Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer 1:

1. Overall count of total rental bikes has increased from 2018 to 2019.
2. Fall season is the most favorable time for bike rentals in both years, with spring being the least favorable.
3. June, September, and August are the months with the highest rentals, while December, January, and February have the least.
4. Clear weather has the highest bike rentals, with none during heavy rain and very few during light rain.
5. Bike rentals are frequent from Thursday to Sunday, but the number drops between Monday to Wednesday.
6. Interestingly, the total rental count comes down for weekends and holidays in 2019 compared to 2018.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer 2:

To avoid multicollinearity. When creating dummy variables for categorical variables without dropping one category, it introduces multicollinearity. By dropping one category, we can avoid multicollinearity as the information from the dropped category is inherently included in the remaining ones.

Interpretability: Dropping one category (typically the reference category) makes the interpretation of the model coefficients more intuitive.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer 3:

Atemp with 63%, temp with 62.7%, and year with 56.9%.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer 4:

- We used residual analysis to validate the assumptions of linear regression.
- We plotted the histogram of the error terms and found that the "Error Distribution" is normally distributed around 0, which indicates that our model has handled the assumption of Error Normal Distribution properly.
- Assumption of Error Terms Being Independent: We see that there is almost no relation between residual and predicted value.

- Homoscedasticity: We can see that the variance is similar from both ends of the fitted line.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for the shared bikes?

Answer 5:

- Winter - 0.3333
- Fall - 0.3273
- Summer - 0.2922

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer 1:

Linear regression is a supervised machine learning technique. There are two types:

- **Simple linear regression:** This involves one predictor variable.
  - Example: $y = b_0 + b_1x$, where b_0 is the intercept and b_1 is the coefficient (slope) of x (the predictor).
- **Multiple linear regression:** This involves more than one predictor variable.
  - Example: $y = b_0 + b_1x_1 + b_2x_2 + ... + b\square x\square$, where b_0 is the intercept, and b_1, b_2, ..., $b\square$ are the coefficients (slopes) of predictors x_1, x_2, ..., $x\square$.

In linear regression, the target variable is continuous. The goal is to find a fitted line (or plane in case of multiple linear regression) that minimizes the sum of the errors between the target values and the predicted values.

Q2. Explain Anscombe's quartet in detail.

Answer 2:

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the limitations of relying solely on summary statistics. It highlights the importance of using data visualization to identify trends, outliers, and other crucial details that might not be apparent from summary statistics alone.

Q3. What is Pearson's R?

Pearson's R, also called the Pearson correlation coefficient, is a statistic that quantifies the strength and direction of the linear relationship between two continuous variables. It measures how well the data points of two variables fit on a straight line. Pearson's correlation coefficient ranges from -1 to 1.

- **Strength of the Relationship:**

    - The value of r indicates the strength of the linear relationship between the two variables.
        - **r close to 1:** Indicates a strong positive linear relationship. As the value of one variable increases, the value of the other variable also tends to increase.
        - **r close to -1:** Indicates a strong negative linear relationship. As the value of one variable increases, the value of the other variable tends to decrease.
        - **r closer to 0:** Indicates a weaker or even no linear relationship between the variables.
- **Direction of the Relationship:**
    - The sign of r indicates the direction of the relationship.
        - **Positive r:** Indicates a positive linear relationship.
        - **Negative r:** Indicates a negative linear relationship.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer 4:

Scaling is a data preprocessing technique in machine learning that transforms the features (variables) of a dataset to a common scale or range. It's done primarily to address issues arising from features having different scales, which can affect the performance of various machine learning algorithms. Variables can have different measurement units and scales. Some may have values in a small range, while others may have values in a much larger range. Scaling ensures that all variables contribute equally to the analysis or modeling process.

- **Normalized scaling:** This scales the data to a specific range, typically [0, 1]. It's achieved by subtracting the minimum value of the variable from each data point and then dividing by the range (maximum value minus minimum value). Normalized scaling is useful when you want to preserve the original range of the data and are not concerned about the distribution's shape.

- **Standardized scaling:** This transforms the data to have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean of the variable from each data point and dividing by the standard deviation. Standardized scaling is often preferred when the underlying distribution of the data is assumed to be normal and the algorithms are sensitive to feature scale or normality assumptions (e.g., Logistic Regression, Support Vector Machines).

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer 5:

A VIF (Variance Inflation Factor) of infinity can occur when there is perfect multicollinearity in the model. Perfect multicollinearity means that one or more independent variables can be exactly predicted from a linear combination of the other independent variables. This can happen if features are highly correlated with each other, making it difficult to isolate the independent contribution of each variable to the model. Having a VIF of infinity indicates a serious multicollinearity problem that needs to be addressed before proceeding with model analysis.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer 6:

A Q-Q plot (quantile-quantile plot) is a graphical tool used to compare the quantiles of your data to the quantiles of a theoretical distribution, typically a normal distribution. It helps you visually assess how well your data follows the expected distribution.

In linear regression, the Q-Q plot is a valuable tool for evaluating the normality assumption of the residuals (errors). A straight diagonal line in the Q-Q plot suggests that the residuals are normally distributed. Deviations from this line can indicate departures from normality, such as skewness or outliers. Identifying these deviations can help you decide if the model assumptions are met and guide you towards potential model improvements.