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?

The analysis of categorical variables shows the following:

1) The fall season has the highest number of riders.
2) There was a much higher number of riders in 2019 than 2018.
3) The spring months / season has the lowest footfalls.
4) The riders keep increasing month-to-month from Jan and peak during fall season.
5) There is a sharp drop in the numbers at the onset of winter
6) Good weather days has a significant higher number of riders.
7) Working day/ Holiday doesn't seem to have any impact.

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

It helps in reducing the extra column used during dummy variable creation and thus avoiding multicollinearity by having an extra column when (n-1) columns are sufficient to represent n columns.

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

Temp variable seems to be having the highest collinearity, which is confirmed in the heatmap.

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

Plotted the y_test and y_pred values and also got the

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

Year, temperature and weathersit_good seems to be biggest contributors towards explaining the demand