# Optimizing Public Bike Rental Allocation using Demand Forecasting

Tae Han

## Problem Statement:

Public bike rental systems have gained popularity worldwide as a sustainable transportation option. However, ensuring the efficient distribution of bikes across rental stations remains a significant challenge. An imbalance in bike availability can negatively impact operational sustainability and the overall user experience.

Due to various factors such as time, location, and weather conditions, bike rental demand fluctuates, leading to shortages at some stations and surpluses at others. As a solution, the New York City public bike system introduced the 'Bike Angels' program, which rewards riders for moving bikes from crowded stations to those with shortages, helping to rebalance the rental system's availability.

This project aims to replicate a similar Bike Angels program for Seoul City and improve its allocation efficiency and sustainability. The objective is to quantify demand using historical rental data and identify bike rides between stations to reduce the probability of shortages in the overall rental system.

## Data Source:

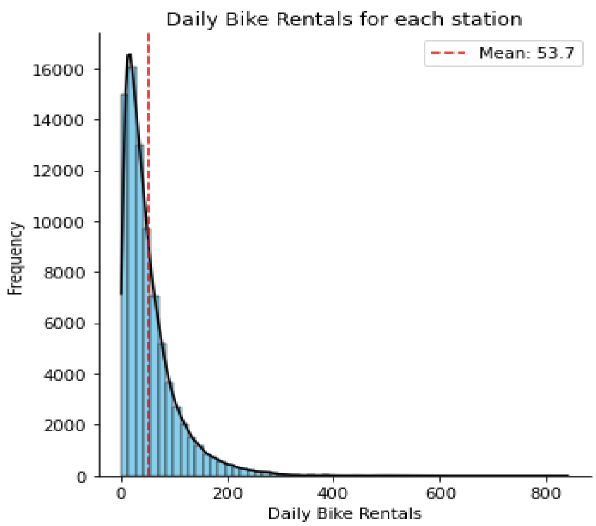
The Seoul City public bike rental system shares its rental records and station information publicly online. This project uses the monthly bike rental dataset from June 2024, which contains 5,004,330 rental records. Each record includes details such as rental and return stations, rental and return timestamps, user information including birth year and sex, distance traveled in meters, duration in minutes, and bike ID. The station information dataset includes 3,340 entries with columns for station names, IDs, addresses, and geographic coordinates in latitude and longitude.

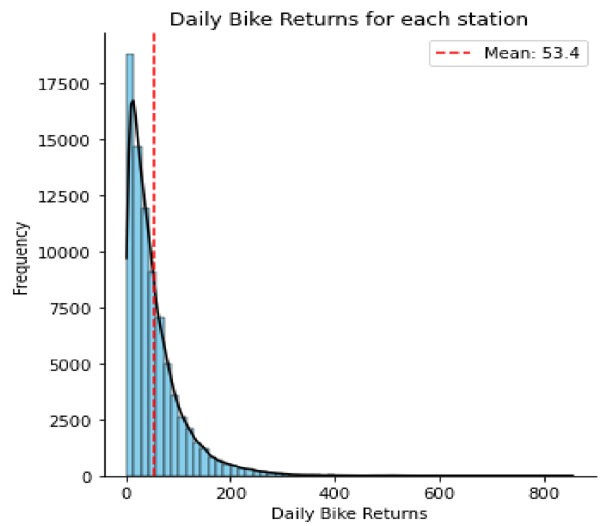
Records with missing return information, zero minutes or zero meters traveled, and long-distance rides exceeding 60 minutes were defined as outliers and removed. After cleaning, 4,359,833 rows remain for analysis, representing 87.1% of the original dataset. These records involve 39,153 unique bikes across 2,740 unique stations.

**Exploratory Data Analysis:**

This section highlights key patterns in rental and return activity, temporal trends, and geographical patterns, which help inform later modeling and optimization processes.

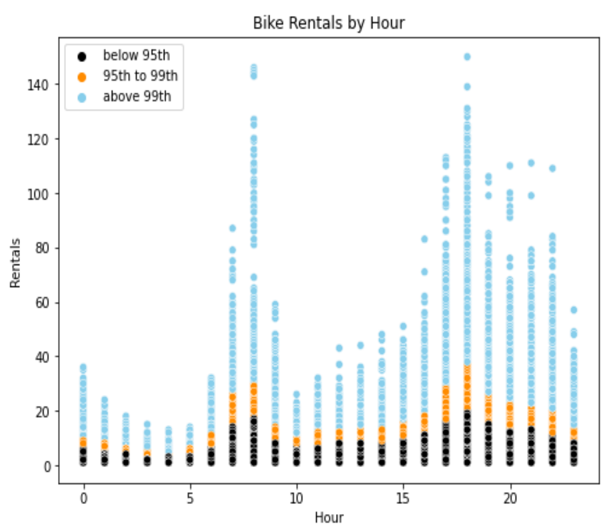
- **Distributions:**

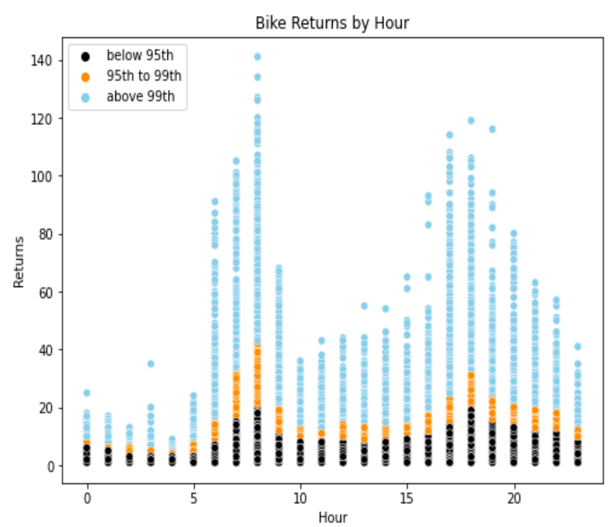



- The station-level daily averages of bike rentals and returns both have right-skewed distributions, indicating an imbalance in rental activity across stations.

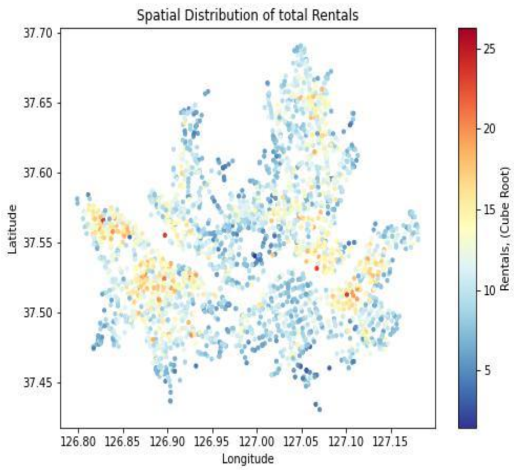
- **Temporal Trends:**

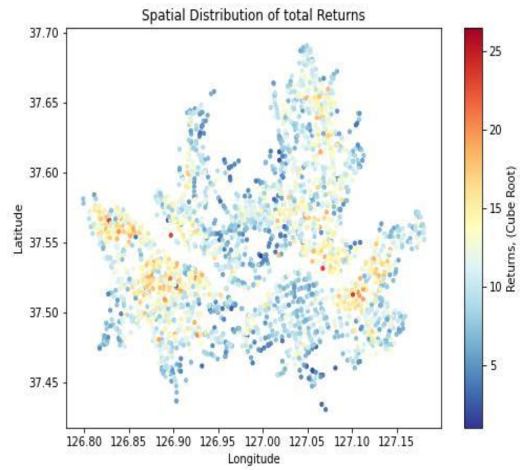



- Hourly rental and return patterns show strong peaks during morning and evening commute hours.

- **Geographical Patterns:**




- High bike-sharing activity is observed near the Han River, which runs through central Seoul and is surrounded by numerous parks.
- Low bike-sharing activity is observed in the central part of the city, where urban areas are surrounded by mountains.

## Methodology:

This project has two main quantitative objectives: (1) to estimate future supply (bike returns) and demand (bike rentals) at each station, and (2) to measure changes in shortage probabilities for all possible bike ride routes from one station to another, based on these estimates.

- **Demand forecasting:**

The hourly counts of bike rentals and returns are used as measures of demand and supply, respectively, and serve as the target response variables for the forecasting models. To capture underlying patterns in the data, various predictor variables are used, including the geographic coordinates of stations, temporal factors (weekday and hours), and recent rental or return activity (e.g. rolling averages over previous days).

To model the nonlinear relationships between these predictors and the response variables, we employ a Generalized Additive Model (GAM) with a Poisson distribution, which is well-suited for modeling count data.

The Poisson GAM is structured as follows.

$$log(\lambda) = \beta_0 + f_1(latitude, longitude) + f_2(weekday, hour) + f_3(past\ rentals/returns)$$

Where:

1) $\lambda$: expected number of bike rentals or returns at a given station for a specific hour.

2) $f_1(latitude, longitude)$: a smooth spatial interaction term capturing location-related effects.
3) $f_2(weekday, hour)$: a smooth temporal interaction term capturing weekly and daily usages.
4) $f_3(past\ rentals/returns)$: a smooth function of recent station-level rental or return activity.

Two separate models are trained for rentals and returns. The first three weeks of data are used to train the models and to tune their smoothing parameters using 5-fold cross-validation. The final week of data is used for evaluation. This approach allows the models to capture complex underlying patterns in the bike-sharing system while maintaining enough generalizability to perform well on future datasets.

- **<u>Calculating Shortage Probability:</u>**

The next step is to simulate the hourly inventory of bikes at each station across an entire day based on the hourly forecasts of rentals and returns, representing the expected inflow and outflow at each station for every hour. Using these simulated flows, calculate the remaining bike inventory and estimate the probability of shortage at each station and hour throughout the day.

The rental and return forecasts at each station and hour are derived from Poisson distributions. To calculate the probability of a shortage (e.g. remaining inventory dropping below zero), model the difference between the rentals and returns using the Skellam distribution, which describes the difference between two independent Poisson variables. Mathematically,

Let:

$$\text{Outflow}_{s,h} \sim \text{Poisson}\left(\lambda_{s,h}^{Rental}\right)$$
$$\text{Inflow}_{s,h} \sim \text{Poisson}\left(\lambda_{s,h}^{Return}\right)$$

Then, the change in bikes at station *s* and hour *h* is:

$$\Delta\text{Bike}_{s,h} \sim \text{Skellam}\left(\lambda_{s,h}^{Return} - \lambda_{s,h}^{Rental}\right)$$

The remaining inventory at station s and hour h is:

$$\text{Bike}_{s,h} = \text{Bike}_{s,h-1} + \Delta\text{Bike}_{s,h}$$

The inflow and outflow of bikes are modeled as count-based variables following Poisson distributions. The difference between these two independent Poisson variables follows a Skellam distribution, which models change in bike inventory at a station for a given hour. Using the Skellam distribution of the difference between inflows and outflows, calculate the probability that a station runs out of available bikes at a given hour.

$$\text{P}\left(\text{Shortage}_{s,h}\right) = \text{P}\left(\text{Bike}_{s,h} < 0\right)$$

Assign the initial number of bikes to each station at 6:00 AM, assuming that bike inventories are reset every morning, by assigning group means of available bikes, split into five groups based on the number of bike rentals. Then, cumulatively calculate the shortage probability at each station for every hour based on the hourly rental and return forecasts and the number of bikes remaining at each hour.

- **Ride Simulation:**

Each bike ride decreases the inventory at the departure station and increases it at the arrival station. These changes in inventory affect the probability of shortage at both stations, as well as the overall distribution of bikes across the rental system. The objective of this section is to calculate the change in overall shortage probabilities resulting from a ride from station A to station B.

For a ride occurring at hour *h*, from station *A* to station *B*:

$$\mathrm{Bike}_{\mathrm{A,h}} \leftarrow \mathrm{Bike}_{\mathrm{A,h}} - 1$$
$$\mathrm{Bike}_{\mathrm{B,h}} \leftarrow \mathrm{Bike}_{\mathrm{B,h}} + 1$$

The change in shortage probability at station S:

$$\Delta \mathrm{P}_{s,h} = \mathrm{P}_{after}\left(\mathrm{Shortage}_{s,h}\right) - \mathrm{P}_{before}\left(\mathrm{Shortage}_{s,h}\right)$$

The overall effect of a ride on the entire system:

$$\Delta \mathrm{P}_{total} = \Delta \mathrm{P}_{A,h} + \Delta \mathrm{P}_{B,h}$$

If the reduction in shortage probability at station B (arrival) is greater than the increase in shortage probability at station A (departure), the ride results in a net decrease in overall shortage probability. Conversely, if the increase in shortage probability at station A exceeds the reduction at station B, the ride contributes to a higher overall shortage probability.

## Evaluation and Final Results:

The objective of evaluation is to ensure that the forecasting models accurately and consistently predict hourly bike rentals and returns across different weekdays, hours, and locations. This allows reliable calculation of shortage probabilities and their changes in each bike ride route.

- **GAM Forecasting Model Performances:**

After tuning the smoothness hyperparameters for the nonlinear terms, the Generalized Additive Models (GAMs) for bike rentals and returns produced the following summaries:

1) Bike Rental Forecasting GAM Model Summary

| Feature | Feature Type | Lambda | P-value | Significance |
|---|---|---|---|---|
| Latitude, Longitude | Tensor | 0.001, 0.001 | 0.00e+00 | *** |
| Weekday | Factor | 0.6 | 0.00e+00 | *** |
| Hour | Smooth | 0.001 | 0.00e+00 | *** |
| Number of Rents (Past Week) | Smooth | 0.6 | 0.00e+00 | *** |

| | | | | |
|---|---|---|---|---|
| Number of Rents (Previous Hour) | Smooth | 0.6 | 0.00e+00 | *** |
| Intercept | - | - | 7.97e-01 | - |

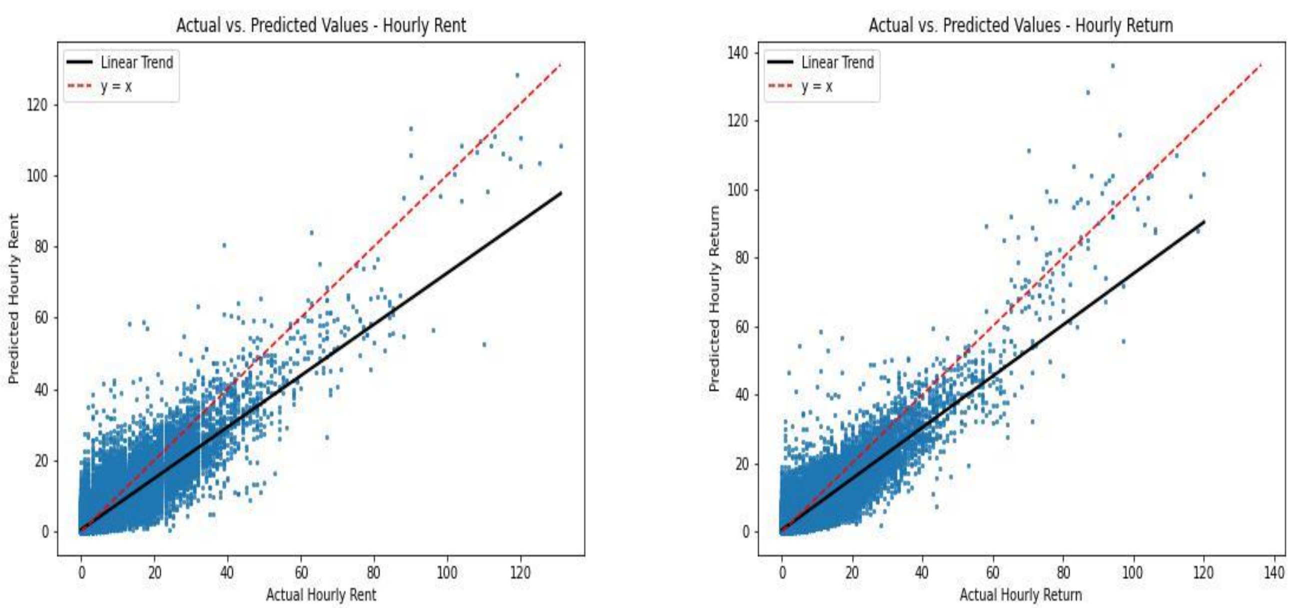
- Pseudo R-squared: 0.663
- Test Mean Squared Error: 4.175

2) Bike Return Forecasting GAM Model Summary

| Feature | Feature Type | Lambda | P-value | Significance |
|---|---|---|---|---|
| Latitude, Longitude | Tensor | 0.001, 0.001 | 0.00e+00 | *** |
| Weekday | Factor | 0.6 | 0.00e+00 | *** |
| Hour | Smooth | 0.001 | 0.00e+00 | *** |
| Number of Returns (Past Week) | Smooth | 0.6 | 0.00e+00 | *** |
| Number of Returns (Previous Hour) | Smooth | 0.6 | 0.00e+00 | *** |
| Intercept | - | - | 9.63e-01 | - |

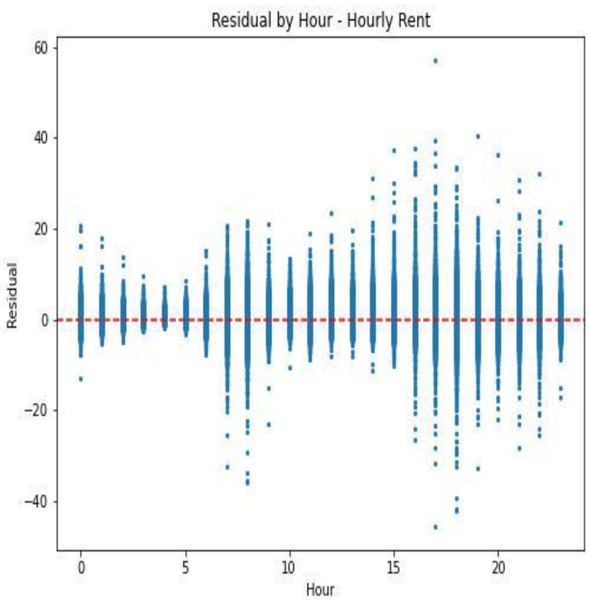
- Pseudo R-squared: 0.701
- Test Mean Squared Error: 3.638

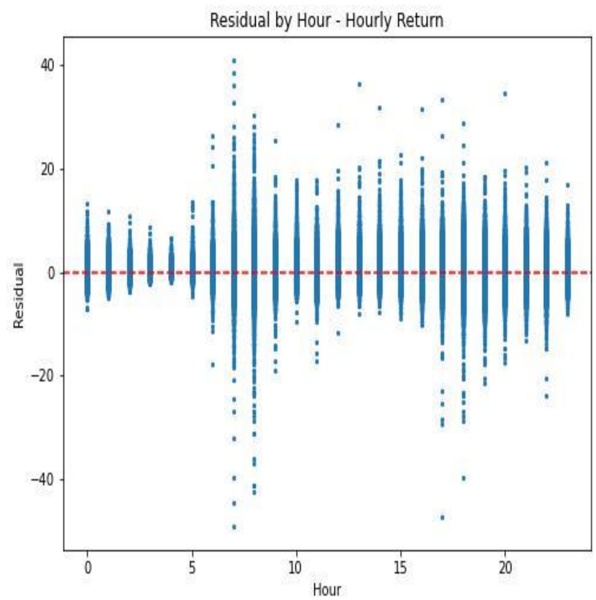
3) Prediction vs. Actual plot




- Actual and predicted hourly rentals and returns are positively correlated. Adding more predictive information into the models could improve their one-to-one alignment.

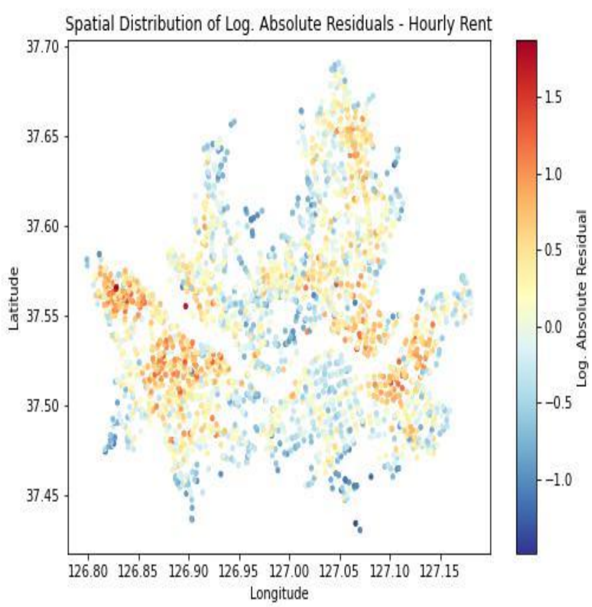
4) Residual by hour

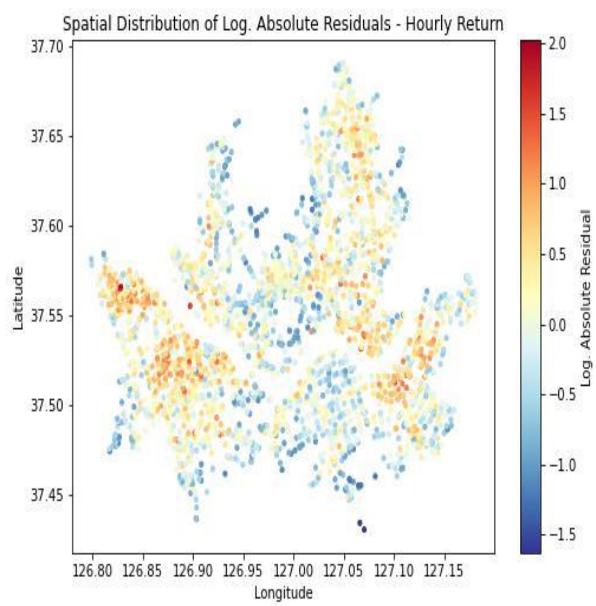



- Larger residuals occur during peak commuting hours around 7,8 AM and 5,6 PM for hourly rentals and returns.

5) Residual by location



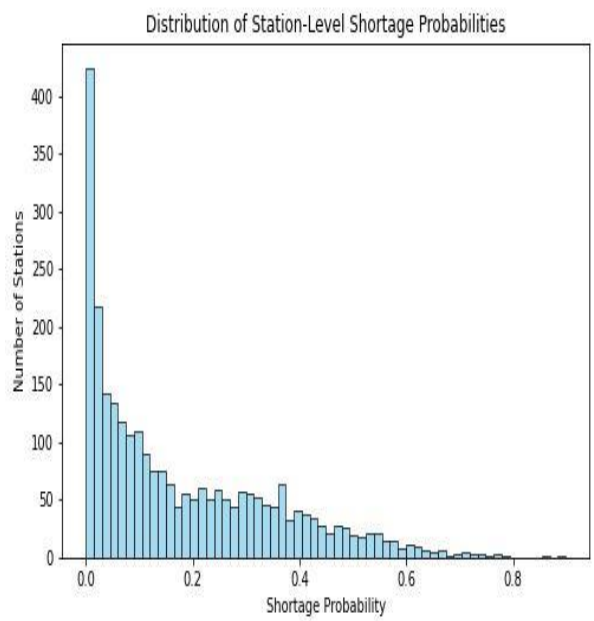




- Larger absolute residuals occur in high-activity areas near the Han River in the city center.

- **Shortage Probability:**

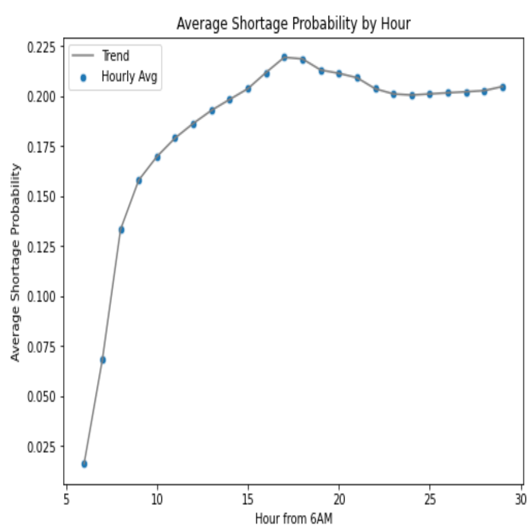
After simulating bike sharing daily operation using the bike rental and return hourly forecasts, evaluate station-level hourly shortage probabilities.

- **Distributions:**



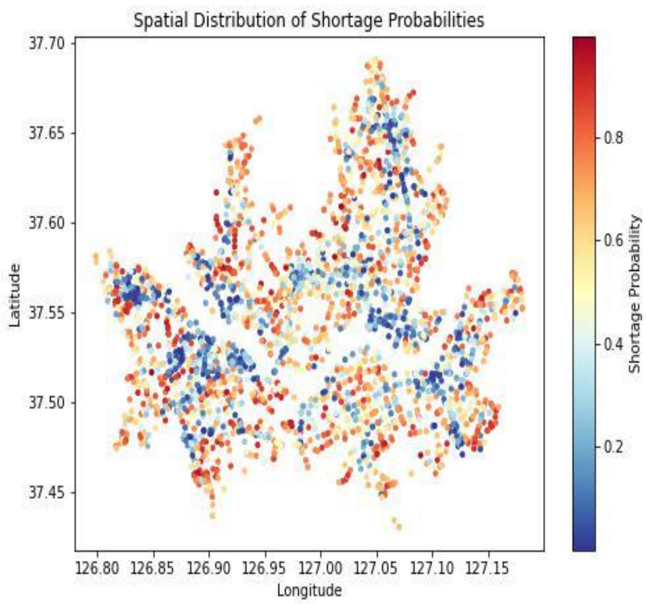

- The average shortage probabilities per station show a right-skewed distribution, indicating supply-demand imbalance across stations.

- **Temporal Trends:**




- As bike-sharing activity increases in the morning, the imbalance between supply and demand worsens throughout the day, with shortage probabilities peaking around 5 PM.

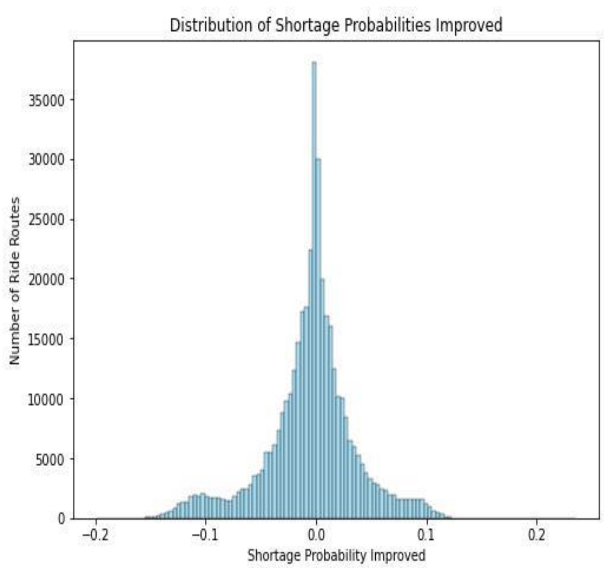
- **Geographical Patterns:**



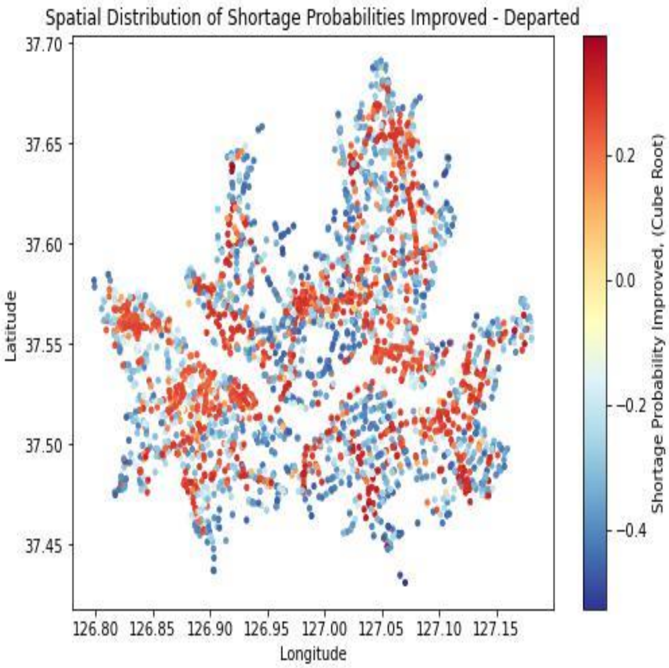

- Higher shortage probabilities occur at stations in outer areas with fewer bike arrivals and in densely populated city centers with high demand.

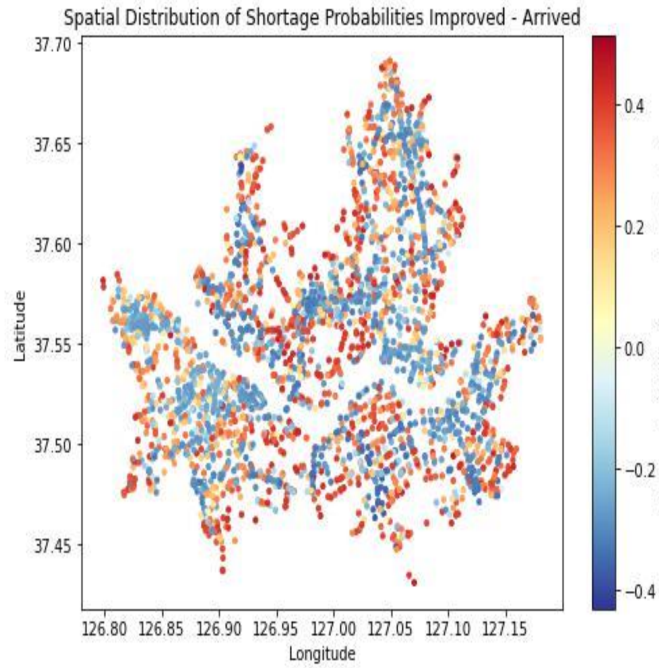
- **Ride Simulation:**

For all possible bike rides from station A to station B, calculate and evaluate the change in shortage probabilities.

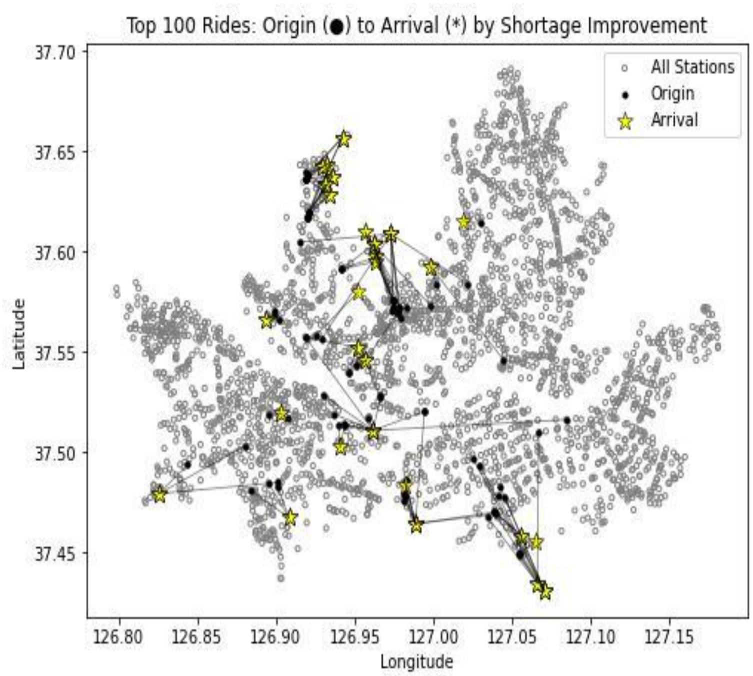



- On average, ride routes remain neutral, though some help improve or worsen station imbalances.




- On the left, bike rides that improved system-wide shortages departed from central, densely populated areas with high rental activity and consistent supply.
- On the right, these rides arrived at outer city areas with low rental activity and limited supply, helping to alleviate shortages.




- The top 100 ride routes that most improve shortage probabilities help redistribute bikes from densely populated inner-city areas to outer parts of the city.

## Conclusion and Implications:

This project examines spatial and temporal imbalances in Seoul's public bike-sharing system using predictive modeling to estimate bike rentals and returns as proxies for supply and demand. We found that shortages are most common in both high-demand central areas and underserved outer regions, especially during peak commuting hours. By forecasting hourly bike rentals and returns and simulating ride-level flows, we quantified the shortage probabilities that stations experience from specific ride routes and identified routes that naturally redistribute bikes across the city.

Future studies can build on these findings by optimizing the initial number of bikes allocated to each station, with the objective of minimizing overall shortage probabilities. Not only the inflows and outflows of bikes, but also the initial daily bike inventory at each station determines the level of shortage a station will experience.

Another direction for future research is to incorporate spatial dependencies between stations. A change in bike inventory at one station often affects neighboring stations due to riders' availability to walk short distances or choose alternate pickup points. Optimizing bike allocations with consideration for nearby station dynamics can lead to more robust system-wide rebalancing strategies.

## References:

Bike Angels. Citi Bike NYC. https://citibikenyc.com/bike-angels

Seoul Bike Ttareungi. Seoul City Public Bike Rental System. https://www.bikeseoul.com/

Seoul Metropolitan Government. (2024). Information on Seoul City Public Bicycle Rental Stations. Seoul Open Data Plaza. https://data.seoul.go.kr/dataList/OA-13252/F/1/datasetView.do

Seoul Metropolitan Government. (2024). Seoul City Public Bicycle Rental History Information. Seoul Open Data Plaza. https://data.seoul.go.kr/dataList/OA-15182/F/1/datasetView.do