Capstone Project : 2

Bike Sharing Demand Prediction

By

Mayank Mishra

Points for Discussion

- Understanding Problem Statement
- Data Set Information
- Feature Summary
- Data Wrangling
- Data Preprocessing
- Exploratory Data Analysis
- Implementing Algorithm
- Challenges
- Conclusion
- Q&A

Understanding Problem Statement

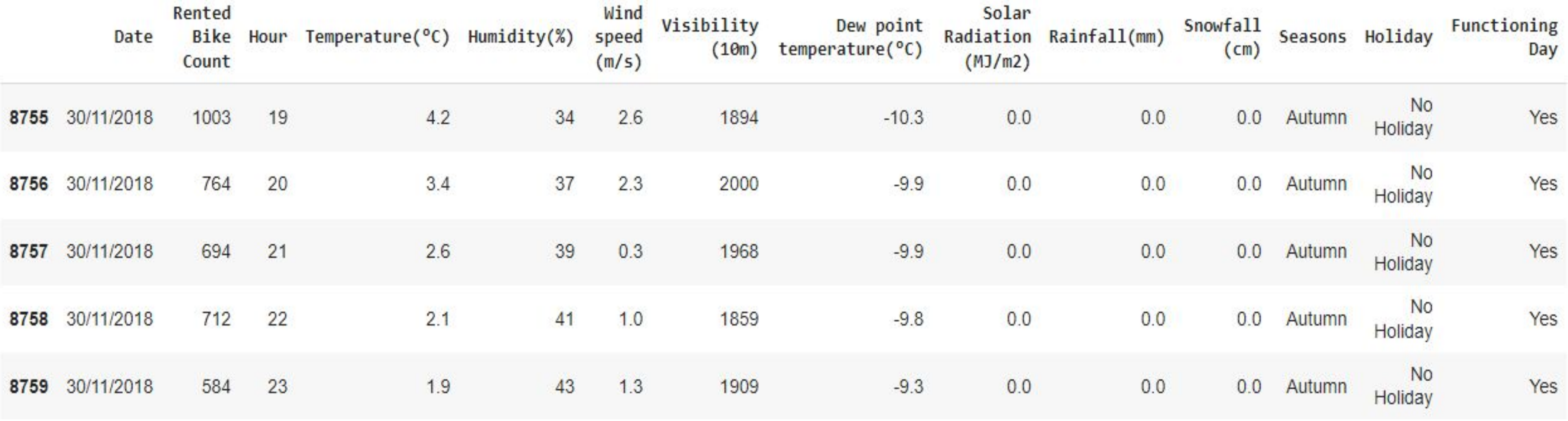
- Topic – “Bike Sharing Demand Prediction”
- Problem Statement -- Explore and Analyze the the prediction of bike count at each hour for stable supply of rental bikes reducing the waiting time.

- Bike rentals are popular service in recent times in many urban cities. Cheap rates and convenience is what making it a popular business. So for more profit it has to be always ready to fulfil demand at different locations at any point of time.

- The goal of this project is to predict number of rental bike demand for each hour of day to enhance the mobility comfort and lessens the waiting time.

Data Set Information

- This dataset contains 8760 observations and 14 features of one year from 01/12/2017 to 31/12/2018.
- Seasons, Holiday & Functioning day are three categorical features.
- This dataset also consists of numerical features namely temperature, humidity, wind speed, dew point temperature, solar radiation, snowfall, rainfall of that particular hour of the day.



	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	Holiday	Functioning Day
8755	30/11/2018	1003	19	4.2	34	2.6	1894	-10.3	0.0	0.0	0.0	Autumn	No Holiday	Yes
8756	30/11/2018	764	20	3.4	37	2.3	2000	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8757	30/11/2018	694	21	2.6	39	0.3	1968	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8758	30/11/2018	712	22	2.1	41	1.0	1859	-9.8	0.0	0.0	0.0	Autumn	No Holiday	Yes
8759	30/11/2018	584	23	1.9	43	1.3	1909	-9.3	0.0	0.0	0.0	Autumn	No Holiday	Yes

Feature Summary

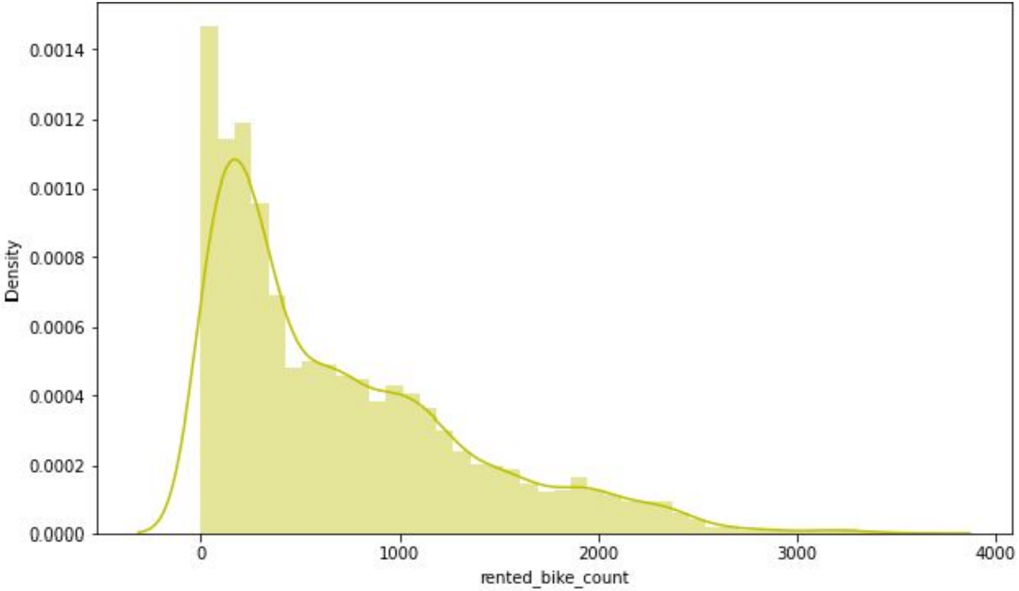
- Date: Date range in dataset is from 01/12/2017 to 30/11/2018
- Rented Bike Count: No. of bikes at each hour of day.
- Hour: Hour is in digital form
- Seasons: Winter, Spring, Autumn, Summer
- Temperature: Temperature is in Celsius.
- Snowfall: cm
- Rainfall: mm
- Wind speed: m/s
- Solar Radiation: It shows radiation intensity in MJ/m2
- Dew Point Temperature: It tells about the temperature in beginning of the day in Celsius.
- Visibility: 10m
- Humidity: %
- Holiday: Holiday / No Holiday
- Functioning Day: NoFunc(Non Functional hours), Fun(Functional hours)

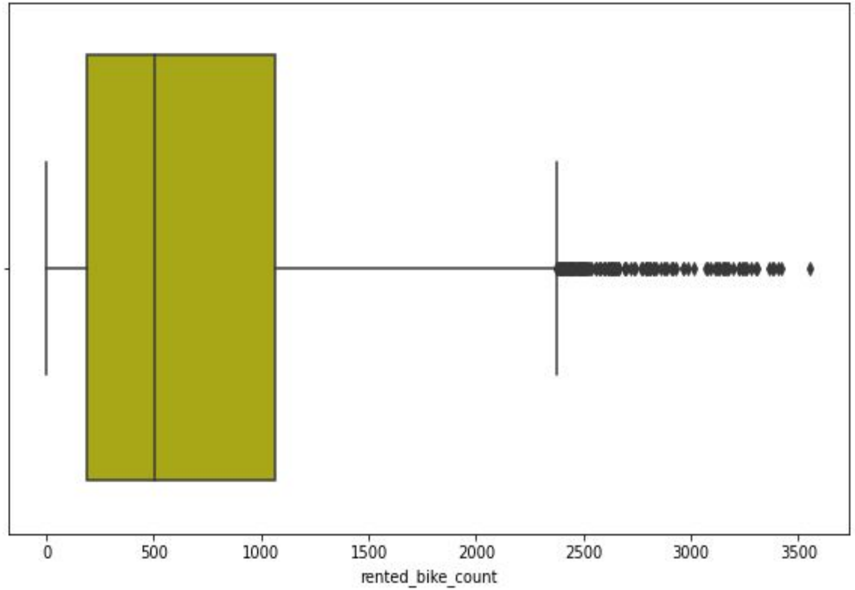
Data Wrangling

- There are no missing value present in our dataset.
- There are no null values present in our dataset.
- There are no duplicate value present.
- Target variable is ‘rented bike count’.
- Convert the date column into three column ‘year’, ’month’, ’day’.
- Name of the features are changed for smooth processing, ‘rented_bike_count’, ‘hour’, ‘temperature’, ‘humidity’, ‘wind_speed’, ‘visibility’, ‘dew_point_temp’, ‘solar_radiation’, ‘rainfall’, ‘snowfall’, ‘seasons’, ‘holiday’, ‘functioning_day’, ‘month’, ‘weekdays_weekend’

Normalisation of Target Variable

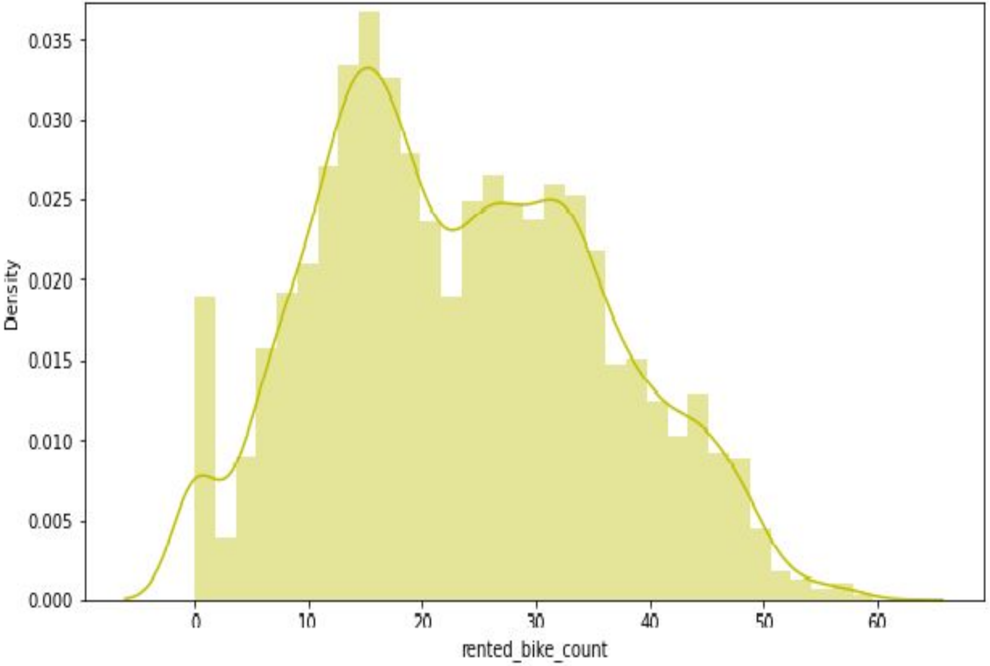
- The distribution plot has moderate right skewness.
- The box plot shows the presence of outliers in target variables.
- As it is assume that distribution of dependent variable has to be normal for linear regression, hence square root operation is performed to make it normal.

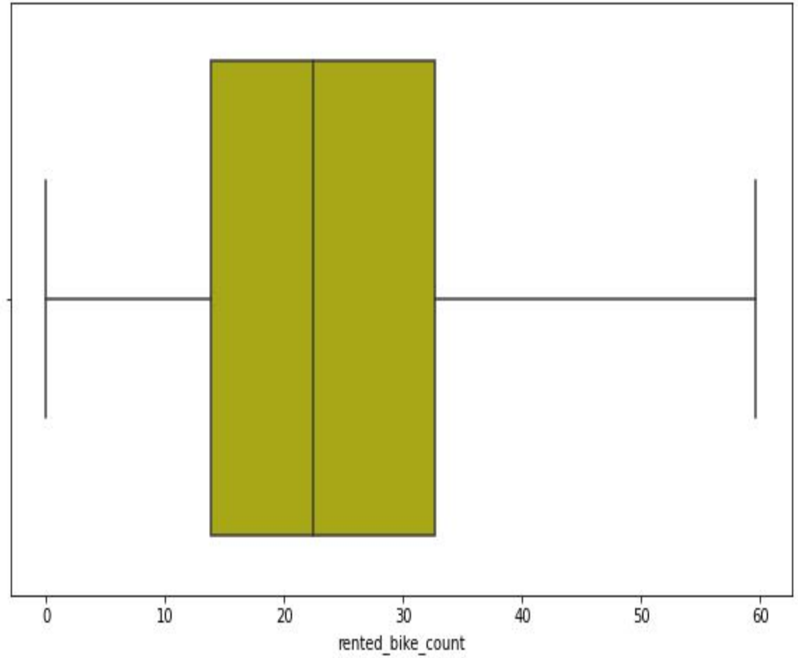




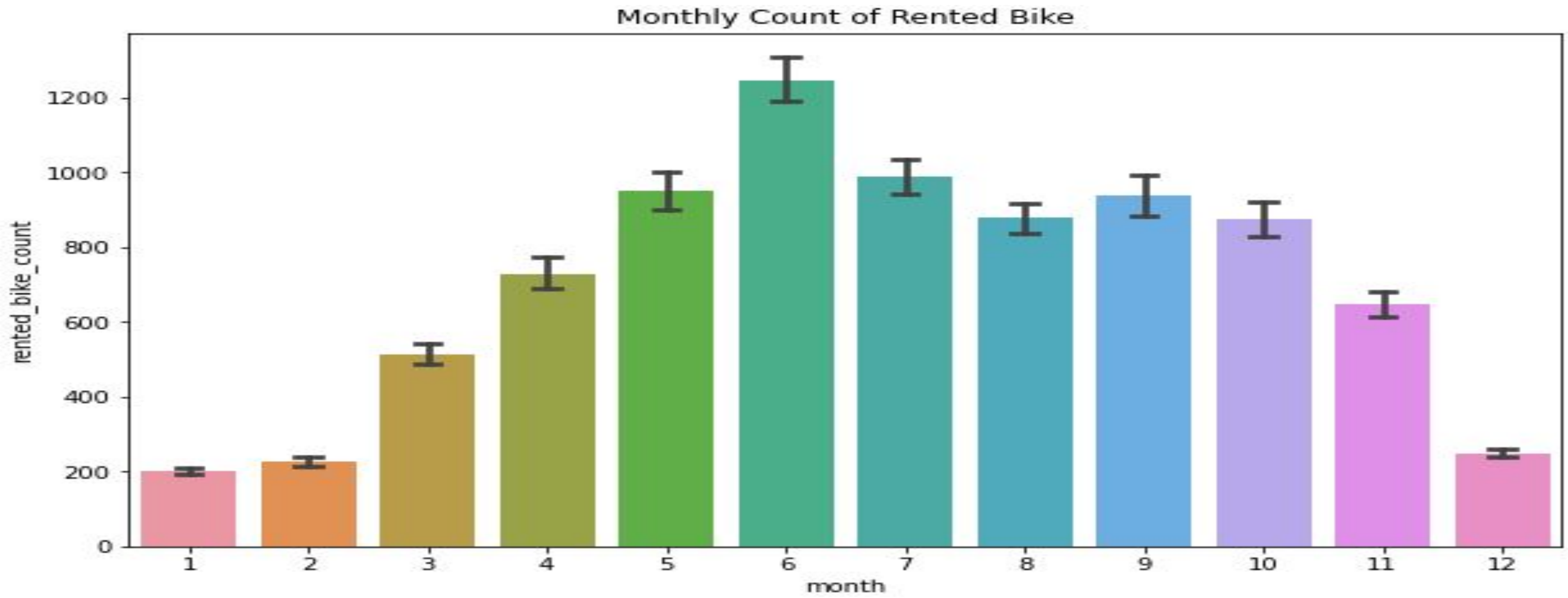
Normalisation of Target Variable(Continued)

- After square root operation, here we get almost normal distribution.
- Also subsequently we see that there is no outliers present after square root operation.



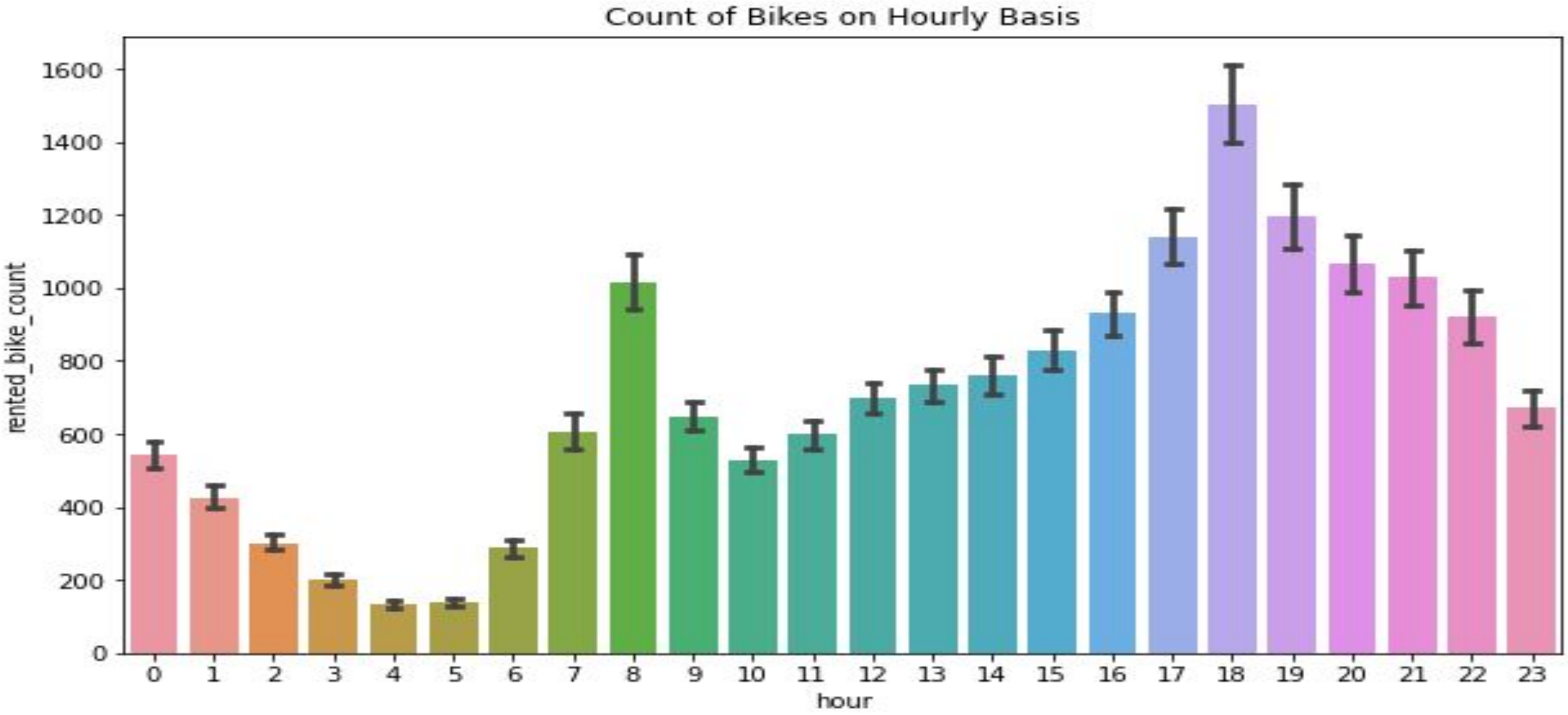


Count of Bikes over a period of Month

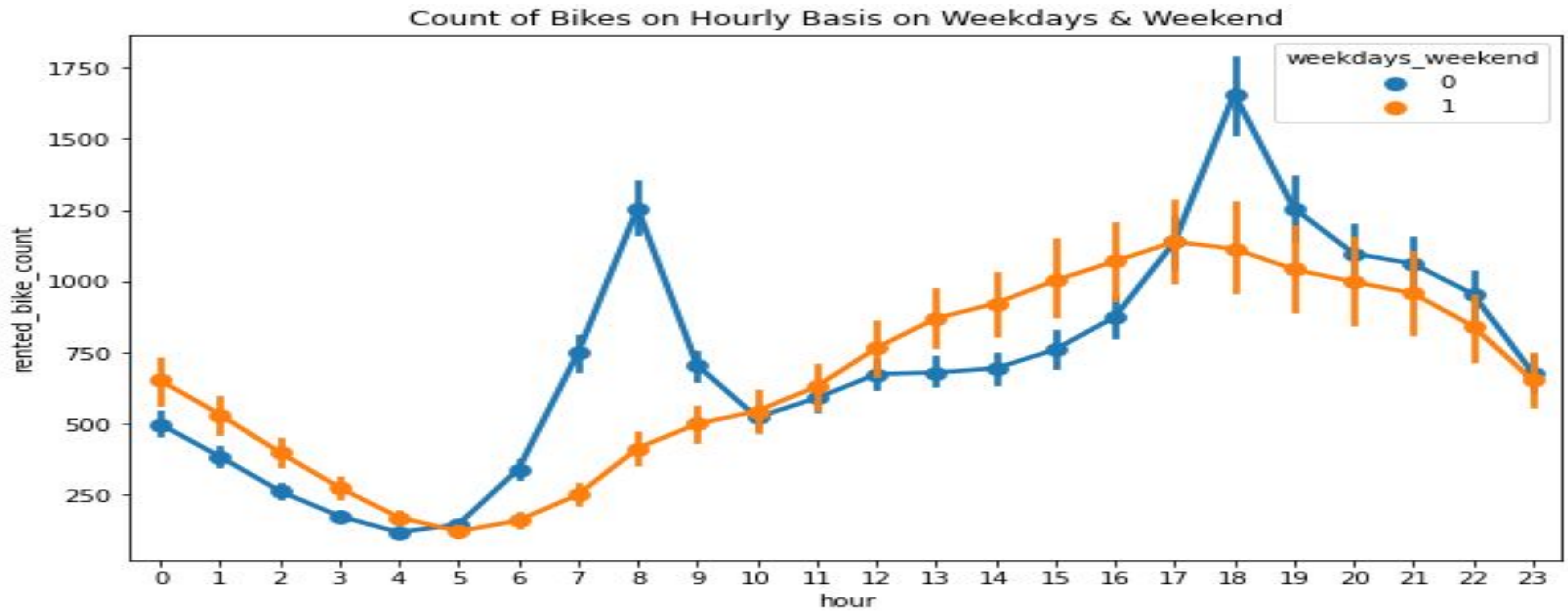


- Demand of rented bike is high in summer season which is from May to October.

Count of Bikes for each Hour in a year

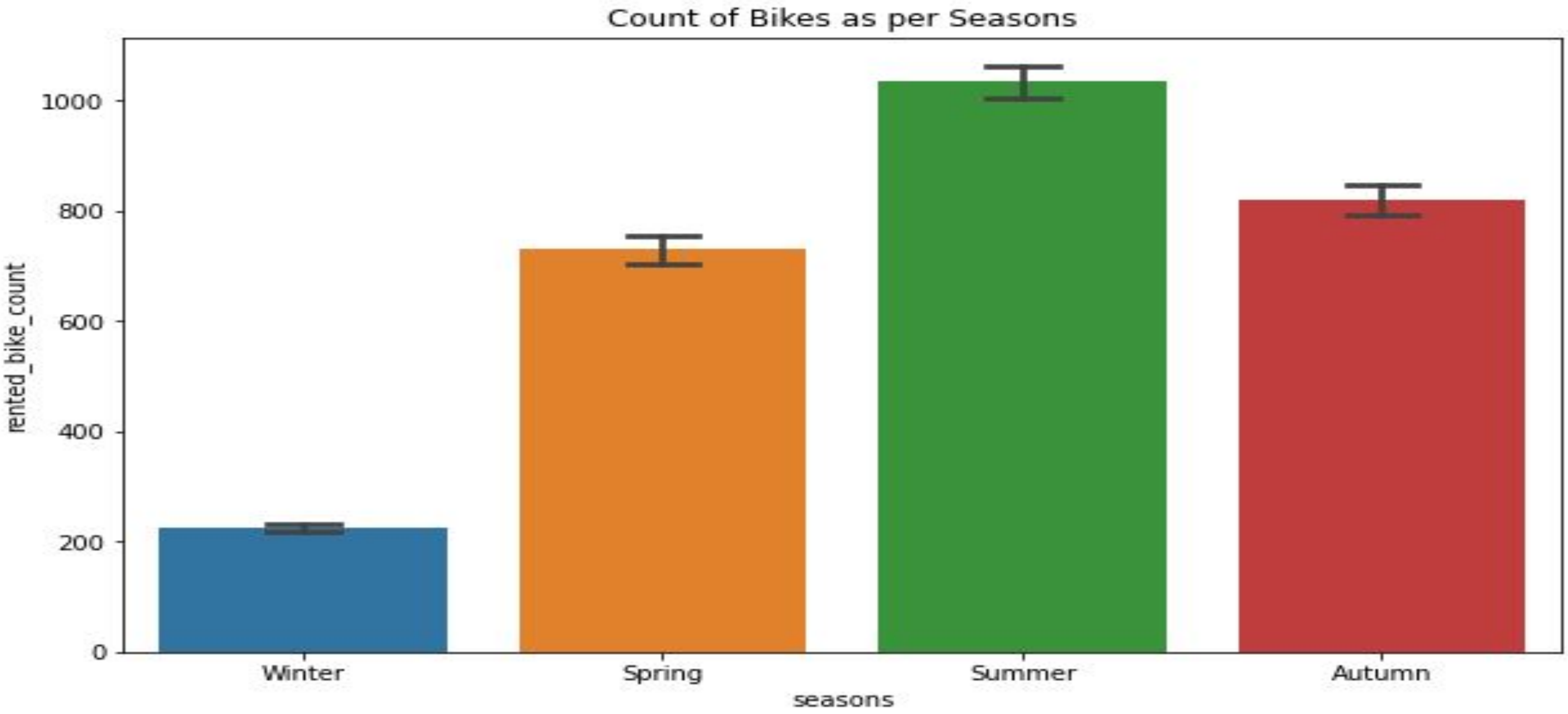


Count of Bikes on weekdays_weekends

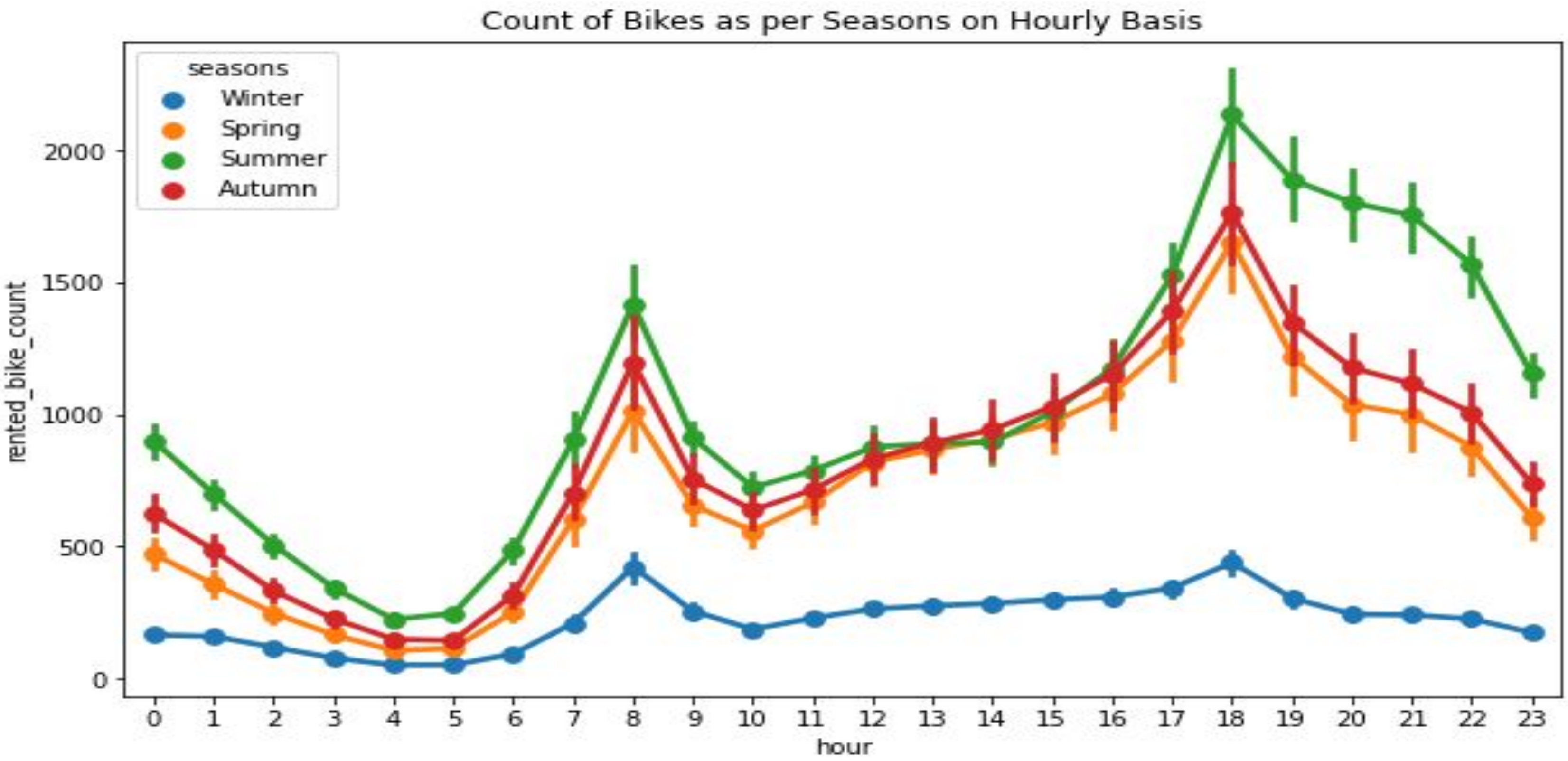


- On Weekdays demand of bike is high between 7 to 9 in morning & 5 to 7 in evening.
- On Weekends demand of bike is low in the morning but as the day progress demand of bike also increases.

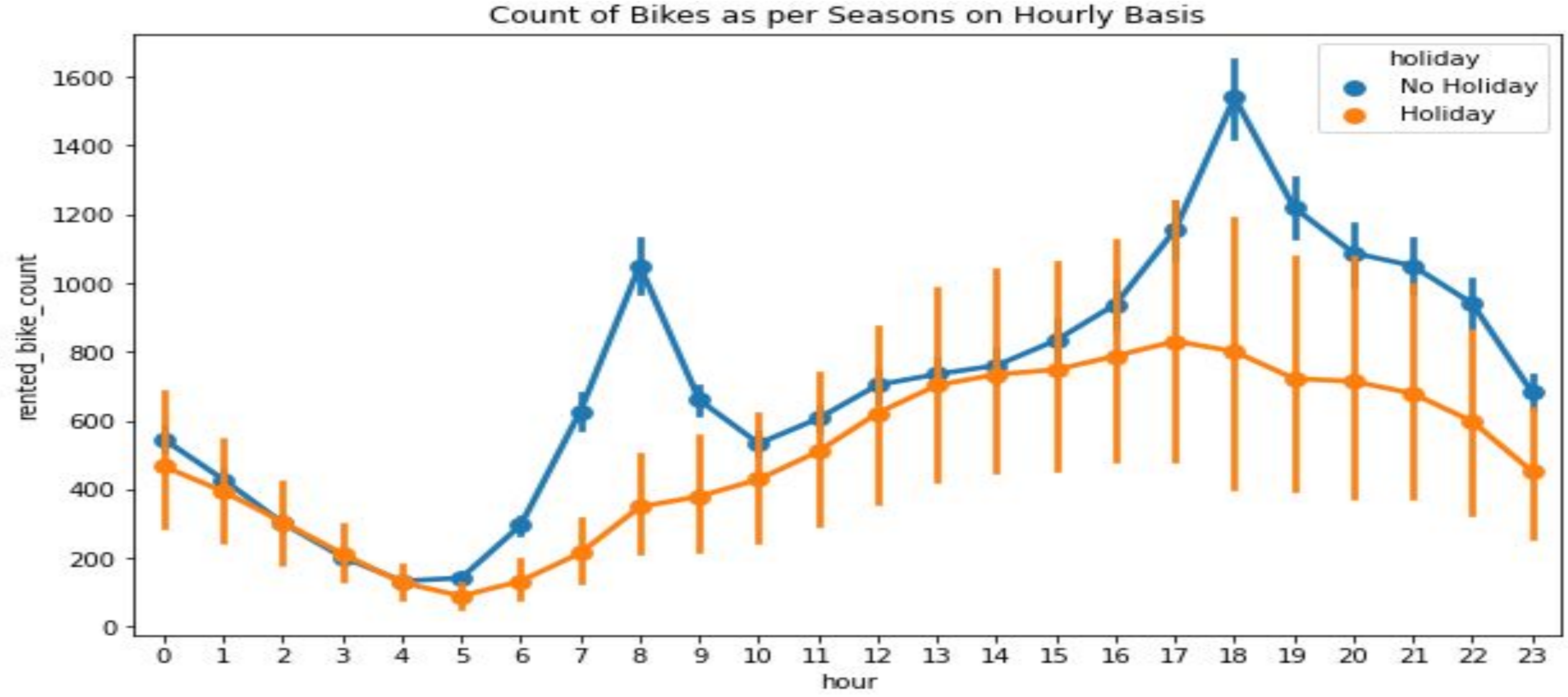
Count of Bikes as per Seasons



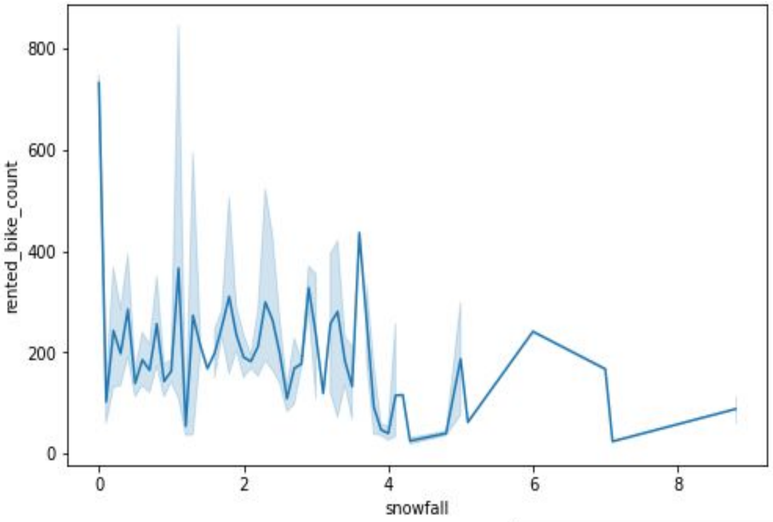
Count of bikes as per Seasons for each Hour

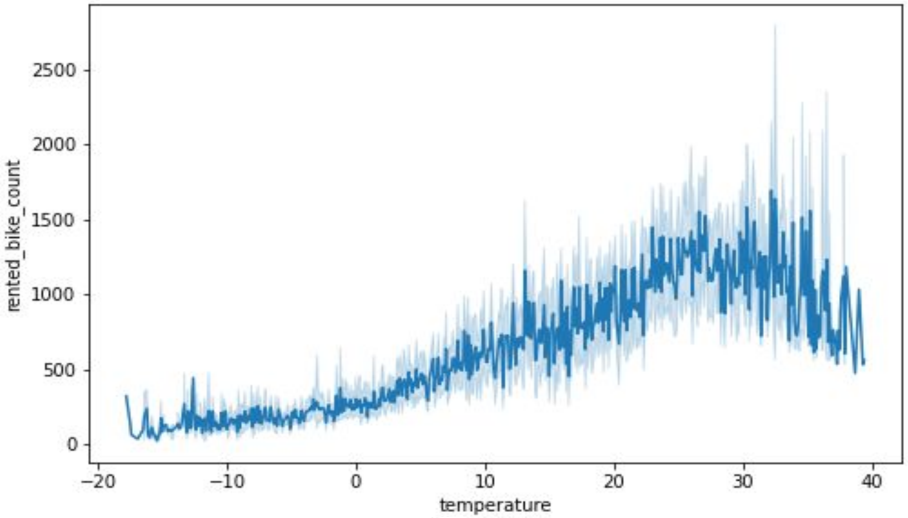


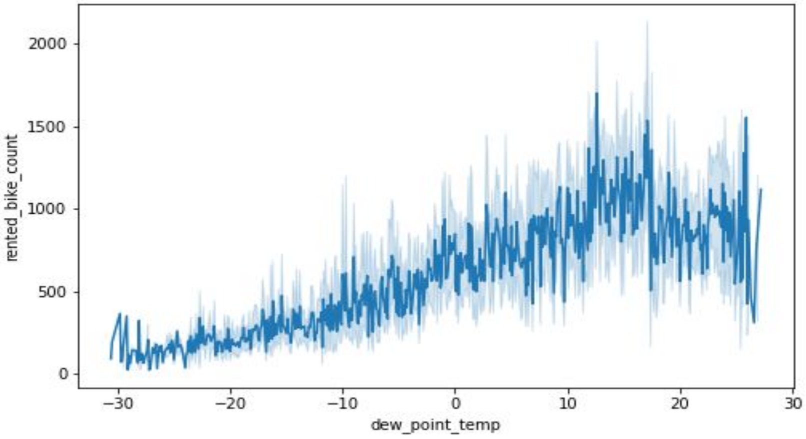
Count of Bikes on Holiday for each Hour



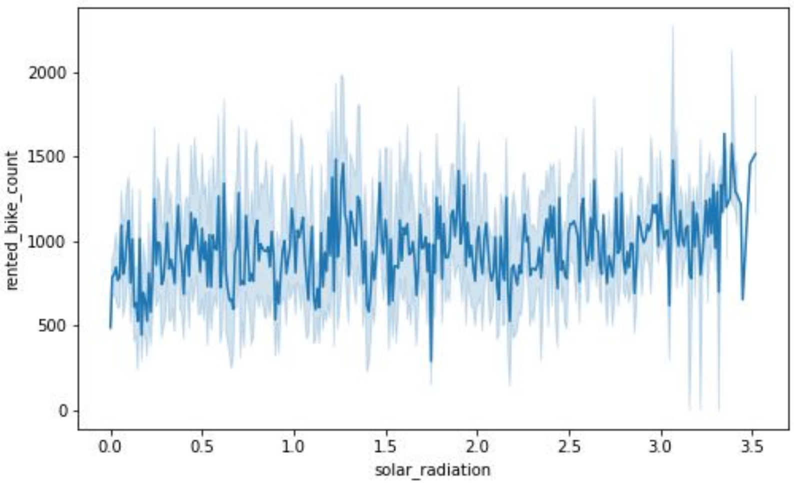
Analysis between Numerical variable & Dependent Variable

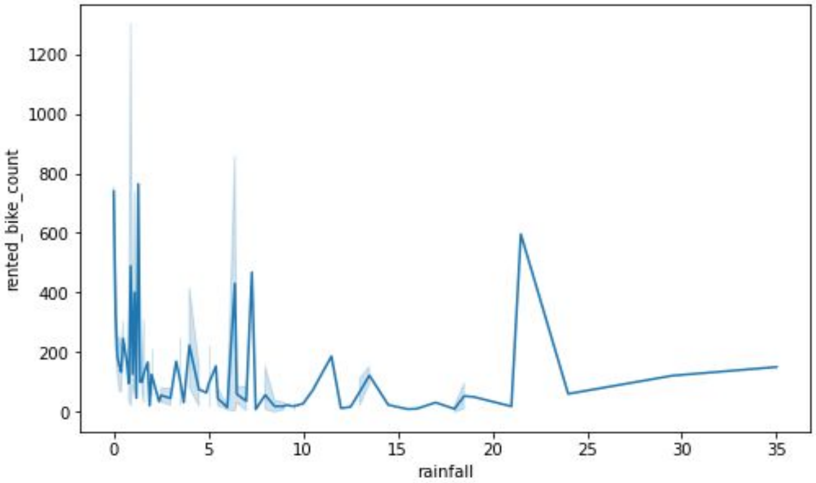


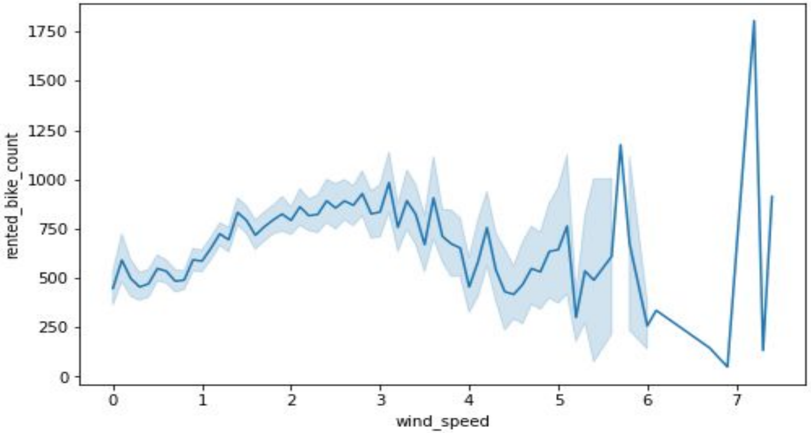




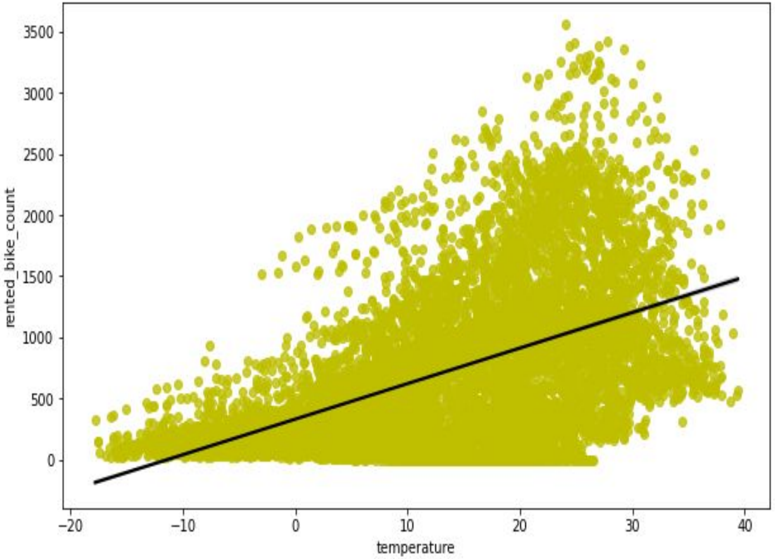
Analysis between Numerical variable & Dependent Variable

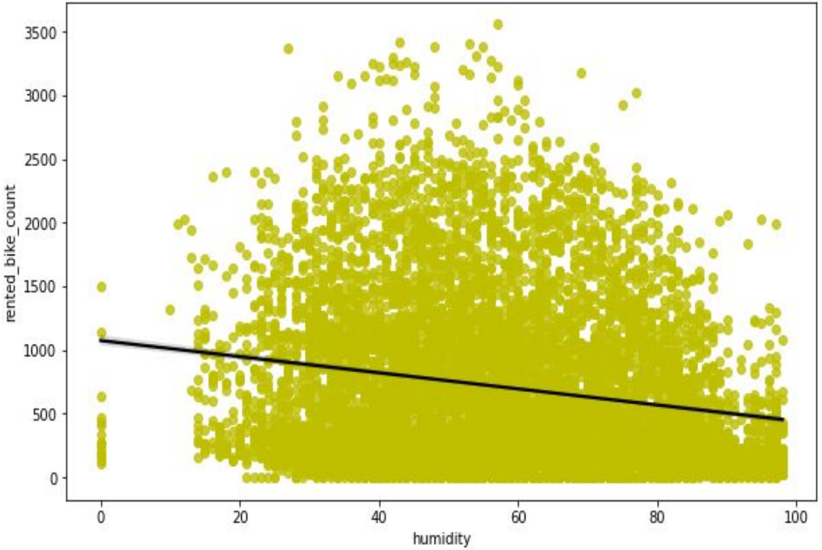


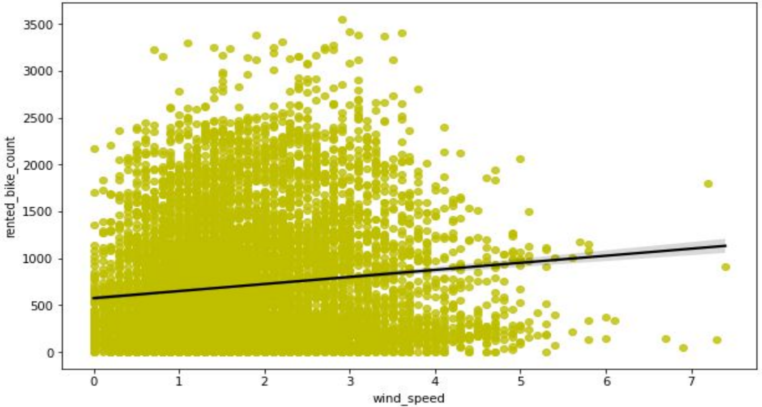


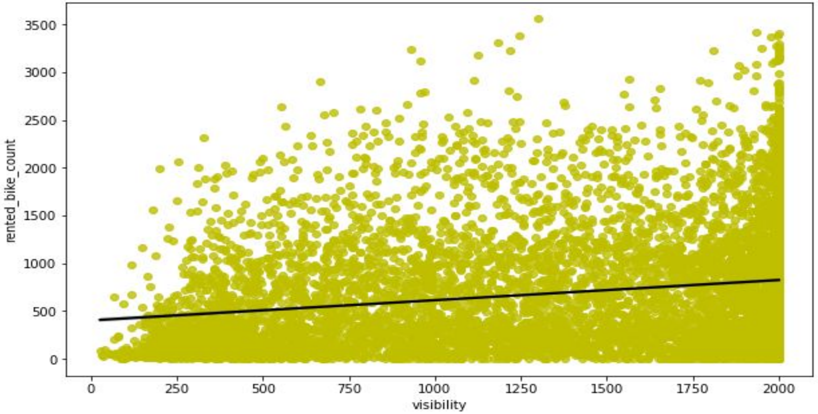


Regression Plot

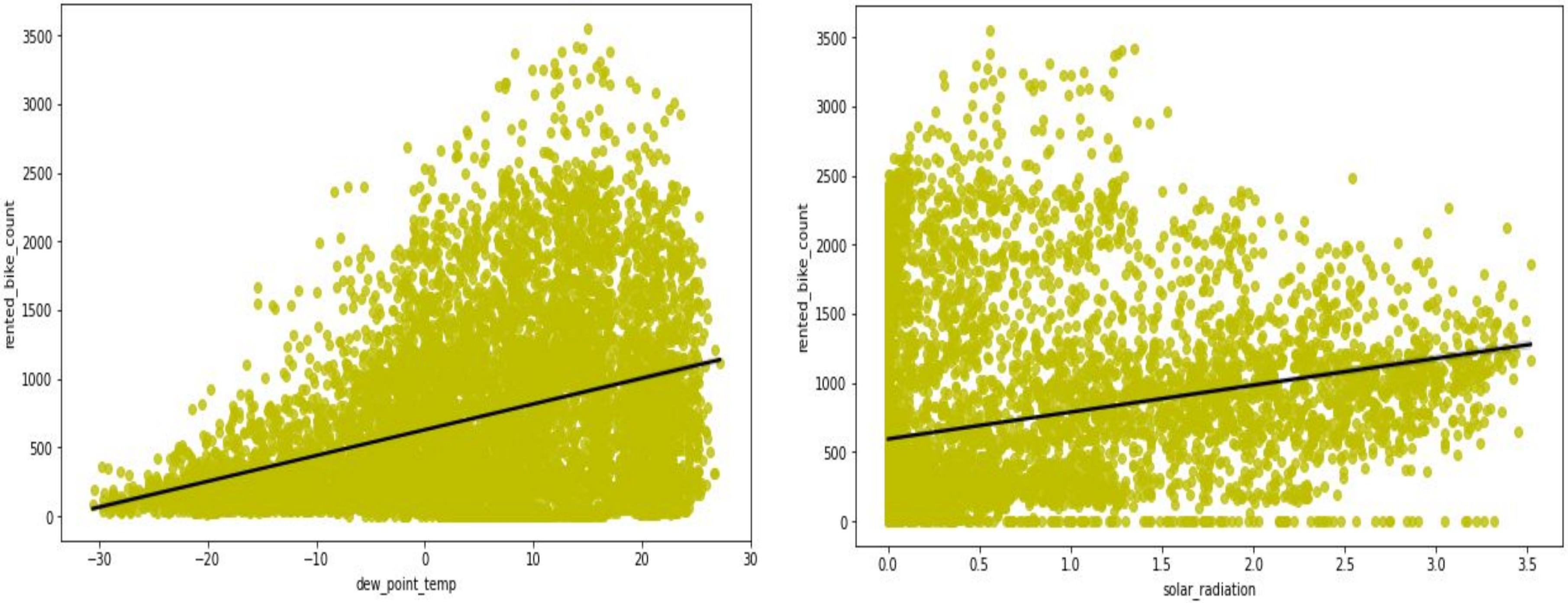




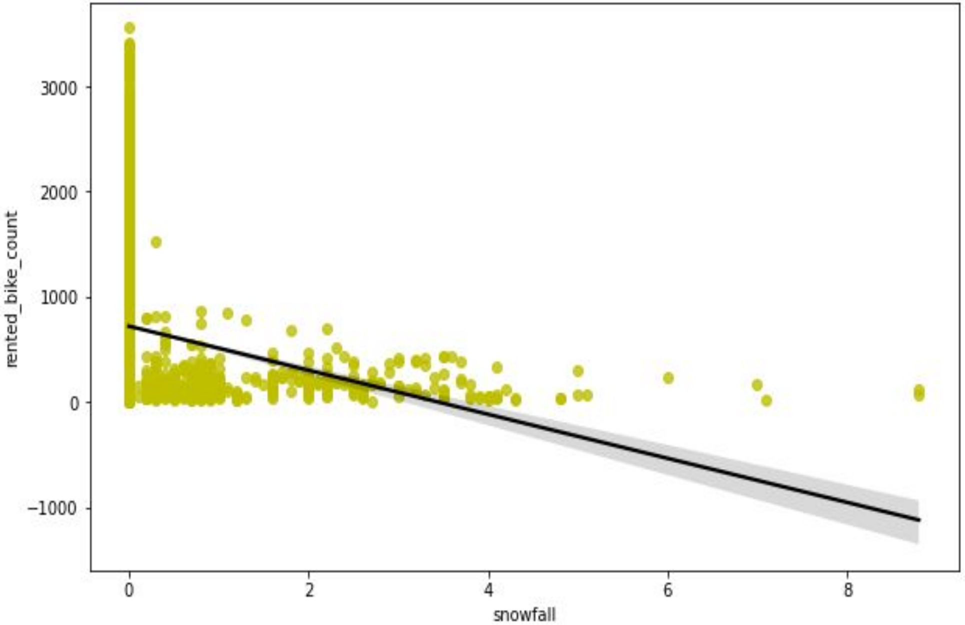


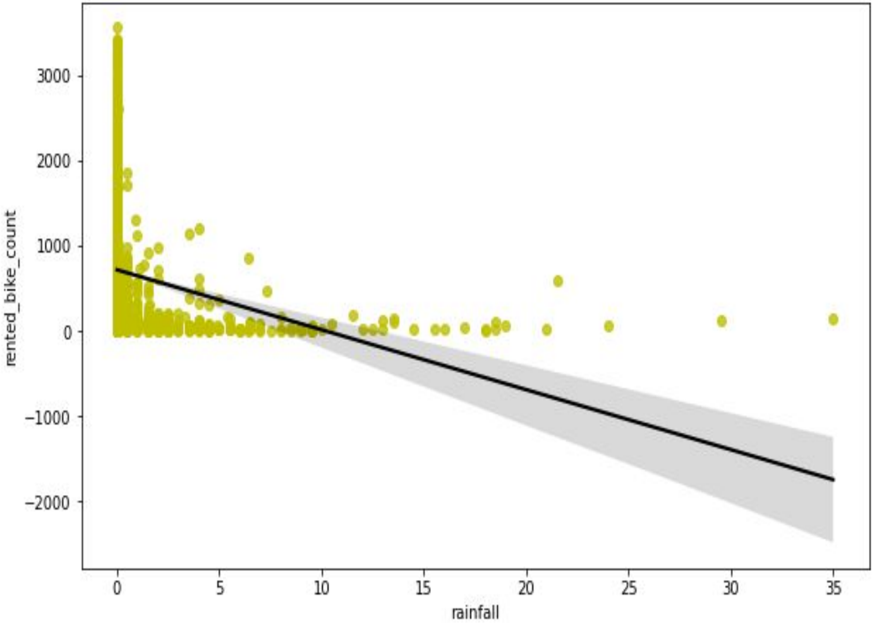


Regression Plot(Contd.)



Regression Plot(Contd.)

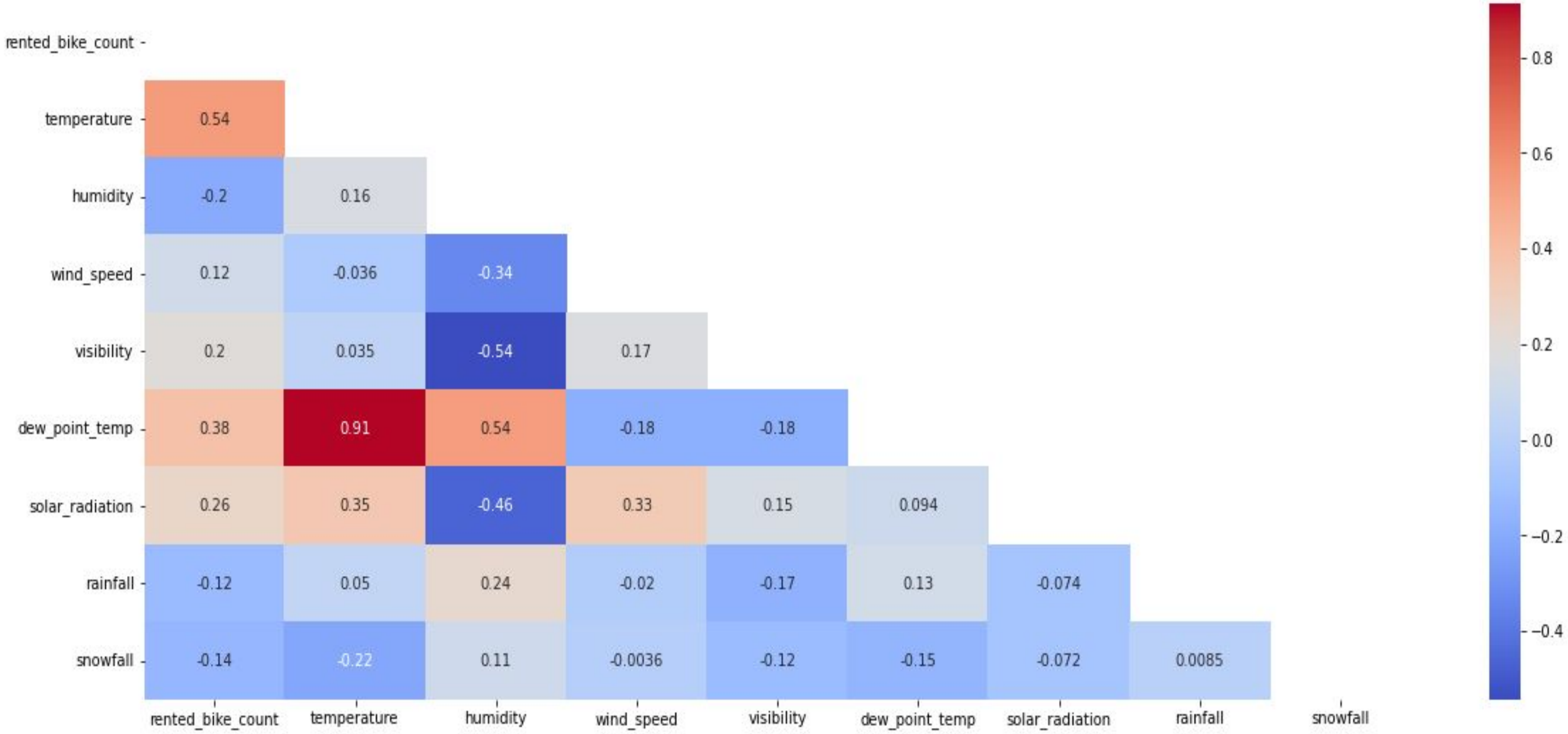




Regression Plot Conclusion

- From regression plot on previous slide we can see that features like temperature, wind speed, visibility, dew point temperature, solar radiation, are positively related to target variable.
- Positively correlated with target variable means the target variable will increase with the increase in value of these features.
- Rainfall, snowfall and humidity are negatively related to target variable,ie, rented bike count decreases with increase in values of these features.

Correlation Matrix



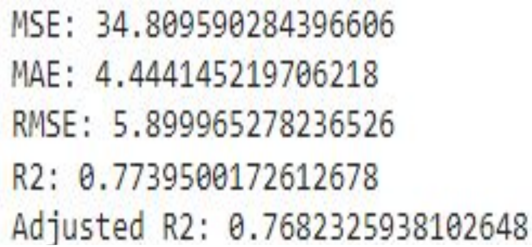
	rented_bike_count	temperature	humidity	wind_speed	visibility	dew_point_temp	solar_radiation	rainfall	snowfall
rented_bike_count									
temperature	0.54								
humidity	-0.2	0.16							
wind_speed	0.12	-0.036	-0.34						
visibility	0.2	0.035	-0.54	0.17					
dew_point_temp	0.38	0.91	0.54	-0.18	-0.18				
solar_radiation	0.26	0.35	-0.46	0.33	0.15	0.094			
rainfall	-0.12	0.05	0.24	-0.02	-0.17	0.13	-0.074		
snowfall	-0.14	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	

Model Building

- Linear Regression
- Elastic Net Regression
- Lasso Regression
- Ridge Regression
- Decision Tree Regression
- Random Forest Regression

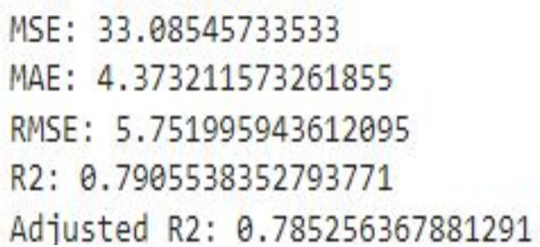
Linear Regression

Train Result

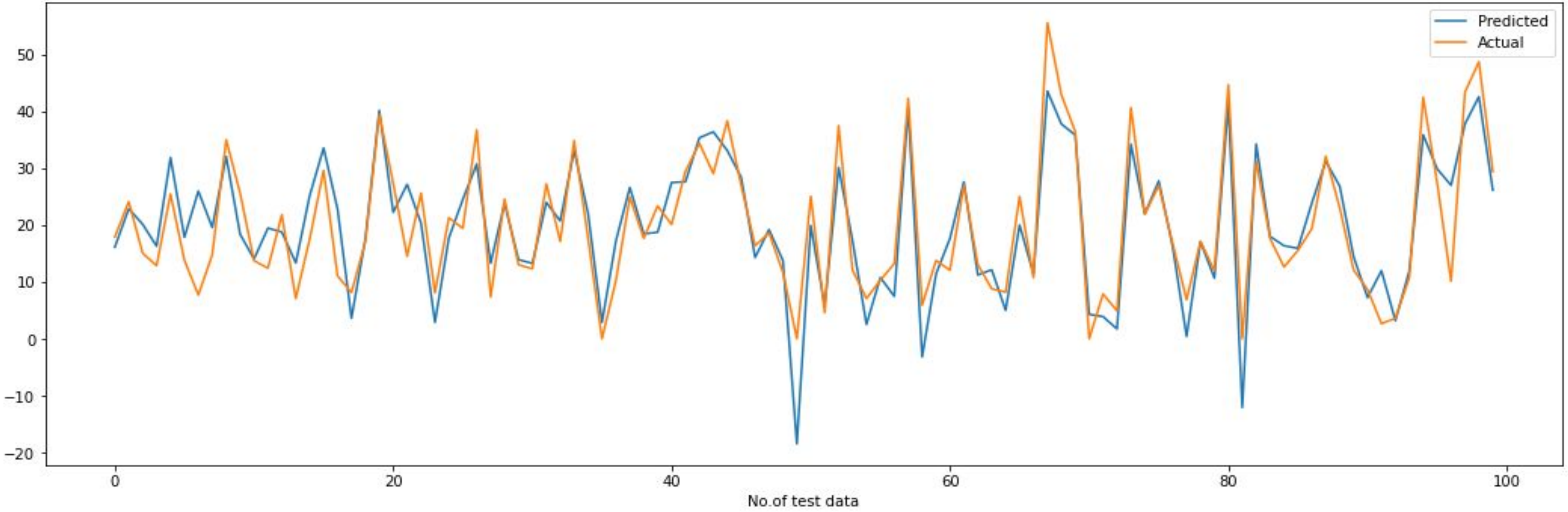


```
MSE: 34.809590284396606
MAE: 4.444145219706218
RMSE: 5.899965278236526
R2: 0.7739500172612678
Adjusted R2: 0.7682325938102648
```

Test Result

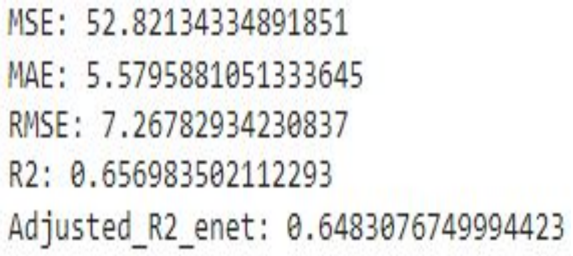


```
MSE: 33.08545733533
MAE: 4.373211573261855
RMSE: 5.751995943612095
R2: 0.7905538352793771
Adjusted R2: 0.785256367881291
```



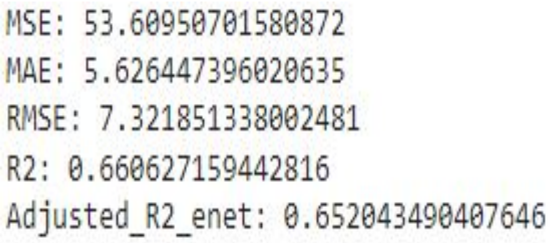
Elastic Net Regression

Train Result

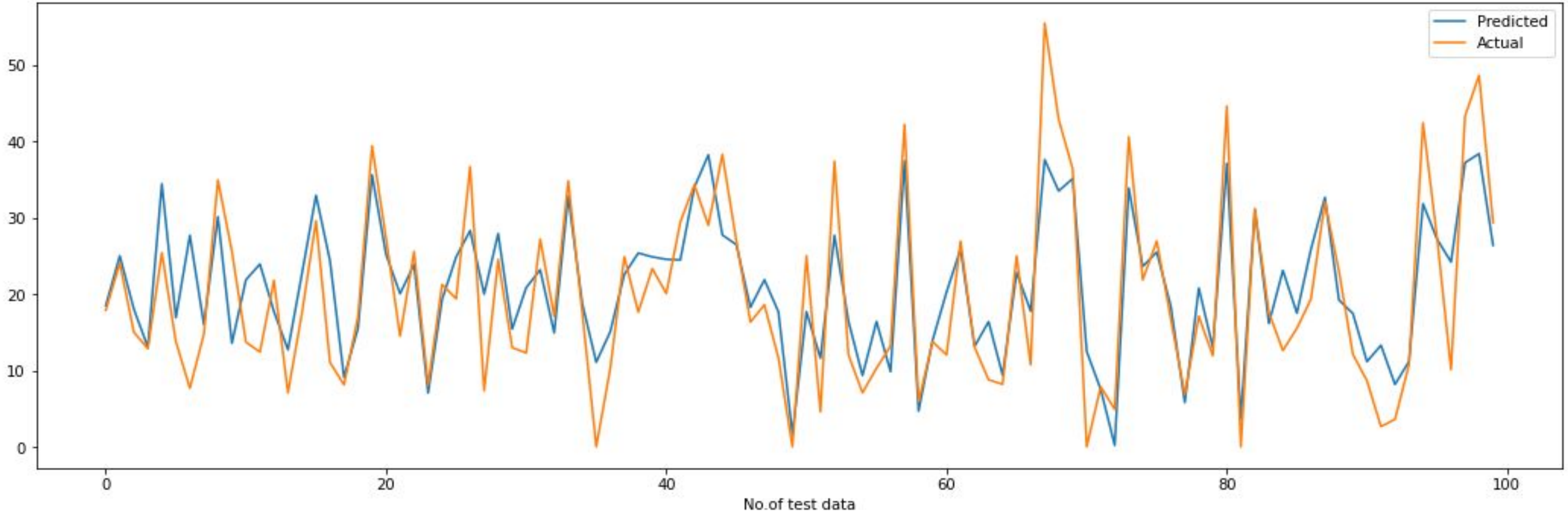


```
MSE: 52.82134334891851
MAE: 5.579588105133645
RMSE: 7.26782934230837
R2: 0.656983502112293
Adjusted_R2_enet: 0.6483076749994423
```

Test Result

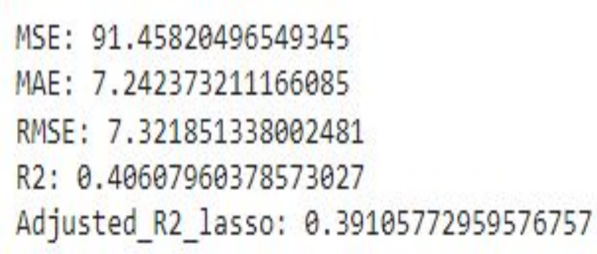


```
MSE: 53.60950701580872
MAE: 5.626447396020635
RMSE: 7.321851338002481
R2: 0.660627159442816
Adjusted_R2_enet: 0.652043490407646
```



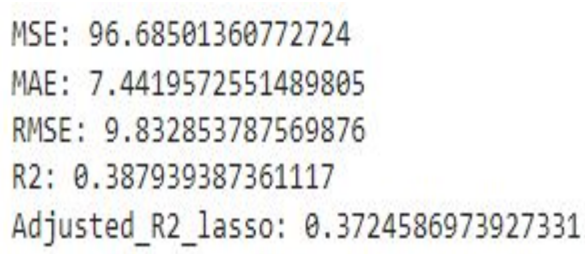
Lasso Regression

Train Result

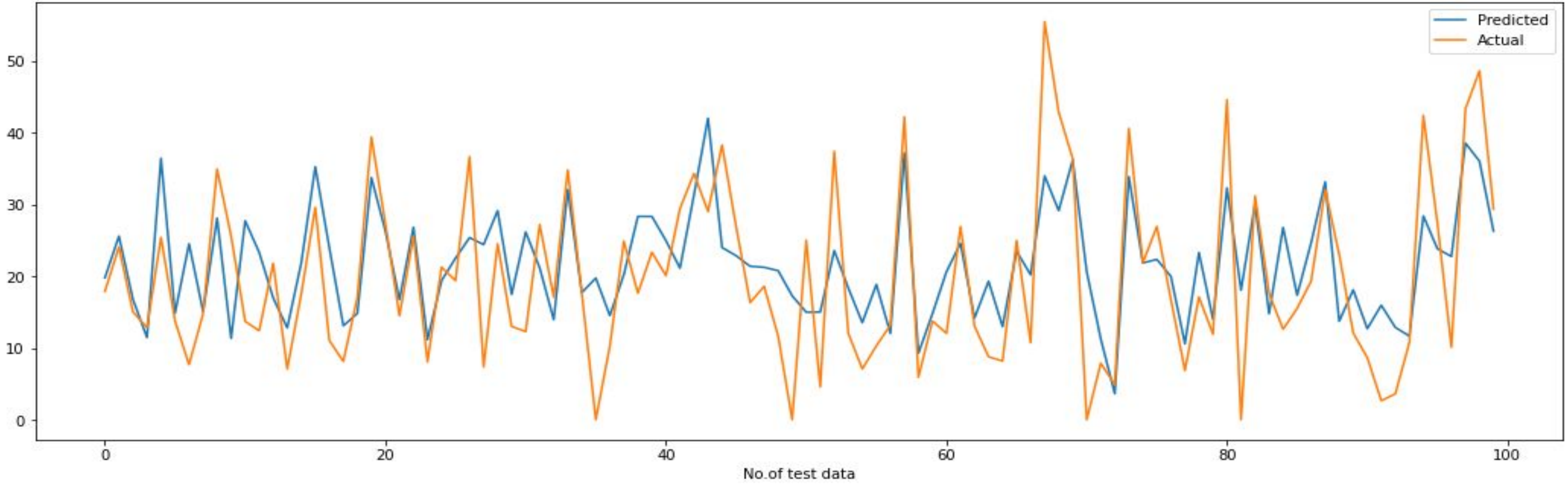


```
MSE: 91.45820496549345
MAE: 7.24237321166085
RMSE: 7.321851338002481
R2: 0.406079603785730 27
Adjusted_R2_lasso: 0.39105772959576757
```

Test Result

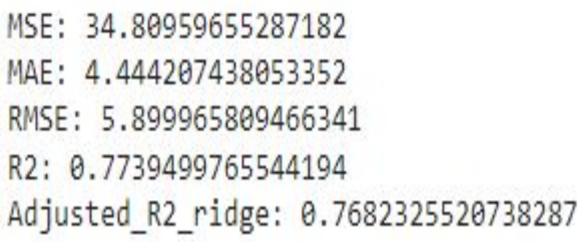


```
MSE: 96.68501360772724
MAE: 7.441957255148980 5
RMSE: 9.832853787569876
R2: 0.387939387361117
Adjusted_R2_lasso: 0.372458697392733 1
```



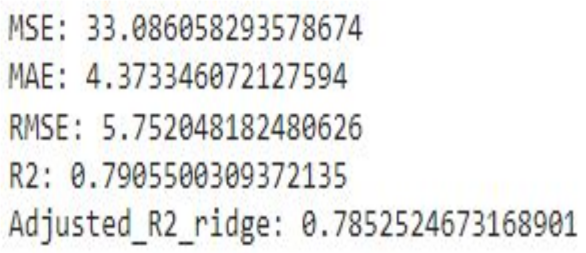
Ridge Regression

Train Result

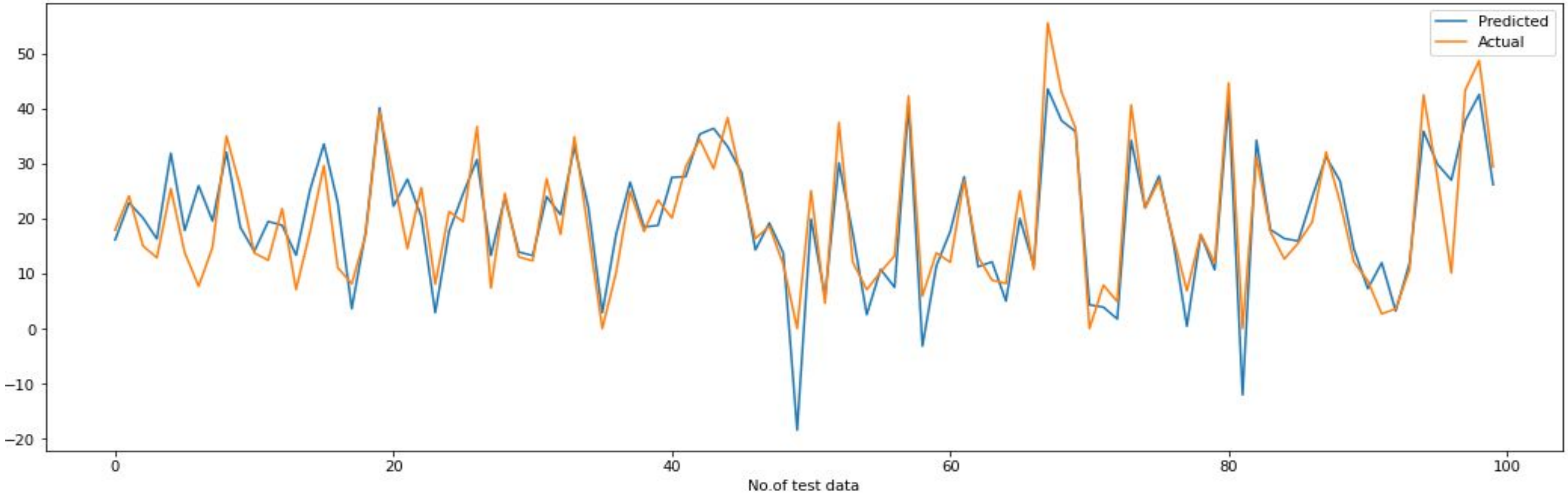


```
MSE: 34.80959655287182
MAE: 4.444207438053352
RMSE: 5.899965809466341
R2: 0.7739499765544194
Adjusted_R2_ridge: 0.7682325520738287
```

Test Result

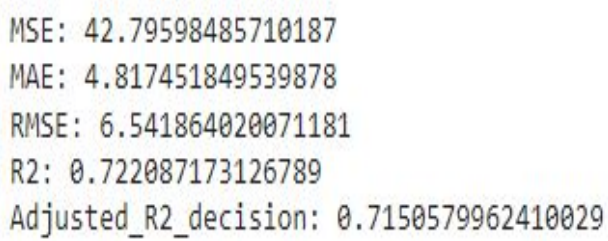


```
MSE: 33.086058293578674
MAE: 4.373346072127594
RMSE: 5.752048182480626
R2: 0.7905500309372135
Adjusted_R2_ridge: 0.7852524673168901
```



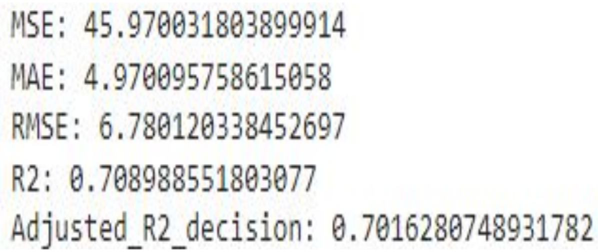
Decision Tree

Train Result

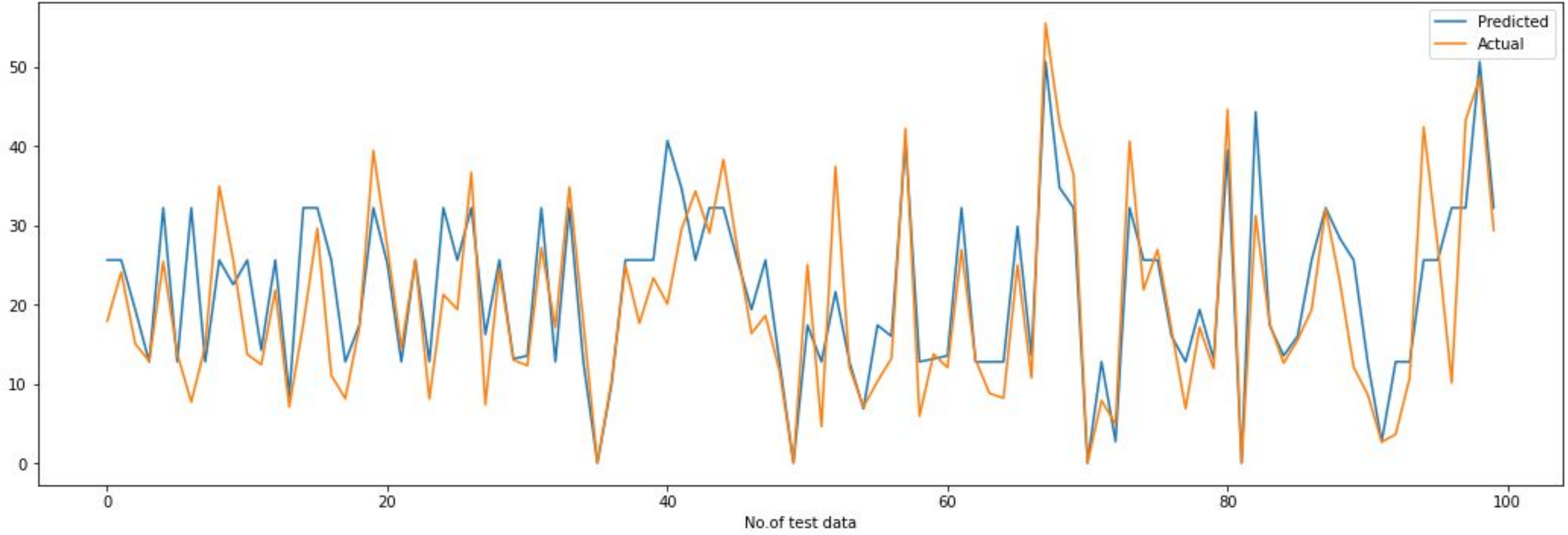


```
MSE: 42.79598485710187
MAE: 4.817451849539878
RMSE: 6.541864020071181
R2: 0.722087173126789
Adjusted_R2_decision: 0.7150579962410029
```

Test Result

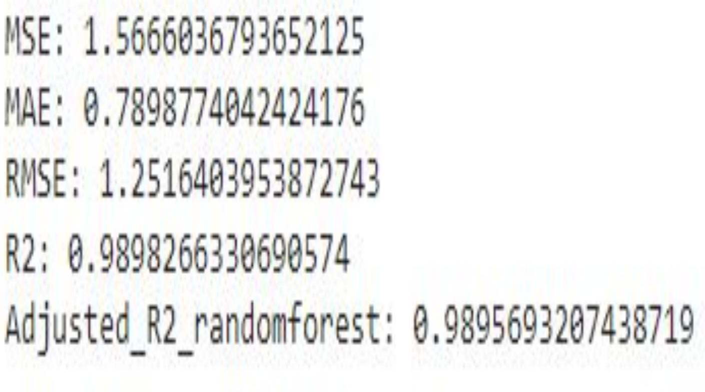


```
MSE: 45.970031803899914
MAE: 4.970095758615058
RMSE: 6.780120338452697
R2: 0.708988551803077
Adjusted_R2_decision: 0.7016280748931782
```



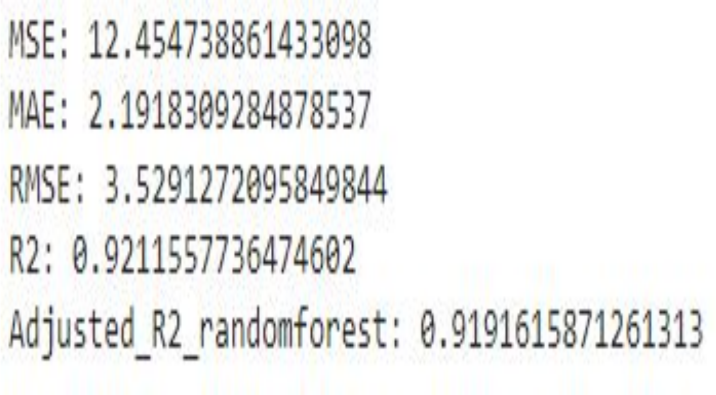
Random Forest

Train Result



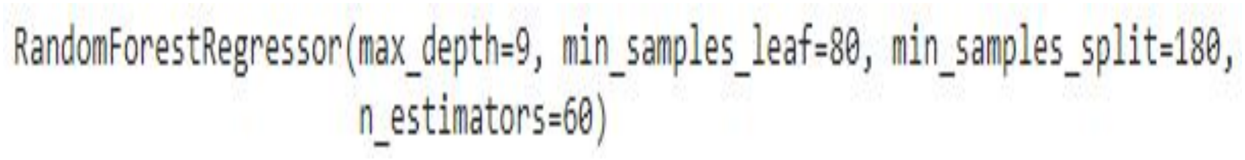
```
MSE: 1.566036793652125
MAE: 0.7898774042424176
RMSE: 1.2516403953872743
R2: 0.9898266330690574
Adjusted_R2_randomforest: 0.9895693207438719
```

Test Result



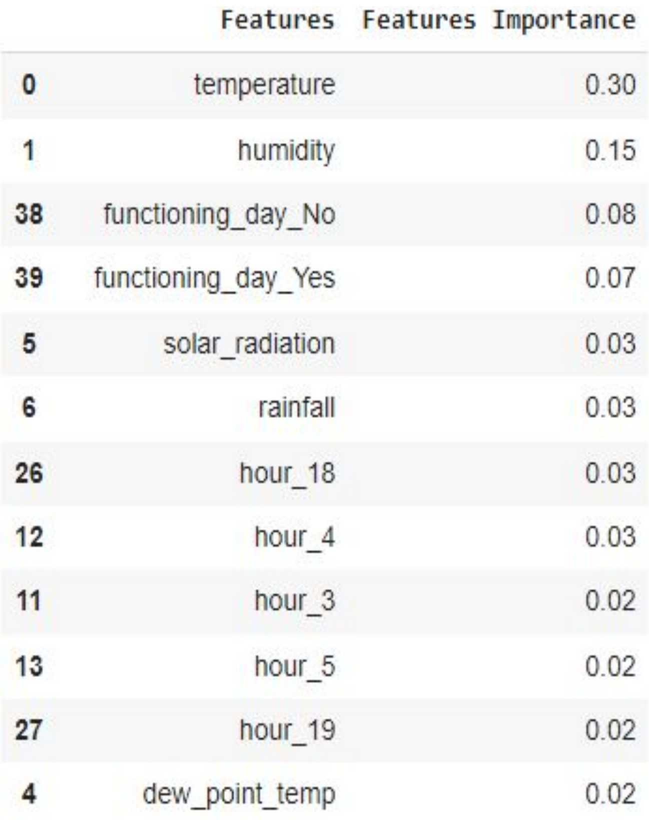
```
MSE: 12.454738861433098
MAE: 2.1918309284878537
RMSE: 3.5291272095849844
R2: 0.9211557736474602
Adjusted_R2_randomforest: 0.9191615871261313
```

Hyperparameter

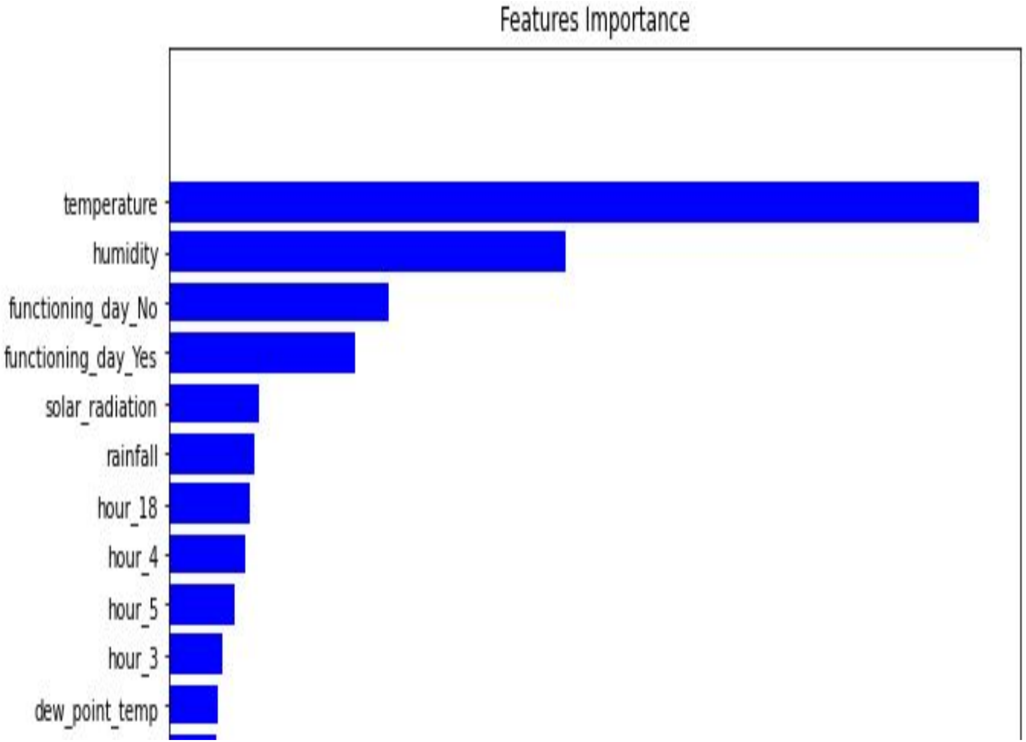


```
RandomForestRegressor(max_depth=9, min_samples_leaf=80, min_samples_split=180,
n_estimators=60)
```

Random Forest



	Features	Features Importance
0	temperature	0.30
1	humidity	0.15
38	functioning_day_No	0.08
39	functioning_day_Yes	0.07
5	solar_radiation	0.03
6	rainfall	0.03
26	hour_18	0.03
12	hour_4	0.03
11	hour_3	0.02
13	hour_5	0.02
27	hour_19	0.02
4	dew_point_temp	0.02

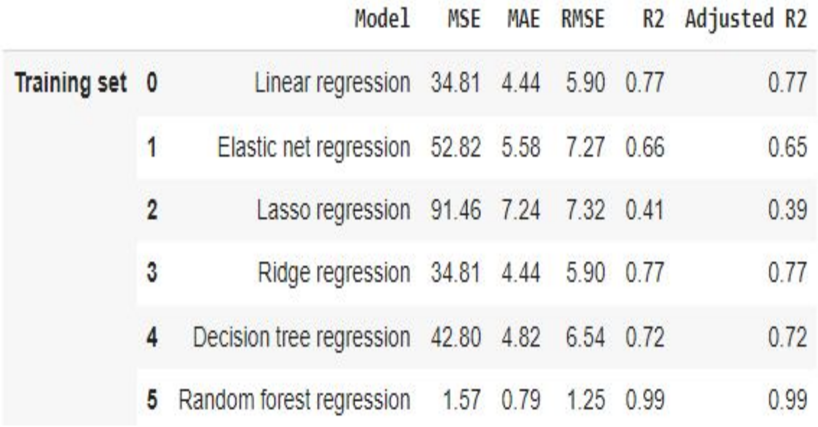


Challenges

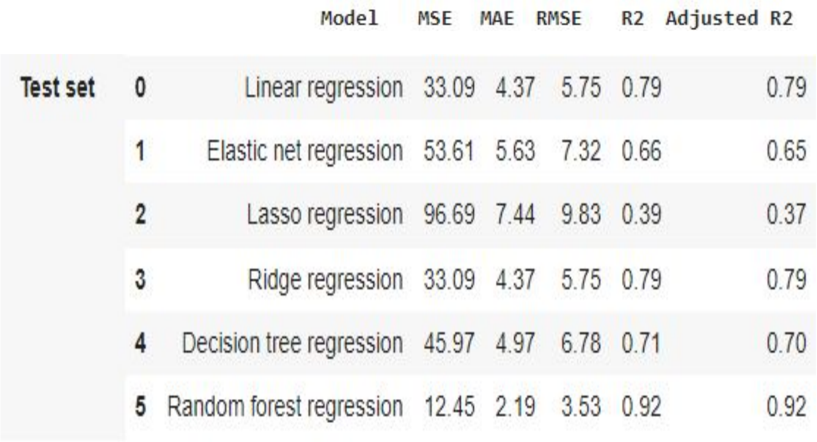
- Data Cleaning
- Data mining
- Feature Engineering
- Feature Selection
- Model optimization
- Hyperparameter Tuning
- Deciding the flow of presentation

Overall Conclusion

- After comparing the root mean squared error and mean absolute error of train and test result of all the applied model on this dataset, I found that Random forest gave the highest r2 score of 99% and 92% on train and test data respectively. Hence, can be concluded that random forest is best model for predicting bike rental for each hour and lessens the waiting time and enhancing the mobility comfort.



		Model	MSE	MAE	RMSE	R2	Adjusted R2
Training set	0	Linear regression	34.81	4.44	5.90	0.77	0.77
	1	Elastic net regression	52.82	5.58	7.27	0.66	0.65
	2	Lasso regression	91.46	7.24	7.32	0.41	0.39
	3	Ridge regression	34.81	4.44	5.90	0.77	0.77
	4	Decision tree regression	42.80	4.82	6.54	0.72	0.72
	5	Random forest regression	1.57	0.79	1.25	0.99	0.99



		Model	MSE	MAE	RMSE	R2	Adjusted R2
Test set	0	Linear regression	33.09	4.37	5.75	0.79	0.79
	1	Elastic net regression	53.61	5.63	7.32	0.66	0.65
	2	Lasso regression	96.69	7.44	9.83	0.39	0.37
	3	Ridge regression	33.09	4.37	5.75	0.79	0.79
	4	Decision tree regression	45.97	4.97	6.78	0.71	0.70
	5	Random forest regression	12.45	2.19	3.53	0.92	0.92

Q&A