1. Explain the linear regression algorithm in detail.

Ans: The Linear Regression algorithm is used to predict the linear relationship between one or more independent variables and a dependent variable.

Here are the steps we follow in Linear Regression:

1. **Importing the data:**
   Linear Regression requires a dataset containing input variables (independent variables) and an output variable (dependent variable). We import the dataset using pandas and the imported data frame will be in pandas dataframe type.
2. **Model representation:**
   The linear regression model represents the relationship between the input variables (X) and the output variable (y) as a linear equation of the form:
   y = b0 + b1*x1 + b2*x2 + ... + bn*xn Here, y is the predicted output, b0 is the y-intercept, b1 to bn are the coefficients or slopes associated with the input variables x1 to xn.
3. **Cost Function:**
   The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted values and the actual values. This difference is measured using a cost function. The most common cost function used in linear regression is the Mean Squared Error (MSE), given by:
   MSE = (1/n) * Σ (y_pred - y_actual) ^2 Here, n is the number of observations, y_pred is the predicted value, and y_actual is the actual value.
4. **Parameter Estimation:**
   The coefficients (b0, b1, ..., bn) are estimated using a method called Ordinary Least Squares (OLS). OLS minimizes the cost function by finding the values of the coefficients that minimize the sum of the squared differences between the predicted and actual values.
5. **Training the Model:**
   Training the linear regression model involves finding the optimal values of the coefficients. This is done by solving the OLS equations, which involve taking partial derivatives of the cost function with respect to each coefficient. The coefficients that minimize the cost function are obtained using various numerical optimization techniques.
6. **Making Predictions:**
   Once the model is trained, it can be used to make predictions on new data. Given a set of input variables, the model calculates the predicted output value by substituting the input values into the linear equation.
7. **Evaluation:**
   The performance of the linear regression model is evaluated using different metrics such as the coefficient of determination (R^2), mean absolute error (MAE), or root mean squared error (RMSE). These metrics provide insights into how well the model fits the data and how accurate its predictions are.

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

Ans: The assumptions which we made for Linear Regression are:

1. Linear relationship between independent variable (X1, X2, …) and dependent variable (y).
2. Error terms are normally distributed (not individual X and y).
3. Error terms have constant variance (homoscedasticity).
4. Error terms are independent of each other.
5. Linear regression assumes that the independent variables are not highly correlated with each other

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

Ans: The correlation of coefficient, also known as the correlation coefficient or Pearson's correlation coefficient (r), measures the strength and direction of a linear relationship between two variables. It measures how close data points in a scatter plot are to a straight line.

The coefficient of correlation, denoted by "r," ranges between -1 and 1. Here's what different values of r indicate:

- If r = 1, it indicates a perfect positive linear relationship, where all the data points lie exactly on a straight line with a positive slope.
- If r = -1, it indicates a perfect negative linear relationship, where all the data points lie exactly on a straight line with a negative slope.
- If r = 0, it indicates no linear relationship or correlation between the variables. The data points are scattered randomly and do not follow a linear pattern.

In a linear regression model, the coefficient of determination, denoted by R2 (R-squared), is a measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variables. It indicates how well the regression model fits the data.

R^2 ranges between 0 and 1. Here's what different values of R^2 indicate:

- R^2 = 0 indicates that the independent variables have no explanatory power on the dependent variable. The regression model does not capture any of the variability in the data.
- R^2 = 1 indicates that the independent variables perfectly explain the variability in the dependent variable. The regression model captures all the variability in the data.

In Linear Regression we mostly use the coefficient of determination (R^2) method in analyzing the model

4. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but exhibit vastly different patterns when visualized.

Ex:

Dataset 1:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5

y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Dataset 2:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5

y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74

Dataset 3:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5

y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Dataset 4:

x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8

y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89

Although all four datasets have the same means, variances, correlations, and linear regression parameters, they exhibit distinct patterns when plotted.

Dataset 1 resemble a relatively linear relationship between x and y, and a simple linear regression model would be appropriate.
Dataset 2 is also fairly linear but with an outlier that has a strong influence on the regression line.
Dataset 3 shows a non-linear relationship that would be better suited to a polynomial regression model.
Dataset 4 consists of a single outlier that drastically affects the correlation and regression analysis, highlighting the importance of identifying and handling outliers.

Anscombe's quartet serves as a reminder that summary statistics alone cannot capture the complexities and nuances of real-world data. It underscores the significance of visualizing data to gain a deeper understanding and avoid making erroneous conclusions based solely on statistical measures.

5. What is Pearson's R?

Ans: The correlation of coefficient, also known as the correlation coefficient or Pearson's correlation coefficient (r), measures the strength and direction of a linear relationship between two variables. It measures how close data points in a scatter plot are to a straight line.
The coefficient of correlation, denoted by "r," ranges between -1 and 1. Here's what different values of r indicate:

- If r = 1, it indicates a perfect positive linear relationship, where all the data points lie exactly on a straight line with a positive slope.

- If r = -1, it indicates a perfect negative linear relationship, where all the data points lie exactly on a straight line with a negative slope.
- If r = 0, it indicates no linear relationship or correlation between the variables. The data points are scattered randomly and do not follow a linear pattern.

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

Ans: Scaling refers to the process of transforming numerical values of variables to a specific range or distribution in the context of data analysis and machine learning. It involves modifying the values of the variables while keeping their relative relationships intact. Scaling is done to ensure that variables are on a comparable scale and to address issues that may arise as a result of differences in magnitude or units.

The main reasons for performing scaling are:

- Ease of interpretation
- Faster convergence of gradient descent methods

There are two common types of scaling techniques

- Standardization
- MinMax Scaling (Normalization)

Standardization:

$x = (x - \text{mean}(x))/\text{sd}(x)$

The variable distribution is centered around 0 with a standard deviation of 1.

MinMax Scaling (Normalization):

$x = (x - \min(x))/(\max(x) - \min(x))$

The minimum value of the variable becomes 0, the maximum value becomes 1, and all other values are scaled proportionally within that range.

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

Ans: Variance Inflation Factor (VIF):

VIF is a measure of how well a predictor variable is correlated with all the other variables, excluding the target variable.

$$VIF(i) = 1/ (1 - R(i)^2)$$

If VIF is infinite that means R^2 value of that particular variable is 1, which is very high. While an infinite VIF indicates a problem with perfect multicollinearity, high (but not infinite) VIF values can still indicate the presence of multicollinearity. A general rule of thumb is that VIF values greater than 5 or 10 are considered high and should be investigated further.

8. What is the Gauss-Markov theorem?

Ans: The Gauss-Markov theorem, also known as the Gauss-Markov assumption or Gauss-Markov conditions, is an important result in linear regression. It defines the conditions under which ordinary least squares (OLS) regression estimators are the best linear unbiased estimators (BLUE) with the least variance among all linear unbiased estimators.

According to the Gauss-Markov theorem, if the following assumptions are met in a linear regression model, the OLS estimators have several desirable properties:

- Linear relationship between independent variable (X1, X2, …) and dependent variable (y).
- Error terms are normally distributed (not individual X and y).
- Error terms have constant variance (homoscedasticity).
- Error terms are independent of each other.
- Linear regression assumes that the independent variables are not highly correlated with each other
- The independent variables are not perfectly correlated with each other.

Violations of these assumptions, such as heteroscedasticity, endogeneity, or omitted variable bias, can result in biased or inefficient estimates. In such cases, other estimation techniques or corrective measures may be required.

9. Explain the gradient descent algorithm in detail.

Ans: Gradient Descent:

Gradient descent is one of the most popular algorithms to perform optimization and the most common way to optimize neural networks. It is an optimization algorithm which is used to minimize the function. The function which is set to be minimized is called an objective function. For machine learning, the objective function is also termed the cost function or loss function. It is the loss function which is optimized (minimized) and gradient descent is used to find the most optimal value of parameters/weights which minimizes the loss function. The loss function is the measure of the squared difference between actual values and predictions.

The equation be: -

$$w = w - \alpha \nabla_w J$$

$$b = b - \alpha \nabla_b J$$

where J (w, b) -> cost/loss function which is to be minimized.
w, b are parameters.

Here's a step-by-step explanation of the gradient descent algorithm:

- Initialize Parameters: Start by initializing the parameters or coefficients of the model with some initial values. These parameters will be iteratively updated during the algorithm.
- Define the Cost Function: Choose a cost function that quantifies the error or mismatch between the predicted output of the model and the actual target values. The goal is to minimize this cost function.
- Compute the Gradient: Calculate the gradient (partial derivatives) of the cost function with respect to each parameter. The gradient indicates the direction of steepest ascent in the cost function.
- Update Parameters: Adjust the parameter values by taking a small step in the opposite direction of the gradient. The step size is controlled by the learning rate hyperparameter, which determines the size of each parameter update. The learning rate should be chosen carefully to balance convergence speed and stability.
- Repeat Steps 3 and 4: Iterate the process of computing the gradient and updating the parameters until convergence is achieved. Convergence is typically determined by reaching a predefined tolerance level or when the change in the cost function becomes very small.
- Output: Once convergence is reached, the optimized parameter values are obtained, which can be used for making predictions or further analysis.

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

Ans: A Q-Q plot, also known as a quantile-quantile plot, is a graphical tool used to determine whether a given dataset follows a specific probability distribution, such as the normal distribution. It compares the observed data quantiles to the quantiles expected from a theoretical distribution. The Q-Q plot is a visual representation of how well the data aligns with the assumed distribution.

The use and importance of a Q-Q plot in linear regression are as follows:

- **Assumption Checking**: In linear regression, it is important to verify whether the residuals (the differences between the observed values and the predicted values) follow a normal distribution. This assumption is required for accurate statistical inference as well as the validity of hypothesis tests and confidence intervals. The Q-Q plot allows us to visually inspect the normality assumption by comparing the observed residuals to the expected quantiles of a normal distribution.
- **Residual Analysis**: The Q-Q plot helps in identifying deviations from normality in residuals. If the plot's data points deviate significantly from the expected diagonal line, it indicates that the residuals are not normally distributed. Deviations from the diagonal line can indicate skewness, heavy tails, or other

irregularities. Such deviations may indicate the presence of influential observations, model misspecification, or assumption violations.

- **Outlier Detection**: Q-Q plots can also be used to identify outliers in a dataset. Outliers are observations that significantly deviate from the expected pattern. Outliers appear in a Q-Q plot as points that deviate significantly from the expected line. Outliers must be identified because they can have an impact on the regression analysis, influencing the estimated coefficients and model fit.
- **Transformation Selection**: If the Q-Q plot shows significant deviations from normality, a linear regression model may not be appropriate for the data because it violates the underlying assumptions. In such cases, data transformations can be used to improve adherence to assumptions. The Q-Q plot can be used to evaluate the effectiveness of various transformations by examining how they affect the residual normality.