Probabilistic Allocation in Urban Bike-Sharing Systems

Roshen Chatwal, Ejike Ike, Marko Isakovic, Andrej Basica

April 2025

1 Abstract

In this project, we addressed inefficiencies in Bluebike's servicing across 12 of Harvard's campus stations. We modeled April 2025 bike demand as a Poisson process and bike movements as a Markov process with transition probabilities empirically derived from April 2024 trip data. Through simulation, we studied daily net changes in bike counts across stations and determined optimal initial allocations under a range of fixed constraints on the total bike population. Our analysis found that maintaining around 104 bikes in circulation leads to a point where earned and missed revenues smoothen out, indicating diminishing benefits for Bluebikes and its riders. Our recommendations provide a data-driven strategy for optimizing Bluebikes' operations, and can hopefully boost the program's social impact and profitability for its joint-owner municipalities.

2 Background

Bicycle-sharing systems mark a key innovation for green "last-mile" personal transportation in dense urban centers. They're rarely profitable on their own, though, usually requiring sponsorships to break even. Relatively unreliable public transport in the Boston area exacerbates this issue in the case of its Bluebikes system. The commonwealth-owned MBTA frequently faces overcrowded buses and closures of vital subway routes. Municipalities surrounding Boston must compensate with measures like wide scale free Bluebikes vouchers, foregoing revenue from their jointly owned venture.

Even during periods of adequate MBTA operation, Bluebikes suffers from a misallocation problem generating massive inefficiency for bikers (i.e., not enough bikes docks or docks available at the right stations) and the company (i.e., missed revenues from the rides that fall through). We aim to help fix this misallocation issue for the benefit of local bikers. In doing so, profitability can hopefully improve and better supplement some foregone earnings during times of MBTA malfunction for the municipalities of Boston, Brookline, Cambridge, Everett, Sommerville, Newton, Arlington, Chelsea, Watertown, and Salem[1].

We use publicly available Bluebikes data from April 2024 as a basis to model demand for bikes at each station and simulate their movements between stations during

April 2025. To simplify computations and our bike allocation recommendations to Bluebikes, we solely focus on rides between 12 stations servicing Harvard's campus (refer to Section 7: Station Map for a visual with relevant numbering). Our base model's key simplifying assumptions include:

- A closed system of bikes (no bikes are added or removed)
- Uniform demand throughout the day and month
- Bikers make independent renting decisions
- A bike can always be docked once it reaches its destination (rarely not the case since docks massively outnumber bikes in circulation per our observations)
- Bike rebalancing occurs once per day right at the start of the day

3 Model

3.1 Bike Demand: A Poisson Process

As Harvard upperclassmen, we've observed that there's always tons of people around campus. Events (a bike getting rented from a station) are rare and participation (the percentage of the crowd using a Bluebike) is minuscule during short time intervals. These conditions make the Poisson model a good fit in the context of renters arriving to claim bikes from a station.

We find estimators for the rate parameter, λ_i, for each station i using a simple process:

1. Define our timestep, Δt to be 10 minutes, which is $\approx$ the average duration of rides between stations servicing Harvard's campus. This is key since bikes are rented and docked all within one timestep in our simulation.
2. We count the number of eligible rides starting from station i over all of April 2024, N_i
3. Calculate the rate parameter (average number of events per timestep), $\lambda_i = \frac{N_i}{\texttt{\# of timesteps in April}} = \frac{N_i}{\frac{60}{\Delta t}\texttt{ timesteps per hour} \times \texttt{24 hours per day} \times \texttt{30 days in April}} = \frac{N_i}{4320}$

Now we can assume $A_i \sim \text{Pois}(\lambda_i)$, where A_i is a random variable representing the total number of potential customers arriving at station i to rent a bike and ride it to another station servicing the Harvard community. Of course, this isn't entirely accurate because potential customers who can't find available bikes at a station aren't accounted for in a dataset that only includes actual rides.

3.2 Net Change in Bikes: A Markovian Simulation

Simulating Bluebike rides between Harvard's stations is a twofold process. We first simulate the event of potential customers (agent) arriving to rent a bike at each station i over a Δt-sized interval using the distribution of A_i assumed above. The flows of bikes between stations can be modeled as a Markov process, where:

- Each of the 12 Harvard stations is a state
- The # of bikes currently available at each station is the state variable
- Potential renters arriving at a station to take a bike are independent agents
- Biking from one station to another (not necessarily distinct) station is an action

We use a canonical transition matrix Q to simulate actions based on empirical results. $Q_{12\times 12}$ is a 12X12 matrix, with each q_{ij} being an estimator for the probability a bike departing from station i arrives/docks at station j. We calculate these pretty simply, setting:

$$q_{ij} = \frac{\texttt{\# April 2024 rides from station i -> station j}}{\texttt{total \# April 2024 intra-Harvard rides departing from station i}}$$

Algorithm 1 Bike Sharing Simulation Pseudocode (100 iterations of the below)

```
Initialize array $s0_{1\times 12}$ where $s[i]$ is the starting number of bikes at station $i$
Initialize array $x_{1\times 12}$ of zeros, where $x[i]$ is the # of bikes inflowing into station $i$ by the end of the current timestep
Initialize arrays $Z_{144\times 12}$ of zeros to track net changes in the state variable, successful rentals, and failed rentals at each station over 24 hours
for each timestep $\Delta t$ in 24 hours do
    for each station $i$ do
        Randomly draw the number of potential customers arriving at station $i$ over the interval using the distribution of $A_i$. Store this as $n_{t_i}$.
        Let $r_{t_i} \leftarrow \min(n_{t_i}, s_{t-\Delta t}[i])$        ▷ actual rentals possible
        Update: $s_t[i] \leftarrow s_{t-\Delta t}[i] - r_{t_i}$
        Track the net change in the # of bikes, the # of successful rentals ($r_{t_i}$), and the # of failed rentals ($n_{t_i} - r_{t_i}$) over the interval
        Map successful customers to destination stations using transition matrix row $Q[i]$, where destination counts are drawn from $\text{Mult}_{12}(r_{t_i}, \vec{p})$ with $\vec{p} = Q[i]$
        Update $x$ to reflect destination inflows of bikes from station $i$
    end for
    Update state: $s_t \leftarrow s_t + x$
    Reset $x$ to all zeros
end for
```

3.3 Rental → Revenue Bridge

We can set a variable representing the price of a Δt-minute Bluebike rental, $p = \$5.50$. Let's call the matrices storing successful and failed rentals from each station over the day Z_s and Z_f, respectively. With these, we have all information necessary to calculate revenue earned and revenue missed over an average day in April.

$$\text{Daily Revenue Earned} = p \cdot \sum_{i=1}^{144} \sum_{j=1}^{12} Z_{s_{ij}}$$

$$\text{Daily Revenue Missed} = p \cdot \sum_{i=1}^{144} \sum_{j=1}^{12} Z_{c_{ij}}$$

Since we assume bikes don't get moved by Bluebike staff over the course of day, the optimal allocation balancing both profitability and ridership demand is the s_0^* that:

- Maximizes average daily revenue earned
- Minimizes average daily revenue missed
- Keeps less total bikes in circulation (avoiding costs for barely used bikes)

4 So Where Should The Bikes Go?

4.1 Gauging Bikes Needed: "Unconstrained" Case

Understanding the daily net change in bikes that must be accounted for requires us to first consider the case where there's more than enough bikes. We began our simulations with $s_0[i] = 100 \quad \forall i$, which exceeds the ≈ 25 docks present at the largest Bluebikes stations and leads to no failed rentals. Then, we plotted the evolution of net changes in the number of bikes at each station over the day.

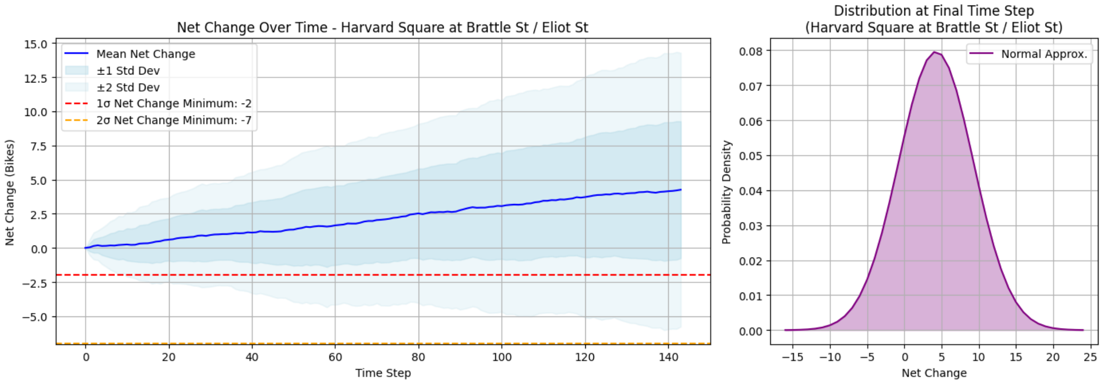

Figure 1: Harvard Square at Brattle St / Eliot St: a net Bluebike importer

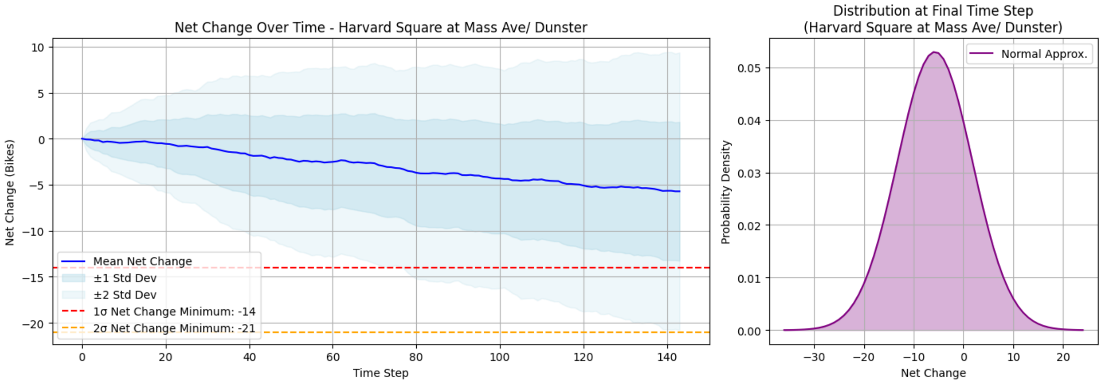

Figure 2: Harvard Square at Mass Ave / Dunster: a net Bluebike exporter

Essentially, we observe two types of Bluebikes stations serving the Harvard communities: importers and exporters. Importers tend to gain more bikes over the day (positive net change) and exporters tend to lose more bikes over the day (negative net change). Given our greater concern for there being a lack of bikes than a surplus, we identify the daily net changes at one and two standard deviations, σ_i, below the mean net change at station i, μ_i. Maintaining enough excess bikes at each station to cover a deficit of $\mu_i - \sigma_i$ leads to a within-Harvard ride provision reliability rate ranging between 77%–88%. Covering a deficit of $\mu_i - 2\sigma_i$ leads to a within-Harvard ride provision reliability rate ranging between $95\% - 99\%$.

We can use these deficit levels to gauge the minimum number of Bluebikes needed in circulation (a real-world constraint) by simply adding them. This is because if each station starts with $|\mu_i + z \cdot \sigma_i|$ bikes, where z is the Z-score of the net-change level we'd like to support, then $\sum_{i=1}^{12} |\mu_i + z \cdot \sigma_i|$ are needed for no station to prematurely run out. To *reasonably service* within-Harvard rides (that 77%–88% reliability range), it takes a total of 74 bikes. To *service most rides* (that $95\% - 99\%$ range), it takes a total of 133 bikes.

4.2 Determining s_0^*: Constrained Case

We now consider a new way to think about bike importers/exporters from stations to model deficits directly without simulation, consistent to our Poisson process. First, we use weighting by expected our earlier calculated departure probabilities to modify our transition matrix to Q', where q'_{ij} is the unconditional probability of a bike actually moving from i to j within a timestep (rather than the conditional probability given a potential renter's arrival at the station):

$$Q = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{12,1} & p_{12,2} & \cdots & p_{12,12} \end{bmatrix} \quad , \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \quad , \quad Q' = \Lambda Q + (I_{12} - \Lambda)$$

We use the transition matrix Q' to calculate λ_i^+ and λ_i^-, the respective bike inflow and outflow rate parameters for bikes at each station. For station i, the import rate is simply the sum of all entries in column i of Q' and the export rate is the sum of all entries in row i of Q' (neither sum includes the diagonal elements because they aren't true inflows or outflows). Using these rate parameters and our independence assumptions, we define a new random variable that's summing our Poisson import/export variables for station i to help model net flow over time:

$$\text{NF}_i = \text{Imports}_i \text{ - } \text{Exports}_i \ \sim \text{Skellam}(\lambda_i^+, \lambda_i^-)$$

We define a loss function over the time horizon using the Skellam$(\lambda_i^+, \lambda_i^-)$ distribution CDF to penalize initial Bluebike allocations B_i that are more likely to experience daily shortages, quantified as:

$$\text{Loss}(B_i) = \sum_{t=1}^{T} \mathbb{P}\left(\text{NX}_{i,t} < -B_i\right) = \sum_{t=1}^{T} F_{\text{Skellam}(\lambda_i^+ t, \lambda_i^- t)}(-B_i - 1)$$

A greedy algorithm then allocates a fixed total number of bikes B by assigning them one at a time to the station i that achieves the greatest reduction in expected loss with one additional bike. This process continues until all bikes are distributed, yielding a near-optimal allocation that minimizes the total expected risk of shortages.

Experimenting with different B values falling within a range guaranteeing reasonable to almost perfect service of within-Harvard rides, we get the approximates s_0^* arrays:

$$\approx \vec{s_0^*} = \begin{bmatrix} \text{Harvard Stadium: N. Harvard St at Soldiers Field Rd} \\ \text{Harvard Kennedy School at Bennett St / Eliot St} \\ \text{Harvard Square at Brattle St / Eliot St} \\ \text{Harvard Univ. River Houses at DeWolfe St / Cowperthwaite St} \\ \text{Harvard Square at Mass Ave / Dunster} \\ \text{Church St} \\ \text{Verizon Innovation Hub 10 Ware Street} \\ \text{Harvard Univ. Gund Hall at Quincy St / Kirkland St} \\ \text{Harvard SEAS Cruft-Pierce Halls at 29 Oxford St} \\ \text{Harvard Law School at Mass Ave / Jarvis St} \\ \text{Harvard Radcliffe Quad at Shepard St / Garden St} \\ \text{Innovation Lab - 125 Western Ave at Batten Way} \end{bmatrix} \implies$$

$$B = 74 : \begin{bmatrix} 5 \\ 4 \\ 7 \\ 3 \\ 12 \\ 5 \\ 6 \\ 5 \\ 8 \\ 9 \\ 5 \\ 5 \end{bmatrix} \quad , B = 89 : \begin{bmatrix} 5 \\ 5 \\ 8 \\ 4 \\ 14 \\ 6 \\ 8 \\ 6 \\ 10 \\ 11 \\ 6 \\ 6 \end{bmatrix} \quad , B = 104 : \begin{bmatrix} 6 \\ 6 \\ 9 \\ 5 \\ 16 \\ 7 \\ 9 \\ 7 \\ 12 \\ 13 \\ 7 \\ 7 \end{bmatrix} \quad , B = 119 : \begin{bmatrix} 7 \\ 7 \\ 10 \\ 6 \\ 18 \\ 8 \\ 10 \\ 7 \\ 14 \\ 15 \\ 9 \\ 8 \end{bmatrix} \quad , B = 133 : \begin{bmatrix} 8 \\ 8 \\ 11 \\ 7 \\ 20 \\ 9 \\ 11 \\ 8 \\ 16 \\ 16 \\ 10 \\ 9 \end{bmatrix}$$

4.3 Simulated Revenues at Optimum

We can translate our $\approx s0^*$ Bluebike allocations to revenues using a slightly modified version of the Markov process in 3.2! We simply set $s0 = \quad \approx s0^*_B$ in line one for B values ranging from 1 to 133. Daily earned and missed revenues are calculated with $p \times$ `total successful rentals` and $p \times$ `total failed rentals`, respectively. We find averages of these values over the 100 simulations and use the simulated 95% confidence interval to obtain the following plot, with vertical red dashed lines located at the B values provided above.

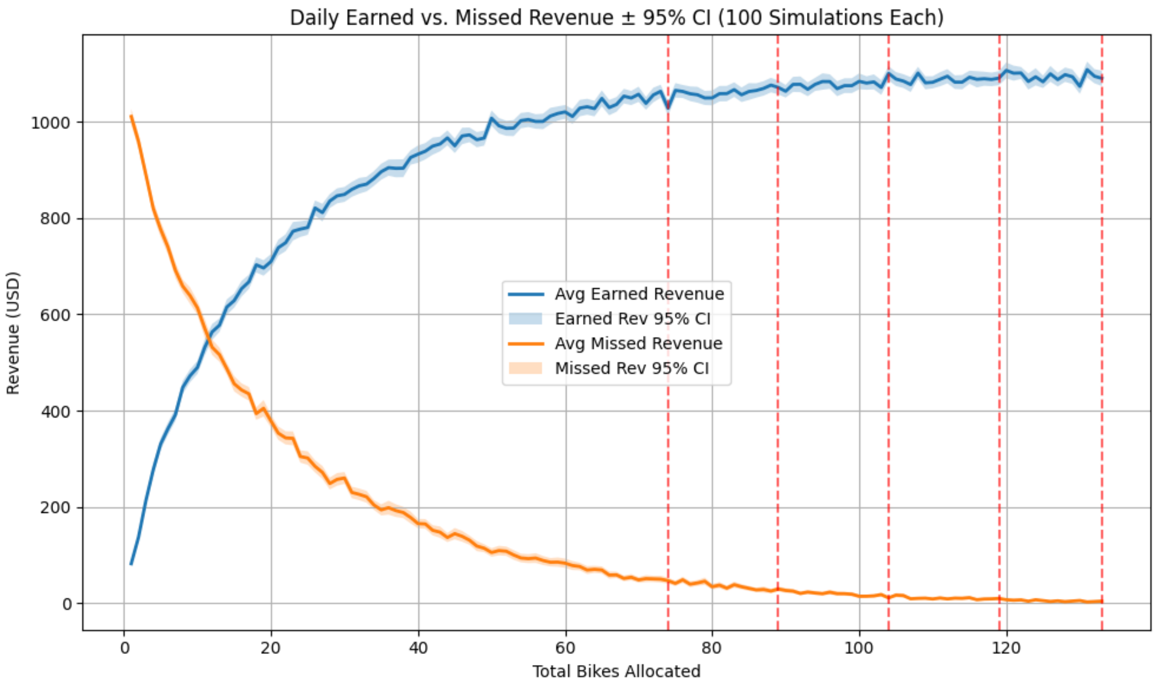

Figure 3: Bluebikes daily earned and missed revenues from within-Harvard rides

5 Discussion

5.1 Starting Allocations and Relative Importance

Per our $\approx s0^*$ findings across different values of B, we can determine that the most important Bluebikes departing stations servicing rides within Harvard are:

1. Harvard Square at Mass Ave / Dunster
2. Harvard Law School at Mass Ave / Jarvis St
3. Harvard SEAS Cruft-Pierce Halls at 29 Oxford St

This ranking matches our expectations, as these Bluebikes stations are located at the intersections of the most densely traveled walkways on campus. We notice that as we continue increasing the value of B by ≈ 15, $\approx s0^*[i]$ jumps by about 2 for $i \in \{4, 9, 8\}$ and only by about 1 for the other stations. This phenomenon also fits expectations, as with more Bluebikes in the system there should be more distributed to stations with higher departure demand. Relative station importance is thus stable across the feasible range of B values, which shows that relative allocations aren't overly sensitive to the final B value chosen.

5.2 Operations: How to Maximize Social Benefit and Profits

Based on our graph of revenues, we see that the Average Earned and Missed Revenue curves heavily flatten near $B = 104$ bikes. Since adding more Bluebikes beyond this quantity provides relatively little marginal benefit to riders and the company in terms of extra transactions facilitated, we believe $B = 104$ lies reasonably close to the true social optimum quantity. Most people who demand rides within Harvard will be able to secure them, and Bluebikes won't have to leave tons of excess bikes out around Harvard. Thus, *we believe Bluebikes can reasonably serve the community best by allocating bikes at Harvard's stations per the* $s0^*_B$ *vector given for* $B = 104$.

According to the canonical microeconomic model, Bluebikes maximizes profit by setting quantity, B, such that marginal revenue, p, equals marginal cost (unknown). Finding this exact quantity is outside our scope (and also isn't the point of our project). However, we believe based on our assumptions about the business model that the primary contributors to a Bluebike's marginal cost are the costs associated with bike transportation at the end of the day back to the optimal allocations. Maintenance and occasional replacement of whole bikes will contribute as well, but probably less so since they're rare and don't involve the relatively high labor cost of paying someone to move bikes.

5.3 Model Shortcomings

Our model is a simplified glimpse at the optimal allocation idea. In order to be computationally efficient, we assume that every ride lasts 10 minutes (our Δt). In reality, ride durations vary depending on factors like start–end station distance and intentionality behind the ride (sight-seeing vs quick transportation). Furthermore, our model is a closed-system only looking at rides within 12 Harvard-serving stations in comparison to Boston's 480 Blue-Bike stations[2]. This isolation adds needed simplicity to the model, but sacrifices the generalization of our results to contexts far different from within-Harvard Bluebike rides. Additionally, we don't consider "self-loop" rides in our $\approx s0^*$ or revenue calculations (they're relatively rare, but we may be losing some insights by ignoring them in 4.2).

5.4 Future Extensions

To account for varying ride durations, we could find the average duration of each of the 144 possible ride combinations. The next steps would be determining where each bike in our simulations intends to go, assigning that ride with the average duration time between those two stations, and only after that time period consider the bike to have arrived at the end stations. A way to do this could be adding a road state which bikes enter and leave according to their respective ride durations. Additionally, we could include more stations in our analysis to make the model more realistic to how the actual Boston Bluebike system operates.

6 Summary

In summary, our Markov chain-oriented model attempts to find the optimal allocation of bikes across 12 Bluebike stations around Harvard for different numbers of total bikes in circulation. We analyzed the April 2024 ride data and considered rides only between our 12 stations of interest. From this, we determined the demand for bikes at each station i via properties of $\text{Pois}(\lambda_i)$ random variables and the simulated bike movements across Harvard's stations using the q_{ij} probabilities in the transition matrix.

We proceeded to look at the "unconstrained" model (1200 bikes) which eliminates unsuccessful rentals and enables a clearer understanding of the variance in daily net changes. Ultimately, the bulk of our analysis relates to the constrained model where we find near optimal allocations of bikes that minimize our loss function for different total numbers of bikes in circulation. We determined that 74 bikes in circulation cover outflows within one standard deviation of the mean net change for each station, and 133 bikes covers outflows within two standard deviations. After simulating and analyzing earned and missed revenues, our model suggests that the social optimal total number of bikes, B^*, is ≈ 104 with an optimal initial allocation across stations of:
[6, 6, 9, 5, 16, 7, 9, 7, 12, 13, 7, 7].

7 Station Map

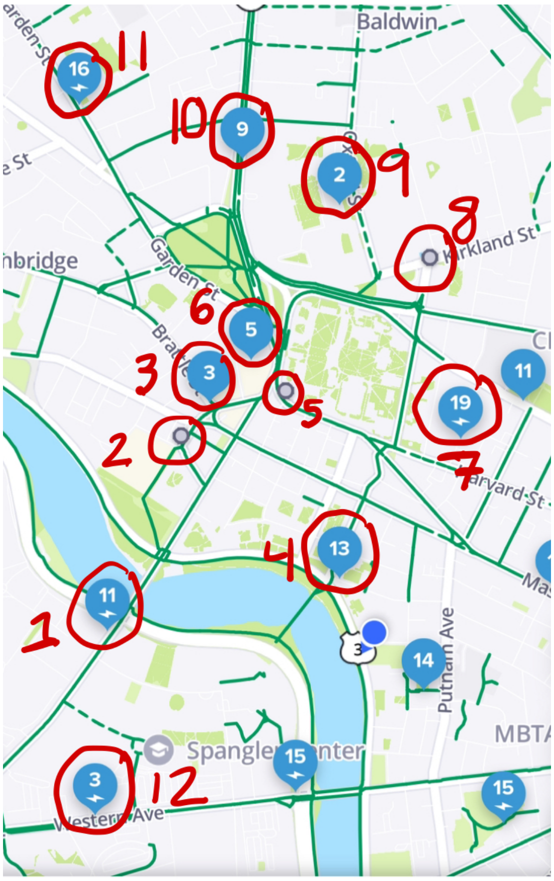

Figure 4: The 12 selected Harvard Bluebike stations (NOTE: i is zero-indexed in code)

8 Sources

1. https://en.wikipedia.org/wiki/Bluebikes

2. https://Bluebikes.com

3. ChatGPT