be positive, negative or zero (means there is no relationship between two variables).
Range of correlation coefficient is from -1 to 1 means **-1 > r < 1**

r = 1 → Perfect Positively Correlated
r = 0 → No Correlation
r = -1 → Perfect Negatively Correlated

Examples of Correlation Coefficient



Strong Positive Correlation r = 0.9

Weak Positive Correlation r = 0.6

No Correlation r = 0

Strong Negative Correlation r = -1

Weak Negative Correlation r= -0.7

Coefficient of Determination: *Coefficient of determination is also known as R-squared, it is number which explains what portion of given data variation can be explained by the developed model.*
1. It is a key output of regression analysis.
2. It helps to determine the predictive power of a regression model
3. It ranges from 0 to 1.
Formula:

R2 = 1- RSS/TSS

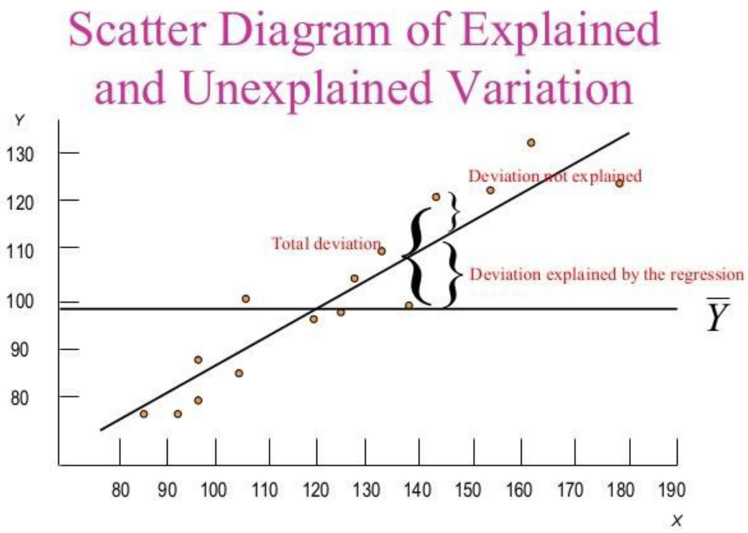
where RSS- Residual of sum of squared and TSS – Sum of error of the data from mean.

As we can see from below image-:

TSS – Sum of Total Deviation

RSS – sum of Deviation not explained

MSS – sum of Deviation Explained by the regression.



Question 4: Explain the Anscombe's quartet in detail.

Answer 4: Anscombe's quartet developed by the statistician Francis Anscombe.

It comprises of four data set and each contains eleven points (x, y) pair. Essential thing to note about data set is that they share same descriptive statistics. By constructing this Francis Anscombe demonstrate importance of graphical data before analyzing it.

Below are the four data set:-

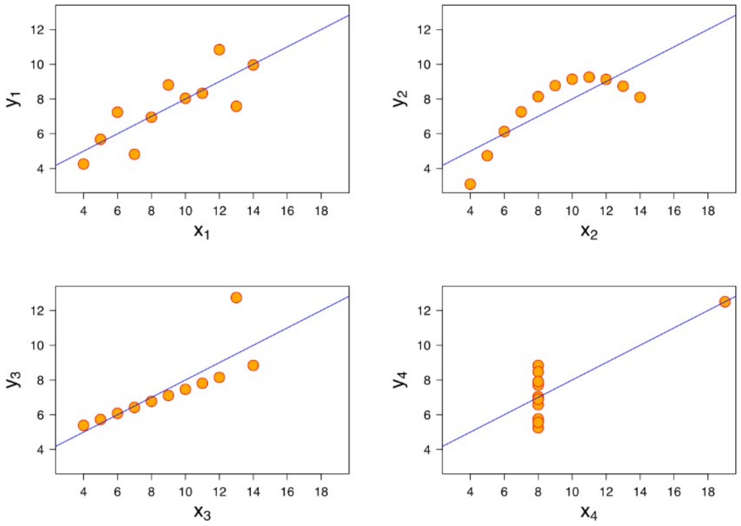
Anscombe's quartet							
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04

6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Property	**Values**	**Accuracy**
Mean of x	9	exact
Sample variance of x: sigma^2	11	exact
Mean of y	7.5	to 2 decimal places
Sample variance of y: sigma^2	4.125	+-0.003
Correlation between x and y	0.816	to 3 decimal places
Coefficient of determination (R^2)	0.67	to decimal places

So from above table we can that mean of x of all the four data set is same and Mean of y is also same. Variance of all the four data set is also same.

Let’s plot these data sets:



We can clearly see that although data set seems like same and the statistics seems also same but the graphical representation is very different.

Dataset I: Appears to have clean and well-fitting linear models.
Dataset II: Is not distributed normally.
Dataset III: The distribution is linear, but the calculated regression is thrown off by an outlier.
Dataset IV: Shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Question 5: What is Pearson's R?

Answer 5: *Pearson's R Is a correlation coefficient designed for linearly related two variables and it is not reliable for non- linearly related variables.* To show the non-linear relationship between two variables Spearman's R is designed.
For Example, Let's take y = X^3 for 100 equally separated values between 1 and 100 then correlation coefficient by technique is given below
Pearson's R = 0.91 (appox.)
Spearman's R = 1 (appox.)

As we increase the power of X the value Pearson's R will go down whereas the Spearmen's R will robust at 1.
If y = X^10 then Pearson's R= 0.66 (appox.) And Speamen's R = 1 (appox.).

So, if we sense that the relationship that is non-linear we should rely on Spearmen's R not on Pearson's R.

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

Answer 6: *Scaling is a technique through which we can bring down the features in machine learning algorithms under the same range.*

Let's say we have features with unit in kilogram, gram and liter and these features are very crucial in our Linear Regression algorithm.
As we can see that these features have different range so, understand the impact scaling by below example of **Gradient Descent algorithm**

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

As we know that Algorithms like Linear Regression, Logistic Regression and Neural Networks uses Gradient Descent Optimization algorithm at the back, and can see the use of X in the formula of Gradient Descent which affect the step size of the Gradient Descent.
To ensure that Gradient Descent moves smoothly toward the minima and that the steps for gradient descent are updated at same rate for all feature, we should scale the data before giving this data to the model.

Normalization vs Standardization: -

Normalization: is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 to 1. It is knowns as Min-Max scaling.

X' = X – Xmin / (Xmax –Xmin)

1. When the value of X in minimum in the column then the numerator become the 0, hence X' is 0.
2. When the value of X in maximum in the column then numerator becomes equal to Denominator,

   hence X' is 1.
3. If the values of X between minimum and maximum, then values of X' lies between 0 and 1.

Standardization: is another scaling technique where the values are centered around the mean with unit standard deviation. This means that means of the attribute becomes 0 and resultant distribution has unit standard deviation.

X' = (X - μ)/σ

Where μ is mean and sigma(σ) is standard deviation of the feature values. In this case values are not restricted to a particular range.

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

Answer 7: As we know that Variance Inflation Factor (VIF) determines that how well an independent variable can be expressed by other independent variables.

Formula of VIF:

VIF = 1 / (1 – R^2) where R^2 is statistical measures which measure that how much variance of dependent variable can be expressed by independent variables it mean, **if all the actual values of dependent variable lies on the Linear Regression Model Line then value of R^2 will be 1 and value of VIF will be infinity means, an independent variable fully expressed by the other independent variables.**

Question 8: What is the Gauss-Markov theorem?

Answer 8: *The Gauss-Markov theorem states that if your linear regression model satisfies the first six classical assumptions, then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators.*

If we satisfy these assumptions, then we can be confident that we are obtaining the best possible estimators.

Assumptions are :-
OLS Assumption 1: The regression model is linear in the coefficients and the error term
OLS Assumption 2: The error term has a population mean of zero
OLS Assumption 3: All independent variables are uncorrelated with the error term
OLS Assumption 4: Observations of the error term are uncorrelated with each other

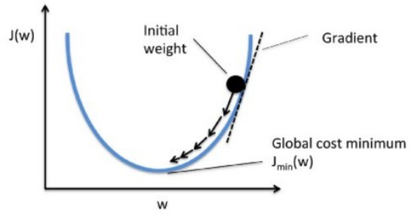
OLS Assumption 5: The error term has a constant variance (no heteroscedasticity)
OLS Assumption 6: No independent variable is a perfect linear function of other explanatory variables
OLS Assumption 7: The error term is normally distributed (optional)

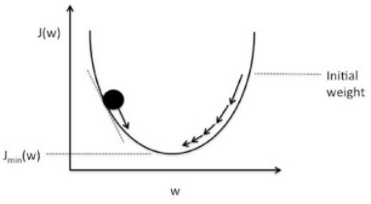
Gauss –Markov theorem says that **OLS is BLUE (Best Linear Unbiased Estimator)**
where 'Best' refers to minimum variance.

Question 9: Explain the gradient descent algorithm in detail.

Answer 9: *Gradient descent is an optimization algorithm. In Linear regression, it is used to optimized the cost function and find the values of the βs (estimators) corresponding to the optimised value of the cost function.*

Gradient descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).





The aim of Gradient Descent for linear regression is to find the solution of ArgMin J(Θ*0*, Θ1)
where J(Θ*0*, Θ1) is the cost function of the Linear Regression.

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Here h, is the linear hypothesis model, h = Θ*0* + Θ1x, y is the true output, and m is the number of data points in training set.

Gradient Descent starts with a random solution, and then based on the direction of the gradient, the solution is updated to the new value where the cost function has a lower value.

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

Answer 10: *Quantile-Quantile (Q-Q) plot, is a graphical tool to help us asses if a set of data plausibly came from some theoretical distribution such Normal, Exponential or Uniform distribution. Also. It helps if two data set came from population with a common distribution.*

This helps in scenario where we have training and test data set received separately and then we can confirm that both data set are from populations with same distribution.

Few advantages:
1. It can be used with sample size also

2. Many distributed aspects like change in symmetry, shift in scale, shift in location and presence of outlier can be detected.

Q-Q Plot fowling scenarios:

If two data sets:

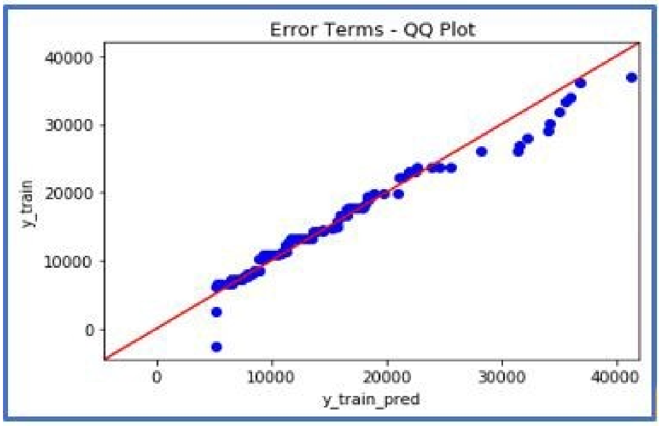
1. Come from population with a common distribution
2. Have common local and scale
3. Have similar distributional shape
4. Have similar tail behavior

Interpretation of Q-Q plot-:

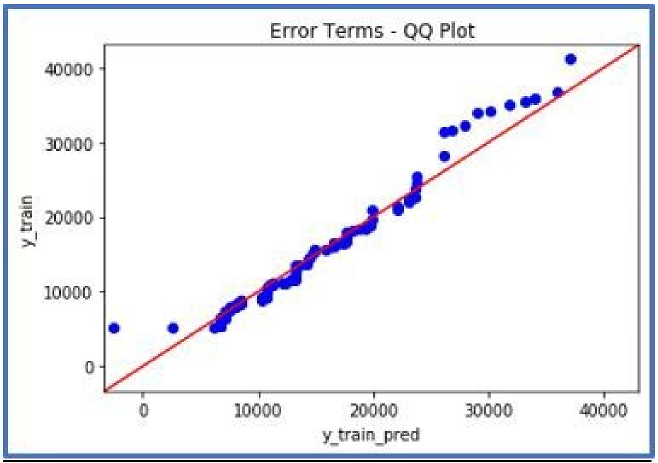
It is a plot of quantile of the first data set against the quantile of second data set.

Possible interpretation for two data sets:

1. **Similar Distribution:** If as point of quantile lies on or close to straight line at an angle of 45 degrees from x-axis.
2. **Y-values < X-values:** If the y-quantile is lower than the x-quantile



3. **X-values < Y-values:** If the x-quantile is lower than the y-quantile



4. **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degrees from x –axis