Name : Nurshafizah Mohd Kamil

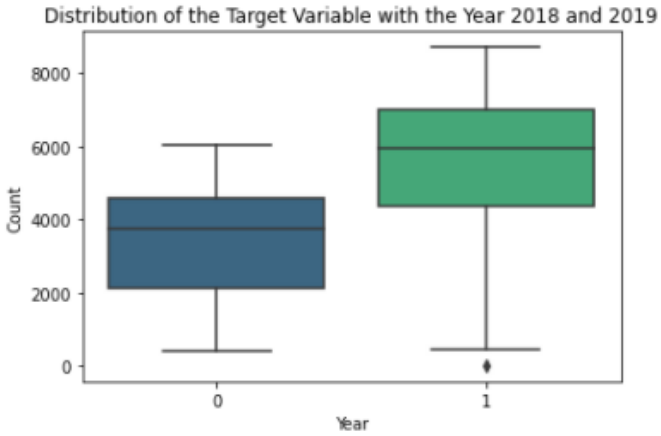
Email : shafizah87@gmail.com

Assignment-based Subjective Questions

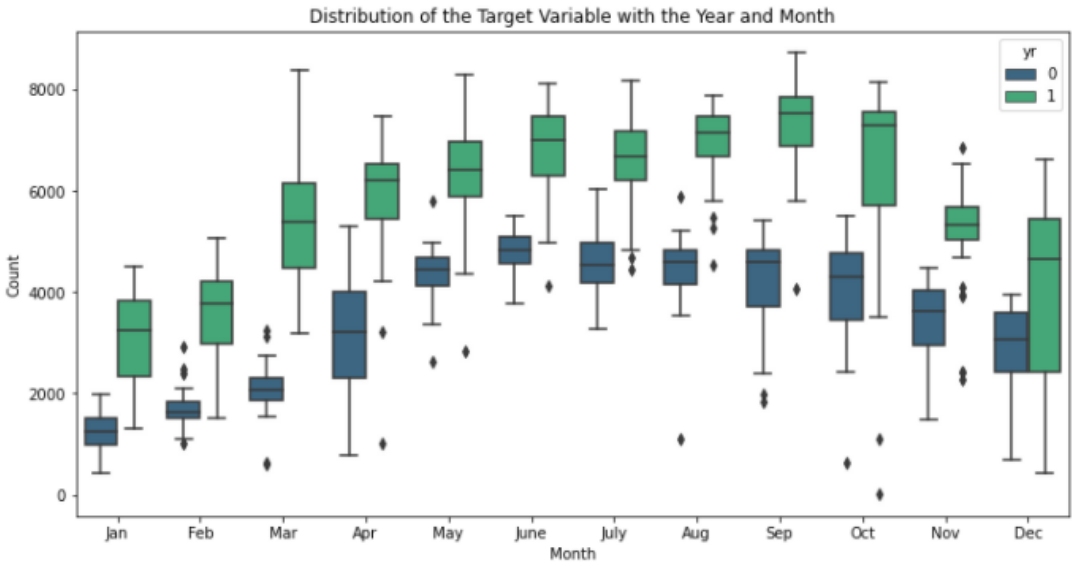
1. *From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?*

The Categorical Variables from the dataset are year, month, season, weathersit, workingday and weekday.

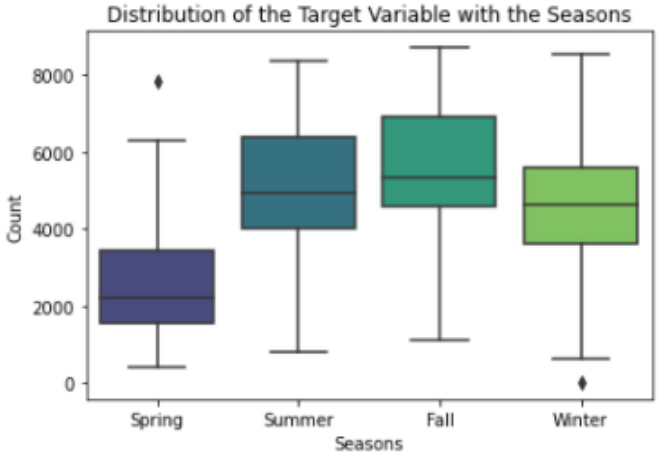
a. Year/ yr – The demand for the year 2019 is greater that the demand in 2018.



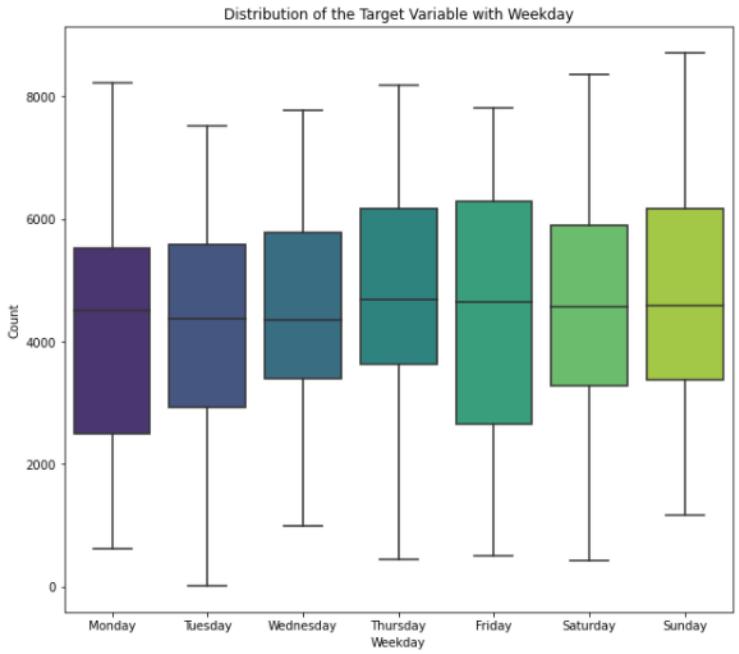
b. Month / mnth – The demand started slow for the month of Jan and Feb. However it is increasing starting from the month of Mar, then subsequently increasing to April to Oct and started reducing starting from the month of Nov and Dec.



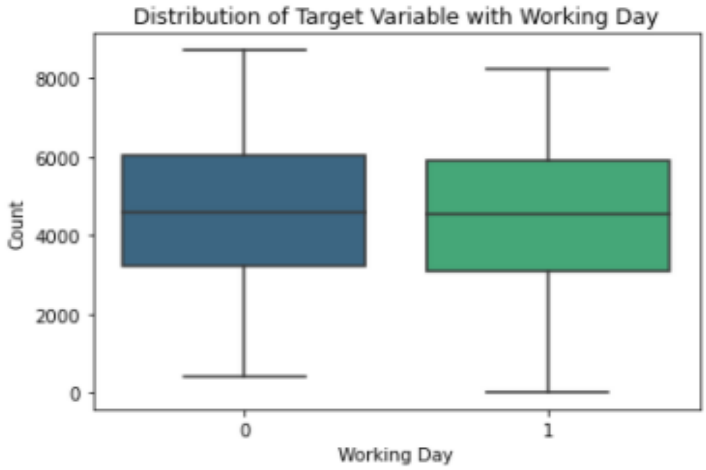
c. Season – The bike rental is greater during the Fall seasons, followed closely by Summer then Winter and the least on Spring.



d. Weekday – There no significant differences for the bike rentals during the weekday, however, a slight increase in the demand can be seen on Friday, Thursday, then followed by Sunday.



e. Workingday – For both workingday and non-workingday there not much differences in the demand. But nonworking day is greater than working day by very small differences.

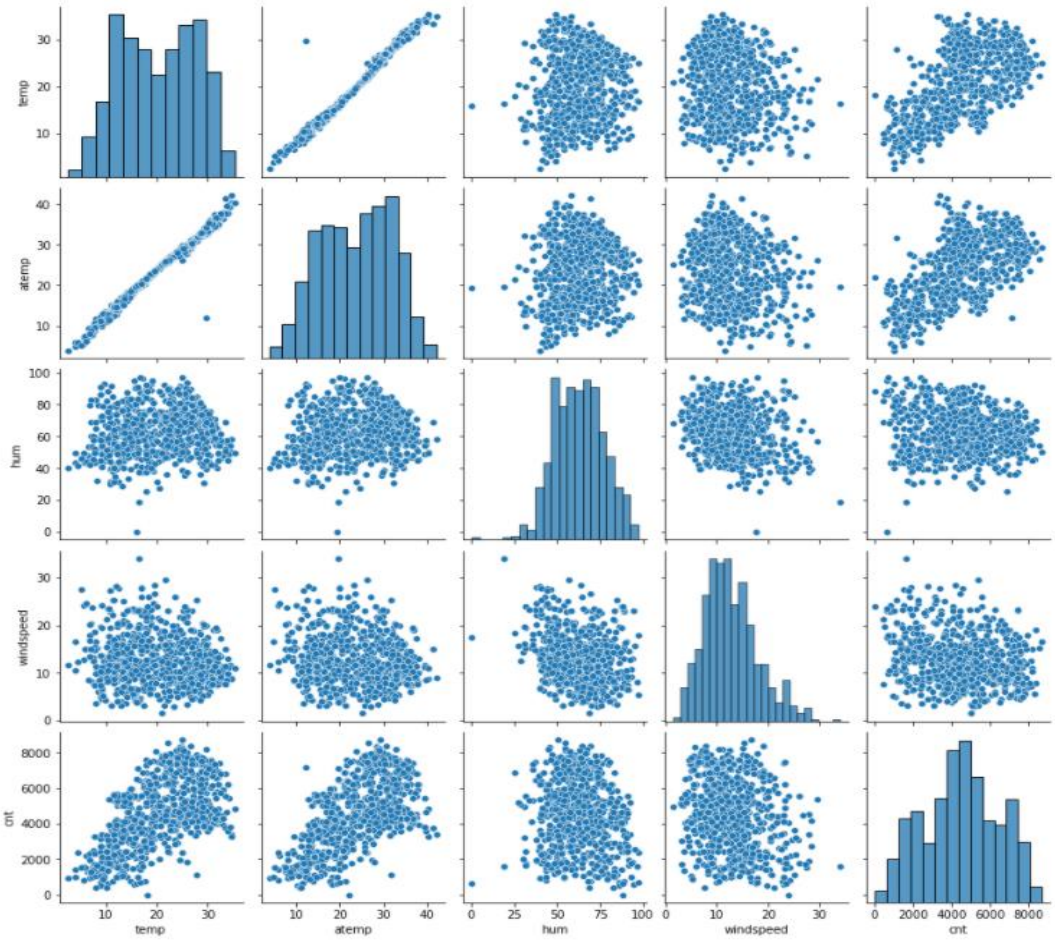


2. ***Why is it important to use drop_first = True during dummy variable creation?***

During the creating of Dummy variable, it is very important to use drop_first = True to drop the reference column after encoding. The default value for this parameter is drop_first = False. On top of that, by using drop_first= True will reduce Multicollinearity in the dataset. Multicollinearity occurs when 2 datasets are correlated with each other and it can be measured by using variance Inflation Factor (VIF).

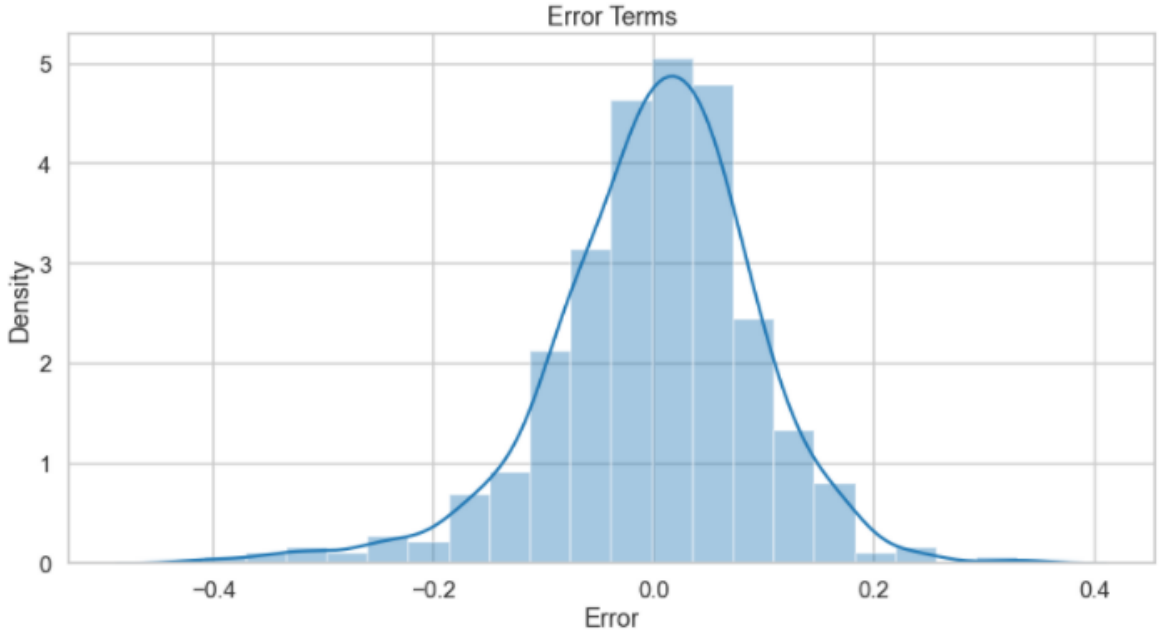
3. ***Looking at the pairplot among the Numerical variables, which one has the highest correlations with the target variable?***

Temp and Atemp are 2 variables which are highly correlated with the target variable(cnt).



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

   To validate the assumptions of the Linear Regression, we can validate by checking the Residual and plotting a distplot. The residual of the regression shall follows the normal distribution and the means shall be centered towards 0.



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

   The top 3 features contribute significantly towards the demands are:

   a. Temperature/ Temp – 0.4515
   b. Year / Yr – 0.2340
   c. September /Sep - 0.0577

General Subject Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical regression method which shows the relationship between the continuous variables and being used for predictive analysis. It shows the linear relationship between the X-axis (independent variable) and the y-axis (dependent variable). Linear regression is as follows:

$$y = mx + c$$

When there is only 1 variable and 1 dependent variable, we shall be referring to the Simple Linear Regression. The formula for the Simple Linear Regression is:

$$Y = b_0 + b_1x_1$$

y : dependent/ predicted variable

b_0 : y-intercept (constant)

b_1 : regression coefficients representing the change in y relative to one-unit change in x_1 and x_2 respectively.

When there are more or multiple independent variables involved, then it will be predicted using the Multiple Linear Regression. It generally explains the relationship between 2 or more predictor variables. Here we will choose our Best Fit Line based from the equation below.

The equation for the Multiple Linear Regression is follow:

The multiple regression equation explained above takes the following form:

$$y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + b_0$$

y : dependent/ predicted variable

b_0 : y-intercept (constant-term)

b_1 and b_2 : regression coefficients representing the change in y relative to one-unit change in x_1 and x_2 respectively.

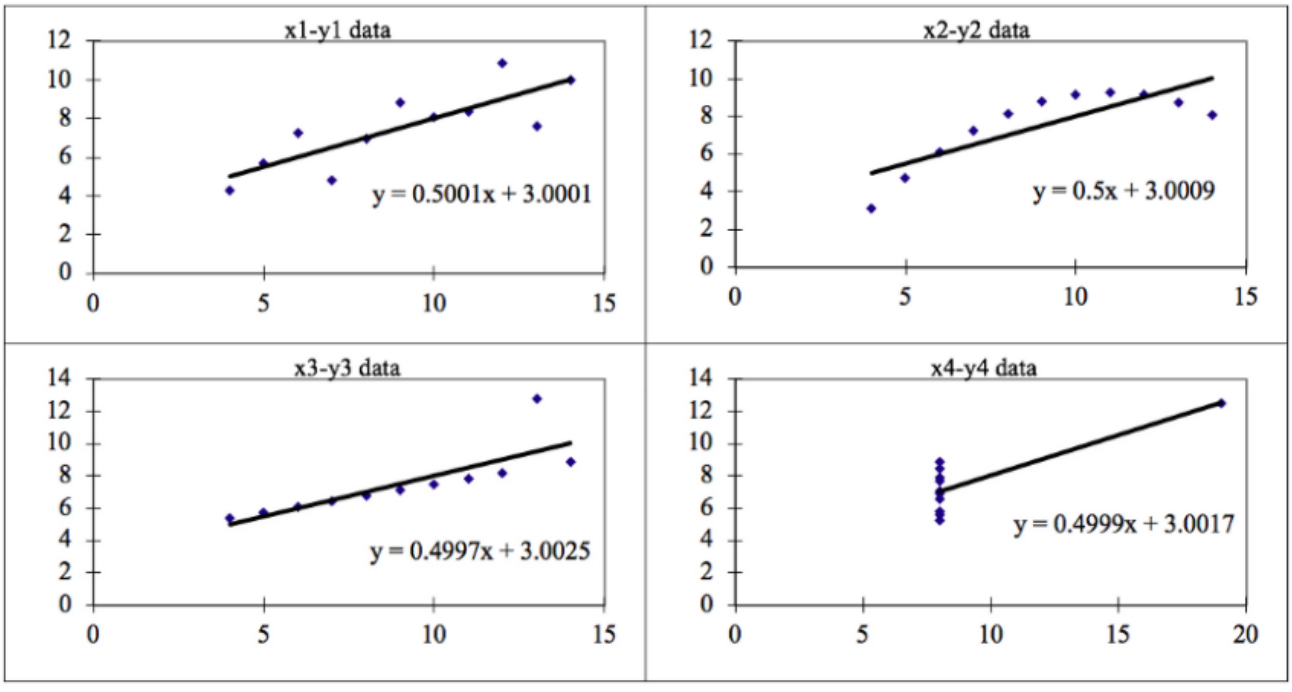
2. Explain the Anscombe's quartet in detail.

Francis Anscombe in 1973 realized that statistics are great for describing general trends and aspects of data, but using statistic alone can't fully depict any data set. He wrote a paper were showed 4 graphs later known as 'Anscombe Quartet'. The set of 11 data points in each data set are almost identical in terms of their basic descriptive statistic including mean, variance and correlation.

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

By looking at the data, it might look and concludes that it is similar, and at that time graphing was almost irrelevant. Anscombe proved that despites looking almost the same basic properties, but look different when plot it out.

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



- The First (top left) dataset shows that it fits the linear regression model.
- The Second (top right) dataset shows that this is not a linear regression model, obviously non-linear.
- The Third (below left) dataset shows a significant outlier which will influence the correlation coefficient from 1 to 0.816.
- The Fourth (below right) dataset shows one high leverage point to produce a high correlation coefficient.

The Anscombe Quartet shows the importance of visualising or graphing the data before start analysing the data.

3. What is Pearson's R?

Pearson's R or also known as Pearson Correlations measure the linearity between 2 set of data which can be ranged from 0 to -1. An r of -1 means there is negative linear relationship between 2 variables, r of 0 means there is no linear relationship between 2 variables and an r of 1 indicates that there is a positive relationship between 2 variables.

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

Scaling is a process where we scale the data within a particular range usually from 0-1 so that no variable is dominated by the other. It is performed during the pre-processing stage when the data carried different weight and magnitude where the data with high magnitude will weigh in a lot more than features with low magnitudes. We need to bring all of the data to the same magnitudes to suppress this weightage effect. This will improve the performance of the machine learning algorithm.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. This means, the minimum value in X is mapped to 0 and the maximum value in X is mapped to 1. It is also known as Min-Max scaling.

Standardization replaces the values by their Z scores. It will bring the data into normal distribution where mean is zero and standard deviation is 1. This means that the means of the attribute becomes zero and the resultant distribution has a unit standard deviation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Variance Inflation Factor (VIF) measure collinearity among predictor variables in a multiple regression analysis. ***When there is a perfect correlation, VIF = infinity. An infinite value of VIF indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.*** Usually, the high value of VIF indicates that there is a high correlation between the variables, this is what we called as multicollinearity. In general, the VIF value of > 10 indicates high correlations. While VIF is 1 indicates that there is no correlations with other variables.

6. What is Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot or the Quantile-Quantile plot is a technique in determining whether the two data sets come from populations with a common distribution such as Normal, Exponential or Uniform distribution. It is a plot of quantile from the first data set against the second data set. If both sets of quantiles came from the same distribution, there is a roughly straight forming points of line.

The Q-Q plot will determine whether:

- The two data sets come form the population with common distribution
- The two data sets have common location and scale
- The two data sets have a similar distributional shape
- The two data sets have similar tail behaviour