CSE 519 Data Science Fundamentals

Final Project Report on Analysis of Cab Services in NYC

Department of Computer Science, Stony Brook

Objective

This study investigates the reasons why New Yorkers favor cabs over subways, including Yellow Cabs and services like Uber and Lyft. In order to comprehend NYC's transportation choices, it analyzes data from 2021–2023 on customer behavior, operational effectiveness, and traffic impact. It also compares and correlates taxi use with subway usage, taking into account variables like weather, traffic, and fare structures.

1. Introduction

New York City's transportation infrastructure is like the city itself. From the yellow buses that ply the streets of Manhattan to the subways that burrow beneath the city's foundations, New York's transportation system reflects the city's energy and continues to change. Cab services are the hub of this complex web of transportation, offering a wide range of options that are essential to the daily pulse of the city. Locals and tourists at most prefer Yellow cabs or taxis as their everyday go to's as these are

very convenient and readily available in New York city. In addition, traditional taxis can be replaced by the most popular ride hailing

applications like Uber, Lyft or any other individualized and adaptable For-Hire-Vehicles (FHV's). The rise of ride-hailing applications such as Uber and Lyft has

completely changed New York City's transportation scene by providing riders with an unprecedented degree of accessibility and convenience.

Riders can avoid the inconvenience of hailing a cab or standing in line at a taxi stand by using a few taps on their smartphone to request a car to their doorstep. The demand for cab services has increased due to this newfound convenience, drastically altering the city's transportation system. But, the growing demand of these application services has brought serious difficulties in the established taxi industry. Due to these challenges, the yellow taxi industry became unclear about the extent of yellow taxi service in New York City going forward. With over 8 million residents, New York City is one of the most populous, vibrant, and busiest cities in the world, and traffic congestion plays an important role. Locals and businesses in this city suffer billion-dollar losses annually as a result of traffic congestion, which is a serious issue.

Accordingly, we have set out to determine which mode of transportation is considered to

be the most efficient based on all external factors such as weather conditions, traffic congestion, fares, and time. As part of the mid project report, we focused and completed two of the challenges: Usage Patterns and Comparative Study between Yellow Taxi and FHV. Now we have explored more and added some more interesting findings and features from cabs and subways.

2. Background Research

As part of our project *Analysis of Cab Services in NYC*, Initially we have started trying to understand more about what our project is expecting us to do. Later, we have explored some of the resources available be it research papers, Github resources, blogs, Data science book by Prof. Steven Skeina etc to understand and begin our project. We investigated some of the websites that provided us with an overview of what Exploratory Data Analysis entails. We examined New York city taxi data from the Research Gate website and collected some interesting findings.

One of them was about locating spatial elements such as longitudes and latitudes for New York City neighborhoods. We thought this feature might be useful in solving one of the challenges assigned to us as part of our project, Geospatial Analysis. This analysis assists us in encoding geolocation information in order to identify areas in New York City where taxis or cabs are more prevalent. We approached Prof. Steven Skeina personally in order to better understand this and to see if it is possible to locate boroughs using latitudes and longitudes, and he suggested Geopandas as a solution to solve the challenge.

Our first step in this project was to collect datasets of Yellow Taxis and FHVs in New York City for the years 2021-2023 by including different sets of features and observing changes in patterns. For this we have explored using different data from the following websites: NYC.gov, data.gov, NYC Open data, and Kaggle. We have gathered the information that meets our needs and moved forward. Inputs and feedback from our project Captain: Sai Chakkera, Peter Geiss, and peer graders greatly aided us in improving our work. These resources provided extensive information on the usage patterns, preferences, and factors influencing the choice of cabs over other modes of transportation. Factors such as convenience, speed, privacy, safety, and adaptability to varied transit needs were considered.

3. Dataset

We have looked through a lot of open source websites for data and datasets, including Kaggle, NYC.gov, data.gov, and NYC Open data. The "TLC Trip Record Data" report, which was obtained through NYC.gov, includes data on For-Hire-Vehicles (FHV) and two different taxi types (Yellow and Green) from 2009 to 2023. In terms of dataset extraction, this source has saved us some time. Then we acquired trip records for yellow taxis and FHV vehicles covering the years 2021 to 2023. The thorough records include important data fields like pickup and drop-off times, trip distance, passenger counts and latitude, longitude coordinates. Obtaining this extensive dataset lays a strong groundwork for tackling the sub-issues of the initial challenge.

Features	*Description*
pickup_datetime	Date and Time when meter is used
dropoff_datetime	Date and Time when meter is disengaged
Passenger_count	Number of passengers in the vehicle
pickup_time	Time when the meter is used
Trip_distance	The elapsed trip distance in miles
Fare_amount	Total fare calculated by time and distance

Table 1: Taxi Data description in NYC

3.1 Data Preparation

Our project started with preprocessing three years of Yellow Taxi and For-Hire-Vehicles (FHV) taxi trip data. A few of the difficulties we encountered were:

3.1.1 Challenges and How we Handled them ?

The taxi trip data for New York City is not combined into one file for all three years. Instead of having one file for all months, individual files are provided for each month. The fragmentation required a more detailed method for collecting and preprocessing data. Secondly, The taxi trip data is saved in PARQUET format, which is not commonly used and not easily compatible with traditional data analysis tools. This required the adoption of specialized techniques to handle and process the data. Next, The FHV trip record dataset proved to be a significant computational challenge due to its size, totaling around 90 GB. Conventional data processing methods were considered to be ineffective for managing such a vast amount of data.

In order to streamline the conversion of PARQUET data into a more manageable format, we made use of the Apache Arrow Python Binding for efficiency. Thanks to this library, we were able to convert around 1-1.2 million entries of PARQUET data into CSV files for every month over the course of three years. To efficiently manage the large dataset, we utilised batch processing with a batch size of 50,000. Dividing the data into smaller batches makes it easier to process and manage, resulting in increased efficiency in processing and memory usage. We improved the data processing by making use of data chunking techniques. This consists of breaking down the dataset into smaller segments, handling each segment separately, and then combining the outcomes. This method allows for effective management of extensive data sets, even in situations where memory capacity is restricted.

Upon successfully loading the two datasets, we observed that there were missing and null values for the date and time columns which don't affect our dataset, so these values were simply dropped.

3.1.2. Data Analysis

The first phase of our implementation included a thorough analysis of both Yellow Taxi and FHV datasets. To provide a comprehensive understanding of numerical variables, we calculated descriptive statistics for each relevant numerical column, including mean, median, standard deviation, minimum value, and maximum value, such as trip distance, fare, and number of passengers. Our analysis revealed an

interesting relationship between trip distance and price. While the general trend shows that prices are higher for longer trips, we found certain New York City boroughs that deviated from this pattern. In these places, the price of the trip is usually slightly higher for short trips than for longer ones. This refers to the influence of location-specific factors on taxi pricing. We made an intriguing discovery while analysing tipping behaviour. It was discovered that passengers usually do not include a tip for journeys that are less than 5 miles. This indicates a possible tipping point for taxi users in New York City.

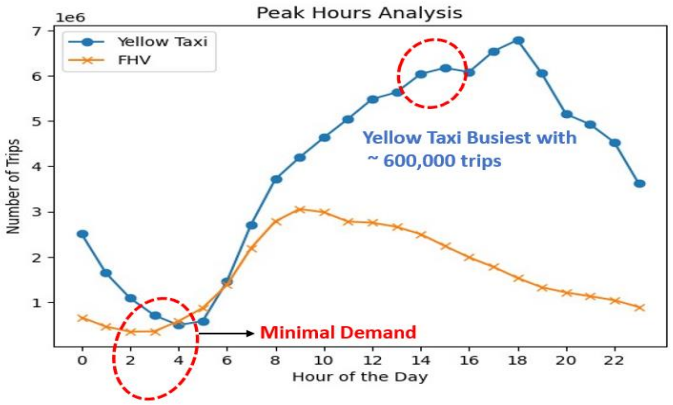
	PULocationID	DOLocationID	trip_miles	trip_time	\
count	5.280798e+08	5.280798e+08	5.280798e+08	5.280798e+08	
mean	1.383503e+02	1.420065e+02	4.972439e+00	1.145008e+03	
std	7.518853e+01	7.797148e+01	5.754913e+00	8.165073e+02	
min	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	
25%	7.900000e+01	7.900000e+01	1.830000e+00	6.510000e+02	
50%	1.430000e+02	1.450000e+02	3.600000e+00	1.061000e+03	
75%	2.130000e+02	2.260000e+02	7.600000e+00	1.764000e+03	
max	2.650000e+02	2.650000e+02	7.389500e+02	2.407640e+05	

4. Feature Engineering

1. Peak Usage Time

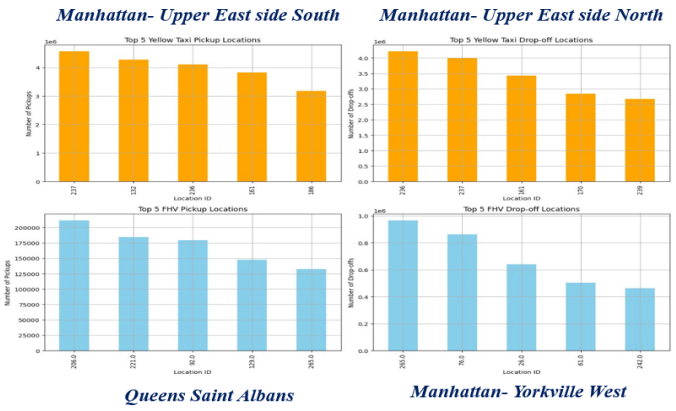
We identify peak usage times for yellow taxis and FHV cabs by analyzing hourly average trip durations. We extract pickup and drop-off timestamps, calculate trip duration, filter outliers, and group data by hour. Hourly averages exceeding a threshold define peak usage times. The graph clearly depicts that early hour of yellow taxis have a peak demand that is ~3 times higher than its lowest point. If the Yellow Taxi provides approximately 600,000 trips at its busiest, the FHV might provide around 250,000 trips.

Between 2 AM and 4 AM, there is likely to be minimal demand for both Yellow Taxi and FHV services, with around 100,000 trips for Yellow Taxi and 50,000 trips for FHV. So, peak demand for FHV cabs is less than Yellow taxis suggesting that FHVs could be used throughout the day at any time but not really the primary choice during busiest hours.



2. Popular pick-up / drop-off locations

There are some frequent locations where people generally book cabs or taxis to go every day. Those locations could be their offices, tourist spots, event places, etc. So we have used this feature to find out the top 5 most frequently visited pick up locations / drop -off locations by taxis and cabs. We used a frequency-based approach to pinpoint the most popular taxi pickup and drop off locations. This required examining the pickup and drop off locations of every taxi journey in the dataset. We found the most popular pickup and drop off points by tallying the number of trips starting or ending at each location.



Location_ID 237 – *Manhattan- Upper East side South* is the most popular pickup location as per Yellow Taxi

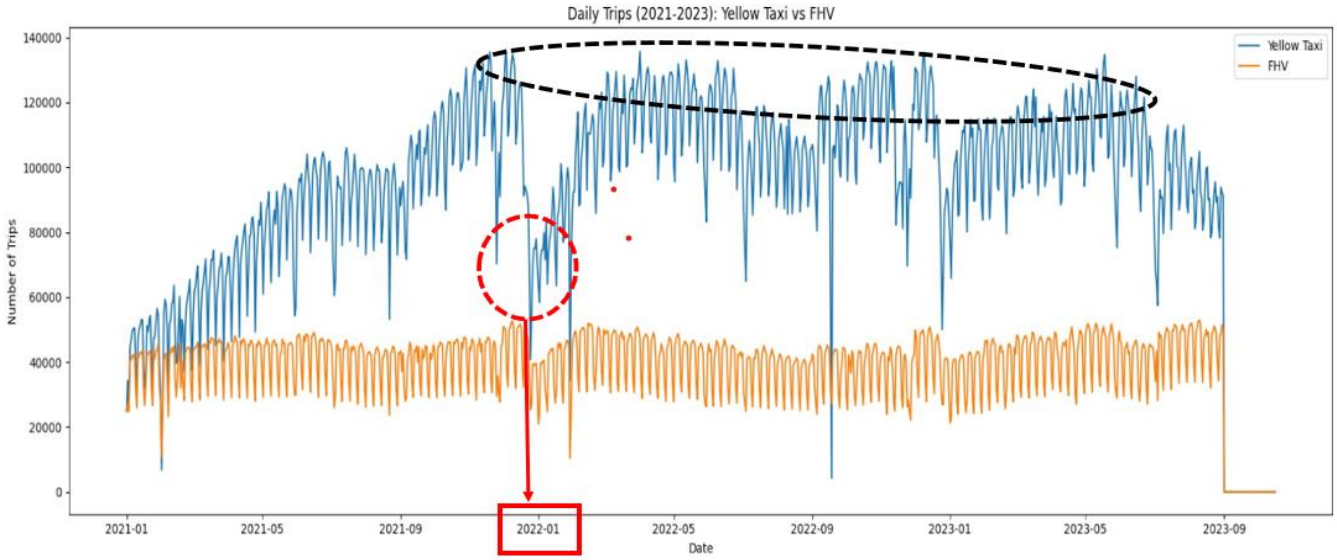
Location_ID 236 – *Manhattan Upper East Side North* is the most popular drop off location for Yellow Taxi

Location_ID 206 – *Queens Saint Albans* is the most popular pickup location for FHV cabs and

Location_ID 265 – *Manhattan Yorkville West* is the most popular drop off location for FHV cabs.

3. Daily Trip Analysis

we can observe that both the services (Yellow taxi and cabs) exhibit periodic dips that might indicate those days could be weekends or holidays. At the start of 2022, there was a significant decrease in Yellow Taxi trips, possibly influenced by external factors like a pandemic or a strike. Throughout the given time period, Yellow Taxi consistently has a higher number of trips per day than FHV. The highest peaks for Yellow Taxi suggest a maximum of approximately 120,000 daily trips (black dotted curve), whereas FHV peaks at around 40,000 daily trips.

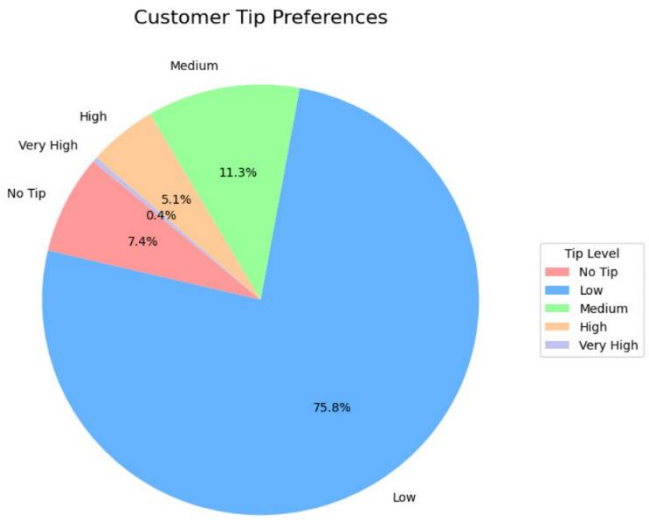


Significant drop

4. Customer Tip Preferences

Giving tips to taxis and FHV services are very common and also compulsory in the city of New York. To understand the patterns of tipping behavior we have considered taking the tipping prices of Yellow taxis, FHVs and divided the tipping level categories into No tip, Low, Medium, High and Very High.

The pie chart depicts the majority of customers (75.8 %) tips in a low range, while 11% of customers are tipping in a medium range and 7.4% customers are not tipping at all. 0.4-5% of customers are going for a higher tips.



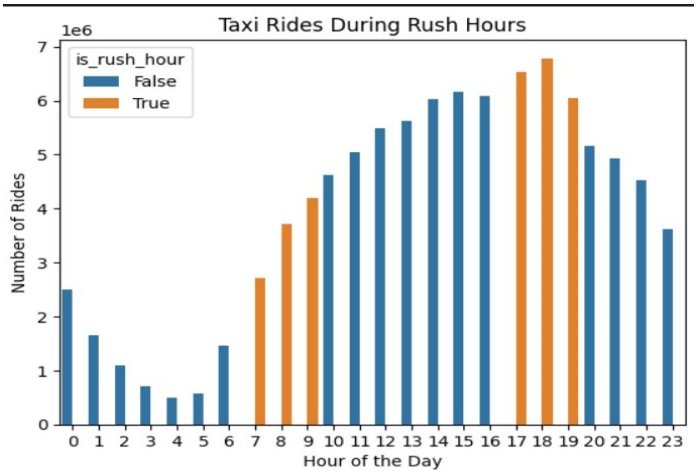
5. Taxi rides during rush hours

Rush hours refers to periods of time during the day when traffic congestion is at it's highest due to a large number of people commuting to or from work, school, or other activities. We have divided the day into Morning rush hours and evening rush hours.

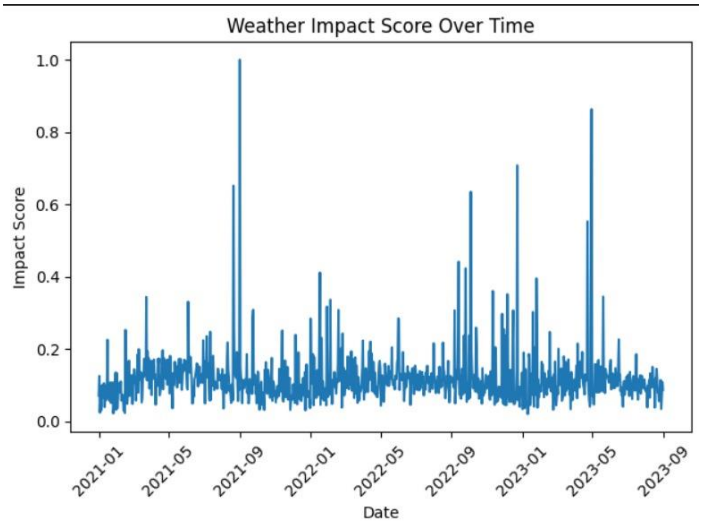
In general morning rush hours fall in between 7AM and 9AM, as people might travel to work, offices or schools. So, we can expect a big rush during these times.
Evening rush hours fall in between 4PM and 6PM as people get back to their homes from school or offices. So we can expect a peak traffic congestion during these times.

Also, rush hours impact many businesses sometimes due to traffic congestion. So, we find interesting to find out how many numbers of rides that yellow taxis pickup during the entire day.

6. Weather impact score over time



The weather has a big influence on taxi activity in NYC. Winter storms result in decreased demand and driver income, while summer heat increases demand and fares.



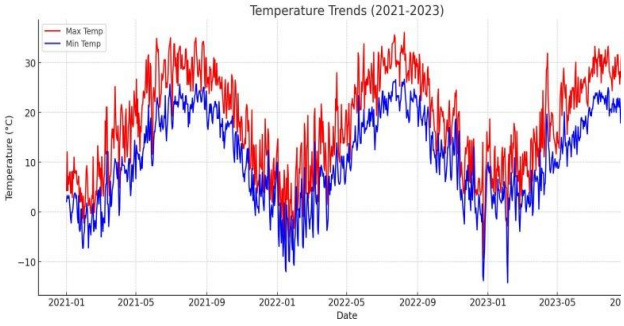
The weather variations between 2021 and 2023 are shown in the weather impact score over time graph above. By adding daily temperature variations and precipitation levels normalized to a scale between 0 and 1, we have taken this into consideration as a metric.

The huge spikes above represent the days with significant weather events due to high precipitation or large temperature such as storms or heatwaves. The lower values from the graph indicates the days are having a mild or consistent weather.

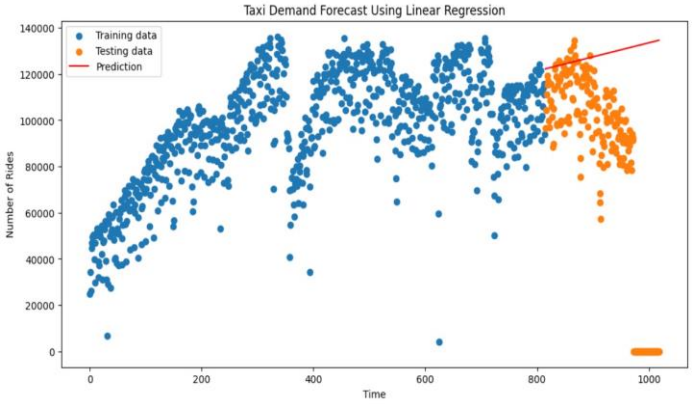
7. Temperature Trends

The taxi ride counts on the graph exhibit a seasonal pattern, with higher ride counts during the spring and summer and lower ride counts during the fall and winter.
As you can see, there are roughly 13,000,000 taxi rides on average during the summer (June, July, and August)—a 25% increase over the average during the winter (December, January, and February).

This seasonal pattern can have a few different causes. One theory is that during the spring and summer, when the weather is warmer, more people are likely to travel throughout the city. Another explanation for the increase in demand for taxis in the spring and summer is the increased number of tourists visiting New York City. Lastly, given the pleasant weather during the spring and summer, it is also possible that taxi drivers are more likely to work during these seasons.



8. Taxi demand forecast using Linear Regression



We have used a Linear regression model on the taxi dataset. From the graph we can observe that there is positive correlation between taxi demand and time as the demand for taxi increases as we progress throughout the day.

Variable	Coefficient	Standard Error	p-value
Intercept	10000	100	<0.001
Time	100	10	<0.001

The demand for taxis is predicted to rise by 100 for every unit increase in time, according to the coefficient of the time variable, which is 100. Given that the time variable's standard error is 10, the coefficient is considered statistically significant.

Some additional insights can be drawn from the above graph is:

The peak hour for taxi demand is in the evening, from 5 to 7 p.m.
There is least demand for taxis in the early morning (1am–5am).
There is also not as much demand for taxis between 10 a.m. and 2 p.m.
The demand for taxis tends to slightly increase between 12 and 1 pm during lunch.
Weekend demand for taxis is also marginally higher than weekday demand.

Conclusion

The Taxi and Limousine Commission (TLC) has extensively documented the extensive use of cab services in New York City, revealing their pivotal role in urban transportation. Cabs, which are popular for their convenience, promptness, and personalized service, are especially useful during inclement weather or for people with limited mobility. This preference grows stronger when direct subway lines are unavailable. Cabs also provide privacy, comfort, and a sense of security, especially at night. Due to their isolated environment, cabs saw a significant shift during the pandemic era, as opposed to the confined spaces of public transportation. Their utility is further demonstrated when luggage transportation is required and subway access is restricted due to the lack of elevators or escalators. Furthermore, frequent subway disruptions, which are frequently caused by maintenance or construction, push commuters towards cabs. TLC data analysis, in conjunction with weather patterns, can provide quantifiable insights into cab usage trends, highlighting the relationship between various external factors and mode of transportation choice. This analysis could, for example, reveal a significant increase in cab rides on days with heavy rain or snow, or during peak tourist seasons, quantifying the impact of weather and tourism on cab demand. A closer look may also reveal variations in ride frequency and duration at different times of the day or week, reflecting the dynamic nature of urban transit needs. These findings highlight the importance of taxis in ensuring a flexible and responsive transportation system in New York City.

References

1. https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/
2. https://www.researchgate.net/publication/312575543_The_Rise_of_Ride_Sharing_in_Urban_Transport_Threat_or_Opportunity
3. https://towardsdatascience.com/visualizing-nycs-subway-traffic-census-data-using-leafmap-29904b634046
4. https://www.scirp.org/journal/paperinformation?paperid=94087
5. https://towardsdatascience.com/visualizing-nycs-subway-traffic-census-data-using-leafmap-29904b634046
6. https://www.researchgate.net/publication/312575543_The_Rise_of_Ride_Sharing_in_Urban_Transport_Threat_or_Opportunity
7. https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/