NY TAXIS

Team members info:

Names: Omer Hagage, Daniel Vishna, Yuval Ussishkin

Email: omer.hagage@mail.huji.ac.il, daniel.vishna@mail.huji.ac.il,

yuval.ussishkin@mail.huji.ac.il

Cs ids: omerhagage, daniel.vishna, yuvalu

Problem description:

For this project we will delve into one of the world's most complex transportation systems – the New York City taxi system. New York is one of the world's biggest cities, with one of the highest percentages of taxi use. Most New Yorkers use cabs on a regular basis, the same as a private car is used in Israel. For this reason, many interesting insights can be gained from looking at data of taxi rides in New York, as these rides can give an accurate description of people movements in the city.

We will attempt to use the data with the aim of identifying successful points for taxi drivers to set up to attract customers. This way we can recommend good locations to the cab drivers by time of day, considering different movement trends in the city.

For this project we will use clustering and node importance methods that we learned in class, while giving them a twist of our own.

Our data:

We used a public dataset of all yellow taxi rides committed in New York City over the years (from 2009 to present). An average month worth of data has approximately 15 million taxi rides and is ~ 1.5 GB. We opted to use the data of 2016, as it Is the latest year that includes exact coordinates of pick up and drop off locations (from 2017 only the taxi zone is given). The exact coordinates give us a much wider view of the taxi routes since the taxi zones indicate neighborhood and not exact location. We also didn't want to explore from 2020 onwards, as the past two years have been years with the Covid19 pandemic, which makes the usual transport patterns erratic (the effects of the pandemic are fascinating to explore, but we will leave that for further research). The data contains the full details on **all** taxi rides in NYC in the relevant time frame, including full coordinates of pick up and drop off location, pick up and drop off time stamp (from which we can derive date and day of the week), the full fare paid, and the tip paid. We also downloaded a shapefile (a data format for representing geographical data) representing the division of NYC to taxi zones (attached on this link https://drive.google.com/file/d/1D8O_jlZGcucdYeQv0SzD5IvEXfdxE6A7/view) .

This is very useful as some of our research would be better conducted on a lower resolution of location than exact geographical coordinates. The shapefile comes in

handy as it gives a lower resolution location definition for pick up and drop off points. The shape file consists of a list of polygons defining the different taxi zones, so given a geographical coordinate we can check in what polygon this coordinate resides and define the corresponding zone. To get a hang of the taxi zones we plotted the rides by geographical location and colored each zone differently. New York is divided into ~250 different taxi zones, and the following visualization gives a good look into their popularity. For instance, JFK airport is easily visible in the bottom right corner of the points. Also, central park is immediately recognizable as the sparse area in upper central Manhattan.

Attached below is a link to the data source, our code downloads the relevant parts (But hold on, it does take a while!):

https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Preprocessing and Sanity Checks:

The main issue in the data preprocessing was removing outliers and corrupted data records. We removed all data with numbers that didn't seem to add up, like a trip time of zero minutes. We also removed trips that start or end out of NYC city limits, as we only want to research trips in the city limits, or trips with more than six passengers (which is illegal by New York State law). The only exception is Newark Airport, which lies in neighboring New Jersey but still is allocated a zone in the New York system. We also edited the time stamp of the trip, separating it out to include the time, trip duration, and day of the week.

After preprocessing we ran a few checks on the data, just to get the hang of it and to give ourselves a quick check that the data adds up and makes sense.

We plotted the number of daily taxi rides per day in the relevant period we are checking. The two dips in the graph are caused by this - https://en.wikipedia.org/wiki/January_2016_United_States_blizzard – and by the annual Memorial Day

Attached below is a plot showing the average duration of a taxi ride by day of the week

The basic outline of the graph makes sense. The duration peaks on Thursdays and Fridays, which makes sense as these are working days closest to the weekend, so usually these days will have the most traffic and congestion, while Sunday which is a rest day has a much lower average trip duration.

Another check we conducted was plotting the average duration of a taxi trip by the time of day:

The results add up nicely, as it makes sense that the duration peaks towards the afternoon hours which are the main rush hour in NYC, while night rides tend to have lower durations. We were slightly perplexed by the fact that the traffic increases so heavily at an early stage of the day (6:00). We made a guess that it might have to do

with that being a busy hour at the airports serving the city, and indeed checking the data after removing rides that serve the airport and comparing their duration to rides that serve the airport seems to prove our guess correct.

We also plotted all rides by trip duration and number of passengers in the car. As you can see below, trips with more passengers have a slight tendency to be quicker rides, probably because long commutes to work are usually done solo or with few people in the cab, while rides with many people tend to be rides for recreation purposes.

Project solution:

We first set out by attempting to recommend hot spots for drivers using the k-means clustering algorithm. This idea makes a lot of sense as our data has geographical coordinates and the time of day, so recognizing the main clusters of data will give us busy locations in the city by time. We decided to use k-means, as the main objective for a taxi driver should be getting as close as possible to the hot pick-up zones, and k- means uses an objective of Euclidean Distance, which is appropriate. While it's true that much of New York is built in a grid scheme, there are many diagonal roads as well (like Broadway). Because of this, the Euclidean Distance shouldn't veer away too much from the actual driving distance between spots. Our method gives us clusters with 3D points as centers, each point representing a 2D Geo coordinate and the time of day. We chose to implement the method with 100 clusters as output.

Attached below is an image of the results but the most informative option is to open the following link which allows interactive exploration of the results - https://htmlpreview.github.io/?https://raw.githubusercontent.com/danielvishna/IDS_project/master/3D_cluster.html .

We also implemented a similar version of clustering, but this time as a cluster of 2D points, containing location data only. For this method we can choose a time of day that spikes our interest and find the top spots for locating a taxicab in this hour, this time using 50 clusters. Attached below is a gif that shows the clusters as calculated every hour - https://imgflip.com/gif/66sbn4. It is fascinating to note how the centroids converge towards Manhattan during the rush hour, while during the night hours the centroids appear more stretched out towards other NYC boroughs like Brooklyn or The Bronx.

Our second attempt of recommending hot spots was based on the following reasoning: while it's true we care about the number of pickups from a taxi zone, we should also look at the popular destinations from this taxi zone. This is because a cab driver would obviously prefer to take passengers to a drop off zone which is busy as well, so the driver could easily find a new passenger after the drop off. This kind of thinking would require us to somehow recursively define if a pickup zone is recommended for a cab driver by looking also at the popular destination zones from this pick-up zone and evaluating them. This reasoning reminded us of the node importance algorithm taught in class, because of the recursive evaluation process.

The natural solution was creating a graph out of the taxi zones as defined in our shapefile, with the zones operating as vertices. The edges of the graph will be

weighted by the number of rides between zones, e.g., if in our data there are ten thousand rides between zone one and two, vertex one will have a directed edge to vertex two, weighted by ten thousand. We will notice that this graph still isn't completely ready for running the node importance algorithm. In the regular node importance procedure, a given node's importance is defined by the nodes which direct to it, e.g., by **incoming** edges. In contrary, in our scheme, we want a node's importance to be defined by nodes it directs to, e.g., by **outcoming** edges. So basically, we need to transpose the graph, so the node importance method can work. We did that by transposing the matrix representing the graph. After gaining the algorithm's grade for each zone, we normalized all grades on a scale from one to ten.

We constructed a heatmap representing taxi zones in New York City by importance that our algorithm assigned them. As would be expected, Manhattan taxi zones tend to get a higher zone ranking than zones located in other boroughs of the city. Apart from that, the zone in the bottom right corner of the city which got a very high ranking (colored in vivid red) is JFK airport. An interesting observation is while central Manhattan is colored red, the southern parts of Manhattan (which is mostly the Wall Street area) aren't really colored and seem to get a much smaller rank. This is surprising because Wall Street is essentially the economic capital of the world, employing huge numbers of employees. The main reason for this result seems to be that many employees of the Wall Street firms reside in affluent suburbs of the city, and these suburbs have a low amount of taxi rides leaving them. Because our method takes the common destinations from a zone into account, the Wall Street areas get penalized.

Presented below are the top twenty taxi zones in New York City by importance as calculated by the algorithm (and normalized). It is interesting to observe that while JFK airport is unsurprisingly top ranking, some areas in midtown Manhattan are ranked higher than LaGuardia Airport, which is also a major airport that serves NYC.

Small sidenote about privacy: Dealing with this large amount of data got us thinking about some privacy issues that arise from the database. Since all taxi rides conducted by New York taxi companies appear in the database, and exact coordinates are given, it makes sense we can learn some things about individuals using this data. Most coordinates in the city center don't tell us much, because a location in midtown New York is among many different buildings and residences, so it is difficult to tell who was using the taxi service. Some big businesses (like big financial institutions in Wall Street) can be identified by their coordinates but as they have thousands of workers not much can be learnt from looking at those rides (maybe with another hack into those businesses we can cross examine the data). It does become interesting though when looking at some of the taxi rides that serve suburbs of the city. In those instances, the address of the customer can be narrowed down to a few houses, and sometimes even to a single residence. For instance, we attached below a row in the data with a very interesting drop-off point.

This entry is interesting because it seems to point to a very specific residence in Rocky Point, a town in Long Island. Attached next is the exact point pinned to a map.

A quick look at google street view confirms that this is a private residence, so there doesn't seem to be much doubt that either the residents of this house or their guests were the customers of this taxi. In the case where it was the individuals residing in this house, we can know exactly where and when they were in other spots around the city, and this seems like a big breach of privacy.

Evaluation:

For evaluation, we set aside some of our data, so we can check our forecasts on unforeseen data. In the following section, when we refer to "data", we mean this set-aside data. We based our forecasts on the months of January to April 2016 and evaluated the forecasts on the months of May and June 2016.

Clustering – Our cluster method will be evaluated by attempting to estimate the number of pick-ups our clustering recommendation has managed to forecast. For this purpose, we will need to define what pick-ups in the data will be "caught" by our clustering prediction. We defined a maximal distance error and maximal time error such that any pickup with lower error rates for both from a cluster centroid will be considered "caught". The main challenge we encountered was the need to normalize the error of the distance with the error of time, as units for time don't have the same meaning as distance units. After some lookup on common New York driving times, we normalized it so that a mistake of one kilometer would be penalized like a mistake of ten minutes. Our evaluation method will check what percentage of all rides conducted in the data are targeted successfully (according to the error rates we defined) by our clusters. We compared this with points with random location coordinates and random time, and compared which points fared better. Plotted below is the comparison between the results, with a different number of clusters (and random points) for each time. As you can see, the clustering results fare far better than a random choice of location (for instance using 100-point clustering we achieve ~9 percent of rides as opposed to ~0.002 percent in a random assortment of points)

Node Importance – We set out to evaluate our node importance by committing a simulation of taxi rides based on node importance versus taxi rides based on a random route and comparing the amount of money made in the two different methods. We simulated a taxi that is positioned in each of the top ten taxi zones by node importance, at 7:00 AM. For the sake of the simulation, our taxi has a chance of p to serve any given customer in the database, so it follows that if in a given minute there are n rides radiating out of the given taxi zone, there is a chance of $1 - p\text{^}n$ to fail to attract a customer. We set $p = 0.7$, after simulating various options and checking their credibility. The simulation is conducted as follows: we calculate the probability that our taxi will take a customer in the given time of day (minute). We utilize a flip coin method with this probability and based on the result rule if the driver attracted a customer. If it is ruled that the taxi didn't attract any customer, we do the same for the next minute of day, until the taxi is ruled to have attracted a customer. The simulation is halted at 20:00. If the taxi is a random taxi and it is ruled to take a customer, we will choose one of the rides at random, otherwise we will choose the ride to the best ranking taxi zone by node importance. This simulation will carry on for 13 hours (of taxi rides time, not execution time) and we will compare the two

methods by revenue, total rides conducted and time “wasted” seeking a customer. Shown below are figures comparing total number of rides, revenue and time “wasted” between the two methods, initialized from the top ten zones by our zone importance algorithm.

As shown, the revenue is higher over all top ten zones, showing that our zone importance method does indeed predict what zones are both busy and tend to have rides that lead to other busy zones. In contrast, the total number of rides doesn’t have a clear difference between the two methods, but in the random walk much more time is wasted waiting for customers than in the case where high-ranking zones are targeted, where no time is wasted at all. This is because our method gives high priority to rides to busy zones, where most rides from that zone lead out to other busy zones.

Future research:

We researched a huge database, and this enables us to develop this project much further. There are countless possibilities to conduct further research. One important development in the past couple of years has been the COVID19 pandemic. This has had a huge impact on all fields of life and on the transport specifically. Even after the pandemic is over, most probably some of the changes it has brought will be here to stay, so researching changes over the past couple of years in taxi trends can be very useful for understanding the changes in the industry.

Another trend in the past years has been the emergence of rideshare companies like Uber. New York has a database for these companies as well, so we can also conduct research on those companies and their increasing influence on the market. Many comparisons can be made to make forecasts about the competition between yellow taxis and rideshare companies.

Further work can also be done exploring working, living and nightlife trends using the database and geographic location. For instance, an influx of rides leading to a certain area in the city at night can lead us to deduce that the area is a nightlife zone with rising popularity. The same goes for popular areas for families to hangout on Sunday or any other type of similar research. This can be used to aid billboard advertising targeting specific population or any other kind of research geared towards people’s whereabouts trends.

Conclusion:

We had a fascinating time investigating this topic and found the data to be very fun to work with. The data amounts were huge, which added various challenges to our project, as we had limited computing ability. Also preprocessing the data took some time, as data of such size contains many mistakes and errors that need to be identified. After tackling those problems, it was interesting to analyze the data and delve into the results (especially with a prior knowledge of NYC layout). Our results showed a huge difference between different taxi zones, and we were surprised at how much Manhattan zones got a higher ranking than other boroughs.