Applied Data Science with R Capstone project

Brian Seko

August 2023

Outline

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

Executive Summary

- Focus: Optimize bike rental supply for better availability and accessibility.
- Data Methods: Used standardized and normalized data, scraped Wikipedia for bike-sharing trends.
- Key Findings:
- Peak Months: June and September see highest rental rates.
- High Demand Hours: 6am to 8pm.
- Key Variables: Weather conditions, specifically temperature and rainfall, influence demand the most.
- Demand Fluctuations: Typical rentals range from 200-300, but can spike to 3500.
- City Trends: Similar cities in terms of population and climate show comparable rental patterns.
- Impact: Supply adjustments based on predictive analytics can reduce operational costs.

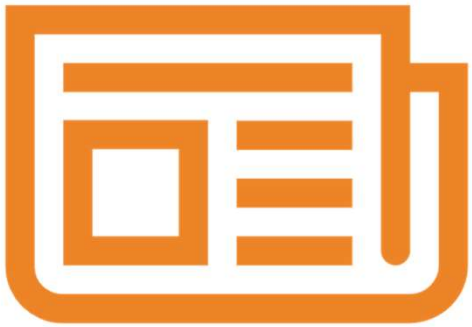
Introduction

- Urban Bike Rentals
  - Critical for city mobility
  - Need to balance availability & accessibility
- Challenge
  - Oversupply = High Costs
  - Undersupply = Lost Revenue
- Solution
  - ML model to predict hourly demand

Key Predictors

- Weather
- Season
- Hour of Day

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

Methodology

Data collection

- Data Sources for Bike Sharing Demand
  - Seoul Bike Sharing Dataset

- Weather Metrics: Temp, Humidity, Windspeed, etc.
- Data Retrieval: Utilized download.file function for dataset acquisition.
- Source: IBM Cloud Storage
- Open Weather API Data

- Provides: 5-day forecast for multiple cities
- Method: HTTP requests to API, stored in data frame
- Tools: JSON parsing, HTTP GET loop, saved to CSV
- Source: OpenWeather
- Global Bike Sharing Systems Dataset

- Content: Active bike-sharing programs globally

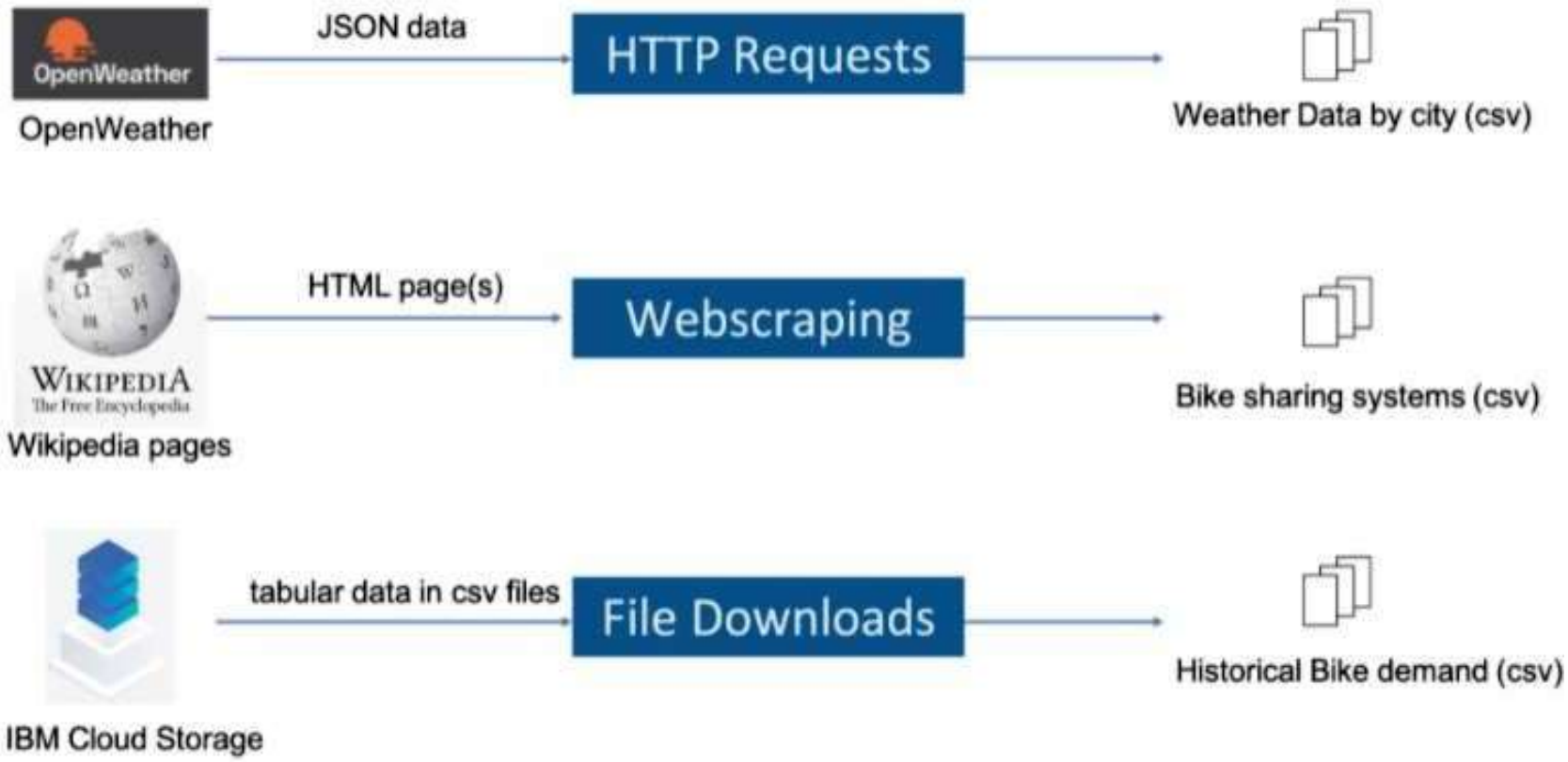
Data collection

Data Collection Methods

- Wikipedia Bike-Sharing Data
  - Approach: Web scraping using read_html, html_node, and html_table
  - Output: Data saved to CSV
  - Source: Wikipedia

- World Cities Data
  - Approach: Used download.file for data acquisition
  - Content: Major cities' name, latitude, longitude
  - Source: IBM Cloud Storage

Data collection



Data wrangling

Advanced Data Wrangling Techniques

- Tools & Libraries
  - Utilized stringr library and Regular Expressions for text manipulation and cleaning.
- Column & Variable Normalization
  - Unified column names across datasets.
  - Used str_replace_all to replace spaces in variable names with underscores, ensuring data consistency.
- Cleaning 'raw_bike_sharing_systems.csv'
  - Purged unwanted reference links using tailored regular expressions.
  - Employed str_replace_all to substitute reference link markers [...] with blank spaces, simplifying the data.
- Numerical Data Extraction
  - Isolated numerical values embedded within text in the 'raw_bike_sharing_systems.csv' dataset.
  - Utilized str_extract to capture and isolate the first number in each relevant field for easier analysis.
- Goal: To create a clean, standardized dataset for efficient and accurate downstream analysis.

Data wrangling

Data Wrangling in 'raw_seoul_bike_sharing.csv'

- Tools Used: dplyr library
- Handling Missing Values
  - Target Variable: TEMPERATURE
  - Method: Used is.na filter to identify missing values.
  - Imputation: Replaced missing values with the mean of existing values.
- Data Normalization
  - Approach: Min-Max Normalization
  - Execution: Utilized mutate to apply normalization to selected variables.

Goal: To create a standardized, complete dataset for more accurate analysis.

EDA with SQL

- Metrics
  - Total & active rental hours
- Peak Time
  - Max rental date & hour
- Weather Data
  - 3-hr interval focus
- Seasonal Scope
  - Dataset's date range & covered seasons
- Rentals & Weather
  - Sort by avg. bike count
- Target Cities
  - 15-20K bikes, with geo & pop data
- Seasonal Trends
  - Top 10 avg. hourly rentals & temps
- Seoul Bike Inventory
  - Joined city & bike tables, key stats

Goal: Uncover key insights for optimizing bike rentals.

EDA with data visualization

- Scatter Plots
  - Rented Bike Count vs. Date, colored by Hour
  - Correlation between Bike Count & Temp, segmented by Seasons
- Histogram
  - With Kernel Density Overlay
- Boxplots
  - Bike Count vs. Hour, grouped by Seasons
- Aggregated Metrics
  - Daily Rainfall & Snowfall via summarize()

Objective: Unveil patterns and correlations for informed decisions.

Predictive analysis

- The primary goal here is to construct a robust linear regression model that accurately predicts hourly bike-sharing demand, focusing mainly on weather and datetime variables.

- The Basics: Linear Models
  - First things first, we split our data into training and testing datasets. With this data, we go ahead and build two distinct models:
- Model 1: This one uses only weather-related variables.
- Model 2: A more comprehensive model, incorporating both weather and datetime variables.
- Taking It Up a Notch: Model Enhancement
- No model is perfect right off the bat, so we improve ours by adding some complexities:

- Polynomial Terms: For capturing curvilinear relationships.
- Interaction Terms: To understand how combined variables affect the output.
- Regularization: To penalize overly complex models and prevent overfitting.

Predictive analysis cont.

- The Metrics Game: Evaluating Success
  - To determine the best model among the contenders, we deploy some time-tested evaluation metrics like R-squared and RMSE. We even add a little flair by generating a bar graph to compare the predictor variable coefficients.

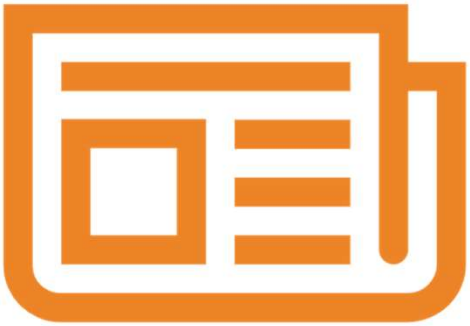
- The Final Lap: Optimization
  - It's all about tweaking and fine-tuning to get that perfect model. Through experimentation and maybe a bit of luck, we aim to land on the most efficient and accurate model to predict hourly bike demand.

- The End Goal
  - The entire exercise aims to pinpoint the optimal model, allowing for highly accurate hourly bike demand predictions. This could be a game-changer in optimizing bike-sharing services, especially in cities where this is a popular form of transport.

Build a R Shiny dashboard

- R Shiny App for Bike Forecasting
- Dive into an interactive Leaflet map that comes alive with a city-specific dropdown.
- Visualize the scene with ggplot graphs, showing not just bike demand but also temperature trends for your chosen city.
- Enhance your experience with city detail plots and straightforward trend lines for temperature and demand.
- Get hands-on with dynamic demand predictions and peek into how humidity correlates with bike-sharing needs.
- End Game: A one-stop, user-friendly app offering razor-sharp insights into bike-sharing demand.

Results



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

EDA with SQL

Busiest bike rental times

The peak in bike rentals was observed in June 2018, hitting a staggering 3,556 rentals at 6 pm.

Summer months consistently show elevated levels of bike rentals, reinforcing seasonal trends in user behavior.

Takeaway: The data suggests that strategic planning for bike availability is crucial, especially for high-demand periods like summer evenings.

```
> query <- "SELECT DATES, HOUR
+ FROM SEOUL_BIKE_SHARING
+ WHERE RENTED_BIKE_COUNT = (SELECT MAX(RENTED_BIKE_COUNT) FROM SEOUL_BIKE_SHARING) ;"
> view <- sqlQuery(conn,query)
> view
      DATES HOUR
1 19/06/2018   18
```

```
> query <- "SELECT RENTED_BIKE_COUNT, DATES, HOUR
+ FROM SEOUL_BIKE_SHARING
+ ORDER BY RENTED_BIKE_COUNT DESC
+ FETCH  FIRST 10 ROWS ONLY ;"
> view <- sqlQuery(conn,query)
> view
   RENTED_BIKE_COUNT      DATES HOUR
1               3556 19/06/2018   18
2               3418 21/06/2018   18
3               3404 12/06/2018   18
4               3384 20/06/2018   18
5               3380 04/06/2018   18
6               3365 22/06/2018   18
7               3309 08/06/2018   18
8               3298 10/09/2018   18
9               3277 17/09/2018   18
10              3256 12/09/2018   18
```

Hourly popularity and temperature by seasons

Bike rentals hit their stride at two key times across all seasons: the morning rush at 8 am and the evening commute at 6 pm.

These time slots consistently rack up the highest average rentals, making them critical hours for supply optimization.

Takeaway: Whether it's winter or summer, 8 am and 6 pm are the hotspots for bike rentals. Adjusting supply for these peak times could maximize both availability and revenue.

```
> query <- "SELECT * FROM (
+ SELECT SEASONS, HOUR, AVG(TEMPERATURE) AVG_TEMP, AVG(RENTED_BIKE_COUNT) AVG_RENTALS
+ FROM SEOUL_BIKE_SHARING
+ GROUP BY SEASONS, HOUR
+ ORDER BY AVG_RENTALS DESC)
+ WHERE ROWNUM <= 10;"
> view <- sqlQuery(conn,query)
> view
   SEASONS HOUR AVG_TEMP AVG_RENTALS
1   Summer   18 29.38804    2135.141
2   Autumn   18 16.03210    1983.333
3   Summer   19 28.27391    1889.250
4   Summer   20 27.06630    1801.924
5   Summer   21 26.27826    1754.065
6   Spring   18 15.97222    1689.311
7   Summer   22 25.69891    1567.870
8   Autumn   17 17.27778    1562.877
9   Summer   17 30.07717    1526.293
10  Autumn   19 15.06420    1515.568
```

Rental Seasonality

```
> query <- "SELECT SEASONS, AVG(RENTED_BIKE_COUNT) AVG_RENTALS
+ FROM SEOUL_BIKE_SHARING
+ GROUP BY SEASONS
+ ORDER BY SEASONS ;"
> view <- sqlQuery(conn,query)
> view
  SEASONS AVG_RENTALS
1  Autumn    924.1105
2  Spring    746.2542
3  Summer   1034.0734
4  Winter    225.5412
```

Across all seasons, analyzing the average hourly bike count reveals distinct trends.

Summer stands out as the prime time for bike rentals, boasting the highest average hourly counts.

In contrast, winter sees the lowest rate of bike rentals, indicating a seasonal dip in demand.

Takeaway: Seasonality plays a significant role in bike rentals, with summer being the peak period. This underscores the need for dynamic supply strategies tailored to each season.

Weather Seasonality

```
> query <- "SELECT SEASONS, AVG(TEMPERATURE), AVG(HUMIDITY), AVG(WIND_SPEED), AVG(VISIBILITY),
+ AVG(DEW_POINT_TEMPERATURE), AVG(SOLAR_RADIATION), AVG(RAINFALL),  AVG(SNOWFALL), AVG(RENTED_BIKE_COUNT)
+ FROM SEOUL_BIKE_SHARING
+ GROUP BY SEASONS
+ ORDER BY AVG(RENTED_BIKE_COUNT) DESC ;"
> view <- sqlQuery(conn,query)
> view
  SEASONS AVG(TEMPERATURE) AVG(HUMIDITY) AVG(WIND_SPEED) AVG(VISIBILITY) AVG(DEW_POINT_TEMPERATURE) AVG(SOLAR_RADIATION) AVG(RAINFALL)
1  Summer        26.587772      64.98143        1.609420        1501.745                 18.750136            0.7612545    0.25348732
2  Autumn        13.821683      59.04491        1.492101        1558.174                  5.150594            0.5227827    0.11765617
3  Spring        13.021759      58.75833        1.857778        1240.912                  4.091389            0.6803009    0.18694444
4  Winter        -2.540463      49.74491        1.922685        1445.987                -12.416667            0.2981806    0.03282407
  AVG(SNOWFALL) AVG(RENTED_BIKE_COUNT)
1    0.00000000              1034.0734
2    0.06350026               924.1105
3    0.00000000               746.2542
4    0.24750000               225.5412
```

Key variables that seem to be tightly linked with bike rentals are temperature, humidity, and dew point temperature.

Other variables don't align well with the patterns observed in average bike rentals.

Takeaway: If you're looking to predict bike rental demand, focus on temperature, humidity, and dew point as your leading indicators. The other factors may not offer the same predictive power.

Bike-sharing info in Seoul

1. Seoul, situated at 37.58° latitude and 127° longitude in South Korea, is a bustling city with a population of approximately 21.8 million people.
2. The city is equipped with 20,000 rental bikes to cater to its residents and visitors alike.
3. Takeaway: With a sizeable population and a substantial number of rental bikes, Seoul presents a unique landscape for studying bike rental trends and demand.

```
> query <-"SELECT a.CITY, a.COUNTRY, a.LAT, a.LNG, a.POPULATION, b.BICYCLES
+ FROM WORLD_CITIES a, BIKE_SHARING_SYSTEMS b
+ WHERE a.CITY_ASCII=b.CITY AND
+ a.CITY_ASCII = 'Seoul'
+ ;"
> view <- sqlQuery(conn,query)
> view
   CITY      COUNTRY   LAT LNG POPULATION BICYCLES
1 Seoul Korea, South 37.58 127   21794000    20000
```

Cities similar to Seoul

This query highlights cities with a total bike count ranging between 15,000 and 20,000.

Included are details such as city and country names, geographical coordinates (LAT, LNG), population figures, and the exact number of available rental bikes.

Takeaway: This focused query offers a snapshot of cities with similar bike infrastructure as Seoul, aiding in comparative analysis and trend prediction.

```
> query <-"SELECT a.CITY, a.COUNTRY, a.LAT, a.LNG, a.POPULATION, b.BICYCLES
+ FROM WORLD_CITIES a, BIKE_SHARING_SYSTEMS b
+ WHERE a.CITY_ASCII=b.CITY AND
+ BICYCLES BETWEEN 15000 and 20000
+ ;"
> view <- sqlQuery(conn,query)
> view
      CITY      COUNTRY   LAT    LNG POPULATION BICYCLES
1 Shanghai        China 31.17 121.47   22120000    19165
2    Seoul Korea, South 37.58 127.00   21794000    20000
3  Beijing        China 39.91 116.39   19433000    16000
4  Weifang        China 36.72 119.10    9373000    20000
5   Ningbo        China 29.88 121.55    7639000    15000
6    Xi¿an        China 34.27 108.90    7135000    20000
7  Zhuzhou        China 27.84 113.15    3855609    20000
```

EDA with Visualization

Bike rental vs. Date

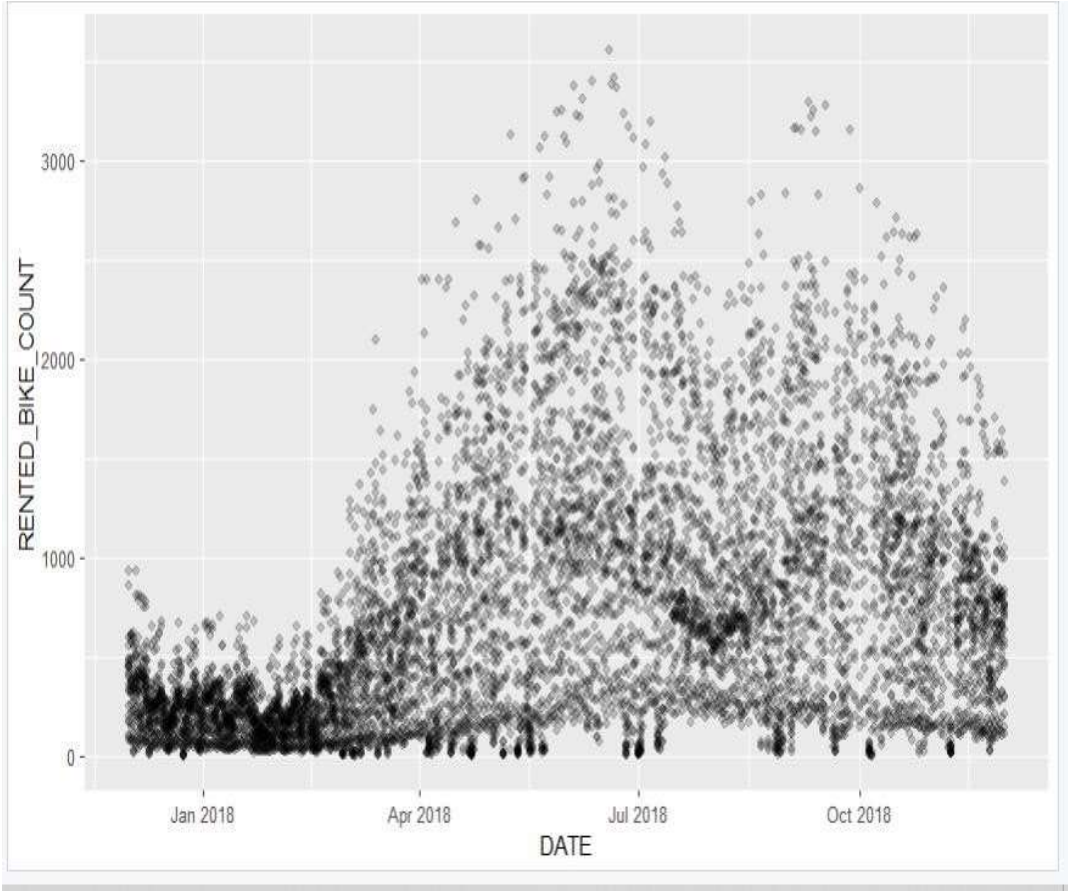
A clear seasonal pattern emerges in the data for hourly rented bike counts.

Winter months record the lowest usage, often falling below 1,000 rentals per hour.

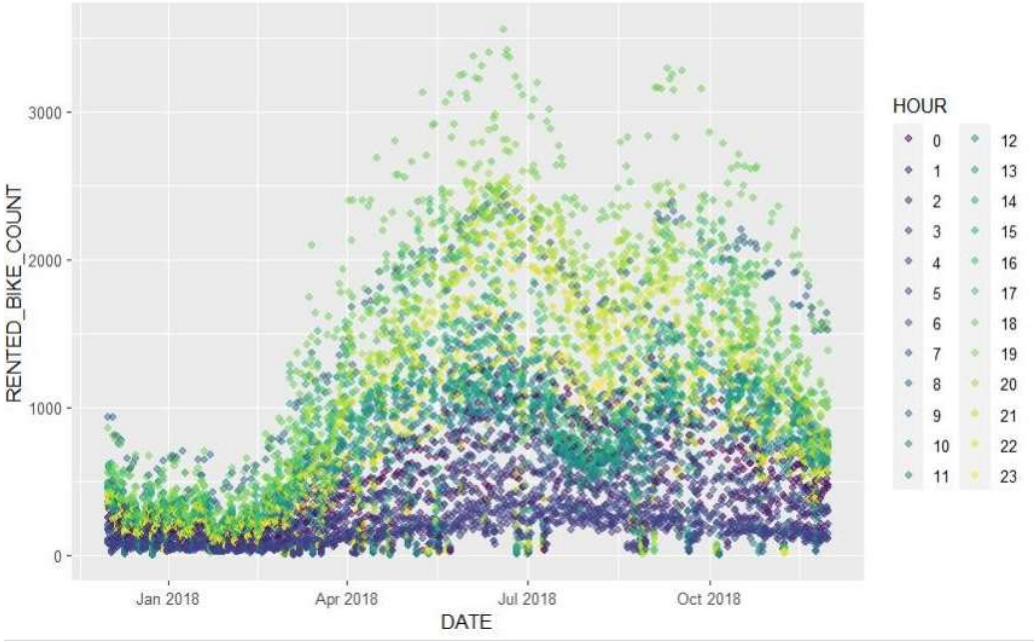
A noticeable uptick begins in spring, peaking during the June-July summer months.

Though there's a slight dip in August, a secondary peak occurs in September.

Takeaway: Understanding these seasonal trends can guide supply optimization, allowing for better planning and reduced operational costs.



Bike rental vs. Datetime



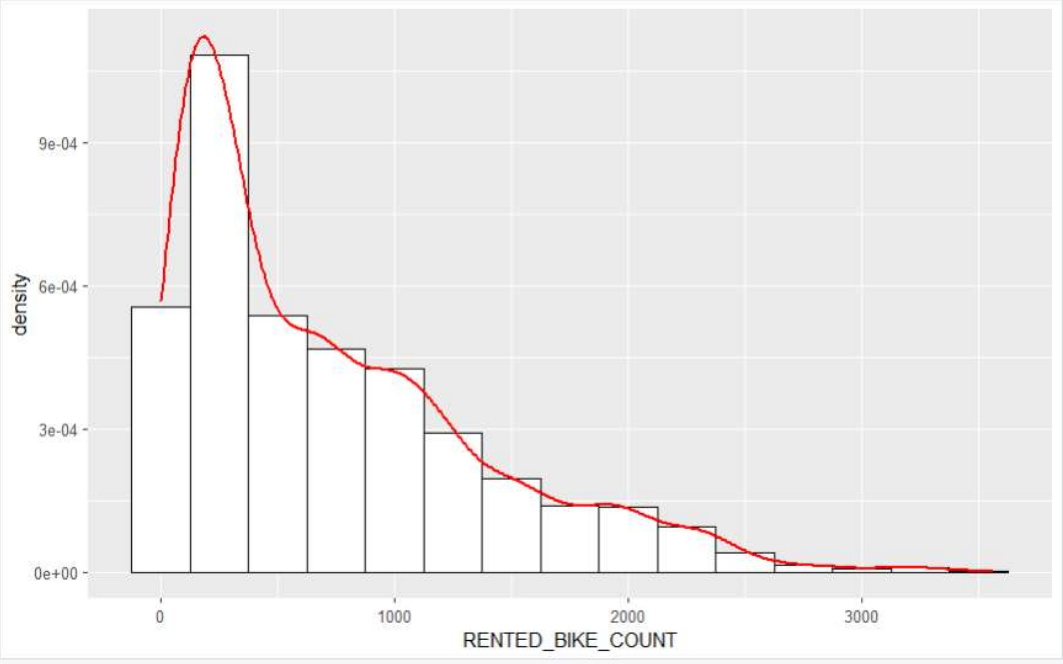
Bike rentals are predominantly concentrated during daytime hours (12 PM onwards) and taper off significantly after midnight.

This trend is consistent and logical, as late-night and early-morning hours are typically rest periods for most people.

This pattern holds true across all seasons.

Takeaway: Awareness of these daily rental cycles can assist in optimizing bike availability and maintenance schedules.

Bike rental histogram



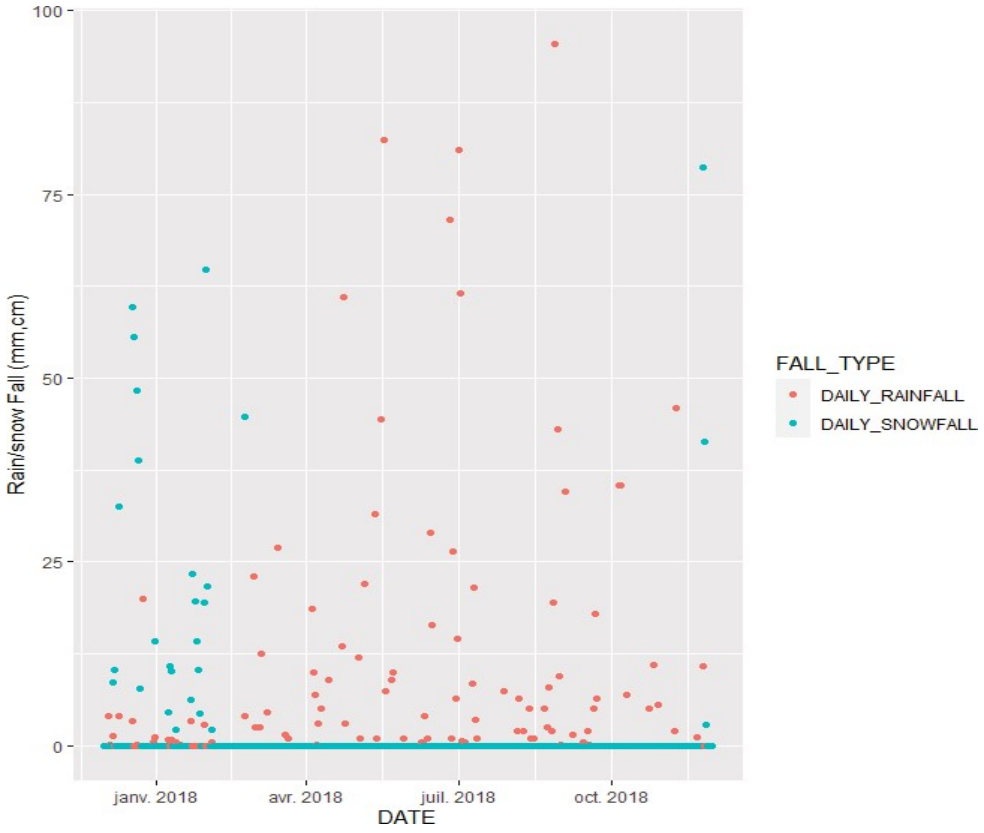
The histogram reveals a mode around 250 bikes, indicating that most of the time, bike rentals hover around this number.

However, we also spot some irregularities—additional 'bumps' at about 700, 900, 1900, and 3200 bikes suggest multiple modes or peaks within subgroups.

The tail of the distribution hints at outlier events where bike rentals shoot up far above the norm.

Takeaway: While the typical demand is moderate, there are specific instances or conditions that trigger significantly higher bike rentals. This suggests room for further investigation to pinpoint these demand drivers.

Daily total rainfall and snowfall



As expected, snowfall mainly occurs in winter, while rainfall is spread throughout the rest of the year.

Generally, the climate is dry; there's often no rain or snow.

In the given year, about 30 days experienced snowfall and around 100 days had rainfall.

Takeaway: The majority of days are precipitation-free, but when it does happen, it's more likely to be rain than snow. This could influence bike rental patterns and should be considered in demand forecasting.

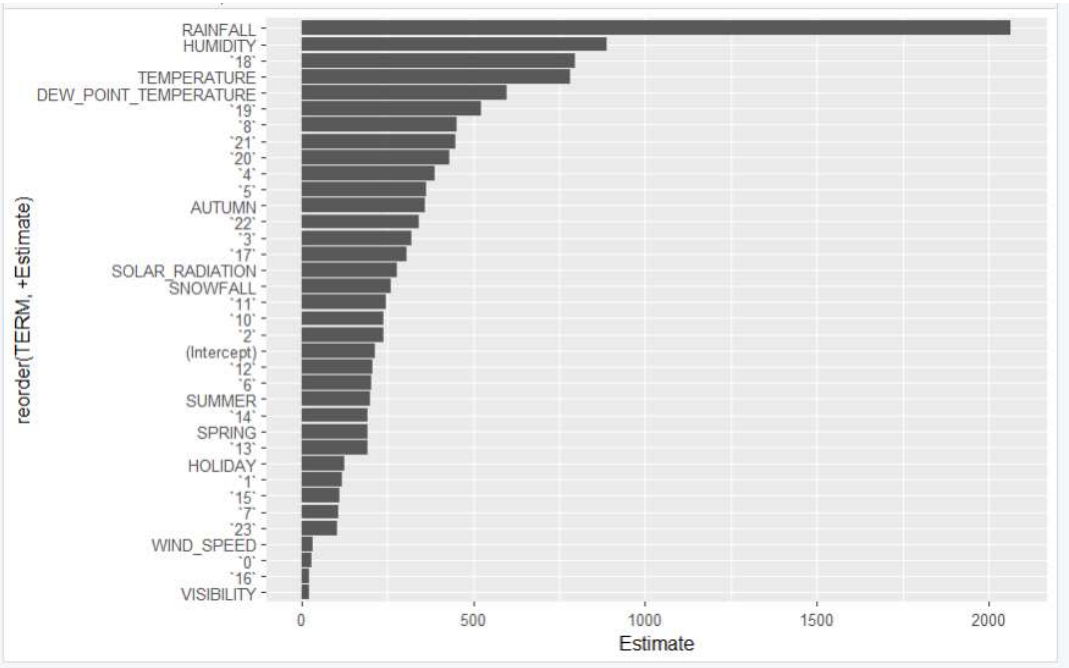
Predictive analysis

Ranked coefficients

Top factors for bike rental prediction in Seoul: Rainfall, 6pm time slot, Humidity, and Temperature.

These factors vary across seasons, adding a layer of complexity to the prediction model.

Takeaway: These dynamic variables partly explain the seasonal patterns in bike rentals. For more accurate predictions, these should be key considerations in the model.



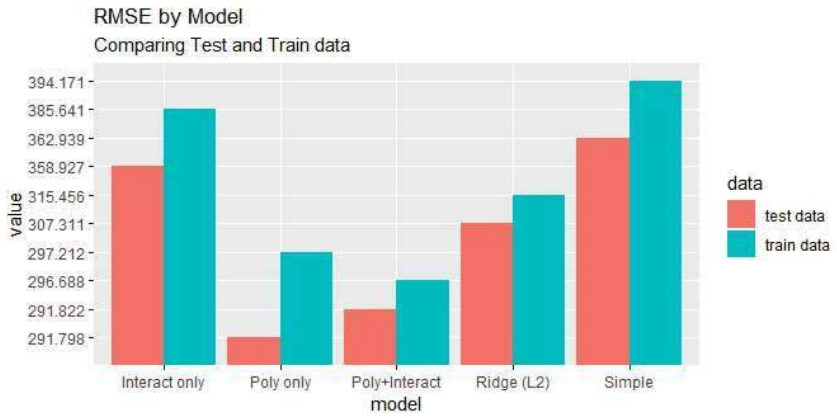
Model evaluation

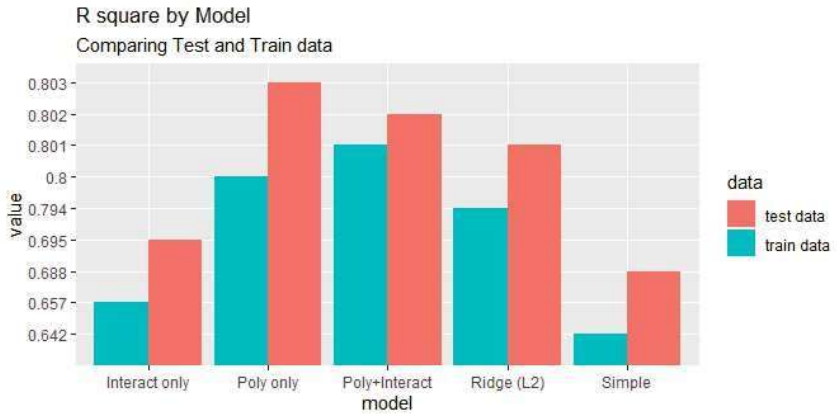
A deep dive into the residuals indicated the need for a log transformation on the predicted variable for better model fit.

Post-transformation, linear regression assumptions are better met.

Going forward, "RENTED_BIKE_COUNT" will represent the log-transformed counts.

Takeaway: Model 2 performs the best, particularly after the log transformation. Keep this in mind for subsequent analysis.





Find the best performing model

- Top Performers: Three models—Poly only, Poly+Interact, and Ridge (L2)—have an RMSE under 330 and R-squared above 0.72.
- Our Pick: Going for simplicity and effectiveness, the Polynomial-only model stands out.

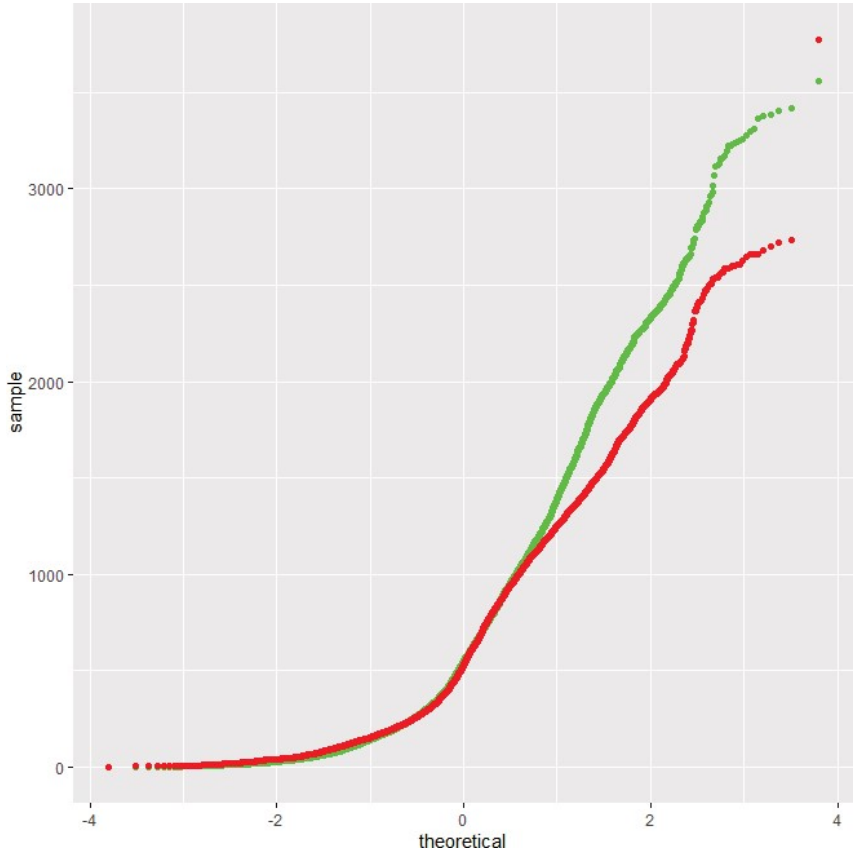
- Model Formula:
- RENTED_BIKE_COUNT ~ . + poly(RAINFALL, 3) + poly(HUMIDITY, 3) + poly(TEMPERATURE, 3) + poly(DEW_POINT_TEMPERATURE, 3)

- Key Takeaway: The Polynomial-only model offers the best balance of accuracy and simplicity.

Q-Q plot of the best model

- Model performance

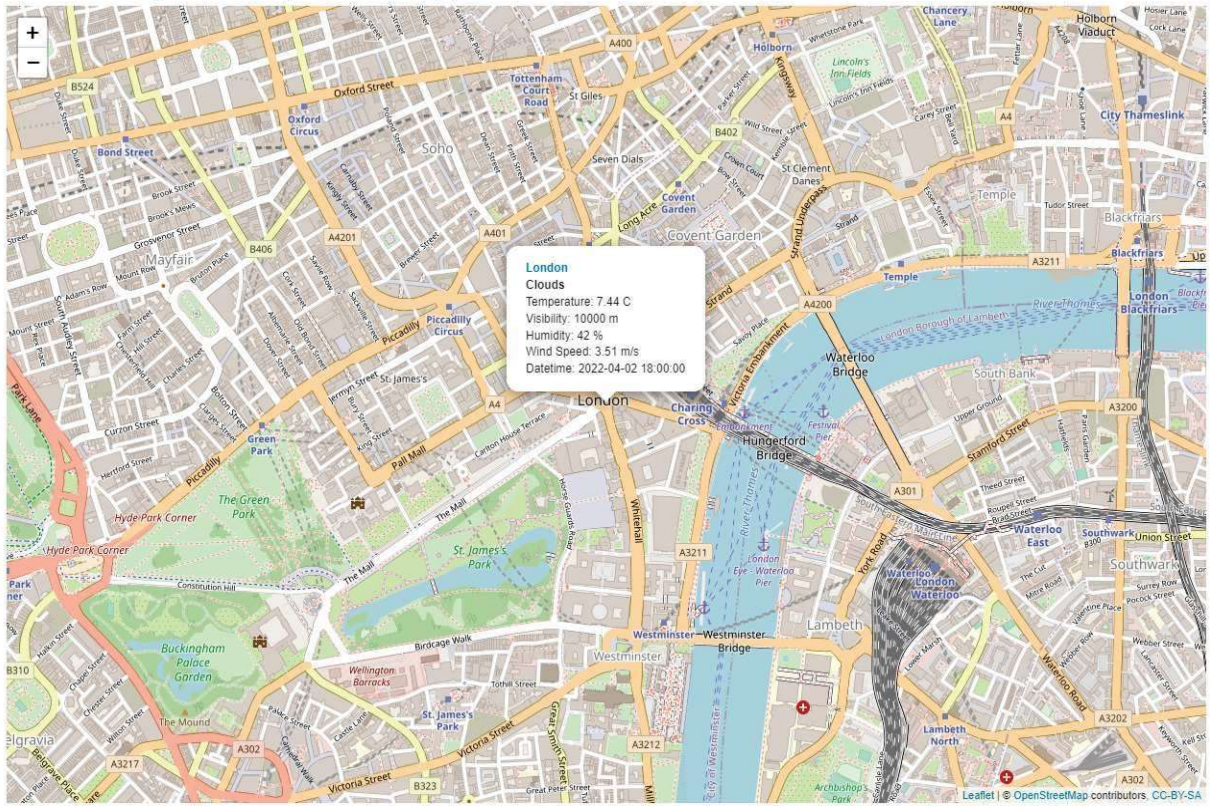
Prediction
Truth

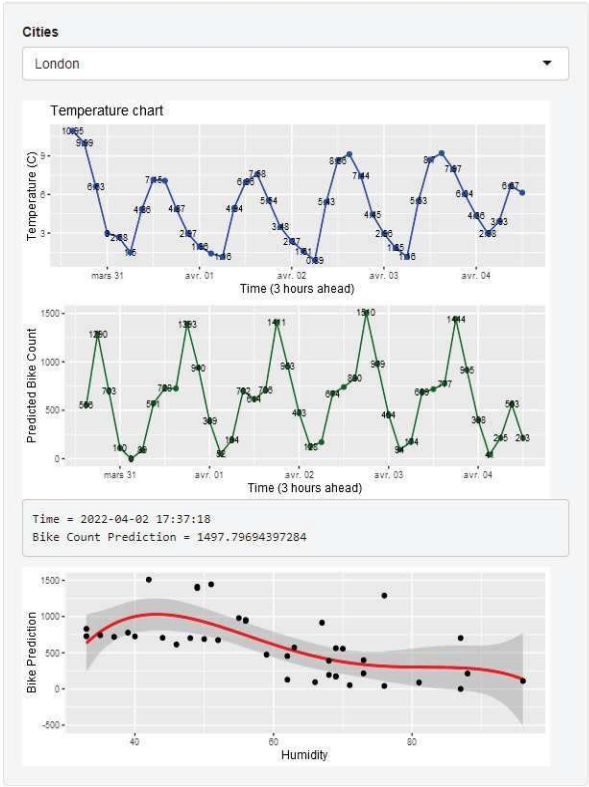


Dashboard

Bike-sharing demand prediction for city : London

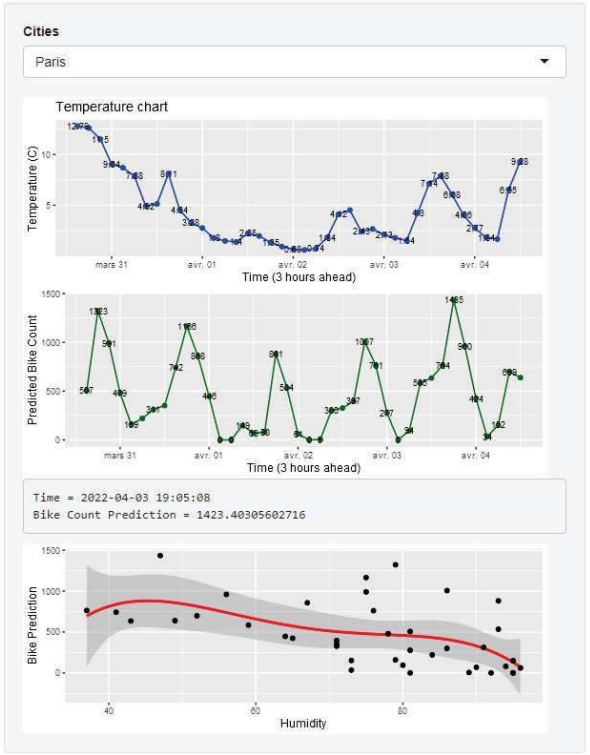
Bike-sharing demand prediction app





Bike-sharing demand prediction for city : Paris

Bike-sharing demand prediction app



CONCLUSION

- Typical Scene: Rentals usually hover around 200-300 bikes.
- Weather & Time: Favorable conditions and peak hours can spike rentals to 3,500.
- Seasonal Trends: Summer and Autumn see the most action, with June and September as peak months. Winter's a slow season.
- Prime Time: 6am and 8pm are the hotspots for rentals.
- Key Factors: Temperature and weather conditions like Humidity/Rainfall heavily influence rentals.
- One Size Doesn't Fit All: Similar cities may have similar patterns, but always check the local data.
- Business Impact: Aligning bike supply with our dashboard predictions could cut costs and better meet customer demands.