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

Analysis of Categorical Variables The investigation into categorical variables revealed their varying impact on the dependent variable, which is the count of bike rentals. Variables like season and weather condition showed a notable influence. For instance, certain seasons with better weather conditions observed an increase in bike rentals, indicating users' preference to ride in comfortable weather. Conversely, adverse weather conditions saw a decrease in rental numbers, reflecting the natural aversion to biking under such circumstances.

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

Importance of drop_first=True in Dummy Variable Creation Utilizing drop_first=True in dummy variable generation is a strategic move to avoid the trap of multicollinearity, which can skew the outcomes of linear models. This approach omits one category from the dummy-encoded variables, ensuring the remaining variables are independent of each other, enhancing model stability and interpretability.

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

Through the pair-plot examination, the variable 'temperature' emerged as having the highest correlation with the target variable. This relationship underscores the intuitive understanding that weather conditions, particularly temperature, significantly influence biking preferences.

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

Post-model construction, assumptions of linear regression were validated through:
Residual Analysis: Ensuring residuals are normally distributed and homoscedastic.
Independence of Errors: Confirming no autocorrelation among residuals.
Multicollinearity Check: Using VIF (Variance Inflation Factor) to ensure predictors are not highly correlated. These steps are vital for confirming the model's reliability and predictive accuracy.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Based on the final model, the top three features significantly affecting bike demand included: Yearly Growth: Reflecting an increasing trend in bike-sharing popularity. Temperature: Highlighting the direct impact of weather conditions on rental behavior. Weather Situation: Indicating the influence of clear versus adverse weather on demand.

General Subjective Questions

1.Explain the linear regression algorithm in detail?

Linear regression is a fundamental statistical approach used to model the relationship between a dependent variable and one or more independent variables. The goal is to find a linear function that predicts the dependent variable values as accurately as possible. Key components include calculating the slope and intercept that minimize the sum of squared differences between observed and predicted values.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s Quartet Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. Each dataset illustrates the importance of visualizing data before analyzing it and the effect of outliers and other anomalies on statistical properties.

3. What is Pearson’s R?

Pearson's R, or the Pearson correlation coefficient, measures the linear correlation between two variables, providing insight into the strength and direction of their relationship. Its value ranges from -1 to 1, where 1 indicates a perfect positive linear correlation, -1 a perfect negative linear correlation, and 0 no linear correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling adjusts the values of numeric columns in the dataset to a common scale without distorting differences in the ranges of values. It's crucial for models that depend on the magnitude of variables. Normalization scales data between 0 and 1, while standardization scales data to have a mean of 0 and a standard deviation of 1, addressing outliers more effectively.

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

An infinite VIF indicates perfect multicollinearity among independent variables, meaning one variable can be perfectly predicted from the others. This scenario typically arises when variables are duplicated or when one variable is a precise combination of others.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a dataset follows a particular distribution, such as the normal distribution. In linear regression, it's used to validate the assumption of normality in the distribution of residuals, which is crucial for the reliability of regression analysis.