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? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Categorical variables like season, weathersit, yr, and mnth could have a significant impact on the dependent variable, cnt (the total count of bikes shared). For instance, seasons and weather conditions likely play a crucial role in bike demand. During more favorable seasons (e.g., spring or summer) and better weather conditions (weathersit indicating clear weather), we can expect a higher bike-sharing count. Workingday and holiday might also influence demand, with working days typically showing higher bike-sharing activity than holidays.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: Using drop_first=True helps to avoid the dummy variable trap, where multicollinearity arises due to the inclusion of all categories. By dropping one category, the model can use the remaining categories without redundancy, ensuring that the independent variables remain independent, thus making the regression more stable.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: By plotting the pair-plot, one would observe that the temp,atemp variable has the highest correlation with the cnt variable.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Linearity: Scatterplots of residuals vs. predicted values should show no pattern.
Homoscedasticity: Residual plots should show constant variance.
Normality: A Q-Q plot of residuals should approximate a straight line to ensure the residuals are normally distributed.
No multicollinearity: Checking the Variance Inflation Factor (VIF) for each predictor to ensure multicollinearity is not a problem.
Independence: Durbin-Watson test can be used to check the independence of residuals.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)

Answer: temp, weathersit_3 , yr.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression models the relationship between a dependent variable and one or more independent variables. It fits a line (or hyperplane) that minimizes the sum of squared differences between observed and predicted values. Assumptions include linearity, independence, homoscedasticity, and normality of residuals.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet consists of four datasets with identical summary statistics but different visual patterns. It highlights the importance of data visualization to detect patterns or anomalies that numerical metrics alone cannot reveal.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R measures the linear correlation between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). A value of 0 means no linear correlation.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling adjusts feature values to a similar range. Normalization rescales data to [0, 1], while standardization makes data have a mean of 0 and a standard deviation of 1. Scaling is important for models sensitive to feature magnitude.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

An infinite VIF occurs due to perfect multicollinearity, where one variable can be exactly predicted by others, causing the model to fail in estimating coefficients.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot compares data quantiles to a theoretical distribution (usually normal). In regression, it checks if residuals are normally distributed, which is important for making valid inferences.