Examining the For–Hire Vehicle Trips in New York City

Narendra Yadav Pottipati
School of Computing and EngineeringUniversity of Missouri-Kansas City Kansas City, MO, USA
npdbd@umsystem.edu

Srinivas Gowtham Majeti
School of Computing and EngineeringUniversity of Missouri-Kansas City Kansas City, MO, USA
smqbn@umkc.edu

Vinay Reddy Baddam
School of Computing and EngineeringUniversity of Missouri-Kansas City Kansas City, MO, USA
vbk7r@umsystem.edu

Sai Teja Reddy Dundra
School of Computing and EngineeringUniversity of Missouri-Kansas City Kansas City, MO, USA
sdvbh@umsystem.edu

Sri Ram Abhinav Reddy Dudi
School of Computing and EngineeringUniversity of Missouri-Kansas City Kansas City, MO, USA
sdgwh@umsystem.edu

ABSTRACT

The goal of this project is to gain insights into the demand, supply, and business goals of for-hire vehicle services like Uber and, Lyft, Via. These insights gained from this project can help optimize their operations and improve their services to meet better the needs of their customers in New York City. The dataset contains information on various aspects of each trip, such as pick-up and drop-off times, location, distance traveled, fare, and additional charges like tolls and surcharges. Through exploratory data analysis, visualization, and statistical modeling, this project aims to identify patterns, trends, and insights that can shed light on the behavior of passengers and drivers, the impact of regulations, and the efficiency of the FHV system. Possible topics of investigation include the spatial and temporal distribution of trips, the factors affecting the trip duration and cost, the prevalence and effects of shared and wheelchair-accessible trips, and the relationship between FHV and other transportation modes. The results of this project can inform policy decisions, improve the quality of service for passengers, and support the sustainability of the For-Hire Vehicles industry.

KEYWORDS:

Python, MatLib, Pandas, Seaborn

1. INTRODUCTION

The For-Hire Vehicle (FHV) industry in New York City has undergone significant changes in recent years, particularly with the rise of services such as Uber, Lyft, and Via. As a result, there has been growing interest in analyzing FHV trip data to better understand the usage patterns of these vehicles, identifying popular pickup and drop-off locations, and examining the impact of various factors such as time of day, day of the week, and weather on trip demand. By analyzing this data, we can better understand the demand, supply, and business goals, optimize their operations, and improve their services to meet better the needs of their customers in New York City.

2. RELATED WORK

When we were searching for projects to implement, we came across below stated article on the internet which later inspired us to dig for more projects related to this. And then we found a dataset from Kaggle related to for-hire vehicle trips and then we decided to analyze and virtualize this dataset which brings very interesting insights that help the for-hire vehicles companies to enhance their services for their customers.

The Cost of Convenience: Ridesharing and Traffic Fatalities in New York City" by Liu et al. (2019) - This study analyzed the impact of ridesharing services on traffic fatalities in New York City. The authors found that the growth of ridesharing services was associated with an increase in traffic fatalities, but this effect was reduced when ridesharing was combined with public transportation.

3. IMPLEMENTATION

Our project implementation involved several steps. We began by downloading and setting up Python's latest version from the official website, followed by installing essential libraries (e.g., Pandas, Matplotlib) via pip command in the command prompt.

We then loaded the dataset into Pandas data frame using read_csv() method, cleaned it by removing duplicates, irrelevant columns, and missing values with drop_duplicates(), dropna(), and drop() methods, and

performed exploratory data analysis using Pandas and Matplotlib by utilizing describe() method and plotting various graphs and charts. Afterwards, we preprocessed the data using Pandas and NumPy, incorporating feature scaling, normalization, and encoding of categorical variables, and customized the chart created with Matplotlib by adding labels, titles, legends, and other annotations.

4. DATA AND METHODOLOGY

Our dataset comprises of nearly 12 million records of for-hire vehicle (FHV) trips in New York City, and it provides extensive information on each trip such as the FHV company license number, pickup and drop-off locations identified by a unique Location ID, fare details. Additionally, timestamps are provided for different stages of the trip. This dataset can be used to gain insights into the behavior of passengers and drivers, the demand and supply of FHV services, and the effects of regulations on the FHV industry in New York City.

```
data_jan.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11908468 entries, 0 to 11908467
Data columns (total 24 columns):
 #   Column                Dtype
---  ------                -----
 0   hvfhs_license_num     object
 1   dispatching_base_num  object
 2   originating_base_num  object
 3   request_datetime      datetime64[ns]
 4   on_scene_datetime     datetime64[ns]
 5   pickup_datetime       datetime64[ns]
 6   dropoff_datetime      datetime64[ns]
 7   PULocationID          int64
 8   DOLocationID          int64
 9   trip_miles            float64
 10  trip_time             int64
 11  base_passenger_fare   float64
 12  tolls                 float64
 13  bcf                   float64
 14  sales_tax             float64
 15  congestion_surcharge  float64
 16  airport_fee           float64
 17  tips                  float64
 18  driver_pay            float64
 19  shared_request_flag   object
 20  shared_match_flag     object
 21  access_a_ride_flag    object
 22  wav_request_flag      object
 23  wav_match_flag        object
dtypes: datetime64[ns](4), float64(9), int64(3), object(8)
memory usage: 2.1+ GB
```

Fig 4.1: Dataset Columns names and datatypes

Total number of row & columns in dataset

```
data_jan.shape

(11908468, 24)
```

Fig 4.2: Total number of columns and rows in dataset.

You can find the dataset in the below link:
https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021

5. ANALYSIS AND RESULTS

Total miles traveled by passengers in comparison to morning, afternoon, and night in a month (Uber, Lyft, via)?
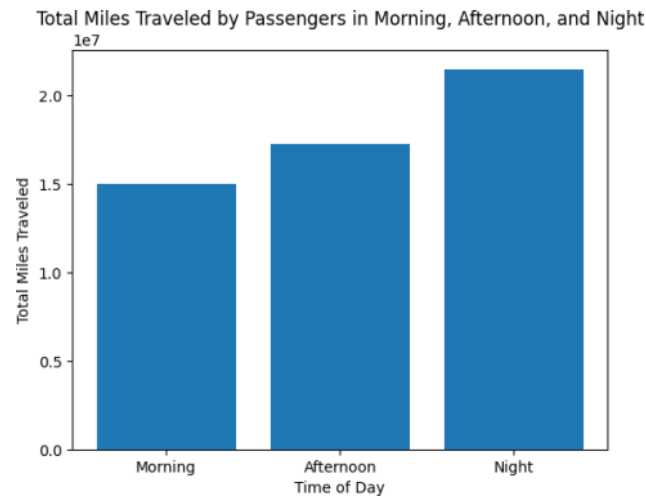

Fig 5.1: Visualization for miles traveled by passengers in comparison to morning, afternoon, and night in a month.

In a month, the most visited destinations were traveled by passengers in (Uber and Lyft, Via) NYC?
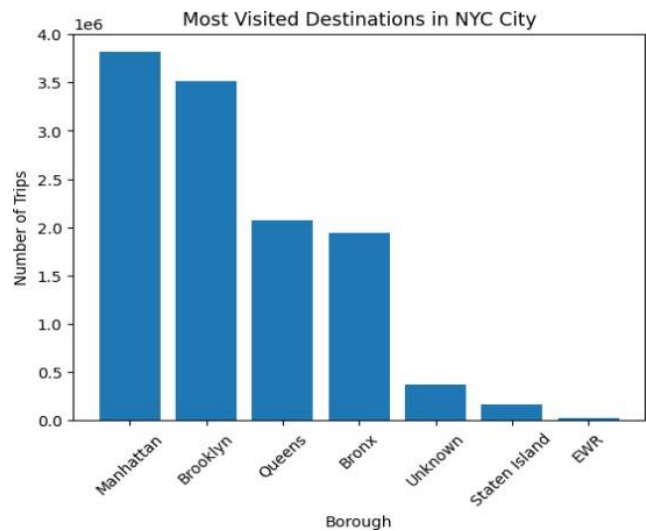

Fig 5.2: Visualization for most visited Destinations in NYC.

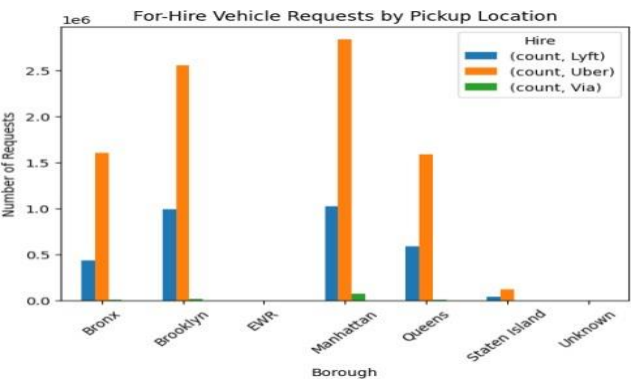Requests by Pick-up locations for For-Hire services

Fig 5.3: Visualization of requests by Pick-up locations for For-Hire services

Average fare amount across different types of for-hire vehicles at pick-up location
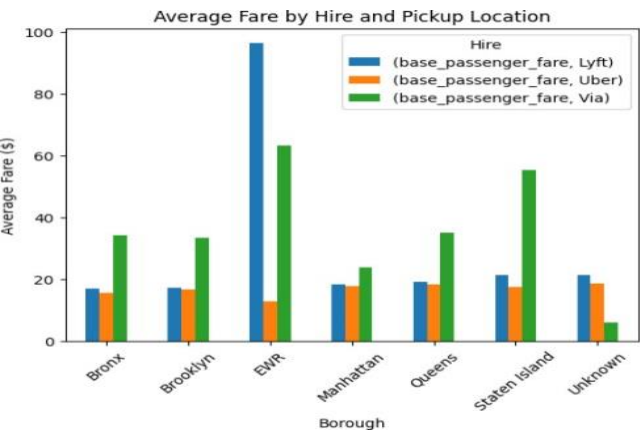

Fig 5.4: Visualization for Average fare amount across different types of for-hire vehicles at pick-up location

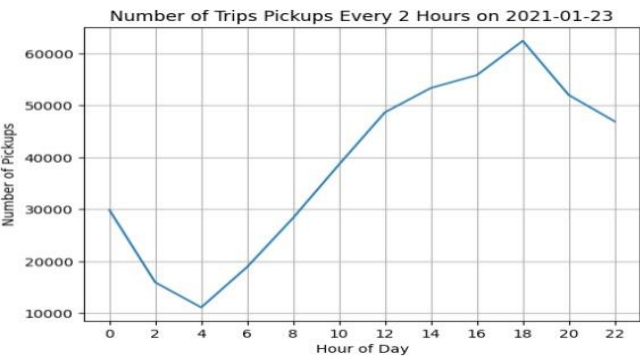The number of trip pickups occurs every 2hrs on a random day.

Fig: 5.5: Visualization for the number of trip pickups occurs every 2hrs on a random day.

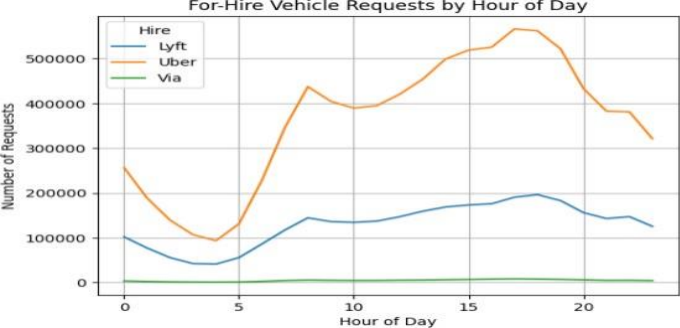For-Hire Vehicle Requests by random Hours of Day.

Fig : 5.6 Visualization of For-Hire Vehicle Requests by random Hours of Day.

Average passenger fare for each hire vehicle and how it varies by time or day, or week?
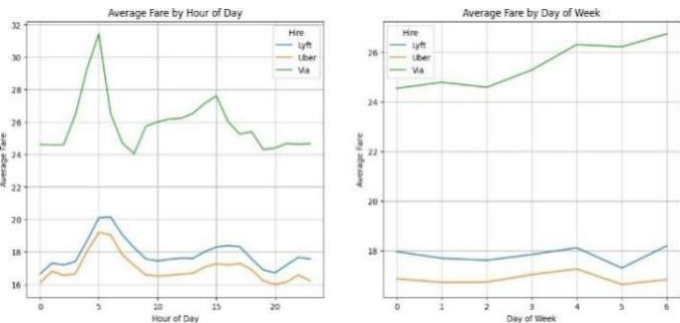

Fig 5.7: Visualization for Average passenger fair for each hire vehicle and how does it vary by time or day or week.

Hire Vehicles with the highest number of trips in each period.
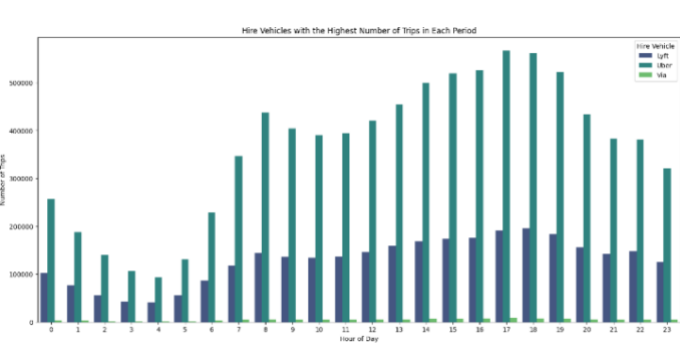

Fig: 5.8 Visualization of Hire Vehicles with the highest number of trips in each period.

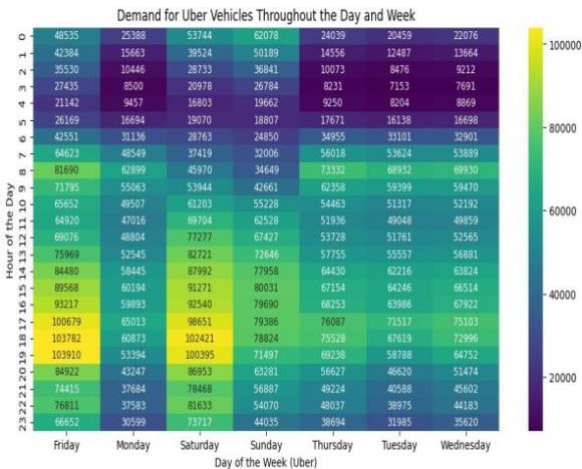Demand for hire vehicles changes throughout the day & week in a month.

Fig: 5.9.1 Visualization of Demand for hire vehicles changes throughout the day & week in a month for uber.
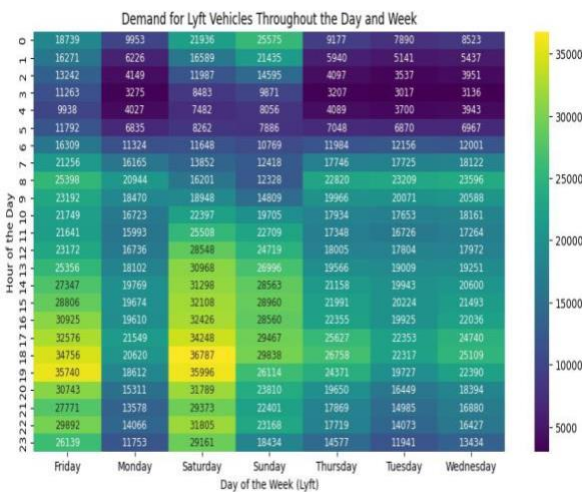
Fig: 5.9.2: Visualization of Demand for hire vehicles changes throughout the day & week in a month for Lyft.
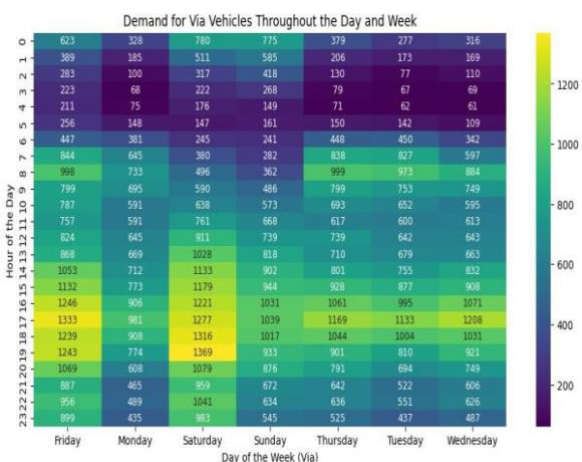
Fig: 5.9.3 Visualization of Demand for hire vehicles changes throughout the day & week in a month for Via.

Hire (Uber, Via, Lyft, Juno) is used by the customer based on distance and price.
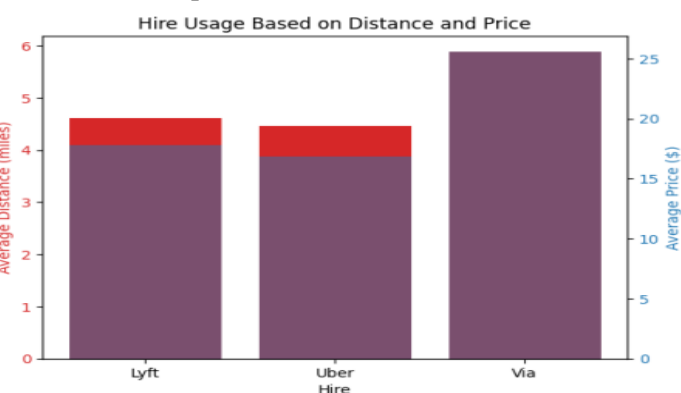
Fig: 5.10 Visualization of Hire (Uber, Via, Lyft, Juno) is used by the customer based on distance and price.

The amount of tolls, taxes, and fees collected vary across different hire vehicles and periods.
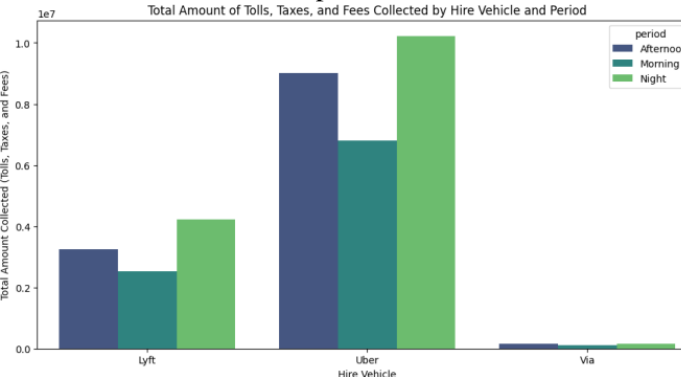
Fig:5.11 Visualization of amount of tolls, taxes, and fees collected vary across different hire vehicles and periods

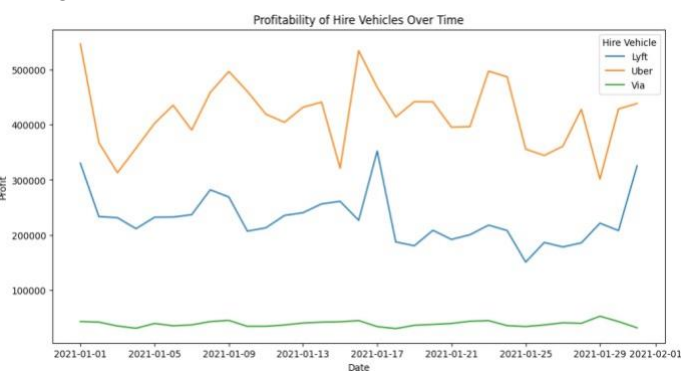Overall profitability of each hire vehicle and how has it changed over time.

Fig: 5.12 Visualization of Overall profitability of each hire vehicle and how has it changed over time

6. CONCLUSION

Our project delved deep into the various aspects of the for-hire vehicle industry in New York City, which has undergone significant changes with the advent of services like Uber, Lyft, and Via. By analyzing the pickup and drop-off locations, we identified the most popular areas where passengers hail for-hire vehicles, enabling these services to better target their services in these areas. Additionally, our investigation of the factors affecting trip duration and cost provided insights into how for-hire vehicle services can optimize their operations to improve profitability while also providing quality service to their passengers.

Overall, our project demonstrated the value of data-driven decision-making in the for-hire vehicle industry and emphasized the importance of continued research to drive innovation and improve the quality of service for passengers. By providing valuable insights into various aspects of the for-hire vehicle industry, our project has the potential to inform policy decisions and improve the sustainability of the industry.

7. AUTHOR'S CONTRIBUTION

As a group, we distributed the workload evenly and ensured each member took responsibility for different tasks during the project. Throughout the different phases of the project, we rotated the leadership role to ensure that everything was going according to plan.

Narendra Yadav and Srinivas Gowtham investigated various datasets to find the best one and ultimately found a suitable dataset on Kaggle. They also took charge of the data visualization part and worked on the midterm presentation.

Vinay Reddy and Sai Teja Reddy investigated different datasets during the initial stages of the project and carried out data- cleaning activities, such as removing unwanted data. They also prepared the project proposal presentation.

Sri Ram Abhinav Reddy worked on creating the project. presentations. He investigated datasets related to Uber For-Hire Vehicles, but we continued with the same dataset as we needed help finding suitable ones.

We shared our progress regularly and later came together to discuss everything. Finally, we collaborated on the documentation aspect, with each member contributing to its completion.

8. REFERENCES

[1].Exploring the taxi and Uber demands in New York City: An empirical analysis and spatial modeling
D Correa - Available at SSRN 4229042, 2017 - papers.ssrn.com … account for the spatial dependence of Uber and taxi **trips** are developed.

*[2]. Understanding the surprising and oversized use of ridesourcing services in poor neighborhoods in New York City*C Atkinson-Palombo, L Varone… - Transportation …, 2019 - journals.sagepub.com… **For**-**hire vehicle trips** in the five boroughs of **New York City** … data on **for**-**hire vehicles** are compiled by the **New York City** Taxi … TLC), we use factor **analysis** and cluster **analysis** to create a …

[3]. Examining Changes in New York's Taxi and For-Hire Vehicle Industry, One Hour at a Time
https://medium.com/@NYCTLC/factbook-deep-dive-1-e4dd57075945

[4]. Value of convenience for taxi trips in New York City
https://www.sciencedirect.com/science/article/pii/S096585642030762X

[5] Pandas documentation.
https://pandas.pydata.org/docs/user_guide/index.html

SOURCE CODE LINK:

https://github.com/gowthammajeti9/Examining-the-For-Hire-Vehicle-Trips-in-New-York-City.git