Catch Me... If You Can: Stat 222 Final Report

Isaac Schmidt ISCHMIDT20 @ BERKELEY.EDU
Department of Statistics

Abstract

This report presents our group's work on a public dataset of taxi rides, provided by the New York City Taxi & Limousine Commission. First, we present our motivating question of interest and potential consequences for stakeholders. Then, some exploratory data analysis and cleaning strategies are discussed, along with a formulation of a kernel density estimate of the spatiotemporal distribution of taxi trips. Next, we incorporate public data from New York's CitiBike bike sharing system, which we use to build a comparison between the distribution of taxi and bike trips. Finally, we use the comparison and other methods to make recommendations on the future of selected bike stations.

1. Background and Problem Formulation

The provided data include every trip taken by New York's yellow and green taxis, broken down monthly from January 2009 through July 2021 [1]. This is a rich dataset, containing over 1.7 billion records in its raw state. Our group's original guiding question, in the broadest sense, is **What is the distribution of taxi trip pickups, across time and space, in New York?** Once this question is answered, it raises many potential followup questions, such as "during which time(s) of day are more trips taken?" or "are there more trips taken on weekdays or weekends?" Of course, each of these questions can be answered through individual queries, but estimating the full spatiotemporal density allows nearly *any* question to be answered easily.

After getting a grasp on the taxi data, our next, more precise question, is "How can we determine which bike stations in New York might be appropriate for shutdown or promotional offers?" The dataset of bike trips comes from New York's CitiBike System, and contains all trips taken from June 2013 through December 2021 [2]. A number of stakeholders could potentially be interested in this analysis:

Lyft is the provider of the bikesharing service, in addition to operating its rideshare service. Of course, Lyft would be interested in efficiently allocating the resources of the bikesharing system. In addition, it could leverage bike data to improve its rideshare offerings, and use previous taxi trip data to build dynamic pricing models [3].

Taxi Companies would be interested to know how and where their services are under- or over-performing different modes of transportation, such as bikes. Areas where bikes are a much more popular mode can be potential targets for taxi companies to improve their service. If there are locations where lots of taxi trips are already being taken, the company can send more drivers to search through those areas.

Government is always interested in what its constituents are up to. This analysis could allow public officials to identify certain neighborhoods that could be under- (or over-)served in terms of transportation infrastructure. A community that receives a larger amount of taxi or bike trips compared to its surrounding neighborhoods may be a more desirable place to be in terms of job prospects, availability of resources, etc [4]. Additionally, transportation information such as this can guide public transit agencies, such as the MTA, on places to increase or decrease existing service.

Customers and the General Public would be directly affected by any changes we recommend. For example, the closure of a bike station may require a commuter to take a longer walk, or switch to an alternative mode altogether. People may make personal decisions based on our analysis—for example, a prospective business owner would ideally like to pick a location and choose operating hours with lots of transportation activity.

2. The Datasets

2.1 Taxi Data

As mentioned previously, the downloaded taxi dataset contains every trip taken by yellow and green taxis, from January 2009 through July 2021, which was a total of 1,732,817,071 trips. The fields of interest include the time and location of the pickup and dropoff, along with the number of passengers, the length of the ride, and the cost of the ride. There were other fields as well, such as fare breakdowns and rate codes, but these were discarded.

2.1.1 Data Challenges

One challenge was the size of the data, which made it difficult to store on disk and read it into memory. Our solution here was to store the data as compressed files, and read them in month-by-month. For any operations performed on the entire dataset, such as model fitting, we took advantage of a remote compute cluster for parallel processing. However, the largest challenge faced was the different reporting of pickup and dropoff locations. Up until June 2016, these locations are given as latitude and longitude coordinates. Beyond that time, locations were only reported as "zones," which are areas roughly equivalent to a neighborhood. This unfortunately meant that we could not find exact streets or intersections for each trip.

2.1.2 Exploratory Data Analysis

Figure 1 shows the number of trips in the dataset for each month. Clearly, the vast majority of trips began in Manhattan, with Brooklyn and Queens adding a small amount, and the other boroughs contributing very little. There are a few other patterns to note here, one being the seasonality of the total number of trips, which seems to peak in the spring and crash in the fall. For the long-term, there is constant decline beginning around 2014, which roughly coincides with the introduction and subsequent rise of ridesharing services. Second, the number of trips falls to almost zero in March and April 2020 due to COVID-19 restrictions, followed by a gradual increase.

Figure 2 shows the pickup locations across the city for two particular dates—one for which we do have precise locations, another for which we don't. Again, it is obvious that most trips come from Manhattan, and very few come from the outer neighborhoods. While

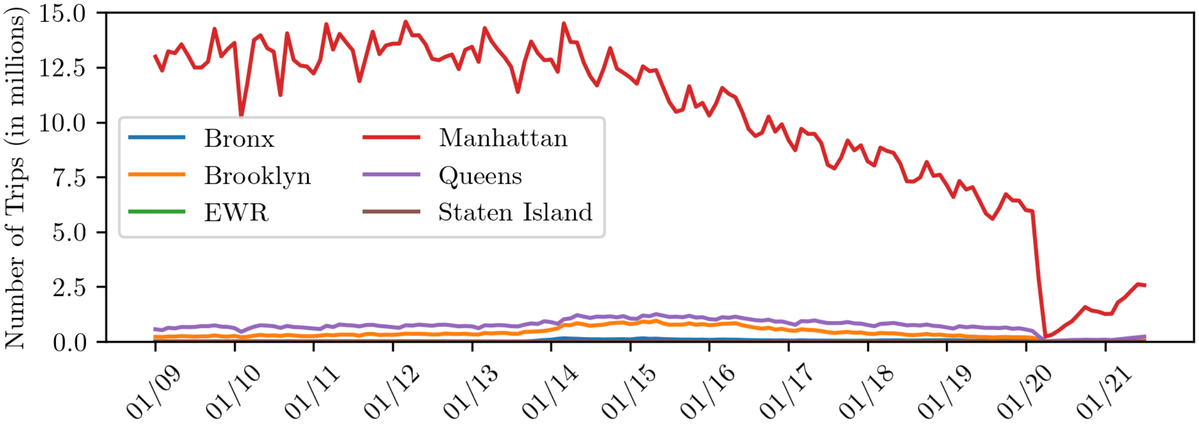

Figure 1: The number of trips per month, broken down by borough of origin.

some zones in Manhattan have close to 10,000 trips in a single day, some areas in Brooklyn or Queens have very few, if any. Another interesting observation is that there are some trips which do not begin in New York at all. Many of these locations are in fact across the Hudson River in New Jersey, but some are in the middle of the water.

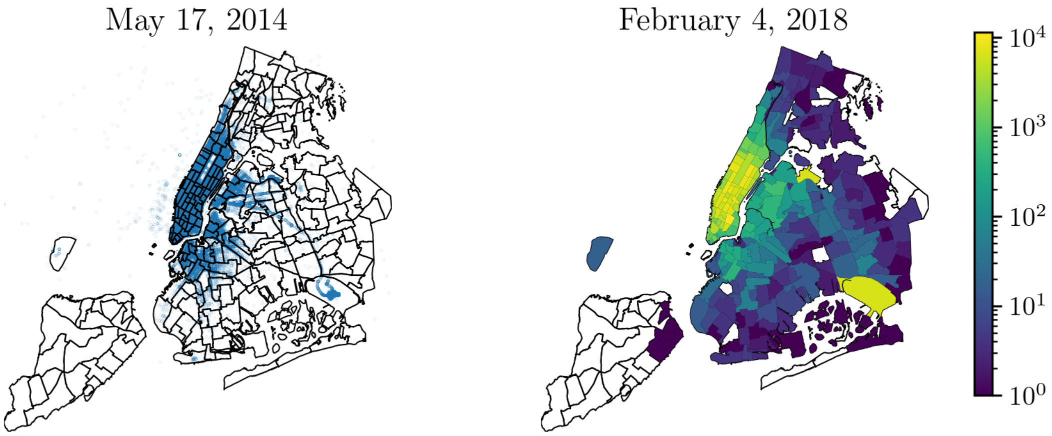

Figure 2: Pickup locations for two selected dates.

2.1.3 Data Cleaning

Apart from the odd locations discussed above, our EDA revealed other data issues. For example, when looking at trips for a particular month broken down by time of day, we found that the average trip cost was negative for certain hours of the day. This meant that we could not take the data completely at face value, and we would have to do some cleaning.

Two main components of our cleaning process were standardizing the column names and orderings across the different months, and transforming the provided latitude and longitude coordinates from the WGS84 geographic coordinate system to a state plane projected coordinate system, to allow for meaningful comparisons between locations. Then, we removed all trips containing unfathomable information, such as a listed duration of six or more hours, or distances of 30 miles or more. Some of these decisions were made based on columns that we did not incorporate into our analysis, such as cost, but we believed nonsensical values in these columns to be a sign that we could not trust any information about such trips. Overall, these cleaning steps removed about 3 percent of the total trips in the database, with bad locations being the most common criterion for removal.

2.2 Bike Data

Similar to the taxi data, the bike data contains every trip taken in the CitiBike system, starting from its launch in June 2013, through December 2021. Included fields are the start and end times of the trip, the start and end locations, as well as some demographic features about the user, such as whether the user is a subscriber to the service, and in some cases, the year in which the user was born. It is worth noting that bike trips can only begin and end at a discrete set of stations, in contrast to taxi trips, which can do so at arbitrary points. Another interesting feature is a unique ID for the specific bike used in each ride, which would theoretically allow one to track each bike as it made its way around the city. Regrettably, we were not able to incorporate this field into our analysis.

2.2.1 Data Challenges

As with the taxi data, the bike data was provided in separate files for each month. In this case, fortunately, the data was a lot easier to work with, as each individual file contained at most a few million trips, in contrast with the taxi data, which in some cases had more than ten million records per month. The biggest problem with the bike data was inconsistency among station names, station IDs, and station locations. The same station name would correspond to different station IDs and locations, meaning it was hard to find a true primary key.

2.2.2 Exploratory Data Analysis

Because our end goal was to use our comparison between taxi and bike trips to make some sort of policy recommendation, we decided to limit our analysis of the bike data to trips that took place in 2021. Trips from previous years, especially pre-COVID-19, may not reflect current transportation patterns, which is why they were excluded. The reason we did not do this with the taxi data was simply because the more recent data did not have the precise pickup and dropoff locations that we were interested in.

Figure 3 shows some information about the number of trips throughout 2021. The plot on the left shows the number of trips per month. Note the small number of trips taken in the winter months compared to those in the summer, probably due to cold temperatures, snow on the ground, and other weather patterns. The plot on the right shows a histogram of the number of trips leaving each station over the entire year. Unsurprisingly, there is a significant right skew, indicating that some stations see far more trips than most others.

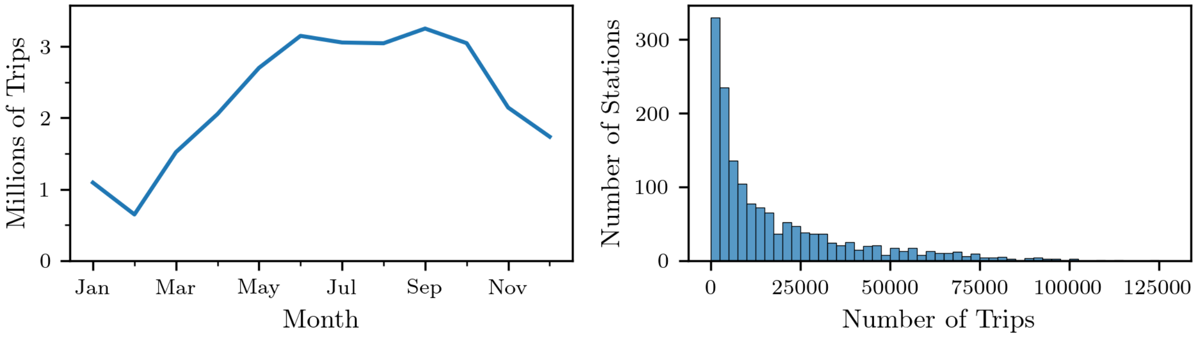

Figure 3: Bike trip counts throughout 2021.

2.2.3 Data Cleaning

Because the bike data was already in relatively clean shape, the only filtering we did was to remove trips containing null values for any field, as well as removing any trips longer than two hours. While it is feasible that a user could be riding the bike for a full two hours, it is more likely that such trips occurred either when a user simply forgot to return a bike, or stored it somewhere other than a station while running some errand. We felt these trips were not representative of travel in the bike system, which is why we excluded them.

The other key component of data cleaning was reconciling the different station identifiers, as described in Section 2.2.1. The station name was determined to be the primary key as it was most definitive. That is, different IDs for the same name might correspond to two stations directly across the street from each other, and we decided there was no point to differentiating such stations. Thus, each name was given the ID and location that were most common amongst all trips with that station name.

3. Taxi Kernel Density Estimate

3.1 Feature Representation

Our original goal was to estimate the true density of taxi trips across space and time. However, questions remained over how exactly we would represent the "features" in our model—i.e. the variables in our joint distribution. Ultimately, we decided on the following:

- x: the x-coordinate of the trip pickup location, in feet
- y: the y-coordinate of the trip pickup location, in feet
- t: the time of the trip, measured as seconds from 12:00 AM Sunday morning

The spatial variables x and y were then linearly transformed such that they fell on the $[0, 100]$ interval, and the time variable t was transformed to be the $[-\pi, \pi]$ interval. Note that t is measured as a point in time throughout a week. This means that we are concerned with any periodicity over the course of a week, but not with the long-term year-on-year trend, or even any seasonality over each month.

3.2 Model

As our goal was to estimate density, a natural choice for a model was a kernel density estimate (KDE). Our KDE takes the following form:

$$\hat{f}\left(x,y,t\right) \propto \sum_{i=1}^{n} \exp\left[-\left(\frac{x-x_i}{h_s}\right)^2 - \left(\frac{y-y_i}{h_s}\right)^2 + \frac{\cos\left(t-t_i\right)}{h_t}\right] \tag{1}$$

where $\hat{f}$ is our estimate of the density at a given x, y, and t, and the sum is over all points in the training set. The "proportional to" symbol is there to indicate that the given expression is scaled by a value such that the KDE integrates to 1 across the entire domain. Note that x and y follow a Gaussian kernel, compared to t, which has the form of the von Mises kernel. This is because while x and y are values that fall along a number line, t is a periodic variable, meaning that the density drawn around a trip taken at 11:55 PM Saturday night *should* wrap around to Sunday morning, which this model accounts for. The von Mises distribution is an approximation to the wrapped normal distribution defined on a unit circle, and is the standard choice for such periodic random variables.

Note also the two different bandwidth hyperparameters, h_s and h_t, which control the "smoothness" of the estimate. Smaller values mean the estimate is more sensitive to the specific points in the dataset, and vice versa for larger values. We tuned each hyperparameter through cross-validation, by training the KDE with different choices for each, and picking the combination that maximized the estimated log-likelihood for a sample of millions of trips. The optimal values selected after cross-validation were $h_s^* = .6$, and $h_t^* = .0003$.

Because we only had precise coordinates for trips before July 2016, all data beyond then was excluded, but we were still left with more than one billion trips. To speed up processing time, these trips were binned by time and location. Specifically, about 1,000 street intersections were randomly chosen from across the city, with a higher density of evaluation points in Manhattan, and a Voronoi diagram was drawn between them. Each real trip in the data was assigned to the bin of the Voronoi cell in which it fell, for the hour of the week in which it started. In total, there were about 180,000 bins. While this decision sacrificed some of the granularity in the data, it sped up processing time immensely.

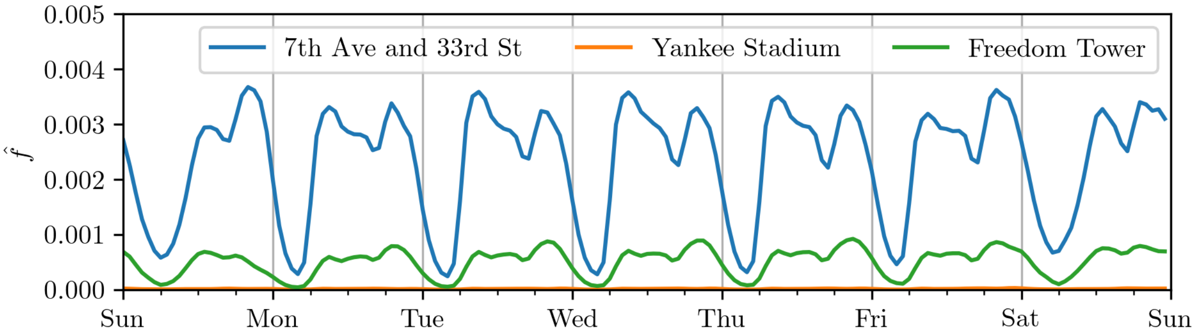

Figure 4: $\hat{f}$ for three selected locations.

3.3 Results

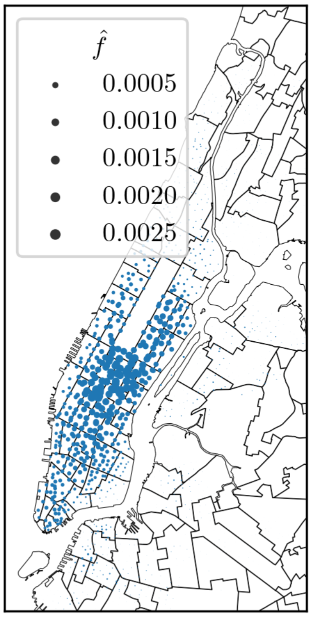

Figure 5: $\hat{f}$ at Mon, 12 PM.

Figure 4 shows the value of $\hat{f}$ across the week for three separate locations. The blue line, 7th Ave and 33rd St, is very close to Pennsylvania Station, a popular transportation hub. One obvious observation is the relative height of the different curves. Penn Station sees more than twice as many trips as a spot close to Freedom Tower, and Yankee Stadium in the Bronx sees almost zero trips. Another point is the daily periodicity, with morning and afternoon peaks visible.

There are some additional subtleties that are worthwhile discoveries. For example, the Freedom Tower's afternoon peaks are only visible during the week and not on weekends, whereas Penn Station sees more trips in the evening compared to the afternoon on most days. While the early peaks are at around 8 AM during the week, they come around noon on the weekends, perhaps due to differences between business travel and leisure activity.

Figure 5 on the left shows the value of $\hat{f}$ for different points in the city at noon on a Monday. Most of the activity is located in Midtown in Manhattan, the Upper East Side sees more activity than the Upper West Side, and anywhere north of Central Park sees very few trips. The same is true for any location outside of Manhattan—yes, there are plotted points there, but the density is so small that they are almost impossible to see.

4. Hypothesis Testing

The kernel density estimate above allowed us to get a handle on the distribution of taxi trips. The natural next step was to compare this distribution with that of the bike data. A big difference between the bike and taxi data, as mentioned before, is that bike trips begin at a finite set of locations. Therefore, we look only at such bike station locations, and our goal is to find stations which see far different bike traffic compared to some measure of taxi traffic.

We decided to run a set of two-sided hypothesis tests, one for each station i:

- H_O: $b_i = t_i$, with b_i and t_i measures of bike and taxi traffic at station i, respectively
- H_A: $b_i \neq t_i$

Stations for which we reject the null hypothesis would be eligible for further analysis for policy recommendations, such as closure, price incentives, increased capacity, and so on.

4.1 Procedure

The first step was to obtain the observed b_i and t_i for each station. b_i is simply the proportion of bike trips that originated at station i, as plotted in Figure 3. To obtain t_i, we calculate the value of the kernel density estimate for taxi trips at station i's location, and then marginalize over time. Then, we normalize these values across all stations such that they sum to 1, to obtain an estimate of the probability of a taxi trip beginning at at bike station i.

To conduct the test, we simply simulate the distribution of bike trips under the null hypothesis. In 2021, there were a grand total of $n = 27,443,540$ bike trips, originating at 1,547 unique stations. Each iteration samples a set of n trips, *assuming the null is true*—that is, $b_i = t_i$. After 10,000 simulations, we obtain the sampling distribution of the bike proportion for each station. Figure 6 shows two examples of such distributions. In both cases, the yellow line displaying the observed t_i is in the middle of the distribution, which makes sense, as the simulation is under the null hypothesis. On the left, at Congress St and Hicks St, the observed bike proportion b_i is well within the distribution, indicating that bike traffic is not much different than taxi traffic at this station. On the right, at Broadway and W 25 St, b_i so far exceeds t_i that the null must be rejected—we conclude that $b_i \neq t_i$ for this station.

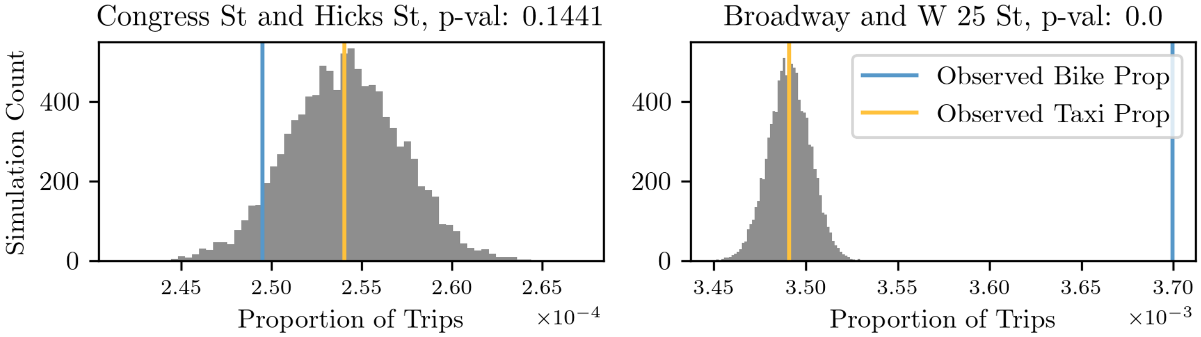

Figure 6: Sampling distributions for two selected stations.

Note that the test statistic here is $|b_i - t_i|$, as the tests are two-sided. The plots above show the null distribution of b_i, not the test statistic. To properly visualize the p-value, consider a second blue line mirrored across the yellow line. The p-value would be all area in the histogram outside of these two blue lines.

4.2 Results

Surprisingly, the p-values were exactly 0 for 1,516 out of the 1,547 stations. This means that for almost every station, we rejected the null hypothesis and concluded that bike traffic was really different than taxi traffic. It is important to recognize that we tried variations on the above procedure, for example by considering one-sided alternative hypotheses, or even limiting the scope of stations to those that received the most traffic. In all cases, we obtained similar results, and to avoid entering into the realm of "p-hacking," we were forced to stick with our original results. There are a couple plausible explanations for this, one being the very large "sample size" of observed bike trips. With over 27 million trips, the variance around the observed proportions will be very, very small, and the sampling distributions under the null will be very, very narrow. The other reason is that it is likely that taxi traffic really is that much different from bike traffic, and it was silly to expect much similarity between the two. Perhaps the relative closeness of the two for a handful of stations may be more coincidental than anything else.

5. Final Recommendations

As stated at the beginning of the previous section, the motivation behind the hypothesis testing was to pinpoint "interesting" stations that we could further analyze and suggest policy on. However, if everything is interesting, really nothing is. Therefore, we looked to an alternative method of highlighting stations. Ultimately, we decided to investigate bike stations which are underperforming relative to their neighbors, so we developed the following procedure:

1. For each bike station i, find its closest station j.
2. Calculate the ratio $\frac{b_i}{b_j}$. A ratio above 1 indicates that i is more popular than j
3. If $\frac{b_i}{b_j} \geq 4$, recommend closing down station j
    - If the ratio $\frac{t_j}{t_i} \geq 1.25$, then we can recommend offering a discount to ride a bike from station j

Our reasoning for the above is that if a bike station is really close to another bike station that is much more popular, there is little point in continuing to operate the less-used station. The motivation behind the last bullet point is that if such an underperforming station happens to see significantly more taxi usage than its neighbor, then that indicates that there is still some demand for transportation in general at that specific location. Therefore, it may not make sense to close the station entirely, but instead to offer some sort of price incentive to entice prospective riders.

Just because a station j is closest to station i does not necessarily mean that i and j are sufficiently close to be comparable. Fortunately, the concentration of bike stations in New York is very high, and in all cases, the closest stations were never more than a quarter mile apart. This means that if any particular station was closed, a regular user of that station could easily switch to another station without much hassle.

Using this methodology, we suggest 33 stations for closure and an additional 14 stations for price discounts, which are plotted in Figure 7. However, there are a couple caveats with these determinations. The first point is that we only analyzed the single closest station. It is possible that there are cases where station A is closest to station B, which itself is closest to station C, which sees far more traffic than both A and B. Here, it might make sense to make a recommendation about both A and B, but our procedure would only recognize B. The other point is far less technical, which is that the closeness of two stations may not be synonymous with their relative accessibility. There may be a very good reason to keep a certain station open, which Lyft and city planners may be well aware of. Our suggestions completely abstract away from any and all context, some of which may be important.

6. Conclusions

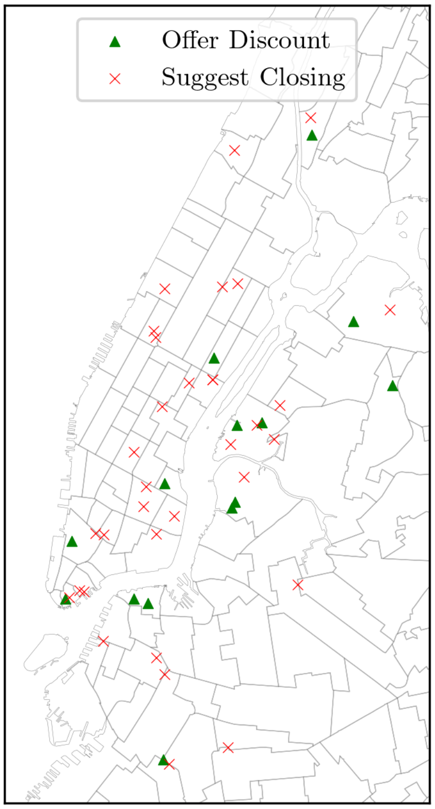

Figure 7: Our recommendations.

This report presents our group's work first on analyzing a dataset of taxi trips, by creating a kernel density estimate for the spatiotemporal distribution of taxi pickups in New York. Then, we introduced a dataset of trips from the CitiBike bike sharing system, and compared taxi and bike travel patterns at selected locations through multiple hypothesis testing, with the aim of determining a set bike sharing stations as candidates for policy recommendations. Unfortunately, this comparison did not lead to results worth analyzing, so we developed an alternate procedure to identify underperforming stations, by comparing the performance of each station to its neighbor. Ultimately, we made a recommendation to either close or offer price incentives for 47 stations in the CitiBike system.

There are a number of steps that could be taken to extend this work. One is to increase the complexity of the KDE, which at the moment is only a joint distribution over three random variables. In addition to pickup location and time, we could incorporate information on dropoffs, which would enable analysis on both the duration and distance of trips. Other features could be included as well, such as hourly temperatures. In the context of this analysis, an improved KDE may lead to a better comparison of bike and taxi traffic at each station, but there could be a number of other uses as well. Another path would be to explore the journey of each individual bike as it makes its way around the city, using the provided "BikeID" variable. Tracking individual bikes could provide more practical insight on which stations are working well and which are not, or better optimize the manual recycling of bikes from station to station, which is another key expense of bikesharing programs.

References

[1] “TLC Trip Record Data,” Aug. 2021.

[2] “Citi Bike System Data,” Feb. 2022.

[3] B. Marr, “The Amazing Ways Uber Is Using Big Data Analytics,” May 2015.

[4] C. Xie, D. Yu, X. Zheng, Z. Wang, and Z. Jiang, “Revealing spatiotemporal travel demand and community structure characteristics with taxi trip data: A case study of New York City,” *PLOS ONE*, vol. 16, p. e0259694, Nov. 2021. Publisher: Public Library of Science.