Assignment-based Subjective Questions

1. Analysis of Categorical Variables and Their Impact on the Dependent Variable:

From the analysis of categorical variables in the dataset, such as Season, Workday, and Year of Booking, it was evident that these variables significantly impact revenue generation. For instance, different seasons showed varying demand patterns for bike rentals, likely due to changes in weather and public preference. The Workday variable indicated a variation in bike usage between regular working days and weekends or holidays, reflecting the lifestyle patterns of potential customers. Moreover, the Year of Booking illustrated a trend in bike-sharing services' growing or declining popularity over time. This type of trend could be linked to broader market dynamics, such as technological advancements, increased environmental awareness, or economic factors.

2. Importance of Using drop_first=True in Dummy Variable Creation:

When creating dummy variables, using drop_first=True is crucial to avoid multicollinearity, a situation where one or more predictor variables in a regression model are highly correlated. This parameter drops the first category level of the categorical variable, reducing the number of dummy variables by one for each categorical feature. It helps prevent the “dummy variable trap,” a scenario where dummy variables are highly correlated (perfect multicollinearity). By dropping one dummy variable, we eliminate redundancy: the information that it would have conveyed is still present in the remaining dummies, ensuring that each piece of information is represented uniquely in the model.

3. Highest Correlation with the Target Variable in Pair-Plot Analysis:

The pair-plot analysis among numerical variables revealed that the temperature attribute had the highest correlation with the target variable, which in this context is likely the demand for shared bikes. This strong correlation suggests that temperature is a key factor influencing bike rental decisions. Typically, more favorable temperatures could lead to increased demand as people are more likely to prefer outdoor activities, including biking, under comfortable weather conditions. This insight can be particularly useful for predicting demand and planning inventory and marketing strategies accordingly.

4. Validating Assumptions of Linear Regression Post-Model Building:

After building the linear regression model on the training set, validation of its assumptions was undertaken primarily through statistical metrics like p-values, R-squared, and Adjusted R-squared values. P-values were used to assess the statistical significance of each predictor, ensuring that the variables included in the model had a meaningful impact on the dependent variable. The R-squared value indicated the proportion of variance in the dependent variable explained by the model, while the Adjusted R-squared provided a more precise measure by adjusting for the

number of predictors in the model. Additionally, residual analysis, checking for normality, homoscedasticity, and independence of residuals, played a critical role in validating these assumptions.

5. Top 3 Features Contributing Significantly to Bike Demand in the Final Model:

Based on the final regression model, the top three features contributing significantly to the demand for shared bikes were identified as Temperature, Weathersit_Light_Snow_Rain, and the pre-COVID Year. The prominence of Temperature aligns with the intuitive understanding that weather conditions significantly impact outdoor activities like biking. The variable Weathersit_Light_Snow_Rain likely captured the adverse effects of challenging weather conditions on bike demand. The ride counts significantly increased before the COVID factor could indicate changes in public behavior and preferences before the pandemic, potentially reflecting an increased inclination towards personal and outdoor modes of transport or shifts in work and travel patterns.

General-based Subjective Questions

1. Explain the Linear Regression Algorithm in Detail

Linear regression is one of the fundamental algorithms for machine learning. It's used to predict a continuous target variable based on one or more predictor variables. The idea is to find a linear relationship between these variables. We try to draw a line (in simple linear regression) or a plane (in multiple linear regression) that best fits our data points. The algorithm involves finding the line that minimizes the difference (error) between the predicted values and the actual values. This process, called Ordinary Least Squares, calculates coefficients (slope and intercept for simple linear regression) that define the line of best fit. The beauty of linear regression lies in its simplicity and interpretability, making it a great starting point for anyone diving into the world of predictive modeling.

2. Explain Anscombe’s Quartet in Detail

Anscombe's quartet consists of four different datasets with nearly identical simple descriptive statistics, yet have very different distributions and appearances when graphed. Each dataset in the quartet highlights different aspects of why visualizing data is as important as running statistical summaries. For instance, one dataset looks like a perfect linear relationship, another is a curve, and another has an outlier affecting the linear relationship. Relying solely on statistical summaries can be misleading, and it's crucial to visualize the data for a comprehensive understanding.

3. What is Pearson’s R?

In my studies, Pearson's R, or Pearson correlation coefficient, has been a key concept for understanding the strength and direction of a linear relationship between two variables. It's a number between -1 and 1 where 1 means a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 implies no linear correlation at all. It's particularly useful in exploratory data analysis to identify relationships or in feature selection to avoid too correlated features.

4. What is Scaling? Why is Scaling Performed? Difference Between Normalized and Standardized Scaling

Scaling is crucial in data preprocessing, especially when dealing with features that vary widely in magnitude or units. It’s about transforming data so that it fits within a specific scale, like 0-1. We do this because many machine learning algorithms perform better or converge faster when features are on a relatively similar scale. Normalization (Min-Max scaling) and Standardization (Z-score normalization) are two common methods. Normalization rescales data to a fixed range - typically 0 to 1. Standardization transforms data to have a mean of zero and a standard deviation of one. It’s more about making data comparable across features.

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

Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases if predictors are correlated. If VIF is infinite, it typically indicates perfect multicollinearity in the dataset. This happens when one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In such cases, the denominator in the VIF calculation becomes zero because the R-squared value is 1, leading to an infinite VIF.

6. What is a Q-Q Plot? Explain the use and importance of a Q-Q plot in linear regression

The Q-Q (Quantile-Quantile) plot is a graphical tool that compares two probability distributions by plotting their quantiles against each other. In the context of linear regression, we often use it to check the normality of residuals. If the residuals are normally distributed (a key assumption in linear regression), the points in the Q-Q plot will roughly lie on a straight line. Understanding the distribution of residuals helps validate the use of certain statistical measures and the reliability of the model, making Q-Q plots an essential tool in regression diagnostics.