Assignment-based Subjective Questions:

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer 1: It helped in understanding the relationship between the variables and how it affects the dependent variables across different categories.

Question 2: Why is it important to use drop_first=True during dummy variable creation?

Answer 2: When creating dummy variables, we need to drop one of the categories to avoid a statistical issue called multicollinearity, it can also be inferred from the other categories.

Therefore, we use the **drop_first=True** parameter to drop one of the categories, which avoids multicollinearity and makes the statistical analysis easier to interpret.

Question 3: Looking at the pair plot among the numerical variables, which one has the highest correlation with the target variable?

Answer 3: Temperature has the highest correlation with the target variable “cnt”.

Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer 4: The assumptions are validated through the residual analysis technique.

Question 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the shared bike demand?

Answer 5: Temperature, Year, and Seasons.

General Subjective Questions:

Question 1: Explain the linear regression algorithm in detail.

Answer 1: Linear regression is a method to predict a continuous output variable based on one or more input variables by finding the best-fitting straight line through a set of data points. The algorithm splits the data into a training and testing set, fits the model to the training data, evaluates its performance on the testing set, and then uses it to predict new data. It assumes a linear relationship between the input and output variables, and outliers in the data can affect its performance.

Question 2: Explain the Anscombe’s quartet in detail.

Answer 2: Anscombe's quartet is a group of four datasets that have the same statistical properties but look different when visualized. It shows the importance of visualizing data and not relying only on summary statistics. The quartet reminds us to examine data visually before making conclusions based solely on statistical analysis.

Question 3: What is Pearson’s R?

Answer 3: Pearson's R is a statistical measure that shows the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1, where values close to +1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation and values close to 0 indicate little or no linear relationship. It is calculated by dividing the covariance of the two variables by the product of their standard deviations. It assumes that the relationship is linear, normal, and homoscedastic.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer 4: Scaling is the process of adjusting the range of values of a variable so that they fall within a specific range, and it is done to make sure that all variables are comparable and that the magnitude of each variable does not affect the analysis results. Normalized scaling and standardized scaling are two commonly used scaling techniques.

Question 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer 5: The VIF value becomes infinite when one predictor variable can be perfectly predicted from a linear combination of the other predictor variables, which is known as perfect multicollinearity, and this can occur when a variable is a linear combination of one or more other variables or when one variable has the same values as another variable.

Question 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer 6: A Q-Q plot is a graphical method used to assess whether a set of data is normally distributed, and it compares the distribution of the data to the expected normal distribution by plotting the quantiles of the data against the quantiles of the normal distribution.
