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?

   Please find the below analysis on analysis of the categorical variables from the dataset using box plots:

   - June, July, Aug, Sep and Oct had high demands for the bike
   - Holiday had less demand in average, maybe because less people going out
   - 2019 had a significant demand compared to 2018
   - Summer and Fall had more demand then Spring and Winter

2. Why is it important to use **drop_first=True** during dummy variable creation?
   - It helps in reducing extra variable creation during dummy variables creation
   - It helps in reducing correlations among dummy variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   - We can see temp and atemp having the highest correlation with the target variable

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

   Below points were taken care for Linear Regression:

   - Checking p-value for each feature which should be less than 0.05
   - Checking VIF value for each feature should be definitely less than 10 but between 5 and 10 also be checked (Valid VIFs are generally less than 5)
   - Error rates should be normally distributed

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

   Top 3 features contributing are:

   - temp
   - winter
   - light snow rain

General Subjective Questions

1. Explain the linear regression algorithm in detail.