**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/>)

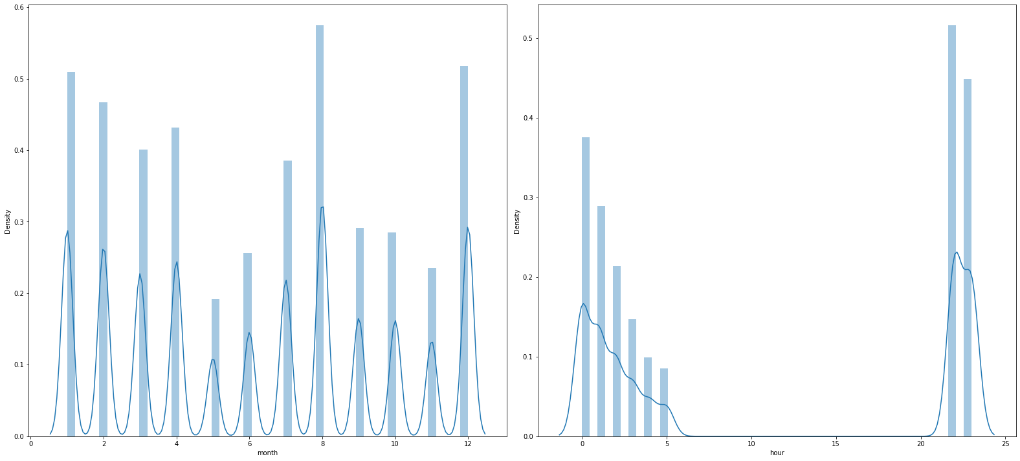
1. 2011-2021 Taxi Trips   
   From each year we queried the necessary rows and columns for the specific questions. Each year is ~8 GB of data, about 114 M taxi rides, where each row is a single ride and the corresponding columns are pick-up and drop-off time, passenger count, trip distance, etc.
2. NYC Taxi Zones – 4 KB GeoJSON file

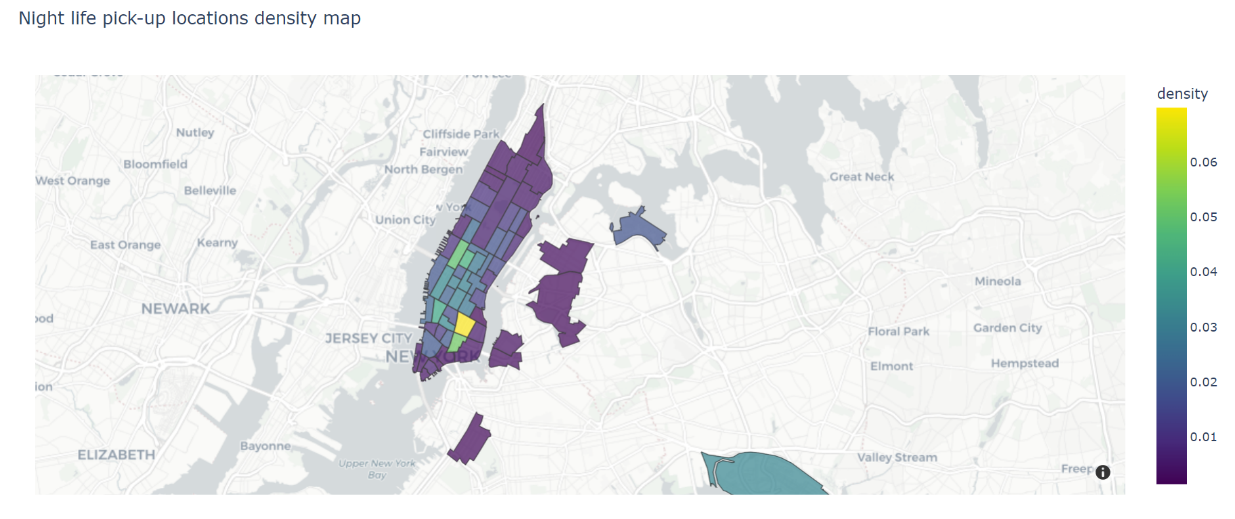
New York is divided into 263 zones, each for which the json file gives information on area, id, borough, geometry, etc.

**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.

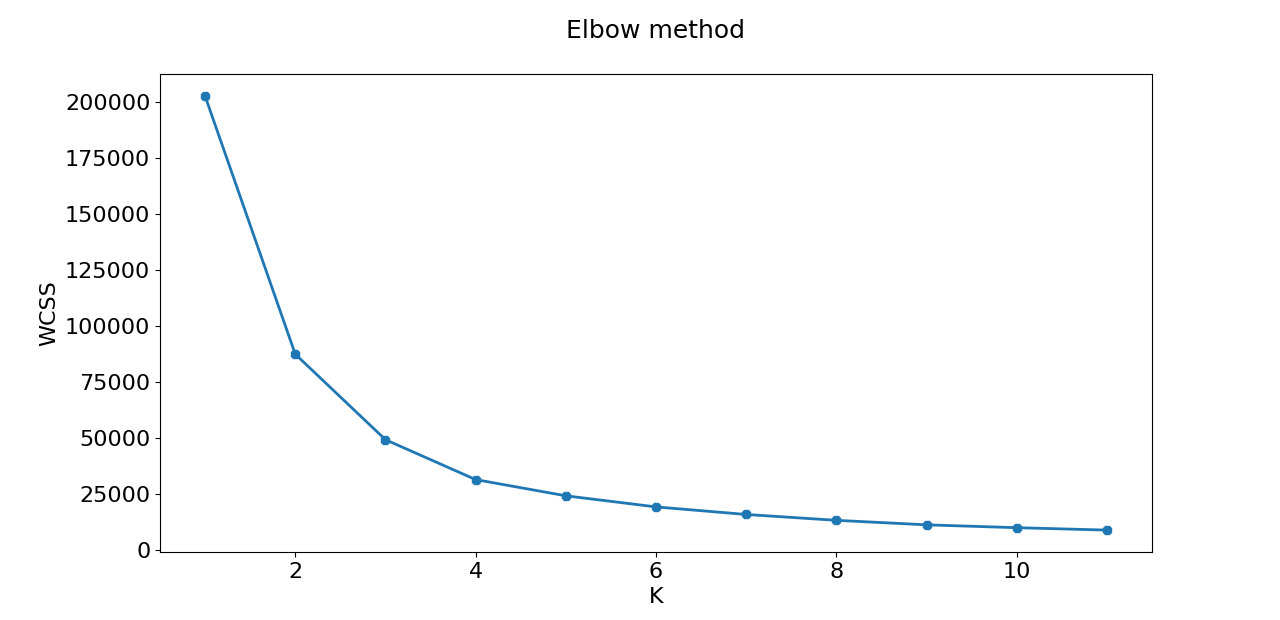
Sanity check: sample data spreads across all months in the hour slots that we defined.

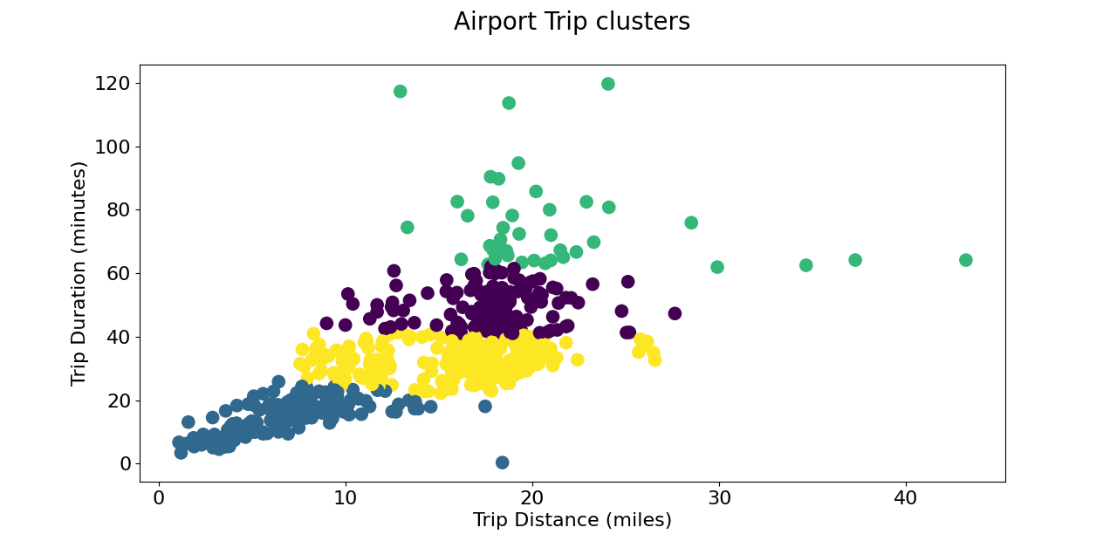
Interactive map link: <https://www.cs.huji.ac.il/w~a_lahat1/NYCTaxi.html>

If you want to live in Manhattan neighborhood that in the center of night life, from this data the recommendation would be that you should lean towards **area 1** or **area 2** where **high percentage** of taxi rides were from 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 **area 3** or **area 4** where only **low percentage** of taxi rides were there throughout the night-life hours.

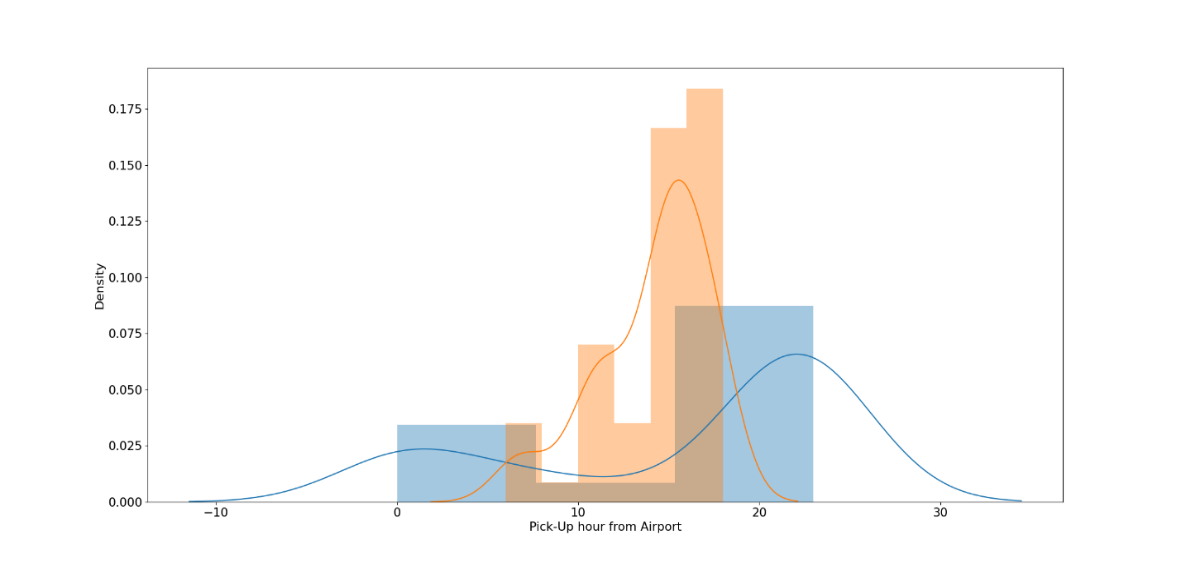
Part 2: Where should I land?

We want to use clustering in order to test if the airport I’m landing in will affect other factors of my trip home, for example, cost, duration, distance etc. Since our data is very large, we decided to take 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.

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

Next, 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 mostly during other hours.



Group A

Group B

Pick-Up hours density plot for clusters 2 + 3

**Evaluation:**

**Future Work:**

**Conclusion:**