Identifying Ride Service Hotspots in New York City

Prince Abunku
CUSP, New York University
New York, USA
pa1303@nyu.edu

Isha Chaturvedi
CUSP, New York University
New York, USA
ic1018@nyu.edu

Gaurav Bhardwaj
CUSP. New York University
New York, USA
gb1877@nyu.edu

***Abstract*—**

With the increase in urban population, city agencies have been struggling to manage their resources efficiently. In most of the metropolitan cities, number of people who use public transportation has increased drastically and city agencies don't have enough resources to suffice every need. This paper addresses one urban issue that has been a major challenge for most of the cities in the world. Since more and more people from the metropolitan cities are relying shared and solo cab services, we wanted to identify the areas that have the maximum number of trips either by Uber or the yellow and green cabs in NYC. By analyzing these hotspots, we can give recommendations for the optimization of trip routes, especially in the case of shared cab services. We also analyzed the ridership patterns and how weather can affect user preferences towards their commute options.

Keywords*— *Analytics, Uber, taxi, location intelligence, TLC, weather effect, ridership patterns, hot-spot analysis

I. Introduction

Today, 55% of the world's population lives in urban areas, a proportion that is expected to increase to 68% by 2050 [1]. With this increase in population, the resources allocated for by the government have gone for a toss and it has become a chaotic scenario. There are more people in a city than the infrastructure can handle and therefore traffic congestions and overloaded subways are common sights in most of the metropolitan cities. New York City alone is home to 8.6 Million [2] people with mere 305 square miles area. With this we can estimate how densely populated NYC is and how difficult it would for city agencies to maintain and optimize the resources efficiently.

Public Transit is one of the key problems that New Yorkers complaint about the most since the roads are not wide enough and have no scope of extension. At the same time the subway system is old and faces technical issues on regular basis causing further problems for the commuters. To avoid this more and more residents of NYC are opting for economical alternative of using shared cab services. Firms like Uber, Lyft and Via have been providing such services where you can book a seat in the cab and pay a fixed amount depending on your pickup and drop locations. Although this is great for people, it has increased the number of cars on the roads and at the same time has decreased the popularity of yellow and green cab services in the city [3].

But the story doesn't end there. We know that the cab services have been trending upwards due to lack of good public transit system, but at the same time these shared trips are not optimized. In many instances, people have complained about spending long hours in the cabs since they were the first ones to get picked and last ones to be dropped off. This has caused serious backfire to these private cab firms. Therefore, it has become absolutely necessary for these firms to develop algorithms so that each person spends acceptable amount of time in the commute.

II. Motivation

The factors discussed above has given rise to a high order problem in the urban ecosystem. We need a mode of commute that is less painful and exhausting and at the same time we need to conserve our resources think about the environment. Bringing more and more cars on the roads is increasing the pollution level in every city and also creating traffic congestions. On average, a typical New Yorkers spends 36 minutes on one-way commute to work using subway and spends approximately 90 hours per year in traffic congestion [4]. This is the highest number in the United States and this shows the plight the New Yorkers have to go through every day. Another reason why people prefer to drive their own car over the shared ride service is that they take longer than personal car to reach the destination.

Ridesharing gained its popularity because of affordable smartphones and rapid development of GPS technology. The user just has to open the mobile app and request a ride to the service provider and the service provider in turn contacts nearest cab driver and assigns them the task. It also allows the customer or the driver to contact each other without giving out personal information. These platforms take advantage of GPS to arrange for the ride and help determine a driver's best route. They also provide other benefits for riders and drivers, including measures of rider and driver quality to foster trust [5], and an efficient payment system, frequently using a credit card that is entered into the platform's database.

The rideshare service has another application for the less dense areas the city. The lesser dense areas of the city are not very well connected with the public transit systems and therefore people have to plan their travel in two phases: from home to subway station/bus stop and then public transit to their workplace. For outer areas of Queens, Bronx and Staten Island, people usually take cabs to reach one of the public transit systems and then continue with next phase of their travel. For example, [6] found that ridesharing users were less likely to have a car than taxi customers, while the APTA report argued that the use of transport modes like ridesharing is associated with less car ownership and more use of public transportation.

III. Related Work

The paper, Disruptive Change in the Taxi Business [7] is about the competition traditional taxi services are facing against ride-sharing services. The paper compares UberX rides to that of taxis in several major cities such as New York City, San Francisco, Boston and others. One of the measures this paper looks at is capacity utilization which is either the rate a passenger is in the car while the vehicle is driving, or the number of miles traveled with a person in the vehicle. The paper finds that utilization is about 30% higher for UberX compared to taxi drivers when looking at time. There are several possible reasons for this. These include the ability for Uber drivers to turn off their apps when no longer looking for passengers, efficient driver-passenger matching and inefficient taxi regulations that prevent taxis from picking up customers after a drop off in other jurisdictions. In the discussion section the paper also mentions that UberX drivers have a potential to earn more income compared to traditional taxi drivers ignoring fixed costs.

The paper [8] talks about the impact of rainfall on the temporal and spatial distribution of taxi passengers. It studies about the effects of weather on transportation. The datasets used in the study are a large-scale, real-world taxi GPS dataset comprising of period of 5 months, of two large cities in China, and the rainfall data of these two cities. For the study, it extracts the state changes points, calculates the characteristics of taxi services, and matches the taximeter data with state change points. It conducts a temporal analysis of taxi trajectory by considering different rain conditions: heavy rain, moderate rain, showers, light rain and no rain. Nearest neighbor method and kriging method packages in ArcGIS version 10.0 are used to develop spatial analysis of taxi service study. The results show that precipitation affects the distribution and volatility of taxi service demand. As rainfall intensity increases, the taxi service demand for evening rush hour increases whereas for non-rush hour periods, opposite trends are seen.

The paper, Sensing Urban Mobility with Taxi Flow [9] talks about the exploring the relationships between pick-up and drop-off locations; the behavior between the previous drop-off to the following pick-up; and the impact of area type in taxi services by analyzing 177,169 taxi trips data collected for Lisbon, Portugal. They start by stating that the other modes of public transports are, although efficient, have limitations when it comes to analyzing the user patterns in space and time. Therefore, with advent of new and precise technologies, taxi data is a good data to study as we get exact pick-up and drop-off locations with start and end time available, along with the route taken. They created a 0.5x0.5 km2 grids on Lisbon's map for analysis and mapped the origin and destination of trips collected. They categorized the Points of Interest(POI) into eight different categories to see the distribution of those categories. Interestingly, Education facilities, Recreation and Services are the dominant in the city. They observed strong links between public transportation terminals and taxis tend to avoid making long trips to suburban areas for pick-up.

In this paper, we conduct a spatial study of taxi and Uber ridership across NYC and draw a comparison between the two. We also compare the ridership patterns with the weather conditions.

IV. Design

The design that we followed for our projects constitutes of publicly available datasets, Hadoop file storage system and Hadoop programs for analysis of data. We received our data from various open source websites such as NYC open data for Taxi and Limousine Commission (TLC) data (2014) (~6GB) and NYC shapefiles (~1.5 MB). Uber trip data was retrieved from a freedom of information request to NYC's TLC (2014) (~4 GB).

The data was then extracted and stored on Dumbo's distributed file storage system. We used python, Map Reduce and Hive to extract information from the dataset. Since we wanted to analyze the data spatially, we also used the NYC taxi zones shapefile. Additionally we used weather data from Weather Underground to compare the effect of weather conditions on ridership.

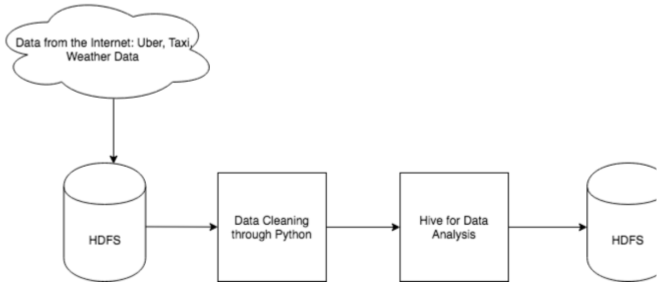

Data from the Internet: Uber, Taxi, Weather Data
HDFS
Data Cleaning through Python
Hive for Data Analysis
HDFS

Figure 1: Schema design of the project

The analysis on Hive majorly included aggregating the data by location and getting the number of trips in those particular areas. Those locations were then extracted out of Hadoop and we created hotspot maps on python using Geopandas. The aggregation tells us how many rides were taken using the yellow taxi and Uber individually in those particular areas.

We tried another approach of aggregating the data by census tracts. We took the shapefile file and then converted it into geojson file for our analysis on hive. However, while applying spatial join with the taxi and Uber data, the result returned 0 values and therefore it wasn't effective for our use. To find the workaround, we converted the shapefile into dataframe in python and then exported it as a csv file. This csv file was then used to spatially join the taxi/Uber data with the geometries on hive, and conduct analysis. For our plots, however, used the original shapefiles and not the converted version of it.

The methodology that we adopted for spatial analysis was majorly based on the Map Reduce and Hive technology. Since for spatial analysis, we need the number of trips starting from a particular point, we had to break the data in such a way that we could get the counts per taxi zone. For NYC, there are five taxi zones defines, one for each borough, but from the TLC report [1] we found that about 90% of the taxi trips pickups are contrained to Manhattan. Therefore, for our analysis we thought it would be a good idea to constrain the analysis to Manhattan as we would get better comparison platforms. Other than Manhattan, the next major hub for taxis is the Airport areas as stated in the aforementioned report.

The maps were created on Python using spatial analysis packages. We exported out the data from the HDFS and used it as an input to our Python script. We joined the shapefile with this data so that we have polygons required for creating the maps. Once the data was joined, we used geopandas tools to create the maps that we displayed in the results section.

V. Results

We used hive to get the number of trips for taxi/Uber across different taxi zones of NYC. Figure 2 shows Taxi Hotspot of NYC. It shows that number of trips are highest for Manhattan, particularly for midtown (Figure 3). This is probably because subway commute options are weaker for midtown Manhattan as compared that of downtown Manhattan, even though both are business centric area, and thus people might rely on taxi for the commute. Upper Manhattan has relatively low dependency on taxis as well, as the subway system is pretty decent in that part and it's not a central workforce area. The other places that have decent number of taxi ridership is Astoria, which is a middle-class and commercial neighborhood in NYC borough of Queens, and JFK airport (Figure 2).

We conducted the same analysis for Uber and found similar results (Figure 4). The one specific area that have high Uber ridership is the Chinatown area. This is probably because the subway transportation is very weak around the area, and most people living in that area don't belong to high income category, thus depending on Uber, which is generally economical as compared to the taxis.

We used Hive to compare the average ratio of taxi to Uber ridership in Manhattan, since most of the hotspots are located in Manhattan. The average ratio came out to be very large (~1819), which is probably because of huge difference in the outliers. Thus, we went on to calculate the median ratio, which came out to be decent as compared to the mean ratio (13). This ratio show that people still rely more on Taxis as compared to Uber. This was quite unexpected as we thought people would use Uber more as it is more economical than taxis. However, since the time-reliability is one of the biggest issues of Uber, it is likely that people rely more on the Taxis as compared to Uber.

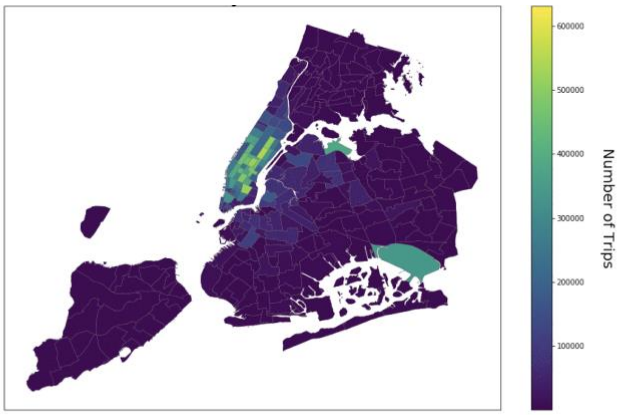


Figure 2: Taxi Hotspot Map of NYC

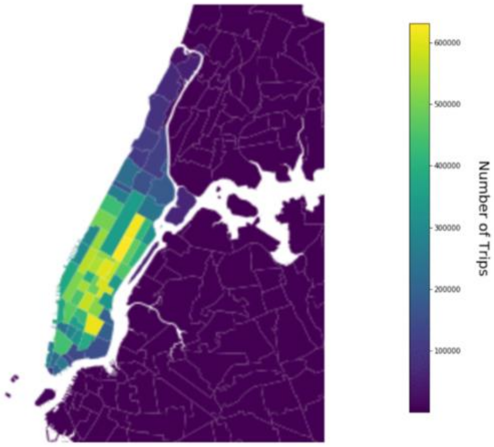


Figure 3: Manhattan Only Taxi Hotspot Map

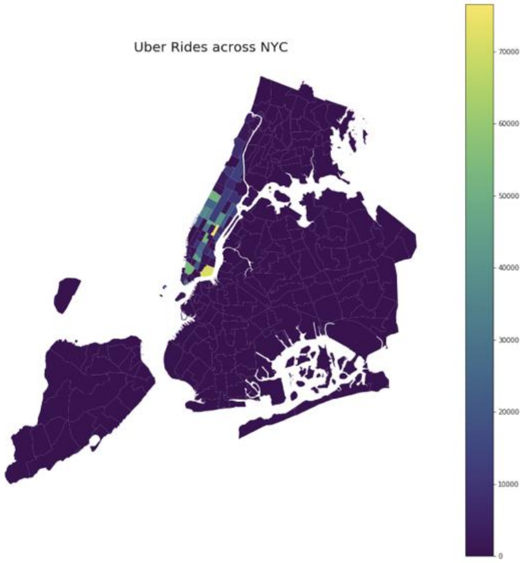
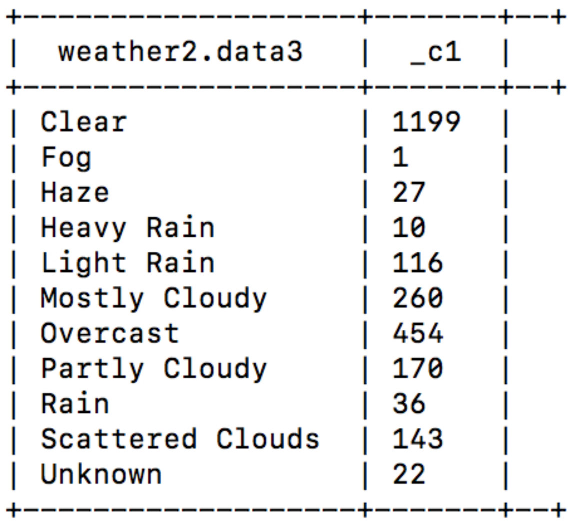


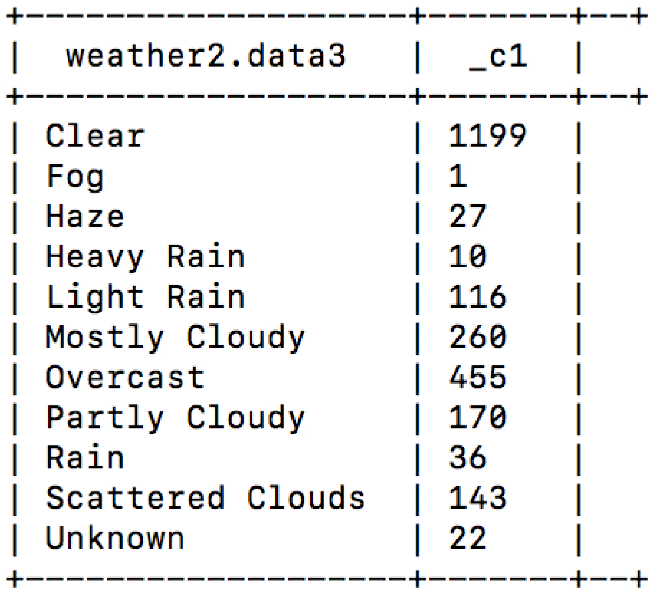
Figure 4: Uber Hotspot Map of NYC

For further analysis, we used Hive to join weather data and Uber ridership data to compare the effect of weather conditions on Uber ridership. Table 1 shows the number of Uber trips made in different weather conditions. The tables show that most number of trips are made when the weather is clear. The results are inconsistent with our expectations as we initially hypothesized that people will use Uber more on rainy days. One reason for the failure of hypothesis here would be insufficient Uber data. Table 2 shows similar results as the Uber data, except in the case of Overcast. This is probably because the taxi and Uber data is limited as compared to the weather data, and thus after joining the tables, the results are not promising.



```
+-------------------+-------+--+
|  weather2.data3   |  _c1  |
+-------------------+-------+--+
| Clear             | 1199  |
| Fog               | 1     |
| Haze              | 27    |
| Heavy Rain        | 10    |
| Light Rain        | 116   |
| Mostly Cloudy     | 260   |
| Overcast          | 454   |
| Partly Cloudy     | 170   |
| Rain              | 36    |
| Scattered Clouds  | 143   |
| Unknown           | 22    |
+-------------------+-------+--+
```

Table 1: Number of Uber trips in different weather conditions



```
+-------------------+-------+--+
|  weather2.data3   |  _c1  |
+-------------------+-------+--+
| Clear             | 1199  |
| Fog               | 1     |
| Haze              | 27    |
| Heavy Rain        | 10    |
| Light Rain        | 116   |
| Mostly Cloudy     | 260   |
| Overcast          | 455   |
| Partly Cloudy     | 170   |
| Rain              | 36    |
| Scattered Clouds  | 143   |
| Unknown           | 22    |
+-------------------+-------+--+
```

Table 2: Number of taxi trips in different weather conditions

One of the biggest challenges in the project is the lack of Uber and taxi data. Secondly, it took a lot of time to figure how to work with the spatial data on Hadoop. As mentioned in the Section IV, initially we converted shapefiles into Geojson format for the analysis on Hive, but the results gave 0 results, and multiple errors keep coming up during the analysis on Hive. Further, spatial joins on Hive took a lot of time (~3 hours and more each file), which hindered the progress of the project.

The results showed that Manhattan has the highest ridership count for both Taxis and Uber, which is what we expected initially. However, we didn't expect that Taxi ridership will be greater than the Uber ridership. Similarly, the dependency of the two modes (Taxi, Uber) on commute on weather conditions are unclear and difficult to analyze.

VI. Future Work

While the spatial analysis does a good job to understand what part of the city is affected the most by a certain event, it is equally important to understand the temporal changes of that event. So for the future work, we would definitely include the temporal data and analyze the pickups in space and time. Also, since more and more ride share services are getting introduced to the commuters, they now have wide range of options to choose from. Therefore, the data from these sources such as Lyft, Via and green taxis could be interesting datasets to look at. These datasets may bridge the current gap between the yellow taxi and Uber, since all of these services offer some perks over Uber.

As further extension, we would work on analyzing this data further and come with a dynamics system of route optimization. Since many users still refrain themselves from using these services because they end up spending more time, an efficient way of routing the trips could potentially help grow the business of these firms. There's a lot of work going on in this domain, but most of the methods fail in real world scenario and they create their hypothesis for the ideal world scenario. These hypotheses are bound to break the real-world issues are more complex and therefore needs systematic approach for choosing the right variables and using efficient techniques to regularize the routes.

Finally, we would want to analyze the traffic patterns in certain parts of the city that often suffer from serious congestions. We would like to analyze alternate routes for such areas so that the ride-share services don't hinder with the usual commute traffic and the riders using these services would have added advantage. This would potentially help the city planners as well as it would better guide them where to invest more and these alternate routes can come to their priority list as they're diverging the traffic from main roads.

VII. Conclusion

This analytic helps in identifying the Uber and taxi ridership hotspots. The Uber and taxi were retrieved from NYC TLC data, which is quite trustable. Further the ridership results are consistent with the TLC report as mentioned in

Section IV. However, the weather data from weather underground appears a bit sketchy, and thus needs to be tested with other weather data sources. The fact that taxi ridership came out to be more than the Uber ridership is probably because of several reasons. At high surge times, like 3 am, taxis are generally cheaper. Secondly during high traffic, Uber can take up to 5 minutes to go to a block and get super expensive, specifically for the case Uber XL, which have excessive wait-times and are expensive as well. Thirdly, Taxi's get access to special red lanes, shared with the bus, which Uber's can't use. This special access helps in speeding up during the traffic quite a bit. Lastly, the rideshare options like Uber and Lyft are not regulated in some cities, hence there is a lack of people's trust in these commute options.

Apart from the future works mentioned in Section VI, the dependency of ridership with the weather conditions also requires to be studied further, to test the goodness of these weather analytic results. Nevertheless, this analytic is a good study for setting the data-driven research grounds for rideshare commute options. Rideshare options are need of the hour as they are generally more economical than normal Taxi options. They occupy low carbon footprint, and thus very useful in controlling the pollution of the urban cities. Lastly, they help in improving traffic situations in highly congested urban areas.

Acknowledgment

We want to thank Professor Suzanne McIntosh for her valuable support and guidance throughout the project. We also want to thank Shenglong Wang from NYU HPC for helping us debugging on HPC, and installing the geo-spatial packages as per our demand on HPC within a timely manner.

References

[1] UNITED NATIONS DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, "2018 revision of world urbanization prospects," 16 05 2018. [Online]. Available: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.

[2] James Barron, "New York City's Population Hits a Record 8.6 Million," 22 03 2018. [Online]. Available: https://www.nytimes.com/2018/03/22/nyregion/new-york-city-population.html.

[3] Nick Lucchesi , "Inverse," 15 03 2018. [Online]. Available: https://www.inverse.com/article/42291-Uber-is-king-chart-shows-the-slow-death-of-the-nyc-yellow-taxi.

[4] Michael Kolomatsky, "The New York Times," 22 02 2018. [Online]. Available: https://www.nytimes.com/2018/02/22/realestate/commuting-best-worst-cities.html.

[5] M. Luca, "Designing Online Marketplaces: Trust and Reputation Mechanisms," *Innovation Policy and the Economy,* vol. 17, 2017.

[6] S. S. N. C. D. D. R. C. Lisa Rayle, "University of California Transportation Center (UCTC) Working Paper," 11 2014. [Online]. Available: https://www.its.dot.gov/itspac/dec2014/ridesourcingwhitepaper_nov2014.pdf.

[7] A. B. K. Judd Cramer, "Disruptive Change in the Taxi Business: The Case of Uber," *American Economic Review,* vol. 106(5), pp. 177-182, 2016.

[8] Y. Z. L. G. N. G. X. L. Dandan Chen, "PLOS," 05 09 2017. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0183574

[9] S. P. C. L. B. Marco Veloso, "ResearchGate," 11 2011. [Online]. Available: https://www.researchgate.net/publication/232318142_Sensing_Urban_Mobility_with_Taxi_Flow.

[10]Taxi and Limousine Commission, "Taxi and Limousine Commission," 2014. [Online]. Available: http://www.nyc.gov/html/tlc/downloads/pdf/2014_tlc_factbook.pdf.