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? (3 marks)

Ans. 1. Spring has the lowest cnt
2. cnt is highly dep on weathersit (people don't prefer biking in light rain)
3. cnt is also independent of workingday & weekdays.
4. on holidays cnt is low.(i.e people prefer biking on workingdays)
5. there is significant rise in cnt in 2019 as compared to 2018.

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

Ans. Dropping first column is to reduce the number of dummy variable in multi levelled, categorical data; its important because more variables causes difficulty In algorithm/misfits or even overfitting issues. Also it doesn't effect the data content like for month, if jan to nov are false it is clear that it is December. Therefore separate olumn for dec is not necessary.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans. 1. temp is positively related to the cnt
2. windspeed is negatively related to cnt
3. humidity is not specifically related but people don't prefer biking when humidity is very low.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans. We will predict the dependent variable for the test set based on the model we built and check the R-sq & other parameters to be similar to that of the train set. And calculating the residuals of test set.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans. Hum-
Temp-
Windspeed-

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans. LR algorithm is as follows:

- We categorize the data into categorical and Numerical data, and convert all the data into meaningful data by converting string or other types of categorical data into Numerical (dummy variable) data.

- We visualise the data points wrt the dependent variable, to understand the relations and decide the model building process.
- Then we distribute the given data into test and train data sets.
- Then we scale the data points of the training data set, to get their coefficients comparable to stabilize the model hence created. This is done mostly using Minmax Scaling. This is done after splitting the datasets so that authenticity of the created model could be verified, by checking the coefficients of test data set.
- Then we create the model, including a column of all 1's to indicate the column of constant.
- We check the developed model's parameters like R-sq, VIFs and p-values, and then modify the model to achieve best fitted model.
- Then we check the residuals, it should come normally distributed around 0.
- Then we do the predictions on the test data with the model we trained on the training set.
- And check the parameters to be as close to the parameters achieved in the training, to verify the model to be a good one.

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

Ans. It has described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

This was to illustrate the importance of visualisation of the variables, and how relying on calculations can give inappropriate mis-fitting results.

3. What is Pearson's R? (3 marks)

Ans. Pearson's R is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Therefore, infinite or large VIF means that multicollinearity is very high.

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

Ans. *Q*uantile-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