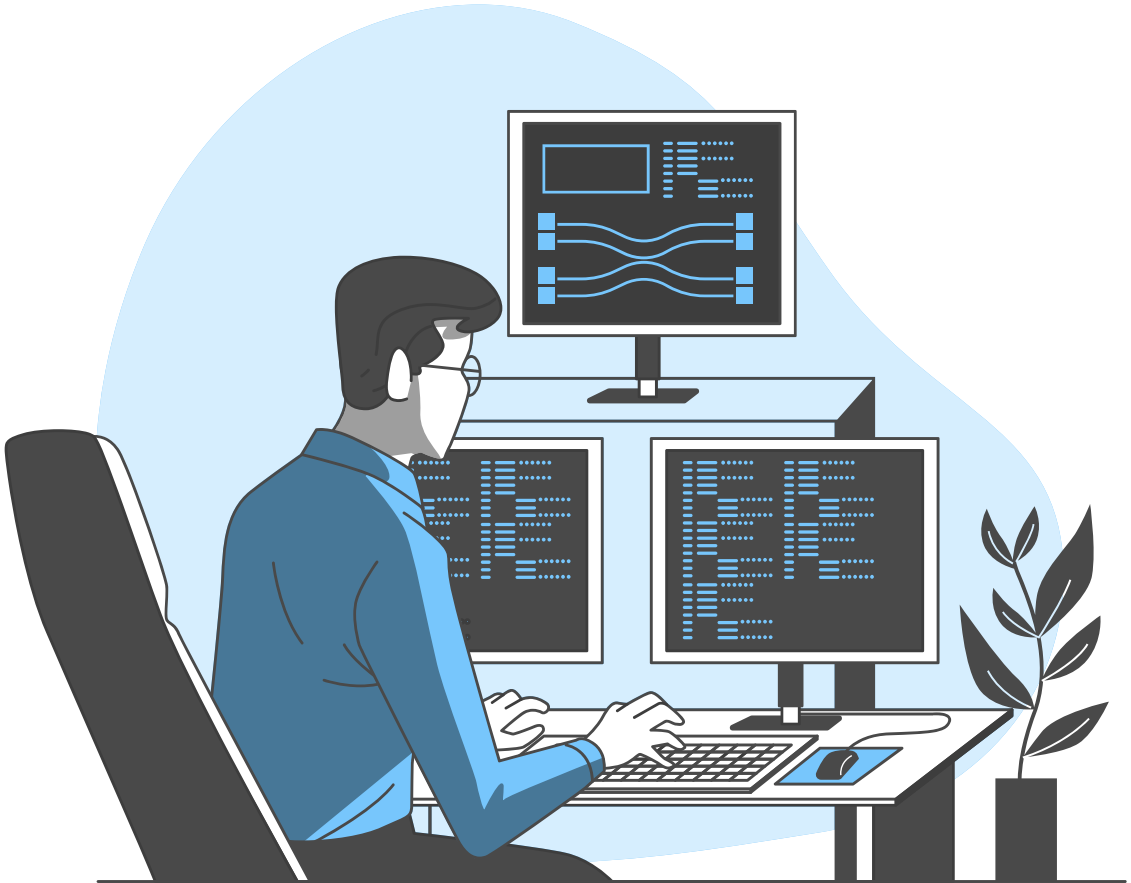
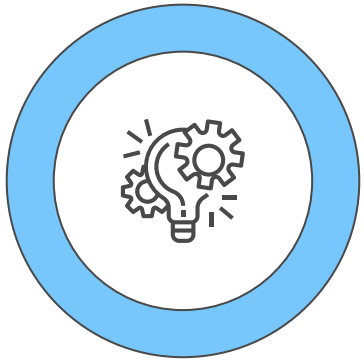


London Bike Sharing

Statical Data Analytics Course



Introduction

The London Bike Sharing Dataset is a comprehensive collection of data related to the bike-sharing system in London, United Kingdom. It provides valuable insights into the usage patterns, trends, and dynamics of the bike-sharing service, offering researchers, analysts, and enthusiasts an opportunity to explore and analyze various aspects of urban transportation.

Chapters

01

Data Understanding

We need to know what we have

02

Data Pre processing

Transformation, manipulation, reduce noise and make data suitable

03

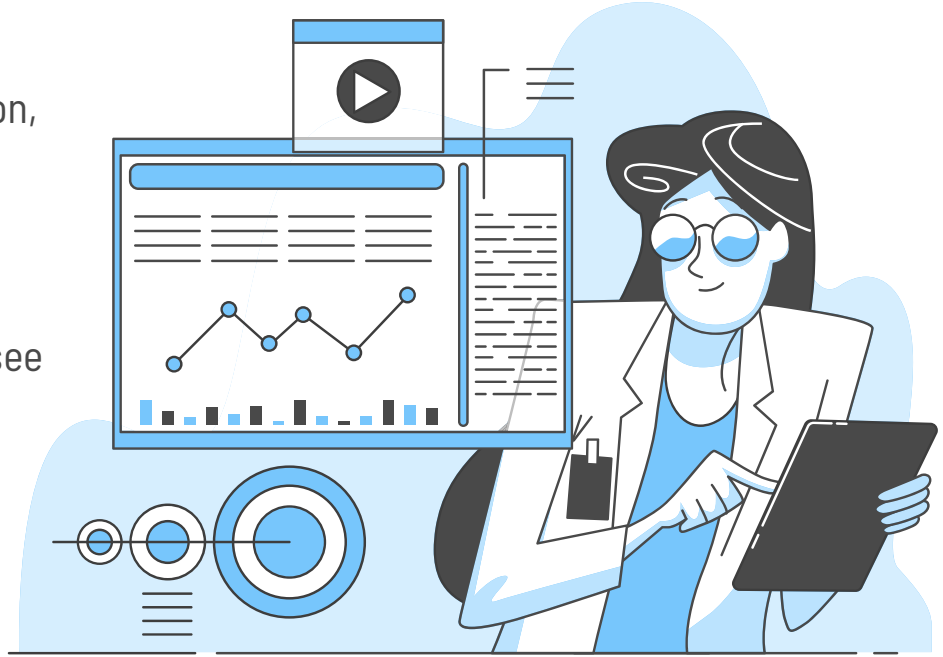
Modeling

Use statistical models and see the results

04

Conclusion

Key points and outcome



About Dataset

The dataset contains comprehensive information on bike trips, weather conditions, and corresponding time and location data.

The London Bike Sharing Dataset includes variables like trip start and end stations, duration, bike ID, user types, and timestamps.

Weather-related data, including temperature, humidity, wind speed, and conditions, are often included in the dataset.

Problem Statment

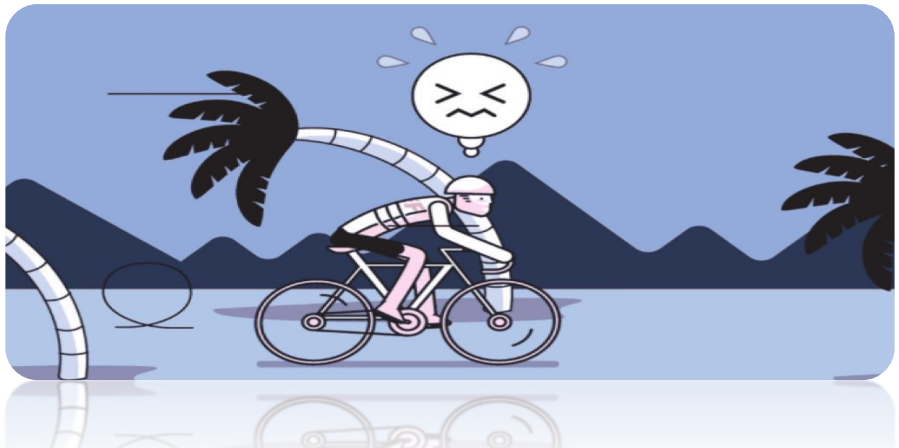
In this project, we are trying to predict the number of bicycles that
can be rented every hour of the day in London.

01
Data Understanding

We need to know what we have

Understanding Data

The dataset contains over 17,414 records spanning from 2015-01-04 to 2017-01-03. This data includes information about the weather conditions at the time of each ride, such as temperature, humidity, wind speed, and precipitation. Additionally, it includes information about public holidays and events that may have affected bike usage. The description of each variable in this data is provided as below:



Understanding Data

Meta Data

timestamp	Representing timestamp of bike share
cnt	Representing total number of bike shares
t1	The temperature in Celsius
t2	The apparent ("feels-like") temperature in Celsius
hum	The humidity level
wind_speed	The wind speed
weather_code	A categorical value indicating the weather situation (1:clear, 2:mist/cloud, 3:light rain/snow, 4:heavy rain/hail/snow/fog)
is_holiday	binary value indicating whether or not the day is a holiday
is_weekend	A binary value indicating whether or not the day is a weekend
season	A numerically encoded value indicating the season (0:spring, 1:summer, 2:fall, 3:winter)

02

Data Pre Processing

Transformation, manipulation, reduce noise and make data suitable

Exploratory Data Analysis

01 Data cleaning

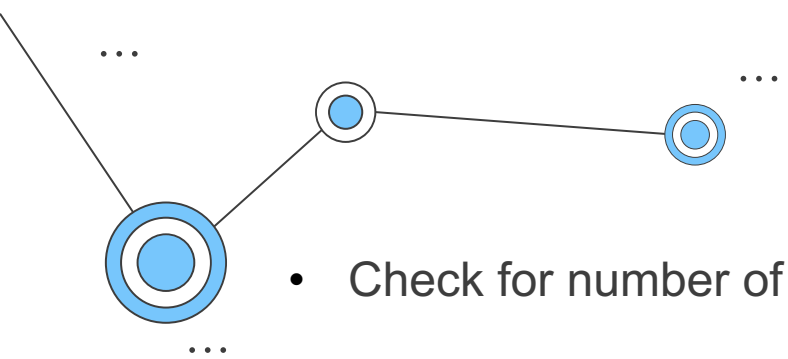
02 Exclude Unused Variables

03 Check missing value

04 Data Visualization

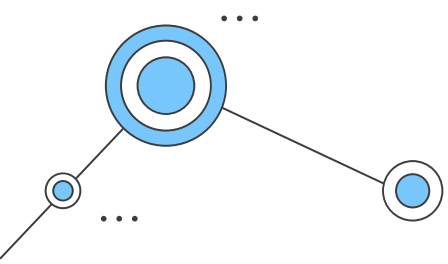
- Extract year, month, day and hour variable from the timestamp column

- And exclude timestamp column from data set



- Check for number of missing value

As we can see in the above table, there are no missing values in the data.



Variable	Number of Missing
timestamp	0
cnt	0
t1	0
t2	0
hum	0
wind_speed	0
weather_code	0
is_holiday	0
is_weekend	0
season	0
year	0
month	0
day	0
hour	0

Take a look to Dataset again

Head of Dataset

cnt	t1	t2	hum	wind_speed	weather_code	is_holiday	is_weekend	season	year	month	day	hour
182	3.0	2.0	93.0	6.0	3	0	1	3	2015	1	4	0
138	3.0	2.5	93.0	5.0	1	0	1	3	2015	1	4	1
134	2.5	2.5	96.5	0.0	1	0	1	3	2015	1	4	2
72	2.0	2.0	100.0	0.0	1	0	1	3	2015	1	4	3
47	2.0	0.0	93.0	6.5	1	0	1	3	2015	1	4	4
46	2.0	2.0	93.0	4.0	1	0	1	3	2015	1	4	5
51	1.0	-1.0	100.0	7.0	4	0	1	3	2015	1	4	6
75	1.0	-1.0	100.0	7.0	4	0	1	3	2015	1	4	7
131	1.5	-1.0	96.5	8.0	4	0	1	3	2015	1	4	8
301	2.0	-0.5	100.0	9.0	3	0	1	3	2015	1	4	9
528	3.0	-0.5	93.0	12.0	3	0	1	3	2015	1	4	10

Data Summarization

Data summary	
Name	**bike**
Number of rows	**17414**
Number of columns	**13**
Column type frequency:	
factor	**8**
numeric	**5**
Group variables	**None**

Factor variables

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
weather_code	0	1	FALSE	4	1: 6150, 2: 4034, 4: 3679, 3: 3551
is_holiday	0	1	FALSE	2	0: 17030, 1: 384
is_weekend	0	1	FALSE	2	0: 12444, 1: 4970
season	0	1	FALSE	4	0: 4394, 1: 4387, 3: 4330, 2: 4303
year	0	1	FALSE	3	201: 8699, 201: 8643, 201: 72
month	0	1	FALSE	12	5: 1488, 1: 1487, 8: 1484, 12: 1484
day	0	1	FALSE	31	6: 576, 14: 576, 21: 576, 22: 576
hour	0	1	FALSE	24	16: 730, 12: 729, 15: 729, 13: 728

Numeric variables

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
cnt	0	1	1143.10	1085.11	0.0	257	844.0	1671.75	7860.0	
t1	0	1	12.47	5.57	-1.5	8	12.5	16.00	34.0	
t2	0	1	11.52	6.62	-6.0	6	12.5	16.00	34.0	
hum	0	1	72.32	14.31	20.5	63	74.5	83.00	100.0	
wind_speed	0	1	15.91	7.89	0.0	10	15.0	20.50	56.5	

Dataset Summary

cnt = Count	
Min.	0.00
1st Qu.	260
Median	846
Mean	1146
3rd Qu.	1676
Max.	7860

t1 = Temperature 1	
Min.	-1.5
1st Qu.	8.5
Median	12.5
Mean	12.5
3rd Qu.	16.0
Max.	34.0

t2 = feels like temperature	
Min.	-6.00
1st Qu.	6.00
Median	12.50
Mean	11.56
3rd Qu.	16.00
Max.	34.00

hum = humidity level	
Min.	20.50
1st Qu.	63.00
Median	74.50
Mean	72.28
3rd Qu.	83.00
Max.	100.00

wind_speed = wind speed	
Min.	0.00
1st Qu.	10.00
Median	15.00
Mean	15.92
3rd Qu.	20.50
Max.	56.50

weather_code	
Min.	1.000
1st Qu.	1.000
Median	2.000
Mean	2.721
3rd Qu.	3.000
Max.	26.000

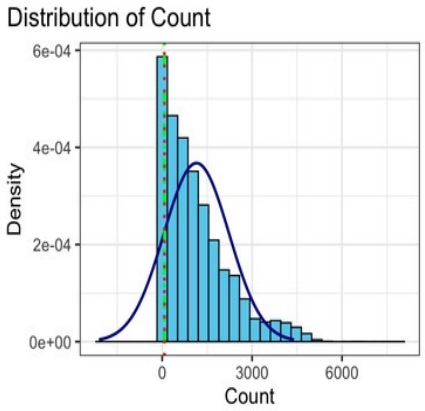
is_holiday	
Min.	0.00000
1st Qu.	0.00000
Median	0.00000
Mean	0.02076
3rd Qu.	0.00000
Max.	1.00000

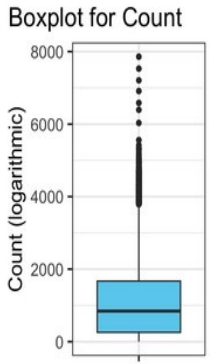
is_weekend	
Min.	0.0000
1st Qu.	0.0000
Median	0.0000
Mean	0.2852
3rd Qu.	1.0000
Max.	1.0000

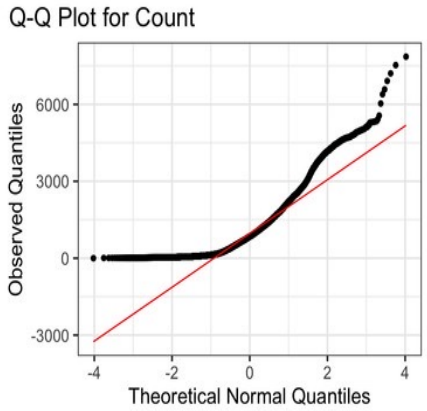
season	
Min.	0.000
1st Qu.	0.000
Median	1.000
Mean	1.486
3rd Qu.	2.000
Max.	3.000

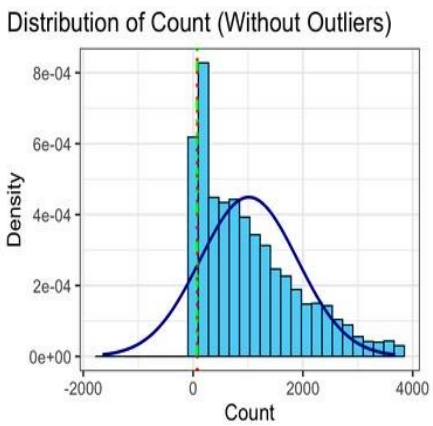
year	
Min.	2015
1st Qu.	2015
Median	2016
Mean	2016
3rd Qu.	2016
Max.	2017

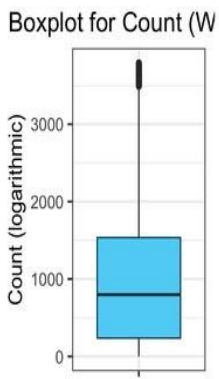
Data visualization

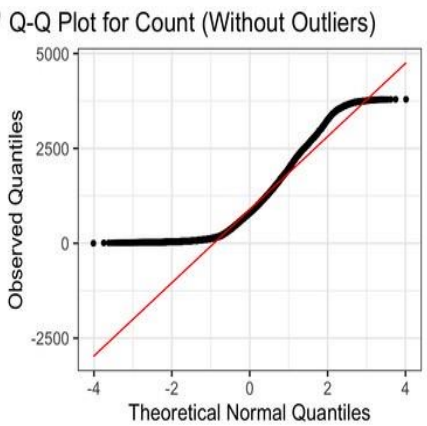

Distribution of Count
Density
6e-04
4e-04
2e-04
0e+00
0
3000
6000
Count

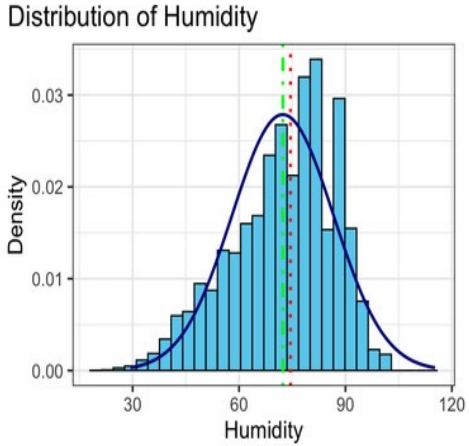

Boxplot for Count
Count (logarithmic)
8000
6000
4000
2000
0

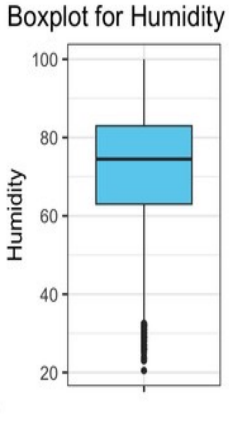

Q-Q Plot for Count
Observed Quantiles
6000
3000
0
-3000
-4
-2
0
2
4
Theoretical Normal Quantiles

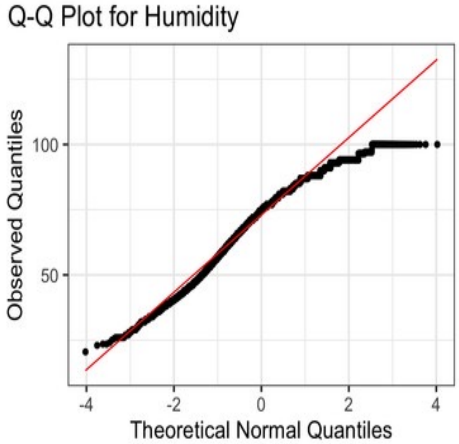

Distribution of Count (Without Outliers)
Density
8e-04
6e-04
4e-04
2e-04
0e+00
-2000
0
2000
4000
Count

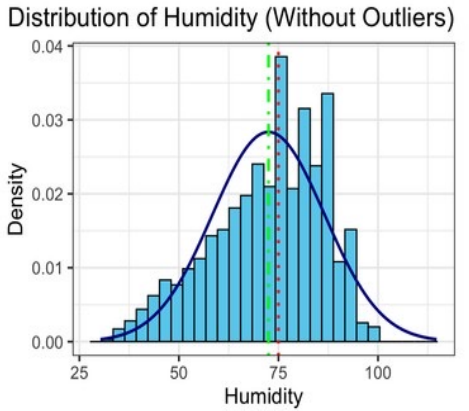

Boxplot for Count (W
Count (logarithmic)
3000
2000
1000
0

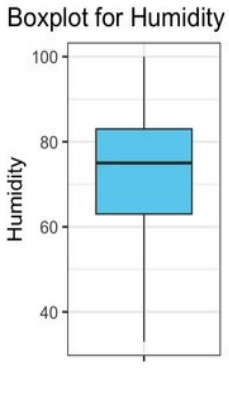

Q-Q Plot for Count (Without Outliers)
Observed Quantiles
5000
2500
0
-2500
-4
-2
0
2
4
Theoretical Normal Quantiles

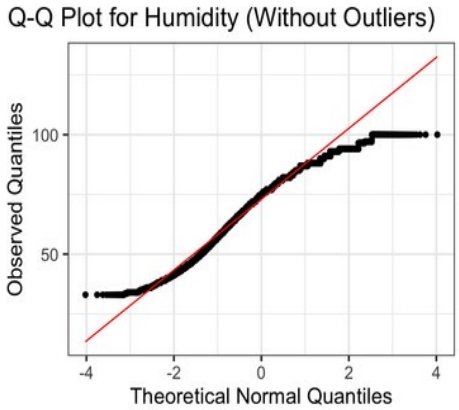

Distribution of Humidity
Density
0.03
0.02
0.01
0.00
30
60
90
120
Humidity

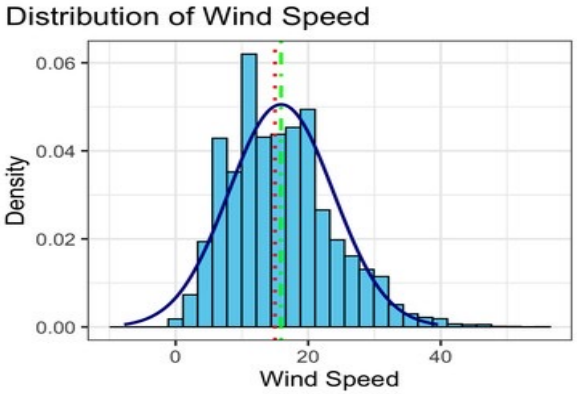

Boxplot for Humidity
Humidity
100
80
60
40
20

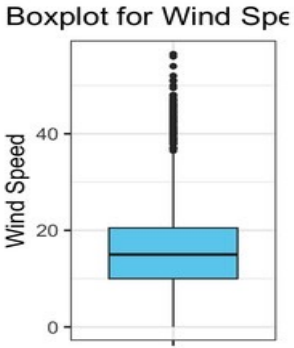

Q-Q Plot for Humidity
Observed Quantiles
100
50
-4
-2
0
2
4
Theoretical Normal Quantiles

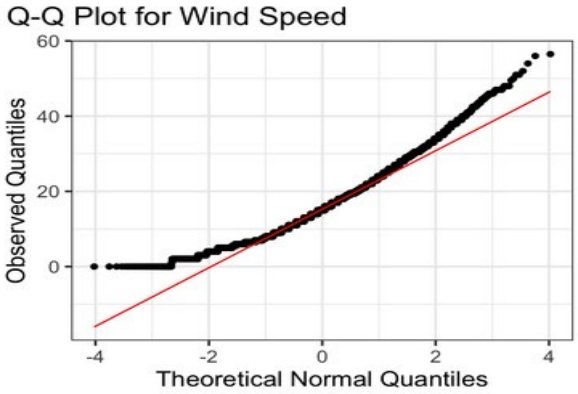

Distribution of Humidity (Without Outliers)
Density
0.04
0.03
0.02
0.01
0.00
25
50
75
100
Humidity

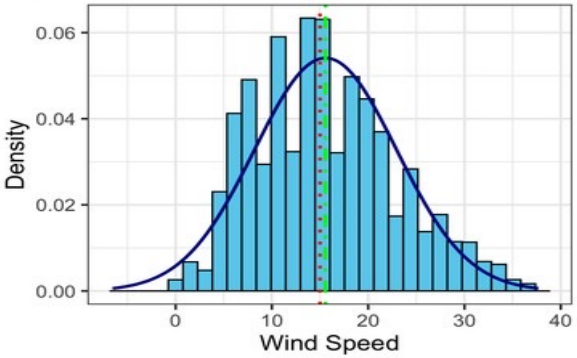

Boxplot for Humidity
Humidity
100
80
60
40

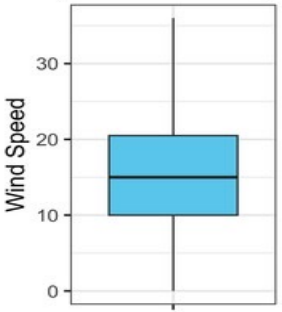

Q-Q Plot for Humidity (Without Outliers)
Observed Quantiles
100
50
-4
-2
0
2
4
Theoretical Normal Quantiles

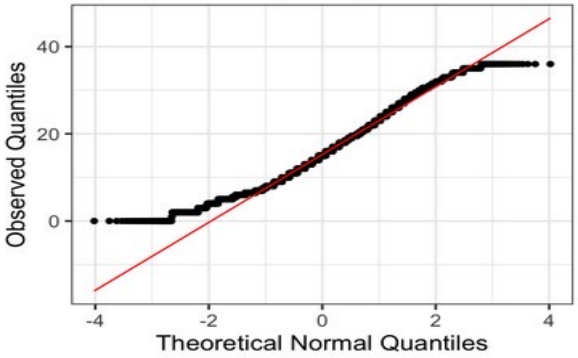

Distribution of Wind Speed
Density
0.06
0.04
0.02
0.00
0
20
40
Wind Speed

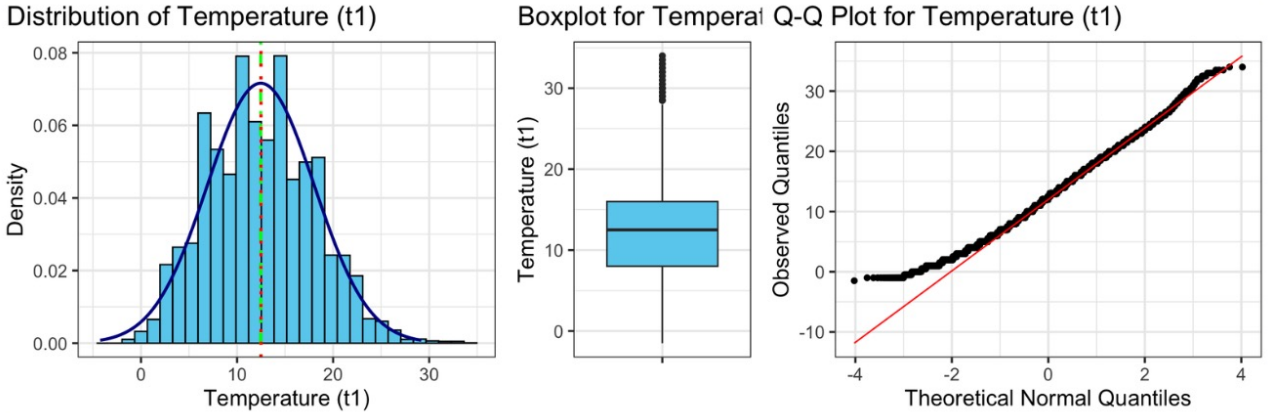

Boxplot for Wind Spe
Wind Speed
40
20
0

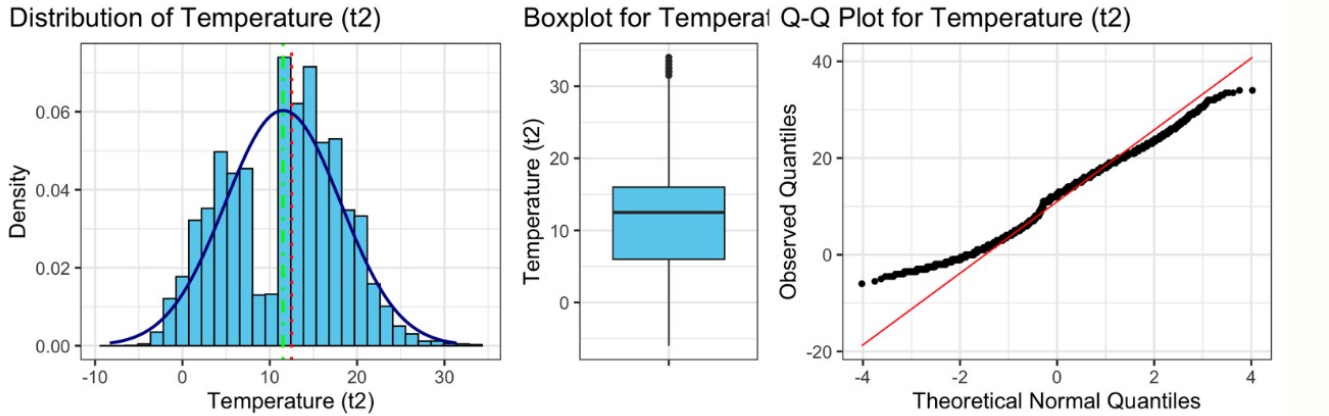

Q-Q Plot for Wind Speed
Observed Quantiles
60
40
20
0
-4
-2
0
2
4
Theoretical Normal Quantiles


Distribution of Wind Speed (Without Outlie
Density
0.06
0.04
0.02
0.00
0
10
20
30
40
Wind Speed

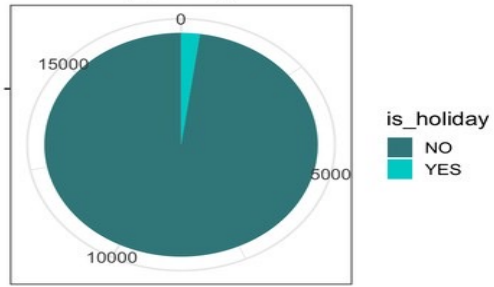

Boxplot for Wind Spe
Wind Speed
30
20
10
0


Q-Q Plot for Wind Speed (Without Outliers
Observed Quantiles
40
20
0
-4
-2
0
2
4
Theoretical Normal Quantiles

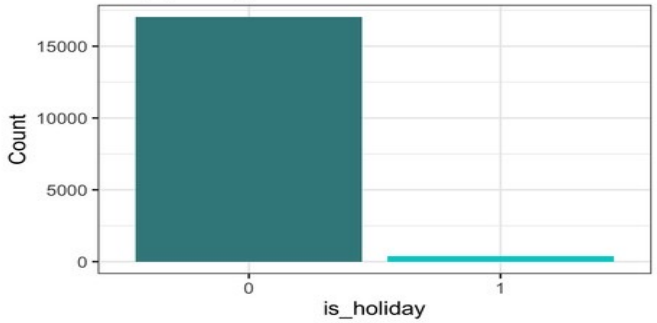

Distribution of Temperature (t1)
Density
Temperature (t1)
Boxplot for Temperat
Temperature (t1)
Q-Q Plot for Temperature (t1)
Observed Quantiles
Theoretical Normal Quantiles


Distribution of Temperature (t2)
Density
Temperature (t2)
Boxplot for Temperat
Temperature (t2)
Q-Q Plot for Temperature (t2)
Observed Quantiles
Theoretical Normal Quantiles

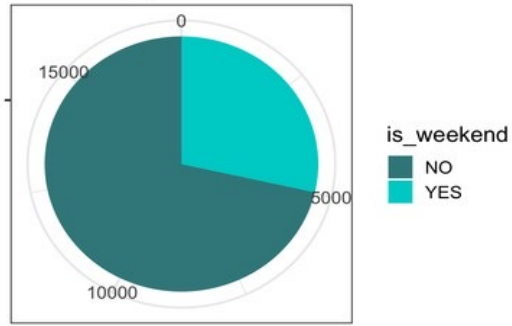
Pie Chart - is_holiday



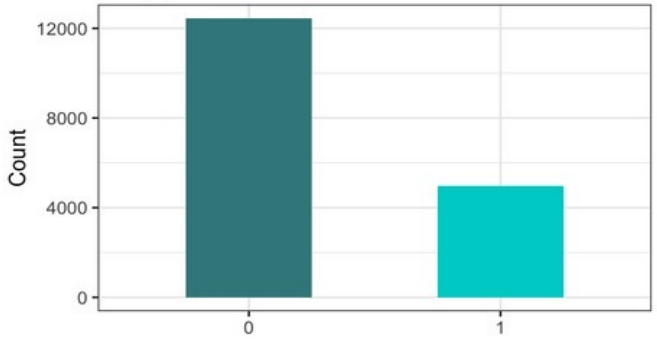
Bar Plot - is_holiday

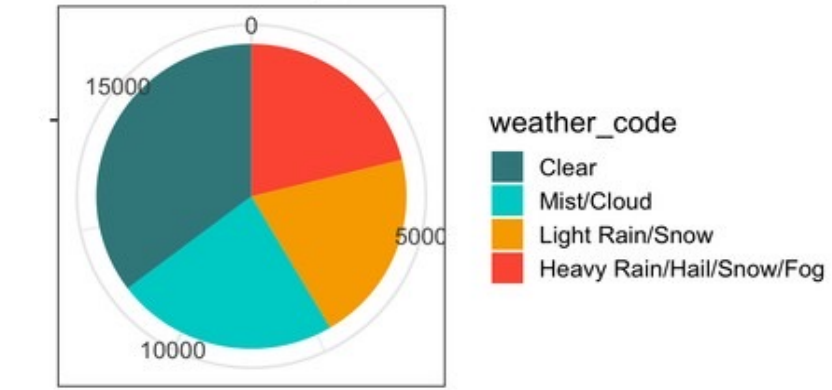


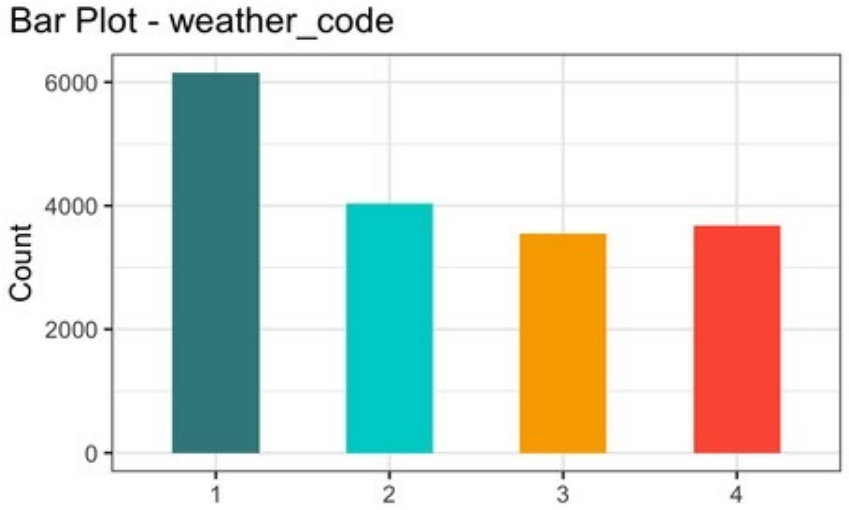
Pie Chart - is_weekend

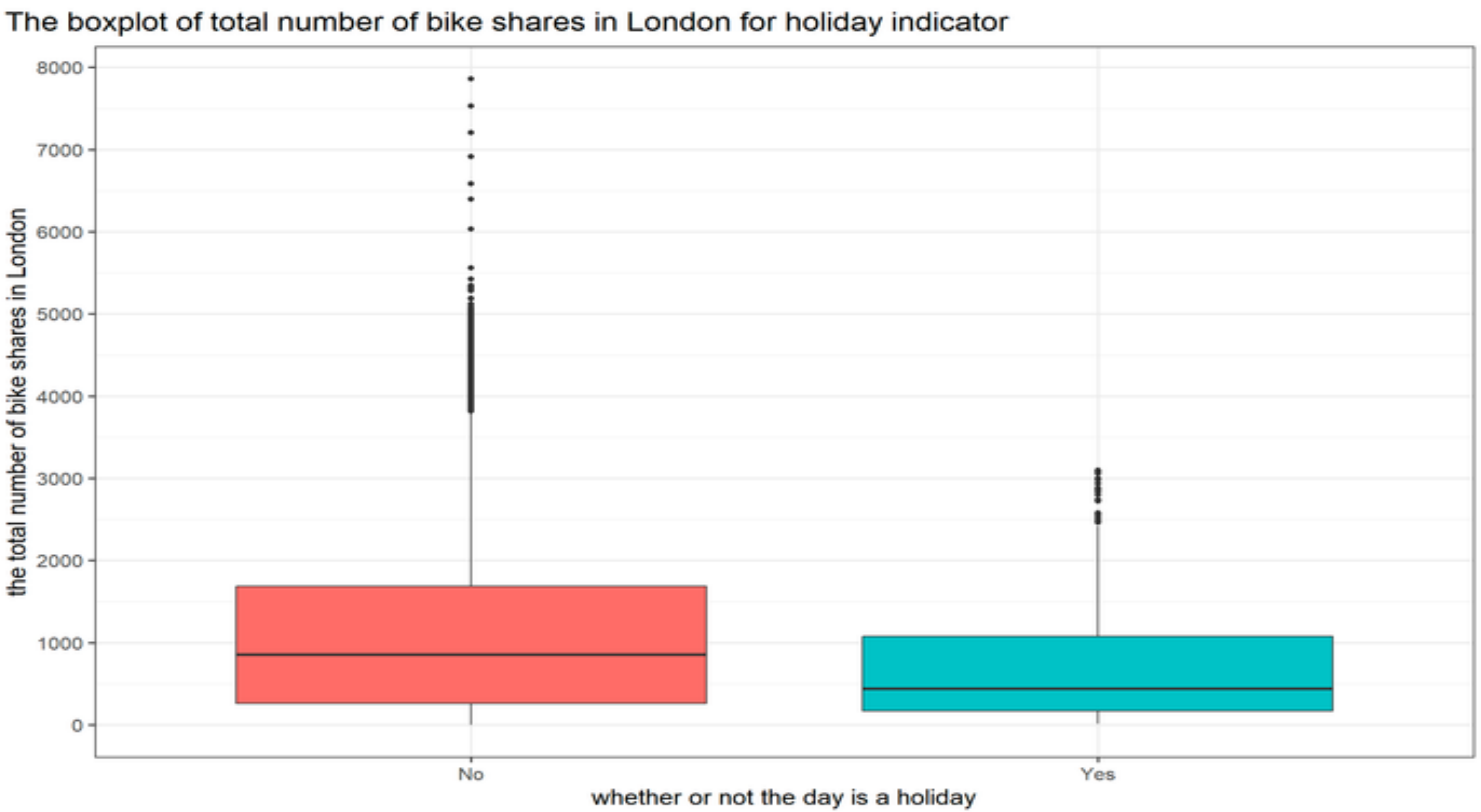


Bar Plot - is_weekend



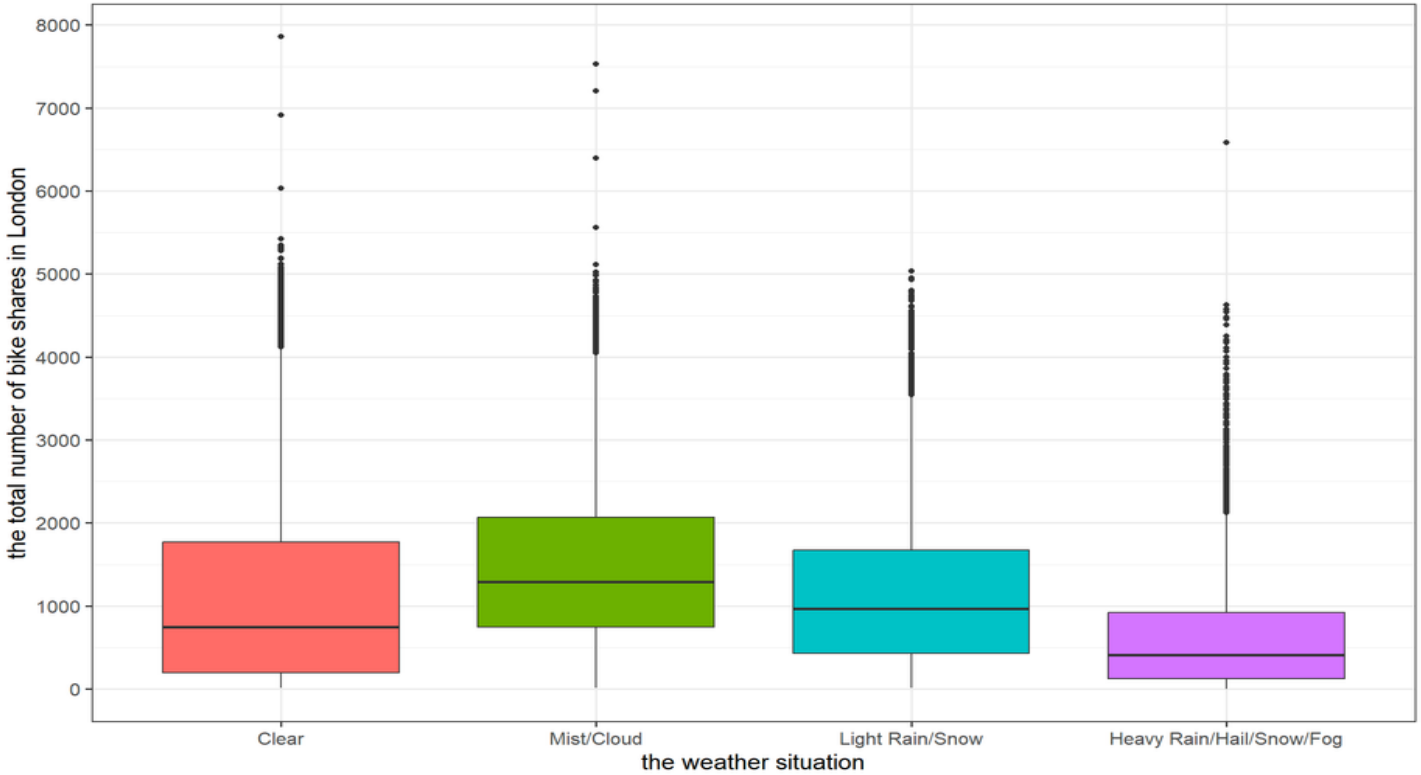

Pie Chart - weather_code
0
15000
5000
10000
weather_code
Clear
Mist/Cloud
Light Rain/Snow
Heavy Rain/Hail/Snow/Fog


Bar Plot - weather_code
Count
6000
4000
2000
0
1
2
3
4

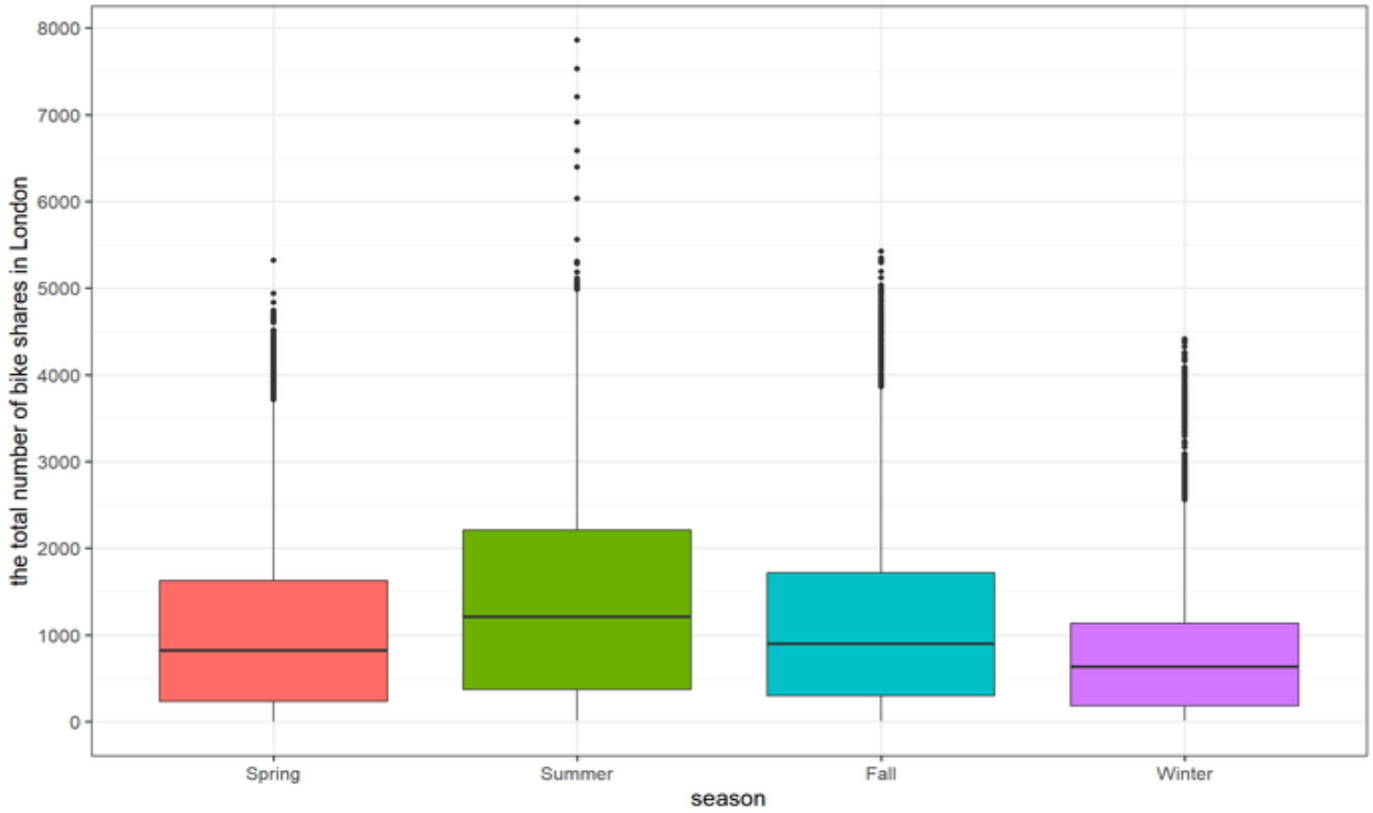

The boxplot of total number of bike shares in London for holiday indicator
the total number of bike shares in London
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
whether or not the day is a holiday


The boxplot of total number of bike shares in London for weekend indicator
the total number of bike shares in London
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
whether or not the day is a weekend

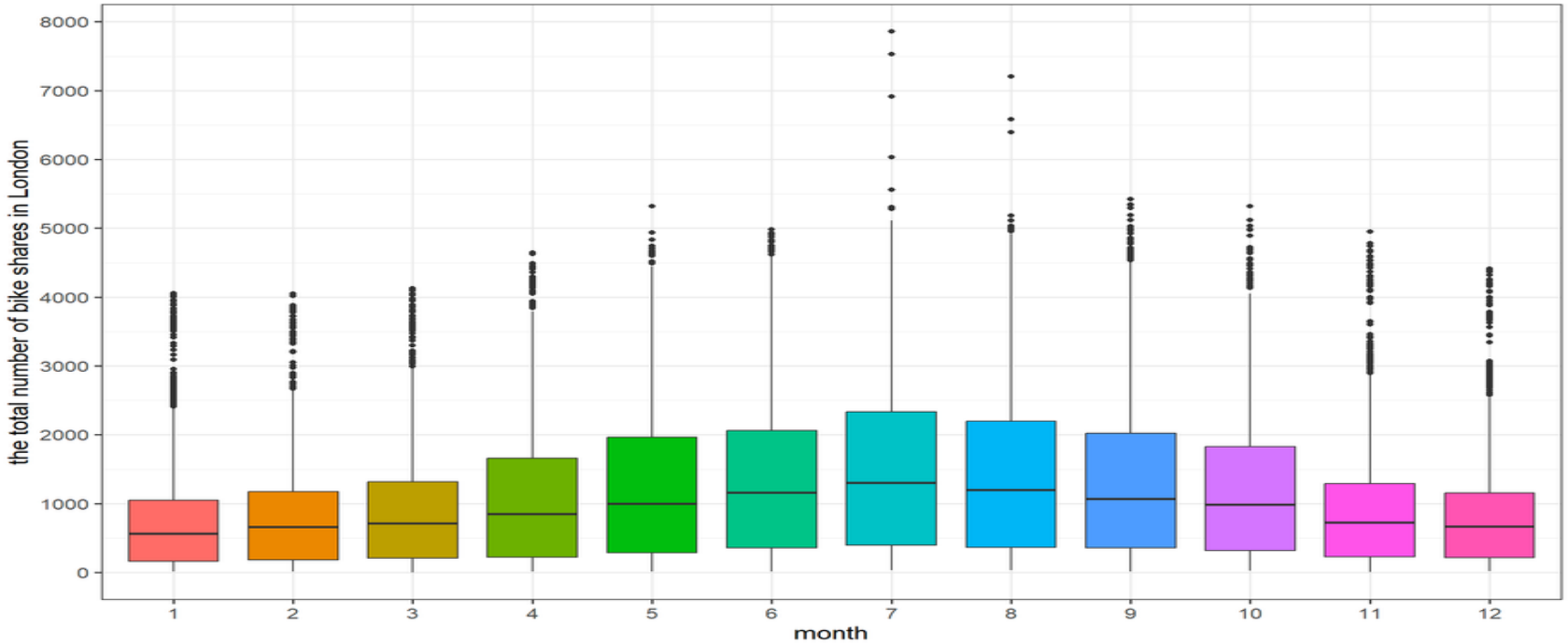
The boxplot of total number of bike shares in London for different weather situations

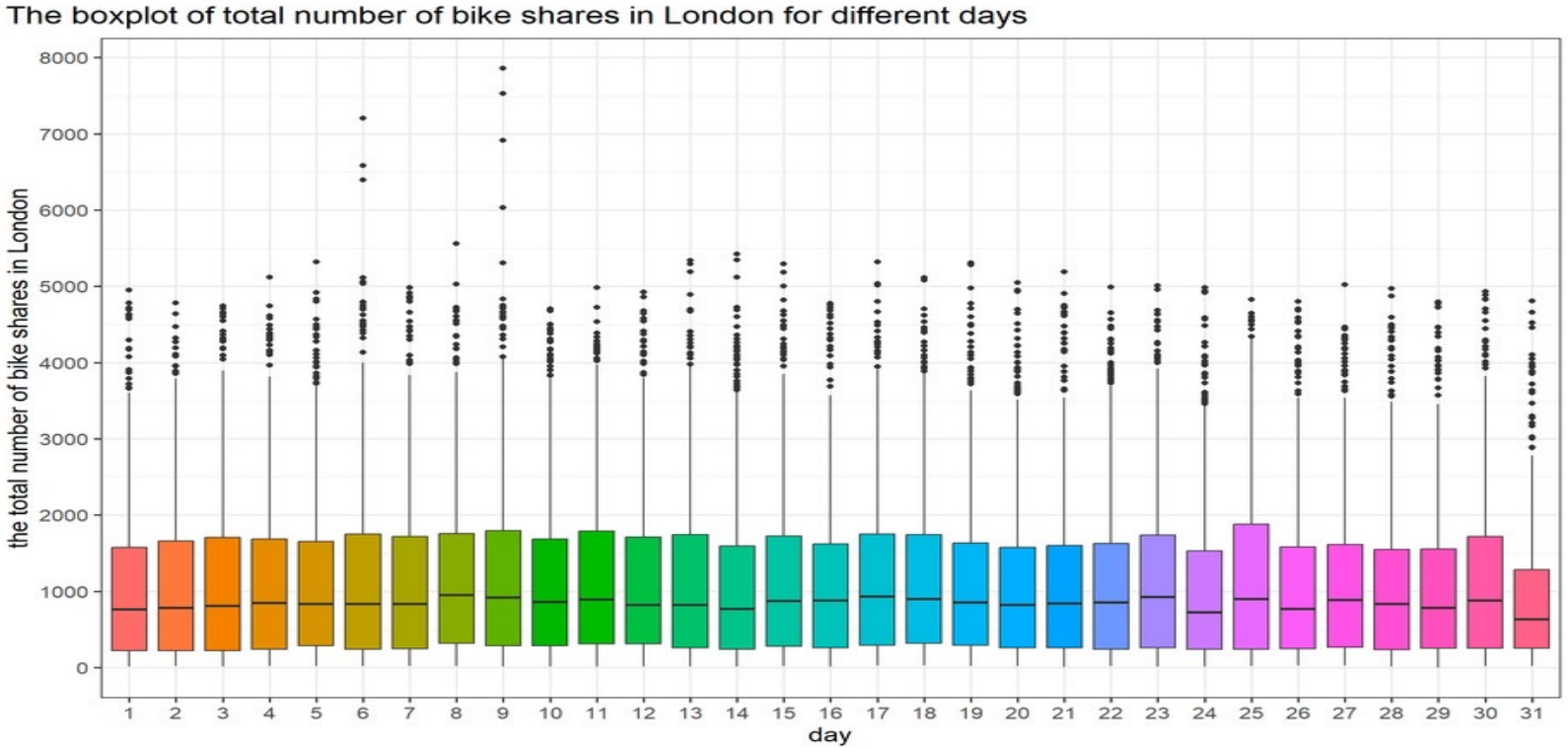


The boxplot of total number of bike shares in London for different seasons

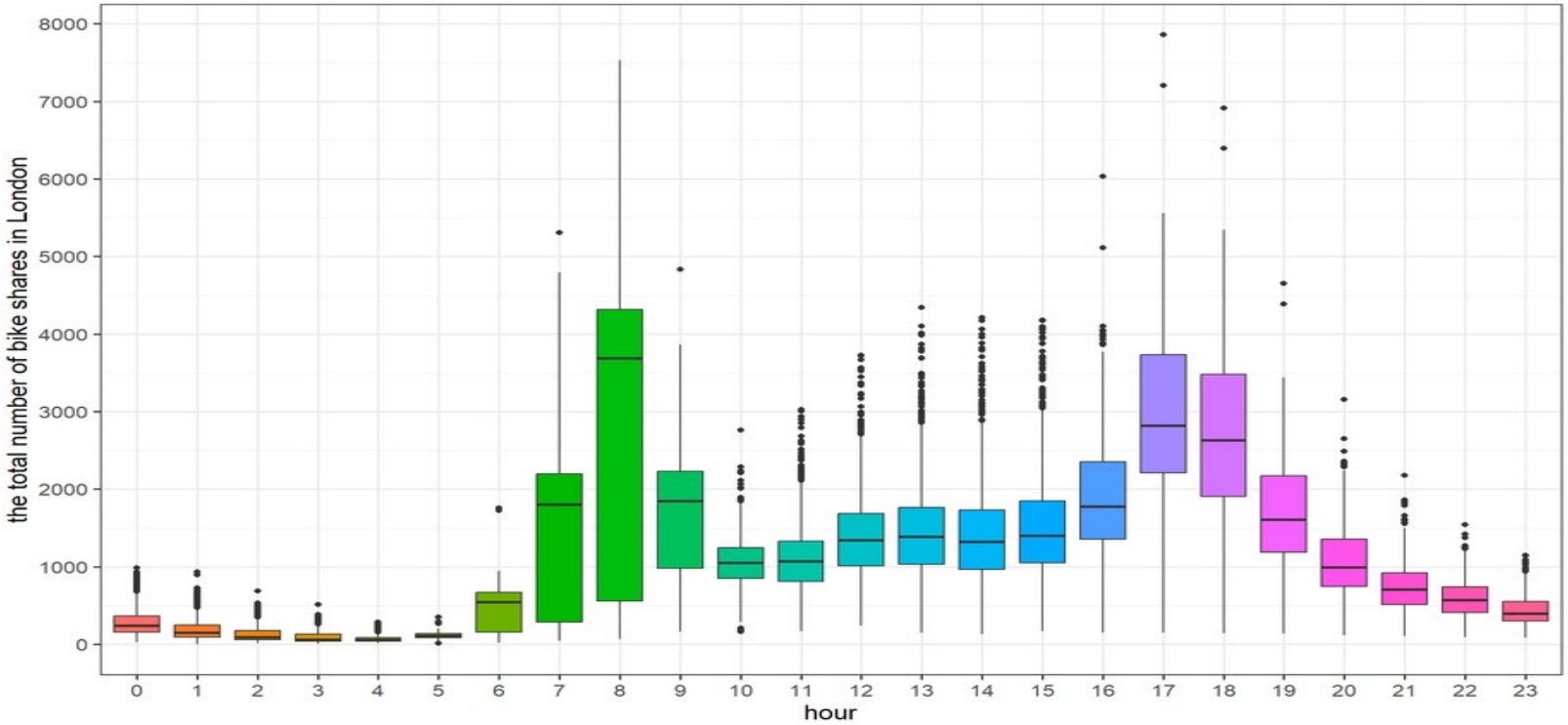


The boxplot of total number of bike shares in London for different months

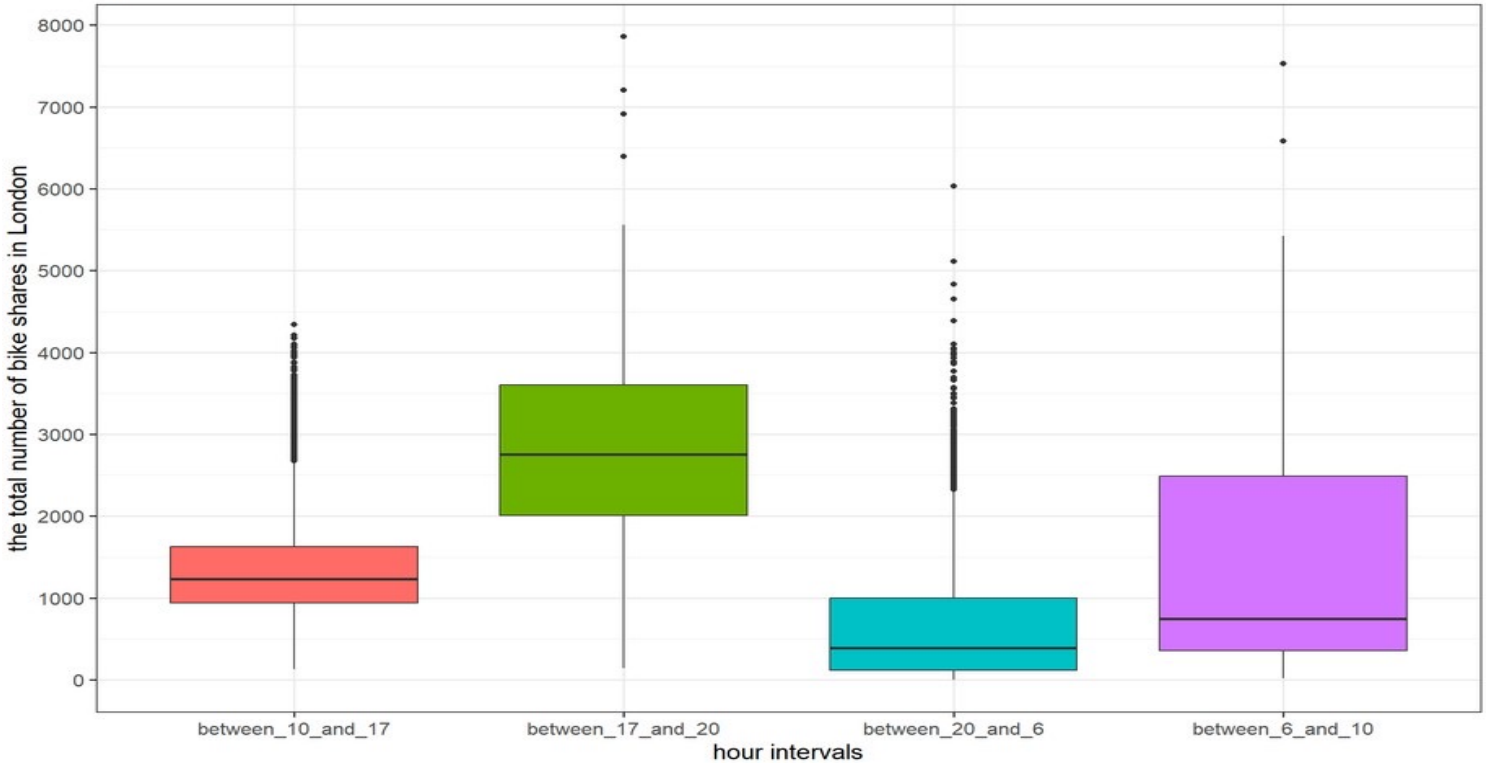



The boxplot of total number of bike shares in London for different days
the total number of bike shares in London
8000
7000
6000
5000
4000
3000
2000
1000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
day

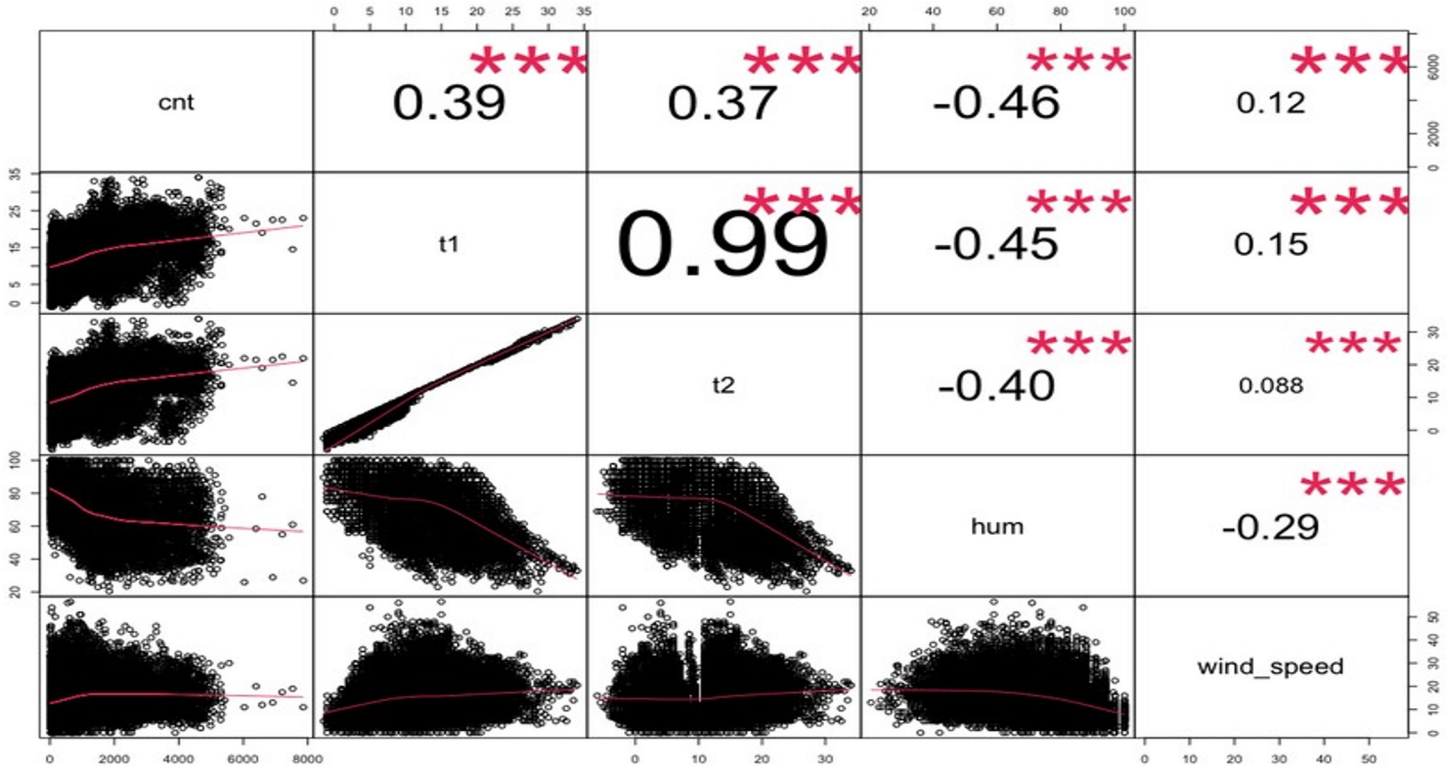
The boxplot of total number of bike shares in London for different hours

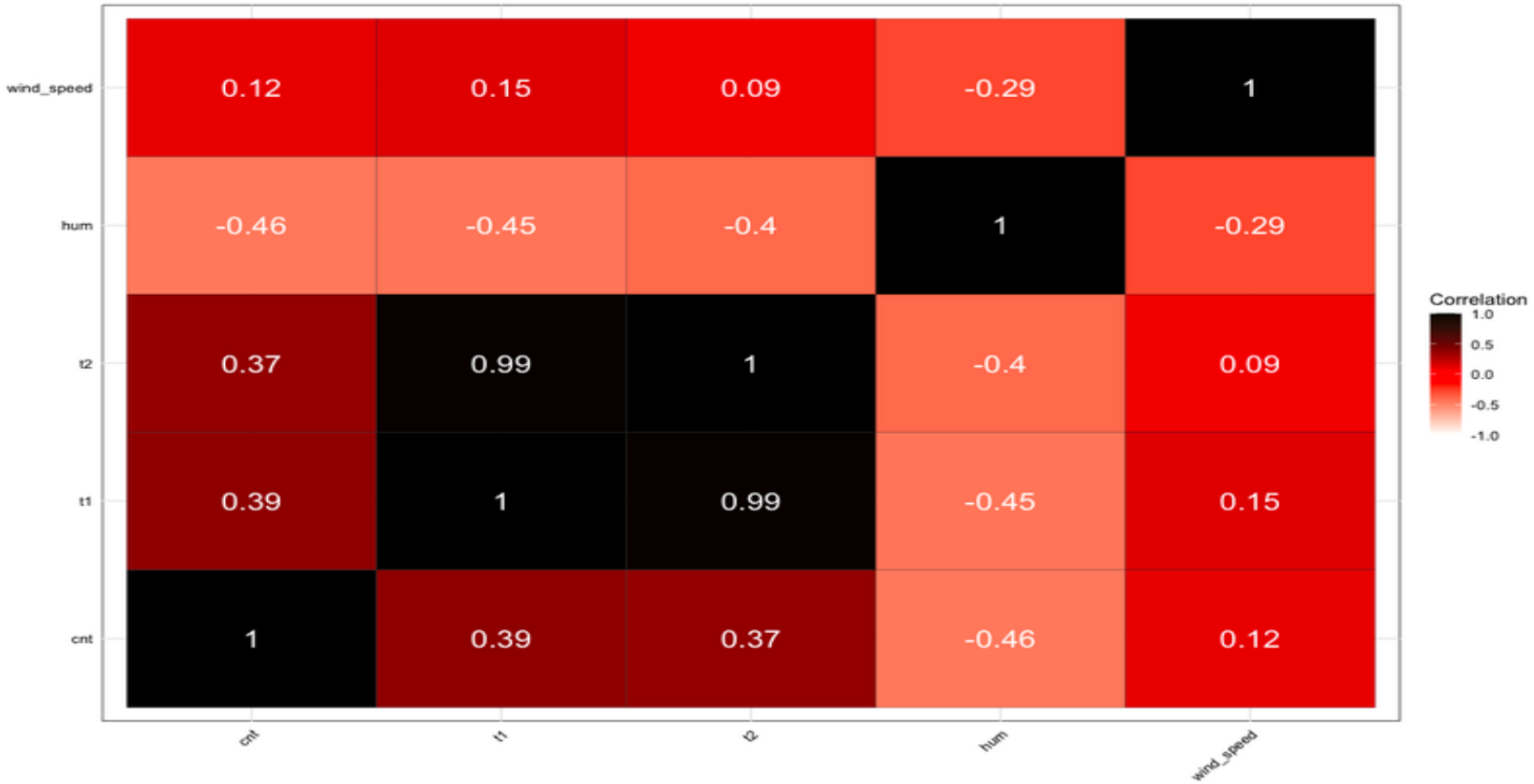


The boxplot of total number of bike shares in London for different hour intervals



Bivariate Data, Scatter Plots and Corr values

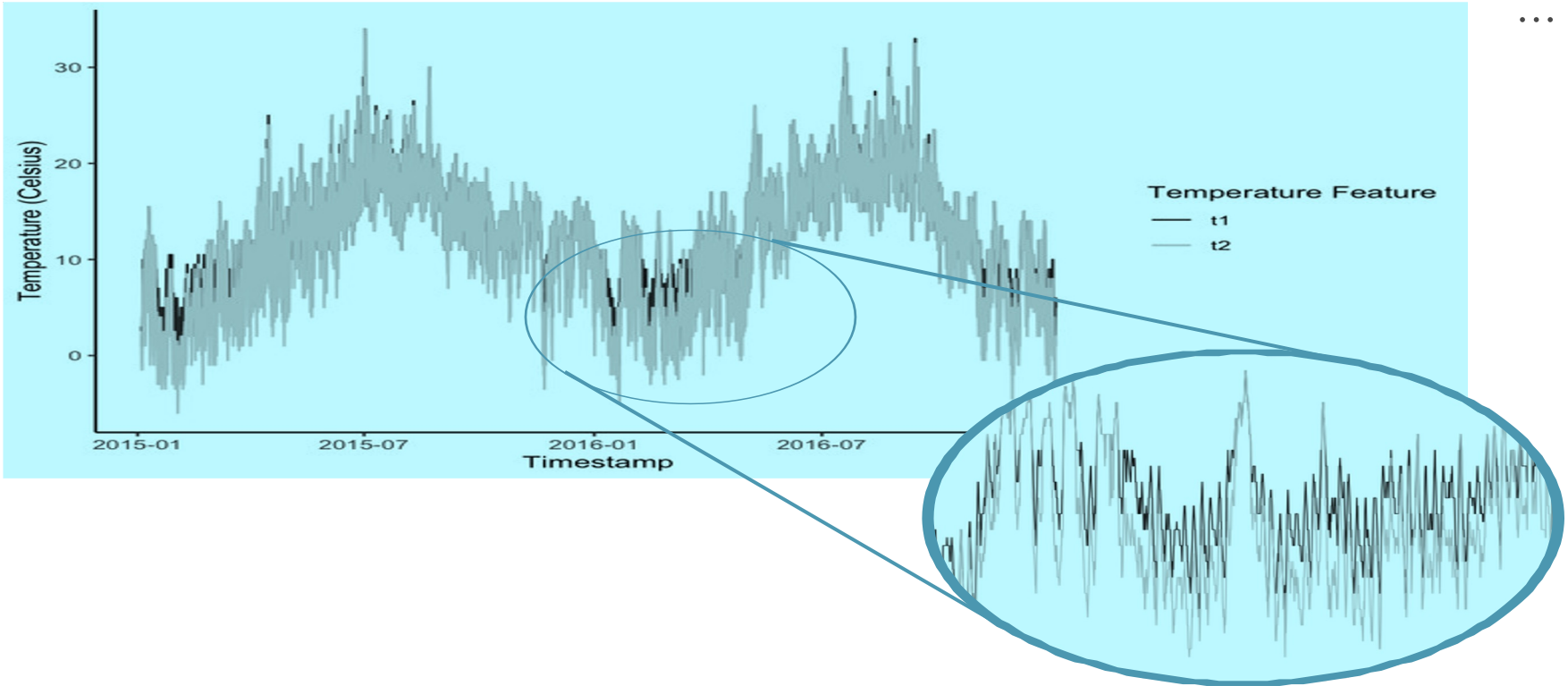



wind_speed
hum
t2
t1
cnt
0.12
0.15
0.09
-0.29
1
-0.46
-0.45
-0.4
1
-0.29
0.37
0.99
1
-0.4
0.09
0.39
1
0.99
-0.45
0.15
1
0.39
0.37
-0.46
0.12
cnt
t1
t2
hum
wind_speed
Correlation
1.0
0.5
0.0
-0.5
-1.0

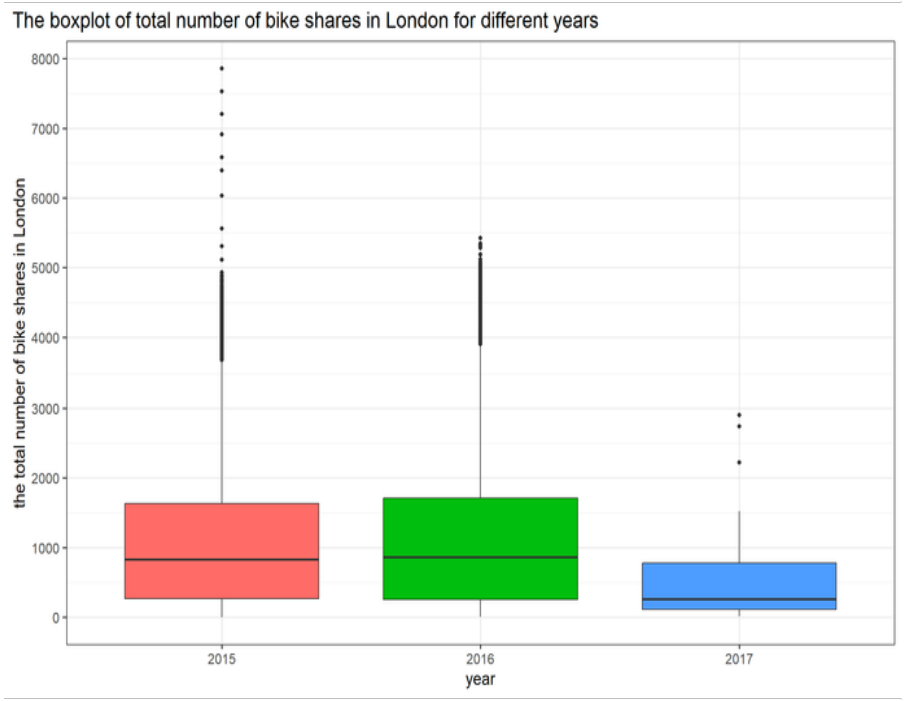
Dropping t2 variable

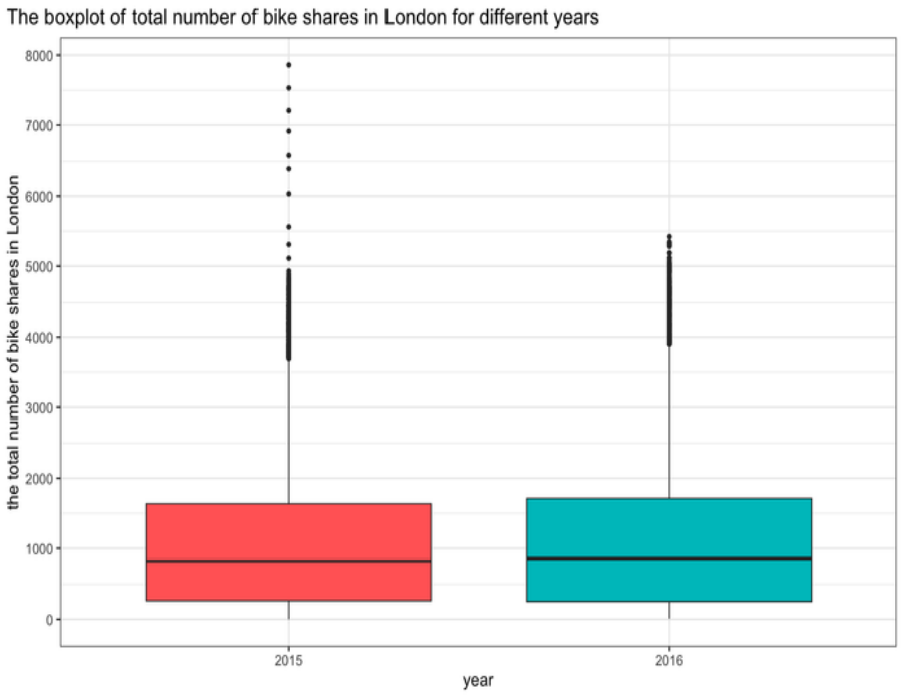
"t1' - the real temperature in C

"t2" - the temperature in C "feels like"



Remove '2017' year from our Timeframe

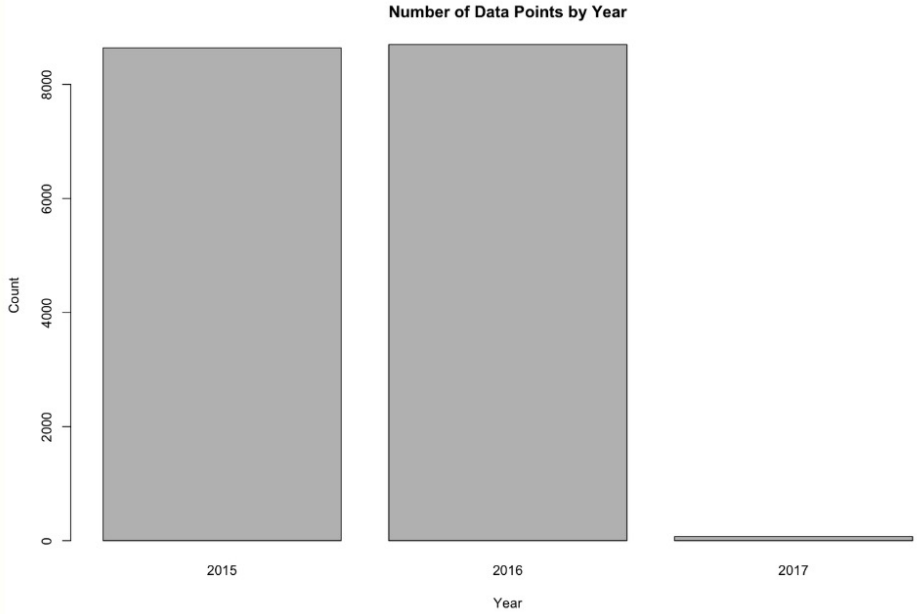




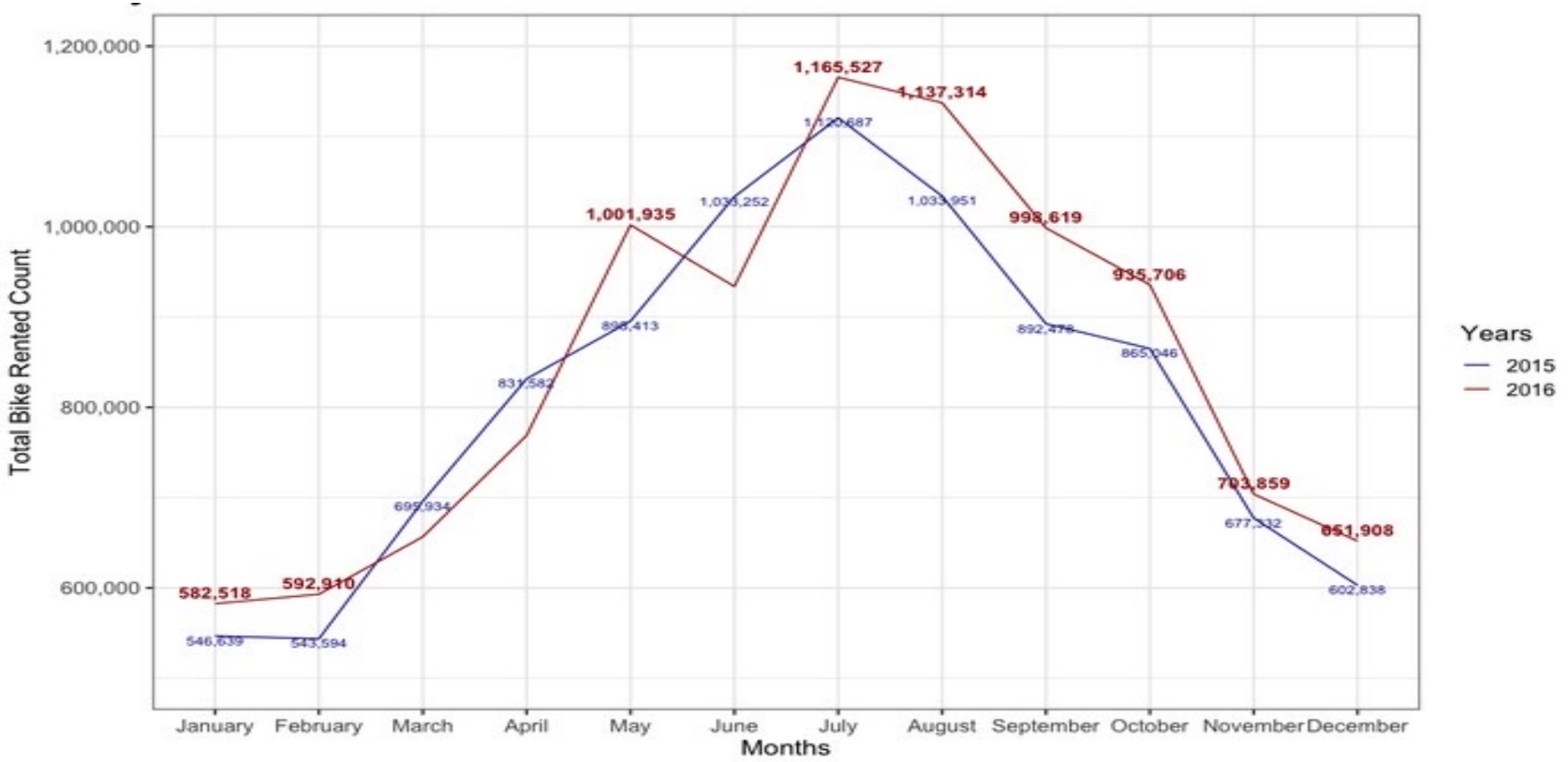
Remove '2017' year from our Timeframe

Checking the amount of recorded data

Year	Number of Data points	Number of days recorded per year
2015	8643	362
2016	8699	365
2017	72	3



Comparing Monthly Rentals: 2015 vs 2016



03

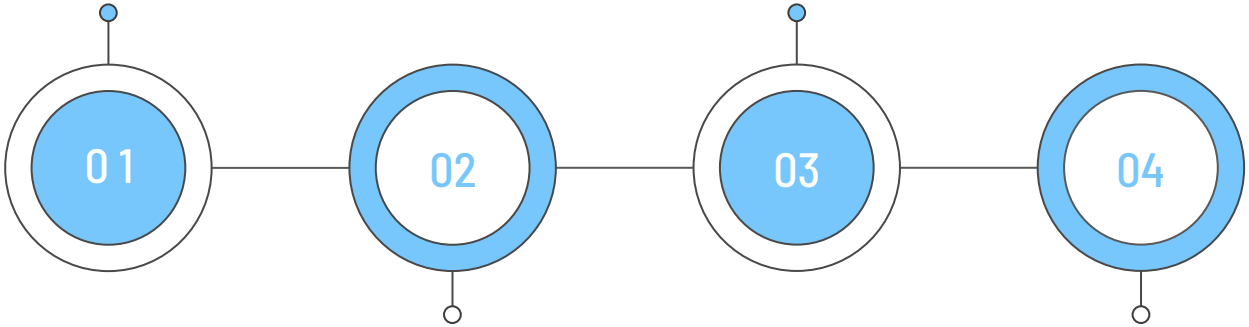
Data Modeling

Use statistical models and see the results

Our Models

Linear Regression

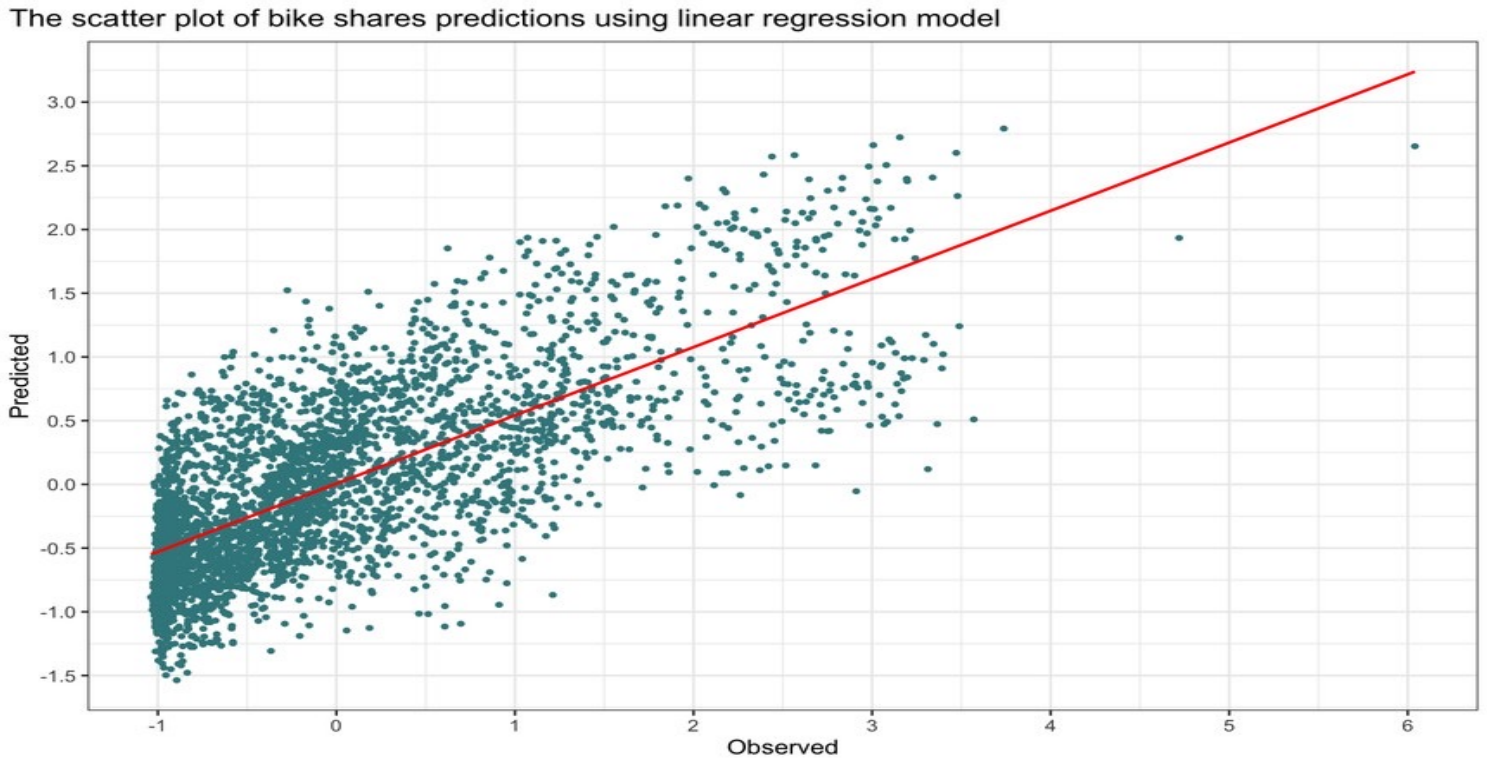
XG Boost



Decision Tree

Random Forest

Linear Regression



Model Evaluation(With Outlier)

0.49456409493

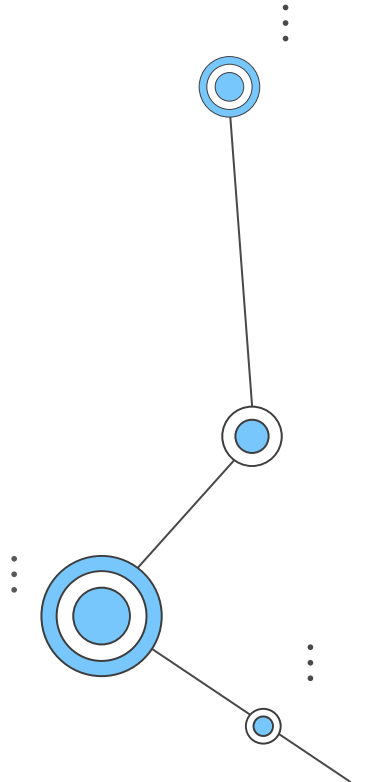
Mean Squared Error

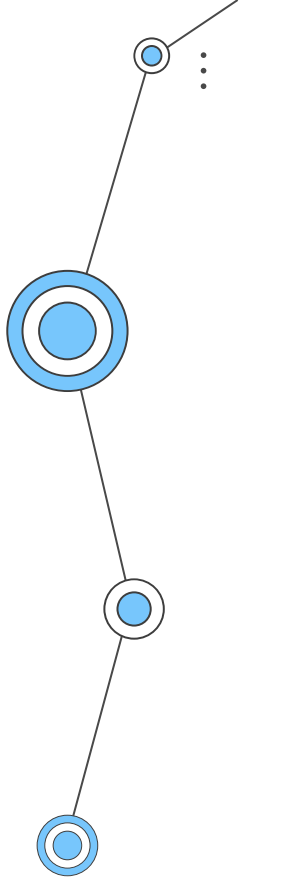
0.703252511505

Root Mean Squared

0.519638506136

Coefficient of Determination





Model Evaluation(Without Outlier)

0.275364299889

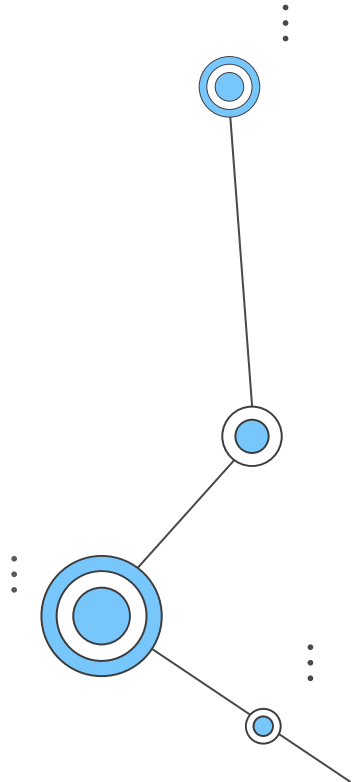
Mean Squared Error

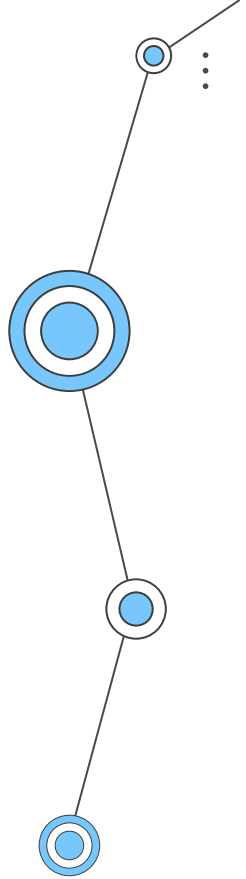
0.524751655123

Root Mean Squared

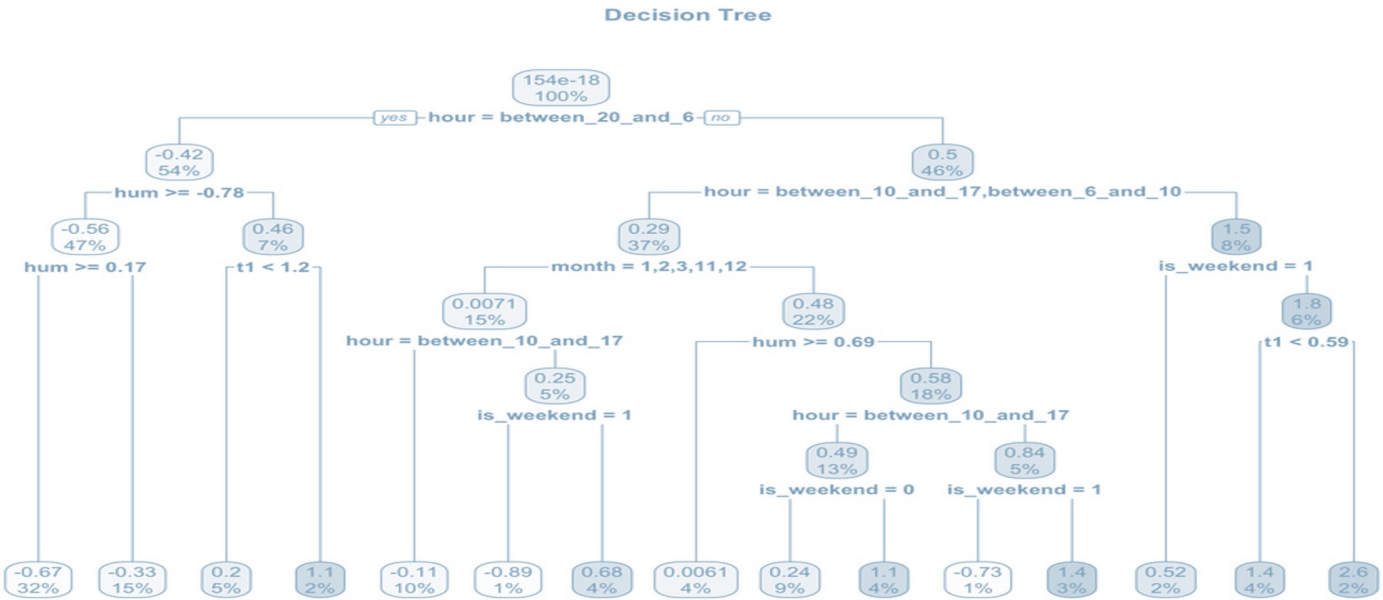
0.716759655213

Coefficient of Determination



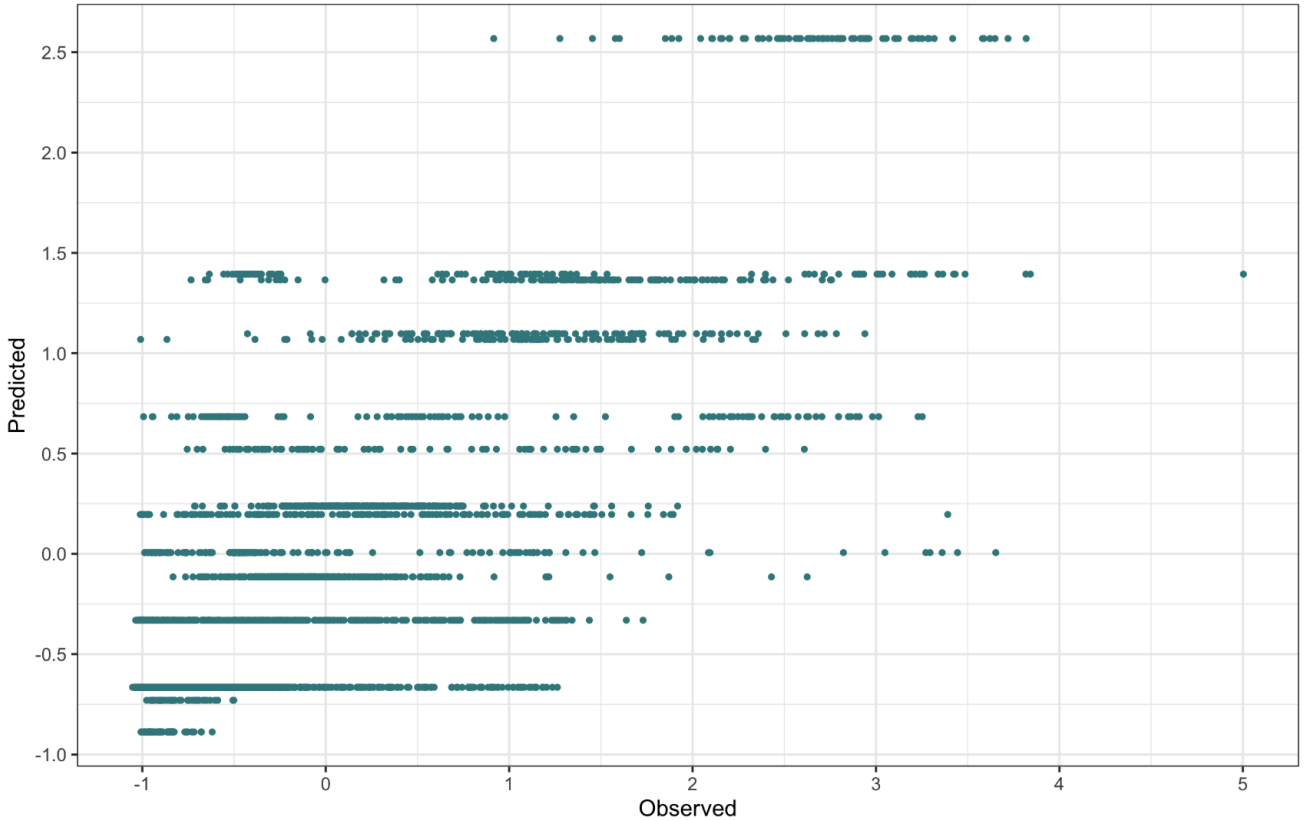


Decision Tree



Decision Tree

The scatter plot of bike shares predictions using decision tree model



Model Evaluation(With Outlier)

0.439149462423

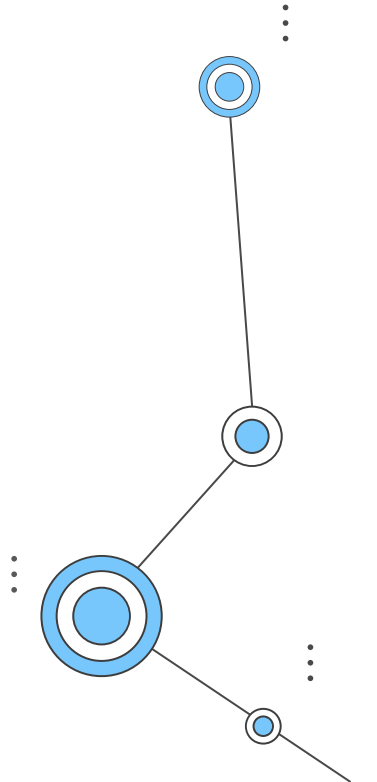
Mean Squared Error

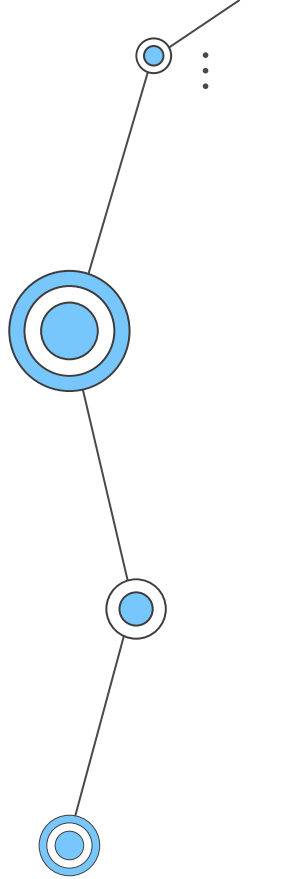
0.662410356645

Root Mean Squared

0.439149462423

Coefficient of Determination





Model Evaluation(Without Outlier)

0.233574247508

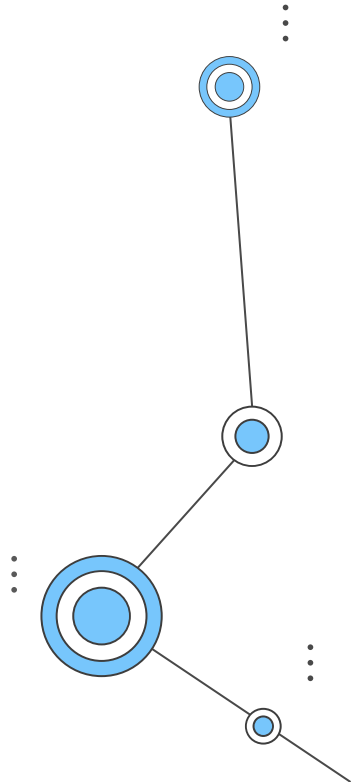
Mean Squared Error

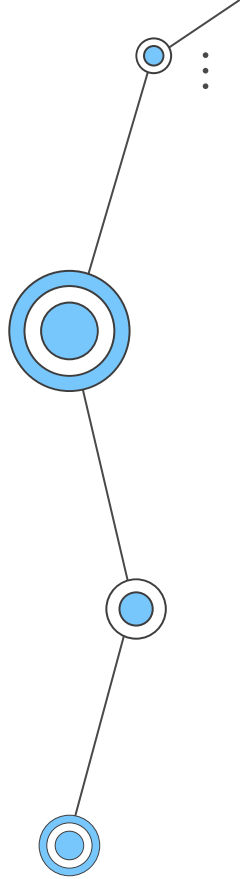
0.483295197067

Root Mean Squared

0.235088971031

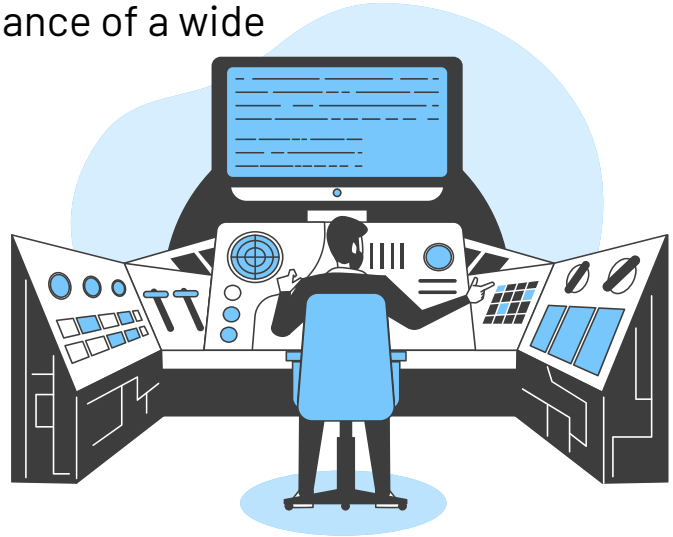
Coefficient of Determination





XG Boost

XGBoost is a powerful gradient boosting algorithm that uses an ensemble of decision trees to optimize the prediction performance of a wide range of machine learning tasks.



XG Boost

```
[1] train-rmse:0.892293 test-rmse:0.892860
[2] train-rmse:0.753810 test-rmse:0.758449
[3] train-rmse:0.668027 test-rmse:0.674478
...
...
[98] train-rmse:0.390820 test-rmse:0.540456
[99] train-rmse:0.389857 test-rmse:0.540917
[100] train-rmse:0.389426 test-rmse:0.541569
```

Model Evaluation(With Outlier)

0.293296661870

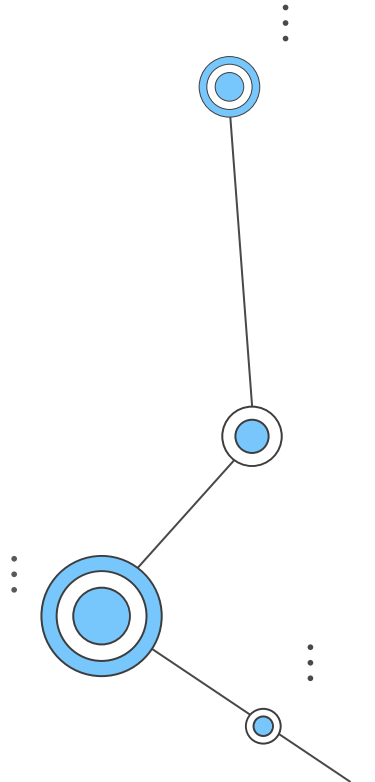
Mean Squared Error

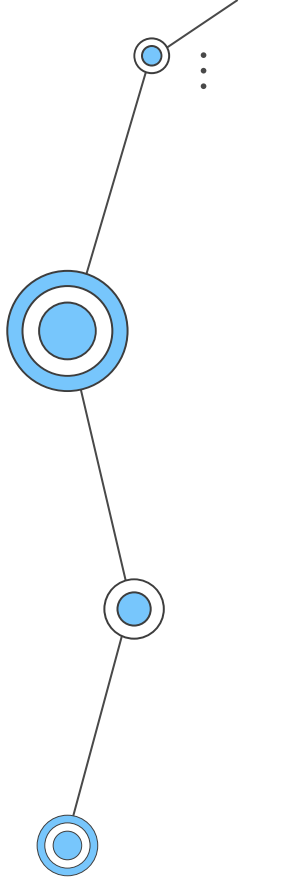
0.541568704662

Root Mean Squared

0.706703338129

Coefficient of Determination





Model Evaluation(Without Outlier)

0.038030960619

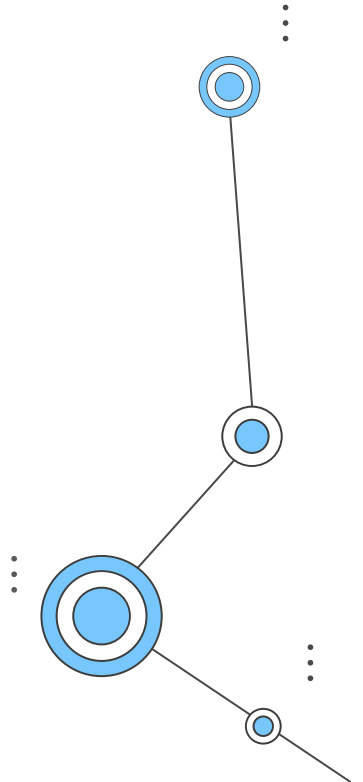
Mean Squared Error

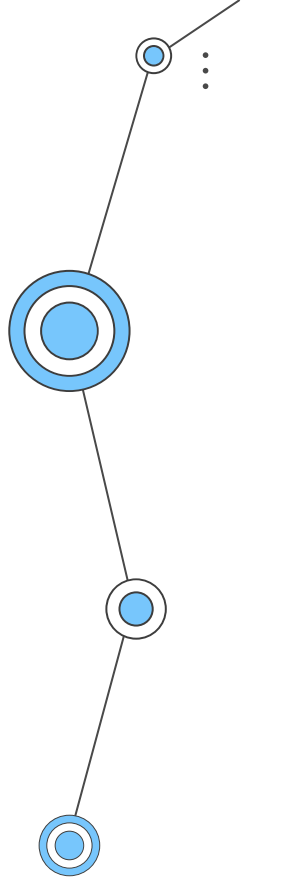
0.195015283040

Root Mean Squared

0.961969039380

Coefficient of Determination





Whoa!

XG Boost model has an unbelievable result until now

Random Forest

Random Forest is a versatile ensemble learning algorithm that constructs multiple decision trees and combines their predictions to achieve accurate and robust results in classification and regression tasks.

Model Evaluation(With Outlier)

0.293296661870

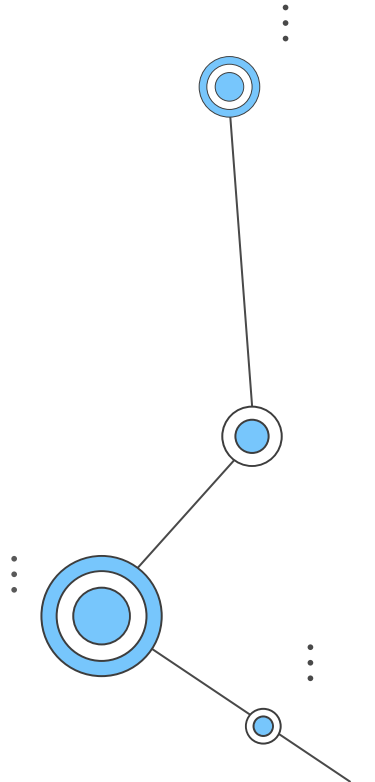
Mean Squared Error

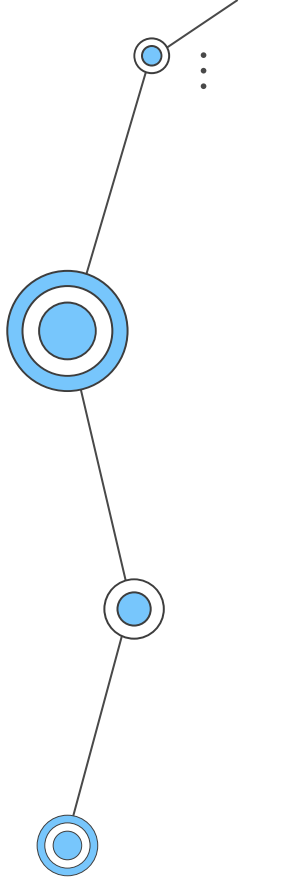
0.541568704662

Root Mean Squared

0.706703338129

Coefficient of Determination





Model Evaluation(Without Outlier)

0.066795829691

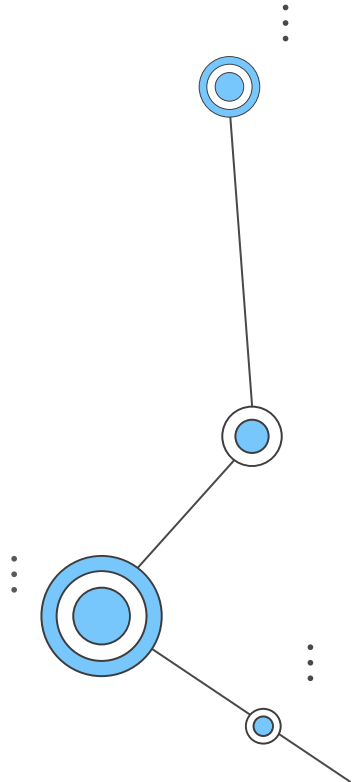
Mean Squared Error

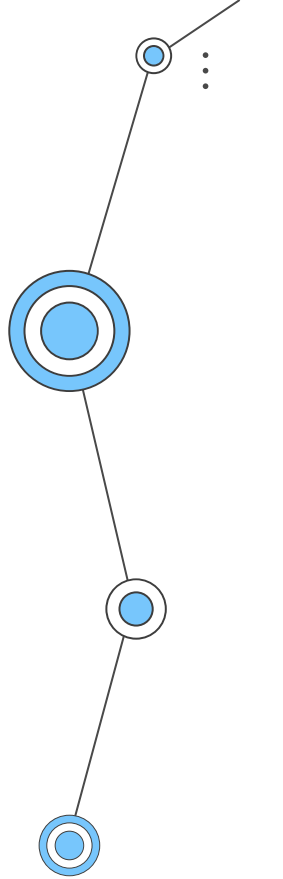
0.258448891836

Root Mean Squared

0.933204170308

Coefficient of Determination



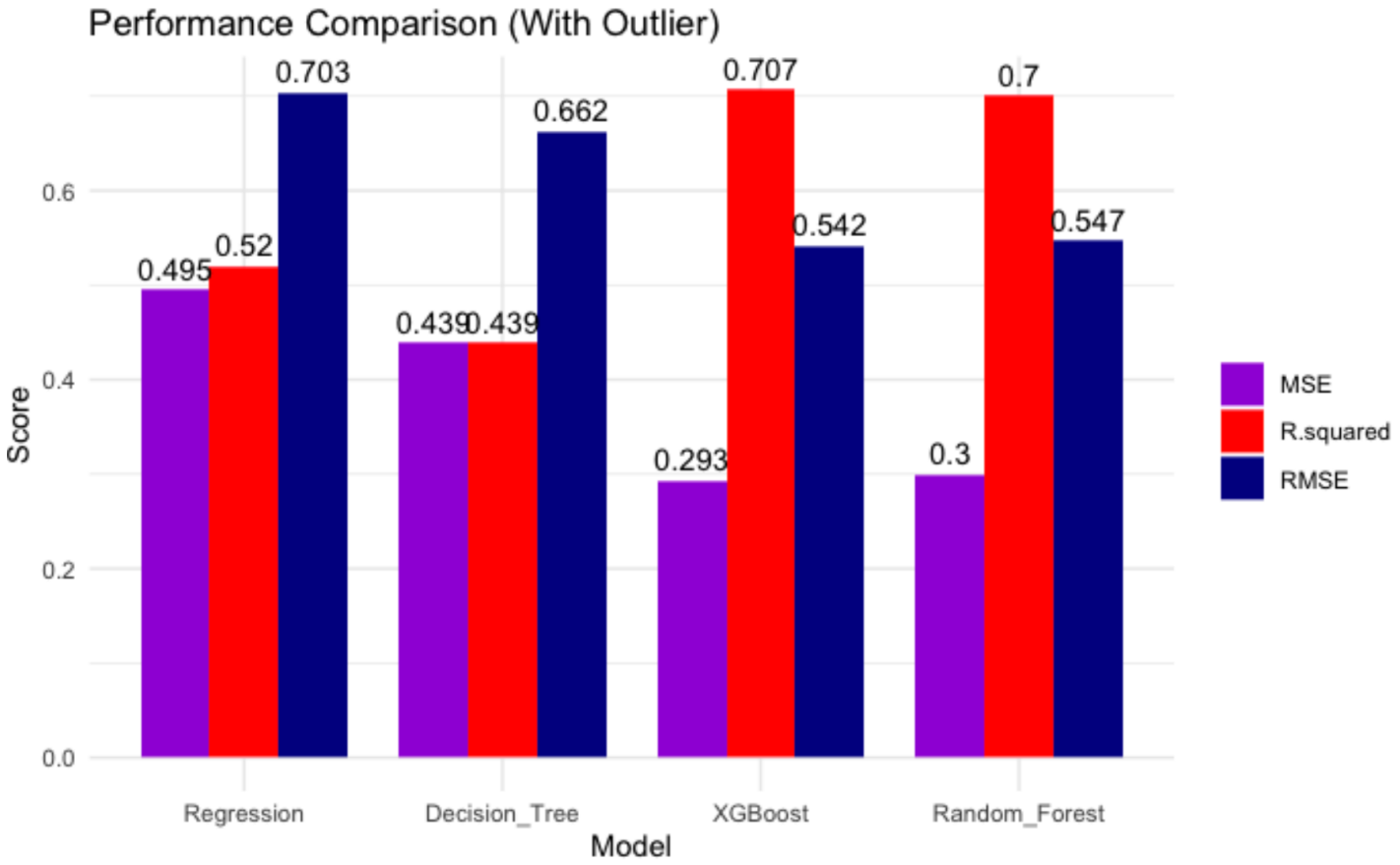


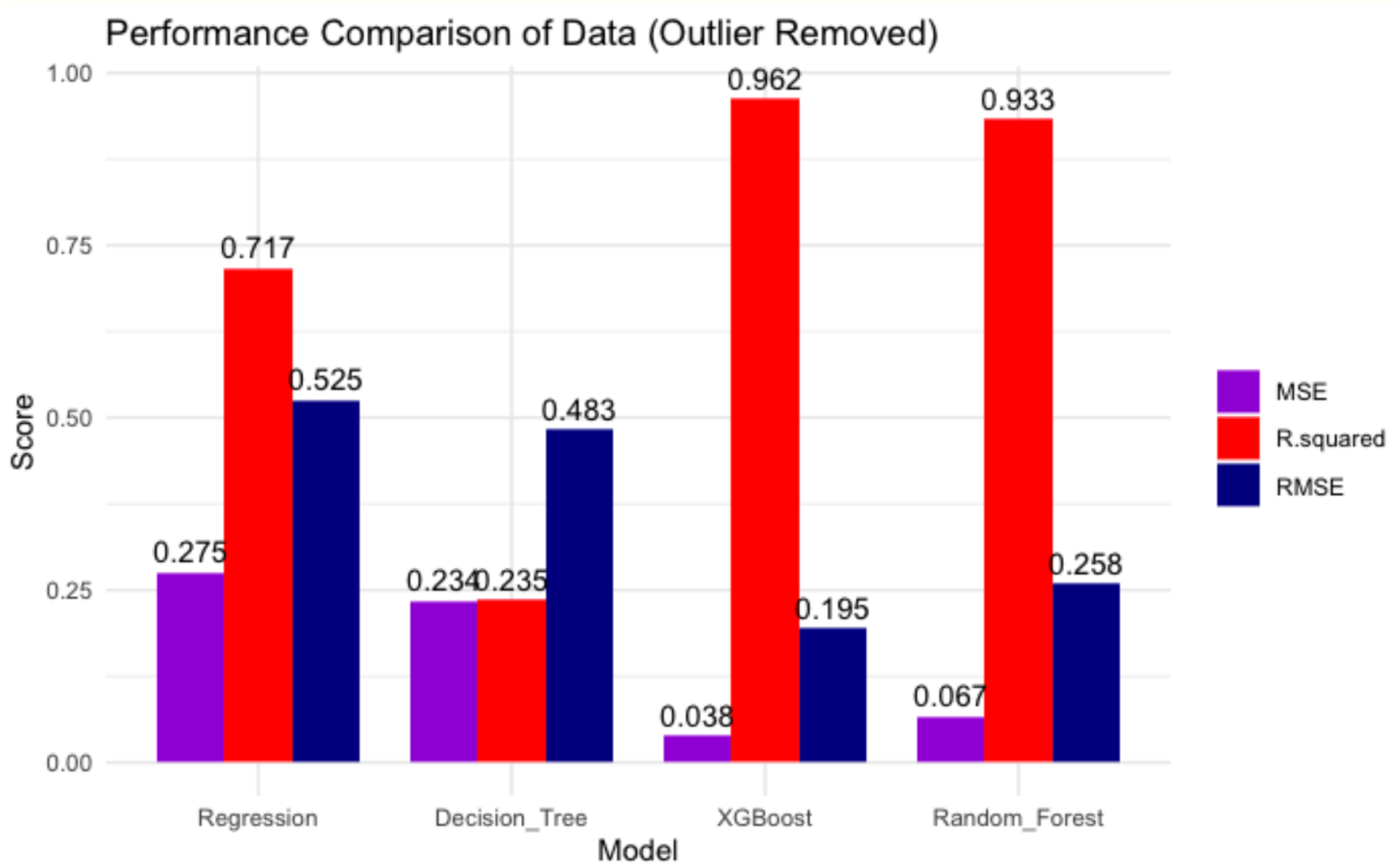
Model Compare with outlier

	Linear Regression	Decision Tree	XG Boost	Random Forest
MSE	0.4945641	0.4387875	0.2932967	0.2996687
RMSE	0.7032525	0.6624104	0.5415687	0.5474200
R_Squared	0.5196385	0.4391495	0.7067033	0.7003313

Model Compare without outlier

	Linear Regression	Decision Tree	XG Boost	Random Forest
MSE	0.2753643	0.2335742	0.0380309	0.0667958
RMSE	0.5247517	0.4832952	0.19501528	0.2584488
R_Squared	0.7167597	0.2350890	0.96196904	0.93320417


Performance Comparison (With Outlier)
Score
0.0
0.2
0.4
0.6
0.495
0.52
0.703
0.439
0.439
0.662
0.293
0.707
0.542
0.3
0.7
0.547
Regression
Decision_Tree
XGBoost
Random_Forest
Model
MSE
R.squared
RMSE


Performance Comparison of Data (Outlier Removed)
1.00
0.75
0.50
0.25
0.00
Score
0.275
0.717
0.525
0.234
0.235
0.483
0.038
0.962
0.195
0.067
0.933
0.258
Regression
Decision_Tree
XGBoost
Random_Forest
Model
MSE
R.squared
RMSE

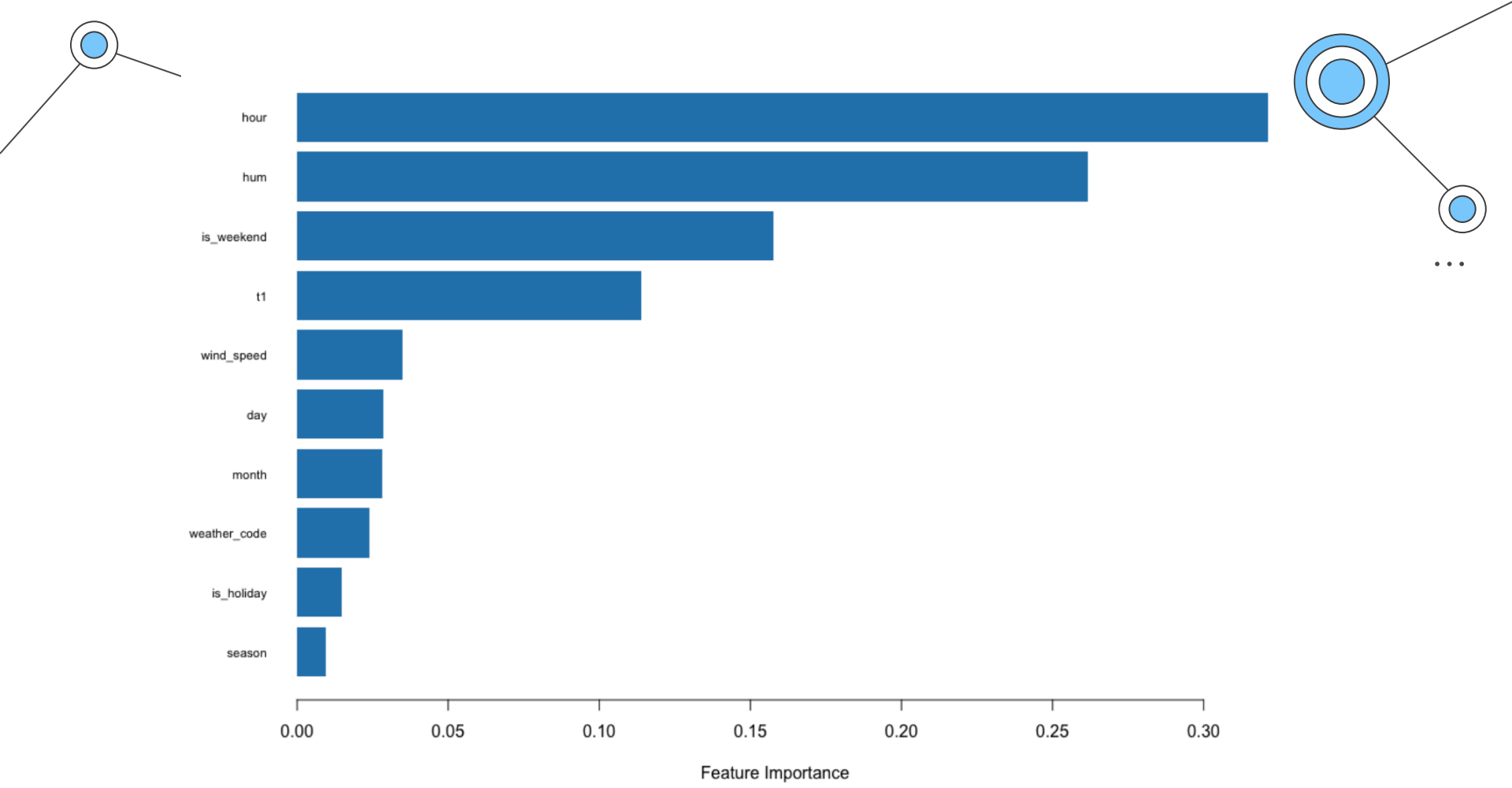
Variable Importance in XGBoost

Feature	*Gain*	*Cover*	*Frequency*
hour	0.321195910	0.09534628	0.102326461
hum	0.261786644	0.18237488	0.175146586
is_weekend	0.157476870	0.02178167	0.038585209
t1	0.113746150	0.22903085	0.191602043
wind_speed	0.034998969	0.14660270	0.172876868
month	0.028453157	0.10147147	0.069982977
weather_code	0.023984173	0.04036193	0.054094950
day	0.028453157	0.12361489	0.134102516
is_holiday	0.014802405	0.01236841	0.009457159
season	0.009436469	0.02939568	0.027614904
year	0.005979181	0.01765125	0.024210327

Gain: emphasizes the predictive power of a feature

Cover: highlights the frequency of using a feature for splitting.

Frequency: reflects the prevalence of a feature in the dataset.


hour
hum
is_weekend
t1
wind_speed
day
month
weather_code
is_holiday
season
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Feature Importance

04

Conclusion

Key points and outcome

Conclusion

The study aimed to predict the number of rented bicycles in London using regression models such as Regression, Decision Tree, XGBoost, and Random Forest. Evaluation metrics showed that XGBoost and Random Forest consistently outperformed the other models, exhibiting higher accuracy with lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values, and higher R-squared value. The analysis also emphasized the impact of outliers on model performance, highlighting the importance of removing outliers during data preprocessing. Overall, XGBoost and Random Forest models, applied to outlier-free data, offer accurate predictions and can aid in optimizing bicycle rental system management for better resource allocation.

Grazie!

Thanks for generous attention

Prof. Roberta Siciliano
Prof. Michele Staiano

Statistical data analysis
2022-2023