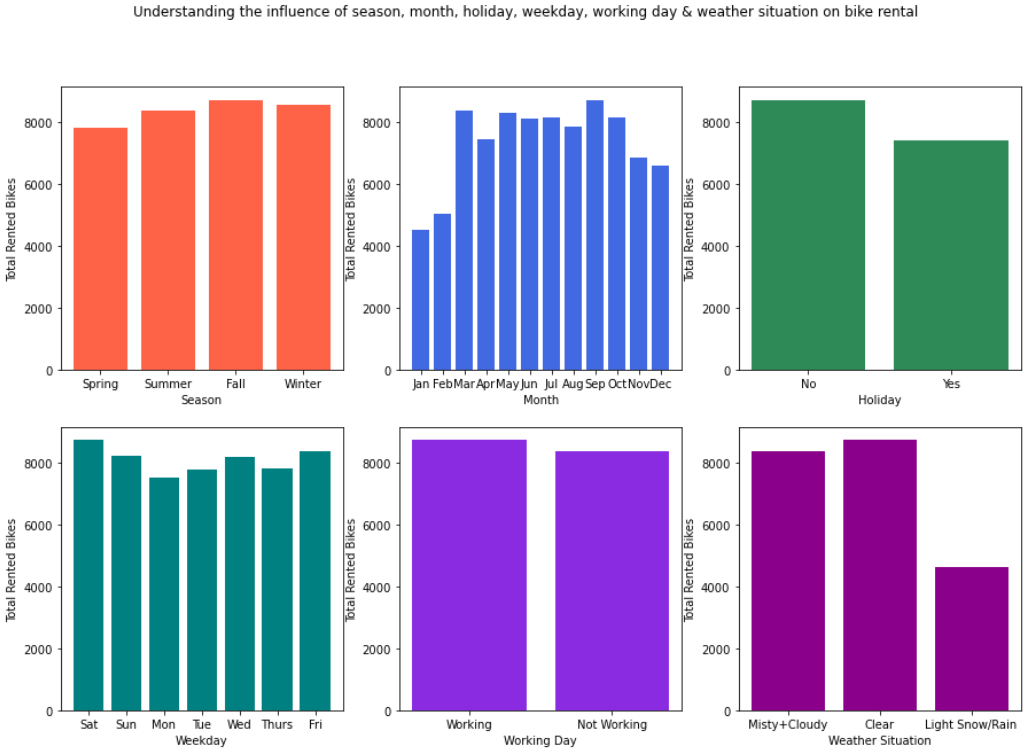
Abhirami S –Answer Sheet

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? (3 marks)

The dependent variable here is the 'cnt' column. The bike rental count is dependent on various environmental factors.

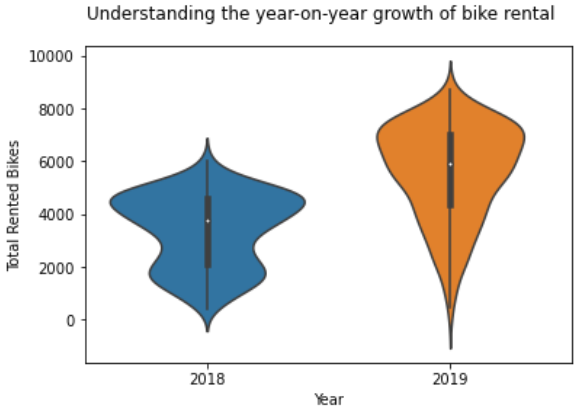
We have 7 categorical variables in the dataset, which have been analysed against the dependent variable. The graphs has been enclosed below for easy understanding.



We can draw the following inference from the above graph:

1. **Season** – We observe good rental rates during fall, followed by summer, attributing to the good weather.
   We also can notice that the graph is displaying an abnormality. The bike rental count during winter is observed to be higher than spring, which is unusual. The reason for this could be the months are not clearly defined for the seasons. We do not have a clear idea on when winter starts and ends.
2. **Month** - We observe that, rentals during Nov,Dec,Jan,Feb are low. This is due to the low temperature.
3. **Holiday** - We see that holiday does not affect the rental count. It has a very slight difference.
4. **Weekday** - We see that weekends have a slightly higher rental count. But can safely assume that, weekday does not have a large impact on bike rental count
5. **Working day** – We see that the bars are almost the similar for both working & non-working days. We can concur that, this factor is not related to rental count.
6. **Weather Situation** - We observe high rentals during clear days followed by misty days. With mild onset of rains, the rental count dips significantly. We also see zero rental count during heavy rain.

Abhirami S –Answer Sheet



7. We observe that the bike rental has increased significantly in the year 2019. This shows that Boombikes has had a good growth in 2019.

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

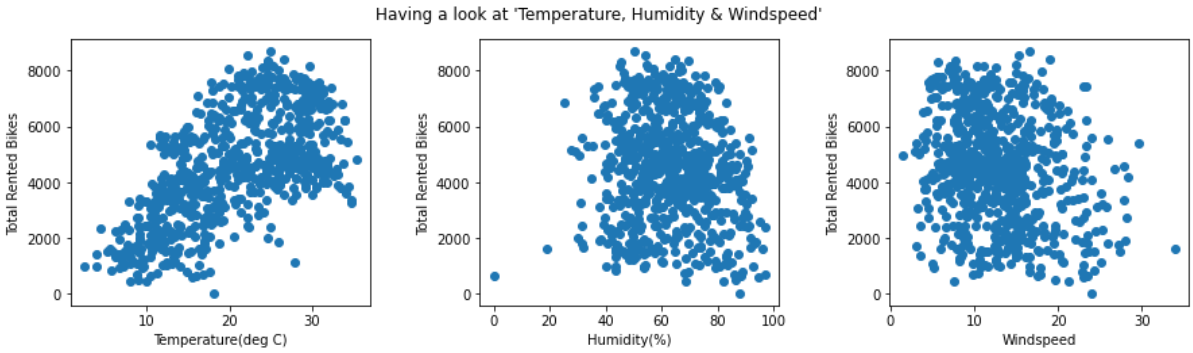
We use drop_first=True to avoid redundancy during dummy variable creation.

When we create dummy variables, we concatenate additional rows to the existing dataset. Each column will correspond to a unique category value in the original column. When the values are dummified, the dummy value under the unique category column will represent the presence or absence of the particular category in the original column.

Assuming we have three categorical value 'A', 'B' & 'C'. When we dummify this, if the sample has 'A' in it, 'B' & 'C' will be denoted by 0. This way 'B' & 'C' will define the value of 'A'. So, once we create dummies for n-1 categories, we will know the value of the nth category.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

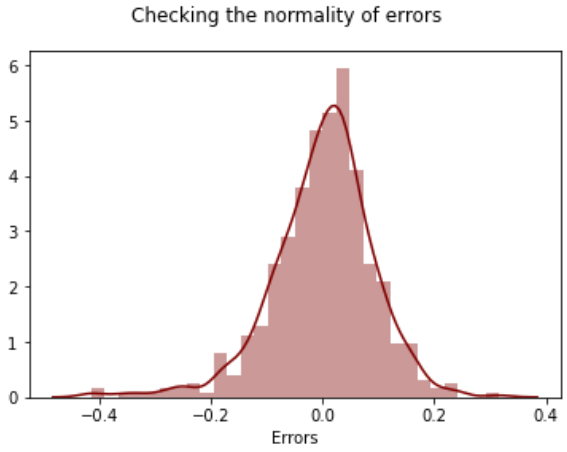
We can observe that temperature has the highest correlation with the target variable.



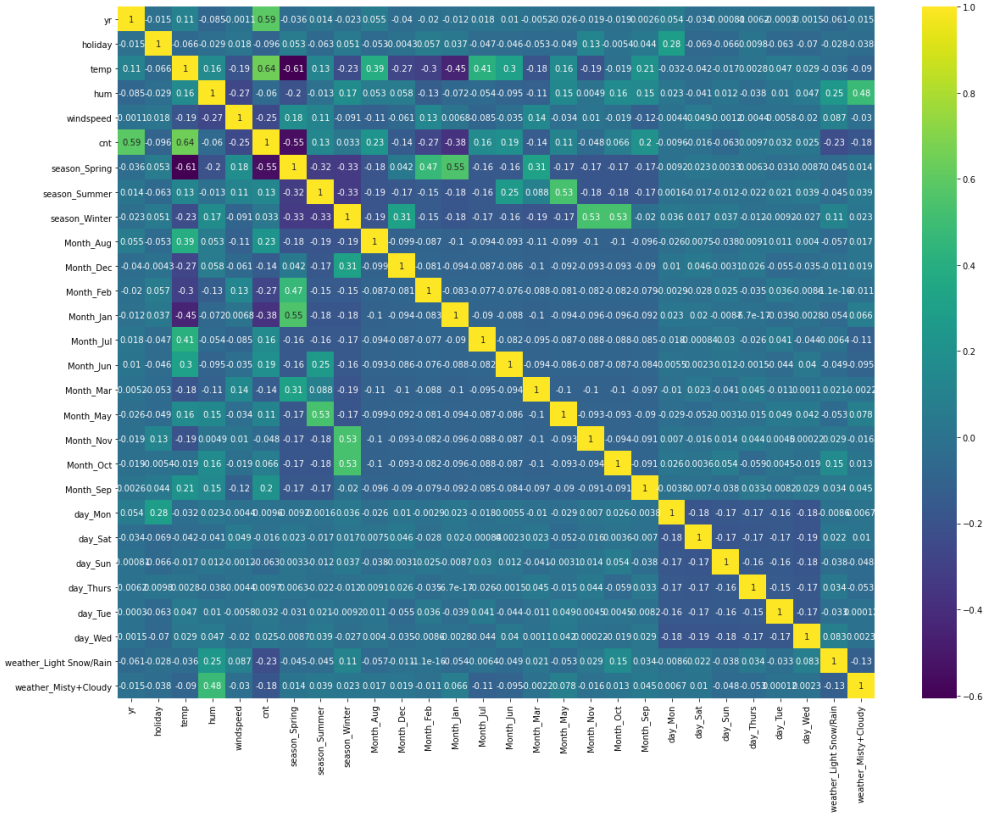
Abhirami S –Answer Sheet

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

1. By checking the normality of errors:



2. By checking the multicollinearity: There is no perfect multicollinearity



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top three features would be:

1. Temperature
2. Weather situation
3. Windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression algorithm is a machine learning technique under supervised learning method. This techniques is used to find the linear correlation between an independent & dependent variables.
Depending on the number of variables, there are two types of linear regression, Simple regression & Multiple Regression.
Simple regression is represented by a best fit line equation of, $y = \beta_0 + \beta_1 x + e$, where β_0 is the intercept & β_1 is the coefficient.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet has been constructed in 1973 by Francis Anscombe. This model was developed by him to emphasize the importance of graphical representation & removing outliers.

The Anscombe's quartet comprises four datasets, with each dataset consisting of eleven (x,y) points. You can see the datasets below.[1]

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

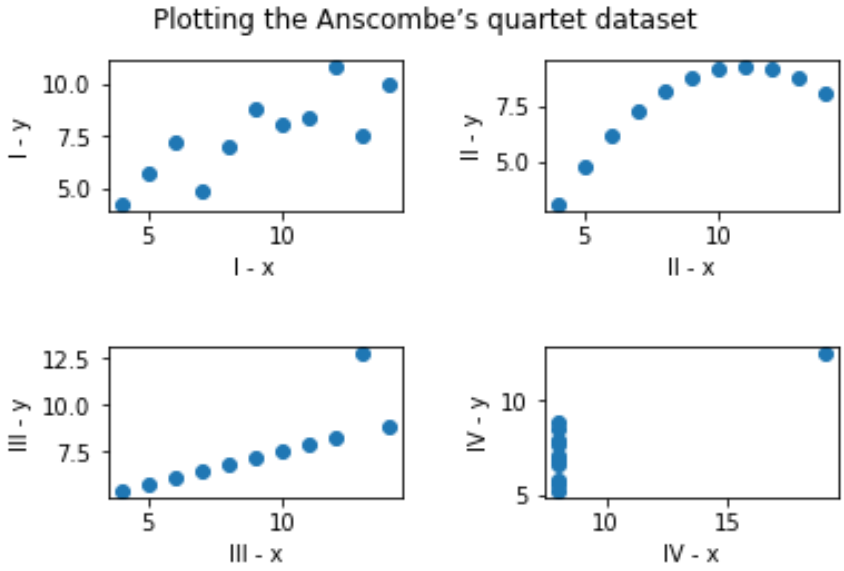
When descriptive analysis is done on these 4 datasets, they present nearly identical mean, standard deviation, and correlation between x and y. You can see the values below.[2]

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

[1] Image Source: https://www.geeksforgeeks.org/anscombes-quartet/
[2] Image Source: https://www.geeksforgeeks.org/anscombes-quartet/

When the given datasets are plotted as graphs, we observe the below variation in the datasets.



Importance: The Anscombe's quartet visually shows the significance plotting a data graphically as compared to observing the descriptive analysis.

3. What is Pearson's R? (3 marks)

Pearson's R is a technique used to identify the correlation between two variables. The variables are required to be quantitative and continuous variables and the data provided is assumed to have normal distribution.

The Pearson's R ranges from -1 to +1, with -1 having a negative correlation & +1 having a positive correlation. When Pearson's R is 0, it implies that the continuous variables are not correlated.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is the process of normalizing the values of the given data. This is done to regularize the units and range of the data to prevent error. For example, the data given might contain different measurement units like, 5cm & 5m, which will throw irregularities when we continue without scaling.

Normalized Scaling or Min-Max Scaling is used to remove outliers from the data. It has a fixed range of 0-1 and values below and above that are removed. This will help us have smaller values of standard deviation. Whereas, standardized scaling is used to equalize the range. Here the scaling is done with a mean value of 0 & standard deviation of 1. In standardized scaling, the outliers in the data will not be impacted.

Another important difference, we can use standardized scaling only when we assume that the data follows Gaussian distribution, whereas normalized scaling can be when data does not follow a Gaussian distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

We know that, VIF = $1 / (1- R^2)$.

The value of VIF tends to infinity, when the value of R^2 tends to 1. This happens only when the regression model perfectly fits the dataset.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q plots or Quantile-Quantile plots is a graphical method used to understand the distribution of a two datasets. It helps assess if the data of the sets are of normal, exponential or uniform distribution. This plots quantile of the data presented to derive an inference.

Generally in a Q-Q plot, a 45^o line is drawn from the bottom of x-axis to the top. The value of quantiles of the data are plotted on x-axis & y-axis. If the graph points fall along the line, this means both the data have similar distribution. As the deviation increases, it denotes the amount of deviation in the data sets.

In linear regression, the Q-Q plot look similar to a scatterplot. This is used as a visual aid to determine the correctness of a distributional assumption taken for a data set. This will be useful for a quick check on our assumption on normal distribution.