2015 Yellow Taxi Trip Data | NYC OpenData

Project report

Lise HAIE
Exchange student
University of Helsinki
lise.haie@helsinki.fi

Yoav ATLAN
Exchange student
University of Helsinki
yoav.atlan@helsinki.fi

María Lina RIERA FERRER
Exchange student
University of Helsinki
maria.l.rieraferrer@helsinki.fi

ABSTRACT

The 2015 Yellow Taxi Trip Data describes all the trips performed in NYC from January to June 2015 with several parameters. In this report, we present the network we studied, its visualization, and the conclusions we can make about the communities and the behavior of the inhabitants of NYC, regarding the graphs and data we obtained.

1 INTRODUCTION

There are more than 8 million inhabitants NYC, among them, numerous persons are travelling via the yellow taxis to cross the 738 km^2 city for work or personal activities. The 2015 Yellow Taxi Trip Data describes all the trips performed in NYC from January to June 2015 with several parameters. By studying this network, we want to identify the inhabitants' travel habits, their daily routines, the hot-spots in NYC. To do that, we are going to analyze the structure of the network and its parameters (degree distribution for instance), and then to see the taxi service needs thanks to an analysis of the communities. To finish, we will try to do some prediction of the evolution of the characteristics of the network by predicting the edges and hot-spots. Our software program is written with Python.

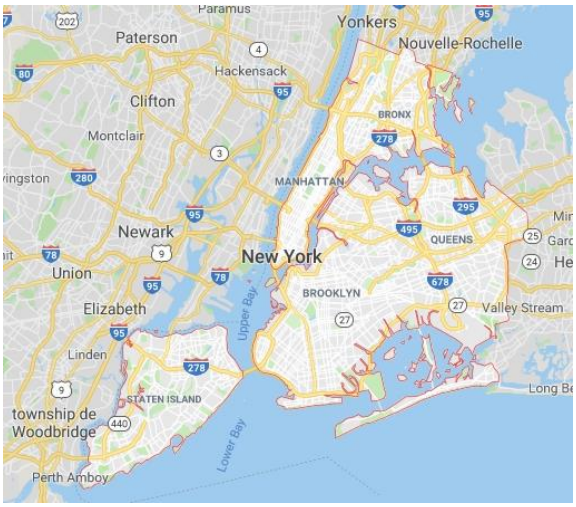

Figure 1: NYC map from Google Maps.

2 DATA DESCRIPTION

2.1 Raw data

We have at our disposal over 146 million trips completed in NYC's yellow taxis. The data was collected from January to June 2015 by technology providers. Records include many parameters, some of them are not relevant for our study, so we chose to work mainly with the following parameters: trip distance in kilometers, pickup and drop-off locations (described with latitudes and longitudes).

trip_distance	pickup_longitude	pickup_latitude
15.1	-73.98356628417969	40.749881744384766
2.9	-74.00045776367188	40.727294921875
1	-73.99409484863281	40.74161148071289
1.2	-73.9753646850586	40.78733444213867
3.6	-74.00634002685547	40.73313522338867

Figure 2: Extract of the dataset

2.2 Transformed data for analysis

To build our graph and to analyze it, we had to transform the data; as we did not have any nodes or edges already operational to put on a graph. First, we decided that departures and arrivals would represent the nodes, and that the trips performed between the nodes would be the edges.

We decided to build a grid pattern that would cross all the locations (departure and arrival of every taxi trip). Each square represents an area of 100x100 m. We have converted the latitudes and longitudes in meters. To finish, we had to remove the unused nodes.

All this done, we had our grid pattern and we generated the nodes and the edges in it. The problem we encountered is that we had 146 million trips and when we ran the code, it took about 13h to create nodes and edges for 40 million trips. That is why we decided to not run all the dataset, but to work with some samples. We considered that with 40 million trips we had enough data to analyze the network as a lot of trips are the same or represent the same kind of travel. We also picked a sample of 1000 trips to see the differences with the numerous 40 million trips.

To conclude about the data and its transformation, at the end we have our grid pattern that crosses all NYC, and we have our two samples to analyze. After the generation of the nodes and edges, we have:

- For the 40 million trips sample: 88 612 nodes and 6 861 573 edges.
- For the 1000 trips sample: 1296 nodes and 981 edges.

3 ANALYSIS TASKS AND RESULTS

Once the data is transformed and now that we have our two samples, our nodes and our edges, we can begin the analysis of the network and interpret it to answer the questions we asked ourselves.

3.1 Visualization of the network

First, let's try to analyze the 40 million trips sample. When we run our Python software, we can plot the following graph of the network:

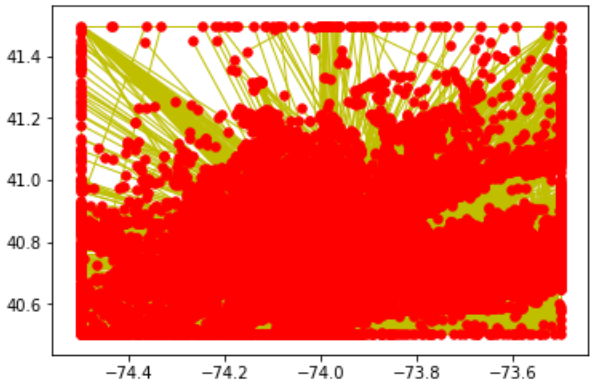

Figure 3: visualization of the network for 40 million trips

The ordinate is the longitude and the abscissa is the latitude. In red you can see the nodes and in yellow the edges.
So, due to the number of nodes (88 612) and of edges (nearly 7 million), we cannot really see if there are some particular groups or communities, nor make some statements. That is why we also analyzed a more reduced sample that we will see later.

Once we had this graph, we calculated the degrees of the nodes. The following graphics are about both in and out degrees as here we analyze our undirected graph. In ordinate there is the number of nodes and is abscissa there are the degrees. This is what we obtained for the degrees of the nodes:

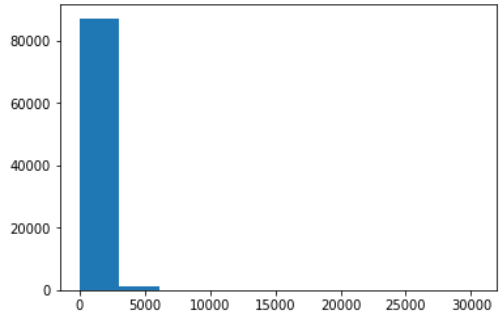

Figure 4: Degrees of the nodes for the 40 million trips sample

So, we can see that there are a lot of nodes which have a degree between 0 and 5000. Then, when we zoom to see clearer, we can see that there is a majority of nodes which have a degree between 0 and 2350, meaning there are lots of edges adjacent to these nodes. The connected components will be analyzed later with the communities.

To finish with the analysis of the 40 million trips sample, we have plot the locations of the different nodes with a different color depending of their intensity; meaning that the more the degree of the node increases, the more intensity the node gets. This is the graph we obtained:

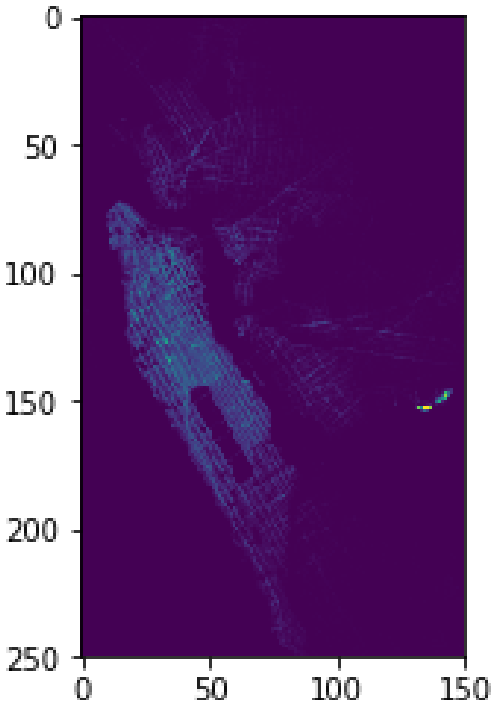

Figure 5: Nodes intensity linked to their degree

So, the axes represent the location of the nodes, and this network of nodes represents in fact the map of NYC.

We can see some groups way better than with the first graph (Figure 3). We can also see that some zones of NYC are way less crossed by the taxis than others. But let's analyze it in the next part, along with the communities.

Now, let's see what the graph and the degrees become with a sample of 1000 trips. When we plot the graph, we obtain this:

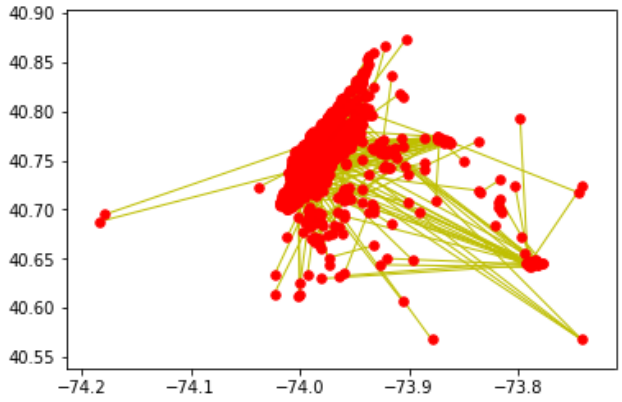

Figure 6: visualization of the network for 1000 trips

As a reminder: the ordinate is the longitude and the abscissa the latitude. In red you can see the nodes and in yellow the edges.
With this reduced number of trips, we can see better the network and the connections between the nodes. We can also see some groups that represent different neighbors of NYC (Manhattan, Queen, Bronx…). So, we can see better the communities of the network. Now, let's see the degrees of the nodes for this sample:

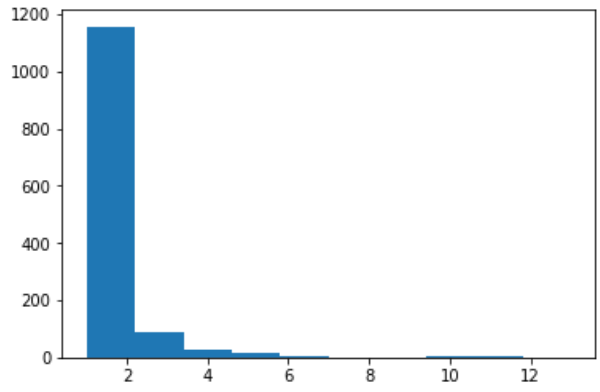

Figure 7: Degrees of the nodes for the 1000 trips sample

As the number of trips diminishes, the degrees of the nodes follow the same path. We can see that almost all the nodes of the network (there are 1296 nodes in this sample) have a degree between 1 and 2.
To conclude on this part, we answered our first question as we visualized the network by plotting two graphs, one for the sample of 40 million trips, and one for the 1000 trips sample. We can visualize better the network with the second sample because in the first one there are too many nodes. We also calculated the degrees of the nodes and started seeing some communities in the network that we are now going to interpret.

3.2 Taxi needs

In order to understand the taxi needs of NYC, we need to analyze the different communities that we can find in the network's graph:

```
In [315]: connected_comp_lengths[:50]
Out[315]:
[83729,
 26,
 20,
 17,
 16,
 12,
 8,
 7,
 7,
 6,
 6,
 5,
 5,
 5,
 5,
 5,
 4,
 4,
```

Figure 8: Example of the connected components of a community for 40 million trips

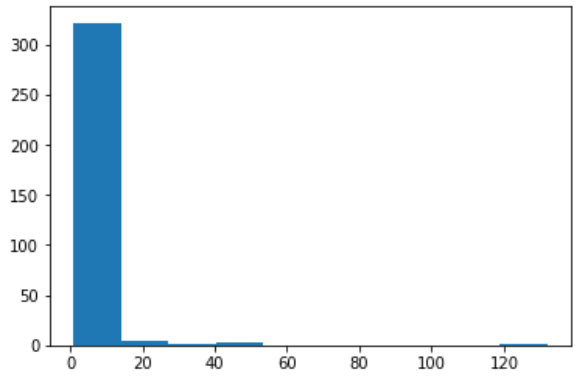

Figure 9: Number of nodes in a community 1000 trips sample

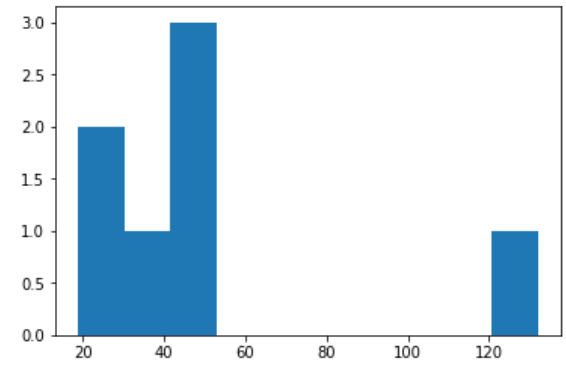

Figure 10: Zoomed version of Figure 9

We can see in Figure 9 that the majority of communities that we can find in the 1000 trips sample are made of an average of 10 connected components. In Figure 10, we can see that the majority of the communities that are not included in the range that goes from 0 to 10 nodes are formed mainly by either an average of 25 or 45 connected components (there are two different peaks in the graphic). The reason as to why this happens is probably that we only used 1000 trips to make these graphs, so the communities will be smaller because there are a lot less nodes and edges being analyzed. However, the fact that there are actually communities made of 120 components, even if they are only a few, shows us some tendency to have relatively big clusters, that applied to the situation would mean the equivalent of hot-spots.

However, this information is not enough to really understand the taxi use in NYC because we cannot actually contextualize the information we have presented. That is why we have the map with a representation of the number of taxis, made using yellow nodes, that shows us the number of taxis that are in the whole city (Figure 11), that actually matches the graphs obtained by the 1000 trips sample and by the nodes' degree graph of 40 million trips and it can help them contextualize both.

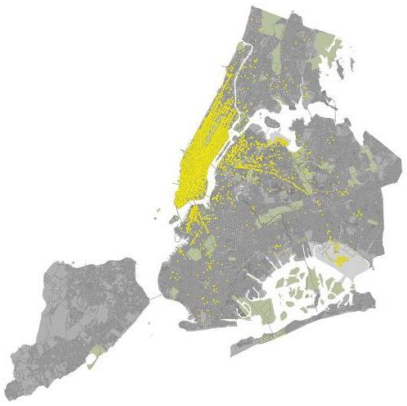

Figure 11: Density of taxis in NYC

Looking to this map, we can see that despite being five different big neighborhoods in NYC (Manhattan, Queens, Staten Island, Bronx and Brooklyn), the vast majority of taxis in NYC are in Manhattan. We can see that there are two other smaller communities in Queens and in Brooklyn and even some smaller communities at the suburbs, where public transport is not as easy to use. The biggest community is clearly in Manhattan.

This information made us think about the reasons why there is such a big difference in taxi availability in the different parts of NYC and we found out that there are several.
First of all, despite not being the neighborhood with the more inhabitants, Manhattan has the highest population density, which is an advantage for taxi drivers because they do not have to drive for a long time to get new clients, there is a constant flow of clients in small zones.
Moreover, Manhattan is also the economical center of the city because it holds 78 % of its economy. That fact actually affects directly to its inhabitants' salary, making it more possible for them to pay for a taxi instead of public transport. Furthermore, Manhattan is the most touristic part of NYC.

So, in conclusion, despite the seemingly unfair distribution of taxis in NYC, the taxi needs of the majority of users are covered.

3.3 Further analysis: prediction of the evolution of the network

We have presented two questions in order to be able to predict the evolution of the network:

Will the network's number of edges increase?

This question has actually been answered by time, because the dataset that we used is from 2015 and nowadays, in 2019, we have already seen that the number of edges not only has not increased. In fact, they have decreased.
There are several reasons for that, the most important being the creation of new companies, such as Uber, Cabify and Lyft that provide the same services that NYC taxis do but with better customer service, lower fares and, most importantly, the customers ask for a trip from an app. There are many cities throughout the world where the taxi services use apps (for example, in Spain there is MyTaxi, in Finland Yango…) but in NYC taxis did not have an app for their users (Curb) until 2019. Apps are an advantage because nowadays everyone uses their phone for almost anything and it is more comfortable to plan and schedule your trip from home than having to call to the radio-taxi services or having to wait in the street for a taxi to come by. Moreover, apps are easy to advertise on internet and younger people and internet users in general will be easily exposed to said advertisements.
Other reasons for the decrease in the use of taxi are that it is a service that is basically used in Manhattan, there are not as many available taxis in the rest of the city. Moreover, there is a really big network of public transport (both bus and subway) that local people can use on daily basis for a lot less money than they would spend on taxis.

To conclude, we should think about the taxi drivers' point of view, because taxi loans are really expensive and they provoke many taxi drivers to be in debt, making it less attractive for them to try to get a taxi license.

How would the construction or collapse of a building (or a group of buildings) affect the network?

This second question is the equivalent of asking how would change the network if we added or removed nodes in the network.

And, after all the information presented, we can give a straight answer: it wouldn't change it at all.

This answer might seem a bit exaggerated, but, if we look again at Figure 11, we can see that there is so much taxi concentration in Manhattan that it would be impossible to move the said concentration somewhere else.
Let's analyze the facts: in 2015 there were 58.3 million tourists visiting NYC, the majority of which stayed, or at least visited Manhattan because it is the most touristic and famous area in NYC. Moreover, as we stated before, Manhattan is the center of the city's economy.
Considering that the two biggest groups of people who use taxis are adults with higher salaries than average and tourists, there would be only three situations that would have an impact on the network's structure: the creation of a unique touristic place that would make people from all around the world to come and visit it, the creation of a new Wall Street that would move the economical center of the city away from Manhattan or just the general collapse of Manhattan.

4 DISCUSSION

We have managed to answer all the main questions we asked ourselves at the beginning of the project. However, when we chose this subject, we did not realize that a 146 million dataset would be really "heavy" for our computers and would take that much time to run with our Python codes. As we previously said, we ran our code for more than thirteen hours and at this point it has analyzed about 40 million trips. We already explained our choice to analyze these 40 million trips, and it remains relevant for our study and the analysis of the network. However, with more time or with a better organization, maybe we could have analyzed more trips and have an even more relevant study.
Moreover, it would have been interesting too to study the time dimension of this dataset. With an analysis of the trips depending on the hours of the day, and depending of the days of the week, we could have analyzed other things such as which communities have more "work persons" taking yellow taxis. This is the study that the other group working on this dataset made.

5 CONCLUSION

In summary, we have performed the analysis of the 2015 Yellow Taxi Trip dataset by answering three main questions: what the structure of the network is, what can this structure say about the taxi needs in NYC, and can we predict the evolution of this network in terms of most used routes and hot-spots. The 146 million trips caused us some trouble while coding and running our code, so we sampled it in a set of 40 million trips and then in a set of 1000 trips. The structure of the network was visualized with a graph where we can see some groups or communities. We also calculated the nodes' degrees and analyzed the communities and the connected components. From this analysis, it turns out that there is a tendency to have relatively big clusters, meaning hot-spots, in several zones of NYC such as Manhattan or the Queens. The distribution of taxi services is quite unbalanced, but our study shows that the needs of the majority of users are still covered, due to the characteristics of the zones where there are lots of taxis. To finish, about the evolution of the network, we showed that the number of edges is decreasing and will not increase in the future, which is mainly due to the creation of new taxi companies using apps and lower prices to attract the clients, but is also due for instance to the network of public transports. Moreover, if we added or removed some nodes to the network; the structure of the network would not change, except in three cases: the creation of a unique touristic place, the creation of a new Wall Street, or the general collapse of Manhattan.

ACKNOWLEDGMENTS

We wish to thank our teachers Pan Hui and Michael Mathioudakis for providing the network analysis course and supporting us with our project.

REFERENCES

[1] NYC Open Data: 2015 Yellow Taxi Trip Data, https://data.cityofnewyork.us/view/ba8s-jw6u
[2] Nilesh Patil, "Analyzing transportation graph from NYC data", 2018 blog: https://nilesh-patil.github.io/blog/transportation-graph-nyc-taxi-data/
[3] NetworkX documentation for drawing and analyzing graphs: https://networkx.github.io/documentation/stable/reference/introduction.html
[4] The New York times: Record Number of Tourists Visited New York City in 2015, and More Are Expected This Year: https://www.nytimes.com/2016/03/09/nyregion/record-number-of-tourists-visited-new-york-city-in-2015-and-more-are-expected-this-year.html
[5] Reuters.com: New York City tourism climbs to record high in 2015 for sixth year: https://www.reuters.com/article/us-new-york-tourism-idUSKCN0UZ1OA
[6] Wikipedia: Demographics of New York City: https://en.wikipedia.org/wiki/Demographics_of_New_York_City
[7] The Guardian: There's no future for taxis': New York yellow cab drivers drowning in debt: https://www.theguardian.com/us-news/2017/oct/20/new-york-yellow-cab-taxi-medallion-value-cost
[8] The New York times: Yellow Cab, Long a Fixture of City Life, Is for Many a Thing of the Past: https://www.nytimes.com/2017/01/15/nyregion/yellow-cab-long-a-fixture-of-city-life-is-for-many-a-thing-of-the-past.html
[9] Bestplaces: Economy in Manhattan, New York: https://www.bestplaces.net/economy/city/new_york/manhattan
[11] Wikipedia: Manhattan: https://en.wikipedia.org/wiki/Manhattan