Subjective Questions

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: From the analysis of the categorical variables from the dataset it could be inferred the bike rental rates are likely to be higher in summer and the fall season, are more prominent in the months of September and October, more so in the days of Sat, Wed and Thurs and in the year of 2019. Additionally, we could discern that bike rental are higher on holidays.

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

Ans: By setting drop_first=True, one dummy variable is dropped from the set of dummy variables. This ensures that the model can differentiate between the effect of each category and the intercept term. Additionally, it also helps to reduce the number of variables in the model, which can help to reduce overfitting and improve model performance.

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

Ans: The temp variable has the highest correlation with the target variable.

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

Ans: Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a predictor variable.

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 contributing significantly towards the demand of the shared bikes are the temperature, the year, and month.

General Subjective Questions

Explain the linear regression algorithm in detail.

Ans: Linear regression is a type of predictive modeling technique used to establish a relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as predictor variables or features). The goal of linear regression is to find the best-fit line that predicts the dependent variable based on the values of the independent variable(s). Overall, linear regression is a simple, yet powerful algorithm used for predicting continuous values. It is widely used in various fields such as economics, finance, and social sciences for modeling and forecasting purposes.

Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four datasets that have identical statistical properties but exhibit vastly different properties when graphed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphing data before analyzing it and to demonstrate the limitations of relying on summary statistics alone. Each of the four datasets in

Anscombe's quartet consists of 11 (x, y) pairs of data. When these datasets are graphed, they appear visually different from each other, despite having identical summary statistics such as mean, variance, and correlation coefficient.
The main purpose of Anscombe's quartet is to demonstrate the importance of data visualization in understanding the underlying patterns and trends in the data. It shows that relying on summary statistics alone can be misleading and can lead to incorrect conclusions about the data.

What is Pearson's R?

Ans: Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

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

Ans: Scaling is the process of transforming data so that it falls within a specific range. The goal of scaling is to standardize the range of features or variables in a dataset, so that they can be more easily compared and analyzed. Scaling is performed to avoid issues that arise when working with variables that have different units or different scales. When the variables are on different scales, some variables may end up dominating the others in the analysis, even though they may not be more important. Scaling can help to remove this bias and ensure that all variables are given equal importance in the analysis.
Normalized scaling, also known as min-max scaling, involves transforming the data so that it falls within a specific range, typically between 0 and 1. Normalized scaling preserves the shape of the original distribution, but it may not be appropriate for variables with extreme outliers.
Standardized scaling, also known as z-score scaling, involves transforming the data so that it has a mean of 0 and a standard deviation of 1. Standardized scaling preserves the shape of the original distribution and is more appropriate for variables with extreme outliers.

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

Ans: The value of VIF is infinite when there is a perfect correlation between the two independent variables. The Rsquared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped to define a working model for regression.

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

Ans: The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform, or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.