

Data Science with R

Capstone project

Mohd Amzad
M.Tech(CSE)

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary



Purpose of the Project

- Collects relevant bike-sharing and weather data to investigate how weather influences urban bike-sharing demand.
- Creates a prediction model to address future demand trends and improve the performance of bike rentals.

Key Findings

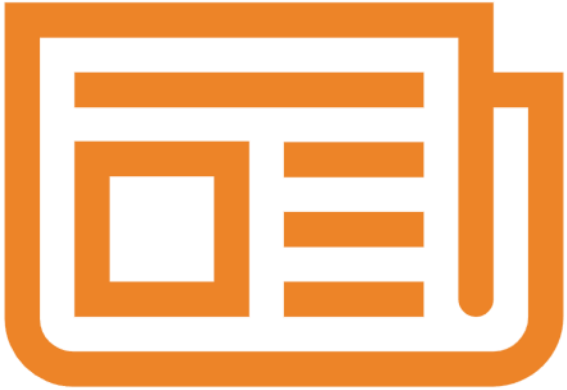
- Bike rental statistics show a rise in demand between 8 a.m. and 6 p.m.
- During the winter, bike-sharing demand is extremely low. It demonstrates the unparalleled role of 'Seasonality' in bike rentals.
- Visually, we can detect some strong relationships as roughly linear patterns.

Introduction



- The bicycle is a sustainable and inexpensive mode of transportation. Its popularity changes from season to season. Days with rain and snowfall will result in a reduction in bike rentals. Bicycles cannot be used in such conditions. This leads us to believe that bike-sharing demand is quite low these days.
- It is apparent that weather conditions have a significant impact on bike-sharing demand. So it is critical to understand the weather conditions in order to match the supply of bikes to the demand.
- In this project, we are gathering datasets of weather forecasts and historical bike sharing demand and developing a model for more accurate bike-sharing demand prediction while accounting for weather conditions.

Methodology



- Data collection
- Data wrangling
- Exploratory data analysis (EDA) using SQL
- Exploratory data analysis (EDA) using data visualization
- Predictive analysis using regression models
 - Build the baseline model
 - Improve the baseline model
- Build a R Shiny dashboard app

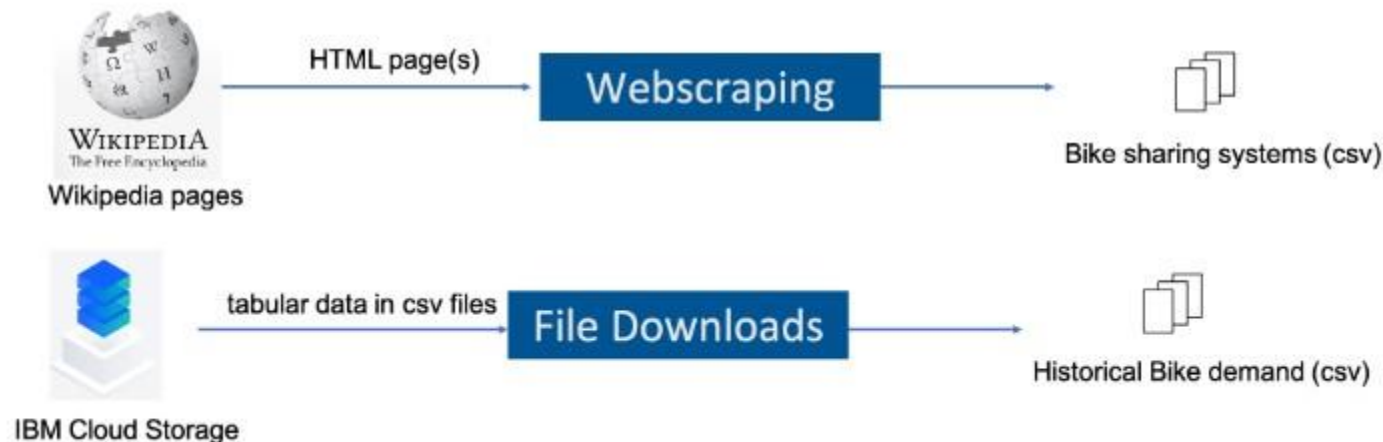
Methodology

Data collection

- ▶ The relevant data is collected from various sources to analyze and predict a city's bike-sharing demand. You will use three main data collection methods, including:
 1. Using OpenWeather REST API to request weather data,
 2. Web scraping information about global bike-sharing systems from Wikipedia web pages,



3. Downloading and aggregating tabular data from cloud storage.



Data wrangling

- Data wrangling removes noise and converts undesirable data to a more suitable format for analysis.



Data wrangling

- ▶ Data wrangling with stringr and regular expressions
 - ▶ Standardize column names for all collected datasets
 - ▶ Remove undesired reference links using regular expressions
 - ▶ Extract numeric values using regular expressions
- ▶ Data wrangling with dplyr
 - ▶ Detect and handle missing values
 - ▶ Create indicator (dummy) variables for categorical variables
 - ▶ Normalize data

- Standardize column names for all collected data sets
 - `'city' -> 'CITY'`
- Remove undesired reference links
 - `'Bike sharing system {123}'`
- Remove textual annotations in numerical strings
 - `'1000 (Updated from 1050)'`

EDA with SQL

- ▶ Count how many records are in the seoul_bike_sharing dataset
- ▶ Determine how many hours had non-zero rented bike count.
- ▶ Query the weather forecast for Seoul over the next 3 hours.
- ▶ Find which seasons are included in the seoul bike sharing dataset.
- ▶ Find the first and last dates in the Seoul Bike Sharing dataset determine which date and hour had the most bike rentals
- ▶ Hourly popularity and temperature by season
- ▶ Find the average, minimum, maximum and standard deviation of hourly bike count during each seasons.
- ▶ Weather seasonality in bike rentals.
- ▶ Total bike count and city information for Seoul.
- ▶ Find all city names and coordinates with comparable bike scale to Seoul's bike sharing system

EDA with data visualization

- In EDA with data visualization, we generated a wide range of charts to visualize the outcomes of the exploratory data analysis.

They are;

- 1) Scatter plots :- for finding correlation between variables
- 2) Box plots :-for finding outliers and irregular behaviors.
- 3) Histogram :- for understanding the distributions of data

Predictive analysis

Predict Hourly Rented Bike Count using Basic Linear Regression Models

- ▶ Split data into training and testing datasets
- ▶ Build a linear regression model using only the weather variables
- ▶ Build a linear regression model using both weather and date/time variables
- ▶ Evaluate the models and identify important variables

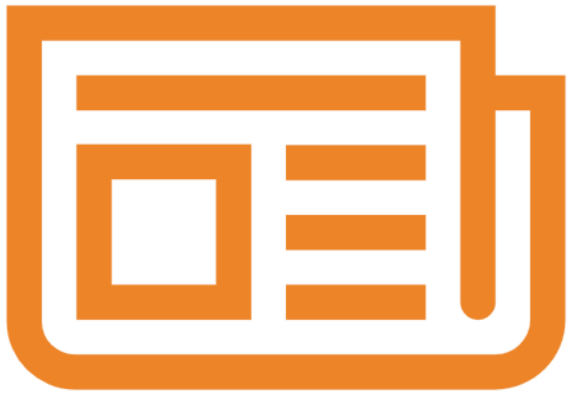
Refine the Baseline Regression Models

- ▶ Add higher order terms
- ▶ Add interaction terms
- ▶ Add regularization
- ▶ Experiment to find the best performed model

Build a R Shiny dashboard

- ▶ Summarize what plots and interactions you built into the dashboard using bullet points
- ▶ Leaflet Map :- For plotting bike prediction overview map
- ▶ Correlation Plot :- For showing static humidity and bike-sharing demand prediction correlation plot.
- ▶ Line charts : - For adding a static temperature trend line and bike-sharing demand prediction trend line
- ▶ Drop-down list :- For selecting from listed cities
- ▶ Plot click : - provides viewers with click points on the plot to know the time and respective bike prediction.

Results



- Exploratory data analysis results
- Predictive analysis results
- A dashboard demo in screenshots

EDA with SQL

Busiest bike rental times

- Date with most bike rentals = 19/06/2018
- Hour = 18
- which means June 19th had the highest peak demand from the date category, and the 6 PM is the peak demand time for the bike rentals.

```
>  
> dbGetQuery(con, "SELECT DATE, HOUR, RENTED_BIKE_COUNT FROM SEOUL_BIKE_SHARING  
+               WHERE RENTED_BIKE_COUNT = (SELECT MAX(RENTED_BIKE_COUNT) FROM SEOUL_BIKE_SHARING)")  
      DATE HOUR RENTED_BIKE_COUNT  
1 19/06/2018  18             3556  
>
```


Hourly popularity and temperature by seasons

- The range of the average hourly temperature is 0 to 2 degrees Celsius.
- Customers clearly choose the season with the least amount of temperature variance for their bike rides.

```
dbGetQuery(con, "SELECT SEASONS,AVG(TEMPERATURE/HOUR), AVG(RENTED_BIKE_COUNT) FROM  
SEOUL_BIKE_SHARING GROUP BY SEASONS,HOUR  
ORDER BY AVG(RENTED_BIKE_COUNT) DESC LIMIT 10")
```

```
> dbGetQuery(con, "SELECT SEASONS,AVG(TEMPERATURE/HOUR), AVG(RENTED_BIKE_COUNT) FROM  
+ SEOUL_BIKE_SHARING GROUP BY SEASONS,HOUR  
+ ORDER BY AVG(RENTED_BIKE_COUNT) DESC LIMIT 10")
```

	SEASONS	AVG(TEMPERATURE/HOUR)	AVG(RENTED_BIKE_COUNT)
1	Summer	1.6326617	2135.141
2	Autumn	0.8906584	1983.333
3	Summer	1.4880937	1889.250
4	Summer	1.3533152	1801.924
5	Summer	1.2513458	1754.065
6	Spring	0.8873457	1689.311
7	Summer	1.1681324	1567.870
8	Autumn	1.0163399	1562.877
9	Summer	1.7692298	1526.293
10	Autumn	0.7928135	1515.568

Rental Seasonality

- Summer is the most popular season for bike rentals, with the largest bike rental count and standard deviation, followed by the autumn season.

```
45 dbGetQuery(con, "SELECT SEASONS,AVG(RENTED_BIKE_COUNT/HOUR) AS HOURLY_BIKE_COUNT,  
46 MAX(RENTED_BIKE_COUNT) AS maximum_of_bike_count, MIN(RENTED_BIKE_COUNT) as minimum_of_bike_count,  
47 SQRT(AVG(RENTED_BIKE_COUNT*RENTED_BIKE_COUNT) - AVG(RENTED_BIKE_COUNT)*AVG(RENTED_BIKE_COUNT))  
48 AS STANDARD_DEVIATION_OF_BIKE_COUNT FROM SEOUL_BIKE_SHARING GROUP BY SEASONS")  
49
```

```
> dbGetQuery(con, "SELECT SEASONS,AVG(RENTED_BIKE_COUNT/HOUR) AS HOURLY_BIKE_COUNT,  
+ MAX(RENTED_BIKE_COUNT) AS maximum_of_bike_count, MIN(RENTED_BIKE_COUNT) as minimum_of_bike_count,  
+ SQRT(AVG(RENTED_BIKE_COUNT*RENTED_BIKE_COUNT) - AVG(RENTED_BIKE_COUNT)*AVG(RENTED_BIKE_COUNT))  
+ AS STANDARD_DEVIATION_OF_BIKE_COUNT FROM SEOUL_BIKE_SHARING GROUP BY SEASONS")  
SEASONS HOURLY_BIKE_COUNT maximum_of_bike_count minimum_of_bike_count  
1 Autumn 102.21540 3298 2  
2 Spring 77.64831 3251 2  
3 Summer 119.24764 3556 9  
4 Winter 27.28696 937 3  
STANDARD_DEVIATION_OF_BIKE_COUNT  
1 617.3885  
2 618.5247  
3 690.0884  
4 150.3374
```

Weather Seasonality

- Weather Seasonality
- Present your query result with a short explanation here

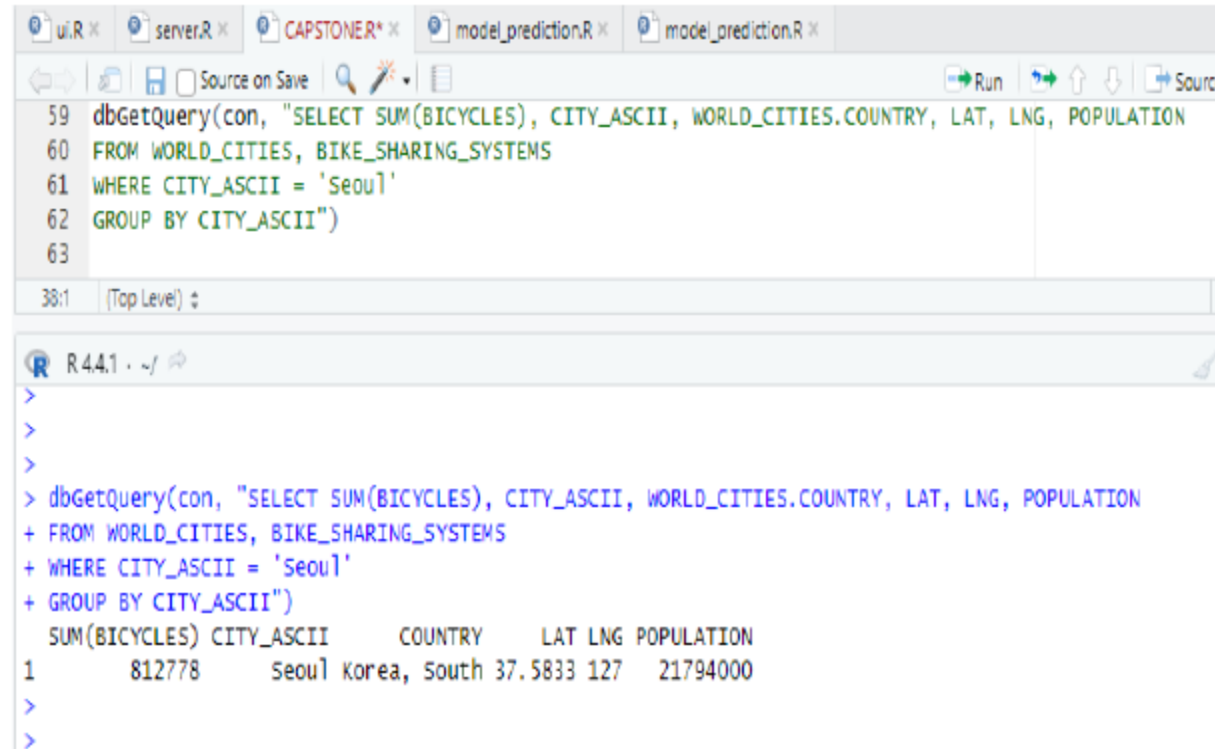
```
52 dbGetQuery(con, "SELECT SEASONS,TEMPERATURE,HUMIDITY,WIND_SPEED,VISIBILITY,DEW_POINT_TEMPERATURE,
53 SOLAR_RADIATION,RAINFALL,SNOWFALL,AVG(TEMPERATURE/HOUR), AVG(RENTED_BIKE_COUNT) FROM
54 SEOUL_BIKE_SHARING GROUP BY SEASONS
55 ORDER BY AVG(RENTED_BIKE_COUNT) DESC")
56
```

38:1 (Top Level) ↕ R S

```
R 4.4.1 ~\
>
> dbGetQuery(con, "SELECT SEASONS,TEMPERATURE,HUMIDITY,WIND_SPEED,VISIBILITY,DEW_POINT_TEMPERATURE,
+ SOLAR_RADIATION,RAINFALL,SNOWFALL,AVG(TEMPERATURE/HOUR), AVG(RENTED_BIKE_COUNT) FROM
+ SEOUL_BIKE_SHARING GROUP BY SEASONS
+ ORDER BY AVG(RENTED_BIKE_COUNT) DESC")
  SEASONS TEMPERATURE HUMIDITY WIND_SPEED VISIBILITY DEW_POINT_TEMPERATURE SOLAR_RADIATION RAINFALL
1 Summer      20.1      63      1.1      1698      12.8      0      0
2 Autumn      23.2      83      0.5      1887      20.1      0      0
3 Spring       2.0      96      1.3      1894       1.4      0      2
4 Winter      -5.2      37      2.2      2000     -17.6      0      0
  SNOWFALL AVG(TEMPERATURE/HOUR) AVG(RENTED_BIKE_COUNT)
1      0      4.0819329      1034.0734
2      0      2.0076295      924.1105
3      0      1.8539948      746.2542
4      0     -0.5761727      225.5412
>
```

Bike-sharing info in Seoul

- Find Seoul's overall bike count as well as city information.
- Seoul between 8 lakh and 2 crore.
- The total number of bikes in Seoul, the capital city of South Korea, indicates how often people ride bicycles there.



The screenshot shows an RStudio window with several tabs: 'ui.R', 'server.R', 'CAPSTONE*', 'model_prediction.R', and 'model_prediction.R'. The active tab is 'CAPSTONE*', which contains a SQL query. The query is as follows:

```
59 dbGetQuery(con, "SELECT SUM(BICYCLES), CITY_ASCII, WORLD_CITIES.COUNTRY, LAT, LNG, POPULATION
60 FROM WORLD_CITIES, BIKE_SHARING_SYSTEMS
61 WHERE CITY_ASCII = 'seoul'
62 GROUP BY CITY_ASCII")
63
```

The output of the query is displayed in the console below the editor. It shows the result of the SQL query, which is a single row of data for Seoul.

```
>
>
>
> dbGetQuery(con, "SELECT SUM(BICYCLES), CITY_ASCII, WORLD_CITIES.COUNTRY, LAT, LNG, POPULATION
+ FROM WORLD_CITIES, BIKE_SHARING_SYSTEMS
+ WHERE CITY_ASCII = 'seoul'
+ GROUP BY CITY_ASCII")
  SUM(BICYCLES) CITY_ASCII  COUNTRY  LAT LNG POPULATION
1      812778      Seoul Korea, South 37.5833 127 21794000
>
>
```

Cities similar to Seoul

- Some Chinese cities, like Seoul, have comparable bike sharing schemes.

```
66
67 dbGetQuery(con, "SELECT WORLD_CITIES.CITY_ASCII, BIKE_SHARING_SYSTEMS.CITY, WORLD_CITIES.COUNTRY,
68 LAT, LNG, POPULATION, BICYCLES
69 FROM BIKE_SHARING_SYSTEMS
70 LEFT JOIN WORLD_CITIES ON BIKE_SHARING_SYSTEMS.CITY = WORLD_CITIES.CITY_ASCII
71 WHERE BICYCLES BETWEEN 15000 AND 20000
72 GROUP BY WORLD_CITIES.CITY_ASCII")
```

72:35 (Top Level) ↕

R 4.4.1 · ~/

```
+ GROUP BY WORLD_CITIES.CITY_ASCII")
```

	CITY_ASCII	CITY	COUNTRY	LAT	LNG	POPULATION	BICYCLES
1	<NA>	Kunshan	<NA>	NA	NA	NA	20000
2	Beijing	Beijing	China	39.9050	116.3914	19433000	16000
3	Ningbo	Ningbo	China	29.8750	121.5492	7639000	15000
4	Seoul	Seoul	Korea, South	37.5833	127.0000	21794000	20000
5	Shanghai	Shanghai	China	31.1667	121.4667	22120000	19165
6	Weifang	Weifang	China	36.7167	119.1000	9373000	20000
7	Xi'an	Xi'an	China	34.2667	108.9000	7135000	20000
8	Zhuzhou	Zhuzhou	China	27.8407	113.1469	3855609	20000

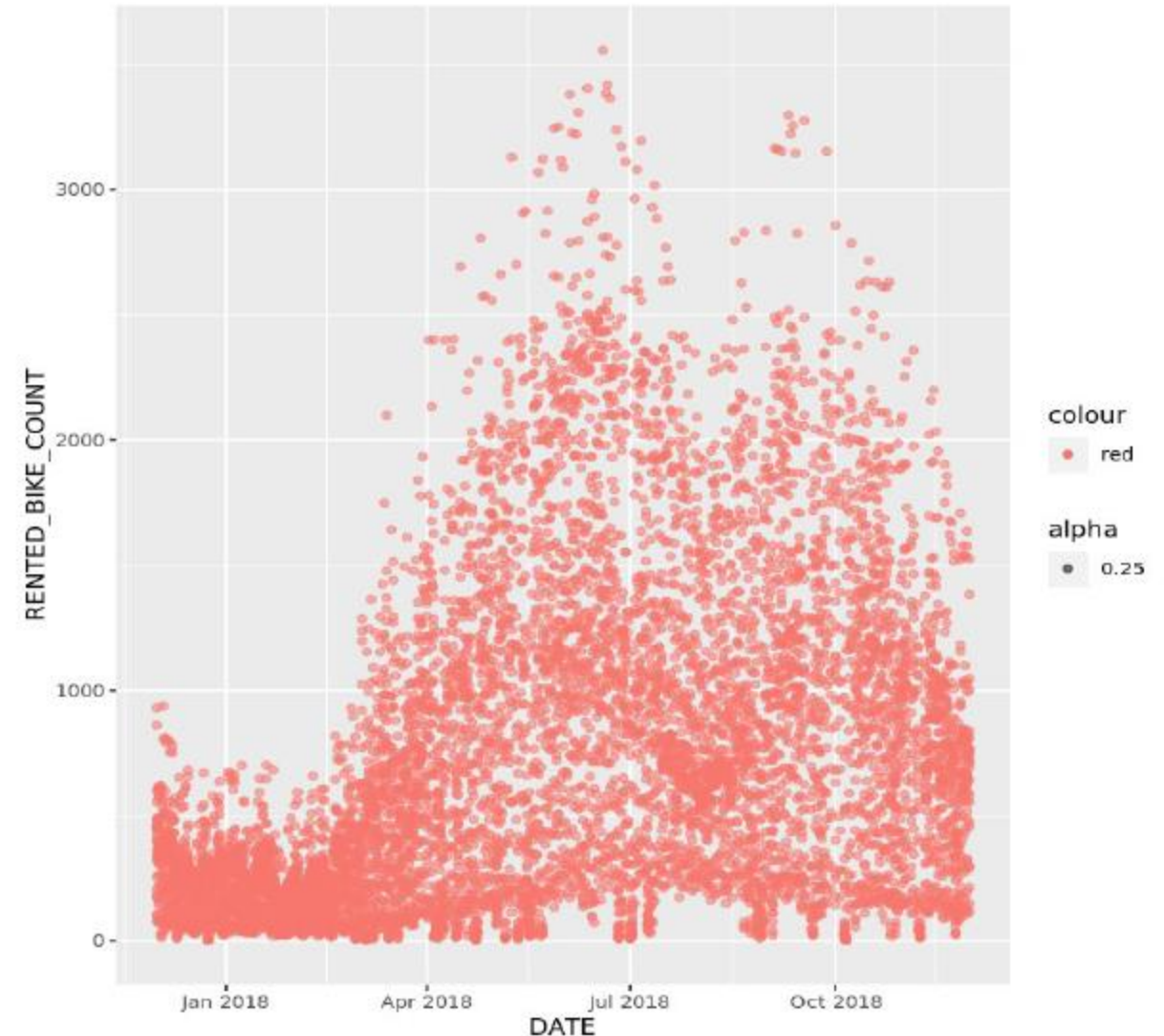
EDA with Visualization

Bike rental vs Date

Here shown a scatter plot of showing the relationship between the variables RENTED_BIKE_COUNT and DATE

The scatter figure indicates a rising trend in rental bike count as days pass, indicating a positive association between factors.

From this we can analyze that DATE has big effect on bike rental.

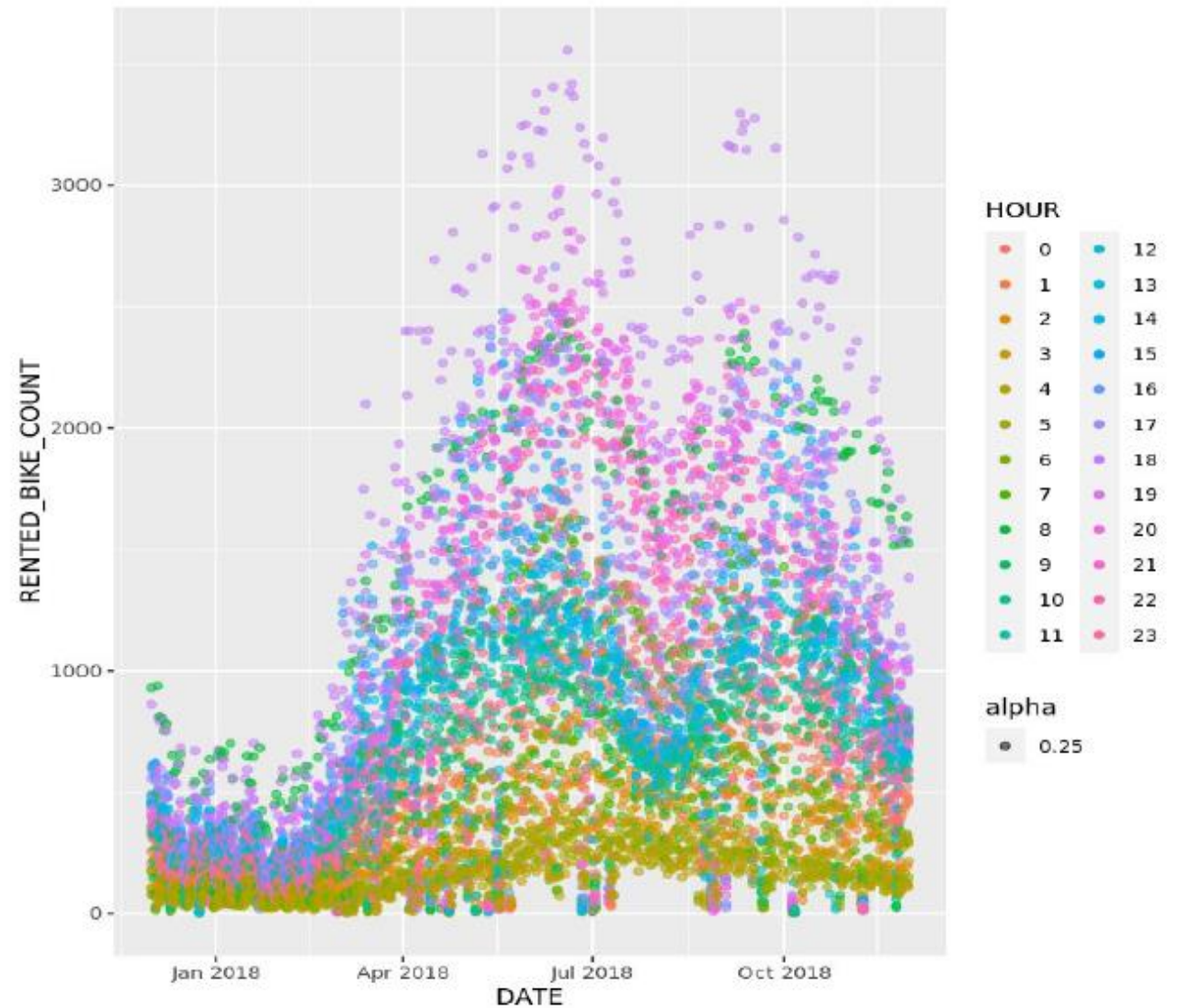


Bike rental vs. Datetime

Display the RENTED_BIKE_COUNT time series in the same plot with the addition of HOURS as the color.

Display the scatter plot screenshot along with the descriptions.

The majority of the rentals happen about six o'clock at night.



Bike rental histogram

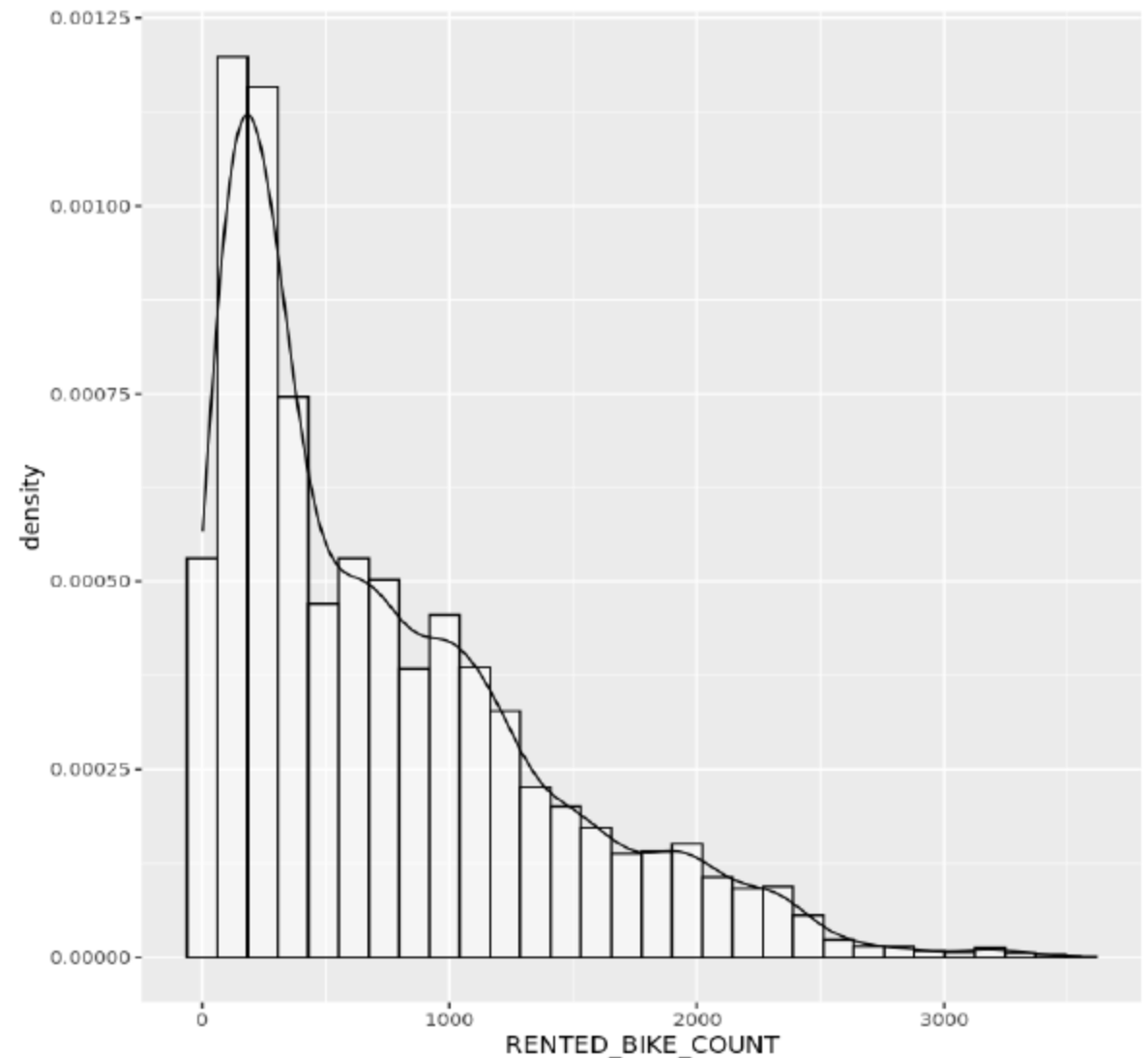
Present a histogram with a kernel density curve superimposed.

The histogram shows us that there are typically not many bikes rented out.

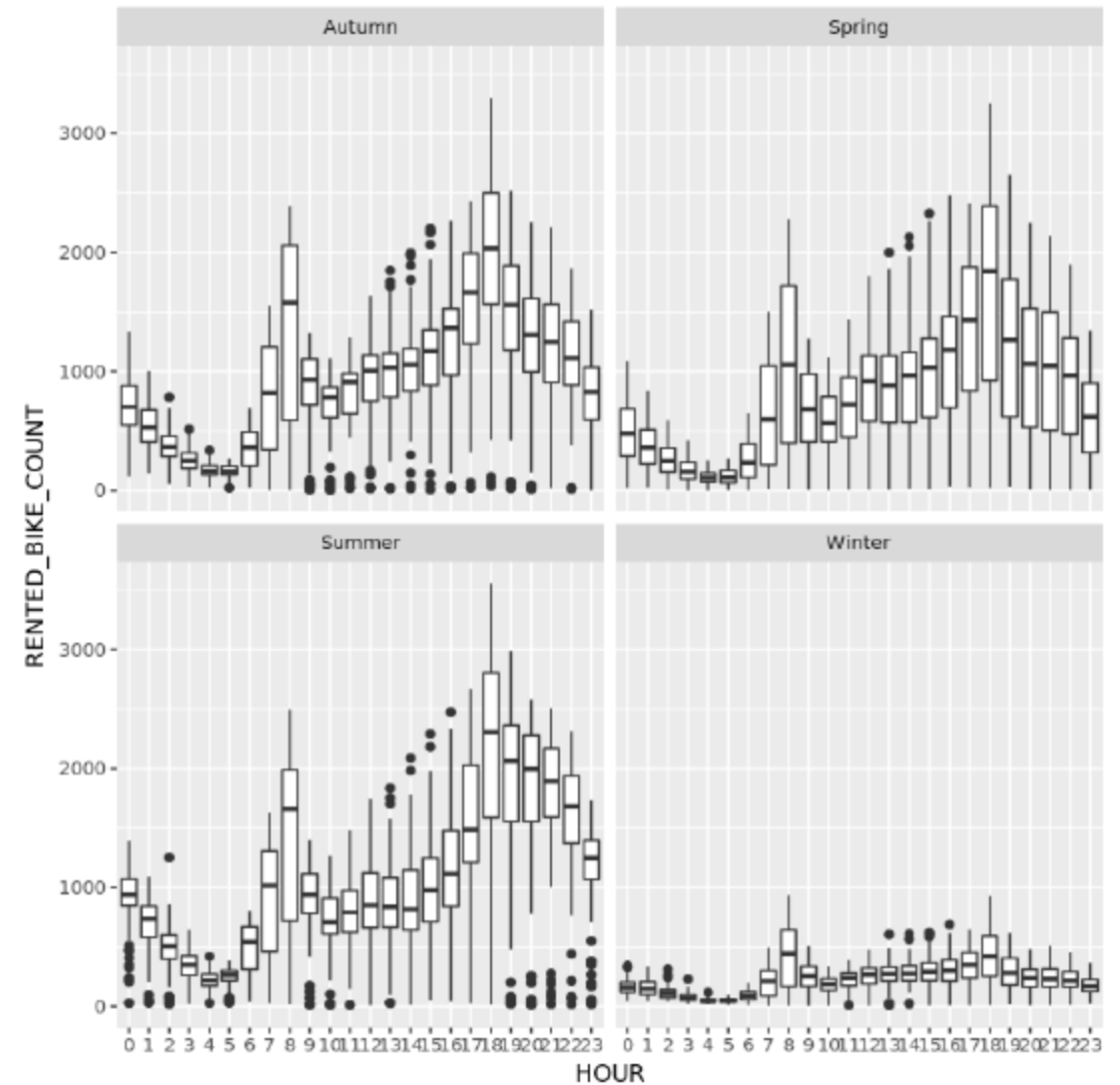
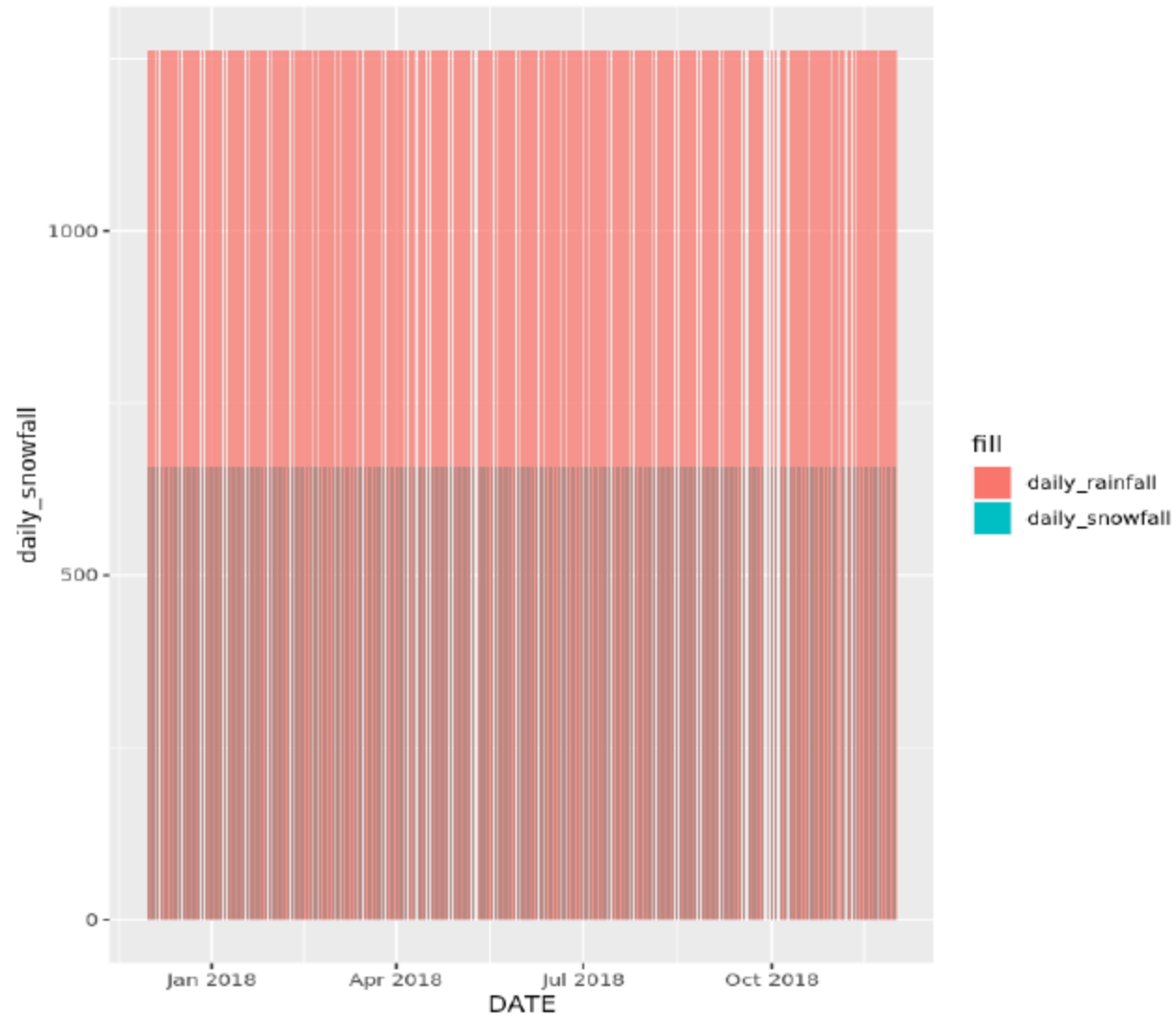
In fact, 250 bikes are leased the most frequently—this is known as the "mode."

Based on the 'bumps' observed at approximately 700, 900, 1900, and 3200 bikes, it appears that additional modes could be concealed within data subgroups.

It's interesting to note that, based on the tail of the distribution, there are occasionally far more bikes hired out than typical.



Daily total rainfall and snowfall



Predictive analysis

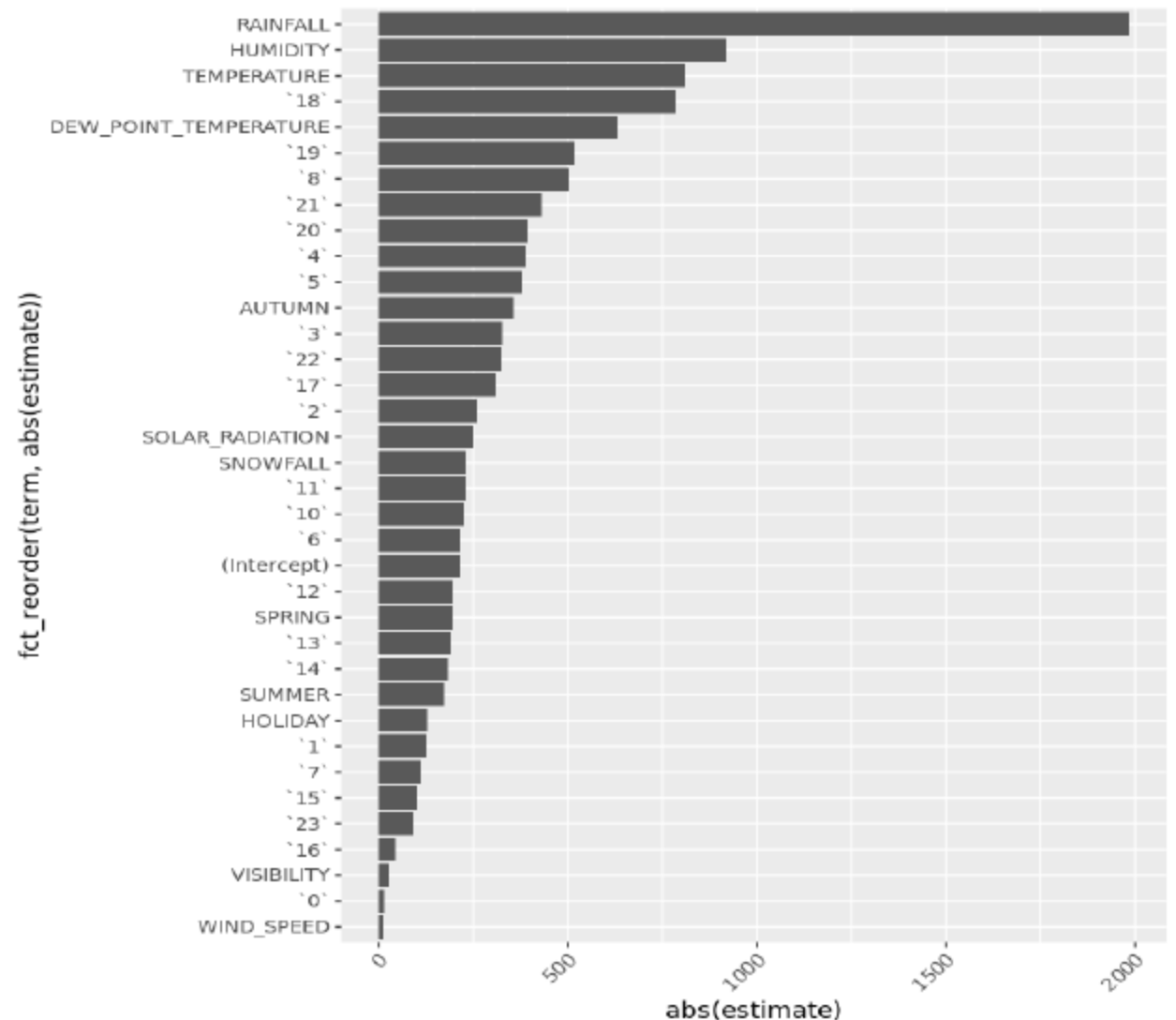
Ranked coefficients

Try to explain why, in order to forecast the demand for bike sharing, some characteristics are crucial and others are not, and how these differences actually effect the rental of bikes.

For example, riding a bike during wet or snowy seasons might be challenging. People prefer to travel in covered vehicles during these seasons since they are more convenient and safe.

However, many favor motorcycles over other vehicles when the weather is sunny. Seasons therefore affect the demand for bike sharing.

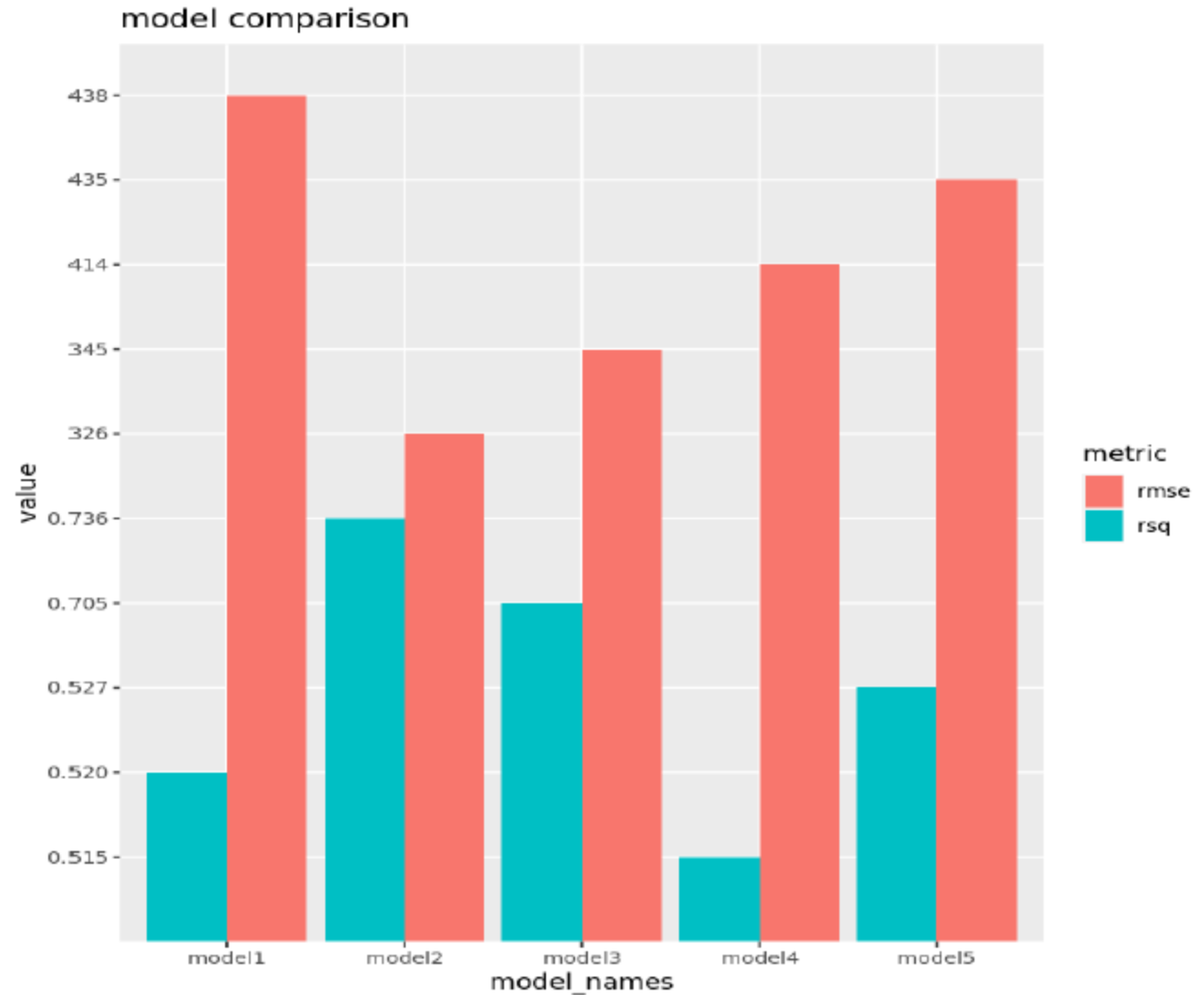
The demand for bike sharing is heavily influenced by the weather, hence variations in the weather have a significant impact on demand.



Model evaluation

Built 5 different models namely model1, model2, model3, model4 and model5 using polynomial terms, interaction terms, and regularizations

- This is a visual representation of the built-model comparison between R-squared and RMSE. For each model, the RMSE is displayed by peach bars, while the R-squared is displayed by teal blue bars.



Find the best performing model

- ▶ Model with high R-squared and low RMSE is known as best performing model.
- ▶ From this comparison chart we can analyze that the best performing model is model 2 with rmse 326 and rsq 0.736.

Formula of best performing model

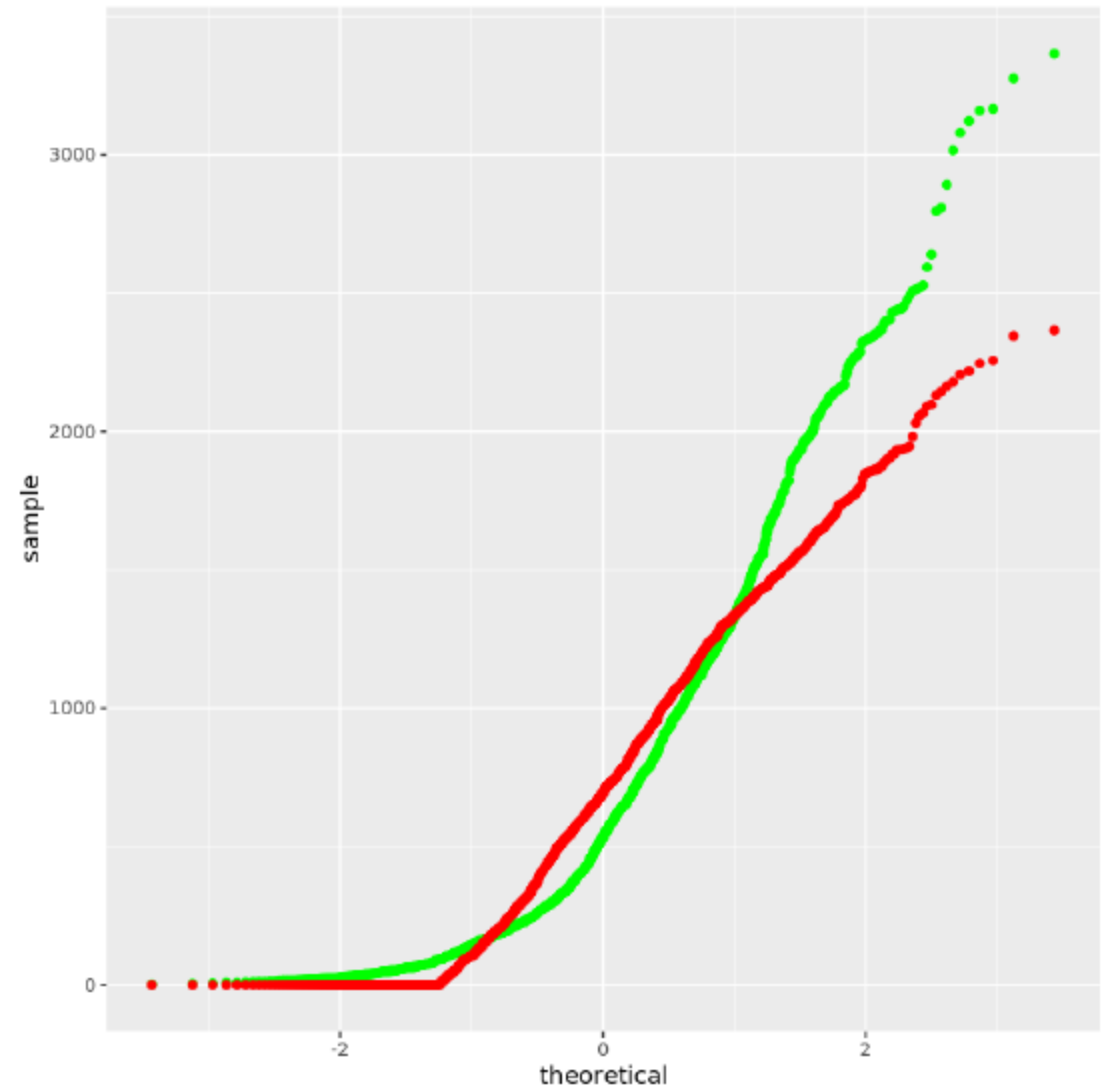
RENTED_BIKE_COUNT ~ . + poly(TEMPERATURE, 6) + poly(HUMIDITY,4) +
poly(SNOWFALL, 5) + HUMIDITY*TEMPERATURE + SUMMER* HOLIDAY + WINTER *
HOLIDAY)

Q-Q plot of the best model

Plot the test results of the best model against the truths in a Q-Q plot.

The discrepancy between the true values on the test dataset and the predictions made by the best-performing model (model 2) is depicted in the Q-Q figure.

Green indicates the actual values of the test data, while Red indicates the test results.

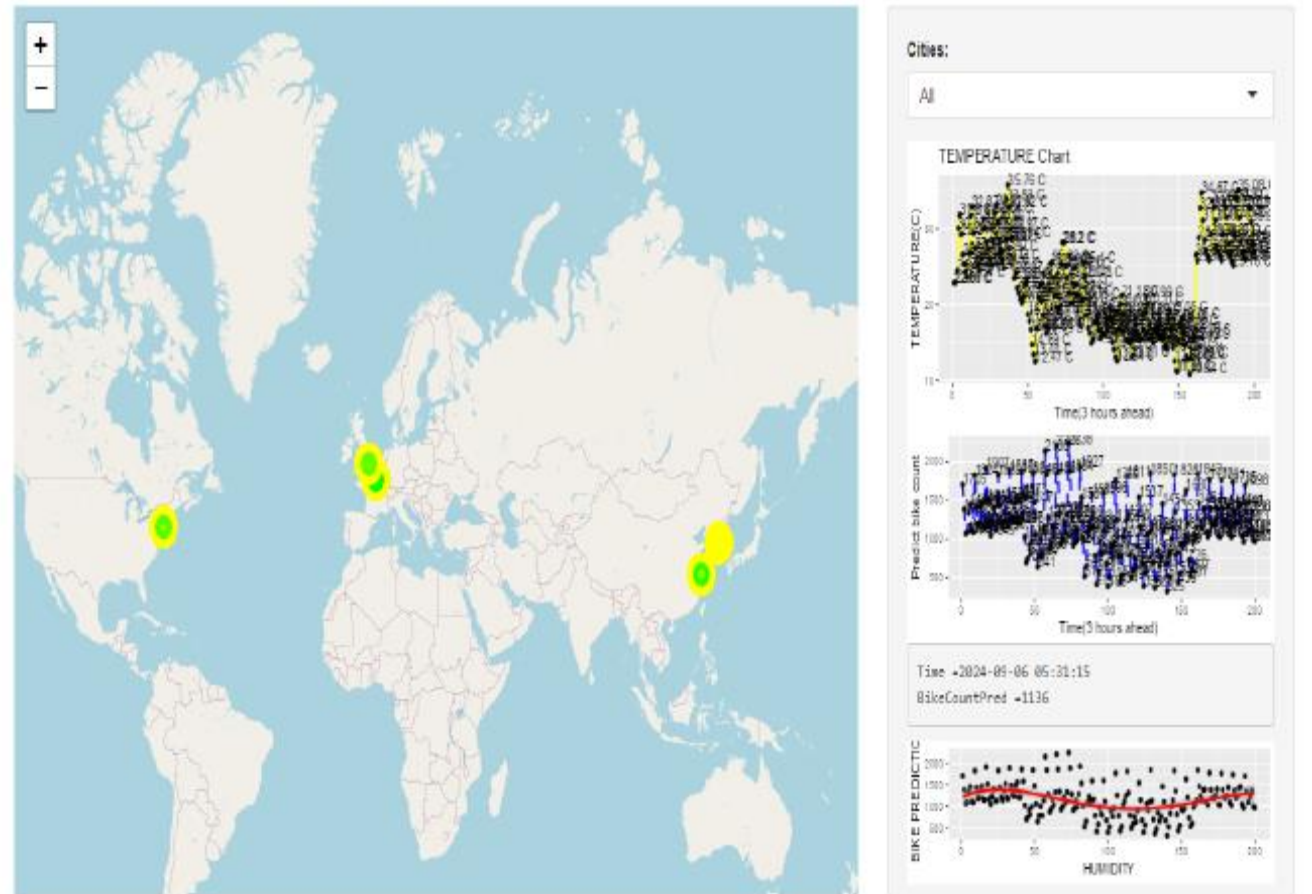


Dashboard

Cities' max bike-sharing Prediction

- ▶ The chosen cities are indicated on the leaflet map by the circle marks. Since we chose "All" in this case, all five cities are displayed.
- ▶ Next, there is a temperature chart with a yellow line representing the temperature trend three hours ahead of current time, and another blue line representing the demand prediction trend for bike sharing for the next three hours across all cities.
- ▶ When we click on the points on the bike prediction line, a text box next to the bike-sharing demand prediction trend chart displays the time and corresponding bike prediction.
- ▶ A correlation plot illustrating the association between bike prediction and humidity is displayed in the final image.

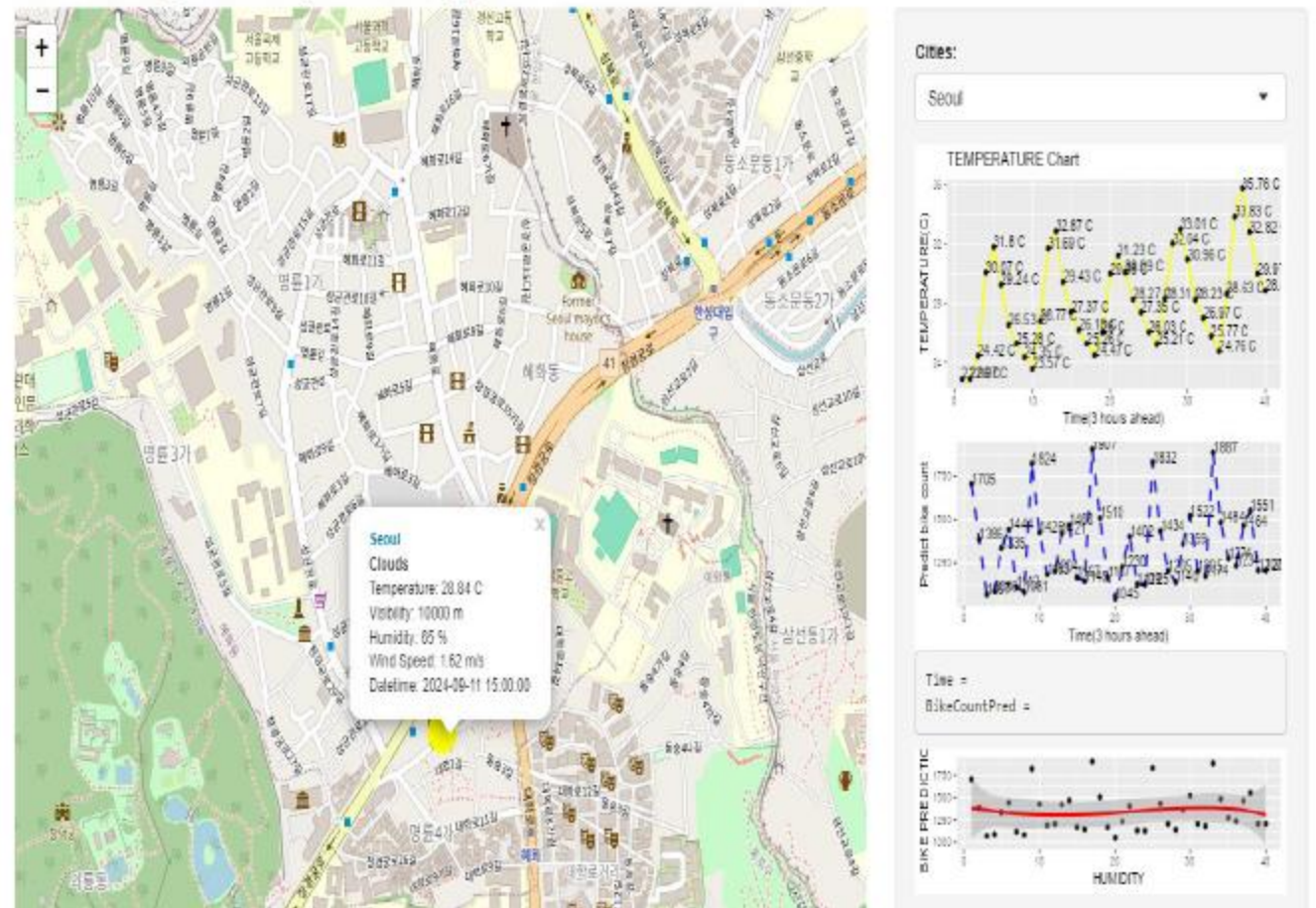
Bike-sharing demand prediction app



Bike-sharing Prediction of Seoul

- ▶ The chosen city, "Seoul," is marked on the pamphlet map with a circle and a thorough label.
- ▶ Next, there is a temperature chart with a yellow line representing the temperature trend three hours ahead of current time and a blue line representing the demand prediction trend for bike sharing in Seoul for the ensuing three hours.
- ▶ When we click on the points on the bike prediction line, a text box next to the bike-sharing demand prediction trend chart displays the time and corresponding bike prediction.
- ▶ A correlation plot illustrating the association between bike prediction and humidity is displayed in the final image.

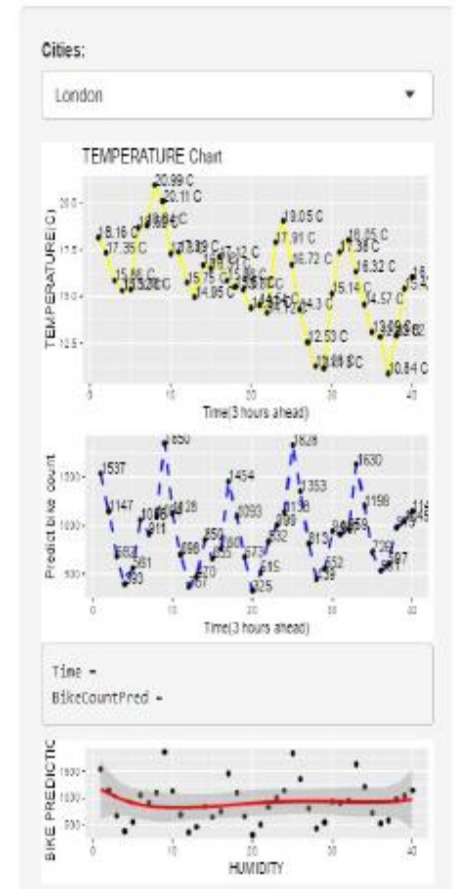
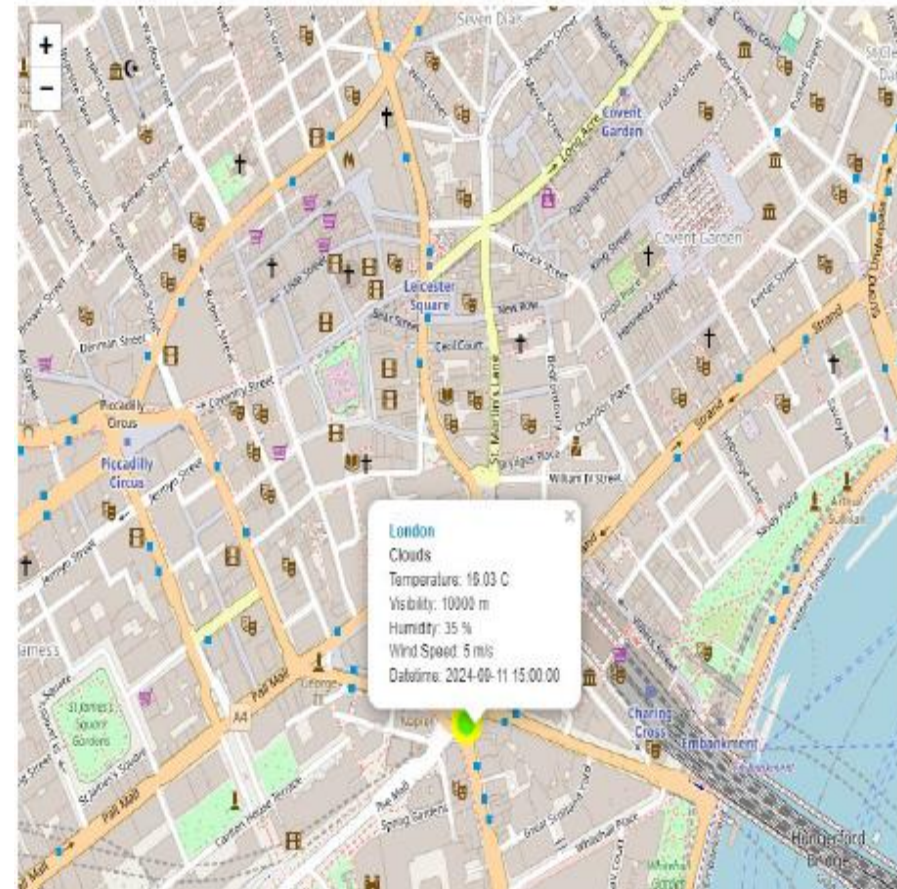
Bike-sharing demand prediction app



Bike_sharing Prediction of London

- ▶ The leaflet map's circle markers display the selected city, "London," along with a comprehensive label.
- ▶ Next, there is a temperature chart with a yellow line representing the temperature trend three hours ahead of current time, and another blue line representing the demand prediction trend for bike sharing in London for the next three hours.
- ▶ When we click on the points on the bike prediction line, a text box next to the bike-sharing demand prediction trend chart displays the time and corresponding bike prediction.
- ▶ A correlation plot illustrating the association between bike prediction and humidity is displayed in the final image.

Bike-sharing demand prediction app

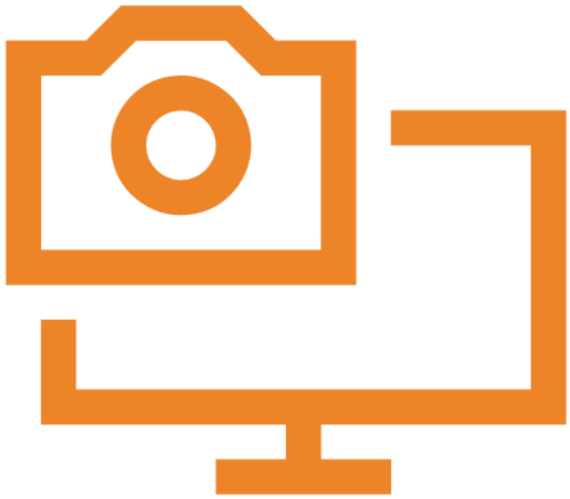


CONCLUSION



- ▶ In summary, the results of our investigation indicate that a variety of elements, including temperature, humidity, seasons, and rainfall, affect the demand for bike sharing.
- ▶ Out of all the elements determining bike rentals, the best influencing factors are season, humidity, temperature, and hour.
- ▶ Bicycle rentals peak in the summer and peak again in the fall and spring. Bike rentals are at an all-time low throughout the winter.
- ▶ Thus, it can be understood that more motorcycles should be made available in the summer and fewer in the winter.
- ▶ Time-related variables The hour that has the biggest impact is 6 PM, when bike rentals are in high demand.

APPENDIX



- ▶ Submit any pertinent files you may have developed for this project, such as charts, R code snippets, SQL queries, and Notebook outputs.

Thank You