

# RideTime: Enhancing Bike Availability Through Statistical Modeling of Usage and Environmental Factors

By Group 5

林昱安, 陳子昊, 謝舜卿, 楊智博

# Introduction

## **Background:**

YouBike plays a key role in sustainable mobility in Taiwan, but suffers from bike availability imbalances, especially during peak hours.

## **Method:**

We built predictive models using historical YouBike data + temperature/humidity + station info.

## **Impact:**

- More accurate forecasting → better bike redistribution → higher user satisfaction.
- Annual CO<sub>2</sub> reduction > 16,000 tons → ≈ NT\$45 million in carbon credits.

Data-driven optimization can simultaneously improve user experience, operational efficiency, and environmental sustainability for public bike systems.

## **Keywords:**

Bike-sharing systems; Predictive modeling; Sustainable urban mobility; Data-driven operations; CO<sub>2</sub> emission reduction

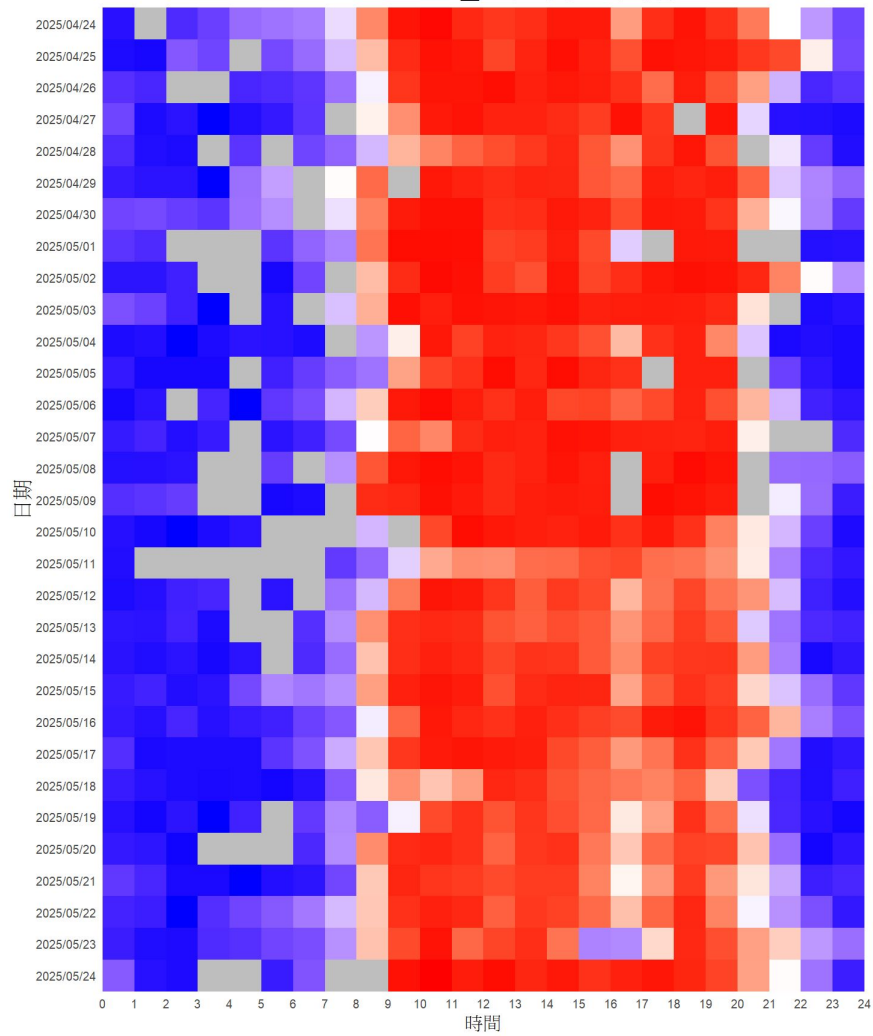
## **Research Question:**

How can the application of data-driven predictive modeling improve bike availability and service reliability in public bike-sharing systems, with specific reference to the case of YouBike in Taiwan?

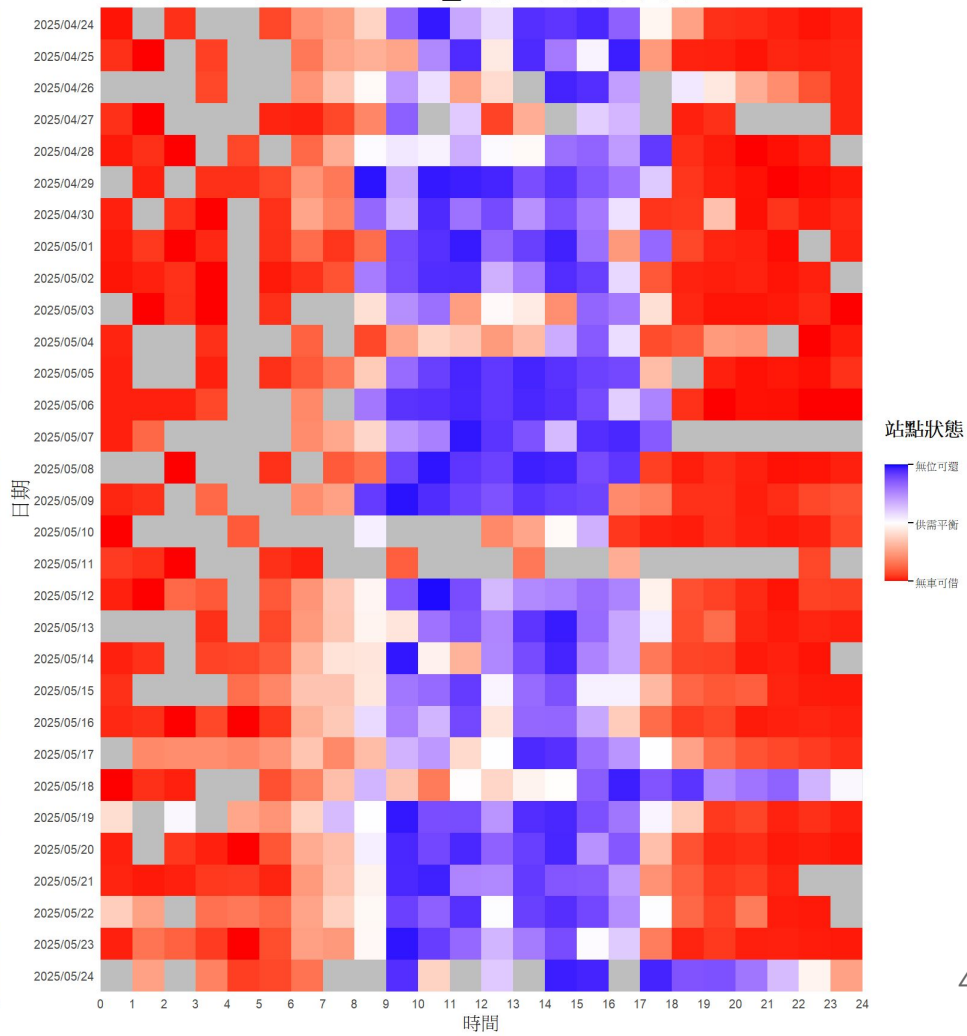
# How can data-driven modeling enhance YouBike availability?



# YouBike2.0\_臺大男一舍前



# YouBike2.0\_臺大小福樓東側



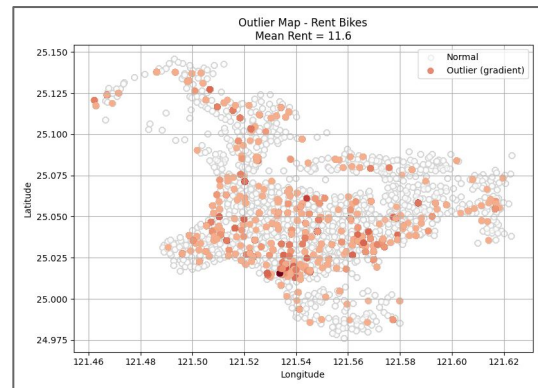
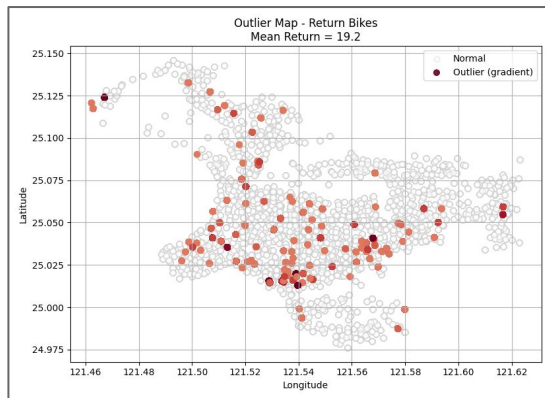
# Why ?

## Motivation:

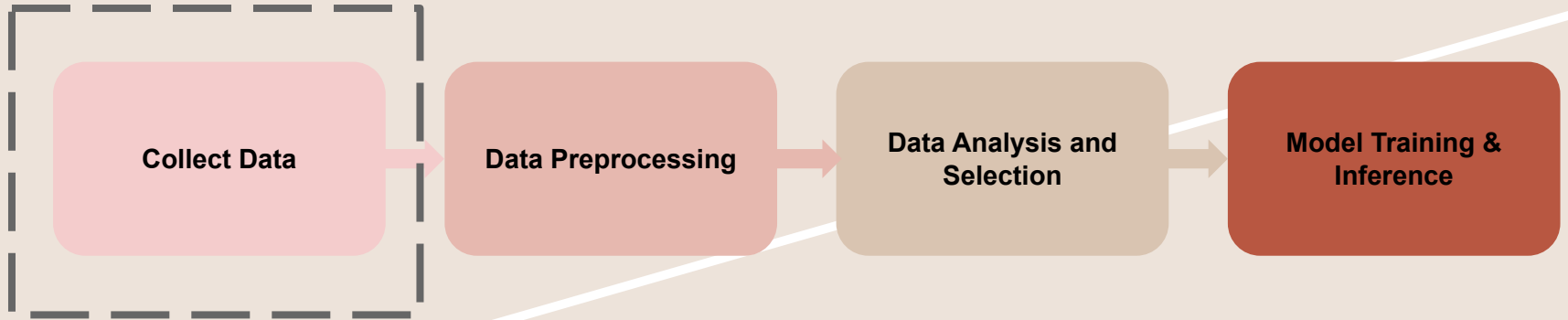
- Problem: Uneven rentable bike and returnable space distribution
- Sustainable Value: YouBike reportedly offsets over 16,000 tons of CO<sub>2</sub> annually. With Taiwan's carbon credit valued at NT\$300 per ton, this represents an environmental benefit worth up to NT\$45 million per year

( Lai, W.-A. (2024) 未來騎YouBike 既環保又可以賺錢 Taiwan Carbon Sustainability and Innovation Foundation.

[https://www.tcsif.org/news\\_detail/TCSIF-NEW11](https://www.tcsif.org/news_detail/TCSIF-NEW11) )



# Data Collection



# Data Sources

## Data Sources:

- Taipei City Open Data:
  - Real-time YouBike updates
  - Station data
- Transport Data eXchange (TDX) :
  - Historical YouBike availability (Minute-level)
  - Station data
- OpenWeather:
  - Temperature (hourly)
  - Humidity (hourly)



# Data Sources

## Data Sources:

- Taipei City Open Data:
  - Station data
- Transport Data eXchange (TDX) :
  - Historical YouBike availability (Minute-level)
  - Station data
- OpenWeather:
  - Temperature (hourly)
  - Humidity (hourly)





# Integration: Multiple data sources

- Packages:

- `dplyr`, `tidyr` : for data manipulation
- `lubridate` : for handling date-time
- `geosphere` : for distance calculations (Haversine)
- `ggplot2`, `scales`, `ggpubr` : for plotting and visualization

- Fetch Data (by Python)

- Data spans a 30-day period from April to May
- Filter out invalid temperature/humidity values (-99)
- Convert ObsTime to POSIXct

3 Source Data

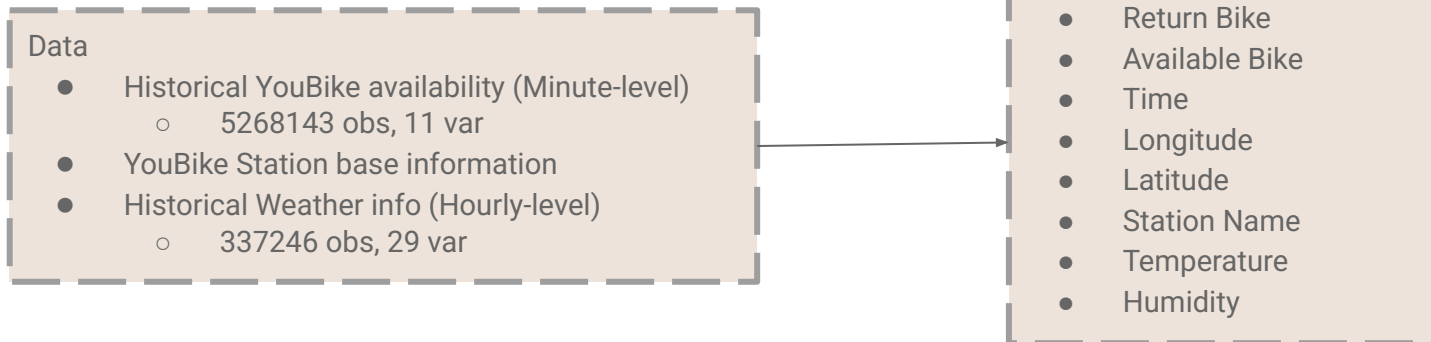
- Historical YouBike availability (Minute-level)
  - 5268143 obs, 11 var
- YouBike Station base information
- Historical Weather info (Hourly-level)
  - 337246 obs, 29 var

- Merge Data:

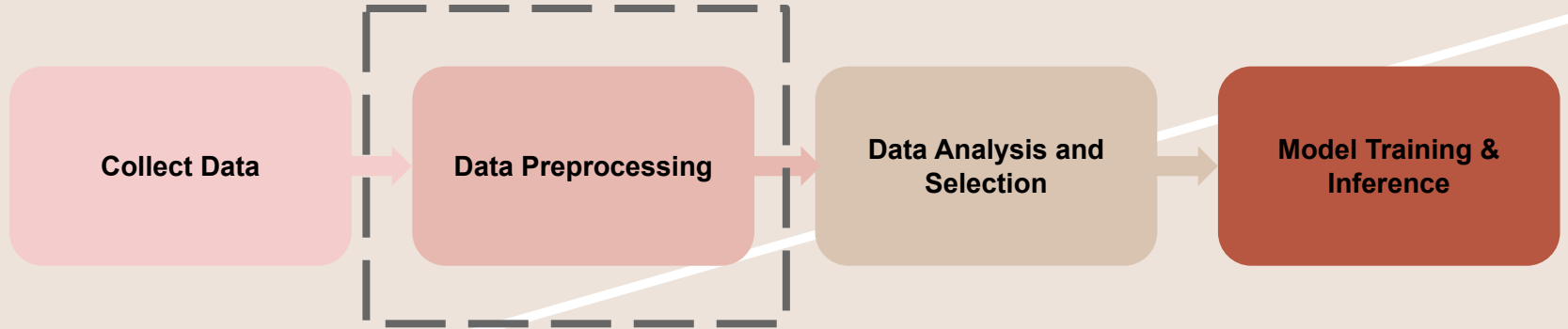
- Historical YouBike availability & YouBike Station base information >> Youbike info
- Youbike info & Weather info.

# Integration: Multiple data sources

- Merge Data (Youbike info & Weather info.):
  - Match the nearest weather observation station: Using the Haversine formula
  - Time Alignment: to hour



# Data Preprocessing



# Data Preprocessing

- Data Cleaning
  - Missing Data: Removes rows with any missing values after merge
    - 5268143 obs → 4327776 obs
  - Outlier
- Data Transformation & Feature Extraction
  - hour (0–23)
  - weekday (Monday to Sunday, Categorical Encoding: 1~7)
  - date

# Data Preprocessing

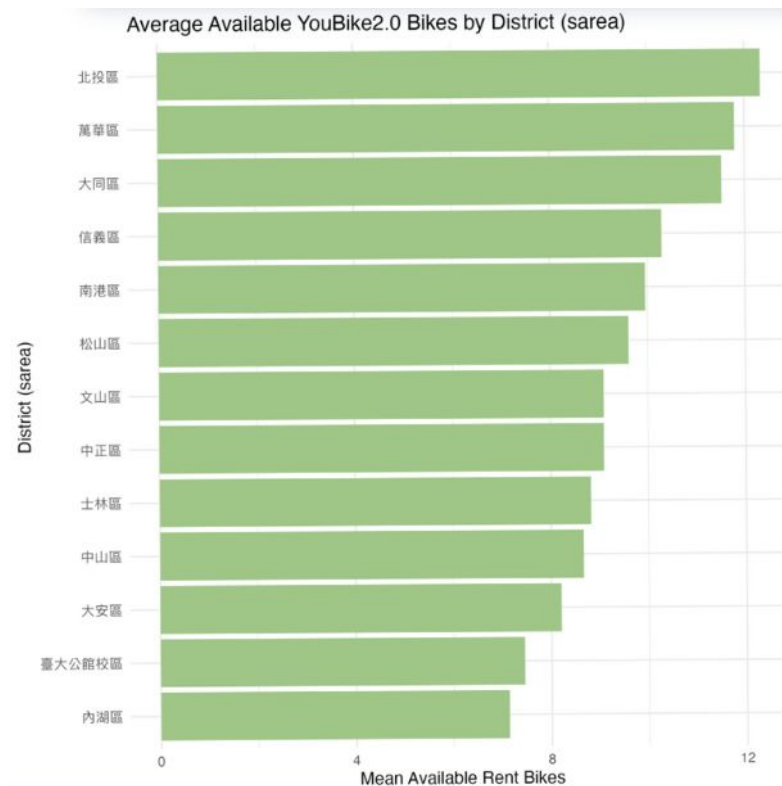
- Time → Cyclic Encoding
  - Representing Hour and Weekday as Sin and Cos values
    - hour (0–23)
    - weekday (Monday to Sunday, Categorical Encoding: 1~7)

hour	weekday	weekday_sin	weekday_cos	hour_sin	hour_cos
9	3	0.433883739117558	-0.900968867902419	0.707106781186548	-0.707106781186548
11	3	0.433883739117558	-0.900968867902419	0.258819045102521	-0.965925826289068
13	3	0.433883739117558	-0.900968867902419	-0.258819045102521	-0.965925826289068
9	3	0.433883739117558	-0.900968867902419	0.707106781186548	-0.707106781186548
9	3	0.433883739117558	-0.900968867902419	0.707106781186548	-0.707106781186548
12	3	0.433883739117558	-0.900968867902419	1.22464679914735E-16	-1
11	3	0.433883739117558	-0.900968867902419	0.258819045102521	-0.965925826289068
5	2	0.974927912181824	-0.222520933956314	0.965925826289068	0.258819045102521
14	3	0.433883739117558	-0.900968867902419	-0.5	-0.866025403784439
14	3	0.433883739117558	-0.900968867902419	-0.5	-0.866025403784439
11	3	0.433883739117558	-0.900968867902419	0.258819045102521	-0.965925826289068
13	3	0.433883739117558	-0.900968867902419	-0.258819045102521	-0.965925826289068
18	3	0.433883739117558	-0.900968867902419	-1	-1.83697019872103E-16
18	3	0.433883739117558	-0.900968867902419	-1	-1.83697019872103E-16
20	3	0.433883739117558	-0.900968867902419	-0.866025403784439	0.5
12	3	0.433883739117558	-0.900968867902419	1.22464679914735E-16	-1
13	3	0.433883739117558	-0.900968867902419	-0.258819045102521	-0.965925826289068
20	3	0.433883739117558	-0.900968867902419	-0.866025403784439	0.5
18	3	0.433883739117558	-0.900968867902419	-1	-1.83697019872103E-16
11	3	0.433883739117558	-0.900968867902419	0.258819045102521	-0.965925826289068
9	3	0.433883739117558	-0.900968867902419	0.707106781186548	-0.707106781186548
22	3	0.433883739117558	-0.900968867902419	-0.5	0.866025403784438
20	3	0.433883739117558	-0.900968867902419	-0.866025403784439	0.5

# Data Preprocessing

Bike availability varies by district:

- Reflecting differences in station density and land use.



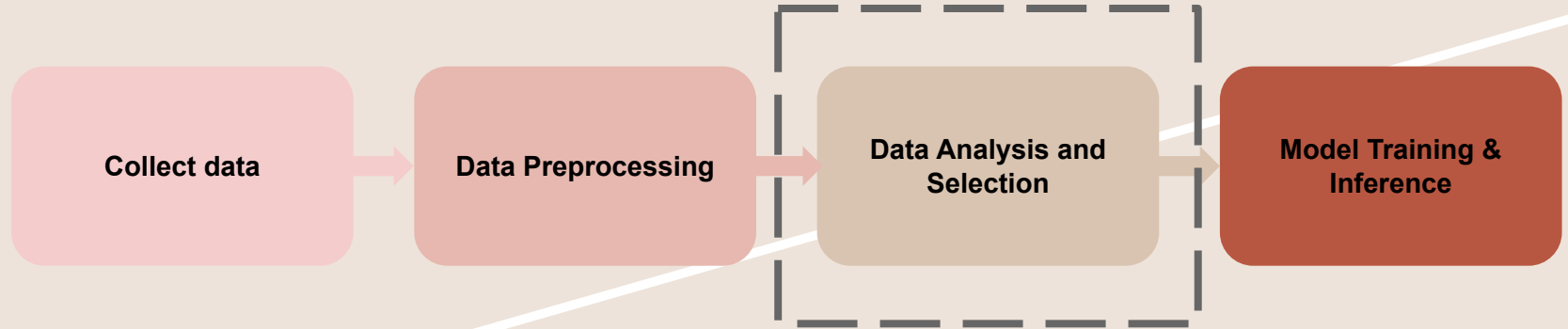
# Data Preprocessing

- District & Location Names

- Marking District and Location names based on one-hot encoding

[illegible]

# Data Selection

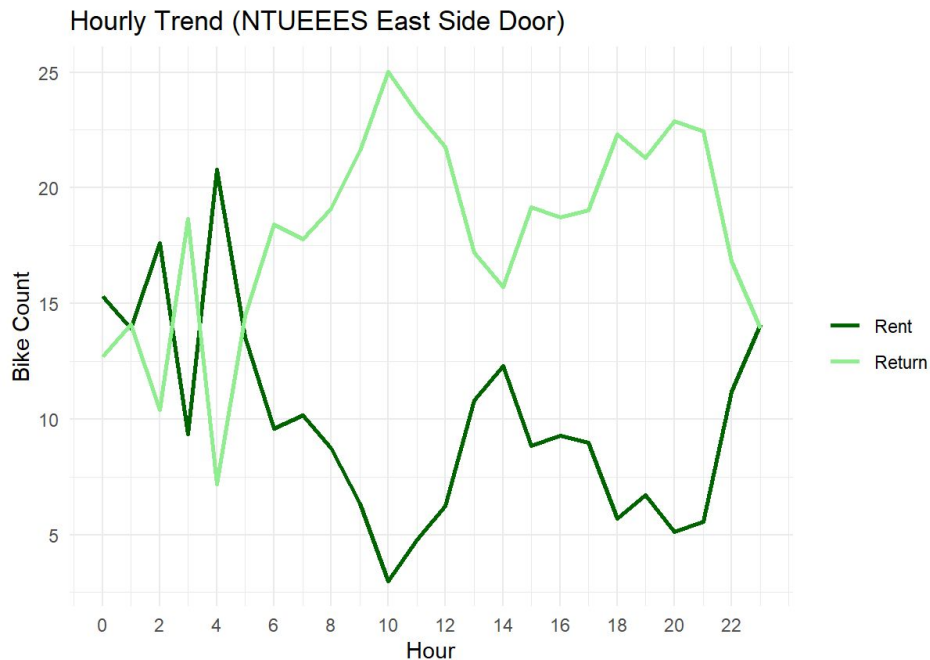
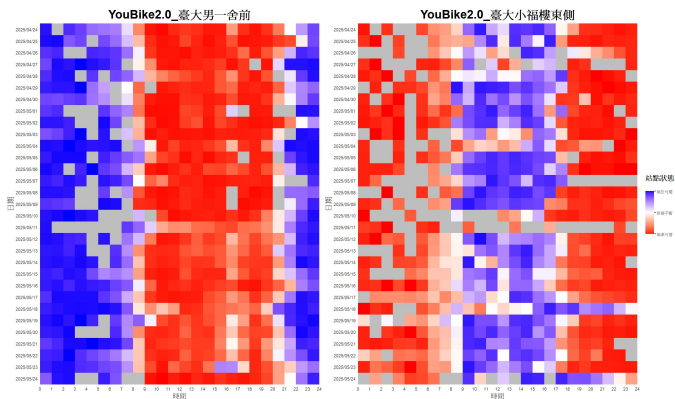




# Hourly Trend

Varying usage patterns is visible between weekdays and weekends,

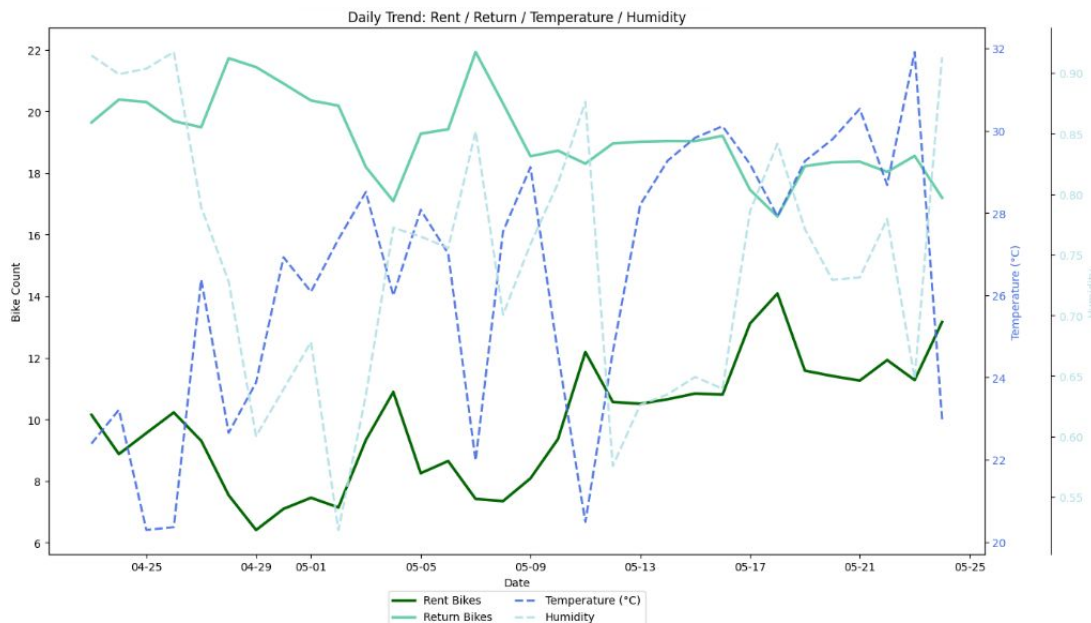
- Highlighting significant hourly fluctuations



# Daily Trend

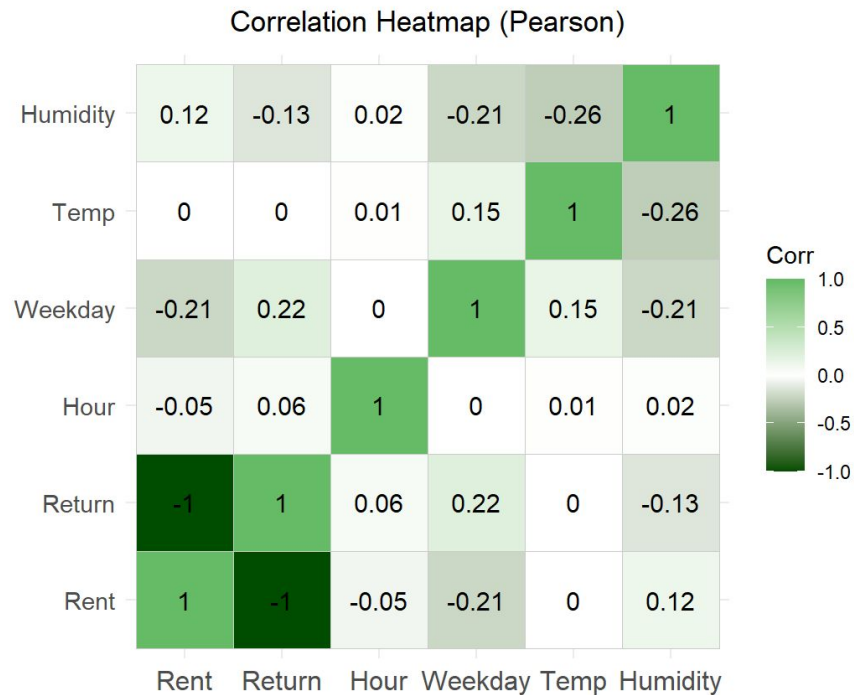
From late April to late May, rising temperature and humidity is correlated w/:

- Increased rentable bikes
- Decreased returnable ones.



# Correlation

Correlation coefficients between individual features and availability are relatively low



# Various variables vs Bike Availability

Return Bike Availability				
Source	Df	Mean Square	F	p
Hour	23	86,283	402.00	< .001 ***
Weekday	6	212,307	989.17	< .001 ***
Date	31	90,734	422.74	< .001 ***
Area	12	1,650,908	7691.82	< .001 ***
Temp	1	120,196	560.01	< .001 ***
Humidity	1	954	4.45	.035 *
Residual	4,327,701	215	—	—

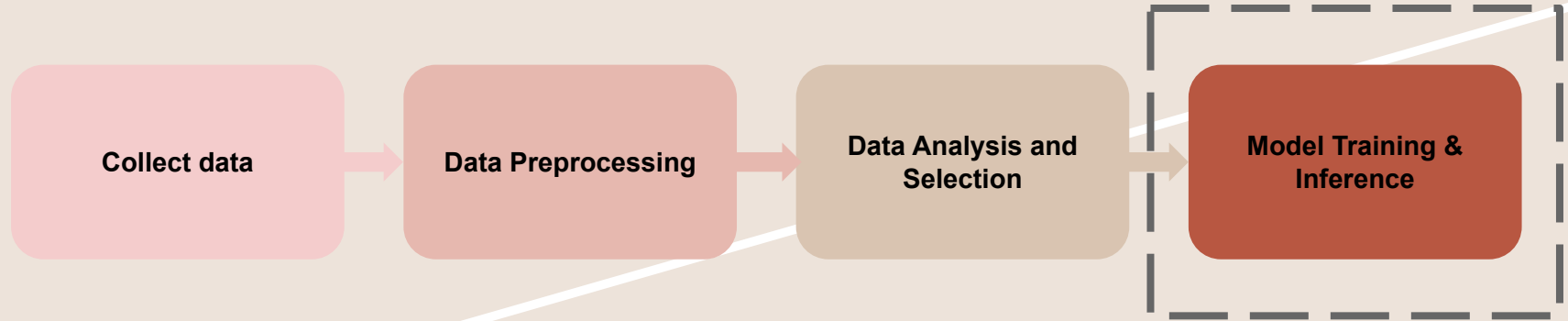
Rent Bike Availability				
Source	Df	Mean Square	F	p
Hour	23	122,965	1066.00	< .001 ***
Weekday	6	385,802	3343.00	< .001 ***
Date	31	252,642	2189.00	< .001 ***
Area	12	515,582	4468.00	< .001 ***
Temp	1	37,500	325.00	< .001 ***
Humidity	1	174,547	1513.00	< .001 ***
Residual	4,327,701	115	—	—

# Feature Selection

Based on these findings, we selected these features to build our predictive model:

- Temporal (hour, weekday)
- Spatial (longitude, latitude, district, location type)
- Environmental (temperature, humidity)

# Model & Prediction

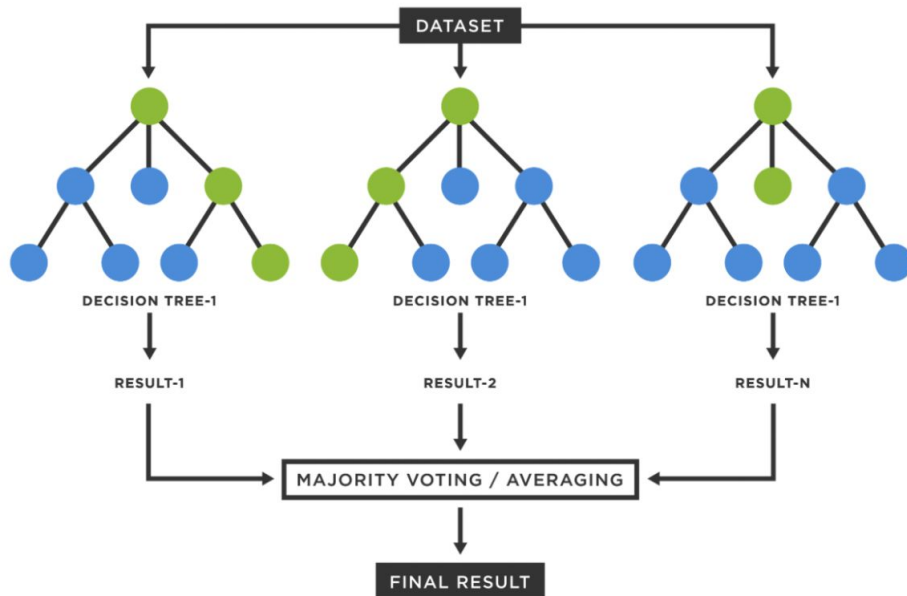


# Model Architecture

Four models are used:

- Random Forest
- CatBoost
- XGBoost
- LightGBM

These models consists of multiple trees w/  
combined outputs.



# Experimental Setup

There are 3 objectives for prediction:

- # of Rent Bike
- # of Return Bike
- Total # of Bikes



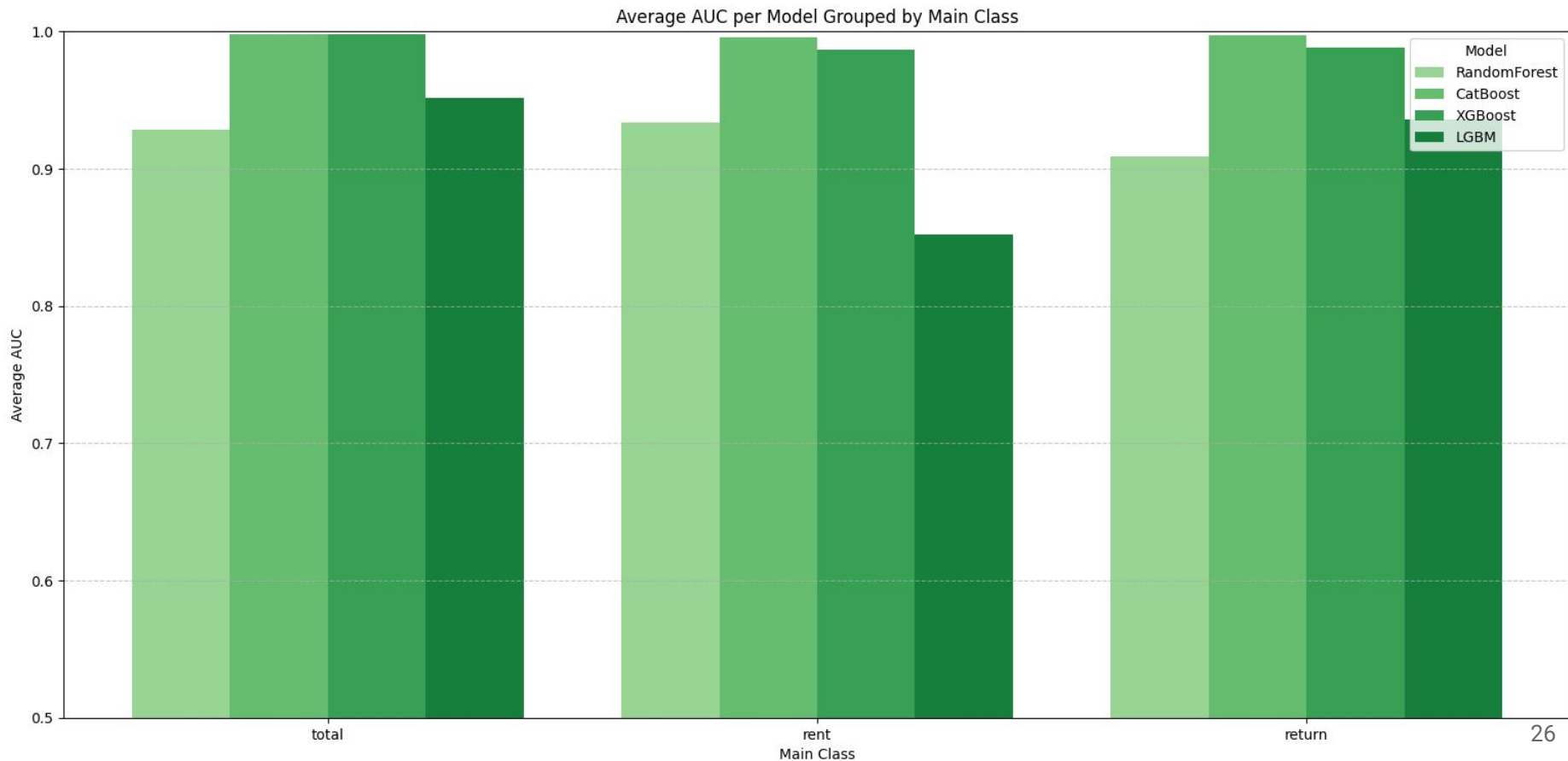
# Experimental Setup

To assess performance, Multi-class classification is conducted:

- Input:
  - Spatial: Longitude, Latitude,
  - Temporal: Hour, Day, Day of week,
  - Climatological: Temperature, Humidity
  - Textual: Location Names
- Goal: Output the correct unit interval
  - Each interval spans 5 bike units

Results are reported in AUC score.

# Model Comparison



# Ablation Study

AUC (Area Under the Curve) scores for each model across four feature exclusion scenarios.

	All Data	w/o Weather	w/o Location Names
RandomForest	0.9237	0.9262	<b>0.9286</b>
CatBoost	<b>0.9970</b>	0.9948	0.9943
XGBoost	<b>0.9911</b>	0.9902	0.9896
LGBM	<b>0.9134</b>	0.9132	0.9101
Average	<b>0.9563</b>	0.9561	0.9556

Consistently high performance across all class indicates:

- spatial and temporal features alone can provide sufficient robustness for this task.

# Low Data Prediction Results

Trial	Total 0 - 30	Total 30 - 60	Total 60+	Rent 0 - 30	Rent 30 - 60	Rent 60+	Return 0 - 30	Return 30 - 60	Return 60+
Baseline	0.6662	0.6101	0.6233	0.5681	0.5127	0.4921	0.6450	0.6506	0.3615
Baseline + & Location	1	1	1	0.8226	0.8106	0.7647	0.8918	0.8660	0.9182

In condition w/ low data, adding location name is very helpful (Baseline model is Random Forest)

# Conclusion



# Conclusion

This study demonstrates the effectiveness of tree-based ensemble models in predicting YouBike availability using temporal, spatial, and environmental features.

By identifying key usage patterns and leveraging high-frequency data, the application of tree-based models can enhance operational forecasting and support smarter bike redistribution.

These improvements not only boost user satisfaction but also contribute to significant environmental and economic gains.

# References

- Lai, W.-A. (2024) 未來騎YouBike 既環保又可以賺錢 Taiwan Carbon Sustainability and Innovation Foundation.  
[https://www.tcsif.org/news\\_detail/TCSIF-NEW11](https://www.tcsif.org/news_detail/TCSIF-NEW11)

# How can data-driven modeling enhance YouBike availability?

**Objective: To visualize the average number of available YouBike2.0 bikes in each Taipei district (sarea).**

## 1. Data Integration: Multiple data sources were combined:

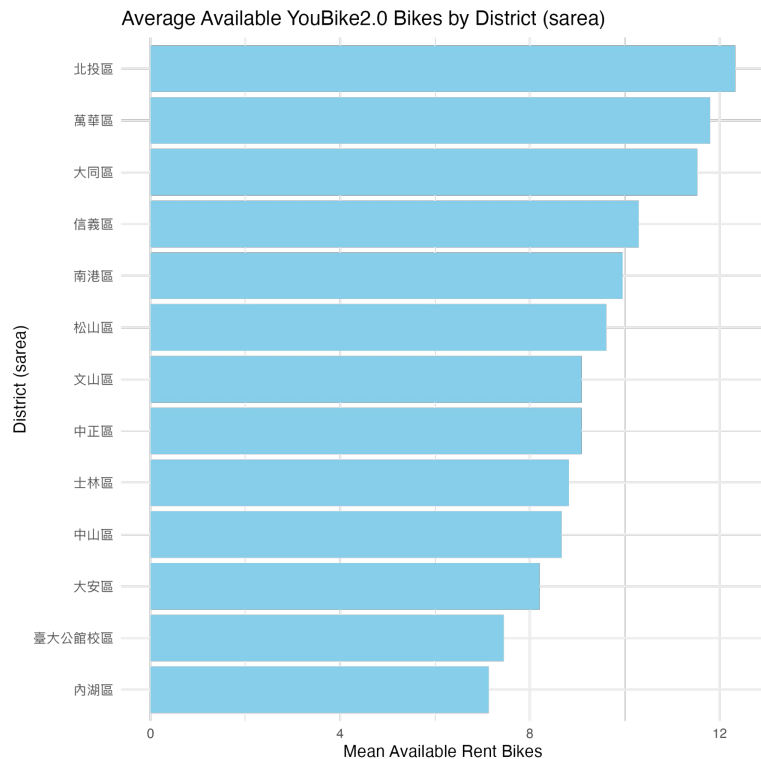
1. YouBike Station Information (names, locations, capacity)
2. Hourly Bike Availability Data (bikes to rent, spaces to return)
3. Weather Data (temperature, humidity, etc.)

## 2. Data Aggregation (**dplyr** library):

The combined data was grouped by the sarea (district) column, **calculated the mean of available bikes for each district.**

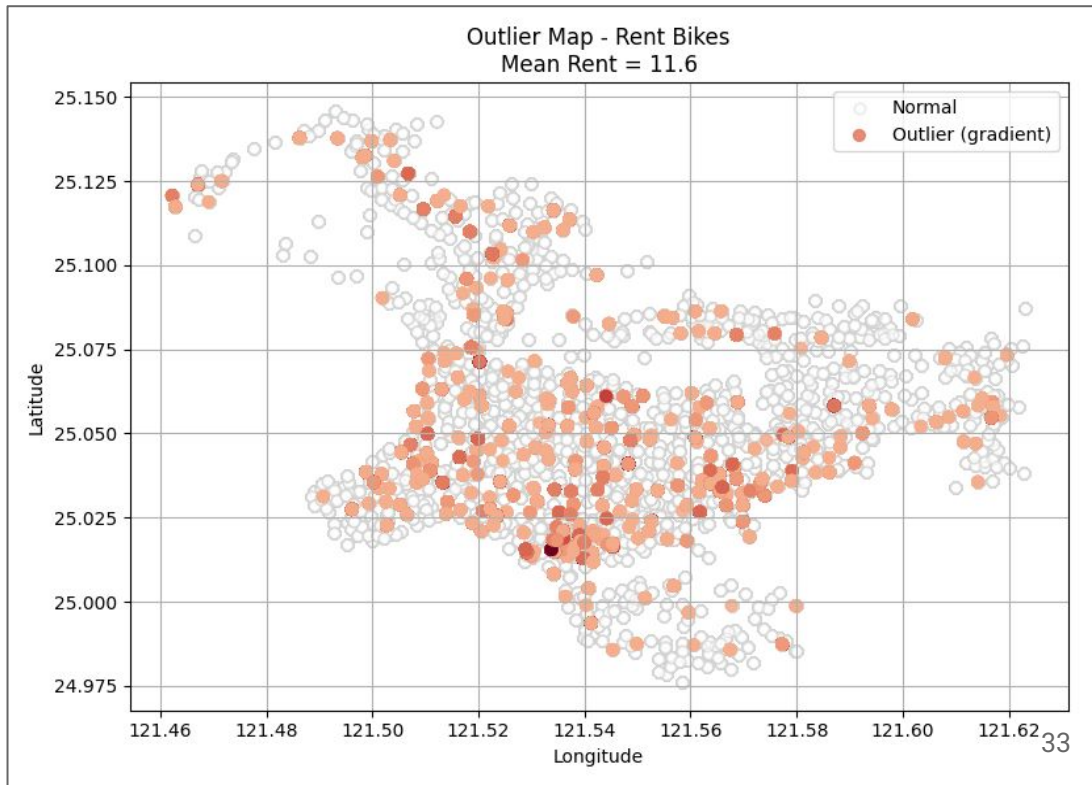
## 3. Visualization (**ggplot2** library):

A horizontal bar chart was generated from the aggregated data.

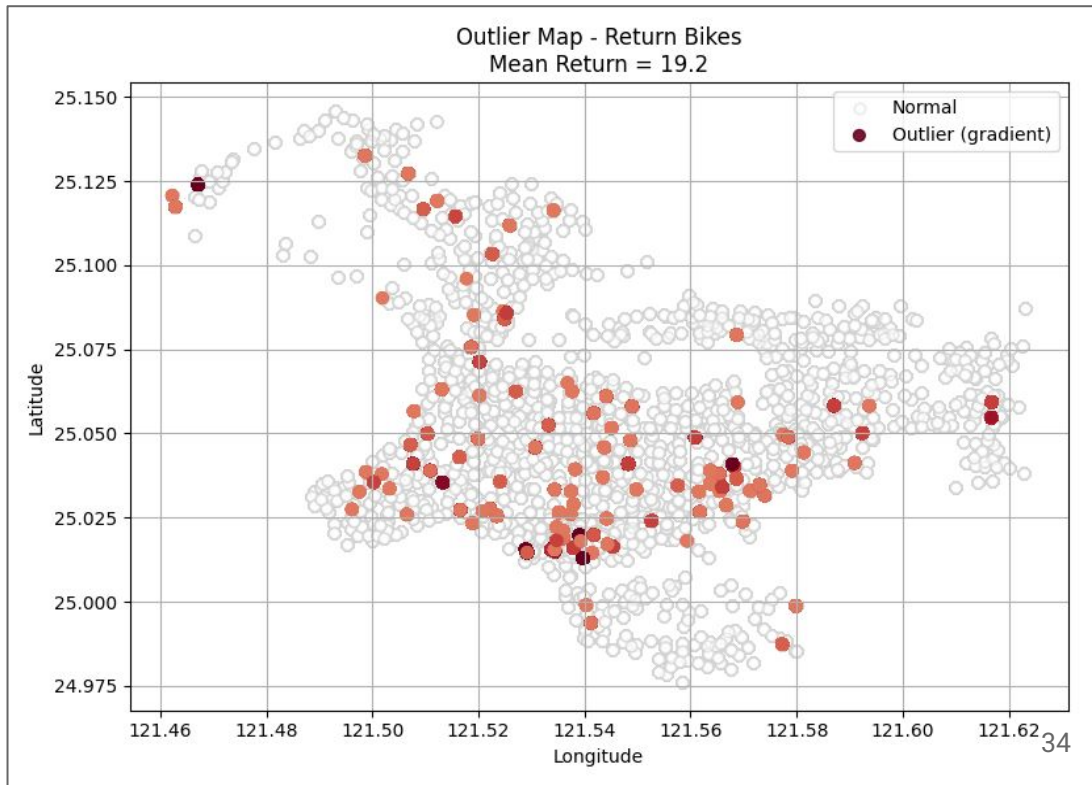





# How can data-driven modeling enhance YouBike availability?



# How can data-driven modeling enhance YouBike availability?



# Dataset

- Packages:
  - `dplyr`, `tidyr` : for data manipulation
  - `lubridate` : for handling date-time
  - `geosphere` : for distance calculations (Haversine)
  - `ggplot2`, `scales`, `ggpubr` : for plotting and visualization
- Fetch Data (by Python) 
  - Filter out invalid temperature/humidity values (-99)
  - Convert `ObsTime` to `POSIXct`
- Merge Data:
  - Historical YouBike availability & YouBike Station base information >> Youbike info
  - Youbike info & Weather info.

## Data

- Historical YouBike availability (Minute-level)
  - Rent Bike, Return Bike, Station ID
- YouBike Station base information
  - Longitude, Latitude, Area, , Station ID
- Historical Weather info (Hourly-level)
  - temperature, humidity

# Notes

These models belong to tree ensemble methods, often called forests. They construct multiple decision trees and combine outputs to make accurate and robust predictions than any single tree could achieve alone

The bar chart compares the prediction accuracy of all models across three target variables (number of returned bikes, rented bikes, and total bikes), with CatBoost emerging as the top-performing model.