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?
    - Seasonal Variation: Higher demand during summer and fall, with winter showing reduced but still positive demand compared to spring.
    - Weather Conditions: Adverse weather conditions significantly reduce demand, highlighting the importance of favorable weather for bike rentals.
    - Monthly Trends: Certain months, particularly in summer and early spring, show higher demand, while winter months see a decline.
    - Weekly Patterns: Weekends generally see higher demand, likely due to recreational use.
2. Why is it important to use drop_first=True during dummy variable creation?
    a. Drop_first = True is very important while creating dummy variable because if not deleted, it will create a multi collinearity.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? - Registered users has the highest correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? Based on Linearity, Q-Q Plot, Multicollinearity, Durbin Watson test and Residuals, assumptions can be validated to verify that assumptions are good for training set.

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 coefficients and p-values from the model summary, the top 3 features contributing significantly to the demand for shared bikes are:

Year (yr): The coefficient is 2143.3551, with a very low p-value (0.000), indicating a strong positive influence on bike demand.

Fall Season (season_fall): The coefficient is 2554.6163, with a very low p-value (0.000), indicating a significant increase in demand during the fall season.

Summer Season (season_summer): The coefficient is 2369.4371, with a very low p-value (0.000), indicating a significant increase in demand during the summer season.

Subjective Questions

1. . Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting line through the data points that can predict the value of the dependent variable based on the values of the independent variables. Here's a detailed explanation of the linear regression algorithm:

a. Understanding the linear equation

   b. **Assumptions of Linear regression**
   c. **Finding the best fit line**
   d. **Evaluating the model**
   e. **Interpreting the coefficient**
   f. **Residual analysis**
   g. **Handling Multi collinearity**

2. Explain the Anscombe's quartet in detail.
Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics but appear very different when graphed. This highlights the importance of graphing data before analyzing it statistically. The datasets were constructed in 1973 by the statistician Francis Anscombe to demonstrate the effect of outliers and the importance of visualizing data.

3. What is Pearson's R?
Pearson's R, also known as the Pearson correlation coefficient, measures the linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation. The value of Pearson's R shows both the direction and strength of the relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Scaling in data preprocessing adjusts the range of data values to facilitate better performance of machine learning algorithms. Normalized scaling (Min-Max scaling) transforms data to a fixed range (e.g., 0-1), preserving the shape of the distribution but can be sensitive to outliers. Standardized scaling (Z-score normalization) centers data around a mean of 0 and a standard deviation of 1, making it less sensitive to outliers and suitable for algorithms assuming normally distributed data. The choice between them depends on the data distribution and algorithm requirements.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Infinite VIF values occur when there is perfect collinearity among variables in a regression model, meaning one variable can be exactly predicted from others. This leads to a situation where the variance inflation factor (VIF) calculation involves division by zero or near-zero, resulting in an infinite value. Perfect collinearity should be addressed by removing or combining variables to ensure reliable regression analysis.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a certain theoretical distribution, such as the normal distribution. In linear regression, Q-Q plots are crucial for checking the assumption of normality in the residuals. If the residuals (differences between observed and predicted values) deviate significantly from a straight line on the Q-Q plot, it suggests that the normality assumption might be violated, impacting the reliability of statistical inferences drawn from the regression model. Identifying deviations prompts further investigation or transformation of data to meet regression assumptions.