Proposal

Demand Prediction

summer 2023

CHAUFFERUR

RAHNEMA COLLEGE

New folder

Table of Contents

1. Introduction

1.1 Background and context

The rise of online taxi services, epitomized by industry giants like Snapp, Tapsi and ... , has transformed the way people travel, revolutionizing the transportation landscape. These platforms have provided individuals with a convenient and reliable means of transportation, offering quick and efficient rides at the touch of a button. The success of such services is largely attributed to their ability to seamlessly connect passengers with available drivers in real-time, creating a dynamic and responsive transportation ecosystem.

However, as the popularity of online taxi services continues to soar, the need for innovative solutions to address the challenges associated with fluctuating demand patterns becomes increasingly evident. For service providers like CHAUFFERUR, accurately predicting and meeting customer demand is crucial for ensuring high levels of customer satisfaction, operational efficiency, and driver utilization.

Traditionally, the transportation industry has relied on historical data, experience-based decision-making, and intuition to manage supply and demand. However, these conventional approaches are limited in their ability to capture the intricate dynamics of the market. Demand for rides can be influenced by a multitude of factors, including time of day, day of the week, weather conditions, local events, and even traffic patterns. Without a comprehensive understanding of these variables and their impact on demand, service providers may struggle to allocate resources optimally, leading to issues such as excessive wait times, driver unavailability, and compromised customer experiences.

To address these challenges and unlock new opportunities for growth, we propose the implementation of advanced predictive analytics to accurately forecast demand for our online taxi service. By harnessing the power of data, we can develop sophisticated models that leverage historical trip data, external factors, and real-time insights to generate accurate demand predictions. Such predictive capabilities will enable us to proactively respond to fluctuations in demand, strategically position drivers, and provide a seamless experience for our passengers.

In summary, the incorporation of demand prediction capabilities through advanced analytics represents a significant step forward for our online taxi service. By understanding the contextual factors that influence customer demand and leveraging data-driven insights, we can revolutionize the way we operate, ensuring that our passengers receive prompt and reliable transportation services while maximizing the efficiency of our driver network. Through this proposal, we aim to position ourselves at the forefront of innovation in the transportation industry and set new benchmarks for service quality and customer satisfaction.

1.2 Problem statement

In the past months, the marketing team has been working on providing incentive plans for chauffeur drivers by predicting demand at specific times. This is to ensure that the company avoids supply shortages in different regions. Pricing in the realm of internet taxis has always been a challenging issue. Determining the trip fare requires finding a balance where the passenger is willing to request the trip, while also ensuring the price is reasonable for the driver. This conflict of interest often complicates the pricing process. The marketing team aims to enhance the value of prices for drivers by offering incentive schemes.

As part of this plan, the team seeks to identify areas with higher travel demand to increase the number of available drivers in those locations. Incentivizing drivers is a marketing initiative, and predicting demand during different time frames is crucial for implementing this process. If it becomes possible to forecast high-demand hours or other peak periods within a day, the proposed process can be effectively put into action.

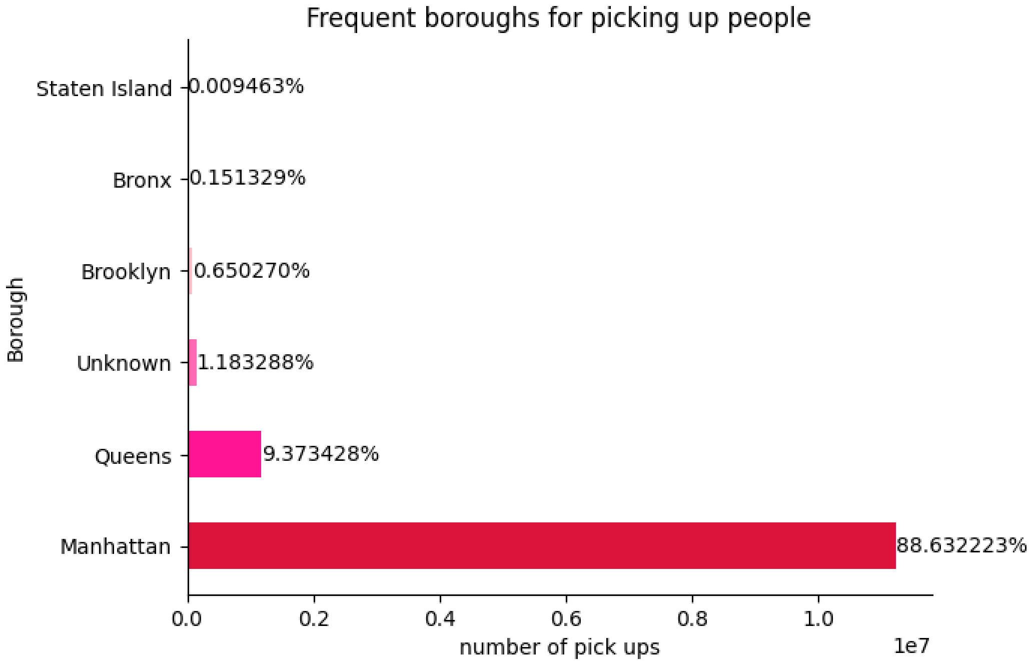
1.3 Importance and relevance of the project

This project offers the capability to predict demand in various time periods and urban areas, enabling the implementation of incentive plans and other marketing programs. Ultimately, it aims to increase the company's profitability while adopting a fair and intelligent pricing approach. By utilizing this method, the project can ensure a high level of accuracy in determining a fair price for both passengers and drivers, thus benefiting all parties involved in the trip.

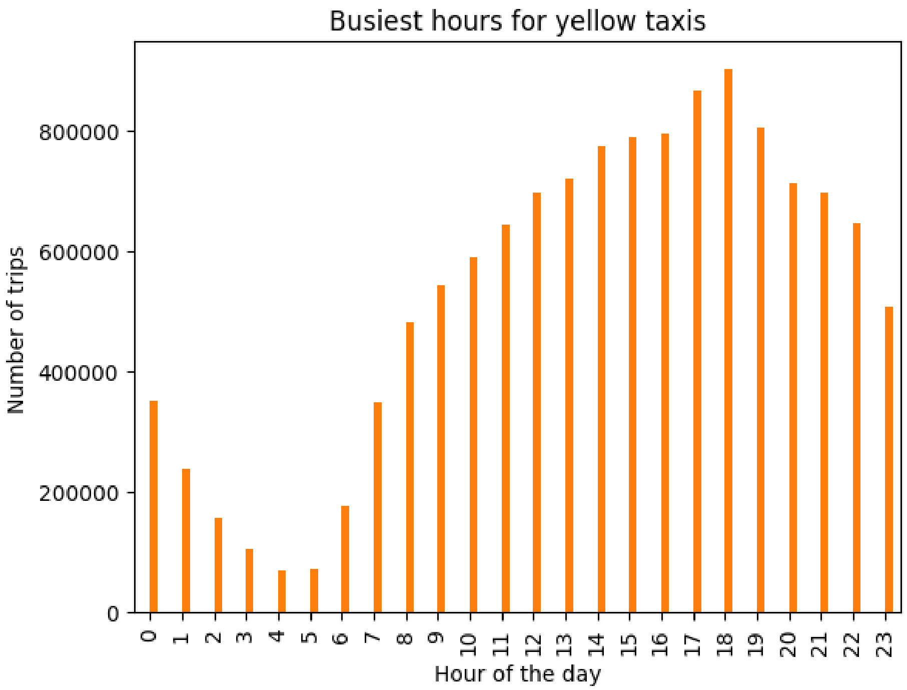
1.4 Data Review

Our project's main goal is to perform demand prediction on Chaufferur dataset that collected in 4 first months of 2023. This dataset contains the trip data about 12,672,737 trips which are made between 262 different zones in NYC. The trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The pick-up and drop-off contains the date and time for each trip respectively which are about trips made between 2023/1/1 to 2023/4/30 and the times follow the HH:MM:SS format even so the are about 62 records in dataset which are dated between 2001 - 2022 which are considered to be outlier data.

The pick-up and drop-off locations are mainly the number indicating a zone in NYC and each of these zones belong to a specific borough. Trip distance is the elapsed trip distance in miles reported by the taximeter and passenger count is the number of passengers in the vehicle. Rate types demonstrate the final rate code in effect at the end of the trip and can have 6 different values, which are 1 to 6 and mean Standard rate, JFK, Newark, Nassau or Westchester, Negotiated fare and Group ride respectively. Furthermore, payment type is a numeric code signifying how the passenger paid for the trip which can indicate Credit card, Cash, No Charge, Dispute, Unknown and Voided trip. Itemized fares are features of data which contain the The time-and-distance fare calculated by the meter, miscellaneous extras and surcharges, MTA tax, improvement surcharge, tip amount, tolls amount, Total amount collected in trip for NYS congestion surcharge, airport fee and the total amount charged to passengers per trip.

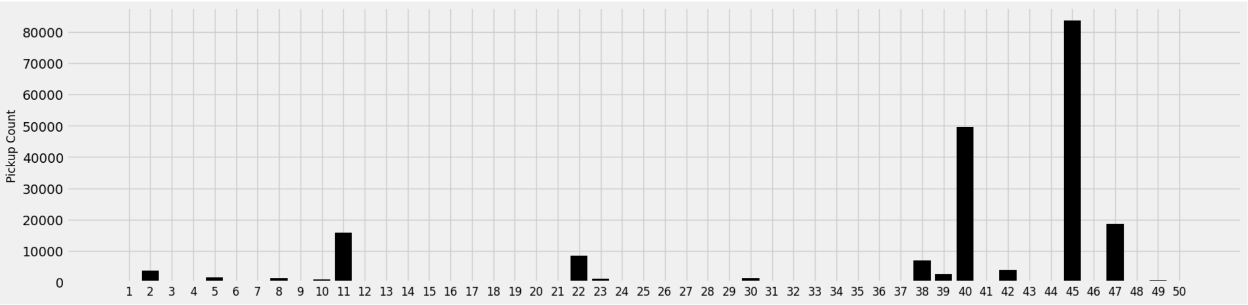


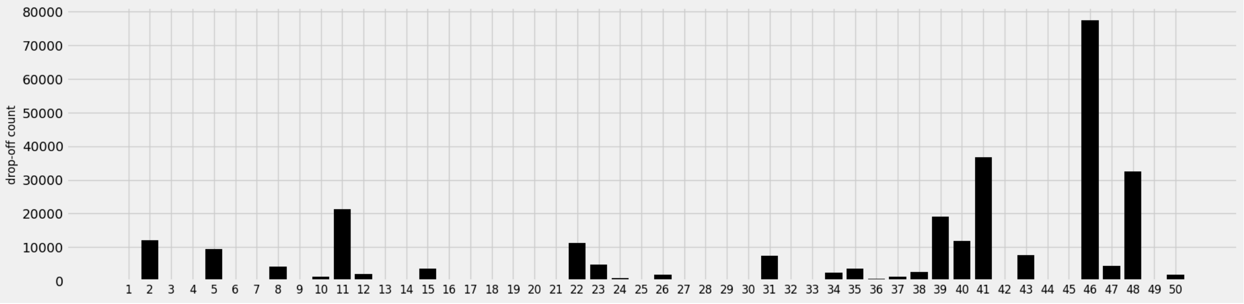
It is shown that 88.6% of trips start in Manhattan. Also 93.5% of trips that started in Manhattan will be finished in Manhattan.



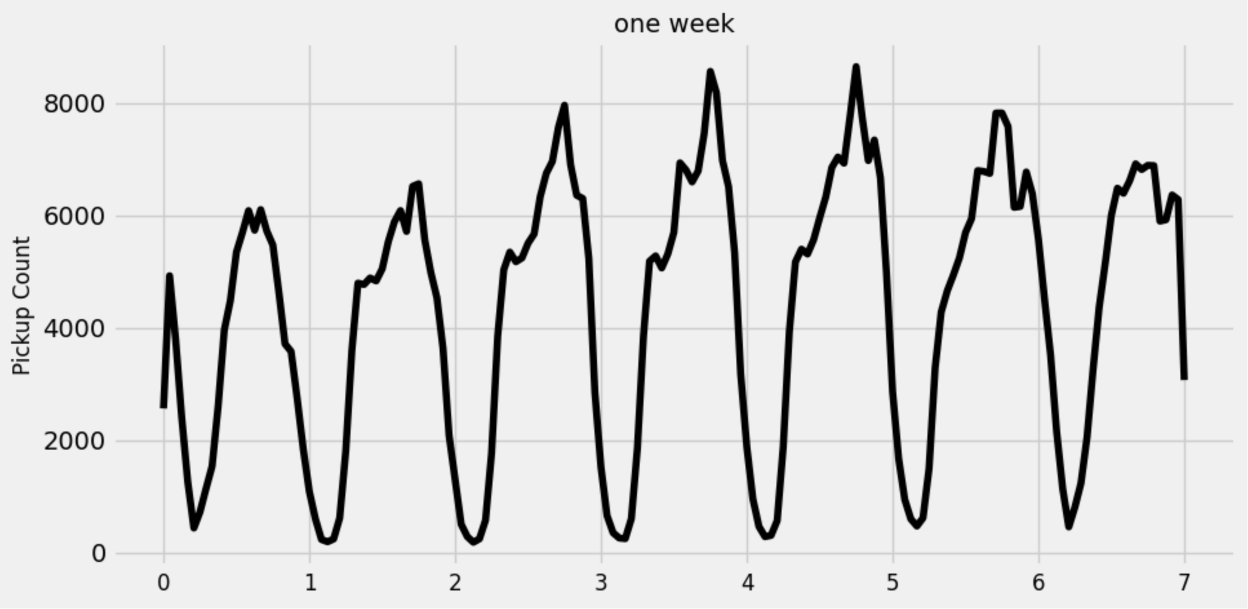
Also the coefficient of correlation between pickup borough and dropoff borough is 0.3, indicating that the connection between these two variables is weak. So there must be many factors that influence where people choose to be picked up and dropped off ,such as distance, time of day and traffic conditions.
The coefficient of correlation between pickup zone and dropoff zone is even lower, only 0.08.

The following histograms depict that the number of pickups and drop-offs in certain districts can be extremely low, and in some cases, the difference between the two can be substantial. For instance, in the 45th district, the number of pickups exceeded 8,000, while the number of drop-offs was less than 500.

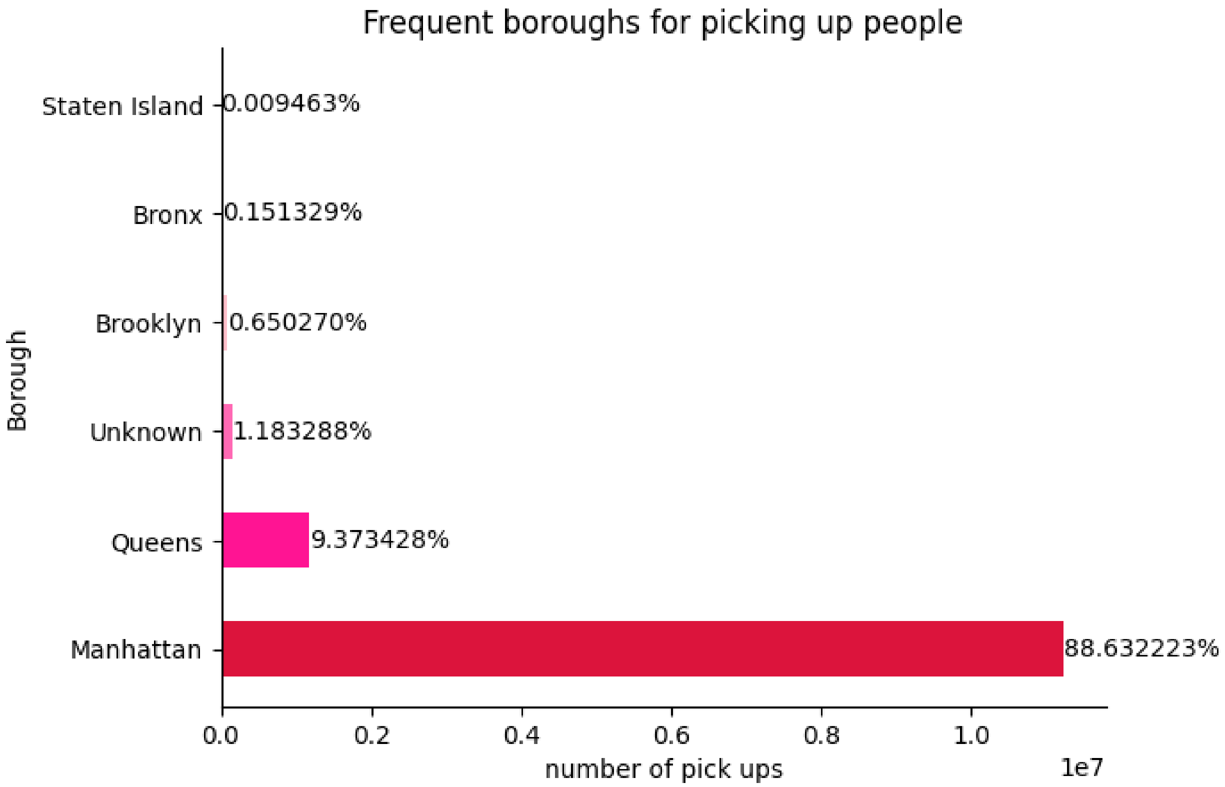


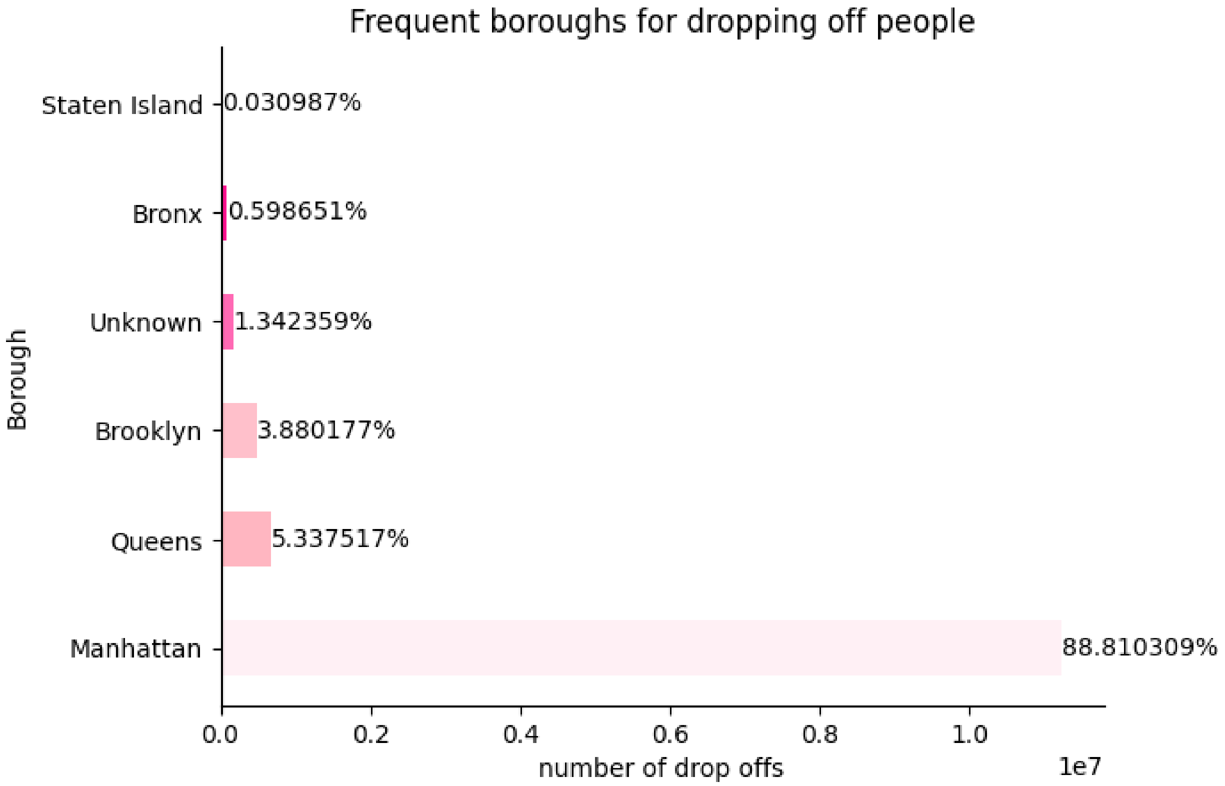


To provide a better understanding, we analyzed the number of pickups on a weekly basis. The next graph displays the frequency of pickups throughout the week and indicates that the highest number of rides were provided on the fourth and fifth days. This trend remains consistent across other weeks as well.



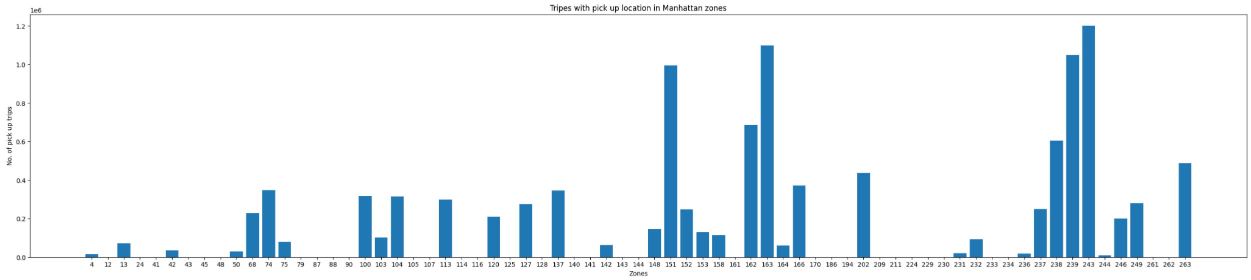
An extensive look to the data shows us that in total there are 3 zones which no pick ups has happened in any of them which are 103,104 which **both are islands** and 110 and about 4 zones which no drop off has happened in them which are 103,104,199 which are **islands** and 110. Now let's look at the pickups and drop offs in each borough. First of all Manhattan has the most number of pickups and drop offs in comparison to other boroughs:

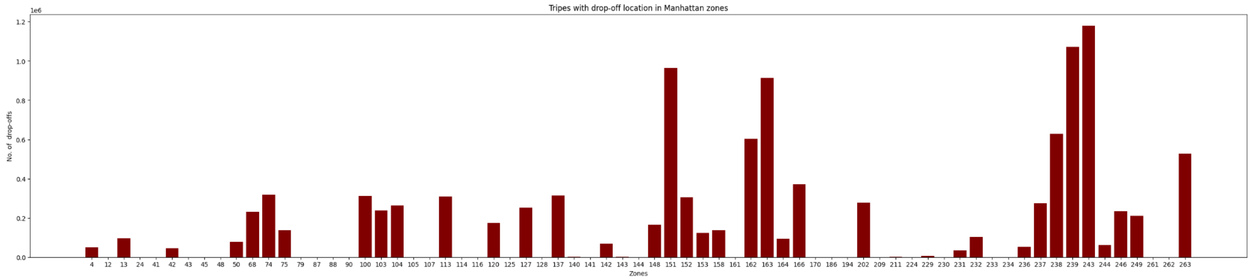

Frequent boroughs for picking up people
Borough
Staten Island
Bronx
Brooklyn
Unknown
Queens
Manhattan
0.009463%
0.151329%
0.650270%
1.183288%
9.373428%
88.632223%
0.0
0.2
0.4
0.6
0.8
1.0
1e7
number of pick ups

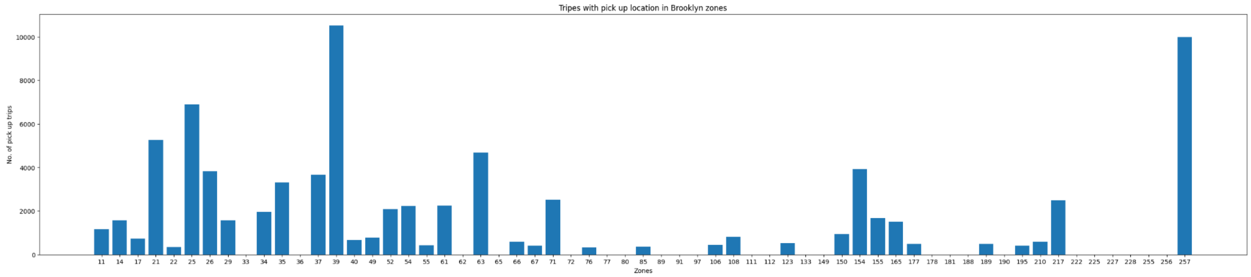

Frequent boroughs for dropping off people
Borough
Staten Island
Bronx
Unknown
Brooklyn
Queens
Manhattan
0.030987%
0.598651%
1.342359%
3.880177%
5.337517%
88.810309%
0.0
0.2
0.4
0.6
0.8
1.0
1e7
number of drop offs

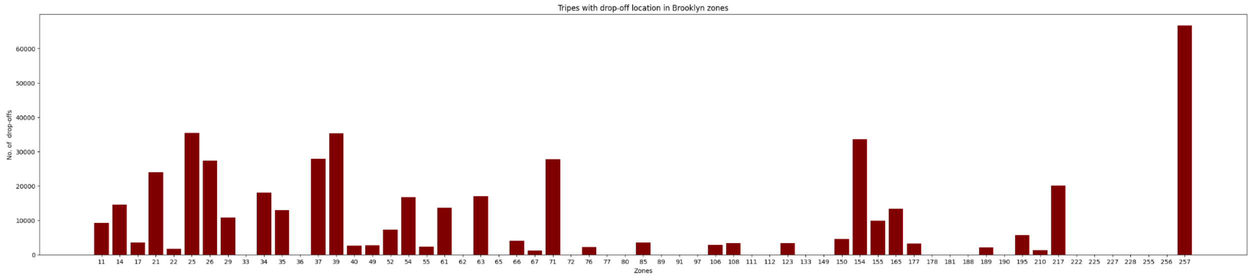
Which contains about 88 percent of each action. After Manhattan, Queens stand as second for both actions while the unknown is third place in the number of pickups but Brooklyn stands in the third place in the number of drop offs, as result data with drop off or pick ups in the unknown borough cannot be ignored. The only zones that should be considered as wrong ones are the ones that end in an island or strats in one, since our data contains trips made by cars.

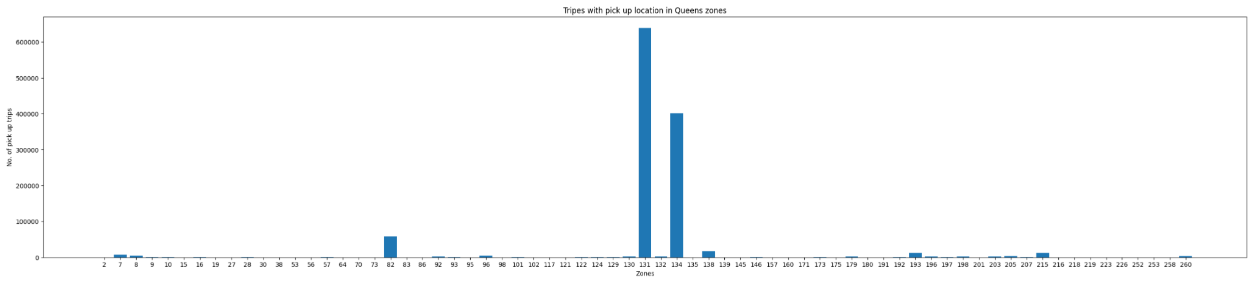
The density of trips in different boroughs vary from one another too. The figures below illustrate the number of trips made in different zones in each boroughs:

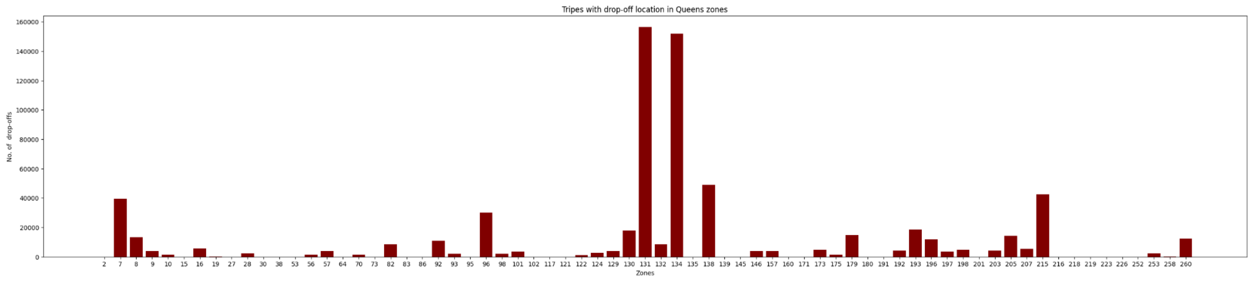

Tripes with pick up location in Manhattan zones

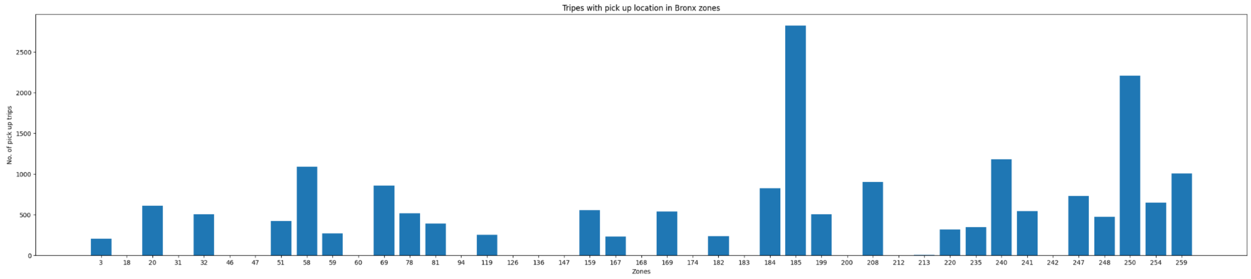

Tripes with drop-off location in Manhattan zones

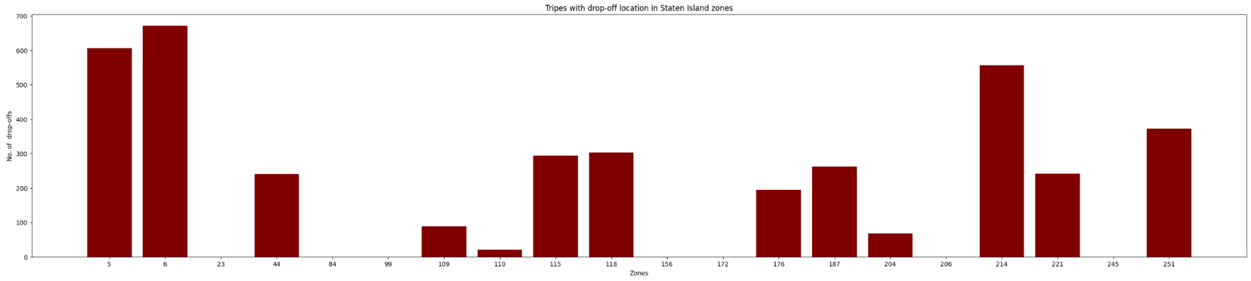

Tripes with pick up location in Brooklyn zones


Tripes with drop-off location in Brooklyn zones


Tripes with pick up location in Queens zones
No. of pick up trips
0
100000
200000
300000
400000
500000
600000
2 7 8 9 10 15 16 19 27 28 30 38 53 56 57 64 70 73 82 83 86 92 93 95 96 98 101 102 117 121 122 124 129 130 131 132 134 135 138 139 145 146 157 160 171 173 175 179 180 191 192 193 196 197 198 201 203 205 207 215 216 218 219 223 226 252 253 258 260
Zones


Tripes with drop-off location in Queens zones
No. of drop-offs
0
20000
40000
60000
80000
100000
120000
140000
160000
2 7 8 9 10 15 16 19 27 28 30 38 53 56 57 64 70 73 82 83 86 92 93 95 96 98 101 102 117 121 122 124 129 130 131 132 134 135 138 139 145 146 157 160 171 173 175 179 180 191 192 193 196 197 198 201 203 205 207 215 216 218 219 223 226 252 253 258 260
Zones


Tripes with pick up location in Bronx zones
No. of pick up trips
0
500
1000
1500
2000
2500
3 18 20 31 32 46 47 51 58 59 60 69 78 81 94 119 126 136 147 159 167 168 169 174 182 183 184 185 199 200 208 212 213 220 235 240 241 242 247 248 250 254 259
Zones


Tripes with drop-off location in Staten Island zones
No. of drop-offs
0
100
200
300
400
500
600
700
5 6 23 44 84 99 109 110 115 118 156 172 176 187 204 206 214 221 245 251
Zones

Also, this matrix shows what percentage of trips that started from one Borough went to other Boroughs.

	Manhattan	Brooklyn	Queens	Bronx	Staten Island	Unknown
Manhattan	93.27	37.38	56.70	32.62	22.35	25.15
Brooklyn	2.40	49.25	14.66	7.78	22.85	2.74
Queens	3.36	10.28	23.56	8.20	15.09	4.66
Bronx	0.35	1.84	2.02	49.92	5.92	0.54
Staten Island	0.01	0.42	00.12	0.33	32.36	00.05
Unknown	0.32	0.63	2.75	1.10	1.08	0.66

1.5 Assumptions

1.5.1 Demand

Demand can be contained with a series of definitions. All the pickup locations which are recorded in our dataset can be viewed as a demand for each location and zone. Another thing to consider is the drop off locations which may indicate that those who are dropped at one location may need to be picked up at a later time.
We defined demand of a zone in a 10-min timestamp as:

$$demand = passenger_{count} + (dropoff_{last6hours} - pickup_{last6hours})$$

With this function, demand is defined with pickups count and passengers of that zone in that specific time, and also the number of people who are still in that zone. There is a possibility that they want to travel to the other zones with a taxi. So we add those to our demand. If “dropoff - pickup” becomes negative, because of the relu function we will consider it as zero.

1.5.2 Timestamp

The mean duration(dropoff time - pickup time) of 72% of trips is around 10 minutes. Also the mean duration of trips that are in Manhattan, is near to 15 minutes. As a result We can use a 10 minute time stamp for predicting demand in our project.

2. Objectives

2.1 Primary objectives

Pricing in online taxis is a complicated issue. The price must be determined in such a way that the passenger requests a trip and the driver accepts the trip at the same time. It is also important that the number of drivers is proportional to the number of travel applicants in each area.

Sometimes, this conflict of interests between the passenger and the driver causes the pricing issue to be challenged. Therefore, this problem should be solved by creating incentives for drivers.

In this issue, we aim to predict the demand in each region and specific time period by building and developing a machine learning model. This machine learning model should be capable of predicting demand with acceptable accuracy in certain time intervals so that its results can be used for future analysis.

2.2 Expected outcomes and benefits

The expected output of this project is a machine-learning product capable of predicting demand by considering various features within a specific time frame and location. This model can assist the marketing team in devising incentive plans.

3. Proposed Timeline

3.1 Milestones and deliverables

Data Preprocessing & Exploratory Data Analysis

- Exploratory Data Analysis & Identify Patterns
- Choose target and feature variables
- Noise, delusive and erroneous samples should be identified and removed
- Remove extreme outliers
- Missing values should be spotted and either removed and imputed by proper
- Restructure Dataset

Feature Engineering

- Feature Extraction
- Feature Selection

Data Splitting

- Split to training, validation, and test

Train Model on Data

- Choosing a Machine-Learning Method
- Visualize model progress during the training job
- Reduce overfitting with regularization

Evaluation

- Choosing evaluation metrics
- Comparing our model to other models
- Hyperparameter Tuning – run a bunch of experiments with different settings and see which works best

Making Predictions

Deploy

4. Conclusion

4.1 Recap of the proposal's key points

The competition among different companies for intra-city taxi trips is a significant issue, and pricing can pose a serious challenge. The price should be set in a manner that satisfies both the passenger and the driver, despite variations in demand and supply across different regions and time frames. These factors can influence the expected price for beneficiaries.

The marketing team aims to incentivize drivers by identifying high-demand areas during different time periods of the day, thereby increasing supply and resulting in higher income. By analyzing trip data from the past four months, it is possible to uncover patterns and trends that can help predict demand and facilitate more informed and optimal decision-making for the marketing team.

4.2 Reiteration of the project's value and potential impact

In this project, we are looking to build and deploy a machine-learning model that can predict demand in a certain period for a specific area with acceptable accuracy. This model, if the results are correct, can have the most financial benefit by creating a balance in the supply and demand process and creating incentive plans, according to the time trends.

4.3 challenges

We have a huge dataset containing 12M rows of information with a variety of features. First of all, there are some odd inputs in our dataset, for instance passenger_count is zero for some rows, or pickup time is later than drop off time. We have to decide if these rows are fixable and reasonable, or we have to find a way to impute them to avoid any errors in the prediction.

Also to train and test this amount of data with Deep Learning Models or time-series analyzing, we need a good GPU to accelerate the speed of the process. For this purpose, we should use Colab Pro.