## 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?

**Ans**: Below information can be inferred from Box plots drawn

- The more bikes are on rent during weekdays or working days compared to holidays and week ends
- More bike rents during the 'fall' season compared to other seasons
- Year 2019 has more bike rents compared to year 2018
- More bike rents during July and September months compared to other months. However, July has slightly more bike rent compared to September
- Saturday has more bike rent compared to other days

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

**Ans**: It helps in reducing the extra column created during the dummy variable creation and thus reduces the corelation created among the dummy variables

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

**Ans**: 0.95, cnt variable has highest (0.95) correlation with registered. Other than this

- temp has highest correlation (0.63) with cnt
- cnt has highest correlation (0.67) with casual

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

**Ans:** By creating the scatter plot (X vs Y). If the data points falls on the straight line in the graph then it indicates that there is linear relationship between dependent and independent variable. This can be found in the last scatter plot drawn in the bike case study.

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

**Ans:** Below are the top 3 features contributing significantly towards demand of the shared bikes

- Temp
- Windspeed
- Year

## General Subjective Questions

1. Explain the linear regression algorithm in detail

Ans. It performs the task to predict dependent variable value using the independent variable. It tries to find the the linear relationship between dependent and independent variables.

2. Explain the Anscombe's quartet in detail?
   Ans: It consists of 4 datasets having identical simple statistical properties but they appear quite different when plot. It demonstrate the effect of outliers on statistical properties.

3. What is Pearson's R?
   Ans: Its also known as bivariate correlation. It measures the linear correlation between two variables. The range of it is -1.0 to 1.0. It can not differentiate between two variables dependent and independent variables nor can not capture the nonlinear relationship between two variables

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   Ans: It is data pre-processing and is applied to independent variable in order to normalize the data within the particular range.

   Normally, data set which is collected has varied magnitude. If we do model without scaling then it would be incorrect modelling as data set has varying magnitude. To solve this need to do scaling to bring all the data or variables to the same level of magnitude.

   This does not impact the p-value, F-statistics ,R-squared values but it affect only coefficients.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans:
   If there is perfect correlation between two independent variables then VIF is an infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
   Ans: It is the plot of the quantities of the first data set against the quantities of the second data set. This tool helps to assess if the set of data probably came from some theoretical distribution, i.e. normal or uniform distribution.