Capstone project -2

Seoul Bike Sharing Demand Prediction

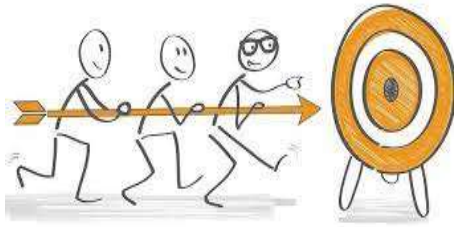
By –Ganesh Walimbe

Seoul Bike Sharing

- The public bicycle rental service in Seoul was started in April 2000. At that time, rental facilities were established in the center of two subway stations in Chang-dong and Yeouido. In 2004, Songpa District was designated as a special bicycle district, and the bicycle-free rental service was operated mainly in the park. In 2008, a full-fledged unmanned service was launched at subway stations and the city.
- In November 2007, Seoul announced its bicycle policy, introducing unmanned public bicycles such as Velib in Paris. The plan was to build 5102 bicycle stations at an interval of 300m within the bike-dedicated road network area and to have 82,400 bicycles. In October 2008, Seoul again announced the bicycle master plan and confirmed again that it intends to review the introduction of public rental bicycles by 2012. However, both of the presentations focused on the expansion of bicycle roads, and there was no progress in public bicycle rental service throughout the city. At the end of the year, the Seoul Facilities Corporation launched an unmanned rental service for the Cheonggyecheon.

Topics of discussion

- ❑ Objectives
- ❑ Introduction to data
- ❑ Data preprocessing
- ❑ Exploratory data analysis
- ❑ Building prediction models
  - Multiple linear regression
  - Lasso regression
  - Ridge regression
  - Elastic net regression
  - Decision trees
  - Random forest
- ❑ Top 10 important factors according to the best performing model .
- ❑ Conclusions



Objectives

My objectives for this project is to

- Perform exploratory data analysis on the dataset to gain insights from the data and to find hidden patterns from the data
- The crucial part of the project is to the prediction of bike count required at each hour for the stable supply of rental bikes using supervised learning models .

Introduction to data

There are 14 variables present in the data set and those variables are :

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - 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)

Data pre-processing

- Checking if there are missing values in the dataset .

There were zero missing values present in the data.

Date	0
Rented Bike Count	0
Hour	0
Temperature(°C)	0
Humidity(%)	0
Wind speed (m/s)	0
Visibility (10m)	0
Dew point temperature(°C)	0
Solar Radiation (MJ/m2)	0
Rainfall(mm)	0
Snowfall (cm)	0
Seasons	0
Holiday	0
Functioning Day	0

- There were no duplicate values in the data ..

Data pre-processing

- **Modifying date column**

We modified the date column in hour, month and weekend or not

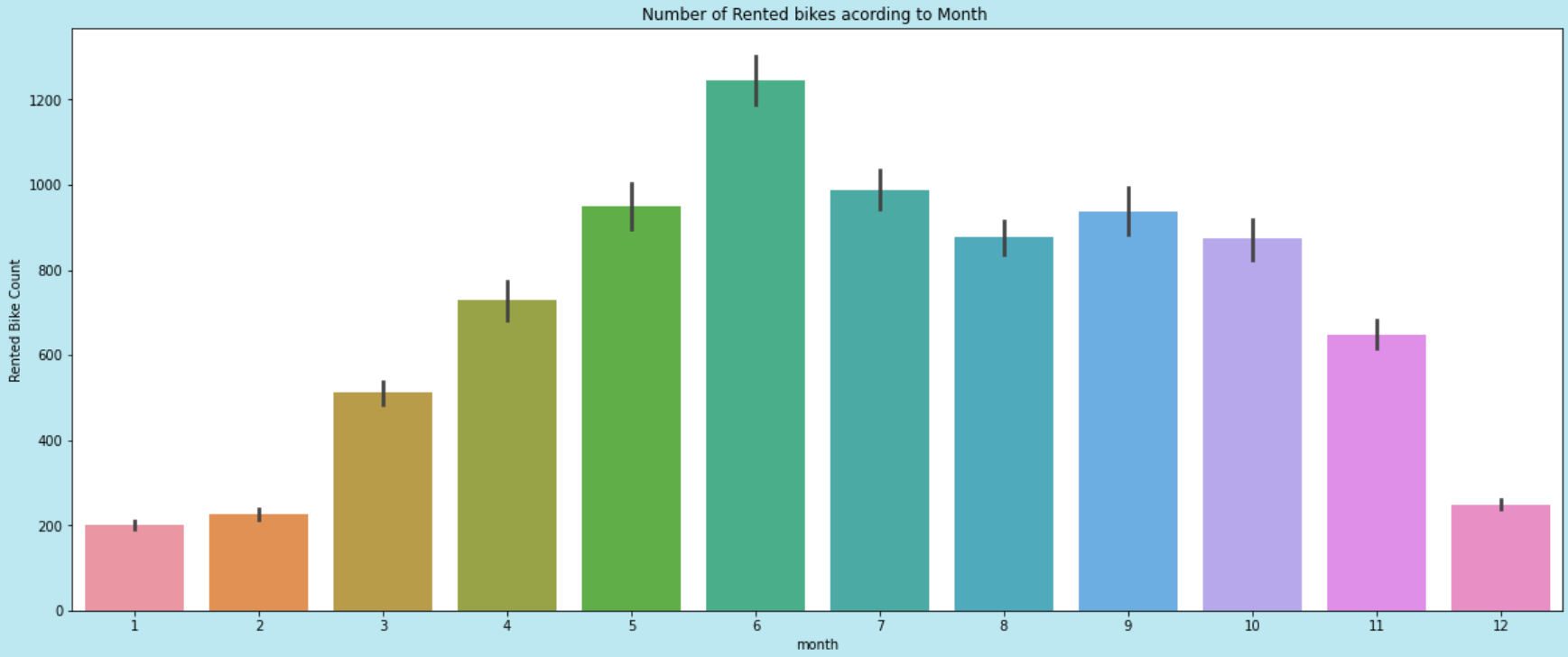
- **Feature engineering**

I performed feature engineering on categorical variables I used label encoding On variable functioning day and holiday. And i used dummies on seasons ,hour and month.

.

Exploratory data analysis(EDA)

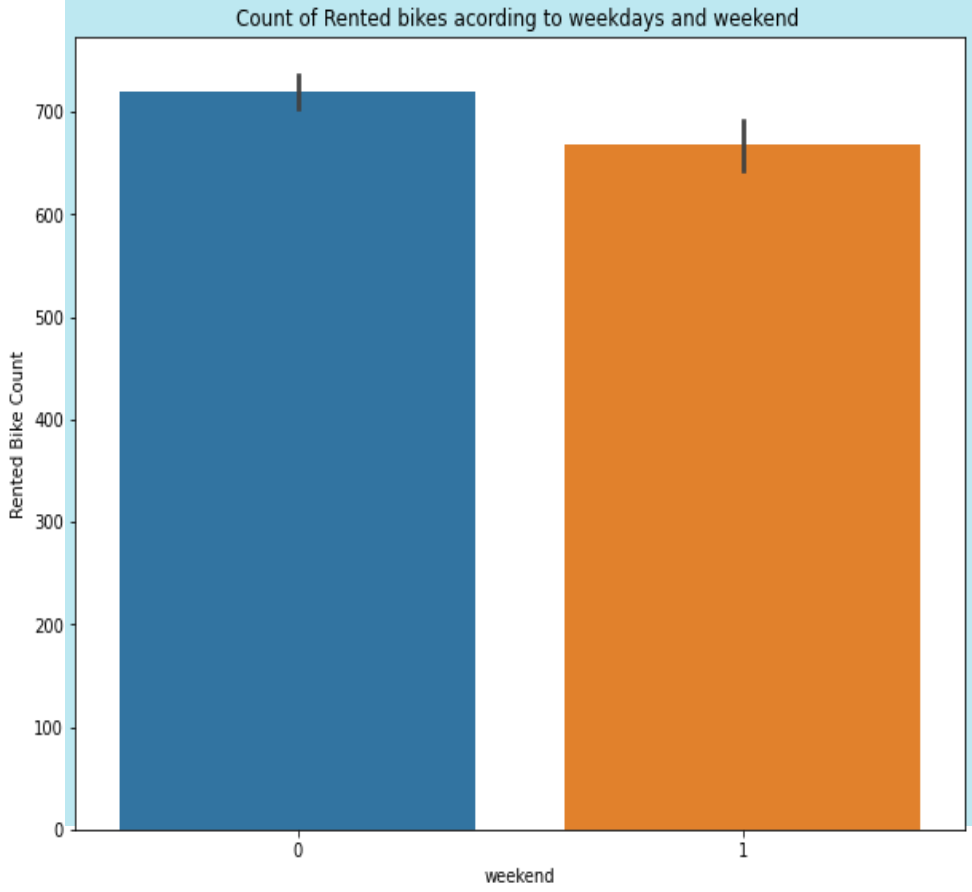
- Monthly rental of bikes analysis



- From the above barplot we can see that maximum number of bikes rented in the month of June and followed by May ,July ,August,september and October.

EDA

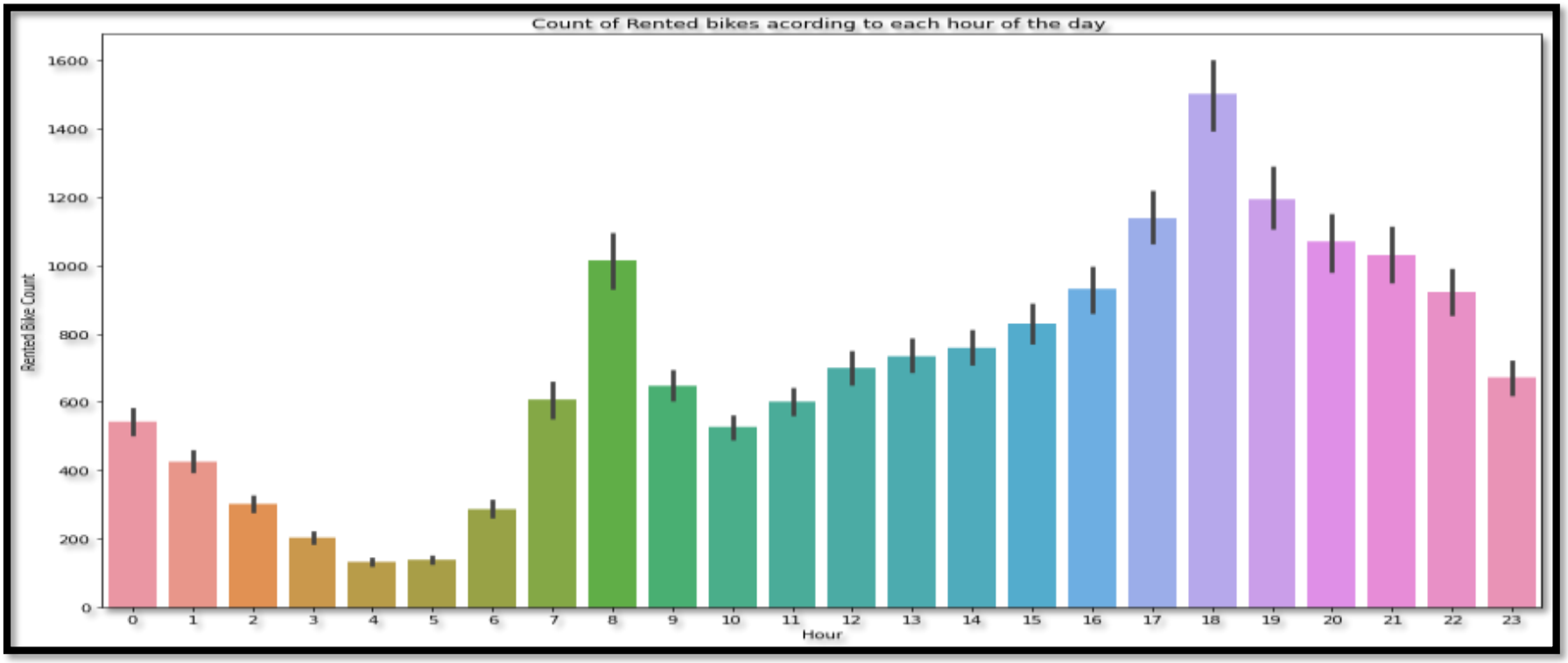
- weekend v/s rental bikes



From the bar plot we can say that people rent bikes slightly more bikes in weekdays as compared to weekend

EDA

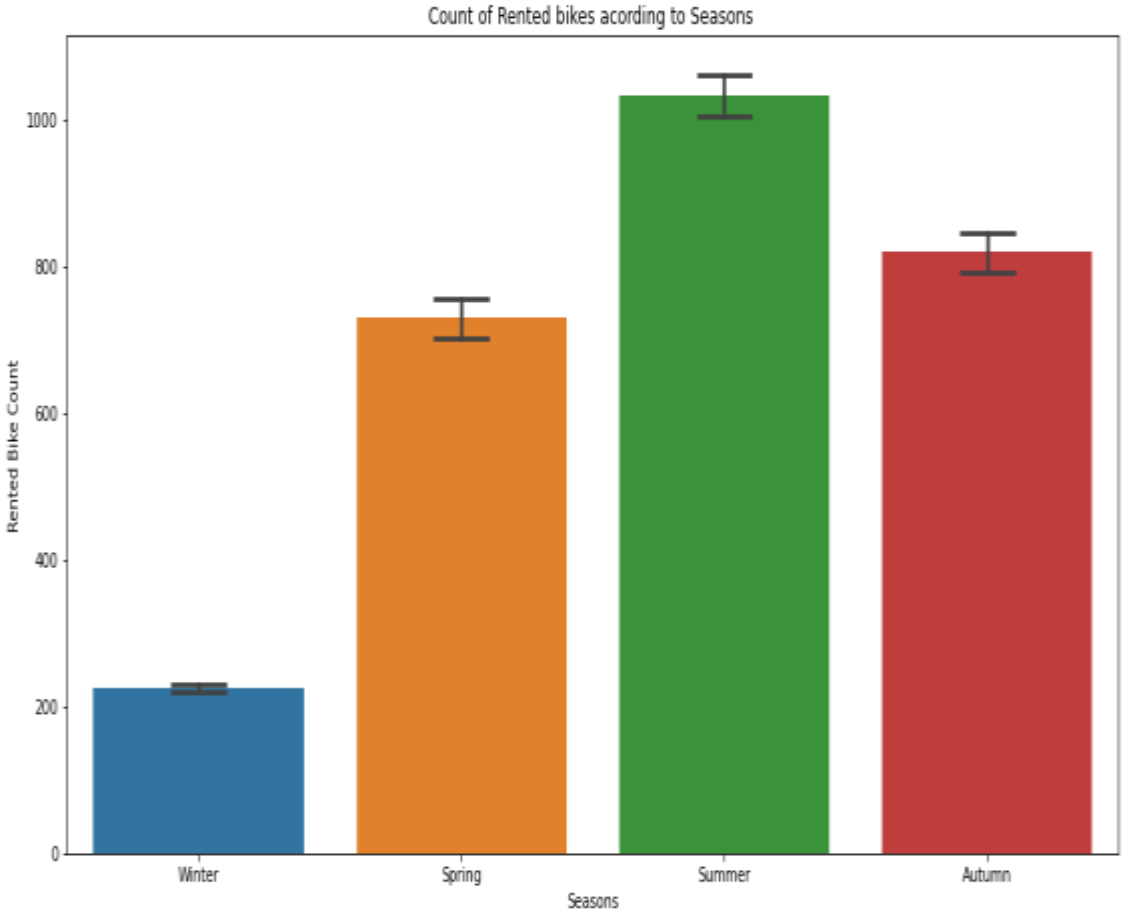
Rental bike count v/s hour



From the above plot we can say that peak time of renting bike at 7am to 9am in the morning and from 5pm to 10pm in the evening

EDA

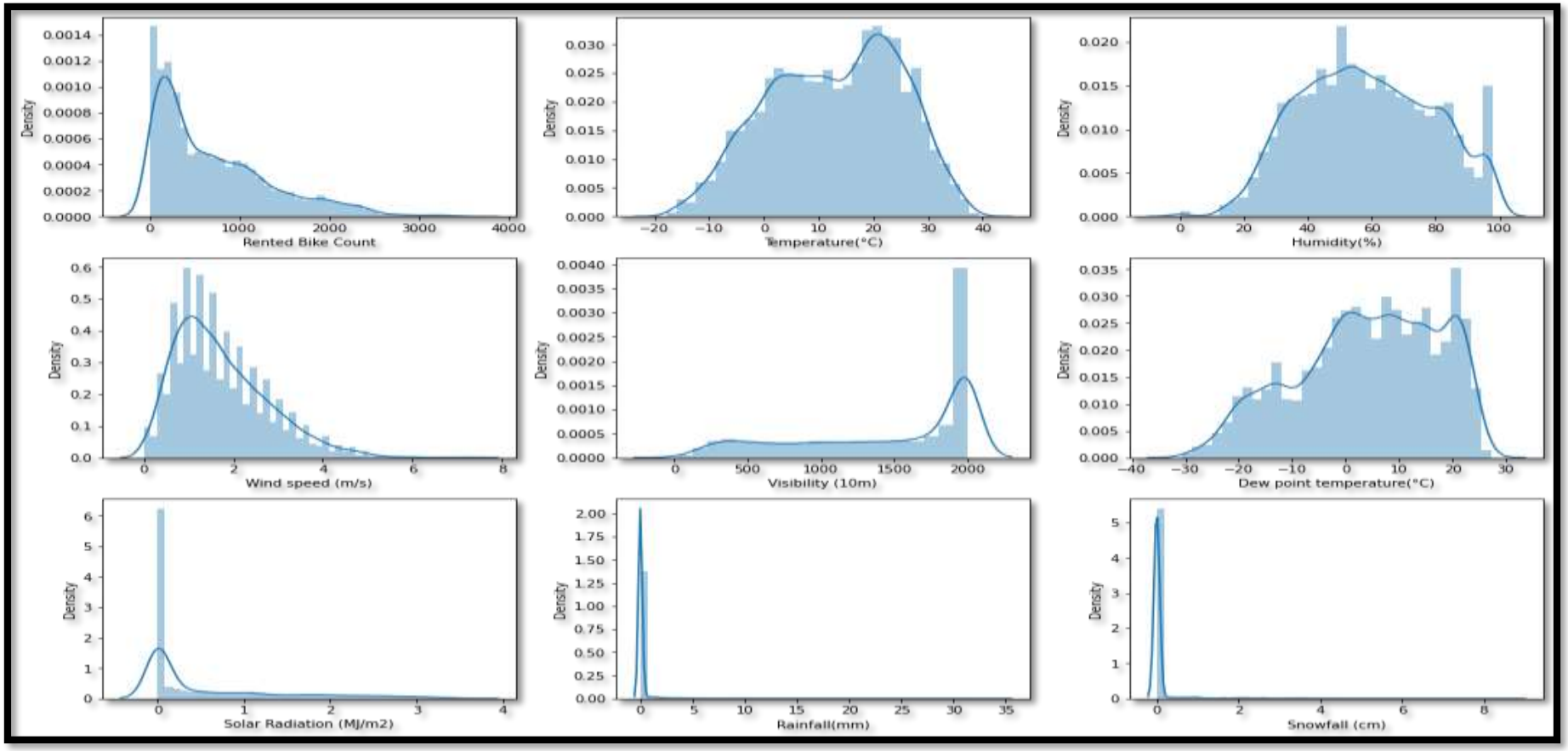
Count of Rented bikes according to Seasons



- Bar plot shows that the use of rented bike in in four different seasons, and it clearly shows that,
- In summer season the use of rented bike is high
- In winter season the use of rented bike is very low because of snowfall.

EDA

Analysis of numerical variables



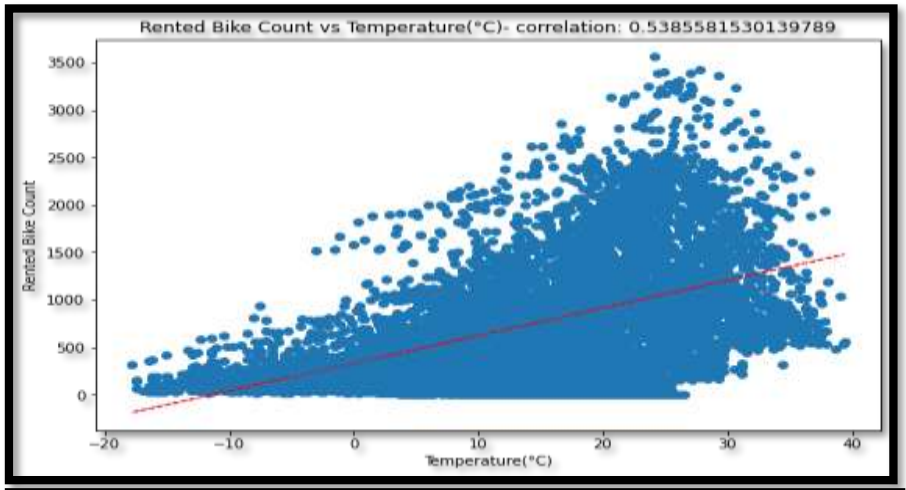
EDA

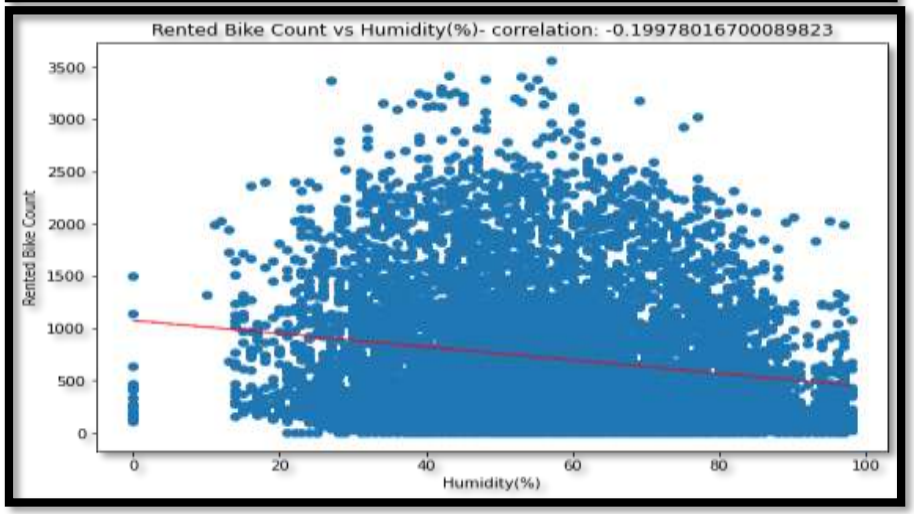
From above distribution plots we can say that,

- Rented bike count , Wind speed ,solar radiation ,rainfall ,snowfall are right skewed variables
- Visibility ,Dew point temperature are left skewed variables.
- Temperature ,Humidity and wind speed are approximately normal.

EDA

Correlation between rented bike count and different variables

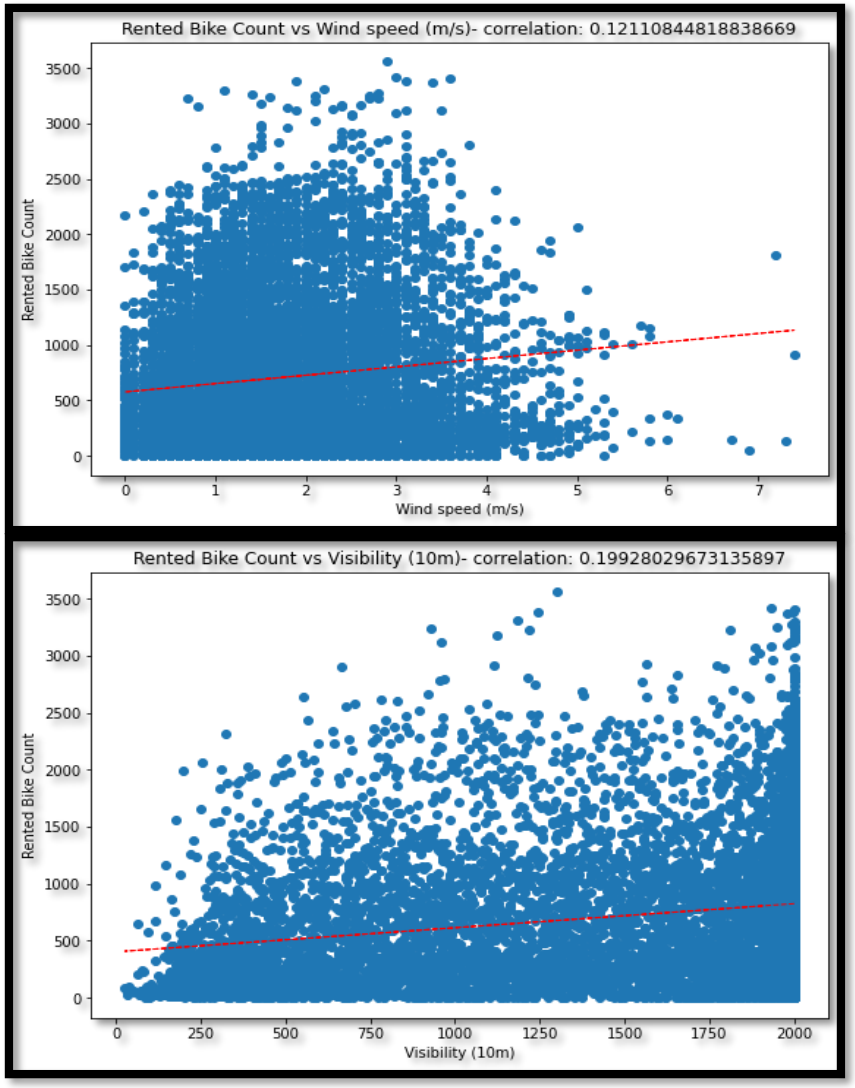




From scatter plot we can see that :

- Temperature is highly positively correlated with rental bike count with correlation 0.53 .
- Humidity is negatively moderately correlated with rental bike count with correlation -0.19

EDA

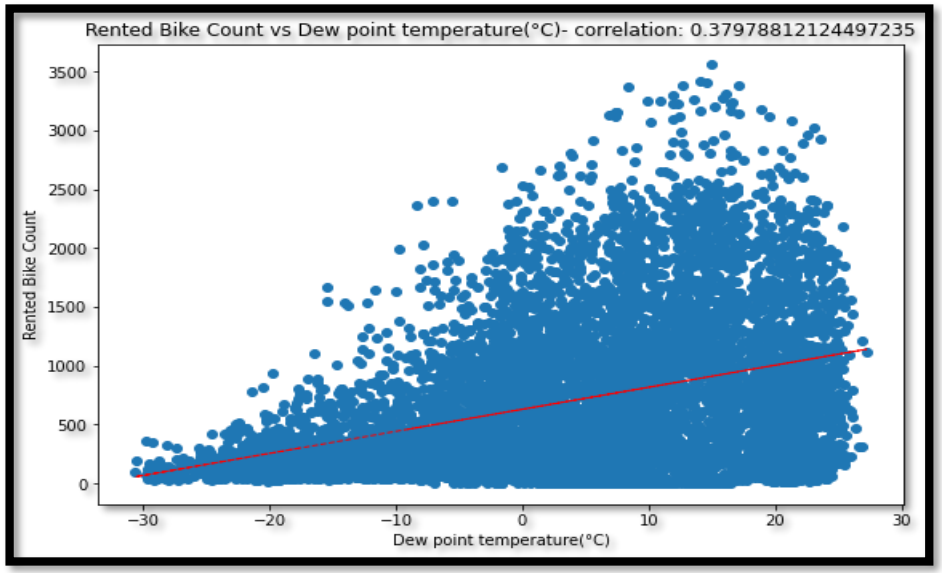


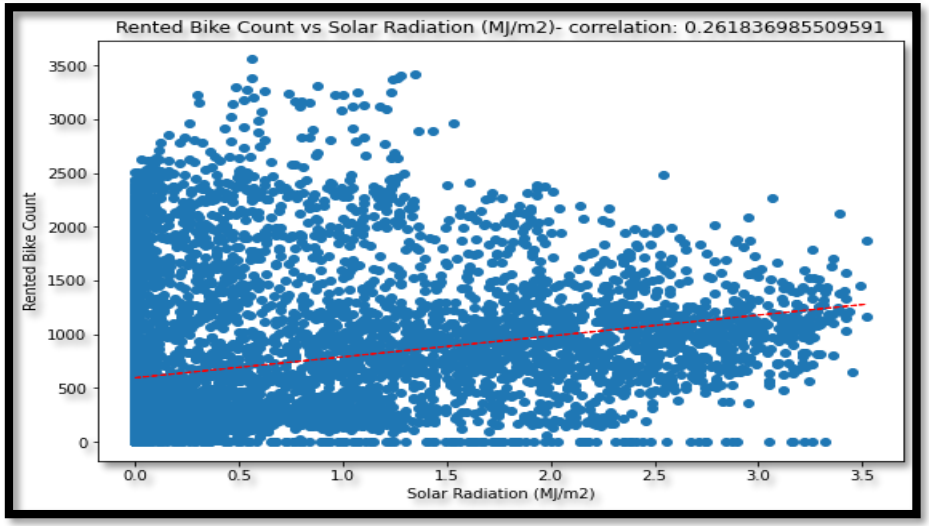
From scatter plot we can see that :

•Wind speed is positively correlated with rental bike count with correlation 0.12 .

•Visibility is moderately positively correlated with rental bike count with correlation 0.19 .

EDA

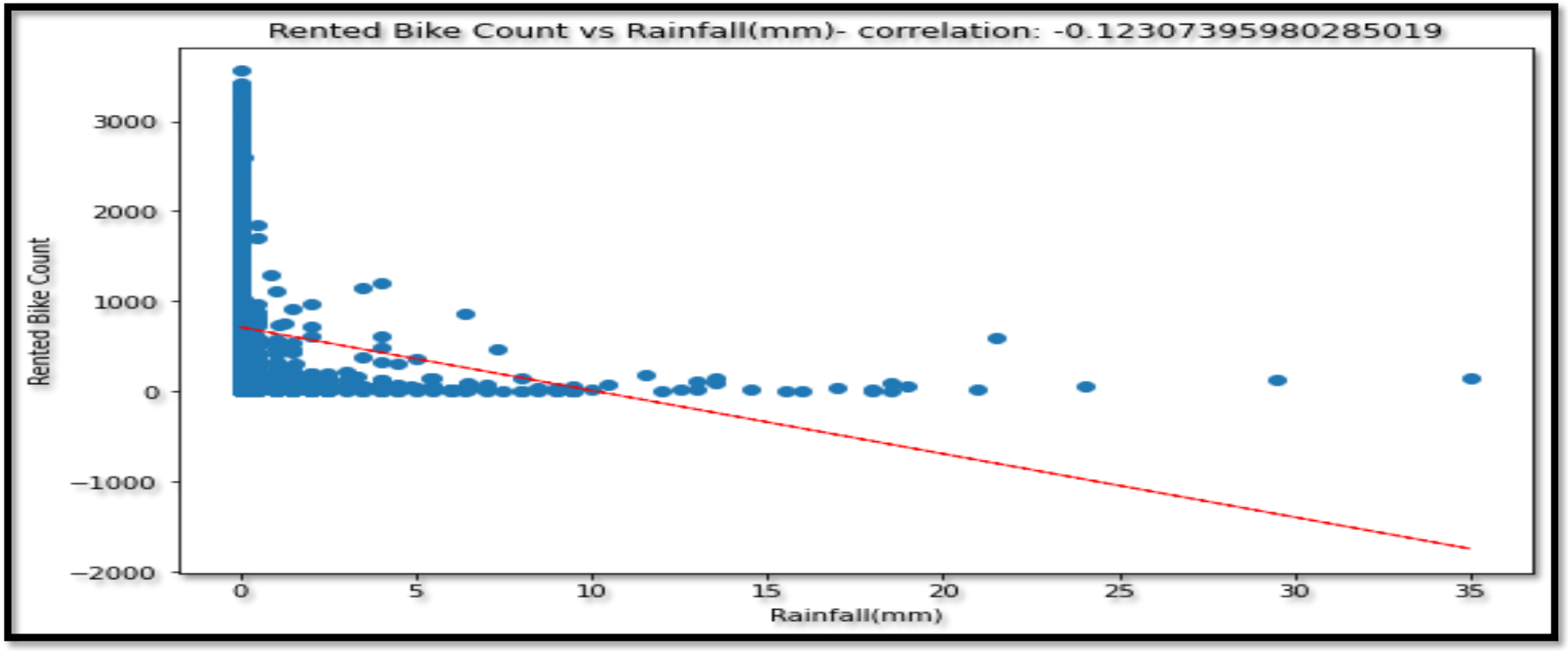




From scatter plot we can see that :

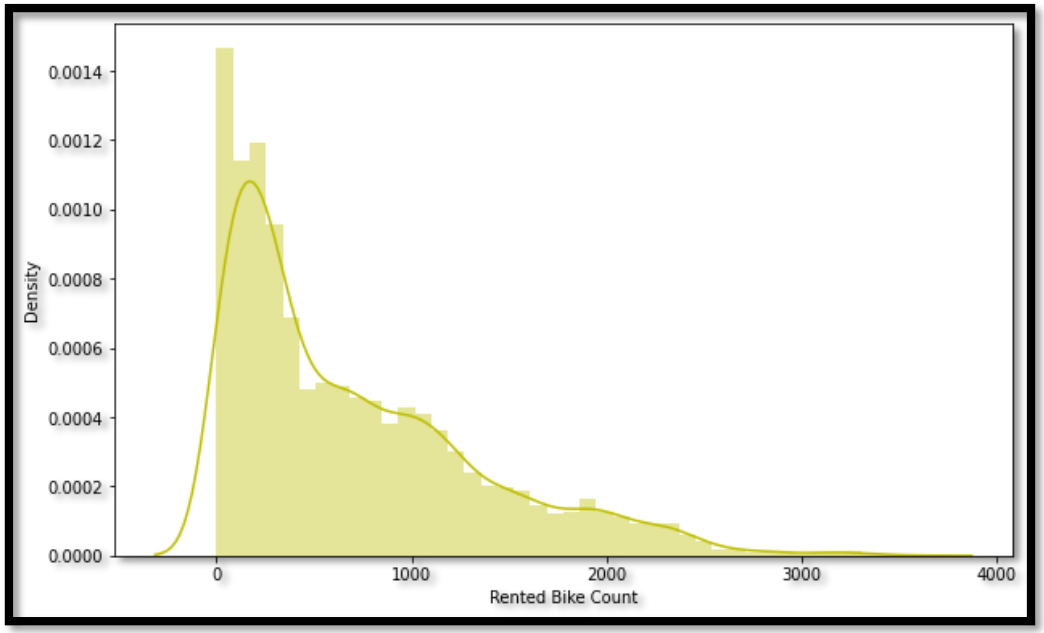
- Dew point Temperature is moderately positively correlated with rental bike count target variable rented bike count with correlation 0.37.
- Solar radiation is moderately positively correlated with rental bike count with correlation 0.26.

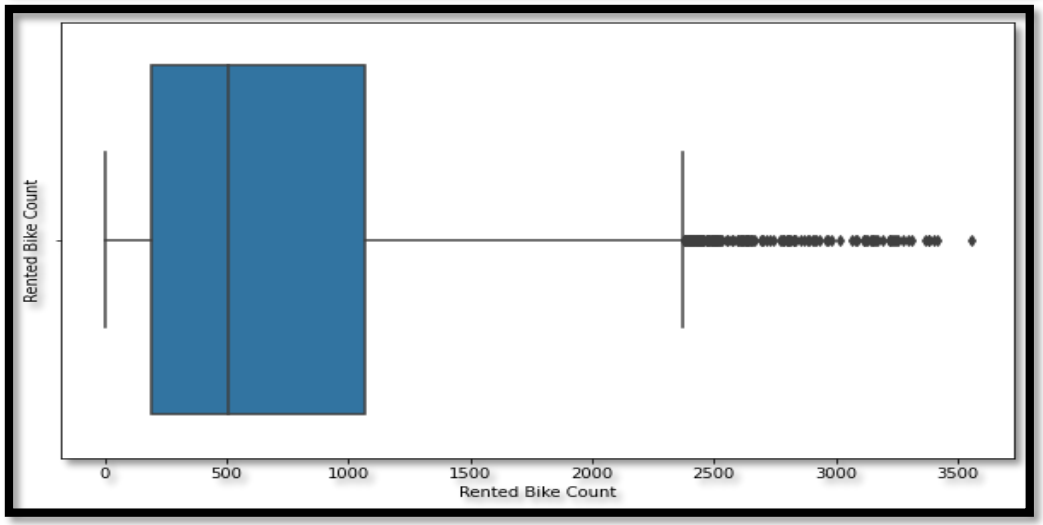
EDA



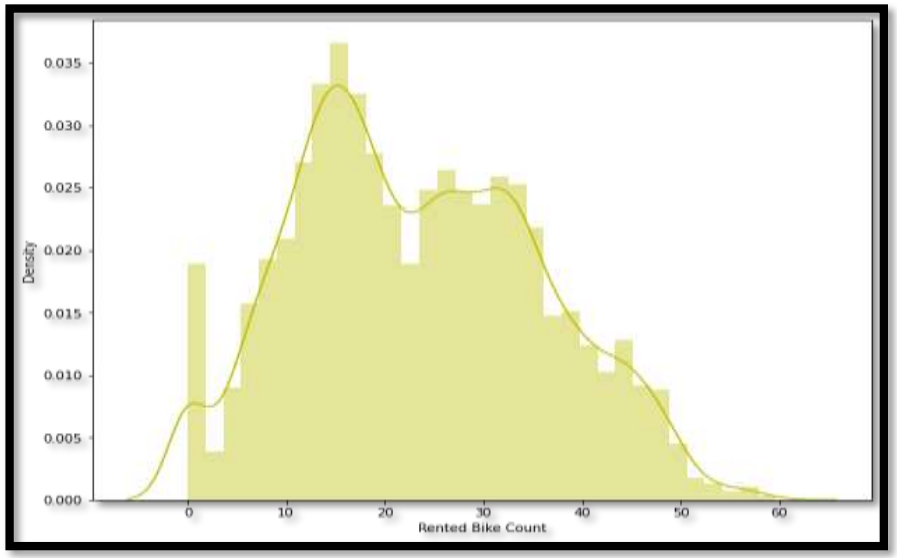
- From above scatter plot we can see that :
- Rainfall is moderately negatively correlated with rental bike count target variable rented bike count with correlation -0.12

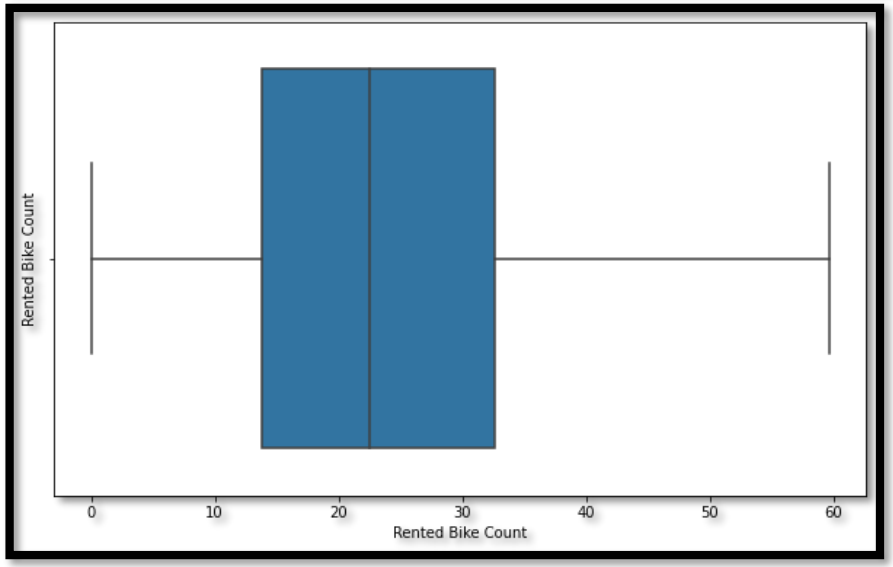
Checking if target variable follows normal distribution





- From distribution plot and scatter plot we can say that our target variable is right skewed and contains some outliers.
- So we will use take the square root of variable to improve the skewness and approximate it to follow normal distribution.





After applying square root on target variable we can see that it approximately follows normal distribution .

Building prediction models

To build the prediction we first have to perform training and test on the independent and dependent variables .I am going split data into 80% for training and 20% for testing ,after that i am going to apply various regression models on training and testing data i will find out which model is performing best on the dataset

.

The models i am going to use are as follows :

- Multiple linear regression
- Lasso regression
- Ridge regression
- Elastic net regression
- Decision trees
- Random forest

Multiple Linear regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.

Assumptions:

- The relation between the dependent and independent variables should be almost linear.
- Mean of residuals should be zero or close to 0 as much as possible. It is done to check whether our line is actually the line of “best fit”.
- There should be homoscedasticity or equal variance in a regression model. This assumption means that the variance around the regression line is the same for all values of the predictor variable (X)
- There should not be multicollinearity in regression model. Multicollinearity generally occurs when there are high correlations between two or more independent variables.

Checking assumptions of Multiple linear regression

Detecting Multicollinearity
By using VIF

	variables	VIF
0	Temperature(°C)	99.007405
1	Humidity(%)	21.243244
2	Wind speed (m/s)	1.426298
3	Visibility (10m)	2.309550
4	Dew point temperature(°C)	125.389013
5	Solar Radiation (MJ/m2)	4.773381
6	Rainfall(mm)	1.112106
7	Snowfall (cm)	1.144230
8	Holiday	1.042744
9	Functioning Day	1.084080
10	Seasons_Autumn	inf
11	Seasons_Spring	inf
12	Seasons_Summer	inf
13	Seasons_Winter	inf
14	Hour_0	inf
15	Hour_1	inf
16	Hour_2	inf
17	Hour_3	inf
18	Hour_4	inf
19	Hour_5	inf
20	Hour_6	inf
21	Hour_7	inf
22	Hour_8	inf
23	Hour_9	inf
24	Hour_10	inf
25	Hour_11	inf
26	Hour_12	inf
27	Hour_13	inf
28	Hour_14	inf
29	Hour_15	inf
30	Hour_16	inf
31	Hour_17	inf
32	Hour_18	inf
33	Hour_19	inf
34	Hour_20	inf
35	Hour_21	inf
36	Hour_22	inf
37	Hour_23	inf
38	month_1	inf
39	month_2	inf
40	month_3	inf
41	month_4	inf
42	month_5	inf
43	month_6	inf
44	month_7	inf
45	month_8	inf
46	month_9	inf
47	month_10	inf
48	month_11	inf
49	month_12	inf

- By using VIF we can see that there is high multicollinearity present
- Here, we use feature selection to remove variables with high multicollinearity. We remove variables with highest multicollinearity one by one by again checking multicollinearity after each removal.

Checking assumptions of Multiple linear regression

	variables	VIF
0	Temperature(°C)	19.815457
1	Humidity(%)	17.948959
2	Wind speed (m/s)	5.238728
3	Visibility (10m)	10.275980
4	Solar Radiation (MJ/m2)	6.514491
5	Rainfall(mm)	1.109775
6	Snowfall (cm)	1.163885
7	Holiday	18.630845
8	Functioning Day	24.887680
9	Seasons_Autumn	5.183991
10	Seasons_Summer	7.306907
11	Seasons_Winter	6.883460
12	Hour_0	1.859057
13	Hour_2	1.848625
14	Hour_3	1.846821
15	Hour_4	1.846095
16	Hour_5	1.849623
17	Hour_6	1.850058
18	Hour_7	1.850558
19	Hour_8	1.878729
20	Hour_9	1.969476
21	Hour_10	2.160973
22	Hour_11	2.378352
23	Hour_12	2.556697
24	Hour_13	2.625966
25	Hour_14	2.554847
26	Hour_15	2.434029
27	Hour_16	2.261677
28	Hour_17	2.116061
29	Hour_18	2.010819
30	Hour_19	1.956650
31	Hour_20	1.923646
32	Hour_21	1.893658
33	Hour_22	1.873650
34	Hour_23	1.860264
35	month_2	1.969334
36	month_5	1.892620
37	month_7	2.223979
38	month_8	2.359351
39	month_9	2.359259
40	month_11	2.256229
41	month_12	2.064071

- I removed variables having high multicollinearity variables those Variable are
  Dew point temperature(°C),Hour_1, month_6,month_3,month_4,month_1,month_10,Seasons_Spring
- From the data frame we can see that multicollinearity has been reduced so we can use multiple linear regression on remaining variables.

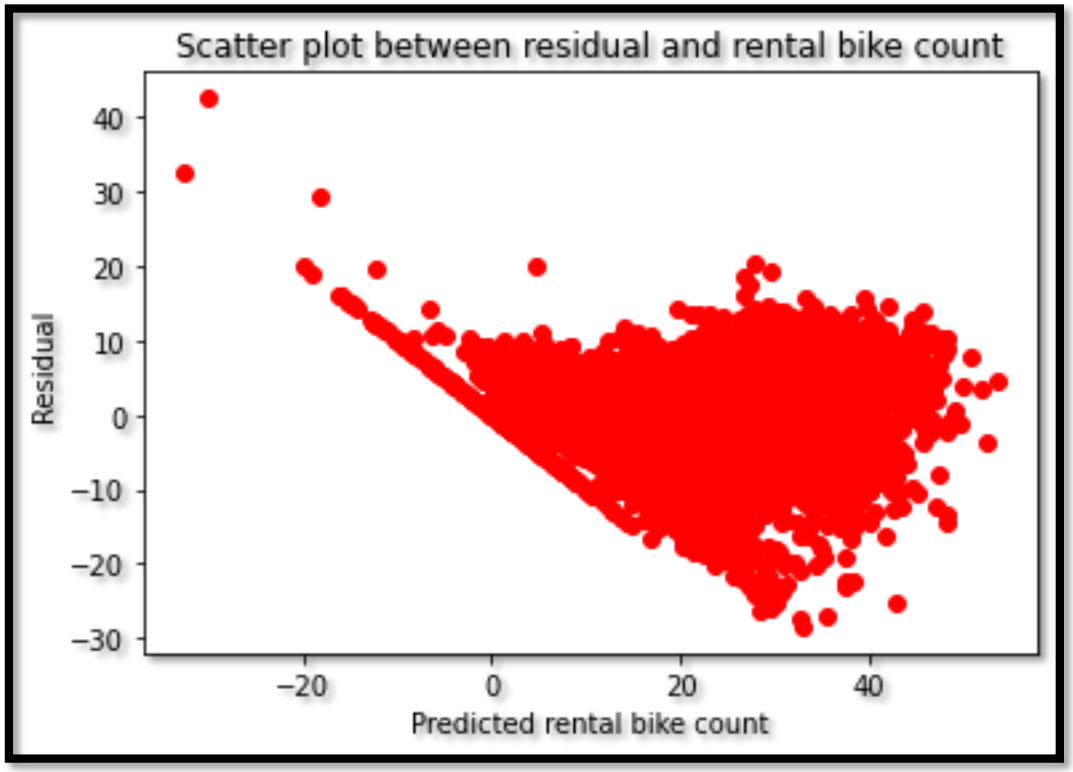
Multiple linear regression

```
#Training_dataset_metrics
print_metrics(y_train, y_train_pred)

MSE is 34.817741769952825
RMSE is 5.900656045725155
R2_score is 0.7743987734371468
MAE is 4.4639451242331765
Adj_R2_score is 0.7792869931920543
```

```
#Test_dataset_metrics
print_metrics(y_test, y_test_pred)

MSE is 33.905668316490555
RMSE is 5.822857401352927
R2_score is 0.784707129852672
MAE is 4.444111348257234
Adj_R2_score is 0.7792869931920543
```

- After applying multiple linear regressions on training and testing data i got the r2 score of the model .
- The r2_score for the training set is 0.77.
- The r2_score for the test set is 0.78 This means our linear regression model is performing well on the data

Heteroscedasticity




```
round((np.mean(residuals_train)))
```

```
0
```

- There is no visible relationship between residuals and predicted rental bike count
- Therefore there is no heteroscedasticity present in the data.
- We can also see that mean of residuals is zero
- Therefore all the assumptions of linear regression model are satisfied .

Lasso regression

```
# Training dataset metrics
print_metrics(y_train, y_pred_train_lasso)
```

```
MSE is 91.85465398123709
RMSE is 9.584083366772072
R2_score is 0.4048286433798308
MAE is 7.267171202255601
Adj_R2_score is 0.3681798790716583
```

```
# Testing  dataset metrics
print_metrics(y_test, y_pred_test_lasso)
```

```
MSE is 97.05945184518272
RMSE is 9.851875549619104
R2_score is 0.38369573583917327
MAE is 7.4573120369945896
Adj_R2_score is 0.3681798790716583
```

Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients

- After applying lasso regressions on training and testing data i got the r2 score of the model
- The r2 score for lasso regression for training data is 0.40 .
- The r2_score for the test set is 0.38 . This means our lasso regression model is not performing well on the data..
- I will perform cross validation and hyperparameters tunning to improve r2 score.

Lasso regression

```
MSE : 33.91094156718748
RMSE : 5.8233101898479935
R2 : 0.7846736459741962
Adjusted R2 :  0.7792526663353733
```

Cross-Validation & Hyperparameter Tuning

- After performing cross validation and hyperparameters tunning of lasso regression
- The R2_score is for testing data is improved to 0.78 .This means our lasso regression model is performing well on the data.

Ridge regression

```
# Training dataset metrics
print_metrics(y_train, y_pred_train_ridge)
```

```
MSE is 34.81776234567281
RMSE is 5.900657789236113
R2_score is 0.7743986401169649
MAE is 4.464062543032918
adj_r2_score is 0.7792742823039116
```

```
# testing  dataset metrics
print_metrics(y_test, y_pred_test_ridge)
```

```
MSE is 33.90762094793536
RMSE is 5.823025068461869
R2_score is 0.7846947311108401
MAE is 4.444422895352889
adj_r2_score is 0.7792742823039116
```

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization

- After applying ridge regressions on training and testing data i got the r2 score of the model .
- The r2 score for ridge regression for training data is 0.77 .
- The r2_score for the test set is 0.78 . This means our ridge regression model is performing well on the data..
- I will perform cross validation and hyperparameters tunning to improve R2 score.

Ridge regression

```
MSE : 33.90586231679181
RMSE : 5.822874059842941
R2 : 0.7847058979972407
Adjusted R2 :  0.7792857303238692
```

Cross-Validation & Hyperparameter Tuning

- After performing cross validation and hyperparameters tunning on ridge regression
- The R2_score is for testing data is improved to 0.78 .

Elastic net regression

```
# Training dataset metrics
print_metrics(y_train, y_pred_train_el)

MSE is 58.44799286228148
RMSE is 7.645128701485769
R2_score is 0.6212867862887405
MAE is 5.8264410123812
Adj_R2_score is 0.6047665429467163
```

```
# Testing dataset metrics
print_metrics(y_test, y_pred_test_el)

MSE is 60.71529130174544
RMSE is 7.792001751908519
R2_score is 0.6144724473746381
MAE is 5.9027708550477564
Adj_R2_score is 0.6047665429467163
```

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models

- After applying elastic net regressions on training and testing data i got the r2 score of the model .
- The r2 score for elastic net regression for training data is 0.62 .
- The r2_score for the test set is 0.61 . This means our elastic net regression model is performing well on the data..
- I will perform cross validation and hyperparameters tunning to improve R2 score.

Elastic net regression

```
MSE : 33.91417832175735
RMSE : 5.823588096848655
R2 : 0.784653093358191
Adjusted R2 :  0.7792315962940237
```

Cross-Validation & Hyperparameter Tuning

- After performing cross validation and hyperparameters tunning on elastic net regression
- The R2_score is for testing data is improved to 0.78 .This means our elastic regression model is performing well on the data.

Decision trees

```
#Training data metrics
print_metrics(y_train1, y_pred_train_dt)

MSE is 47.93620841653809
RMSE is 6.923597938683188
R2_score is 0.6893977936019883
MAE is 5.050593674132065
Adj_R2_score is 0.6538266735147509
```

```
#Testing data metrics
print_metrics(y_test1, y_pred_test_dt)

MSE is 52.9296519639074
RMSE is 7.275276762014446
R2_score is 0.6639093917618941
MAE is 5.219255022241323
Adj_R2_score is 0.6538266735147509
```

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

- After applying decision tree model on training and testing data i got the r2 score of the model .
- The r2 score for decision tree for training data is 0.68 .
- The r2_score for the test set is 0.66 . This means our Decision tree model is performing well on the data..

Random forest

```
#Training data metrics
print_metrics(y_train1, y_pred_train_rf)

MSE is 1.5603016966938077
RMSE is 1.2491203691773693
R2_score is 0.9898900400000669
MAE is 0.7839264771527887
Adj_R2_score is 0.9185181327717464
```

```
#Testinf data metrics
print_metrics(y_test1, y_pred_test_rf)

MSE is 12.458518735540261
RMSE is 3.5296626943010097
R2_score is 0.9208913910405305
MAE is 2.1903766546083574
Adj_R2_score is 0.9185181327717464
```

Random forests or **random decision forests** is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

- After applying random forest model on training and testing data i got the r2 score of the model .
- The r2 score for random forest model for training data is 0.98.
- The r2_score for the test set is 0.92 . This means our random forest model is performing well on the data..
- The Random forest model is giving the highest r2_score among all the models so we can deploy this model for prediction.

Random forest

Feature	Feature Importance
Temperature(°C)	0.311149
Functioning Day	0.151433
Humidity(%)	0.147534
Hour_18	0.032824
Hour_4	0.031021
Rainfall(mm)	0.030744
Solar Radiation (MJ/m2)	0.028924
Hour_5	0.027313
Hour_3	0.020815
Dew point temperature(°C)	0.019573

- Top 10 most important features according to random forest model are temperature, functioning day,humidity,rainfall,solar radiation,hour_18 ,hour_4 ,hour_5,hour_3 means 6pm 4am 5am 3am respectively and dew point temperature.

Conclusions

- From exploratory data analysis of the given data we can see that
- The maximum number of bikes rented in the month of June and followed by May ,july,August,september and October .
- The people rent bikes slightly more bikes in weekdays as compared to weekend
- peak time of renting bike at 7am to 9am in the morning and from 5pm to 10pm in the evening .People mostly use the rental bikes to go their workplace so that's why company should increase the availability of the bikes during the peak hours.
- In summer season the use of rented bike is high and In winter season the use of rented bike is very low because of snowfall .

Conclusions

- From the scatter plots we can that ,
- Temperature is highly positively correlated with rental bike count target variable.people prefer renting bike when the temperature normal not to hot and not too cold outside
- Dew point temperature ,solar radiation,visibilty,Wind speed is moderately positively correlated with rental bike count target variable
- humidity and rainfall are negatively moderately correlated with rental bike count variable. people does not prefer to rent when there is raining outside.

Conclusions

➢ **Prediction model insights :**

I implemented 6 machine learning algorithms Linear regression ,lasso regression ,ridge regression ,elastic net regression ,decision tree, Random Forest . The results of our evaluation are:

- The r2_score of multiple linear regression model for test set is 0.78 . This means our multiple linear regression performing well on the data.
- The r2_score of lasso regression for test set is 0.38 . This means our lasso regression model is not performing well on the data.
- The r2_score of ridge regression for test set is 0.78 This means our ridge regression model is performing well on the data.
- The r2_score of elastic net regression for test set is 0.61 . This means our elastic net regression model is performing well on the data.
- The r2_score of Decision Tree model for the test set is 0.66 . This means our Decision Tree model is performing well on the data.
- The r2_score of Random forest model for the test set is 0.92 . This means our Random forest model is performing very well on the data.

Conclusions

i also performed hyperparameter tunning on lasso, ridge and elastic net regression ,

a)The R2_score of lasso regression after hyperparameters tunning is 0.78 .This means our lasso regression model is performing well on the data.

b)The r2_score of ridge regression for the test set after hyperparameter tuning is 0.78 This means our ridge regression model is performing well on the data.

c)The r2_score of elastic net regression for the test set after hyperparameter tunning is 0.78. This means our elastic net regression model is performing well on the data.

- The Random forest model is giving the highest r2_score among all the models so we can deploy this model for prediction. According to random forest model the most important 10 features are temperature, functioning day,humidity,rainfall,solar radiation,hour_18 ,hour_4 ,hour_5,hour_3 means 6pm 4am 5am 3am respectively and dew point temperature.

Thank you