Seoul Bike Sharing Demand Prediction

Sourav Chowdhury(Data science trainee)
AlmaBetter, Bangalore

Abstract:

Public rental bike sharing is becoming popular because of its increased comfortableness and environmental sustainability. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. Data used include Seoul Bike Share data, have weather data associated with it for each hour.Seoul Bike sharing system was set up in 2015.
The data used in analysis is collected for the year 2017 and 2018.

1. Problem Statement:

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Variable breakdown:

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Attribute Information:

- Date : year-month-day
- Hour - Hour of the day
- Temperature-Temperature in Celsius

- Humidity - %
- Wind Speed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)
- Rented Bike count - Count of bikes rented at each hour(Dependent Variable)

Steps involved:

- Importing libraries
- Exploratory Data Analysis
- Extracting features from date
- Data Preparation
- Model Building Using Hyper Parameter Tuning
    1. Linear Regression
    2. Lasso Regression
    3. Ridge Regression
    4. ElasticNet Regression
    5. DecisionTree Regressor
    6. RandomForest Regressor
    7. Gradient Boost
    8. Xg Boost
- Model Evaluation through Metrics :
    1. Mean Squared Error
    2. Root Mean Squared Error
    3. R-Squared
    4. Adjusted R-Squared

2. INTRODUCTION :

Having the data of Seoul Bike of 2017 and 2018 with 8760 entries.Depending upon various factors mentioned in the variable section, we have analyzed the data with Exploratory data analysis. There are various factors on which the demand for the Bike in a particular area depends.The aim of this Project is to develop machine learning models to predict the number of bikes required in a Bike Sharing System. Bike-sharing systems allow anyone to hire bicycles from one of the city's numerous automated rental stations, ride it for a short distance, and then return it to any station in the city. Many cities across the world have recently implemented similar systems to encourage citizens to utilize bicycles as an environmentally sustainable and socially equitable means of transportation, as well as a good complement to other forms of public transportation.

In this Project we have :

- builds regression models using bike data from Seoul. Most methods use computational models, while my approach is based on machine learning. Due to the availability of data today and the ever increasing amount of data that can be collected.

3. Methods:

3.1 Data set:

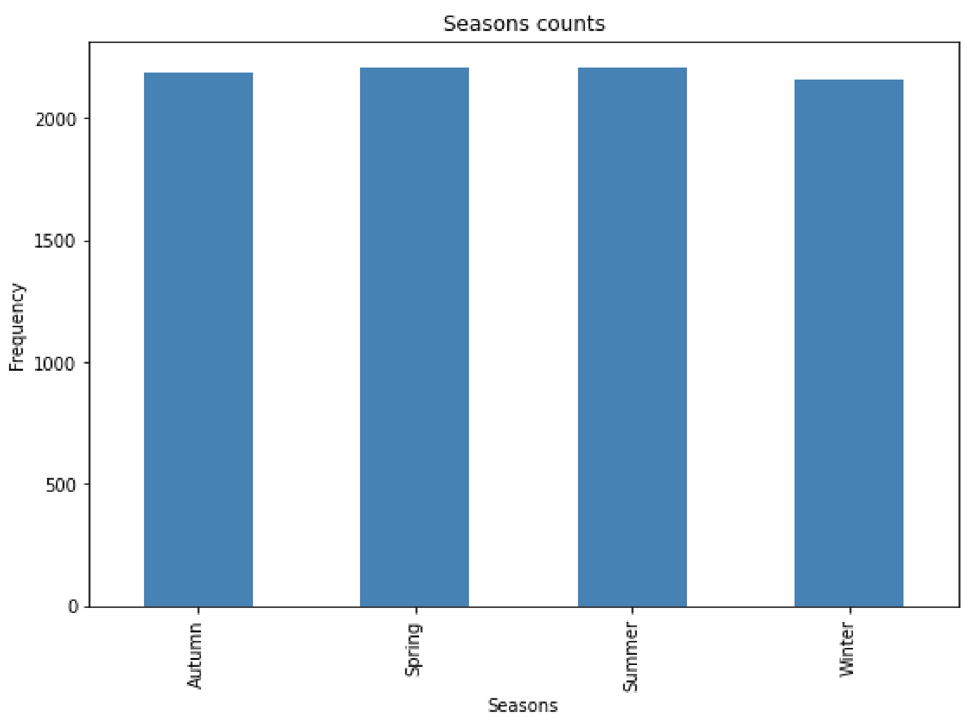
For this Project, the dataset of bike-sharing demand in Seoul was used.It is composed of 14 attributes: Date, Rented Bike Count, Hour, Temperature(°C),Humidity (%), Wind Speed (m/s), Visibility (10m), Dew Point Temperature(°C), Solar Radiation (MJ/m2), Rainfall (mm), Snowfall (cm), Seasons, Holidays, Functioning Day with 8760 instances.

```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Date                       8760 non-null   object
 1   Rented Bike Count          8760 non-null   int64
 2   Hour                       8760 non-null   int64
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object
 12  Holiday                    8760 non-null   object
 13  Functioning Day            8760 non-null   object
```

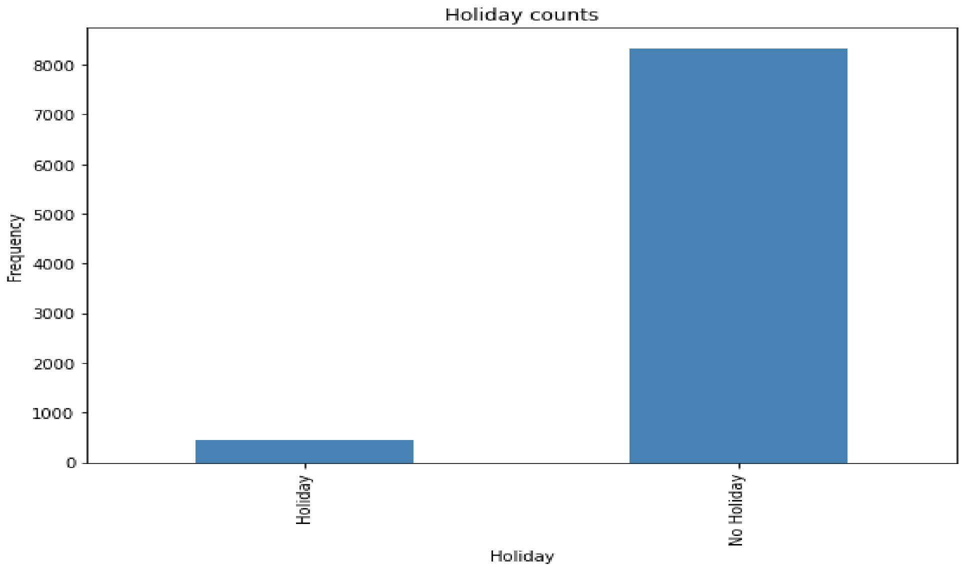
3.2 Data exploration and pre-processing:

Methods for data exploration and preprocessing will be presented in order to improve prediction for machine learning models.

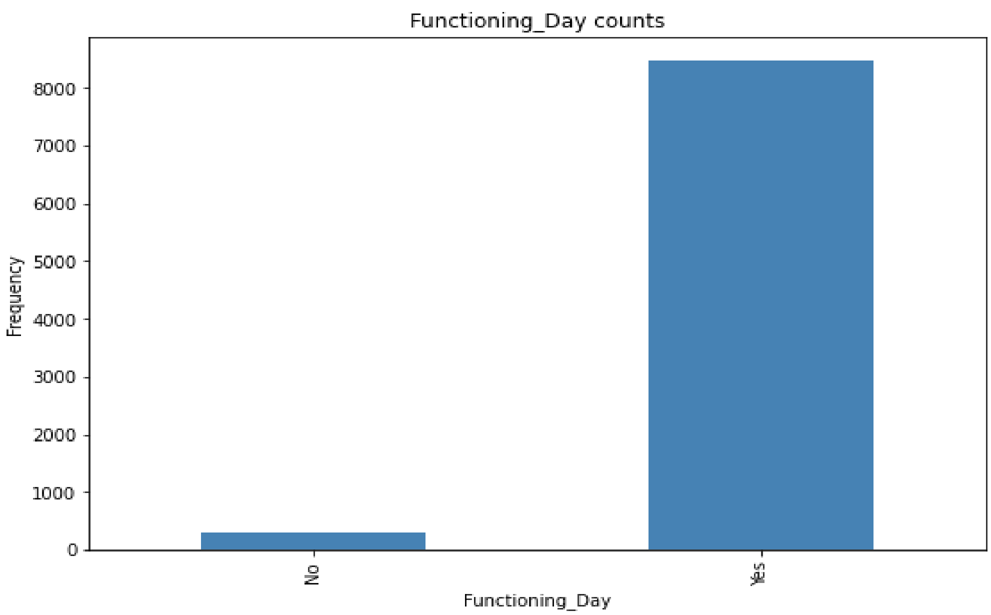
- **Univariate Analysis:**




```
Name: Seasons
Spring    2208
Summer    2208
Autumn    2184
Winter    2160
```



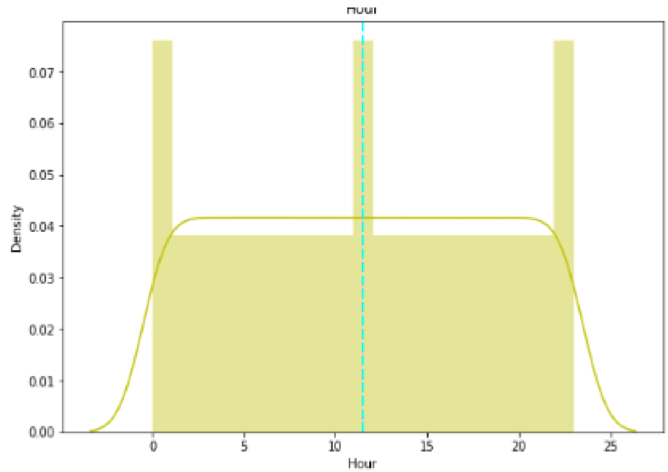
Name: Holiday
No Holiday 8328
Holiday 432

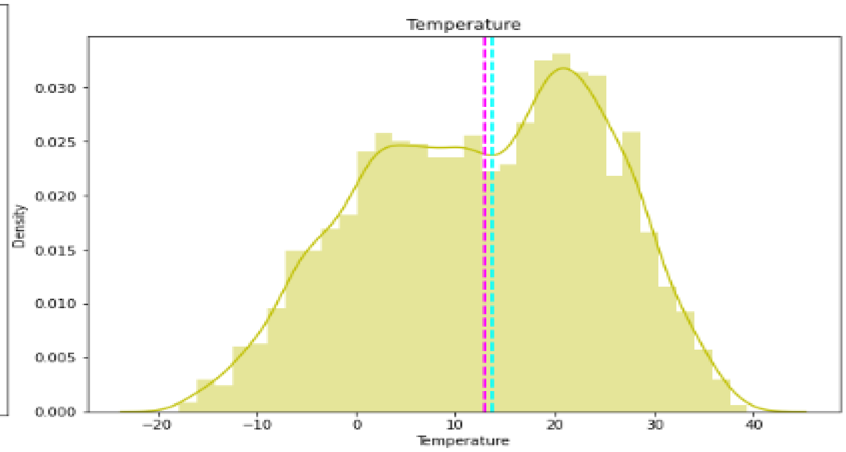


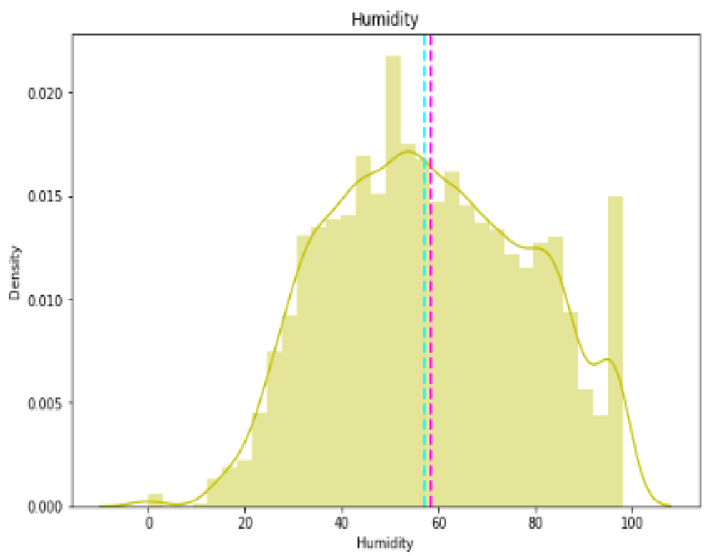
Functioning_Day

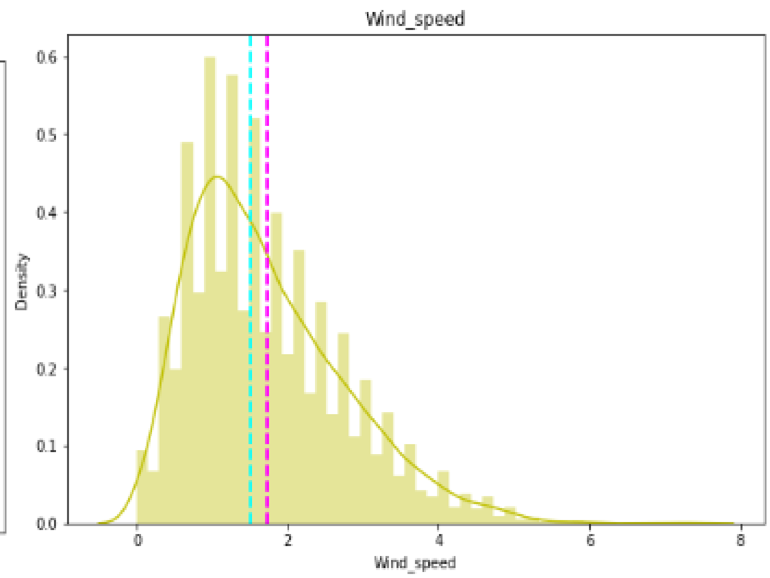
Yes 8465

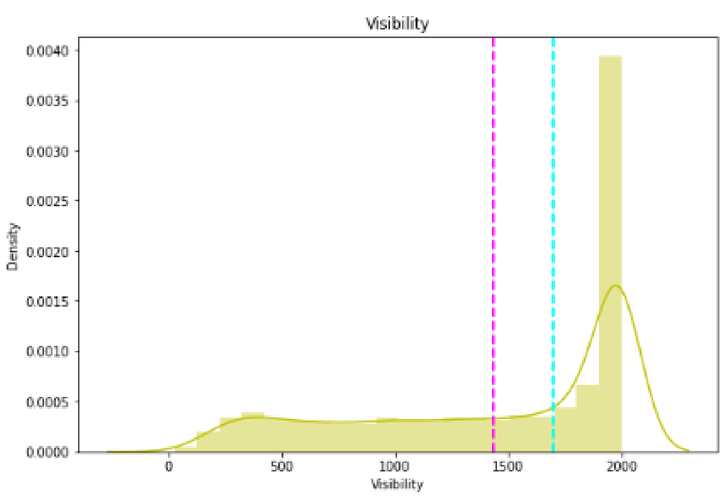
No 295

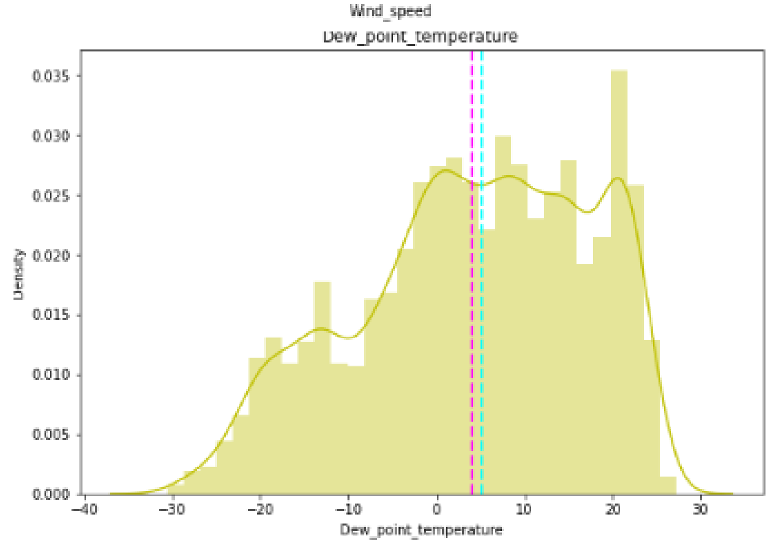

Density
Hour

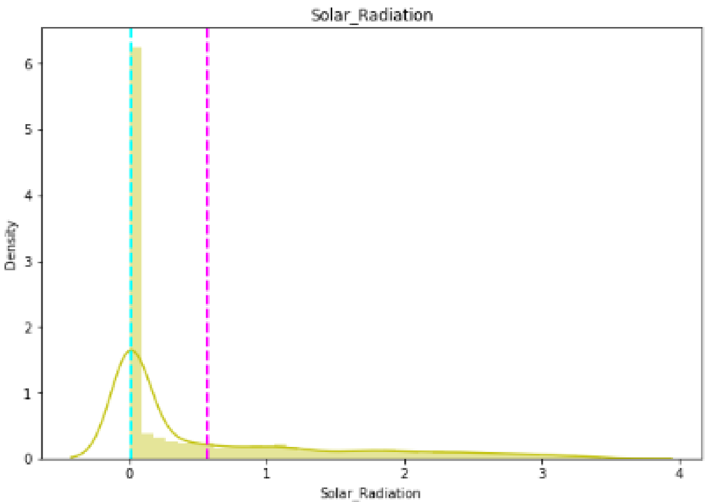

Temperature
Density
Temperature

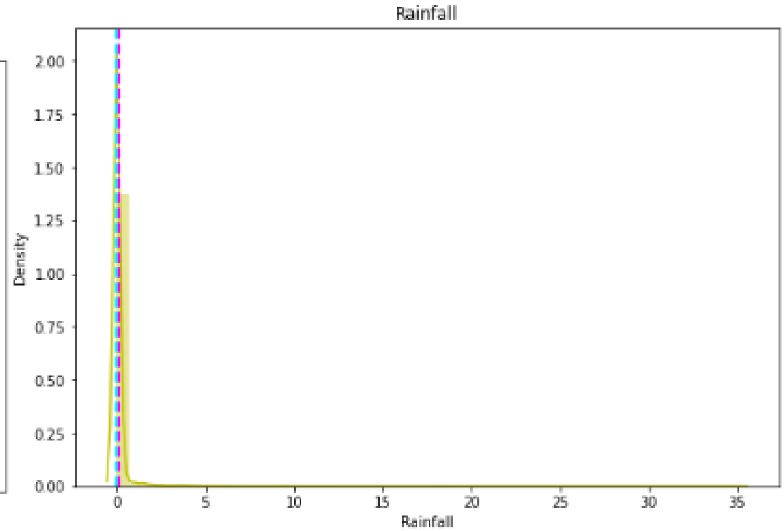

Humidity
Density
Humidity

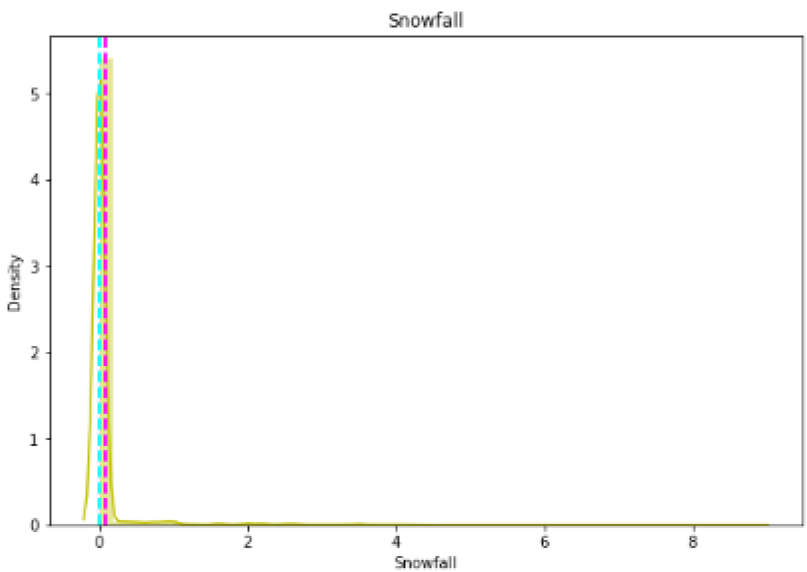

Wind_speed
Density
Wind_speed

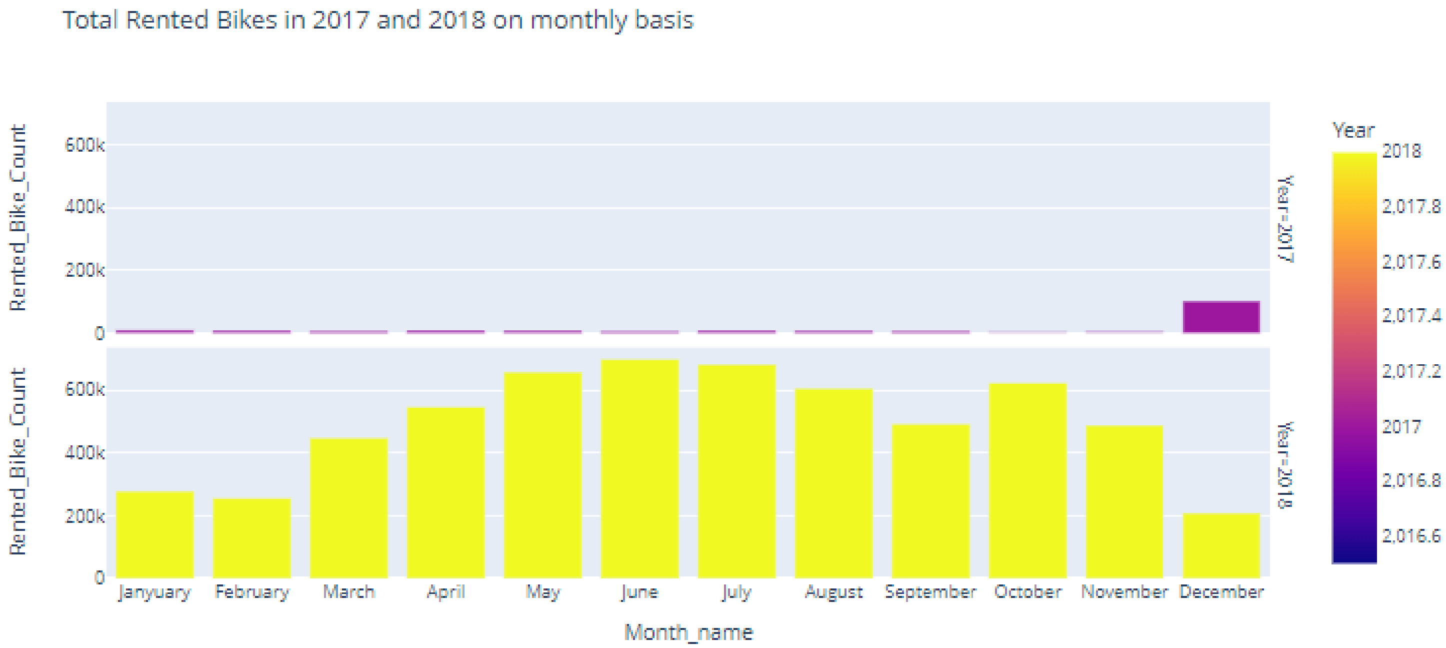

Visibility
Density
Visibility


Dew_point_temperature
Density
Dew_point_temperature


Solar_Radiation
Density
Solar_Radiation
0
1
2
3
4
5
6

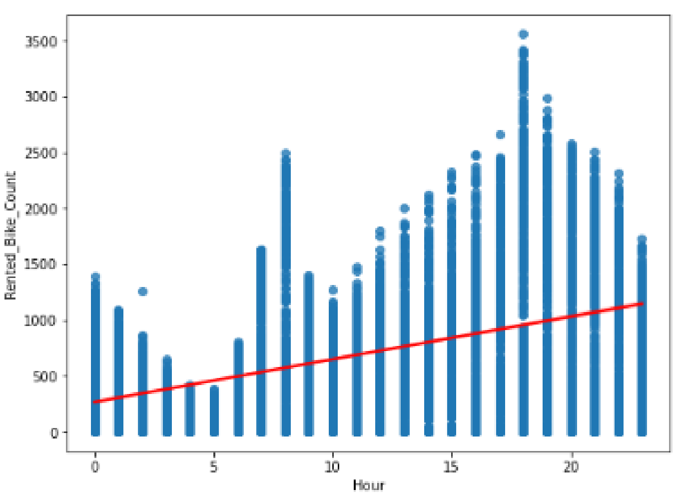

Rainfall
Density
Rainfall
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0
5
10
15
20
25
30
35

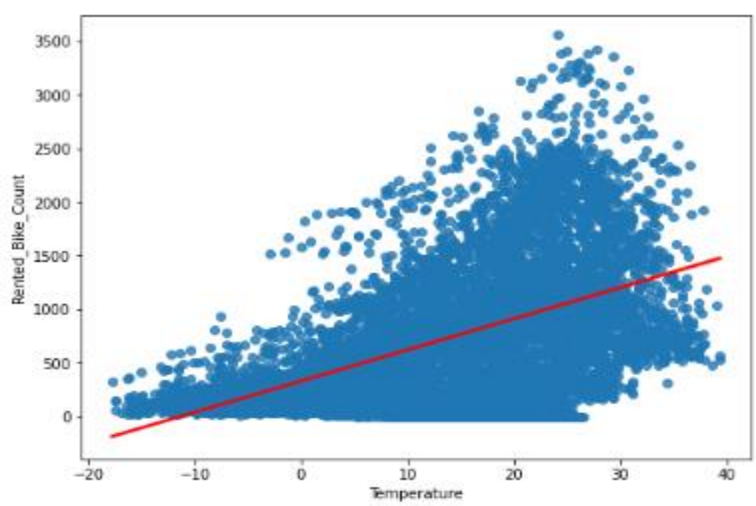

Snowfall
Density
Snowfall
0
1
2
3
4
5
0
2
4
6
8

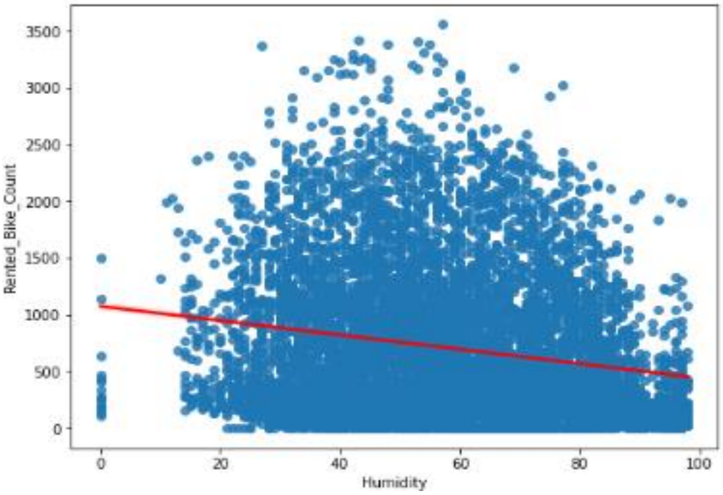


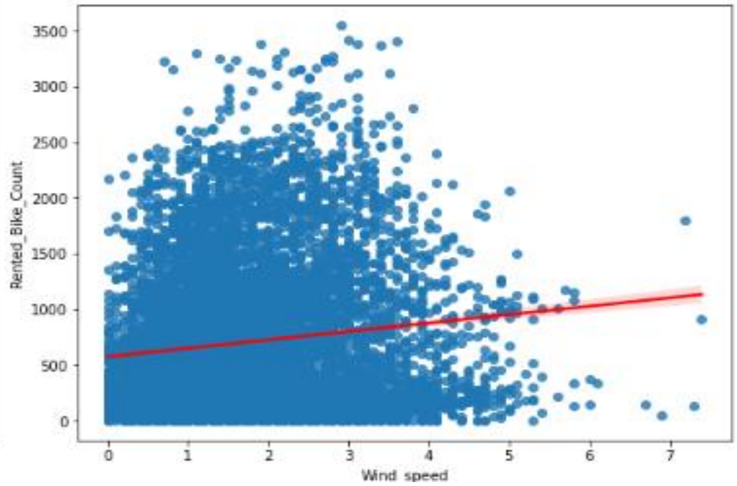
Thus from Univariate analysis we can deduce that the Distribution among different variables is not perfectly Normal.

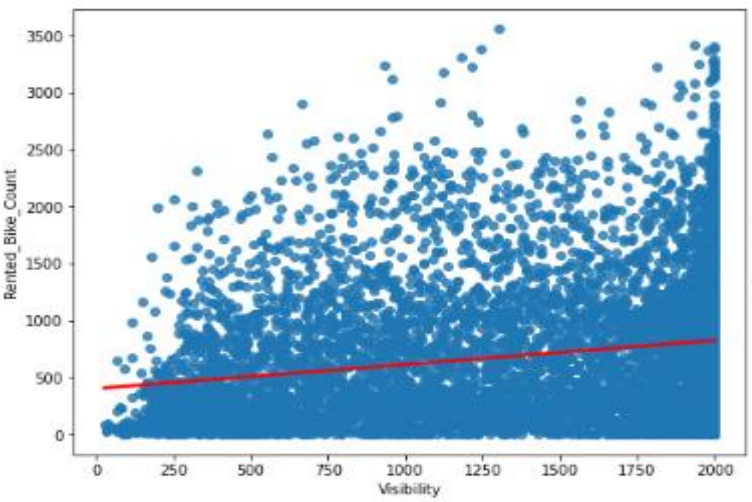
- **Bivariate Analysis:**

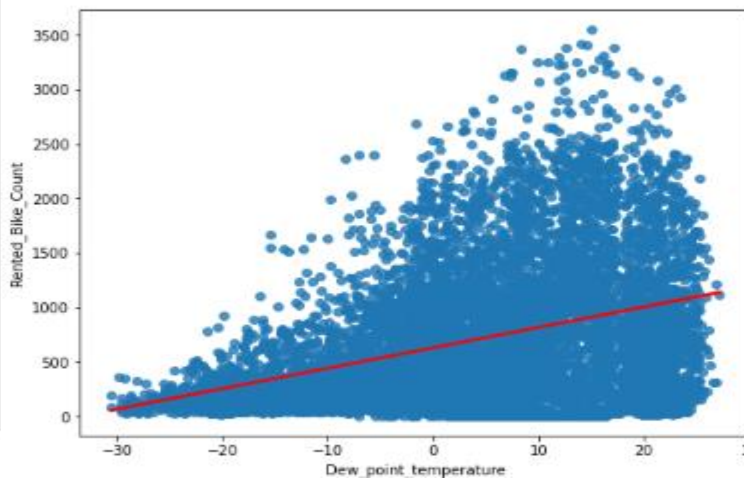


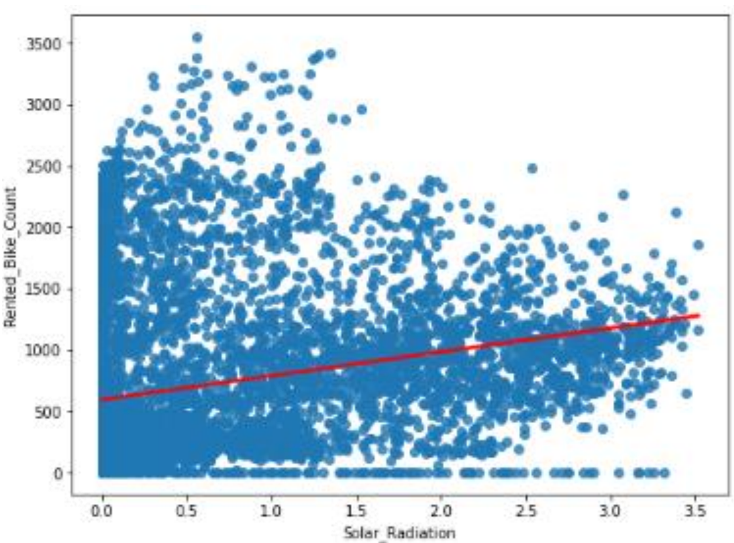


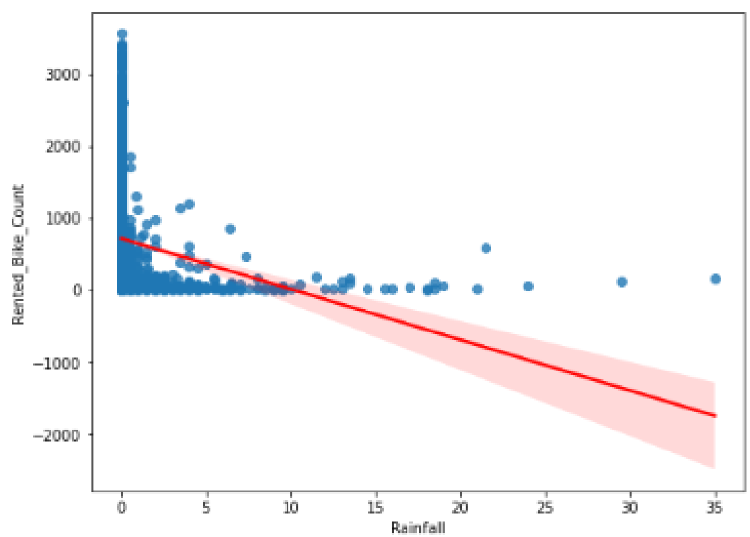

Rented_Bike_Count
Humidity

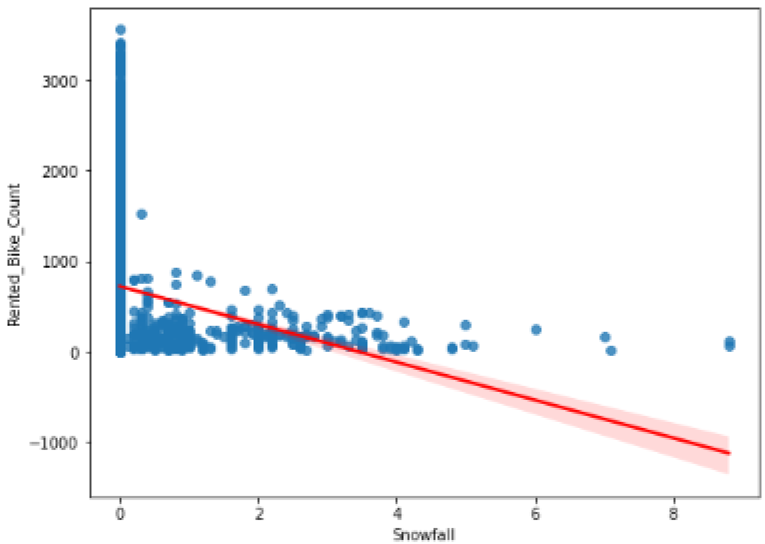

Rented_Bike_Count
Wind_speed

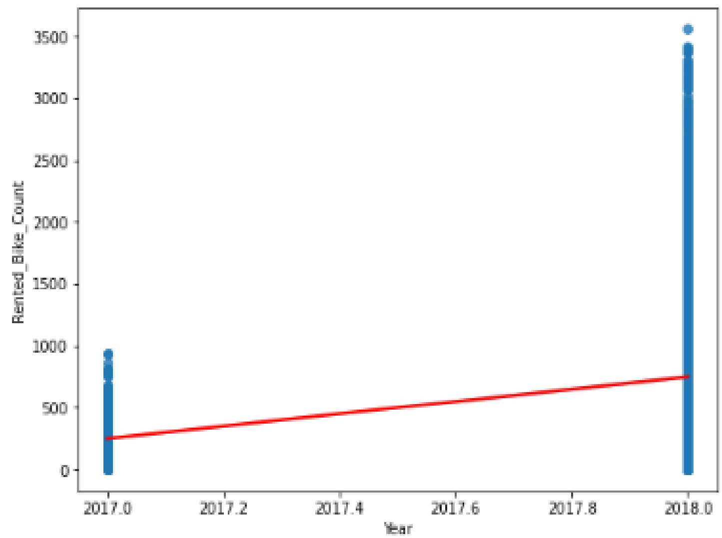

Rented_Bike_Count
Visibility


Rented_Bike_Count
Dew_point_temperature


Rented_Bike_Count
Solar_Radiation


Rented_Bike_Count
Rainfall

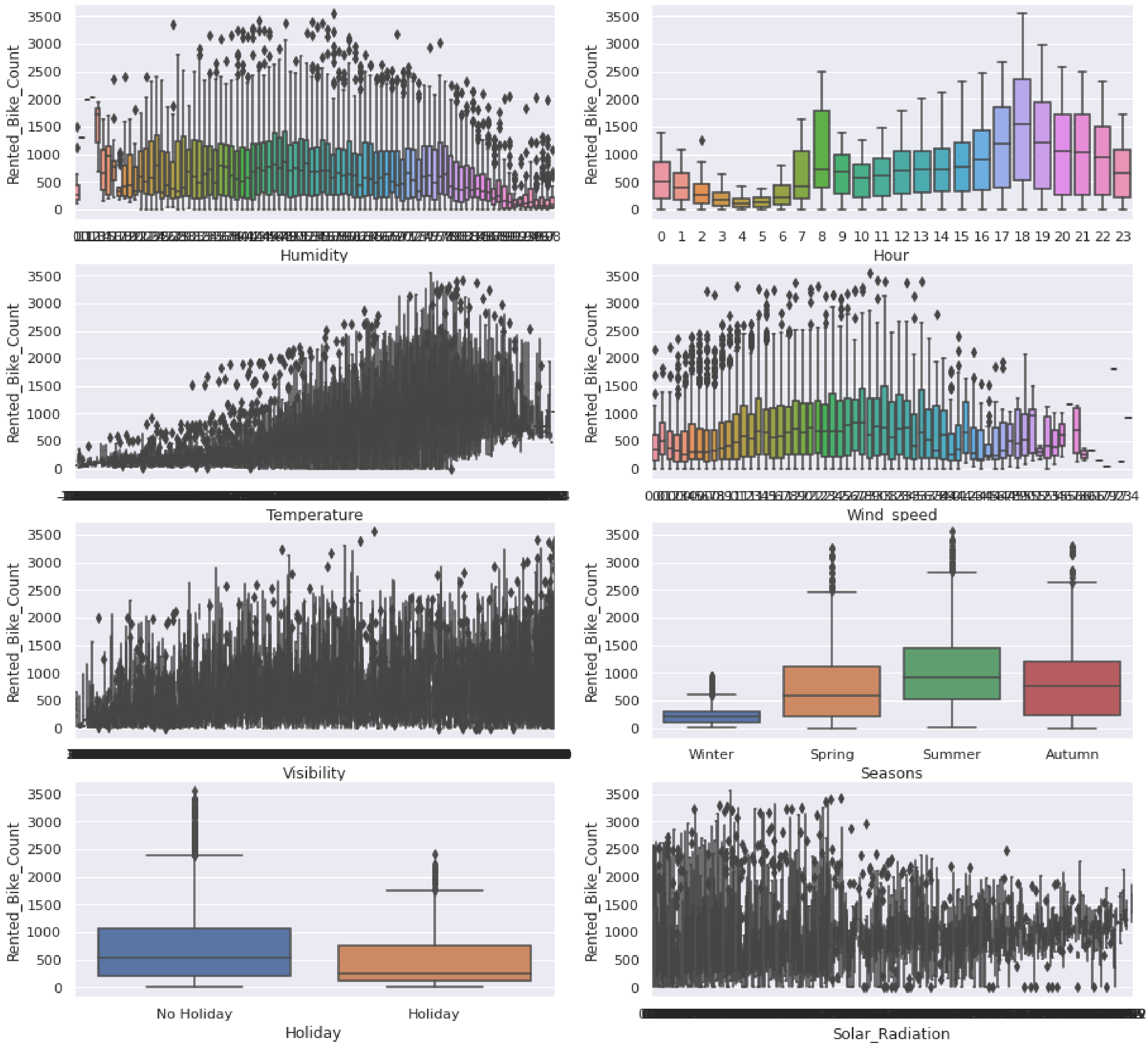




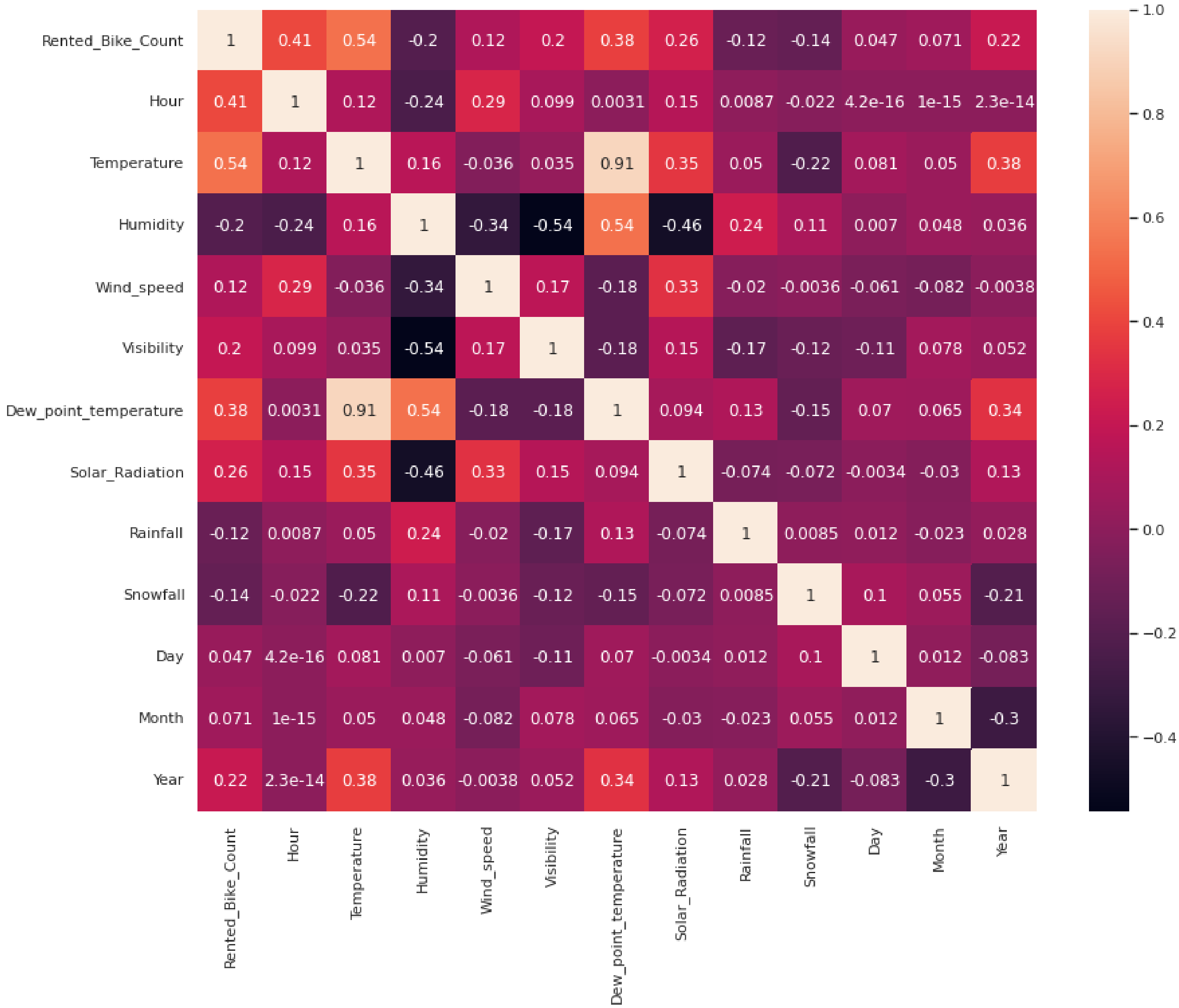
From Bivariate Analysis We have found The different Features have different impact on Rented Bike count.

- **Hour:** Demand for bikes is mostly in the evening between 3 to 8 pm, also the least demand is at morning 5pm.
- **Temperature:** People prefer to rent bikes at normal temperature of 20°C. to 30°C. Hence it is positively related to Rented Bike.
- **Humidity:** It is negatively correlated , as people prefer to rent a bike less if there is more moisture in the air.
- **Wind_speed:** Wind Speed doesn't affect much for renting a bike but is slightly positively correlated.
- **Visibility:** It does not affect, similar to wind speed, it is positively correlated.
- **Dew point temperature:** The dew point is the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity. It is positively correlated with data.
- **SnowFall and Rainfall**: People don't prefer to rent a bike, when there is rainfall or snowfall.

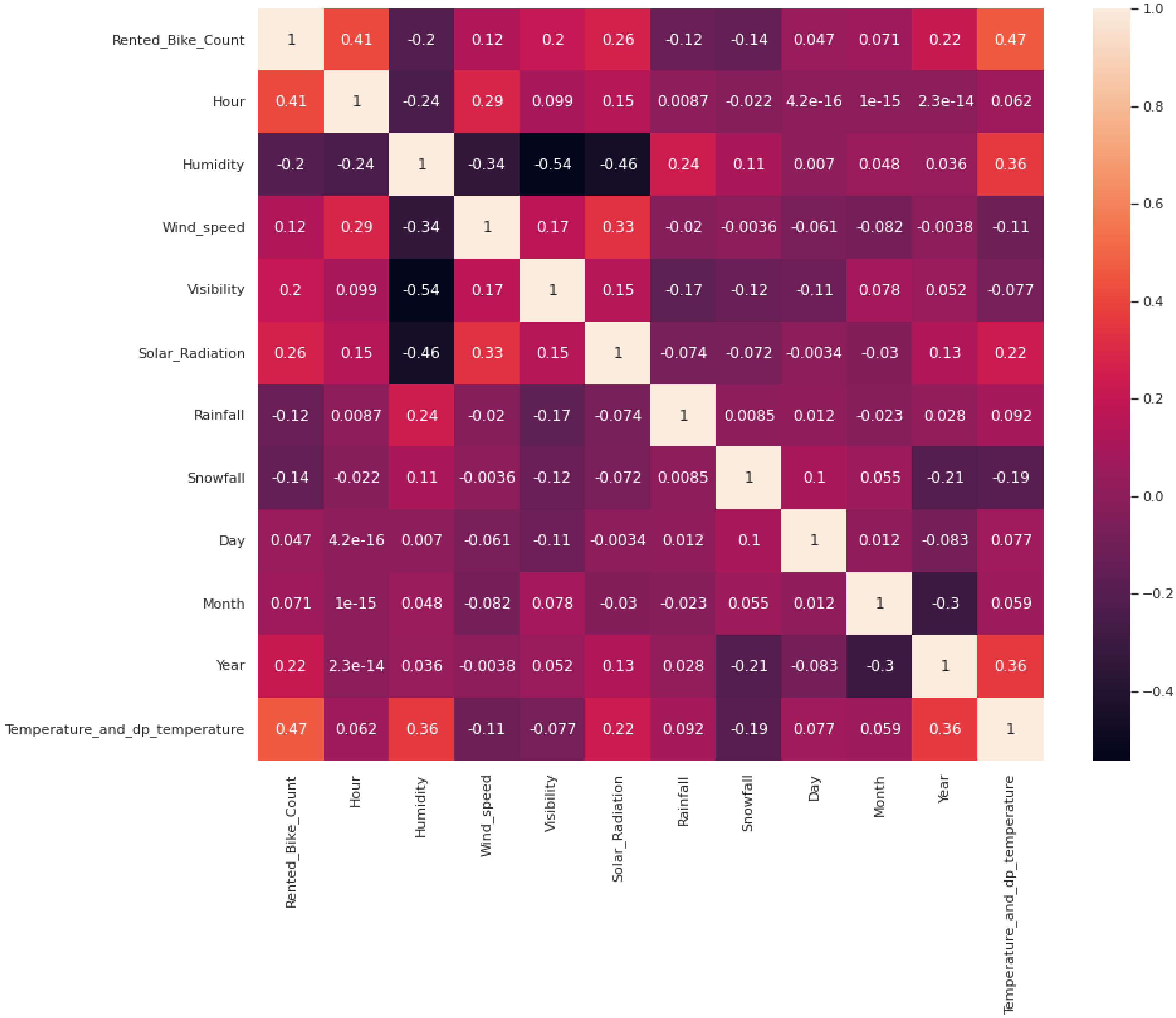
Outlier detection



Correlation



Temperature and Dew Point temperature are highly correlated.We can add them to make one single column Thus removing them and then analyzing the data.



Highest correlation is shown by Temperature and Dp Temperature.

3.3 Data Preparation :

One hot encoding

creating dummies column for the given feature

	Rented_Bike_Count	Hour	Humidity	Wind_speed	Visibility	Solar_Radiation	Rainfall	Snowfall	Day	Month	Year
0	254	0	37	2.2	2000	0.0	0.0	0.0	12	1	2017
1	204	1	38	0.8	2000	0.0	0.0	0.0	12	1	2017
2	173	2	39	1.0	2000	0.0	0.0	0.0	12	1	2017
3	107	3	40	0.9	2000	0.0	0.0	0.0	12	1	2017
4	78	4	36	2.3	2000	0.0	0.0	0.0	12	1	2017

Temperature_and_dp_temperature	Seasons_Spring	Seasons_Summer	Seasons_Winter	Holiday_No Holiday	Functioning_Day_Yes
-22.8	0	0	1	1	1
-23.1	0	0	1	1	1
-23.7	0	0	1	1	1
-23.8	0	0	1	1	1
-24.6	0	0	1	1	1

4. Model Building

4.1 Implementing Linear Regression

For Linear Regression to be implemented we have to take certain assumptions.

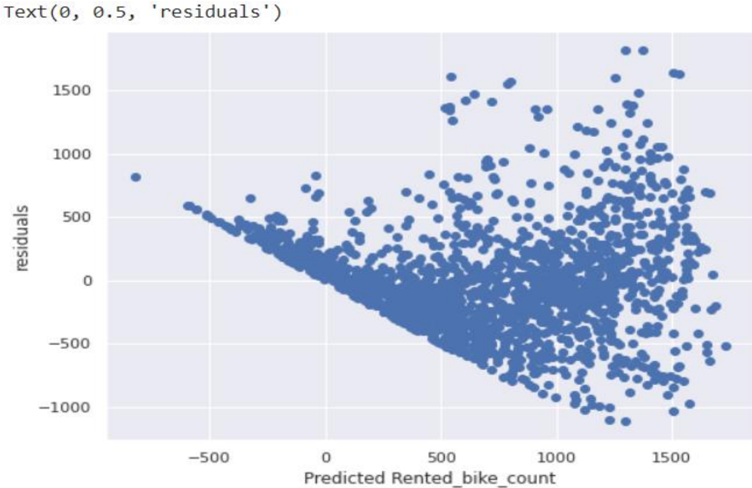
1. **Linear relationship** - There should be a linear relationship between feature variable and dependent variable.
2. **Little or no-multicollinearity** - There should not be multicollinearity among variables.
3. **Little or no auto-correlation** - Another assumption is that there is little or no autocorrelation in the data. Autocorrelation occurs when the residual errors are not independent from each other.
4. **Homoscedasticity** - Variance should be the same, i.e. error term should be same across all values of the independent variable.

Regression Test data metrics:

```
MSE : 169871.70934024057
RMSE : 412.15495792267325
R2: 0.5624656403341544
Adjusted R2 : 0.5584307413401177
```

As we can see Using Regression only 56% Score can be achieved and the graph is slightly matching the actual parameters. Also Heteroscedasticity is much i.e variance is not same across all points.


```
Text(0, 0.5, 'residuals')
```



4.2 Implementing Lasso Regression using cross validation.

The best fit alpha value is found out to be : 0.01

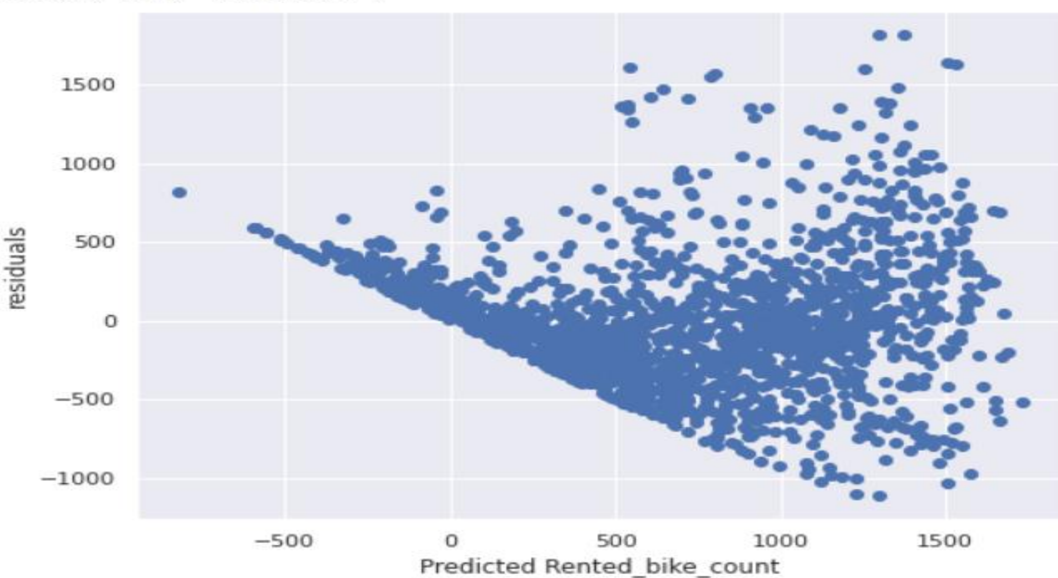
the negative mean squared error is: **- 191791.758**

Evaluation Metrics:

```
MSE : 169873.39789560612
RMSE : 412.15700636481495
R2: 0.5624612911638689
Adjusted R2 : 0.5584263520622101
```

We only able to achieve 56.246% of score.also predicted graph is not perfectly overlapping to actual graph, heteroscedasticity exist in data


```
Text(0, 0.5, 'residuals')
```



4.3 Implementing Ridge Regression Using Cross Validation.

The best fit alpha value is found out to be : 0.01

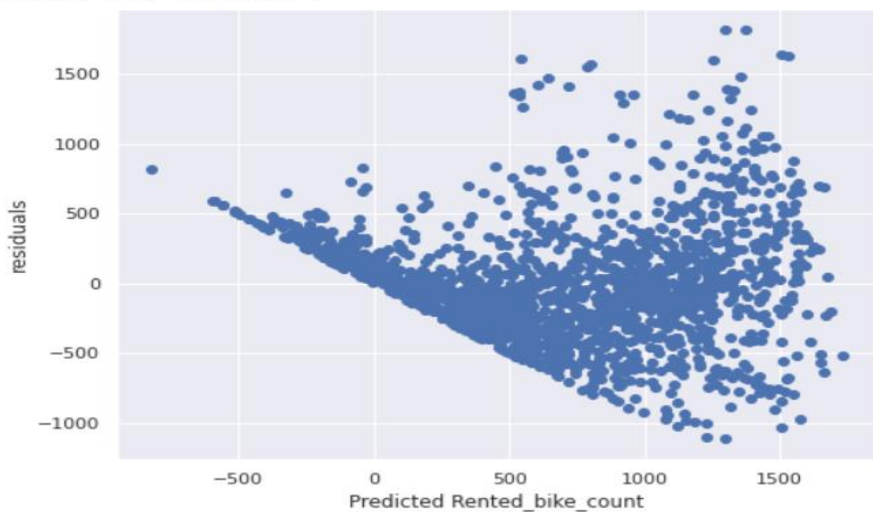
the negative mean squared error is: - 191771.27

Evaluation Metrics:

```
MSE : 169871.71502247467
RMSE : 412.1549648159957
R2: 0.5624656256985645
Adjusted R2 : 0.5584307265695599
```

Achieved 56.246% of R2 Score quite good comparative to Lasso.
Still the Predicted line of graph is not matching perfectly with actual one.
Heteroscedasticity exists in data.


```
Text(0, 0.5, 'residuals')
```



4.4 Implementing Elastic Net Regressor Using Cross validation

The best fit alpha value is found to be :

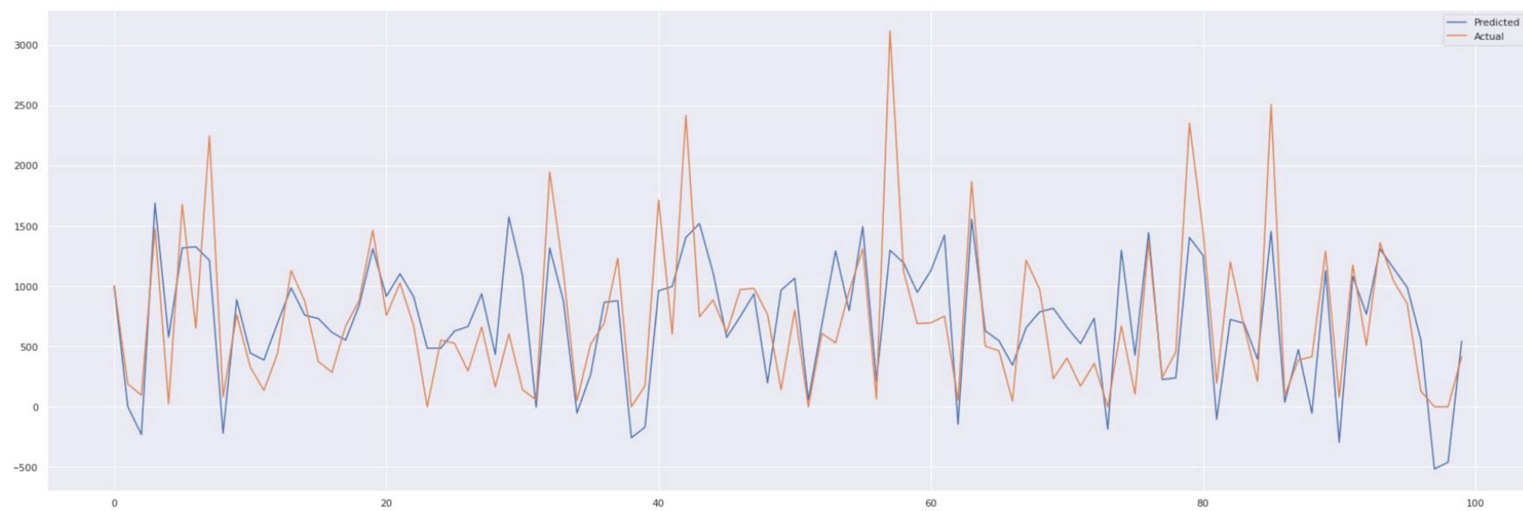
Alpha : 0.0001, L1 ratio : 0.6

The negative mean squared error is : - `191856.19`

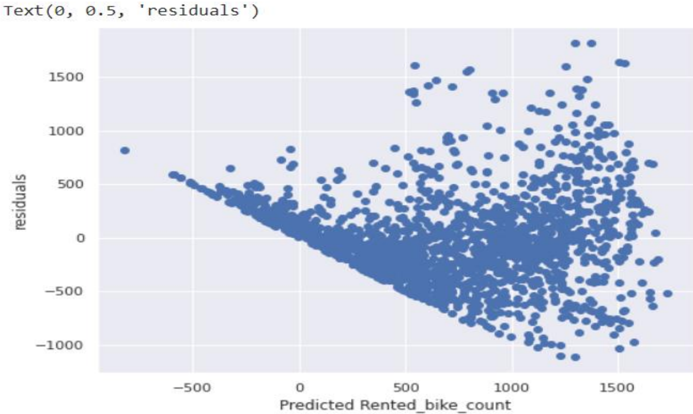
Evaluation Metrics:

```
MSE : 169871.9503256486
RMSE : 412.1552502706336
R2: 0.5624650196340002
Adjusted R2 : 0.5584301149159276
```

Using ElasticNet our score decreases.hence the predicted graph also not to the exact of Actual one, Heteroscedasticity exists.




```
Text(0, 0.5, 'residuals')
```



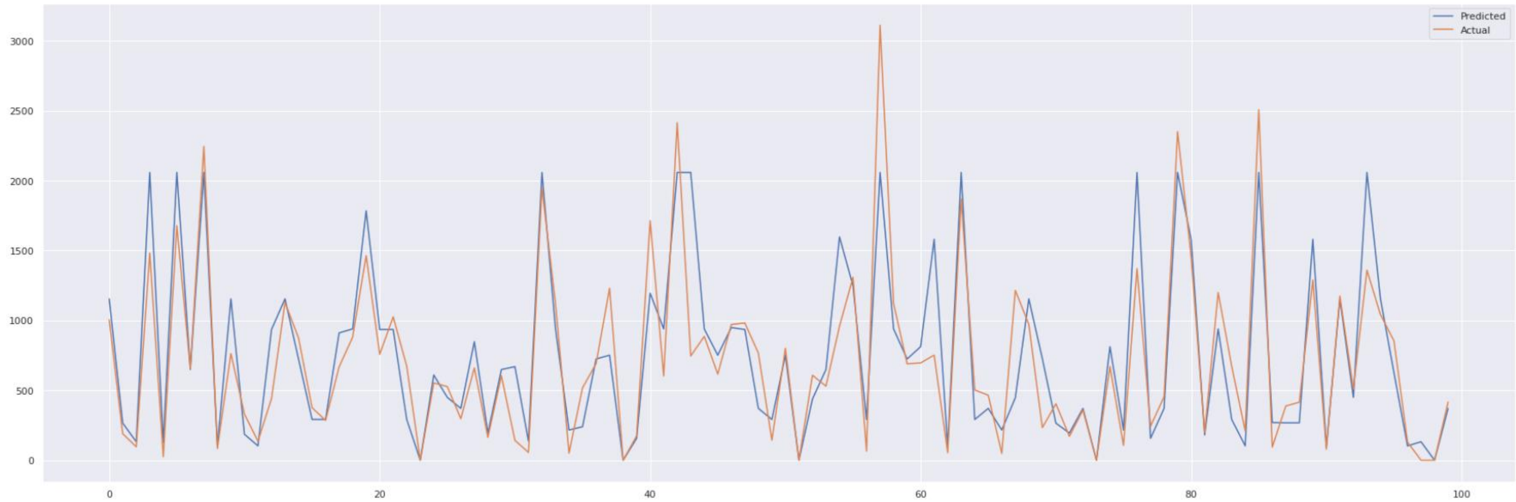
4.4 Implementing of decision tree by using decision tree regressor

DecisionTreeRegressor(criterion='mse', max_depth=8, max_features=9,
max_leaf_nodes=100)

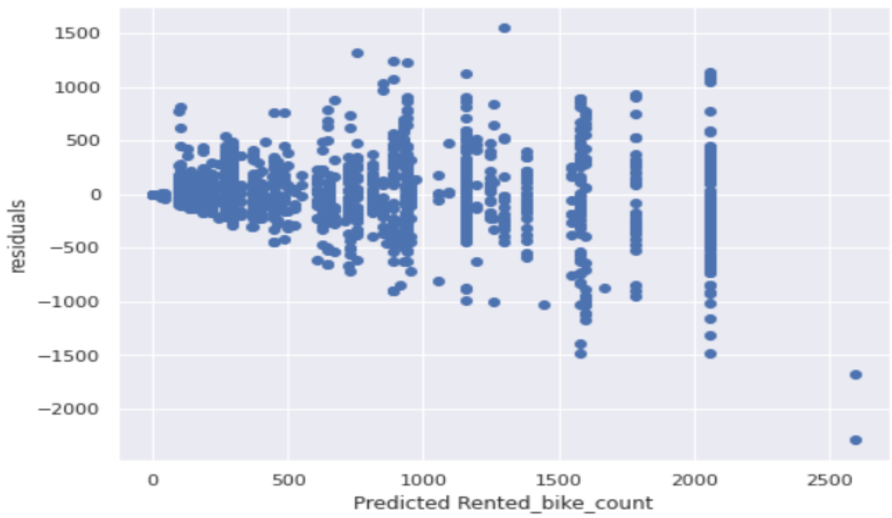
Evaluation Metrics :

```
MSE : 95207.6730809099
RMSE : 308.55740645933275
R2 : 0.7547759515782825
Adjusted R2 : 0.752514519431454
```

Using Decision Tree our score jumped to **75.477**%, Hence the predicted graph comes closure to the Actual one, Heteroscedasticity also gets improved.



Text(0, 0.5, 'residuals')

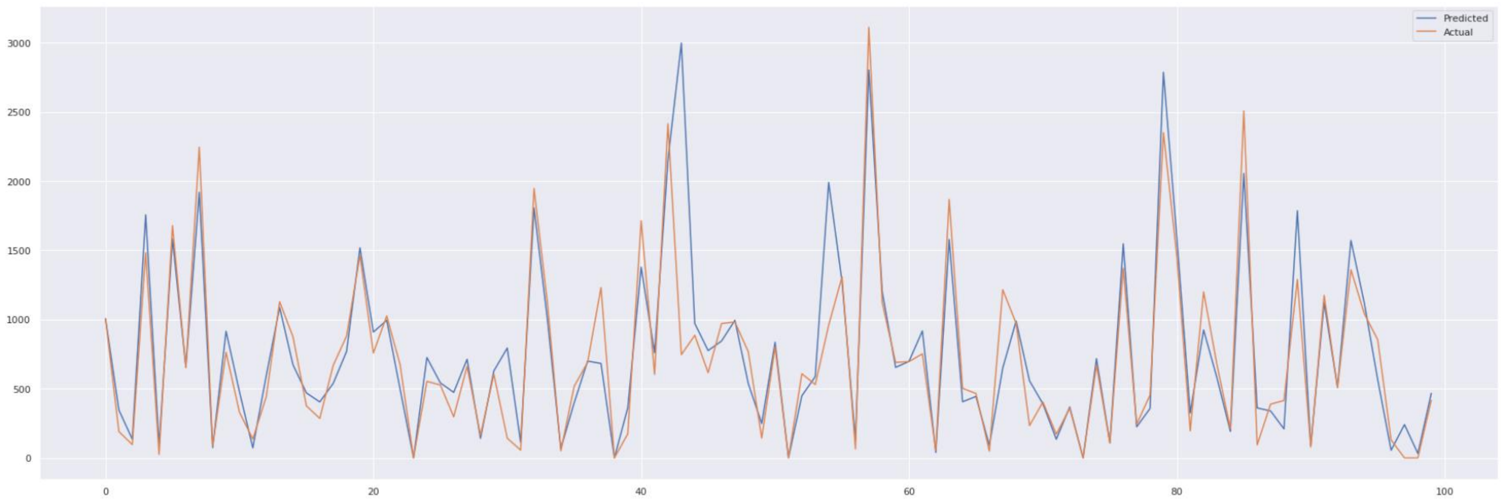


4.5 Implementing Random forest Regressor Using Grid Search cross validation

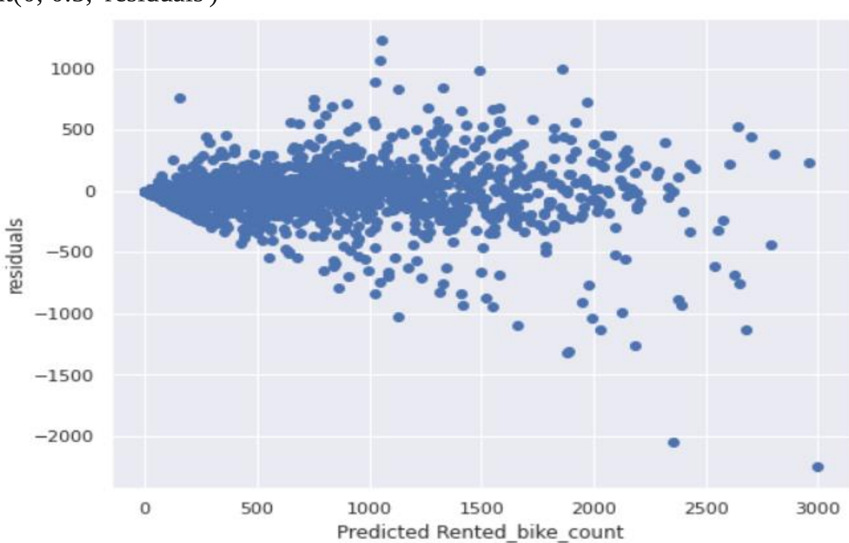
Evaluation Metrics :

```
Model Score: 0.8645485542753845
MSE : 52588.712428025116
RMSE : 229.3222894269659
R2 : 0.8645485542753845
Adjusted R2 : 0.8632994343148117
```

Using Random Forest Regressor our score jumped to 86.454%, Hence the predicted graph comes closure to the Actual one, Heteroscedasticity also gets improved.



Text(0, 0.5, 'residuals')

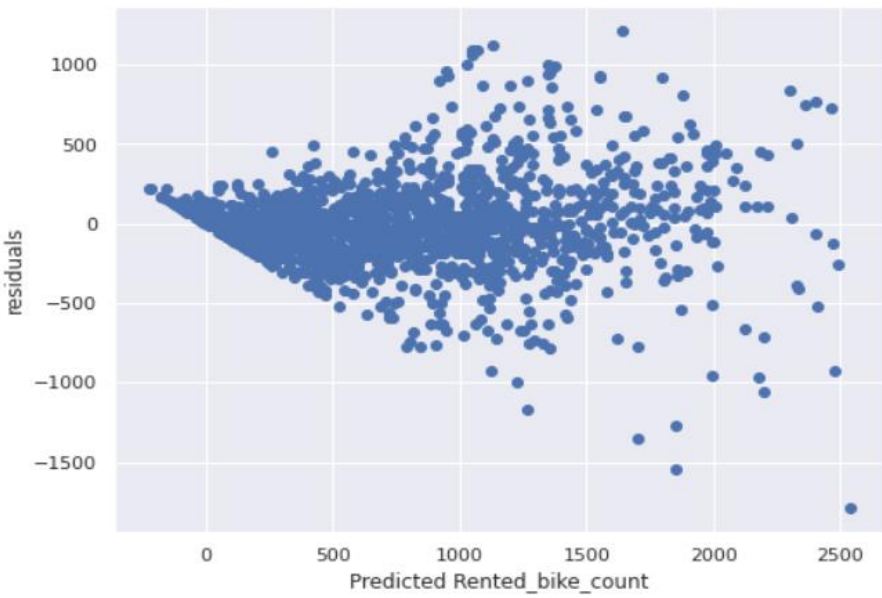


4.6 Implementing Gradient Boost

Evaluation Metrics :

```
Model Score: 0.8557050771607113
MSE : 67959.29753281914
RMSE : 260.69004110786267
R2 : 0.8249589184399939
Adjusted R2 : 0.8233447067368469
```

Using Gradient Boost we were able to achieve 82.49% of the score.And the predicted graph also seems more accurately overlaps Actual one.



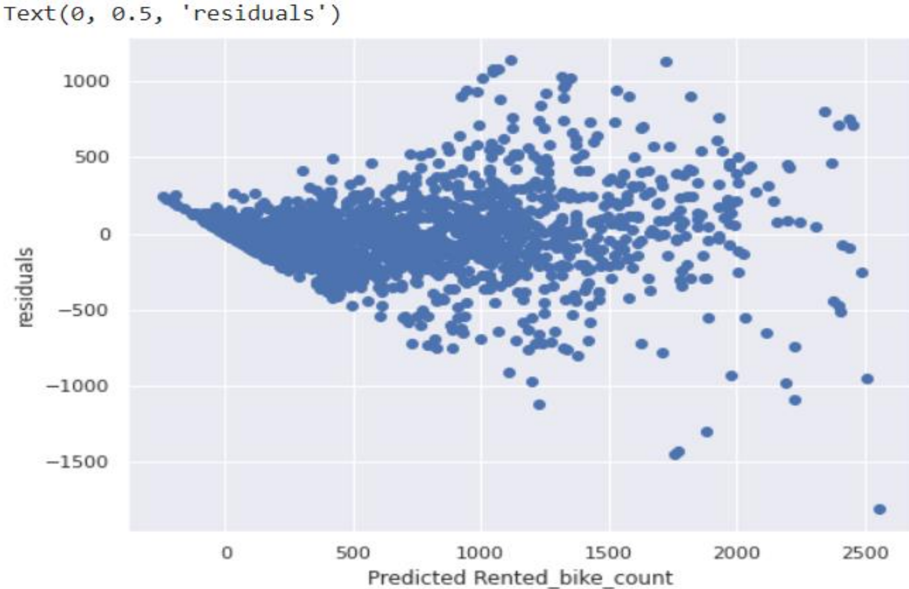
4.6 Implementing Xg Boost using GridSearchCV

Evaluation Metrics

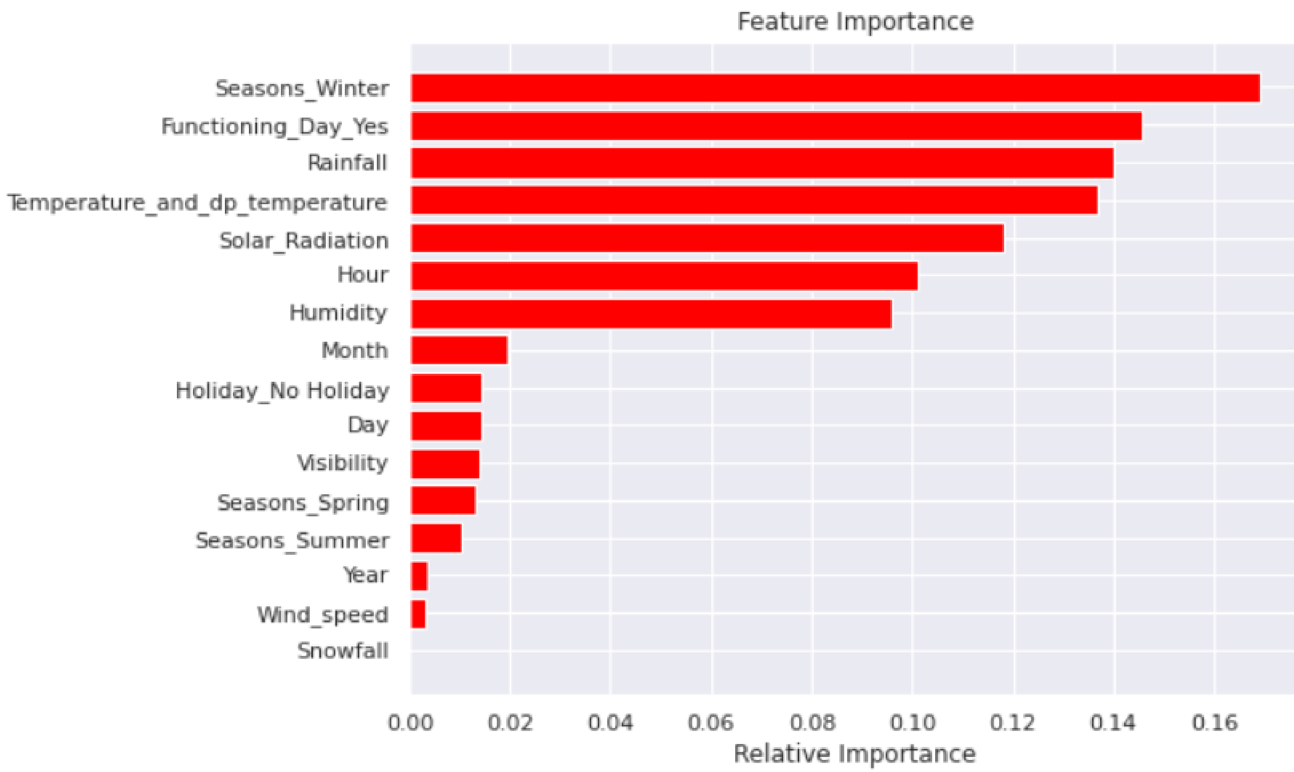
```
MSE : 68250.19807949613
RMSE : 261.2473886558412
R2 : 0.8242096530978654
Adjusted R2 : 0.8225885317431483
```

Using Xg Boost the score rose up to 82.42%. And this is the best algo to use.


```
Text(0, 0.5, 'residuals')
```



Feature Importance in XgBoost:



The most important Feature for Rental Bike

Model Summary:

For Train Data :

SL NO	MODEL_NAME	Train MSE	Train RMSE	Train R^2	Train Adjusted R^2
1	Linear Regression	190710.62548259995	436.70427692272483	0.5488899632663061	0.5478575271931064
2	Lasso Regression	190710.6254850217	436.7042769254976	0.5488899632605777	0.5447298707027501
3	Ridge Regression	190710.63414152752	436.7042868366734	0.5488899059814547	0.5447298128954048
4	ElasticNet Regression	203220.4874674924	450.7998308201683	0.5192989308565615	0.5148659526973137
5	DecisionTree Regressor	76328.2706429576	276.27571489900737	0.8194518586133761	0.8177868613441046
6	RandomForest Regressor	7014.866722873859	83.75480119296958	0.983406919373109	0.9832538996094028
7	Gradient Boost	61001.9124998751	246.9856524170485	0.8557050771607113	0.8543744035206948
8	Xg Boost	61295.43648699404	247.57915196355697	0.855010770714616	0.8536736942485836

For Test Data:

SL NO	MODEL_NAME	Test MSE	Test RMSE	Test R^2	Test Adjusted R^2
1	Linear Regression	169871.70934024057	412.15495792267325	0.5624656403341544	0.5584307413401177
2	Lasso Regression	169871.72597325977	412.1549781007865	0.5624655974928983	0.5584306981037839
3	Ridge Regression	169871.77465071727	412.1550371531534	0.5624654721155751	0.5584305715702432
4	ElasticNet Regression	183719.51037262668	428.6251396880807	0.5267982017652659	0.5224343811475394
5	DecisionTree Regressor	91626.47632392391	302.6986559664973	0.7639999514779179	0.7618235821543713
6	RandomForest Regressor	52363.04539891553	228.8297301464902	0.8651297992599828	0.863886039483706
7	Gradient Boost	67959.97028150507	260.69133142761973	0.8249571856578451	0.823342957975151
8	Xg Boost	68250.19807949613	261.2473886558412	0.8242096530978654	0.8225885317431483

Conclusion :

- As it was stated in the problem, rented bike count was low in 2017 untill november. After that rented bike count started increasing.
- There was sharp increase in demand from the end of 2017 that too in winter season of the year. The demand however decrease at the end of 2018.
- Bike count rent is highly correlated with 'Hour', which seems obvious. Demand for bike is mostly in morning (7 to 8) and in the evening (3 to 9).
- After doing exploratory data analysis, applying Linear Regression model didn't go quite well as it gave only 56% accuracy.
- Lasso and Ridge Regression helps to reduce model complexity and prevent over-fitting which may result from simple linear regression. with Lasso, ridge and ElasticNet regressor We got r squared value of 0.5624, 0.5624, 0.5267 respectively.
- With Decision Tree we are able to achieve the r2 score of 0.7639.
- Gradient Boost gave r squared value of 0.8249 on test data.
- XG Boost gave r squared value of 0.8242
- RandomForest Regressor gives higher value of R squared metric in train data 0.9834 and on test data it is 0.8651
- RandomForest Regressor came with best accuracy to approximate numbers of rented bikes demand. It gives amazing results of training r-square at 0.9834 and test r-square value at 0.8651 also with adjusted r-square with 0.8638.

References:

- Geeksforgeeks
- Wikipedia
- Kaggle