LINEAR REGRESSION ASSIGNMENT

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?

- The box plots for the year suggest that bike rentals peaked in 2019.
- During the fall season, according to the seasonal box plots, bike rentals saw a significant increase.
- The box plots for working days versus holidays and weekends indicate that bike rentals are higher on regular workdays compared to weekends or holidays.
- In terms of months, September stands out as the period with the highest bike rental numbers, as shown in the month box plots.
- On weekdays, particularly Saturdays, bike rentals are notably higher, as indicated by the weekday box plots.
- Furthermore, the weather conditions depicted in the weathersit box plots suggest that bike rentals are more frequent during clear, partly cloudy, or few clouds weather conditions.

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

Failure to use drop_first=True could lead to correlation among the dummy variables, rendering them redundant, which goes against the expectations of our analysis.

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

The scatter plot clearly shows that both 'atemp' and 'temp' have the strongest correlation with the target variable.

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

We observed that the error terms conform to a normal distribution when plotted on a histogram, which validates one of the fundamental assumptions of a linear regression model.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Temperature (temp): With a coefficient of '0.5499' temperaturen emerges as a crucial factor affecting bike hiring. Temp leads to a rise in bike hire by '0.5499' units. This underscores the importance temperature in bike rental planning.

Year (yr): The coefficient value of '0.2331' suggests that for each unit increase in the year variable, there is an increase in bike hire numbers by 0.2331 units. This indicates a positive correlation between the year and bike rental booking.

Light Snow and Rain (light snow and in): Notably, the coefficient value of '-0.2880' highlights the impact of weather conditions on bike rentals. Specifically, in comparison to Weathersit1, a unit increase in Weathersit

(indicating light snow and rain) decreases bike hire numbers by 0.2880 units. This emphasizes the significance of weather considerations in optimizing bike rental strategies.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is an algorithm used for predicting the relationship between variables by fitting a linear equation to observed data. The process involves several steps:

1. **Data Preprocessing**: The dataset is cleaned and explored through techniques like exploratory data analysis to understand the relationships between variables.
2. **Data Splitting**: The dataset is divided into a training set, used to train the model, and a testing set, used to evaluate the model's performance.
3. **Variable Selection**: Relevant variables are chosen based on their correlation and collinearity to train the model effectively.
4. **Model Training**: The linear regression model is trained using the training dataset, adjusting coefficients to minimize the error between predicted and actual values.
5. **Model Evaluation**: The trained model is evaluated using the testing dataset to assess its predictive accuracy. Metrics like R-squared and p-values are examined to gauge the model's performance.
6. **Feature Elimination**: If necessary, features may be dropped or adjusted to improve the model's accuracy.
7. **Testing Assumptions**: The model's assumptions, such as normality of errors, are tested to ensure its validity.
8. **Model Deployment**: Once validated, the model can be used to make predictions on new data points within the range of the model.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet refers to four datasets that have nearly identical descriptive statistics but differ significantly when graphed. This highlights the limitations of relying solely on summary statistics and emphasizes the importance of visualizing data.

3. What is Pearson's R?

Pearson's correlation coefficient, or Pearson's R, measures the strength and direction of the linear relationship between two variables. It ranges from -1 to +1, with values closer to ±1 indicating stronger correlations.

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

Scaling is the process of transforming data to fit within a specific range or distribution. It ensures that all variables contribute equally to the model and prevents numerical instability. Normalized scaling adjusts data to a Gaussian distribution, while standardized scaling compresses data into a specific range.

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

Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between independent variables, resulting in a perfect correlation.

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) visually compares two datasets to assess if they come from the same distribution. It is useful in linear regression to validate assumptions about the distribution of errors and ensure the model's reliability.