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?

Ans: The dataset includes categorical variables such as season, year, month, holiday, weekday, and weather situation. The analysis reveals the following insights:

- The fall season, specifically the month of September, attracts the highest number of active customers.
- The year 2019 witnessed a higher sales volume compared to 2018.
- The number of active users tends to decrease during holidays.
- Weather conditions significantly impact user activity. For instance, heavy rainfall results in no users, while a partly cloudy or clear sky corresponds to the maximum user count.

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

   Ans. If we don't use drop_first=True, the dummy variables could become correlated with each other, leading to redundancy. This is not desirable for our analysis.

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

   Ans. *atemp* and *temp* has the highest correlation with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans. One of the fundamental assumptions of a linear regression model is that the error terms should correspond to a normal curve, when plotted on histogram. Hence, this assumption is validated.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans. The top 3 features directly influencing the count are the features with highest coefficients. These are: Temp, Year (positively influencing) and snowy and rainy weather (negatively influencing)

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans: Linear regression is a predictive technique used to determine the relationship between variables and how a dependent variable is influenced by one or more independent variables.

   The process begins with data examination and cleaning through exploratory data analysis. The dataset is then divided into a training set (used to train the model) and a testing set (used to evaluate the model's accuracy).

   The collinearity of variables is checked, and the necessary variables are used to train the model. The R-value of the model and the p-values of dependent variables are evaluated. If necessary,

certain columns are dropped and the steps are reiterated (feature elimination) until a final model is obtained.

One of the conditions of linear regression is that the error curve must follow a normal distribution. After ensuring this, the model is tested with the test dataset. The resulting model can then be used to provide valuable insights or make predictions on datapoints within the model's range.

2. Explain the Anscombe's quartet in detail.
   Ans: Anscombe's Quartet comprises four datasets, each with eleven (x, y) points, created by statistician Francis Anscombe. Despite having similar descriptive statistics, the datasets appear different when graphed. They highlight the impact of outliers and influential observations on statistical properties. The quartet includes:

   - Dataset I: Well-fitted by a linear regression model.
   - Dataset II: Non-linear, hence not fitting a linear model.
   - Dataset III: Contains outliers not handled by a linear model.
   - Dataset IV: Has a high-leverage point influencing the correlation coefficient.
   - In essence, the quartet underscores the importance of data visualization before model building to avoid misleading results from regression algorithms and to identify anomalies.

3. What is Pearson's R?
   Ans: Pearson's R, or the Pearson correlation coefficient, quantifies the linear relationship between two variables. It ranges from -1 to 1, where 0 indicates no correlation, -1 a perfect negative correlation, and +1 a perfect positive correlation. It's calculated as the covariance of the variables divided by the product of their standard deviations. It's widely used in statistics and data analysis.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans: Scaling is crucial for a model to operate with suitable coefficient ranges. For instance, if a car's sale depends on two independent variables, price and months, the price range could be excessively high as there are only 12 months in a year. In such cases, appropriately scaling the price variable can prevent decimal errors in the model. There are two primary types of scaling:

   Normalized Scaling: This type of scaling transforms the data distribution into a Gaussian one. It doesn't have a fixed range and is commonly used in Neural networks.
   Standardized Scaling: The given example pertains to standardized scaling. Here, the variable values are compressed into a specific range to fit the model. This method ensures that all features contribute equally to the model's outcome.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: If there is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Ans: A Q-Q plot is a graphical method used to determine if different data sets originate from the same statistical distribution. This tool is especially useful in linear regression when we have separate testing and training datasets. In such cases, it's crucial to verify that both datasets come from the same distribution to ensure the validity of the model.