Soul Bike Sharing Prediction Report

Created By Alice Fu

2024/12/8

CONTENT

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

❖ Rising Importance of Bike Sharing:

Over the past few decades, bike sharing has become increasingly significant as more people seek healthier and more livable cities where such activities are readily accessible.

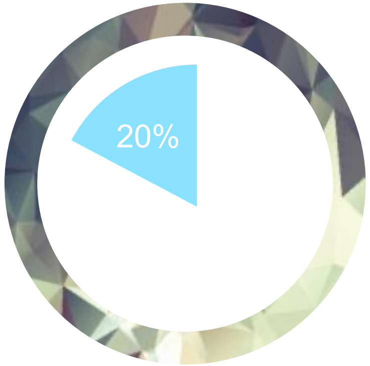
❖ Prediction of bike rental number：

- We discovered that a polynomial model with more terms and interactions, achieved the best performance.
- Key factors include temperature, rainfall, humidity, peak hour, and weekdays, indicating weather's significant impact on bike rentals.

❖ Dataset Overview:

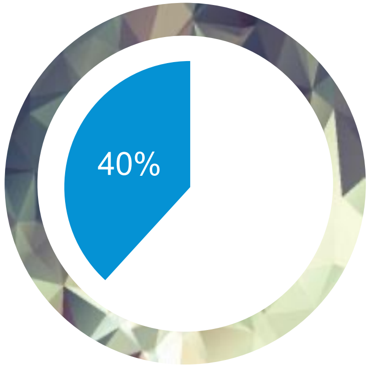
The dataset encompasses weather information (including temperature, humidity, windspeed, etc.), hourly bike rental counts, and date details for the Capital bike share system from 2017 to 2018.

INTRODUCTION



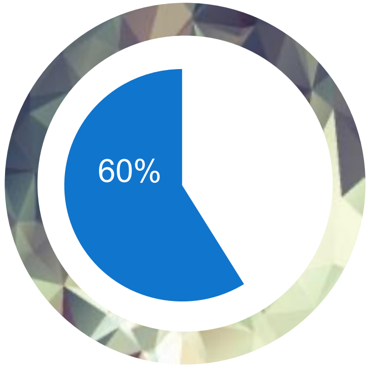
Problems Defining

It is important for each of these cities to provide a reliable supply of rental bikes to optimize accessibility at all times.



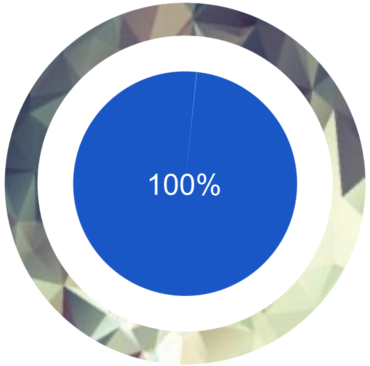
Solid Background

The Global Bike Sharing Cities Dataset is an HTML table on the Wikipedia page List of bicycle-sharing systems



Supportive Tool

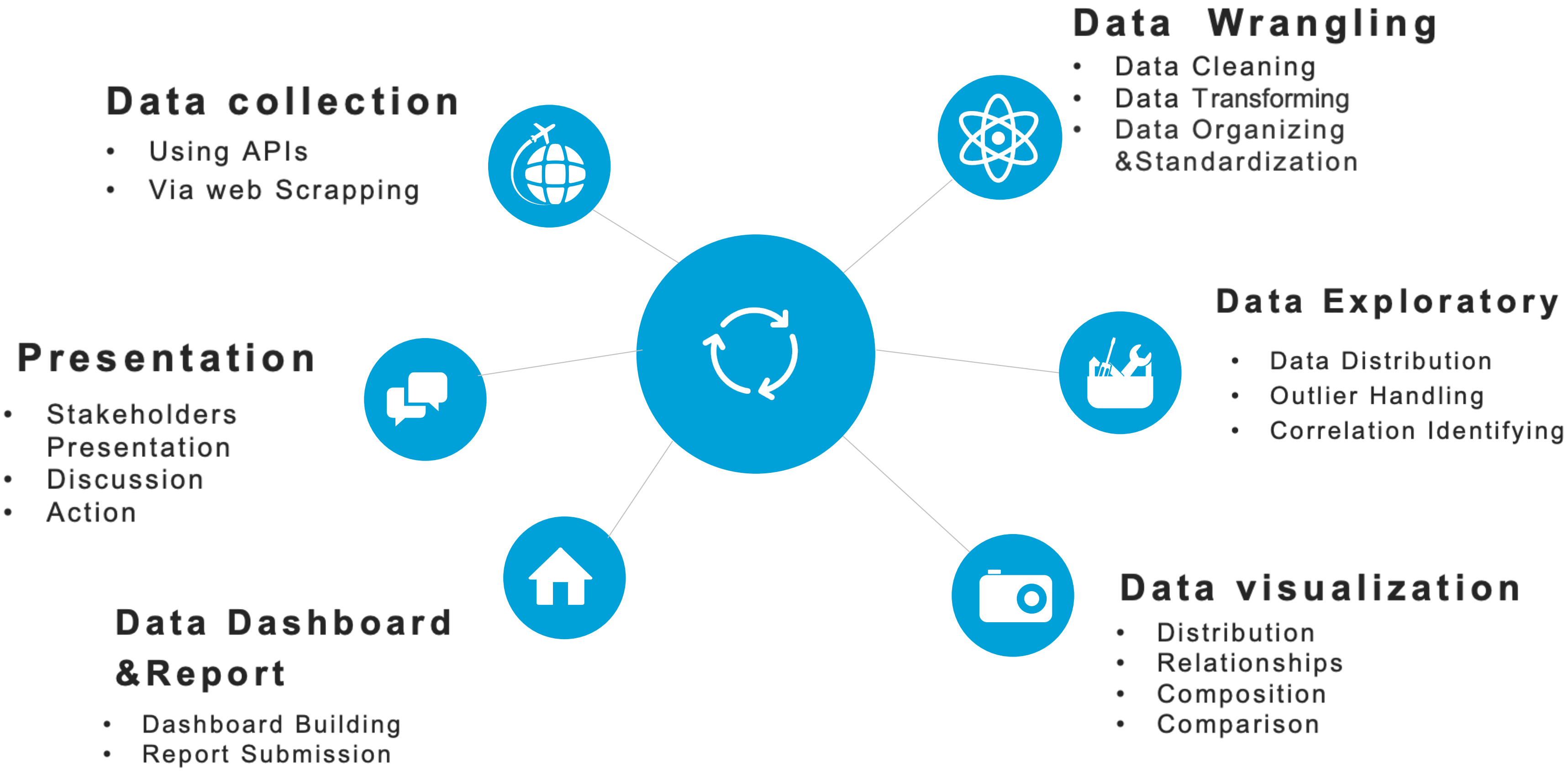
The Open Weather API allows users to access current and forecasted weather data for any location including over 200,000 cities.



The Goal

Minimizing program costs, including bike supply to meet demand, is important. Predicting hourly bike needs based on weather helps optimize supply.

METHODOLOGY



Data collection

- Using APIs
- Via web Scrapping

Data Wrangling

- Data Cleaning
- Data Transforming
- Data Organizing &Standardization

Data Exploratory

- Data Distribution
- Outlier Handling
- Correlation Identifying

Data visualization

- Distribution
- Relationships
- Composition
- Comparison

Data Dashboard &Report

- Dashboard Building
- Report Submission

Presentation

- Stakeholders Presentation
- Discussion
- Action

RESULTS

DATA WRANGLING WITH DPLYR

Wrangling the Seoul bike-sharing demand historical dataset using Dplyr

1 Summary

8760 rows

14 columns

2 Detect and handle missing values

- Detect and handle missing values in RENTED_BIKE_COUNT
- Impute missing values for the TEMPERATURE column using its mean value

3 Create indicator

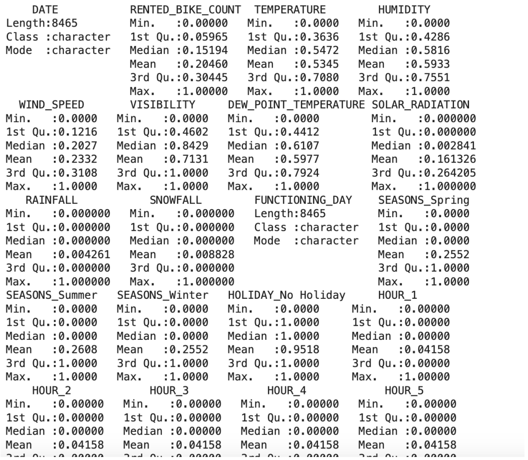
- Convert HOUR column from numeric into character
- Convert SEASONS, HOLIDAY, FUNCTIONING_DAY, and HOUR columns into indicator columns

4 Normalize data

Apply min-max normalization on RENTED_BIKE_COUNT, TEMPERATURE, HUMIDITY, WIND_SPEED, VISIBILITY, DEW_POINT_TEMPERATURE, SOLAR_RADIATION, RAINFALL, SNOWFALL

5 Standardize the column names

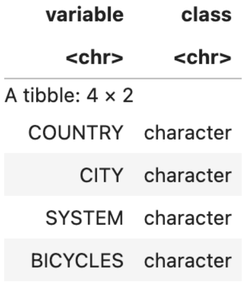
standardize their column names of the related files



```
     DATE           RENTED_BIKE_COUNT  TEMPERATURE        HUMIDITY     
 Length:8465        Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
 Class :character   1st Qu.:0.05965   1st Qu.:0.3636   1st Qu.:0.4286  
 Mode  :character   Median :0.15194   Median :0.5472   Median :0.5816  
                    Mean   :0.20460   Mean   :0.5345   Mean   :0.5933  
                    3rd Qu.:0.30445   3rd Qu.:0.7080   3rd Qu.:0.7551  
                    Max.   :1.00000   Max.   :1.0000   Max.   :1.0000  
   WIND_SPEED       VISIBILITY     DEW_POINT_TEMPERATURE SOLAR_RADIATION   
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000        Min.   :0.000000  
 1st Qu.:0.1216   1st Qu.:0.4602   1st Qu.:0.4412        1st Qu.:0.000000  
 Median :0.2027   Median :0.8429   Median :0.6107        Median :0.002841  
 Mean   :0.2332   Mean   :0.7131   Mean   :0.5977        Mean   :0.161326  
 3rd Qu.:0.3108   3rd Qu.:1.0000   3rd Qu.:0.7924        3rd Qu.:0.264205  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000        Max.   :1.000000  
    RAINFALL            SNOWFALL        FUNCTIONING_DAY    SEASONS_Spring  
 Min.   :0.000000   Min.   :0.000000   Length:8465        Min.   :0.0000  
 1st Qu.:0.000000   1st Qu.:0.000000   Class :character   1st Qu.:0.0000  
 Median :0.000000   Median :0.000000   Mode  :character   Median :0.0000  
 Mean   :0.004261   Mean   :0.008828                      Mean   :0.2552  
 3rd Qu.:0.000000   3rd Qu.:0.000000                      3rd Qu.:1.0000  
 Max.   :1.000000   Max.   :1.000000                      Max.   :1.0000  
 SEASONS_Summer   SEASONS_Winter   HOLIDAY_No Holiday     HOUR_1       
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000     Min.   :0.00000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000     1st Qu.:0.00000  
 Median :0.0000   Median :0.0000   Median :1.0000     Median :0.00000  
 Mean   :0.2608   Mean   :0.2552   Mean   :0.9518     Mean   :0.04158  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000     3rd Qu.:0.00000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000     Max.   :1.00000  
     HOUR_2            HOUR_3            HOUR_4            HOUR_5       
 Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000  
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000  
 Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000  
 Mean   :0.04158   Mean   :0.04158   Mean   :0.04158   Mean   :0.04158  
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000  
```

DATA WRANGLING WITH REGULAR EXPRESSIONS

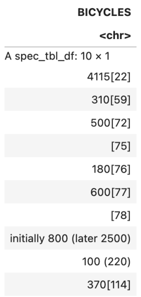
Clean up the bike-sharing systems data using Tidyverse



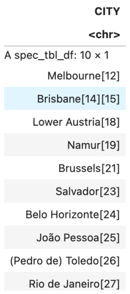
variable	class
<chr>	<chr>
A tibble: 4 × 2	
COUNTRY	character
CITY	character
SYSTEM	character
BICYCLES	character

1 Type Check

Character Type



BICYCLES
<chr>
A spec_tbl_df: 10 × 1
4115[22]
310[59]
500[72]
[75]
180[76]
600[77]
[78]
initially 800 (later 2500)
100 (220)
370[114]



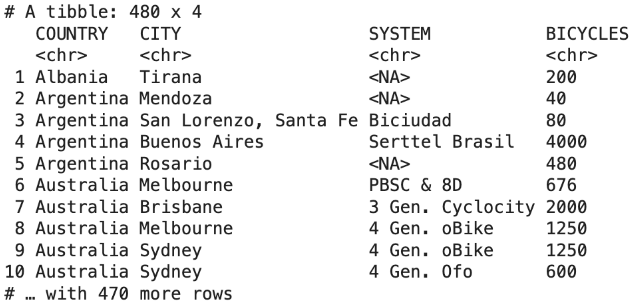
CITY
<chr>
A spec_tbl_df: 10 × 1
Melbourne[12]
Brisbane[14][15]
Lower Austria[18]
Namur[19]
Brussels[21]
Salvador[23]
Belo Horizonte[24]
João Pessoa[25]
(Pedro de) Toledo[26]
Rio de Janeiro[27]

2 Remove undesired reference links

Find any elements in the column containing non-numeric characters

3 Remove reference links

Use the dplyr::mutate() function to apply the remove_ref function to the CITY and SYSTEM columns



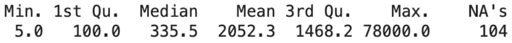
```
# A tibble: 480 x 4
   COUNTRY   CITY                   SYSTEM           BICYCLES
   <chr>     <chr>                  <chr>            <chr>
 1 Albania   Tirana                 <NA>             200
 2 Argentina Mendoza                <NA>             40
 3 Argentina San Lorenzo, Santa Fe Biciudad         80
 4 Argentina Buenos Aires           Serttel Brasil   4000
 5 Argentina Rosario                <NA>             480
 6 Australia Melbourne              PBSC & 8D        676
 7 Australia Brisbane               3 Gen. Cyclocity 2000
 8 Australia Melbourne              4 Gen. oBike     1250
 9 Australia Sydney                 4 Gen. oBike     1250
10 Australia Sydney                 4 Gen. Ofo       600
# … with 470 more rows
```

4 Extract the numeric value

Use the mutate() function to apply extract number on the BICYCLES column

5 Summary

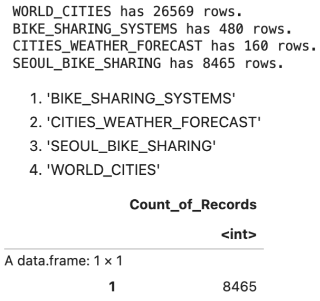
Use the summary function to check the descriptive statistics of the numeric BICYCLES column



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
    5.0   100.0   335.5  2052.3  1468.2 78000.0     104 
```

EDA WITH SQL

Perform exploratory data analysis using SQL queries with the RSQLite R



```
WORLD_CITIES has 26569 rows.
BIKE_SHARING_SYSTEMS has 480 rows.
CITIES_WEATHER_FORECAST has 160 rows.
SEOUL_BIKE_SHARING has 8465 rows.
```

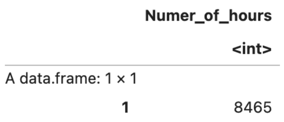
1. 'BIKE_SHARING_SYSTEMS'
2. 'CITIES_WEATHER_FORECAST'
3. 'SEOUL_BIKE_SHARING'
4. 'WORLD_CITIES'

A data.frame: 1 × 1

	Count_of_Records <int>
1	8465

1 Record Count

8465 rows

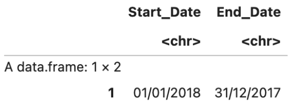


A data.frame: 1 × 1

	Numer_of_hours <int>
1	8465

2 Operational Hours

- Determine how many hours had non-zero rented bike count

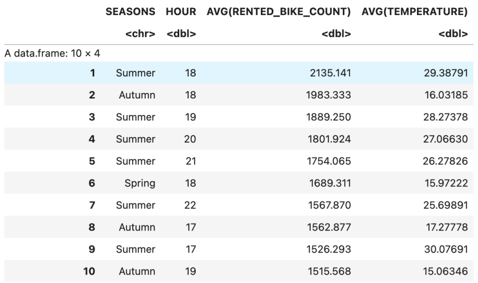


A data.frame: 1 × 2

	Start_Date <chr>	End_Date <chr>
1	01/01/2018	31/12/2017

3 Weather Outlook

- Query the 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.



A data.frame: 10 × 4

	SEASONS <chr>	HOUR <dbl>	AVG(RENTED_BIKE_COUNT) <dbl>	AVG(TEMPERATURE) <dbl>
1	Summer	18	2135.141	29.38791
2	Autumn	18	1983.333	16.03185
3	Summer	19	1889.250	28.27378
4	Summer	20	1801.924	27.06630
5	Summer	21	1754.065	26.27826
6	Spring	18	1689.311	15.97222
7	Summer	22	1567.870	25.69891
8	Autumn	17	1562.877	17.27778
9	Summer	17	1526.293	30.07691
10	Autumn	19	1515.568	15.06346

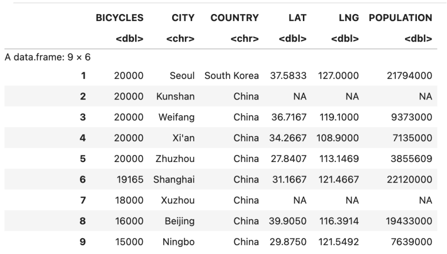
4 Popularity Explore

•Determine which date and hour had the most bike rentals.

•Determine the average hourly temperature and the average number of bike rentals per hour over each season. List the top ten results by average bike count.

•Find the average hourly bike count during each season.

•Consider the weather over each season



A data.frame: 9 × 6

	BICYCLES <dbl>	CITY <chr>	COUNTRY <chr>	LAT <dbl>	LNG <dbl>	POPULATION <dbl>
1	20000	Seoul	South Korea	37.5833	127.0000	21794000
2	20000	Kunshan	China	NA	NA	NA
3	20000	Weifang	China	36.7167	119.1000	9373000
4	20000	Xi'an	China	34.2667	108.9000	7135000
5	20000	Zhuzhou	China	27.8407	113.1469	3855609
6	19165	Shanghai	China	31.1667	121.4667	22120000
7	18000	Xuzhou	China	NA	NA	NA
8	16000	Beijing	China	39.9050	116.3914	19433000
9	15000	Ningbo	China	29.8750	121.5492	7639000

5 Summary

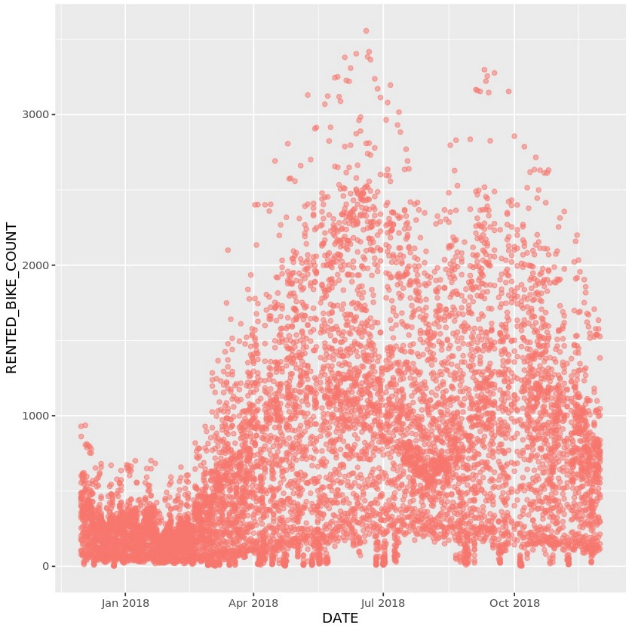
•Total Bike Count and City Info for Seoul

•Find all city names and coordinates with comparable bike scale to Seoul's bike sharing system

EDA WITH VISUALIZATION

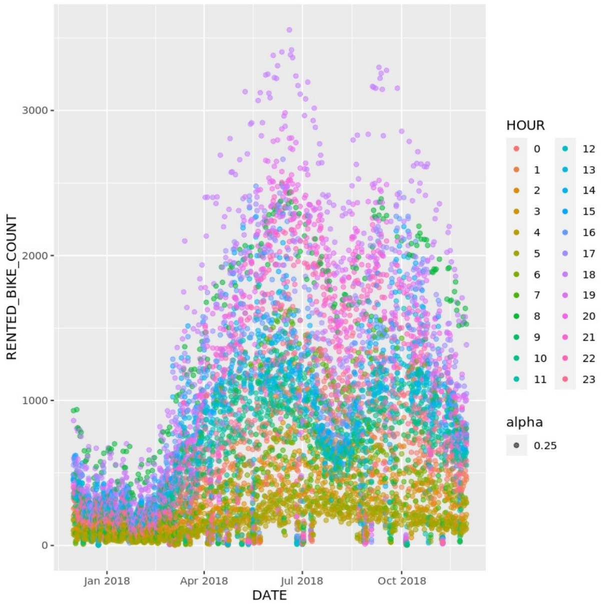
Perform exploratory data analysis using Visualization with tidyverse & ggplot

Create a scatter plot of RENTED_BIKE_COUNT vs DATE



We can see the rented bike count start to increase around FEB/March and reach the **max on June** then decrease little bit towards AUG then increase around **SEP** and then start decreasing again towards the end of the year.

Create the same plot of the RENTED_BIKE_COUNT time series but now add HOURS as the colour.

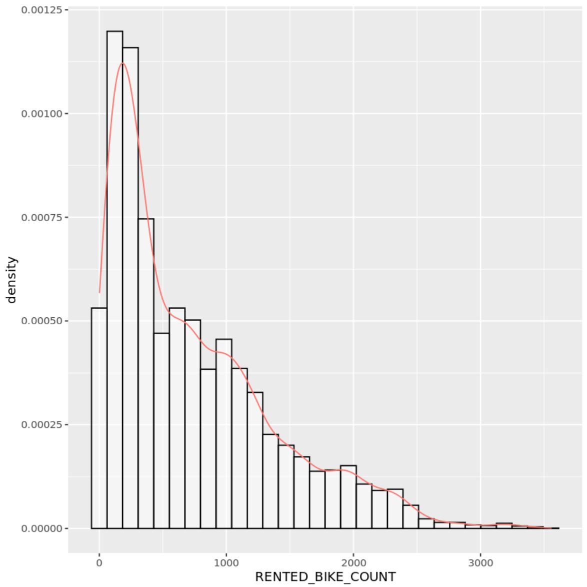


We can see the rented bike count are to low at the dawn and start to increase slowly during the early hours of the morning to reach to max at the evening in 6 or 7 then start decreasing again.

EDA WITH VISUALIZATION

Perform exploratory data analysis using Visualization with tidyverse & ggplot

Create a histogram overlaid with a kernel density curve

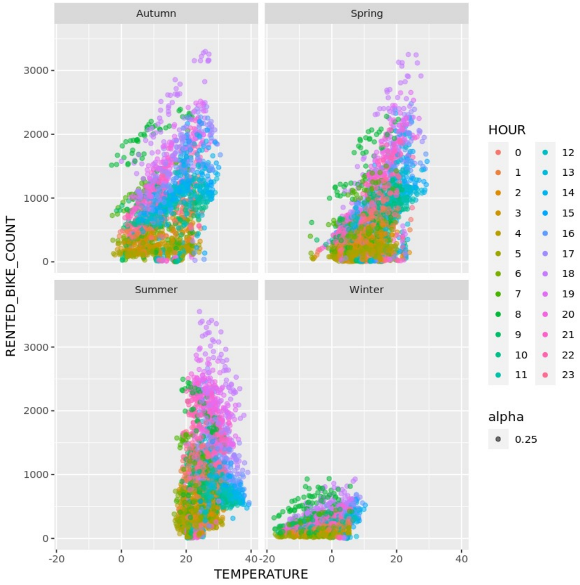


- Histogram Observation: Most times, few bikes are rented; the mode is about 250.
- Modes in Subgroups: Bumps at 700, 900, 1900, and 3200 bikes suggest hidden modes in subgroups.
- Rare Occasions: Occasionally, many more bikes are rented than usual.

EDA WITH VISUALIZATION

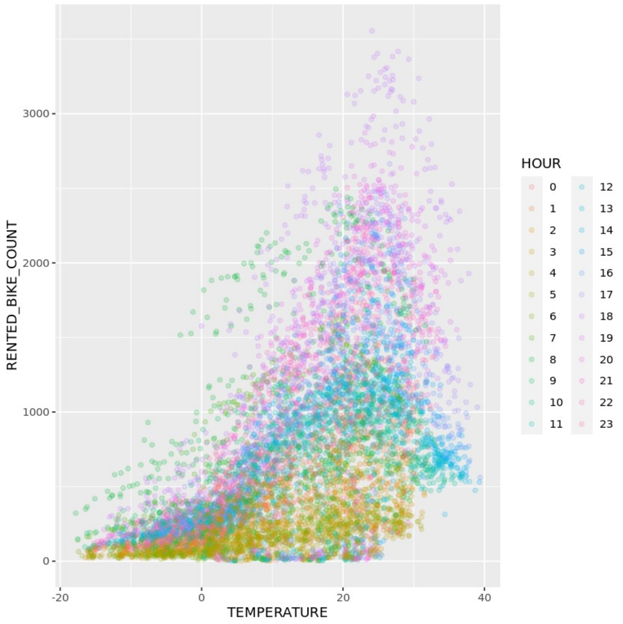
Perform exploratory data analysis using Visualization with tidyverse & ggplot

Correlation between two variables (RENTED_BIKE_COUNT and TEMPERATURE by SEASONS)



Visually, strong correlations are evident as approximately linear patterns.
Autumn & Spring: Similar bike usage patterns with temperatures between 0-20°C; higher usage in warmer weather.
Summer: Consistent usage hours but reduced bike counts in hotter weather.
Winter: Significant drop in bike rentals, with a max of 1000 bikes; peak usage in early morning and evening (6-7 PM).

Create a scatter plot of RENTED_BIKE_COUNT vs TEMPERATURE by Hour as Color

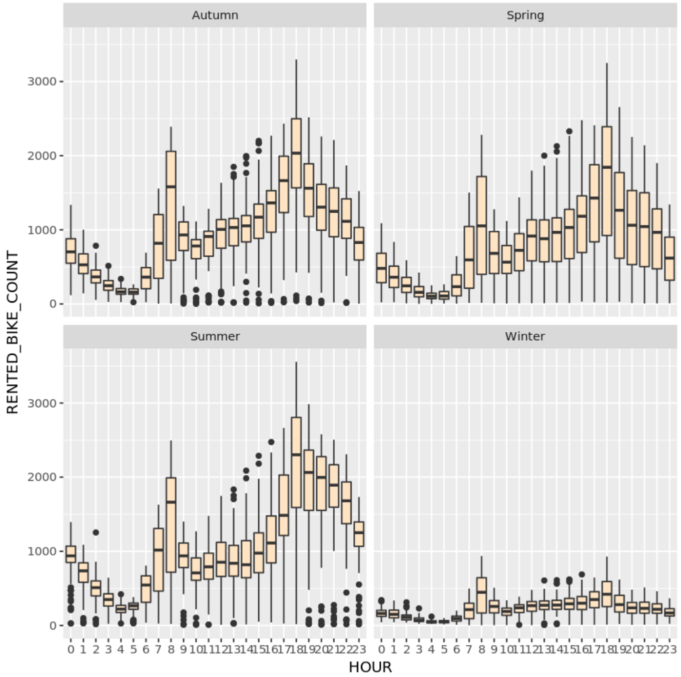


higher usage in warmer weather with peak hour in morning and evening 6-7PM.

EDA WITH VISUALIZATION

Perform exploratory data analysis using Visualization with tidyverse & ggplot

Create a display of four boxplots of RENTED_BIKE_COUNT vs. HOUR grouped by SEASONS



Seasonal Variations:
Bike rental counts vary by season but key features remain similar.

Peak Demand:
Peak demand times are consistent across all seasons, at 8 am and 6 pm.

Outliers in Data:
Many outliers in bike count data during different seasons.

Usage Patterns:
People generally use bikes at similar times in different seasons, with slight variations in counts.

Winter Drop:
Significant drop in bike rentals during Winter.

Predict Hourly Rented Bike Count using Basic Linear Regression Models

1 **Data Split**

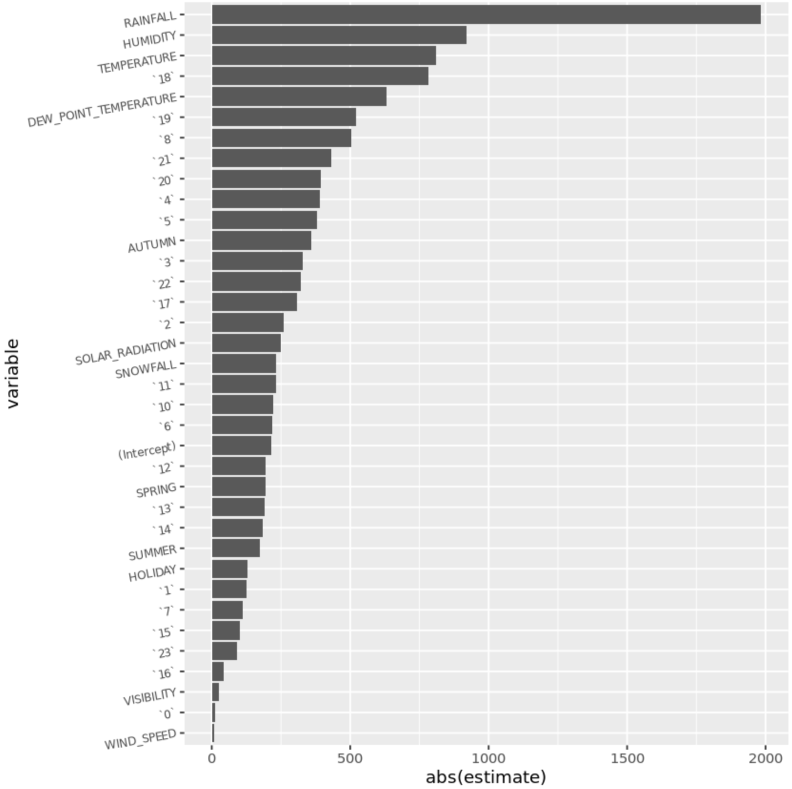
Split data into training and testing datasets

2 **Model Build**

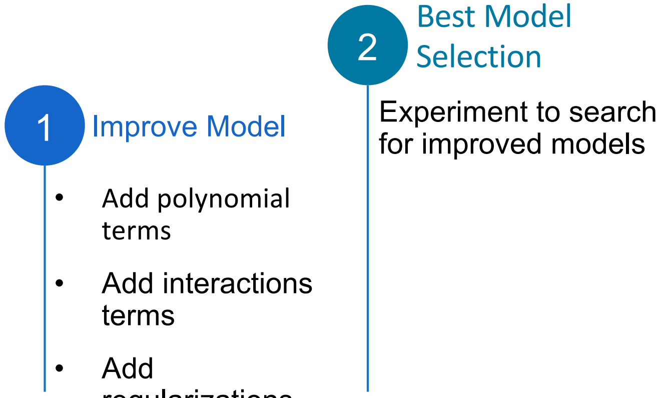
- Build a linear regression model using only the weather variables
- Build a linear regression model using both weather and date variables

3 **Coefficient Identity**

Evaluate the models and identify important variables



Refine the Baseline Regression Models



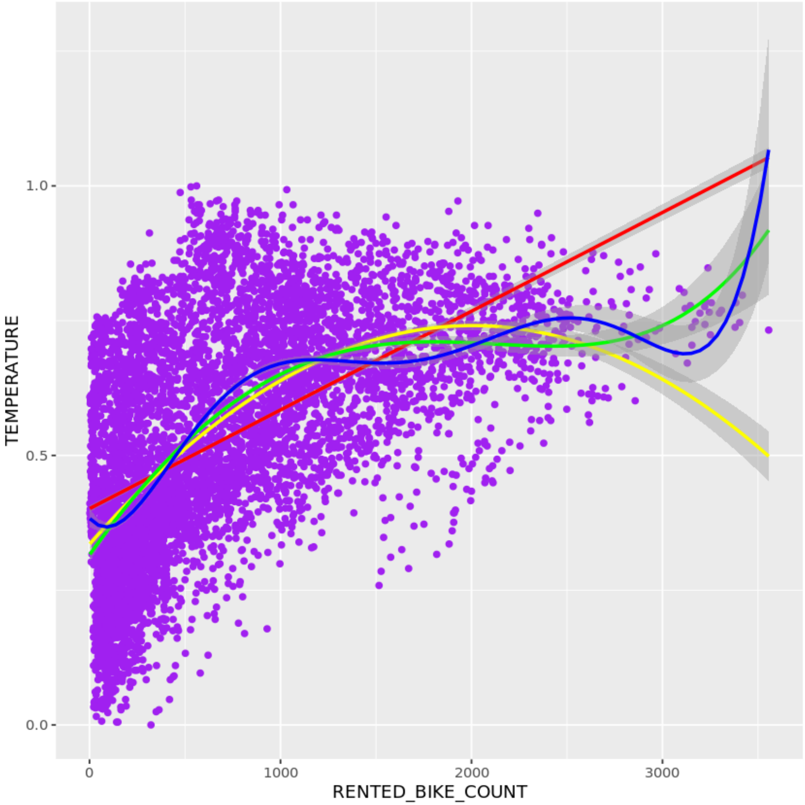
Improve Model

- Add polynomial terms
- Add interactions terms
- Add regularizations terms

Best Model Selection

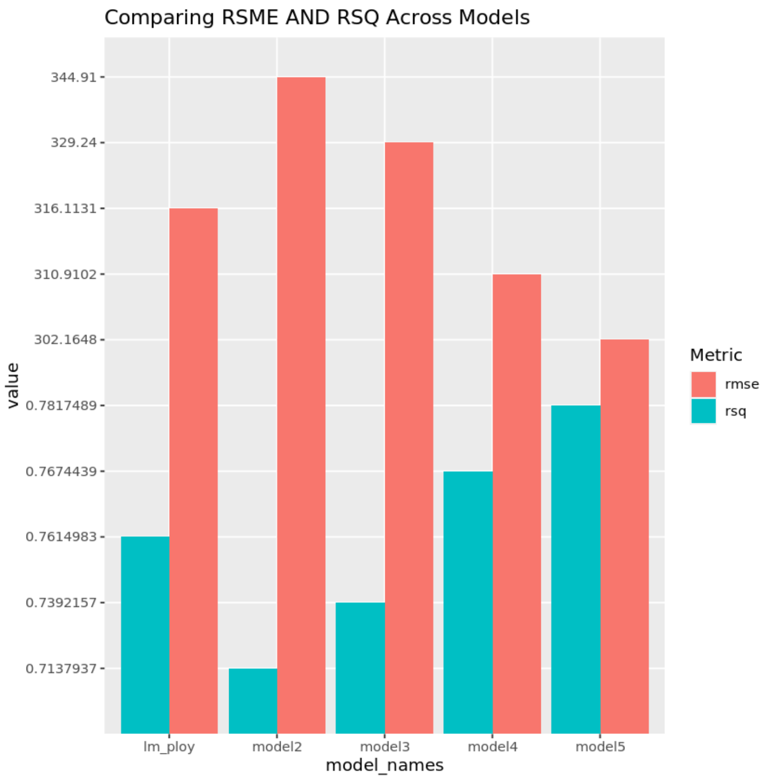
Experiment to search for improved models

Correlation between RENTED_BIKE_COUNT and TEMPERATURE with the higher order polynomial fits

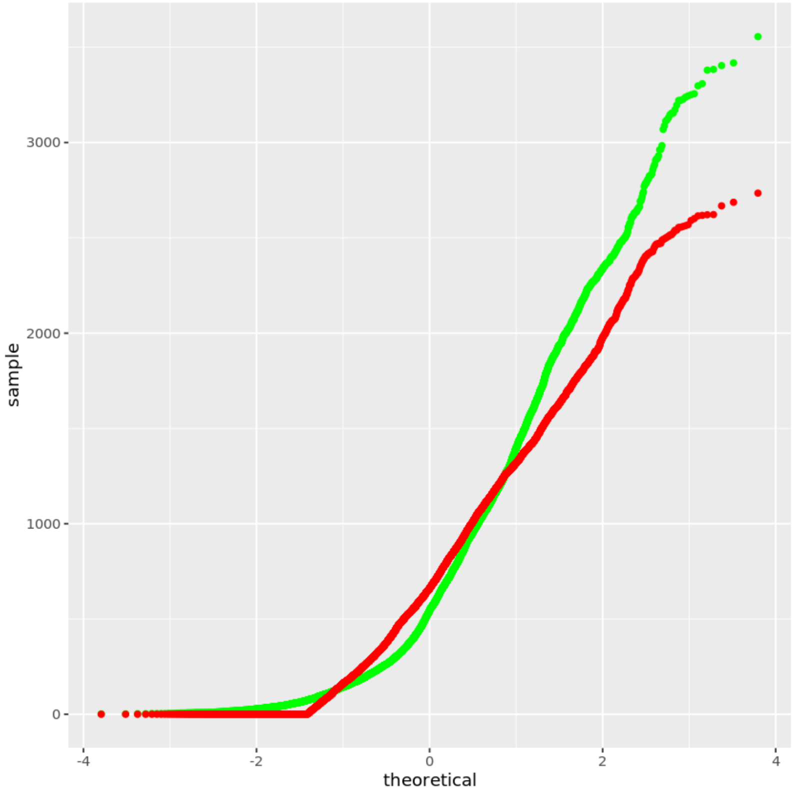


Refine the Baseline Regression Models

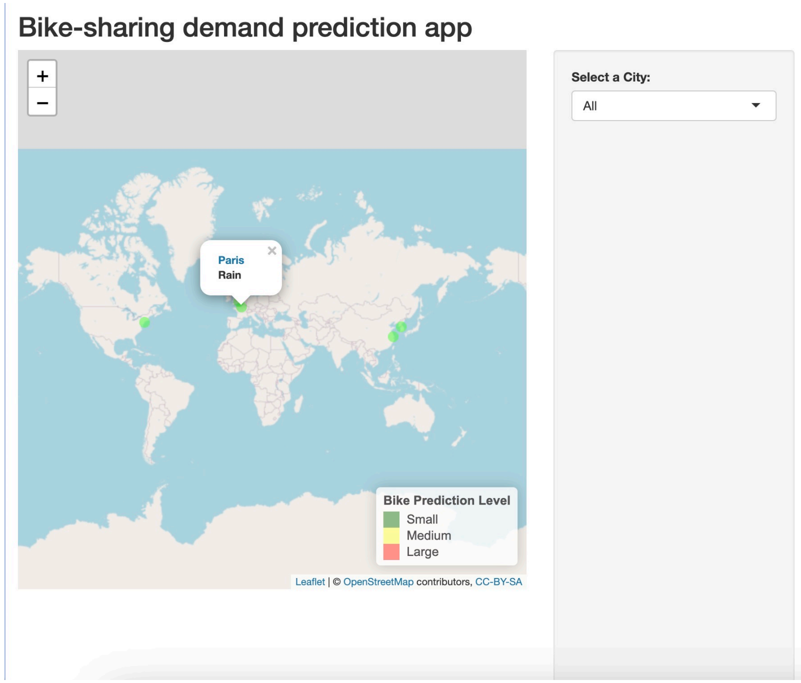
Model 5 Reported the best performed model in terms of rmse and rsq



Plotting the distribution difference between the predictions generated by your best model vs the true values on test dataset



R Shiny dashboard



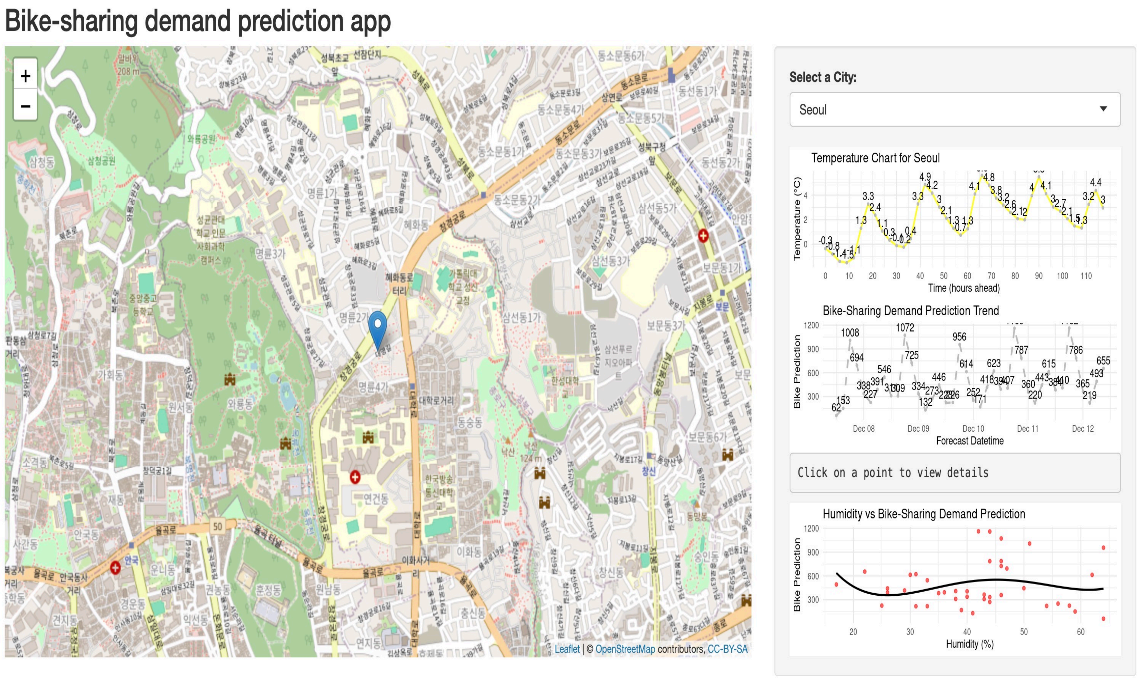
Bike-sharing demand prediction app

Select a City:

All

Bike Prediction Level

- 🟩 Small
- 🟨 Medium
- 🟥 Large



Bike-sharing demand prediction app

Select a City:

Seoul

Click on a point to view details

+
−
Temperature Chart for Seoul
Temperature (°C)
Time (hours ahead)
Bike-Sharing Demand Prediction Trend
Bike Prediction
Forecast Datetime
Dec 08
Dec 09
Dec 10
Dec 11
Dec 12
Humidity vs Bike-Sharing Demand Prediction
Bike Prediction
Humidity (%)
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Discussion

FINDINGS

Weekday vs. Weekend: Bike rental count is higher during weekdays than weekends.

Peak Hours: Rental bike counts peak at 8 AM and 6-7 PM, with demand gradually increasing from 5 AM to 8 AM, then dipping, and rising again until 6-7 PM.

IMPLICATIONS

Temperature & Wind: People prefer renting bikes at moderate to high temperatures and even with light winds, suggesting a need for comfortable weather conditions.

Seasonal Trends: Highest bike rentals in Autumn and Summer, lowest in Winter, indicating seasonal preferences.

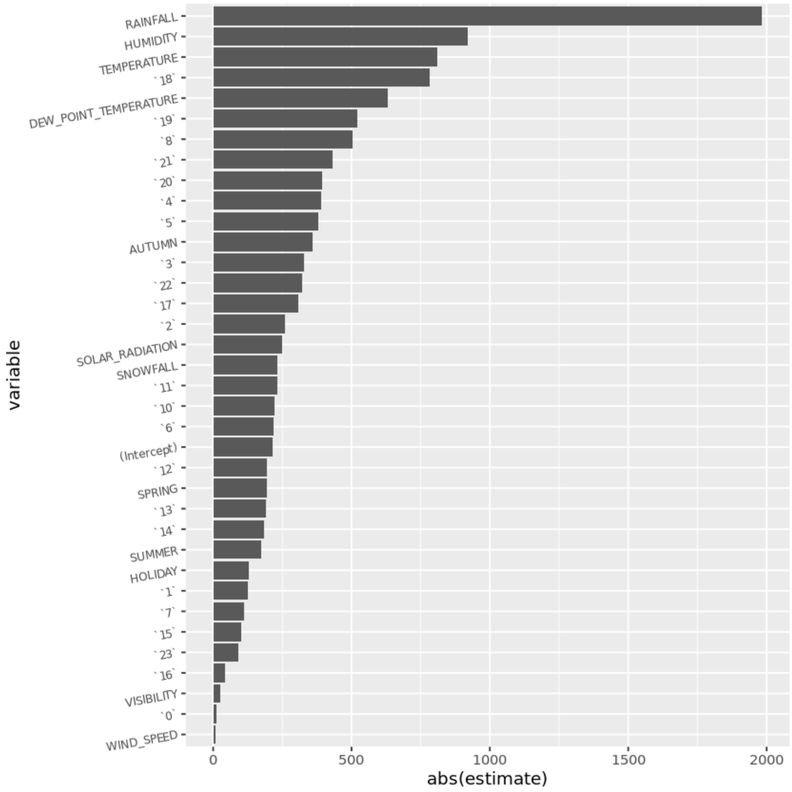
Weather Conditions: Bike rentals are highest on clear days and lowest on snowy and rainy days, impacting rental decisions.

Conclusion

- These observations suggest that bike rentals are influenced by various factors including time of day, weather conditions, hours and seasons.
- Understanding these patterns can help bike rental companies optimize their services and better meet customer demand.
- Bike rentals peak during morning and evening commutes. Bike rental services should place more bikes at popular stations and increase redistribution frequency during these periods.
- They may consider increasing the availability of bikes during peak hours and seasons, and adjusting prices based on weather conditions to attract more customers. This ensures smoother commutes, enhances user satisfaction, and encourages continued bike usage.

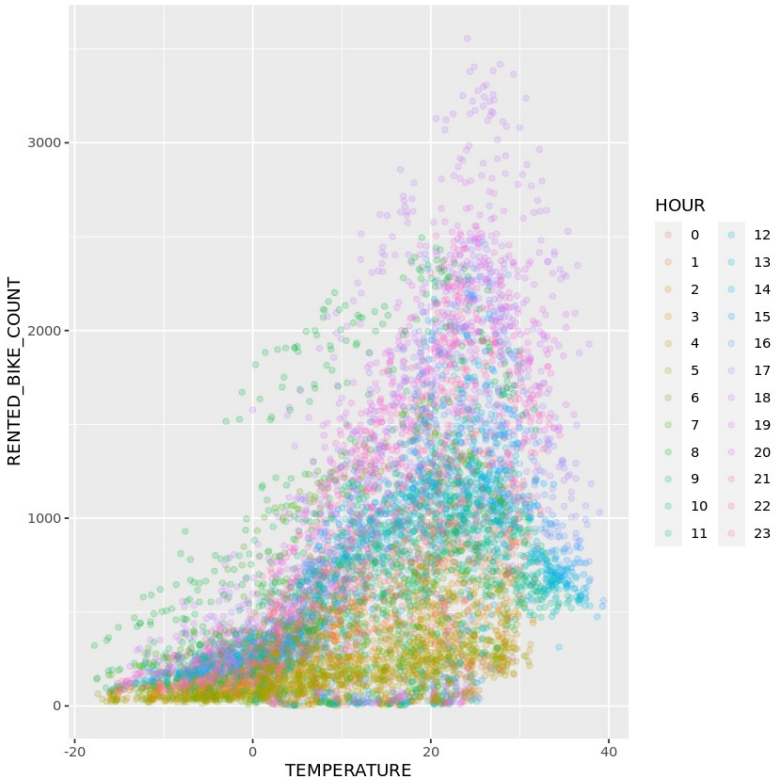
Appendix

Mark down the 'top-ranked variables by coefficient'



Appendix

Create a scatter plot of RENTED_BIKE_COUNT vs TEMPERATURE by Hour as Color



SCOURCE

Weather & Bike-Sharing

OpenWeather APIs Calls -Current & Prediction of Weather
https://home.openweathermap.org/users/sign_up

Web scrape a Global Bike-Sharing Systems Wiki Page
https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems

THANK YOU