Assignment-based Subjective Questions:

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

Ans: Important categorical variables, such as season, weather, month, and weekday, significantly influence the demand for shared bikes. For example, demand is more sensitive in summer and fall than in winter and spring. Similarly, more bike rentals are witnessed during clear weather conditions than heavy rainfall.

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

Ans: Having drop_first=True will ignore the multicollinearity by dropping the first level of each categorical variable. This, in turn, ensures the independence of the dummy variables, hence stabilizing and improving the interpretability of the model.

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

Ans: The variable 'registered' has the highest correlation with the target variable 'cnt', indicating that the number of registered users is a strong predictor of total bike demand.

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

Ans: We validated the assumptions of linear regression by performing residual analysis. This included checking for homoscedasticity (constant variance of residuals), normality of residuals using Q-Q plots, and multicollinearity using Variance Inflation Factor (VIF).

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?

Ans: The top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. 'yr' (year) - indicating an increasing trend in demand over time.
2. 'temp' (temperature) - higher temperatures are associated with increased bike demand.
3. 'atemp' (apparent temperature) - similar to temperature, it significantly impacts bike demand.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The goal is to predict the dependent variable (Y) using a linear combination of the independent variables (X1, X2, ..., Xn). The equation of the line is given by:

$Y=\beta0+\beta1X1+\beta2X2+...+\betanXn+\epsilon$

where β0 is the intercept, β1,β2,...,βn are the coefficients, and ϵ\epsilonϵ is the error term. The coefficients are estimated by minimizing the sum of the squared differences between the observed and predicted values (least squares method).

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but very different distributions and visual patterns. It illustrates the importance of visualizing data before analyzing it, as relying solely on summary statistics can be misleading. Each dataset demonstrates different anomalies like outliers, non-linearity, and the effect of a single influential point.

3. What is Pearson's R?

Ans: Pearson's R, or Pearson correlation coefficient, is a measure of the linear correlation between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation.

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

Ans: Scaling adjusts the range of independent variables to ensure they are on a comparable scale, which improves the performance and stability of machine learning models. Normalized scaling adjusts the range of values to [0, 1], while standardized scaling adjusts the values to have a mean of 0 and a standard deviation of 1. Normalization is useful when the distribution of data is unknown, while standardization is preferred when the data follows a normal distribution.

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

Ans: An infinite VIF occurs when there is perfect multicollinearity, meaning one independent variable is an exact linear combination of other independent variables. This indicates redundancy in the predictors, leading to instability in the regression coefficients.

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

Ans: A Q-Q (quantile-quantile) plot is a graphical tool to assess if a dataset follows a particular theoretical distribution, typically the normal distribution. In linear regression, it is used to check the normality of residuals. If the residuals follow a straight line in the Q-Q plot, it indicates that they are normally distributed, which is an assumption of linear regression.