Name: Chidambaram Subramanian **Batch:** ML-C71

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 like season, weathersit, and mnth significantly impact bike demand. For **Example**:

1. **Season**: Bike rentals may be higher in summer than winter.
2. **Weathersit**: Poor weather conditions negatively impact bike rentals.
3. **Month**: Some months may have higher rentals due to holidays or temperature variations.

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** when creating dummy variables is important because it removes one of the dummy variables for each categorical feature, thereby reducing redundancy and avoiding multicollinearity between the newly created dummy variables.

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)

From numerical **features**, **temp** (or **atemp**) has the highest correlation with **cnt** since bike rentals increase with higher temperatures.

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)

The model's assumptions were validated using:

- Linearity Check: Residual plots show a random scatter.
- Multicollinearity Check: VIF values below 5 confirm low collinearity.
- Homoscedasticity Check: Residual variance remains constant across predictions.
- Normality of Residuals: Q-Q plots verify normal distribution.

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 Recursive Feature Elimination (RFE) and p-values, the most significant features are:

- temp (temperature)
- season (seasonality effect)
- weathersit (weather conditions)

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 statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The equation is:

$Y = B0 + B1.X1 + B2.X2 + ... Bn.Xn + E$

where

- B0 is the intercept,
- B1) are coefficients, and
- E represents error.

The model minimizes the sum of squared residuals using Ordinary Least Squares (OLS). Assumptions include linearity, independence, normality, and homoscedasticity. It helps in prediction and understanding feature impacts, widely used in economics, business, and science.

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 created by statistician Francis Anscombe in 1973 to illustrate the importance of data visualization. These datasets have nearly identical summary statistics (mean, variance, correlation, and regression line) but display drastically different distributions when plotted.

It highlights the limitations of relying solely on numerical statistics and emphasizes the need for data visualization to understand data patterns properly.

Each dataset has:

- Same mean (X, Y)
- Same variance (X, Y)
- Same correlation coefficient (~0.82)
- Same regression equation: y = 3 + 0.5X

However, their scatter plots reveal very different relationships, demonstrating that summary statistics alone are insufficient for data analysis.

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 correlation coefficient (R) measures the linear relationship between two variables. It

ranges from -1 to 1:

- +1 ⟶ Perfect positive correlation
- -1 ⟶ Perfect negative correlation
- 0 ⟶ No correlation

Correlations of all independent variables are measured against the dependent variable, i.e cnt.

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 numerical features to a similar scale, improving model performance.

- Min-Max Scaling: Rescales values between [0,1].
- Standardization (Z-score Scaling): Transforms data to have a mean of 0 and standard deviation of 1.

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)

The Variance Inflation Factor (VIF) quantifies multicollinearity in a regression model. A VIF value of infinity occurs when:

1. **Perfect Multicollinearity:** One independent variable is an exact linear combination of another (or multiple) independent variables.
2. **Duplicate or Highly Correlated Variables:** If two variables are identical or nearly identical, dividing by zero in the VIF formula causes an infinite value.
3. **Dummy Variable Trap:** If dummy variables for categorical features are incorrectly handled (e.g., not using drop_first=True), perfect correlation can occur.

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 Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution (usually normal distribution). It plots the quantiles of the sample data against the quantiles of a normal distribution.

Importance in Linear Regression:

- Checks Normality of Residuals – A key assumption in linear regression.
- Detects Skewness & Outliers – Deviations from the diagonal line indicate departures from normality.
- Validates Model Assumptions – If residuals are normally distributed, hypothesis tests and confidence intervals are valid.
