1Q. Explain the linear regression algorithm in detail.

A: Linear regression is a method of finding the best fit line between dependent and independent variables of given data. This is done by Residual Sum of Squares (RSS) method generally. Linear regression comes under supervised machine learning algorithm. Supervised learning algorithm means the model is trained with data which has label for all the variables. So, when we try to predict for the test data, this data will be evaluated against the model which it trained with labeled data set.

We start with the scatter plot of the two variables. Here, one variable is predictor which is aligned in x-axis and another one is output variable which is aligned in the y-axis. Now, we try to fit a best fit line in this scatterplot by minimizing the residual sum of square (OLS – ordinary least squares). The output variable is always a continuous variable. We use gradient descent optimization technique to minimize the cost function. Then we try to find the equation of the best fit line ($y = b0 + b1X$) and optimized b0 and b1.

After we get the optimal b0 and b1, we can calculate the predicted values. And after subtracting from actual and squaring the difference of the actual and predicted value, we get the RSS.

We generally follow below steps:

1. Load data and data visualization using EDA
2. Prepare data using dummy variable or label
3. Split and scale the train data using standardizing or normalizing
4. Build model using RFE and manual OLS approach to find the optimal betas
5. Perform residual analysis
6. Evaluate the model with test data

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

A: We have below assumptions of linear regression regarding residuals:

1. Error terms are normally distributed.
2. Error terms have mean value of 0.
3. Error terms have same and constant variance.
4. Error terms are independent of each other

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

A: Coefficient of correlation is the degree of relationship between two variables. The value of correlation can be between -1 to 1. 0 correlation coefficient means, there is no relationship between two variables. Negative sign says that the variables are inversely correlated and positive value says that they are moving the same direction.

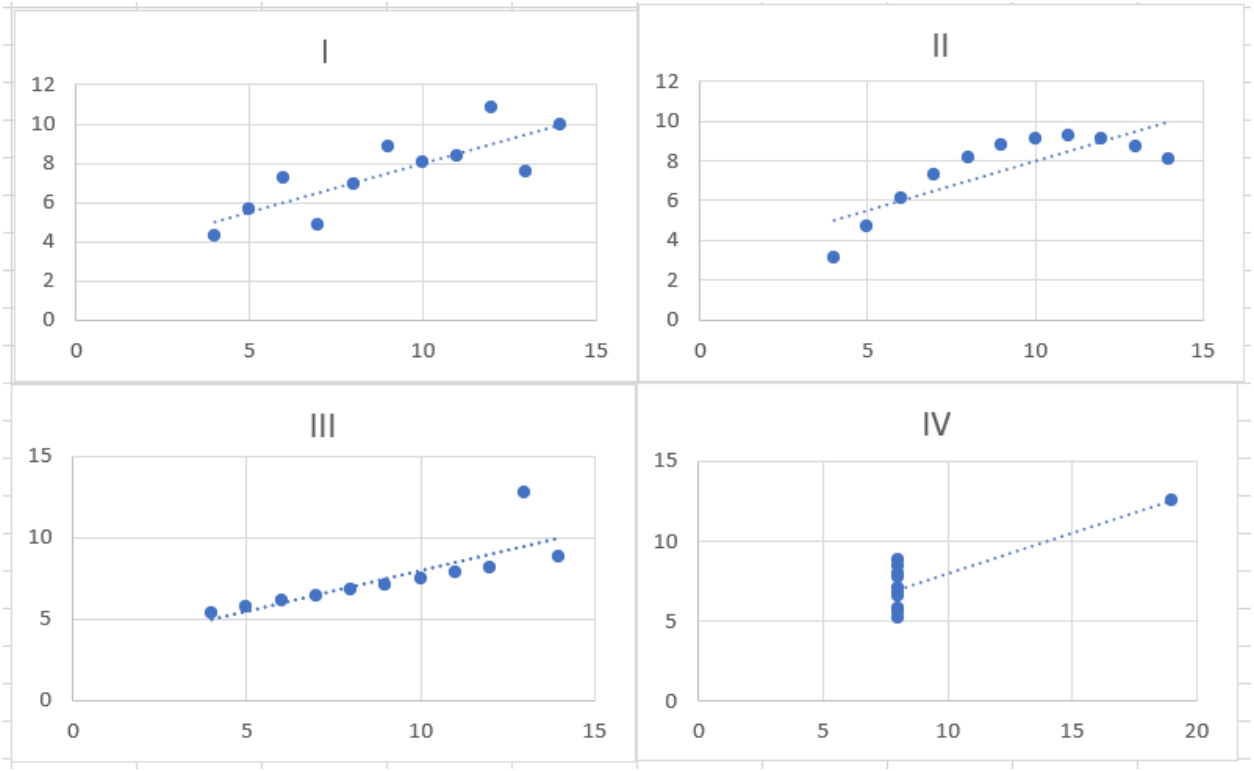
Coefficient of determination is the square of the coefficient of the correlation. This is called R square. As, this is the square value of R, the value cannot be negative and lies between 0 and 1. The strength of a linear regression model is also expressed as R square (1 – RSS/TSS).

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

A: Anscombe's quartet contains four data sets. Each data set has 11 pairs. All these 4 data sets have same Sum, Average and Standard Deviation and variance. But the data points are plotted in a graph, they tell different story each even though their descriptive statistics are same.

	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
SUM	99.00	82.51	99.00	82.51	99.00	82.50	99.00	82.51
AVG	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
STD	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

When these data sets are plotted, we get same regression lines as below.



First graph has simple linear relationship.

Second one is not normally distributed.

Third one is in linear distribution but one outlier offset the calculated regression.

Fourth one has one outlier that created high correlation coefficient even though other data points do not indicate any relationship between the variables.

5Q: What is Pearson's R?

A: The Pearson correlation coefficient is called as Pearson's R. This is also known as bivariate correlation. The value of the Pearson's R can be between -1 to 1. -1 indicates that the variables are in total negative linear correlation, 0 means no correlation and 1 means total positive correlation.

This is calculated as the covariance of the two variables divided by product of the standard deviation of the two variables.

Pearson's R(x, y) = cov(x,y) / sd(x) * sd(y)

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

A: Scaling is the process to standardized/normalized the data to fit in certain range. This is part of data preprocessing.

Dataset may contain features with high variance in magnitude, unit and range. So, it will be problematic to read and understand the statistics if two quantities are in different range or unit. This why scaling is performed to make the whole dataset into a single scale.

Standardized scaling replaces the values by their z-score. So, all the value of a particular column can be read by its mean = 0 and standard deviation = 1. Formula behind: z = (x – u)/sigma

Normalized scaling replaces all the value of a column ranging between -1 and 1. Essentially, the whole dataset will have range of -1 to 1. Formula behind: x' = x – mean(x) / max(x) – min(x)

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

A: VIF = 1 / (1-R2)

So, if R2 = 1, VIF is infinite (undefined)

We can get R2 = 1 when the variable is totally correlated with all other variables. In that case, we will get coefficient of correlation as 1 and eventually R2 = 1.

8Q: What is the Gauss-Markov theorem?

A: Gauss-Markov theorem states that if the errors in the linear regression model are uncorrelated with mean value zero and constant variance, the ordinary least squares estimate the coefficients gives the best linear unbiased estimate (BLUE).

Assumption of the theorem:

1. Parameters are in a linear relationship
2. Data is randomly sampled from the population.

3. Regressors are not perfectly correlated with each other.
4. Regressors are not correlated with error term.
5. The errors have constant variance.

Q9: Explain the gradient descent algorithm in detail.

A: Gradient descent is an optimization technique to find out the value of the coefficients of a function that minimize the cost function of the linear regression.

We have straight line equation: y = b0 + b1x

We need to find the optimal b0 and b1 in order to find the best fit line of the data points.

Here, we will use the gradient descent algorithm to find the optimal b0 and b1. The cost function thus becomes the function of J(b0, b1).

Gradient descent stats with a random solution and based on the direction of the gradient, it updates the new value. With this approach, it tries to find out the minimum value of the function.

Q10: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A: Q-Q plots are graph of two quantiles against each other. A 45 degree is plotted on the QQ plot. If the plot follows the 45-degree line, then it can be assumed that the sample data is normally distributed or the data set is coming from a common distribution.

If we want to fit a linear regression model and if all the points lie approximately on the line, then we can say that the they have common distribution. If the points don't fit, they we can say the residuals are not Gaussian. So, for small sample size, we can assume that the estimator is not Gaussian. So, the standard confidence interval and significance tests are invalid.