Modeling Citi Bike Availability

A Markov Chain Based Approach

Elliot Pickens Alexander Pizzirani

November 27, 2021

New York City's Citi Bike system serves thousands of riders each day across the city. It is quite a feat just to keep it up and running, but is there any way to improve performance through data analysis? Using Citi Bike trip data for the month of July 2021, we developed several Markov chain based models to see whether we could use them to model the number of available bikes at stations around the city.

Station Selection

Before we can model bike availability, we had select our target Citi Bike stations. Citibikes may be added or removed from stations for two reasons: riders starting or ending their rides at a dock, or Citibike staff adding or removing bikes as part of system re-balancing or bike maintenance operations (henceforth staff operations). Our dataset captures rider activity, but does not explicitly capture staff operations. It is possible to see evidence of these movements when a bike's next ride begins at a different station than the previous ride's end station, although it is not possible to determine when the bike was removed from the previous station.

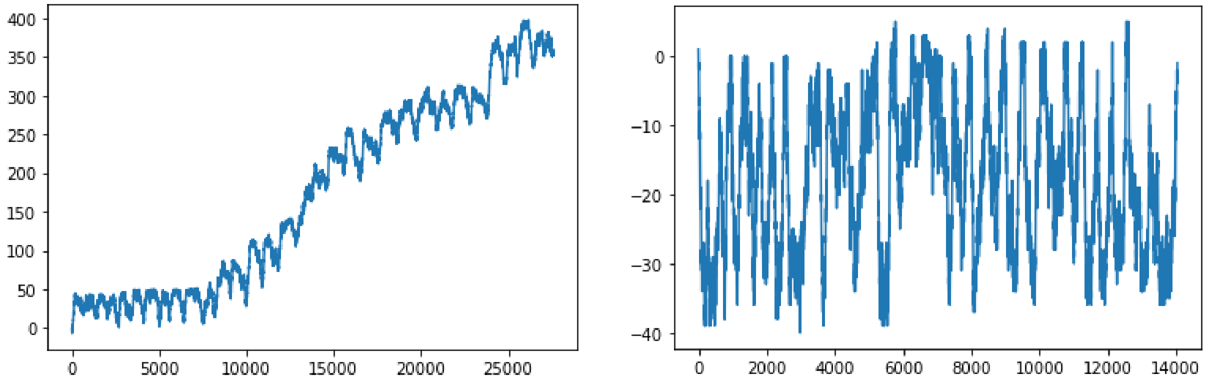
We seek to model available bikes for the selected stations, so we favor stations with limited evidence of staff operations in order to reduce the error associated with these less-observable changes in available bikes. A simple heuristic for the level of staff operations associated with a given station is the difference between the number of rides started and ended at the station. Intuitively, bikes being routinely re-balanced away from a station would result in an imbalance skewed towards rides ended at the station (staff remove bikes if a station sees too much incoming traffic), and vice versa.

We calculate this imbalance metric for stations with more than 5000 outgoing rides in the dataset (subset of resulting data frame shown):

Station	Bikes Out	Bikes In	Difference
8 Ave & W 16 St	6490	6489	1
E 11 St & 1 Ave	7016	7015	1
W 47 St & 10 Ave	5043	5045	2
⋮	⋮	⋮	⋮
Albany St & Greenwich St	5136	5144	8
Great Jones St	7201	7209	8
Sullivan St & Washington Sq	5025	5033	8

This metric is not perfect, as some of the low imbalance stations still show some non-stationarity in the bikes available over the month covered by the dataset. For an example of this behavior, consider the 11th St & 1 Ave station, which drifts upwards before reverting, yet has nowhere near the intense trend characterizing some of the most heavily trafficked stations, such as E 17th St & Broadway:

We ultimately selected the 8 Ave & W 16 St, E 11 St & 1 Ave and Henry St & Grand St stations for their low level of staff operations, relatively high level of activity and limited drift over the period covered by the data set.



(a) 11th St & 1 Ave Station

(b) E 17th St & Broadway

Figure 1: Bike Availability Over Time

A Basic Rate of Arrival Based Model

The data provides information on exactly when and where rides begin an end. Given this we decided to start out our modeling process by trying to build a stationary distribution based upon the rate of customer arrival.

To do this we used the following process to find each station's transition matrix:

1. Split the data into morning and evening periods, and break it down into five minute bins.
2. For both morning and night:
   (a) Calculate the overall change in capacity during each interval as count(Customers ending their ride at station A)− count(Customers starting their ride at station A).
   (b) Count the number of times each change in capacity value occurs.
   (c) Calculate $P(\text{capacity change} = i) = \dfrac{\text{count}(\text{capacity change} = i)}{\sum_{j \in \{\text{capacity changes}\}} \text{count}(\text{capacity change} = j)} \ \forall\, i.$
   (d) Build an ordered $1 \times K$ vector $S = [min(i), \dots, max(i)] \ \forall\, i \in \{\text{capacity changes}\}$.
   (e) Create a $(\text{max station capacity} + 1) \times (\text{max station capacity} + 1)$ zero matrix P that will become the transmission matrix.
   (f) For each row i in the matrix, $P_{i(i+k)} = S_k \ \forall\, k \leq K$

   Note: It is possible that $i + k < 0$ or $i + k >$ max station capacity. We can handle this in several ways.

   I. If the sum of probabilities in a row is < 1 divide the row by the row sum to normalize it.

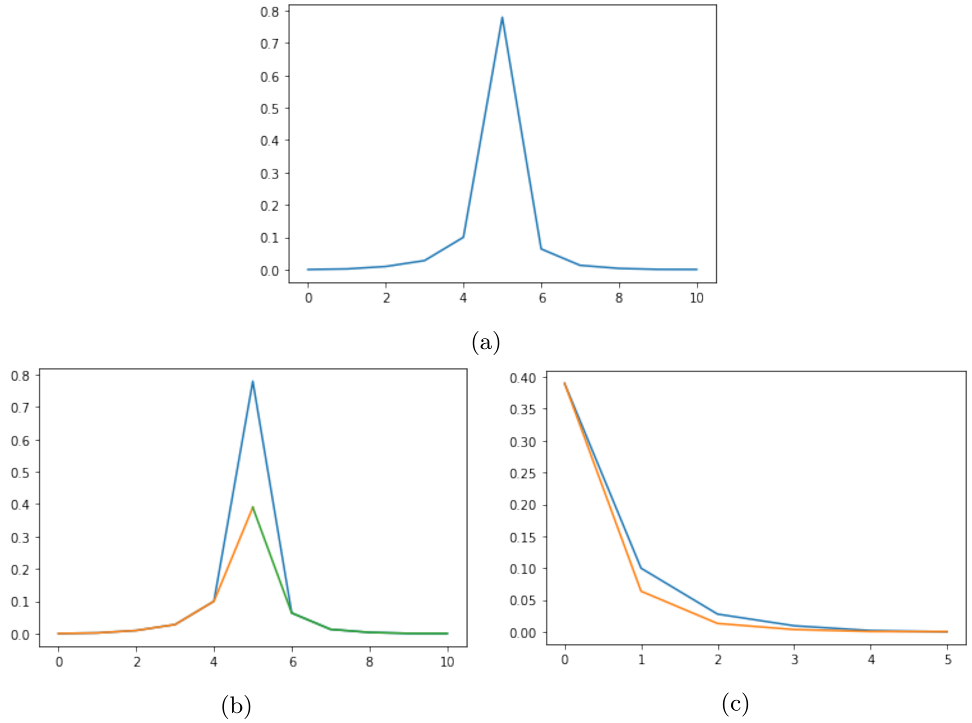
   II. Take the missing probability ($p = 1 - \sum_k P_{ik}$) and redistribute it evenly among the points in the direction of the cut off i.e. if in the infeasible states are P_{ik} where $k < i - j$ we redistribute the probability among the states P_{ik} where $i - j \leq k \leq i$

   III. Take the missing probability ($p = 1 - \sum_k P_{ik}$) and redistribute it evenly among the points in the opposite direction of the cut off i.e. if in the infeasible states are P_{ik} where $k < i - j$ we redistribute the probability among the states P_{ik} where $i \leq k \leq$ max capacity

   (g) Return the (morning or night) transition matrix.

Discussion

With this model we are trying to create a model based on a stationary distribution of customer "flow" within any given five minute period. We do this by subtracting departures from arrivals, and then counting instances of changes. To get a better feel for what this looks like, we can examine the distribution of net bike availability changes over the course of a morning (note that this is not the stationary distribution associated with the transition matrix).



(a)

(b)

(c)

Figure 2: Customer Flow Breakdown (E 11 St & 1 Ave (Morning))
Note: a) and b) chart x-values should be adjusted by subtracting 5 (i.e. they should be centered at 0)

Looking at the graph above we see that the distribution looks strikingly similar to a Laplace distribution. Doing a little decomposition informed by the relationship between the Laplace and exponential distributions we can see that what this model is really capturing is a linear combination of random variables describing the number of customers in or out from the station during a given time period. In this case, the combination is the difference of two exponential distributions. What is most interesting about this is how naturally this relationship arose in the data. This suggests that this stationary distribution of changes is almost a discrete bidirectional special case of a Poisson process.

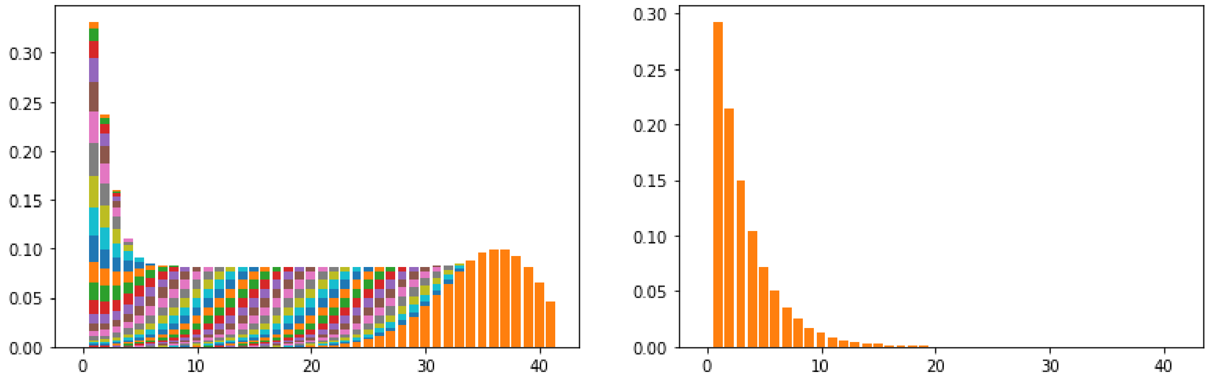
While this distribution of changes is interesting, it is not quite the same as the distributions present in the rows of the initial transition matrix, or the stationary distribution that the transition matrix eventually converges to.

Utilizing this distribution to determine the steps of a random walk, we can estimate the distribution of available bikes at the station after n steps. After 1000 steps, the distribution converges to what looks to be an exponential distribution with a mean around 3.3. Clearly this is a problematic model. This model estimates that the capacity will hover around a level that is nearly empty, which simply does not line up with reality. But why is this the case? If we take a look at 2c we can see that the estimated arrival distribution favors departures. Over time this causes a slow shift toward the favored side, skewing the expectation of the available bikes at the station downward.[1]

If we look at the distributions of available bikes at the station after 144 steps (half a day) we can see what the outcomes of using this distribution as a transition matrix to simulate the number of available bikes after half of a day would be for the full range of different starting values. This gives us some additional insight as to how that drift develops.[2]

[1] We can see the inverse of this occur in 9b

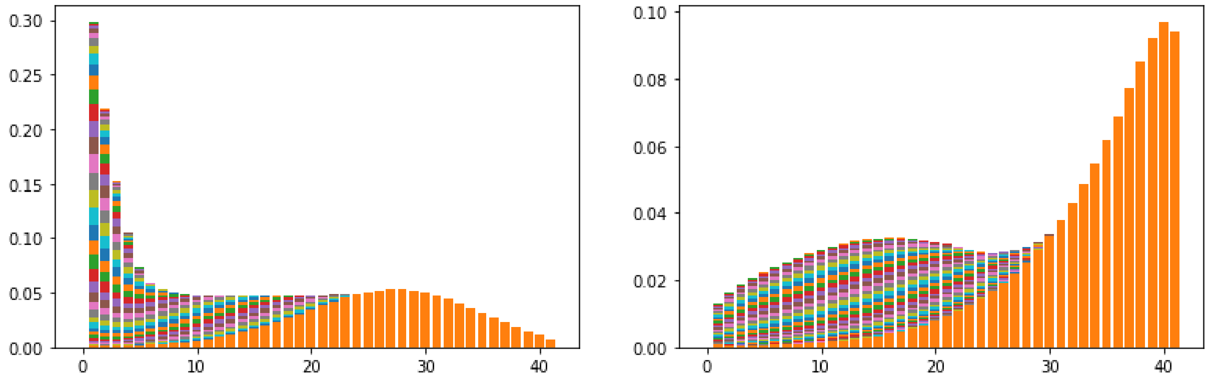
[2] While in 4b we see a heavy left skew this is not commonly the case among the evening models produced by this method. The majority still show a right skew, suggesting bike removals are more common in most cases.



(a) After 50 Steps

(b) After 1000 Steps

Figure 3: Customer Flow Model (E 11 St & 1 Ave (Morning))



(a) After 144 Steps (Morning, E 11 St & 1 Ave)

(b) After 144 Steps (Evening, E 11 St & 1 Ave)

Figure 4: Customer Flow Model (Half a Day Results)

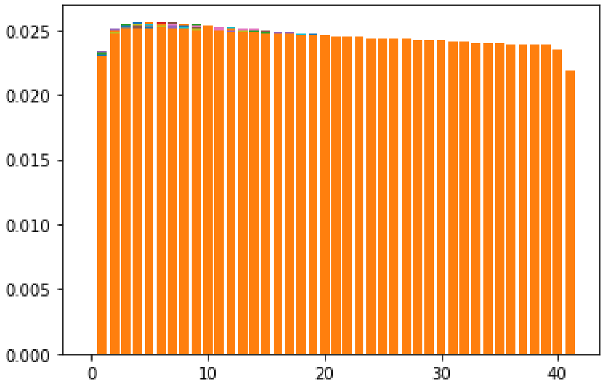


Figure 5: Customer Flow (With Forced Symmetry) (E 11 St & 1 Ave (Morning))

It is possible to prevent this drift by forcing the distribution of changes to be symmetric in both arrivals and departures, but this naturally turns our model into a true random walk neutering any representative power it had.

Taking all this into account it is clear that we need a better model to properly model the behavior of the system. We considered a few different options including expanding upon this change-based strategy, but ultimately we settled upon trying to calculate state changes by estimating the capacity and working from there. In the next section we present this model.

A Capacity Aware Model

The previous model, while useful, is not a capacity-first model. That is to say that we begin by trying to understand customer arrival and then graft that onto the potential state changes. The weakness here is that we might be over-aggregating things and failing to recognize relationships present in the data. To account for this we developed a second model based on first estimating capacity at each time step and then observing transitions between those estimated capacity states.

For this setup, our method for computing the transition matrices and stationary distributions is as follows:

1. Build a "backbone" of the week, weekday, hour and five minute bucket for each of the 3 stations examined, in order to capture data on the net change in available bikes at each station in each five minute period over the course of July, and account for station/five minute intervals when no bikes were moved.

2. Map trip starts and ends at our stations of interest to -1 and +1 changes, calculate some intermediate features and determine the net change that occurs in each five minute interval at each station.

   Note: At this point, we have a full chronology of the five minute interval net changes for each station.

   Caveat: However, we do not know how many bikes are available initially or at any point in time.

3. To adjust for this relative definition of available bikes, we take the cumulative sum of the five minute changes over the course of the month, then adjust it by re-basing the series with the minimum value reached in the first day set to zero. This adjustment is only preliminary and we account for additional minor drifts by de-trending the series using the rolling 24 hour minimum as zero. We also cap the number of bikes at each station's capacity.

4. Having estimated the full series of available bikes, we remove weekends and split the data into morning and evening data frames.

5. For each station, morning/evening data frame, we calculate the transition matrix as:

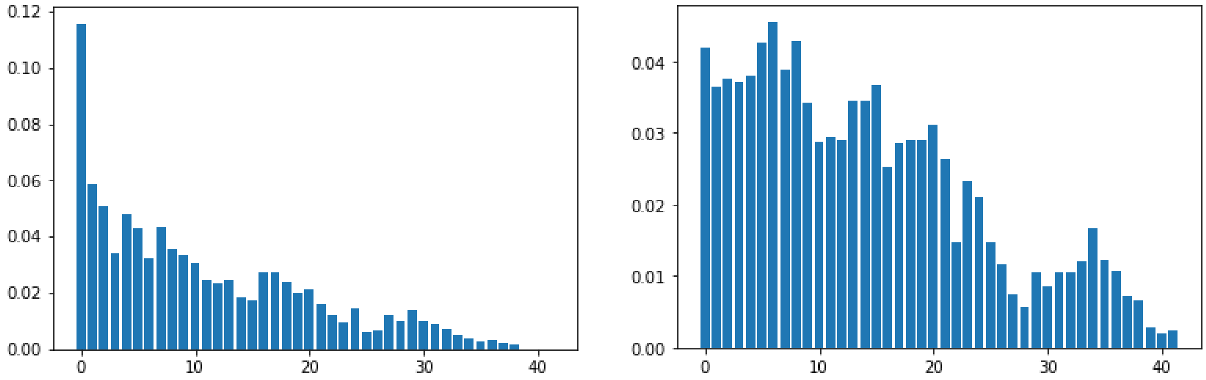
$$P_{ij} = \frac{\text{count}(X_{t+1} = j \cap X_t = i)}{\text{count}(X_t = i)} \tag{1}$$

6. We also estimate the stationary distribution for each station by raising the morning and evening transition matrices to the thousandth power.

This approach produces the stationary distributions shown in 6 for the stations of interest.

For the E 11 St & 1 Ave station, we see that in the morning, the most likely state is an empty station (occurring in approximately 12% of five minute intervals). In contrast, during the evening, we can expect slightly more available bikes at this station. States between 0 and 11 available bikes occur just over 50% of the time in the evenings, whereas states between 0 and 9 available bikes occur 50% of the time in the morning. Overall, this suggests that demand for outgoing bikes from this station is skewed towards the morning. Intuitively, this makes sense given the largely residential character of the station's East Village surroundings, which make it a logical starting point for commuters.

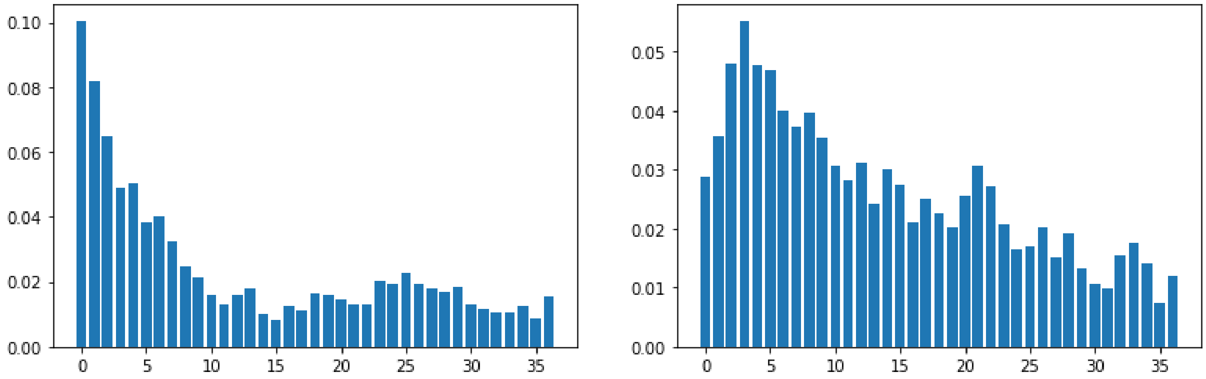
Similarly, for the Henry St & Grand St station, we see that in the morning, the most likely state is an empty station (occurring in approximately 10% of five minute intervals). In contrast, during the evening, we can expect slightly more available bikes at this station. Interestingly, during evening hours, the most likely states are two or three available bicycles, with each of these states approximately 50% more likely than zero available bicycles States between 0 and 11 available bikes occur just over 50% of the time in the evenings, whereas states between 0 and 7 available bikes occur 50% of the time in the morning. Overall, this suggests that demand for outgoing bikes from this station is skewed towards the morning. Intuitively, this makes sense



(a) E 11 St & 1 Ave (Morning)

(b) E 11 St & 1 Ave (Evening)

Figure 6: Stationary Distributions



(a) Henry St & Grand St (Morning)

(b) Henry St & Grand St (Evening)

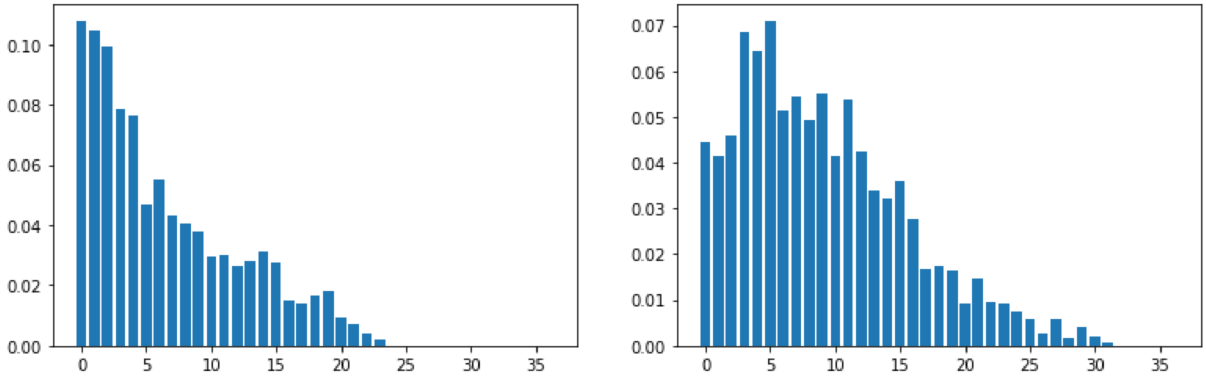
Figure 7: Stationary Distributions

given the largely residential character of the station's Lower East Side surroundings, which make it a logical starting point for commuters.

For the 8 Ave & W 16 St station, we see yet again that the most likely state during morning hours is an empty station (occurring in approximately 12% of five minute intervals). In the evening, we can expect slightly more available bikes at this station, with between three and five bicycles appearing about 22% of the time. States between 0 and 8 available bikes occur about 53% of the time in the evenings, whereas states between 0 and 4 available bikes occur 54% of the time in the morning. Overall, this suggests that demand for outgoing bikes from this station is skewed towards the morning. Intuitively, this makes sense given the largely residential character of the station's surroundings, which make it a logical starting point for commuters. However, this station is also markedly less likely to be at capacity during the morning or evening hours relative to the other stations considered. This may be due to the fact that Chelsea is also home to a number of points of interest such as the Whitney Museum, Chelsea Market and Meatpacking District nightclubs that induce more traffic over the entire course of the day.

Discussion

Overall, we feel that this model does an admirable job of modeling the system. Although the shapes of these distributions might not look too different from those of the first model from afar, the distributions are actually very different upon inspection. There are two main reasons behind this: their tails, and their shape. The tails on these distributions are significantly fatter i.e. there is more density located within the tails of

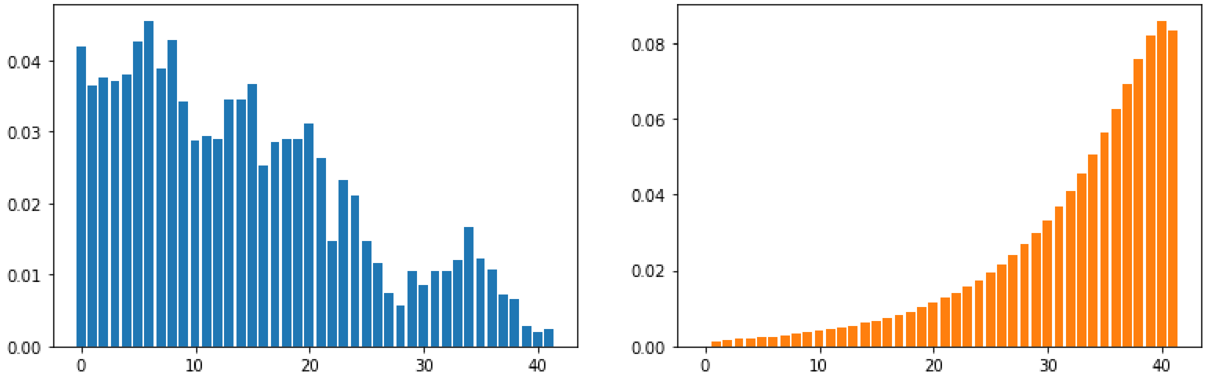


(a) 8 Ave & W 16 St (Morning)

(b) 8 Ave & W 16 St (Evening)

Figure 8: Stationary Distributions

these distributions. Even in the morning distributions where the distributions heavily favor 0 the density of 0 never exceeds 12%, which is small compared to the 30+% of the time the original model resulted in zero available bikes at the end of a morning or evening.



(a) E 11 St & 1 Ave (Evening, Capacity Aware)

(b) E 11 St & 1 Ave (Evening, Arrival)

Figure 9: Stationary Distributions

We can also see that the shapes are notably different with the evening distributions being particularly disparate. The two graphs above show an extreme example of this, but even among those those that are not so strikingly different, we can see that the distributions created by the new method are more moderate. There is also possibly some bimodal behavior occurring in some of these models. The stationary distributions in 7 are a good example of this.

Overall, we believe that this model better captures the dynamics of the system, but there are still some questions we have about these models. Most important among these is how we understand the magnitude of the right skew. It seems possible that zero is too heavily weighted by these models. This could be caused by a number of factors including a simple internal underestimation of the rate at which these stations have zero available bikes due to our limited personal anecdotal evidence, or a minor calculation error in the rolling drift adjustment. It should also be noted that since we deliberately chose stations with low rates of re-balancing we see strong right skews most of the time, because these stations consistently operate with a positive, but fairly low number of available bikes. These dynamics allow the stations to serve incoming or outgoing riders, and do so consistently without requiring staff intervention.

Conclusion

In this report, we presented two alternative Markov chain based models of Citibike station capacity over time. Neither is perfect, but we believe that both capture some important information about the system. The customer arrival model gave use a good idea of the rate at which customers are actually showing up to these stations, but was possibly too reductive. The capacity-based model gives a better understanding of the transition between bike capacity states, but involves some tricky estimation of what is essentially a latent variable.

Ultimately, it appears that modeling the capacity of an individual station using a Markov model is quite a difficult task. We attempted to examine a few possibilities within the scope of the initial question, but plenty remains to be modeled here. It might be, for example, more effective to use a Markov model to model the movement of bikes between stations and understand capacity on a city-wide level rather than looking at such a confined snapshot of the broader Citi Bike ecosystem. It may also be possible that the one state memory of Markov models is not the best method for modeling capacity when there are clear and repeated daily trends. These trends imply serial correlations between the states that may or may not clash with the memoryless-ness of the Markov chain. That being said there is a clear power in the easy estimation of stationary distributions that these models provide - although due care is required to use them effectively.

This project also highlights some of the challenges faced by Citibike staff in maintaining adequate balance in the system, providing insight into the motivation of non-staff-driven rebalancing efforts such as the Bike Angels program.