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: As per the final model I have build I can say that the categorical variable weather has significant effect, precisely **weathersit_3**. **seasons_4** is also a good predictor mnth_9 is again the good predictor. Basically we can see the coefficients of the variables in our model summary and based on the higher value of coefficient can say that this variable will be the good predictor.

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

   Answer: drop_first drops the first dummy variable from the list of new dummy variables, as if the variable has n categories than the dummy variables count should be (n-1)

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

   Answer: temp and atemp, both has 0.65.

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

   Answer: I will validate it on the rest 30% of the test set by these steps:

   1. Scaling the numeric variables as we did for the train set
   2. Adding the constant as we have added in the train set as the 'statsmodels' library didn't add constant itself
   3. Finally predicting the values for test set using the LR-train model which we have build using train set.

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

   Answer: Top 3 features are :

   1. yr
   2. temp
   3. mnth_9

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer: The linear regression algorithm is the one by which we can predict the target variable using the independent variable. It basically means that two variables x and y for a dataset will be linearly correlated.

2. Explain the Anscombe's quartet in detail. (3 marks)

   Answer: Anscombe's quartet is the combination of four datasets which have identical statistics. They seems to be identical but they have very different distributions and appear differently when we plot it on scatter plots.

It focus onto the visualization of the data set before applying any kind of algorithms on the data to build a model so that one can identify the various anomalies that are present in the data set by seeing the distribution.

3. What is Pearson's R? (3 marks)
   Answer:
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Answer: what : Scaling is a way to normalizing the data in a particular range(0-1).
   Why: Scaling is performed to the data where the values are varying in unit, range and magnitude, if we won't perform the scaling with the data having very high magnitude and different unit than the model prepared by this data will be useless. So to solve this issue we have to perform scaling with the data of high magnitude.
   Difference between normalizing scaling and standardized scaling
   - if we normalize the data, it will get compressed between 0 and 1. No datapoint can go beyond the boundary 0to 1, the min-max scaling takes care of outliers. It can be used for the data doesn't having Gaussian distribution

   - the spread of the data in the standardized version is more, the mean will be 0 but it will not be centered around zero, it won't take care of outliers, It can be used for the data having Gaussian distribution

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer: When we get R-square for a any set of variables as 1, that means it's a perfect correlation between two independent variables, because as per the formula of vif = 1/(1-R-square) this will lead to infinity as the denominator becomes 0.
   To overcome this we need to drop any of the variable as I did in my case in my python notebook

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Answer: before answering the Q-Q(Quantile-Quantile) plots, we should know what is quantile
   Quantile is the fraction or the limit where certain values falls below that quantile or fraction or limit
   As an example we have a continuous data like 1,4,6,8,10,12,13,15,18, 30,35,40,42,50
   If we say 1st quantile is 15 that means all the values less than 15 fall in this quantile.

   Now we can say that Q-Q Plots are plots of two quantiles against each other.

   If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.
   If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location,
scale, and skewness are similar or different in the two distributions.