Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset what could you infer about their effect on the dependent variable? (3 marks)

Answer: The analysis showed that categorical variables, particularly `season` and `weathersit`, significantly impact the demand for shared bikes. For example, bike rentals tend to increase during warmer seasons such as summer and fall and decrease in colder seasons like winter. Additionally, weather conditions play a crucial role, with clear and mild weather conditions favoring higher rental demands compared to days with heavy rain or snow. These variables are critical in predicting bike rental demand due to their direct influence on users' willingness to rent bikes under different environmental conditions.

2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

Answer: Using `drop_first=True` during dummy variable creation is important to avoid multicollinearity, a situation where one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. This option drops the first category level of the categorical variable, reducing the number of dummy variables by one. It prevents the dummy variable trap, a scenario where dummy variables are highly correlated, by ensuring that the set of dummy variables are not overly redundant, improving model stability and interpretation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: The pair-plot analysis among the numerical variables indicated that `temp` (temperature) has the highest correlation with the target variable `cnt` (total bike rentals). This suggests that temperature is a significant predictor of bike rental demand, with warmer temperatures likely encouraging more people to rent bikes.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: The assumptions of Linear Regression were validated through:

- Linearity: Checking scatter plots of residuals vs. predicted values to ensure a linear relationship.
- Homoscedasticity: Observing residuals spread evenly across the regression line to confirm constant variance.
- Independence: Leveraging the Durbin-Watson statistic to check for independence of residuals.
- Normality of Residuals: Using Q-Q plots to verify that residuals follow a normal distribution.

This comprehensive validation ensures that the linear regression model's assumptions hold, contributing to its reliability and predictive power.

5. Based on the final model which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: The final model revealed that the top 3 features contributing significantly to explaining the demand for shared bikes are:

- Temperature (`temp`): Demonstrating a positive correlation, indicating higher temperatures lead to increased bike rentals.
- Year (`yr`): Showing an upward trend in bike rentals over time, reflecting growing popularity or expansion of the service.
- Season (`season_spring`): Specifically, the presence of spring (compared to the omitted category) negatively affects bike rentals, likely due to variable weather conditions.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The basic form of a linear regression model is (Y = b0 + b1X1 + …… + bnXn + c), where Y is the dependent variable, Xi are independent variables, bi are coefficients to be estimated, and c is the error term. The goal of linear regression is to find the best-fitting line through the data points that minimizes the sum of the squared differences between observed and predicted values (residuals). This is typically achieved using the Ordinary Least Squares (OLS) method, which calculates the optimal coefficients that reduce the residual sum of squares (RSS).

2. Explain the Anscombe’s quartet in detail. (3 marks)**

Answer: Anscombe's quartet consists of four distinct datasets that have nearly identical simple statistical properties (mean, variance, correlation, and regression line), yet look very different when graphed. Each dataset highlights the importance of visualizing data before analyzing it and demonstrates how different data distributions can lead to the same statistical conclusions. Anscombe's quartet serves as a powerful reminder that statistical metrics alone cannot capture the nuances of data and that graphical exploration is essential for a comprehensive data analysis.

3. What is Pearson’s R? (3 marks)

Answer: Pearson's R, or the Pearson correlation coefficient, is a measure of the linear correlation between two variables, X and Y,ranging from -1 to 1. A Pearson's R value of 1 indicates a perfect positive linear relationship between variables, a value of -1 indicates a

perfect negative linear relationship, and a value of 0 suggests no linear correlation. It's used to assess the strength and direction of a linear relationship between two continuous variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling is a technique used to normalize the range of independent variables or features in data modeling. It is performed to ensure that each feature contributes equally to the analysis and helps algorithms converge faster. Normalized scaling transforms data to fall within a specified range, typically [0, 1], ensuring that outliers have less impact. Standardized scaling transforms data to have a mean of 0 and a standard deviation of 1, known as z-score normalization. This method is useful when data follows a Gaussian distribution and is essential for algorithms that assume data is centered around zero.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: An infinite value of Variance Inflation Factor (VIF) occurs when there is perfect multicollinearity in the data, meaning one or more independent variables can be exactly predicted from the others with no error. This perfect correlation among variables causes the denominator in the VIF formula to be zero, leading to an infinite value. Infinite VIF indicates that the concerned variables provide redundant information, which can distort the results and interpretations of regression models.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: A Q-Q (Quantile-Quantile) plot is a graphical tool to compare two probability distributions by plotting their quantiles against each other. In the context of linear regression, a Q-Q plot is often used to assess whether the residuals of the model are normally distributed, an assumption underlying many statistical tests used in linear regression analysis. If the points in the Q-Q plot fall approximately along a straight line, it suggests that the residuals have a normal distribution. This is important for the validity of the statistical inferences made from the regression model.