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 demand of bike is less in the month of spring when compared with other seasons

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.

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

ANSWER :- atemp and temp both have same correlation with target variable of 0.63 which 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 :- According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the feature and the target**.**

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

ANSWER :- The top 3 features contributing significantly towards the demands of share bikes are :-
a.) weathersit_Light_Snow(negative correlation)
b.) yr_2019(Positive correlation)
c.) temp(Positive correlation)

General Subjective Questions

1.) Explain the linear regression algorithm in detail.

ANSWER :- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

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 :- Its full form is Pearson correlation coefficient.is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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

ANSWER :-
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.

Nomalized scaling brings all the data in the range of 0 and 1.sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardized scaling replaces the values by their Z scores. It brings all data into standard normal distribution. 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 :- VIF becomes infinite only when a perfect correlation is achieved between two independent variables.In case of perfect correlation, we get R2=1 , which lead to 1/(1-R2) infinity

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

ANSWER :- 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