Capstone Project

Bike Sharing Demand Prediction

Submitted by:

Vikas Chaudhary

Flow of the Presentation

- Introduction
- Data Preparation
- EDA
- Algorithms Implementation
- Challenges
- Conclusion

Introduction

The bike rental business is growing rapidly for the last few years in major cities across the globe. Below are several important points about this business:

- It reduces the traffic problem which is prevalent in almost all the cities.
- People with no personal transport can opt for the service for their movement.
- It may also help in mitigating pollution since most of the vehicles used in this business are electric or will be replaced by electric.
- This business involves large investments both in physical and non-physical infrastructure.

Objective: It is essential to develop systems that help in flawless operations in this business. One of the sections of the entire system is to know the approx. number of bikes that must be available. On the basis of the given dataset that contains previous records, we can develop a machine learning (ML) model to predict the number of required bikes at a given point of time for better customer experience and overall revenue or profitability.

Methodology: Machine Learning (ML) Linear Regression (Supervised ML Model)

Database Summary:

- It is a public bike rent dataset of Seoul the capital city of South Korea.
- It has 8740 rows and 14 columns
- The values in the dataset are of 3 categories i.e. Numerical, Categorical and Datetime.
- The dataset has broadly four pieces of information in different columns, these are: column about the number of bikes, columns having information about date, time, month etc, columns about days of operations and columns about the climate or weather.

Data Preparation

AI

The database has the following features:

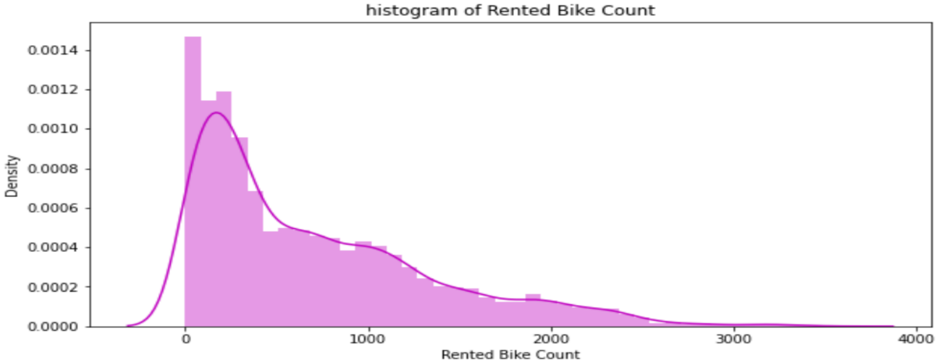
- Date - Day/Month/Year
- Rented Bike Count - Number of Bikes rented every hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

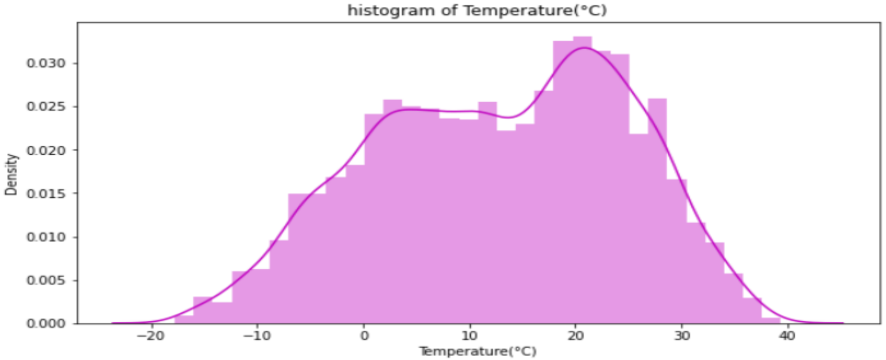
Overview of Dataset

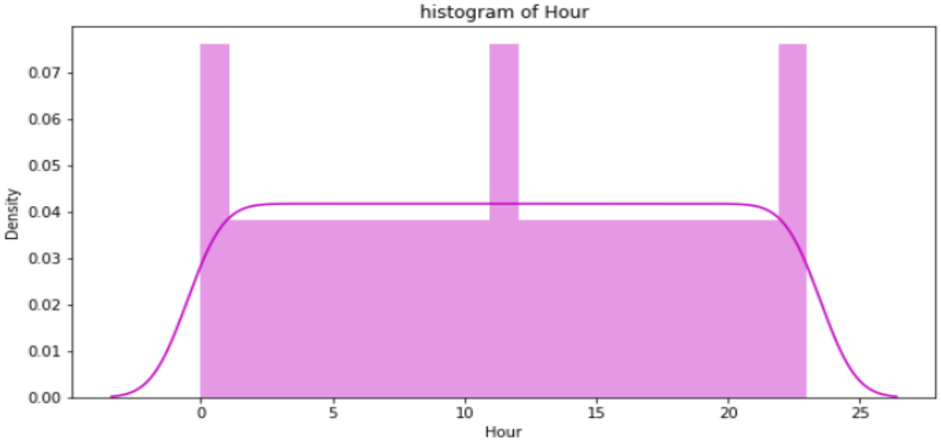
- It has no 'NaN' or 'Null' values
- It has no duplicate values
- It contains the hourly record of bike rentals from December 1, 2017, to November 30, 2018, i.e. of an entire year.
- 'Rented Bike Count' is a dependent feature.
- Day (number and name), Month (number and name) and Year were extracted from the 'Date' column.
- After creating new features from the 'Date' feature the entire dataset contains only 'numerical' and 'categorical' datatype.
- 'wind_speed', 'solar_radition', 'rainfall', 'snowfall' and 'Rented Bike Count' have outliers.

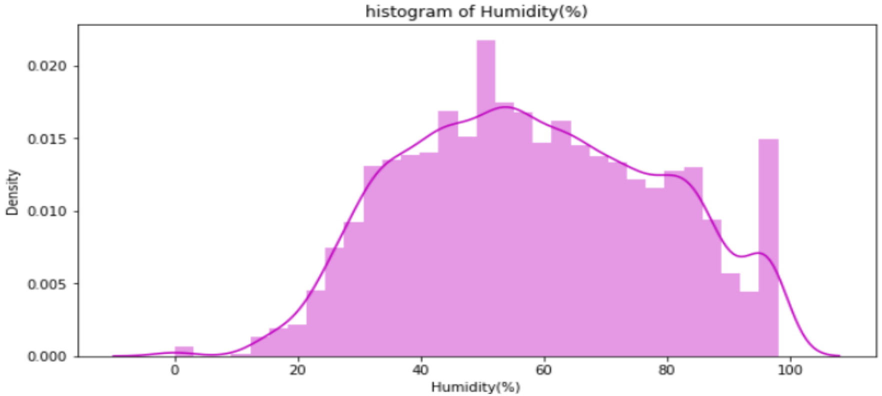
EDA

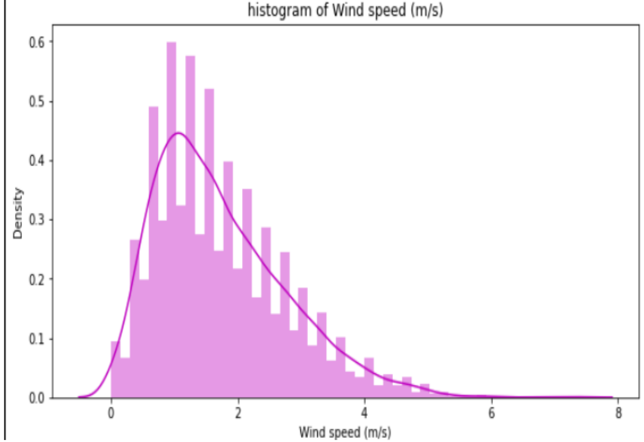
Plots of the Distribution of Numerical Features

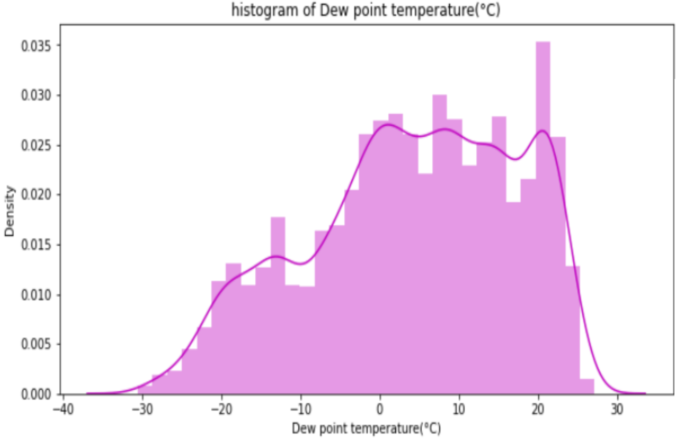


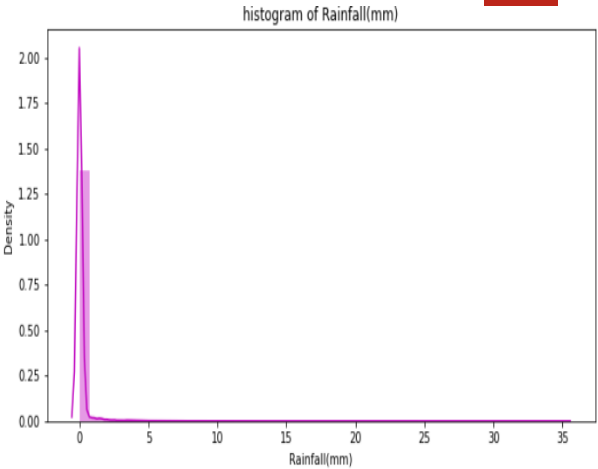


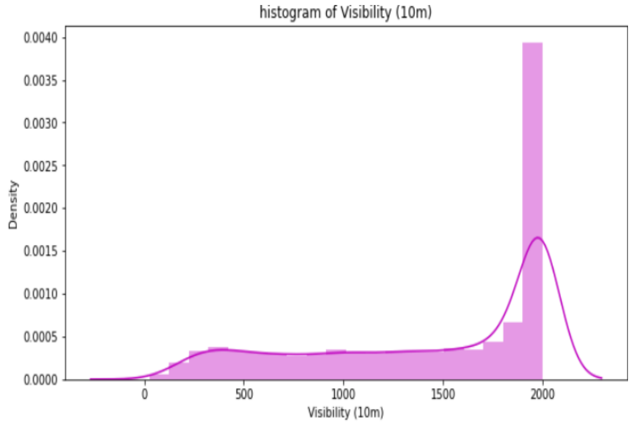


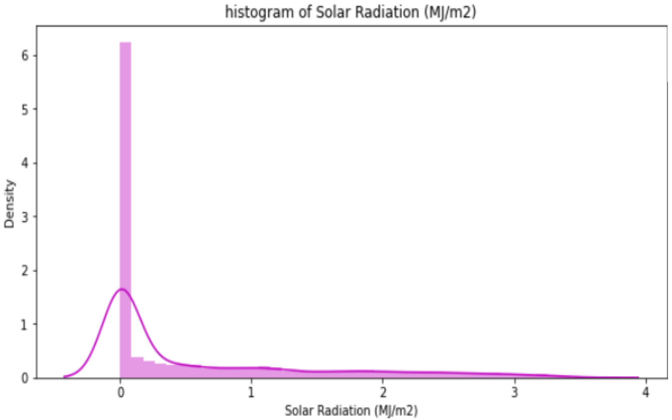


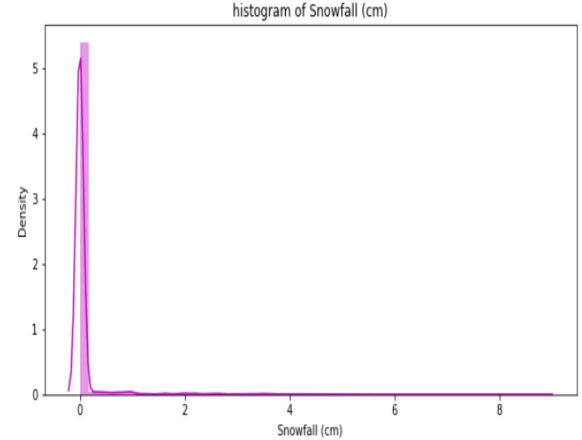






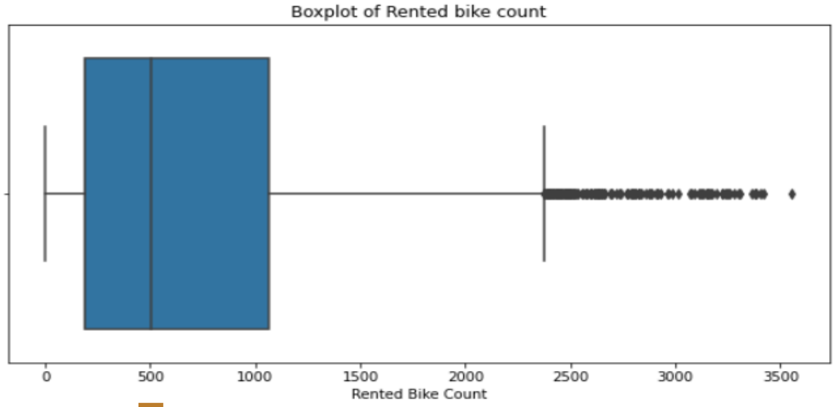


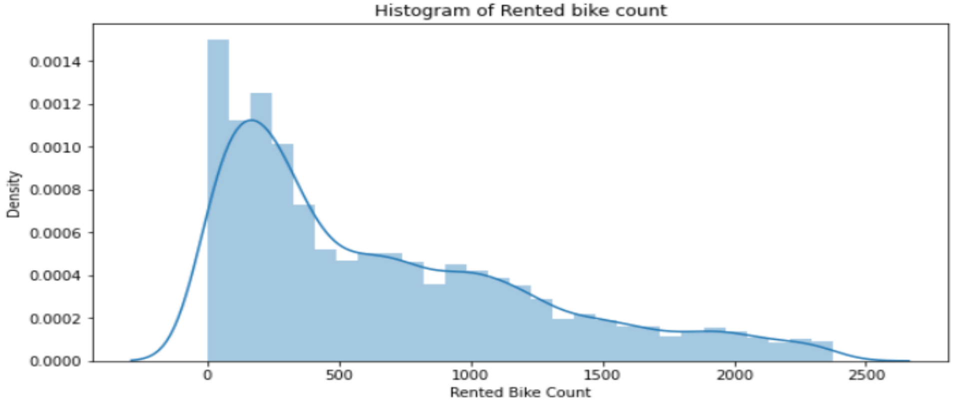




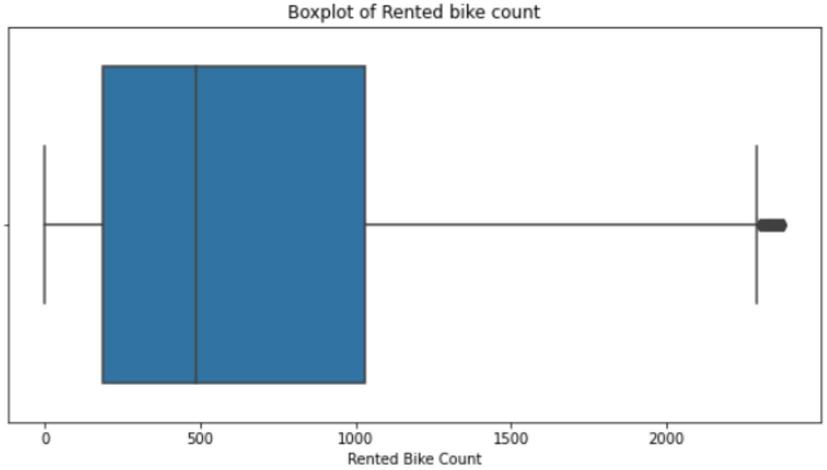
From the above histograms we can conclude that most of the histograms are skewed either on the right side or left side apart from ‘Dew point temperature’, ‘temperature’, ‘hour’ and ‘humidity’ these are normally distributed.

Outlier – Dependent Feature



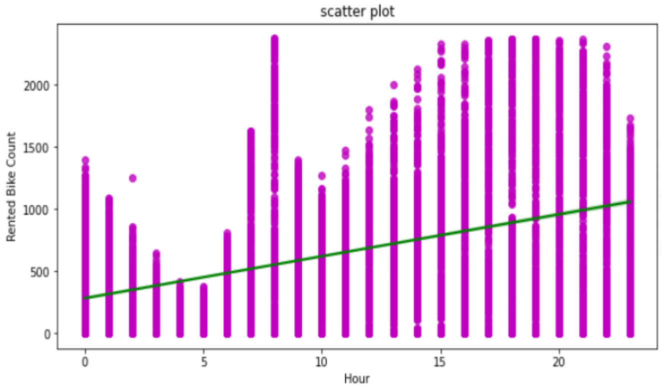


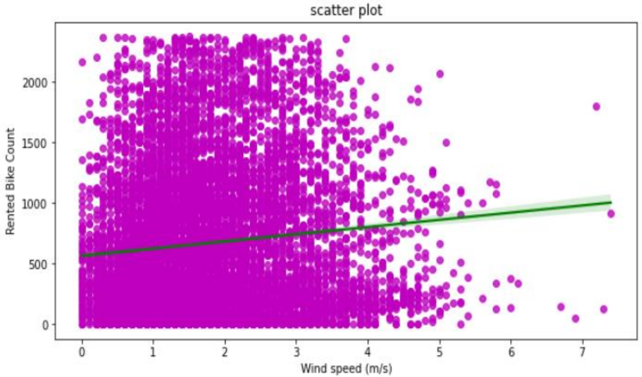
With outliers

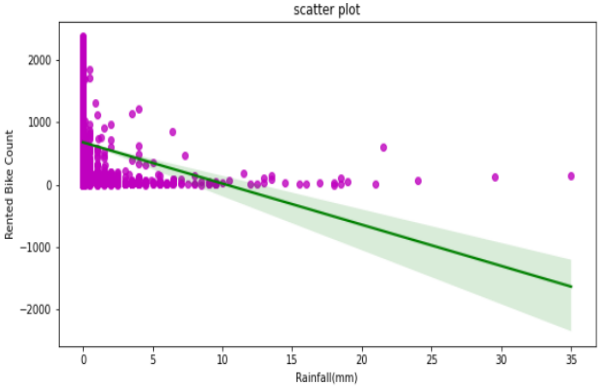


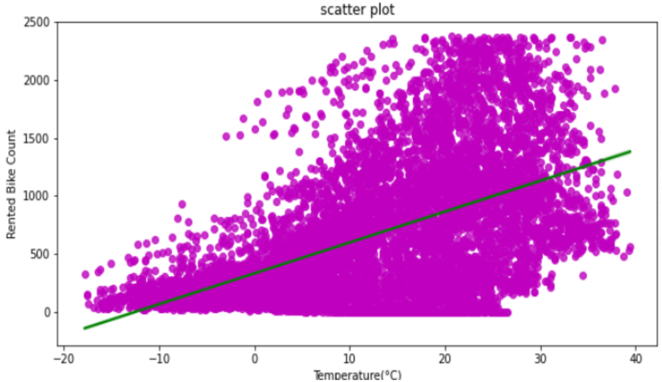
After removing some outliers

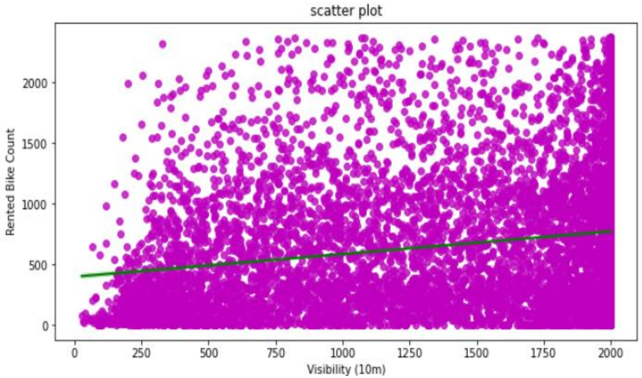
Scatter Plot with Linear Regression

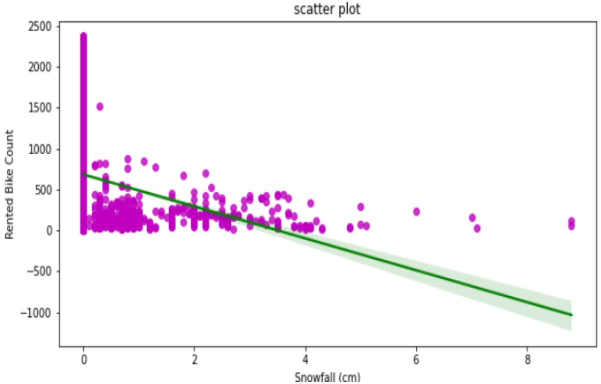


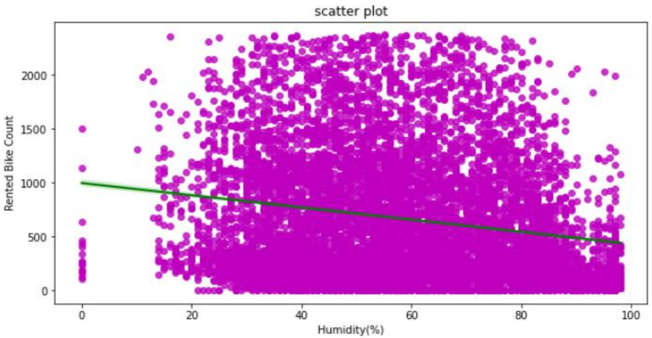


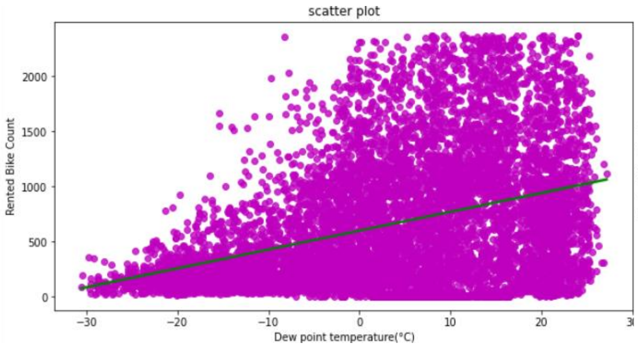


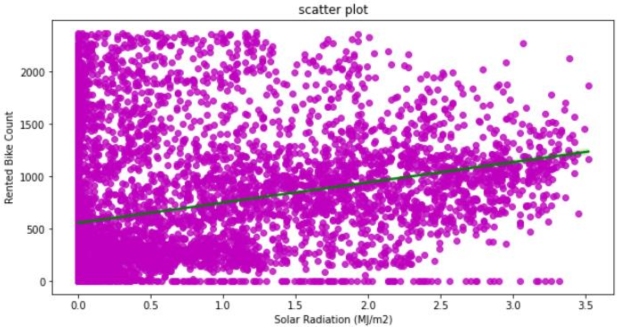






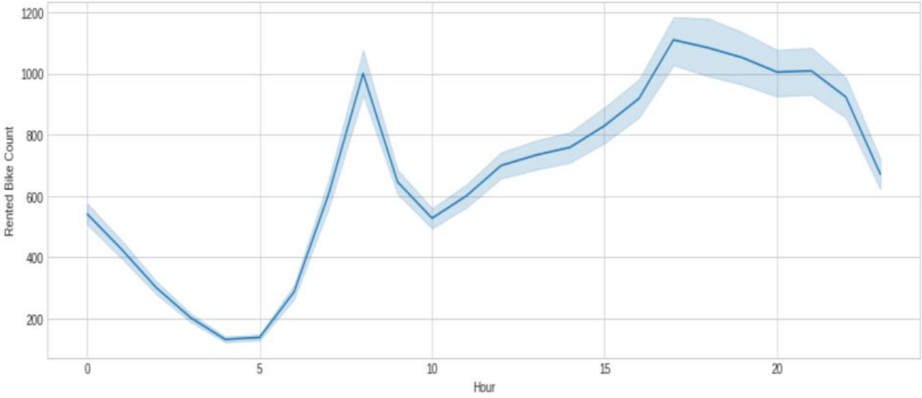


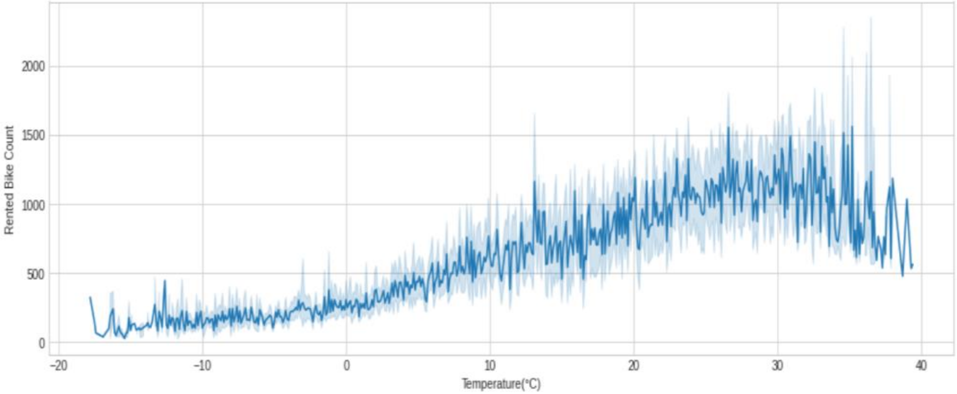


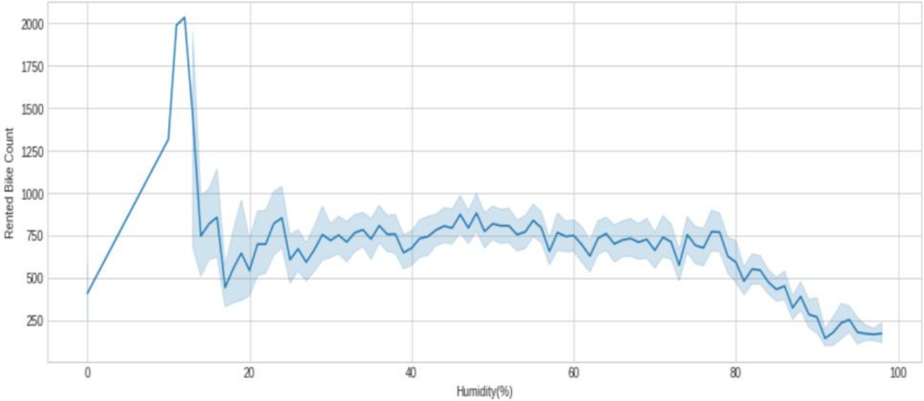


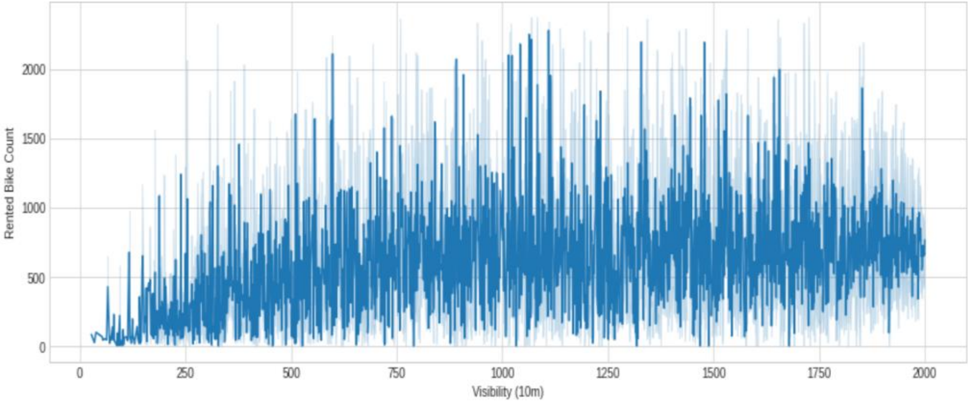
i. Hour, Wind speed, Temperature, Visibility, Dew Point Temperature and Solar Radiation have positive co-relation with the number of Rented Bikes.
ii. Rainfall, Snowfall and Humidity have negative co-relation with the number of Rented Bikes.

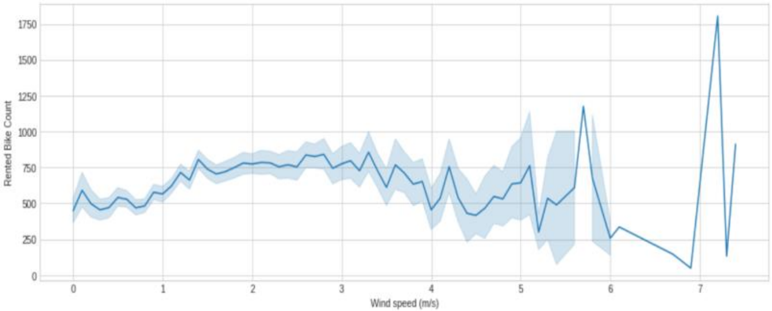
Graph to show the distribution of numerical features

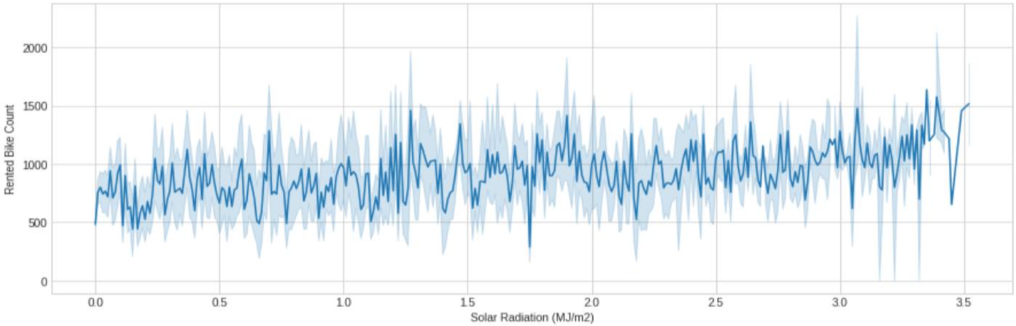


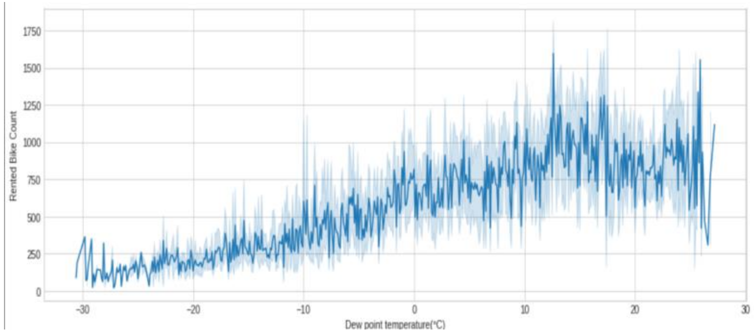


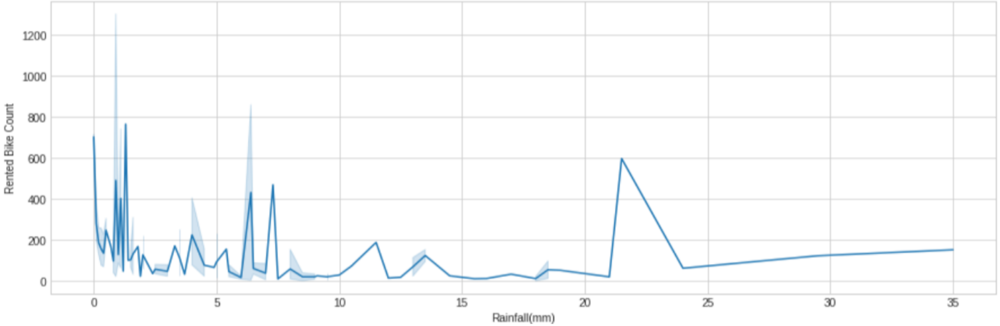


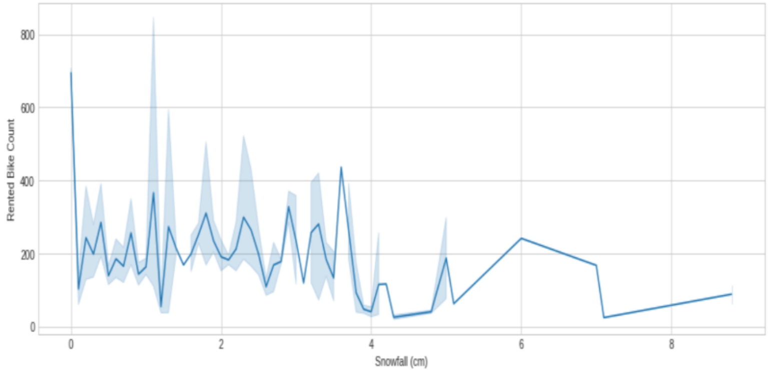






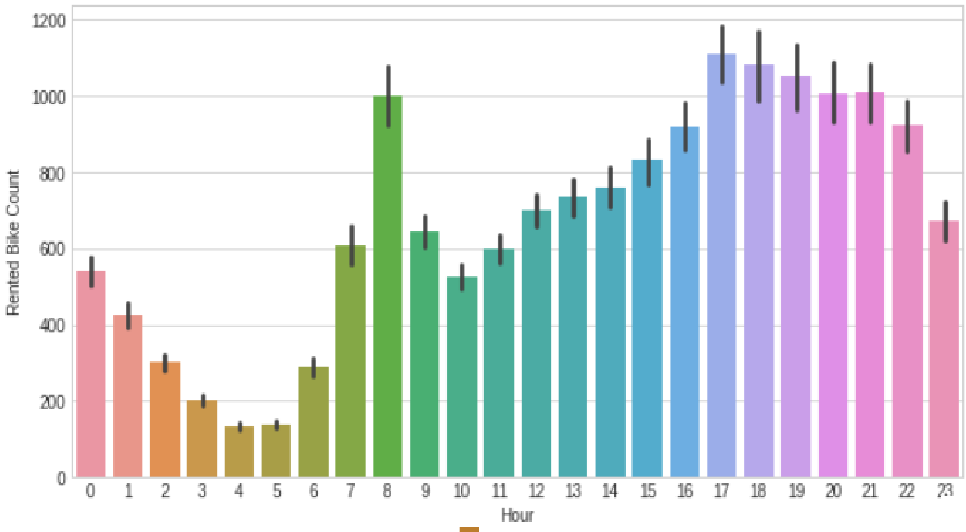




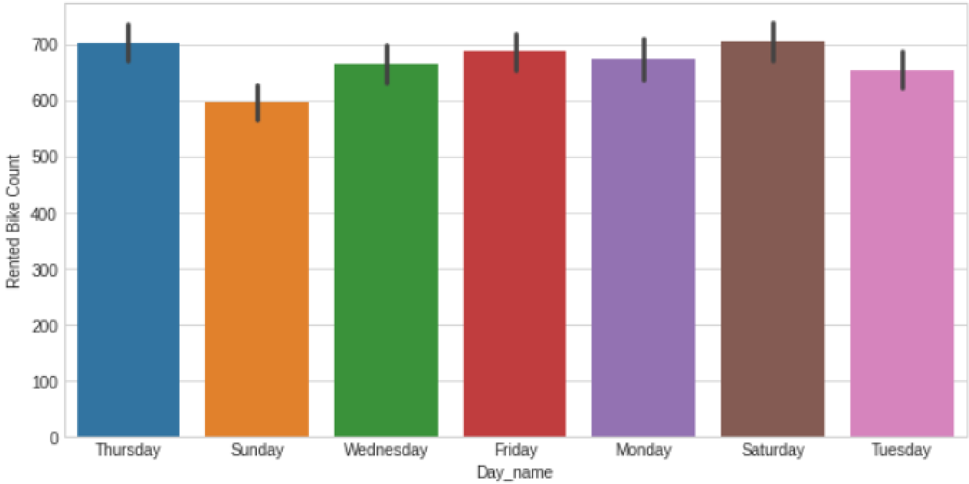


The graphs related climate or weather are showing some uncommon events which appears to be rare.

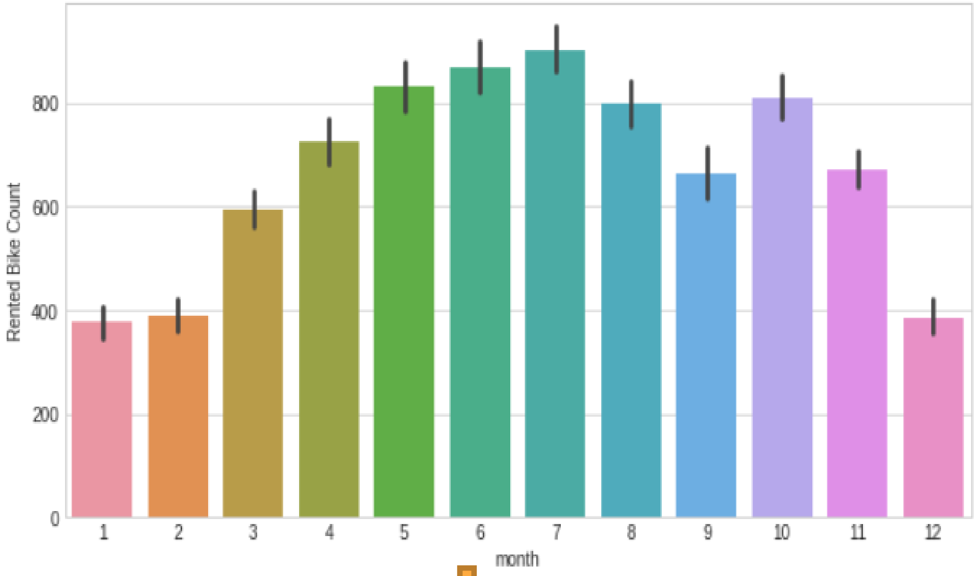
Rented Bikes on Different Occasions



The average number of rented bikes on different days not showing any significant variation, it is nearly the same except for the Sunday.

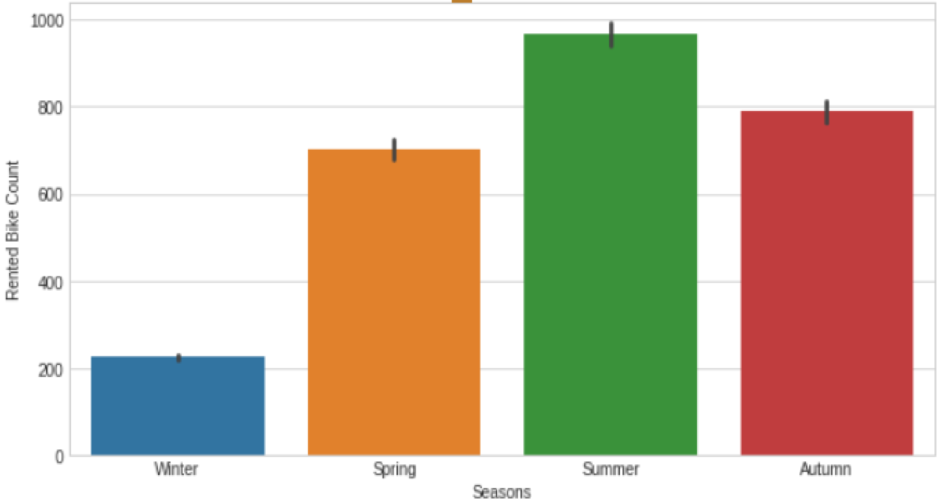


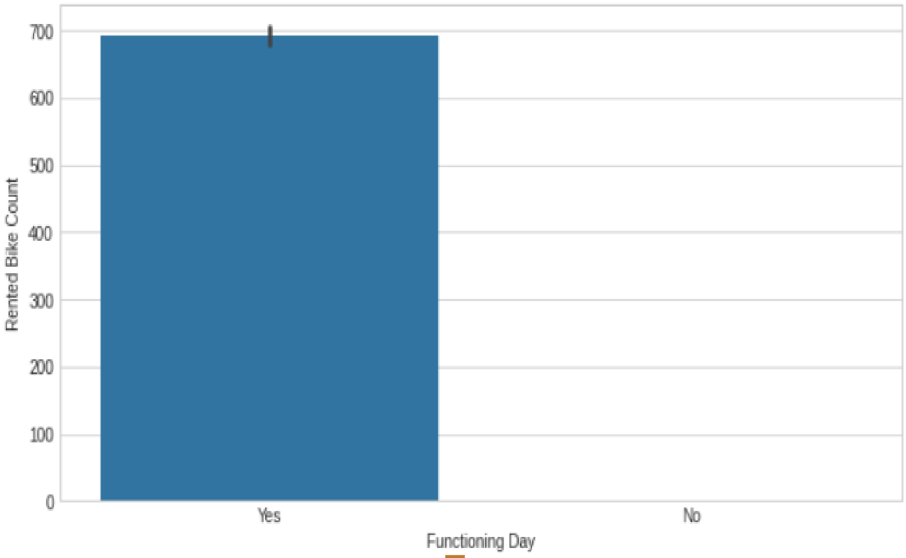
- 8 A.M. has the highest number of rented bikes in morning while later at 5 P.M. it is highest (overall).
- We can say that working hours has higher traffic.



July (7) month has the highest while January (1) has lowest number of rented bike record.

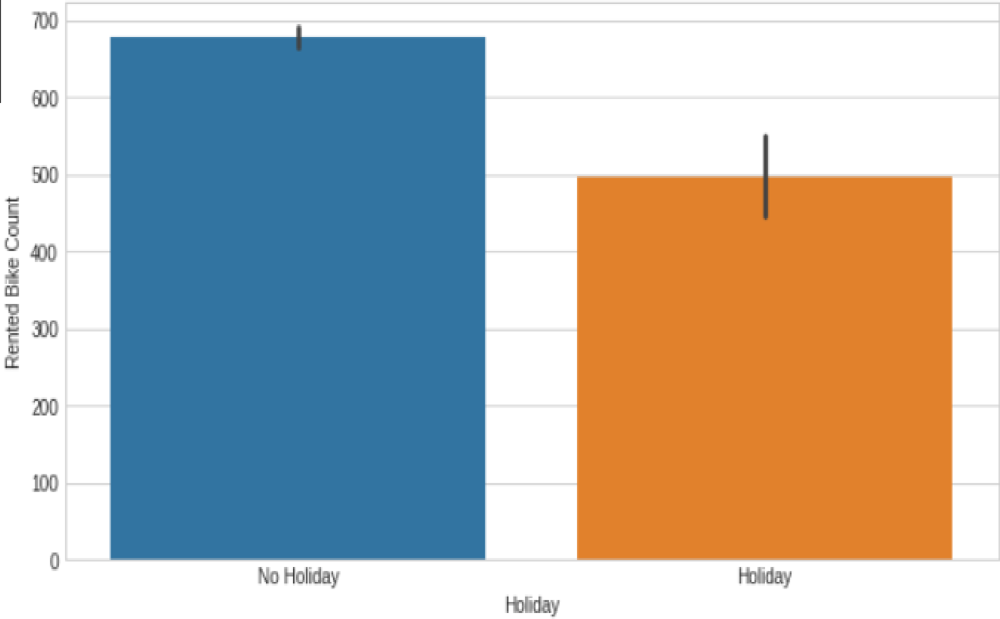
- Summer has the highest while winter has the lowest number rented bikes.
- The weather or climatic condition is directly affecting the overall number of rented bikes.



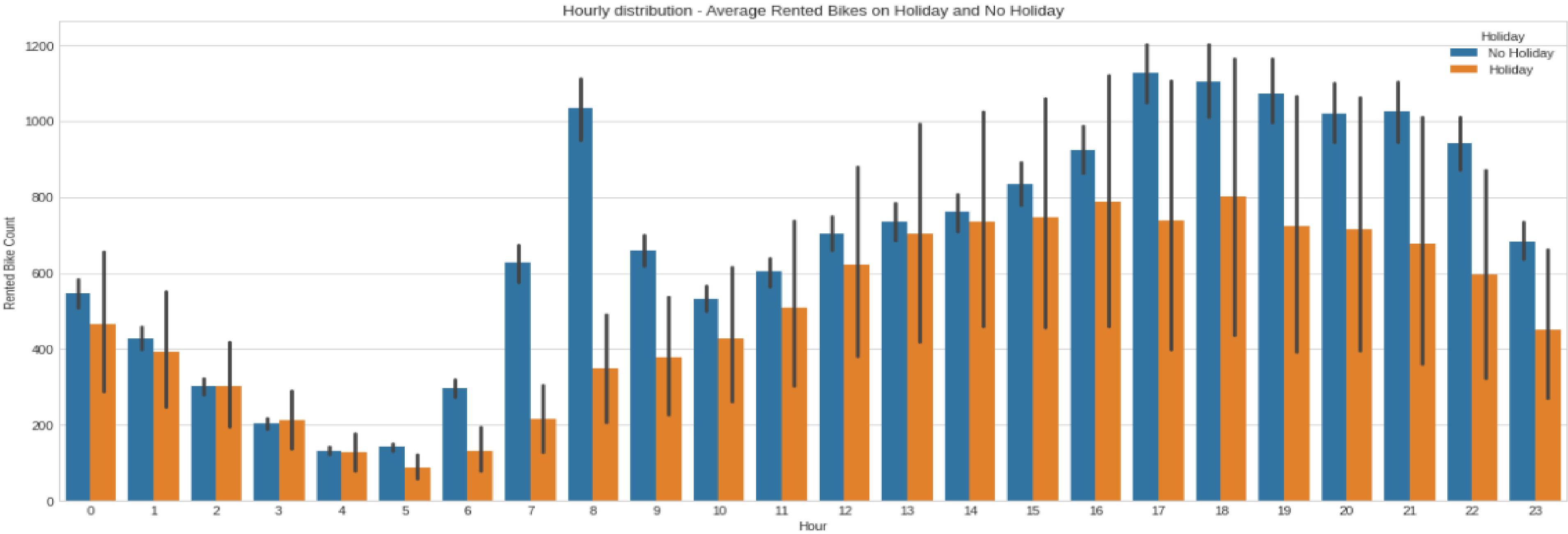


There are no rented bikes on non-functioning day.

The number of rented bikes are higher on No-Holidays as compare to Holiday.

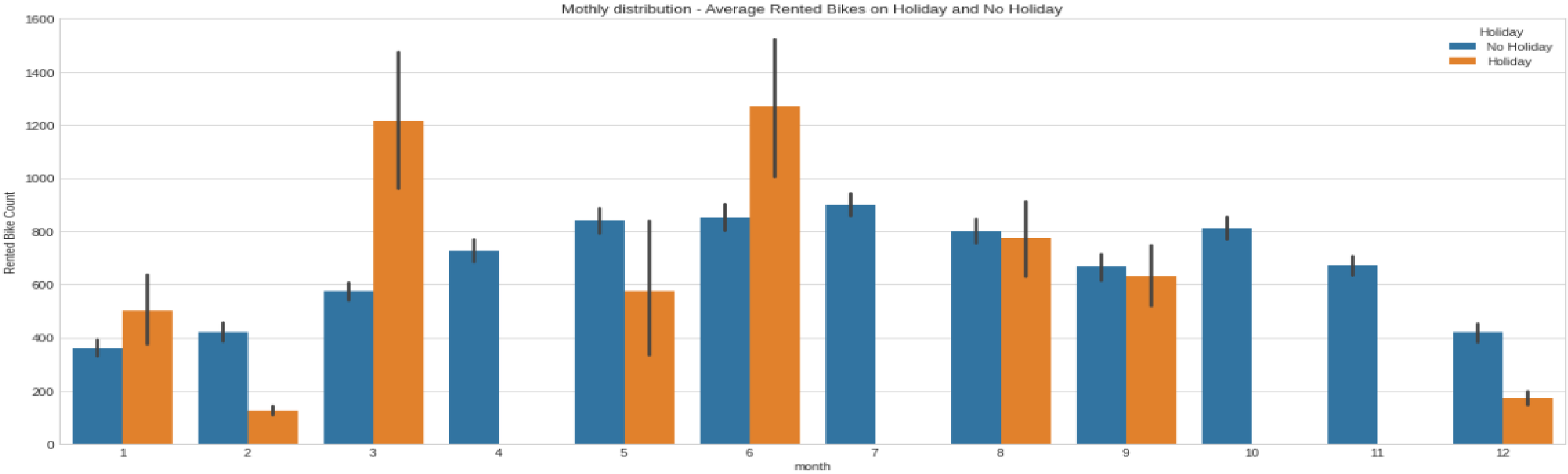


Plot to show the hourly average of rented bikes on Holiday and No Holiday



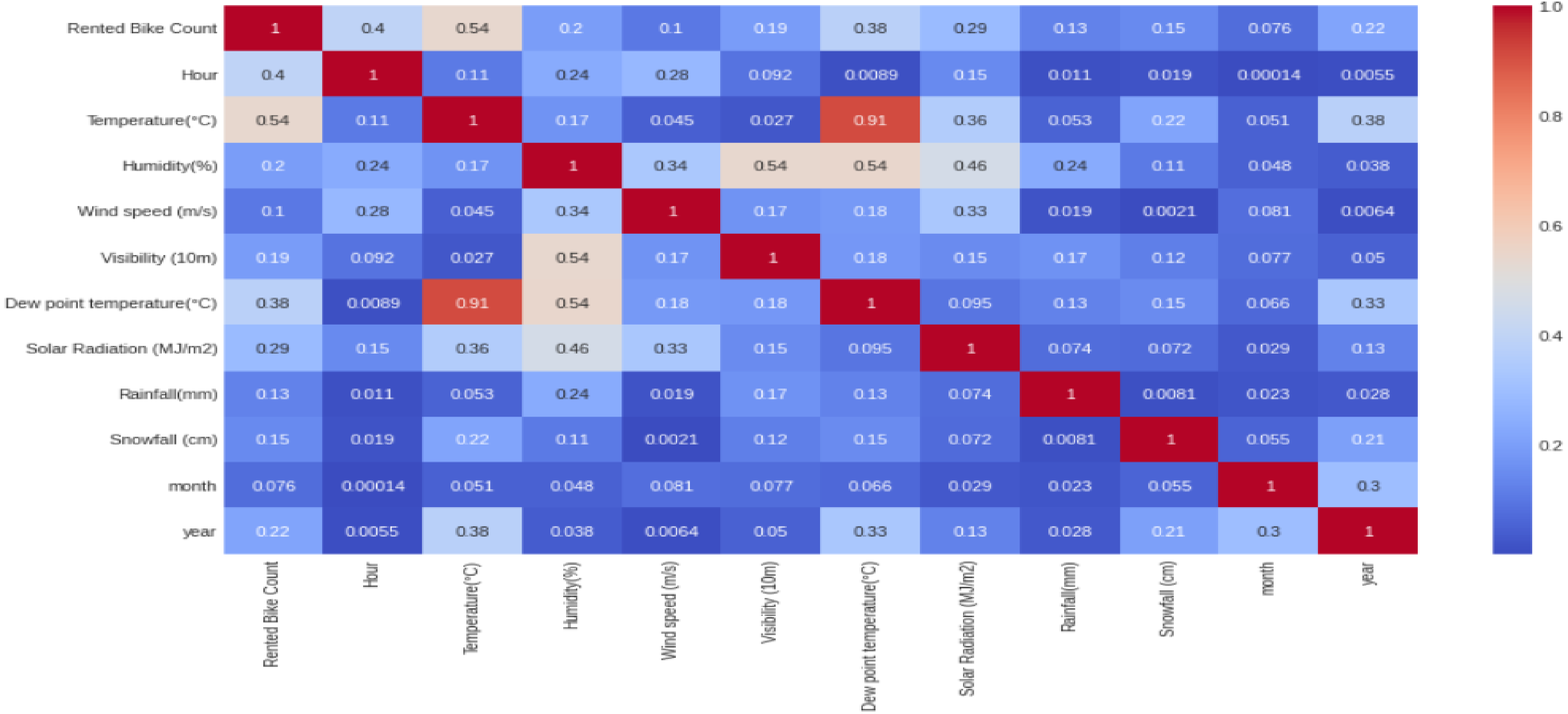
- Overall the day without holiday has highest rented bikes across the day.
- Only at 3 A.M. has the number of rented bikes on holiday are slightly higher as compare to the no-holiday.

Plot to show the Monthly distribution of average rented bikes on Holiday and No Holiday



- The month of January (1), March (3), June (6) have recorded more bike rents on holiday than no-holiday.
- The month of February (2), May (5), August (8), September (9) and December (12) have recorded more bike rents on no-holiday than the holiday.
- The month of April (4), July (7), October (10), November (11) only have records of no-holiday.

Heat Map



‘Temperature’ and ‘Dew point temperature’ are highly correlated so we can remove the ‘Dew point temperature’.

Regression Algorithms Implementation

- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic Net Regression
- Decision Tree
- Gradient Boosting Machine (GBM) Algorithm
- Random Forest

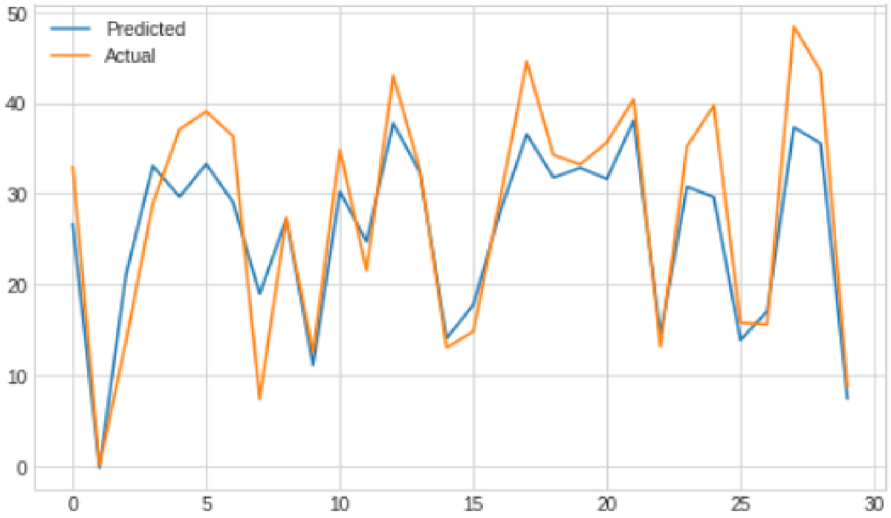
Linear Regression

Train Set Result

MSE: 112025.77624181836
RMSE: 334.70251902520596
R2: 0.680463123643269
Adjusted R2: 0.6778660749451524

Test Set Result

MSE: 104362.71856230534
RMSE: 323.052191700204
R2: 0.6915679926964599
Adjusted R2: 0.6839233480921776



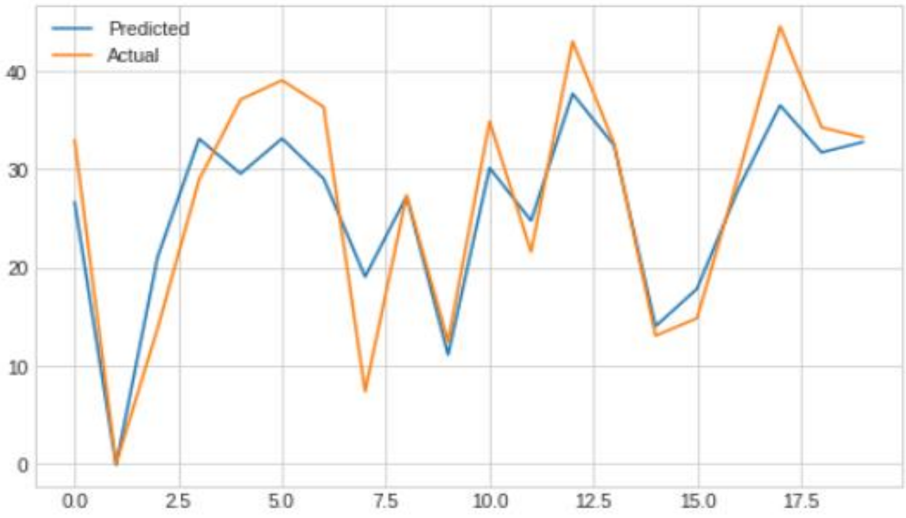
Lasso Regression

Train Set Result

MSE: 112330.94010853478
RMSE: 335.1580822664654
R2: 0.6795926890699168
Adjusted R2: 0.676988565880113

Test Set Result

MSE: 104699.4186057251
RMSE: 323.57289535083913
R2: 0.690572914457010З
Adjusted R2: 0.682903606331064



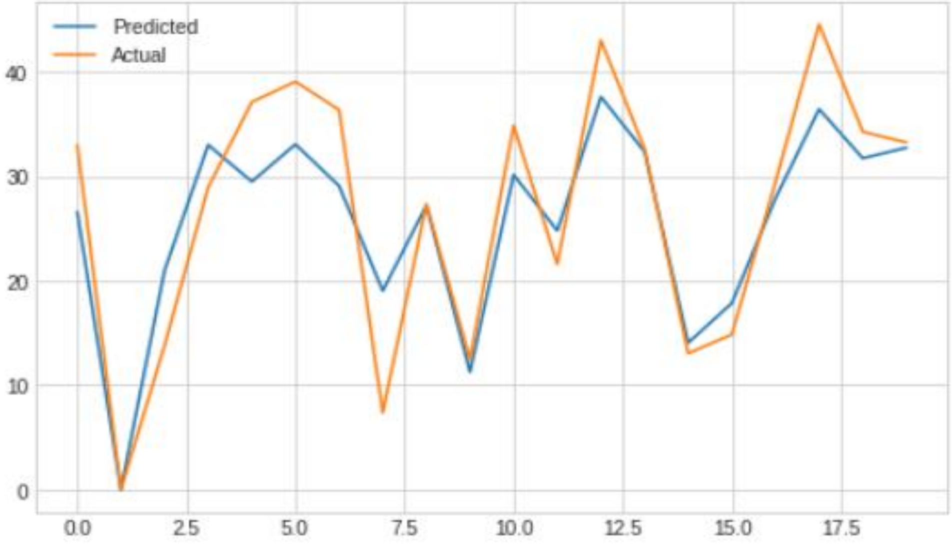
Ridge Regression

Train Set Result

MSE: 112780.35564991992
RMSE: 335.82786610095343
R2: 0.6783107980346751
Adjusted R2: 0.6756962562243911

Test Set Result

MSE: 105169.11181943778
RMSE: 324.2978751386413
R2: 0.6891847902042316
Adjusted R2: 0.6814810767107236



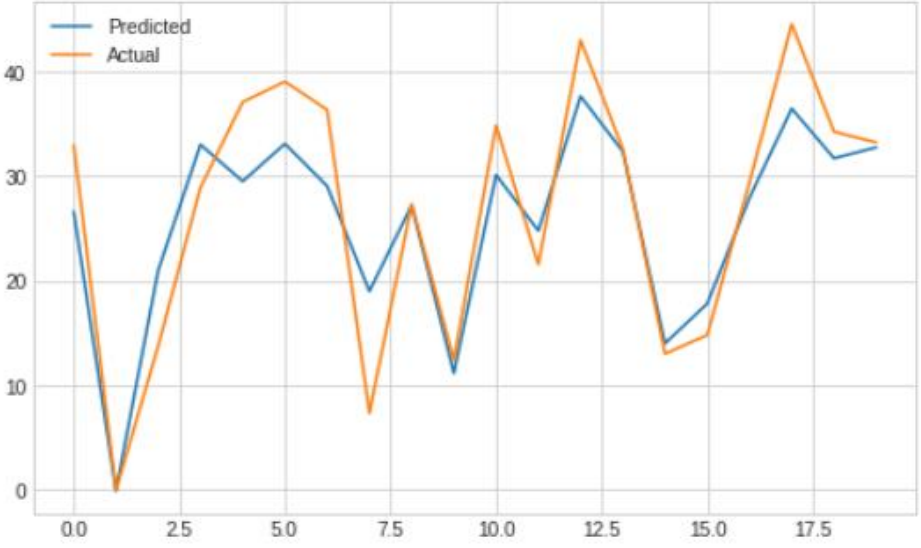
Elastic Net Regression

Train Set Result

MSE: 112495.71025126922
RMSE: 335.40380178416166
R2: 0.6791227067275265
Adjusted R2: 0.6765147637375033

Test Set Result

MSE: 104898.82803654196
RMSE: 323.88088556835515
R2: 0.6899835828271967
Adjusted R2: 0.6822996678162407



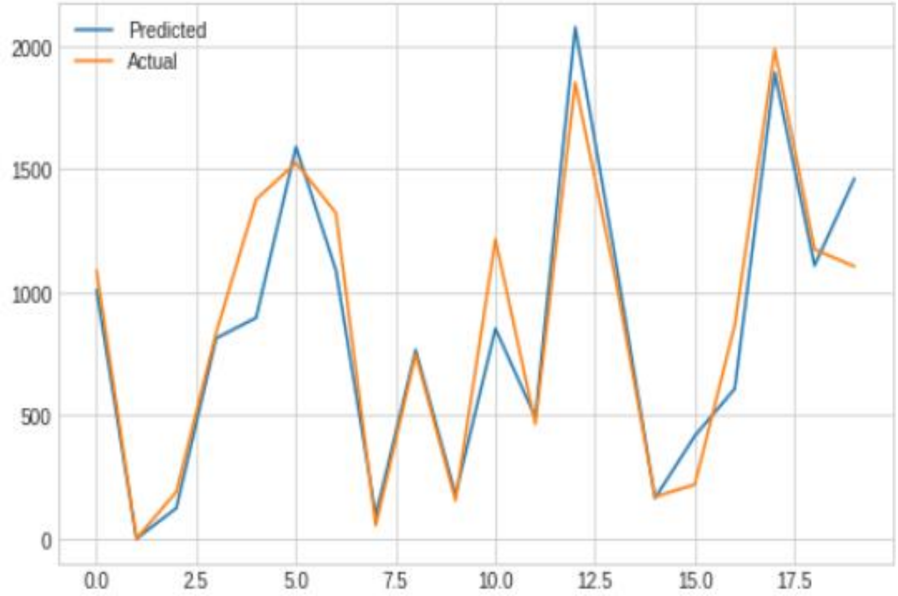
Decision Tree

Train Set Result

MSE: 60452.605487521316
RMSE: 245.87111560230355
R2: 0.8275679279078496
Adjusted R2: 0.8261664793694327

Test Set Result

MSE: 88098.91248256624
RMSE: 296.8146096177987
R2: 0.7417694596021831
Adjusted R2: 0.7353690839583867



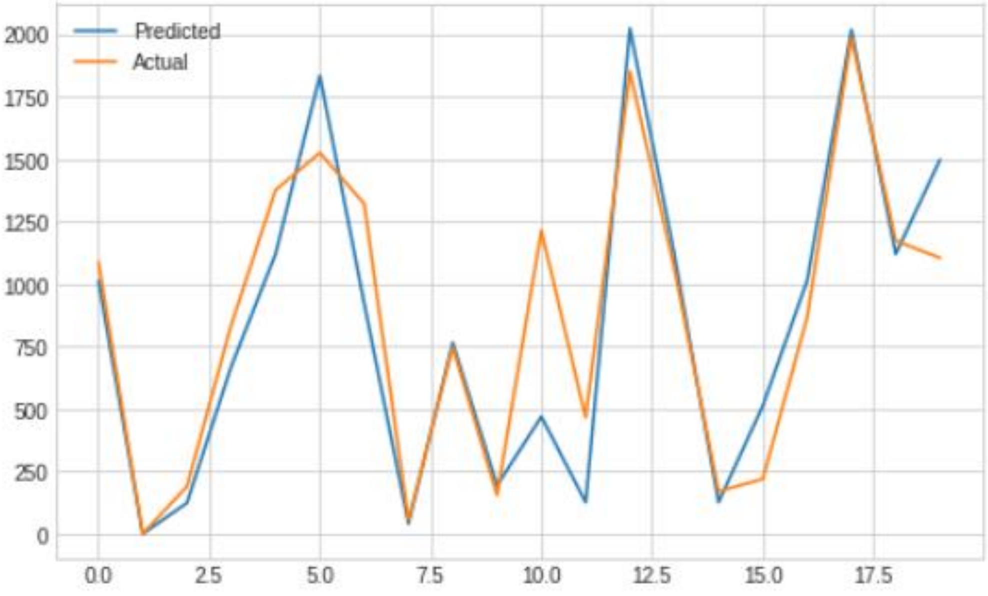
Random Forest

Train Set Result

MSE: 8313.886129407461
RMSE: 91.18051397863175
R2: 0.9762858755074199
Adjusted R2: 0.9760931380154514

Test Set Result

MSE: 50036.10882855877
RMSE: 223.68752497302742
R2: 0.8521240372400101
Adjusted R2: 0.848458856084853



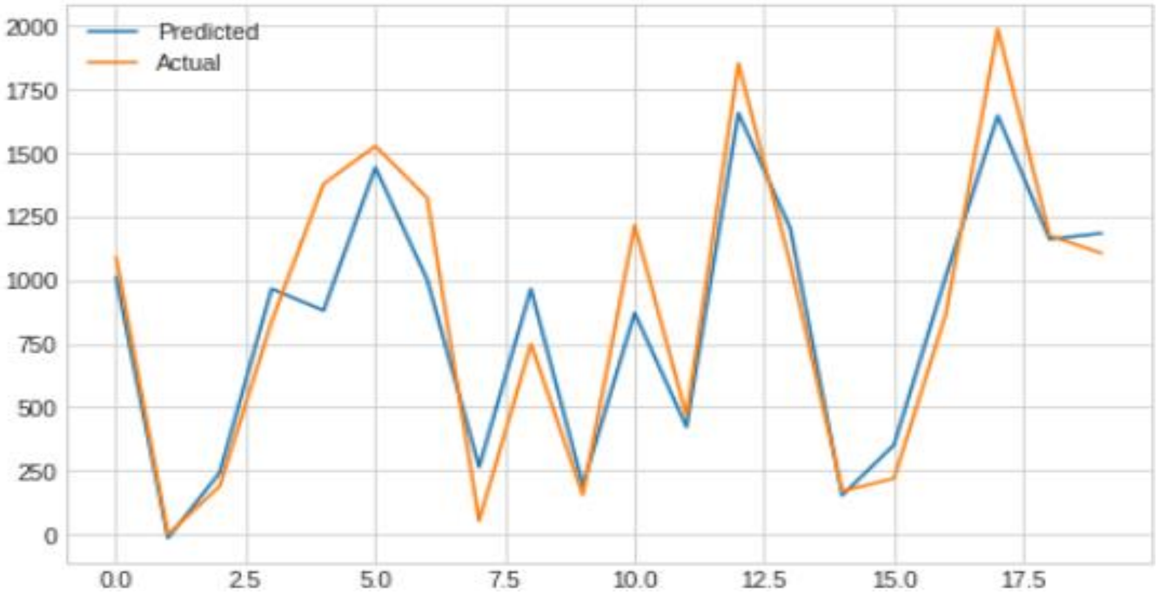
Gradient Boosting Machine (GBM) Algorithm

Train Set Result

MSE: 66303.73504725948
RMSE: 257.4951165503134
R2: 0.8108784504911393
Adjusted R2: 0.8093413575598387

Test Set Result

MSE: 73333.25394781605
RMSE: 270.80113357926706
R2: 0.7832720052829799
Adjusted R2: 0.7779002914005753



Challenges

- **The Dataset is large.**
- **The project requires some domain knowledge.**
- **Some features are interconnected so the feature selection was challenging.**
- **Large graphical representation was required.**
- **Instances of rare events esp. related to climate or weather were mentioned in the dataset.**
- **Different algorithms were giving different performance scores and the variation was large.**
- **The number of features was affecting the overall model performance.**

Conclusion

- There are several features that are impacting the overall bike rental trend, these features are ‘Hour’, ‘Season’, ‘Holiday’ or ‘No Holiday’, and those covering climate esp. ‘rainfall’ and ‘snowfall’.
- There are no operations of agency on non-functioning days so no bikes were rented during those days.
- Among different seasons ‘Summer’ has the highest while ‘winter’ has the lowest number of bike rentals.
- The outliers that the graphs of ‘wind speed’, ‘solar radiation’, ‘rainfall’ and ‘snowfall’ show are not the outliers.
- Snowfall and rain do affect the bike rentals as the number reduces during such and days with clear weather have higher number rented bikes.
- Non-Holidays or working days have a higher number of rented bikes during 7-9 AM and later 5-10 PM. While on holidays there is not any specific peak time.
- Out the all the models that are used Random Forest has the highest R2 Score i.e. Train-0.976 (approx.), Test-0.848(approx.).

Thank You