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?

season: summer and fall season has the highest number of bookings and hence it can be a good predictor.

mnth: May, June, July, August, September & October have the highest number, hence it can also be a good predictor.

weathersit: clear weather has the highest number of bookings. It can also be a good predictor as it also shows some sort of trend.

holiday: Most of the bike bookings were happening when it is not a holiday. It can be the good predictor

weekday: weekday variable has medians between 4000 to 5000 bookings. It can either have a good or no effect on the dependent variable.

workingday: mostly bookings were made in the working day. So, a working day can be a good predictor.

yr: 2019 is a year with the highest number of bike bookings.

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

During dummy variable creation we use drop_first= True because it reduces the extra column created during the dummy variable creation and also reduces the correlation among the dummy variables.

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

There is a correlation of 0.63 between cnt-temp and cnt-atemp.

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

By performing residual analysis of train test i.e By Testing the normality of error terms. If they are following normal distribution our model is correct, Test of homoscedasticity- no visible patters would mean our model is correct, correct regression function, a test of uncorrelated error terms

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

   Light_snowrain – negative correlation
   Yr- positive correlation
   Temp- positive correlation

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   It is a machine learning alo that is based on supervised learning. It performs a regression task where a target prediction value is based on some independent variables. For the most part, we use a regression algorithm for finding out the relationship between variables and the prediction

2. Explain the Anscombe's quartet in detail.

   It is a group of data sets that was built in 1973 to highlight the importance of data visualization before using any alogrithm as these data sets are nearly identical in statistical view but if we scatter plot these data sets, we can notice the difference. And also, it fools the linear regression algorithm

3. What is Pearson's R?

   We can define it as a number to describe the strength of linear association between the variables i.e. the variable goes up and down together so it has a positive coefficient and if it goes in the opposite direction, it is negative.

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

   Scaling is used on the independent variable to normalize the data within a particular range as step of data- processing and it helps in speeding up the calculation.

   It is applied because most of the time data is highly varying in magnitude, units and range. And if scaling is not implied algo would only take magnitude in consideration and not units hence it would be termed as incorrect modeling. Thus, to avoid this problem we use scaling as it brings all the variables to same magnitude.

   Normalized scaling

It brings all the data in the range of 0 and 1. Sklearn.preprocessig.Minmaxscaler is used to implement it in python.

standardized scaling
It brings all data into a normal distribution which has mean zero and S.D. one. sklearn.preprocessing.scale is used to implement it in python.

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

   It means the two independent variables are perfectly correlated. Here, R*2 is 1 thus giving us 1/(1-r*2) infinity. And if we want to solve this problem, we need to drop one of the variables.

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

   Q-Q plots known as Quantile-Quantile plots are plots of two quantiles against each other. The main purpose of a Q-Q plot is to find out if two sets of data come from the same distribution, it is also used to compare the shapes of distributions, providing a graphical view of how properties are similar or different in two distributions