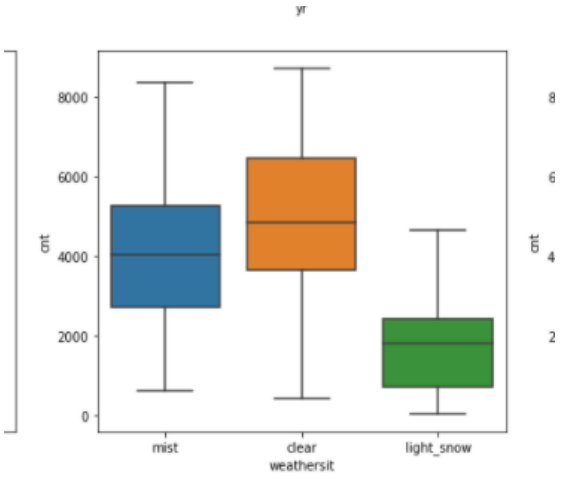
Assignment-based Subjective Questions

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

Ans. From the Exploratory Data Analysis performed on the dataset we can conclude that the categorical variables have a greater effect on the dependent variable.
From the boxplot made during EDA of the Bike Sharing dataset we can visualise the relationship between the categorical variables and the dependant variable.
For example if we consider the categorical variable weathersit the cnt value varied based on the weathersit(mist,clear,light_snow) which can be inferred from the box plot.It was high when the value was clear and low for the light_snow.



Similarly after building the model from the equation(given below) we can infer that the weathersit variable has a considerable effect on cnt
cnt=0.426729-(0.136521* holiday)-(0.100401*windspeed)+(0.497460*casual)-(0.162943*season_spring)+(0.070032*mnth_Sep)-(0.147933*weekday_Mon)-(0.076515*weekday_Sun)-(0.077957*weekday_Tues)-(0.212319*weathersit_light_snow)-(0.058941*weathersit_mist)+(0.200237*yr_2019)
The inclusion of categorical features such as weathersit,yr,season,weekday, etc we saw a significant growth in the value of R-squared and adjusted R-squared.It implies that these categorical features have a major influence and are useful to explain the dependent variable.

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

Ans: drop_first=True is important during the dummy variable creation as it helps in reducing the extra column created. Hence it reduces the correlations created among dummy variables.
For example:
Suppose we have Gender column. It has two values Male and Female.
Now using get_dummies we created two columns each for Male and Female. But one is sufficient for analysis other is unwanted/redundant feature and increases the correlation.
If the Female column is 1 it means Gender is Female .If 0 it means Male.
It is exactly opposite for the Male column. if 1 means Male 0 means Female.
Hence we can drop one of the column and take the remaining column.
Therefore 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?

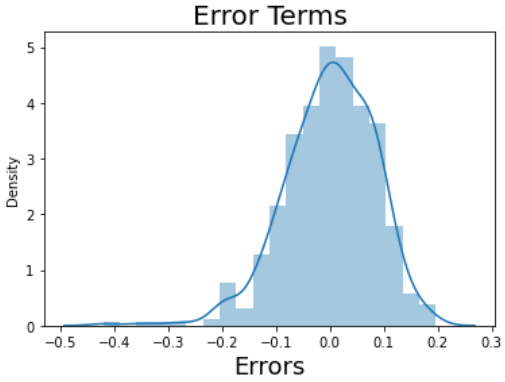
Ans: In the pair plot the 'registered' column(0.95) has the highest correlation with the target variable(cnt).

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

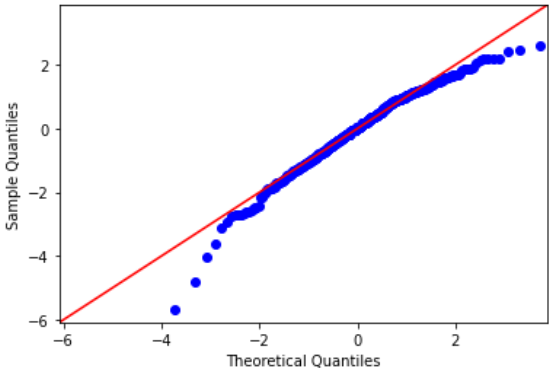
Ans:Validation of Linear Regression assumptions:

- Verify the error terms follow normal distribution:

```
# The error terms are normally distributed with mean 0 -Plot the same using distplot
sns.distplot((y_train-y_train_cnt),bins=20)
plt.title('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)
plt.show()
```



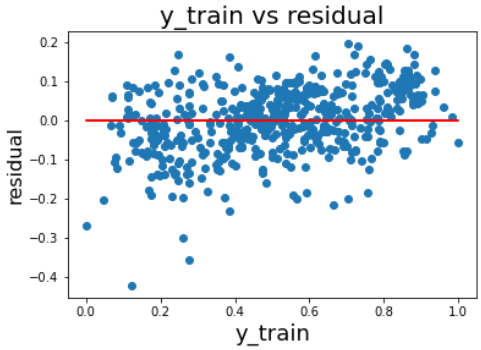

```
#Error Terms are normally distributed using Q-Qplot
import scipy.stats as stats
res = lm5.resid # residuals
fig = sm.qqplot(res,stats.t, fit=True,line="45")
plt.show()
```



From the above two graph it is clearly evident that the residuals are normally distributed.

- To verify the error terms are independent of each other and Homoscedasticity

```
# Plotting y_train and residual to understand the spread.
fig = plt.figure()
plt.scatter(y_train,y_train-y_train_cnt)
plt.plot(y_train,(y_train - y_train), '-r')
plt.title('y_train vs residual', fontsize=20)              # Plot heading
plt.xlabel('y_train', fontsize=18)                         # X-label
plt.ylabel('residual', fontsize=16)
```



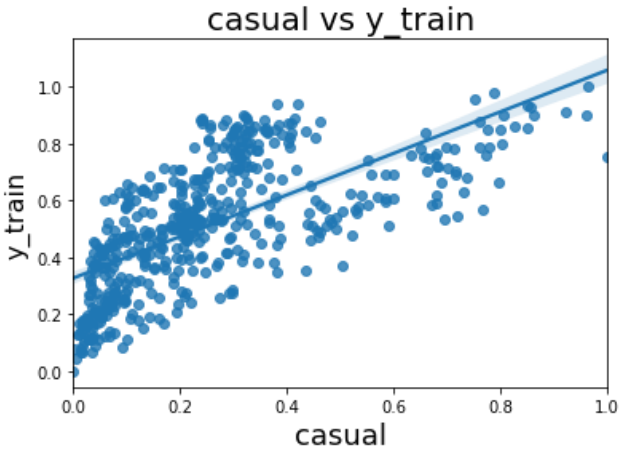
There is no visible pattern and the error terms have constant variance.

- Multicollinearity doesnot exists
  #Get the VIF for the model 5 after dropping const
  vif=pd.DataFrame()
  X=X_train_5
  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

  output:
  FeaturesVIF0const10.203casual1.996weekday_Mon1.424season_spring1.377weekday_Sun 1.268weekday_Tues1.1411yr_20191.122windspeed1.099weathersit_light_snow1.0910weath ersit_mist1.095mnth_Sep1.061holiday1.02

  The VIF <5 for all the predictors which proves that there is no multicollinearity.

- The linear Regression between X,y
  The highest coefficent is of casual for linear regression model .Plotting the value wrt y_train will show the plot be linear



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans: The equation of the best fitted line is
cnt=0.426729-(0.136521* holiday)-(0.100401*windspeed)+(0.497460*casual)-(0.162943*season_spring)+(0.070032*mnth_Sep)-(0.147933*weekday_Mon)-(0.076515*weekday_Sun)-(0.077957*weekday_Tues)-(0.212319*weathersit_light_snow)-(0.058941*weathersit_mist)+(0.200237*yr_2019)
From the above we can see that the top 3 features that contribute significantly are:

column	coeff value
Casual	0.49746
weathersit_light_snow	-0.212319
yr_2019	0.200237

The variables are Casual,Weathersit,yr

General Subjective Questions

1. **Explain the linear regression algorithm in detail**

Ans: Linear Regression Algorithm is a machine learning algorithm based on supervised learning.In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

For example :In case of a when running a sales promotion we expect the number of customers to increase.For this we use historical data and plot the same on graph and see if the no of customers increase with the promotion.

In the plot if the value are linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation.
or else there will be a linear downward relationship which means when increase in X there is decrease in y.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$y=a+bx$

where b is slope
a is intercept

In case of the multiple linear regression the value of y depends on multiple predictor variable.

$y=a+b1x1+b2x2+b3x3....+bnxn$

where b1,b2,....bn are slope of corresponding predictor
a is intercept

Steps in Implementation of Linear Regression

Step-1: Data Pre-processing

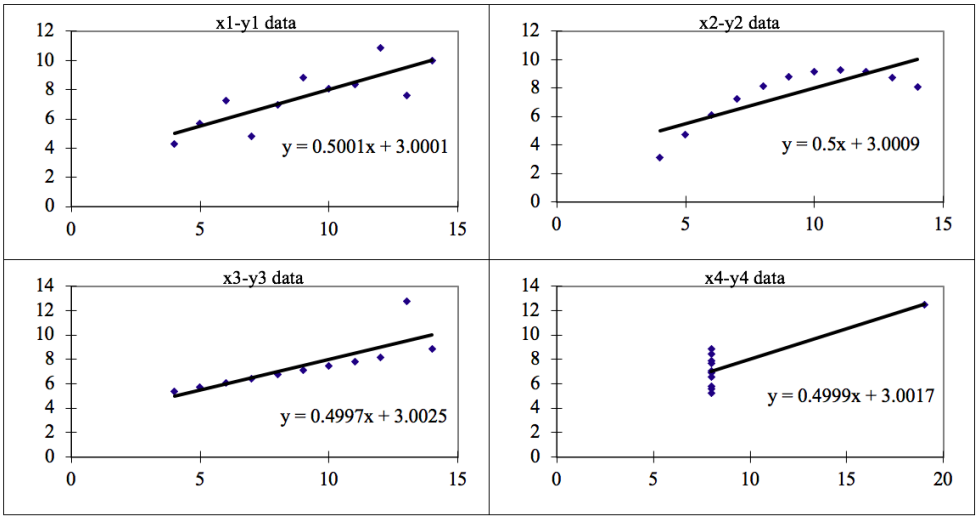
Import the necessary libraries and dataset.And then perform exploratary Data Analysis on the dataset.

Step 2:Dealing with multicollinearity by dropping varaibles
Also Convert the categorical variable using get_dummies to dummy variable .
Create X and y
Create train and test sets
Scaling of the Training set data
Step 3:Build a model using sklearn/Statsmodel on the training.Use any Feature selection(manual or automated) till a proper model is acheived.
Step4:Evaluate the model on training set using(R2,Advanced R2,F statistics,VIF)
Step 5: Verify the linar regression assumptions
Step 6:Use the linear model on test set and make predictions.Evaluate the results using ((R2,Advanced R2,F statistics,VIF))

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises of four datasets that have nearly identical simple descriptive statistics but when they are plotted on the Scatter plot they have a very different distribution.
Consider the given example :

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	



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

Though the four datasets have the same mean variance still the graphs are different for them. And the regression model can be fooled by the values .Hence it is always necessary to visualise the data set before implementing the Regression model.

3.What is Pearson's R?
Ans: Pearson's R measures the strength of the linear relationship between two variables.
Pearson's r always lies between -1 and 1.

When x increases if y increases in exactly the same way the pearson's r value is 1
When x increases if y decreases in exactly the same way the pearson's value is -1
r=0 the data is not related in any way(random points)

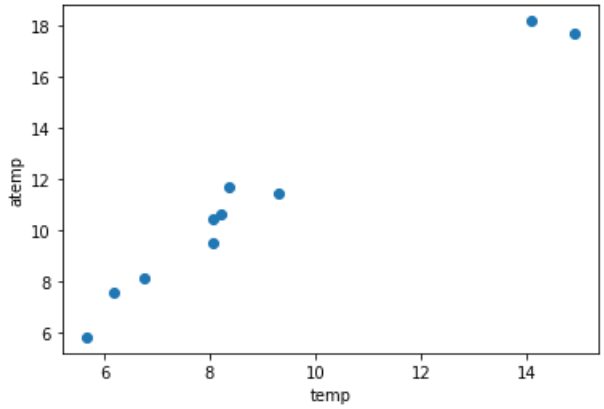
The formula for pearson's r is given below:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Consider the below data (sample) from the bike sharing dataset:

temp	atemp
14.11085	18.18125
14.9026	17.68695
8.050924	9.47025
8.2	10.6061
9.305237	11.4635
8.378268	11.66045
8.057402	10.44195
6.765	8.1127
5.671653	5.80875
6.184153	7.5444

Scatter plot



Calculation of the pearson's r value in python for the temp and atemp

```
from scipy.stats import pearsonr
corr, _ = pearsonr(temp.temp, temp.atemp)
print('Pearsons correlation: %.2f' % corr)
```

Pearsons correlation: 0.98
We have received a value of 0.98 for pearson's r.It means that both the temp and the atemp are high positively correlated.

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

Ans: 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.

Scaling is done because of the following reason:

Mostly the 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.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

MinMax Scaler can be used in presence of outliers as it is not affected by the same.

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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python

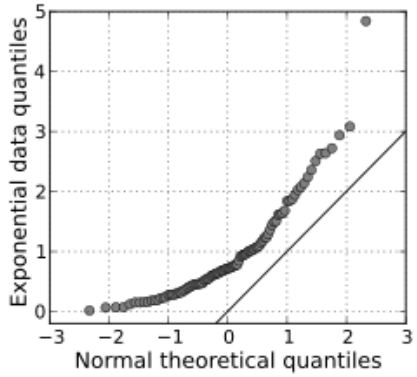
StandardScaler cannot guarantee balanced feature scales in the presence of outliers

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: 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).
The formula for VIF =1/(1-R2)
When there is s perfect correlation between the two independent variables then R2 =1

Then VIF=1/0=infinity

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q-Q plot is called a normal

quantile-quantile (QQ) plot. The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

Consider the Q-Q plot for bike sharing data

```
import scipy.stats as stats
res = lm5.resid # residuals
fig = sm.qqplot(res,stats.t, fit=True,line="45")
plt.show()
```