New York Taxi & Limousine Commission:
a “Big Data Technologies” Course Data Analysis

Fabio Fusato (ID 204980), MSc in Mathematics for Life and Data Science, *fabio.fusato@studenti.unitn.it*
Federico Sparapan (ID 204580) MSc in Data Science, Curriculum B, *federico.sparapan@studenti.unitn.it*
Francesco Maria Marrone, (ID 205181) MSc in Informatics, *francesco.marrone@studenti.unitn.it*

ABSTRACT

The paper features a Data Analysis study on New York City Taxi service from 2011 to 2018. The work is focused on Clustering methods, statistical comparison between yellow and green cabs as well as taximeter providers.

Moreover, it features specific analyses performed grouping traffic data by location and borough, with particular attention to airports and key locations with a high number of pickups and drop-offs.

The number of trips changes through single days, choosing common days and eventfully days, were also a subject of analysis.

Keywords

NYC, Taxi, Spark, DataBricks, Jupyter Notebook, Visualization, Statistics, Trend, Seasonality, Clustering, Dataset, SQL, Data Analysis, Machine Learning,

1. INTRODUCTION

In New York City, taxicabs could be yellow or green, and they are widely recognised as symbols of the city. Taxis painted in canary yellow are able to pick up passengers anywhere in the five boroughs. Those painted in green apple, which began to appear in August 2013, are allowed to pick up passengers only in Upper Manhattan, the Bronx, Brooklyn, Queens (excluding La Guardia Airport and J.F. Kennedy International Airport), or Staten Island. Both types apply the same fare rules. Also, all taxicabs are operated by private companies and licensed by the New York City Taxi and Limousine Commission (TLC)[1].

2. METHODOLOGY

Since this work features some Machine Learning implementations, we used both DataBricks to work with Spark and Jupyter Notebook to access datasets rapidly and to draw better plots.

We used several Spark libraries, which included VectorAssembler, SQL and Functions. We also imported Python libraries such Pandas, Geopandas, MatPlotLib and Numpy.

As of the datasets, we downloaded data from every month between years 2011-2018 of NYC Taxi Trips, both from Yellow and Green Cabs, that are available at: “https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page”. From the same website we also downloaded the file “*Taxi Zone lookup Table*”. Due to hardware limitations, we extracted a random 30k rows dataset from each month, to be used as a sample for our study. We started from 2011 since the 2010 dataset was difficult to randomize due to problems inside the dataset itself.

To carry out a more realistic analysis, we also selected some complete-month datasets.

The code in the attached folder is organized in files, named as the section their code refers to.

3. DATA ANALYSYS

Starting from a Data Cleaning process, we were able to increase the manipulability of our datasets.

3.1 Data Cleaning

After having downloaded all data, we randomly extracted 30k rows from each set and then we matched all months, resulting with 8 yellow datasets and 6 green ones. Unfortunately, some dataset’s column names changed over the year, so we had to rearrange them in order to make them fit the matching process.

One of the biggest challenges was due to the fact that until June 2016 pickups and drop-offs coordinates were marked as longitude and latitude, while since July 2016 they are identified with an ID number. We converted 2011-2016 coordinates into ZoneID using a function written in Python.

Since there is a total of 265 zones, we decided to add the Boroughs to our datasets, in order to make more general analyses and to improve data visualization. To do this we joined the file “*Taxi Zone lookup Table*” with our datasets. The file was downloaded from the same website and it features the ZoneID as well as the correspondent Borough. We ordered them into *zona_prelievo* and *zona_scarico* columns.

After this, we modified all datasets following the structure of 2018 dataset, changing the values of the columns *VendorID* and *payment_type* and also renaming some columns that changed name over the years. This allowed us to smoothly match all months.

In order to make an analysis on single days and hour, we added the columns *anno, mese, giorno, ora* to our datasets.

We finally decided to drop rows with *zona_prelievo* and *zona_scarico* marked as “Unknown” and with *payment_type* marked as “Dispute”, “No Charge” or “Unknown”. In the first case, due to our kind of analysis, it did not make sense to analyse trips with unknown locations. In the second one, the number of that cases was significantly low, making statistics not relevant to this study.

After the Cleaning process, our datasets featured about 350k rows and 21 columns per year, with the exception of Green 2013 (the service started on August, so the dataset has about 150k rows) and Green and Yellow 2018 (Data are available until June, hence about 180k rows).

3.2 Clustering

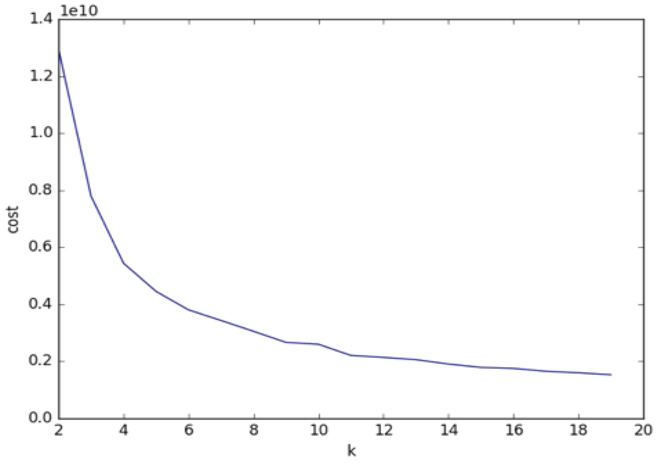
3.2.1 Green Taxis

In order to perform a clustering algorithm over the whole green dataset we joined the annual dataset in a unique one called “*total_verdi*”.

Implementing the experience acquired during lectures, we chose the clustering method "*K-Means*". Firstly, we defined a dictionary to transform the string values in '*store_and_fwd_flag*' in integer type values and we were then able to launch the algorithm.

As "*features vector*" we used all the columns of the dataset, with the exception of "*pickup_datetime*", "*dropoff_datetime*", "*zona_prelievo*" and "*zona_scarico*". We could not exclude *a priori* which column to use or not to use, so we decided to exclude only columns with redundant information.

We tested it with various values of *k*, from 2 to 20, that are the number of groups in which the dataset will be divided. In order to identify the best *k* value, we used an *elbow function* which allowed us to choose the right *k* in relation to the computation cost[2]. We then extracted three values of *k*, equal to 10,11,12. These values correspond to the moment where the elbow function becomes flat, so the increase of *k* is not decisive in relation to the reduction of the computational cost. The graph of the function is showed below.

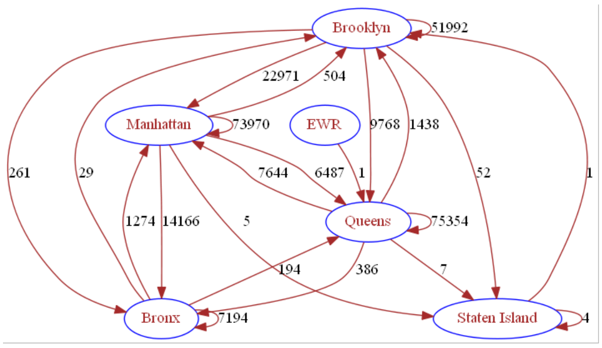


For these three values we performed the *silhouette performance*[3] value, obtaining the following values:

- K = 4, silhouette value: 61.6%
- K = 5, silhouette value: 57.2%
- K = 6, silhouette value: 57.8%

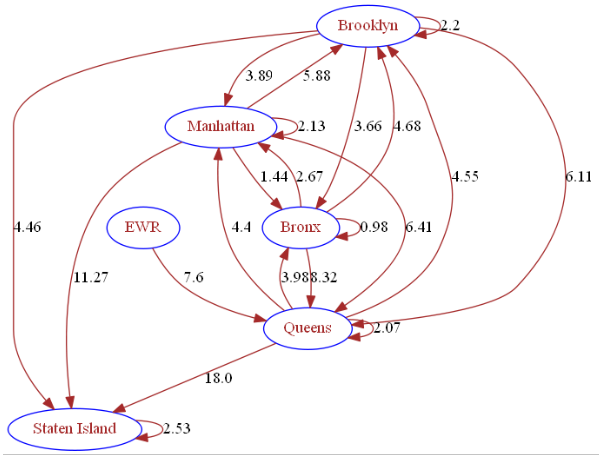
We choose 6 as *k* value due to the fact that the computational cost of *k* equal to 4 is not worthy to pick this as value. At the end, we computed a few analyses over the record in the "zero" set prediction.

The following graph illustrates the trips average between boroughs:



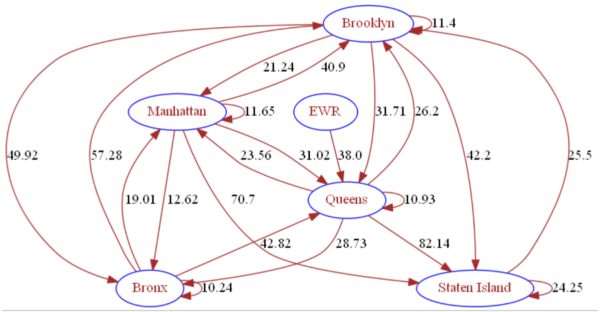
It is clear that Queens, Manhattan and Brooklyn are the boroughs with most traffic in terms of taxi trips, while Staten Island, Bronx and Newark Airport (EWR) record a smaller number of trips. Ahead in the study, we will check if these group data correspond to the entire dataset.

Following this, we analyzed the tip amount between boroughs:



It is clear that the trips inside one of the three most busy borough correspond to the same number of tips, which is around 2 $. It is also interesting to notice that routes with lower traffic, like Manhattan-Staten Island or Manhattan-Brooklyn are at the same time the routes with the highest tips.

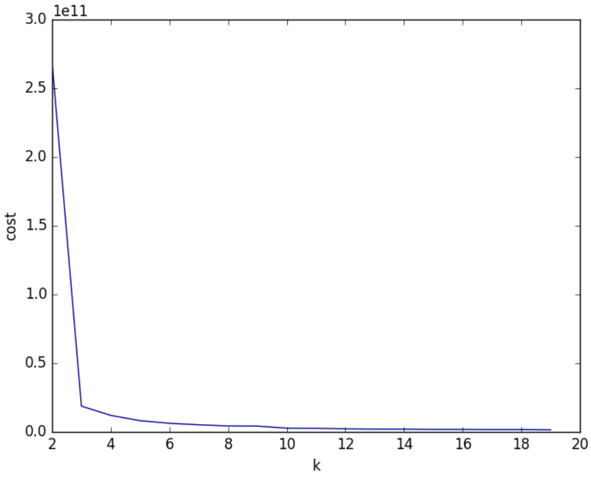
The graph below shows the average fare amount between boroughs:



As expected, routes with pickup and drop-off within the same borough are the cheapest, while routes between different boroughs are the most expensive.

3.2.2 Yellow Taxis

Using same algorithm for the yellow dataset, we detected an issue with the *null values* in the dataset. In order to overcome it, we used the method "*setHandleInvalid("skip")*" to make null values irrelevant. The result of the algorithm was the following *elbow function*:



We set a value of *k* equal to 3 but the clustering was meaningless since almost all records were included in the same subset, leaving the other two almost empty. For this reason, the silhouette performance presented a value near 100%:

- K = 3, silhouette value: 99.9%

Setting parameter *k* equal to 6 in order to obtain similar results as the green clustering, a silhouette of 55.1% was obtained.

Due to the similarity with the green one, the analyses done for the yellow cluster were omitted.

3.3 Vendor and Company rankings: which is the best?

The VendorID marks the record's provider for each trip. The NYC Taxi & Limousine features two types of Vendors: the first one, marked as "1", is from the company "*Creative Mobile Technologies, LLC*"; while the second one, from "*VeriFone Inc.*", is marked as "2". Each has different installations, removing procedures, prices etc[4].

We detected that the yellow cabs feature 49% - 51% of the vendors, while the green ones correspond to a percentage of 80% of "*VeriFone Inc.*"

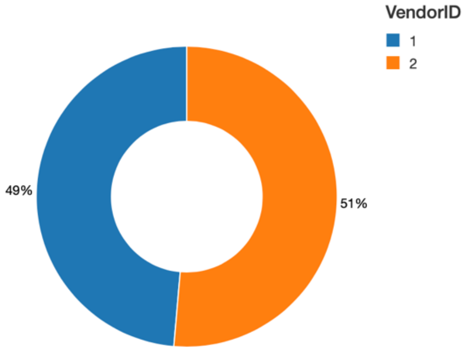


Figure 1a. Yellow Taxis Vendor comparison

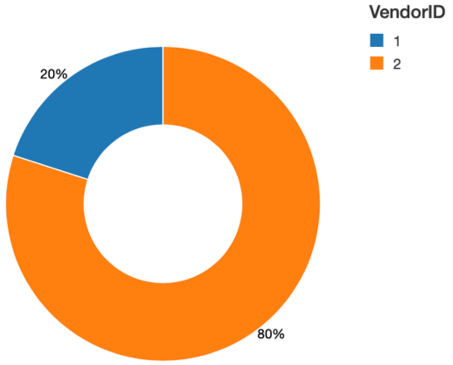


Figure 1b. Green Taxis Vendor comparison

3.3.1 Distance Analysis

The first analysis performed was a comparison of the average distance:

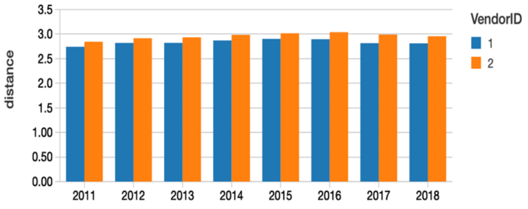


Figure 2a.. Yellow average Distance

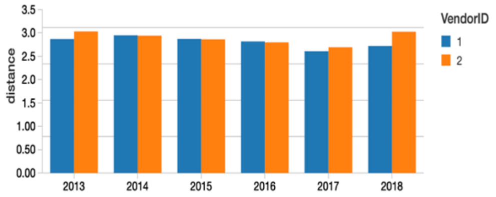


Figure 2b. Green Average Distance

The analysis showed that both have the same average, that is around 2.8 miles. It is also clear that it remains stable over the year, with no significative changes.

3.3.2 Who earns more?

An interesting parameter to check is which company, between yellow and green taxis, earns more money. The analysis was performed excluding tips, taxes and tolls.

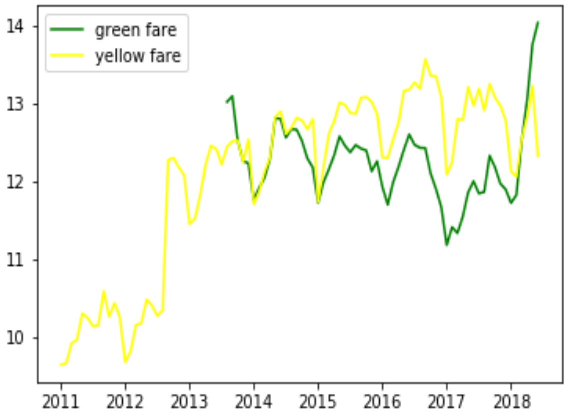


Figure 3. Monthly average fare amount

There is a variation of about one dollar between yellow and green, with the yellow reaching a peak in 2016 and greens with a low point in 2017. The same number of trips per year was taken for both companies. For this reason, it is possible to conclude that, until 2018, a trip with the green taxis costed less than the yellow.

The graph shows another important characteristic: there are significant differences between months of the same year. In fact, during the winter season, fare amounts are quite lower than in the summer season. In statistic, this phenomenon of Time Series is called "*seasonality*"[5]. We decided to check whether this is also a feature of the trips distance:

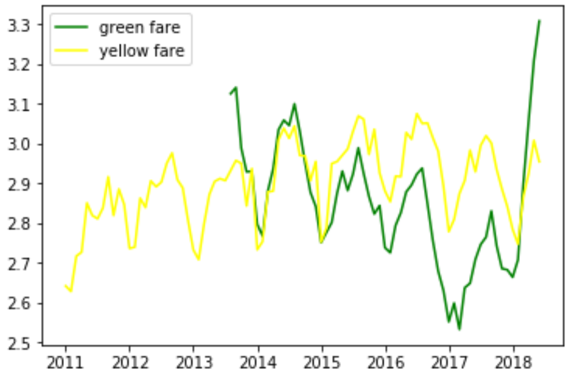


Figure 4. Monthly average trip distance

The graph shows that there is a seasonality correlation between fares amount trends and trip distance. Despite distance variations are not as significant as fare amounts at the end of 2012, both graphs present similar seasonality trends.

It is also interesting to notice that the trends of the companies, both for the fare amount and the distances, follow opposite directions: as the yellow trend grows, the green one decreases, while from 2017 it is the opposite. These trends are better shown in the graphs below, grouped by years:

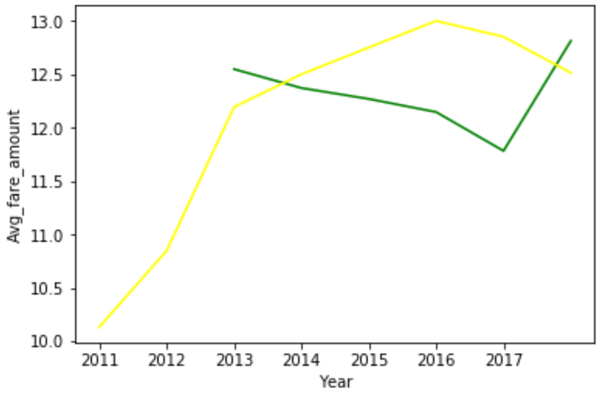


Figure 5a. Annual average fare amount

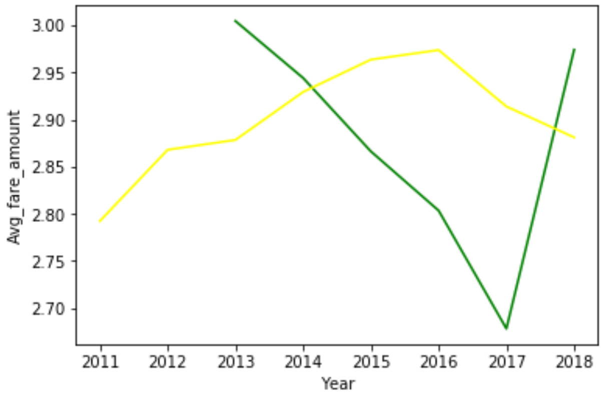


Figure 5b. Annual average trip distance

We want to see if there are differences between vendors, we already showed that there are no difference concerning the distances, let's see if there are for the fare amount:

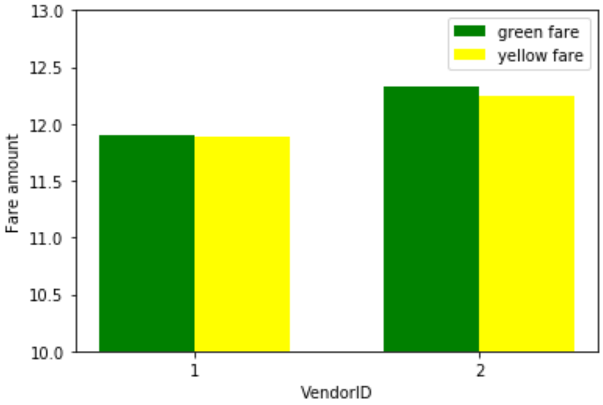


Figure 6. Fare amount by Vendor

We see that results are quite the same, without significant difference between *Creative Mobile Tech* (marked as "1") and *VeriFone Inc*. ("2").

3.3.3 Which one is the most common payment type?

It is interesting to understand, between the two possible kinds of payment, which one is the most used across the years. In order to perform this analysis, we grouped green and yellow according to years and payment type. We then counted the number of trips corresponding to each payment. The results are showed below.

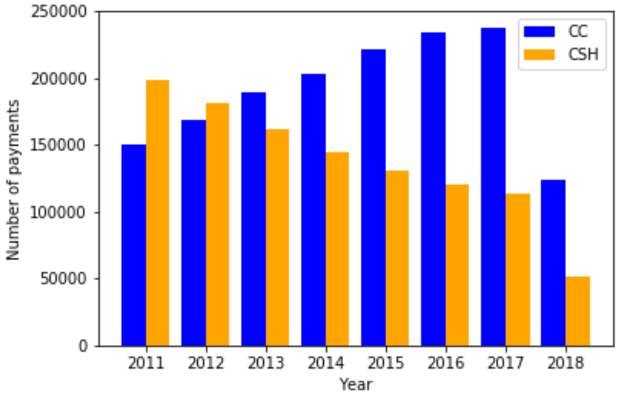


Figure 7a. Yellow Taxis payment type

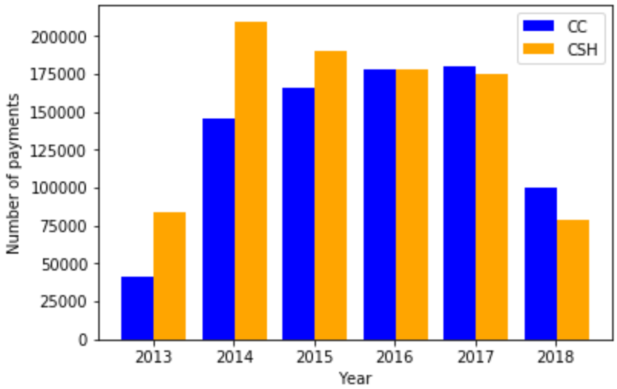


Figure 7b. Green Taxis payment type

There is a relevant difference between yellow and green: while in the first case we notice an inverted trend throughout the years, which starts with a big amount of cash payments decreasing in favour of credit card payment, as for the green cabs we notice consistent trends for both payment solutions.

3.3.4 Which one receives more tips?

Finally, to identify which company performs best among the two NYC taxi service providers, it is needed to analyse which one receives higher tips. Payment type by cash doesn't allow tips, that are set to 0. Tips are instead are delivered only with a payment by credit card. In order to conduct a more relevant analysis, all kind of payments are included.

Year 2017 was chosen for both companies (350k rows each), and the analysis showed that yellow cabs received 227,688 tips, while green 148,660 tips. The graph below illustrates which ones received the highest tips:

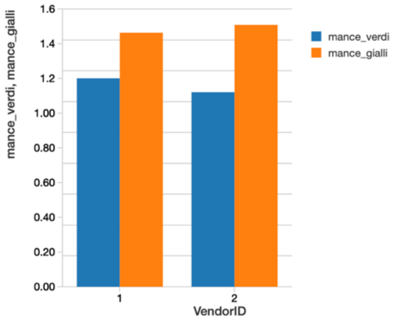


Figure 8. Tips by vendor

Again, yellow cabs (in orange) are better than green cabs. The tip amount is not related to the Vendor, but it is significantly related to the Taxi company.

3.3.5 Comparison conclusions

Following this analysis, observing that with the same fare distance fare amount and tips are higher for the yellow taxis, it is possible to conclude that is better to be a yellow taxi driver. The Vendor company does not affect the trips.

It is noticeable that, starting from 2018, data related to the average distance and the fare amount is changing in favour to green taxis. A future analysis might be useful to identify the duration of this trend.

3.4 Single-day analysis

3.4.1 Introduction

In order to improve the study, single days from the complete dataset were also analysed. The numbers are referred to real records. For this analysis, only yellow cabs were used.

First, we chose some significant days. Then, we observed the average of the trips per day of the same months, and picked two days, one in weekend and the other within the week, in order to compare a regular day to a holiday one.

The significant days chosen are: SuperBowl 2012 (February 5th), Hurricane Sandy hitting New York (30 October 2012), Trump election day (8 November 2016), 2016 Thanksgiving Day (24 November).

3.4.2 February 2012: Super Bowl

This Super Bowl event choice is due to the popular match between the New York Giants team versus New England Patriots.

A check of the average of February 2012, showed that the 5th was not significantly low in numbers of trips: 466,000 versus an average of 516,000. The 5th of February was not day with the lowest number of trips (which was February 19th with 420,000 trips). For this reason, it is possible to state that the SuperBowl, even when the city's team is playing, does not significantly influence the mobility of the city. Therefore, there was no reason to proceed with further observations.

3.4.3 October 2012: Hurricane Sandy

On October 29th Hurricane Sandy, the deadliest and most destructive Hurricane of 2012, hit New York City flooding streets, tunnels and subway lines and cutting power in and around the city.

In the graph below, it is shown how it has affected the city, by observing the number of taxi trips. Excluding the last three days, October 2012 presented a trip average of 510,000, that went down to 468,000, last three days included. To allow an easier interpretation, October 10th was chosen as comparison date (500,000 trips).

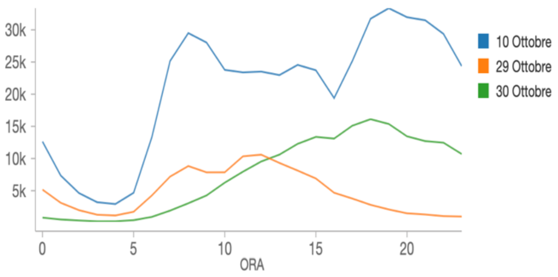


Figure 8a. 2012, October

The difference between a regular day and the other two appears clear. First of all, the number of trips on the 29th and 30th October were one-fifth compared to the other; secondly, from 12 a.m. of the 29th the trips slowly decrease in number, reaching a minimum of 243 trips on the 30th at 3 a.m., in contrast with the 3203 trips at the same time on October 10th.

3.4.4 November 2016: Trump Elections and Thanksgiving Day

Computing the average of the trips per day, we saw that it corresponded to about 336,000 trips per day. The random days selected were Sunday the 13th and Wednesday the 30th, due to the fact that the trips were the same as the mean. The day of the elections, the 8th, the number of trips was 319,000 while on Thanksgiving Day it was 220,000. Below the distribution grouped by hour:

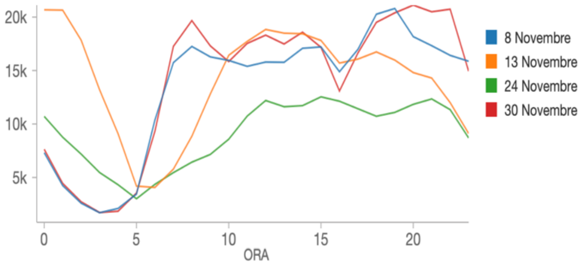


Figure 8b. 2012, November

The 8th and the 30th follow similar trends, while the 13th reaches a peak in the first hours of the nigh, because people use cabs after midnight to come back home after a Saturday night. Thanksgiving Day is between the two trends, with a significantly lower number of trips, more linearly distributed throughout the day. It is possible to conclude that that, while there is a difference between weekends and week days, there is no difference between a regular day and the election day.

3.5 Location analysis: which one is the best airport in NYC?

We compared the three airports reachable by taxi, that are JFK, Newark and LaGuardia, to check which is the one that generates more traffic. The biggest one is JFK, with a 2017 traffic of 59 million people, followed by Newark (EWR) with 43 million and LaGuardia with 30 million[6]. Due to the fact that green taxis do not serve JFK and LaGuardia, as we said in paragraph 2 of this paper, the 2017 yellow sample was used to perform this analysis. This will be useful to confirm whether airport passenger traffic corresponds with the number of trips.

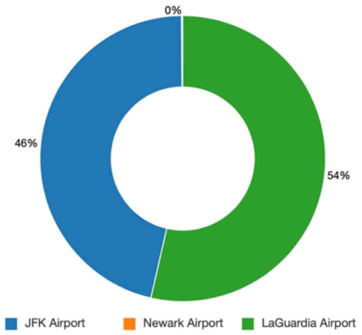


Figure 9a. 2017 airports pickups

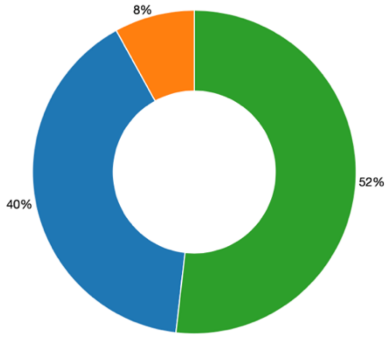


Figure 9b. 2017 airports drop-offs

Surprisingly, the trips show a completely different trends: despite all airports being reachable by cabs, LaGuardia is the most chosen by the people, even if it is the less travelled. Has this something to do with the prices?

Table 1. Average airport trips prices

Airport	Avg. Fare amount	Avg Total amount
JFK	46.67$	56.62$
EWR	73.34$	88.97$
LaGuardia	32.37$	42.73$

The analysis confirms the fact that prices (analysed excluding tips and tolls), could be a reason for not taking the taxi to JFK and EWR, preferring other kinds of transport (the train for and from JFK, for example, costs only 7.5$, versus an average of 47$ + tip and tolls of a taxi trip).

3.6 Analysis by Borough

The analysis has been conducted both on yellow and green dataframes to understand which one, between the two companies, recorded the highest number of trips over a period of seven years. We use as sample the six zones of Manhattan in which either green or yellow cabs recorder more pickups. The choice of Manhattan as borough comes from the fact that for the yellow taxis it is by far the borough with most pickups and drop-offs, while for the green taxi, even if it is not the most popular borough, it presents similar number of pickups to the first one (Brooklyn). Below are the relatives pie graphs:

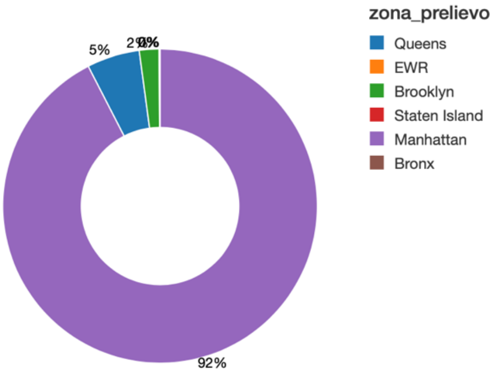


Figure 10a. Yellow pickups percentages by borough

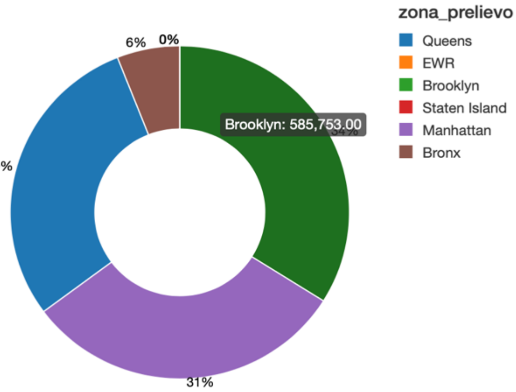


Figure 10b. Green pickups percentages by borough

We then performed the analysis of the three zones inside Manhattan which have the most pick-ups per company: these are, for the yellow, the Upper East Side South, Midtown Center and the Upper East Side North, while for the green are East Harlem North, Central Harlem and East Harlem South.

As illustrated in the bar graph below, it is possible to notice how in both cases a frequent pick-up zone corresponds to a very infrequent for the other company. This might be related to a non-compete clause between the two companies.

According to the official site of the NYC TLC[1], indeed, yellow taxis are allowed to operate in south Manhattan, while the green taxis cover a small part of north Manhattan and the other boroughs. This analysis, by showing the trips distribution between boroughs and zones, confirms it.

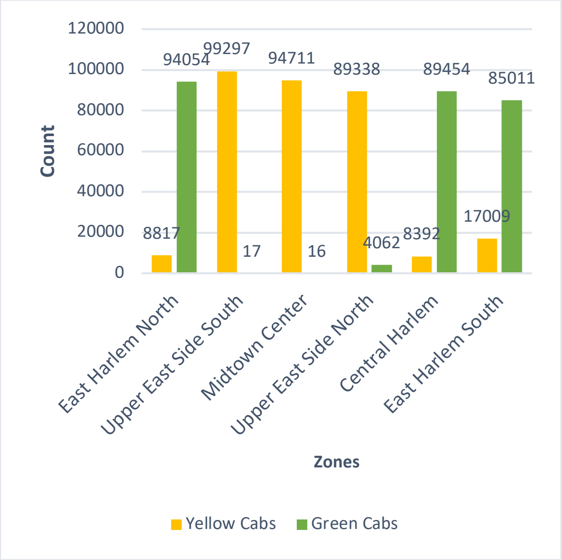


Figure 11. Zones with the most pickups

4. POSSIBLE FURTHER ANALYSES

Concerning the Clustering method, it was worth to try a PCA (Principal Component Analysis) and a Dimensionality Reduction[7].

The first would allow us to properly decide which columns are the most important to be included in the "*features*" vector, while with Dimensionality Reduction it would be possible to reduce the dimensions of the "*features*" vector and perform a better subsets visualization and comprehension.

The use of tools from statistics of stochastic processes (trend and seasonality detection, fitting an ARIMA model on data) is advisable to make more advanced forecasts on future values of time series.

An application to this case of study is the prediction of future values of fare amount and trip distances analysed in paragraph 3.3.2 or in the payment type analysed at 3.3.3.

By observing the borough's analysis (3.5), it is noticeable that, in Manhattan, green taxis cover the poorer zones of the borough[8]. Therefore, a social analysis of the phenomenon could be also very relevant.

5. REFERENCES

[1] NYC-TLC. New York City Taxi & Limousine Commission. Retrieved from: https://www1.nyc.gov/site/tlc/vehicles/get-a-vehicle-license.page

[2] Kodinariya, T.M., Makawana, P.R. *Review on determining number of Clòuster in K-Means Clòustering.* International Journal of Advance Research in Computer Science and Management Studies, Nov. 2013.

[3] Rousseuw, P.J. Silhouettes: *A Graphical Aid to the Interpretation and Validation of Cluster Analysis.* Computational and Applied Mathematics. 20: 53-65. 1987.

[4] NYC-TLC. New York City Taxi & Limousine Commission. Retrieved from: http://www.nyc.gov/html/tlc/html/industry/authorized_tpep_providers.shtml

[5] Brockwell, P.J., Davis, R.A. *Introduction to Time Series and Forecasting.* Edited by Springer, New York, 1996.

[6] The Port Authority of NY & NJ. *Airport Traffic Record.* 2017

[7] Keogh, E., Chakrabarti, K., Pazzani, M,.Mehrotra, S. Dimensionality Reduction for Fast Similarity Search in Large Time Series Database. *In Knowledge and Information System*, Springer, Irvine, California, 2001.

[8] Wallace, D., Wallace, R. *Life and Death in Upper Manhattan and the Bronx: Toward an Evolutionary Perspective on Catastrophic Social Change.* Research Article, July 2000