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: <Your answer for Question 1 goes below this line> (Do not edit)

Categorical variables such as season, weathersit, mnth, yr, weekday, and holiday have logical relationships with the bike demand (cnt). Variables like season and weathersit strongly correlate with environmental factors that impact user behavior, while yr captures growth trends in bike-sharing popularity.

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: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True during dummy variable creation helps to avoid multicollinearity in the regression model by eliminating one category from each categorical variable. This ensures that the dummy variables are not perfectly correlated, which could otherwise distort the model's coefficients and lead to redundancy in the feature set.

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: <Your answer for Question 3 goes below this line> (Do not edit)

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: <Your answer for Question 4 goes below this line> (Do not edit)

To validate the assumptions of Linear Regression, the following steps were taken:

1. **Linearity**: Checked by plotting the predicted values vs. residuals to ensure no clear pattern or trend, indicating a linear relationship between features and the target.
2. **Normality of Residuals**: Validated by plotting a histogram or Q-Q plot of residuals to confirm they follow a normal distribution.
3. **Homoscedasticity**: Verified by plotting residuals vs. predicted values to check for constant variance (no funnel shape or patterns).

These checks ensure that the model meets the core assumptions of linear regression.

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: <Your answer for Question 5 goes below this line> (Do not edit)

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

1. Temperature (temp)
2. Year (yr_1)
3. Weather Situation (weathersit_2)

These features show strong influence on bike demand based on their impact in the regression model.

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 is a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting straight line (in the case of simple linear regression) or hyperplane (in the case of multiple linear regression) that minimizes the difference between predicted and actual values.

Different steps involved are Hypothesis, Cost Function, Optimization and Assumptions.

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 is a set of four datasets that appear identical when analyzed using basic statistical methods like mean, variance, and correlation but differ significantly when visualized. Created by statistician Francis Anscombe, it highlights the importance of data visualization alongside statistical summaries.

Each dataset in the quartet has:

- The same mean and variance for both x and y variables.
- The same correlation between x and y.
- The same linear regression line.

However, when plotted, the datasets show vastly different patterns, such as non-linear relationships, outliers, or variations, emphasizing that relying solely on summary statistics can be misleading.

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, also known as the Pearson correlation coefficient, measures the strength and

direction of the linear relationship between two continuous variables. Its value ranges from -1 to 1:

+1 indicates a perfect positive linear relationship.
-1 indicates a perfect negative linear relationship.
0 indicates no linear relationship.

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 transforms features to a common range to ensure no variable dominates due to its magnitude. It's essential for algorithms like linear regression, k-NN, and SVM, which are sensitive to feature scales.

Normalized scaling (Min-Max scaling) transforms data to a fixed range, usually [0,1]:

$$Xnorm = X - Xmin / (Xmax - Xmean)$$

Standardized scaling (Z-score normalization) centers data around the mean and scales it to unit variance:

Normalization is suited for algorithms needing bounded data, while standardization is preferred when data has outliers or varying distributions.

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)

A VIF (Variance Inflation Factor) value becomes infinite when perfect multicollinearity exists, meaning one predictor variable is a perfect linear combination of one or more other predictor variables. In this case, the linear regression model cannot compute unique coefficients because the predictors are highly correlated, causing the denominator in the VIF formula to become zero, leading to an infinite VIF value.

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 (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution (e.g., normal distribution). It plots the quantiles of the observed data against the quantiles of the expected distribution. In linear regression, a Q-Q plot is crucial for

assessing the normality of residuals, which is an assumption for valid inference. If the residuals follow a straight line, it suggests that the data is approximately normally distributed. This is important because non-normal residuals can indicate problems such as model misspecification, outliers, or non-linearity, which could affect the regression results' validity and accuracy.