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 - The demand of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 2018.

The demand is equally distributed among the weekdays

The demand is lower when the weather is bad

Data inferred Atemp and Temp are more or less the same

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

Answer - Assuming you have a column with 3 unique values dropping your first categorical variable is possible because if every other dummy column is 0, then this means your first value would have been 1. What you remove in redundancy.

For example a house is furnished semi-furnished or non furnished in the column called Furnished, we make use of drop_first to create dummy variables with only 2 columns (semi furnished and Furnished) if a house is either of these two then it is automatically assumed that it is not unfurnished hence an unfurnished value would be 0 for semi furnished and 0 for furnished as well hence we can remove the column called unfurnished.

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

Answer – Looking at the pair plot it is evident that Temp are the highest correlated to the target variable (cnt) - 0.63

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

- we then performed the prediction on the training data using a regression model and found the y train pred value, subtracting the y train value to the y train pred value we found that the Errors are normally distributed here with mean 0. So everything seems to be fine
- We also found the r_sqared value using this test analysis and noticed it was very close to the value from our final OLS model
- We then created a y test pred and found that the r2_square value was almost similar to the values from the training data which suggested a training set of 70-30 with random state set to 42 percent was an ideal fit

- R2 value for predictions on test data (0.7987479140295531) is almost same as R2 value of train data(0.7909321666618335). This is a good R-squared value, hence we can see our model is performing good even on unseen data (train data)
- we lastly checked for homoscedasticity and found that the model passed the requirement.

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

weathersit_bad -1579.615781**holiday** -733.95825 and **mnth_mar** 476.054416 are the top three features contibuting to this model with the parameters.