Capstone Project - 2

Bike Sharing Demand Prediction

Team Members:

V Bhavya Reddy

Introduction

From a business perspective, predicting the number of rented bikes is a crucial part of the revenue generation process because having excess bikes results in resource waste (bike maintenance, bike parking, and security costs), and having fewer bikes leads to revenue loss (losing customers due to non-availability). By estimating the bikes to be rented, the company can work more efficiently.

The goal of this project is to provide a model to predict stable supply of bike rentals to predict demand at any hour. This dataset contains hourly bike rental counts along with weather information and dates spanning an entire year.

Problem Statement

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.

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Attributes

- 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)

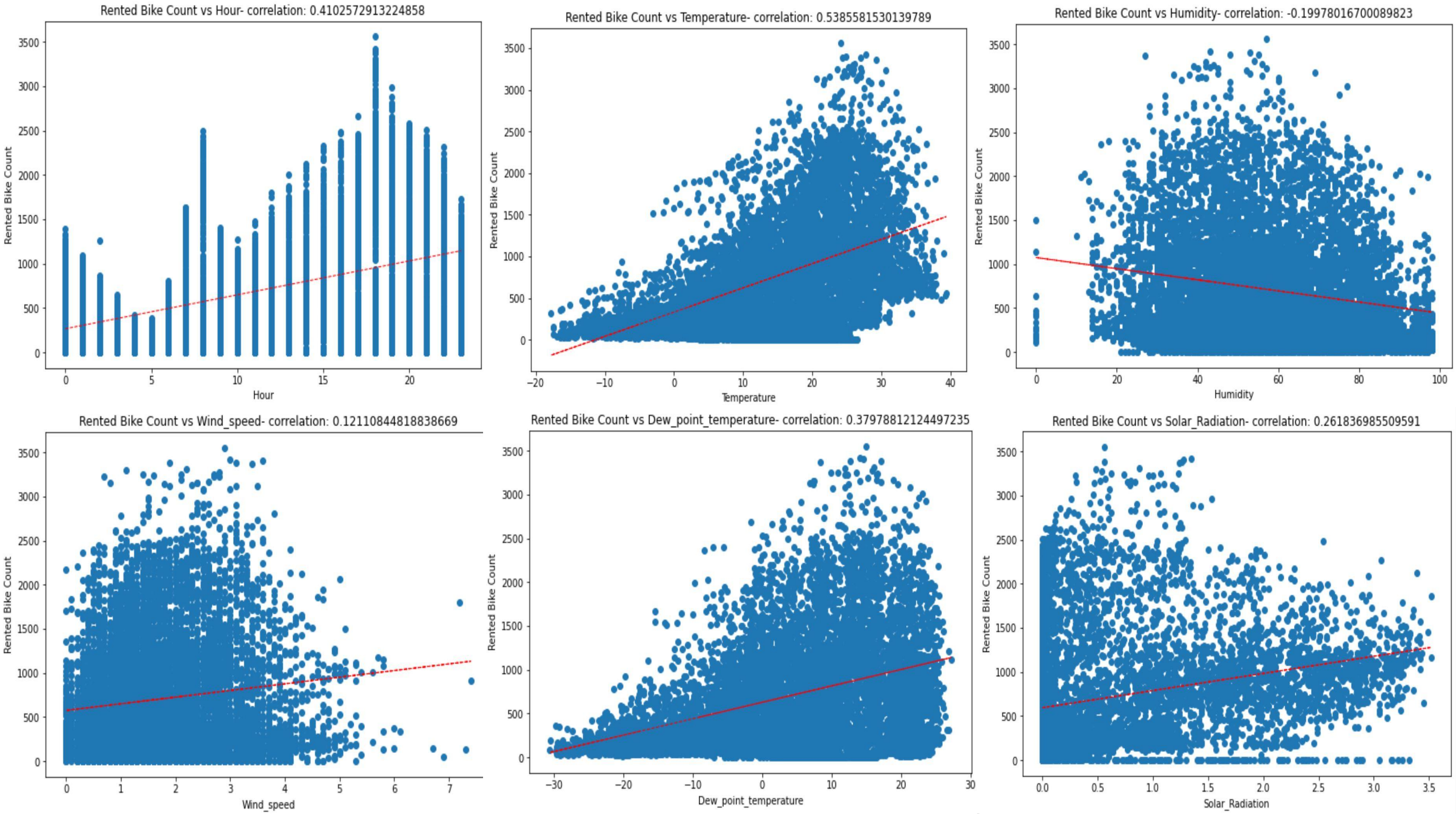
Column Name	Format	Range	Explanation
Date	dd/mm/yyyy	01/12/2017 To 30/11/2018	Date
Rented Bike count	int64	0 - 3556	Count of bikes rented at each hour
Hour	int64	0 - 23	Hour of the day
Temperature	float64	-17.8 to 39.4	Temperature in Celsius
Humidity	int64	0 - 98	Humidity as %
Windspeed	float64	0 - 7.4	Wind Speed
Visibility	int64	27 - 2000	Visibility
Dew point temperature	float64	-30.6 - 27.2	Dew point temperature in Celsius
Solar radiation	float64	0 - 3.52	Solar radiation in MJ/m2
Rainfall	float64	0 - 35	Rainfall in mm

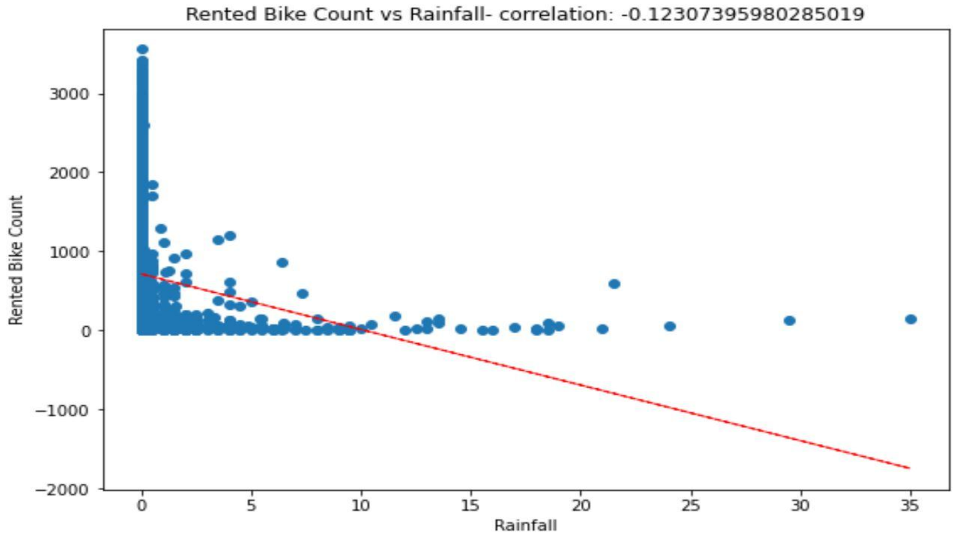
Column Name	Format	Range	Explanation
Snowfall	float64	0 - 8.8	Snowfall in cm
Seasons	object	Winter, Spring, Summer, Autumn	Which season is it?
Holiday	object	Holiday/No holiday	Is it a Holiday or not
Functional Day	object	NoFunc(Non Functional Hours), Fun(Functional hours)	

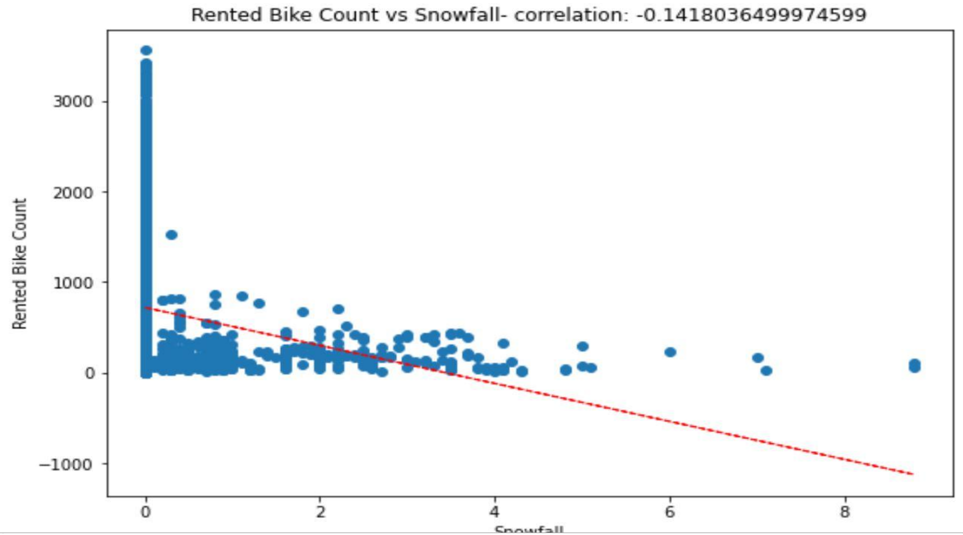
Observations on attributes:

- The **target column** Rented Bike Count per hour, ranges between 0 and 3556 over the 1 year span.
- Mean of Rented Bike Count = 704.6, with median and 75% percentile = 504.5 and 1065.25, respectively. This suggests that the 'Rented Bike Count' distribution is more denser at lower values. This is expected as out of 24 hours, we would expect the bike demand/usage to be high for maximum of maybe 6 hours or so.
- Hence, we shall expect a strong correlation with hour column.

Bivariate Analysis of numeric features

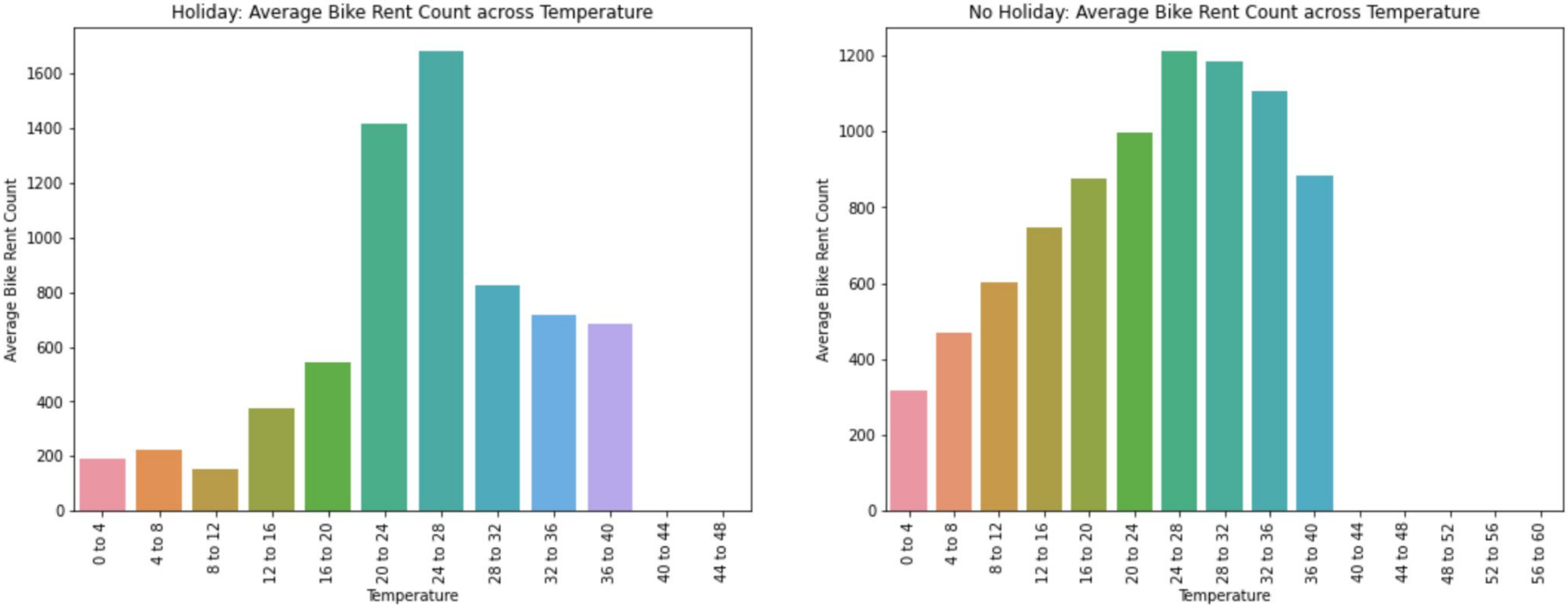

Rented Bike Count vs Hour- correlation: 0.4102572913224858
Rented Bike Count
Hour
Rented Bike Count vs Temperature- correlation: 0.5385581530139789
Rented Bike Count
Temperature
Rented Bike Count vs Humidity- correlation: -0.19978016700089823
Rented Bike Count
Humidity
Rented Bike Count vs Wind_speed- correlation: 0.12110844818838669
Rented Bike Count
Wind_speed
Rented Bike Count vs Dew_point_temperature- correlation: 0.37978812124497235
Rented Bike Count
Dew_point_temperature
Rented Bike Count vs Solar_Radiation- correlation: 0.261836985509591
Rented Bike Count
Solar_Radiation





Overall Numeric feature observation:

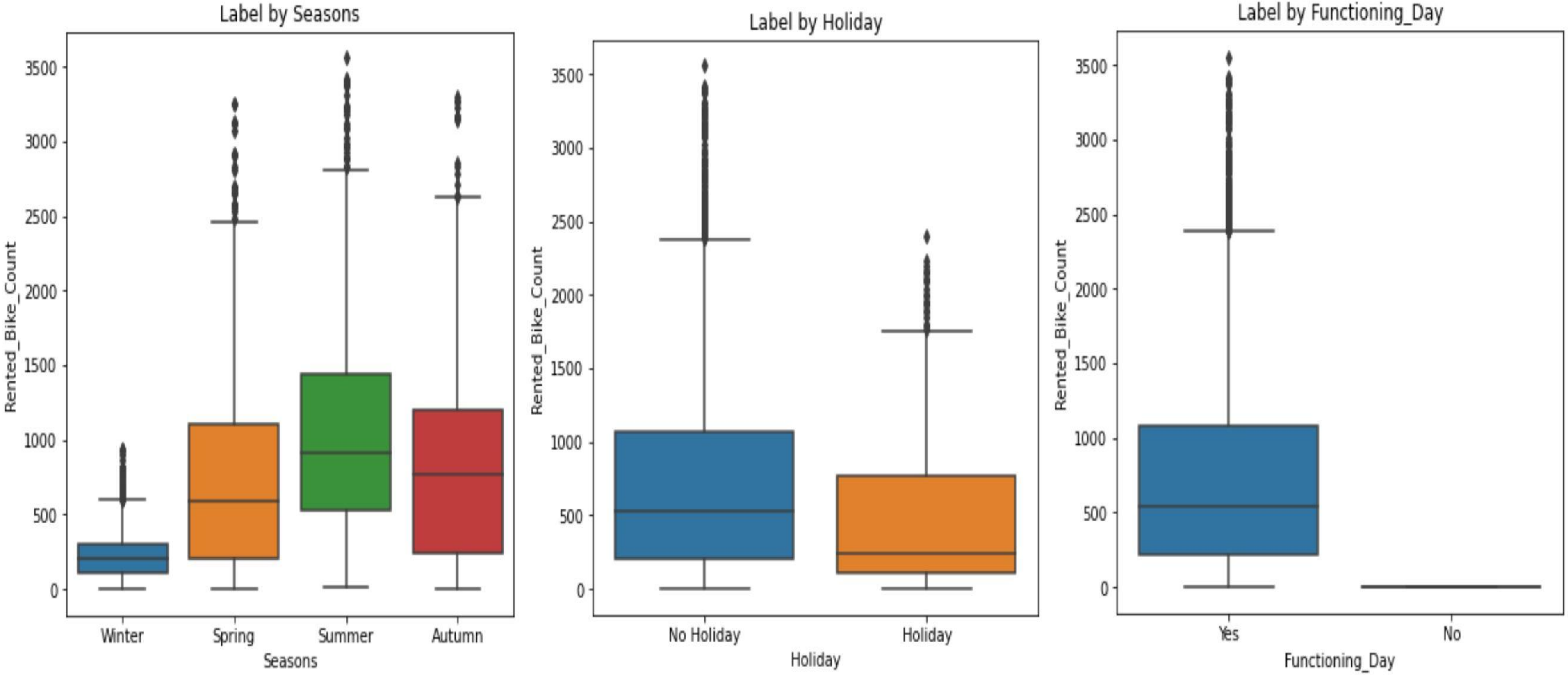
* Higher reservations can be seen at around 8am and 6pm (office hours) and very few reservations at very early in the morning.
* We can notice that in general, more people tend to prefer biking at moderate to high temperatures; however, if the temperature is too hot there is a small decline in count.
* Temperature, Windspeed, Visibility, Dew point temperature, Solar radiation have a positive correlation with bike rents.
* Humidity, Rainfall, Snowfall have a negative correlation with bike rents.



INFERENCE:

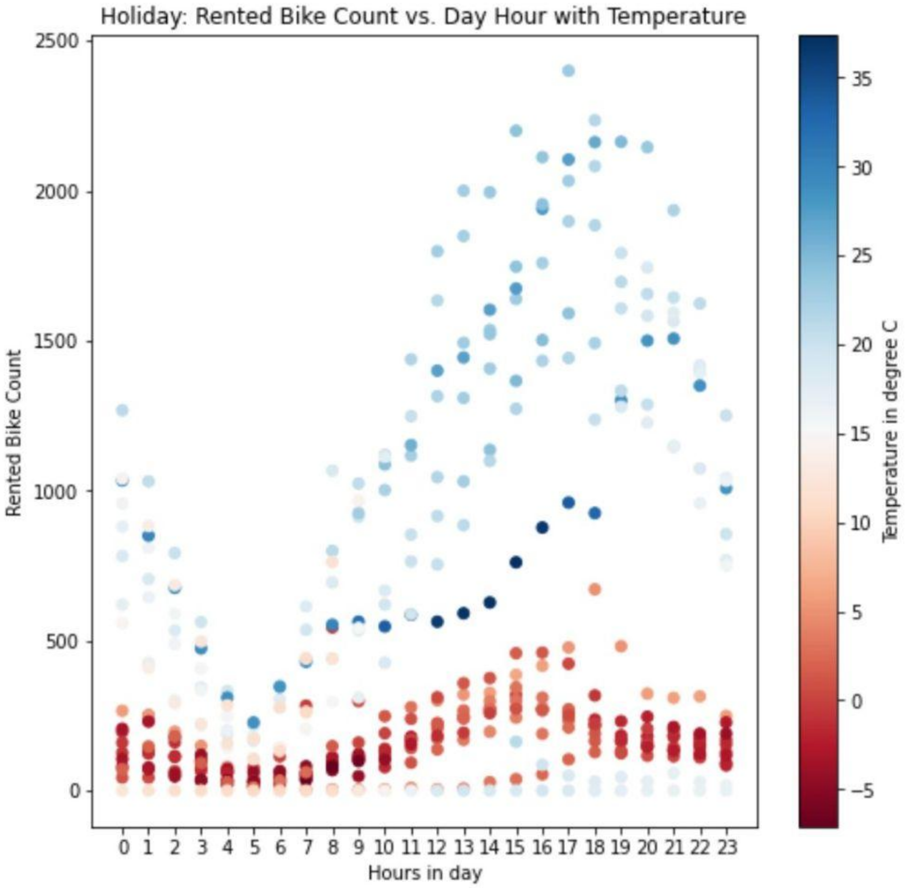
From the above bar plot, we notice that there is a increase in the average bikes rented with temperature and a small decrease at the highest temperature bin.

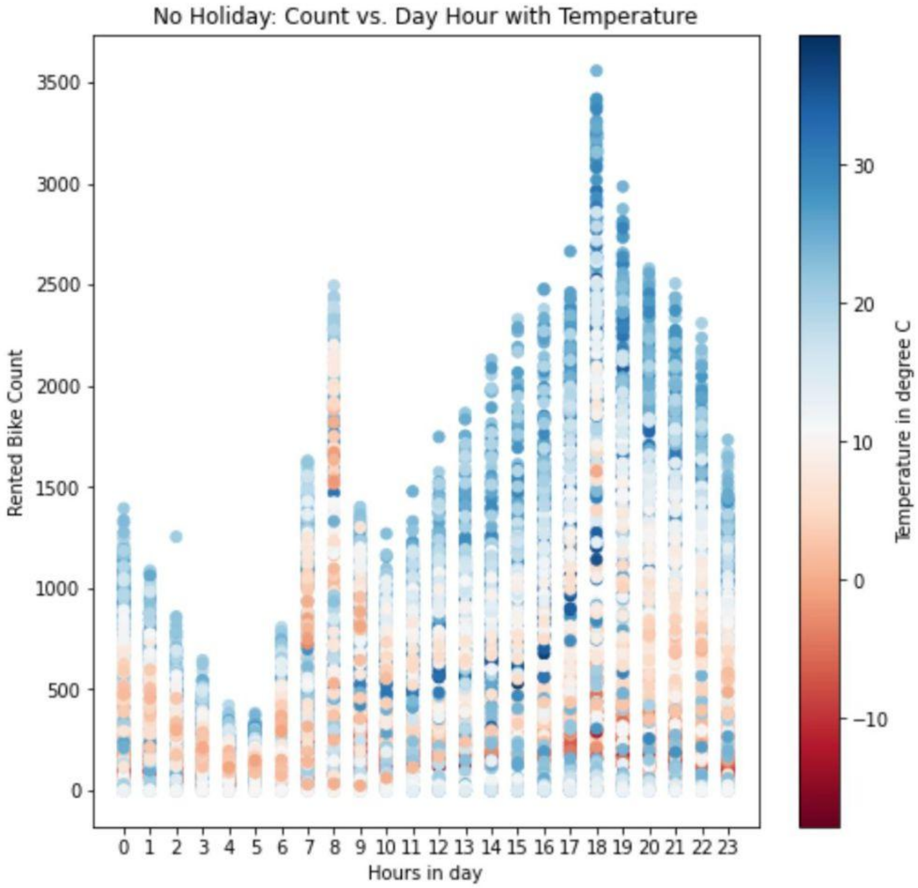
Rented Bike Count vs. Seasons, Holiday, Functioning Day



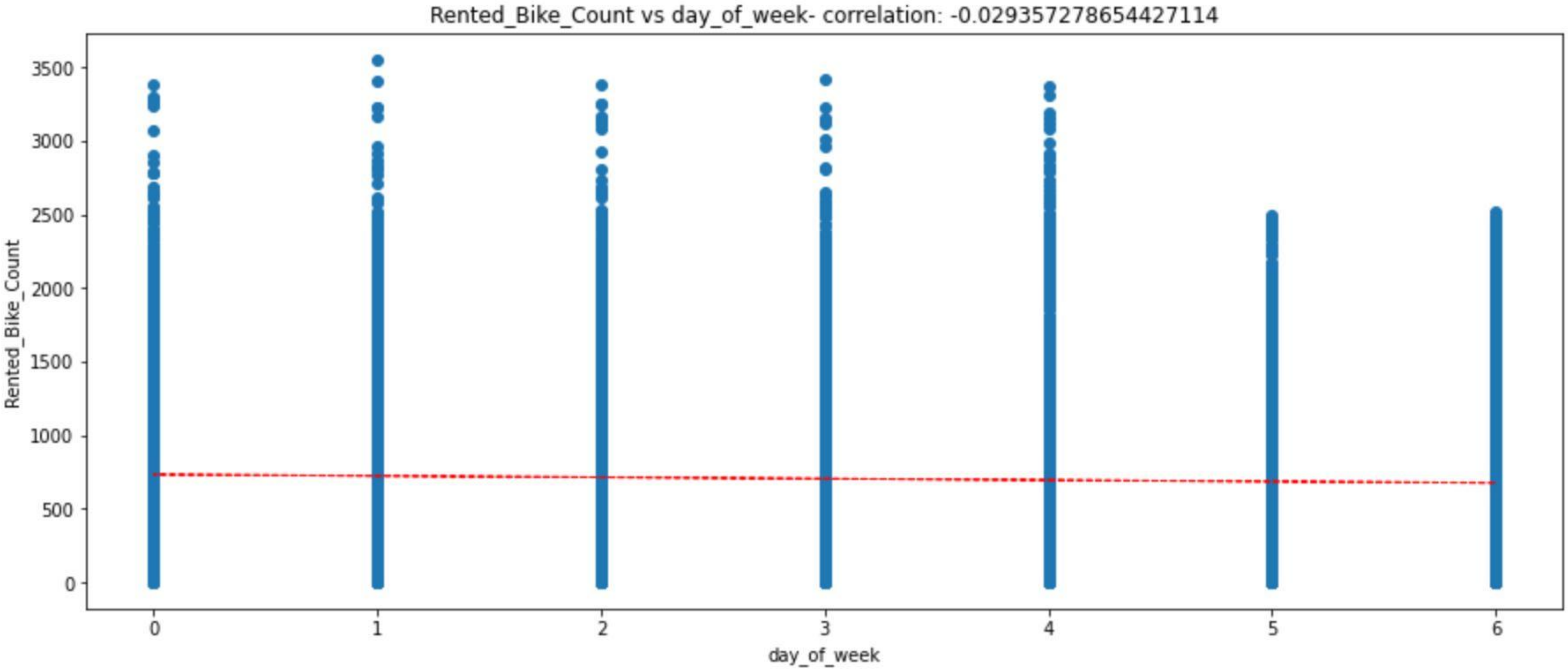
Rented Bike Count vs. Seasons, Holiday, Functioning Day

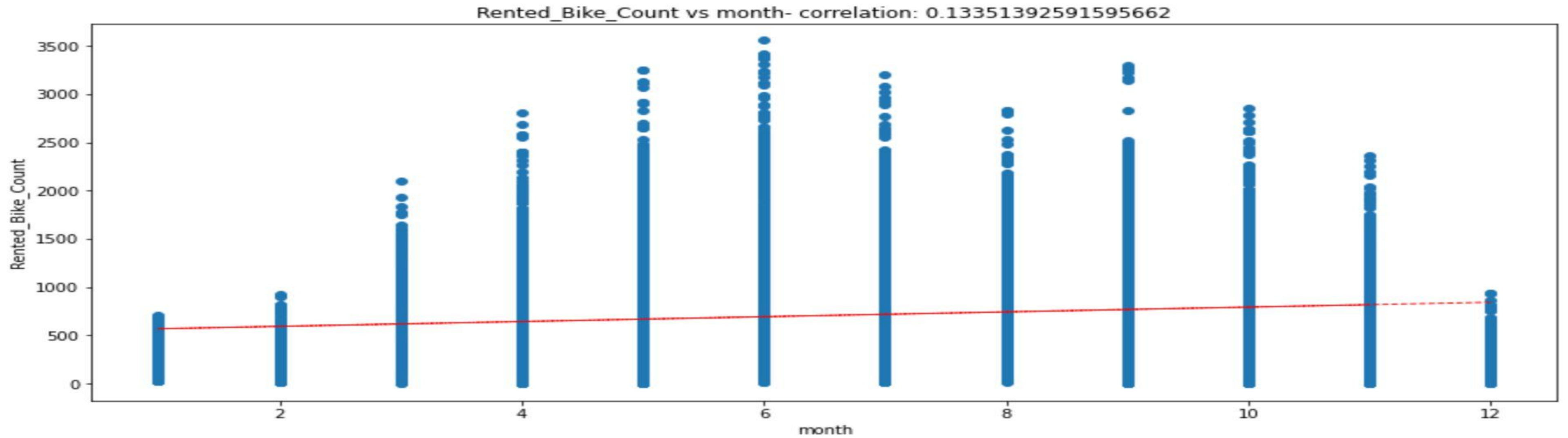
- Rented Bike Count are lesser in Winter season compared to other seasons.
- Lots of outlier points in every season and for 'No Holiday'.
- Many bikes were rented on working days.
- On the non functioning day there are no bike rented.

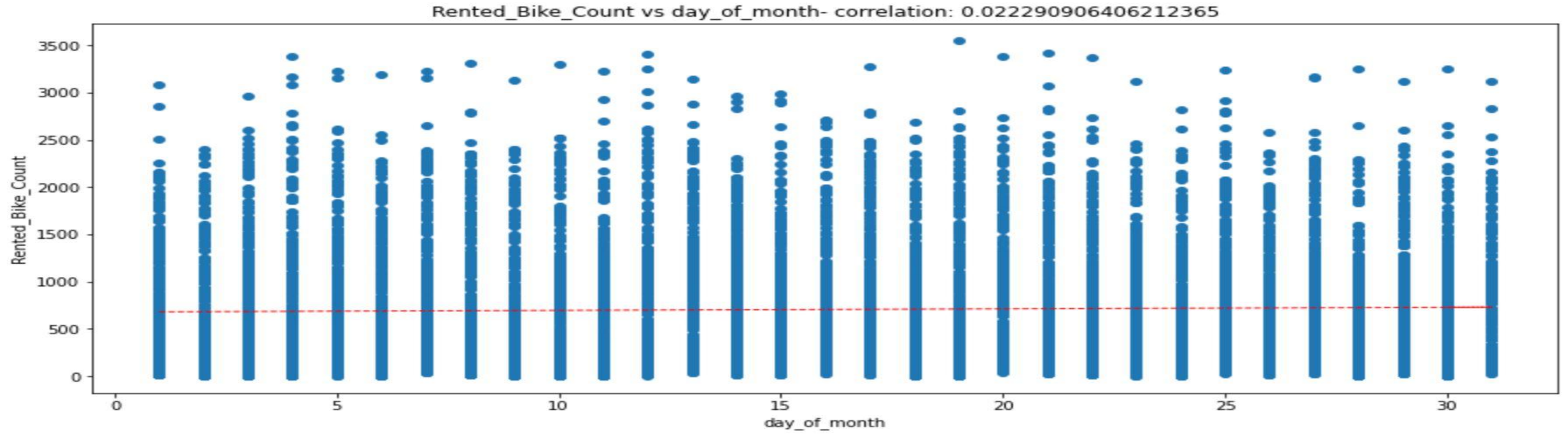

Holiday: Rented Bike Count vs. Day Hour with Temperature
Rented Bike Count
Hours in day
Temperature in degree C

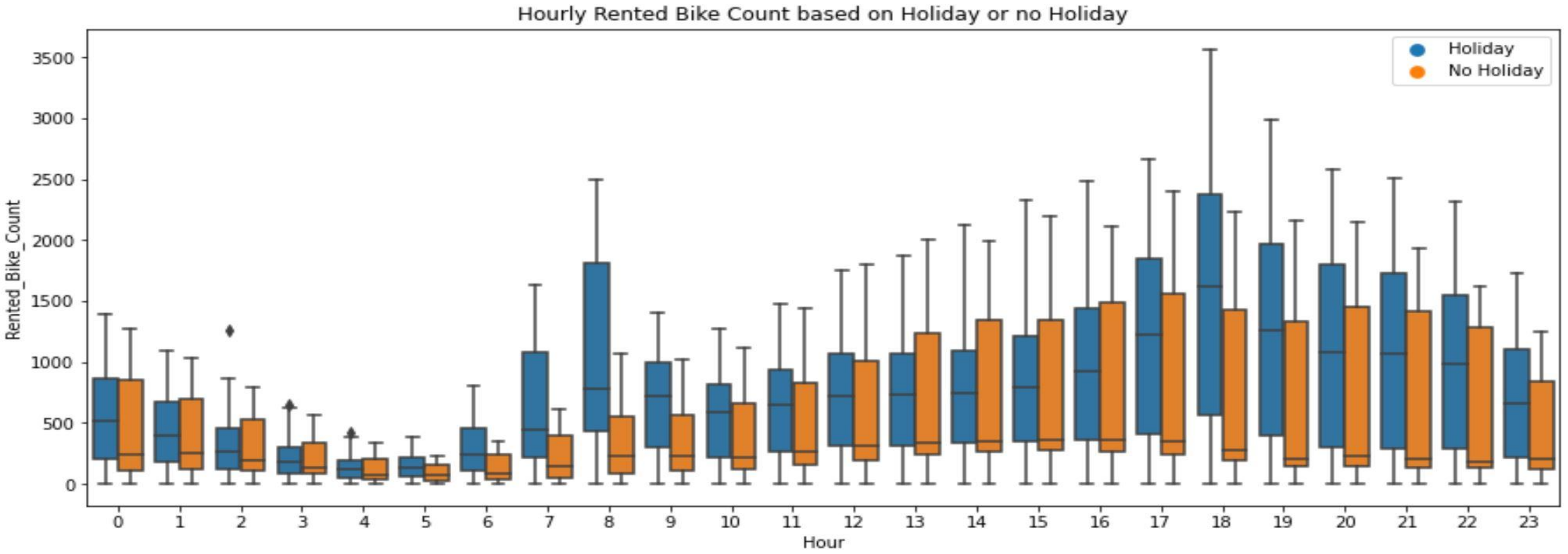

No Holiday: Count vs. Day Hour with Temperature
Rented Bike Count
Hours in day
Temperature in degree C

Date Feature:

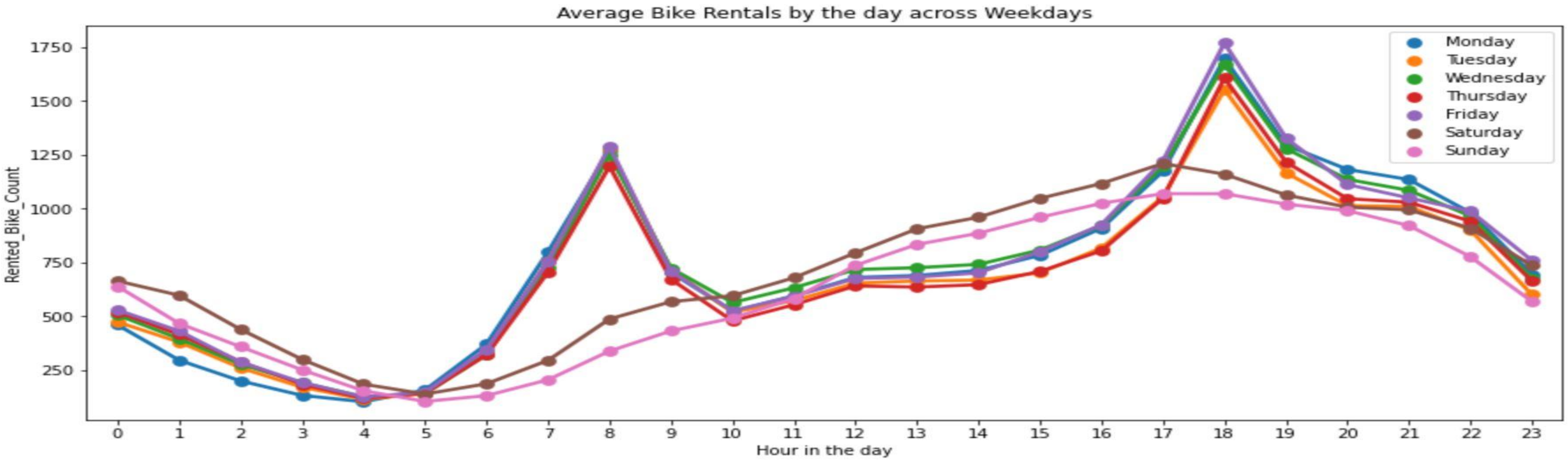



Rented_Bike_Count vs month- correlation: 0.13351392591595662
Rented_Bike_Count
month
0
500
1000
1500
2000
2500
3000
3500
2
4
6
8
10
12

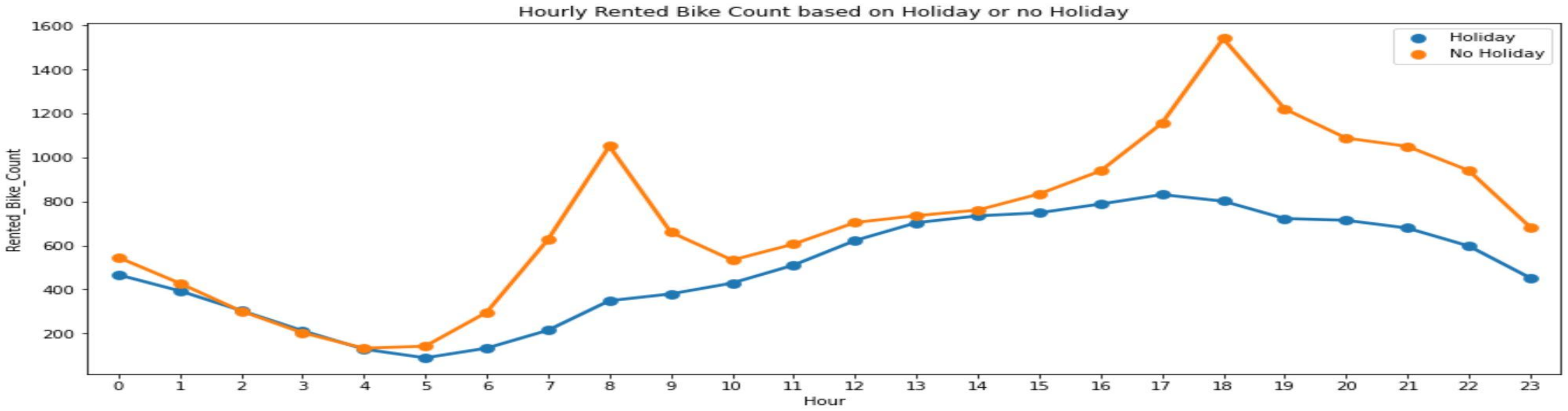

Rented_Bike_Count vs day_of_month- correlation: 0.022290906406212365
Rented_Bike_Count
day_of_month
0
500
1000
1500
2000
2500
3000
3500
0
5
10
15
20
25
30



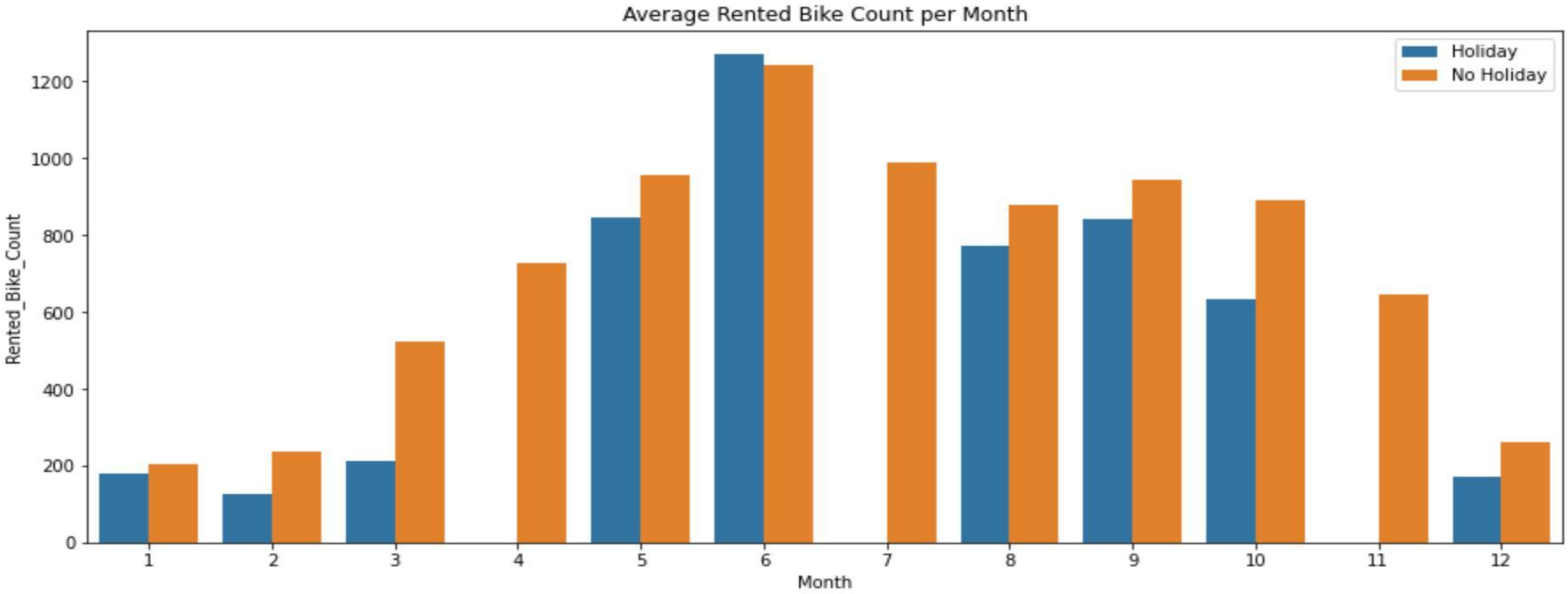
Very few number of outliers can be seen in the seaborn box plot across hours.



Average Bike Rentals by the day across Weekdays
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Rented_Bike_Count
Hour in the day



Hourly Rented Bike Count based on Holiday or no Holiday
Holiday
No Holiday
Rented_Bike_Count
Hour



we can see that we have highest bike rents in June. No holidays in 4th, 7th, and 11th month.

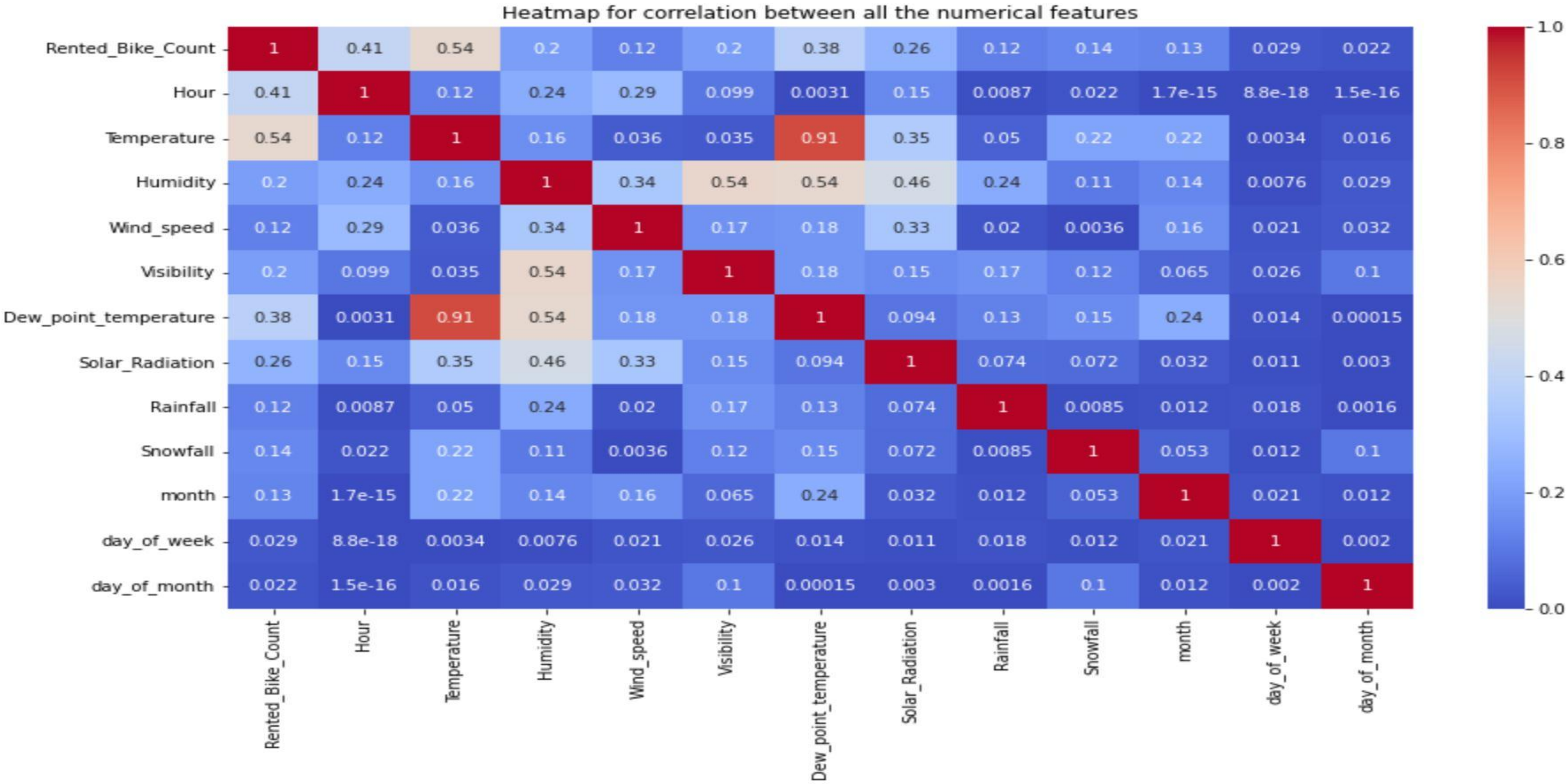
Observation

Higher reservations can be seen at 8am and 6pm (office hours) and very low reservations at very early in the morning.

No Holiday: There is a peak in the rentals at around 8am and another at around 6pm.

Holiday: There is steady increase in rentals from 1pm to 5pm. These correspond to probably tourists.

Heat map for numeric feature column



Inferences from the heatmap:

- Temperature and Dew point temperature are highly correlated.
- We see a positive correlation between Rented Bike Count and Temperature (as seen in the scatter plot). This is probably only true for the range of temperatures provided.
- We see a negative correlation between Rented Bike Count and Humidity. The more the humidity, the less people prefer to ride.
- Rented Bike Count has a weak dependence on day_of_month, day_of_week.

VIF

Multicollinearity check

	variables	VIF
0	Hour	4.019774
1	Temperature	3.308515
2	Humidity	7.407425
3	Wind_speed	4.669663
4	Visibility	5.600624
5	Solar_Radiation	2.301785
6	Rainfall	1.082041
7	Snowfall	1.141194
8	month	5.041744
9	day_of_week	3.124912
10	day_of_month	3.798687

Summary

Season: Month column has a direct mapping with season (Winter:December, January, February Summer: June to August, Autumn: September to November and Spring: March to May). Hence we will drop Seasons column.

Functioning Day: The bikes rented on Non-functioning days are zero, so we remove the rows of non-functioning day because we do not want to create any bias and later drop Functioning Day column.

Temperature: Temperature and Dew point temperature are highly correlated. Hence retain only the Temperature column.

Date: Intuitively, there should be no dependency on date. Hence drop this column.

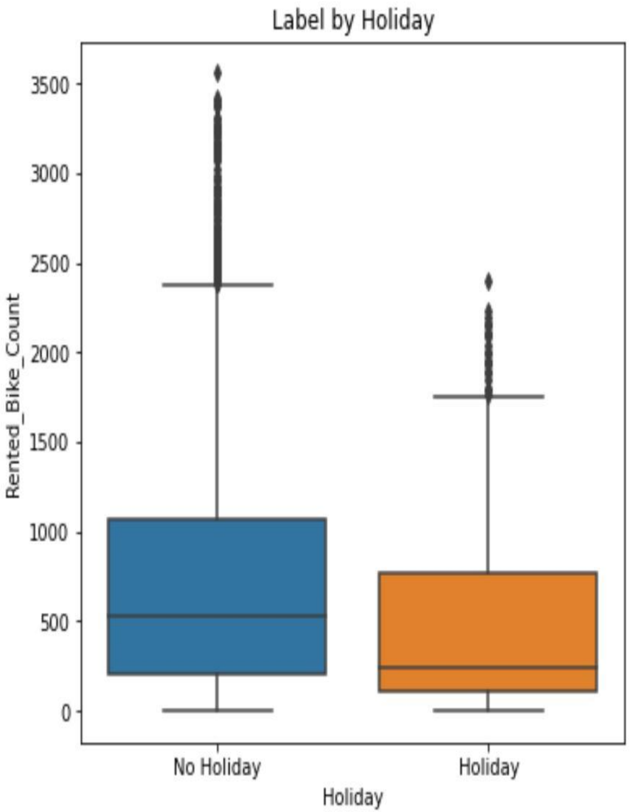
Hour: Split hour column to hour_0, hour_1, ..., hour_23. Drop hour_23 since it is a function of the rest of the hour columns.

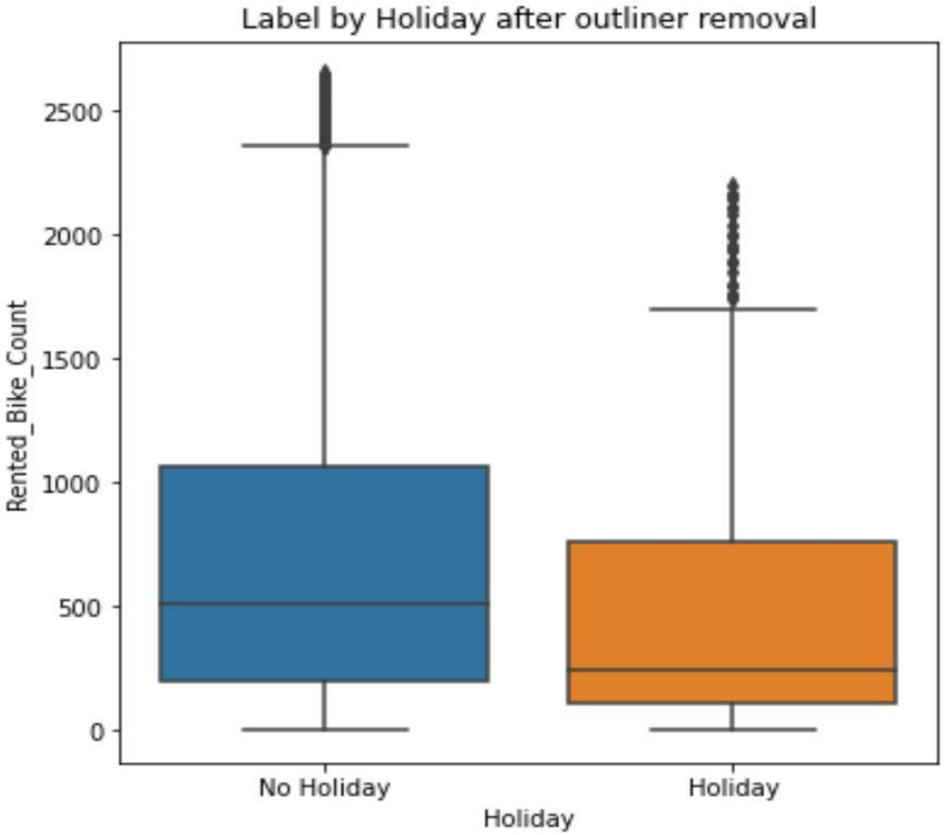
Dropped columns:

1. Dew_point_temperature, Date, day_of_week, day_of_month, Seasons, Functioning_Day - columns are dropped.
2. Outliers removal - Holiday
3. Data encoding on - Holiday, Hour, month.
4. Hour, Hour_23, Seasons, Seasons_Winter - columns are dropped.

Feature Engineering

Removing Outliers Holiday





Dropping columns

1. Dew_point_temperature
2. Date
3. Day_of_week
4. Day_of_month
5. Seasons
6. Functioning_Day

Also, remove entries of Functioning_Day == 0.

1. Linear Regression

Linear Regression Score : **0.6925**

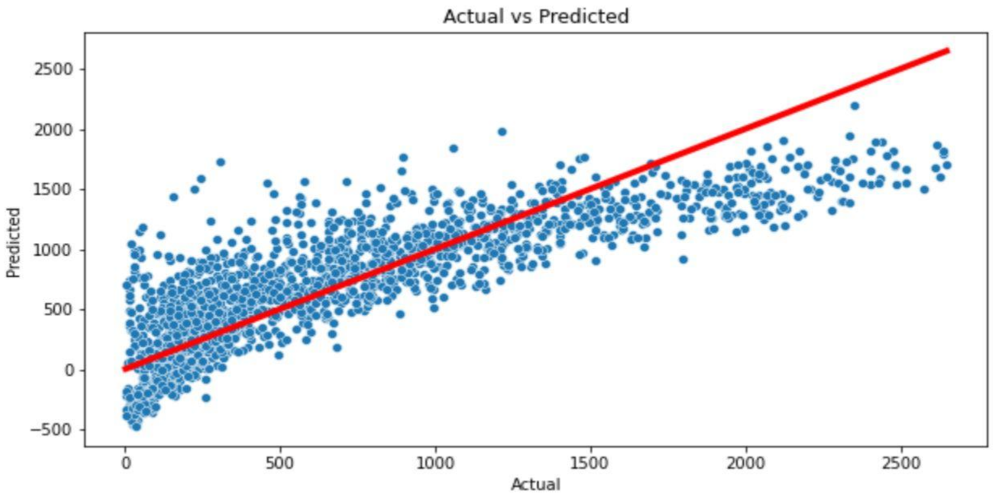
Linear Regression metrics	
MSE for train dataset:	114837.751
MSE for test dataset:	119490.275
RMSE for test dataset:	345.673
R2 for test dataset:	0.685
Adjusted R2 for test dataset:	0.677

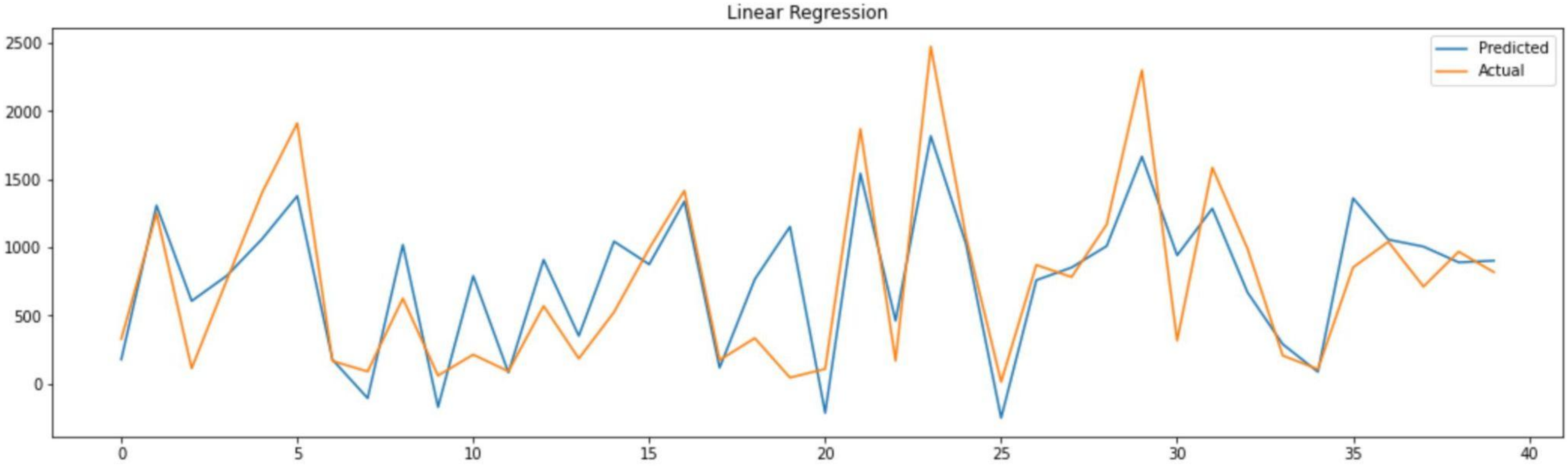
Observations:

1. Since the RMSE value of train and test data are quite close, the model doesn't seems to be an overfit model.
2. Overall, a good initial model.

Linear Regression cont..

Predicted vs. Actual shows a significant error.





Lasso Regulation

1. The best fit alpha value is found out to be : {'alpha': 0.2}

2. Using {'alpha': 0.2} the negative mean squared error is: -116364.53331421837

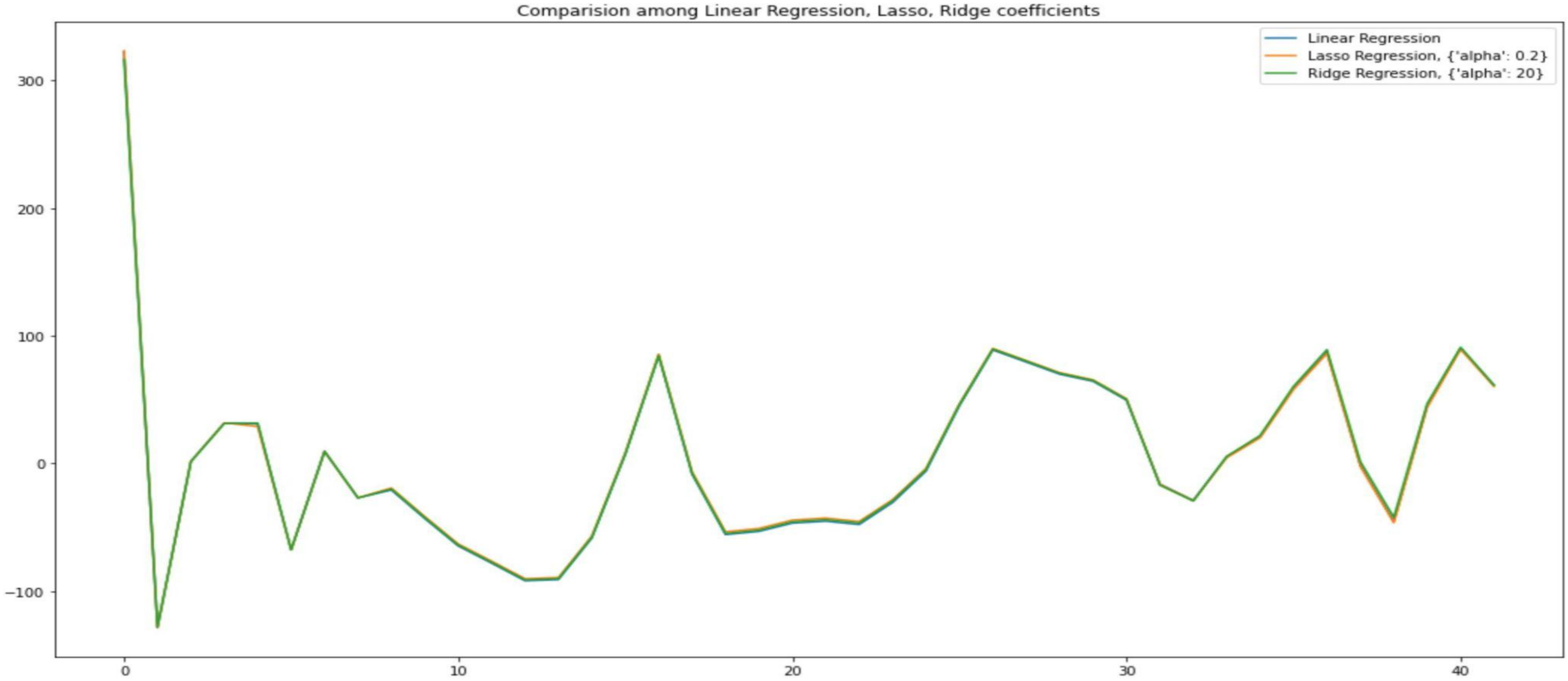
Ridge Regulation

1. The best fit alpha value is found out to be : {'alpha': 20}

2. Using {'alpha': 20} the negative mean squared error is: -116375.60205590996

Observation:

Similar performance as linear regression. It was quite expected because the linear regression didn't seem to be overfitted.



1. The linear Regression, Lasso, and Ridge coefficients are nearly identical.
2. No need in regularising the linear regression.

Decision Tree Regression

Min_samples_leaf = 21

Score = 0.758

Random Forest Regressor

Parameters used:

n_jobs=-1 -> All the cores are used.

max_depth = 50

n_estimators=1000

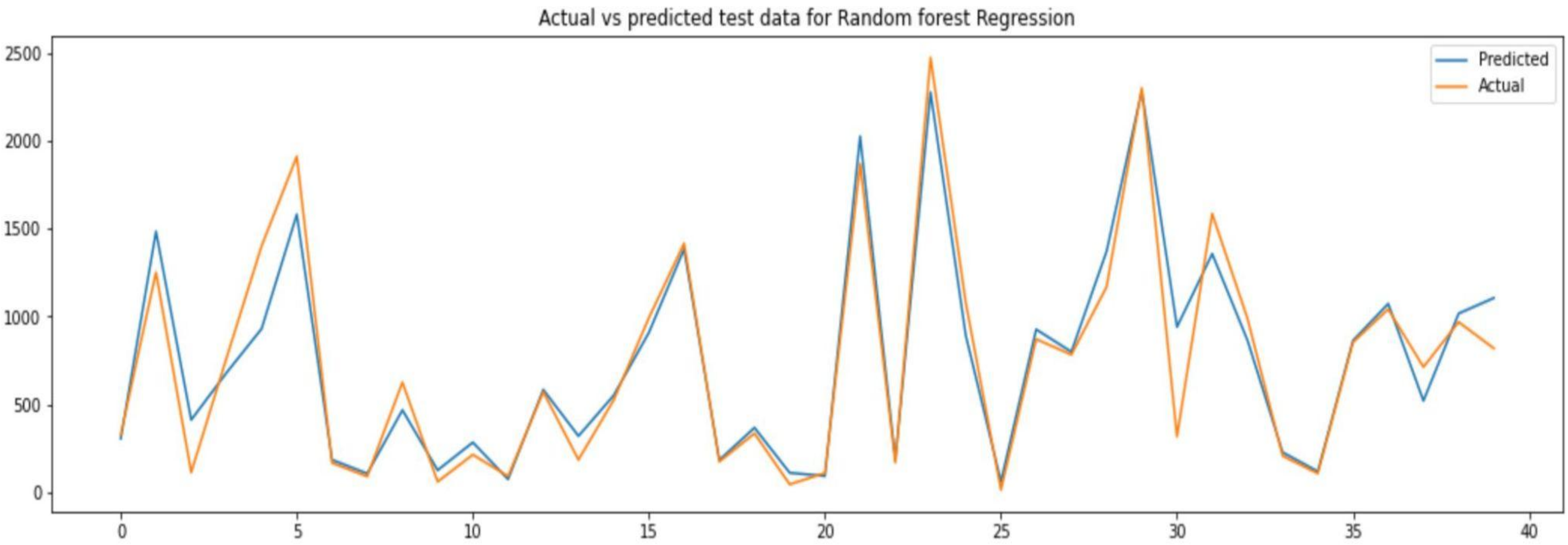
min_samples_leaf=1

min_samples_split = 0.002

These are the best parameters chosen for Random Forest Regression.

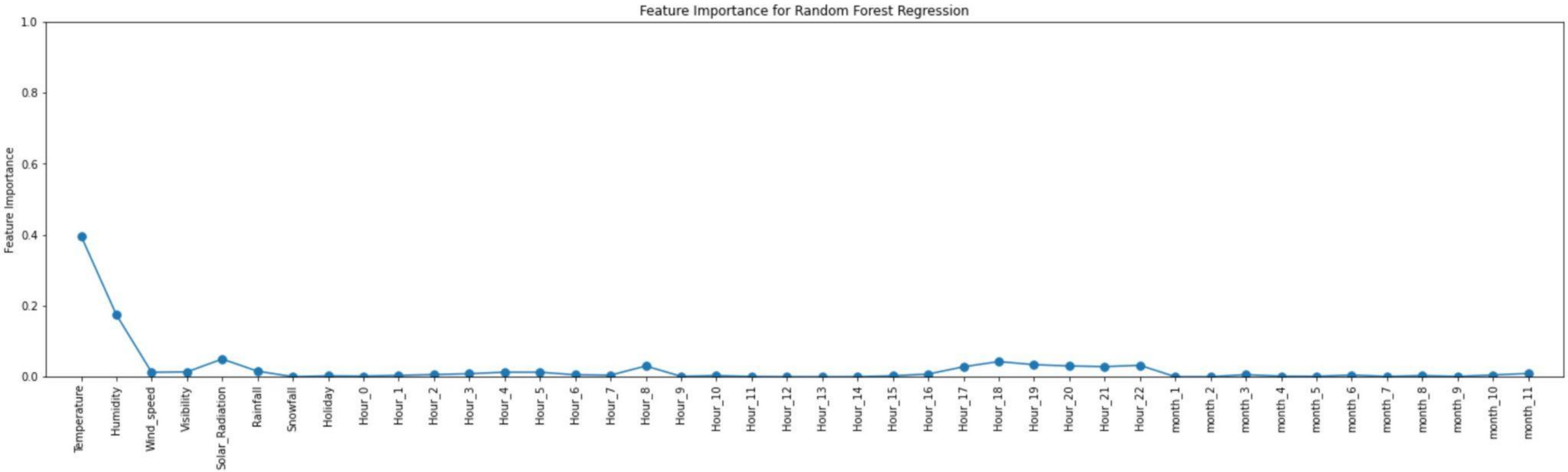
Random forest regression score : 0.848

Random Forest Regression cont..

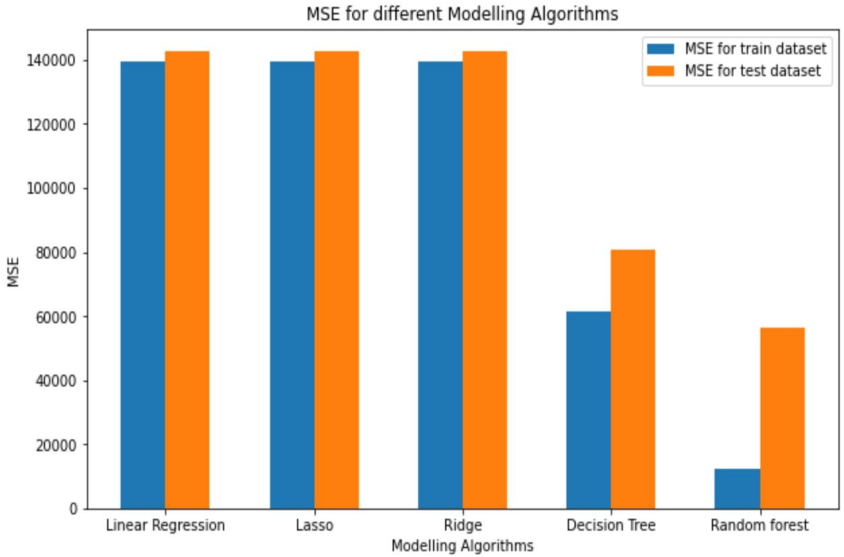


Random forest regression model has better fitting than the earlier models.

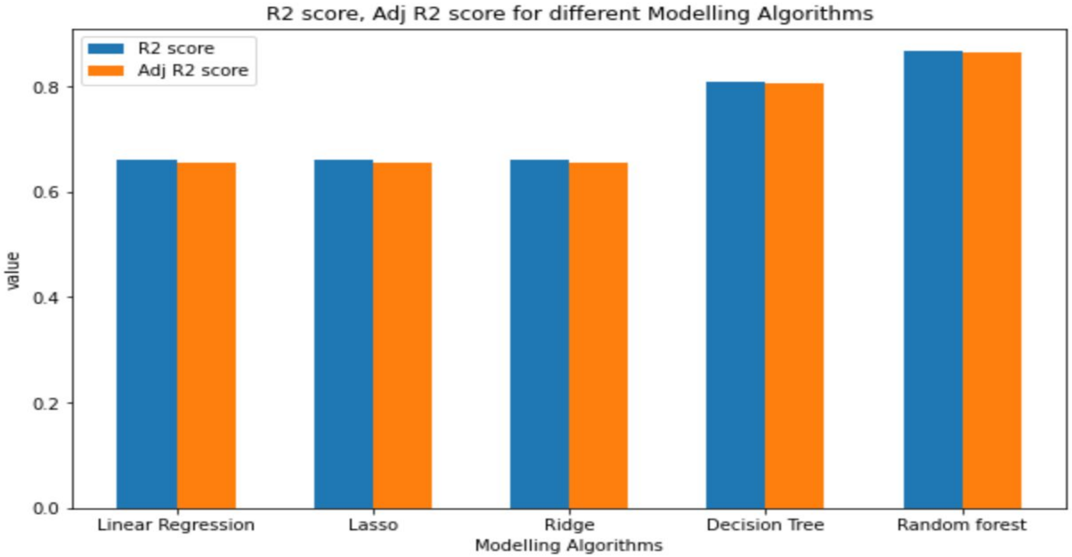
Feature importance for random forest regression



Comparison among various Metrics



The MSE and R2 score for Random forest regression is significantly improved.



Data Exploration Conclusion:

- Temperature: People generally prefer to bike at moderate to high temperatures. We see highest rental counts between 20 to 32 degree Celsius.
- Humidity: With increasing humidity, we see decrease in the number of bike rental count.
- Hour: Bike rental count is mostly correlated with the time of the day. As indicated above, the count reaches a high point during peak hours on a no holiday and is mostly uniform during the day on a non-holiday.
- Temperature, Windspeed, Visibility, Solar radiation: They have a positive correlation with bike rents.
- Rainfall, Snowfall: They have a negative correlation with bike rents.
- Seasons: We see highest number bike rentals in Summer and the lowest in Winter season.

Modeling Conclusions:

- We use 5 Regression Models to predict the hourly rented bike count - Linear Regression, Lasso, Ridge, Decision Tree, Random Forest.
- Among all the 5 models, Random Forest Model has the best metric analysis.
- Lasso or Ridge regularisation did not provide any improvement to the regular linear regression.