Analysis of cab services in NYC

1 Introduction

In recent years, there has been a noticeable shift in transportation preferences in metropolitan areas, such as New York. Despite a well-established subway system, more commuters are choosing cab services. Our research aims to uncover the driving factors behind this shift and understand the socio-economic dynamics contributing to the preference for cabs.

Understanding commuter trends in metropolitan areas like New York poses a significant challenge. Addressing socio-economic dynamics requires considering factors like income levels, employment patterns, and urban demographics. This involves examining data on household incomes, employment rates, geographical distribution of commuters, and understanding the needs and preferences of demographic groups like the working class, students, and tourists. External factors such as traffic conditions, weather, crime records, and special events also influence transportation choices.

To comprehend the increasing preference for cabs over the subway system, we acknowledge challenges in data handling. Collecting and integrating diverse data sources with differences in format, quality, and granularity is a primary challenge. Ensuring accuracy and consistency of this data is crucial for deriving meaningful insights.

Our investigation focuses on a comprehensive analysis of this evolving phenomenon, offering a nuanced understanding of contemporary urban commuter choices. We structure our approach to perform a comparative analysis of transportation modes and private options (FHVs, yellow taxis, Uber, Lyft). The analysis includes safety measures, population density, congestion, and other factors.

We conducted a comprehensive analysis, comparing temperature variations and taxi frequency across time, days, districts, and boroughs. This allowed us to formulate hypotheses on advantageous locations and optimal times for taxi trips. We developed a neural network model predicting ride fares based on predetermined features known to drivers. This model assists drivers in making informed decisions for more profitable trips.

The subsequent sections will detail our proposed analysis and methodology, outlining the datasets used to achieve research objectives. The document is structured into Proposed Analysis and Methods, Dataset, Goals, and References.

2 Datasets

For our analysis purposes, we required variety of datasets across different domains. This included data on taxi trip in NYC, crime statistics in NYC, weather data of NYC and demographic data. This section provides a detailed description of the datasets employed and highlights the significant updates made since our proposal's last iteration.

2.1 Dataset Description

1. **NYC TLC Dataset:**The NYC TLC Dataset[2] records various attributes, including pick-up and drop-off dates/times, locations, trip distances, fares, rate types, payment methods, and driver-reported passenger counts. For our project, we have updated the dataset from 2018 to 2022, encompassing four types of taxi trip records: yellow taxi, green taxi, FHV (For Hire Vehicle)[6], and FHVHV (High Volume For Hire Vehicle)[7]. Each taxi type in this dataset possesses distinct characteristics that differentiate them from one another.

2. **Taxi Zone Dataset:** The NYC Taxi and Limousine Commission (TLC) data mentioned above include pick-up and drop-off locations, identified by numerical values ranging from 1 to 263[8]. These numerical identifiers correspond to taxi zones and include geometric information for each taxi zone. Additionally, the data provides a comprehensive list of TLC taxi zone location IDs, location names, and the respective boroughs associated with each zone. By integrating this dataset with NYC TLC data, a more detailed insight into the operational locations and areas frequented by these taxis is obtained.

3. **NYC Weather Dataset:** The weather data [3], depicting year-long weather conditions in the Central Park Area of

NYC, was sourced from Kaggle. This dataset is instrumental in observing the comprehensive weather patterns across New York City. It encompasses essential features such as temperature, precipitation, and wind speed. The primary objective of utilizing this dataset is to investigate and understand the correlation between weather conditions and taxi usage in NYC.

4. **NYPD Crime Statistics Data:** Utilizing NYPD crime statistics across various transit districts provided insights into the patterns of taxi and cab usage among individuals. This includes monthly and annual reports which contain the number of complaints and arrests within the transit system and on buses disaggregated by transit district and precinct.

5. **Demographic Data:**

2.2 Data Preprocessing

In any data science project, data preprocessing stands out as a pivotal task. In the context of our project, we endeavored to incorporate data from various sources, a challenging task given the disparate formats and timelines of the initial datasets. The following tasks outline our efforts in curating the dataset:

1. **Merging and Filtering Data:** Focused on consolidating data for the year 2022, we addressed the monthly format of the NYC TLC data by scripting in Python using the Pandas framework to merge monthly data into a yearly format across all taxi types. Simultaneously, we filtered the weather dataset, originally spanning 2016-2022, to exclusively include data from the year 2022. Regarding crime statistics from the NYPD, the detailed data encompassed precinct-level statistics for each transit district, prompting us to incorporate data for all precincts categorized under each district. Additionally filtering the required features from the dataset in order to correlate it with the trip frequency in different boroughs.

2. **Data Entry and Manipulations:** To unite disparate datasets based on common parameters, we ensured consistent formatting of time records. Manipulations involved aligning time records to facilitate the joining of NYC TLC data and weather data. Additionally, lacking direct online data for the relationship between each NYC TLC zone and transit district, we manually updated the NYC TLC lookup table with corresponding transit district information for each zone.

3. **Joining Datasets:** The final stage of data preprocessing involved integrating the various datasets. Initially, we joined the NYC TLC data and NYC TLC lookup table based on zone IDs, providing trip records in each zone and transit district. Subsequently, we combined this curated data with NYC weather data over time, yielding trip records in each zone and transit district with corresponding weather conditions at specific times.

4. **Chunking Datasets:** One of the prominent challenges encountered in our data processing endeavors was the need to aggregate substantial volumes of big data within the constraints of limited resources, specifically an 8GB RAM capacity. In order to address this challenge effectively, a strategic approach involving the segmentation of data into manageable chunks was adopted. This facilitated the sequential processing of data in batches during the training of our model, thereby ensuring the preservation of data integrity throughout the entire procedure.

3 Analysis and Modeling

The primary goal of our project is to conduct a comprehensive analysis of NYC cab services, with a specific focus on identifying optimal areas for cab drivers to enhance their profitability. During the initial phase, our efforts were concentrated on completing a substantial portion of the comparative analysis as outlined in the proposal. However, in the second phase, our project shifted its focus towards the development of a neural network model designed to predict trip fares based on given features. The subsequent sections will provide a detailed account of the analysis conducted and the methodologies employed in the project.

3.1 Comparative Analysis and Visualizations

We have undertaken a comprehensive comparative analysis of diverse private taxi services, assessing them across various parameters. The findings derived from these analyses will inform the identification of key features for the development of a statistical model. The ultimate objective is to construct a predictive model that can identify optimal locations for taxi drivers, thereby enhancing opportunities for earning tips.

1. **Geospatial Analysis:** This analysis explores the pickup and drop-off patterns of yellow taxis, green taxis, and For-Hire Vehicles (FHVs) in New York City, utilizing data sourced from The New York City Taxi and Limousine

Commission (TLC). The primary objective is to derive insights into transportation trends and spatial dynamics. The overarching hypothesis being examined is the notion that green taxis are confined to picking up passengers exclusively in Northern Manhattan (north of West 110th Street and East 96th Street), while yellow taxis enjoy more expansive pickup privileges.

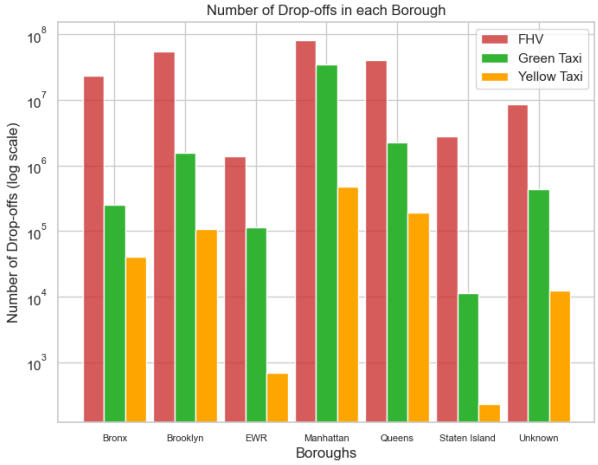

Fig 1: Number of Drop-off in Boroughs

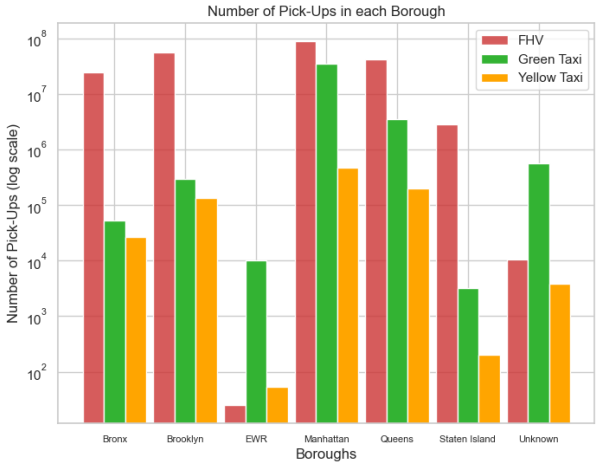

Fig 2: Number of Pick-Ups in Boroughs

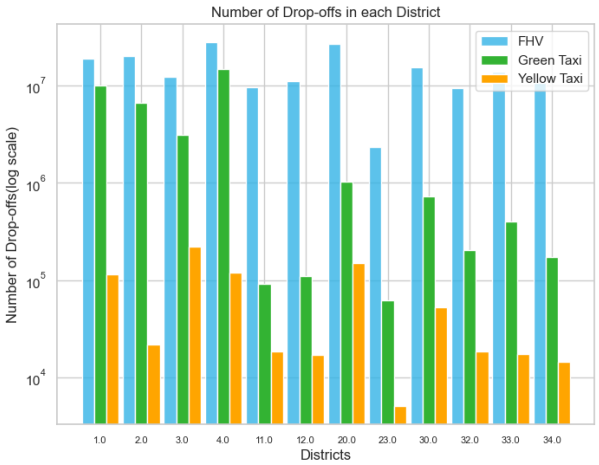

Fig 3: Number of Drop-off in Transit Districts

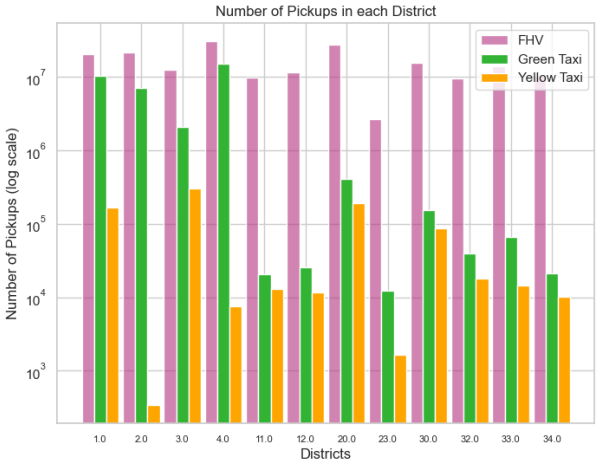

Fig 4: Number of Pick-Ups in Transit Districts

Based on Figures 1 and 2, it is evident that Manhattan surpasses other boroughs in NYC in the utilization of cab services. Despite EWR not being a borough, it is considered a separate zone for TLC, exhibiting the lowest usage of taxi services, following Staten Island. Additionally, a significant disparity is observed between the usage of FHVs, green and yellow taxis. Similar patterns emerge in Figures 3 and 4, where transit districts 1, 2, 3, and 4, falling under Manhattan, show higher taxi usage. Moreover, there is a notable preference for green taxis across all districts, despite their limited pickup authorization in Northern Manhattan.

2. **Time and Days Analysis**: In this part of examination, our aim is to scrutinize the relationship between time and the

days of the week. Through this analysis, we seek to gain insights into optimal times and days for cab drivers.

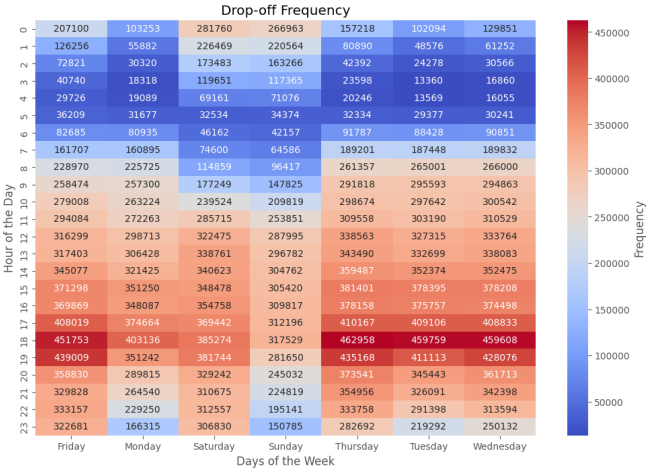

Fig 5: Drop-off Frequencies per Day per Hour for Green/Yellow Taxis

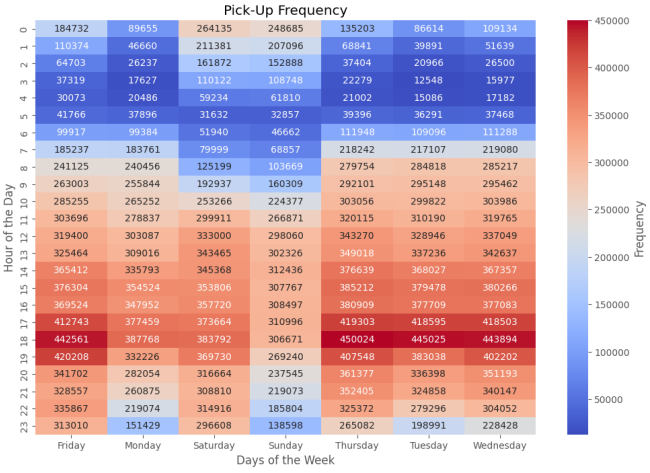

Fig 6: Pick-up Frequencies per Day per Hour for Green/Yellow Taxis

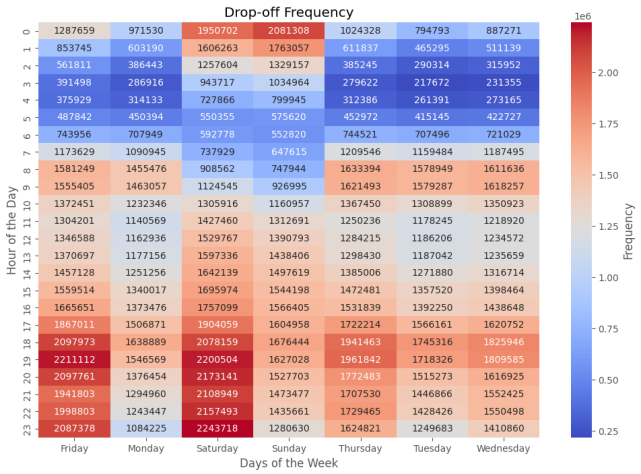

Fig 7: Drop-off Frequencies per Day per Hour for FHVs

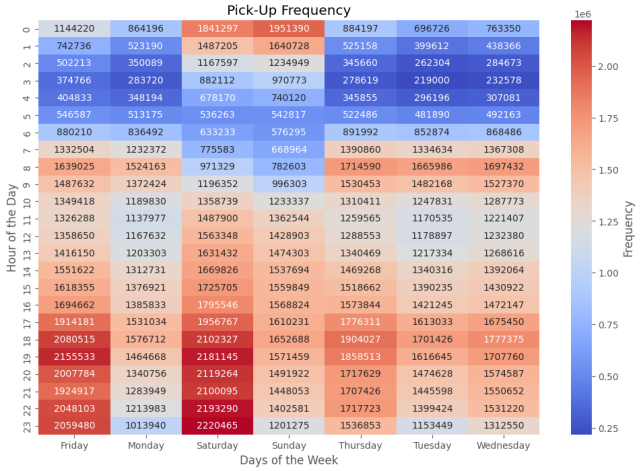

Fig 8: Pick-Up Frequencies per Day per Hour for FHVs

Figures 5 and 6 illustrate that there are approximately 400,000 pickups and drop-offs between 5 pm and 7 pm, particularly on weekdays. This suggests that individuals are utilizing taxi services to reach their homes after a lengthy and exhausting day at work. Additionally, there is a notable increase in trip frequency around midnight on Saturdays and Sundays, indicating that people are decompressing after a demanding workweek and responsibly opting for taxi travel. Conversely, the least frequency is observed between 4 am and 6 am. These figures distinctly portray the patterns of cab service utilization throughout the entire week. Based on this we can say 5pm - 7pm is good time for cab drivers to make profit. Figures 7 and 8, conversely, depict a noteworthy change in commuters' taxi

preferences, particularly on Fridays and Saturdays from 10 PM to 12 AM. This shift may be attributed to the heightened social activities and nightlife during these days in the city. The increased demand for transportation during late-night hours is influenced by the desire for secure and flexible travel options as individuals extend their outings over the weekends. In summary, For-Hire Vehicles (FHVs), characterized by their extended service hours, offer a reliable and convenient choice for late-night transportation.

3. **Weather and special events**: Weather and special events can profoundly influence individuals' travel preferences. Adverse weather conditions, for example, may prompt more individuals to opt for the shelter and convenience offered by cabs, while special events could result in heightened demand for cab services due to increased traffic congestion and limited public transportation options. These elements play a crucial role in shaping the decision-making process of cab users. To analyze the weather's impact, we will utilize Central Park weather records, providing essential insights into how travel preferences evolve with varying weather conditions and special events. This data serves as a valuable resource for comprehending the influence of these factors on people's transportation choices.

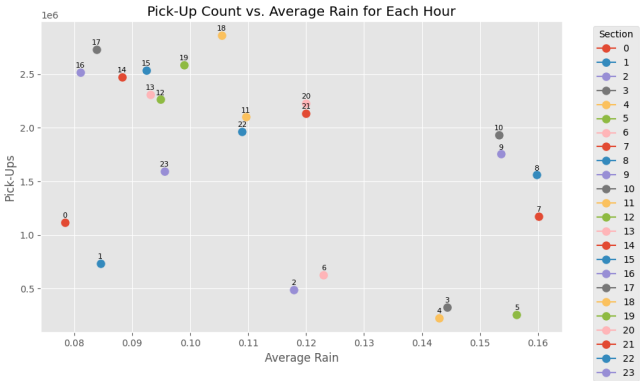

Fig 9: Pick-Up Count vs Average Rain per Hour

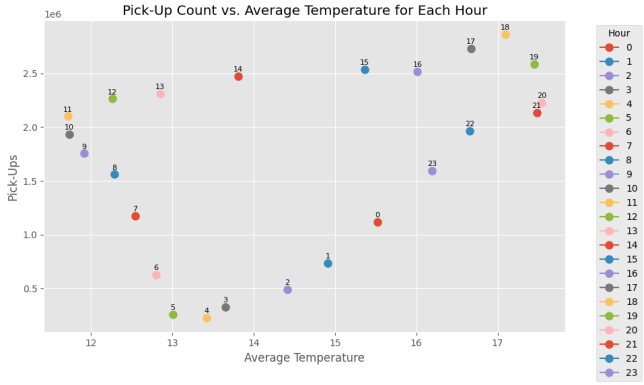

Fig 10: Pick-Up Count vs Average Temperature per Hour

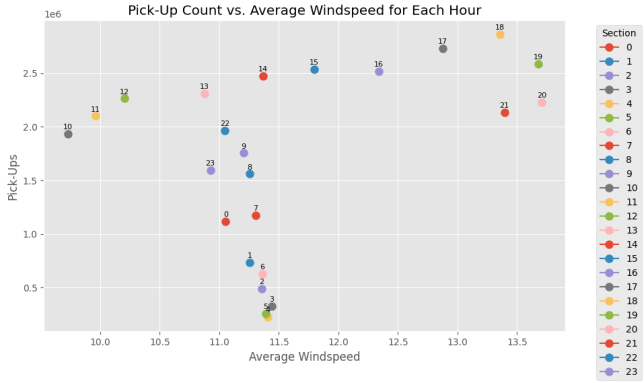

Fig 11: Pick-Up Count vs Average Windspeed per Hour

Figure 10 illustrates a temperature decrease between midnight and 12 am, coinciding with a drop in pick-up counts. However, even as the temperature decreases from 5 am to 10 am, there is an increase in pick-ups, suggesting people still need transportation for work during these times, regardless of the temperature. A similar trend is observed in Figure 11, where, despite average rainfall exceeding 0.15 mm between 7 am and 10 am, pick-up counts remain high, indicating continued travel during these hours. Notably, from 12 pm to 7 pm, there is a substantial increase in pick-ups with less rain. Figure 9 also reveals an increase in pick-up counts with higher wind speed from 10 am to 9 pm. However, between 10 pm and 3 am, the count decreases, resuming an upward trend from 4 am to 9 am with constant wind speed. Despite varying weather conditions, the data in Figure 11 underscores the consistency in morning taxi service frequency.

4. **Safety Analysis** : Safety is a crucial factor influencing transportation choices, and assessing subway crime data

is essential for identifying stations with elevated reported crime rates. By correlating this information with taxi pick-up and drop-off data in these areas, we can discern whether individuals are opting for more cab services, potentially to evade public transportation at these specific stations due to safety concerns. Our research relies on NYPD crime statistics, specifically focusing on the number of crimes in buses and public transit. This dataset will serve as a pivotal tool in our examination of crime and transportation trends in New York City[1].

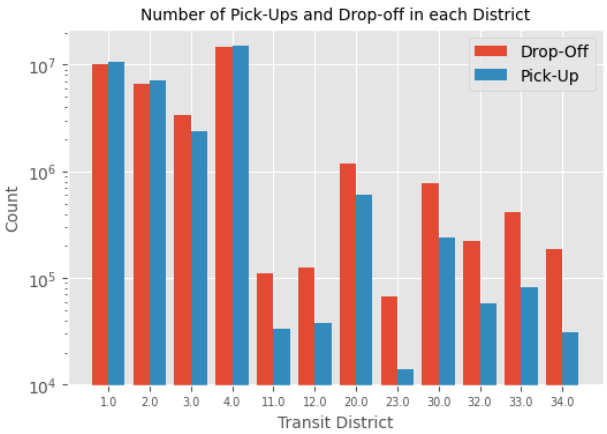

Fig 13: Number of Pick-Up and Drop-off per District

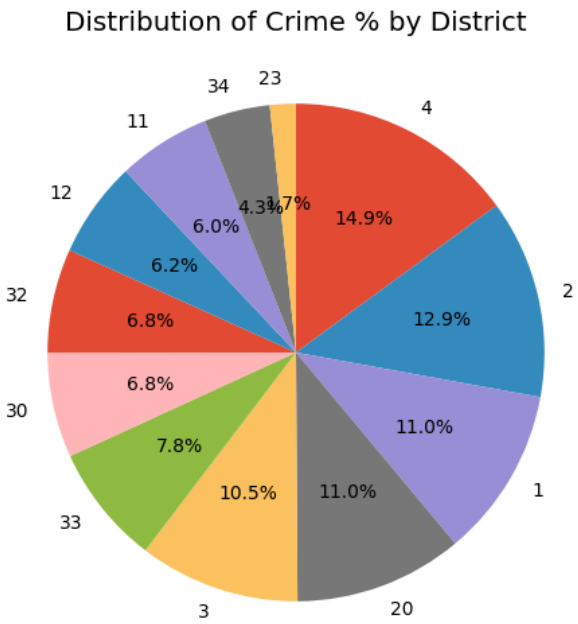

Fig 12: Distribution of Crime Rate by District

Figure 12 depicts a notable discrepancy in crime rates within transit and buses, particularly in districts 1, 2, 4, and 20, with rates recorded at 14.9%, 12.9%, 11.0%, 11.0%, and 10.5%, respectively. Notably, districts 1, 2, and 4 are situated in Manhattan, an area previously identified for its heightened taxi frequency. This correlation is supported by the data presented in Figure 13, which confirms the highest taxi frequency in districts 1, 2, 3, 4, and 20. District 20, located in Queens, exhibits significantly greater taxi usage compared to District 23, despite the latter having the lowest recorded crime rate at 1.7%. This observation implies a tendency for increased reliance on taxi services over public transport in areas with elevated crime rates, highlighting a noteworthy transportation preference dynamic influenced by safety concerns. Figure 12 reveals a similar pattern for FHVs as observed with Green and Yellow taxis.

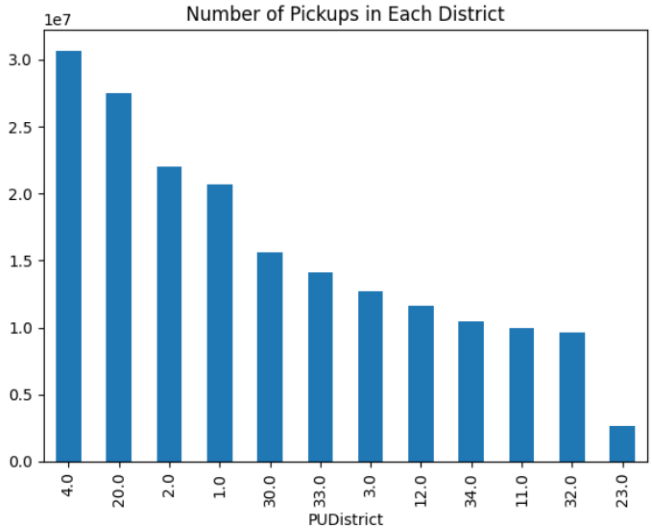

Fig 14: FHV highest Pickup District

5. **Analysis of Cross Borough Trips** :

Figure 15 illustrates that Manhattan exhibits the highest volume of pickups for inter-borough journeys. This phenomenon can be attributed to several factors, including Manhattan's central geographical location, the presence of numerous business districts, and its role as a prominent entertainment hub. This observation aligns with our prior analysis, reinforcing the conclusion that Manhattan stands as the borough with the highest level of activity among all.

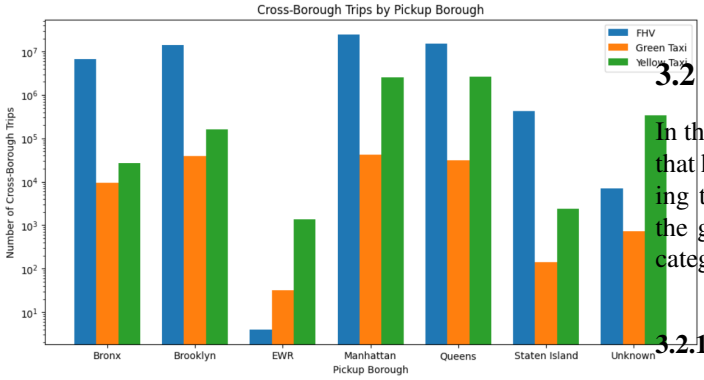

Fig 15: Cross Borough Trips

6. **Demographics** :

Observing Figure 16, it becomes evident that Brooklyn, Queens, and Manhattan stand out as the most populous boroughs in New York City. Consistent with our prior analyses, these three boroughs also emerge as the most frequented areas for taxi services. Notably, although Manhattan may not hold the top position in terms of population, it exhibits the highest taxi usage. This phenomenon could be attributed to the prevalence of numerous offices, restaurants, and other establishments that attract significant foot traffic in the borough.

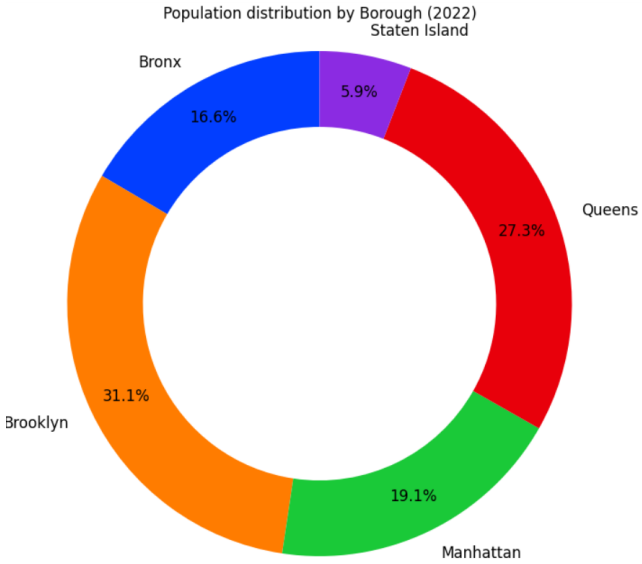

Fig 16: Population Distribution of NYC

3.2 Modelling

In this section, we will expound upon the neural network model that has been meticulously developed for the purpose of predicting trip prices across various taxi datasets, including those of the green, yellow, and high-volume for-hire vehicle (HVfHV) categories.

3.2.1 Feature Engineering

While the majority of features across these datasets were consistent, certain distinct features necessitated engineering to construct the training dataset. The training features for each dataset are delineated below:

- **Yellow Taxi :** 'VendorID', 'tripTime','tripDistance', 'passengerCount','PULocationID', 'DOLocationID', 'RatecodeID', "dayOfWeek", "hourOfDay" , "tripFair" .
- **Green Taxi :** 'VendorID', 'tripTime','tripDistance', 'passengerCount','PULocationID', 'DOLocationID', 'RatecodeID', "dayOfWeek", "hourOfDay" ,"tripType", "tripFair"
- **HvFhv Taxi :** 'VendorID', 'tripTime','tripDistance', 'passengerCount','PULocationID', 'DOLocationID', 'RatecodeID', "dayOfWeek", "hourOfDay" ,"tripType", "tripFair"

While there exists varying input features, the singular shared feature under prediction is the "trip fare." All the aforementioned features are utilized in the training of our model.

3.2.2 Architecture and Parameters

We employed a five-layer linear neural network for the task of predicting trip fares, aiming to capture the intricate relationships within the dataset's features. Given the absence of a straightforward linear relationship among these features, a multi-layer neural network was deemed suitable for this purpose. The architectural configuration of the model is depicted in the accompanying figure.

During the training process and parameter optimization, specific layer configurations and neuron counts were employed in the network. Additionally, key parameters such as the mean squared error (squared L2 norm) were utilized as the loss function. The Adam Optimizer, configured with a learning rate of 0.001, facilitated the optimization process, and a batch size of 128 was employed for training efficiency.

3.2.3 Results

The performance metrics, including Loss, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), for three distinct models across three datasets are presented in Table 1. The reported scores, achieved through extensive fine-tuning, indicate the efficacy of these models in predicting trip fares based on a limited set of input parameters. The scores demonstrate the models' proficiency in predicting trip fares. However, it is worth noting that the inclusion of additional features, such as weather data, has the potential to enhance the relationships among the variables, thereby leading to more accurate predictions. This suggests an avenue for further improvement in predictive accuracy by expanding the feature set. Understanding these scores is crucial for optimizing driver profit. A lower MAE and RMSE generally indicate more accurate fare predictions, supporting drivers in maximizing profits by efficiently setting fares based on reliable model outputs. The choice between models should consider the specific trade-offs in loss functions and the emphasis on accuracy or precision within the context of profit optimization.

Models	Loss	MAE	RMSE
Yellow-NN	8.0831	3.4024	11.1364
Green-NN	4.1184	2.3983	9.2765
HvFhv-NN	13.0831	2.5264	6.5234

Table 1: Result and Scores

4 Conclusion

In conclusion, following a thorough analysis encompassing all three taxi services, it is evident that For-Hire Vehicles (FHVs) consistently demonstrate the highest frequency of pickups and drop-offs across nearly every borough and district, surpassing corresponding figures for both green and yellow taxis. This analytical insight presents an opportunity to glean valuable information regarding prevalent transportation trends and spatial dynamics within the city. Specifically, Manhattan emerges as the borough with the highest traffic volume, attributed to factors such as office spaces, restaurants, and other establishments.

Furthermore, our examination reveals that areas experiencing higher rates of crimes on public transit prompt individuals to opt for taxis as a safer mode of transportation. Weather conditions also play a pivotal role in influencing transportation choices, with notable variations during specific time windows. Notably, during morning hours (6 AM-10 AM) and evening hours (5 PM-7 PM), individuals exhibit heightened taxi usage, likely due to commuting to schools or offices.

Additionally, our analysis underscores increased taxi usage during late nights on weekends. The developed model holds significant utility, as it empowers drivers to consider various trip parameters and identify potentially profitable routes. Notably, each model tailored for different types of taxis demonstrates commendable performance, thereby providing drivers with a robust tool to enhance their profitability.

5 References

1. https://www.nyc.gov/site/nypd/stats/crime-statistics/borough-and-precinct-crime-stats.page
2. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
3. https://www.kaggle.com/datasets/ecboxer/nyc-weather
4. https://www.researchgate.net/publication/335504977_Exploring_the_Taxi_and_Uber_Demand_in_New_York_City_An_Empirical_Analysis_and_Spatial_Modeling
5. https://medium.com/@haonanzhong/new-york-city-taxi-data-analysis-286e08b174a1
6. https://www.nyc.gov/assets/tlc/downloads/pdf/fhv_congestion_study_report.pdf
7. https://catalog.data.gov/dataset?q=High+Volume+FHV+trip+records&sort=score+desc
8. https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc