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)

-We observe that for the variable 'Season,' people are more likely to opt for bike sharing during the Summer, Fall, and Winter seasons compared to the Spring season.

-We observe that for the variable 'yr,' 2019 shows the highest usage compared to 2018.

-We observe that for the variable 'weathersit,' people tend to use the boom bikes slightly more during weather conditions such as mist, cloudy skies, and clear skies compared to light rain, light snow, and thunderstorms.

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)

-Avoid Multicollinearity: Including all dummy variables can create linear dependency. Dropping one ensures unique coefficients for the rest.

-Improved Model Stability: Reducing multicollinearity makes model coefficients more stable and interpretable.

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)

The 'temp 'and the 'atemp' variables exhibits a slightly stronger positive correlation with the target variable 'cnt' compared to the others.

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)

Linearity: Ensure linear relationship.

Check: Residuals vs. predicted values, scatter plots.

Normality of Residuals: Residuals follow a normal distribution.

Multicollinearity: Predictors not highly correlated.

Check: VIF, correlation matrix.

Model Fit Metrics: Assess overall model performance.

Check: R^2, Adjusted R^2

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)

Based on the final model, the top three features significantly contributing to explaining the demand for shared bikes are 'atemp,' 'yr,' and 'working day.'

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)

<Your answer for Question 6 goes here>

Linear regression identifies the relationship between a dependent variable and one or more independent variables.

Goal: Minimize the difference between actual and predicted values using the Ordinary Least Squares (OLS) method.

Assumptions: Linear relationship, constant variance of errors, normal residuals, independent errors, and no strong multicollinearity among predictors.

Types:

Simple Linear Regression: One independent variable.

Multiple Linear Regression: Two or more independent variables.

Evaluation: Use R-squared, Mean Squared Error (MSE), and residual analysis to measure model effectiveness.

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)

<Your answer for Question 7 goes here>
Anscombe's quartet consists of four datasets that share identical summary statistics (mean, variance, correlation, and regression line) but display strikingly different distributions when visualized.

Dataset 1: Exhibits a typical linear relationship.
Dataset 2: Shows a clear non-linear (curved) relationship.
Dataset 3: Includes an outlier that distorts the regression line.
Dataset 4: Strongly affected by one outlier, obscuring the lack of correlation in other points.

Key points: Always visualize your data, as statistical summaries alone can conceal critical patterns such as non-linearity, outliers, or influential points.

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)

<Your answer for Question 8 goes here>
Pearson's R measures the strength and direction of a linear relationship between two continuous variables, ranging from -1 to +1.
+1: Perfect positive correlation (both variables increase together).
-1: Perfect negative correlation (one variable increases as the other decreases).
0: No linear correlation.

The closer r is to ±1, the stronger the relationship. Positive r indicates a direct relationship, while negative r indicates an inverse 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)

<Your answer for Question 9 goes here>
What is Scaling? Scaling adjusts the range of features in a dataset to ensure all variables contribute equally to a machine learning model.

Why is Scaling Performed?
Ensures fair feature contribution.
Improves model performance and convergence speed.
Handles variance in feature magnitudes.
Essential for distance-based algorithms (e.g., k-NN, clustering).

Normalized vs. Standardized Scaling:
Normalization: Scales values to a [0, 1] range.
Standardization: Centers data to a mean of 0 and standard deviation of 1.
Output Range: Normalization typically [0, 1], while Standardization has no fixed range.

Usage:

Normalization: Ideal for bounded input algorithms like neural networks.

Standardization: Suitable for models assuming normal data distribution like SVM or PCA.

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)

<Your answer for Question 10 goes here>

VIF becomes infinite in the presence of perfect multicollinearity, where one variable is an exact linear combination of others. This occurs when R2=1 in the following formula VIF=1/(1-R2), leading to a zero denominator.

Implications:
Infinite VIF signifies redundancy among predictors, making regression coefficients unreliable.

Solutions:
Remove or combine collinear variables.
Employ feature selection or dimensionality reduction.
Apply regularization techniques like ridge regression.

Why Does This Matter?
Interpretation Issues: Infinite VIF highlights redundant information among predictors, leading to difficulties in interpreting individual variable effects.
Unstable Estimates: Multicollinearity inflates standard errors of coefficients, making them highly sensitive to small changes in data.

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)

<Your answer for Question 11 goes here>
What is a Q-Q Plot? A Q-Q (Quantile-Quantile) Plot is a graphical tool used to compare the distribution of observed data with a theoretical distribution, like a normal distribution. It plots the quantiles of the observed data against the quantiles of the expected distribution.
If the points align closely with the 45-degree reference line, it indicates that the data follows the expected distribution. Deviations from the line suggest departures from the assumed distribution, such as skewness, heavy tails, or outliers.

Use of a Q-Q Plot in Linear Regression: In linear regression, a Q-Q plot is used to check the normality assumption of the residuals (differences between observed and predicted values). This normality is crucial for valid hypothesis testing and confidence intervals.

Steps in Use:
Compute the residuals of the regression model.
Generate a Q-Q plot to compare residuals to a normal distribution.
Assess whether the points follow the 45-degree line:
Points near the line indicate normality.
Systematic deviations suggest non-normality.

Importance of a Q-Q Plot in Linear Regression:
Check Assumptions: Ensures residuals meet the normality assumption for valid statistical inferences.
Detect Outliers: Deviations at the extremes may indicate outliers.
Model Diagnostic: Helps evaluate if a transformation of the dependent variable or residuals

needed to improve model fit.