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: The demad of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 2018.

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

Answer: It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

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

Answer: Temprature has highest correlation with target variable count.

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

Answer:
1. Linear functional form: The response variable y should be a linearly related to the explanatory variables X.
2. Residual errors should be i.i.d.: After fitting the model on the training data set, the residual errors of the model should be independent and identically distributed random variables.
3. Residual errors should be normally distributed: The residual errors should be normally distributed.
4. Residual errors should be homoscedastic: The residual errors should have constant variance.

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

Answer:

weathersit_Light_Snow(negative correlation).
yr_2019(Positive correlation).
temp(Positive correlation)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:
Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given

independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.

2. Explain the Anscombe’s quartet in detail.

Answer:
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots

3. What is Pearson’s R?

Answer:
Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive

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

What?
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why?
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1.
sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

MinMax Scaling: x= x-min(x)/max(x)-min(x)

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

Standardisation: x= x-mean(x)/sd(x)

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some

information in the data, especially about outliers.

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

Answer:
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution