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)

1. Year 2019 has higher Bike Rental than the year 2019.
2. The fall season has a high number of bike rentals.
3. People prefer bike rental when weather situations is good.
4. May to Oct is having a high number of bike rentals.
5. People don't prefer much to rent bike on weekends as compare to weekdays.
6. People don't prefer renting bike 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: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True in dummy variable creation prevents multicollinearity, known as the "dummy variable trap." By dropping one category (the reference category), it avoids redundancy and ensures variable independence.
This also simplifies interpretation: each coefficient reflects the change in the dependent variable relative to the dropped category, making the model more intuitive.
Setting drop_first=True is thus both a practical and interpretive best practice in regression models.

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)

*With the pair-plot**, registered*** variable is having high correlation with the target ***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: <Your answer for Question 4 goes below this line> (Do not edit)

Using Residual Analysis, we validated the Linear Regression assumptions:
Error Normality: Histogram shows error terms are normally distributed around 0, confirming the assumption of normal error distribution.
Independence: Little to no relationship exists between residuals and predicted values, indicating independent error terms.
Homoscedasticity: Variance is consistent across both ends of the fitted line, meeting the homoscedasticity requirement.

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)

temp
weathersit_bad
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 is a supervised learning algorithm for predicting a continuous target variable Y using one or more predictor variables X.

It establishes a linear relationship represented by the equation $Y=\beta0+\beta1X$. where $\beta0$ is the intercept and $\beta1$ the slope for predictor X.

The algorithm relies on assumptions: linearity, independence, homoscedasticity, and normally distributed errors. It optimizes by minimizing the sum of squared residuals (Least Squares Estimation) to find the best-fit line, typically using Gradient Descent, an iterative process that updates parameters in the direction that reduces the cost function.

The model's performance is assessed through metrics such as R-squared, which shows the proportion of variance explained by the model, and Adjusted R-squared, which adjusts for the number of predictors. Additionally, Mean Absolute Error (MAE) and Mean Squared Error (MSE) provide measures of prediction accuracy. Linear regression's simplicity makes it widely used for predictive modeling and trend analysis.

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 collection of four datasets that have identical statistical properties but vastly different distributions and appearances when graphed. Their scatter plots reveal distinct patterns - one is linear, one is quadratic, one has an outlier, and one is constant. This highlights the importance of visualizing data, as relying solely on statistical metrics can be misleading. The quartet emphasizes the necessity of exploratory data analysis in understanding data behavior.

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 Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Its value ranges from -1 to +1.

A Pearson's R of +1 indicates a perfect positive correlation, meaning as one variable increases, the other also increases proportionally. Conversely, -1 indicates a perfect negative correlation, where one variable increases while the other decreases.

A value of 0 signifies no linear correlation. Pearson's R is commonly used to assess relationships between variables.

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 is the process of adjusting the range of independent variables or features in a dataset to ensure they contribute equally to model training. It is essential because many machine learning algorithms, especially those based on distance measurements (like k-nearest neighbors), are sensitive to the scale of input data.

Normalized Scaling (min-max scaling) transforms features to a fixed range, typically [0, 1], by subtracting the minimum value and dividing by the range.

Standardized Scaling (z-score normalization) centers the data around the mean with a standard deviation of 1, effectively converting it into a distribution with a mean of 0 and a variance 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) can become infinite when a predictor variable is perfectly collinear with other predictors in a regression model.

This means that one predictor can be expressed as an exact linear combination of others, leading to a situation where the R-squared value, used to calculate VIF, becomes 1.

Consequently, the formula for VIF, which includes (1−R2)(1 - R^2)(1−R2) in the denominator, results in division by zero, yielding an infinite VIF. This scenario indicates severe multicollinearity, suggesting that the predictor should be removed or re-evaluated to enhance the model's stability and interpretability.

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 quantiles of a dataset against those of a theoretical distribution, usually normal.

In linear regression, it assesses the **normality of residuals**. If the points closely follow the diagonal line, it indicates normally distributed residuals, validating model assumptions and ensuring accurate statistical inferences.
