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

- Linear regression is a supervised machine learning algorithm that models the relationship between a dependent variable and one or more independent variables. It establishes a linear connection between the input (x) and the output (y), helping predict outcomes. It uses a cost function to optimize model performance and provide the best-fit line between x and y.

2. **Explain Anscombe’s quartet in detail.**

- Anscombe's quartet consists of four datasets that have similar simple statistical properties but differ significantly in visual representation. The concept illustrates the importance of plotting data before analysis to uncover anomalies or characteristics such as outliers, non-linearity, and data diversity.

3. **What is Pearson’s R?**

- Pearson’s R measures the strength and direction of a linear relationship between two variables. The value ranges from -1 to +1, where positive values indicate a positive correlation, negative values indicate a negative correlation, and a value of 0 signifies no correlation.

4. **What is scaling? Why is it performed? What is the difference between normalized scaling and standardized scaling?**

- Scaling adjusts data to a standard range to ensure algorithms perform better. Normalization scales data to a range of 0 to 1, while standardization shifts data to have a mean of 0 and a standard deviation of 1. Normalization is prone to information loss about outliers compared to standardization.

5. **Why is the value of VIF sometimes infinite?**

- VIF, or Variance Inflation Factor, measures the correlation among predictors in a model. A VIF value of infinity occurs when there is perfect multicollinearity, meaning two or more predictors are perfectly correlated, making it hard to determine their individual contributions.

6. **What is a Q-Q plot? Explain its use and importance in linear regression.**

- A Q-Q plot compares two distributions by plotting their quantiles against each other. It helps in identifying if a dataset follows a specific distribution. In linear regression, it is used to verify if the residuals follow a normal distribution.

Assignment-Based Questions

1. **What can you infer about categorical variables from the dataset?**

   - Key inferences include:

     - Season and month significantly affect bike bookings.

     - The holiday variable shows no significant effect on bike bookings, making it a poor predictor.

     - Other variables like weather and working day show trends that suggest their influence.

2. **Why is it important to use drop_first=True during dummy variable creation?**

   - Using `drop_first=True` helps in preventing multicollinearity by dropping one level of categorical variables. This ensures that the remaining dummy variables are independent.

3. **Which numerical variable has the highest correlation with the target variable?**

   - Temperature (temp) shows the highest correlation with the target variable, but it has multicollinearity with 'atemp', so only one should be used.

4. **How did you validate the assumptions of linear regression after building the model?**

   - Assumptions were validated by checking multicollinearity, p-values, residual normality (via histogram), and overall model significance (F-statistics). The model's coefficients were also checked for significance.

5. **Which are the top 3 features contributing significantly to bike demand in the final model?**

   - The top 3 features are:

     - Temperature (temp): Positively influences bike bookings.

     - Weather Situation 3 (weathersit_3): Negatively affects bike bookings.

     - Year (yr): Positively impacts bike bookings.