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: a. Demand is higher in the year 2019 than in 2018

b. Demand is higher in between month 5(May) and in month 10 (Oct) than in rest of the months

c. Demand is higher during No holiday than in holidays

d. Weekly distribution of demand does not vary much

e. Demand is higher when the weather situation is clear

f. Demand is higher during summer and fall season

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

Ans: 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?

Ans: Variable “temp” has the highest correlation with the target variable

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

And: I use the same regression model on the test set and calculated the R-squared. The R-square for the test came close to the R-square value for the training set.

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

Ans: i. Temperature (temp),

ii. Weathersit_3(light snow, light rain+….)

iii. yr (year)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: 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. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering, and the number of independent variables being used.

2. Explain the Anscombe's quartet in detail

Ans: 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?

Ans: It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

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

Ans: Scaling 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.

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.

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

Normalised scaling=(x-xmin)/(xmax-xmin)

Standarised 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).

Standarised scaling= (x-xmean)/SD(x)

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

Ans: Recall the formula of VIF,

VIF= 1/(1-R2).

When R2=1, then VIF tends to infinity. i.e. it occurs when there is perfect correlation between the two independent variables.

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

Ans: 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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.