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?

Answer 1:

- The Month Jan is the "lowest demand month". Whereas, months from 'Jun to Sep' is the period when bike demand is high.

- The demand of bike increased in the year 2019 when compared to year 2018.

- Bike demand is less during holidays.

- There is no variation in bike demand during working non-working days.

- Bike demand is more on clear weather day, when the humidity & temperature is less.

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

Answer 2:

If we do not use **drop_first = True**, then 'n' dummy variables will be created, and these predictors ('n' dummy variables) are themselves correlated, this in turn, leads to Dummy Variable Trap.

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

Answer 3:

With target variable of 0.63, 'atemp' and 'temp' both have same correlation. This is the highest among all numerical variables.

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

Answer 4:

- By checking that the variance of the residuals (error terms) is constant across predictions [Homoscedasticity].

- By verifying that the R2 value for the predictions on the test data is almost same as the R2 value of the trained data.

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

Answer 5:

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:-

- temp

- weathersit

- season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer 1:

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.

When the number of the independent feature, is 1 then it is known as Simple Linear regression, and in the case of more than one feature, it is known as Multiple linear regression.

2. Explain the Anscombe's quartet in detail.

Answer 2:

Anscombe's quartet comprises of 4 data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.

3. What is Pearson's R?

Answer 3:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation.

It can take on values between -1 and 1. The further away 'r' is from zero, the stronger the linear relationship between the two variables.

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

Answer 4:

Scaling:

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 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.

Difference between normalized scaling and standardized scaling:

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?

Answer 5:

If R squared value is 1 [i.e., there is a perfect correlation], then VIF = infinity.

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

Answer 6:

The Quantile – Quantile plots or Q-Q plots are used to plot the quantiles of a sample distribution against quantiles of a theoretical distribution.