**NYC Taxi Trip Data – A Data Science Project**

Aviv Lahat, aviv.lahat@mail.huji.ac.il, a\_lahat1

Haya Riesel, haya.riesel@mail.huji.ac.il, hrizel

Miriam Goldstein, Miriam.goldstein1@mail.huji.ac.il, miriamgo

**The Idea:** In NYC, it’s very likely that one’s form of transportation will be in a yellow cab. Observing taxi traffic can provide city planners, citizens and taxi companies with information in diverse aspects, such as pricing, hourly and season patters, insights on where to direct taxi drivers and more.

We aim to analyze these aspects and draft a recommendation for city planners based on NYC taxi trip data that spans over several years.

**The Data:** From the NYC OpenData (<https://opendata.cityofnewyork.us/>)

1. 2011-2021 Taxi Trips   
   From each year we queried the necessary rows and columns for the specific questions. Each year is ~8 GB of data, about 100M taxi rides, where each row is a single ride and the corresponding columns are pick-up and drop-off time, passenger count, trip distance, etc.
2. NYC Taxi Zones – 4 KB GeoJSON file

New York is divided into 263 zones, each for which the json file gives information on area, id, borough, geometry, etc.

1. DOF: Summary of Neighborhood Sales by Neighborhood Citywide by Borough

The Department of Finance (DOF) maintains records for all property sales in New York City, including sales of family homes in each borough. This list is a summary of neighborhood sales for Tax Class 1, 2 and 3 Family homes.

About ~6,500 rows of real estate sales by year and neighborhood

**The Solution:**

Part 1: Hey Party People!

When planning new neighborhoods in the city, city planners and different groups of citizens like families, young couples, students, etc. want to know where they should hunt for housing. Having information on the night life is a crucial consideration. To visualize this question, we mapped out taxi pick-up hot spots from 10pm to 5am, which we label as night-life hours.

Chart, histogram

Description automatically generatedFor this part, we sampled 10M out of 83M records from the taxi trips taken in 2019.  
Sanity check: sample data spreads across all months in the hour slots that we defined.

Map

Description automatically generatedThe interactive map shows areas in NYC with at least 50K taxi pick-ups during night life hours, and is colored by how popular it is for a taxi to pick-up riders from an area in NYC compared to all other areas.

Interactive map link: <https://www.cs.huji.ac.il/w~a_lahat1/NYCTaxi.html>

For this next part we will focus on neighborhoods in Manhattan with more than 50K taxi pick-ups in this time frame. From our data, if you want to live in the center of night life, the recommendation is that you should lean towards East Village, Clinton Eastor Lower East Side where 20% of taxi rides started here during the night-life hours. Otherwise, for example, if you’re a family with children, you’d want something quieter and the recommended areas are Upper East Sideor Clinton Westwhere only 2%of taxi rides were there throughout the night-life hours.

Part 2: Where should I land?

We want to use clustering in order to test if the airport a person lands in will affect other factors of their trip home, for example, duration, distance, time of day etc. We took a representative sample of trips home from JFK, Newark and LaGuardia Airport by sampling 20 rides in each hour of the day. We defined two features to cluster upon - trip duration and trip distance, and checked what other features are similar in each resulting clustered group.

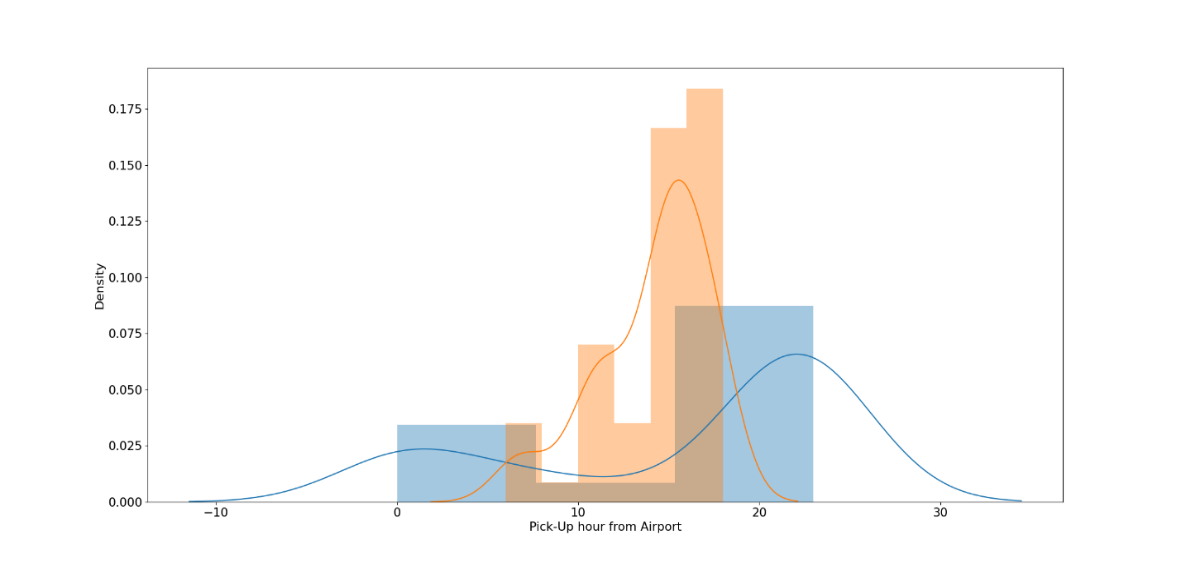
Chart, line chart

Description automatically generatedFirst, to find the best k to cluster by, we used the Elbow method and chose k to be 4.

Chart, scatter chart

Description automatically generatedNext, we used K-Means algorithm for the clustering.

One feature that correlates nicely with the resulting clusters is the hour of the taxi trip. Let us define group A as the taxi trips with a trip duration between 20 and 40 minutes, and group B as the trips with a trip duration between 40 and 60 minutes. Focusing on trip distances between 15 and 20 miles, we can see that group B was picked up from the airport mostly during rush hours between 17:00-20:00, and group A during other hours.



Group A

Group B

Pick-Up hours density plot for clusters 2 + 3

Part 3: PageRank to CityRank

Another approach we used to find the popular zones in New York City was with the PageRank algorithm.

To find these areas, we revised the algorithm so that the nodes are the different zones as defined by the taxi data, and the directed and weighted edges are the percentage of rides between each two zones.

Map

Description automatically generatedWe ran the revised version, nicknamed CityRank on 5M taxi rides from 2017 and mapped out the algorithm output in this visualization.

To evaluate our ranking, we used a separate data base of NYC Real Estate and hypothesized that central zones in New York would have corresponding higher demand in real estate. We looked at the median price of a family estate in NYC and saw that there exists a correlation between the zones ranked highly by the CityRank algorithm and the Real Estate demand for that area.

This map visualizes the median prices of each neighborhood:

Map

Description automatically generated

The cross between the two sources databases was done by the names of the neighborhoods. Neighborhood names is a non-injective value and there can be a neighborhood that is listed slightly differently between the two databases. To do this we used a function that calculates the degree of match between two words and did a crossover only when the score crossed a certain threshold. The function works so that it calculates the edit distance.

For an addition evaluation, we also checked multiple years and similar findings were found there as well.

Continuing this train of thought, it is interesting to see if the areas where there is an increase in CityRank along the years are also increasing in the value of real estate in the area.

We calculated the CityRank of the years 2019 and 2017. We took 2 years distant enough years so that if a significant enough change occurs in that period, we would detect it.

Since real estate has risen significantly over the years we would like to calculate the percentage of increase in New York in general and treat it as an estimate of whether there has been an increase or decrease in real estate. (It is possible that the price of the property increased, but it increased less than the general rate of increase in real estate, so this is considered a decrease in the value of the real estate.) In addition, it will be seen whether the rank of the area decreased or not.

We calculated the appreciation (calculation of the percentage increase compared to the initial price in percent) and the difference between the ranks (the difference can be negative or positive). We have interpreted the points on the graph and indeed you can see the change:

(Note that most of the points with a positive ranking are above an appreciation of 80)

Chart, scatter chart

Description automatically generated

The neighborhoods whose ranking did not rise or fall do not interest us (because they may have been central enough before so there is no increase in rank, but in appreciation there will be a drastic increase because in central areas there is a very high increase in price). Let's just look at the points whose ranking went up or down to see what happened there:

Chart, scatter chart

Description automatically generated

Indeed, it can be seen that there is an uptrend in the graph as the rating rises.

**Evaluation:** The main part of our evaluation was making sure our sample data was representative of the whole taxi trip data for the year we used in each part.

To ensure this, we created the distribution plots featured throughout the report that show the sample data spreads across the months, hours, days etc. needed for each different part of the project.

For example: in Part 1 we focused on night life hours between 10pm to 5am, and the corresponding distribution plot shows that indeed the data is taken from these hours and across all months of 2019.

Impediments: We would have wanted to run these algorithms on the whole collection of taxi data spanning across multiple years instead of using representative samples of a specific year for each part of the project. This would have required large cloud storage, therefor for the framework of this project we were content with samples and checked that they were representative of the population.

**Future Work:** The taxi trip data set has a number of features that can add useful data to a prediction model for NYC citizens. Incorporating it with other NYC data bases such as weather, population spreads and more would result in a very informative model.

**Conclusion:**