Citi Bike Sharing in NYC

Xiangyu Zeng

December 4, 2024

1 Introduction

Urban transportation systems have been rapidly evolving to meet the demands of a growing environmentally conscious population. Bike-sharing programs have emerged as a sustainable solution for urban mobility. Among these programs, Citi Bike in New York City has become an emblematic example. With thousands of bicycles and docking stations strategically positioned throughout the city, Citi Bike has enabled countless New Yorkers and visitors to navigate the city quickly and conveniently. However, managing such a vast and dynamic system efficiently is no easy task.

This case report addresses the challenge of optimizing Citi Bike's operational strategies. It aims to answer key questions about fleet size, bike repositioning, daily outages, and point value recommendations for encouraging customer bike repositioning.

Our primary objectives are the following:

1. Optimal Fleet Size: determine the ideal number of bikes in the fleet at the beginning of the day (5:00 am) for each station.
2. Overnight Bike Movements: calculate the number of bikes that should be moved overnight to meet demand for the next day to maintain system balance.
3. Predicted Outages: estimate the number of outages, or instances when bikes or empty docks are unavailable.
4. Bike Angels Point Values: provide recommendations for setting point values for Bike Angels, a feature that encourages and incentive riders to reposition the bike and thus improving the availability of bikes and docks for fellow riders.

Our primary findings are the following:

1. We optimize bike allocation for 1851 stations in NYC at 5 am, requiring a total of 29,401 bikes out of the 60,217 docks.
2. Overnight, we will need to relocate 3161 bikes for optimal positioning the next morning.
3. Following the previously mentioned ideal placement, the city experiences an average of 438 daily outages, where customers can't find an available bike or an empty dock.
4. We set the Bike Angel points for each stations, using 2 pm as an example. Out of 1851 stations, 1742 stations need no relocation, 69 stations need a bike removal, and 40 stations need an extra bike. We visualized our results through graphs and an NYC map to illustrate the best course of action for each station.

2 Preliminary

2.1 Data Selection

We utilize two datasets in this analysis. The first dataset, named `202307_citibike_tripdata.csv` and hereinafter referred to as "trip data," is the raw data obtained from the official Citi Bike website https://www.citibikenyc.com/system-data, capturing all customer biking information in July 2023. The second dataset, named `citibike_station_information.json` and hereinafter referred to as "capacity data," gives the capacity of each station.

2.2 Assumption

- The trip data includes extensive information about Citi Bike in NYC, encompassing data from New Jersey due to cross-state commuting patterns. However, relocating bikes overnight between stations in different states presents logistical challenges that could impact our analysis. To address this, we will focus solely on stations within NYC and exclude those spanning both states.
- The trip dataset covers biking data for the entire month. However, we'll focus exclusively on weekday biking information, as weekday patterns differ from weekends.
- The trip dataset records biking activity 24/7, but the period from 12:00 am to 5:00 am experiences significantly fewer riders. Thus, we'll exclude biking records during this timeframe.
- We assume rides initiated at each station follow a Poisson process in time, with the same rate at any given time across weekdays but differing rates throughout the day. We will ignore the variation between weekdays.
- We will not account for any form of censoring and exclusively rely on the trip data. This means we won't retain information about customers who were unable to access a bike due to a shortage, and we won't keep records of those customers who must seek another station to return a bike.

2.3 Data Preprocessing

Given the presence of imperfections in the trip data and the considerations of the assumptions mentioned earlier, it was imperative to perform data preprocessing. The initial dataset have 3,776,256 data rows of recording in trip data.

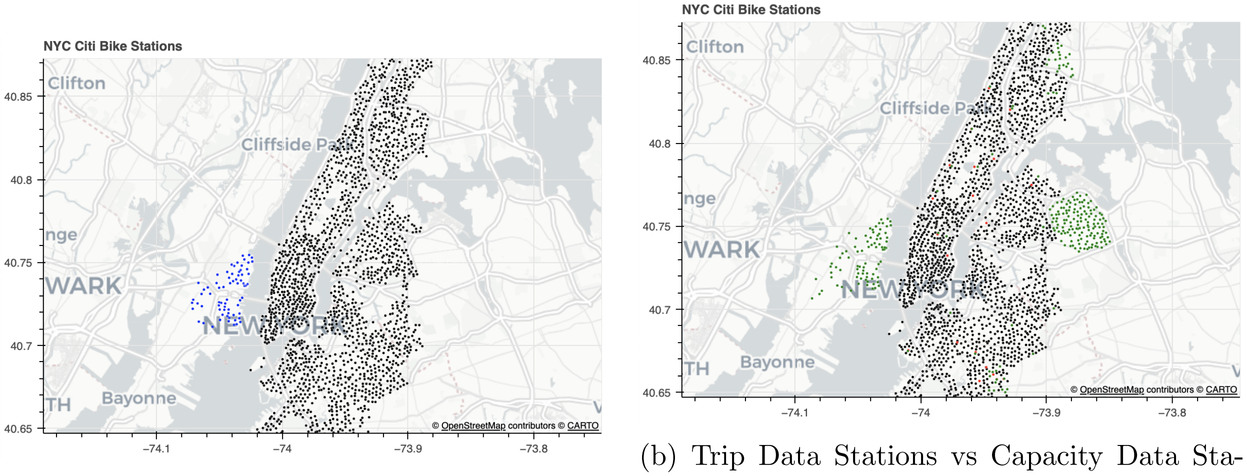
The data preprocessing procedures for trip data are the following:

1. Eliminating incomplete entries, resulting in 3,767,484 rows of data, removing 0.2323% of data.
2. Filtering the dataset to retain data from New York City while excluding entries from New Jersey, leaving us with 3,766,648 rows of data, removing additional 0.0222% of data.
3. Cleaning the dataset by converting station IDs from strings to floats to ensure consis-

tent data types.

4. Comparing the current station list of trip data to the capacity data, deleting any station that does not have a docking capacity associated with. This step left us with 3701027 rows of data, removing additional 1.7421% of data.
5. Excluding data from weekends and the time period between 12 am and 5 am, the dataset has been reduced to 2537448 rows, removing additional 31.4394% of data.

Now, we have a more refined trip dataset, referred as "updated data" hereinafter. We will perform subsequent analysis using this updated data. There are 21 weekdays in July 2023, and since there are 2537448 rows in the updated data, it can be inferred that roughly $\frac{2537448}{21} \approx 120831$ rides using Citi Bike in New York City each day. We create visualizations to illustrate the differences between the original data and the updated data.



(a) New Jersey Stations vs to New York Stations

(b) Trip Data Stations vs Capacity Data Stations

Figure 1: Comparison of Stations Between Different Datasets.

The left graph above illustrates the comparison between stations in New York and New Jersey. The 70 blue points represent stations in the trip data located in New Jersey, which are not utilized in our analysis. The right graph compares stations from trip data and those from capacity data. The 298 green points indicate stations present in the capacity data but absent in the trip data, with 70 stations in New Jersey illustrated on the left, and 130 stations refers to a community with a Citi Bike system not recorded in the trip data on the right side of the graph. The 20 red points signify stations in the trip data that are not present in the capacity data, possibly due to recent additions of new stations. In both graphs, 1851 black points signify stations included in our analysis.

2.4 Preliminary

2.4.1 All Stations

With the updated data, we now have a list of a total of 1,851 unique stations, named "all_station_id".

2.4.2 Dock Capacity

We loaded the capacity data and extracted the docking capacity for each station in "all_station_id", creating a capacity list named 'capacity' that matches each station in "all_station_id". There are a total of 60217 docks in NYC.

2.4.3 Customer Arrival Rate μ

We computed the mean arrival rate for each station between 5 am and midnight, constructing a matrix μ in which μ_{ik} represents the mean arrival rate in a weekday for station i from time k to time $k+1$. To determine the mean arrival rate for a particular station at a given time, we tallied the total number of rides departing from station i in time $(k, k+1)$ over the entire month, and then divided by the number of weekdays in our dataset, which encompassed 21 weekdays in July 2023.

$$\mu_{ik} = \frac{\text{total number of rides initiated from station } i \text{ in time } (k, k+1)}{21}$$

2.4.4 Transitional Probability Matrix P

We computed 19 transitional probability matrices, each corresponding to a specific time staring from 5 am to midnight, to illustrate the likelihood of a rider traveling from one station to another. Within each transitional probability matrix P, the element P_{ij} represents the probability of a rider returning to station j given that the rider coming from station i. In instances where there are no departures from station i during the entire hour, the corresponding entries in the transition matrix for that time are defined as $P_{ii} = 1$, and $P_{ij} = 0 \ \forall i \neq j$. This makes sure that all rows of the probability matrix sum up to 1.

If total number of rides initiated from station i in time $(k, k+1) \neq 0$,

$$P_{ij}(k) = \frac{\text{total number of rides initiated from station } i \text{ to station } j \text{ in time } (k, k+1)}{\text{total number of rides initiated from station } i \text{ in time } (k, k+1)}$$

If total number of rides initiated from station i in time $(k, k+1) = 0$,

$$P_{ij}(k) = 0 \ \forall i \neq j \text{ and } P_{ii}(k) = 1$$

2.4.5 Bike Return Rate λ

We derived the bike return rate matrix λ from the data on customer arrival rates and the transitional probability matrix. The element in bike return rate matrix λ_{jk} represents the number of rides returning to station j in time $(k, k+1)$. It's important to note that we did not directly calculate the bike return rate from the dataset as we did for the customer arrival rate; instead, we utilized the data on customer arrival rates and the transitional probability matrix. This approach assumes that each bike leaving in time $(k, k+1)$ returns to another station within the same hour. While this assumption may not always hold true, it has a negligible impact on the results compared to a direct calculation from the dataset. To compute the return rate for station j in time $(k, k+1)$, our approach involves summing over all 1851 stations, taking the product of the customer arrival rate at station i in time $(k, k+1)$ with the transitional probability of a ride traveling from station i to station j.

Alternatively, this calculation can be visualized as performing the dot product between the kth column of the customer arrival rate matrix μ_{ij} and the jth column of the transitional probability matrix at time k $P_{ij}(k)$.

$$\lambda_{jk} = \sum_{i=1}^{1851} \mu_{ik} \cdot P_{ij}(k)$$

2.4.6 Outage

The objective of this analysis is to minimize the number of dissatisfied users in order to enhance the overall performance of Citi Bike system throughout the city. Dissatisfied users include individuals fail to find an available bike due to bike unavailability, as well as those attempting to return a bike but encountering a full fleet. To determine the number of dissatisfied users, we partition the day into periods and compute the dissatisfied users within each time period, subsequently aggregate these numbers to obtain the total dissatisfied users for the entire day. In this analysis, we will use one hour as the time period. Initially, we must compute the daily bike flow for each hour of station i, denoted as $X(i)$. $X(i)$ represents a list of the number of bikes available at the beginning of each hour throughout the day. $X(i) = [x_i(5), x_i(6), x_i(7), ...]$

To calculate this flow, we rely on two essential factors: the customer arrival rate μ and the bike return rate λ. Let's assume that the bike return rate at time $(k, k+1)$ for station i is denoted as λ_{ik}, and the customer arrival rate for riding bikes away from the station i is represented by μ_{ik}.

Mathematically, the bike count at the end of each hour, denoted as $x_i(k)$, is determined by the following formula:

$$x_i(k) = x_i(k-1) + \lambda_{ik} - \mu_{ik}$$

It is important to note that the resulting value should always remain within the bounds of 0 and the station's capacity d_i to ensure that it accurately represents the number of bikes available. Hence, we adjust the formula to consider this fact:

$$x_i(k) = \text{ max } (\text{min } (x_i(k-1) + \lambda_{ik} - \mu_{ik}, d), 0)$$

Next, we need to compute the expected number of bikes remaining at the end of hour k (at time $k+1$) if we account for all demand. We denote this quantity as $y_i(k)$, which can be computed using the following formula:

$$y_i(k) = x_i(k) + \lambda_{ik} - \mu_{ik}$$

Now, with the above information, we can compute the number of unsatisfied users, which includes two distinct categories. The first category comprises individuals unable to access a bike due to an insufficient supply of bikes. If the expected number of remaining bikes at the end of time k falls below zero, i.e., $y_i(k) < 0$, the number of unsatisfied customer is equal to the absolute value of $y_i(k)$, the number of customers who cannot ride away a bike.

The second category consists of those unable to return a bike because there are no available docks. If the expected number of remaining bikes at the end of time k surpasses the docking capacity d_i, i.e., $y_i(k) > d_i$, the number of unsatisfied customer is equal to the value of $y_i(k) - d_i$, the number of customer who cannot return a bike.

Therefore, the number of unsatisfied users at time k, denoted as $N_i(k)$ can be determined by the sum of the above two situations:

$$N_i(k) = \max\ (y_i(k) - d, 0) - \min\ (y_i(k), 0)$$

And the cumulative count of unsatisfied users over the entire day for station i assuming initial number of bike $x_i(5) = j$ is determined by summing $N_i(k)$ over all k, the each time interval within that day, which is from 5 am to midnight.

$$N_i(x_i(4) = j) = \sum_{k=5}^{23} N_i(k)$$

the number of bikes at each station at a specific time, we can calculate the total outages from the time to the end of the day recursively.

3 Methodology and Result

3.1 Optimal Fleet Size

We calculate the Optimal Fleet Size for each station at 5 am by minimizing the outages. Let $N_i(x_i(4) = j)$ denote the number of total unsatisfied users (outages) over the entire day at station i when assigning j bikes at 5 am (the end of hour 4), where $N_i(x_i(4) = j)$ can be calculated using the method discussed in 2.4.6. For each station i with full capacity d_i, we enumerate the starting number of bikes j from zero to full capacity ($j = 0, 1, ..., d_i$) and find the list of corresponding $N_i(x_i(4) = j)$.

$$N_i = [N_i(x_i(4) = 0)], [N_i(x_i(4) = 1)], ..., [N_i(x_i(4) = d_i)]$$

With this list, we then find the corresponding list of j value that minimizes $N_i(x_i(4) = j)$. Usually, there are multiple fleet sizes j that have the same minimal number of dissatisfied users.

$$J = [j_0, j_1, ...j_i] \text{ where each j in list J minimizes } N_i(x_i(4) = j)$$

In this case, we will take the median of those fleet sizes. This is because we are using the deterministic flow when calculating the arrival and return rate, but with stochastic flow in reality, it is better to take the middle value to allow for some randomness. When the median is not an integer, we will round it down, because we believe that having customers not be able to return a bike is worse than having customers not able to ride a bike.

$$j^* = \text{floor (median } (J))$$

The total optimal number of bikes in the fleet is 29401 bikes, which accounts for 48% of the total capacity. Figure 2 (a) shows the optimal fleet size compares to capacity of each station.

3.2 Overnight Bike Movements

Given the ideal station placement and deterministic traffic, we can compute the daily net flow for each station. This net flow represents the overnight bicycle movements required to maintain the optimal fleet size for the following day. These calculations rely on the customer arrival rate and the bike return rate derived from the previously mentioned data. As the notation described above, let $x_i(k)$ represent the number of bikes at station i at the end of hour k; let λ_{ik} represent the bike return rate in time $(k, k+1)$; let μ_{ik} be the customer arrival rate in time $(k, k+1)$. Therefore, $x_i(4)$ is the optimal fleet size of station i at the beginning of 5am, and $x_i(23)$ is the number of bikes at the end of the day. The net flow of station i is $x_i(23) - x_i(4)$, where

$$x_i(k) = \max\left(\min\left(x_i(k-1) + \lambda_{ik} - \mu_{ik}, d\right), 0\right), \quad k = 5, 6, \ldots, 23$$

By this recursive formula, we are able to calculate each $x_i(k)$ during the day for a specific station i, and the net flow is the difference between the number of bikes at the end of the day and the number of bikes in the beginning of the day, namely $x_i(23) - X_i(4)$. After calculation, the total net flow for all stations is 3161 bikes involve of 1838 stations, and the maximum number of bike movements in one station is 34. Figure 2 (b) shows the number of bike movements of each station, where a positive value indicate bike removal, and a negative value indicate bike addition.

3.3 Predicted Outages

The outages represent instances when bikes are temporarily unavailable to users due to high demand, and the instances when bikers cannot return a bike because all docks are full. We calculate the predicted outages based on our deterministic customer arrival rate and bike return rate, as well as assuming optimal positioning described above. The way we calculated outages at station i in time $(k, k+1)$ is detailed described in 2.4.6.

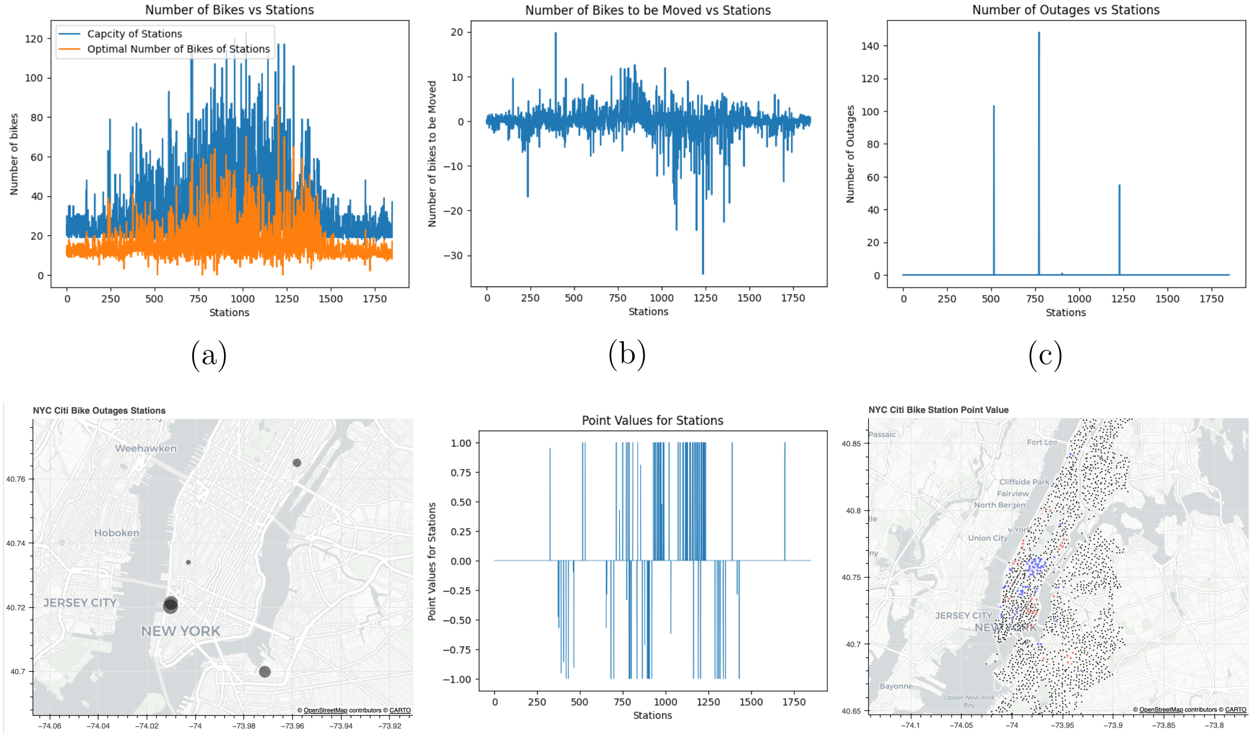
After summing up the outages at each station, our model predicts that over a typical weekday, there may be 438 outages within the Citi Bike system. These 438 outages happens in 5 stations as shown in Figure 2 (c) and (d): "Dock 72 Way & Market St" has 103 outages; "Greenwich St & Hubert St" has 131 outages; "North Moore St & Greenwich St" has 148 outages; "W 4 St & 7 Ave S" has 1 outage; "Ave & E 68 St" has 55 outages. At these stations, the influx and efflux are so substantial and asymmetric during a specific hour that, no matter how meticulously we allocate the initial positioning, we cannot further reduce the number of outages. Compared with there are 120831 rides in NYC each day, only 0.362% of the rides are dissatisfied if employed under optimal positioning.

3.4 Bike Angels Point Values

"Bike Angel" is a Citi Bike program where riders earn rewards for helping redistribute bikes to improve system efficiency. We set the point values for Bike Angels at Tuesday 2 pm by comparing the change of outages when adding/removing a bike. Since our earlier analysis overlooked weekday variations, we continue to rely on this assumption and use the previously discussed rate (λ and μ) when analyzing Tuesday. Assuming each station is half full at 2pm, let $N_i(x_i(13) = \frac{d_i}{2})$ be the number of outages for station i at the end of 13:00 pm (the beginning of 2 pm). We compute both $N_i(x_i(13) = \frac{d_i}{2} + 1)$ and $N_i(x_i(13) = \frac{d_i}{2} - 1)$ to compare whether adding a bike or removing a bike could reduce outage.

Figure 2 (e) and (f) shows our result. In Figure 2 (e), the graph illustrates the outage change associated with each station's response: a positive y-value suggests that adding a bike is better, while a negative y-value indicates removing a bike is better; the magnitude of the y-value corresponds to the number of the outage reduced. In Figure 2 (f), the blue dots indicate stations that should add a bike to reduce outages; the red dots indicate stations that should remove a bike; the black dots indicate stations that adding or removing a bike does not affect the outages. The size of the dots indicate the impact on reducing outages when bikes are added or removed. In determining point values, we can use the change in outages, or proportional to the change in outages, indicated by the y-value in Figure 2 (e), to ensure that the point value corresponds to the system improvement achieved by adding or removing a bike.

3.5 Graphic Result



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2: Graphic Results.

4 Limitations of the Model

4.1 Deterministic Flows

In all our analyses, we rely on deterministic flow patterns observed in the datasets and perform our calculations based on these deterministic arrival and return rates. However, our model may not fully capture certain aspects of stochastic flows, which inherently involve randomness beyond the scope of our model. Therefore, while our model estimates only 438 outages per day happened in 5 stations, in reality, the presence of randomness could result in a considerably higher number of outages.

Addressing this limitation is challenging. An enhanced approach would involve creating simulations that replicate the arrival and return processes using the deterministic values from our analysis as inputs. These simulations would incorporate random variables following exponential or Poisson distributions to model customer behaviors, providing a more comprehensive view of stochastic elements.

4.2 Simplifying Assumptions

We made simplifying assumptions in our model, although these simplifications may bear significance in real-world scenarios:

1. Holiday Oversight: Our model does not factor in holidays that fall on weekdays, treating them as regular weekdays. For instance, July 4th is a national holiday when many individuals do not need to commute to work, leading to a decrease in bike usage for that day.

2. Weekday Emphasis: Our analysis focuses exclusively on weekdays, omitting weekends, and assumes a uniform arrival rate for the same time across different weekdays. While this simplification is reasonable, it disregards the potential disparities in rider behavior between various weekdays, such as the distinction between Monday and Friday, and the influence of weekend utilization on the entire system.

3. Midnight to 5 am Exclusion: The model assumes negligible ride activity from midnight to 5 am. While this simplification may be accurate for certain stations, it may not hold true for others, particularly in 24-hour urban environments

This limitation can be addressed through the following approaches: identify unusual days or anomaly customer behaviors, with the possibility of excluding holidays from the dataset to account for deviations from regular weekday patterns; conduct a parallel analysis focused on weekend data and incorporating insights from both weekday and weekend analyses when implementing strategies in practice; reintegrating data for nighttime biking and potentially treating this period as a unified 5-hour time segment to account for activities that occur in the early hours.

4.3 Limitation of Bikes

In our model, we enumerate and explore the range of outages for each station, starting from an initial state of zero bikes and extending to full bike capacity. Because there are 60217 docks in NYC, this approach implies that we operate under the assumption of having the same amount of bikes available for positioning. However, it is important to acknowledge that in reality, New York City has finite limitations on the number of available bikes, which may differ from our computed values.

To overcome this limitation, there are several strategies that can be employed. One option involves using the sequential greedy algorithm to search for the next best station to increment or decrement the number of bikes when our calculated optimal quantity exceeds or falls short of the actual bike count in NYC, respectively. Alternatively, we use Lagrange multiplier to modify the model by introducing constraints or penalties that account for the practical limitations of the number of bikes available in the real NYC context.

4.4 Censoring

Our model overlooks censoring, relying solely on available trip data. This omission fails to account for outages arising from bike or dock shortages, potentially skewing customer demand representation. This oversight can lead to a cycle of shortages fueled by inaccurate demand information within our optimization model.

Addressing this issue is challenging. Directly recording these outages is impractical. One approach is to gather feedback from customers. For example, implementing a feedback feature in the Citi Bike app would allow customers to report outages, providing a better understanding of outage occurrences and actual demand.

5 Conclusion

In summary, our analysis of Citi Bike sharing in New York City focused on key aspects of service optimization and enhancing efficiency. While our model isn't perfect, it serves as a robust starting point for fleet planning, determining optimal bike numbers, overnight bike repositioning, and customer repositioning incentives. These insights, based on July 2023 Citibike data, provide a foundation for improving Citi Bike operations in New York City. Our findings offer actionable recommendations to enhance the system's efficiency and reliability, and we've included the code for replication or broader analyses.