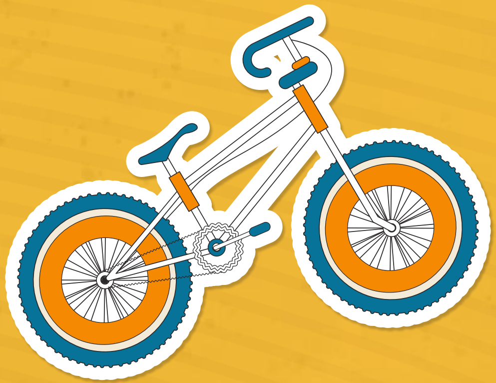
Capstone Project

Bike Sharing Demand Prediction



Sumbitted By

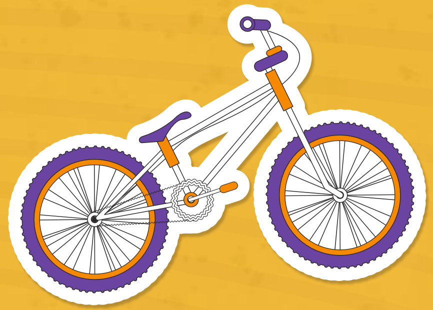
Poovarasan
Shipfriend0368@gmail.com

TABLE OF CONTENTS

Problem Description:

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.

Data description:



Dependent variable

- Rented Bike count - Count of bikes rented at each hour

Independent variables

- Date : year-month-day
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10 m
- 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 Summary:

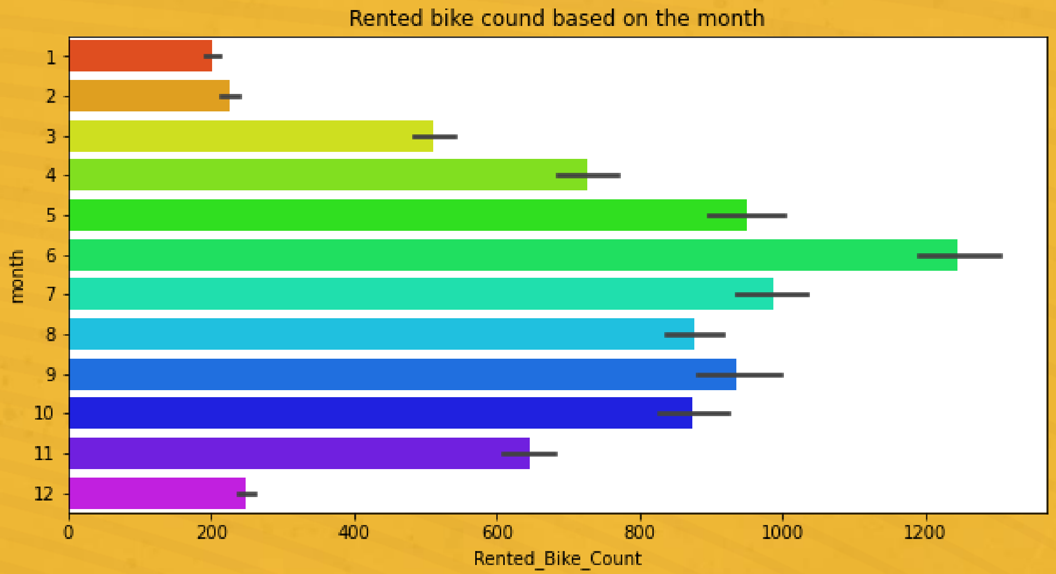
	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

- This dataset contains 8760 lines and 14 columns
- Numerical variables - temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall
- Categorical variables - seasons, holiday and functioning day
- Rented bike column - which we need to predict for new observations

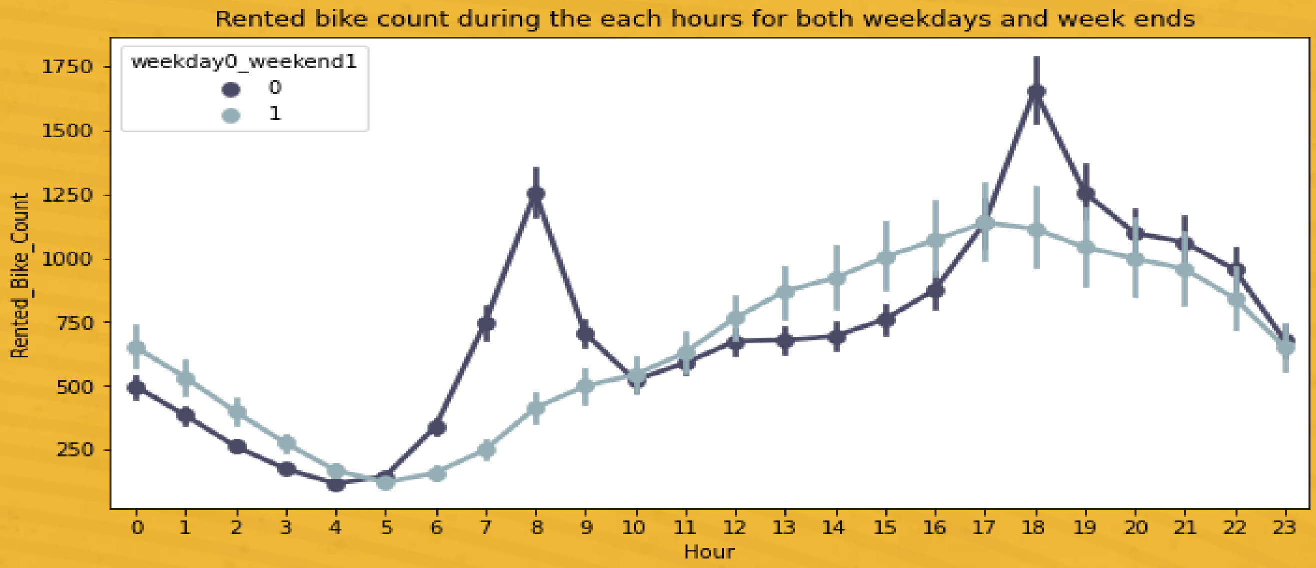
Exploratory Data Analysis:

Month

The demand for leased bikes is higher from months 4 to 10 compared to other months, as seen by the above bar plot. These months fall during the summertime.

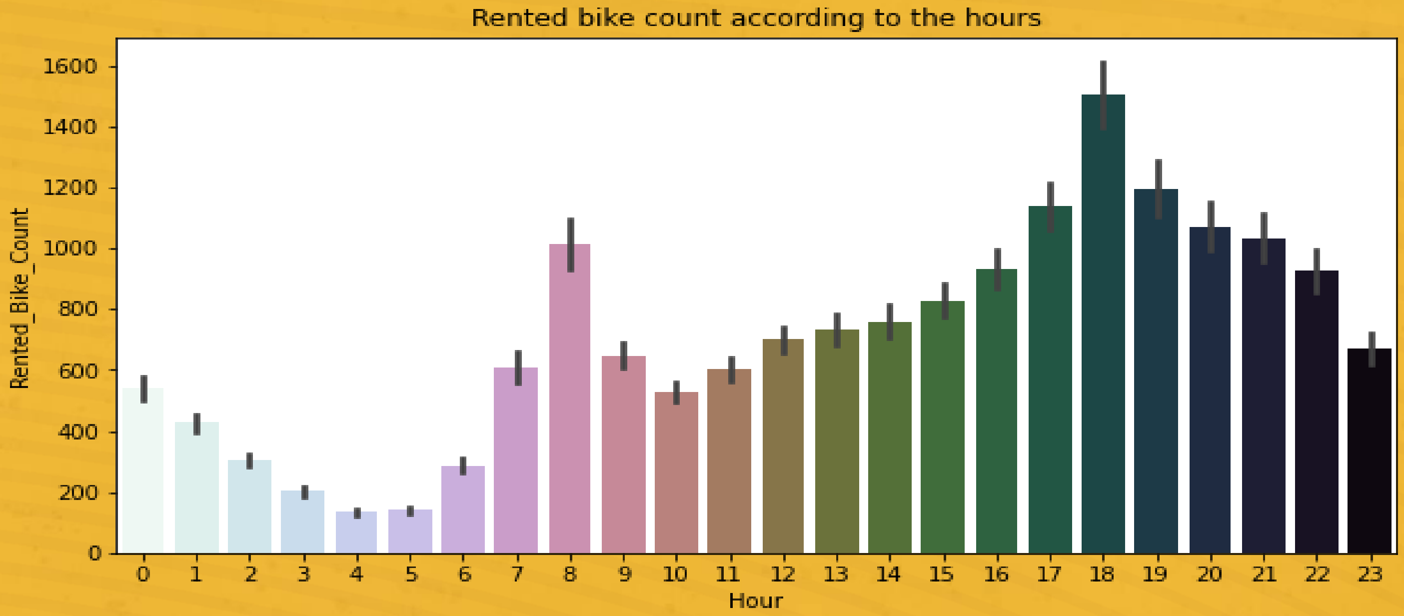


Weekend and Week days



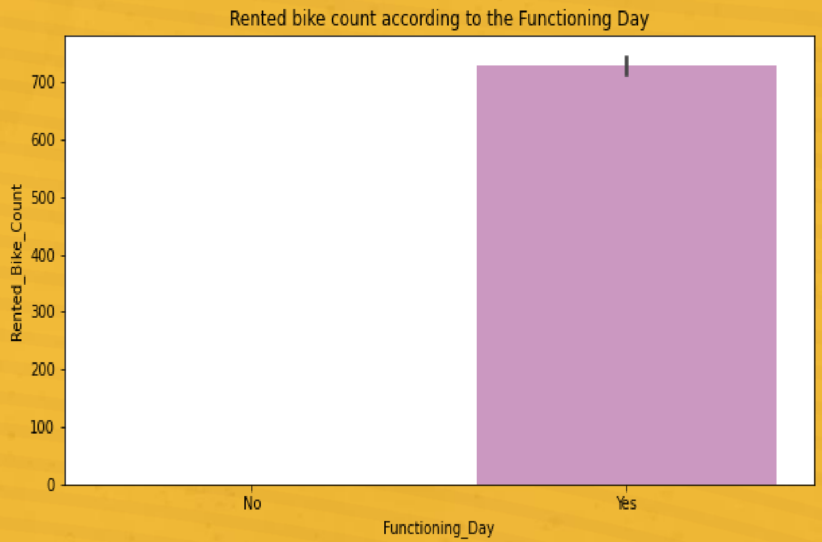
In the point plot above, the weekdays are depicted as blue and the weekends as sky blue, with the weekdays' rental bike count rising during business hours (7 to 9 and 17 to 19)

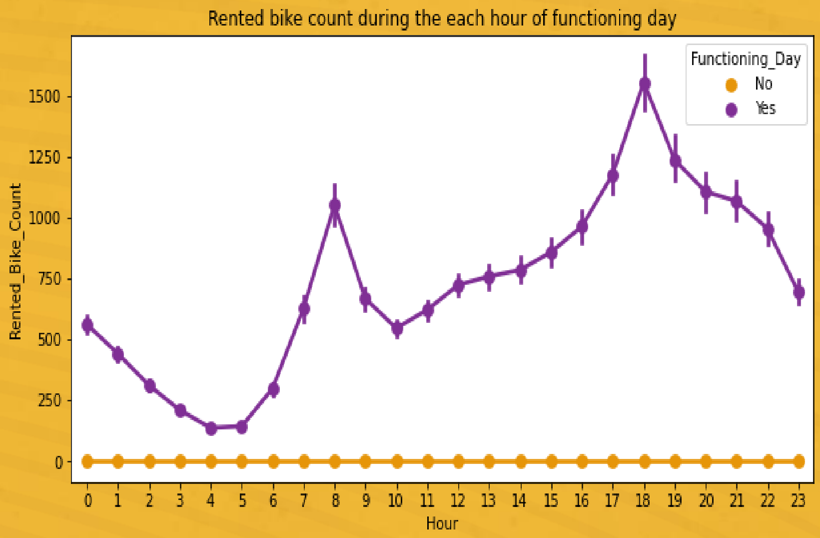
Hours



The above chart show the peak hours of bike renting, morning 7 to 8 and evening 16 to 19 are high peaks hours

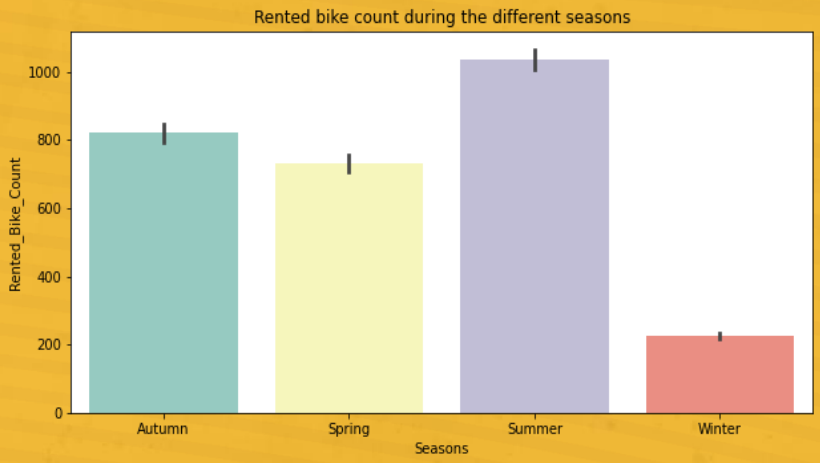
Functioning Day

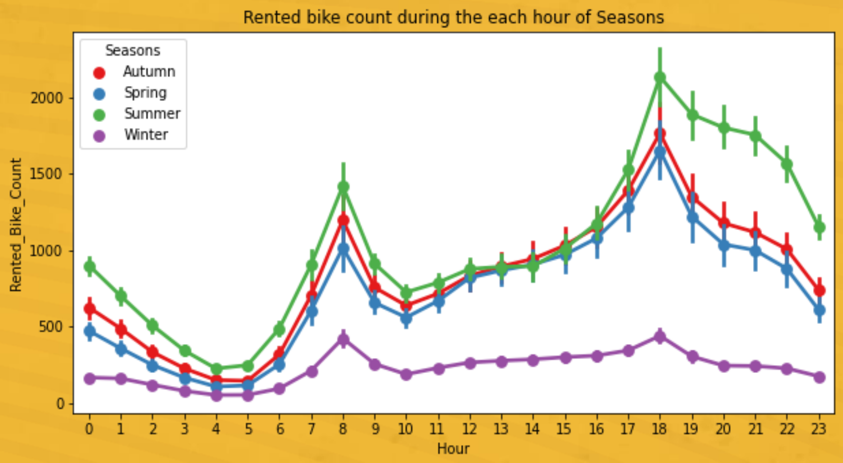




As per the above graph analysis, During none functioning day bikes are seems not rented.

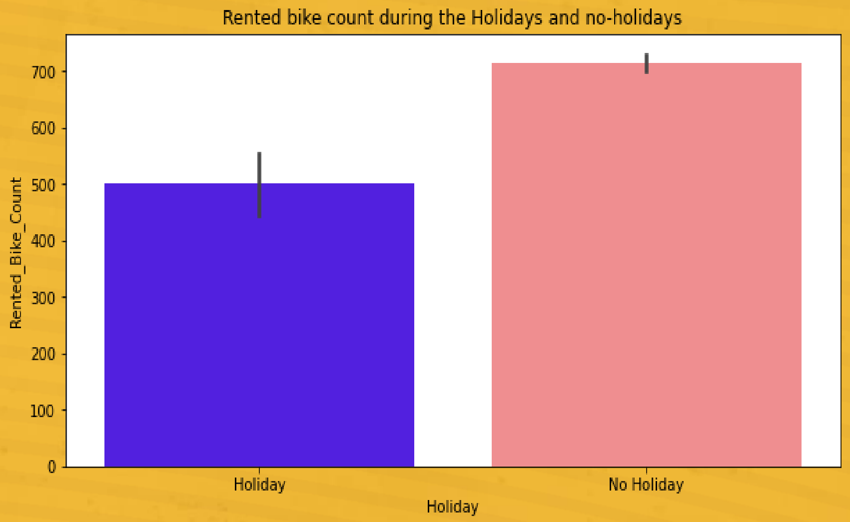
Seasons

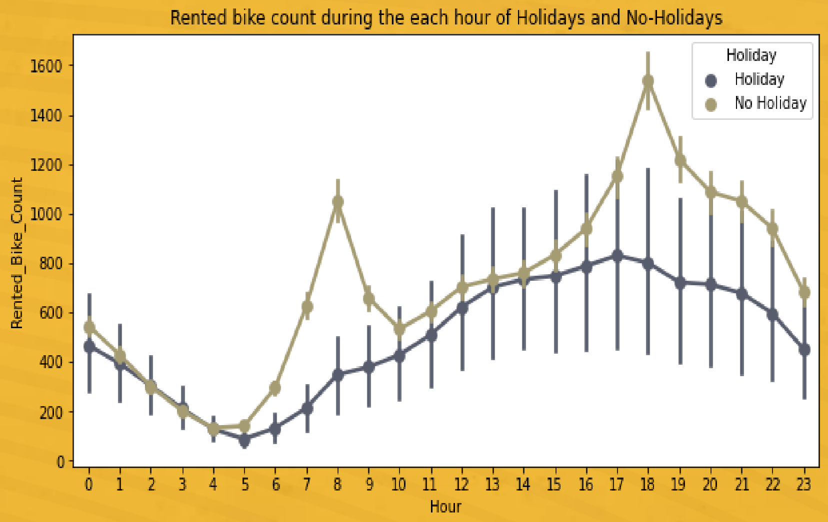




In comparison to all other seasons, the performance of the leased bike count during the winter is relatively low, and the bike counts are much raised during the summer.

Holiday

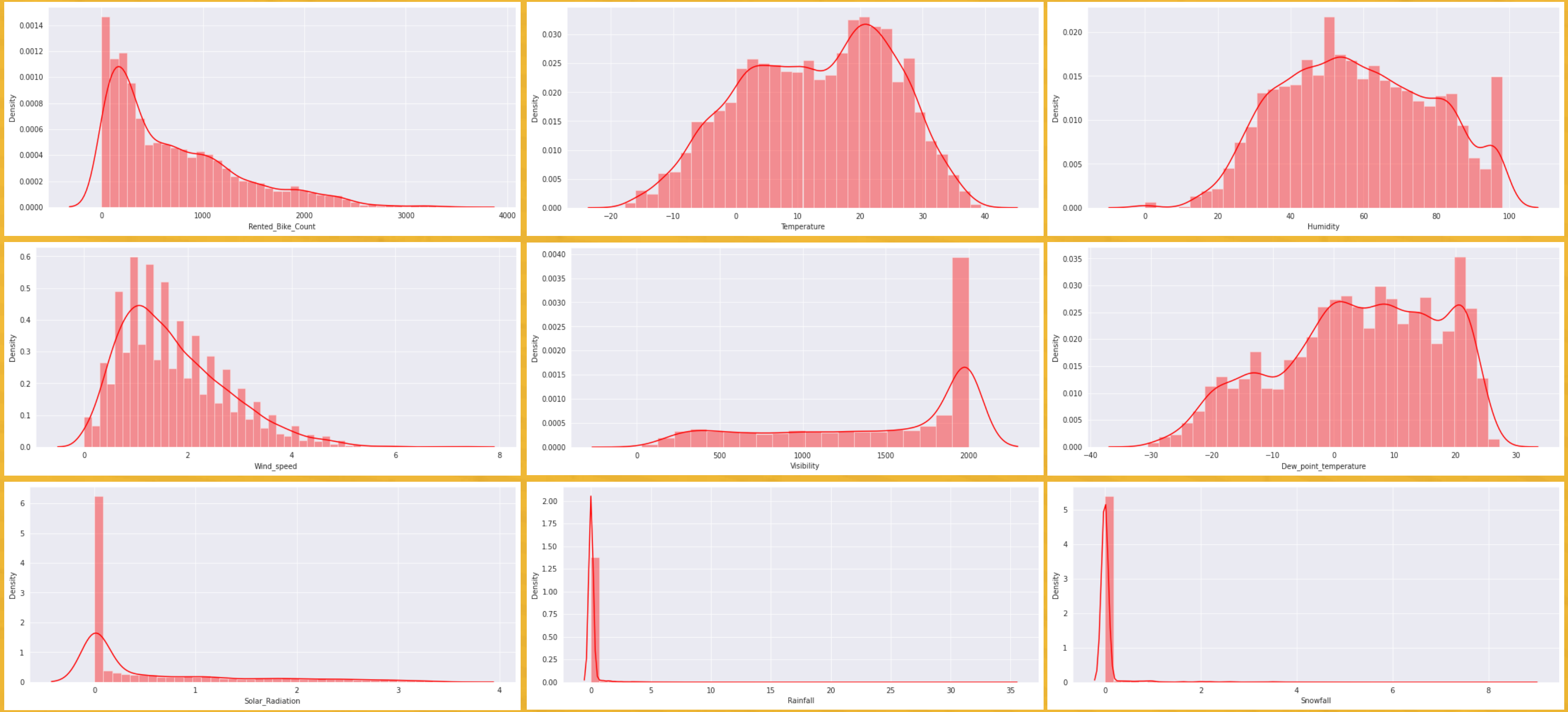




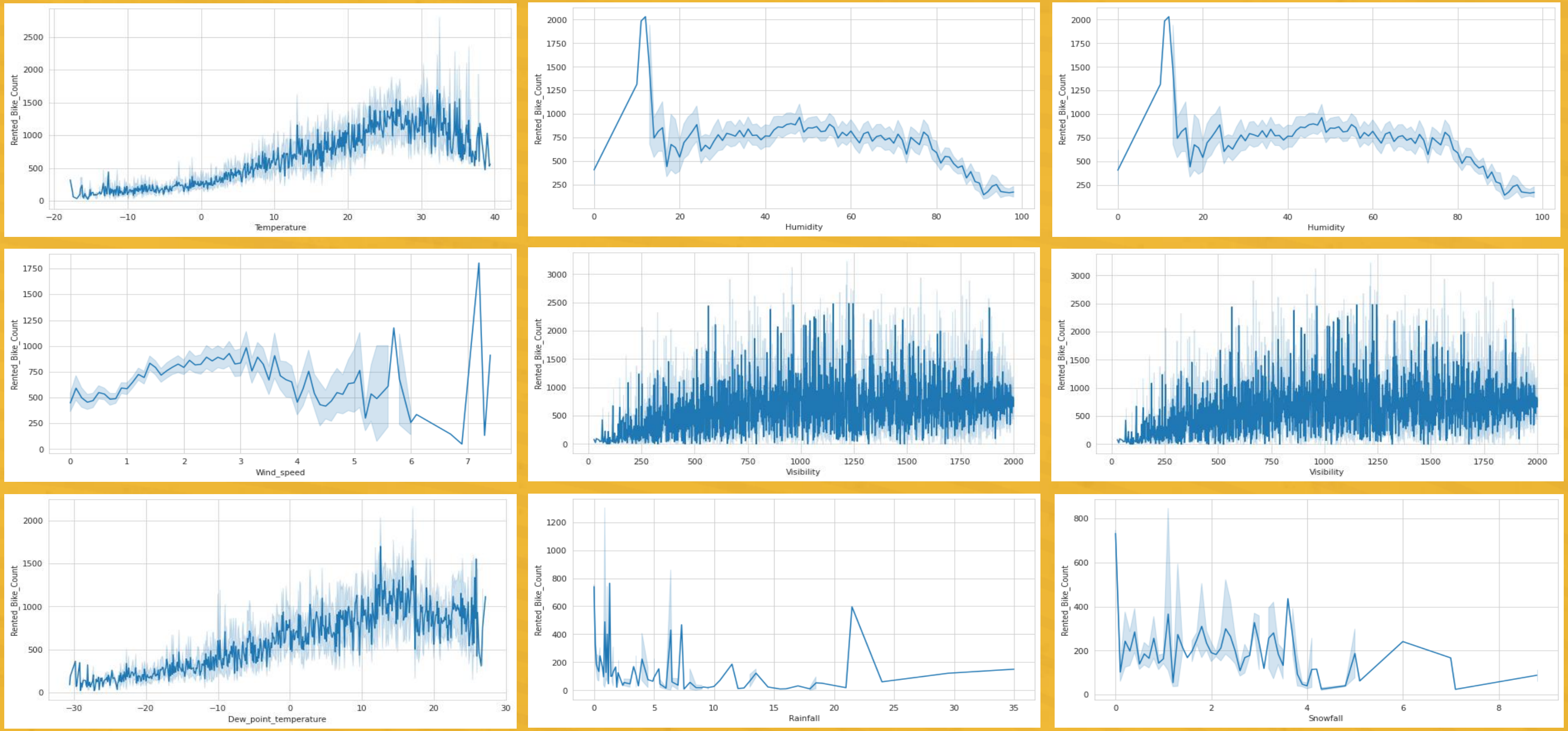
When compared to holidays, bike counts are significantly higher on non-holiday days; this may be because of office hours.

Numerical Data Analysis

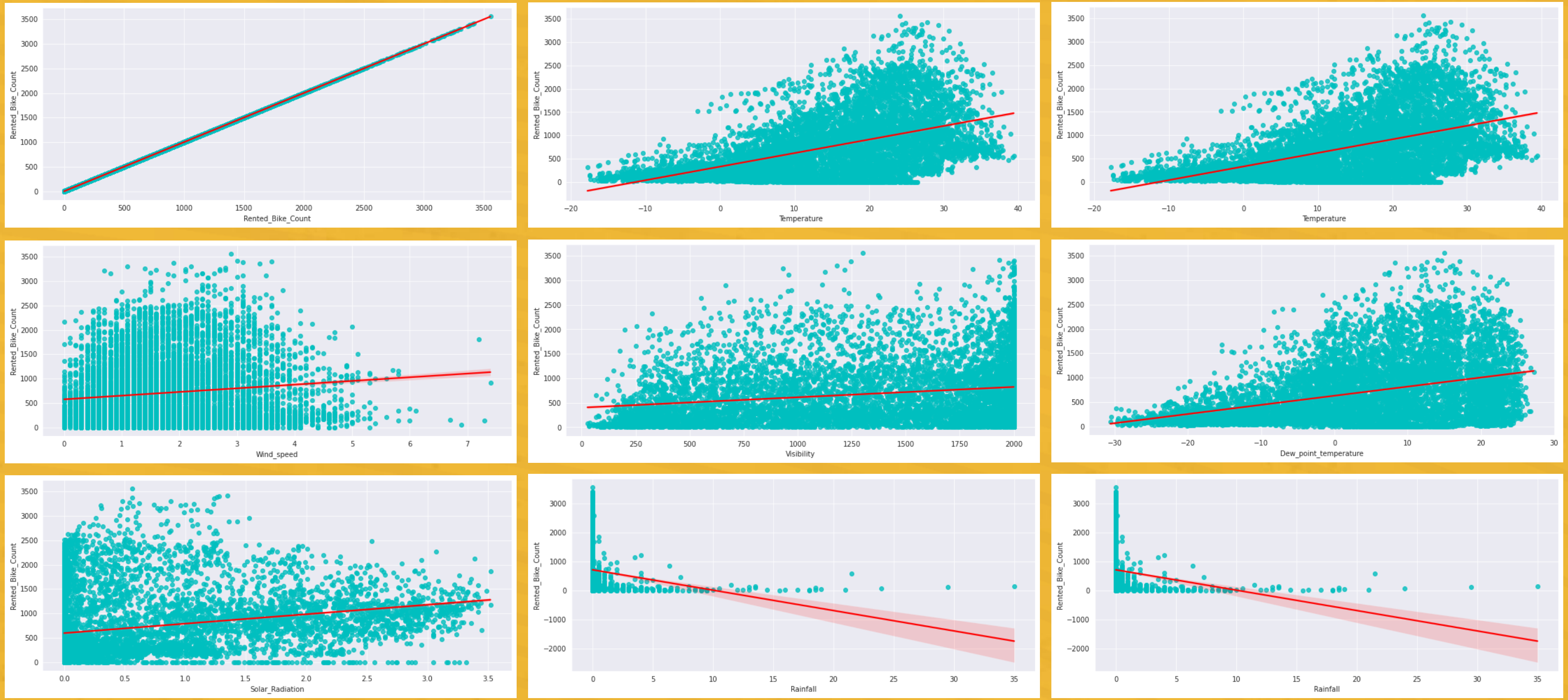
1. Displot



2. Line plot


Rented_Bike_Count
Temperature
Humidity
Wind_speed
Visibility
Dew_point_temperature
Rainfall
Snowfall

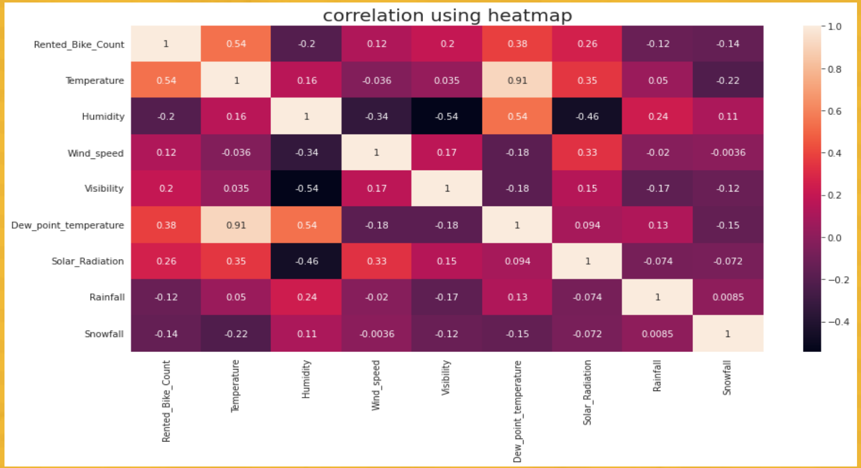
3. Regplot



Rented_Bike_Count
Rented_Bike_Count
Temperature
Temperature
Wind_speed
Visibility
Dew_point_temperature
Solar_Radiation
Rainfall
Rainfall

Data Preprocessing:

1. Feature selection




```
Bike_df.corr()
```

	Rented_Bike_Count	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Snowfall
Rented_Bike_Count	1.000000	0.538558	-0.199780	0.121108	0.199280	0.379788	0.261837	-0.123074	-0.141804
Temperature	0.538558	1.000000	0.159371	-0.036252	0.034794	0.912798	0.353505	0.050282	-0.218405
Humidity	-0.199780	0.159371	1.000000	-0.336683	-0.543090	0.536894	-0.461919	0.236397	0.108183
Wind_speed	0.121108	-0.036252	-0.336683	1.000000	0.171507	-0.176486	0.332274	-0.019674	-0.003554
Visibility	0.199280	0.034794	-0.543090	0.171507	1.000000	-0.176630	0.149738	-0.167629	-0.121695
Dew_point_temperature	0.379788	0.912798	0.536894	-0.176486	-0.176630	1.000000	0.094381	0.125597	-0.150887
Solar_Radiation	0.261837	0.353505	-0.461919	0.332274	0.149738	0.094381	1.000000	-0.074290	-0.072301
Rainfall	-0.123074	0.050282	0.236397	-0.019674	-0.167629	0.125597	-0.074290	1.000000	0.008500
Snowfall	-0.141804	-0.218405	0.108183	-0.003554	-0.121695	-0.150887	-0.072301	0.008500	1.000000

The correlation between temperature and dew point temperature is 0.91 its very high to avoid error in model I have dropped dew_point_temperature column, and all other columns are taken for Machine learning Model Creation.

2. One hot encoding

```
[ ] # Create dummy variables

    categorical_features = Bike_df.select_dtypes(include=['category'])

    bike = Bike_df
    bike = bike.drop(categorical_features, axis=1)
    data = pd.get_dummies(categorical_features,drop_first=True)
    data = pd.concat([bike, data], axis=1)
    data.head()
```

	Rented_Bike_Count	Temperature	Humidity	Wind_speed	Visibility	Solar_Radiation	Rainfall	Snowfall	Hour_1	Hour_2	...	month_4	month_5	month_6	month_7
0	254	-5.2	37	2.2	2000	0.0	0.0	0.0	0	0	...	0	0	0	0
1	204	-5.5	38	0.8	2000	0.0	0.0	0.0	1	0	...	0	0	0	0
2	173	-6.0	39	1.0	2000	0.0	0.0	0.0	0	1	...	0	0	0	0
3	107	-6.2	40	0.9	2000	0.0	0.0	0.0	0	0	...	0	0	0	0
4	78	-6.0	36	2.3	2000	0.0	0.0	0.0	0	0	...	0	0	0	0

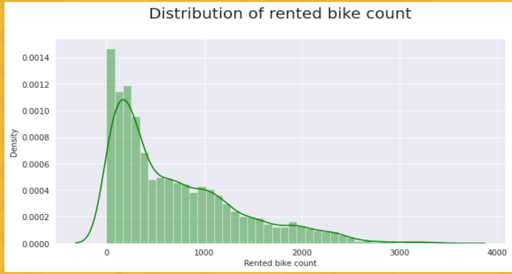
5 rows × 48 columns

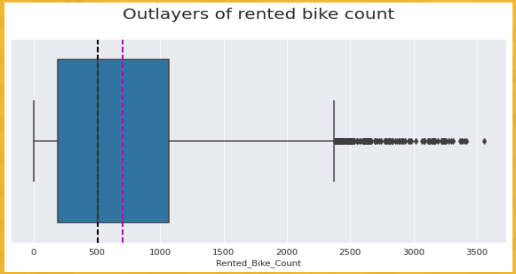
to perform the regression model the categorical variable values are converted in numerical using dummies method

3. Normalizing Depending Variable data's

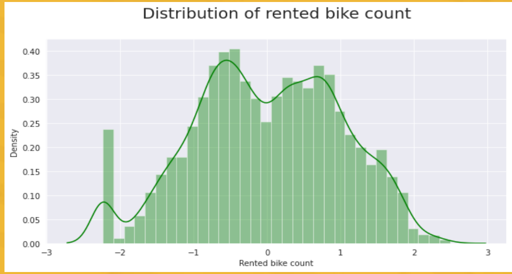
In this normalization process initially rented bike count data's slightly right skewed and out layers are found, for regression model data should be normally distribution is very required so, I converted the rented bike count data's as normal distribution using power transformer method and out layers are removed by square root method.

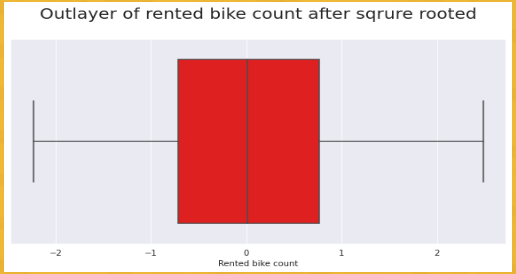
Before





After

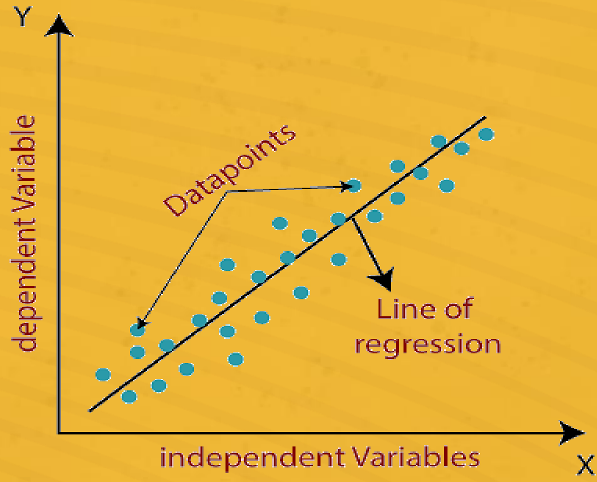




Regression models:

What is Regression?

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.



Linear Regression:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.

Random forest Regression:

Random forest is a supervised learning algorithm that uses an ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. The trees in random forests run in parallel, meaning is no interaction between these trees while building the trees.

Gradient Boosting Regression:

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels.

Evaluation criteria

Mean Squared Error

The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.

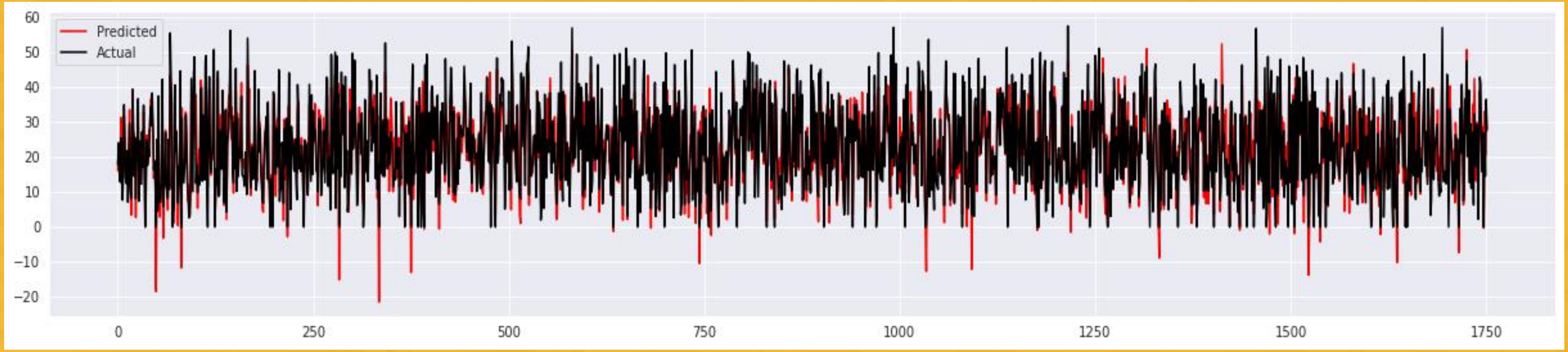
Root mean square error

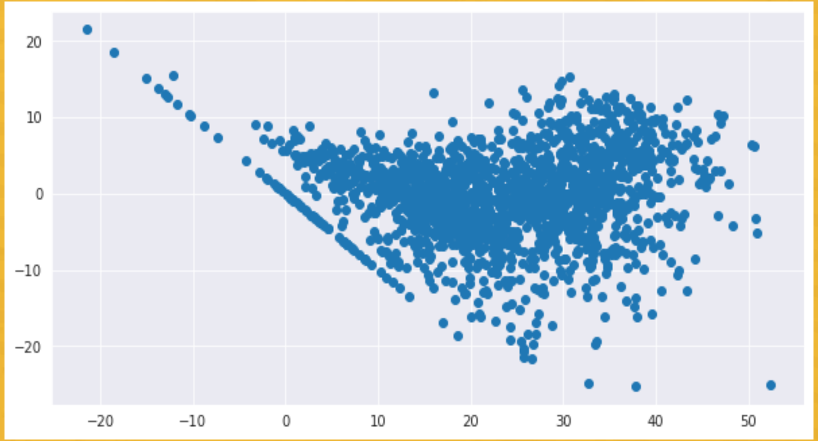
RSME (Root mean square error) calculates the transformation between values predicted by a model and actual values. In other words, it is one such error in the technique of measuring the precision and error rate of any machine learning algorithm of a regression problem

Mean Absolute Error

In the context of machine learning, absolute error refers to the magnitude of difference between the prediction of an observation and the true value of that observation. MAE takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group. MAE can also be referred as L1 loss function.

Linear Regression

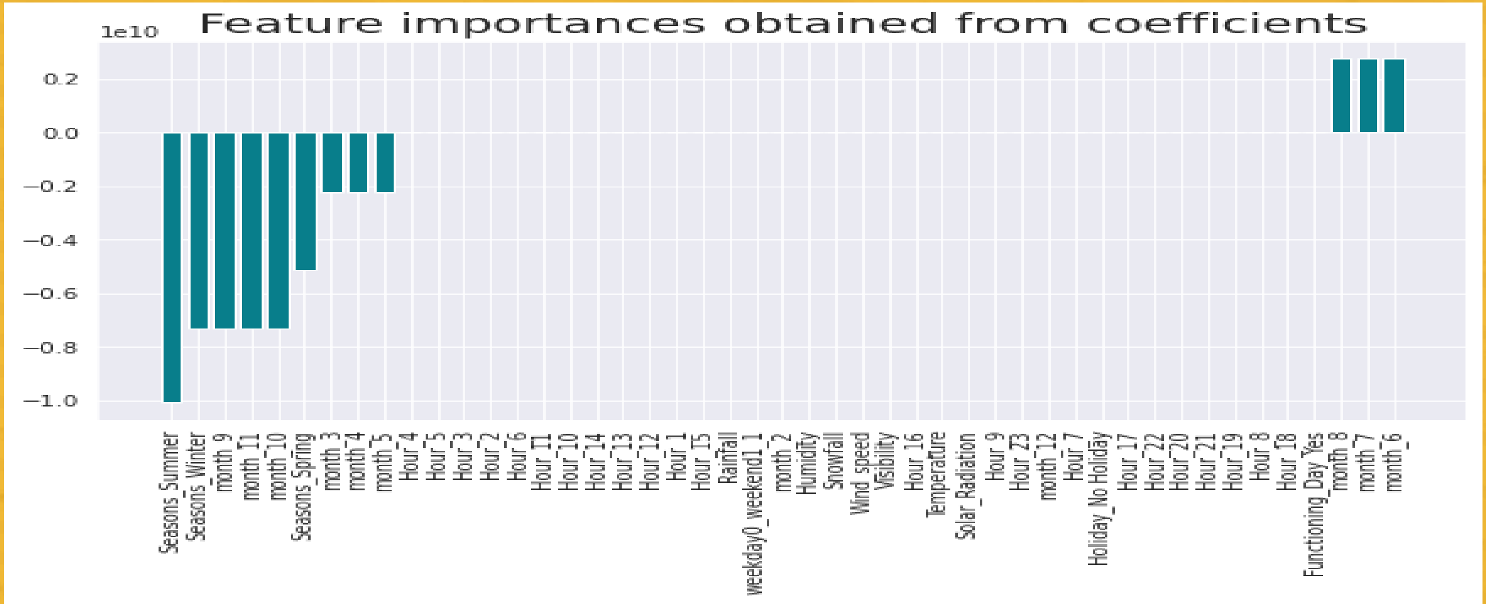




Linear_train_MSE : 34.793980690219946
Linear_train_RMSE : 5.898642275152812
Linear_train_MAE : 4.459079727099164
Linear_train_r2 : 0.774552733127227
Linear_train_adjusted_r2 : 0.7683344106254546

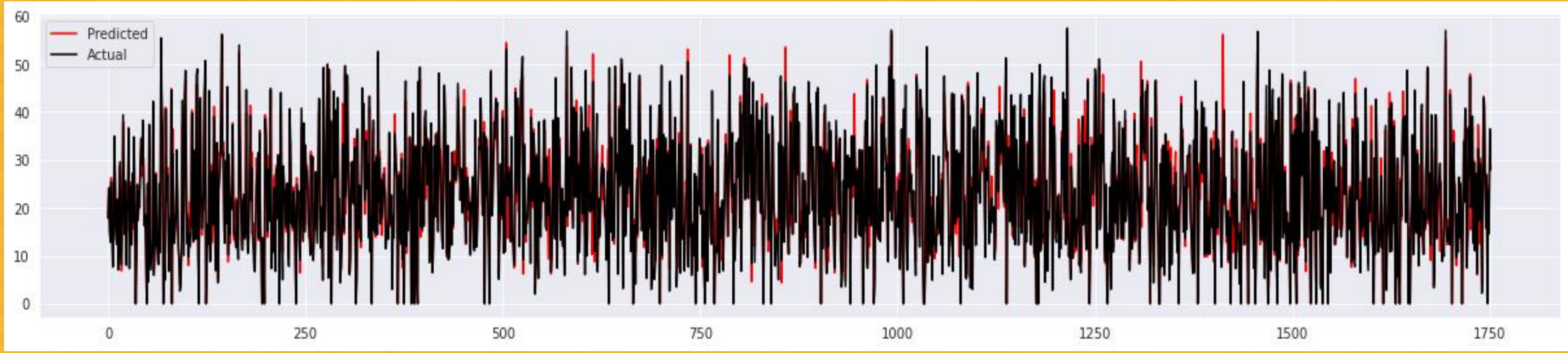
Linear_test_MSE : 33.894124066686764
Linear_test_RMSE : 5.821866029606553
Linear_test_MAE : 4.442357194447212
Linear_test_r2 : 0.78478043307297
Linear_test_adjusted_r2 : 0.7788442126236915

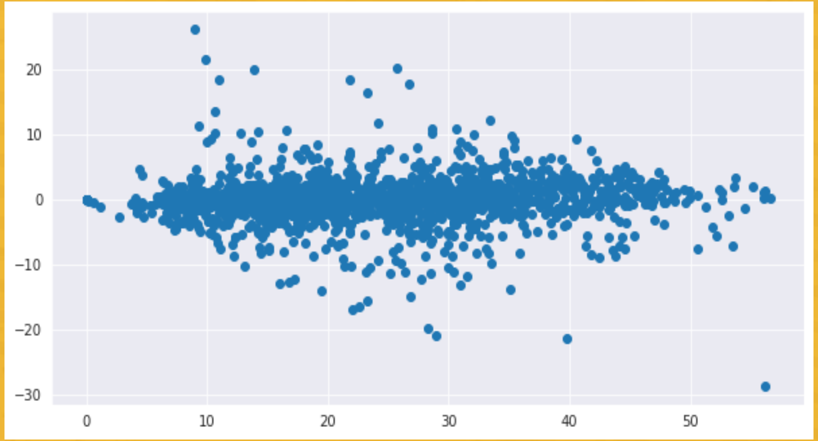
Importance Features



As per linear regression model the importance features are got as Summer, winter, spring seasons and 6 to 9 months, these are very important for bike rent.

Random forest Regression

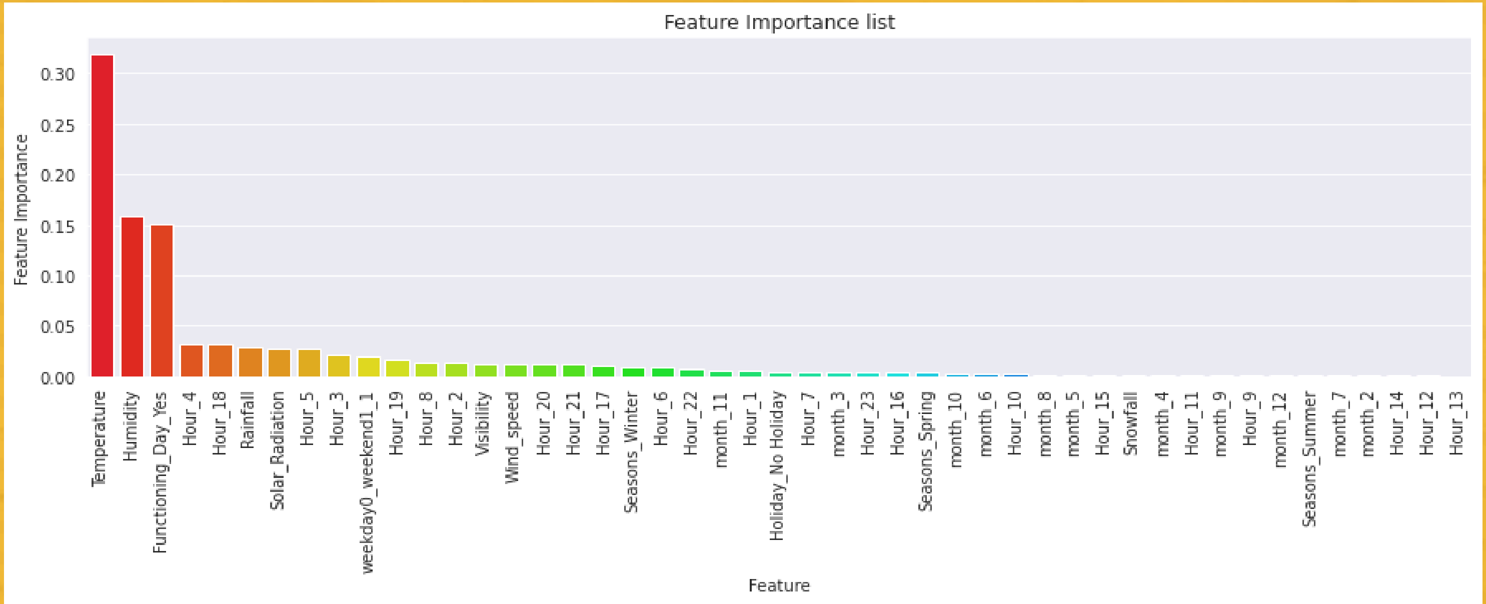




Random_forest_train_MSE : 1.564292460843556
Random_forest_train_RMSE : 1.2507167788286666
Random_forest_train_MAE : 0.7936095276654537
Random_forest_train_r2 : 0.9898641818817243
Random_forest_train_adjusted_r2 : 0.989584614128462

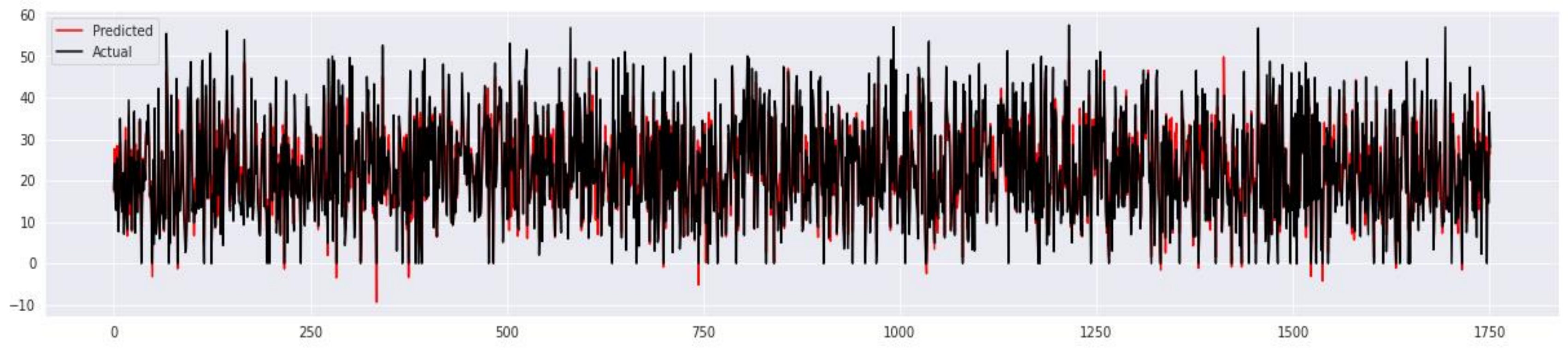
Random_forest_test_MSE : 12.86120728783719
Random_forest_test_RMSE : 3.586252541001148
Random_forest_test_MAE : 2.201294788433046
Random_forest_test_r2 : 0.91833441521601
Random_forest_test_adjusted_r2 : 0.916081902020677

Importance Features



The above Chart shows, Random Forest regression model results as, highly importance features are Temperature, Humidity, Functioning day.

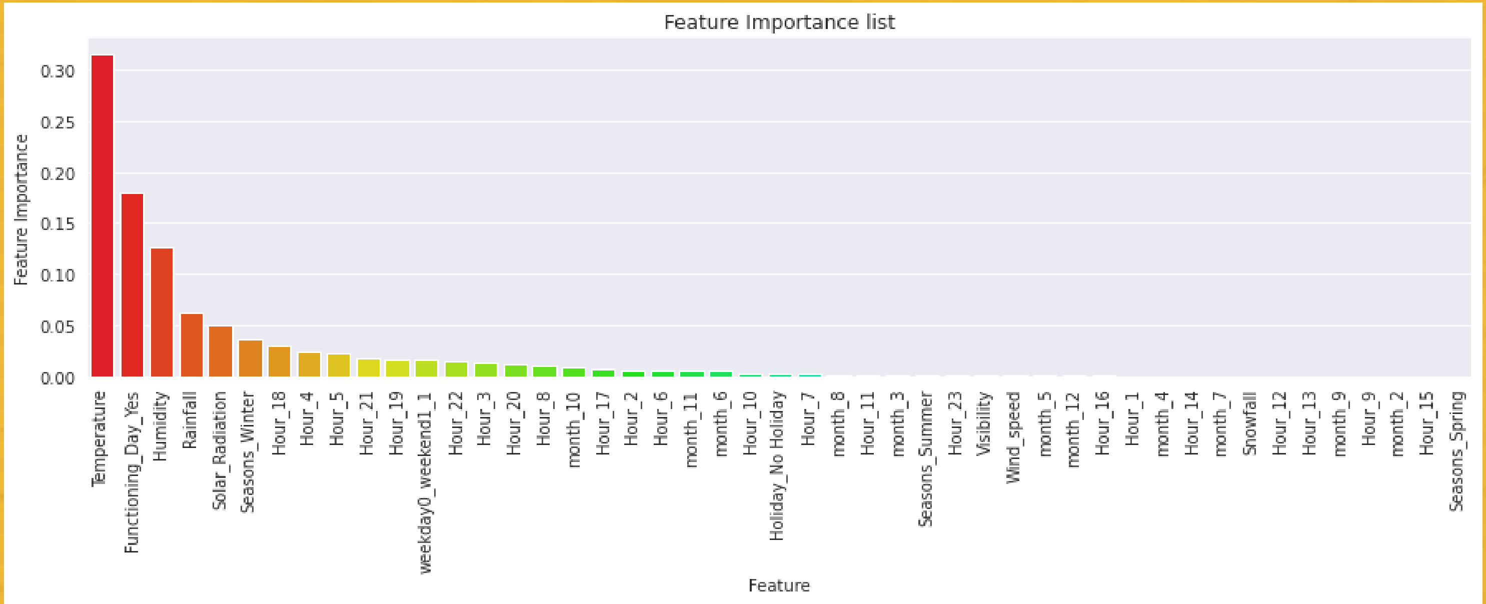
Gradient Boosting Regression



Gradient_Boosting_train_MSE : 18.730814047267177
Gradient_Boosting_train_RMSE : 4.3279110489088355
Gradient_Boosting_train_MAE : 3.28122151688366
Gradient_Boosting_train_r2 : 0.8786338685747016
Gradient_Boosting_train_adjusted_r2 : 0.8752863285647315

Gradient_Boosting_test_MSE : 21.48922367586383
Gradient_Boosting_test_RMSE : 4.635647061183997
Gradient_Boosting_test_MAE : 3.487368803906541
Gradient_Boosting_test_r2 : 0.863548578390225
Gradient_Boosting_test_adjusted_r2 : 0.8597849534984061

Importance Features



Based on above Chart, Gradient Boosting regression model results as, highly importance features are Temperature, Functioning day, Humidity, Rainfall and Solar radiation these are highly decides Bike renting.

Conclusions:

		Model Name	MSE	RMSE	MAE	R2 score	Adjusted R2
Training Data	0	Linear regression train	34.793981	5.898642	4.459080	0.774553	0.768334
	1	Gradient_Boosting regression train	18.730814	4.327911	3.281222	0.878634	0.875286
	2	Random_forest regression train	1.564292	1.250717	0.793610	0.989864	0.989585
Test Data	0	Linear regression test	33.894124	5.821866	4.442357	0.784780	0.778844
	1	Gradient Boosting regression test	21.489224	4.635647	3.487369	0.863549	0.859785
	2	Random forest regression test	12.861207	3.586253	2.201295	0.918334	0.916082

Comparing to three Models the Random Forest algorithm has highest R2 Score 98 % and 91 % respectively for Train and Test data's and Gradient Boosting algorithm also had good R2 Score 87 % and 86 % respectively for Train and Test data's then, there is no overfitting found in all these models. The Feature importance of both models are slightly different from each others. So better result We can deploy this models.

- The demand for rented bikes is higher from months 4 to 10 compared to other months, these months fall during the summertime.
- Use of rental bike count rising during business hours (7 to 9 and 17 to 19), these are consider as peak hours.
- During none functioning day bikes are seems not rented.
- Performance of the rented bike count during the winter is relatively low, and the bike counts are much raised during the summer.
- When compared to holidays, bike counts are significantly higher on non-holiday days; this may be because of office hours

THANKS!