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?

- Fall season has the highest mean of rental bikes
- People rent bike in Year 2019 more than in Year 2018
- Peak season of the year is in Summer (June, July, August, September)
- People rent more bikes on clear days.
- Holidays have a lower demand for bikes than non-holiday days.
- When the weather is clear and there are few clouds, demand for bikes is strong; nevertheless, demand is lower when there is light snow and light rain.
- Every weekday has a comparable demand for bikes.
- With regard to working days and non-working days, there is no discernible difference in bike demand.

2. Why is it important to use drop_first=True during dummy variable creation?
drop_first=True is crucial since it minimizes the extra column that is formed when a dummy variable is created. As a result, it lessens the correlation that dummy variables create.

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

Registered has the highest correlation with the target variable cnt.

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

You must make sure the model is accurate before making any forecasts. In order to achieve this, you must first carry out a residual analysis of the error terms. Error terms should follow normal distribution

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

According to the final model, the top three variables that significantly contribute to explaining the demand of the shared bikes are: temperature (0.4758), year(0.235) and windspeed(-0.1325)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

An explanation of the relationship between a dependent variable and an independent variable using a straight line is the goal of a simple linear regression model. The independent variable is also known as the predictor variable, and the dependent variables are also known as the output variables.

A linear regression equation is represented as: y = a+bx where b = Slope of the line, a = y-intercept of the line, x = Independent variable, y = Dependent variable

2. Explain the Anscombe’s quartet in detail.

Anscombe's Quartet is a set of four data sets that, while they appear to be almost equal in simple descriptive statistics, have certain anomalies that, if a regression model is developed, would deceive it. When displayed on scatter plots, they have significantly different distributions and show up differently.

The four datasets are as follows:

Dataset 1: This reasonably matches the linear regression model.

Dataset 2: Due to the non-linear nature of the data, a linear regression model could not be fitted to the data very successfully.

Dataset 3 displays the dataset's outliers that the linear regression model cannot account for.

The outliers in dataset 4 that the linear regression model cannot account for are displayed.

3. What is Pearson’s R?

Pearson’s R - known as Pearson correlation coefficient (r) is the most popular method for determining a linear connection is the. The intensity and direction of the link between two variables is expressed as a number between -1 and 1.

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

It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Also, 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. If scaling is not done, the algorithm will only consider magnitude and not units, which will result in inaccurate modeling. We must scale all the variables to the same degree of magnitude in order to resolve this problem.

Normalization usually entails rescaling the values into the [0, 1] range. Normalization is used when features are of different scales and it is really affected by outliers.

Data are often rescaled during standardization so that the mean is 0 and the standard deviation is 1. Standardization is not much affected by outliers and it is not constrained to a certain range.

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

VIF = infinity if there is a perfect correlation. This demonstrates that the two independent variables have an exact association.

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

Quantile-Quantile plots, often known as Q-Q plots, are scatter plots made by contrasting two different quantiles. The variable for which you are testing the hypothesis is represented by the first quantile, and the distribution itself is represented by the second.

You can visually compare the sample distribution of the relevant variable with any other potential distributions using a Q-Q visualization. When comparing the morphologies of two distributions, a Q-Q plot is used to show how characteristics like location, scale, and skewness are the same or different in the two distributions.