New York City Uber Pickup Analysis

Parth Mishra*
University of Colorado
parth.mishra@colorado.edu

Leif Waldner
University of Colorado
leif.waldner@colorado.edu

1. ABSTRACT

This project aimed to find the correlations between neighborhood demographics and Uber pickup frequency. We wanted to know if there were certain demographic profiles that could potentially affect pickup frequency. By integrating demographic data for New York's various neighborhoods, we cross-referenced them with Uber pickup data gathered from January 2015 to June 2015 and displayed the results in various univariate, bivariate, and multivariate relationships. Our results were heavily skewed towards higher population areas such as Downtown Manhattan. That being said, there were interesting but not unexpected relationships between median neighborhood income and the amount of pickups that occur. People with more money tend to get picked up more. While our results do not provide insight into any truly groundbreaking results, they set the stage for further research that builds upon the results we gathered.

2. INTRODUCTION

The city of New York is home to the largest concentrated population in the entire country and as such is a haven for the growing trend of ride sharing companies as a direct competitor to Taxis and other mass transit options. The convenience and cutting edge technology that these companies rely on have allowed companies such as Uber and Lyft to flourish. With their rise has come ethical concerns revolving around the dynamics of the driver-passenger relationship. All across the country, there have been reports of how drivers may abuse the pickup system to cancel rides on people for various reasons including, but not limited to: race, neighborhood safety, distance, etc. These issues have prompted multiple iterations of driver policies from these ride sharing companies based on this kind of feedback.

This project aims to highlight some of the socioeconomic relationships when looking at pickup frequency in certain neighborhoods. New York city is the perfect candidate city to do this analysis for several reasons. The first reason is that there is just a large number of pickups that occur relative to other cities. This allows us to draw more solid patterns based on a larger amount of data. Rural areas are harder to draw conclusions about due to scarcity of ride sharing in these areas. Additionally, while many cities may fit the bill of having large populations of riders, New York City distinguishes itself due to the heavy diversity its very concentrated population. This stands in contrast to a city such as San Francisco, which, while being the home of Uber, has a more homogeneous demographic profile in its various neighborhoods. New York city has 5 major boroughs and smaller neighborhoods within those boroughs that feature people from all walks of life. The diversity in socioeconomic status creates a microcosm from which we can observe trend propagation on a very concentrated level. The results of

this analysis could prove useful when looking at other cities such as Los Angeles which has similar qualities to New York City.

3. PRIOR WORK

Most of the prior work in this area has revolved around analyzing Uber pickup data with regards to its impact on traditional public forms of transportation such as bus systems and taxis. Analysis in these areas compared pick up rates, locations, and times, to compare the service models. Using the integration of various transportation services with Uber's data, they were able to paint a comprehensive picture of the differences and similarities between these services from geospatial perspective. Information from these various studies has been used to inform both public opinion as well as the legal landscape of ride sharing services as it pertains to competing with more traditional forms of transportation.

Much to our surprise, very few data-centric approaches have been made regarding the relationship of Uber with socioeconomic variables. Mostly reports in this area are with regards to the business model itself. While we were unable to find specific studies that examined the data behind this topic, the general interest in the potential ethical implications of Ubers business model reinforced our desire to examine the data and find the relationships, if any.

4. DATA SET

Our main Uber data set comes from a Kaggle Dataset[2] we found. This dataset was previously created and used by data analytics company fivethirtyeight. This dataset contains 18.5 million data points. This dataset contains part of all we need to know about Uber rides in NYC. The dataset is a simple record of the time, location, and dispatching base for an individual pickup. The location is encoded with a "locationID" which, confusingly, refers to a taxi-zone pickup. These locationIDs are loosely correlated with individual neighborhood in the greater New York City area. The time attribute encodes both the nominal date of the pickup as well as the the time of the pickup down to the hour and minute. The value of this dataset lied mostly with the location as it is the basis for our analysis. We also used the time data too but it is mostly irrelevant for our particular brand of analysis. The dispatching base attribute was disregarded completely.

The second major source of data was our demographic information dataset[1] provided by the New York Department of City Planning. We used this as part of our integration steps. We initially were looking at data that just encompassed the five major boroughs but pivoted to using neighborhood data instead since there was large variance in community profiles within the boroughs. For example, parts of Brooklyn are very gentrified and present drastically different socioeconomic profiles than neighborhoods that are less than a few miles away which are poor, low-income communities. This dataset contains a very large amount of demographics information that are segmented by various

5. ANALYSIS TECHNIQUES

The following is a comprehensive look at the various techniques we used in our analysis. The general overview of the process steps include data integration, preparation, analysis, and visualization. We will examine each step in chronological order:

1. Data Integration

Here the goal was to find and collect all of our data into a singular store from various sources outlined previously. These steps were quite involved and required pulling data from various sources. The goal of this project is to examine insights related to the Uber pickup distribution with regards to location and demographics. New York has very diverse neighborhoods that can vary greatly from each other even though they happen to be in the same overall borough (Queens, Bronx, Manhattan, etc.) Getting fine tuned demographic information that corresponded to the locations provided by the Uber data proved to be difficult as there is no one clear community definition/boundaries that have demo-

graphic information available. The location IDs in the Uber data set are very arbitrary and do not easily correspond to any area definition standard that we would be able to find demographic information for. As mentioned previously, they do loosely correspond to colloquially known neighborhoods within boroughs. In order to integrate with existing demographic data, we needed to map these locations to standardized areas. We ultimately had to select one standard out of many for defining neighborhoods that would capture the most granular information possible. We ended up choosing a standard known as the "Public Use Microdata Area" or "PUMA" for short. This seemed to be the most standardized of the area definitions we could find. Other standards such as Neighborhood Tabulation Areas (NATs) were considered but ultimately not used.

This mapping of location IDs to PUMAs was not a perfect process and thus some manual decisions needed to be made regarding some location IDs in the Uber data that did not fit perfectly into the PUMAs. While not an issue for the majority of mappings, we had to manually encode some areas for which there was no clear mapping. These areas could have potentially influenced our results but generally speaking, these areas were mostly in the outskirts of Brooklyn and Queens whereas the more well defined areas were in Manhattan and Staten Island.

Interestingly enough, we ran into a situation in which we found out that airports are popular locations for Uber pickups but in most cases, they are not their own tabulated region with appropriate demographic and socioeconomic statistics. This makes sense since airports are distinct, transient places that share little with the surrounding community. We were not exactly sure how to treat these locations (LaGuardia, JFK, and Newark) and may needed to modify them in our future analysis as they have the potential to skew some statistics due to their large volume of pickups relative to other zones. Eventually, we ended up keeping them in since they ended up not really affecting the overall statistics generated when excluding them versus including them.

Ultimately our final integrated dataset has several fields including:

- Uber Location ID
- PUMA ID
- Median income
- Total income
- Percent of Population Primarily Walk to Work
- Percent of Population Primarily Take Public Transportation to work

The augmented data we pulled from has a lot more fields (25+) so we may revisit and add additional data fields now that the initial process for integration is solidified. For example, there is additional data faceted by race and ethnicity that might be worth looking into as well. For the purposes of this analysis, we decided not to take on too many variables at once. Later in this report we discuss potential further areas of research and potential modifications to our procedure that incorporate these data.

2. Data Preparation

This step was focused on making sure the data is as "clean" as possible by going through an extensive auditing process.As mentioned in the data integration section, our data had some instances in which certain data values could not be programmatically added to our dataset such as when we were encoding Uber's location ID with a more standard PUMA. Hand encoding some values required some minor auditing to make sure that the dataset had integrity. We sampled our categorical and nominal value types to make sure there were the expected number of unique values for their fields. Similarly, we did some quick sampling of our numerical categories to ensure there were no

strange values or outliers that we needed to take care of. Fortunately the data did not have many encoding errors and the data types were all correct. In the Uber dataset specifically, we ran into some missing values for time and dispatch base but did not handle the data point any differently since they were not essential to the analysis. If we encountered data with no location ID then we would have had to drop it but we did not.

3. Data Storage

Since our analysis is based off of historical data, we were able to utilize an OLAP model of data storage. These data were not multidimensional in the capacity that we utilized them for and thus putting them into a DBMS store was not necessary. The processing of these millions of data points was very slow however and in retrospect, a batch processing system such as Hadoop or Apache Spark that read in our data from a non-relational store would have been a much better idea. Instead, we created a separate aggregate table using the Python library Pandas which was very inefficient for the amount of data we were dealing with.

6. EVALUATION METRICS

To evaluate the results of our data mining on the Uber data set, our team plans to begin our mining on smaller portions of the data set. Since we are creating a predictive model, we can compare our predicted user behavior with actual user behavior that occurs later in our data set. Using this technique, our group will be performing cross-validation to evaluate how accurate our predictive model can get. We could see if our user behavior predictions grow in accuracy as the size of our data set being mined increases. This would tell us that our predictive model is accurate and would become more accurate with an increased amount of user data.

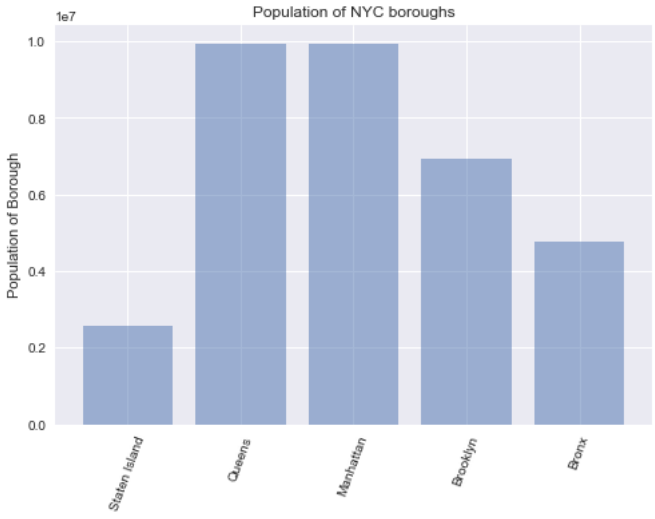
4. Analysis

Following the complete integration, preparation, and storage of our data. We leveraged the use of Python's Pandas library to handle the aggregated data for each PUMA and compare it with their respective demographic data. Using data-frames we were able to compose graphs of various relationships and present them in graph form using Python's matplotlib. We did not have a preconceived hypothesis that we wanted to test and instead focused on representing the relationships we found and use them as a basis for forming a hypothesis in any additional research in the future.

7. RESULTS

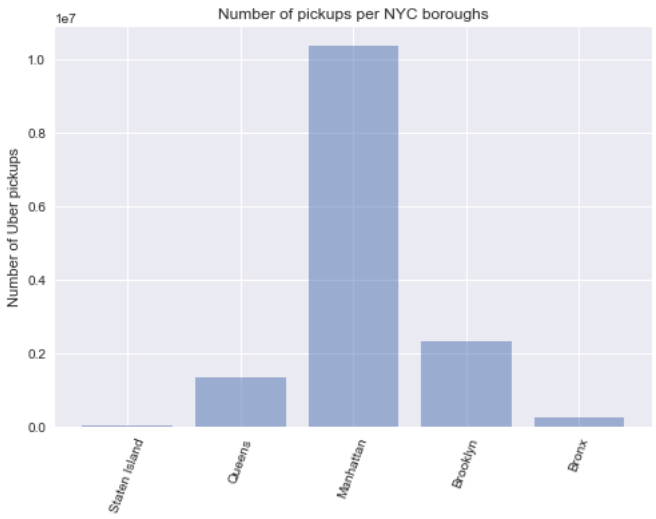
Our findings were generated in two phases, one looking at the relationships of boroughs as a whole and the other looking at relationships within each borough. We wanted to see the relationships at both a macro and micro level since we had all of the data available to us to facilitate this.

We started by looking at basic information such as the population of each of the boroughs. It was important to do this as it provides important contextual information for any further findings, especially as they relate to pickup frequencies. The populations can be seen in figure 1 below:



From this we can see that Manhattan and Queens feature the highest populations while Staten Island is the least populated. Next, looking at the number of pickups per borough (Figure 2), we see a very interesting result. From these

data, it's clear that Manhattan is by far the most popular area for Uber pickups with just over a million pickups recorded. This is in stark contrast to the amount of pickups in Queens, which, while having similar populations, is dwarfed by Manhattan. This is to be expected as Manhattan is the central point of commerce in New York and is home to some of the most widely known and recognized areas of New York City.



Additionally we would surmise that an area with that much commerce should have more people taking Uber rides as they can afford it compared to other systems of transportation. If we use average median income of the boroughs as show in Figure 3, we can see that this does hold true for Manhattan but interestingly does not for Staten Island. With the second highest median income, it has a minuscule amount of Uber pickups. This relationship indicates some underlying factor for the Uber pickups that is not easily explainable by the particular data we gathered.

8. LIMITATIONS

In the previous section we found many interesting relationships, mtny of which seemed easily explainable and others, not so much. The nature of this project was to begin to use a data centric approach to finding the relationships of Uber pickup frequency and socioeconomic factors. The results we found can not and should not be used to draw definitive conclusions as there are several major limitations with our analysis based on several assumptions. One assumption that these data make is that people getting an Uber in a location are a representative sample of people in that location when this is not necessarily true as commuting is very popular. We also assume that the effect caused by that would be mitigated by the fact that most Uber rides are not very far. This assumption is not backed up by any evidence outside of our anecdotal evidence and thus makes conclusions about transient Uber riders not very definitive. The basis for this assumption is also borne from the fact that it's generally more cost efficient for the average commuter to use mass public transit for longer distances e.g. between boroughs in the morning rather than going out for lunch while at work.

The time frame of data looked at also presented some issues with regards to drawing conclusions at a macro, societal level. There simply isn't enough longevity in these data to account for seasonal fluctuations, major events, etc. Ideally, these data would be comprised of many years worth with at least 3-5 years before drawing any sort of definitive conclusions.

Spatially speaking, these data do not explain certain characteristics of the pickups frequency that are a result of geographic factors. The prime example is one we alluded too in the previous section of Staten Island. The reason for the low amount of Uber pickups is because of a geographic constraint, namely, it's an island and a small one too. The ferry is the most common method of transportation to and from but that does not show up in our data. While this may be a trivially obvious example, it serves as a reminder to be cautious about making inferences from these data in isolation. These studies must be done holistically and with a lot of context when interpreting the results.

9. APPLICATIONS

While our exact findings do not and should not be the cause for any sweeping policy change from Uber, it does present an interesting method-

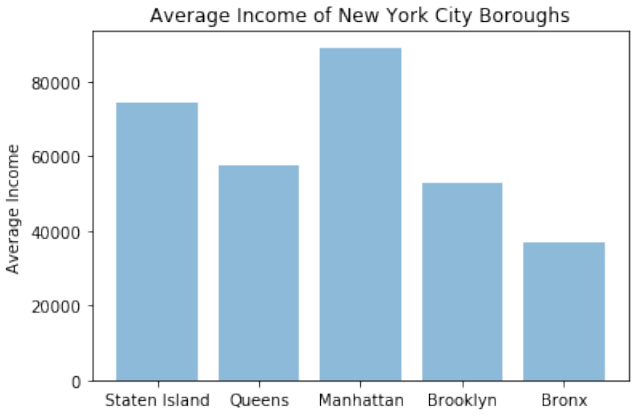
ology for examining behavior trends in Uber pickups. Companies such as Uber and Lyft could use their presumably large wealth of tracking information combined with the markets they operate in and find patterns of where and when their rides are most taken which could influence pricing, surge pricing and timing, marketing, hiring, etc. With recent controversies in ride sharing driver ethics, these companies can try to find causal factors that can better inform their response to these situations when they arise e.g. is surge pricing disproportionately affecting ridership in a community?

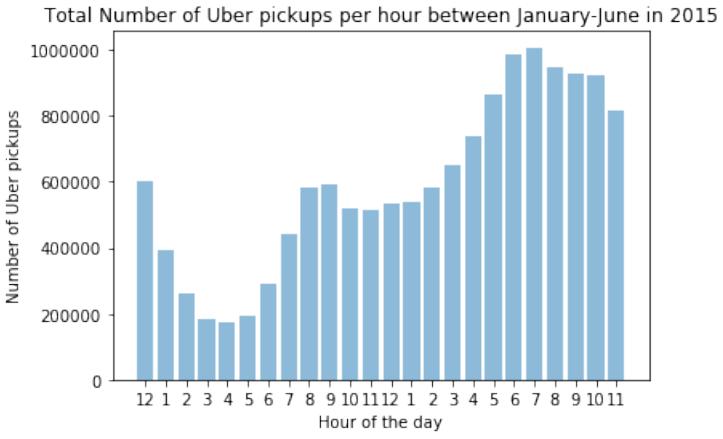
Traditional transit companies such as Taxis and mass public transit systems can use similar methodologies to examine their rider profiles and find new opportunities such as competing with Uber and Lyft, expansion projects, renovations, route adjustments, etc.

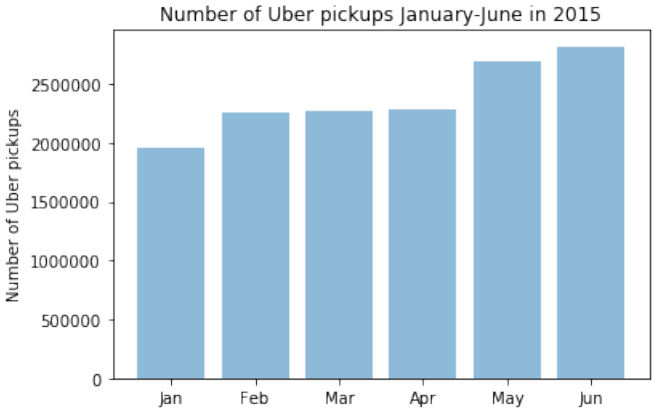
10. VISUALIZATION

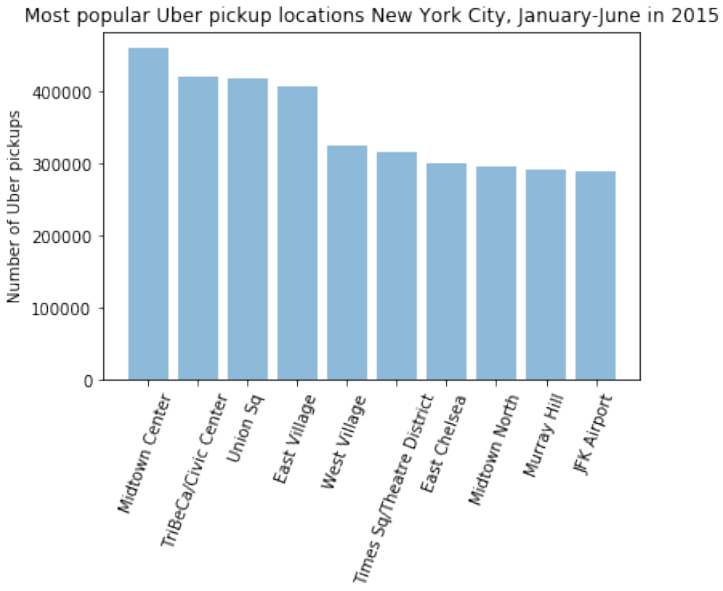
The geological nature of our investigation lent itself to having a visualization of our dataset that would allow us and a viewer to see the overarching trend in Uber pickups and include additional data that one could look at to enhance their understanding of what may be causing the visualization being presented. To do this, we utilized the D3 javascript library to display a map of New York City with subsections that outline the PUMAs we were looking at in our data. Each PUMA was filled with color corresponding the frequency of Uber pickups in that PUMA. Darker colors corresponded to higher amount of total pickups and vice versa for lighter colors. Additionally, hovering over these areas gives a tooltip that presents some basic information such as the colloquial name of the PUMA, total number of pickups, and median income of the area. The intent of this visualization is to show the data we were looking at while also allowing viewers to scan through the data and find interesting patterns for themselves or prompt a further area of research.

The live look at our D3 visualization can be found here:









11. REFERENCES

[1] Demographic, social, economic, and housing profiles by community district/puma. `https://data.cityofnewyork.us/City-Government/Demographic-Social-Economic-and-Housing-Profiles-b/kvuc-fg9b`, 2014.

[2] Uber pickups in nyc. `https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city`, 2015.