Capstone Project - II

Bike Sharing Demand Predication

NAVED MANSURI
Data Science Trainee, Almabetter

Point of Discussion

- Problem statement
- Data summary
- EDA
- Feature engineering
- Machine learning model
    - Linear regression
    - Decision tree
    - Polynomial regression
    - Random forest
    - Gradient boosting
- Model validation
- Model comparison
- Conclusion

Problem statement

- Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort.
- Making the rental bikes available at the **right time** will reduce the **waiting time** and make them more accessible to the public.
- Eventually, providing the city with a stable supply of rental bikes becomes a major concern
- The goal is prediction of the number of rental bikes necessary each hour to maintain a stable supply.

Data summary

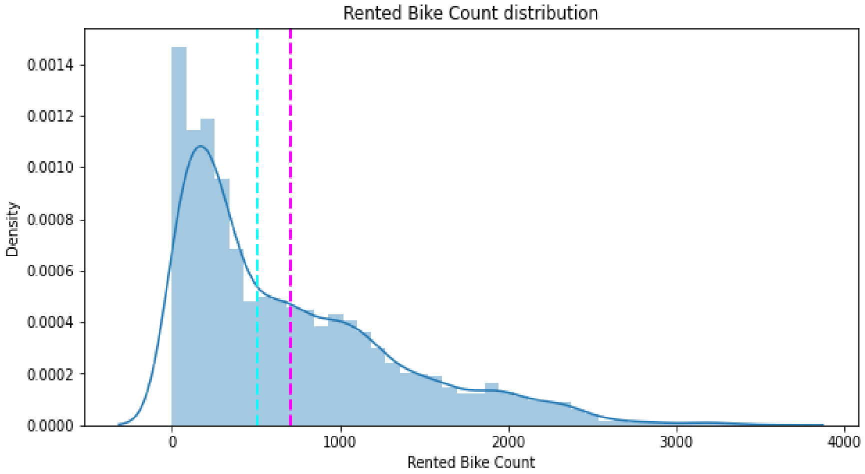
- There are 8760 rows and 14 columns in the data set and 10 are numeric features and 4 categorical.
- **Rented Bike Count** is dependent variable.
- Data information
  - **Rented Bike count** - Count of bikes rented at each hour
  - **Hour** - Hour of per day
  - **Temperature**-Temperature in Celsius
  - **Humidity** - %
  - **Windspeed** - m/s
  - **Visibility** - 10m
  - **Dew point temperature** - Celsius
  - **Solar radiation** - MJ/m2

Data summary

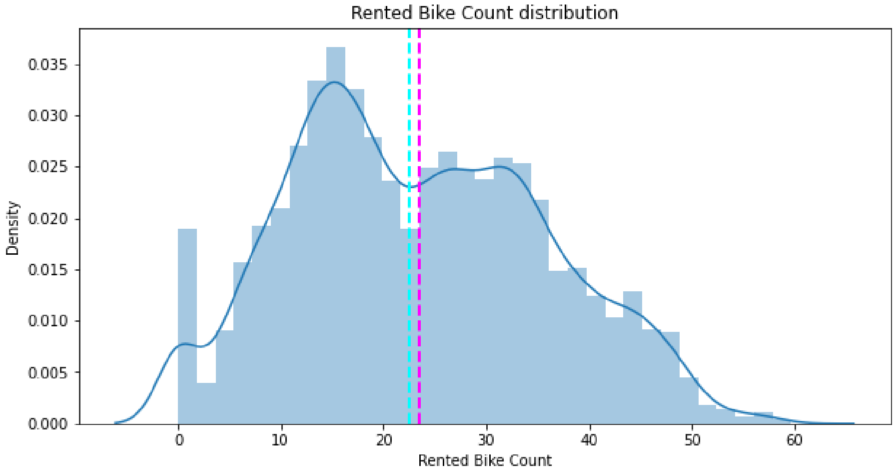
- **Rainfall** - mm
- **Snowfall** - cm
- **Seasons** - Winter, Spring, Summer, Autumn
- **Holiday** - Holiday/No holiday
- **Functional Day** - No Func(Non Functional Hours), Fun(Functional hours)

Exploratory Data Analysis

Analyzing dependent variable



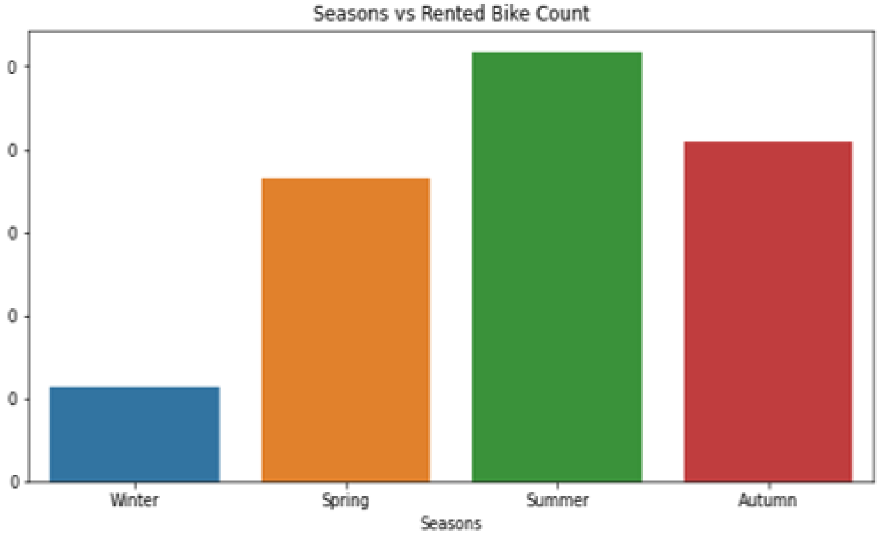
Dependent variable is positively skewed

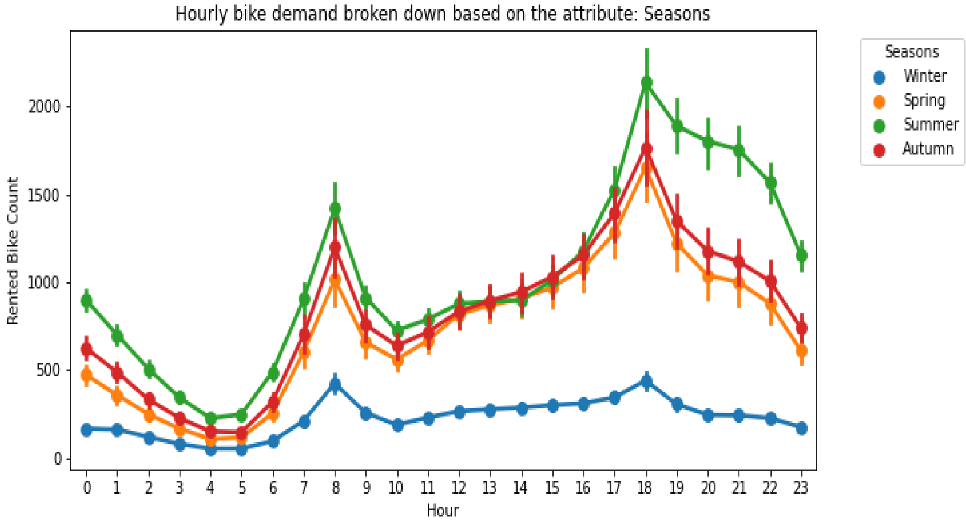


Dependent variable after square root transformation

Dependent variable is right skewed. To get better predictions dependent variable is almost normally distributed. To achieve this, we can transform the data by log and sqrt transformation.

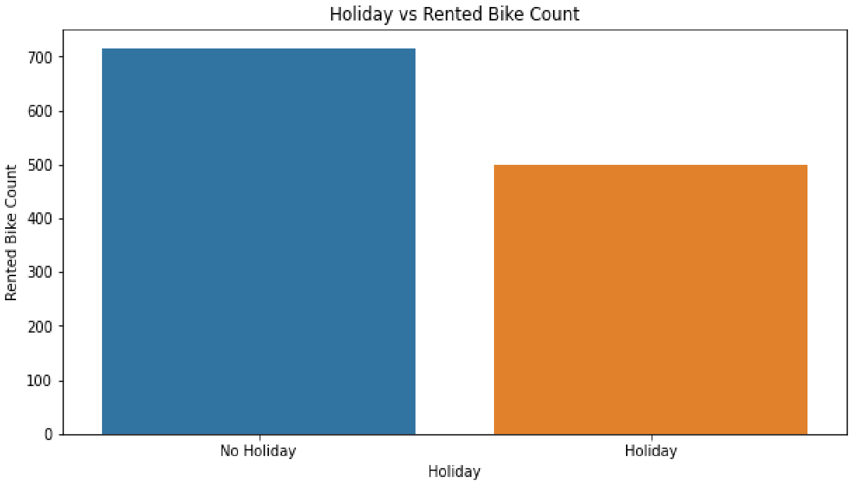
Analyzing ranted bike count by seasons

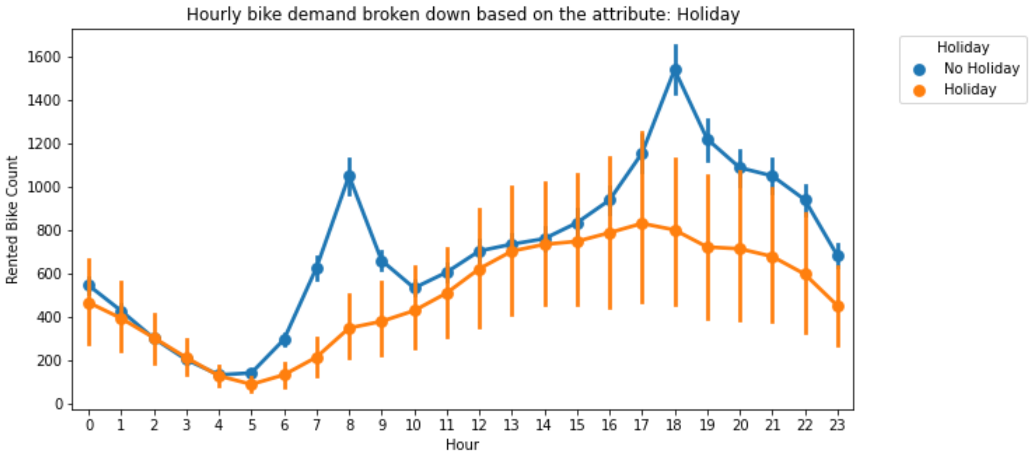




The demand for bikes is highest in summer and less demanding in winter seasons.
We can see that the most bikes were rented at 18:00 Hr. (6:00 PM) as opposed to 5:00 Hr. (5:00 AM). result, people tend to rent bikes very less in the morning.

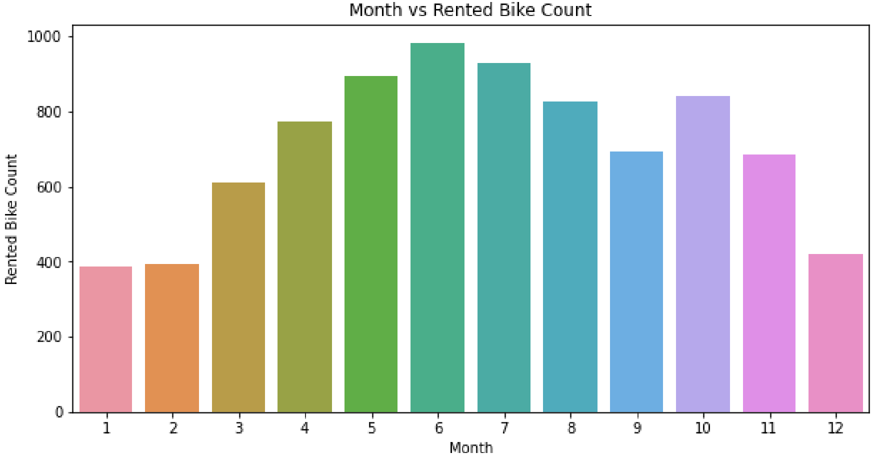
Analyzing ranted bike count by holidays

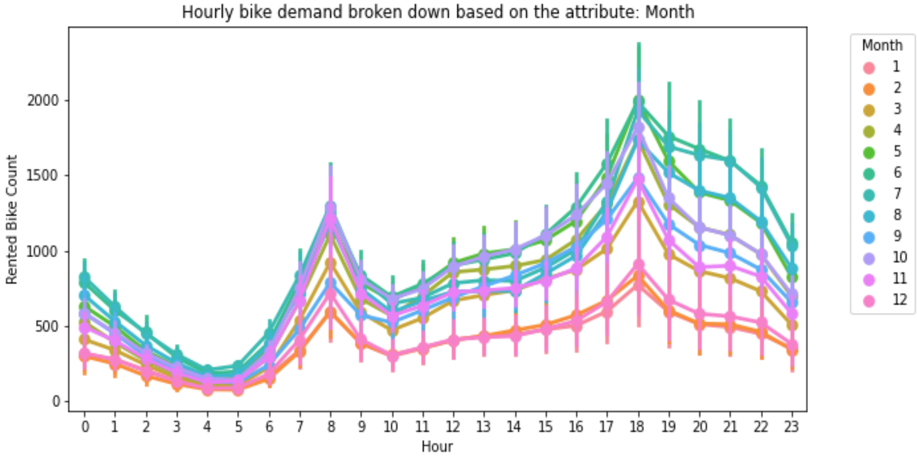




During working day people tend to rent more bike this we can assume that on holidays people tends to rent less bike.

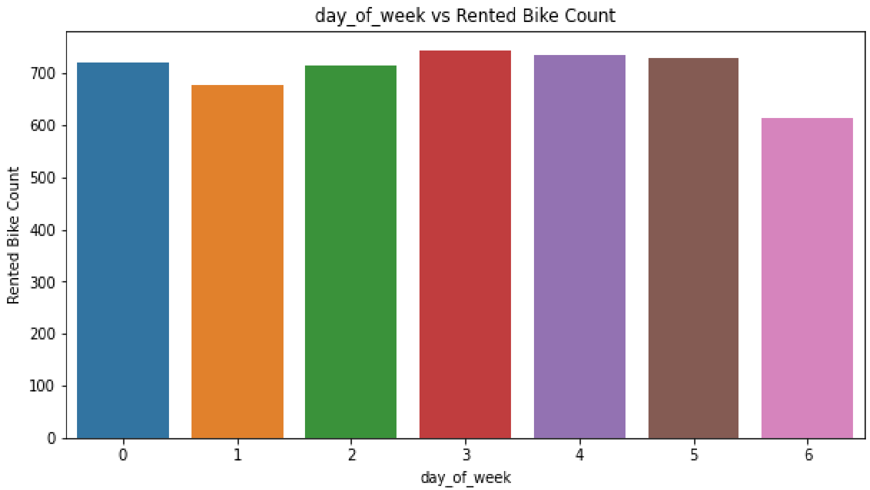
Analyzing ranted bike count by months

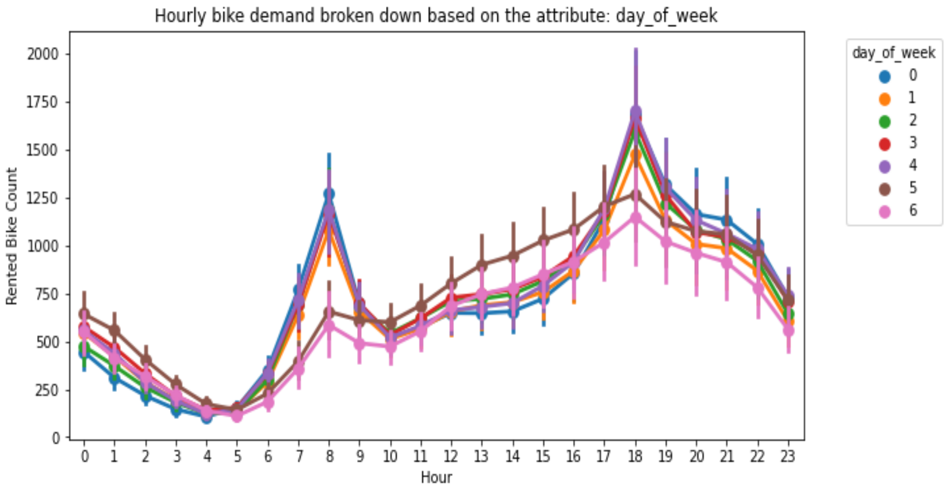




When we compare month to the number of bikes rented, we can see that people tend to hire more bikes in June (6) rather than less bikes in December(12) or January(1). We can say this that people tend to rent more bikes in the summer than they do not in the winter.

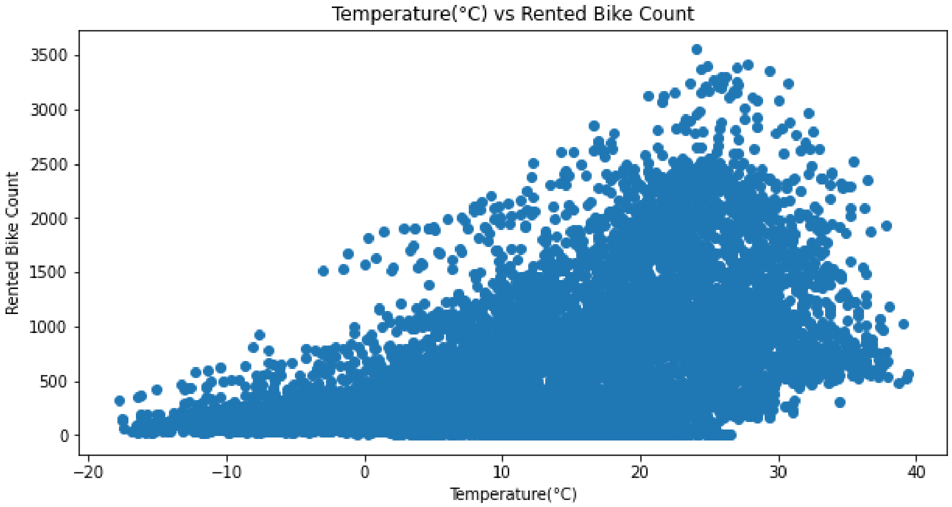
Analyzing ranted bike count by day of week

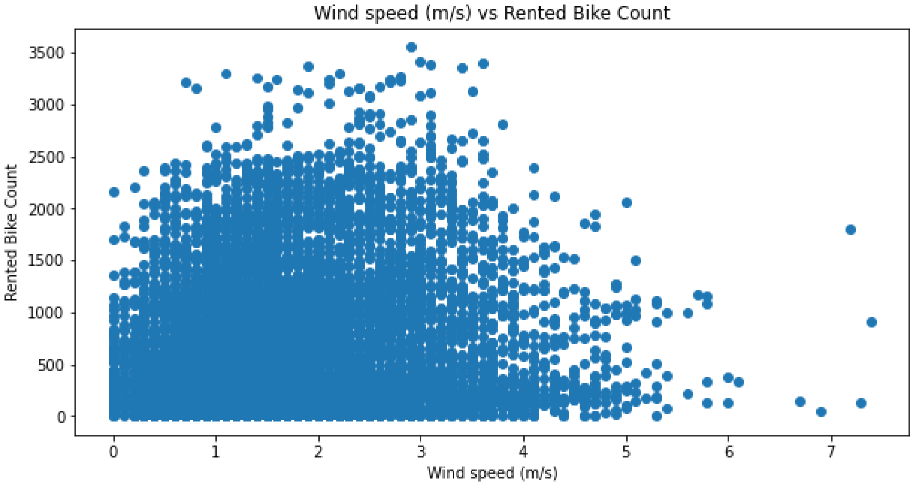




we can see that people tends to rent more bike during weekdays as compared to weekends.

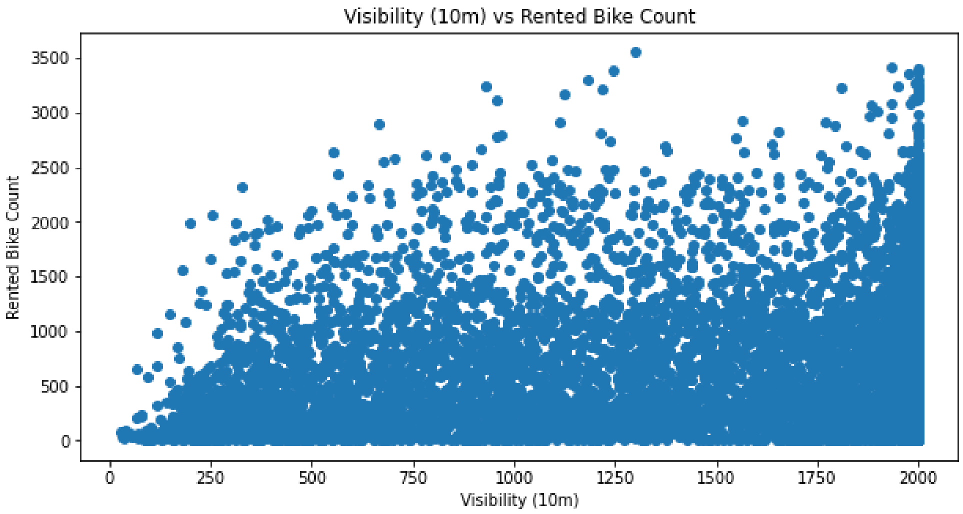
Analyzing relationship between dependent and independent variables

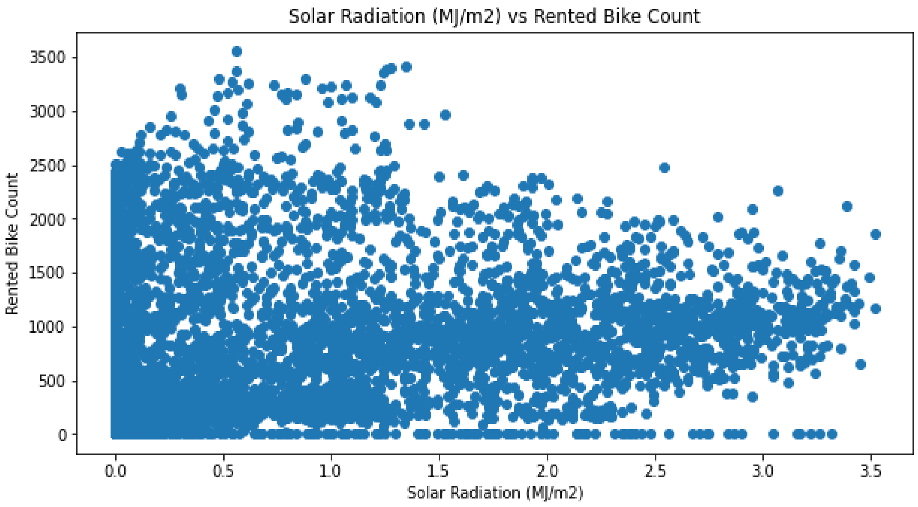




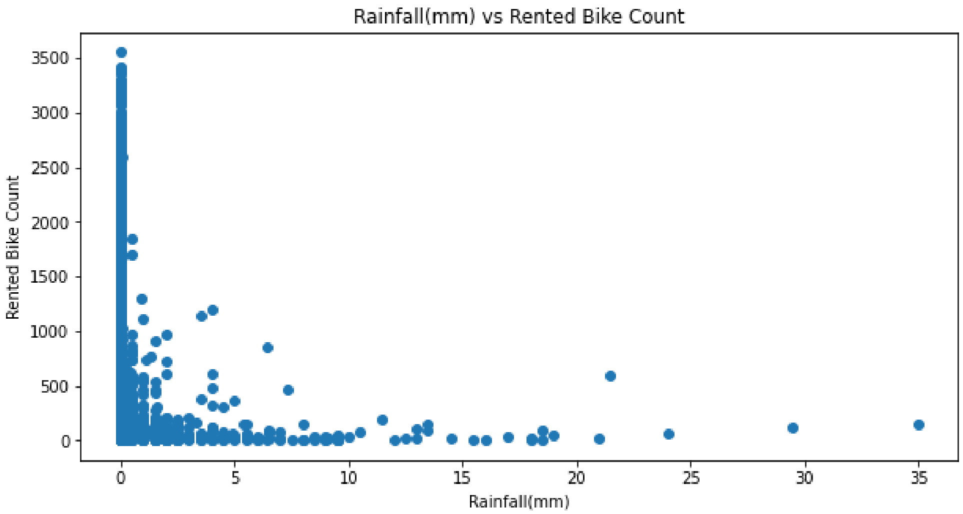
bike count is higher when the weather is between 10 and 30 degrees, and when it is below zero, a few people bike ranted.
When the wind speed is higher, there is less demand, when the wind flow is between 0 and 4, the ranted bike count is higher.

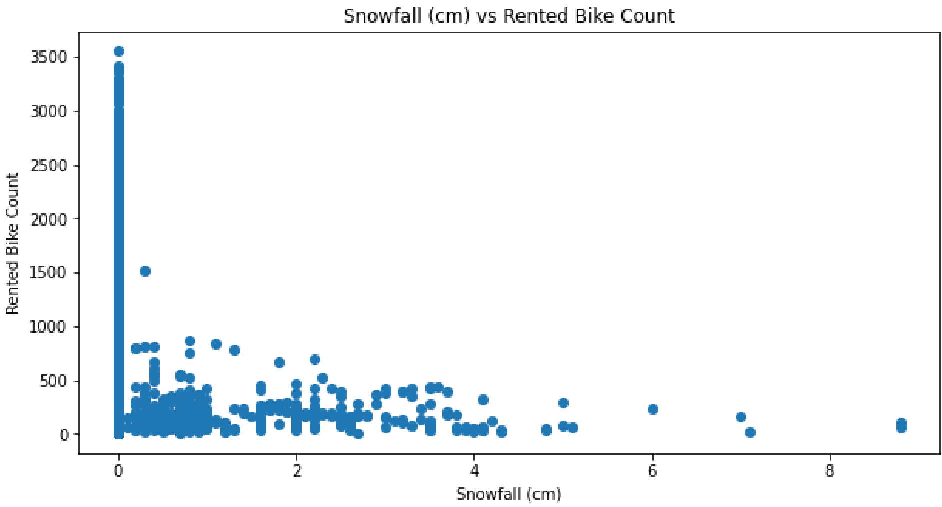
Analyzing relationship between dependent and independent variables





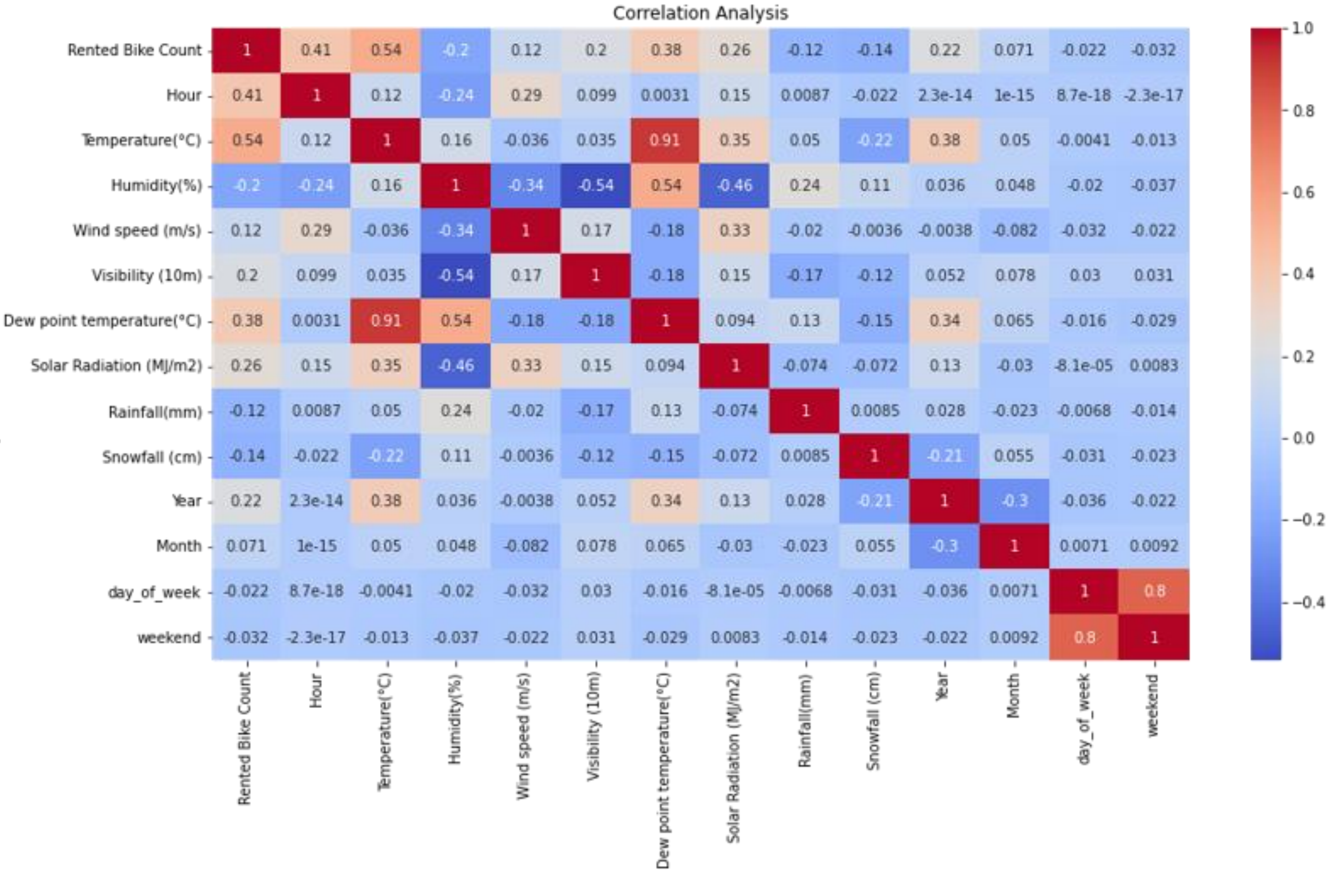
Analyzing relationship between dependent and independent variables





Correlation map

- **Temperature** is highly corelated with **dew point temperature.**
- For removing multicollinearity drop **temperature** or **dew point temperature**.



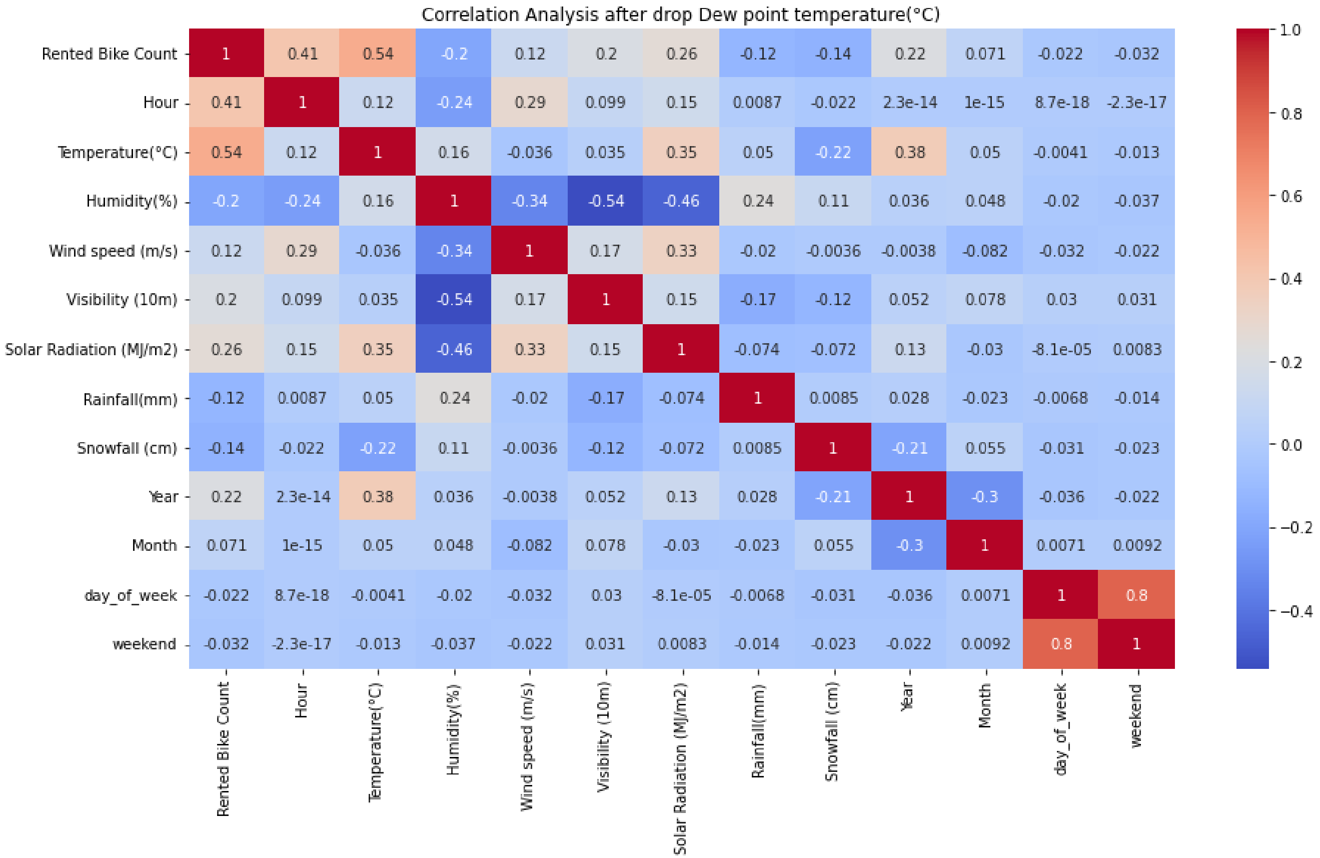
Correlation map

- Now we drop dew point temperature and check VIF.
- After dropping dew point temperature VIF is acceptable range.

	variables	VIF
0	Hour	3.955864
1	Temperature(°C)	3.248065
2	Humidity(%)	5.915106
3	Wind speed (m/s)	4.612145
4	Visibility (10m)	5.085606
5	Solar Radiation (MJ/m2)	2.276399
6	Rainfall(mm)	1.079576
7	Snowfall (cm)	1.122222
8	day_of_week	8.588900
9	weekend	3.808363

Correlation map

- ❑ After dropping multicollinearity now correlation map look like.



Feature engineering

In our data frame we have 3 features "**Season**" , "**Functioning Days**" and "**Holiday**" which contain categorical values , we know that to fit data to our machine learning model we need all numerical features, so we change into numeric features.

And also create new variable using **pd.get_dummies()** to Convert categorical variable into variables.

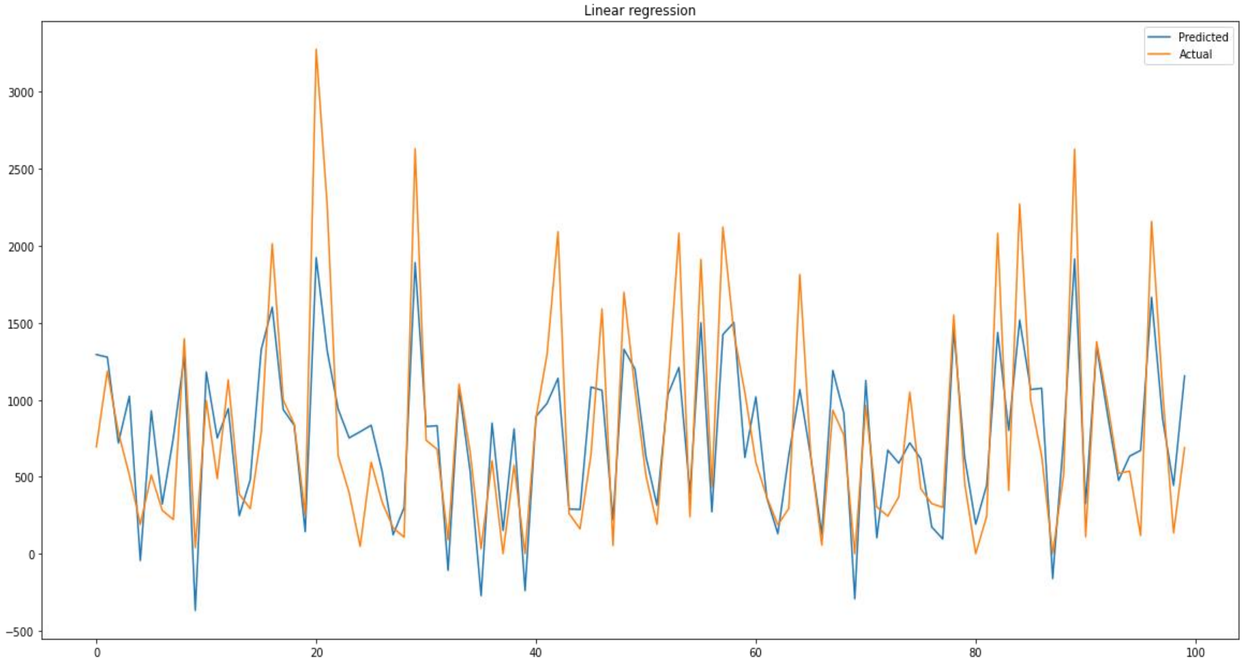
```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Rented Bike Count         8760 non-null   int64
 1   Hour                      8760 non-null   int64
 2   Temperature(°C)           8760 non-null   float64
 3   Humidity(%)             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

Machine learning model

Linear regression model

```
Evalution metrices on train data
MSE : 135302.5358
RMSE : 367.8349
R2 score: 0.675
Adjusted R2 :  0.665
```

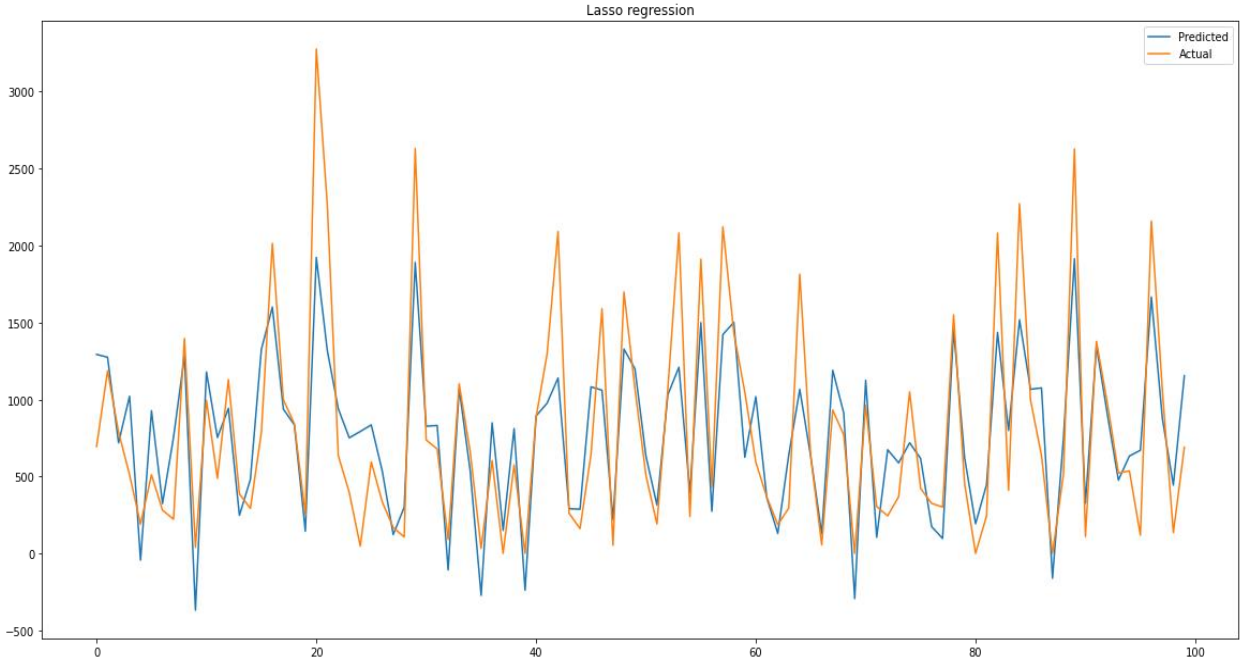
```
Evaluation metrices on test data
MSE : 131577.4113
RMSE : 362.736
R2 score: 0.682
Adjusted R2 :  0.671
```



Lasso regression

```
Evalution metrices on train data
MSE : 135303.2413
RMSE : 368.0
R2 score: 0.675
Adjusted R2 :  0.665
```

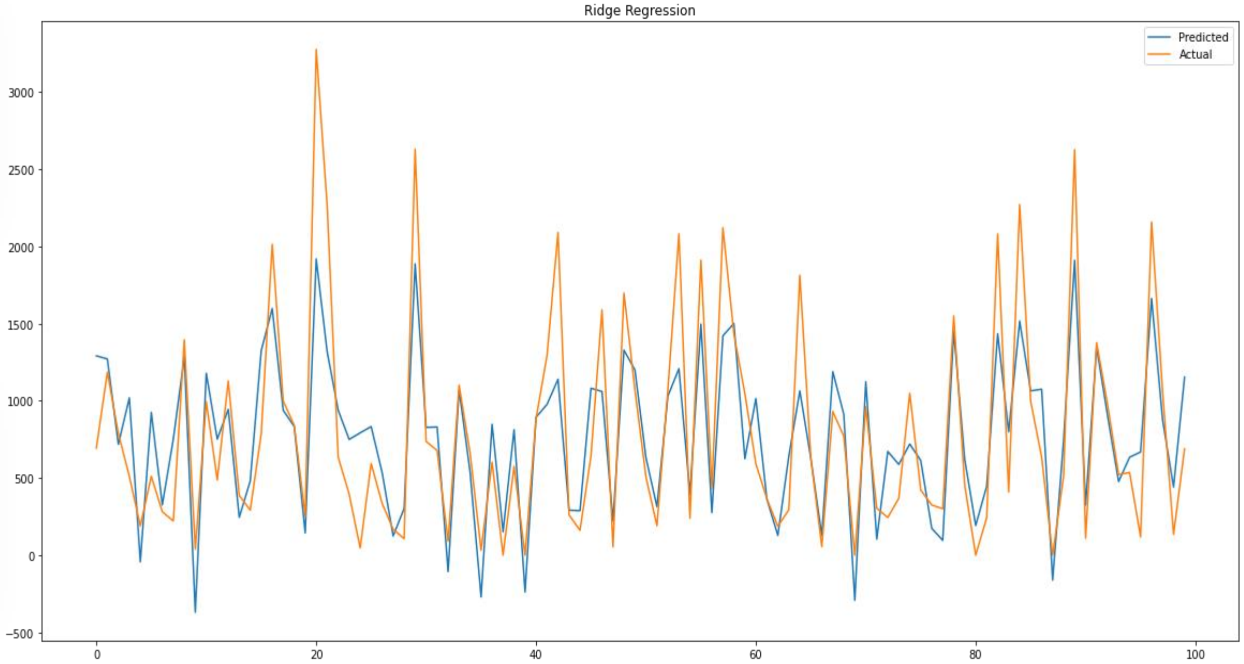
```
Evaluation metrices on test data
MSE : 131579.1089
RMSE : 362.738348
R2 score: 0.682
Adjusted R2 :  0.671
```



Ridge regression

```
Evalution metrices on train data
MSE : 135306.4509
RMSE : 368.0
R2 score: 0.675
Adjusted R2 :  0.665
```

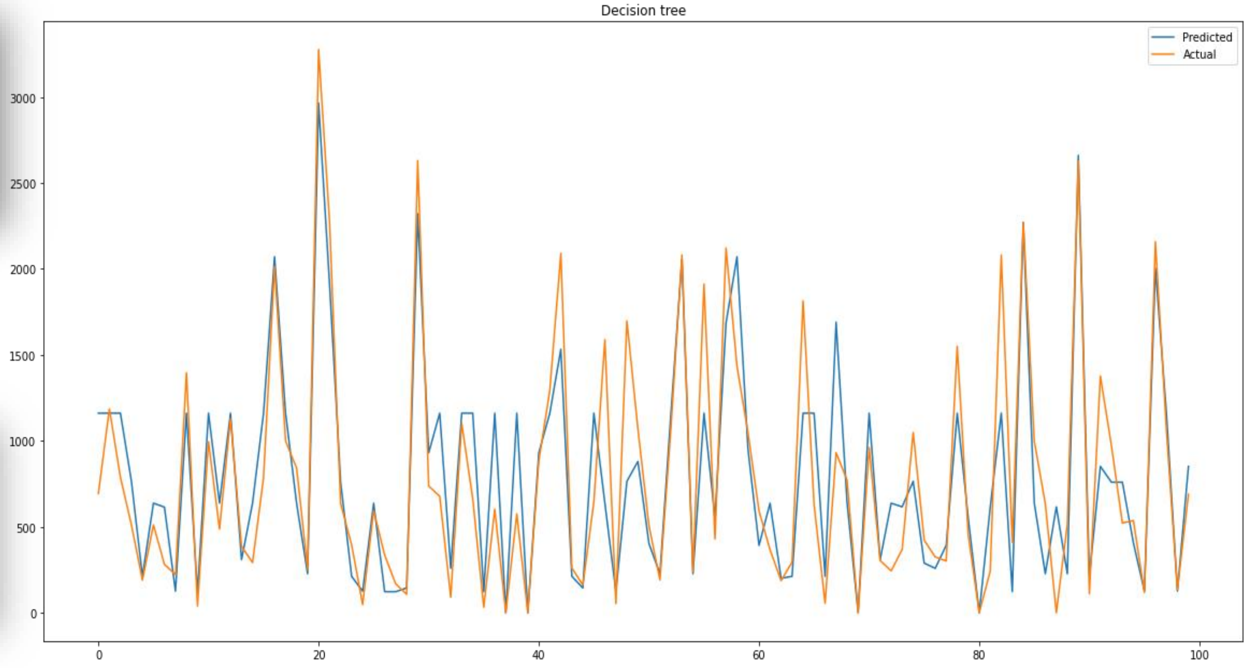
```
Evalution metrices on test data
MSE : 131586.3486
RMSE : 362.748327
R2 score: 0.682
Adjusted R2 :  0.671
```



Decision tree model

```
Evalution metrices on train data
MSE : 73334.356
RMSE : 270.803
R2 score: 0.824
Adjusted R2 :  0.818
```

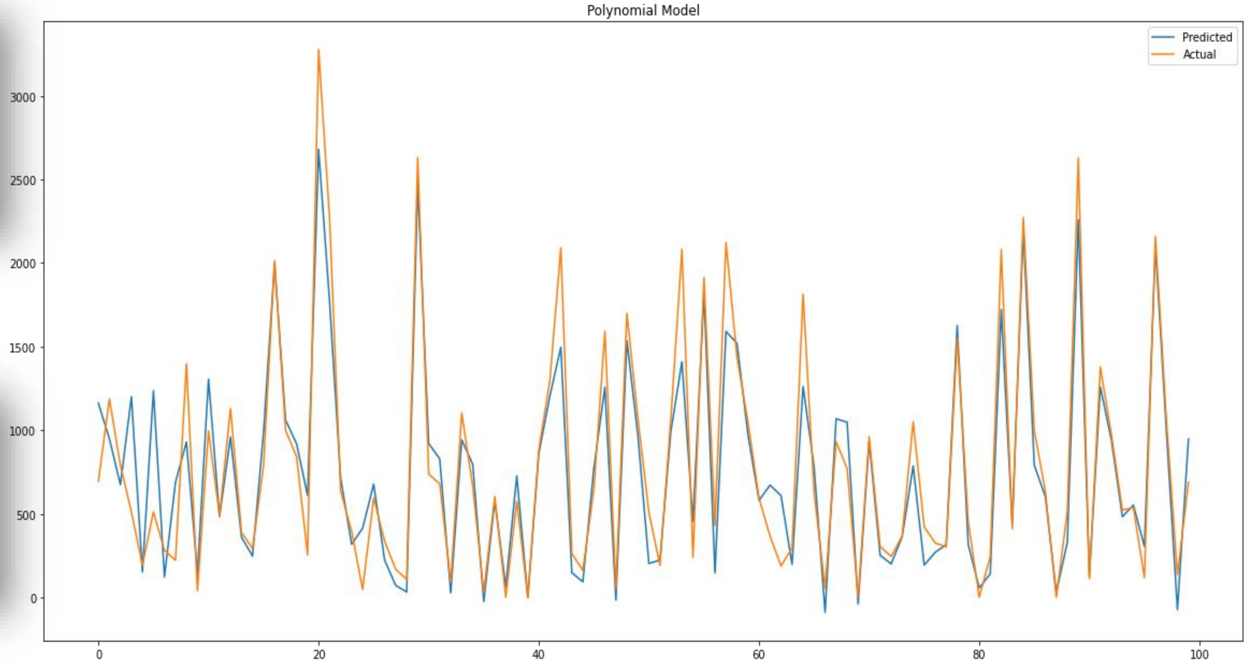
```
Evaluation metrices on test data
MSE : 105779.906
RMSE : 325.238
R2 score: 0.744
Adjusted R2 :  0.736
```



Polynomial regression

```
Evalution metrices on train data
MSE : 44727.6624
RMSE : 211.4892
R2 score: 0.893
Adjusted R2 :  0.889
```

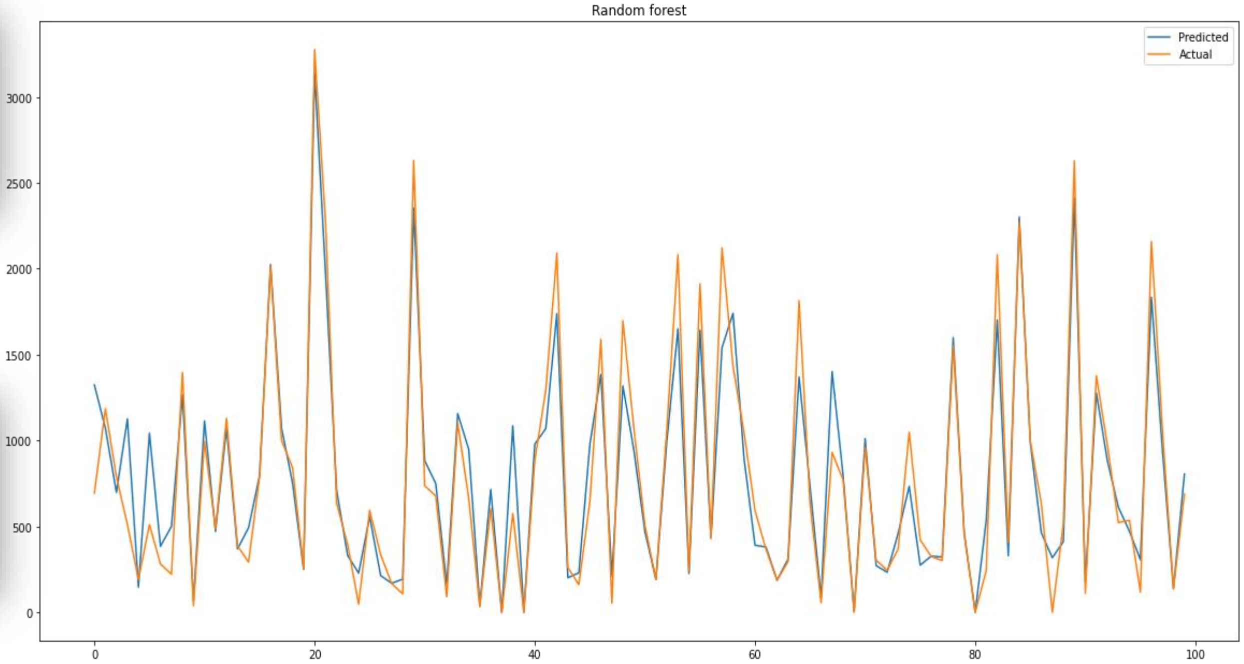
```
Evalution metrices on test data
MSE : 64493.1107
RMSE : 253.955
R2 score: 0.844
Adjusted R2 :  0.839
```



Random forest model

```
Evalution metrices on train data
MSE : 6766.2725
RMSE : 82.257
R2 score: 0.984
Adjusted R2 :  0.983
```

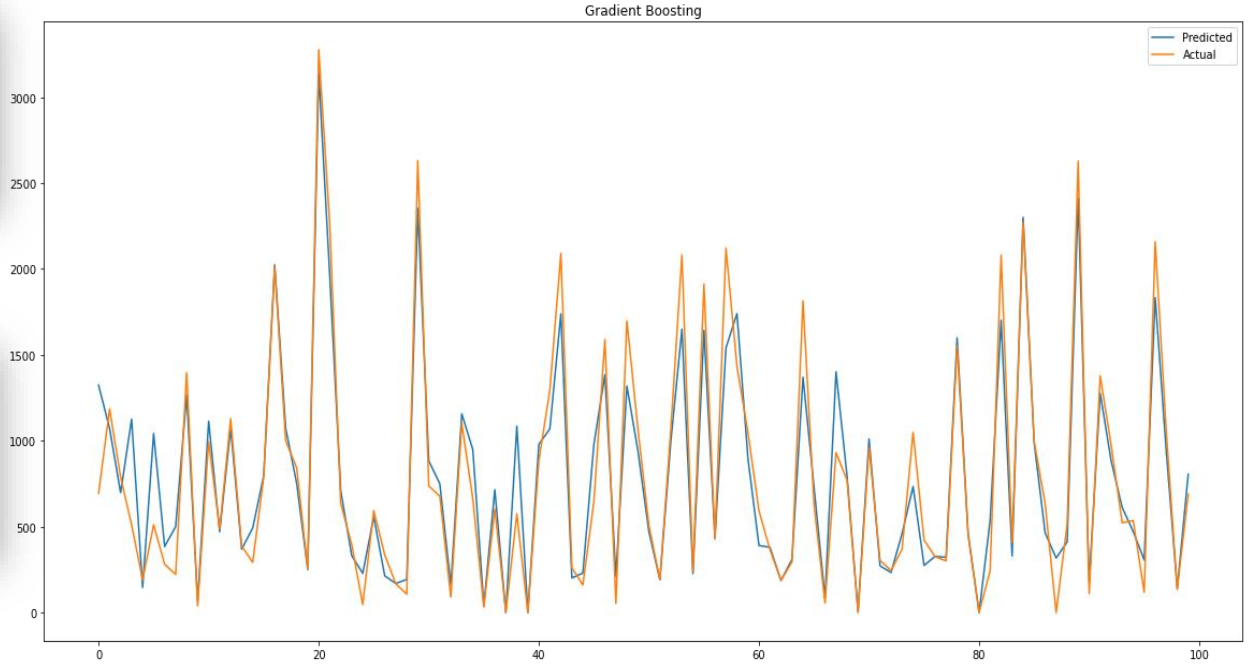
```
Evalution metrices on test data
MSE : 49167.8912
RMSE : 221.738
R2 score: 0.881
Adjusted R2 :  0.877
```



Gradient boosting

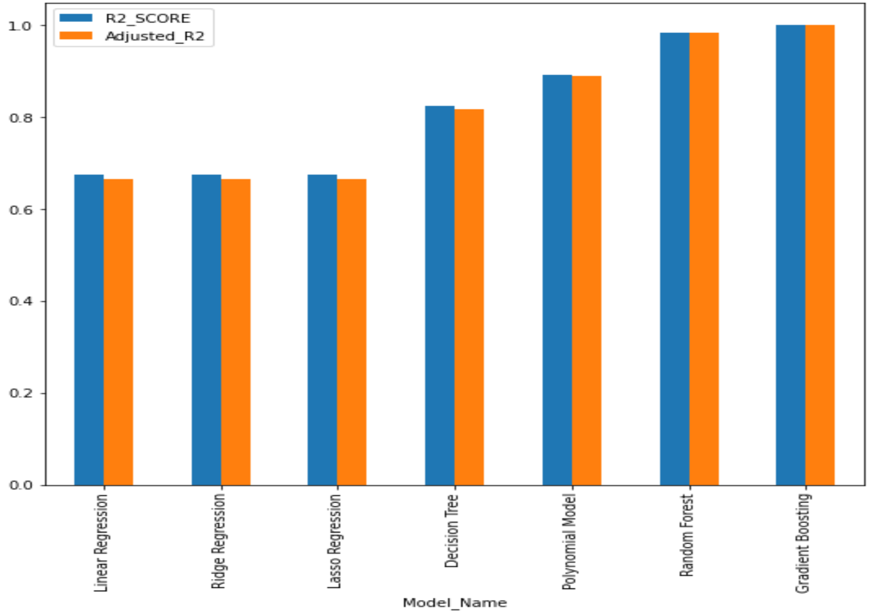
```
Evalution metrices on train data
MSE : 6.0053953
RMSE : 2.4505908
R2 score: 0.9999856
Adjusted R2 :  0.99998512
```

```
Evalution metrices on test data
MSE : 52602.3413633
RMSE : 229.3520032
R2 score: 0.8727231
Adjusted R2 :  0.8685956
```



Model comparison

	Model_Name	Train_MSE	Train_RMSE	R2_SCORE	Adjusted_R2
0	Linear Regression	135302.535800	367.834900	0.6750	0.6650
1	Ridge Regression	135306.450900	368.000000	0.6750	0.6650
2	Lasso Regression	135303.241300	368.000000	0.6750	0.6650
3	Decision Tree	73334.356000	270.803000	0.8240	0.8180
4	Polynomial Model	44727.662400	211.489200	0.8930	0.8890
5	Random Forest	6766.272500	82.257000	0.9840	0.9830
6	Gradient Boosting	68.333355	8.266399	0.9998	0.9998



- ❑ Gradient boosting gives the highest R2 score. On training set the r2 score is 99%.
- ❑ We can use either Random Forest or Gradient Boosting model for the bike rental stations.

Conclusion

- holiday or non-working days there is less bike demand.
- high demand at morning 8am and evening 6pm.
- clear visibility and low solar radiation is increasing bike demand.
- Gradient boosting R2 score is 99% and random forest R2 score is 98%.
- random forest and gradient boosting best model that can be used for the bike demand predication because performance matrix R2 and adjusted_R2 is higher.
- so result is best machine learning model for bike ranted prediction we use gradient boosting or random forest model.

QnA

Thank you