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. PROBLEM STATEMENT

We want to explore an Uber dataset that outlines Uber pickups in New York City. This dataset encompasses location, time, dispatcher, and date. We want to look at trends over time of spacial and temporal uber ride distribution in New York City. If possible, we also want to integrate with other demographic information to get more nuanced insights.

2. LITERATURE SURVEY

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 ridesharing services as it pertains to competing with more traditional forms of transportation.

3. PROPOSED WORK

There are several steps that must be completed in order to create a comprehensive analysis of Uber pickups with respect to demographic information. These steps include:

1. Cleaning: With 18.8 million data points, there is a large possibility of a "dirty" data being present and so a iterative auditing process is necessary to ensure that the dataset contains proper encoding, parsing, and accuracy prior to working with it.

2. Integration: The Uber pickup data examined in isolation will not be sufficient to answer our questions as they pertain to gaining an understanding of various demographic factors' influence on the distribution of pickups in New York. Similar to previous works in this area, we need to integrate with some other data sets or bin existing data in the dataset to encode coordinates at the county or neighborhood level. This step is crucial for then examining the distribution of rides with respect to several known demographic information about these local areas within New York City.

3. Preprocessing: As previously mentioned, the integration process will likely lead to some much needed preprocessing in order to have an effective dataset to work with. re-encoding location or adding an additional column for neighborhood/county will need to be procedurally added in such a way that the consistency and accuracy of our data is retained.

4. Visualization: Once the data has been mined/gathere
cleansed, and fully integrated, we can then start

to generate some early hypotheses about the data that we glean from exploratory data analysis. We will use then use D3 to create an interactive visualization that walks the reader through understanding the insights we found in our analysis.

4. DATA SET

Our data set comes from a Kaggle Dataset [1] we found. This dataset was previously created and used by fivethirtyeight. This dataset contains 18.5 million data points. This dataset contains part of all we need to know about Uber rides in NYC, and for the demographic information weâĂŹd like to incorporate we will be using geographic data we find in the United States Census Bureau. If we are able to find anything useful from this data set, we have the potential to identify interesting characteristics for the use of Uber's versus socio-economic information.

5. 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.

6. TOOLS

For our data mining project, our group plans to utilize several tools. We will use available data mining tools to originally set and do any initial cleaning of the dataset that needs to be done. Mainly Python packages such as Pandas. We would most likely use MongoDB to store our data for easy access. This may be subject to change if we decide to use other storage. Python will be our main programming language as it is great for data computation and analysis. Python also has plenty of useful packages that can be used to simplify analysis. Numpy, ScyPi, and Pandas provide functions to help with numerical computation. scikit-learn provides easy to use tools for data mining and analysis. Additionally R will be used for statistical analysis. For the visualization of the insights gathered, we will utilize D3 to create an interactive author driven narrative that takes the user through our analysis process.

7. MILESTONES

We have several milestones for which we can track our progress:

1. Data Integration: Here the goal is to find and collect all of our data into a singular store from various sources outlined previously. Since there is no clear indication from our problem requirements of the specific data store to use, we are choosing a simple NoSQL store MongoDB although any SQL store would probably work just as well.

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 demographic information available. We ultimately had to select one standard for defining neighborhoods that would capture the most information possible. This was not a perfect fit and thus some manual decisions needed to be made regarding some location IDs in the Uber data that did not fit perfectly into a predefined region for which we had data for. 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 so-

cioeconomic statistics. This makes sense since airports are distinct, transient places that share little with the surrounding community. We are still not exactly sure how to treat these locations (LaGuardia, JFK, and Newark) and may need 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.

Ultimately our final 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 progress report, we decided to not add too much data at once.

2. Data Preparation: This goal is focused on making sure the data is as "clean" as possible by going through an extensive auditing process outlined previously.

As mentioned in the data integration section, our data had some instances in which certain data values could not be programmtically added to our dataset such as when we were encoding Uber's location ID with a more standard PUMA ID (Public Use Microdata Area). 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. Had we been working with messier data, there might have been more to do here.

3. Analysis: Our main methods of analysis will be looking at how ridership varies with different demographics variables as well as location based data. In this step, we might also generate a linear regression model to further examine relationships between various continuous predictor variables that are going to be present in the dataset.

Having completed the integration of all the data we need for analysis, we started to make some exploratory graphs using Python's matplotlib library that are located in the following results section. Since we do not have a clear hypothesis we are testing at this point, we wanted to see what data we get from some simplistic queries and then hone in on some aspects that might be particularly interesting. There is also the possibility that some of our exploratory analysis might prompt us to gather more detailed data to create a better look some trends.

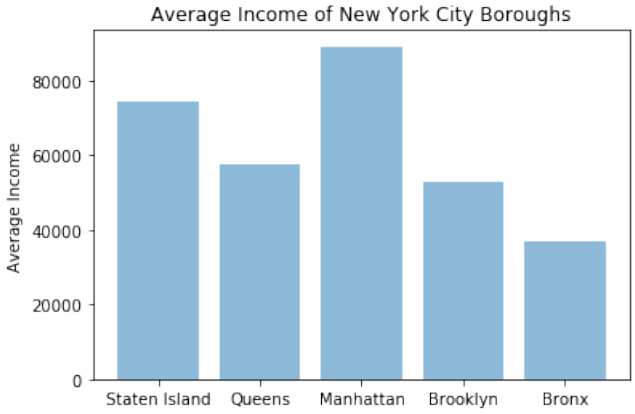
4. Visualization: The goal for this to have an effective interactive visual representation of our hypotheses that clearly walks the reader through our thought process in generating and validating them.

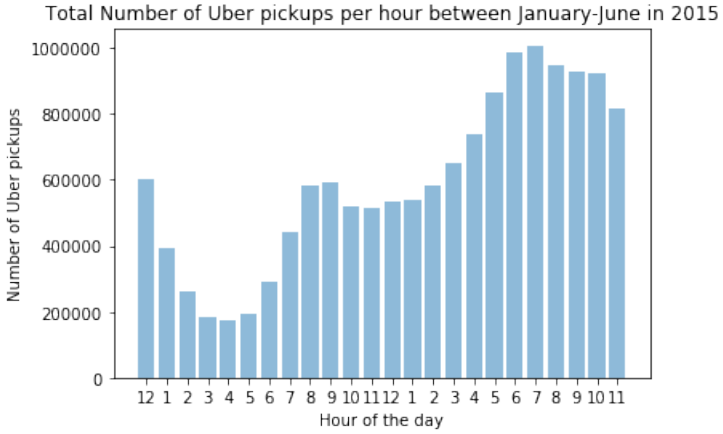
We have not started on the visualization aspect of this project although we have outlined what our vision of this is going to be. Using a one or more of the insights extracted in the analysis portion, we want to create a fully interactive heatmap of Uber pickups on top of a map of New York that has the neighborhood boundaries marked as well. This visualization will have both author and reader driven elements that serve to highlight some of our find-

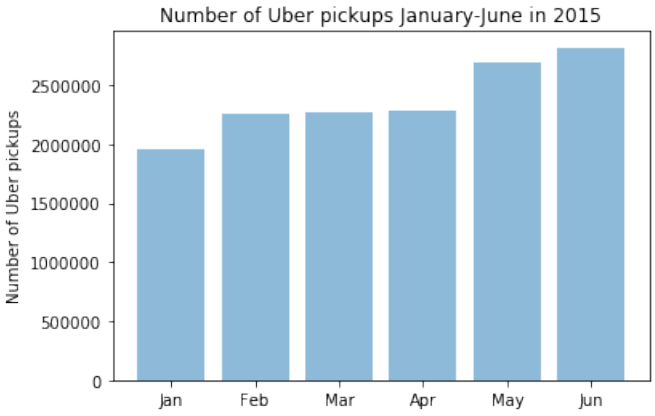
ings but also allow a reader to explore and find their own patterns and discoveries that we may have missed. This visualization will be constructed in D3 with a light front-end display. Our team does not have a ton of D3 experience so some adjustments to the final visual product may change in the future. Some difficulties we potentially foresee are the mapping of locations to an SVG of New York with neighborhoods that may not exist yet in which case we might need to make one.

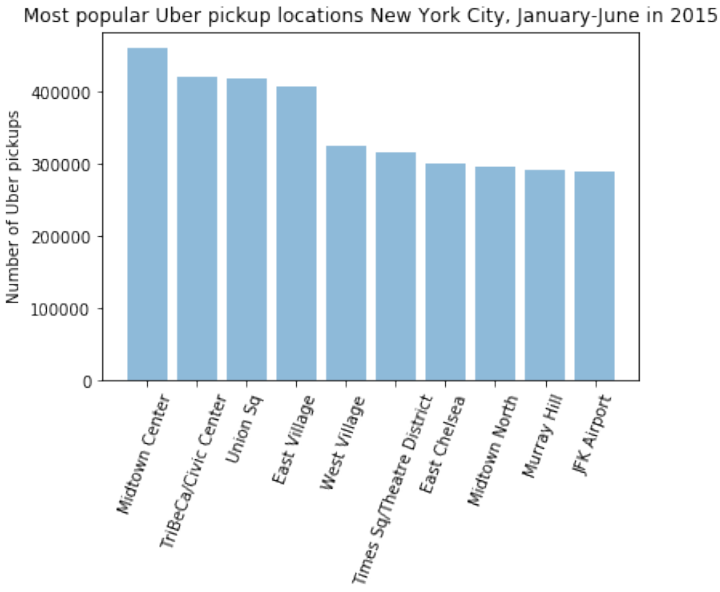
8. RESULTS

As mentioned before, we have begun our exploratory data analysis so our results are limited to some overarching questions and descriptions that we are using as a base for further analysis and/or discover areas where we might need or want additional data. The following is a look at some of our early stage graphs on our newly created dataset:









9. REFERENCES

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