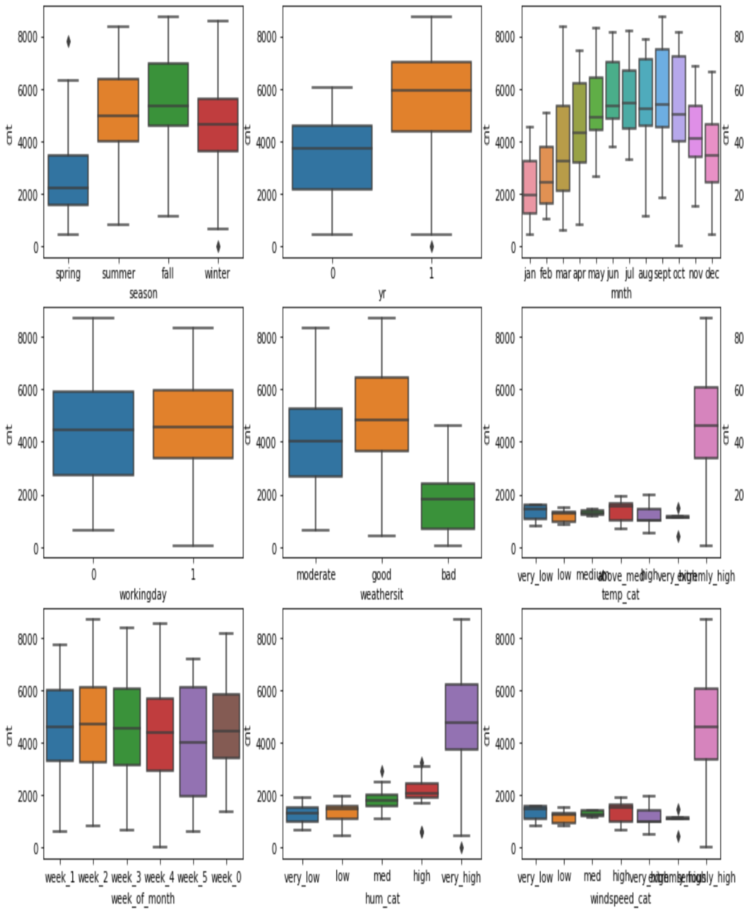
Assignment-based Subjective

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)
Ans:

Categorical values that are considered in the dataset, and few also derived are: 'season','yr','mnth','holiday','weekday','workingday','weathersit', 'temp_cat', 'start_of_month','end_of_month','week_of_month', 'hum_cat', 'windspeed_cat'



Note: Graph can be checked in notebook for better visualization

1. Fall has the Highest Demand, whereas spring has the lowest demand
2. Demand in bike sharing business has increased in 2019 compared to 2018
3. Demand keeps on increasing till September, with September Month being month of Highest demand. And there is decreased demand from Oct to Jan.
4. Higher demand when it is not a Holiday.
5. Weekends have slightly higher demand, but the mean is more or less the same. We can derive that there is not much impact of day of week in the data. And similar is with working day
6. Bad weather has the lowest demand, whereas when the weather is good, there is a high demand.
7. Start of the Month has less deviation in demand compared to the rest of the days.

8. End of the Month also has less deviation in demand compared to the rest of the days.
9. Week of the Month does not have much to say

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

Ans:

drop_first=True parameter in the dummy variable creation is important in order to maintain the N-1 Columns,
Example:
If we have 12 Columns of the Months, then the final number of columns for dummy variable needs to be N-1 which is achieved by drop_first=True

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

Ans:

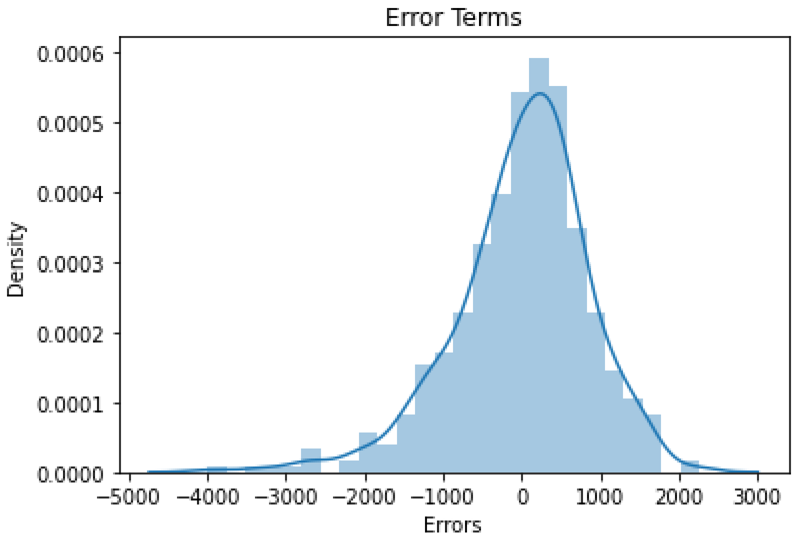
Looking at the pair-plot , variable with highest correlation is variable “temp” with correlation of 0.63 (as seen in heatmap) with respect to “cnt”.

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

Ans:

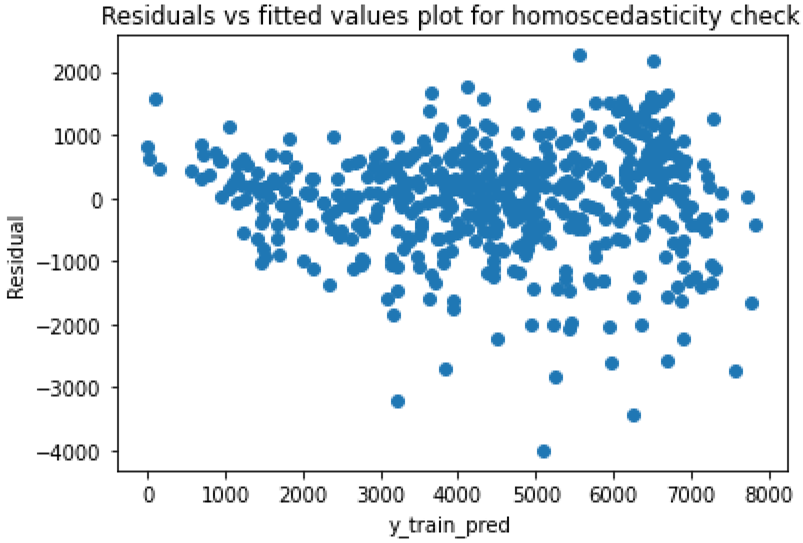
To Validate the assumptions of Linear Regression after Building, It was validated using three parameters:

1. Check for Normally Distributed Error Terms.
   As we can see all the error terms follow Normal Distribution with 0 as mean.

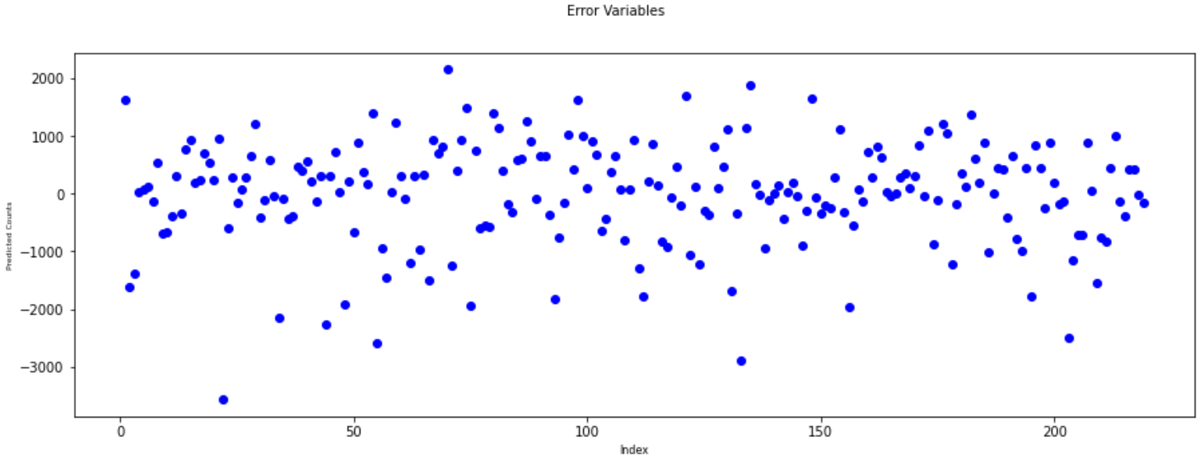


2. Homoscedasticity Check
   Homoscedasticity is also known as absence of heteroscedasticity. Which generally means that there is an impact of outliers in the model. The Homoscedasticity test

shows that there is no variance in the test variables



3. Randomly Distributed Error variables with respect to Predicted Counts
   The residuals are independent of each other. In the plot we can see that the variables are distributed randomly and do not have any correlationship.



4. Selecting a model with low VIF scores of Variable to Avoid Multicollinearity.
   The Independent Variables are not correlated to each other. This tends to keep the model to be built using independent variables and reduce the number of variables required to predict the model.
   As we can see below, all the VIFs in the final model are below 5.0

```
           Features   VIF
5.  2          temp   4.77
6.  1    workingday   3.92
7.  3     windspeed   3.41
8.  0            yr   2.02
9.  7   weekday_sat   1.66
10. 4 season_summer   1.56
11. 5 season_winter   1.38
12. 6     mnth_sept   1.19
13. 8 weathersit_Snow 1.06
```

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

On Based on Final Model, top 3 Features contributing significantly are:

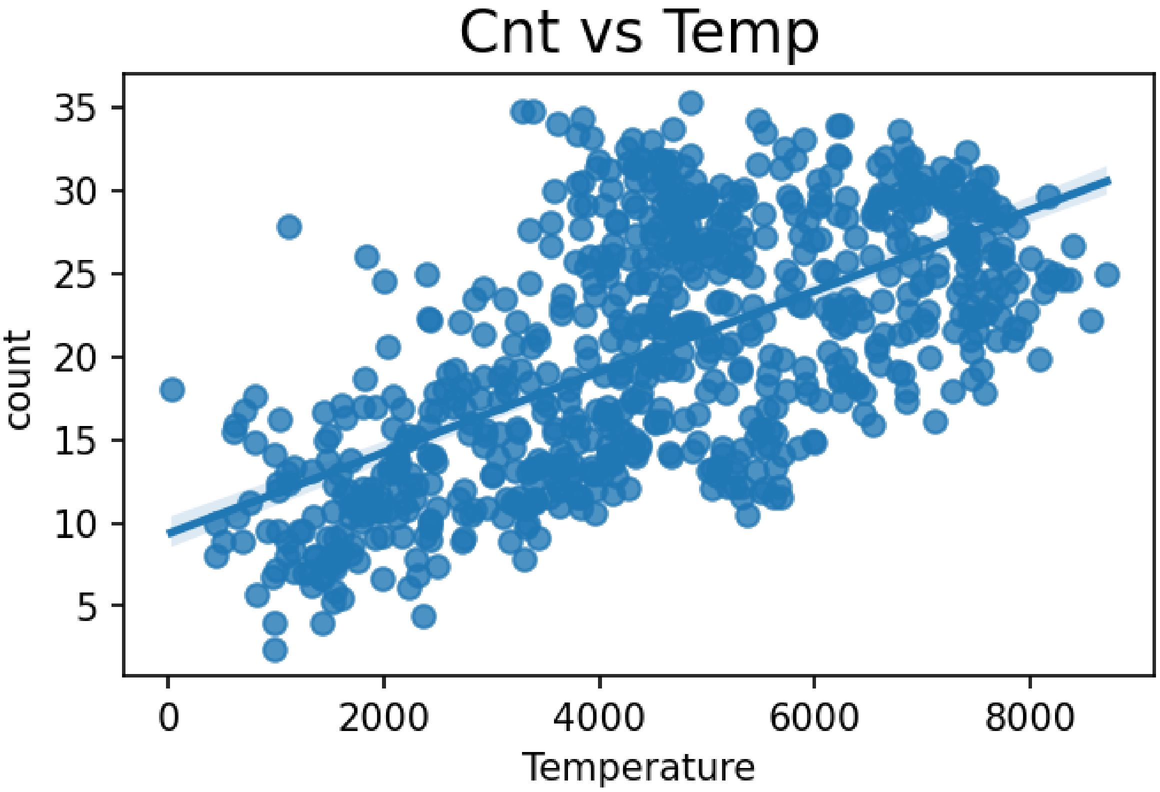
1. Temp , with coefficient of 4960.5648
2. Weathersit category 3 i.e. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (with coefficient of -2236.57)
3. yr , this variable explains the demand based on 2018 and 2019. Here we can clearly see that demand is increasing in consecutive years. (with coefficient of 2037.00)

General Subjective

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

Linear Regression is used to predict the target variable, given some Input data points. The basic principle of linear regression is to create a model which can use correlation of two or more variables and fit a line between the data points to calculate or predict target variable.

1. **Concept of Linearity:**
   When two variables are plotted on a graph, and we fit a line through the points, the resulting correlation between the given points is called linearity.



Here in the graph above, we can see that there is an upward trend in the fit line. This property can be further harnessed to create Multivariable linear regression.

2. **The Model:**
The statistical/ML Model which is created is nothing but a mathematical equation which is similar to the equation of a line.
Y = mx + c

The Equation can be modified to get the the relation between two variables and generally denoted as:
Y = β0+β1X+[ϵ]

Here,
Y is dependent Variable,
X is Independent Variable
ϵ is the error term

3. **Evaluation:**
To Evaluate the Linear Regression, we use certain statistical metrics, like R-Squared Score, Adjusted R-Squared Score, Root Mean Square Error etc.

R-Squared score is a statistical measure that determines how good the fit is between the variables. A R-Squared score lies between 0 to 1

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2	=	coefficient of determination
RSS	=	sum of squares of residuals
TSS	=	total sum of squares

Adjusted R-Squared score is upgraded form of R-Squared Score such that also considers the number of variables which are used to build the model and in turn penalises the model’s R-Square score

$$Adjusted\ R^2 = 1 - \frac{(1-\ R^2)(N-1)}{N-p-1}$$

Where
R^2 Sample R-Squared
N Total Sample Size
p Number of independent variable

Root Mean Squared Error is nothing but standard deviation of the residuals of the predicted variables.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

4. **Assumptions of Linear Regression**

   Linear relationship: This Assumption in Linear Regression is one of the most important assumptions. Where it states that there is some linear relationship between the dataset variables and the target variable.

   No auto-correlation or independence: The residual terms are independent of each other. I.e. there is no correlation between the error terms. The Residual terms need to be randomly distributed.

   No Multicollinearity: It is considered that there is low multicollinearity in the model. This can be achieved using the VIF.

   Homoscedasticity: This assumption states that there is absence of heteroscedasticity. This ensures that the model is not influenced heavily by outlier terms in the data set

   Normal distribution of error terms: This assumption states that the error terms are normally distributed with its mean at 0.

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

3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks