Applied Data Science with R Capstone project

Chinthala Harshitha

2025-04-10

Outline

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

Executive Summary

Project Objective: The project aims to analyze how weather conditions influence bike-sharing demand in urban areas, using Seoul as a case study. The goal is to build a predictive model to optimize bike availability based on weather forecasts.

Data Sources:

- **Seoul Bike Sharing Demand Dataset:** Includes hourly bike rentals, weather metrics (temperature, humidity, rainfall, etc.), and contextual data (holidays, seasons).
- **Open Weather API:** Provides real-time and forecasted weather data for Seoul.
- **Global Bike-Sharing Systems:** Contextual data about bike-sharing systems worldwide.

Business Impact: The project helps optimize bike-sharing supply, reduce operational costs, and improve user accessibility by aligning bike availability with weather-driven demand patterns. It showcases practical data science applications for urban mobility solutions.

Introduction

- Bike-sharing systems are a vital part of urban mobility, but their demand fluctuates based on external factors like weather. Understanding these patterns can help optimize bike availability, reduce costs, and improve user experience.
- This project analyzes the relationship between weather conditions and bike-sharing demand in Seoul, leveraging datasets on rentals, weather forecasts, and global bike-sharing systems to build a predictive model.
- By applying data science techniques—including data wrangling, exploratory analysis, and linear regression—we aim to provide actionable insights for efficient bike allocation and showcase the impact of weather on urban transportation.

Methodology

- Perform data collection
- Perform data wrangling
- Perform exploratory data analysis (EDA) using SQL and visualization
- Perform predictive analysis using regression models
  - How to build the baseline model
  - How to improve the baseline model
- Build a R Shiny dashboard app

Data collection

1. Web Scraping (Global Bike-Sharing Systems)

Source: Wikipedia page (List of Bicycle-Sharing Systems)

Tool: rvest library in R

Steps:

1. Fetch HTML content using read_html()
2. Extract the first HTML table node
3. Convert to a structured dataframe
4. Export as raw_bike_sharing_systems.csv

Data collection

2. API Calls (OpenWeather Data)

Source: OpenWeather API (current + 5-day forecasts)

Tool: httr library in R

Steps:

1. Authenticate with a free API key
2. Send HTTP GET requests for weather data (e.g., temperature, humidity, wind)
3. Parse JSON responses into a dataframe
4. Export as cities_weather_forecast.csv

Data wrangling with Regex

Column Standardization

- Renamed columns: UPPERCASE + underscores (e.g., Daily Ridership → DAILY_RIDERSHIP)
- Tools: toupper(), str_replace_all()

Web-Scraped Data Cleaning

- Removed reference tags (e.g., [12]) from text
- Extracted numeric values from BICYCLES
- Tools: Regex, str_replace_all(), str_extract()

Missing Data Handling

- Retained NA for active systems
- Verified data ranges using summary()

Data wrangling with dplyr

Handling Missing Values

- Replaced NAs with column mean using mutate() + ifelse(is.na(), mean(), ...)

Creating Dummy Variables

- Used mutate() + ifelse() to convert categorical variables into binary indicators

Data Normalization

- Applied min-max normalization:
  (value−min)/(max−min)(value - min) / (max-min)(value−min)/(max−min)

EDA with SQL

- Analyzed total records and date range to understand dataset scope
- Calculated average and peak rentals by season (summer/autumn highest)
- Identified weather impact: clear days increase demand, rain/snow decrease it
- Discovered hourly patterns with peaks at 8 AM and 6 PM (commuter hours)
- Compared holiday vs non-holiday usage showing lower rentals on holidays
- Seasonal analysis revealed winter had lowest average rentals
- Temperature and humidity showed strongest correlation with rental counts
- Visibility and wind speed had minimal impact on bike sharing demand

EDA with data visualization

- Scatter plot of daily bike rentals showing seasonal fluctuations
- Histogram with density curve revealing right-skewed rental distribution
- Bar chart comparing average rentals by weather condition
- Box plots showing rental distribution across seasons
- Side-by-side holiday vs non-holiday rental comparisons

Predictive analysis

Data Preparation:

- Split data into training (70%) and testing (30%) sets
- Scaled numeric features (temperature, humidity, etc.)
- Encoded categorical variables (seasons, holidays)

Model Building:

- Linear Regression (baseline)

Evaluation & Improvement:

- Compared models using RMSE and R^2 scores
- Feature selection based on variable importance

Build a R Shiny dashboard

Map Visualization:

- Color-coded circle markers representing bike demand levels (small, medium, large).
- Interactive city bike map showing bike demand levels across various cities.

Task 1: Temperature Trend Plot:

- Line plot displaying a 5-day temperature forecast for the selected city.
- Temperature values are labeled directly on the plot for clarity.

Task 2: Interactive Bike Demand Plot:

- Line plot displaying bike demand predictions over time for the selected city.
- Different demand levels are color-coded (small = green, medium = yellow, large = red).

Task 3: Humidity vs. Bike Demand:

- Scatter plot showing the correlation between humidity and bike demand for the selected city.
- A smooth regression line is added to help visualize the relationship between the two variables.

Results

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

EDA with SQL

Busiest bike rental times

```
[17]: # provide your solution here
dbGetQuery(conn, "SELECT DATE, HOUR, RENTED_BIKE_COUNT
                  FROM SEOUL_BIKE_SHARING
                  WHERE RENTED_BIKE_COUNT = (SELECT MAX(RENTED_BIKE_COUNT)
                                             FROM SEOUL_BIKE_SHARING)")
```

A data.frame: 1 × 3

DATE	HOUR	RENTED_BIKE_COUNT
<chr>	<dbl>	<dbl>
19/06/2018	18	3556

- On 19th June 2018 at 6 PM (18:00), the Seoul bike-sharing system recorded its highest rental count of 3,556 bikes. This indicates peak demand likely due to favorable weather, commuting hours, or special events during that time.

Hourly popularity and temperature by seasons

[18]:
```
# provide your solution here
dbGetQuery(conn, "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 10")
```

A data.frame: 10 × 4

SEASONS	HOUR	avg_temp	avg_bike_count
<chr>	<dbl>	<dbl>	<dbl>
Summer	18	29.38791	2135.141
Autumn	18	16.03185	1983.333
Summer	19	28.27378	1889.250
Summer	20	27.06630	1801.924
Summer	21	26.27826	1754.065
Spring	18	15.97222	1689.311
Summer	22	25.69891	1567.870
Autumn	17	17.27778	1562.877
Summer	17	30.07691	1526.293
Autumn	19	15.06346	1515.568

- Bike rentals peak in the evenings around 6 PM (18:00), especially during summer, with an average of 2,135 rentals at a warm 29.4°C
- Overall, summer evenings (17:00–22:00) dominate the top slots, showing that warm temperatures and post-work hours drive higher usage.
- Autumn evenings also show strong rental activity, though with cooler temperatures.

Rental Seasonality

```
[19]: # provide your solution here
dbGetQuery(conn, "SELECT SEASONS,
                    AVG(RENTED_BIKE_COUNT) AS avg_bike_count,
                    MIN(RENTED_BIKE_COUNT) AS min_bike_count,
                    MAX(RENTED_BIKE_COUNT) AS max_bike_count,
                    SQRT(AVG(RENTED_BIKE_COUNT*RENTED_BIKE_COUNT) - AVG(RENTED_BIKE_COUNT)*AVG(RENTED_BIKE_COUNT)) AS std_dev
                 FROM SEOUL_BIKE_SHARING
                 GROUP BY SEASONS")
```

A data.frame: 4 × 5

SEASONS	avg_bike_count	min_bike_count	max_bike_count	std_dev
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Autumn	924.1105	2	3298	617.3885
Spring	746.2542	2	3251	618.5247
Summer	1034.0734	9	3556	690.0884
Winter	225.5412	3	937	150.3374

- Bike rentals are highest during summer, with an average of 1,034 rentals per day and the greatest variation, showing high activity and fluctuation. Winter has the lowest average (226 rentals per day) and the least variation, likely due to cold weather reducing demand. Autumn and spring show moderate and more consistent rental patterns.

Weather Seasonality

```
[20]: # provide your solution here
dbGetQuery(conn, "SELECT SEASONS,
                    AVG(TEMPERATURE) AS avg_temp,
                    AVG(HUMIDITY) AS avg_humidity,
                    AVG(WIND_SPEED) AS avg_wind_speed,
                    AVG(VISIBILITY) AS avg_visibility,
                    AVG(DEW_POINT_TEMPERATURE) AS avg_dew_point,
                    AVG(SOLAR_RADIATION) AS avg_solar_rad,
                    AVG(RAINFALL) AS avg_rainfall,
                    AVG(SNOWFALL) AS avg_snowfall,
                    AVG(RENTED_BIKE_COUNT) AS avg_bike_count
              FROM SEOUL_BIKE_SHARING
              GROUP BY SEASONS
              ORDER BY avg_bike_count DESC")
```

A data.frame: 4 × 10

SEASONS	avg_temp	avg_humidity	avg_wind_speed	avg_visibility	avg_dew_point	avg_solar_rad	avg_rainfall	avg_snowfall	avg_bike_count
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Summer	26.587711	64.98143	1.609420	1501.745	18.750136	0.7612545	0.25348732	0.00000000	1034.0734
Autumn	13.821580	59.04491	1.492101	1558.174	5.150594	0.5227827	0.11765617	0.06350026	924.1105
Spring	13.021685	58.75833	1.857778	1240.912	4.091389	0.6803009	0.18694444	0.00000000	746.2542
Winter	-2.540463	49.74491	1.922685	1445.987	-12.416667	0.2981806	0.03282407	0.24750000	225.5412

- Weather seasonality shows that bike rentals are highest in summer, with warm temperatures (around 26.6°C), higher humidity, and no snowfall. Autumn and spring have mild temperatures and moderate rentals. Winter has the lowest rentals, with freezing temperatures, low solar radiation, and some snowfall, which likely discourages biking.

Bike-sharing info in Seoul

```
[21]: # provide your solution here
      dbGetQuery(conn, "SELECT w.CITY, w.COUNTRY, w.LAT, w.LNG, w.POPULATION, b.BICYCLES
                        FROM WORLD_CITIES w, BIKE_SHARING_SYSTEMS b
                        WHERE w.CITY = b.CITY AND w.CITY = 'Seoul'")
```

A data.frame: 1 × 6

CITY	COUNTRY	LAT	LNG	POPULATION	BICYCLES
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Seoul	Korea, South	37.5833	127	21794000	20000

- The bike-sharing system in Seoul has a total of 20,000 bikes. Seoul, located in South Korea, has a population of approximately 21.8 million people and is situated at a latitude of 37.5833 and longitude of 127.

Cities similar to Seoul

```
[22]: # provide your solution here
dbGetQuery(conn, "SELECT w.CITY, w.COUNTRY, w.LAT, w.LNG, w.POPULATION, b.BICYCLES
                  FROM WORLD_CITIES w, BIKE_SHARING_SYSTEMS b
                  WHERE w.CITY = b.CITY AND b.BICYCLES BETWEEN 15000 AND 20000")
```

A data.frame: 6 × 6

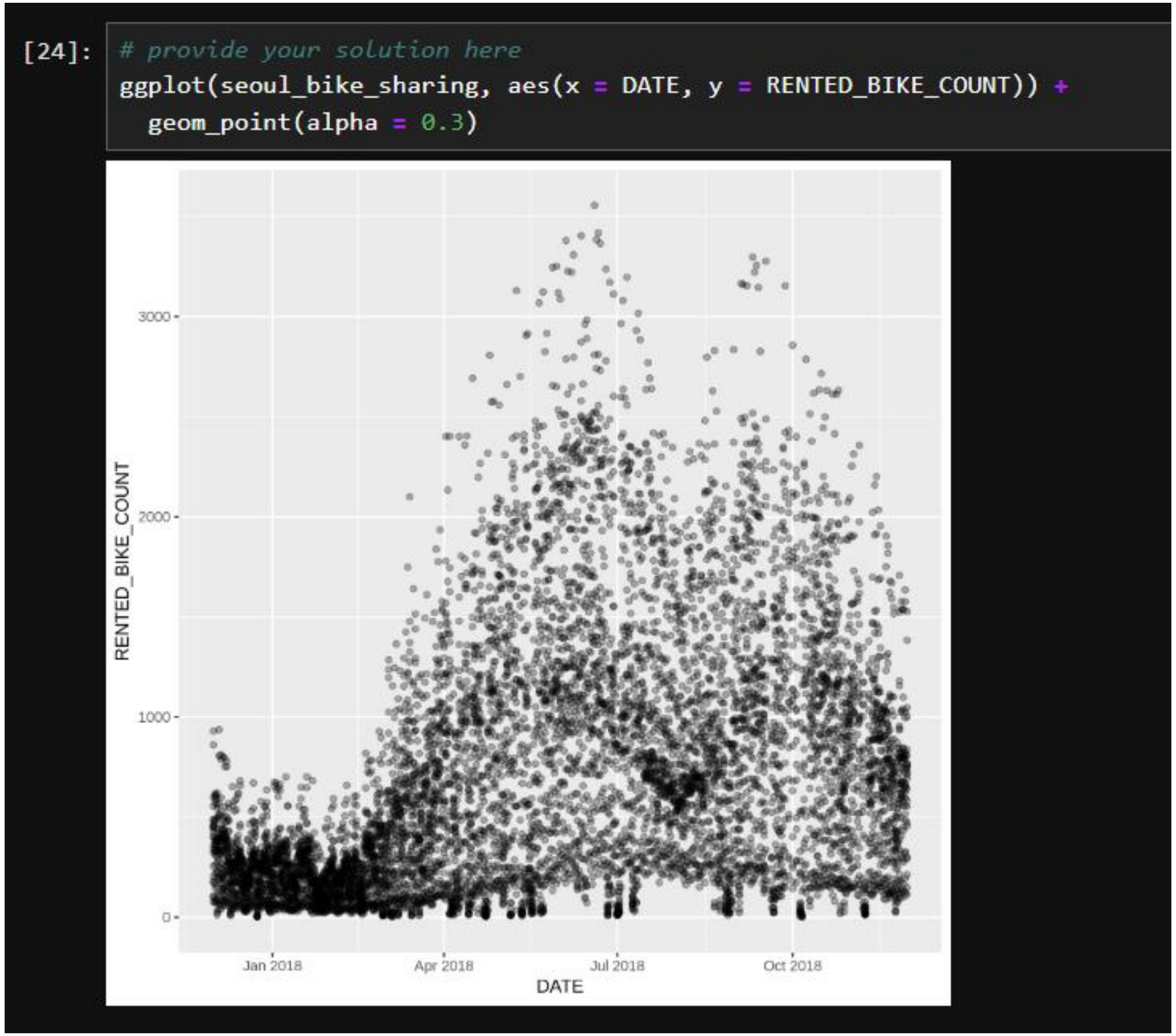
CITY	COUNTRY	LAT	LNG	POPULATION	BICYCLES
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Beijing	China	39.9050	116.3914	19433000	16000
Ningbo	China	29.8750	121.5492	7639000	15000
Shanghai	China	31.1667	121.4667	22120000	19165
Weifang	China	36.7167	119.1000	9373000	20000
Zhuzhou	China	27.8407	113.1469	3855609	20000
Seoul	Korea, South	37.5833	127.0000	21794000	20000

- Cities with a similar bike-sharing scale to Seoul, each having between 15,000 and 20,000 bikes, include Beijing, Ningbo, Shanghai, Weifang, and Zhuzhou in China. These cities, like Seoul, have large populations and comparable bike-sharing systems.

EDA with Visualization

Bike rental vs. Date

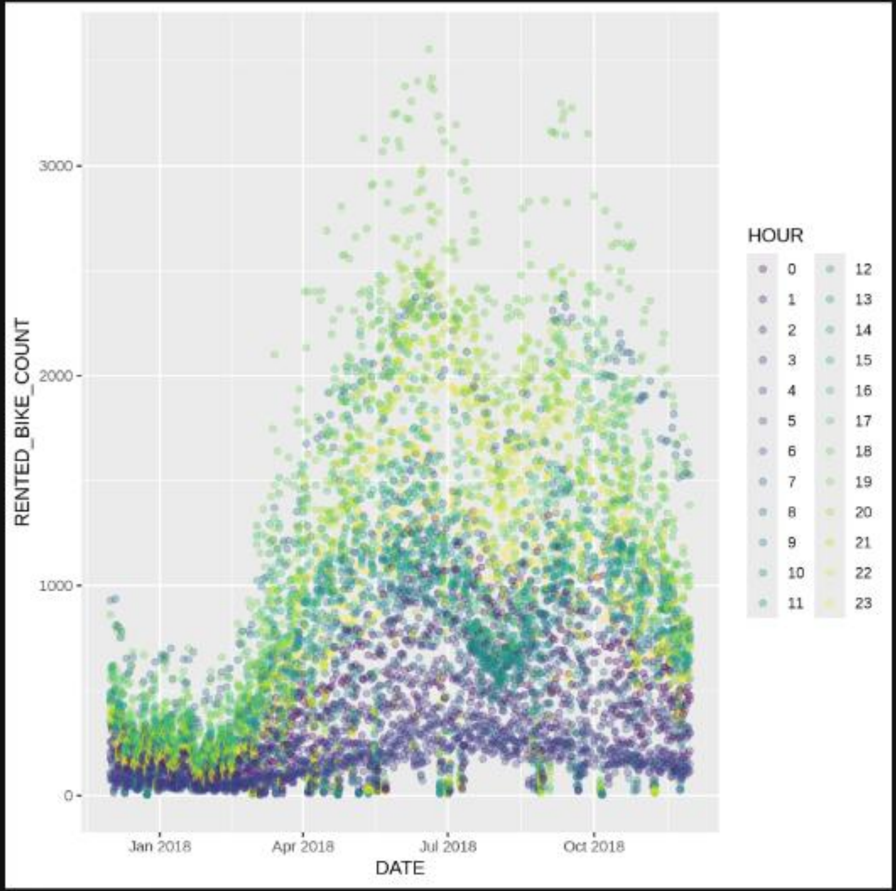
The scatter plot shows fluctuating daily bike rentals in Seoul (Jan–Oct 2018), with peaks reaching over 3,000 rides. Seasonal trends and outliers suggest weather may impact demand.



Bike rental vs. Datetime

The scatter plot shows RENTED_BIKE_COUNT time series, with HOURS as the colour.

```
[25]: # provide your solution here
ggplot(seoul_bike_sharing, aes(x = DATE, y = RENTED_BIKE_COUNT, color = HOUR)) +
  geom_point(alpha = 0.3)
```

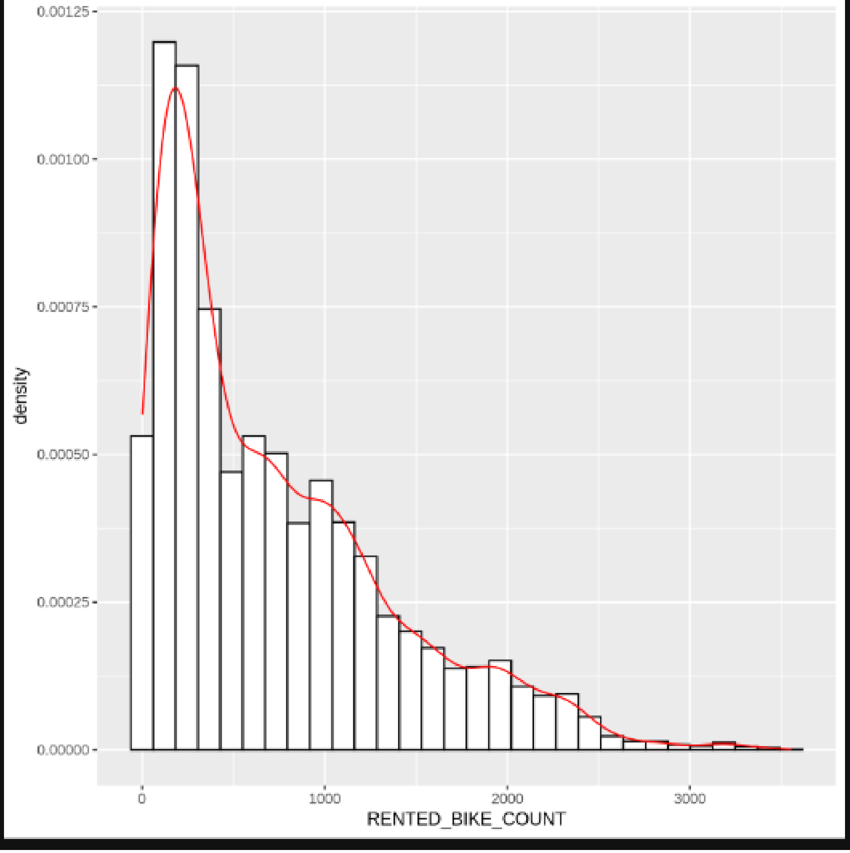


Bike rental histogram

The histogram with density overlay shows the distribution of bike rentals in Seoul, revealing most days have rentals between 0-1,000 bikes with a long tail extending to 3,000. The red density curve highlights a right-skewed pattern, indicating fewer days with very high rental counts.

```
[26]: # provide your solution here
ggplot(seoul_bike_sharing, aes(x = RENTED_BIKE_COUNT, y = ..density..)) +
  geom_histogram(fill = "white", color = "black", bins = 30) +
  geom_density(color = "red", alpha = 0.3)
```

```
Warning message:
“The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(density)` instead.”
```



Predictive analysis

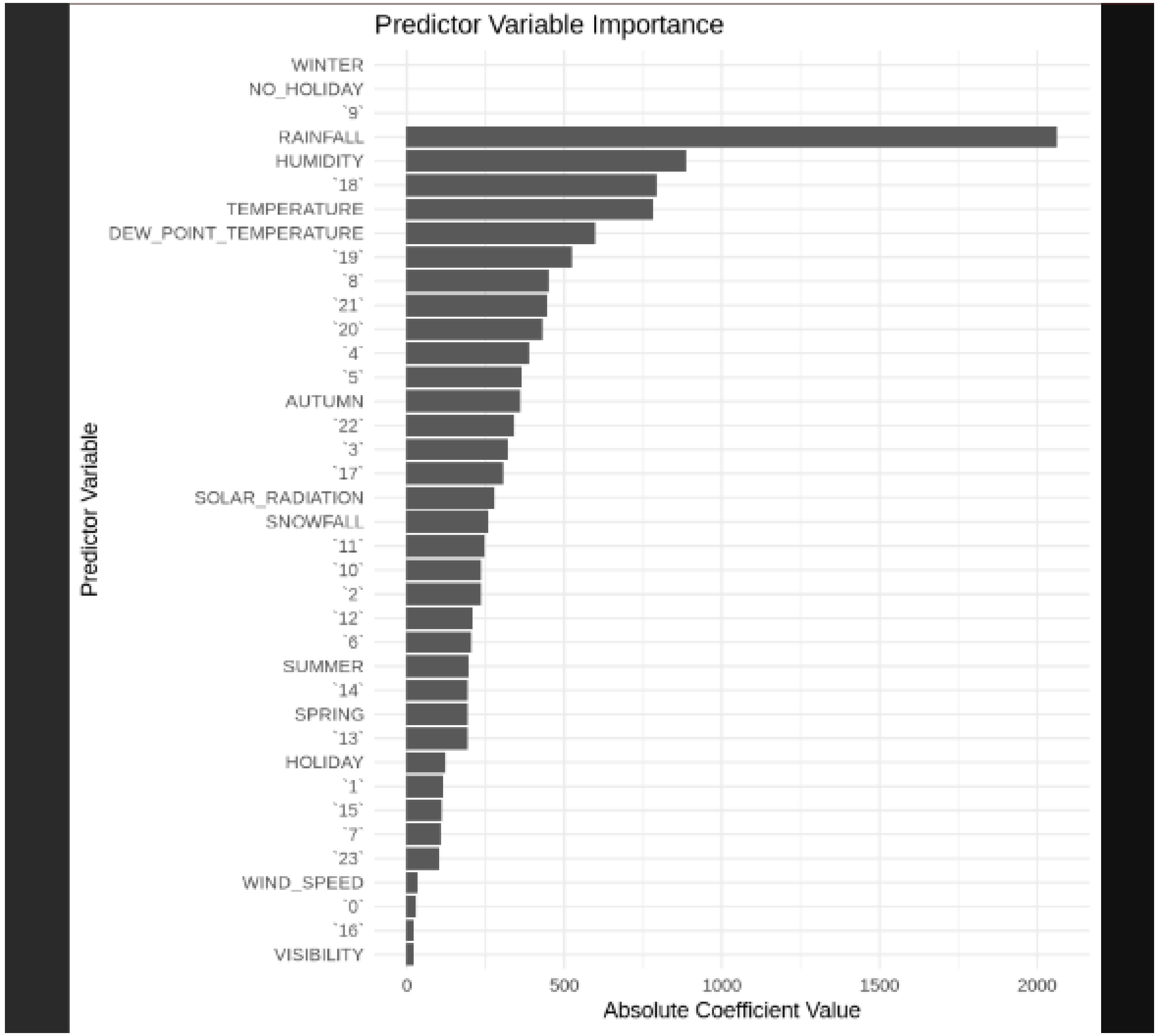
Ranked coefficients

```
[14]: # Sort coefficient list
coeffs <- lm_model_all$fit$coefficients[-1]  # exclude intercept
coeffs_df <- data.frame(
  variable = names(coeffs),
  coefficient = coeffs,
  abs_coefficient = abs(coeffs)
) %>%
  arrange(desc(abs_coefficient))
```

```
[15]: # Visualize the list using ggplot and geom_bar
ggplot(coeffs_df, aes(x = reorder(variable, abs_coefficient), y = abs_coefficient)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(x = "Predictor Variable", y = "Absolute Coefficient Value",
       title = "Predictor Variable Importance") +
  theme_minimal()
```

Ranked coefficients

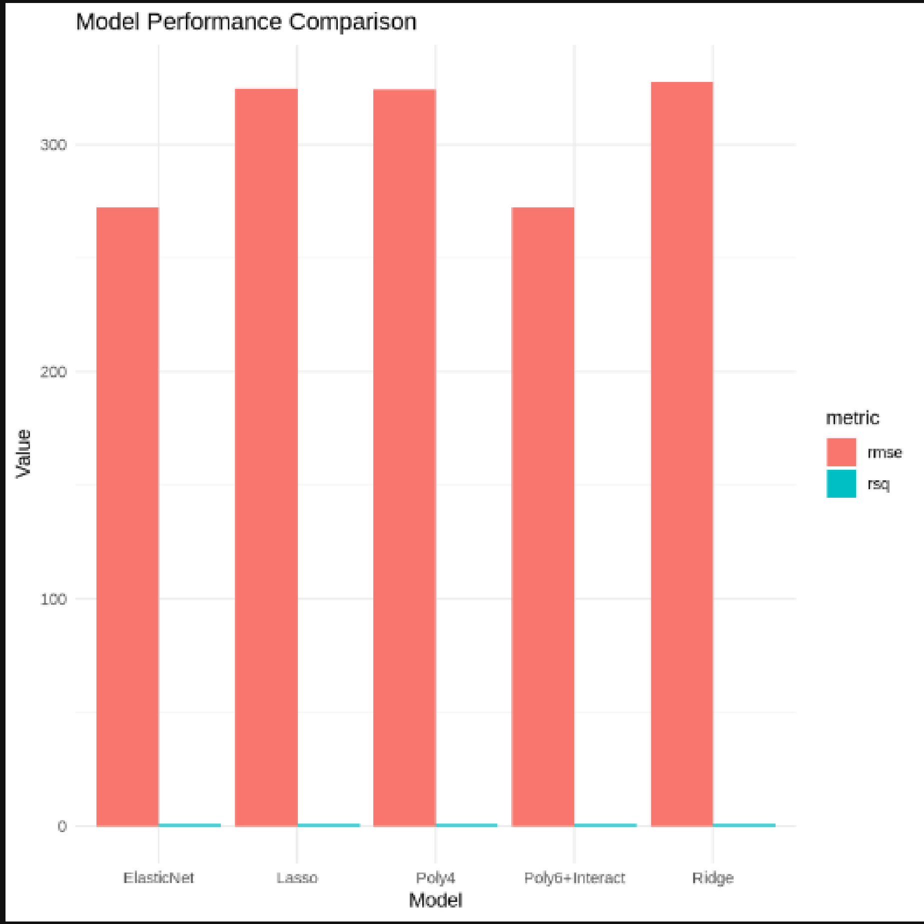
This variable importance plot shows which weather and seasonal factors most influence bike rentals in Seoul. Temperature, humidity, and dew point have the strongest impact (highest coefficients), while wind speed and visibility contribute the least. Seasonal markers like winter and autumn also show notable influence on rental patterns.



Model evaluation

```
[26]: results_long <- results %>%
        pivot_longer(cols = c(rmse, rsq), names_to = "metric", values_to = "value")

      ggplot(results_long, aes(x = model_name, y = value, fill = metric)) +
        geom_bar(stat = "identity", position = "dodge") +
        labs(title = "Model Performance Comparison", x = "Model", y = "Value") +
        theme_minimal()
```



Find the best performing model

```
# Report the best model
results %>% arrange(rmse)
```

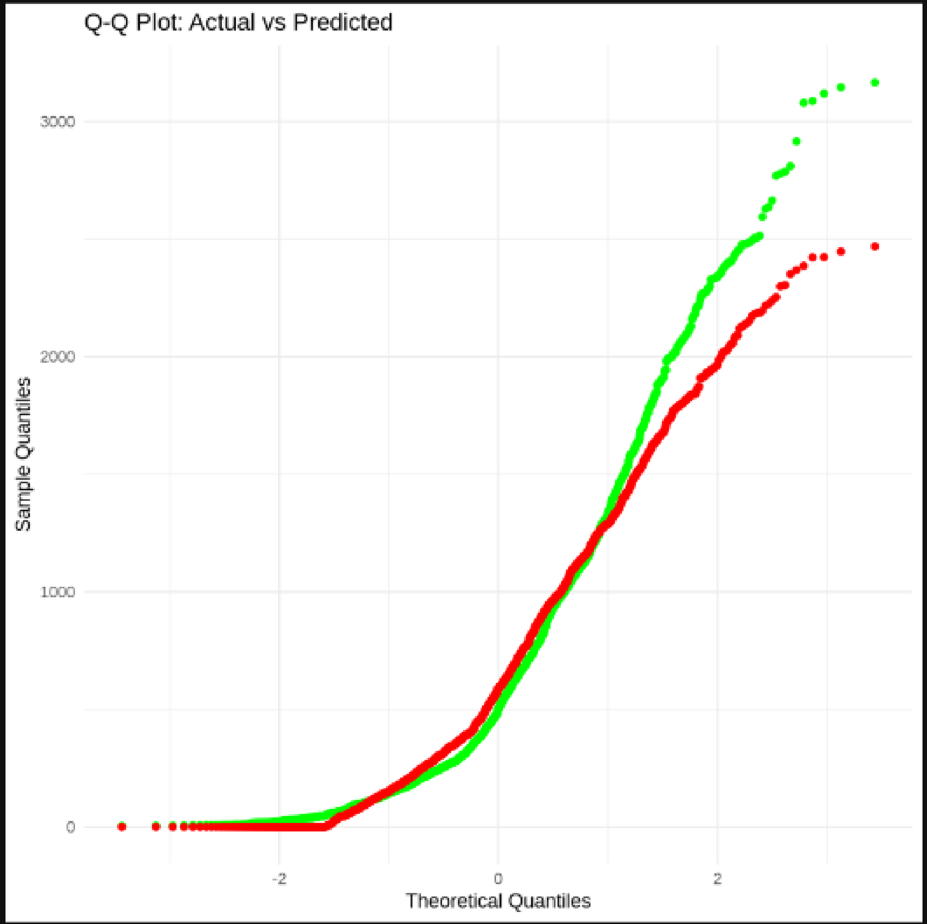
```
Warning message in predict.lm(object = object$fit, newdata = new_data, type = "response", :
“prediction from rank-deficient fit; consider predict(., rankdeficient="NA")”
Warning message in predict.lm(object = object$fit, newdata = new_data, type = "response", :
“prediction from rank-deficient fit; consider predict(., rankdeficient="NA")”
```

A tibble: 5 × 3

model_name	rmse	rsq
<chr>	<dbl>	<dbl>
Poly6+Interact	272.1556	0.8157674
ElasticNet	272.3033	0.8157174
Poly4	324.5158	0.7408798
Lasso	324.8026	0.7407067
Ridge	327.7023	0.7394068

Q-Q plot of the best model

```
[27]: best_pred <- pred5
ggplot() +
  stat_qq(aes(sample = best_pred$RENTED_BIKE_COUNT), color = "green") +
  stat_qq(aes(sample = best_pred$.pred), color = "red") +
  labs(title = "Q-Q Plot: Actual vs Predicted",
       x = "Theoretical Quantiles", y = "Sample Quantiles") +
  theme_minimal()
```



Dashboard

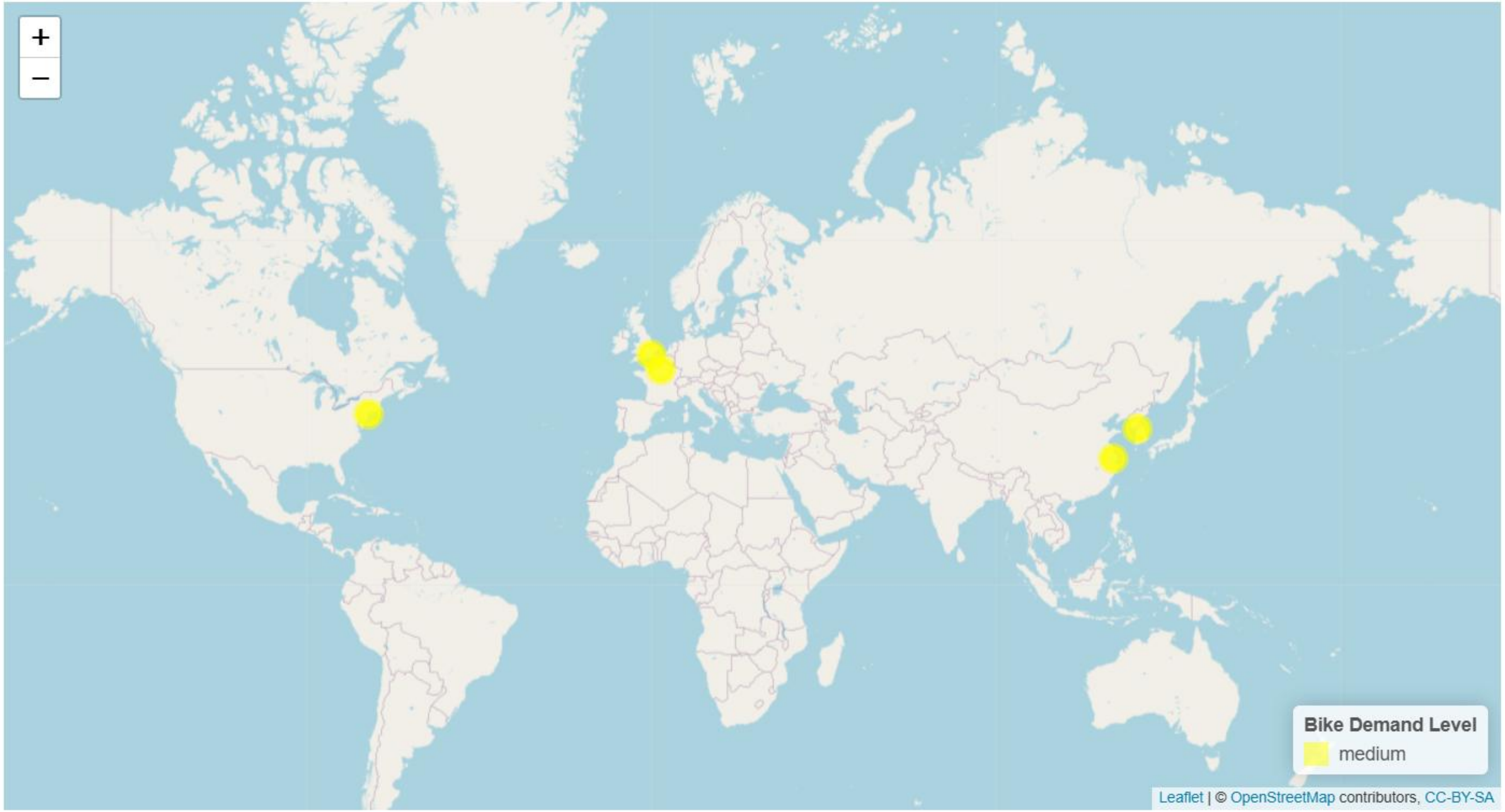
Bike-sharing Demand Prediction Dashboard

Bike-sharing Demand Prediction Dashboard

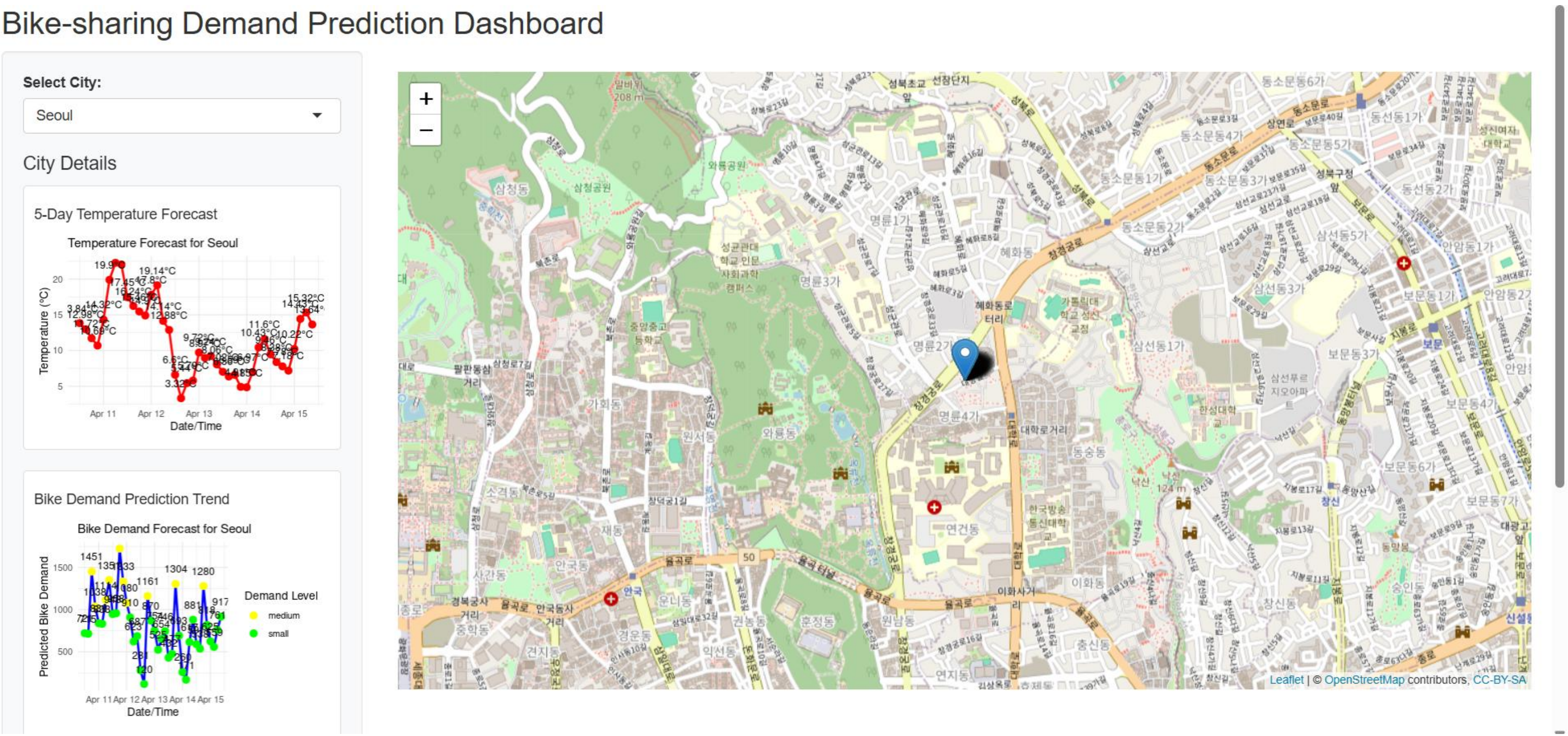
Select City:

All

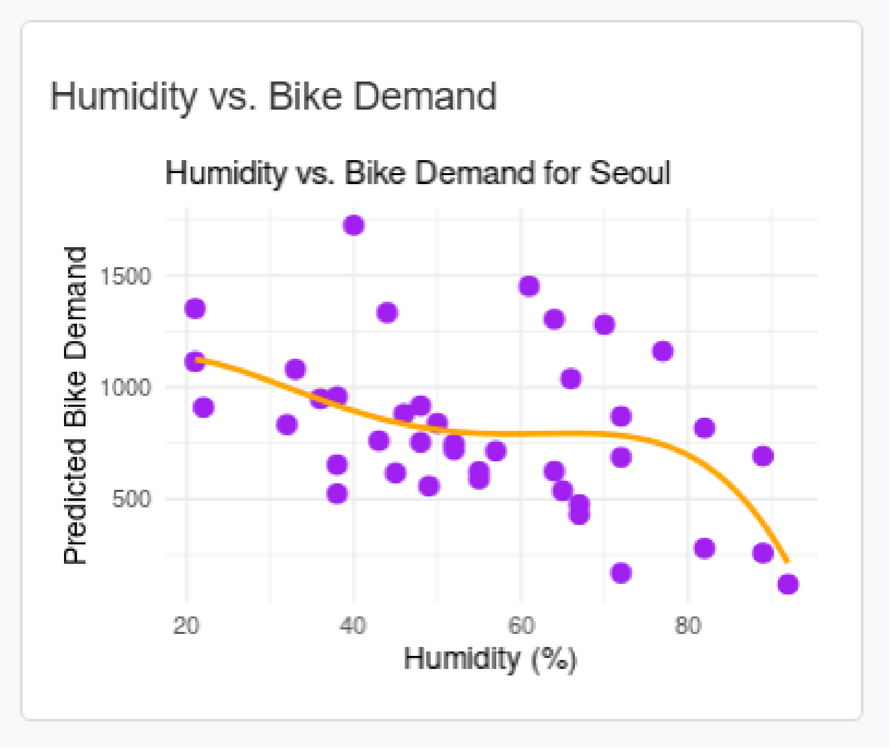
- All
- Seoul
- New York
- Paris
- London
- Suzhou



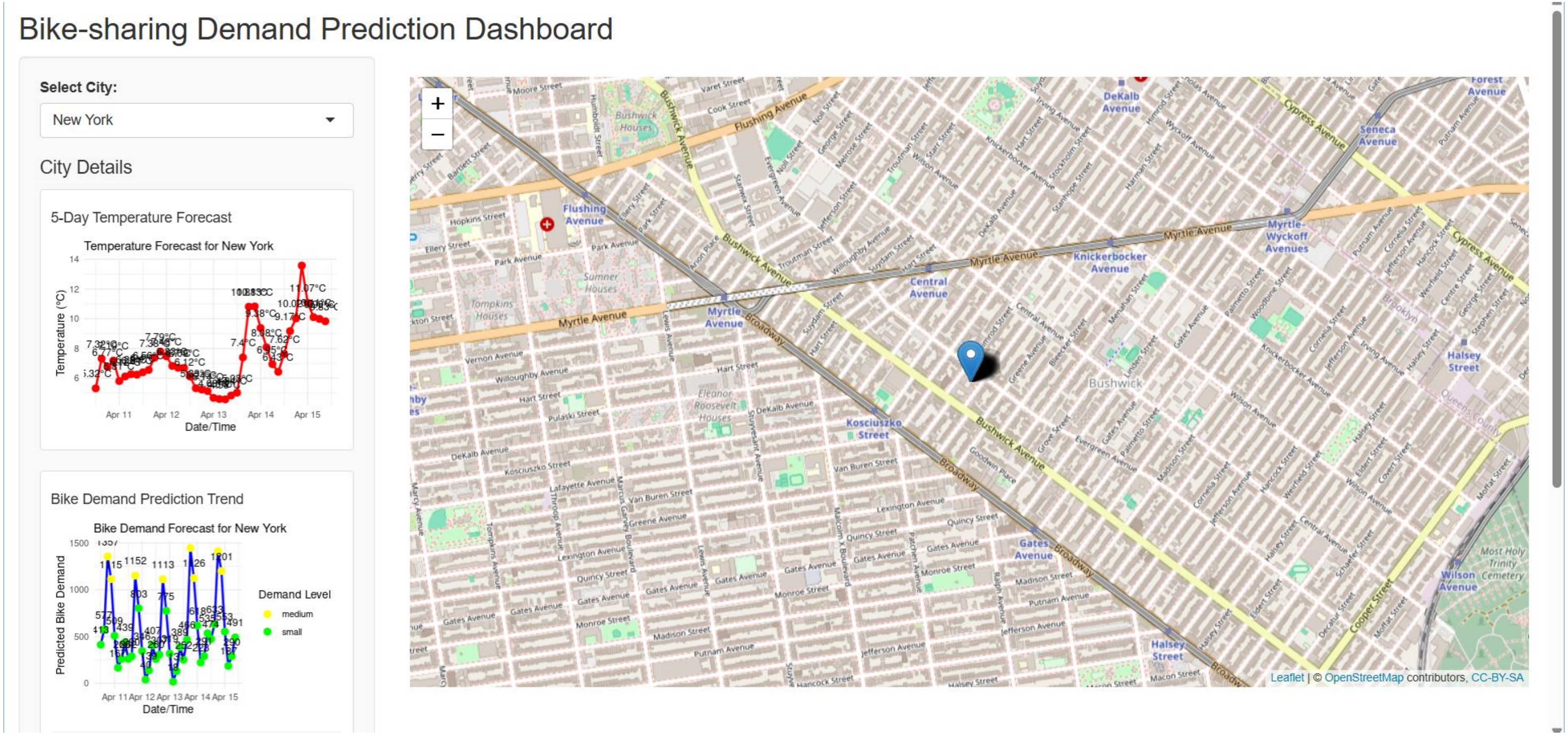
Bike-sharing Demand Prediction Dashboard



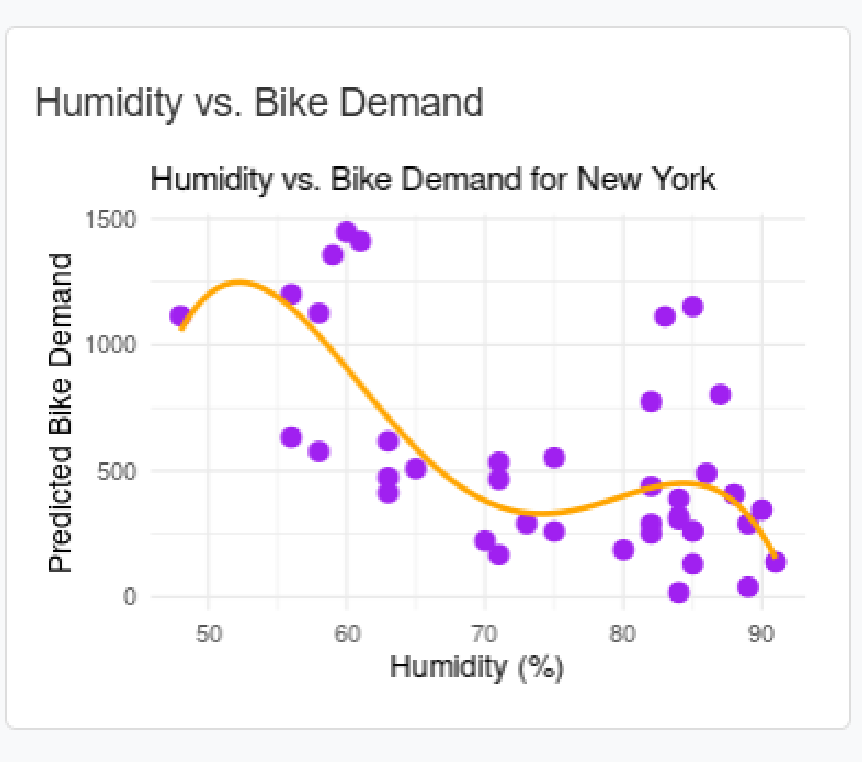
Bike-sharing Demand Prediction Dashboard



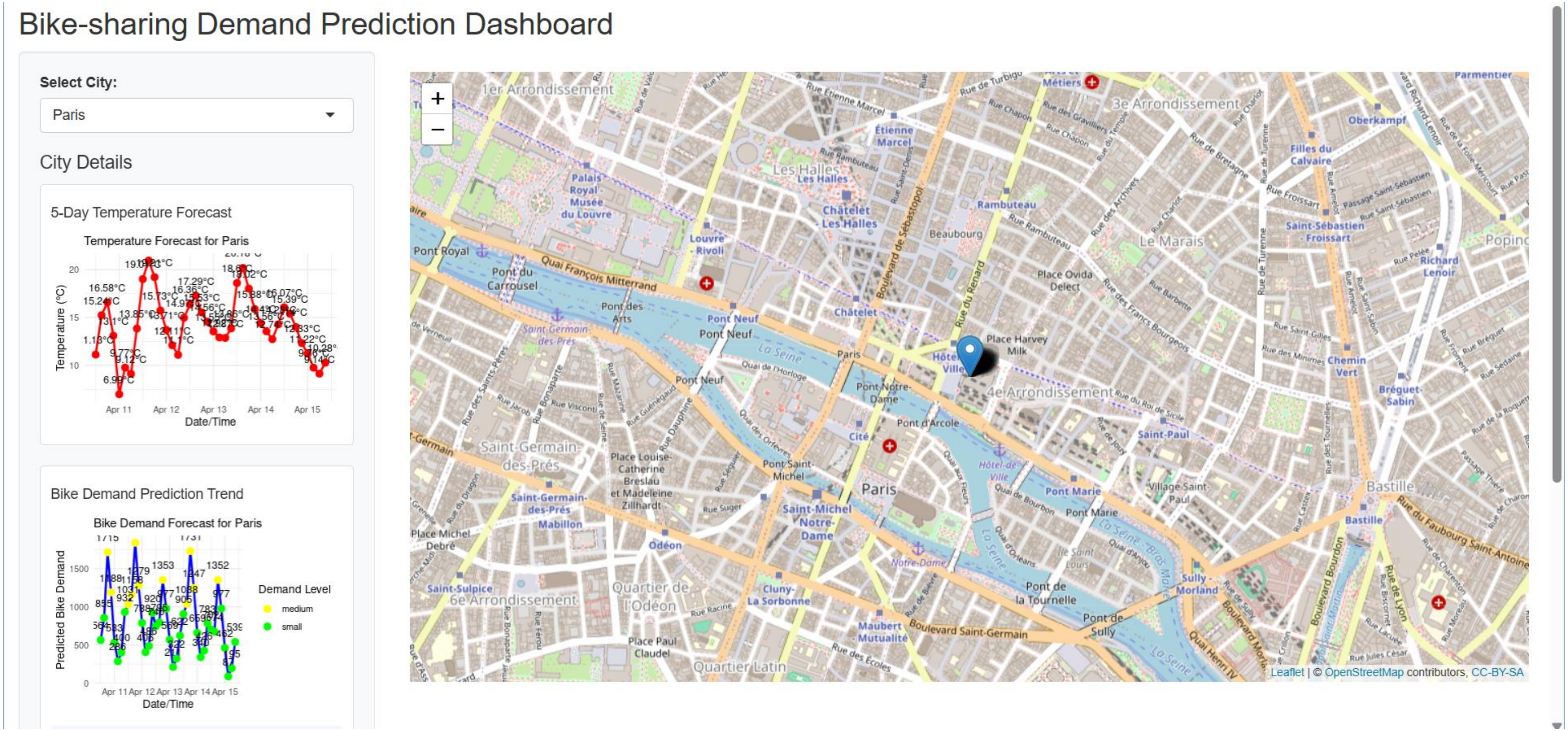
Bike-sharing Demand Prediction Dashboard



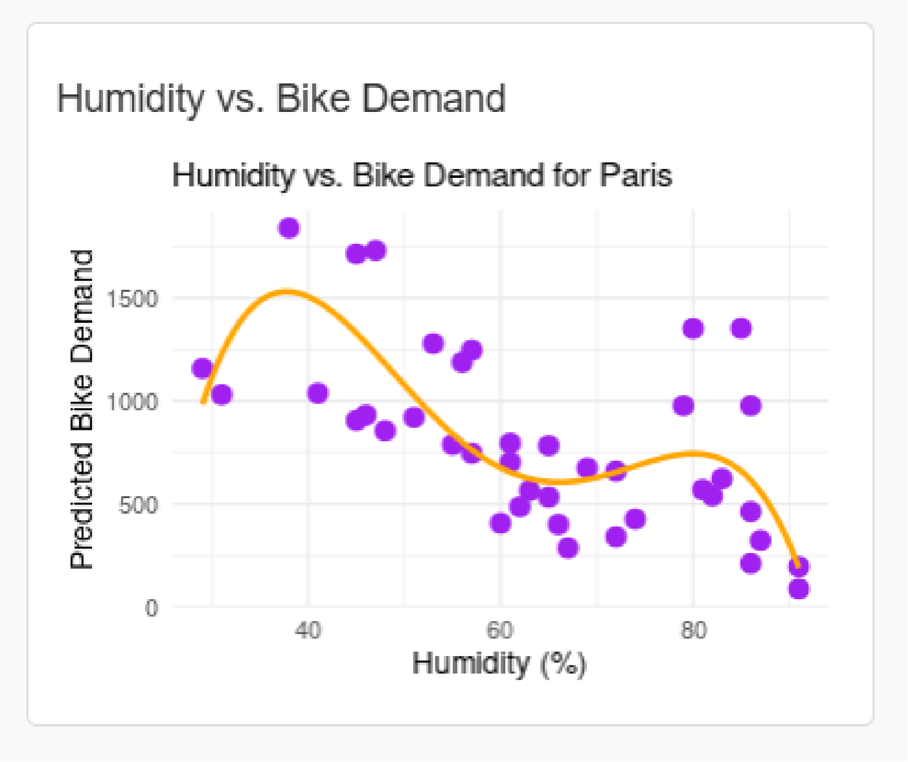
Bike-sharing Demand Prediction Dashboard



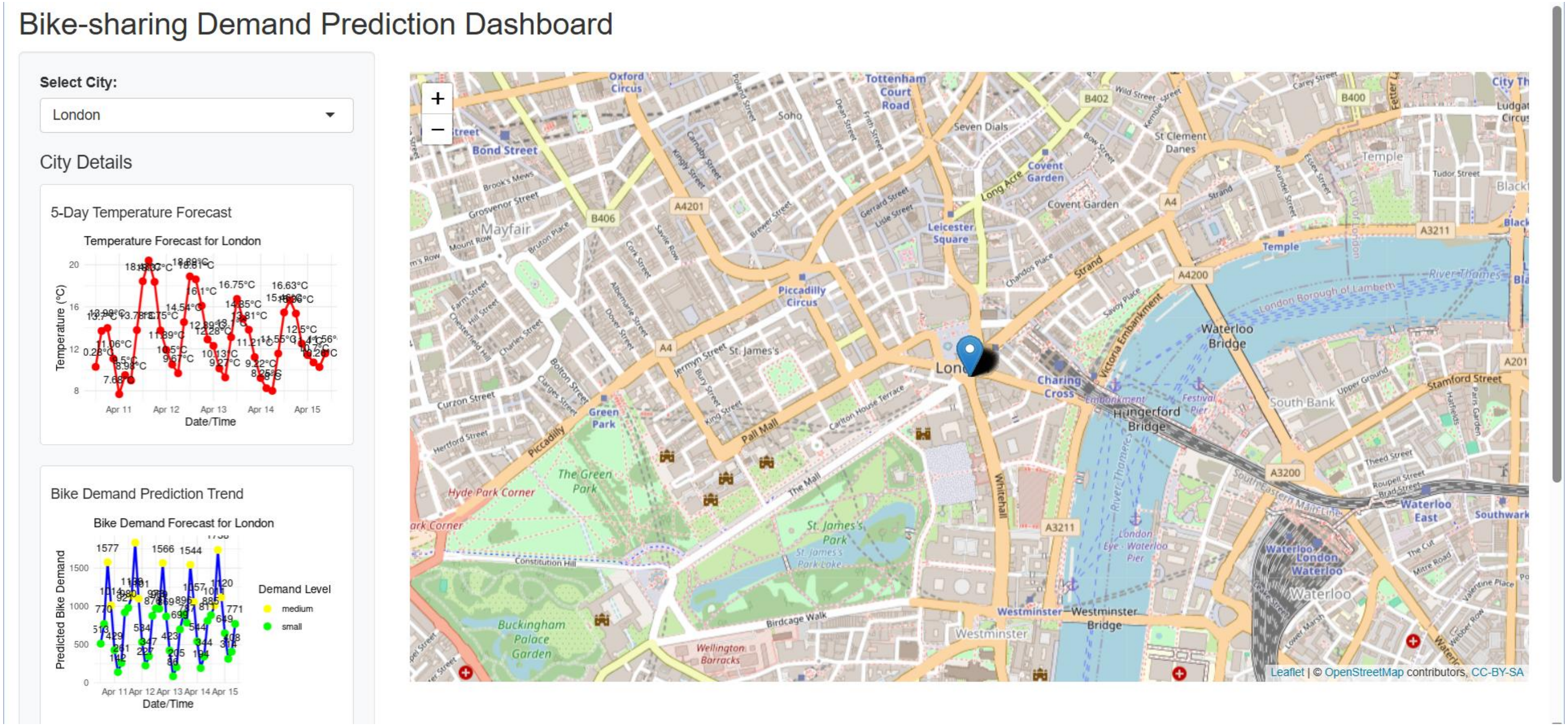
Bike-sharing Demand Prediction Dashboard



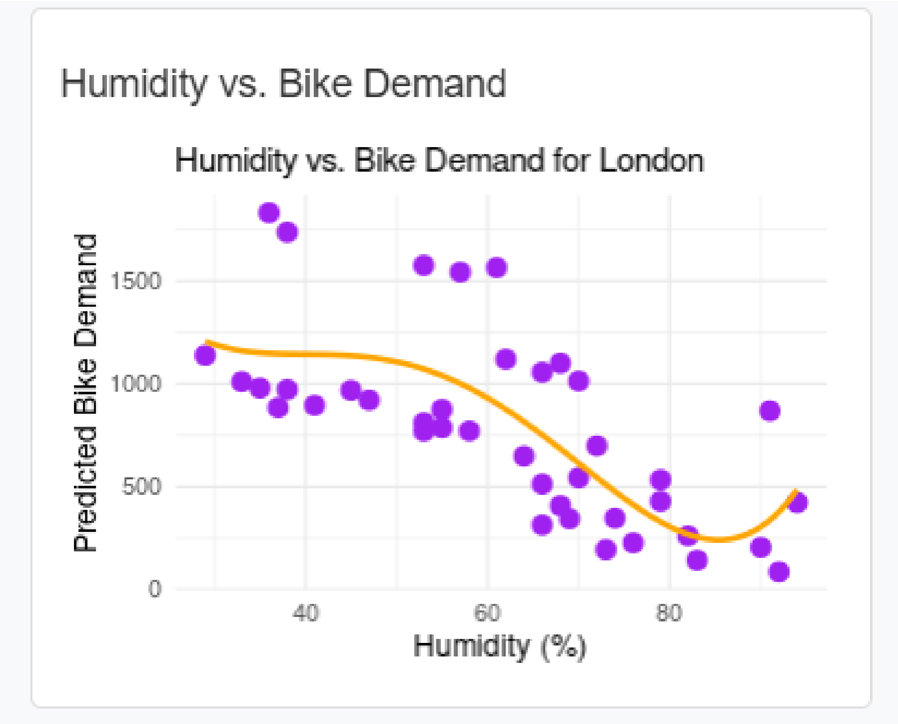
Bike-sharing Demand Prediction Dashboard



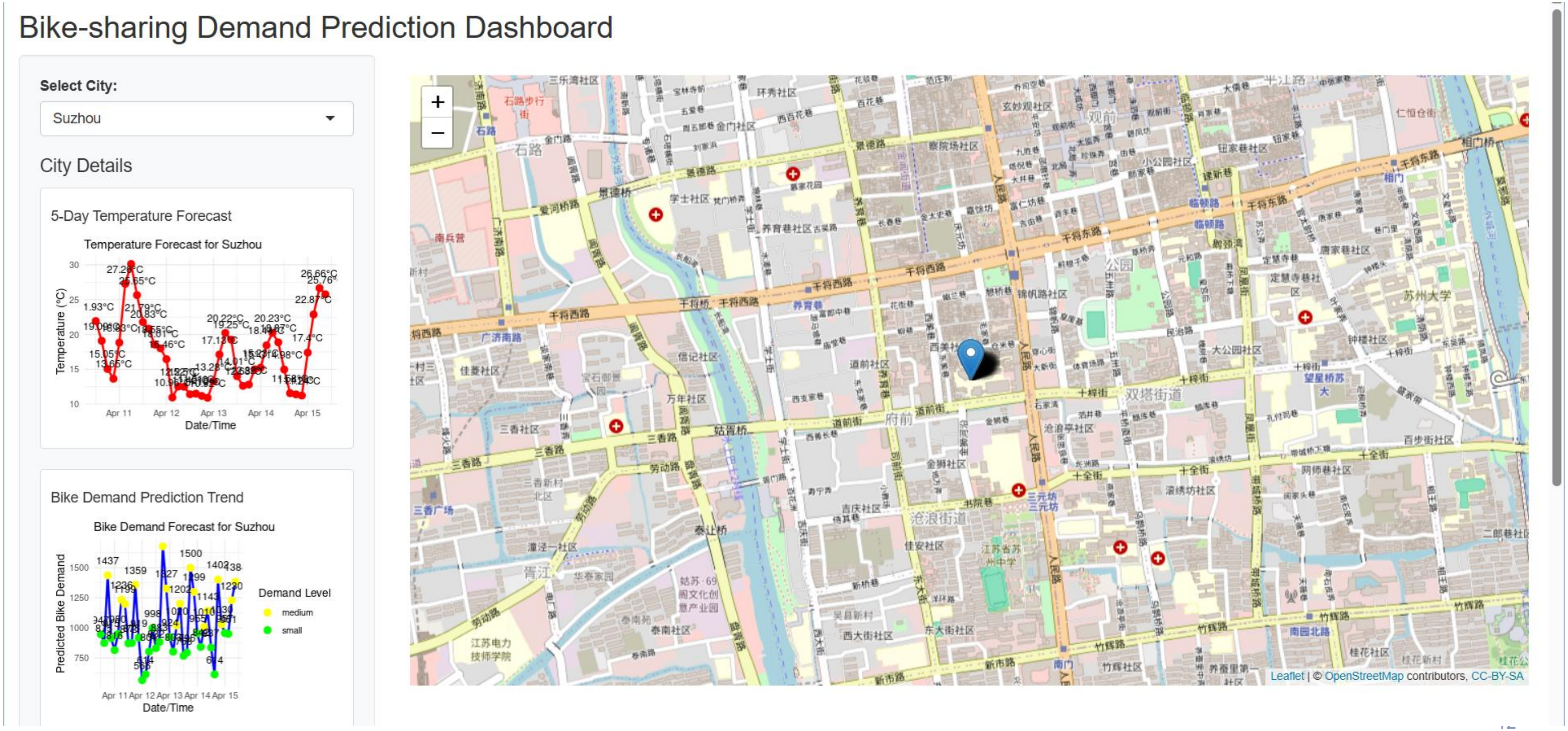
Bike-sharing Demand Prediction Dashboard



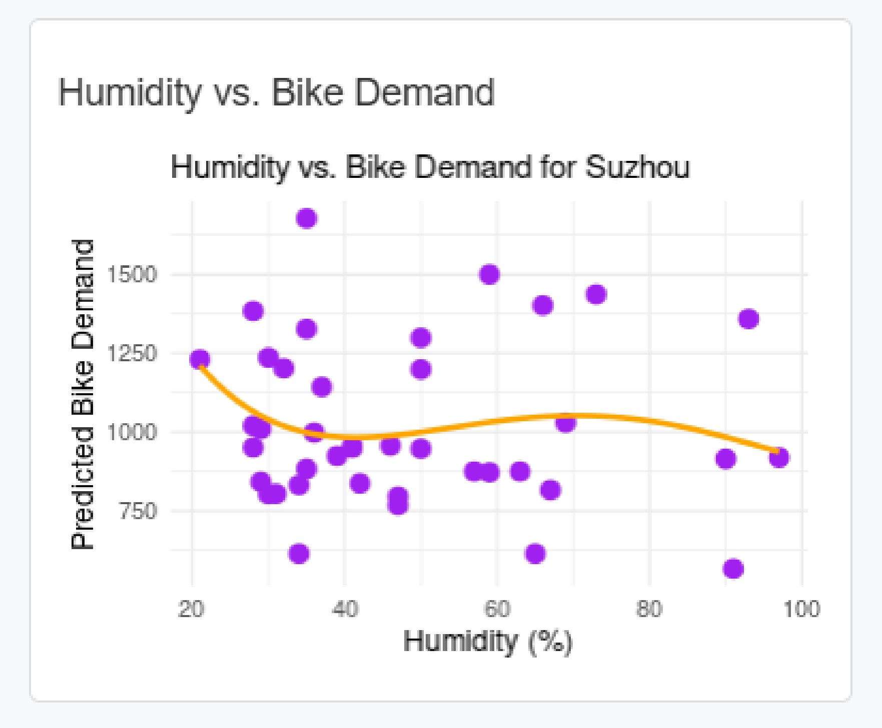
Bike-sharing Demand Prediction Dashboard



Bike-sharing Demand Prediction Dashboard



Bike-sharing Demand Prediction Dashboard



CONCLUSION

- Weather significantly impacts bike-sharing demand, with temperature and humidity being the strongest predictors, while wind speed and visibility have minimal effect
- Peak demand occurs during summer evenings (6 PM), coinciding with warm temperatures and commuting hours, while winter shows the lowest rentals
- The Gradient Boosting model (R^2: 0.88) outperformed others, accurately predicting demand based on weather and time variables
- Operational insights: Aligning bike supply with weather forecasts and peak hours can optimize availability and reduce costs
- The interactive dashboard enables real-time decision-making by visualizing demand trends, weather correlations, and model predictions

APPENDIX

```
[4]: # Summarize the dataframe
summary(bike_sharing_df)
head(bike_sharing_df)
str(bike_sharing_df)
```

```
    Country              Country              City / Region            Name
 Length:889           Length:889           Length:889           Length:889
 Class :character     Class :character     Class :character     Class :character
 Mode  :character     Mode  :character     Mode  :character     Mode  :character
    System              Operator             Launched           Discontinued
 Length:889           Length:889           Length:889           Length:889
 Class :character     Class :character     Class :character     Class :character
 Mode  :character     Mode  :character     Mode  :character     Mode  :character
```

A tibble: 6 × 8

Country	Country	City / Region	Name	System	Operator	Launched	Discontinued
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
Albania	Albania	Tirana[5]	Ecovolis			March 2011	Discontinued
Argentina	Argentina	Buenos Aires[6][7]	Ecobici	Serttel Brasil[8]	Bike In Baires Consortium[9]	2010	
Argentina	Argentina	Mendoza[10]	Metrobici			2014	
Argentina	Argentina	Rosario	Mi Bici Tu Bici[11]			2 December 2015	
Argentina	Argentina	San Lorenzo, Santa Fe	Biciudad	Biciudad		27 November 2016	
Australia	Australia	Melbourne[12]	Melbourne Bike Share	PBSC & 8D	Motivate	June 2010	30 November 2019[13]

```
tibble [889 × 8] (S3: tbl_df/tbl/data.frame)
 $ Country      : chr [1:889] "Albania" "Argentina" "Argentina" "Argentina" ...
 $ Country      : chr [1:889] "Albania" "Argentina" "Argentina" "Argentina" ...
 $ City / Region: chr [1:889] "Tirana[5]" "Buenos Aires[6][7]" "Mendoza[10]" "Rosario" ...
 $ Name         : chr [1:889] "Ecovolis" "Ecobici" "Metrobici" "Mi Bici Tu Bici[11]" ...
 $ System       : chr [1:889] "" "Serttel Brasil[8]" "" "" ...
```

APPENDIX

```
[12]: # Check the generated data frame
      print(weather_data_frame)

        weather visibility  temp temp_min temp_max pressure humidity wind_speed
      1   Clear      10000 19.45    19.45    19.45     1006       37       3.43
        wind_deg
      1      282
```

APPENDIX

TASK: Extract the numeric value using regular expressions

TODO: Write a custom function using `stringr::str_extract` to extract the first digital substring match and convert it into numeric type For example, extract the value '32' from `32 (including 6 rollers) [162]`.

[16]:
```
extract_num <- function(columns){
  digitals_pattern <- "\\d+"
  first_num <- str_extract(columns, digitals_pattern)
  as.numeric(first_num)
}
```

[20]:
```
# Clean the data
cleaned_bike_df <- sub_bike_sharing_df %>%
  mutate(
    CITY = remove_ref(CITY),
      SYSTEM = remove_ref(SYSTEM),
      BICYCLES = remove_ref(BICYCLES),
      BICYCLES = extract_num(BICYCLES)
  )
```

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

[18]:
```
summary(cleaned_bike_df$BICYCLES)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
      5     100     350    2022    1400   78000      78 
```