1.Explain the linear regression algorithm in detail.

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. It is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

Types of applications:

1.Finding out the effect of Input variables on target variables.

2.Finding out the change in Target variable with respect to one or more input variables.

3.To find out upcoming trends.

Model that assumes a linear relationship between the input variables X and the single output variable y. y can be calculated from a linear combination of the input variables X.

When there is a single input variable X, the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

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

1.It is assumed that the error terms are normally distributed. If the residuals are not normally distributed, the model is not able to explain the relation in the data.
2. It is assumed that the residuals have a mean value of zero.
3. It is assumed that the residual terms have the same variance.
4.It is assumed that the residual 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 between two variables, say x and y. It can go between -1 and 1.

It is denoted by r:
The value of r is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations.

Coefficient of determination:

We call it R square.
It shows percentage variation in y which is explained by all the x variables together. Higher the better.

It is always between 0 and 1. It can never be negative – since it is a squared value.
The *coefficient of determination* is such that $0 \leq r^2 \leq 1$.

4.Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe.

It comprises four datasets, each containing eleven (x,y) pairs.
The essential thing to note about these datasets is that they share the same descriptive statistics.
When they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

5.What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables.The Pearson's correlation coefficient varies between -1 and +1 If the variables tend to go up and down together, the correlation coefficient will be positive.

If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Scaling:
It is used to normalise the range of independent variables or features of data.
The range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

Uses:
It can make the analysis of coefficients easier.If your features differ in scale then this may impact the resultant coefficients of the model and it can be hard to interpret the coefficients.

Difference between normalised scaling and standardised scaling:

1.Normalization usually means to scale a variable to have a values between 0 and 1.
2.Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation.

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

Variance inflation factor is a measure of the amount of multicollinearity.

A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

8.What is the Gauss-Markov theorem?

The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate possible.

There are five Gauss Markov assumptions:
1. Linearity: the parameters we are estimating using the OLS method must be themselves linear.
2. Random: our data must have been randomly sampled from the population.
3. Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
4. Exogeneity: the regressors aren't correlated with the error term.
5. Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

9.Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.

In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

1.The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.
2.The cost of the coefficients is evaluated by plugging them into the function and calculating the cost.
3.The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point.

4.A learning rate parameter must be specified that controls how much the coefficients can change on each update.
5.This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

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

It helps to graphically analyse and compare two probability distributions by plotting their quantiles against each other.
If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

1.It can be used with sample sizes.
2.Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
3.A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.