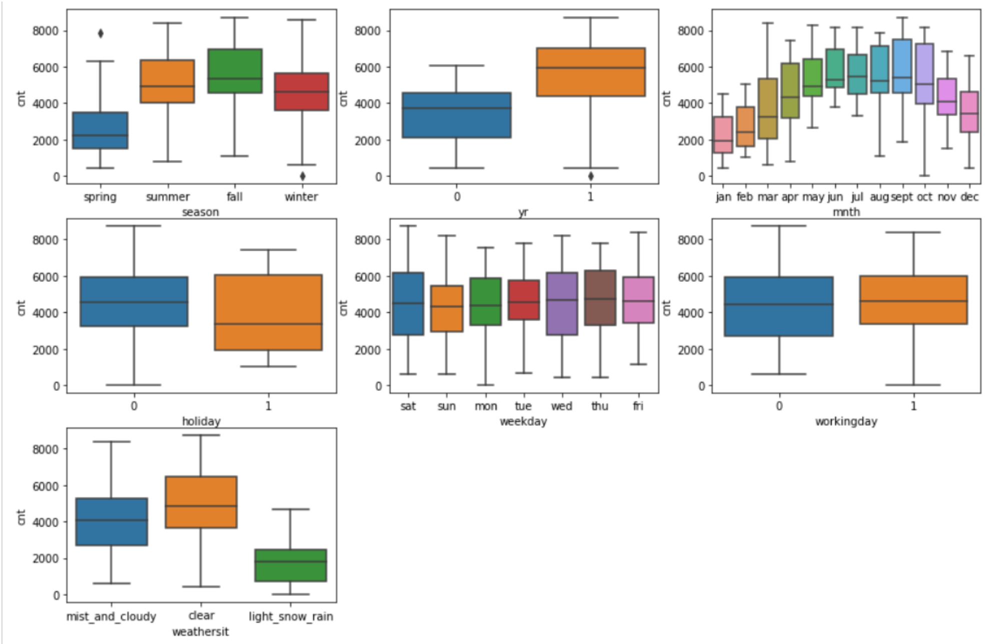
Assignment based Questions

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

Below are the categorical variables in the given data set

1) Season
2) Weathersit
3) Holiday
4) Mnth
5) Yr (Year)

When we have the box plots they have the below effect on each of below mentioned variables.



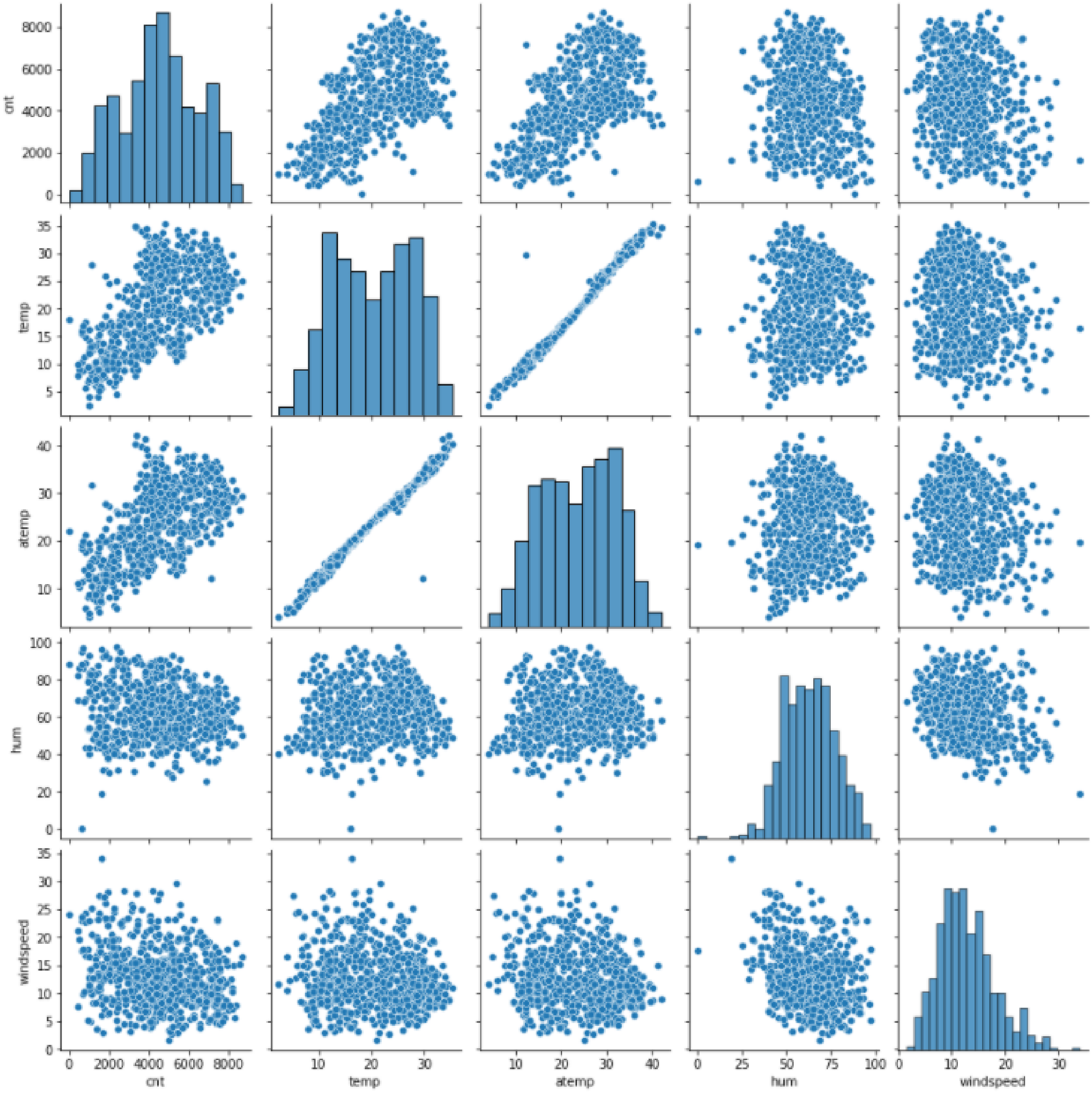
1) **Season** - The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt.Summer and winter had intermediate value of cnt.
2) **Weathersit** - Highest count was seen when the weathersit was' Clear, Partly Cloudy'
3) **Holiday** - rentals reduced during holiday.
4) **Mnth** - September saw highest no of rentals while December saw least.
5) **Yr (Year)** - The number of rentals in 2019 was more than 2018

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

If the column is not dropped then the dummy variables will be correlated (redundant). Dropping the columns also reduces the number of features. .Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables.

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

   As per the analysis the below is the pair plot among the numerical variables from the given data set



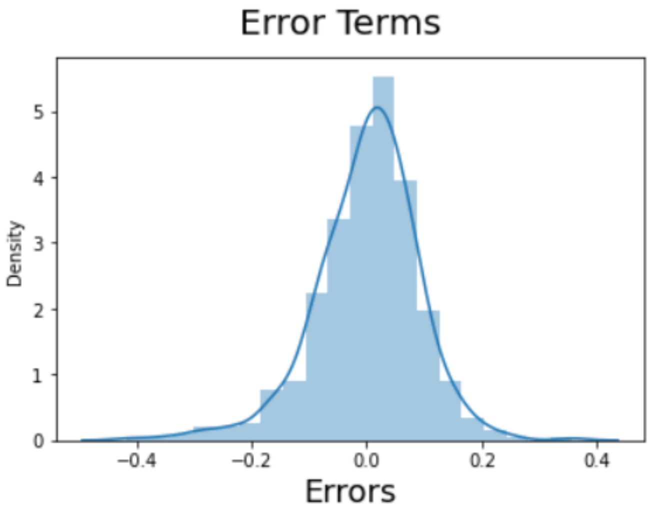
By looking at the above diagrams, the variables **‘temp’** and **‘atemp’** are considerably highly correlated with the output variable ‘cnt’. Also these variables are highly correlated themselves. Hence in the analysis **‘atemp’** is removed by RFE.

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

   Below are the assumptions of Linear regression

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

Once the model is built, we first validate it on the training set, we see the residuals (error terms) following the normal distribution. They should be distributed around '0'. If the graph (dist graph) we plotted shows that it is following normal distribution we can say that assumptions of Linear regression are valid. Below is the plot of residuals after the model is built from the given data set.



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

   Based on a model built on the given data set. Below are the top 3 features contributing significantly for the demand of shared bikes.

   Below are the coefficients of top 3 features

   1) temp : 0.491508
   2) Yr: 0.233482
   3) Weather light snow and rain : -0.285155 It is negatively impacting the demand

General Questions

1. **Explain the linear regression algorithm in detail?**
   - Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values.
   - Linear regression is based on the popular equation "y = mx + c" , an equation of a line.

- A linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).
- In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.
- Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc.

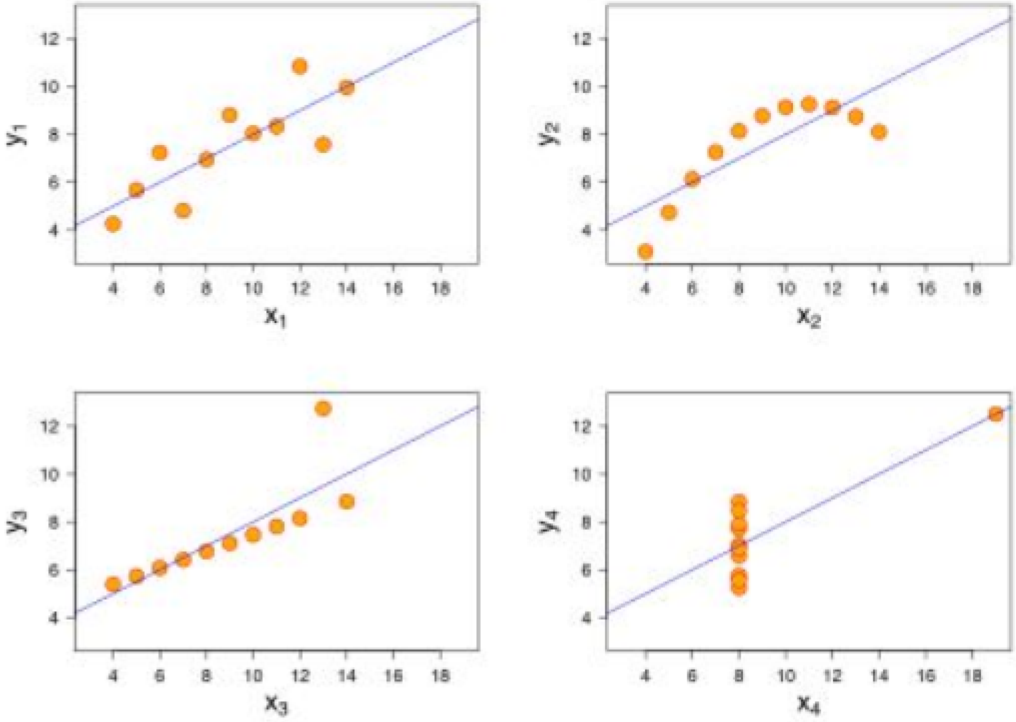
Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable.

2. Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables.

2. **Explain Anscombe's quartet in detail?**

Anscombe’s Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph.It was developed to emphasise both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them,it’s not linear.
- The last graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson’s R?

It is the summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data.

- r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association