Uber NYC For-Hire Vehicles Trip Data (2021)

Amy Wong, Jessie Estrada, Alan Kuang, Brayan Gonzalez Luna, Rodolfo Hoyo
Department of Information Systems, California State University Los Angeles
CIS 4560-01 Introduction to Big Data
awong7@calstatela.edu, jestra126@calstatela.edu, ckuang4@calstatela.edu, bgonza120@calstatela.edu, rhoyo@calstatela.edu

Abstract: The paper explains the method and process used to analyze the time and location in order to find trends in terms of time and location the for-hire vehicles made. Not only will this help us learn more about the trip records, but it will also help analyze the data for drawing conclusions.We can filter this data to know what streets in New York City have the passengers visited through the ridesharing companies. As such, the analysis of this data will make a great impact in how we look at different ridesharing companies.

1. Introduction

This project uses Hadoop and Hive to process the Uber NYC for-hire vehicles trip dataset. This dataset consists of information regarding the different types of taxis in New York City in terms of time, location, mileage, and pay among different ridesharing companies. In this dataset, the ridesharing companies mentioned are Uber, Lyft, Via, and Juno.

We have chosen this dataset because we know that New York City is very busy and there is always a ton of traffic. By analyzing the underlying trip patterns that happened in 2021, we can find ways to improve Uber's ride-sharing services. This could help the companies know how to better place pickup locations in NYC. This could also help the users make decisions on which ride-sharing company they should rely on.

2. Related Work

2.1 Empirical Studies of Ride-Hailing Services

Yan et al. (2020) conducted an empirical study to investigate the adoption factors, travel characteristics, and mode substitution effects of internet-based ride-hailing services, also known as Mobility-on-Demand (MoD). The authors found that the adoption of MoD is influenced by factors such as income, car ownership, and travel distance. They also found that MoD services have a substitution effect on private car use, but little impact on public transportation use.

Sadowsky and Nelson (n.d.) used a discontinuity regression analysis to examine the impact of ride-hailing services on public transportation use in the United States. The authors found that the introduction of ride-hailing services is associated with a reduction in bus ridership, but no significant impact on rail ridership.

2.2 Spatiotemporal Analysis of Ride-Hailing Services

Varone (n.d.) conducted a spatiotemporal analysis of the growth of ride-hailing services in New York City. The author found that the growth of ride-hailing services is concentrated in Manhattan and areas with high population density. The analysis also showed that the demand for ride-hailing services is higher during weekends and evenings, and that the services are more popular in areas with higher incomes and higher levels of car ownership.

Jin et al. (2022) proposed a deep multi-view graph-based network for predicting citywide ride-hailing demand in Beijing, China. The authors found that their model outperformed other state-of-the-art models in predicting ride-hailing demand, and identified key features that impact the demand for ride-hailing services, such as weather, traffic, and events.

2.3 Impact of Ride-Hailing Services on Public Transit

Sturgeon (n.d.) conducted a case study on the impact of ride-hailing services on public transit at the San Francisco International Airport. The author found that the introduction of ride-hailing services led to a decline in public transit ridership, particularly for the airport shuttle service. However, the analysis also showed that the impact was not uniform across different public transit services, and that ride-hailing services could complement public transit for certain trips.

Overall, these studies provide insights into the adoption, usage patterns, and impacts of ride-hailing services on transportation systems, and can inform policy and planning decisions related to urban mobility.

3. Specifications

Table 1 Data Specification

Data Set	Size (MB)
fhvhv_tripdata_2021-01.csv	294.9
fhvhv_tripdata_2021-02.csv	233.7
fhvhv_tripdata_2021-03.csv	142.5

fhvhv_tripdata_2021-04.csv	**149.4**
fhvhv_tripdata_2021-05.csv	**114.8**
fhvhv_tripdata_2021-06.csv	**76.5**
fhvhv_tripdata_2021-07.csv	**73.7**
fhvhv_tripdata_2021-08.csv	**152.3**
fhvhv_tripdata_2021-09.csv	**119.8**
fhvhv_tripdata_2021-10.csv	**118**
fhvhv_tripdata_2021-11.csv	**84.9**
fhvhv_tripdata_2021-12.csv	**194.5**
taxi_zone_lookup.csv	**12 KB**

The below table shows the specifications of the cluster and specifications of our project

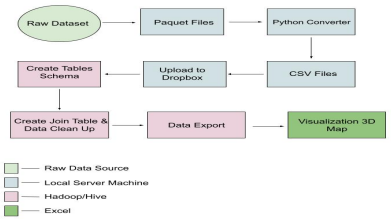
Table 2 H/W Specification

Cluster version	**3.1.2**
Number of nodes	**5**
CPU speed	**2.00GHz**
Memory	**400 GB**
Number CPU Cores	**8**

4. Implementation Flowchart

Our raw data dataset came from Kaggle. The data comprised all for-hire rides requested in the New York region. We had to download the data in .parquet format and use Python to convert it to csv format. After, we uploaded the csv files into Dropbox and began creating the data table schemas. From there, we uploaded it to our Hive server. Once the data was uploaded, we made our tables and joined the table to our city zones. At this point, we exported the data table into our hive server, then downloaded it to our local machines and created our 3D mappings.

Figure 1 - Implementation Flow Chart



5. Data Cleaning

In our data cleaning, we compiled all 12 csv files into our data table. At the same time, in the process of creating the table and importing our data, we replaced Kaggle's zone codes with the proper ride-sharing company. We accomplish this by introducing a switch statement inside our table creation phase.

CASE

WHEN *hvfhs_license_num = "HV0002" THEN regexp_replace(hvfhs_license_num, 'HV0002', 'Juno')*

WHEN *hvfhs_license_num = "HV0003" THEN regexp_replace(hvfhs_license_num, 'HV0003', 'Uber')*

WHEN *hvfhs_license_num = "HV0004" THEN regexp_replace(hvfhs_license_num, 'HV0004', 'Via')*

WHEN *hvfhs_license_num = "HV0005" THEN regexp_replace(hvfhs_license_num, 'HV0005', 'Lyft')*

END AS *ridesharing_company,*

6. Analysis and Visualization

After cleaning our data, and going through the process of further preparing it for more analysis, we extracted them into Excel. The usage of 3D Maps helped us get a better understanding of the data and insight as to what these results mean.

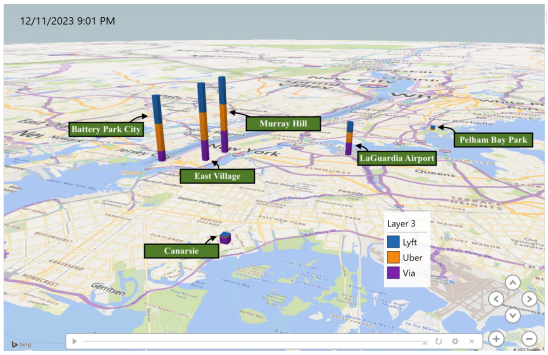


Figure 2 - Pickup Zones and Requests by Time for each Company

This figure shows six different pickup locations and the average congestion surcharges for each company. As seen, the congestion surcharge in Battery Park City, East village, and Murray Hill, have the highest congestion surcharge. These three neighborhoods are very close to one another which supports the idea as to why congestion surcharge is the highest. Congestion surcharges in this area tend to remain around the same throughout the day. The other locations have a congestion surcharge increase in the morning times (8:00am) and another increase during the evening times. Throughout the day these tend to remain fairly low.

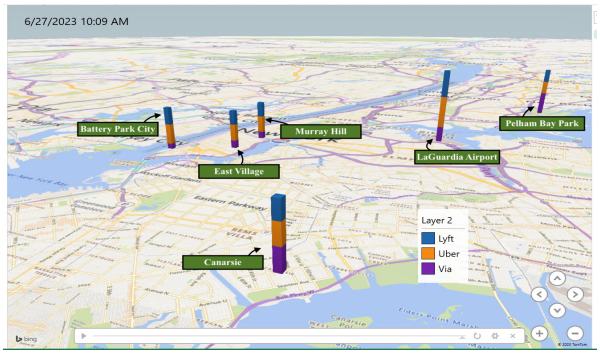


Figure 3 - Pickup Zones and Average Driver Pay for each Company

In this figure, we also have the different pickup locations and the average driver pay from each company. As seen, the average driver pay is increased in LaGuardia Airport, and at Pelham Bay Park. These locations generate the most revenue for these drivers. For the LaGuardia Airport, the driver pay is most likely increased due to the congestion surcharge, and the airport fees. Since there will always be airport fees in this location, this means that pickups or drop offs made here will in turn always make driver pay increase. As we can also see, Uber has the highest average driver pay range because they are the most profitable company. More trips, more customers, and more revenue all contribute to the increase in driver pay.

Figure 4 - Pickup Zones and Requested Trips

The figure above showcases the same pickup locations as well as 3 bars representing each company and their amount of requested trips on an hourly basis. In the mornings and evenings is when the bars are raised. This means that the amount of requested pickups are increased. Still, Uber has the most pickup requests and is the most popular ride service. Similar to figure 2, there is an increase in users at around 8:00 am as well as at 6:00 pm. Throughout the day, the requested trips is fairly low but still rising.

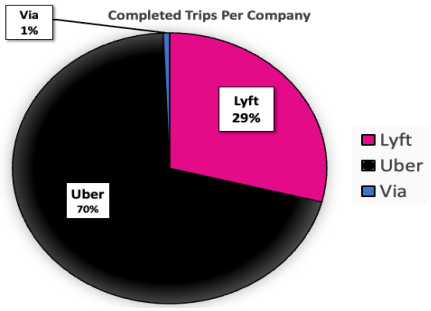


Figure 5 - Percentages of Completed Trips

This figure shows that within our dataset, 70% of the completed trips belonged to Uber. This further solidifies the idea that Uber is the most popular transportation service. Lyft has 29% and Via with only 1%. This also demonstrates that Via is the least used transportation company. From this, we can see how Uber is the most popular and the higher amount of revenue being generated.

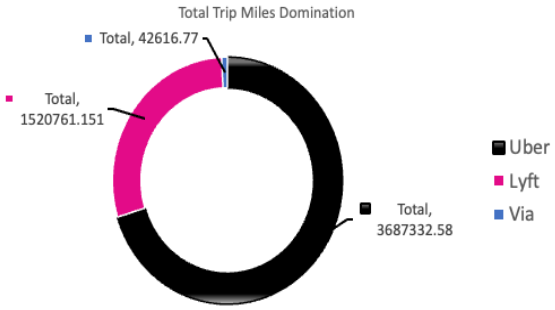


Figure 6 - Trip Miles Domination for Each Company

This figure shows each of our three companies and their total percentage of miles in our dataset. With no surprise, Uber has the most traveled miles with 3687332.58. More trip miles will then lead to more popularity and more revenue. Via has earned the least amount of money and that could also affect company growth. If not many people are using their service, they will not be able to expand and thus, will continue to be at the same level of popularity.

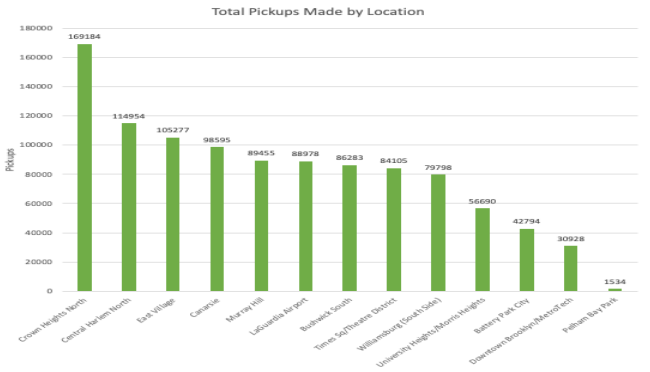
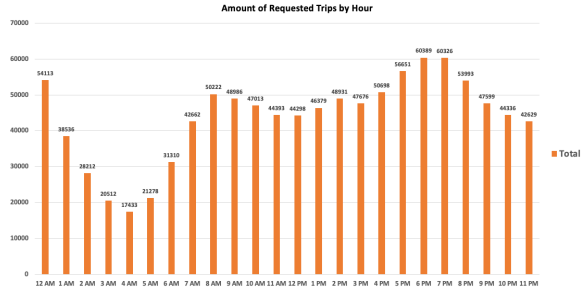


Figure 7 - Pickups made at Every Pickup Location

This figure shows all the different pickup locations and the total amount of pickups made in each. As seen above, Crown Heights North location had the highest number of pickups with a total of 169,184. This demonstrates that this location is more populated than others. This location could be more popular as well and has more people that visit this area. It is also interesting to note that LaGuardia Airport does not have the most or the least amount of pickups made. This goes to show that the high population of New York City leads to a large amount of pickups in regular neighborhoods.

Figure 8 - Trip Miles Domination for Each Company



This figure shows the every hour and the number of requested trips. As we can see, 6 PM and 7 PM are the two times where most requested trips are made. 4 AM is clearly the time in which the least amount of trips are requested. It is interesting to note that there is an increase at around 8 AM and 6 PM. This correlates with the time in which most people go to and from school, work, or any other activity. As such, we can note that at these times congestion surcharges will increase due to the large amount of requests and traffic.

7. Conclusion

Finally, following up the analysis we did, we can conclude the following:

I. That Uber is the most popular ride share service in New York City than Via and Lyft.
II. Uber has a higher congestion surcharge than Via and Lyft.
III. Uber has a better driver pay than Lyft and Via.
IV. The time with the most requested trips was between 6-7pm.

Using the data available from the year 2021,we successfully developed an interactive tool that performs data analysis and manipulation.Further analysis can be done with a larger dataset that has more years and not only limited to 2021.

For more information, dashboards and code, please visit our project's GitHub link[1].

References:

[1] Yan, X., Ward, J. W., Wang, Y., Tan, G. W. H., Rayle, L., Prieto, M., Park, C. K., Okyere, D. K., Nie, Y. M., Lavieri, P. S., Hall, J. D., Haddad, E. A., Flamm, B., … Burkhardt, J. (2020, April 16). Mobility-on-demand: An empirical study of internet-based ride-hailing adoption factors, travel characteristics and mode substitution effects. Transportation Research Part C: Emerging Technologies. https://www.sciencedirect.com/science/article/abs/pii/S0968090X19314573

[2] Sadowsky, N., & Nelson, E. (n.d.). The impact of ride-hailing services on public transportation use: A discontinuity regression analysis. Bowdoin Digital Commons. https://digitalcommons.bowdoin.edu/econpapers/13/

Varone, L. R. (n.d.). Understanding spatiotemporal growth of ride source services in New York City. OpenCommons@UConn. https://opencommons.uconn.edu/gs_theses/1290/

[3] Sturgeon, L. R. (n.d.). The impact of transportation network companies on public transit: A case study at the San francisco international airport. Scholarship @ Claremont. https://scholarship.claremont.edu/scripps_theses/1318/

[4] Jin, G., Wang, Q., Polson, N. G., Zhang, Z., Zhang, Y., Nagy, A. M., LeCun, Y., Mikolov, T., Wang, Y., Gremlin, J., Yu, B., Li, F., Hochreiter, S., & Cui, Z. (2022, September 9). Deep multi-view graph-based network for citywide ride-hailing demand prediction. Neurocomputing. https://www.sciencedirect.com/science/article/abs/pii/S0925231222010931

[5] shuheng_mo. (2023, February 2). *Uber NYC for-hire vehicles trip data (2021)*. Kaggle. https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021.

[1] GitHub Link: https://github.com/JessieEstrada/NYC_Uber_Trip_Data_2021_Project