Urban Bike-Sharing Forecasting and Optimization Worldwide

Thet Khine Lin

22 Jul 2024

Outline

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

Executive Summary

Introduction

- This project aims to predict the hourly rented bike count using weather and date information for Seoul, South Korea, and to build an interactive dashboard for visualizing weather forecast data and predicted hourly bike-sharing demand across five cities: Seoul, South Korea; New York, USA; Paris, France; Suzhou, China; and London, UK.

Objectives

- Predict hourly bike rental demand based on weather conditions and date information for Seoul.
- Develop an interactive dashboard to visualize forecasted weather data and bike-sharing demand in multiple cities.

Executive Summary

Methodology

- The methodology of this project encompasses a comprehensive approach, combining data collection, data wrangling, exploratory data analysis (EDA), the development of predictive models, and the creation of an interactive dashboard.

Recommendations/Implications

- City planners and bike-sharing companies can use the model from this project to optimize bike distribution and improve service availability in Seoul.
- The interactive dashboard can be used for real-time decision-making and strategic planning for renting bikes for the five cities: Seoul, South Korea; New York, USA; Paris, France; Suzhou, China; and London, UK.

Introduction

As a Data Scientist at an AI-powered weather data analytics company, I have undertaken a project to analyze the impact of weather on bike-sharing demand in urban areas. This project is divided into two main parts:

Part – 1 (Weather Impact Analysis on Bike-Sharing Demand in Seoul)

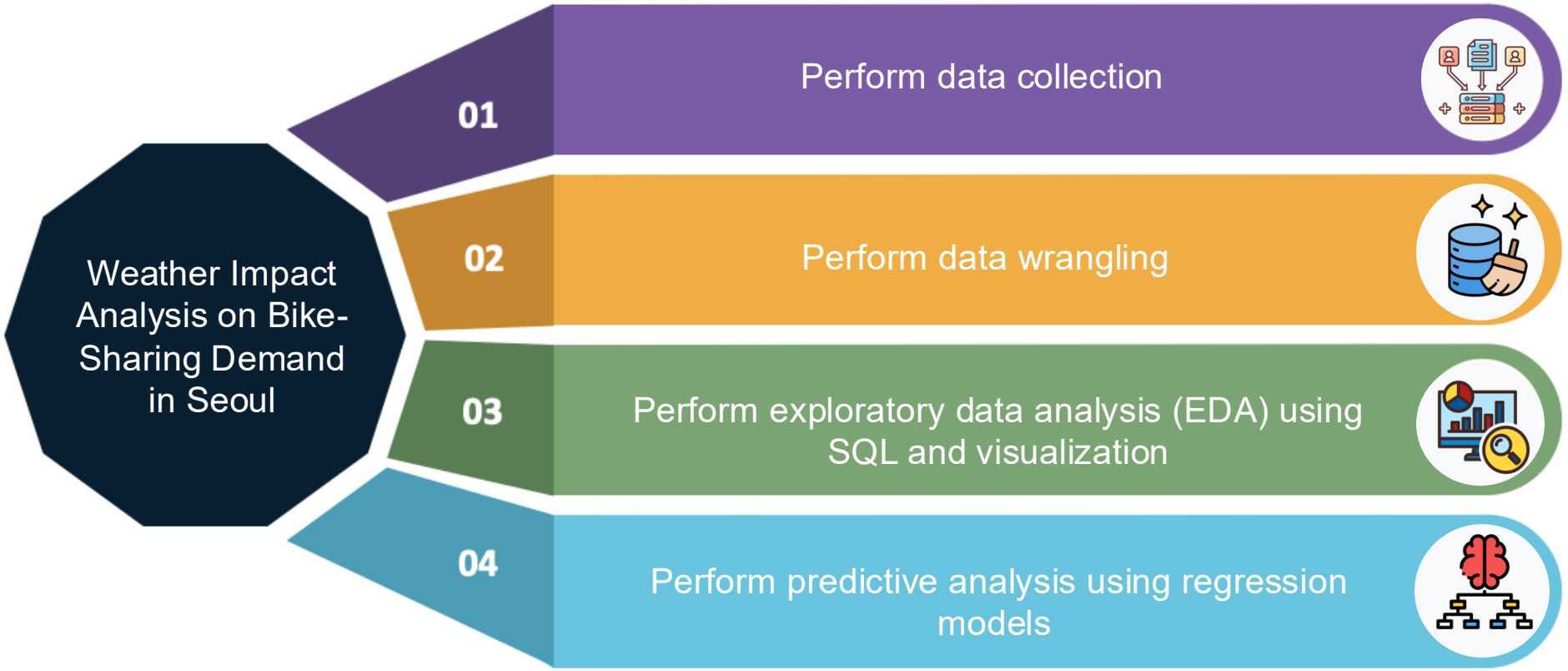
- Data Collection and Processing: Gathered and processed weather and bike-sharing demand data from various sources.
- Exploratory Data Analysis (EDA): Conducted EDA to uncover patterns and relationships in the data.
- Predictive Modeling: Developed predictive models to forecast bike-sharing demand in Seoul based on weather conditions.

Part – 2 (Interactive Dashboard for Global Cities:)

- Utilization of the Trained Model: Employed the trained model, which used Seoul data, to predict rental bike counts for five cities: Seoul, South Korea; New York, USA; Paris, France; Suzhou, China; and London, UK.
- Data Extraction: Extracted weather data from the Open Weather API.
- Integration of Results: Combined the predictive results with existing weather data.
- Dashboard Development: Created a live dashboard displaying:
    - Temperature trends for the next five days at 3-hour intervals.
    - Predicted bike counts for the next five days.
    - Predicted bike counts versus humidity for the next five days.

Methodology

Content of Part-1 Methodology



Data Collection

In the part-1 of this project, data are collected from three sources:

1. Global bike sharing systems data on public web pages
2. Cities Weather forecast data via OpenWeather APIs
3. Aggregated tabular data from cloud storage

Collecting Global bike sharing systems data

- Global bike sharing data contains a list of bicycle-sharing systems, both docked and dockless, around the world.
- It is collected from a Wikipedia page by extracting a required table and save it as data frame.

Collecting Global Weather Forecast Data

- Global Weather Forecast data contains predicted weather condition of specified cities for coming 5 days.
- It is collected from Open Weather API Map website by using an API key.

Collecting Seoul bike sharing and world cities data

- World cities data contains information such as name, latitude, and longitude, about major cities around the world.
- Seoul bike sharing data contains how many bicycles were rented at when.
- Both of them are collected by downloading directly from IBM Cloud Storage.

Data wrangling

- The data wrangling process includes handling missing data, normalizing categorical variables using the One-Hot-Encoding method, standardizing continuous values, and formatting data types and column names.
- These steps were performed using the "dplyr" package from the "tidyverse" and regular expressions in R.

EDA with SQL

Task	Performed Query
1.	Display number of records in Seoul Bike Sharing Dataset
2.	Display number of total hours, which has non-zero rented bike count
3.	Display forecasted weather condition for Seoul over next 3 hours
4.	List the seasons, which are included in Seoul Bike Sharing Dataset
5.	Display the first and last dates in Seoul Bike Sharing Dataset
6.	Display the date and hour, which had the most bike rentals
7.	Display the average hourly temperature and the average number of bike rentals per hour over each season

EDA with SQL

8.	Display average hourly bike count during each season by including minimum, maximum, and standard deviation of the hourly bike count for each season
9.	Display the average of TEMPERATURE, HUMIDITY, WIND_SPEED, VISIBILITY, DEW_POINT_TEMPERATURE, SOLAR_RADIATION, RAINFALL, and SNOWFALL per season
10.	Display total number of bikes available in Seoul by including CITY, COUNTRY, LAT, LNG, POPULATION
11.	List all cities with total bike counts between 15000 and 20000 by including city, country names, coordinates (LAT, LNG), population, and number of bicycles for each city

EDA with data visualization

No	Chart Type	Applied Columns	Purpose
1.	Scatter plot	X = DATE Y = RENTED_BIKE_COUNT	To explore how many bikes are rented on which date
2.	Scatter plot	X = DATE Y = RENTED_BIKE_COUNT COLOR = HOUR	To explore how many bikes are rented on each date, with the points colored based on their starting rental hour.
3.	Histogram	X = RENTED_BIKE_COUNT	To explore the distribution of Rented_Bike_COUNT column
4.	Scatter plot	X = TEMPERATURE Y = RENTED_BIKE_COUNT	To explore the correlation between temperature and the number of rented bikes in each season

EDA with data visualization

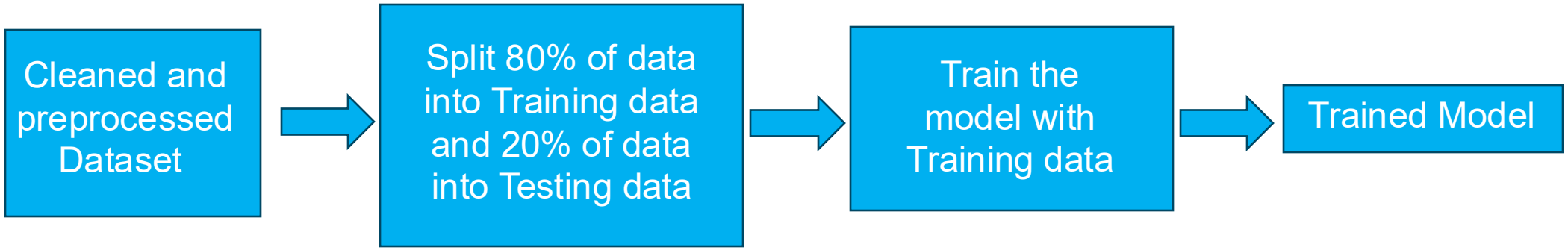
5.	Box plot	X = HOUR Y = RENTED_BIKE_COUNT	To explore the distribution of the number of rented bikes by hour in each season
6.	Box plot	X = DAILY _TOTAL_RAINFALL, DAILY_TOTAL_SNOWFALL Y = Total Values	To explore the distribution of daily total rainfall and snowfall

Predictive analysis (Regression)

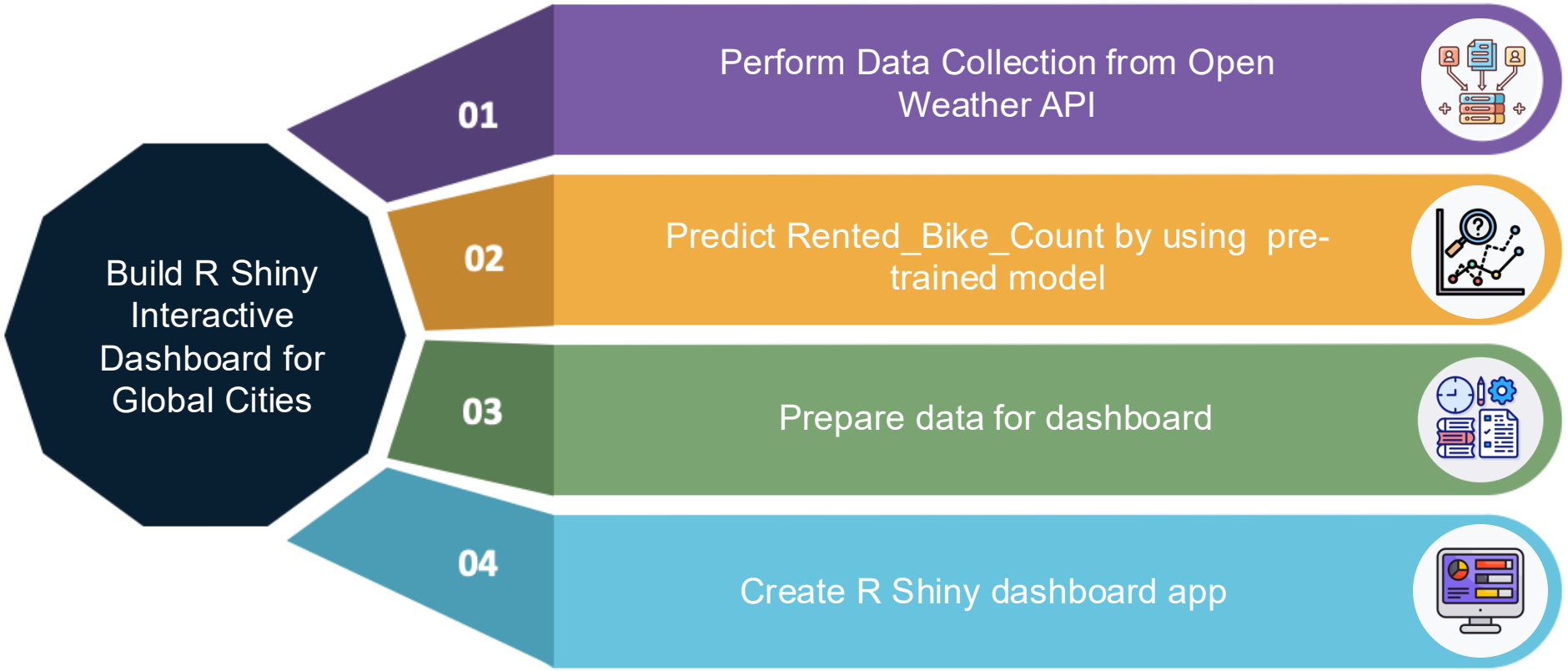
- Before selecting the model for predicting rented bike count in Seoul, six different models were developed and evaluated by using the data that had been collected.
- Those algorithm are Random Forest, Polynomial Linear Regression, Polynomial Linear Regression with interaction terms, Ridge Regularization with grid search, Lasso Regularization with grid search, and Elastic Net Regularization with grid search.
- Once all the above models were trained with trained data, they were tested with test data and evaluate them by using Root Mean Square Error (RMSE), and R-Squared methods.
- Then the model, which gave the lowest Root Mean Square Error (RMSE) and highest R-Squared value, was selected as the best model. (In this project, Random Forest is the best model)
- The following slide will show how each model was developed with a flow chart.

Predictive analysis (Regression)

Flow Chart of Developing a Regression model



Content of Part-2 Methodology

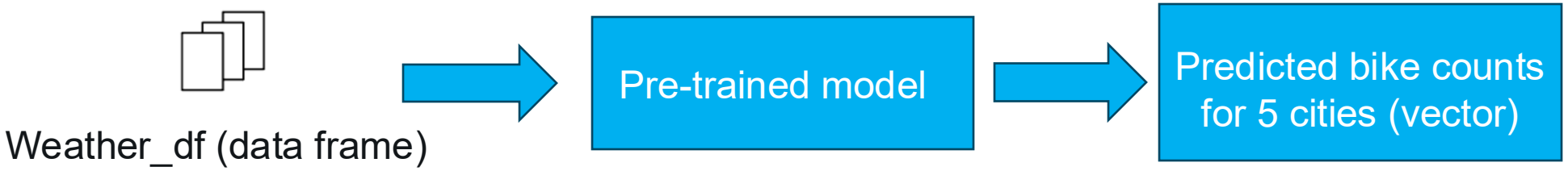


Data Collection

- In the part-2 of this project, data are collected from Open Weather Map website by using API key.
- The data contains the weather condition of 5 different cities: Seoul, South Korea; New York, USA; Paris, France; Suzhou, China; and London, UK.
- The collected data were saved in data frame to use as an input data of pre-trained model.

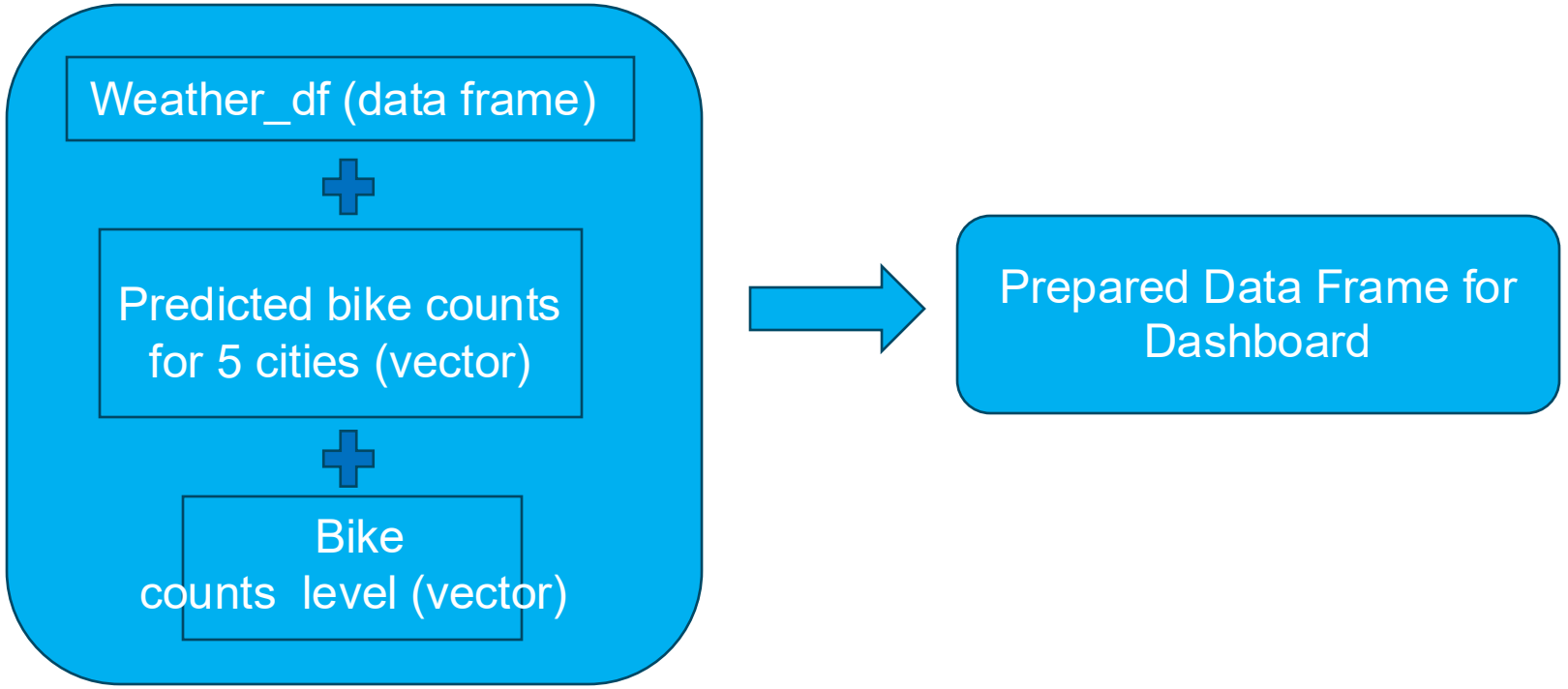
Predict Rented_Bike_Count

- To predict the number of rented bikes for the five cities, Seoul, South Korea; New York, USA; Paris, France; Suzhou, China; and London, UK, the pre-trained model's coefficients were used.
- The numbers were predicted by giving the weather_df data frame as an input to the pre-trained model.



Prepare Data For Dashboard

- In the data preparation for creating dashboard, the required features' columns from the weather data collected from Open Weather Map, the predicted bike count, and the level of predicted bike count, which was created based on the number of predicted bikes, were merged together into a single data frame.



Build a R Shiny dashboard

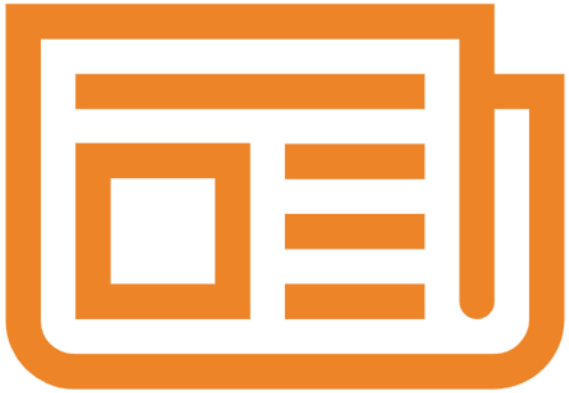
Plots used in dashboard

No.	Plot name	Plot type	Purpose
1.	City_bike_map	Map	To observe maximum bike prediction for next 5 days across five different cities
2.	Temp_line	Line graph	To observe temperature trend in next 5 days of each city
3.	Bike_line	Line graph with interaction	To observe the trend of predicted number of bike for coming 5 days.
4.	Humidity_pred_chart	Scatter plot	To observe how humidity effect on predicted bike count in coming 5 days.

Interaction used in dashboard

No.	Interaction name	Interaction type	Purpose
1.	City_dropdown	Dropdown menu	To select specific city for drilling down about that city

Results



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

EDA with SQL

Busiest bike rental times

SQL Query and Result:

```
> 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
```

Explanation:

The data analysis reveals that the highest number of bike rentals in Seoul occurred on the 19th of June, 2018, at 18:00 (6 PM). This indicates a peak in bike-sharing activity during the early evening, which is typically a busy time as people commute home from work.

Hourly popularity and temperature by seasons

SQL Query and Result:

```
> dbGetQuery(con,
+            "SELECT SEASONS, HOUR, AVG(TEMPERATURE) AS AVG_TEMP, AVG(RENTED_BIKE_COUNT) AS AVG_BIKE_COUNT
+            FROM SEOUL_BIKE_SHARING
+            GROUP BY SEASONS, HOUR
+            ORDER BY AVG_BIKE_COUNT DESC
+            LIMIT(3)")
  SEASONS HOUR AVG_TEMP AVG_BIKE_COUNT
1  Summer   18 29.38791       2135.141
2  Autumn   18 16.03185       1983.333
3  Summer   19 28.27378       1889.250
```

Explanation:
The table shows the average number of bike rentals and average temperature for each hour across different seasons. The data is sorted by the average bike count in descending order. The query result shows that the highest average bike count is observed during the summer season at 18:00 (6 PM), with an average bike count of 2135.141 and an average temperature of 29.39°C. This indicates peak bike rental activity in the early evening during summer.

Rental Seasonality

SQL Query and Result:

```
> dbGetQuery(con,
+            "SELECT SEASONS, HOUR, MIN(RENTED_BIKE_COUNT) AS MIN_BIKE_RENT,
+            MAX(RENTED_BIKE_COUNT) AS MAX_BIKE_RENT, AVG(RENTED_BIKE_COUNT) AS AVG_BIKE_COUNT,
+            SQRT(AVG(RENTED_BIKE_COUNT*RENTED_BIKE_COUNT) - AVG(RENTED_BIKE_COUNT)*AVG(RENTED_BIKE_COUNT) ) AS STDDEV_BIKE_COUNT
+            FROM SEOUL_BIKE_SHARING
+            GROUP BY SEASONS, HOUR
+            ORDER BY AVG_BIKE_COUNT
+            LIMIT(3)")
  SEASONS HOUR MIN_BIKE_RENT MAX_BIKE_RENT AVG_BIKE_COUNT STDDEV_BIKE_COUNT
1  Winter    4            20           120       50.47778          19.05561
2  Winter    5            13           100       51.22222          18.98759
3  Winter    3            21           230       77.81111          37.68256
```

Explanation:
The table shows the minimum, maximum, and average number of bike rentals, as well as the standard deviation of bike rentals, for each hour across different seasons. The data is sorted by the average bike count in descending order. The highest average bike count is observed during the summer season at 18:00 (6 PM), with an average bike count of 2135.141, a minimum rental count of 17, and a maximum rental count of 3556. The standard deviation of bike rentals at this time is 884.0829.

Weather Seasonality

SQL Query and Result:

```
> dbGetQuery(con,
+            "SELECT SEASONS, AVG(TEMPERATURE), AVG(HUMIDITY), AVG(WIND_SPEED),
+            AVG(VISIBILITY), AVG(DEW_POINT_TEMPERATURE), AVG(SOLAR_RADIATION),
+            AVG(RAINFALL), AVG(SNOWFALL)
+            FROM SEOUL_BIKE_SHARING
+            GROUP BY SEASONS")
```

	SEASONS	AVG(TEMPERATURE)	AVG(HUMIDITY)	AVG(WIND_SPEED)	AVG(VISIBILITY)	AVG(DEW_POINT_TEMPERATURE)	AVG(SOLAR_RADIATION)	AVG(RAINFALL)	AVG(SNOWFALL)
1	Autumn	13.821580	59.04491	1.492101	1558.174	5.150594	0.5227827	0.11765617	0.06350026
2	Spring	13.021685	58.75833	1.857778	1240.912	4.091389	0.6803009	0.18694444	0.00000000
3	Summer	26.587711	64.98143	1.609420	1501.745	18.750136	0.7612545	0.25348732	0.00000000
4	Winter	-2.540463	49.74491	1.922685	1445.987	-12.416667	0.2981806	0.03282407	0.24750000

Explanation:
The table provides the average values of various weather parameters for each season. Overall, the summer season experiences the highest temperatures, humidity, dew point temperature, solar radiation, and rainfall, whereas the winter season sees the lowest temperatures, highest snowfall, and notable wind speeds. Autumn has the highest visibility among all seasons.

Bike-sharing info in Seoul

SQL Query and Result:

```
> dbGetQuery(con,
+            "SELECT SUM(BSS.BICYCLES) AS TOTAL_NUM_BICYCLES_SEOUL, WC.CITY, WC.COUNTRY, WC.LAT, WC.LNG, WC.POPULATION
+            FROM WORLD_CITIES WC, BIKE_SHARING_SYSTEMS BSS
+            WHERE WC.CITY = BSS.CITY AND UPPER(BSS.CITY) = 'SEOUL'")
  TOTAL_NUM_BICYCLES_SEOUL  CITY      COUNTRY     LAT LNG POPULATION
1                    20000 Seoul Korea, South 37.5833 127   21794000
```

Explanation:
The table provides information about the total number of bicycles in Seoul's bike-sharing system, along with some geographical and demographic details of the city.

Cities similar to Seoul

SQL Query and Result:

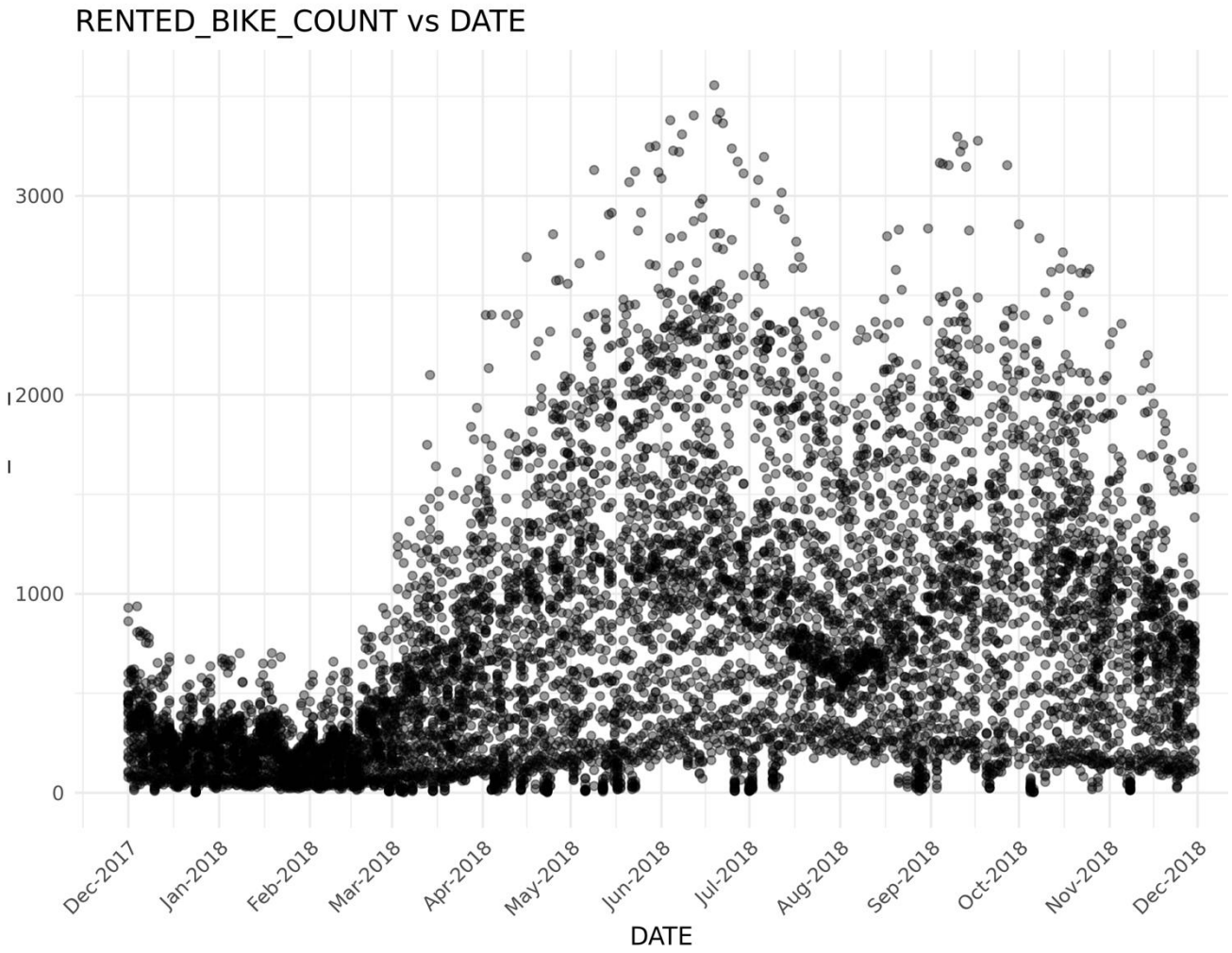
```
> dbGetQuery(con,
+            "SELECT WC.CITY, WC.COUNTRY, WC.LAT, WC.LNG, WC.POPULATION, SUM(BSS.BICYCLES) AS TOTAL_BICYCLES
+            FROM WORLD_CITIES WC, BIKE_SHARING_SYSTEMS BSS
+            WHERE WC.CITY = BSS.CITY
+            GROUP BY WC.CITY
+            HAVING TOTAL_BICYCLES BETWEEN 15000 AND 20000")
      CITY      COUNTRY     LAT      LNG POPULATION TOTAL_BICYCLES
1  Beijing        China 39.9050 116.3914   19433000          16000
2   Ningbo        China 29.8750 121.5492    7639000          15000
3    Seoul Korea, South 37.5833 127.0000   21794000          20000
4 Shanghai        China 31.1667 121.4667   22120000          19165
5  Weifang        China 36.7167 119.1000    9373000          20000
6  Zhuzhou        China 27.8407 113.1469    3855609          20000
```

The query lists cities that have a total number of bicycles in their bike-sharing systems between 15,000 and 20,000. The above query result reflect the scale and distribution of bike-sharing systems in major East Asian cities, highlighting the prevalence and importance of such systems in urban transportation planning.

EDA with Visualization

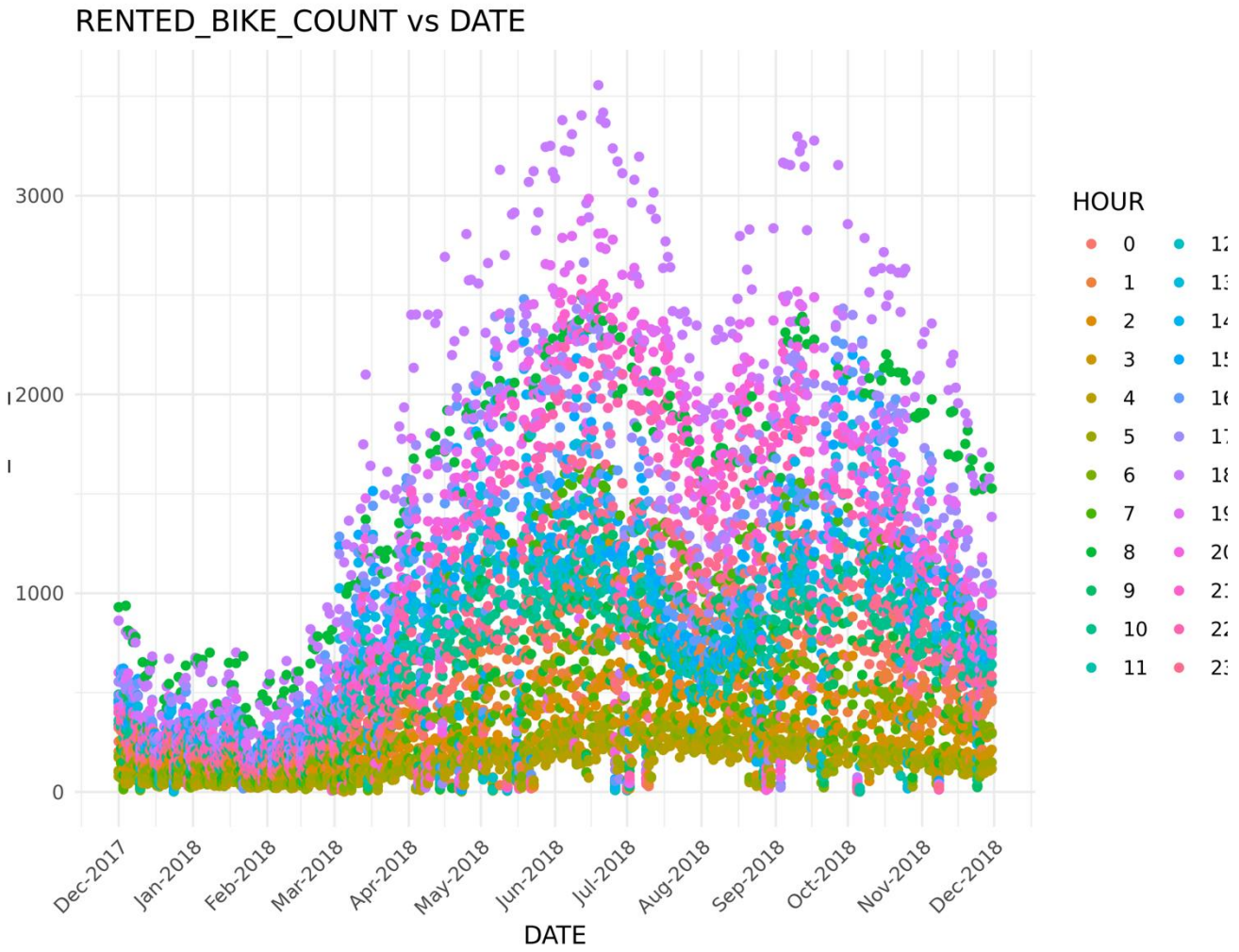
Bike rental vs. Date

- Upon examining the scatter plot of rented bike counts, it is evident that the majority of observations cluster densely below 500 bikes on the RENTED_BIKE_COUNT axis. This suggests that most days experience a rental demand of fewer than 500 bikes
- Besides, it also shows that the number of rented bike went up to over 3000 in the months between May and August, and September and October.



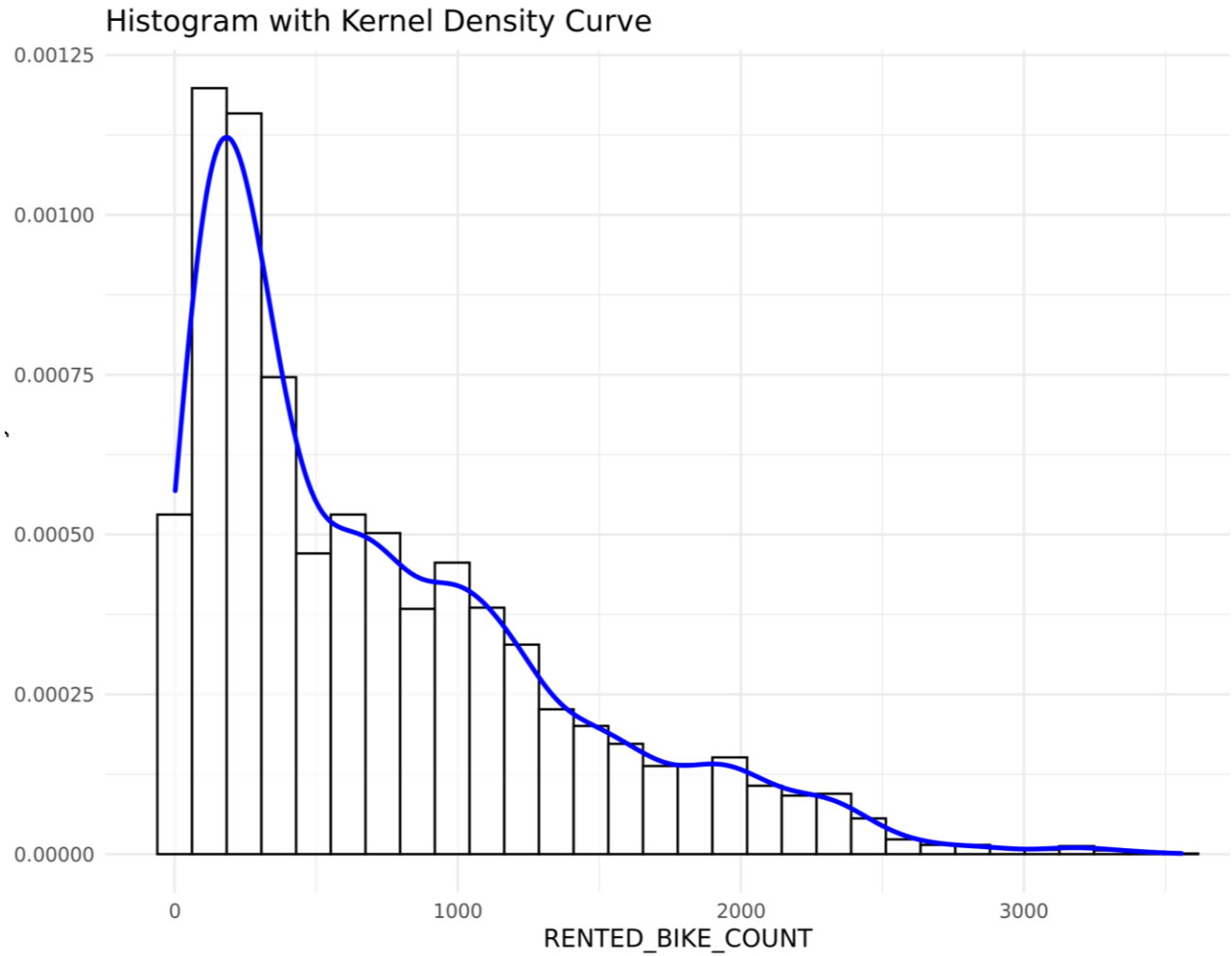
Bike rental vs. Datetime

- Once the HOUR is used as a color, it is clear that the highest bike counts, exceeding 3000, occurred at 18:00 (6 PM).
- The lowest rented bike counts were observed during the hours between 3:00 and 5:00. However, the rented bike count between that times are below 500.



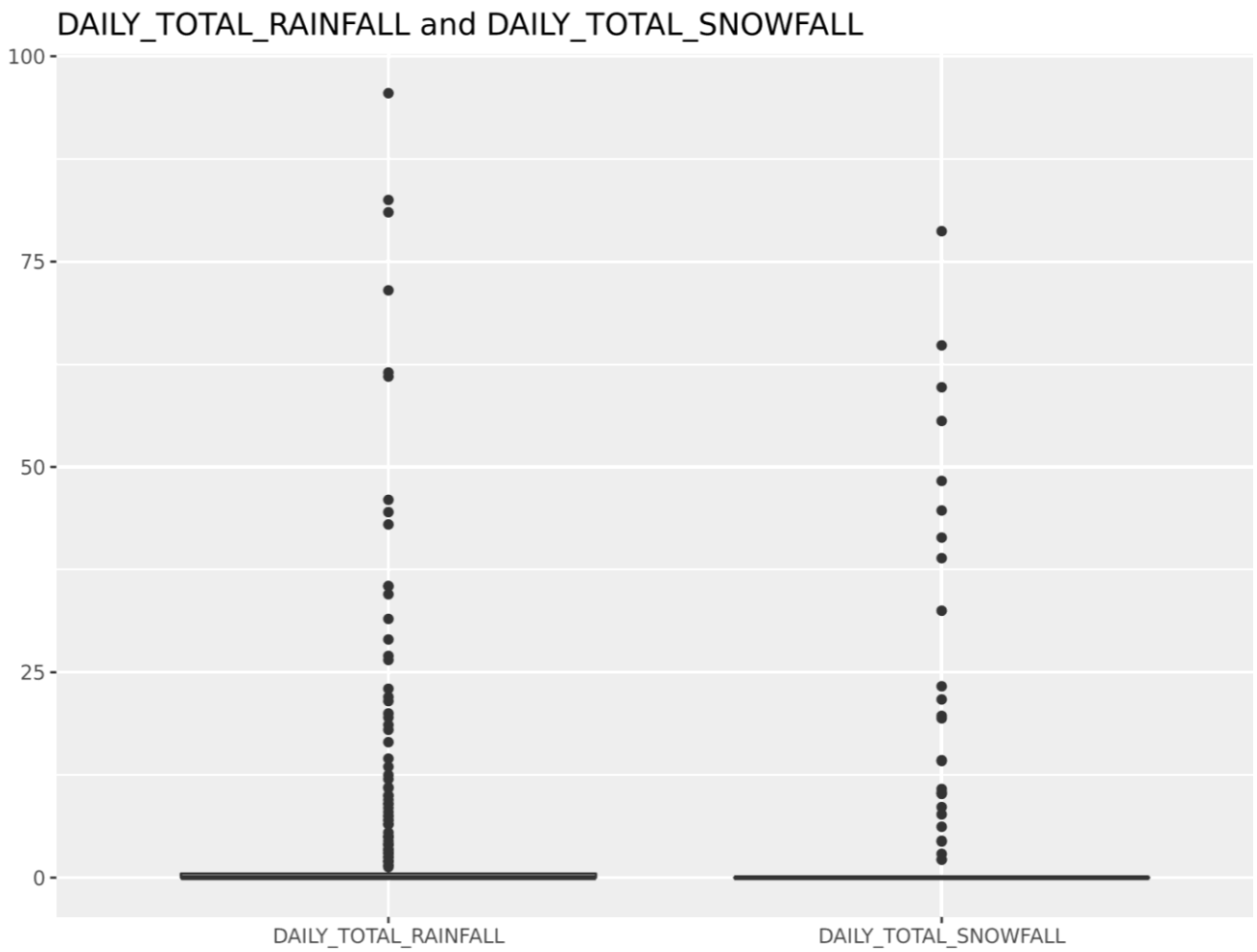
Bike rental histogram

- Based on the histogram, we observe that bike rentals typically involve relatively low numbers most of the time. The 'mode', which represents the most frequent number of bikes rented, appears to be around 250.
- Additionally, there are noticeable peaks ('bumps') in the histogram around 700, 900, 1900, and 3200 bikes rented. These peaks suggest that there may be distinct subgroups within the data where other modes or frequent rental amounts occur.
- Besides, judging from the tail of the distribution, there are occasional instances where significantly more bikes are rented out than what is typically observed



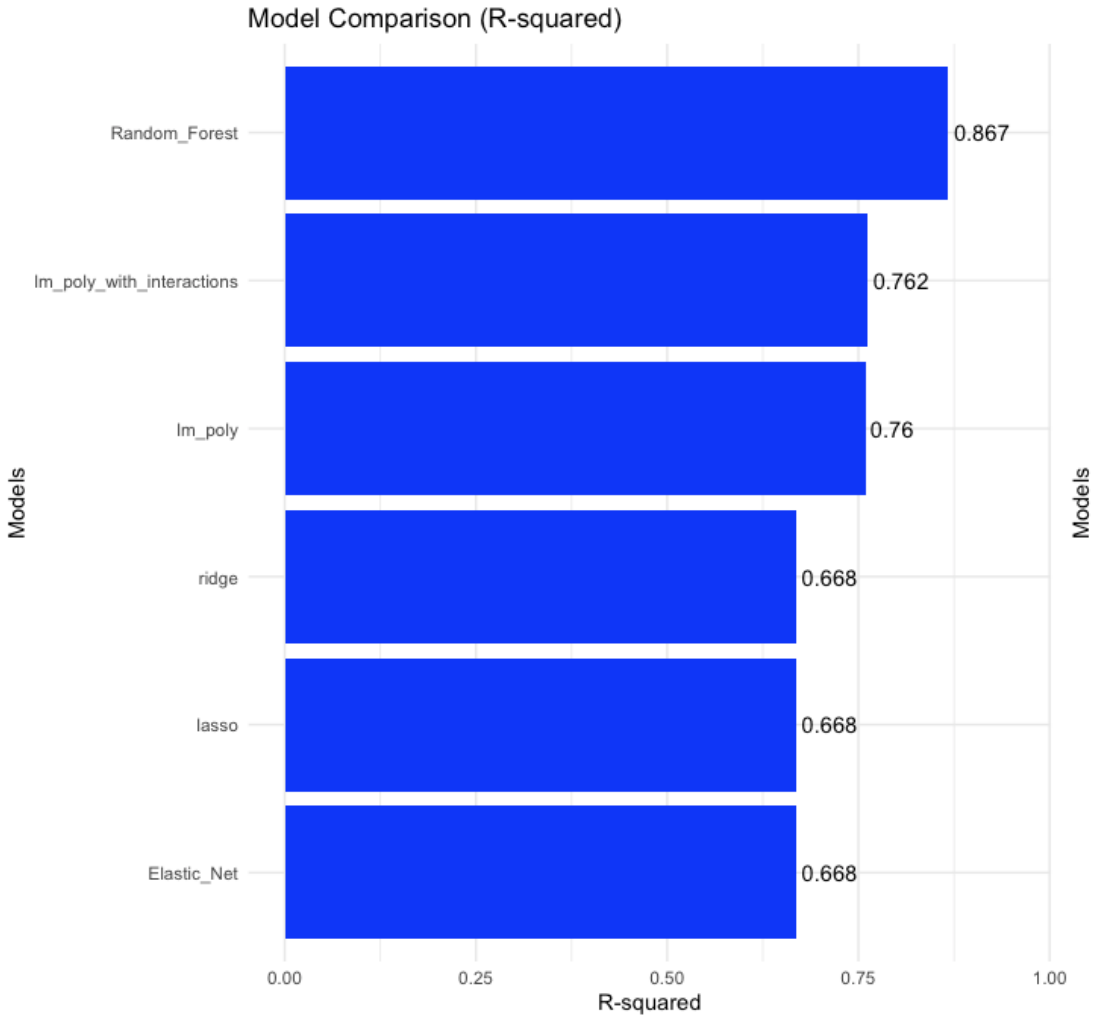
Daily total rainfall and snowfall

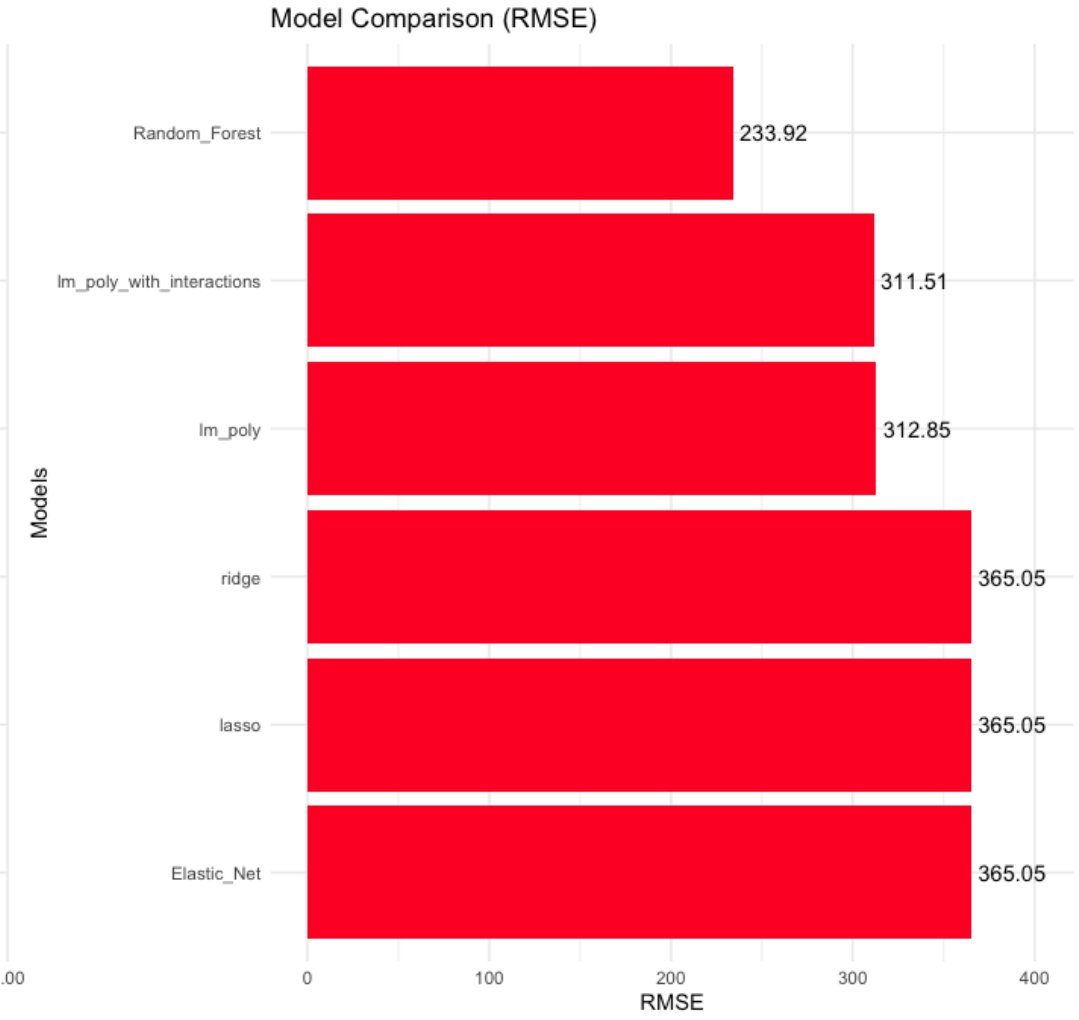
- Based on the box plots, we can clearly see that the interquartile range (IQR) for daily total rainfall is between 0 and 1, while the IQR for daily total snowfall is exactly at 0. This observation is derived from the total data points for both metrics, which amount to 353.
- For daily total rainfall, 100 data points are greater than 0, representing approximately 28% of the total. When ordered in ascending order, only about 3% of these points fall within the IQR of the box plot. In contrast, for daily total snowfall, only 27 data points are greater than 0, representing just 7% of the total. Consequently, the IQR of the box plot for daily total snowfall remains at 0.



Predictive analysis

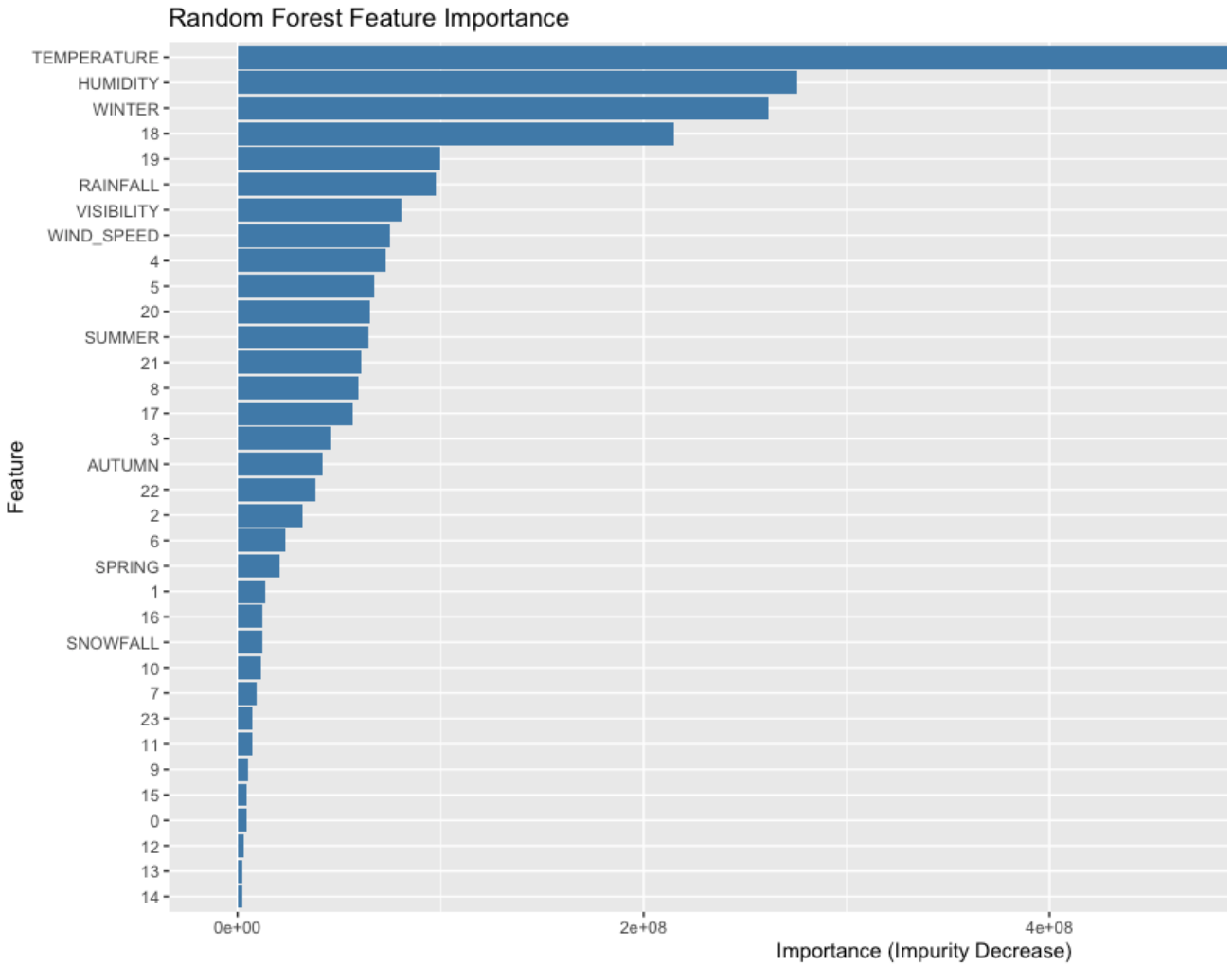
Model evaluation





Random Forest Feature Importance

- This plot shows the relative importance of features in the Random Forest model, measured by how much each variable reduces prediction error across decision trees.
- Temperature is the most influential factor, indicating that rider comfort strongly drives bike-sharing demand. Humidity and winter also have high importance, reflecting the impact of uncomfortable weather and seasonal conditions on usage.
- Weather variables such as rainfall, visibility, and wind speed have moderate influence, while seasonal indicators like summer, autumn, and spring contribute less overall.
- The numeric features 0–23 represent hours of the day. Higher importance for hours such as 8, 17–19, and 18 corresponds to peak commuting periods, while late-night and midday hours (e.g., 0, 1, 12, 14) have low importance due to more stable or consistently low demand.
- Overall, the model shows that weather conditions dominate demand, with time-of-day effects most relevant during peak hours.



Find the best performing model

- In the part 1 of this project, the selected model is a Random Forest. This model was chosen as the best performing model because it produced the lowest RMSE value and the highest R-squared value among the five models tested.
- That Random Forest model is saved in “rf_bike_model.rds” and used to make prediction in dashboard.
- A Random Forest regression model was constructed to predict bike-sharing demand using weather, seasonal, and hourly predictors. The model consists of 500 decision trees, each trained on a bootstrap sample of the training data. At each split, 8 randomly selected predictors (mtry = 8) are considered, with a minimum of 10 observations per terminal node (min_n = 10), promoting model stability and reducing overfitting. The input features include continuous weather variables (temperature, humidity, wind speed, visibility, rainfall, snowfall), one-hot encoded hour-of-day indicators (0–23), and seasonal indicators (spring, summer, autumn, winter). Final predictions are obtained by averaging outputs across all trees, and impurity-based feature importance is used to assess the contribution of each predictor.

Random Forest Model's Architecture Code

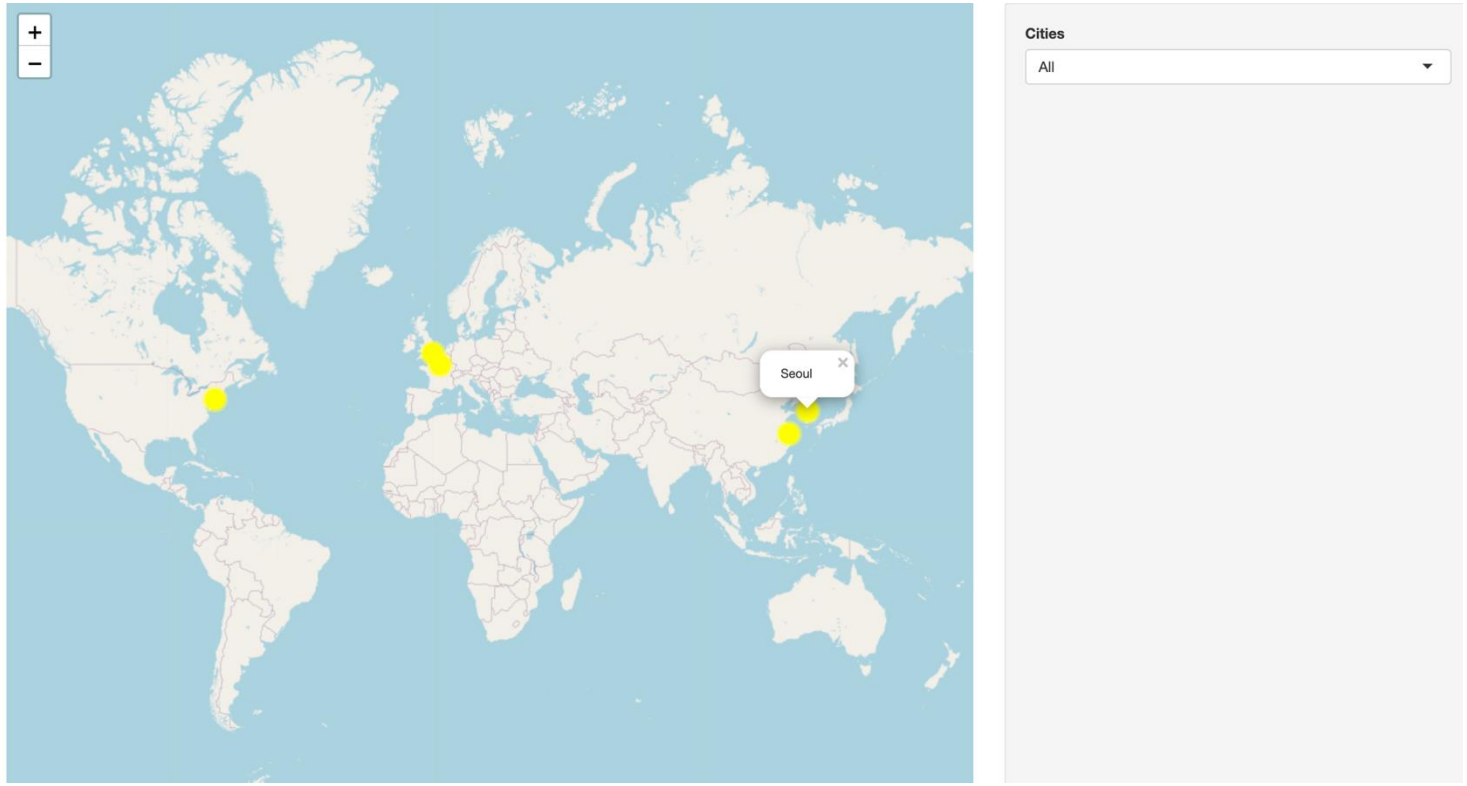
```
#############################################################
# Model-6 (Random Forest Regression)
#############################################################

rf_spec <- rand_forest(
  mode  = "regression", trees = 500, mtry  = 8, min_n = 10 ) %>%
  set_engine("ranger", importance = "impurity")

rf_formula <- RENTED_BIKE_COUNT ~
  TEMPERATURE + HUMIDITY + WIND_SPEED + VISIBILITY + RAINFALL + SNOWFALL +

  `1` + `2` + `3` + `4` + `5` + `6` + `7` + `8` + `9` + `10` + `11` + `12` +
  `13` + `14` + `15` + `16` + `17` + `18` + `19` + `20` + `21` + `22` + `23` + `0`+

  AUTUMN + SPRING + SUMMER + WINTER

rf_fit <- rf_spec %>%
  fit(rf_formula, data = train_data)

rf_test_results <- rf_fit %>%
  predict(new_data = test_data) %>%
  mutate(truth = test_data$RENTED_BIKE_COUNT)

rf_test_results$.pred[rf_test_results$.pred < 0] <- 0

rsq_rf <- rsq(rf_test_results, truth = truth, estimate = .pred)
rmse_rf <- rmse(rf_test_results, truth = truth, estimate = .pred)

rsq_rf
rmse_rf
```

Dashboarding

Max Bike Rental Distribution Across Five Cities

The map highlights five cities using circles, with each circle's size and color representing the number of max rented bikes. Larger circles indicate higher rental counts. The color scheme is as follows:

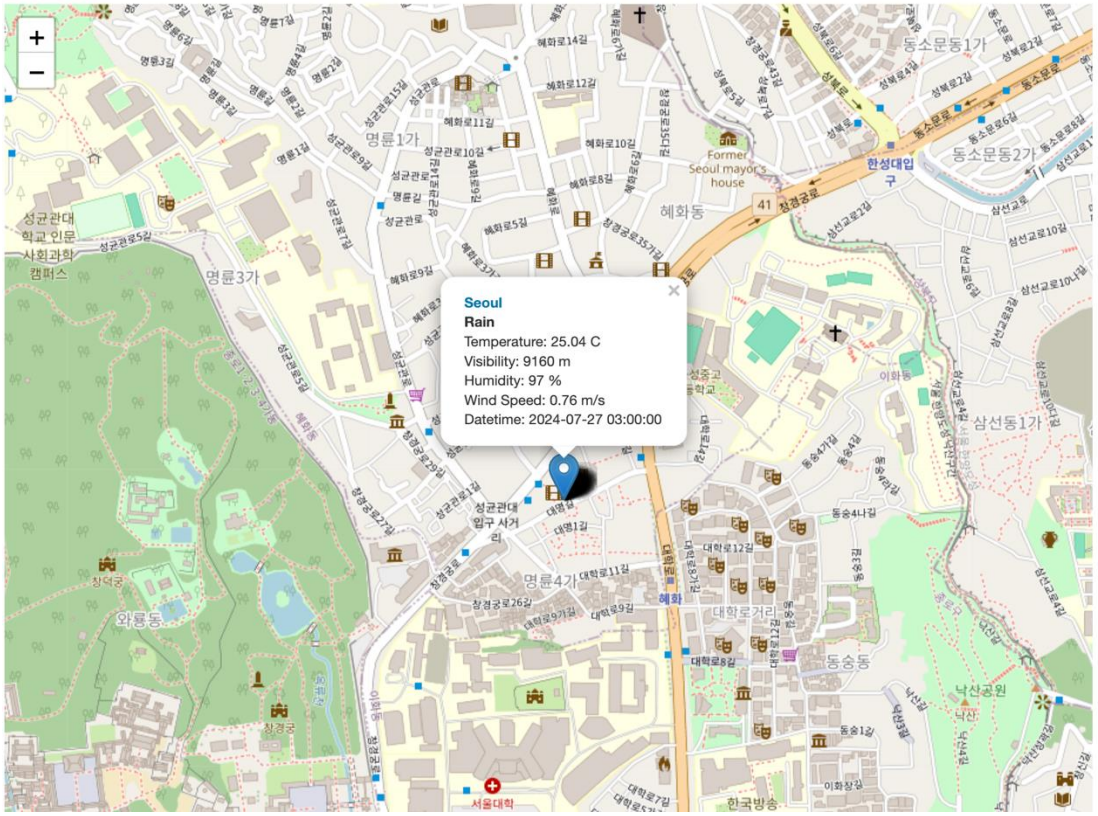
- **Green**: Represents a small rental count (0 to 1000 bikes).
- **Yellow**: Indicates a medium rental count (1000 to 3000 bikes).
- **Red**: Signifies a high rental count (over 3000 bikes).

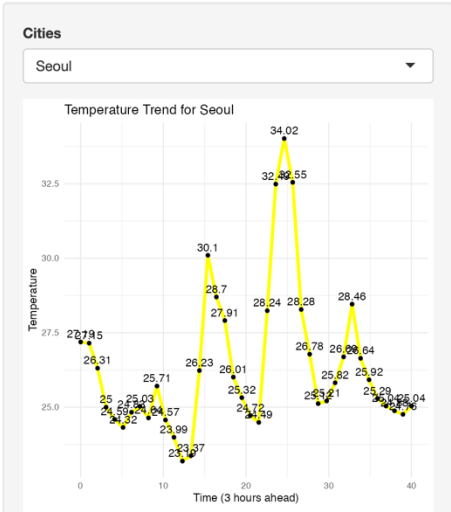


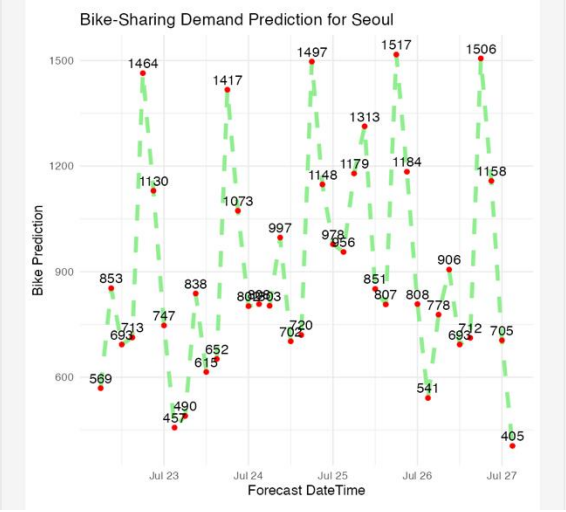
The above map depicts all five cities indicated by yellow circles, meaning the bike rental count in all cities falls within the range of 1000 to 3000.

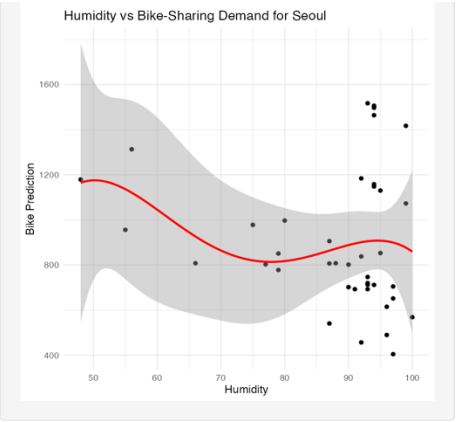
Seoul Bike Rental Trends & Forecasts

Bike-sharing demand prediction app









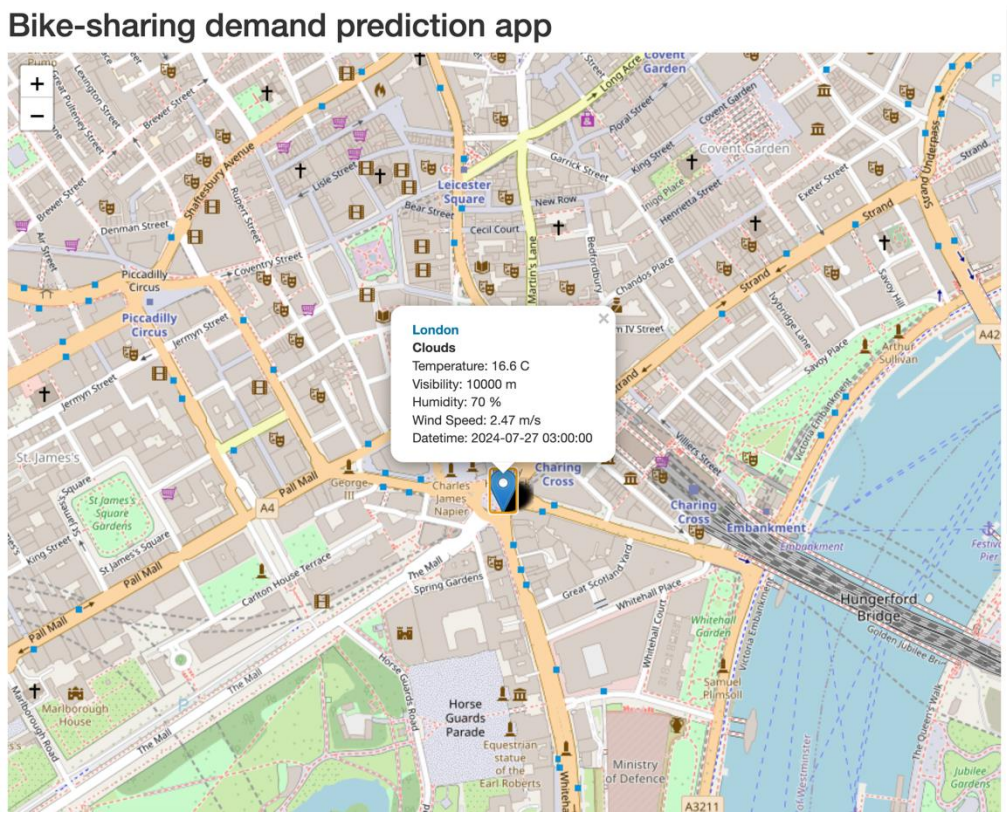
Seoul Bike Rental Trends & Forecasts

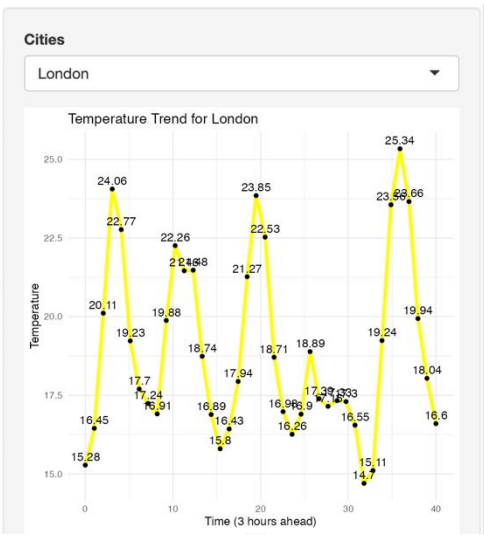
This dashboard provides key insights into bike rental trends in Seoul, including current distribution, future predictions, and weather correlations.

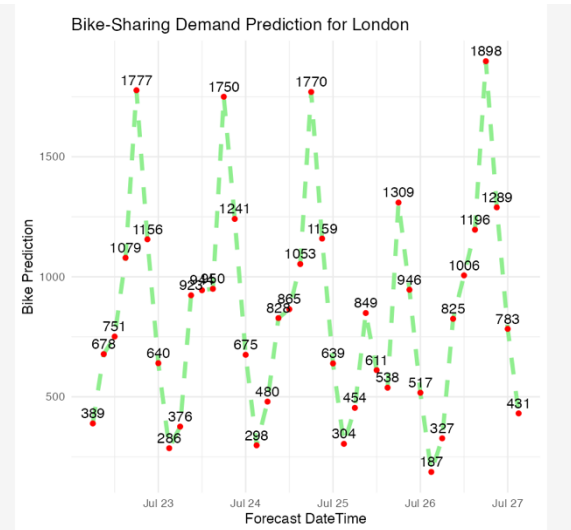
Key Elements:

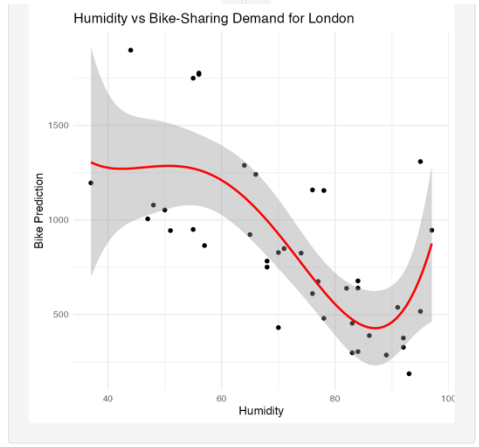
- **Max Bike Rental Distribution:** Visualizes the spatial distribution of bike rentals across Seoul. Clicking on a location reveals the predicted weather and temperature conditions for the last hour of the last day.
- **5-Day Temperature Trend:** A line graph displaying the predicted temperature trend for the next 5 days. Data points indicate daily temperature forecasts ranging from a peak of 34.02°C to a low of 23.19°C
- **5-Day Bike Count Prediction:** A line graph forecasting bike rental counts for the next 5 days. Data points represent daily rental predictions with a maximum of 1517 and a minimum of 405.
- **Humidity vs. Bike Count Correlation:** A scatter plot with a trend line showing the relationship between humidity levels and bike rental counts for the next 5 days. The data suggests minimal correlation, with points showing no clear trend.

London Bike Rental Trends & Forecasts









London Bike Rental Trends & Forecasts

This dashboard provides key insights into bike rental trends in London, including current distribution, future predictions, and weather correlations.

Key Elements:

- Max Bike Rental Distribution: Visualizes the spatial distribution of bike rentals across London. Clicking on a location reveals the predicted weather and temperature conditions for the last hour of the last day.
- 5-Day Temperature Trend: A line graph displaying the predicted temperature trend for the next 5 days. Data points indicate daily temperature forecasts ranging from a peak of 25.34°C to a low of 14.7°C
- 5-Day Bike Count Prediction: A line graph forecasting bike rental counts for the next 5 days. Data points represent daily rental predictions with a maximum of 1898 and a minimum of 187.
- Humidity vs. Bike Count Correlation: A scatter plot with a trend line showing the relationship between humidity levels and bike rental counts for the next 5 days. The data suggests a small negative correlation between them.

CONCLUSION

This project aimed to analyze the impact of weather on bike-sharing demand in urban areas, focusing on Seoul and extending the analysis to a global scale.

Part 1: Weather Impact Analysis on Bike-Sharing Demand in Seoul

- **Data Collection and Processing:** Successfully gathered and processed weather and bike-sharing demand data.
- **Exploratory Data Analysis (EDA):** Uncovered significant patterns and relationships in the data.
- **Predictive Modeling:** Developed robust models to forecast bike-sharing demand in Seoul based on weather conditions.

Part 2: Interactive Dashboard for Global Cities

- **Utilization of Trained Model from Part 1:** Applied a trained model to predict rental bike counts for Seoul, New York, Paris, Suzhou, and London.
- **Data Extraction:** Extracted weather data using the Open Weather API.
- **Integration of Results:** Combined predictive results with current weather data.
- **Dashboard Development:** Created an interactive dashboard showcasing temperature trends, bike count predictions, and bike counts versus humidity for the next five days.

CONCLUSION

Applications

- **For City Planners and Bike-Sharing Companies in Seoul:** The predictive model can optimize bike distribution and enhance service availability.
- **For Real-Time Decision-Making and Strategic Planning:** The interactive dashboard aids in making informed decisions and planning for bike rentals in Seoul, New York, Paris, Suzhou, and London.

By utilizing the outcomes of this project, we can improve bike-sharing services in Seoul and other global cities by adapting to weather variations.

APPENDIX

GitHub Link for the whole project:

https://github.com/thetkhinelin25/Data-Science-R-Capstone-Project.git

Thank You!