NYC taxi trips exploratory analysis

Antonia Donvito (Data Science curr. B, Psychology)
Margherita Fanton (Data Science curr. B, Sociology)
Marco Pakler (Data Science curr. B, Business Administration and Management)

1. INTRODUCTION

The project aims to analyse the data provided by the "NYC Taxi and Limousine Commission", a commission that is responsible for licensing and regulating cabs and for-hire vehicles. Currently in NYC more than 130,000 vehicles are licensed to operate. These are divided into three categories: yellow taxi, green taxi and for-hire vehicles. Yellow taxis are the only vehicles allowed to pick up passengers directly from the street anywhere in the city of New York. Green taxis provide the same service offered by yellow taxis but limited to northern Manhattan and in other boroughs. For-hire Vehicles (Black Cars, Limousine) provide prearranged service reserved in advance (e.g.: Uber). The variables in the dataset are slightly different between the three categories. However, the common ones are the following:

- *VendorID*: a code indicating the provider
- *Pickup_datetime*: starting date and time of the journey
- *Dropoff_datetime*: ending date and time of the journey
- *Passenger count*: number of passengers
- *Trip distance*: length of the journey expressed in miles
- *PULocationID*: place where the journey started
- *DOLocationID*: place where the journey ended
- *Payment_type*: From 1 to 6, indicating (1) Credit card, (2) Cash, (3) No charge, (4) Dispute, (5) Unknown, (6) Voided trip
- *Tip_amount*: tip left only by credit card users
- *Tolls_amount*: tolls paid during the journey
- *Total_amount*: total charged; tips excluded.

Our goal was to work on this massive dataset and consider as much data as possible, in accordance with the capabilities of our personal computers. More specifically, these are the steps that we have taken and that will be described in detail in the next sections: selection of a sample, data cleaning, statistical analysis, exploratory analysis through unsupervised clustering, data visualization with plots, network graphs and maps. The problems we had to deal with were of different types; the choices that we made to solve them shall be described in the following paragraph.

2. DATA CLEANING

For our analyses we decided to take into account the datasets of the years 2016, 2017 and the first six months of 2018, both for yellow and green taxi trips. We did not consider For-Hire Vehicles, as we wanted to focus on the most popular type of vehicles. The earliest step in data cleaning were performed locally with Python.

Yellow and green datasets for the second semester of 2016 had some index and header issues. We managed to fix them with a function for the greens and by performing the same operations one csv file at a time for yellows.

A noticeable issue was the use of two different codifications of the trip location. The datasets of the first six months of the year 2016 provided latitude and longitude coordinates for pick-up and drop-off locations, whereas, in the the following months, the location was recorded using a numeric code, that corresponds to a specific taxi zone of New York. Therefore, in order to make all the data uniform, it was necessary to convert the geographical coordinates into taxi zones. To do so, the *Taxi Zone Shape File* downloadable at https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page has come in very handy. The table provides the coordinates for each taxi zone, described as a polygon. After some manipulation of this type of data, which was of string type, we were able to convert each zone to an actual Polygon object, using the Shapely library, while each pick-up point and drop-off point was transformed into a Point object, using the same library. Then, we just associated each point to the zone that it belongs to, checking which polygons contains the point.

Furthermore, in Databricks, we have taken advantage of another useful table, the *lookup table* (downloadable at the same link provided above), that maps each taxi zone to the borough that it belongs to. This was done in order to assign each location point to larger, and perhaps more meaningful, areas.

This being done, after having checked for inconsistencies in 2017's and in 2018's datasets as well, we performed a sample_and_merge function that takes as input a list of datasets and outputs a single one composed by a combination of a 1% random sample of each dataset. Finally, we set the same column names schema for all the samples. After this operation we ended up with 4 sampled files ('yellow2016(01-06)_cleaned.csv', 'green2016(01-06)_cleaned.csv', 'yellow_merged.csv', 'sampled_greens_2016-07.2018(noEhail).csv') that we have uploaded on Databricks and joined into a single dataset ('taxi') of more than 3M of rows. We performed some more cleaning operations, such as removing all the entries whose *VendorID* was higher than 2 and whose passenger count was 0. We counted 11633 entries with 0$ as total amount, but we have chosen to keep them, as they seemed to provide interesting insights anyway. The column "*Ehail Fee*" was dropped, as the null values were much more abundant than numeric values. On the other hand, we have decided to create some new variables, to get more powerful insights. First, to keep track of the taxi colour, we added the categorical variable "*Colour*". Then, we have extracted time features from the data; hour; a categorical variable to distinguish between morning (from 6 am to 6 pm) and night; day of the week (coded from 1, Monday, to 7, Sunday); a categorical variable to distinguish between work day and weekend.

3. DATA ANALYSIS

3.1 Best vendors

As explained in the introductory section, the taxis are of two types, either Green or Yellow, and the service is provided by two vendor companies, either *Creative Mobile Technologies (1)* or *VeriFone Inc. (2)* recorded as "VendorID". Overall, the Yellow cabs are much more than the Green ones: the Yellow taxis cover 88% of the total. The two companies are more balanced: *Creative Mobile Technologies* represent 42% of the total, while the *VerifFone Inc.* the remaining 48%. As a trivial consequence, the overall gain of Yellow taxis is higher than the overall gain of Green taxis, respectively 30 millions and around 5 millions of dollars; the same can be said for the two companies, with a total amount of 19 millions and 22 millions of dollars respectively (meaning that the proportions are somehow respected). To get more interesting insights about the vendors, we have also calculated the average fare amount, the average tip amount, the average trip distance, shown by **Figure 1**.

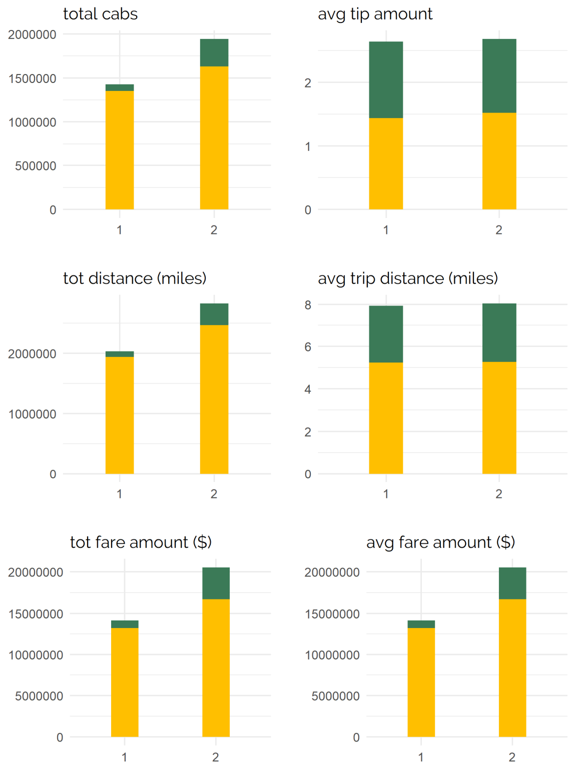

Figure 1: comparison between vendors

On average, regardless the company, a trip costs around 10$, plus 1.5$ of tip. On the other hand, if we consider the colour, yellow cabs tend to cover slightly longer distances than the green ones, respectively 5.2 miles and 2.7 miles. Therefore, the best vendors result to be the yellow cabs of the Verifone Inc. company, although, if controlling for the number of total cabs, there are not huge differences between the vendors.

To get a more general idea of the total amount of money made by the taxi companies, regardless of the type, we have also investigated the trend of the total gain over time, with a focus on 2017 and 2018. As **Figure 2** shows, in the first month of 2018, the gains started to decrease.

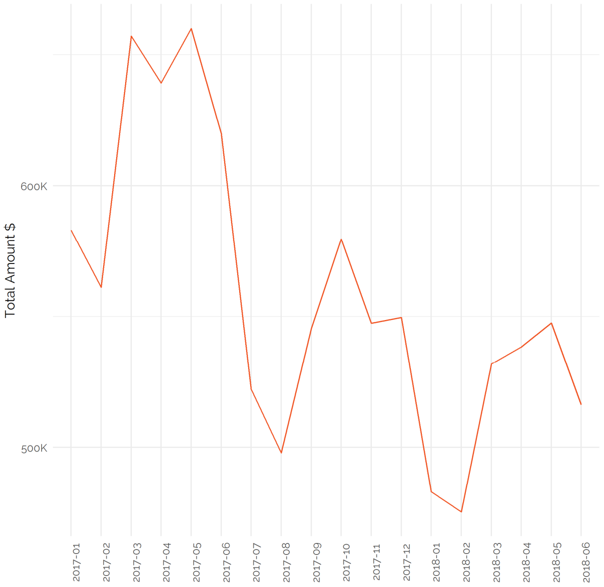

Figure 2: total gain over time

3.2 Best Locations

What are the most popular pick-up and drop-off locations in New York? To answer this question, we have looked at the data from two different perspectives. First, we have taken into account the 5 boroughs of New York, plus the Newark Liberty International Airport (EWR). The busiest area is Manhattan: almost 74% of the cabs either leave from or go to Manhattan. More specifically, here, the drop-offs are more than the pickups, respectively 84% of the total drop-offs and 68% of the total pick-ups, suggesting that a significant amount of taxis come from other boroughs (more about this later on). The second busiest area is the Newark Liberty International Airport, with around 10% of the total trips starting or ending there. Interestingly, the drop-off are much more than the pick-ups. Queens and Brooklyn are equally popular, whereas the trips in Bronx and Staten Island add up to only 2% of the total. Not all the trips are geolocated: for 1% of the data, the pick-up and drop-off location is unknown. We have built a graph in which the node size increases with the number of pick-ups in that node (**Figure 3**).

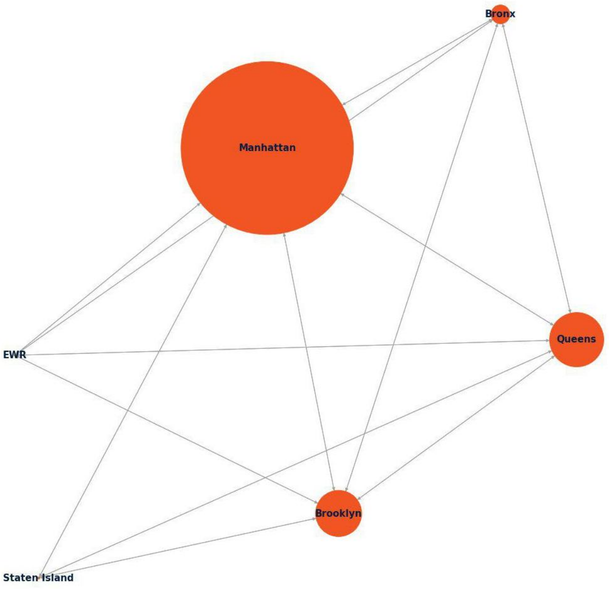

Figure 3: best pick-up locations

3.2.1 A closer look at Manhattan

Since Manhattan resulted to be the busiest borough, we wanted to investigate the trips the take place there, by looking at the most popular zones (**Figure 4**). The top five locations with the most pick-ups and the drop-offs in Manhattan are the same. Penn Station, followed by Times Square, obtains the maximum number of pick-ups; while Times Square and Midtown Center are the ones with the maximum number of drop-offs. Midtown Center is featured at the fourth place for pick-ups, preceded by Upper East Side South and followed by Upper East Side North. The latter is the fourth most popular location for drop-offs, lead up by, again, Penn Station, and followed by Upper East Side South.

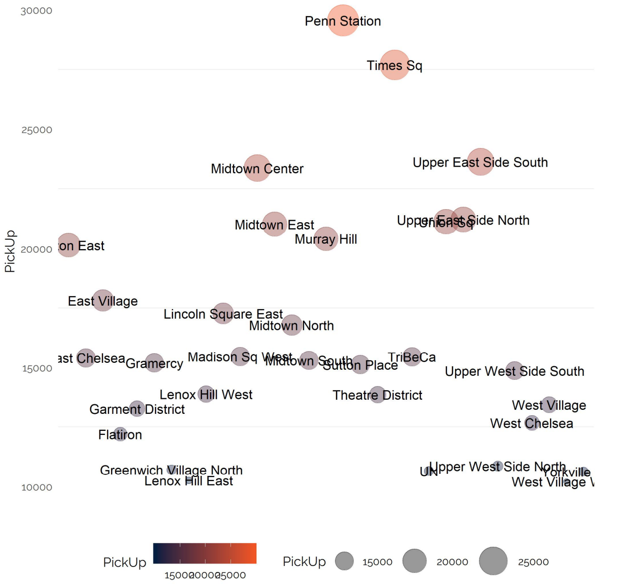

Figure 4: best pick-up zones in Manhattan (>10K pickups)

3.2.2 Borough connections

Which borough do the cabs go to, based on their starting location? We tried to answer this question, building a cross table. As expected, most of the trips end in the same borough where they start, meaning that people take a taxi more often to move inside a borough than outside of it. Cabs from Manhattan tend to go more frequently to Queens and Brooklyn, while the preferred destination of all the other boroughs is, as expected again, Manhattan. A notable exception is represented by Staten Island, as just a few cabs leave the island.

3.3 Total Pick-Ups and Drop-Offs

For each trip, the pick-up and drop-off date was recorded, hence it was possible to determine the peak-hours. As **Figure 5** shows, the trend for pick-ups and drop-offs are similar. The least congested time is between 4 am and 5 am; from 5 am to 8 am, the number of trips increase sharply; from 8 am to 2 pm, it slowly increases; after slightly dropping at 4 pm, the curve quickly goes up until reaching its peak at 7 pm. Overall, the most congested time in New York is between 5 pm and 10 pm. From 10 pm on, the traffic becomes less and less intense.

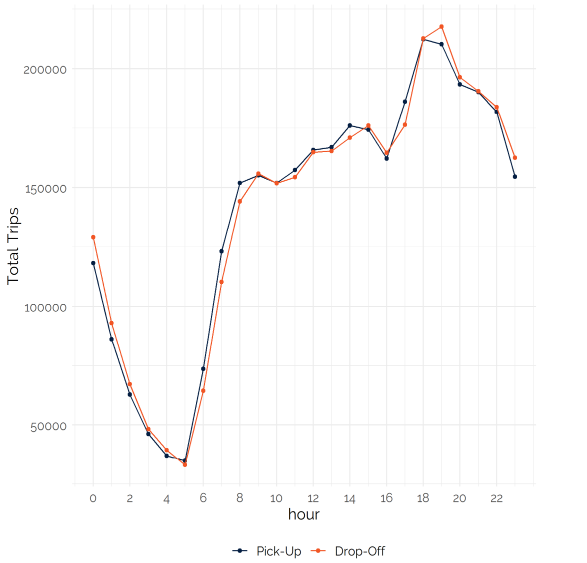

Figure 5: trips by hour

What about the traffic trend in each borough? The peak hour varies, as can be seen in **Figure 6**, and it follows the pattern just described. The only exception is Staten Island, whose peak hour is 11 am.

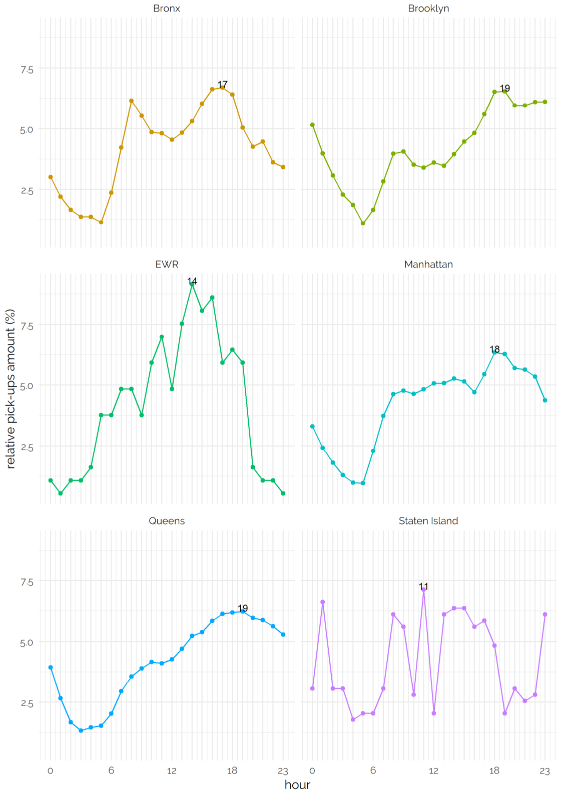

Figure 6: trips by hour and borough

3.3.1 Additional time features

To get a better understanding of the traffic in each borough, as mentioned in the introduction, we created a categorical variable to distinguish between morning week day and weekend. During the weekend, people travel slightly more than during the work days, with an average of 497000 trips during a single work day, and 554000 trips on Saturday or Sunday (**Figure 7**).

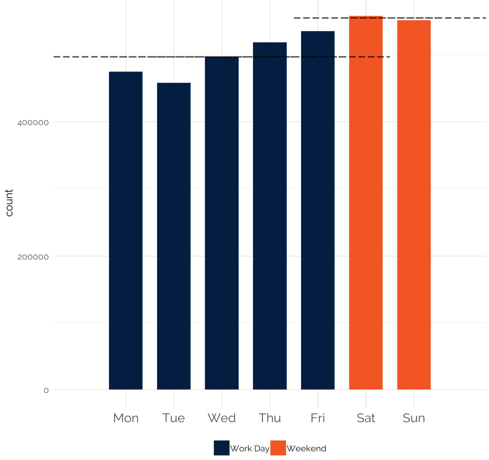

Figure 7: trips by hour and borough

Moreover, if comparing day (6 am – 6 pm) and night (6 pm- 6 am) in each borough, different trends can be discovered. **Figure 8** shows suggests that, in relative terms (meaning, when controlling for the total amount of trips in each borough), nights in Brooklyn and Queens are busier than nights in Manhattan.

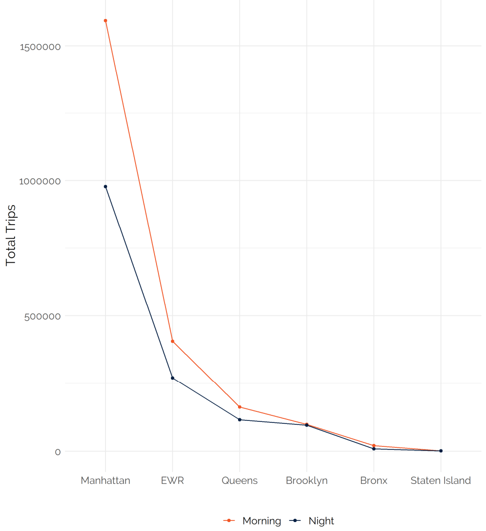

Figure 8: boroughs by day and night

3.4 How fare changes

Lastly, we investigated how fare changes in space and time. There are many elements in determining the price of a taxi trip. It is trivial to consider that two of those are the distance and the duration of the trip. The collected trips have average fare amount 10.28$ (SD 11.42$). To analyse how the average fare amount varies over time and per location, we decided to standardize the amount by dividing it by the mentioned quantities. Trip distance was provided in the original data, while duration has been computed as shown in the time features extraction section, where we chose as unit of measure minutes.

Since a trip is paid when came to end, we selected the drop-off hour to construct a table for the average fare amount through the day (**Figure 9**). The average most expensive trips (more than 15$) are night trips that ended very early in the morning (5am, 6pm); followed by trips belonging to the slot from noon to and 3pm (10$). The lowest part of this ranking is heterogeneous, since it contains non-adjacent hours it is not possible to comment on time slots, and at the same time its average fare amount does not vary much. The average standardized fare amount for morning trips is 8.50$, in the night it seems to be less expensive (5.90$).

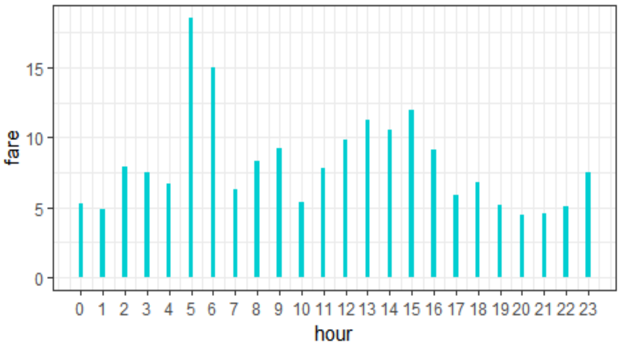

Figure 9: fare by hour

For what concerns the variation of fare amount by locations we average fare amount is wide. There must be some outlier entries in the subset of trips leaving Newark-Liberty Airport (>3.4K$). The lowest average fare amount belongs to trips leaving Manhattan (3.66$). In ascending order, the other boroughs: Queens (13.94$), Brooklyn (18.59$), Staten Island (23.15$), Bronx (56.12$). Similarly, to the pick-up boroughs ranking, the average fare amount is highest when arriving to Newark-Liberty Airport (EWR, 98.90$), lowest when to Staten Island

(4.98$); followed by Manhattan (5.64$), Brooklyn (15.05$), Queens (22.81$) and Bronx (33.93$).

4. DATA EXPLORATION

In order to extract further meaningful information, we have investigated the data through unsupervised techniques, namely k-means clustering. This algorithm partitions the data into groups, by assigning each observation to the cluster with the most similar prototype (the closest mean). To set the parameter k, we have run 10 clustering, with k ranging from 1 to 9, and compared the SSE (sum of squared error) for each k. We have chosen 5 as k, turning to the elbow method, according to which the optimal k is the "elbow" of the error chart, as shown in **Figure 10**.

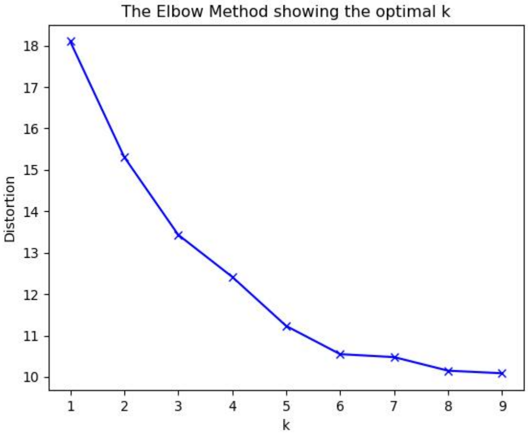

Figure 10: elbow method to choose k

4.1 The features

Then, we have investigated the features of each clusters, summarized by **Figure 11**.

We can identify a cluster of taxis that cover an average distance of 1.5 miles, mostly at 6 pm., at the average price of 9$ per trip (*prediction=0*). This cluster is the largest and is particularly concentrated in Manhattan. There is also a large number of taxis with similar characteristics, except for the most popular pickup hour: these cabs travel around 7 am, hence in the morning, and are very popular in Brooklyn as well as in Bronx (*prediction=3*). Interestingly, there is a cluster of taxis that cover longer distance, with an average of 14 miles per trip, which travel mostly in Manhattan and Queens (*prediction=4*). There is also a smaller group of taxis which travel on average 19 miles (*prediction=1*): most of the taxis leaving from the EWR Airport belong to this group. For this group, the average fare amount is therefore particularly high: 105$ per trip! Finally, there are taxis that cover "medium" distances, 5 miles per trip on average, which are almost evenly spread across all the boroughs (*prediction=2*). As for the colors, it is not possible to catch significant differences between yellow and greens taxis. In fact, for each cluster, yellows cabs are significantly more abundant than the green ones, because of the large imbalance between the two taxi colour distributions.

4.2 The geographical distribution

Using the Python library Basemap, we have also drawn a scatterplot of our data on a NYC map (**Figure 12**). To do so, we have first downloaded 1 million of rows from Databricks (the maximum amount allowed), then we have made a further sampling, by randomly selecting 1% of the data for each borough, except for Staten Island and EWR, for which we have selected 2%. This moderate oversampling in favour of Staten Island and EWR was motivated by the fact that these two areas are so underrepresented in our sample that no point would have shown up on the map.

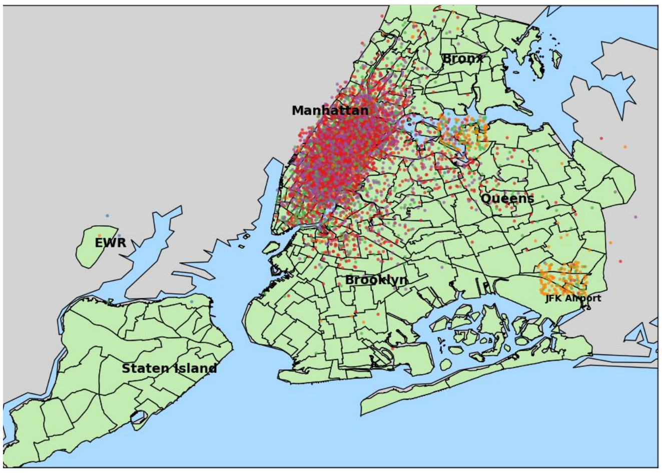

Figure 12: distribution of clusters

Moreover, we have created a graph in which every node represents an area and an edge between two nodes means a set of trips that have been done between them (**Figure 13**).

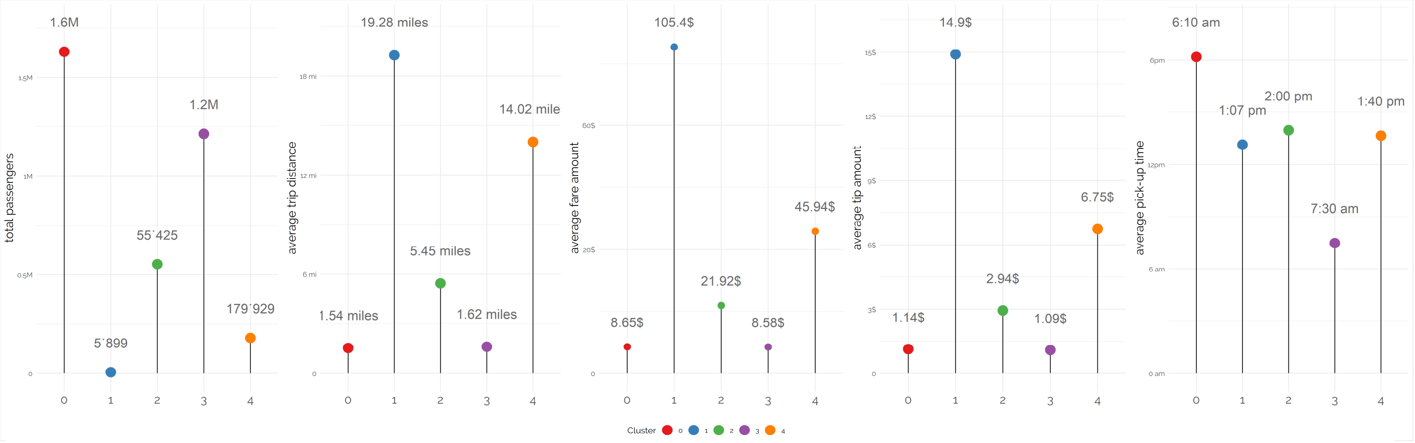

Figure 11: clusters features

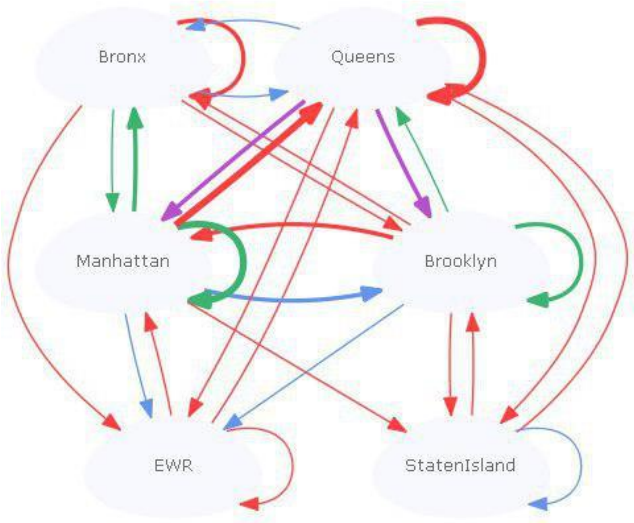

Figure 13: most common type of trip based on clusters

The displayed boroughs graph provides a different analysis perspective from the feature investigation of the clusters, since it is aimed to find the most common kind of trips across and within areas. It has been constructed by placing each borough as a node. The edges coming into a node represent the most prevalent cluster of trips among entries recorded from the given node to its destination. To do so, we extracted the linked areas from the sample with clusters file and used as edges colour attribute the predicted cluster. Self-loops starting from a node and ending in the same one stand for trips with pick-up and drop-off locations within a borough. For what concerns the size of the arrows, we decided to apply different thickness according to the entity of the trip stream. Smaller arrows indicate sets of trips under a thousand records. Thickest ones, above the ten thousand. The medium sized directed edges are between these quantities.

We can see that the most common cluster of trips amongst the represented arrows is the largest one (*prediction=0*, displayed in red). The fifth prediction is absent at this level of data aggregation. The only purple edges (*prediciton=4*) originate from Queens, one towards Manhattan and one to Brooklyn.

5. CONCLUSION

For what pertains clustering, it is interesting to notice that the total passenger number and the average fare amount are two of the features distinguishing trips groups the most. This result is in line with what we expected. Moreover, we presumed that clusters would have gathered in almost distinct locations. Surprisingly, it came out that our hypothesis was naive. Clusters are entangled, as in the geographical distribution map, in the busiest area with almost no clear dispersion pattern. A teasing further analysis of the inspected data could be performing the same classification of trips with other unsupervised approaches and compare the results with what we found. Another possible direction for further analyses might be to investigate whether, in the last months, the overall amount of taxi trips is still decreasing or not, and to investigate the reasons of this trend. In the broader objective of widening the analysis on NYC traffic it would be to merge this data with pick-up records of other transportation network companies, for instance Uber. This would allow us to make some considerations about the traditional taxi service and the effects of the sharing economy.