## 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?
    a. year - 2019 seems to have higher count than 2018
    b. month - count is low at start of the year, gets increased by the mid year and gradually decreases by end.
    c. holiday - on holidays counts seems to have broader distribution but weekdays has higher mean value.
    d. weekday - all weekday have same mean of total count.
    e. weathersit - if the weather is good it has more counts.
    f. season - fall and summer has high counts. fall season has bit higher than summer

2. Why is it important to use drop_first=True during dummy variable creation?
    a. When there are n level categorical values, n-1 dummy variables has to be created. As the remaining one is exceptionally understood.
    b. Eg: there are 3 level of category values good, bad, very bad. Good and bad dummy variables are created. When good and bad are 0 ,0 respectively, it is understood that very bad is 1.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. Using f-statistics. For a good model, f statistics is high and p(F) is very low
    b. Rsquared-value suggests the significance of the model
    c. P value suggests the significance of the variable
    d. Vif gives the multicollinearity

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. Year, Temperature and winter season

# General Subjective Questions

1. Explain the linear regression algorithm in detail.
    a. Linear regression assumes that the dependent variable is linearly correlated to the independent variables.
    b. Y=c+mx
    c. It is used to find the independent variables that are responsible for change in dependent variable.
2. Explain the Anscombe's quartet in detail.
    a. Four similar statitics when seen in descriptive form but looks different when graphed.
    b. i.e their distributions are very different.
3. What is Pearson's R?

    Pearson's r is the covariance of two variables divided by their standard deviation.
    It ranges from -1 to +1.
    It is used to identify the patterns in the data set, linear coefficient.
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Scaling is performed before building a model.
    b. It ranges the value between 0 and 1
    c. Normalized scaling is X-xmin/xmax-xmin, value lies between 0 and 1
    d. Standardized scaling is x-mean/sigma, value lies between -1 and 1
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. VIF is used to identify the multicollinearity among the independent variables.
    b. Higher the value of VIF, higher the relationship among the variables.
    c. For good model VIF value of all the variables should be <5.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
    - Q-Q plot or Quantile Quantile plot, is used to find whether the two variables comes from same distribution.
    - X axis , plotted for quantiles of first variable
    - Y axis, plotted for quantiles of second variables.
    - It is useful when the data is continuous and we have large data set.