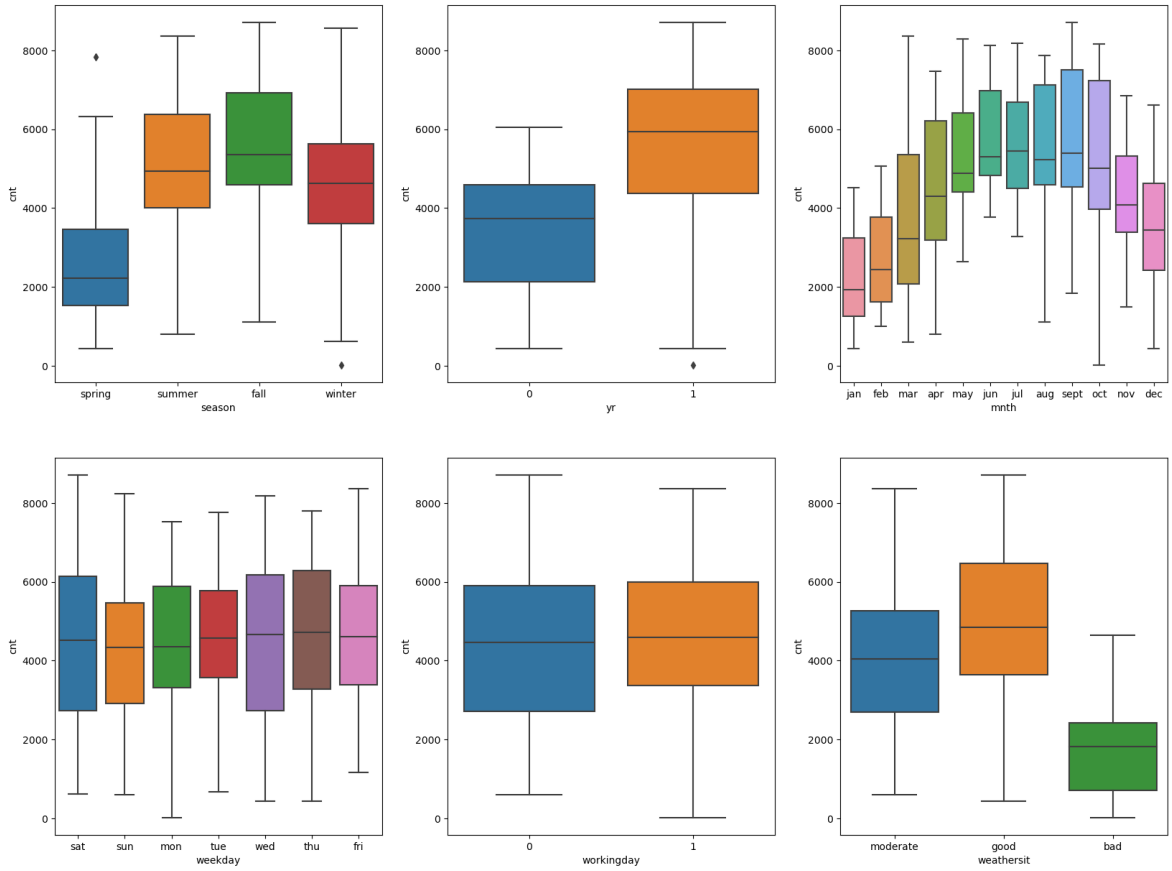
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   Ans:



   Based on the above graph, the categorical variables, have a significant impact on cnt variable.

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

   Ans: Dummy variables are used to map n variables to n-1 columns. The value in the dummy variables is always 1 or 0. For example, if there are 2 categorical variables, we can use just one column to infer both. 1 could indicate the one category, while 0 would indicate the second category. With this we can generalize and define that for a categorical column with n values, we need n-1 columns. For this reason, we are dropping the first and creating only n-1 columns.

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

   Ans: temp and atemp have the highest correlation with respect to cnt variable (which is the dependent variable in this case)

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

   Ans:

   1. Low p-values – Indicate the significance level of the features.
   2. Normal Distribution of errors which is a must for a Linear Regression Model
   3. Lower VIF for the all the independent variables
   4. Handling multi-collinearity

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

   Ans: 1. Year, Temp and windspeed - For all these, the p-values are low.

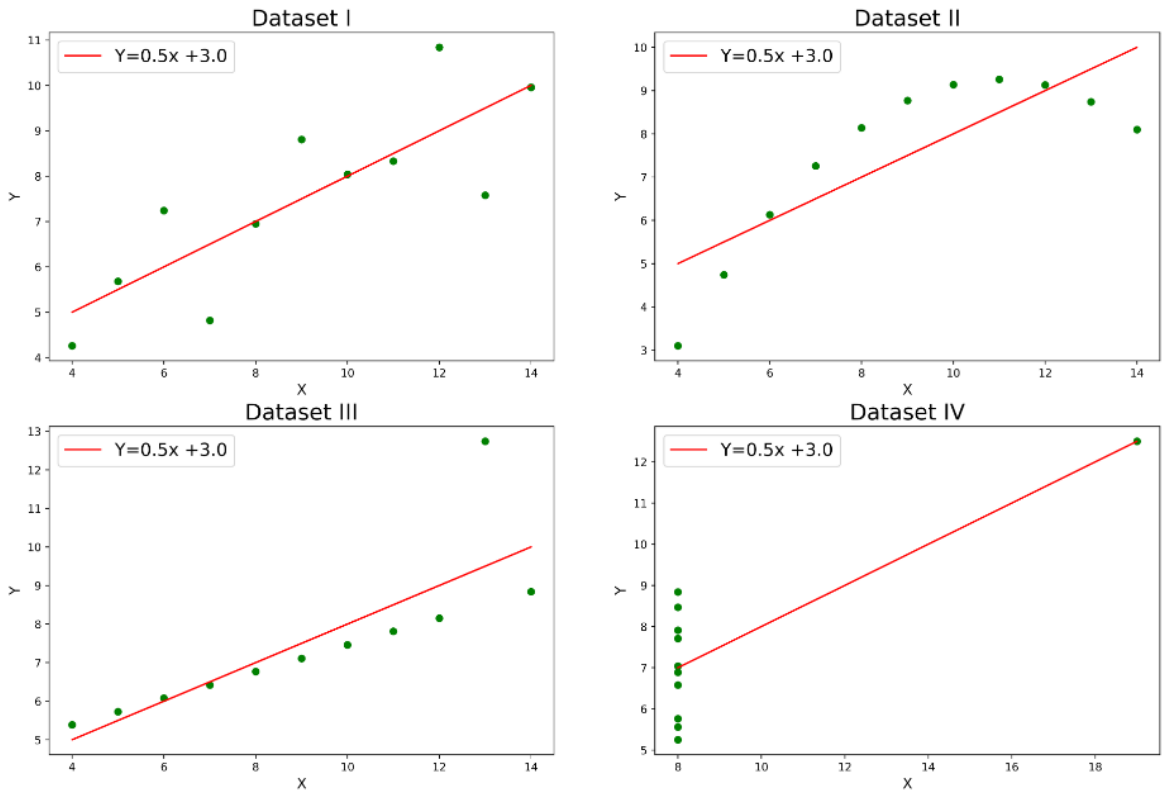
General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans: Linear Regression is used to predict the value of the dependent variable, based on one or multiple independent variables. If the prediction is through one variable, then it is as Simple Linear Regression, otherwise 'MLR – Multiple Linear Regression". The goal of linear regression is to derive the best slope and the co-efficient values which results in a best fit line with minimal errors. The algorithm also states that the residual value (errors) has a normal distribution with the mean always at 0.

2. Explain the Anscombe's quartet in detail.

   Ans: This is a dataset that has the same mean and standard deviation. Though these values are the same, they are qualitatively different. This basically compares 4 datasets and because of the same mean and standard deviation, it is misunderstood to be same features. However, the scatter plots of each of these indicate that these are different parameters/variables. The below graph shows a scatter plot which is different.



3. What is Pearson's R?

Ans: It is a coefficient that measures linear correlation between 2 data sets. It is a value between 0 and 1 and -1 to 0. 0-1 indicates positive correlation and negative values indicate negative correlation.

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

Ans: When numerical data are in different formats, using the same can cause higher weigh values and ignore the other critical parameter. This results in an incorrect model. We need to scale all the numerical values so that it fits in the same scale (or measure). There are 3 types of scaling – Min-max Normalization, Mean Normalization and Standardization

Min-Max Scaling – Scales the numbers from 0 to 1 or -1 to 1.

Formula is

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

Mean Normalization – Here the mean value is used for normalization. Formula is

$$x' = \frac{x - \bar{x}}{\max(x) - \min(x)}$$

Standardization - The general method of calculation is to determine the distribution mean and standard deviation for each feature. Then we subtract the mean from each feature. Then we divide the values (mean is already subtracted) of each feature by its standard deviation. Formula is

$$x' = \frac{x - \bar{x}}{\sigma}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: VIF stands for variance inflation factor. The formula to compute VIF is (1/1-R2). If the computed R2 is 1, then VIF will become infinity. The lower VIF, the better. VIF 10 and above is considered high. 5 and above should also not be ignored.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: Q-Q (quantile-quantile) plots play a vital role in graphically analyzing and comparing two probability distributions by plotting their quantiles against each other. If the two distributions that we are comparing are exactly equal, then the points on the Q-Q plot will perfectly lie on a straight-line y = x. In linear regression, we can use this plot to see if the train and the test set have the same distribution or not.