## 1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a supervised machine learning algorithm used for predicting a continuous target variable based on one or more independent variables. Here's a detailed explanation:

- **Objective**: Linear regression aims to find the best-fitting linear relationship (line) between the independent variables (features) and the dependent variable (target) by minimizing the sum of squared errors.
- **Equation**: The linear regression equation is of the form:

  ```
  makefile
  ```
- `Y = b0 + b1*X1 + b2*X2 + ... + bn*Xn`
  - 
  - Y represents the predicted target variable.
  - b0 is the intercept.
  - b1, b2, ..., bn are the coefficients of the independent variables X1, X2, ..., Xn.
- **Fitting**: The coefficients (b0, b1, b2, ..., bn) are estimated using techniques like Ordinary Least Squares (OLS), which minimizes the sum of squared differences between the actual and predicted values.
- **Assumptions**: Linear regression assumes that there is a linear relationship between the variables, the errors are normally distributed, and there is no multicollinearity or heteroscedasticity.

## 2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. It illustrates the importance of visualizing data and not relying solely on summary statistics. Key points:

- All four datasets have the same mean, variance, correlation, and linear regression line.
- However, when you plot these datasets, you'll see significant differences in their distributions and relationships.
- Anscombe's quartet highlights that summary statistics alone can be misleading and may not capture the full complexity of the data.

## 3. What is Pearson's R? (3 marks)

Pearson's correlation coefficient (often denoted as Pearson's R) is a statistic used to measure the linear relationship between two continuous variables. It ranges from -1 to 1:

- If R = 1, it indicates a perfect positive linear relationship.
- If R = -1, it indicates a perfect negative linear relationship.
- If R = 0, it indicates no linear relationship.

Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It quantifies the strength and direction of a linear association between variables.

## 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

**Scaling** in the context of machine learning refers to the process of transforming features to have a consistent scale. It is performed to:

- Ensure that different features contribute equally to the model.
- Prevent features with larger scales from dominating those with smaller scales.
- Facilitate the convergence of gradient-based optimization algorithms.

**Normalized scaling** (also known as Min-Max scaling) scales features to a specific range, typically [0, 1]. It is calculated as:

```
scss
X_normalized = (X - X_min) / (X_max - X_min)
```

**Standardized scaling** (Z-score scaling) scales features to have a mean of 0 and a standard deviation of 1. It is calculated as:

```
makefile
X_standardized = (X - X_mean) / X_std
```

Normalized scaling retains the original distribution but scales it, while standardized scaling centers the data around the mean and adjusts for the spread.

## 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF (Variance Inflation Factor) measures the extent to which the variance of an estimated regression coefficient is increased due to multicollinearity in the data. VIF can become infinite when a predictor variable can be perfectly predicted by a linear combination of other predictor variables.

In other words, when one predictor variable is a linear combination of others, its VIF becomes infinite because its variance is infinitely inflated by the multicollinearity. This happens because the matrix used in the VIF calculation becomes singular, making it impossible to compute the exact value of VIF.

## 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A **Q-Q (Quantile-Quantile) plot** is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, usually the normal distribution. Here's how it works:

- In a Q-Q plot, the x-axis represents the quantiles (sorted values) from the theoretical distribution (e.g., a normal distribution), and the y-axis represents the quantiles of the observed data.
- If the points on the Q-Q plot approximately follow a straight line (the 45-degree line), it suggests that the data closely follows the theoretical distribution.

Importance in linear regression:

- Q-Q plots are used to check the assumption of normality of residuals in linear regression. Residuals should ideally follow a normal distribution for the linear regression model to be valid.
- If the Q-Q plot deviates from a straight line, it indicates that the residuals are not normally distributed. This can help identify potential issues in the model, such as heteroscedasticity or outliers.
- Q-Q plots are valuable for diagnosing and improving the accuracy of linear regression models by highlighting deviations from normality assumptions in the residuals.