1. Explain the linear regression algorithm in detail.
   Ans - Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting line that minimizes the difference between the predicted and actual values. The algorithm calculates the coefficients (slope and intercept) that define the line using a method called ordinary least squares. It then uses these coefficients to make predictions on new data points by multiplying the independent variables with their respective coefficients and adding the intercept term. Linear regression is widely used for tasks like predicting housing prices, sales forecasting, and trend analysis.

2. What are the assumptions of linear regression regarding residuals?
   1. Linearity: The relationship between the dependent variable and the independent variables is assumed to be linear. The residuals should be randomly scattered around zero, indicating that the model captures the linear trend adequately.
   2. Independence: The residuals should be independent of each other, meaning that there should be no correlation between consecutive residuals. This assumption ensures that the model does not capture any systematic patterns or time-dependent relationships.
   3. Homoscedasticity: The residuals should have a constant variance across all levels of the independent variables. In other words, the spread of the residuals should be consistent throughout the range of predicted values. Deviations from homoscedasticity indicate heteroscedasticity, which violates the assumptions.
   4. Normality: The residuals should follow a normal distribution. This assumption is important for statistical inference and hypothesis testing. If the residuals deviate significantly from normality, it may affect the validity of the p-values and confidence intervals associated with the model's coefficients.
   5. No multicollinearity: The independent variables should not be highly correlated with each other. High multicollinearity can make it difficult to distinguish the individual effects of the independent variables on the dependent variable.

3. What is the coefficient of correlation and the coefficient of determination?
   The coefficient of correlation, commonly denoted as "r," measures the strength and direction of the linear relationship between two variables. It quantifies the degree to which changes in one variable are associated with changes in the other variable. The coefficient of correlation ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation. The coefficient of correlation is calculated using the covariance between the two variables divided by the product of their standard deviations.

   The coefficient of determination, denoted as "R-squared," is a measure of how well the independent variables explain the variation in the dependent variable. It represents the proportion of the total variation in the dependent variable that is explained by the linear regression model. R-squared ranges from 0 to 1, with 0 indicating that the independent variables have no explanatory power, and 1 indicating that they explain all the variation in the dependent variable. R-squared is calculated as the square of the coefficient of correlation (r) between the predicted values and the observed values of the dependent variable.

4. Explain the Anscombe’s quartet in detail.
   Anscombe's quartet is a collection of four datasets, each consisting of pairs of x and y variables. Despite having similar summary statistics, the datasets differ significantly in their patterns and relationships. The quartet serves as a reminder that relying solely on summary statistics can be misleading and emphasizes the importance of visualizing data to gain a deeper understanding. It highlights the need to explore data through various statistical techniques and graphical representations to avoid making incorrect assumptions or drawing misleading conclusions.

5. What is Pearson’s R?
   Pearson's correlation coefficient, denoted as "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It assesses the degree to which changes in one variable are associated with changes in the other variable. Pearson's R ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation. It is calculated as the covariance between the two variables divided by the product of their standard deviations. Pearson's R is widely used in statistical analysis, correlation studies, and feature selection to understand the linear association between variables.

6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Scaling, in the context of data preprocessing, refers to the process of transforming numerical variables to a common scale. It involves adjusting the values of the variables to specific ranges or distributions. Scaling is performed to ensure that variables with different scales or units contribute equally to the analysis, as many machine learning algorithms are sensitive to the scale of the input features.
   Normalized scaling, also known as min-max scaling, rescales the values of a variable to a range between 0 and 1. It preserves the relative relationships and distribution of the data. The formula for normalized scaling is:

   scaled_value = (value - min) / (max - min)

   Standardized scaling, also known as z-score scaling, transforms the values of a variable to have a mean of 0 and a standard deviation of 1. It centers the data around the mean and standardizes the spread of the values. The formula for standardized scaling is:

   scaled_value = (value - mean) / standard_deviation

   The main difference between normalized scaling and standardized scaling lies in the transformation applied to the data. Normalized scaling maintains the original range of the data, while standardized scaling creates values that are standardized based on the mean and standard deviation. Normalized scaling is suitable when the distribution shape and range of the data are

important, while standardized scaling is useful when comparing variables with different means and variances or when using algorithms that assume normally distributed variables.

7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   The occurrence of an infinite value for the Variance Inflation Factor (VIF) typically happens when there is perfect multicollinearity in the data. Perfect multicollinearity means that one or more independent variables can be exactly predicted by a linear combination of other independent variables. This situation leads to a breakdown in the calculation of VIF, resulting in an infinite value. It occurs when the determinant of the matrix of independent variables' correlations is zero. This happens, for example, when one variable is a perfect linear combination of others, or when variables are duplicated or highly correlated. Infinite VIF values indicate a severe issue in the model, requiring remedial actions such as removing correlated variables or addressing the multicollinearity problem through feature engineering or dimensionality reduction techniques.

8. What is the Gauss-Markov theorem?
   The Gauss-Markov theorem is a fundamental result in linear regression analysis. It states that under the assumptions of the classical linear regression model (CLRM), the ordinary least squares (OLS) estimator is the best linear unbiased estimator (BLUE) among all linear unbiased estimators. In other words, the OLS estimator has the minimum variance among all linear unbiased estimators in the CLRM. This theorem holds even if the errors are heteroscedastic and non-normally distributed, as long as other CLRM assumptions are met. The Gauss-Markov theorem highlights the efficiency and optimality of the OLS estimator in providing unbiased estimates of the regression coefficients.

9. Explain the gradient descent algorithm in detail.
   The gradient descent algorithm is an optimization algorithm used to find the minimum of a function. In the context of machine learning, it is commonly employed to minimize the cost function in training a model. The algorithm starts with an initial set of parameter values and iteratively updates them by moving in the direction of steepest descent. At each iteration, the gradients of the cost function with respect to the parameters are calculated. The parameters are then updated by subtracting a fraction of the gradients multiplied by the learning rate, which determines the step size. This process continues until convergence is reached or a maximum number of iterations is reached. Gradient descent is efficient for large-scale problems and is available in different variants such as batch gradient descent, stochastic gradient descent, and mini-batch gradient descent.

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    A Q-Q (quantile-quantile) plot is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution.

    In linear regression, a Q-Q plot is often used to evaluate the assumption of normality for the residuals. By comparing the observed residuals to the expected residuals under a normal

distribution, a Q-Q plot helps assess whether the residuals deviate significantly from normality. If the points in the plot lie approximately along a straight line, it suggests that the residuals follow a normal distribution. Conversely, deviations from a straight line indicate departures from normality.

The importance of a Q-Q plot in linear regression lies in its ability to identify potential issues with the normality assumption. If the residuals deviate substantially from normality, it may impact the validity of statistical inference, such as p-values and confidence intervals. Detecting non-normality in the residuals can guide further analysis, such as applying appropriate transformations to the variables or considering alternative regression models that accommodate non-normal distributions. The Q-Q plot serves as a visual diagnostic tool to assess the assumption of normality and supports the robustness and reliability of the linear regression analysis.