NYC Citibike Project

Abraham Hill, Evan Teng, Salvador Galarza

November 29th, 2021

Filtering and Pre-processing of Data:

By day:

First, we removed all rides that started during Saturday or Sunday, since the ride patterns on weekends are different from those on weekdays.

By duration:

Then, we removed all rides that had a duration of over 6 hours, under the assumption that these users did not use the bike as intended.

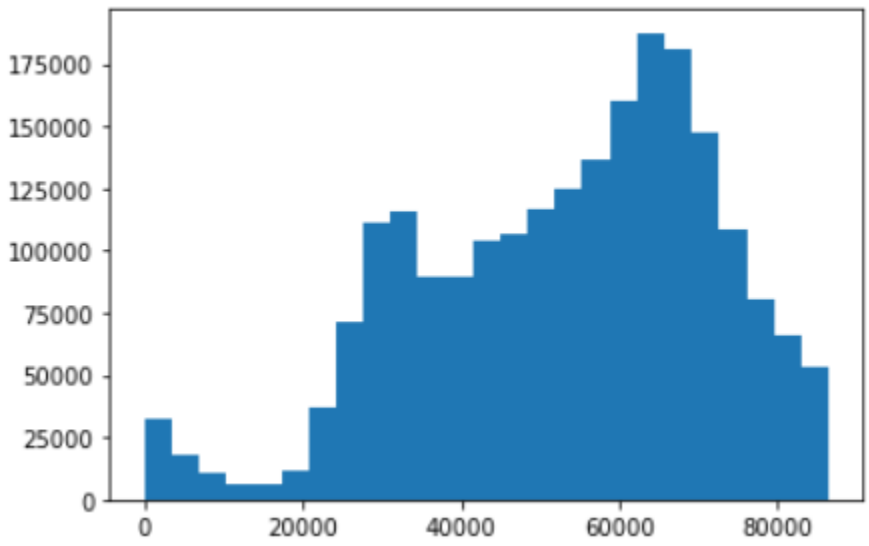
By time:

Lastly, we divided the day into a 'morning' block and an 'evening' block, as patterns are different based on time of day. We then defined 'morning' and 'evening' blocks as follows:

Morning: 5:00am to 2:00pm

Evening: 2:00pm to 12:00am

We decided to choose these time intervals, and not take into account rides between 12:00am-5:00am based off of this histogram:



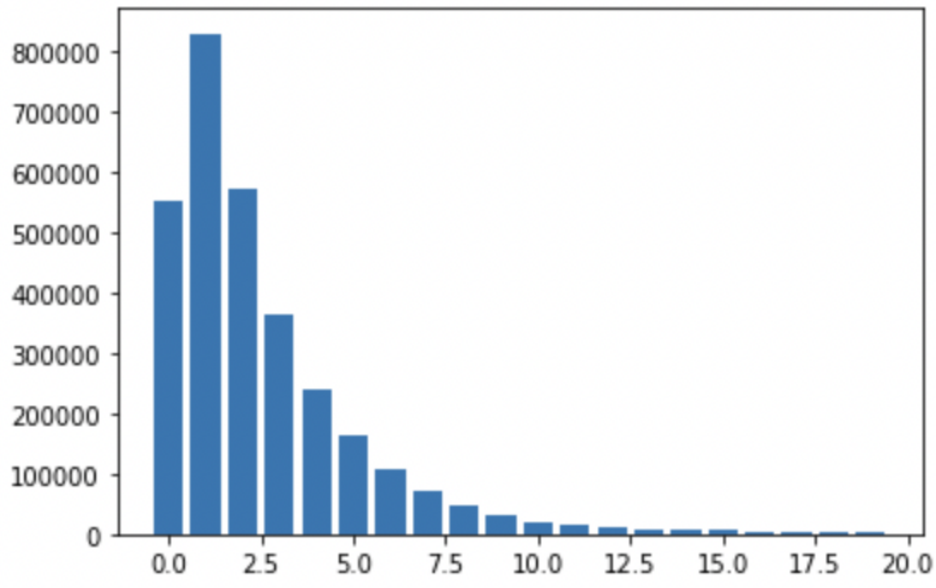
With the y-axis denoting counts (# of rides) and the x-axis denoting time of the day in seconds. Here, we see that between 0 seconds (12:00am) and ~18,000 seconds (5:00am) there is not much biking activity going on, which is expected. At around 18,000 seconds, we see a large spike in bikes being used, so we decided to use that boundary as the start of the morning block. We decided to end the morning block at 2:00pm (50,400 seconds) because at this time we observe a second increase in the frequency of rides, which corresponds to the afternoon/evening rush hour.

Warm-Up Questions

1. Histogram of Ride Durations

Due to changes in the ride patterns on weekends, we filtered out rides that occurred on Saturday or Sunday. We also filtered out outlier rides with a duration above 6 hours, because we assumed that these did not correspond to a user riding the bike as intended.

Histogram: Each bin corresponds to a 5 minute range starting with 0-5min, then 5-10min, theen 10-15min and so on left to right.



2. Expected Ride Duration

The expected ride duration is 16.02 minutes, and the variance is 331.24 minutes. The probability of a given ride lasting longer than 20 minutes is 0.248.

3. Probability of a ride lasting longer than 20 minutes given that the user is a member

P(Duration > 20 | Member) = 0.194

4. Probability that the user is a member given that the ride is longer than 25 minutes

Bayes' rule states that $P(A|B) = \frac{P(B|A)\times P(A)}{P(B|A)\times P(A) + P(B|A^C)\times P(A^C)}$

We computed the quantities P(Duration > 25 | Member), P(Duration > 25 | Casual), P(Member), and P(Casual) and plugged these into Bayes' rule with A representing the event that the rider is a member and B representing the event that ride duration is greater than 25 minutes. The result was P(Member | Duration > 25) = 0.642

Creation of the Transition Matrices:

We chose "5 Ave & E 93 St", "Broadway & W 51 St", and "6 Ave & W 45 St" as our three stations to model because they have a high volume of rides. To model the number of bikes in a station as a discrete time Markov Chain, we first split the time into 5 minute intervals. Then, looping through each 5 minute interval and over every row of the pre-processed and filtered data, we found the net change in the number of bikes in the station. We did this by counting the number of rides that ended in the 5 minute interval at our station, and the number of rides that started in the five minute interval from our station and subtracting the latter from the former. We then stored this into a frequency array `freq` with size = 2*(station_capacity) + 1. The count at index `freq[net_change + station_capacity]` represents the number of observations of that net change. An example representation of the `freq` array is below:

index	0	...						...	2*station capacity
net change	-station capacity	...	-2	-1	0	+1	+2	...	+station capacity
count	#	#	#	#	#	#	#	#	#

We fed this array to the `matrix()` function, which produces the one-step transition matrix for the Markov Chain that models the number of bikes in the station in a given time period. The major assumption of this function is that it doesn't matter how many bikes are in the station, the transition frequencies will be the same. This makes sense in many cases because if a user wants to start a ride from the station, they probably don't care how many bikes are in the station as long as there is at least one. The same is true for rides ending at the station: the user will park their bike there if there is space. The `matrix` function converts the observed number of transitions in the freq matrix to probabilities, then it inserts those probabilities into rows of the transition probability matrix. If there are observed transitions that are impossible given the initial capacity, the matrix function ignores those probabilities and renormalizes the rest of the row so that it sums to 1. For example, if we observe three +6 transitions in a capacity 15 station, and we want to compute the row of the transition matrix corresponding to 10 bikes, then the `matrix` function disregards those +6 transitions because it recognizes that having 16 bikes in the station is impossible. Instead of renormalizing the remaining (feasible) transitions, another approach would be to allocate the probability of all infeasible transitions to the +0 net change. The reasoning behind this is that if users ride in groups, they will not start a ride or end a ride unless everyone in the group can unpark or park a bike in the same station. In the case where every user cannot do so, they may pass up this station and look for another one. We made the assumption

that most users act independently, and so the effect described above should be minimal, which is why we did not take this alternate approach.

Computing the Stationary Distributions:

We computed the stationary distribution by

5 Ave & E 93 St Morning Block:

$\boldsymbol{\pi}$ =

[[3.68218400e-01][2.21111723e-01][1.07396773e-01][6.78708695e-02][9.20455677e-02][7.01407413e-02][3.10647254e-02][1.45881622e-02][1.09412888e-02][8.41423983e-03][4.30134725e-03][1.74888788e-03][8.93225378e-04][6.00749404e-04][3.51390576e-04][1.58696880e-04][7.03687544e-05][3.89477474e-05][2.28203065e-05][1.12743760e-05][4.92270131e-06][2.36502798e-06][1.28082797e-06][6.57189065e-07][3.02220652e-07][1.37489618e-07][6.83194006e-08][3.50795474e-08][1.68278087e-08][7.59545133e-09][3.55097118e-09][1.76236025e-09][8.67568625e-10][4.04025182e-10][1.85300729e-10][8.85971694e-11][4.34162063e-11][2.07936132e-11][9.64133300e-12][4.52209853e-12][2.18510341e-12][1.05435288e-12][4.93949488e-13][2.22869586e-13]]

5 Ave & E 93 St Evening Block:

$\boldsymbol{\pi}$ = [[0.17692856] [0.11195991] [0.06015356] [0.08706676] [0.06011241] [0.0305447] [0.02761552] [0.02246594] [0.01896157] [0.01739827] [0.01567637] [0.01249453] [0.01208396] [0.01225307] [0.01198475] [0.01097191] [0.0113316] [0.00940326] [0.00962928] [0.01182863] [0.01016603] [0.00854206] [0.00867766] [0.00748136] [0.00780817] [0.00889029] [0.00658975] [0.00556389] [0.005748] [0.00555211] [0.00602577] [0.00707011] [0.00544996] [0.00676474] [0.00628432] [0.00625831] [0.00575061]

[0.04814218] [0.03059366] [0.00970131] [0.00867188] [0.0058461] [0.00181981]

[0.04573739]]

Broadway & W 51 St Morning Block:

$\boldsymbol{\pi}$ = [[0.02608339] [0.02831039] [0.02902696] [0.02840421] [0.02809622] [0.02774911]

[0.02868151] [0.02801326] [0.02849688] [0.02909558] [0.02907297] [0.02784099]

[0.02702547] [0.02576406] [0.02584307] [0.02490224] [0.02382278] [0.02400508]

[0.02375701] [0.02359893] [0.02282119] [0.02188379] [0.02099124] [0.0201901]

[0.01997272] [0.02001244] [0.01945943] [0.01883867] [0.01805564] [0.0179923]

[0.01735233] [0.01712378] [0.01646958] [0.01636157] [0.01633719] [0.01578686]

[0.01503993] [0.01419426] [0.01352083] [0.0126852] [0.01207] [0.01145259] [0.01090503]

[0.01014714] [0.00944642] [0.0088392] [0.00821182] [0.00738059] [0.00685782]

[0.00608922] [0.00563114] [0.00527035] [0.00501953]]

Broadway & W 51 St Evening Block:

$\boldsymbol{\pi}$ = [[0.06795702] [0.06568165] [0.063351] [0.06938498] [0.06585285] [0.06265175]

[0.05498152] [0.05698847] [0.0548638] [0.04938491] [0.04614123] [0.03989558]

[0.03494338] [0.03295789] [0.02823476] [0.02444275] [0.02121957] [0.01843027]

[0.01587128] [0.01333687] [0.0116302] [0.01003231] [0.00873183] [0.00757753]

[0.00654103] [0.0057466] [0.00502173] [0.00442713] [0.00688607] [0.00587643]

[0.00516138] [0.0047539] [0.00418836] [0.00372101] [0.003054] [0.00284836] [0.00257698]

[0.00225073] [0.00201869] [0.00171279] [0.00145662] [0.00129528] [0.00108107]

[0.00091809] [0.00078463] [0.00067605] [0.00056532] [0.00046491] [0.00039859]

[0.00033394] [0.00027638] [0.00023116] [0.00018933]]

6 Ave & W 45 St Morning Block:

$\boldsymbol{\pi}$ = [[0.00922321] [0.01008887] [0.01121869] [0.01248822] [0.01327013] [0.01390783] [0.01497936] [0.015956] [0.01628185] [0.01713437] [0.01772292] [0.01820367] [0.01884111] [0.01967416] [0.0200627] [0.02037215] [0.02107265] [0.02152711] [0.02230822] [0.02294481] [0.0238108] [0.02463953] [0.02548073] [0.02605238] [0.02629258] [0.02681115] [0.02687138] [0.02768235] [0.02846058] [0.02858569] [0.02926723] [0.02966894] [0.02880128] [0.0292668] [0.0279282] [0.02778158] [0.02791249] [0.02635366] [0.02598275] [0.02577639] [0.02520616] [0.02533244] [0.02528122] [0.02250713] [0.0209685]]

6 Ave & W 45 St Evening Block:

$\boldsymbol{\pi}$ =

[[8.87732245e-02][8.94472858e-02][8.28500046e-02][8.80265116e-02][8.90320461e-02][7.832 74516e-02][7.31450106e-02][6.54573134e-02][5.79084478e-02][4.82083652e-02][3.99361024e -02][3.29463425e-02][2.71024524e-02][2.19264868e-02][1.77777702e-02][1.44328472e-02][1. 53428820e-02][1.26076075e-02][1.02476059e-02][8.70748871e-03][7.28291729e-03][5.840070 66e-03][4.81783617e-03][3.92472794e-03][3.23529171e-03][2.57687299e-03][2.04448074e-03][1.64196313e-03][1.31170260e-03][1.03555446e-03][8.23862550e-04][6.55905070e-04][5.673 41434e-04][4.50198613e-04][3.53894577e-04][2.83604785e-04][2.27387773e-04][1.79291575e -04][1.43555035e-04][1.13598107e-04][8.99616871e-05][6.92313966e-05][5.36274472e-05][4. 18696105e-05][3.20057317e-05]]

Finals Comments and Insights:

For each of the stations, the morning and evening stationary distributions had different patterns. For 5 Ave & E 93 St, the morning distribution is much more heavily weighted in the beginning, whereas the evening distribution follows a similar trend, but to a lesser degree. For Broadway & W 51 St, the stationary distributions' patterns were reversed compared to 5 Ave & E 93 St. In Broadway & W 51 St, the morning distribution was more evenly spread out, whereas the evening distribution had higher probabilities in the beginning. In contrast to the other 2 stations, the morning block for 6 Ave & W 45 St had higher probabilities at the end, where values steadily increased from start to end.

It can be seen that for 5 Ave & E 93 St station, the majority of the stationary distribution is at the front end, with a gradual, steady decline in probability as we travel towards the end. This indicates that the 5 Ave & E 93 St station is one from which many rides leave from, but is not a destination most of the time. It is important to do this analysis because this tells us that Citibike will have to redistribute bikes so that the demand can be met. For Broadway & W 51 St and 6 Ave & W 45 St stations, the probabilities are much more evenly dispersed throughout the distribution, so less redistribution is necessary.

Overall, it can be seen that each station has its own unique stationary distribution, and that even within the same station, the distributions vary between the morning and evening blocks. These kinds of insights are very valuable to Citibike to help understand the in and out flow of bikes at each station.