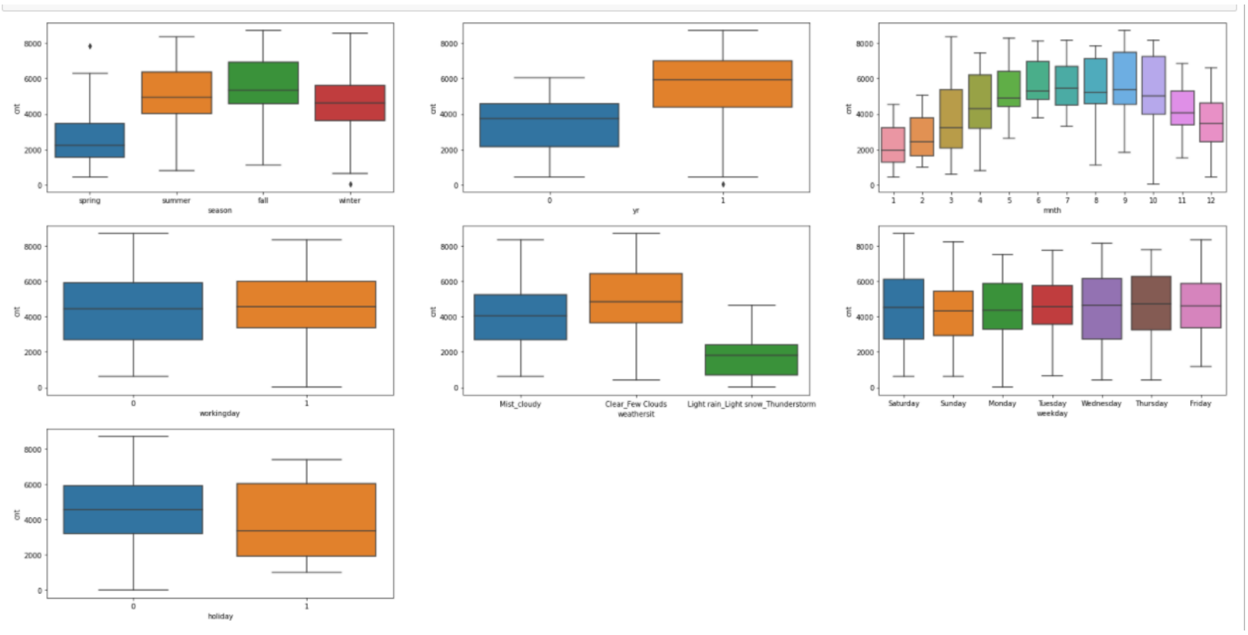
ASSIGNMENT BASED SUBJECTIVE QUESTIONS :-

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

Answer:- There were 7 Categorical variables in the dataset.

We used Boxplot (**refer the figure below**) to study their effect on dependent variable(“**cnt**”).



The inference that we could derive were:

1) **season**: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

2) **mnth**: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

3) **weathersit**: Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

4) **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

5) **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

6) **workingday**: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

7) **yr:-** year also have a very good working trend with cnt which can be seen from the boxplot diagram.

2. Why is it important to use drop_first=True during dummy variable creation?
Answer:- drop_first=True is important to use,as it helps in reducing extra column created during dummy variable creation.Hence,it reduces correlation created among dummy variables.
For example:- Lets say we have 3 types of values in Categorical Column and we want to create Dummy variable for that column. If one variable is not furnished and semi_furnished,then it is obviously unfurnished. So,we do not need any 3rd variable to identify the unfurnished.

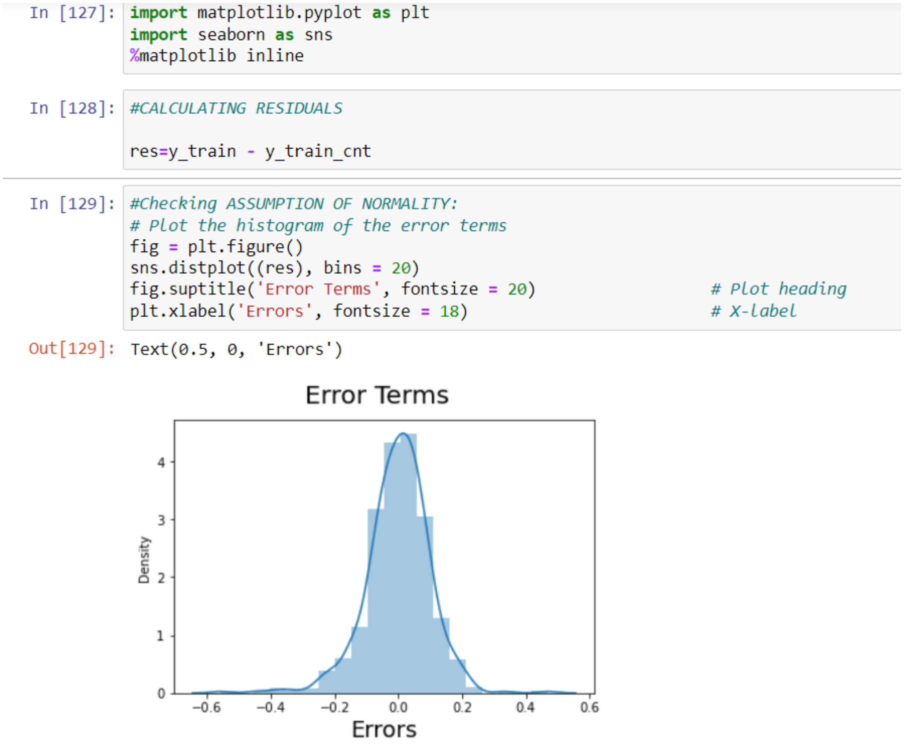
Hence,if we have categorical variable with n levels,then we need to use n-1 columns to represent the dummy variables.

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

Answer:- Looking at the pairplot among the numerical variables,there is a linear relation between **temp**,**atemp** and target variable '**cnt**'.

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

Answer:- There are 3 assumptions of Linear Regression after building the model on Training set.
a) Error terms are normally distributed with mean zero(not X,Y):-



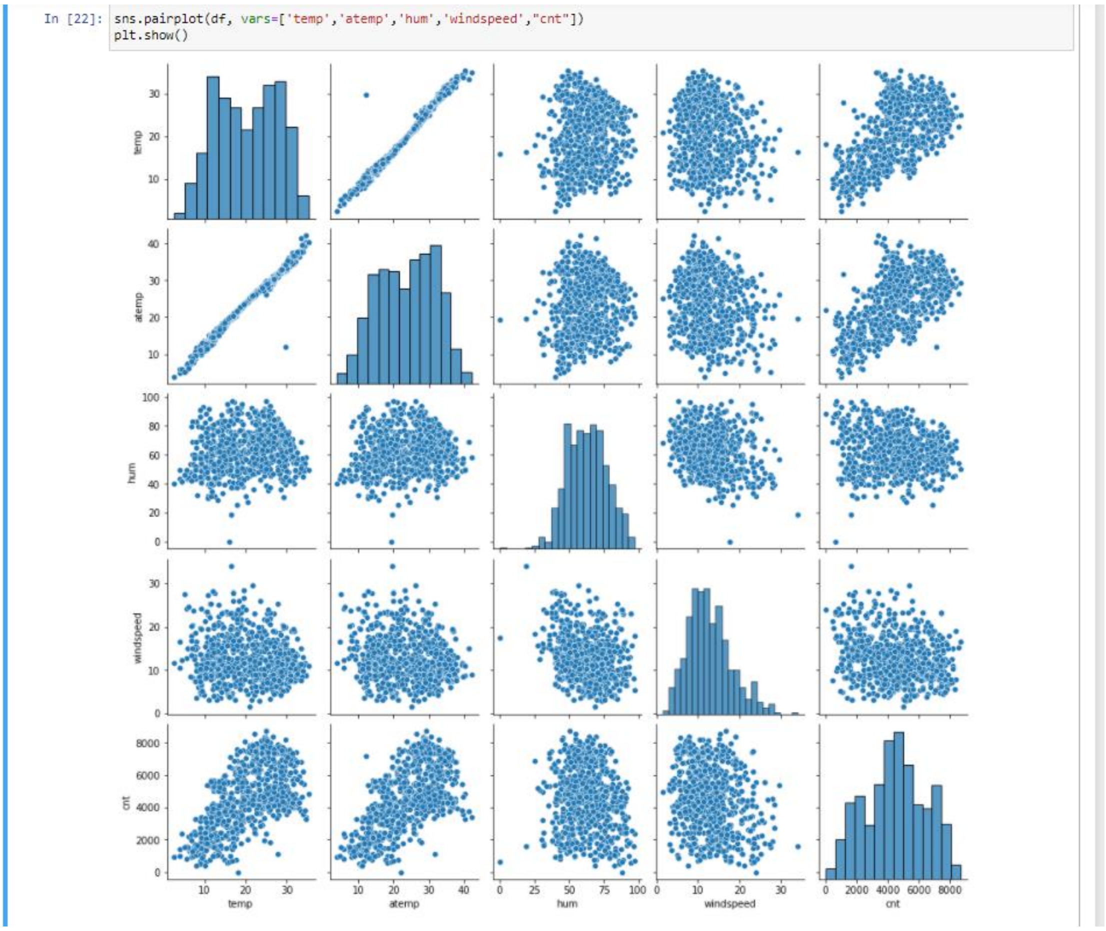
```
In [127]: import matplotlib.pyplot as plt
          import seaborn as sns
          %matplotlib inline
```

```
In [128]: #CALCULATING RESIDUALS

          res=y_train - y_train_cnt
```

```
In [129]: #Checking ASSUMPTION OF NORMALITY:
          # Plot the histogram of the error terms
          fig = plt.figure()
          sns.distplot((res), bins = 20)
          fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
          plt.xlabel('Errors', fontsize = 18)                         # X-label

Out[129]: Text(0.5, 0, 'Errors')
```

Insight :- From the above histogram,we can see that errors are normally distributed. Hence,our assumption for Linear Regression is valid.

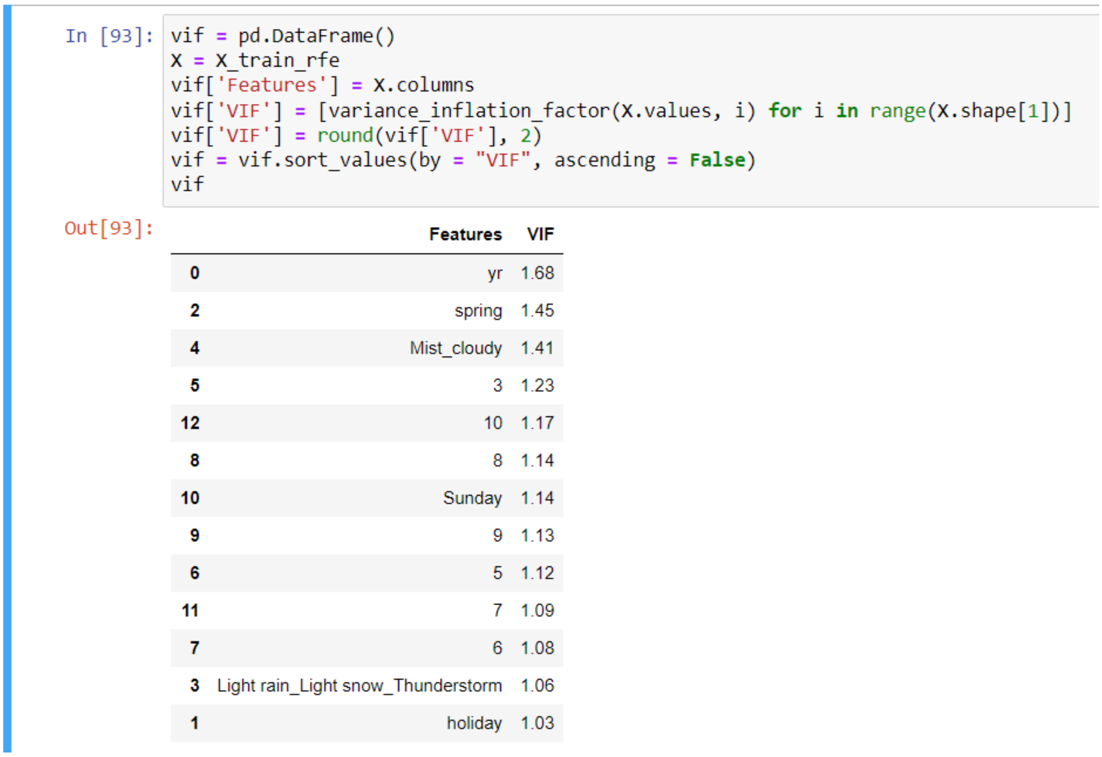
b) There is a Linear relationship between X and Y



```
In [22]: sns.pairplot(df, vars=['temp','atemp','hum','windspeed',"cnt"])
         plt.show()
```

Insight:- Using the pairplot, we could see there is a linear relation between **temp** and **atemp** variable with the predictor ‘**cnt’**.

c) There is no Multicollinearity between the Predictor variables.



```
In [93]: vif = pd.DataFrame()
         X = X_train_rfe
         vif['Features'] = X.columns
         vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
         vif['VIF'] = round(vif['VIF'], 2)
         vif = vif.sort_values(by = "VIF", ascending = False)
         vif
```

Out[93]:

	Features	VIF
0	yr	1.68
2	spring	1.45
4	Mist_cloudy	1.41
5	3	1.23
12	10	1.17
8	8	1.14
10	Sunday	1.14
9	9	1.13
6	5	1.12
11	7	1.09
7	6	1.08
3	Light rain_Light snow_Thunderstorm	1.06
1	holiday	1.03

Insight:- From the VIF Calculation,we could find that there is no multicollinearity existing between the predictor variables,as all the values are within permissible range of below 5.

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

Answer:- Based on the final model,the 3 top features contributing significantly towards explaining the demand of shared bikes are:-

Kindly refer the target variable **'cnt'** is calculated as follows:-

cnt= 0.246*yr-0.083*holiday-0.198*spring-0.321*Light rain_Light snow_Thunderstorm-0.090*Mist_cloudy+0.063*3+0.123*5+0.148*6+ 0.153*8+0.193*9-0.049*Sunday+0.126*7+0.116*10

a) **yr**:- Demand increases(as coefficient is positive) in case of yr.

b) **Light rain_Light snow_Thunderstorm**:- Demand decreases(as coefficient is negative) in case of Light rain_Light snow_Thunderstorm.

c) **8(August month)**:- Demand increases(as the coefficient is positive) in case of 8.

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.
Answer:- In simple terms, linear regression is a method of finding the best straight line fitting to the given data,i.e. finding the best linear relationship between independent and dependent variables. In technical terms,linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data,between independent and dependent variables. It is mostly done by Sum Of Squared Residuals Method.
There are 4 assumptions in **Linear Regression Model**:-
1. There is a linear relationship between X and Y.
2. Error terms are normally distributed about mean.
3. Error terms are independent of each other.
4. Error terms have a constant variance.

2. Explain the Anscombe's quartet in detail.
Answer:- Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.
It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties**.
There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.
After drawing the four datsets using Scatter plots with 11 datapoints of X and Y,we found out that:-

Dataset 1: this **fits** the linear regression model pretty well.

Dataset 2: this **could not fit** linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.

Dataset 4: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.

From these four datasets,we can conclude that the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

Answer:- The Pearson correlation coefficient, *r*, can take on values between -1 and 1. The further away *r* is from zero, the stronger the linear relationship between the two variables. The sign of *r* corresponds to the direction of the relationship. If *r* is positive, then as one variable increases, the other tends to increase. If *r* is negative, then as one variable increases, the other tends to decrease. A perfect linear relationship (*r*=-1 or *r*=1) means that one of the variables can be perfectly explained by a linear function of the other.

The assumptions are as follows: **level of measurement, related pairs, absence of outliers, and linearity**. Level of measurement refers to each variable. For a Pearson correlation, each variable should be continuous.

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

Answer:- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-max scaling:-
It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.
MinMaxScaling:- x= (x-min(x))/(max(x)-min(x))

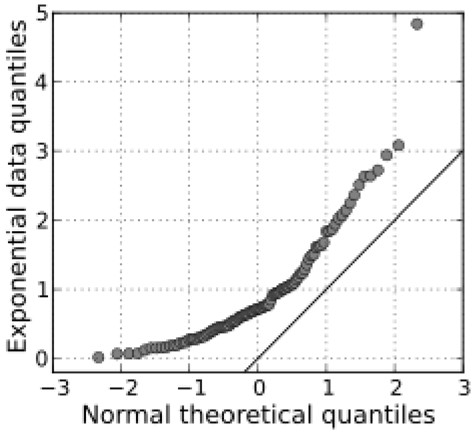
Standardization Scaling:-
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).
Standardisation:- x=(x-mean(x))/std(x)
sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

Answer:- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer:- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.