1. **Explain the linear regression algorithm in detail.**

Linear Regression is one of the supervised learning technique which allows us to model a relationship between a dependent variable (Target variable) and independent variable(s).

Linear Regression Algorithm:

Below are the steps which are used to create a model using linear regression process.

a. Reading and Understanding the data.
b. Split the data into Train and test datasets (Ratio is 70:30 or 80:20).
c. Train your model on training dataset (Scale the variables using Normalization or Standardization).
d. Build Model I.e. Fit the train dataset using OLS (Ordinary Least Square Method)
e. Check for different parameters example – adjusted R^2, p-Values of feature variables and F-statistics and VIF (to check multicollinearity in the model)
f. If everything looks fine, proceed to the next step else drop one highly insignificant feature variable and go back to step e.
g. Check strength of the model using r2_score and MSE(Mean Square Error) methods.
h. Residual Analysis: Plot actual Y and predicted y (using the linear equation we have got from train dataset) of train dataset and analyse the residuals. They should follow assumptions of linear regression algorithm.
i. **Model Evaluation**: Now try to fit test dataset on the model created with train dataset.
j. Once this is done, evaluate the model.
    - For this we can Plot actual y value and predicted y-value of test dataset.
    - We can also plot error terms for this.
k. **Error Analysis:** Check for r2_score of test dataset.
    - A model is said to be good if If the difference between r2_score of train and test dataset is less than 5%.

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

Below are the assumptions of linear regression regarding residuals:

a) The mean of residuals is zero.
b) Homoscedasticity of residuals i.e. constant variance.
c) Error terms are normally distributed.
d) Error terms are independent of each other.

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

Coefficient of correlation: It is the degree of relationship (i.e. strength and direction) between 2 variables. Its value may vary from -1 to 1.

- A Positive Correlation represents that two variables are moving in unison and they fall and rise together.
- A negative correlation represents inverse relationship between both variables means with rise of one variable another variable fall and vice versa.
- Correlation of 0 shows that there is no relationship between 2 variables.

- Correlation of +1 or -1 represents a strong relationship between variables but the direction of relationship depends on + or – symbol as stated for positive and negative correlation respectively.
  It is also called Pearson's R.

Coefficient of determination: It represents the variation in y which is explained by all X variables altogether. Higher the better. It is represented by R-square.

- Its value may vary from 0 to 1.
- An R-sqaure of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from independent variable.
- R^2 value between 0 and 1 indicates the extent to which dependent variable is predicted.
  For example – R2 value of 0.10 means that 10% of variance in Y is predictable from X.
  It is also a measure of how well the regression line represents the data.
  For a simple linear regression, it is square of coefficient of correlation.

4. Explain the Anscombe's quartet in detail.

Answer:

Basically, Anscombe's quartet emphasizes on importance of Data Visualization with along with Summary Statistics. Most often, people find it very difficult to visualize the data and go ahead with the summary statistics to make any prediction about the data.

One Statistician named, Francis Anscombe, realized in 1973 that summary statistics i.e. mean, average and variance only allows us to understand the variance in data but they don't provide us the information on how the data looks like in their native form which if we know, can completely change our prediction.

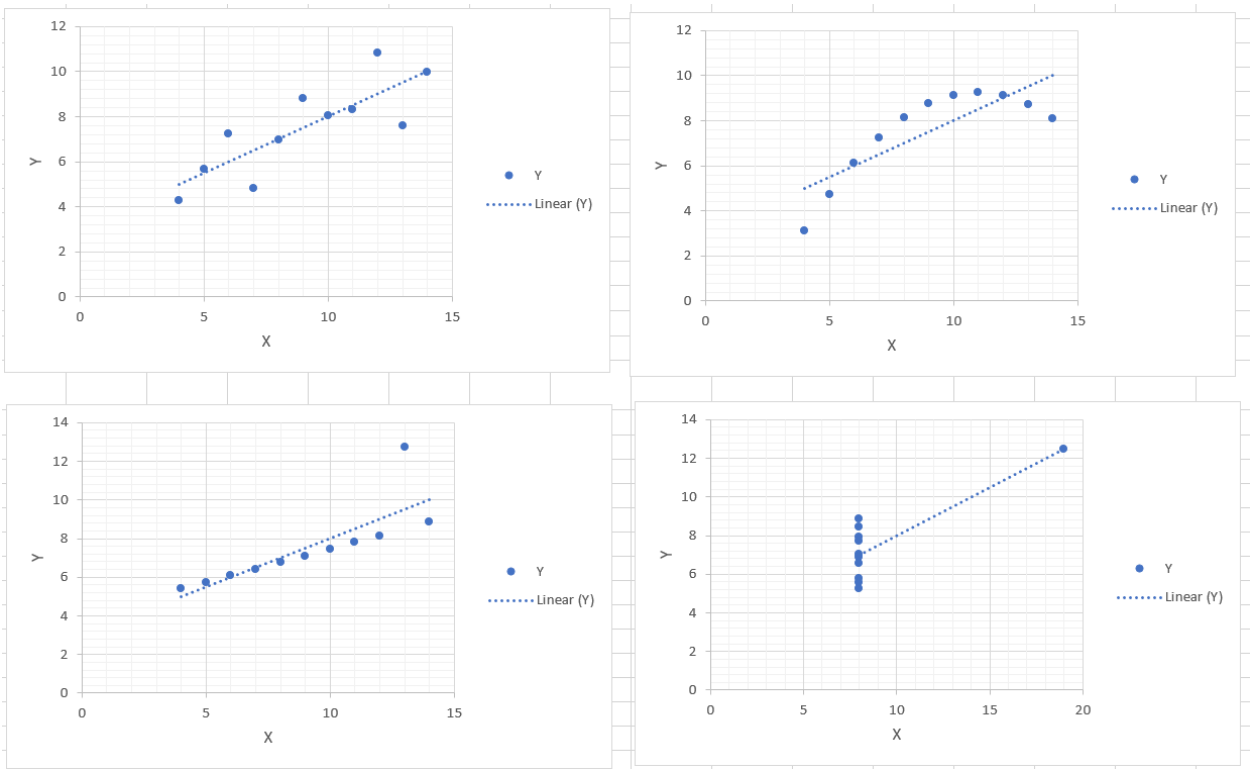
He provided a group of four datasets, which are also known as Anscombe's quartet. Now all these datasets have same variance, mean, sum, and standard deviation. Let's see them:

	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	82.51	99	82.51	99	82.5	99	82.51
Mean	9	7.500909	9	7.500909	9	7.5	9	7.500909
Standard Deviation	3.316625	2.031568	3.31662479	2.031657	3.316625	2.030424	3.316625	2.030579
Correlation		0.816421	Corrleation	0.816237	Correlatio	0.816287	Correlatio	0.816521

Here, if we see summary statistics of all four datasets, we found that:

a. Sum of X is 99 and Sum of Y is 82.51 for all 4 datasets.
b. Similarly, Mean of X is 9 and mean of Y is 7.50 for all 4 datasets.
c. Standard deviation of X is 3.32 and that of Y is 2.03 across all datasets.
d. With that correlation between X and Y is also same across all datasets i.e. 0.82

But If I plot all these four datasets, plots will look as:



Here,

1. First Dataset is corresponding to 1st dataset. It shows a linear relationship between X and Y
2. Dataset II i.e. second plot shows that relationship between X and Y is not linear as the plot is a curve.
3. Dataset III i.e. third plot shows a linear relationship between X and Y. However, it would have been a perfect relationship if the outlier was not there.
4. Dataset IV i.e. fourth plot shows that 1 outlier is sufficient to produce a high correlation value between X and Y. But as we see the best fit line came from this is actually not best.

Hence, we can conclude that visualizing the data is very important before starting on data analysis. Outliers should be removed otherwise it may result into incorrect predictions.

5. What is Pearson's R?

Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two independent variables.

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

Scaling: It is a step of data pre-processing which is applied to independent variables or features of data. It basically helps to normalize the data within a range. Sometimes, it also helps in speeding up the calculations in algorithm.

Why Scaling is performed:

1. Faster convergence of gradient descent methods
2. Ease of interpretation
3. Real world dataset contains features that highly vary in magnitudes, units and range. Because of this, objective functions will not work as expected without normalization in some of machine learning algorithms.

For example - Many classifiers calculate the distance between two data points using Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature.

Let's understand this with help of an example –

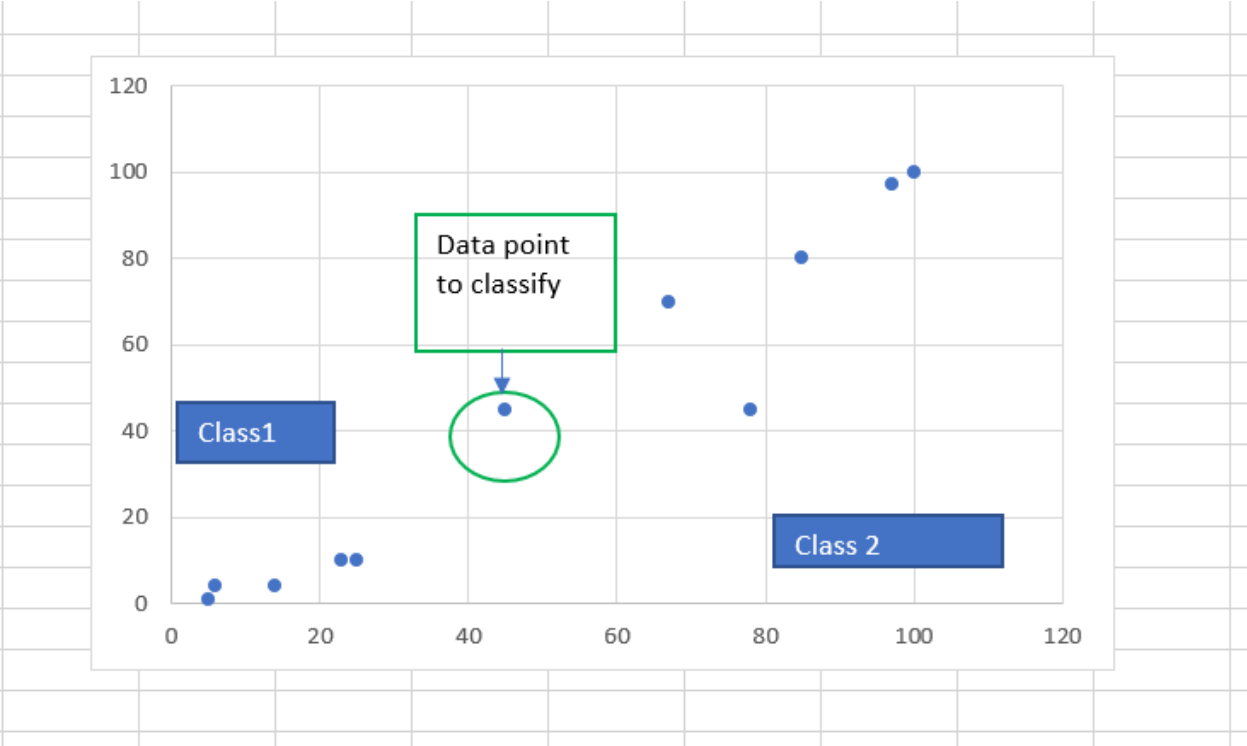
If we have a data-set with features – age, salary, BHK apartment with data size of 5000 people, each having these independent data features.

Now if we label these data points as:

- Class 1 – Yes: means with given age, salary and BHK apartment feature value one can buy the property.
- Class 2 – No: means the given age, salary and BHK apartment feature value one can't buy the property.

Using dataset to train the model, one aims to build a model that can predict whether one can buy a property or not with given feature values.

Once the model is trained, an N-dimensional(N being number of features present in dataset) graph with data points from given dataset, can be created. Representation of that model would be –



As shown in the figure, some of the data points belong to class 1 and higher data points belong to class2.. Now a new data point (encircled in figure) is given and it has different independent values for the 3 features (Age, Salary, BHK Apartment) mentioned above. The model has to predict whether this data point belongs to Yes or No.

Prediction of the class of new data point:

The model calculates the distance of this data point from the centroid of each class group. Finally, this data point will belong to that class, which will have a minimum centroid distance from it.

Suppose centroid of class 1 is [40, 22 Lacs, 3] and data point to be predicted is [57, 33 Lacs, 2].

Using Manhattan Method,

Distance = (|(40 - 57)| + |(2200000 - 3300000)| + |(3 - 2)|)

It is clearly seen that Salary feature will dominate all other features while predicting the class of the given data point and since all the features are independent of each other i.e. a person's salary has no relation with his/her age or what requirement of flat he/she has. This means that the model will always predict wrong.

So, the simple solution to this problem is Feature Scaling. Feature Scaling Algorithms will scale Age, Salary, BHK in fixed range say [-1, 1] or [0, 1]. And then no feature can dominate other.

What is the difference between normalized scaling and standardized scaling?

Normalized Scaling or Min-Max Scaling:

Through this method, all variables are scaled in such a way that all values lie between 0 and 1. This is done using maximum and minimum value of that particular feature.

$x_{changed} = (x - x_{min})/(x_{max} - x_{min})$

Here xchanged : Scaled value of that data point which we want to calculate.

x: Actual value of the data point.

xmin: Minimum value of that feature variable

xmax: Maximum value of the feature variable

Standardized Scaling: The variable are scaled in such a way that their mean is 0 and standard deviation is 1.

$x= (x - mean(x))/sd(x)$

- Here we subtract min value of the population from all values so that the new range of population will be [0, (x_max - x_min)]
- Now we will divide each element by (x_max - x_min) to convert the range of population into [0, 1]

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

Answer:

VIF i.e. Variance Inflation Factor tells us whether the features are correlated to each other in our regression model which could affect significance of that model.
VIF is calculated as: $1/(1-R^2)$

Generally, a VIF of less than 5 is considered good and model is considered free of multi-collinearity.
But there are cases when VIF is INF for our feature variable(s), that means in that case tolerance level$(1-R^2)$ is 0 which indicates that the estimates are most imprecise.
Means the feature variables are perfectly redundant.

8. What is the Gauss-Markov theorem?

Answer:

Gauss-Markov Theorem states that if our Linear Regression model satisfies a certain set of assumption, then OLS i.e. Ordinary Least Squares regression produces the regression coefficients such that it gives us the Best Linear Unbiased Estimate(BLUE). Which means

- Coefficients are unbiased.
- Estimates have the smallest possible variance.

⇨ **Below are the assumptions which should be followed by our linear regression model:**

a. The parameters we are estimating using OLS must be linear in nature with all predictor variables.
b. The error terms should have a population mean of 0.
c. Our data must be randomly selected from the population dataset.
d. Independent variables should not be correlated with each other.
e. Error terms should have a constant variance (Homoscedasticity)
f. Error terms are normally distributed.
g. Independent variables are not correlated with error terms.

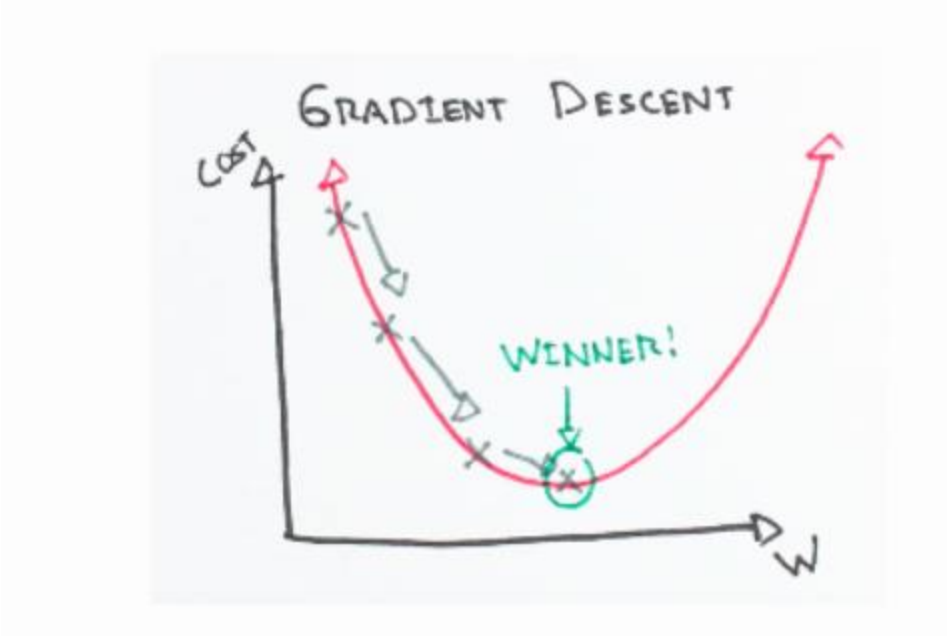
9. Explain the gradient descent algorithm in detail.

Answer:

Gradient Descent is an optimization algorithm which is used to minimize a function by iteratively moving into a direction of steepest descent. In machine learning, we try minimizing the cost function using gradient descent method which is used to update parameters(coefficients) in our model.

Let's understand the approach with the help of a diagram:

Say, you are at highest left position in below diagram and you need to come at the minima position. Now to come to the minima position, you need to take some steps and then again check the direction where you need to go. Now the step size which you take while going from peak position to minima position is called **Learning Rate.**



Learning Rate or alpha cannot be very high as it is associated with the risk of overshooting minima point as slope is constantly changing and
also, it cannot be very small, as it would take long time to converge and become computationally expensive.

So, we plot the cost function with different values of alpha so that we can pick up the right value of alpha or learning rate.

Algorithm:
For a linear regression, our cost function is :
SSE = ½ * Sum(Actual value of target variable - Predicted value of target variable)^2
Where y_{pred} = ax+b (for a simple linear regression where we have only 1 independent variable)

Aim of Gradient descent is to find optimal values of a and b to reduce prediction error.

Steps involve in Gradient Descent Algorithm are:

a. Initialize the coefficients (or weights) with random values and calculate SSE.
b. Calculate the gradient i.e. change in SSE when there is a small change in values of coefficients.
c. Adjust the weights with gradients to reach optimal values where SSE is minimized.(Take partial derivatives of cost function w.r.t. each parameter)
d. Use new weights for prediction and to calculate SSE.
e. Repeat steps b and c till further adjustments to weights doesn't significantly reduce the error.

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

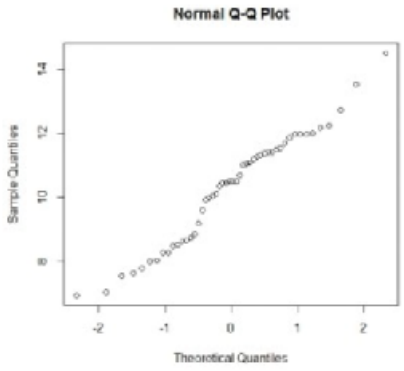
Answer:

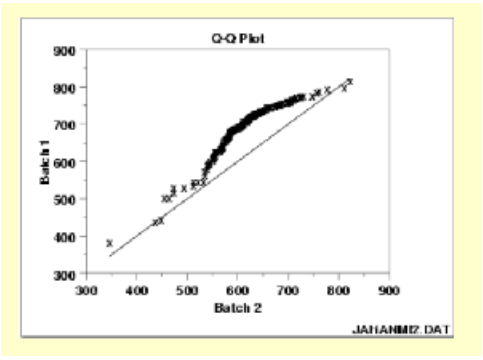
Q-Q plot i.e. quartile -quartile plot is a graphical technique by which we determine if two data sets or samples have come from the population which follows the same normal distribution.

We take a quantile set from both the datasets and plot against each other. By quantile, I mean fraction of those data points which come below the given value.
We also plot a reference line which is a 45 degrees line.

- If two datasets have come from same population or the different population with common distribution, the plotted points should fall approximately close to this reference line.
- If there is a great difference of the points from this reference line, we can conclude that both datasets belong to the population with different distribution.

Example – Lets see below diagrams:





If we see figure1 – it shows that our data points are somewhere along the reference line which clears that both the distribution have come from the population with same distribution.

While figure 2 – shows that data points are at a great distance w.r.t reference line which states that both the datasets have come from population with different distribution.

Uses of Q-Q plot:

Q-Q plot provides us answer of following points:

- If the datasets have come from population with common distribution as explained above.
- If the datasets follow similar distribution shapes.
- If the datasets have common location and scale.
- Q-Q plot also enable us to test various distribution aspects simultaneously. Shifts in scale, presence of outliers, shifts in location and changes in symmetry – can all be detected from this

plot. For example – if two data sets come from populations whose distributions differ only by a shift in scale, the points should lie along a straight line that is displaced either up or down against reference line.

- It has an advantage over histograms as q-q plots doesn't have any design parameters such as number of bins.

Importance of Q-Q Plot in linear regression:

In real world scenarios, being a data analyst or data scientist, one of the important job is to predict the behavior of independent variables and draw conclusions only based on the available dataset as data collection can be a costly process to follow and may result into biased result based on the nature of data collection process.

Now to conclude from the limited data points, we try to find out a common distribution to apply any pre-defined inferences.

We use Q-Q plots to check this common distribution which may ease our work and enable us to make correct prediction even if we had limited data.