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 categorical variables have a very large effect on the target variable.

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

Ans. - drop_first=True helps in reduicng the extra column during the creation of dummy variables. Therefore it reduces the correlations crested among dummy variables.

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

Ans. - "temp" and "atemp" are the two numeric vairables which are highly correlated with the target variable "cnt".

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

Ans. - Distribution of the residuals should be normal and centered around 0. We tested this residual assumption by producing a distplot of residuals to see if they follow a normal distribution or not.

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 predictor variables that influence demand of the shared bikes according to our final model are:

Weathersit: A unit increase in the weather situation variable reduces the number of bike hires by 0.2828 units.

Temperature(temp): A unit increase in the temp variable increases the number of bike rentals by 0.5173 units.

Year(yr): A unit increase in the yr variable increases the number of boke rentals by 0.2324 units.

General Subjective Questions

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

Ans. - Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares.

e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.

2. **Explain the Anscombe's quartet in detail.**

Ans. - Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone. The quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.

3. **What is Pearson's R?**

Ans. - The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation. In layman's terms it asks if we can draw a line graph to represent the data.

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

Ans. - **Scaling** is a geometric change that linearly enlarges or reduces things. A property of objects or rules known as scale invariance is that they remain unchanged when scales of length, energy, or other variables are multiplied by a common factor. Scaling law, a law that explains how many natural phenomena exhibit scale invariance.

Scaling is performed because It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations. The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range.

The difference between normalized scaling and standardized scaling is that the values of a normalized dataset will always fall between 0 and 1. A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

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

Ans. - The VIF indicates how much collinearity has increased the variance of the coefficient estimate. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. A VIF less than 5 indicates a low correlation of that predictor with other predictors. A value between 5 and 10 indicates a moderate correlation,

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

Ans. - A scatterplot generated by plotting two sets of quantiles against each other is known as Q-Q plots. Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

The Q-Q plot answers the following questions :

- Does the two datasets have common location and scale?
- Does the two datasets have similar distributional shapes?
- Does the two datasets have similar tail behaviour?