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?

Season: Fall Season seems to have high bike rentals. Spring seems to have low rentals

Year: 2019 have more rentals than 2018.

Month: April to October seems to have high rentals. Rental trends aparently increases till June, stays steady till October and decreases till december

Weathersit: The bike rentals are higher when the weather is clear or partly cloudy; with the median higher compared to other weather situations

Weekday: Weekday or Weekend doesnt seem to have any impact on the rentals. cnt is almost 4800

Holiday: Holiday or Working day doesnt seem to have any impact on the rental counts.

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

If we do not use drop_first = true, it leads to dummy variable trap. Meaning, n dummy variables are created predicting themselves known as multicollinearity. One dummy variable can be easy explained by the other dummy variables (n-1).

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

The variables **temp** and **atemp** have the highest correlation with the target, **cnt**, variable.

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

Looking for patterns in residual analysis. And. Error is observed to be normally distributed by plotting histogram

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

Temperature with 0.52, Year with 0.23, Light rain_Light snow_Thunderstorm with 0.28 negatively

General 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 by fitting a linear equation to the observed data. The primary goal of linear regression is to find the best-fitting line (or hyperplane, in the case of multiple independent variables) that minimizes the sum of the squared differences between the predicted and actual values. This best-fitting line is often referred to as the regression line. The linear regression model can be represented as:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n+\varepsilon$$

- Y is the dependent variable.

- $X_1, X_2, \ldots, X_n$ are the independent variables.
- β_0 is the y-intercept (constant term).
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables.
- ε represents the error term.

2.Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear quite distinct when graphed. This quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics. The four datasets have the same mean, variance, correlation, and linear regression line, but they exhibit dramatically different patterns when visualized.

3. What is Pearson's R?

Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is named after the statistician Karl Pearson who developed it.

The Pearson correlation coefficient takes values between -1 and 1, where:

r=1: Indicates a perfect positive linear relationship. As one variable increases, the other variable also increases proportionally.

r=−1: Indicates a perfect negative linear relationship. As one variable increases, the other variable decreases proportionally.

r=0: Indicates no linear correlation between the variables.

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

Scaling is a preprocessing step in data analysis and machine learning where the values of different variables are transformed to a standard range. The purpose of scaling is to bring all features or variables to a similar scale, making it easier to compare them and ensuring that no variable dominates the others due to differences in their original magnitudes. Scaling is particularly important for algorithms that are sensitive to the scale of the input features, such as gradient-based optimization algorithms used in many machine learning models.

Here are two common types of scaling: normalized scaling and standardized scaling.

1. Normalized Scaling (Min-Max Scaling):

2. Standardized Scaling (Z-score Normalization):

Key Differences:

Range:Normalized scaling brings values into a specific range, typically [0, 1]. Standardized scaling centers the data around zero with a standard deviation of 1.

Sensitivity to Outliers:Normalized scaling can be sensitive to outliers since it depends on the range of the data. Standardized scaling is less sensitive to outliers because it is based on the mean and standard deviation.

Use Cases:Normalized scaling is often used when the distribution of the data is not assumed to be normal and when the specific range is important. Standardized scaling is commonly used when the distribution is assumed to be normal or when robustness to outliers is desired.

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

The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to determine the individual effect of each variable on the dependent variable. The VIF quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated.

Reasons for Infinite VIF:

Perfect Multicollinearity: If two or more variables in your dataset are perfectly correlated (have a correlation coefficient of ±1), it leads to a situation where one variable can be perfectly predicted from the others. This results in an 2Ri2 of 1 and, consequently, an infinite VIF.

Perfect Prediction: In some cases, the values of one variable can be perfectly predicted using a linear combination of the other variables, leading to a situation of perfect multicollinearity and an infinite VIF.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a given sample of data follows a particular theoretical distribution. In the context of linear regression, Q-Q plots are often employed to check the normality assumption of the residuals (the differences between observed and predicted values) or to assess whether a set of residuals follows a specific distribution, usually the normal distribution.

Use and Importance in Linear Regression:

Normality Assumption: In linear regression, one of the key assumptions is that the residuals are normally distributed. Checking the normality of residuals is important because many statistical inference procedures, such as hypothesis tests and confidence intervals, rely on the assumption of normality.

Detecting Outliers: Q-Q plots can help identify outliers and deviations from normality in the tails of the distribution. Outliers can have a significant impact on the results of a regression analysis.

Model Validity: Assessing the normality of residuals through Q-Q plots contributes to the overall validity of the regression model. If the residuals are not normally distributed, it may indicate that the model assumptions are not met, and further investigation or model refinement may be necessary