Exploration of Uber Pickups in NYC July 2014

Diego Avila

diego.avila.calderon@protonmail.com

July 7, 2020

Table of Contents

I. Introduction

A. Background

Uber is a ride-hailing company that uses location services on phones to track customers and assign a driver in their vicinity . The app relies on location data like geo-coordinates to optimize the speediness of pickups and improve their services. The Uber app permits users to submit a trip request through their phones which is automatically sent to a driver in that user's vicinity.

Uber provides the platform through which customers and drivers can connect. Uber drivers are classified as independent contractors meaning that they are non-employees and they run their own business. Some people use Uber as a part-time job; driving whenever they have time just to earn some extra money.

B. Business Problem

Allocation of resources is an important method through which a business can help minimize inefficiency and cut costs. When resources are not being used or being misused, opportunities are being lost in the process. The internet, new technologies and data science have given us the tools necessary to better our resource allocation and reduce inefficiency.

In a ride-hailing service, idle-time (time with no customer i.e. time searching for a customer) represents a misuse of resources. In the past, a driver's sense of intuition and luck was the best method to secure a passenger. Nowadays, technology and data from technology helps reduce this inefficiency.

Although plenty of data already exists, drivers don't always have this data readily available. This study hopes to allow a driver to know near what locations he/she can expect to find customers and also what days, and during what hours are more lucrative to drive. The exploration will serve give a better idea of where there is demand for drivers and how they can best allocate their time.

II. Methodology

A. Data sources

The dataset that will be used consists of neighborhood data of New York, Foursquare location data and Uber trip data from July 2014. The Uber trip data, that can be found on Kaggle, contains location information (geocoordinates), date and time of pickup. FiveThirtyEight obtained the data from the NYC Taxi and Limousine Commision (TLC) by submitting a Freedom of Information Law request on July 20,2015.

Furthermore, two Geojson files are of relevance in this report. One is a file that has the names, the borough that they belong to, and location info of all the neighborhoods in NYC. The other geojson file consists of boundary limits of the zip codes within NYC. This second file will be useful for the creation of a map.

Using the Foursquare API, location information will be used to gather information about venues near that location. Among other information, the category of the venues will be of relevance to this exploration

B. Data Cleaning

The Uber data from July consisted of 796,121 entries and 4 features namely Date/Time, Latitude coordinate, Longitude coordinate and Base. During the downloading of the file, the Date/Time information is read as a string. Therefore this info had to be formatted into datetime type. After the conversion, features from this column were extracted. The day of the week (Mon, Tue, Wed, etc), the time of day and the day of the month were all separated and put into a new feature of the dataframe. This is meant to facilitate some exploratory data analysis.

A choropleth map can help visualize the distribution of data. In order to visualize the distribution of the pickups, the location coordinates of the pickups could be traced back to a geographical delimitation like postal code or neighborhood and this could then be mapped. However, considering there are almost 800,000 entries, a random sample from the data consisting of

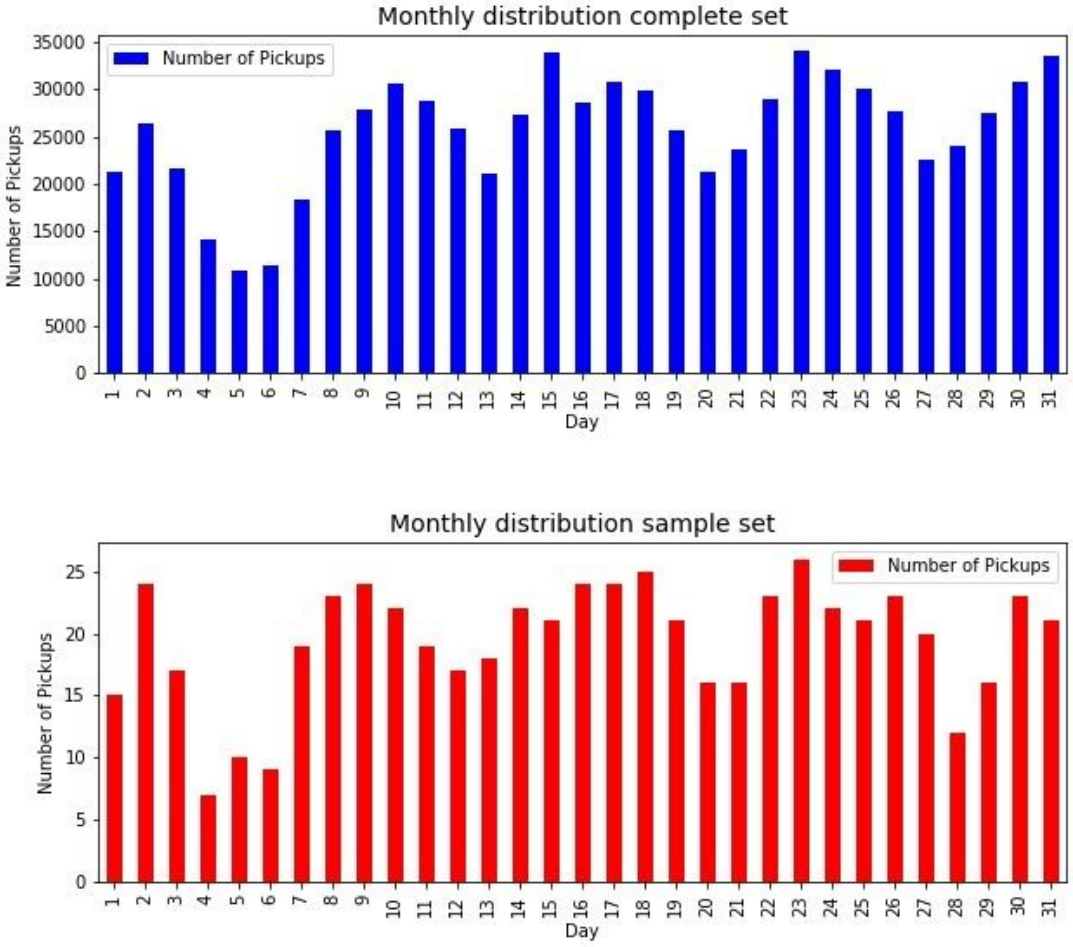
about 600 entries were taken and plotted. The size of this set was determined considering the time it would take to retrieve zip code and other delimitation information using geopy. A maximum limit of 10 minutes of calls with geopy was set.

C. Methods of Analysis

To get a general sense of the pickup entries, date time properties were mapped out by their frequency. This provided a visual representation of what days of the week, during what hours and what days of the month are busier for pickups.

Since a sample set was extracted for the creation of a Choropleth map, the representability of this set was tested through the similarity of the Monthly distribution of Pickups. The following diagram shows that the two datasets are not identical. However, considering that the sample set consisted of .07% of the total data it is important to note that the two sets are fairly comparable in their distribution.

There



There are many neighborhoods within NYC and the area in which you are located can determine the type of uber user that you are going to pickup. Through the exploration of the similarity between neighborhoods, it is possible to aggregate the information and make general judgements about the types of customers, the important venues in that neighborhood, etc. This grouping of neighborhoods was accomplished through a method of unsupervised machine learning known as K-means clustering.

K-means clustering is an unsupervised machine learning algorithm and a method of classifying data into groups known as clusters. The way in which this is achieved consists of three main steps.

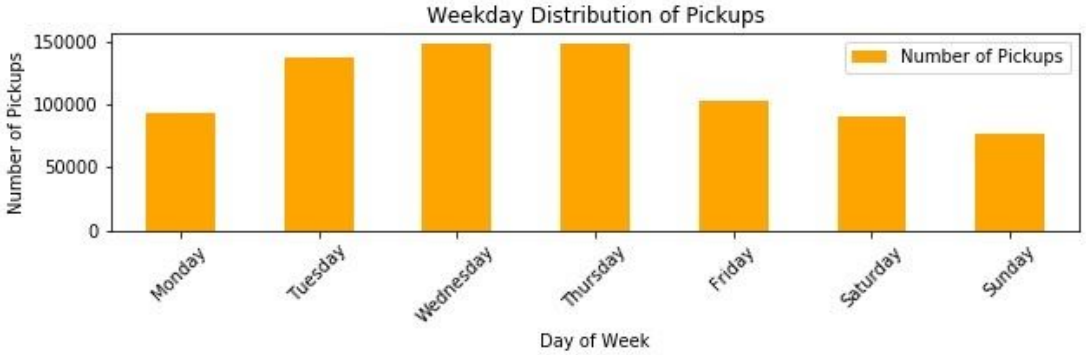
The first step is to determine where the centroid of a cluster lies and how many clusters we want to make. The number of clusters refers to the number of groups of data with similar characteristics. The centroids can be chosen randomly since the algorithm will eventually correct this location. Once the center is determined, the next step is to calculate the distance of the data points to each of the centroids and assign the data point to the nearest centroid. After all data points are assigned, the algorithm states that the centroid should be moved to the center of it’s assigned data points. With new centroid coordinates, the calculation of the distance to the new center is calculated and the process is repeated until the average is fixed meaning that the centroid doesn't move therefore fixing the clusters.

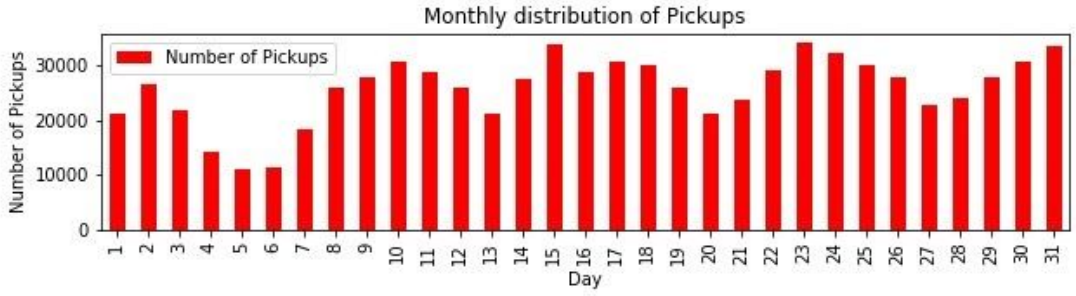
A critical choice in K-means is determining the amount of clusters to use. The elbow method is a widely accepted method of evaluating the correct amount of clusters. In this method the squared mean of the distance of each point to its center is calculated. The elbow is chosen since it is considered the point where diminishing returns are no longer worth the additional cost. Although the elbow method was applied, the resulting graph was a bit inconclusive and a cluster amount of 5 was chosen deliberately.

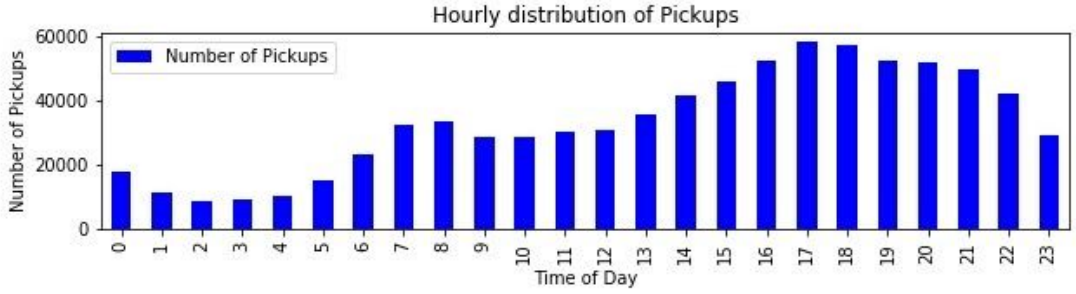
III. Results

A. Pickup Distribution

The entries of all Uber pickups during July of 2020 were graphed by their pickup datetime features. The resulting distributions are showcased in the following diagrams.

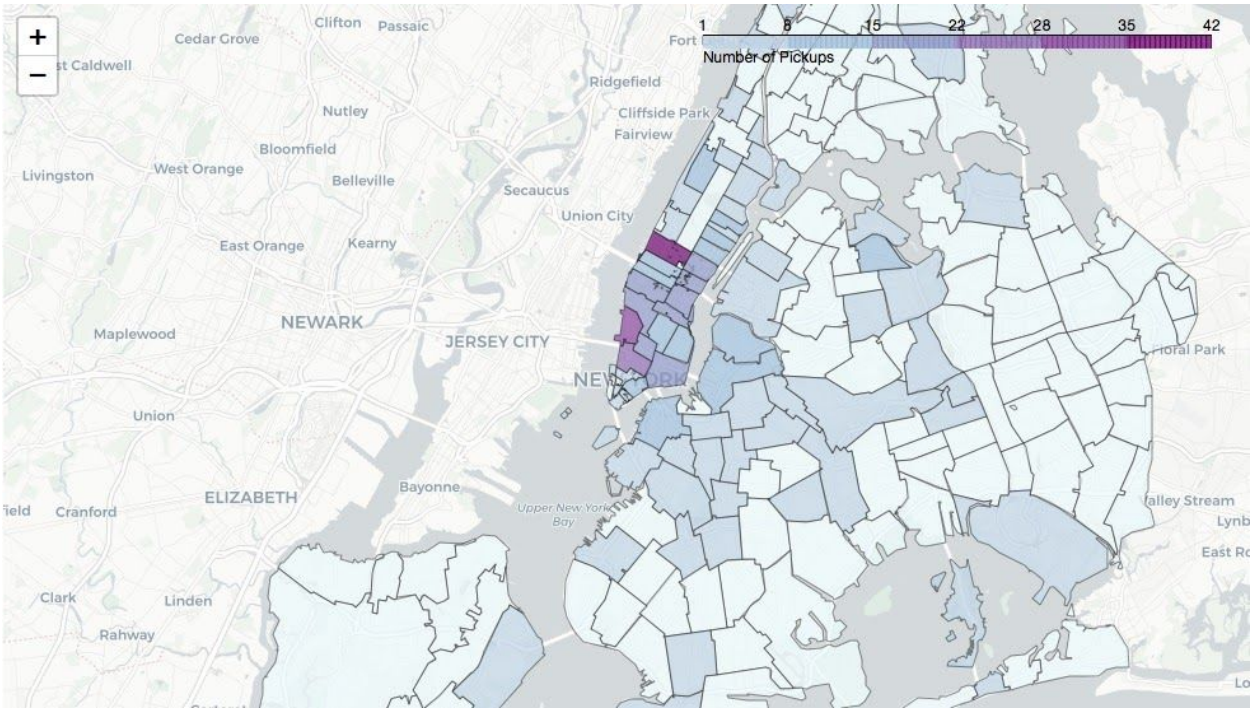




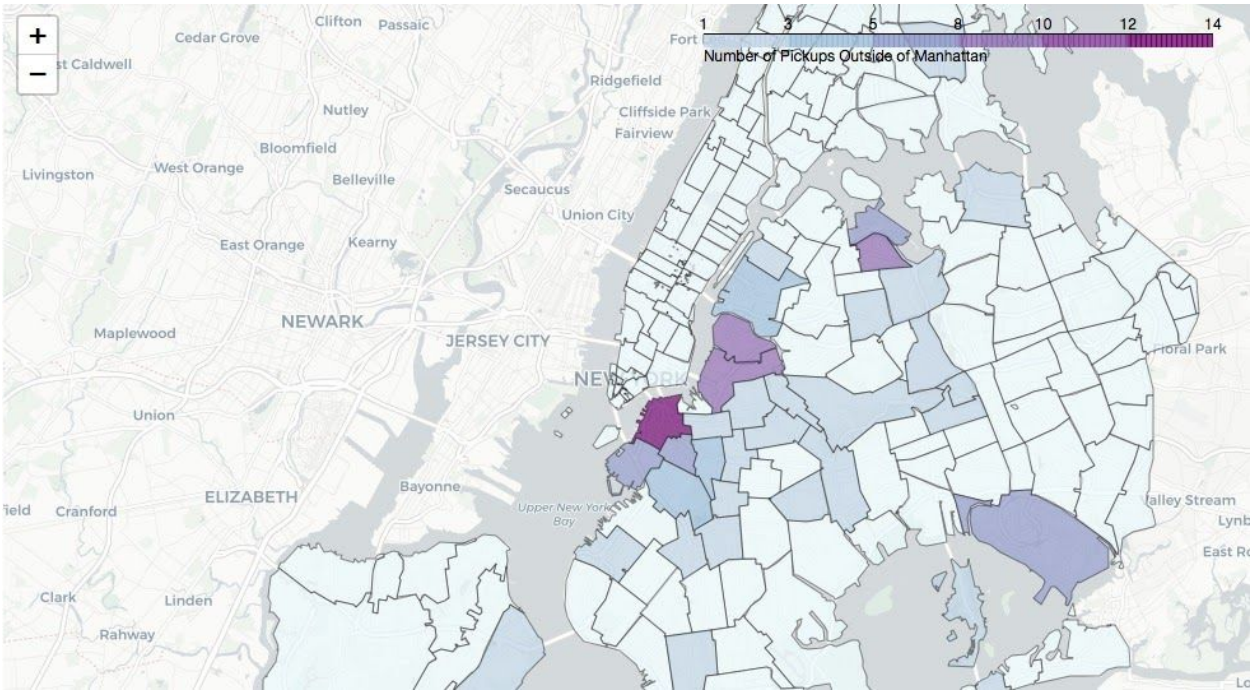


B. Location Distribution

From the sample set, the distribution of pickups were plotted by zipcode in Choropleth maps using a geojson file. Two different maps were generated found in the figures below.

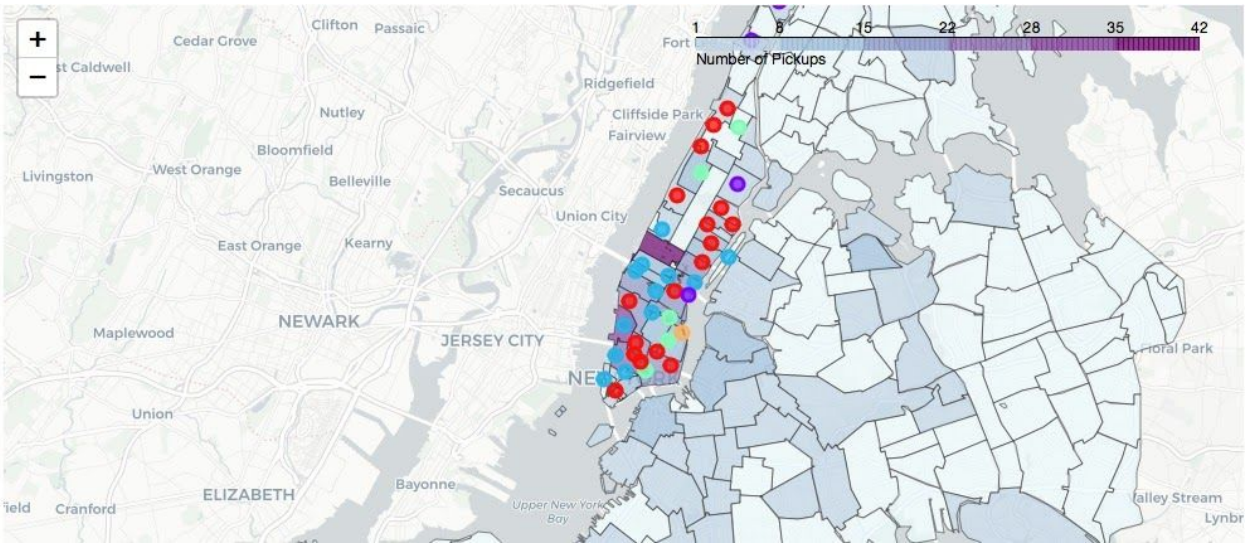


Choropleth with Manhattan



Choropleth without Manhattan

C. Clustering with K.Means



Clustered Neighborhoods in Manhattan Visual (blue: Cluster 0, Red: Cluster 2)

```
manhattan_venues_sorted[manhattan_venues_sorted['Cluster Labels']==0]
```

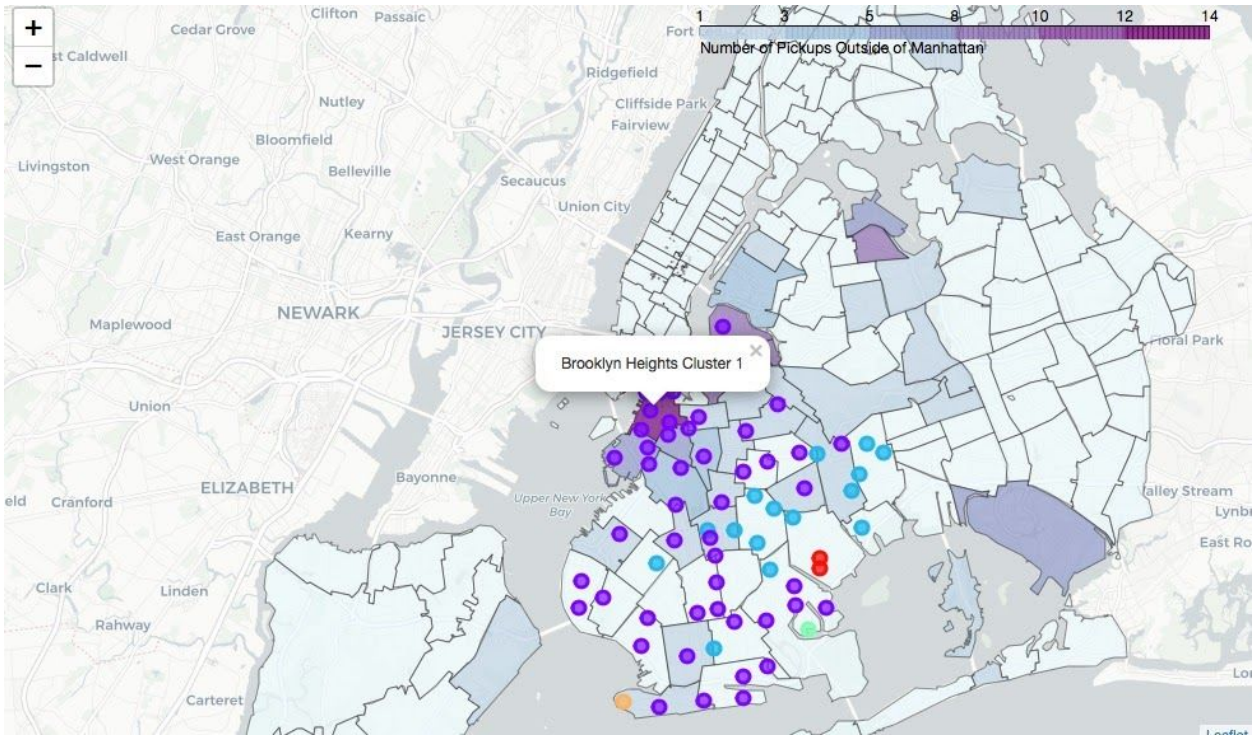
	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Carnegie Hill	0	Coffee Shop	Café	Italian Restaurant	Pizza Place	Yoga Studio	Gym	Bar	Bakery	Gym / Fitness Center	Bookstore
3	Chelsea	0	Coffee Shop	Art Gallery	American Restaurant	French Restaurant	Bakery	Ice Cream Shop	Bookstore	Hotel	Bar	Café
9	Financial District	0	Coffee Shop	Cocktail Bar	Bar	American Restaurant	Pizza Place	Gym	Gym / Fitness Center	Italian Restaurant	Wine Shop	Mexican Restaurant
12	Greenwich Village	0	Italian Restaurant	Sushi Restaurant	Café	Clothing Store	Indian Restaurant	Chinese Restaurant	Ice Cream Shop	Coffee Shop	Bubble Tea Shop	Burger Joint
13	Hamilton Heights	0	Pizza Place	Café	Coffee Shop	Deli / Bodega	Mexican Restaurant	Bakery	Caribbean Restaurant	Chinese Restaurant	School	Sushi Restaurant
16	Lenox Hill	0	Italian Restaurant	Coffee Shop	Sushi Restaurant	Pizza Place	Café	Cocktail Bar	Gym / Fitness Center	Burger Joint	Gym	Deli / Bodega

Example Cluster 0 Manhattan

```
manhattan_venues_sorted[manhattan_venues_sorted['Cluster Labels']==2].head()
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	2	Park	Hotel	Coffee Shop	Gym	Memorial Site	Boat or Ferry	Italian Restaurant	BBQ Joint	Gourmet Shop	Sandwich Place
5	Civic Center	2	Coffee Shop	Hotel	Gym / Fitness Center	Cocktail Bar	French Restaurant	Yoga Studio	American Restaurant	Spa	Park	Italian Restaurant
6	Clinton	2	Theater	Italian Restaurant	Gym / Fitness Center	Wine Shop	American Restaurant	Coffee Shop	Cocktail Bar	Hotel	Sandwich Place	Gym
10	Flatiron	2	Gym / Fitness Center	Vegetarian / Vegan Restaurant	Cycle Studio	Mediterranean Restaurant	Italian Restaurant	New American Restaurant	Yoga Studio	Furniture / Home Store	Spa	Café
14	Hudson Yards	2	American Restaurant	Gym / Fitness Center	Hotel	Italian Restaurant	Café	Coffee Shop	Restaurant	Gym	Dog Run	Park

Example Cluster 2 Manhattan



Clustered Neighborhoods in Manhattan Visual

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
12	Canarsie	0	Deli / Bodega	Caribbean Restaurant	Bus Line	Food	Gym	Asian Restaurant	Filipino Restaurant	Farm	Farmers Market	Fast Food Restaurant
52	Paerdegat Basin	0	Gym	Child Care Service	Bus Line	Food	Asian Restaurant	Filipino Restaurant	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant

Example Cluster 0 Brooklyn

```
brooklyn_venues_sorted[brooklyn_venues_sorted['Cluster Labels']==1]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bath Beach	1	Pharmacy	Bubble Tea Shop	Pizza Place	Donut Shop	Fast Food Restaurant	Italian Restaurant	Gas Station	Chinese Restaurant	Clothing Store	Coffee Shop
1	Bay Ridge	1	Pizza Place	Italian Restaurant	Spa	American Restaurant	Bar	Bagel Shop	Greek Restaurant	Ice Cream Shop	Sandwich Place	Pharmacy
2	Bedford Stuyvesant	1	Coffee Shop	Deli / Bodega	Café	Pizza Place	Bar	Cocktail Bar	Bus Station	Bus Stop	Boutique	Fried Chicken Joint
3	Bensonhurst	1	Italian Restaurant	Chinese Restaurant	Ice Cream Shop	Donut Shop	Park	Sushi Restaurant	Hotpot Restaurant	Sporting Goods Shop	Smoke Shop	Shabu-Shabu Restaurant
4	Bergen Beach	1	Harbor / Marina	Playground	Hockey Field	Baseball Field	Athletics & Sports	Women's Store	Farmers Market	Event Service	Event Space	Factory
5	Boerum Hill	1	Coffee Shop	Dance Studio	Bar	Bakery	Sandwich Place	Furniture / Home Store	Arts & Crafts Store	French Restaurant	Kids Store	Cocktail Bar

Example Cluster 1 Brooklyn

IV. Discussion

A. Pickup Distribution

The bar graphs show us the amount of customer traffic that Uber drivers can expect to receive on a given weekday, hour of day and day of month in July. Wednesday and Thursdays are days of larger demand while weekends (Sat,Sun) and Mondays witness the smallest amount of pickups. This implies that a large portion of the use of Uber comes from work-related movements. Regular work days are significantly busier than non-work days.

When looking at the monthly distribution of pickups it is interesting that the days with least amount of pickups are July 4 and July 5. Seeing as though the 4th is a national holiday, we can expect that there is less movement around the city since more people are at home. The 5th of July had a comparable amount of pickups to the 4th. This is probably due to the fact that it fell on a Saturday and as we saw in the first diagram, weekend days experience relatively smaller

amounts of demand for pickups. Here we can visualize the monthly trend by seeing the way in which the bars peek. The curve is very sinusoidal, experiencing troughs during the weekends (days (5 and 6)+7n) and peaks during workdays.

Lastly, the time of day bar graph demonstrates that rush hour is an important time for pickups. In the morning, between 7 and 8 there are many requests for Uber drivers which steadies out until around 2pm when it starts increasing again. There is a lot more activity in the afternoon, maxing out at 6pm, but still relatively high until 11pm

B. Location Distribution

The two choropleth maps show the areas in which most uber pickups were made. As we can observe from the first figure, an overwhelming amount of pickups occur in Manhattan. This was as expected since it is a big hub for companies and tourists alike. The second choropleth map allowed us to better appreciate what other postal codes are highly requested. The map shows that after Manhattan, Brooklyn receives the most uber pickups along with other neighborhoods scattered around. This implies that if someone doesn't feel in the mood of going to Manhattan for a pickup, their best bet is to go to Brooklyn. However, still to stay in the near vicinity of the bridges. This, however, does not imply that the service itself won't lead the driver into Brooklyn after the pickup, which is very likely to happen.

C. Clustering

After Clustering the neighborhoods in Manhattan, we can observe that a large majority was clustered into two clusters: (two colors visible on map) blue and red . Both clusters have similar properties being that among the top 10 venues, coffee shops are found. The busiest neighborhoods were mainly classified as cluster 2 (blue). These are fairly touristic areas like the West Village where Hotels and fitness centers are common. In comparison to cluster 0, this area is less crowded with food venues.

In Brooklyn, the big majority of neighborhoods were clustered into a single cluster, cluster 1. Cluster 1 included Neighborhoods like Brooklyn Heights and Fulton Ferry which are hotspots for pickups relative to the rest of New York excluding Manhattan. THis cluster, similar to cluster 0 in Manhattan is characterized by the large amount of food-related venues.

V. Conclusion

The clearest take away from the data is that an overwhelming amount of pickups are concentrated in Manhattan. Within Manhattan, there are two main clusters that comprise the majority of the borough: Clusters 0 and 2. Cluster 0 can be characterized as neighborhoods that have coffee shops, pizza places and other restaurants. These are busy neighborhoods with lots of options for food. Neighborhoods classified as Cluster 0 include Little Italy, Greenwich Village and the Financial District. Cluster 2, on the other hand, seems like a more local-friendly cluster, where parks, fitness centers are pretty common. Yet, still very touristic since it includes lots of Hotels and entertainment related venues like Theaters. This Cluster, among other neighborhoods, includes Lincoln Square, Midtown and Flatiron.

Considering a driver is seeking to optimize their work hours, the recommendation would be to work during weekdays, since weekends exhibit less activity and to work between the hours of 4 to 9 pm. Considering a person might not want to work in Manhattan, due to the busy streets and traffic, the best recommendation would be to stay near Brooklyn. Here, places with more food locals than parks are recommended.

In order to be more accurate with the hourly distribution it would have been more reasonable to create two separate charts- one showing the hourly distribution during weekdays and the other during weekends. This would have given a better idea of the distribution on the weekend since this data was hidden because of there being more weekdays than weekend days.

It is important to note that the dataset does not constitute a fixed representation or result. Although trends like weekly distribution of pickups is highly likely to stay the same, markets are dynamic; changing every day and the data for a single month doesn't determine that this distribution will stay that way.

The study provided interesting insight into the world of an uber driver in New York City. This study can definitely be useful for a driver that takes on the job part-time since the driver can plan out what hours and what areas they want to work in. Considering that some uber drivers are

only part-time workers, picking people up as they commute somewhere else, the drivers can plan their routes so as to ensure that a pickup will take place.