MUSA Capstone Proposal 1

Predicting Dockless Bike and Scooter Availability

Elisabeth Ericson

Motivation and background

The motivation for this project is rooted in the rapid emergence of dockless shared mobility as a novel urban transportation mode that was virtually unheard of until in the last five years. Dockless scooters introduced to city streets by private companies quickly drew riders, but also backlash; controversy over growth and management has led many cities to impose increasingly onerous restrictions on the scooters' deployment and operation. Simultaneously, scooter advocates point to the potential for shared micromobility to provide a first- and last-mile car alternative for urban neighborhoods poorly served by transit and traditional bikeshare.

In my literature review, I found research supporting the hypothesis that dockless bikeshare improves accessibility in underserved neighborhoods compared to traditional docked bikeshare systems: In San Francisco, Qian et al. (2020) found that dockless bikeshare provided greater access to bikes in disadvantaged "Communities of Concern" (CoC) than docked bikeshare did, and also helped mitigate the bikeshare usage gap between CoCs and other communities. Lazarus et al. (2020) found that San Francisco dockless electric bikes "more heavily serviced lower-density areas" – which tend to have lower transit accessibility – than docked bikes did. Moreover, analyzing data from 32 U.S. cities with both docked and dockless micromobility, Meng and Brown (2021) found that "the distribution of docked systems is extremely unequal, and that dockless systems greatly reduce geographical inequalities relative to docked." That did not mean dockless systems were devoid of inequality: in a study of dockless bikeshare in the Boston suburbs, Gehrke et al. (2021) found that areas with higher shares of rental housing units and minority residents generated more dockless bike trips, yet had lower access to the bikes.

In that light, I had intended to study the spatial distribution of dockless bikeshare and scooter trips across urban neighborhood, particularly in relation to transit accessibility. Unfortunately, the low spatial resolution of published trip-level data made this analysis less interesting: public bikeshare data tends to round origin and destination coordinates to two decimal places, roughly equivalent to rounding to the nearest kilometer. Given that urban neighborhoods one kilometer apart can have vastly different built environments and transit accessibility, I ultimately chose not to pursue this avenue.

I finally settled on a practical question, inspired by my own experiences as a user of dockless micromobility services: **What are the chances that a bike or scooter listed as "available" in the app will be taken by someone else before I can get to it?** Or to generalize: **What spatial, temporal, weather, and other factors most accurately predict how long an inactive bike or scooter will remain idle between trips?**

Datasets identified

The primary dataset for this analysis will be scraped from at least one API endpoint publishing real-time dockless bike or scooter coordinates in General Bikeshare Feed Specification (GBFS) format. The real-time data includes, most relevantly, a randomized ID for each inactive vehicle, along with its current latitude and longitude. As a proof of concept, and to collect preliminary data for exploratory analysis, I wrote a simple Python scraper and collected this data from Washington, D.C.'s Capital Bikeshare system every minute over a 24-hour period. For the actual data collection, I will want to write a more polished and reliable scraper, and to set up a more robust pipeline to collect data uninterruptedly over the course of several weeks.

Other data sources will cover a range of spatial, temporal, weather, and other features that could plausibly be predictive of how long a vehicle remains idle. Examples might include Census population density data; some measure of employment density; distance to a major road or intersection; distance to rail and bus stations, if possible incorporating some measure of service frequency; distance to traditional docked bikeshare; points of interest (e.g. restaurants, bars, tourist attractions) drawn from OpenStreetMap; day of week and time of day; historical weather data; and availability of other dockless vehicles nearby.

High-level summary of methods

In the briefest possible terms, my proposed methods can be summarized as follows:

1. Scrape real-time bike and/or scooter data every minute for several weeks
2. Clean and manipulate data to compute dependent variable ("minutes until taken")
3. Collect feature data from other sources, as described above
4. Iteratively build machine learning model to predict how much longer a given bike or scooter is likely to remain available
5. Build a basic web front-end to display real-time predictions

This simplification is clearly deceptive, and each step in this process involves substantial technical challenge. By far the biggest risk of this project is that, quite simply, I won't be able to pull it off. It would probably be in my interest to figure out some sort of contingency plan, and I have spent some time thinking about what that might look like, but haven't come to a firm conclusion yet.

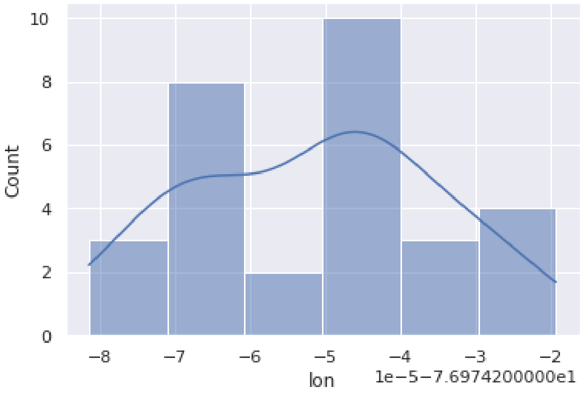
Deliverables

The deliverable for this project would be a simple web app where a user can enter their current location, select a nearby bike or scooter, and see an estimate of how likely it is to still be available when they get there. This interactive deliverable would be accompanied by a written report describing the research process, model development, and findings.

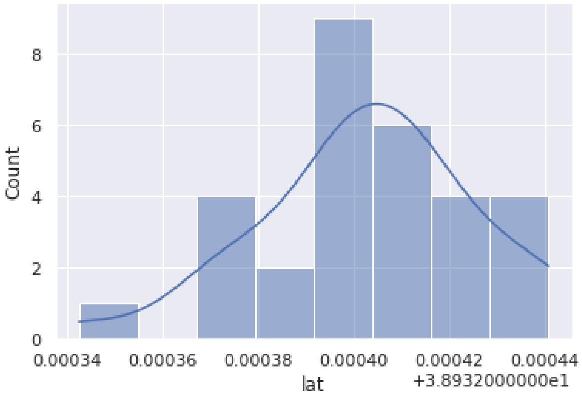
Initial data exploration results

My initial exploration of the preliminary scraped data uncovered some technical complications: first, the dockless bikeshare provider whose data I scraped resets the IDs of its bicycles every half hour; second, due to GPS error, reported coordinates are not always consistent between records for the same bicycle, even when the bike has not actually moved. The data I scraped contained 398,168 rows and 16,883 unique IDs, but the mean and median number of inactive bikes at any given time were only 185.19 and 184, respectively.

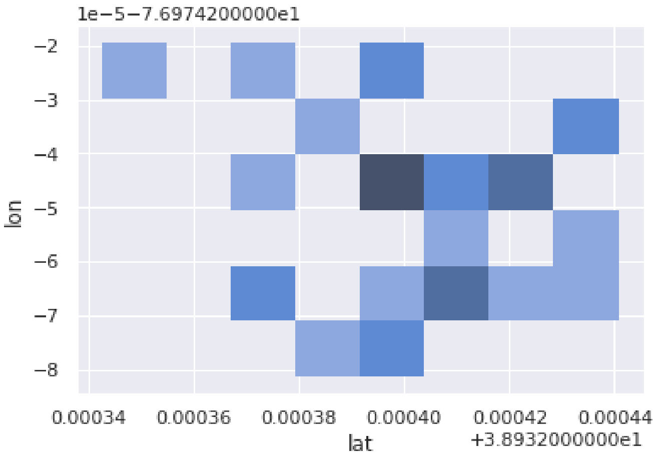
The plots below show the distribution of latitude and longitude readings for a single bike over a 30-minute period (*n.b.* still need to figure out how to make Seaborn show the actual coordinates as axis labels):



Longitude



Latitude



I've started working on possible methods to address both of these issues to be able to manipulate the data into a usable format, but this is absolutely the most significant technical challenge I'm facing right now, and if I can't resolve it in the next week or two, there are aspects of this project I may need to reconsider.

References

Gehrke, Steven R., Bita Sadeghinasr, Qi Wang, and Timothy G. Reardon. 2021. "Patterns and Predictors of Dockless Bikeshare Trip Generation and Duration in Boston's Suburbs." *Case Studies on Transport Policy* 9 (2): 756–66. https://doi.org/10.1016/j.cstp.2021.03.012.

Lazarus, Jessica, Jean Carpentier Pourquier, Frank Feng, Henry Hammel, and Susan Shaheen. 2020. "Micromobility Evolution and Expansion: Understanding How Docked and Dockless Bikesharing Models Complement and Compete – A Case Study of San Francisco." *Journal of Transport Geography* 84 (April): 102620. https://doi.org/10.1016/j.jtrangeo.2019.102620.

Meng, Si'an, and Anne Brown. 2021. "Docked vs. Dockless Equity: Comparing Three Micromobility Service Geographies." *Journal of Transport Geography* 96 (October): 103185. https://doi.org/10.1016/j.jtrangeo.2021.103185.

Qian, Xiaodong, Miguel Jaller, and Debbie Niemeier. 2020. "Enhancing Equitable Service Level: Which Can Address Better, Dockless or Dock-Based Bikeshare Systems?" *Journal of Transport Geography* 86 (June): 102784. https://doi.org/10.1016/j.jtrangeo.2020.102784.