Understanding the NYC Taxi Service

Willi Menapace
203778
willi.menapace@studenti.unitn.it

Luca Zanella
207520
luca.zanella-3@studenti.unitn.it

Daniele Giuliani
203508
daniele.giuliani@studenti.unitn.it

ABSTRACT

Big Data is rapidly becoming a predominant field in ICT. Its importance is continuously growing due to the huge volume and variety of available data. In this work we try to exploit its descriptive and predictive power in order to better comprehend the evolution and current state of the taxi service in New York City, leveraging the service data that the NYC Taxi and Limousine Commission (TLC) made publicly available.

We analyze a wide spectrum of views, starting with a general overview of the service evolution in time and confronting it with with the emerging phenomenon of ride-hailing applications such as Uber, Juno and Lyft. We then proceed to perform a segmentation of the traffic in order to understand better the characterization of taxi customers and later continue by analyzing subjects such as movement patterns of traffic, competition between yellow and green taxis and characteristics of airport traffic.

The results we present should help the reader understand the main trends and characteristics of the taxi service, which may suggest ways to increase the quality and popularity of the service.

1. INTRODUCTION

The NYC Taxi and Limousine Commission (TLC) has publicly made available a dataset containing data about taxi trips performed in New York from January 2009 to June 2018. This study aims at exploiting this dataset to understand how the New York taxi service works.

The report describes step by step how our analysis is performed. It starts from data acquisition, compression, transformation and cleaning and then proceeds analyzing different aspects of the data in order to provide the reader with an overview of how the taxi service in New York works.

The main technology enabling the analysis of such dataset is Apache Spark, coupled with R for data visualization.

2. DATA ACQUISITION AND COMPRESSION

The main dataset used for the analysis is publicly available at `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`. We download the data for the time period from January 2010 to June 2018, both for Yellow and Green taxis which consists of 200GB of comma separated value files containing over a 1 Billion records. The schema of the dataset varies by year and taxi company for a total of 6 different schemas. The main columns available through the datasets refer to pickup and dropoff datetimes, trip distance, number of passengers, fare amount, total amount, tip amount, extras and taxes, tolls, ratecode and payment type. The most notable difference between the data is the information about pickup and dropoff locations which is expressed with latitutes and longitudes until 2016 and with numerical taxi zone identifiers for the following years.

After acquisition of the data we proceed to compress it to .tar.gz format in order to handle them more effectively. The compression is significant, with a resulting dataset size of 43GB. The gzip format however is not particularly suited for big data analysis. The parquet file format provides significant performance advantages such as columnar storage, which allows computations that only need specific columns to access them selectively, specific compression algorithms for each column with dictionary specific encodings and the possibility to decode the file even partially, which is convenient in distributed environments such as Spark. The size of the dataset is slightly reduced to 39GB.

3. DATA TRANSFORMATION

The parquet dataset is then transformed into a common schema format to uniform the data and ease the analysis. We decide to adopt the following schema:

- `taxi_company`, "green" or "yellow" according to the taxi type that served the trip
- `pickup_datetime`, date and time of the pickup
- `dropoff_datetime`, date and time of the dropoff
- `pickup_location_id`, taxi zone id where the pickup happened, -1 if the zone is unknown

- `dropoff_location_id`, taxi zone id where the dropoff happened, -1 if the zone is unknown
- `passenger_count`, number of passengers onboard
- `trip_distance`, the metered distance in miles
- `ratecode_id`, the fare rate used for the calculation of the amount, which includes special rates for JFK or Newark trips.
- `fare_amount`, the fare amount for the trip in $
- `tolls_amount`, the amount of tolls in $
- `total_amount`, the total amount for the trip in $. It includes tips only for credit card payments.
- `mta_tax`, the amount due to MTA tax in $
- `improvement_surcharge`, the amount due to the improvement surcharge in $
- `extra`, the amount due to extra in $ including only rush hour and overnight charges
- `tip_amount`, the tip amount in $. It is set only for credit card payments
- `payment_type`, the used payment method

Note that the common schema codifies the pickup and dropoff locations as the ids of the zone where the pickup or the dropoff happened and not as coordinates. The main challenge for schema conversion is the transformation of dropoff and pickup locations expressed as coordinates into the respective taxi zones. The dataset is accompanied by a shapefile which specifies the geographical boundaries of each zone. Unfortunately, naive algorithms for dataset conversion directly using the shapefile are able to only convert some tens of records per second per processor which would make the conversion of 1 Billion of records unfeasible. Instead, we decide to develop a more efficient algorithm for zone association which makes use of look up tables to increase the performance of the conversion up to the thousands of records per second, allowing the complete conversion of the dataset.

3.1 Coordinates to Zone conversion algorithm

Until 2016, information about pickup and dropoff locations is represented through latitudes and longitudes expressed in the World Geodetic System (WGS) coordinate system. The need to have a consistent schema through the years lead us to develop an efficient algorithm to convert the information about pickup and dropoff locations to numerical taxi zone identifiers.

According to the shapefile provided with the dataset, the city of New York is divided in 265 zones and the geometric locations, representing each of these zones, are expressed through the NAD83 / New York Long Island (ftUS) coordinate reference system. Our first attempt is to divide New York in a 1000x1000 lookup matrix in which each tile is delimited by some coordinates and is associated with a list of possible zones which are present in these boundaries in order to restrict the search. This is achieved by including each shape record, representing a specific zone of New York, in the tiles forming the smallest rectangle surrounding that zone. Due to this coarse approximation, each tile may contain many zones.

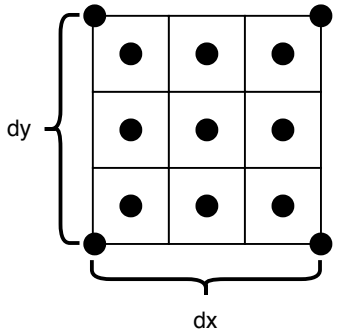


Figure 1: The probing points for a single tile used for checking which zones are actually contained in it. dx and dy, the width and height of each tile, correspond to approximately 70m.

Given the latitude and the longitude of a location in the correct coordinate system, the numerical taxi zone identifier can be obtained by looking at the corresponding tile inside the lookup matrix, scanning each contained shape and retrieving the shape record containing that point. This algorithm is a notable improvement over the naive method, being able to convert some hundreds of records per second for each processor, but still does not allow for the conversion of the whole dataset. The main problem is the need to still linearly search through a list of shapes, although its average length is considerably reduced with respect to the naive implementation. The optimization that allows us to reach the target performance is to refine the coarse list of possible zones for each tile until the point where we directly associate each matrix tile with a unique zone id instead of a list of possible zones. This is possible because by using a 1000x1000 lookup matrix, each tile corresponds to a rectangular region of circa 70x70m, which is granular enough to be associated with a single zone even in the case where it lies across zone boundaries. This zone refinement is achieved by defining a set of probing points, such as the ones represented in fig. 1, and checking which zones in the coarse list for the current tile actually contain at least one of the probing points. Whenever a tile is associated with more than one zone even with this refinement, we randomly choose only one for performance reasons. Due to the high matrix granularity, this situation occurs in a minority of regions and when it happens it causes tolerable approximations. The performance achieved through this optimization allows for the conversion of roughly 4000 records per core per second, which makes it possible for an 8 thread processor to convert the whole dataset overnight.

The dataset in common format for years 2010 to 2018 resulting from the conversion has a size of 24GB.

4. DATA CLEANING

An analysis of the dataset in the common format highlights data quality problems in some of its entries. To increase the quality of the analysis we decide to perform some data

cleaning on the dataset. Null entries are a minority of the dataset, so we decide to drop every entry in the dataset with a null attribute. Every categorical feature is then enforced to assume a value in its set of legal values. In particular, location ids are enforced to be valid ids and each entry for which our coordinate conversion algorithm was not able to identify a correspondent taxi zone is dropped. We then plot the distribution of the values of each numerical feature and conservatively identify a point in its tail after which all entries are dropped, for example tolls amounts greater than 120$ are discarded because very irrealistic according to the distribution of the tolls and probably symptom of poor data quality. We also conservatively drop entries for which the duration is greater than 24 hours or for which the year is not in the range from 2010 to 2018.

The resulting dataset contains 1.07 Billion entries versus the original 1.38 Billion entries and the size of the dataset reduced to 20GB.

5. PRELIMINARY ANALYSIS

The first phase of the analysis is the study of the domain. The New York taxi service is operated by private individuals owning taxi licenses. Two kind of taxi cabs are present, each one with specific characteristics and limitations: yellow cabs, also called medallion taxis, and green cabs, also called boro taxis. The city of New York is divided in five different districts called boroughs, shown in fig. 2: Manhattan, Bronx, Brooklyn, Queens and Staten Island and each one is further divided into taxi zones. Both yellow and green taxis share the same fare system, but, while yellow cabs are allowed to pickup passengers anywhere in the five boroughs, green cabs are not allowed to pick up passengers in South Manhattan, LaGuardia Airport and JFK Airport, the most profitable zones for yellow taxis, but are allowed to drop the passengers anywhere. The reasons for the limitations of green taxis are two. The first is that before the introduction of green taxis, finding a cab in the boroughs outside Manhattan was challenging because yellow taxis prefer to stay in the more profitable Manhattan zone, so this limitation favors the distribution of taxis in the outer zones. The second reason is that the introduction of new licenses for green taxis was seen as a menace from the yellow cab business which strongly opposed to their introduction and obtained that green taxis would not be allowed to pick up passengers at the lucrative LaGuardia and JFK airport zones, with the exception of previously arranged trips by passengers with the driver.

The fare system is shared by both yellow and green cabs. The amount starts from a base of 2.50$ and increases as a function of time when moving slowly and as a function of distance when moving faster. Some extras may be added based on the hour of the day as well as various taxes. The passenger is also requested to pay for any incurred toll and is requested to tip the driver with a variable amount that ranges around 20% of the total. Fares between any zone of Manhattan and LaGuardia airport have a special flat rate of 52$ plus extras. Newark airport also has a special rate which adds a 17.50$ surcharge.

The analysis starts with a high level, exploration of the data. As shown in fig. 4 the number of total trips performed in the

Figure 2: The five boroughs of New York. Notice the position of the three airports: LaGuardia, JFK and Newark. Courtesy of https://nycmap360.com

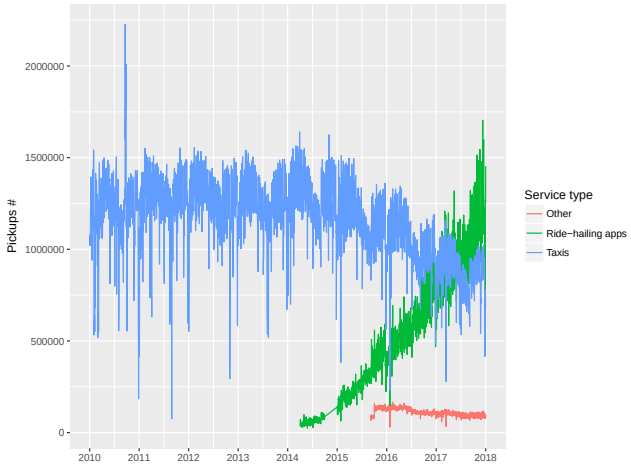


Figure 3: Historical taxi and ride hailing apps activity. Data from https://github.com/toddwschneider/nyc-taxi-data

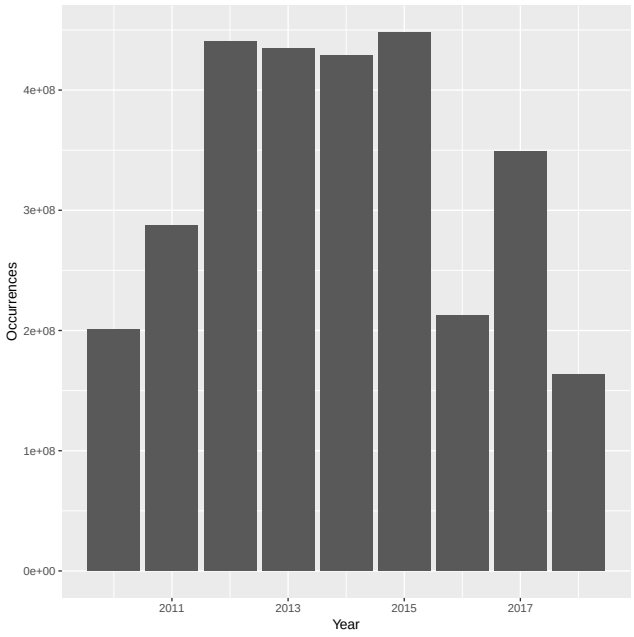


Figure 4: Total trips performed each year. Data for years 2010 and 2016 is missing due to data quality problems, while data for 2018 is only relative for the period January-June. Notice the decrease started in 2015.

years grows slightly until 2015 and decreases in the following years. Note immediately that 2010 and 2016 are anomalous years due to data quality problems that caused the removal of entire months of entries for those year. Looking at the profits, calculated based on the total amount, we can see an increasing trend in the years until 2015, after which there is a drop in the profits because of the reduced number of trips performed by the taxis. The growth of the profits, despite the only slight increase of the number of trips is given by a higher cost per trip in the years, as shown in fig. 4. The decline of the number of taxi trips started in 2015 can be explained by the growth in popularity of ride hailing services such as Uber, Lyft and Juno which, from 2015, are becoming widespread in the city as shown in fig. 3 and in 2017 surpassed taxis in popularity.

The study continues with the analysis of pickups and dropoffs. fig. 10 shows the number of pickups as a function of time. As already suggested by fig. 4 we notice a significant decrease in taxi service popularity starting from 2015. As depicted in fig. 5 we can see that the periods of major activity for pickups are 8-16 and 20-24. The dropoffs follow the same pattern. Interestingly, by looking at the variation of the average speed through the day, as shown in fig. 6, we note that there is an inverse relationship between number of pickups and average speed, suggesting that in the highlighted time frames the city is subject to higher traffic congestions. The most active pickup zones, as we can see from fig. 7, are Manhattan, the part of Brooklyn closer to Manhattan, LaGuardia and JFK airports. The dropoff zones closely follow the pattern of the pickup zones, but are slightly more evenly distributed across all zones. fig. 9 highlights the situation, showing that dropoff locations are more granularly distributed with re-

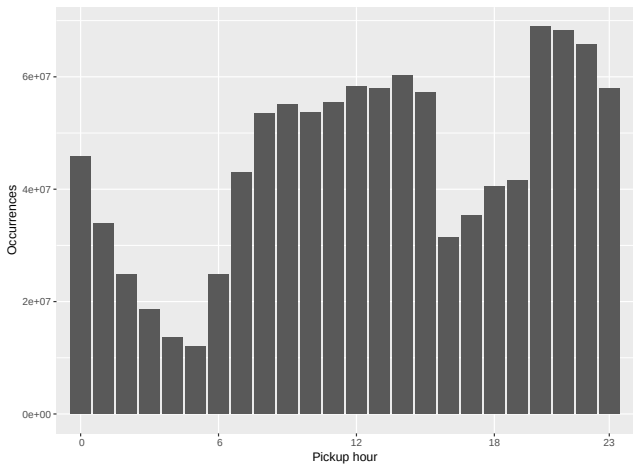


Figure 5: Number of pickups as a function of the pickup hour. Notice the major activity periods at 8-16 and 20-24.

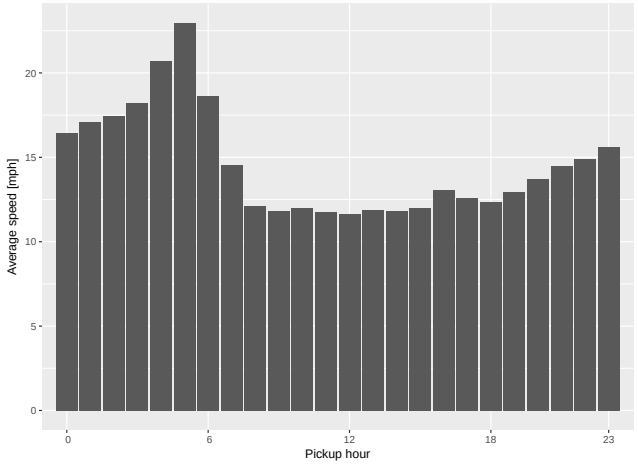


Figure 6: Average speed as a function of pickup hour. Notice how the average speed increases in the periods of minor activity.

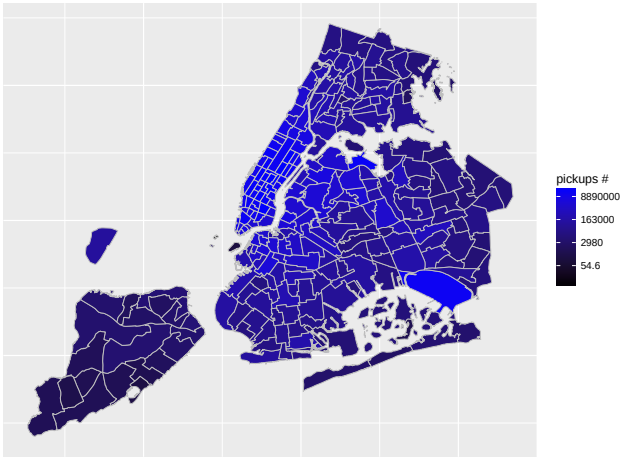


Figure 7: Map in logarithmic scale showing the number of pickups per zone. Dropoff location map is omitted because of close similarity.

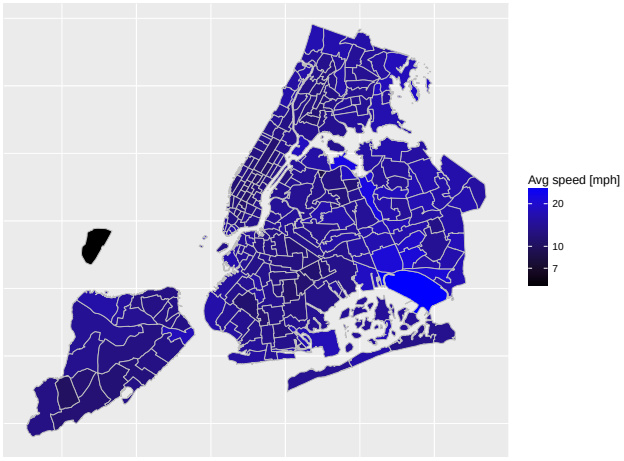


Figure 8: Map showing the average speed per pickup zone. Notice how outer boroughs typically feature higher average speeds, an indicator of the quantity of traffic.

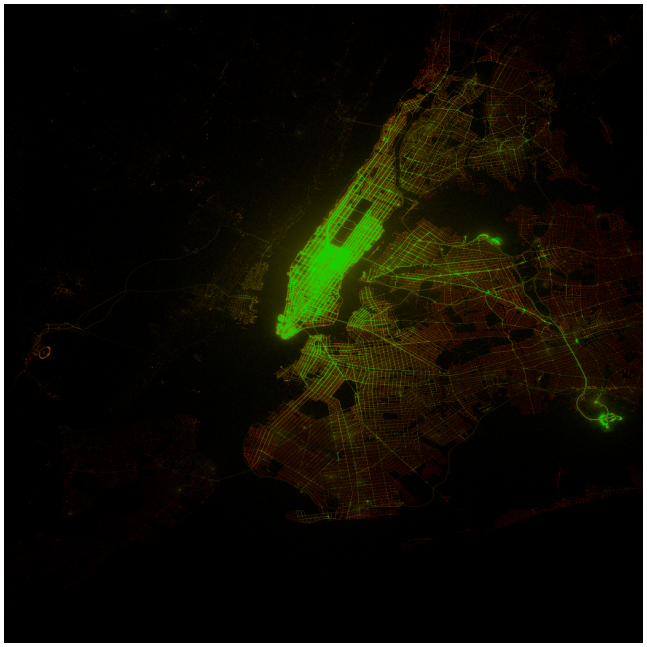


Figure 9: Map showing pickups points in green and dropoff points in red. Notice how dropoff points are granularly distributed across every street, while pickup points concentrate in Manhattan and on the main streets, meaning that people need to move to major roads in order to find taxis. Looking closely, it can also be noted that the most blurry areas, South of Central Park and Southest part of Manhattan, which correspond to high GPS error area, are the ones which indeed have the highest concentration of tall buildings. For the high resolution version see appendix A

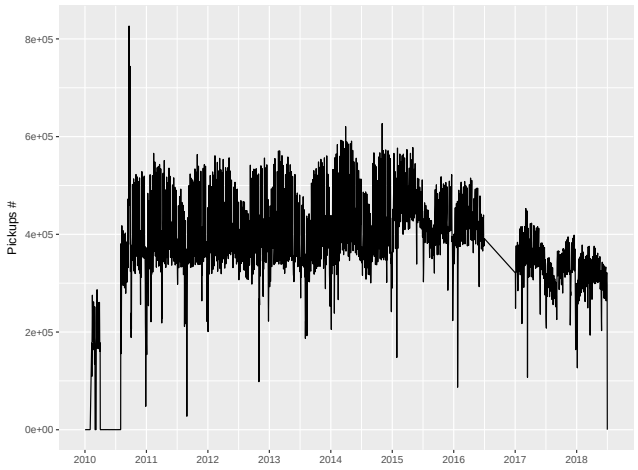


Figure 10: Pickups count by day. Notice the decline in taxi pickups started in 2015.

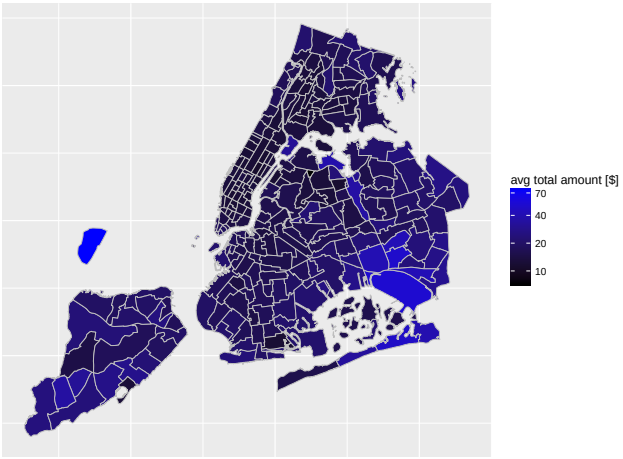


Figure 11: Map showing the average total amount by pickup location. Notice the inverse relationship with the distance from Manhattan.

spect to pickup locations. It must be noted however that the majority of the activity is concentrates in the most profitable Manhattan and airport sectors. We also note that, as illustrated in fig. 8, the average speed per pickup location is inversely proportional to the number of pickups in that location, with the slowest zones being Manhattan, South Bronx and North Brooklyn. We can assume this zones to be the most subject to traffic.

It is also interesting to notice how total amounts relate to the pickup location. As depicted in fig. 11 it can be noted that there is an inverse relationship between average total amount and zone distance from Manhattan. This is an expected outcome of the fact that Manhattan is one of the favourite dropoff locations, so on average the furthest zones from it are the most expensive. fig. 12 shows the average trip distance by pickup location. Here we can note the direct correlation with the average total amount of the preceding picture. The average total amount also varies during the day, as depicted in fig. 13. Two different peak periods are

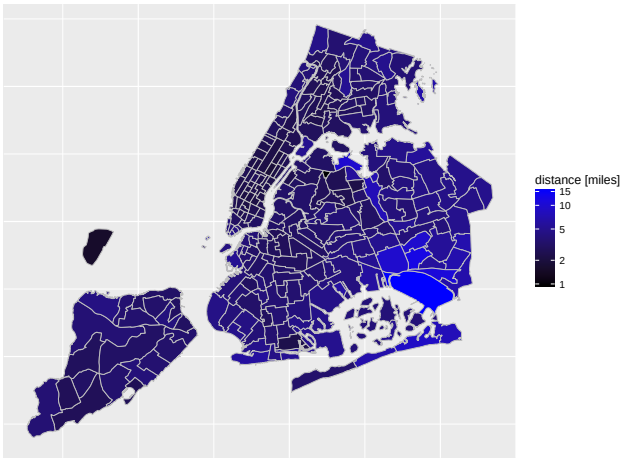


Figure 12: Map showing the number the average trip distance by pickup location. Notice the inverse relationship with the distance from Manhattan and the correlation with the average total amount map.

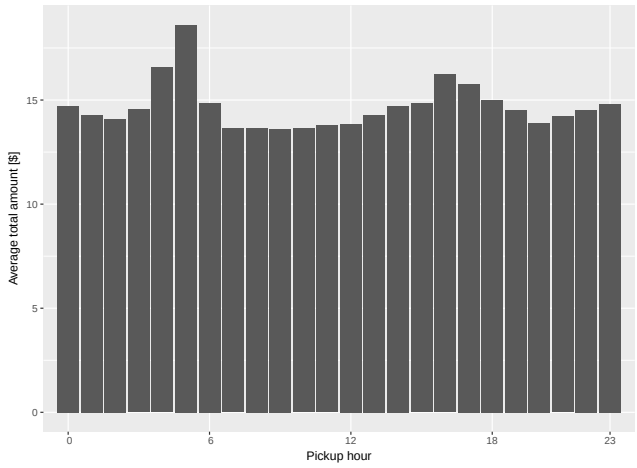


Figure 13: Average total amount divided by pickup hour. Notice the two peaks at 5 and 16.

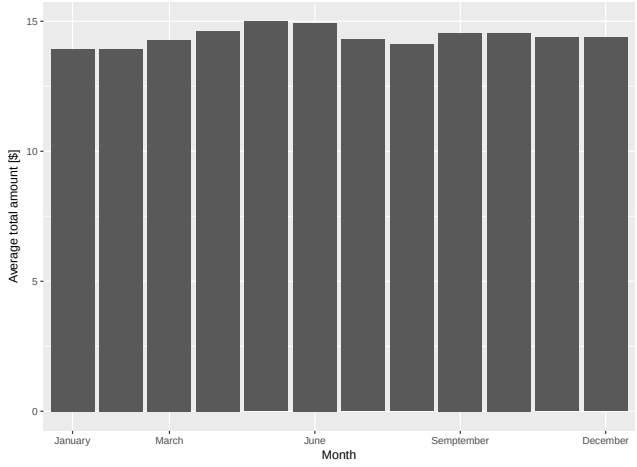


Figure 14: Average total amount divided by month. May and September-October result the most profitable periods.

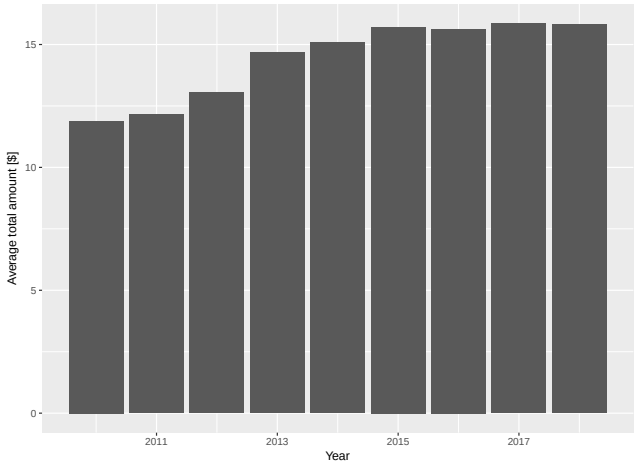


Figure 15: Average total amount by year. In 2012 the New York taxi commission increased the fare rates.

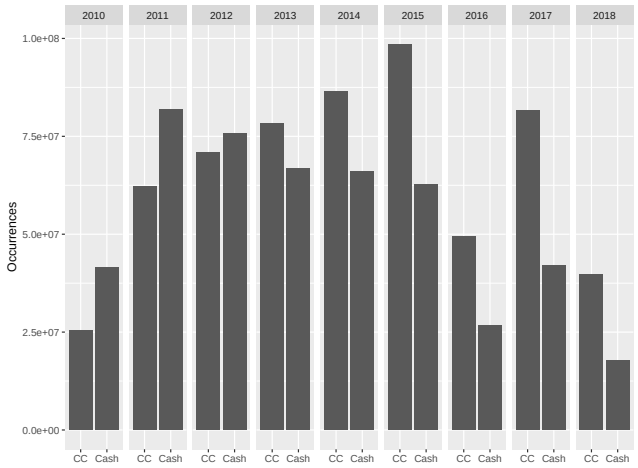


Figure 16: Distribution of payment types during the years. Notice the steady decline of cash payments in favor of credit card payments.

identified around 5 and 16. The discontinuity at 16 is caused by the introduction of a 1$ extra from 16 to 20 as a standard component of the fare amount, while the peak at 5 is caused by a sharp increase in the average trip distance from 2.92 to 4.61 miles due to an increased proportion of people going to the airport early in the morning with respect to the normal traffic. It can also be noted from fig. 14 that the average total amount varies during the months, with most profitable periods being May and September-October.

We also noted a general increase of the total amount during the years, as noted in fig. 15, which is a consequence of many factors. First, the average fare amount passed from 10.18$ of 2010 to 12.61$ of 2018 due to a general revision of fare rates operated by the New York taxi commission in 2012 which can be seen in fig. 18. Other factors include slight increases in the average tolls, extras and improvement surcharge during the years. It must be noted however that the increase is exacerbated by the fact that total amounts in our datasets include tips only for trips paid with credit

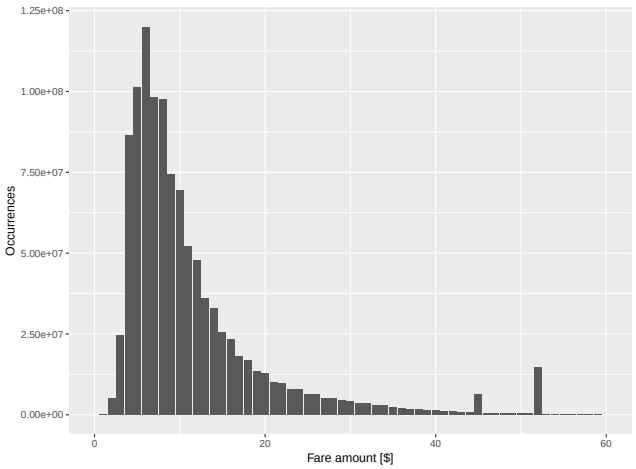


Figure 17: Distribution of fare amounts. Notice the peaks at 45$ and 52$ corresponding to the JFK airport flat rates before and after the 2012 fares revision operated by the New York taxi commission.

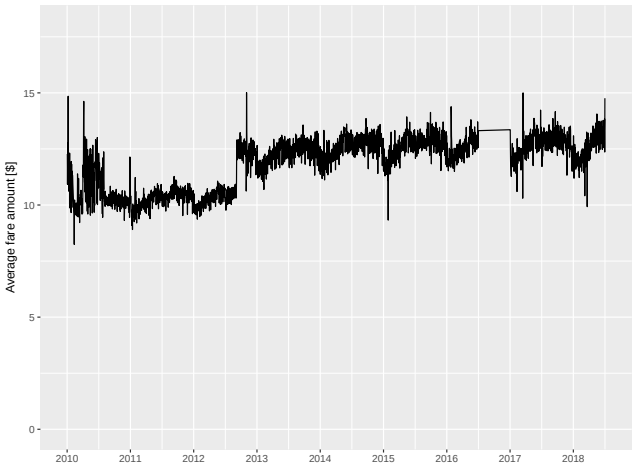


Figure 18: Average fare amount by day in $. Notice how in 2012 the new fare system caused a prominent increase in fare rates.

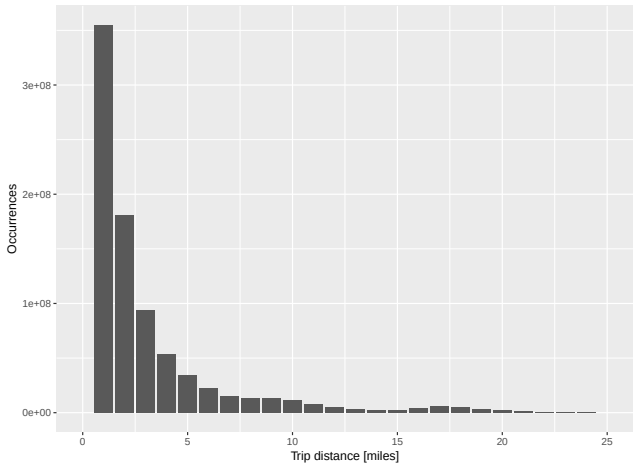


Figure 19: Distribution of trip distances. Notice the three modes at 1, 9 and 17 miles.

card and that during the years there has been a constant increase in the use of credit cards as shown in fig. 16. If our dataset included also cash tips, then we would still notice an increase in the total amount during the years, although more subtle. The overall distribution of fare amounts is shown in fig. 17. Note the sharp discontinuities at 52$ and 45$. These represents the special fare rate of trips linking Manhattan to the JFK airport, before and after a general revision of fare rates operated by the New York taxi commission during 2012.

A consistent part of the total amount is given by tips. A calculation of average tips percentages, calculated only on credit card paid trips, shows that the typical trip is tipped 18.6% and this rate remains stable in all the analyzed years. This value corresponds to the guidelines of many online sources that suggest tipping taxi drivers from 15% to 20% of the total amount.

Other interesting findings regard the distribution of trip distances which is shown in fig. 19. It can be observed that the distribution is trimodal, with the first mode at 1 mile, which represents short trips performed mainly inside Manhattan, the second at 9 miles, which represents trips between Manhattan and LaGuardia airport, and the third at 17 miles, which represents trips between Manhattan and JFK airport.

Looking at the passenger count distribution in fig. 20 we can note that the majority of trips is performed with 1 or 2 passengers, while there is also a component of group rides which causes a peak at 5 passengers.

By looking at the distribution of payment types during the years in fig. 16 it can be noted a strong trend towards paying with credit card versus cash, which was prevalent until 2012.

As a last general remark, we note that the distribution of tolls amount, as shown in fig. 21, is bimodal, with the first mode at 0$ and a smaller mode at 5$, representing the fact that the majority of trips are not subject to tolls, while the ones that are subject to them pay on average 5$ of tolls. These correspond to tolls for Queens-Midtown Tun-

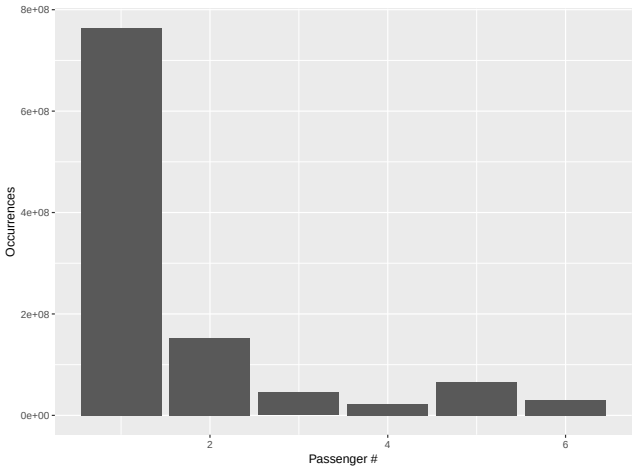


Figure 20: Distribution of the number of passengers.

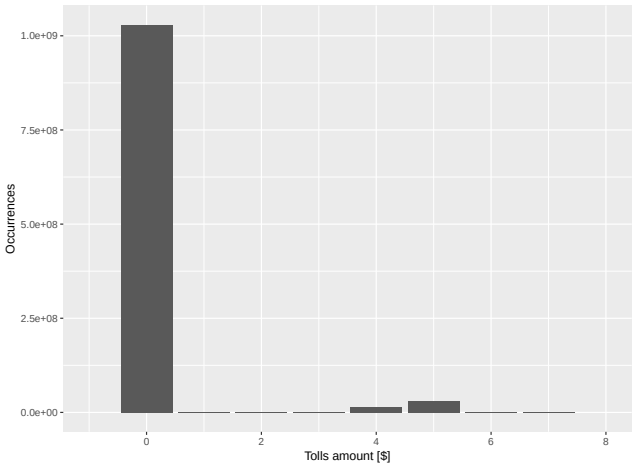


Figure 21: Distribution of tolls amount. Notice the modes at 0$ and 5$.

nel, Brooklyn-Battery Tunnel and the Triboro Bridge, which are convenient ways to avoid traffic for reaching the various parts of Manhattan from Brooklyn and Queens.

6. TRAFFIC SEGMENTATION

After the data cleaning and preliminary analysis phase, we decide to perform a characterization of the different types of trips through k-means clustering. In order to perform an effective clustering, we introduce new features to the dataset and manipulate existing ones in order to guide the model towards finding interesting trends in the data. The following features are given to the clustering algorithm:

- taxi company, boolean field representing whether the trip was served by yellow or green taxis
- pickup hour and dropoff hour, scaled integer fields representing the hour of pickup and dropoff
- weekend, boolean field set to true if the trip happened during the weekend
- passenger count, scaled numerical field representing the number of passengers
- speed, scaled numerical field representing average speed
- distance, scaled numerical field representing metered distance
- ratecode id, one hot encoded field representing the used ratecode
- fare amount, scaled numerical field representing metered fare
- tolls amount, scaled numerical field representing the incurred tolls
- payment type, one hot encoded field representing the payment method

Note that, as our objective is to find a global characterization of trip types which possible span the entire city, we do not include pickup and dropoff zone ids as clustering features because this would cause the algorithm to divide trips based on their starting or ending point which does not carry useful information. Notice also that we scaled every numerical field by its mean and variance. Without doing so, features with a wide range such as the fare amount would be the only ones relevant for the algorithm, which is not the result we want to obtain.

The appropriate number of clusters for the k-means clustering algorithm is detected using the Elbow method, which results are shown in fig. 22. For our dataset, the procedure suggested a number of cluster equal to 5.

We then proceed to identify the characteristics of each of these 5 identified clusters. By looking at fig. 23, we notice that cluster 1 corresponds to early morning trips done from 0 until 7, cluster 4 corresponds to day trip from 8 to 15 and cluster 2 corresponds to trip from 16 to 24. Furthermore, from fig. 24 it can be seen that cluster 3 comprehends trips

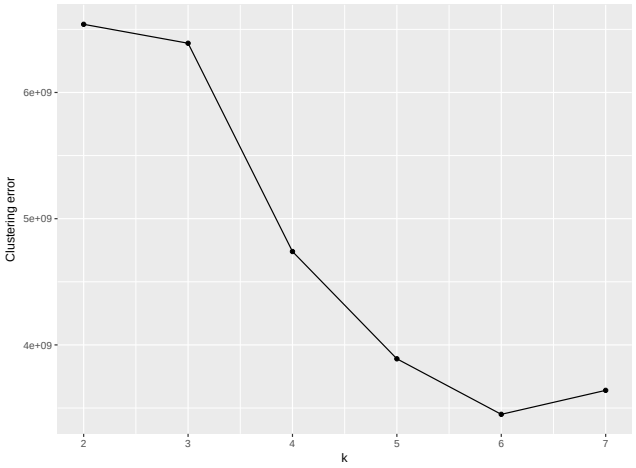


Figure 22: Representation of clustering error as a function of the number of clusters k. For performance reasons the error is estimated on a subportion of the dataset, so some points have an anomalous behavior. For k=5 we note that the error curve starts flattening, so 5 is chosen as the optimal k value.

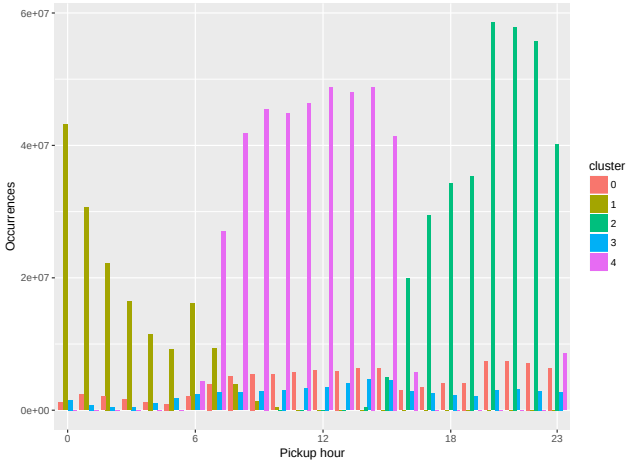


Figure 23: Number of pickups by day hour and cluster. Notice how clusters 1, 4 and 2 are determined based on the pickup hour.

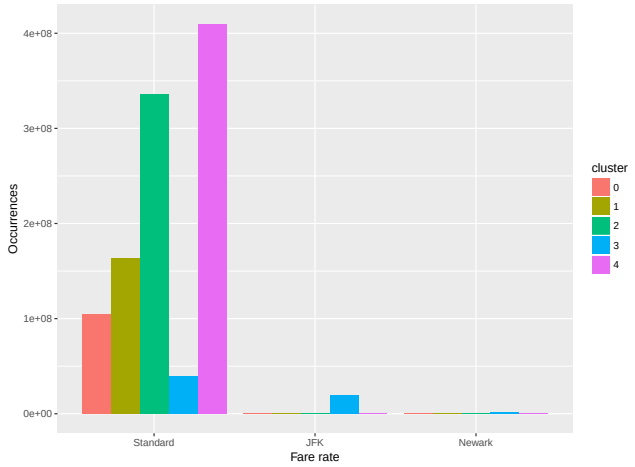


Figure 24: Distribution ratecodes divided by cluster. Notice how the majority of JFK trips are grouped in cluster 3.

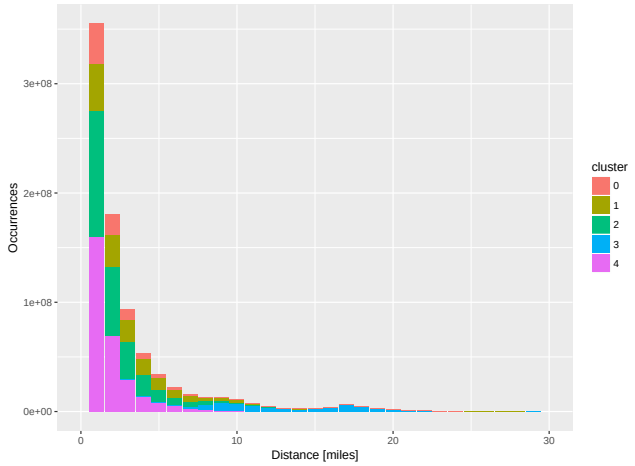


Figure 25: Distribution of trip distance divided by cluster. Notice the modes for cluster 3 at 9 and 17 miles which corresponds respectively to trips to LaGuardia and JFK airport.

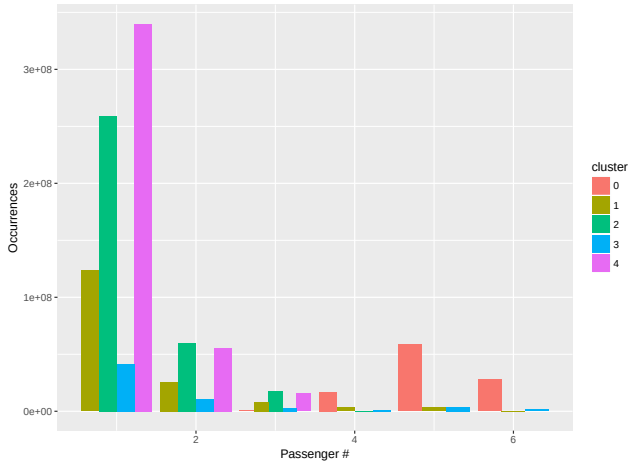


Figure 26: Distribution of number of passengers per cluster.

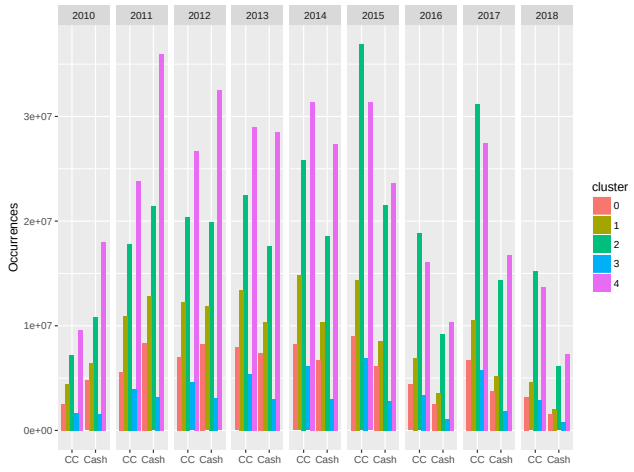


Figure 27: Distribution of payment types in time per cluster.

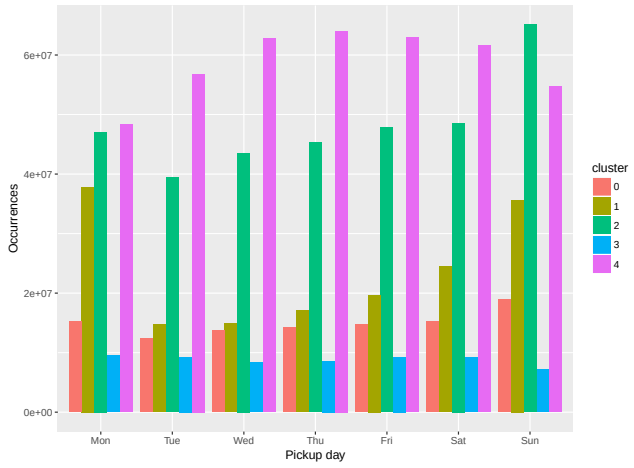


Figure 28: Distribution across the week of the number of pickups per cluster.

to JFK and Newark airports. Moreover, by looking at fig. 25 and by plotting the pickup locations on a map as a function of the cluster, we can also say that cluster 3 comprehends trips to LaGuardia airport. Lastly, by looking at fig. 26 we can see that cluster 0 corresponds to trips performed by groups of 4 or more people. We note also that cluster 1 and 2, that represent respectively early morning and night trips, could be unified in a single cluster with common characteristics. The reason why our clustering algorithm separated them is that the pickup and dropoff hour property cause trips in the evening to have a high distance measure from trips in the morning, while a more reasonable distance metric would use a circular similarity measure and assign to trips in the evening and trip in the morning a smaller distance.

By looking further into the data, we can notice from fig. 25 that day trips in cluster 2 tend to have a light tail, compared to heavy tails of clusters for morning and night trips and for group rides. This difference is probably explained by the fact that people tend to use taxis also for long and expensive trips at nights where the availability of other means of transport is reduced. From fig. 27 it can be noted that group rides and day trips tend to have a lower CC to Cash payment ratios, while airport trips, morning trips and night trips tent to be paid with credit card more frequently. It is also interesting to notice how these type of traffics vary through the week as shown in fig. 28. Day trips tend to have their peak on Thursday and have lower activity along the weekend. On the other hand, group rides, morning trips and evening trips tend to peak in the weekend, probably because of the increased nightlife. Airport trips, on the other hand, have their peak on Monday and register their lowest point on Sunday.

fig. 29 shows the graph constructed to better visualize the different types of traffic inside the city of New York. The number of edges, starting from a node, is variable and depends on the amount of pickups for that zone, this provides an immediate view of the most heavily trafficked areas of New York. Each node has a set of attributes describing: the level of activity, the average distance, duration, speed and total amount associated with the trips departing from that zone. Each of these attributes has a value between LOW, MEDIUM and HIGH calculated by finding the first and third quartiles of the corresponding distribution. The value of a given attribute for a node is considered LOW if it lies below the 25% of the data in the dataset, MEDIUM if it lies between the 25% and the 75%, and HIGH for values above 75%. An important aspect to note is that, in order to obtain a more comprehensible graph, the edges from a node to itself are removed. From this graph it can observed that the majority of the trips tend to remain inside the same borough, also it can be seen that Manhattan is the borough with the highest traffic density, especially during the working hours. Inside this area some other interesting aspects can be noted. First, the zone of East Village is subject to a high amount of traffic especially in the evening, in fact this area is renowned for the nightlife. Second, the Roosevelt Island is not directly connected to any other zone in the same borough and this is confirmed by looking at its geographic location: it can be seen that, even though it is inside of the Manhattan borough, there is no direct path to Manhattan itself, and the only way to leave the island is

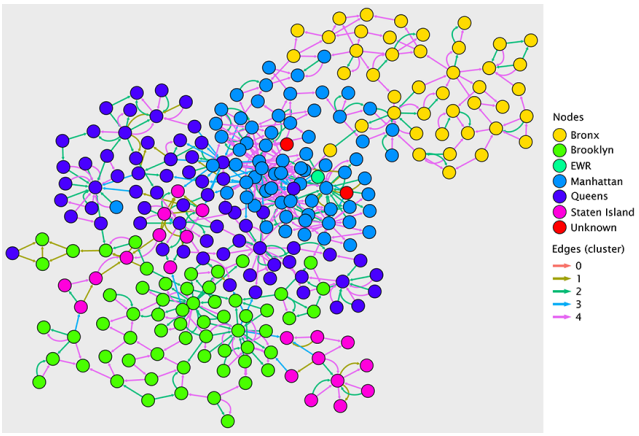


Figure 29: Directed graph showing the different types of traffic found using k-means clustering. Each node represents one of the 265 zones of New York, while each edge, between two zones, represents a traffic flow in the direction of the arrow. The colour of the nodes depends on the borough of the zone they represent. The colour of the edges depends on the type of traffic. For the interactive version of the graph and its legend see appendix A.

through the Roosevelt Island Bridge which leads to Astoria in the Queens. While looking at the Bronx borough, a less interconnected graph, representing a lower volume of activity, can be observed. It can be noticed that night trips are less frequent, this may be due to the level of criminality that discourages people from moving during these hours. Finally, the graph highlights that the clustering correctly identifies long distance trips, the majority of which is traffic towards the airports. For the interactive version of the graph and its legend, please refer to appendix A.

7. TRAFFIC FLOW ANALYSIS

One of the objectives of this report is to better understand how people move inside the city of New York. We partially describe this activity in section 5 by showing the high activity levels of Manhattan and of the airport zones. In order to better understand the flows of traffic, we decide to plot, for each pickup borough, the distribution of dropoff locations as a function of time.

fig. 30 shows the distribution of dropoff locations for pickups happened in the Bronx borough. The majority of dropoffs as expected happen inside the Bronx borough itself, North Manhattan or the two airports.

fig. 31a highlights the distribution of dropoff locations for pickups happened in the Brooklyn borough. Also in this case, the majority of dropoffs happen inside the borough itself, the two airports or the adjacent Manhattan zones. An interesting consideration can be done by looking at fig. 31b which shows the favourite dropoff locations for rides beginning at 9. From 7 to 9 we can see a high rate of dropoffs in the Tribeca district of Manhattan, which is the ending point of the Brooklyn Bridge. This flow vanishes during the day and is probably caused by workers moving into the city early in the morning.

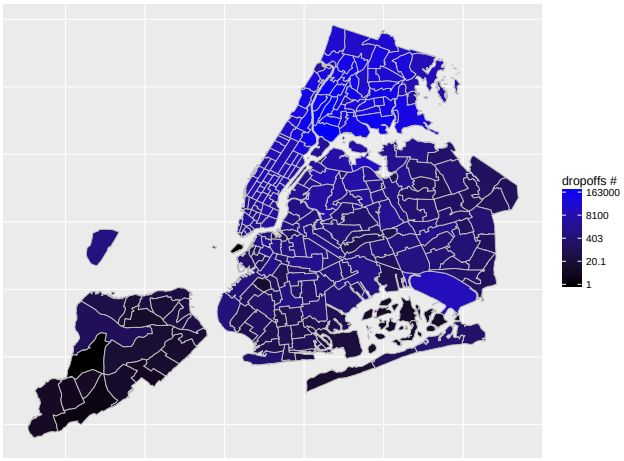
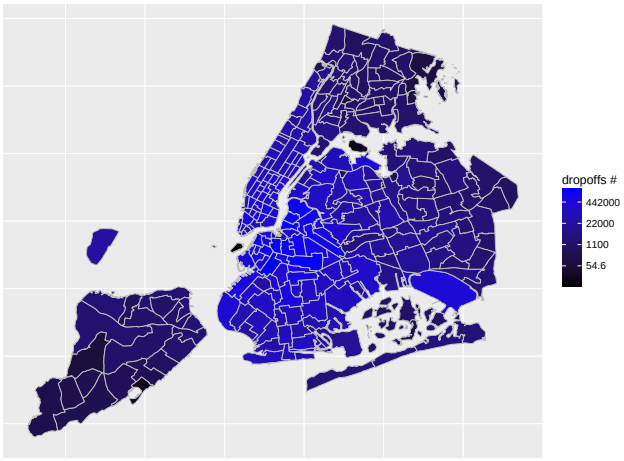
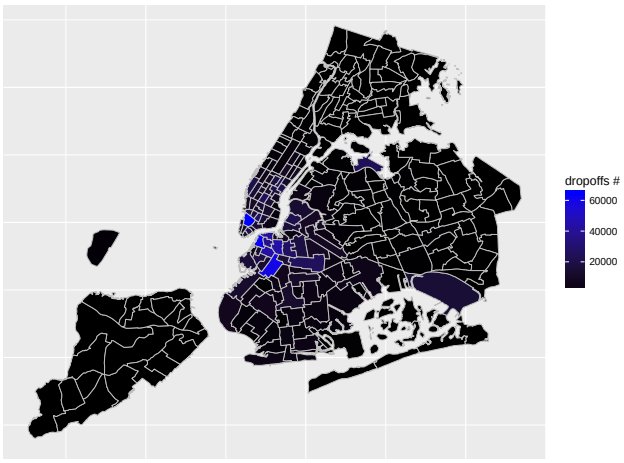


Figure 30: Map in logarithmic scale showing the preferred dropoff locations for rides starting inside the Bronx borough.

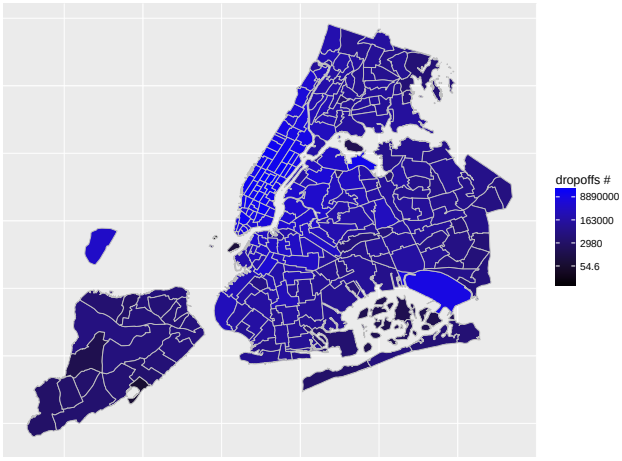


(a) Overall dropoff locations (log scale)

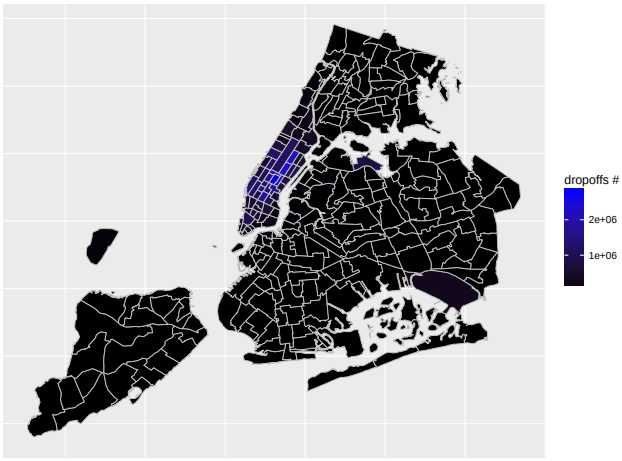


(b) Dropoff locations at 9

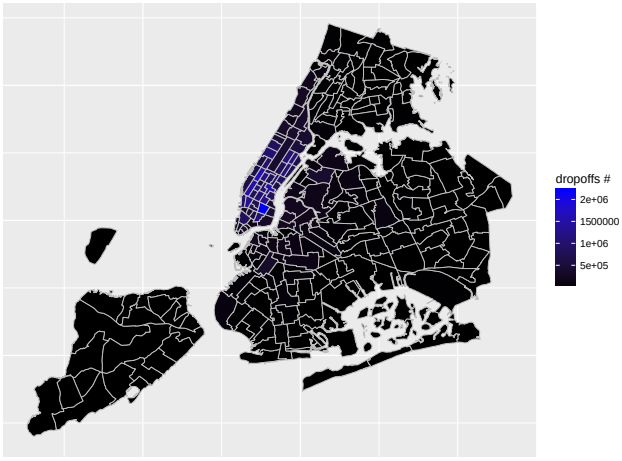
Figure 31: Map showing the preferred dropoff locations for rides starting inside the Brooklyn borough, overall and at 9. Notice the high concentration of dropoffs in the Tribeca Manhattan district at 9.



(a) Overall dropoff locations (log scale)

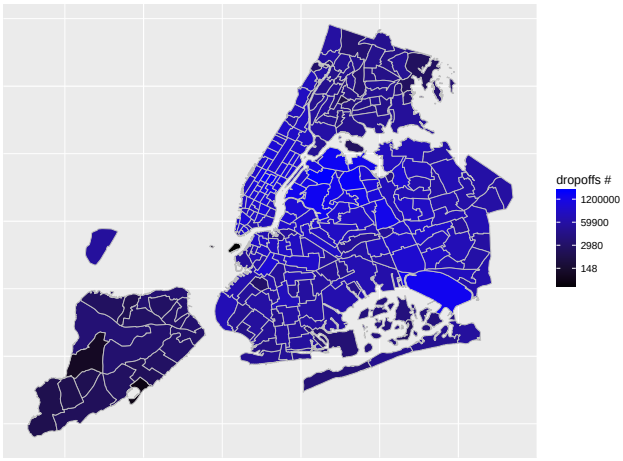


(b) Dropoff locations at 10

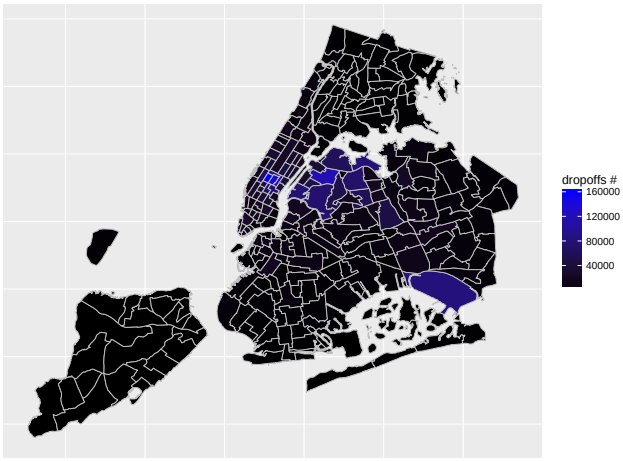


(c) Dropoff locations at 22

Figure 32: Map in showing the preferred dropoff locations for rides starting inside the Manhattan borough, overall, at 10 and at 23.



(a) Overall dropoff locations (log scale)



(b) Dropoff locations at 12

Figure 33: Map showing the preferred dropoff locations for rides starting inside the Queens borough, overall and at 12. Note how at 12 a consistent part of the traffic goes to the Manhattan district under Central Park. This movement remains constant all day from 7 to 16.

fig. 30 depicts the distribution of dropoff locations for pickups happened in the Manhattan borough. Most dropoffs happen within Manhattan itself or in the airports. During the day the most active dropoff zones are those on the South East corner of Central Park, while after 20, most of the dropoffs concentrate in the South West Manhattan area and in East Village. From this fact and from some research on the New York nightlife we infer that this shift is caused by people moving to these zones to experiment the New York nightlife. The fact is cross checked by looking at the locations where most pickups happen at night, which are again South West Manhattan and East Village.

fig. 33a shows the distribution of dropoff locations for pickups happened in the Queens borough. The majority of dropoffs happen in North West Queens, the adjacent part of Manhattan and at the airports. We notice that from 7 to 16 there is a steady movement of people from Queens to

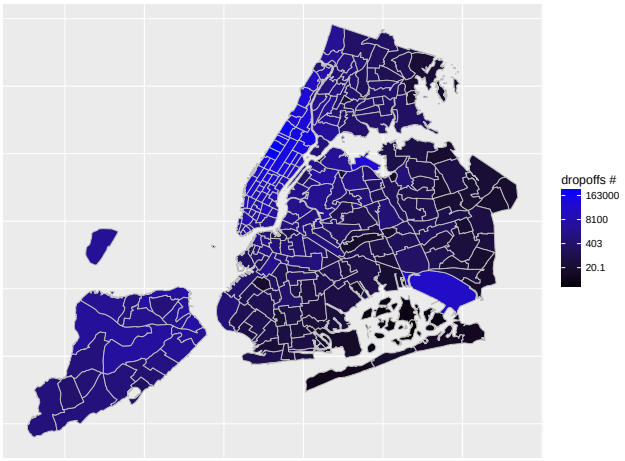


Figure 34: Map in logarithmic scale showing the preferred dropoff locations for rides starting inside the Staten Island borough.

the Manhattan districts South of Central Park, similar to that shown in fig. 33b. By isolating the airport traffic it can be seen that a part of this phenomenon is caused by trips started at airports that seem to prefer the zones below Central Park as destination.

fig. 34 shows the distribution of dropoff locations for pickups happened in the Staten Island borough. This district features an anomalous behavior in that the majority of dropoffs happen in the zones West of Central Park and the airports instead of inside Staten Island itself. This behavior remains constant throughout the whole day.

8. YELLOW VS GREEN CABS

A major change in the taxi business was given by the addition of green taxis along with the traditional yellow cabs. In this section we analyze the impact of this introduction and the difference from the yellow taxis.

fig. 35 shows the distribution of trips between the companies in the years from 2010 to 2018. We can notice that the introduction of green taxis caused a slight reduction of the trips performed by yellow taxis and from 2015 both types of taxi suffered from a substantial decrease due to the competition of companies such as Uber.

Looking at fig. 36, which compares the distribution of fare amounts, we immediately notice that the curves have similar shape, although the longest and profitable airport trips, the ones with 52$ fare rate, are almost always performed by yellow cabs. This is expected, since green cabs can only perform dropoffs at the airport and pick up passengers only if the trip is arranged in advance, which is a severe limitation.

By analyzing fig. 37 we also note that, while the distribution of yellow taxis is trimodal, the distribution for green taxis is not. In particular we notice that green taxis miss the 9 miles and 17 miles mode, which are the trips to the JFK and LaGuardia airports. Trips at LaGuardia airport do not happen often for the same limitations of green taxis

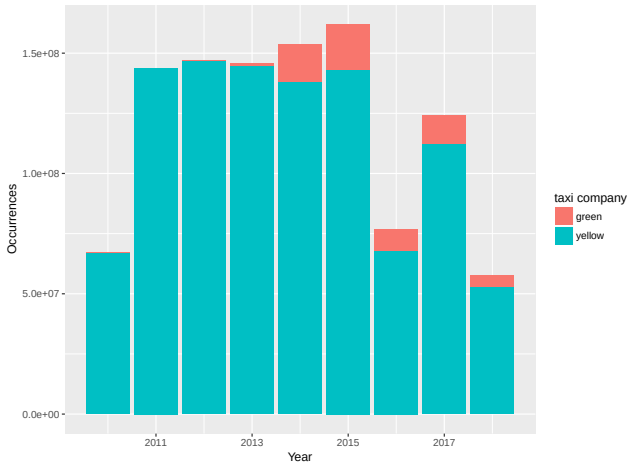


Figure 35: Total trips performed each year divided by taxi type. Data for years 2010 and 2016 is missing due to data quality problems, while data for 2018 is only relative for the period January-June.

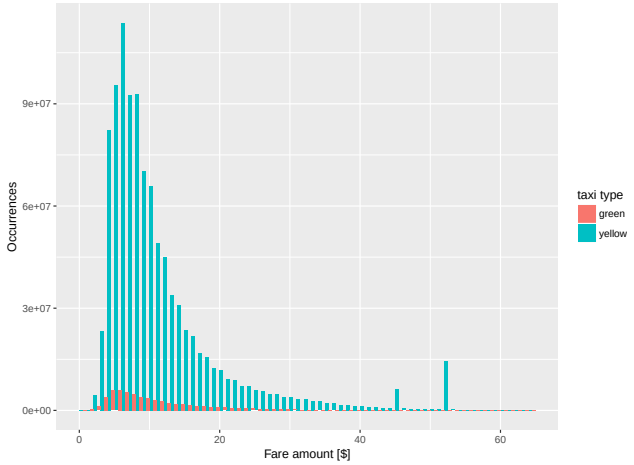


Figure 36: Fare amount distribution divided by company. 45$ and 52$ fares represent special fare rates to the airport.

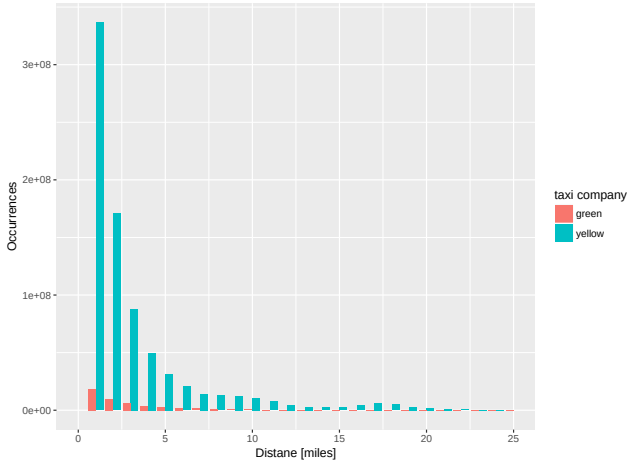


Figure 37: Trip distance distribution divided by company.

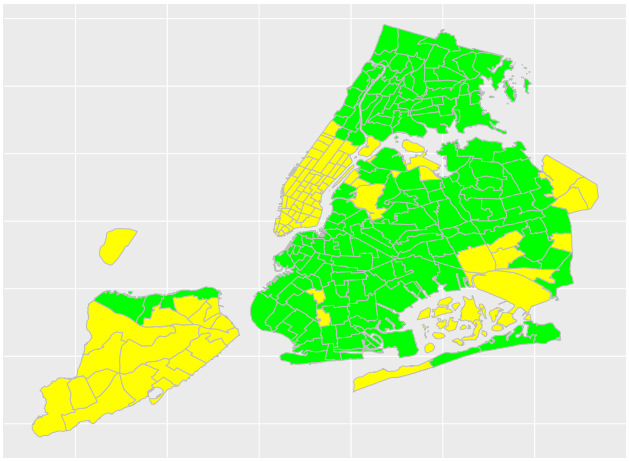


Figure 38: Map showing which taxi type preveals in the various zones since year 2014. Yellow zones show where yellow taxis perform the most pickups, green zones represent the ones where green taxis perform more pickups. Notice how Manhattan and the airport zones are prevalently served by yellow taxis.

explained in the former paragraph.

Green taxis were introduced explicitly to help serving the outer boroughs, where it is difficult to find a yellow cab, because they prefer to remain in the profitable Manhattan zone. fig. 38 shows which taxi type is more popular in the various zones since year 2014. Indeed we can see that green taxis became prevalent in North Manhattan, where they are allowed to pick up passengers, and in every outer borough with the exception of Staten Island and the two airport zones where they are not allowed to perform non arranged pickups. fig. 39 shows an alternative view of the yellow-green taxi presence in New York.

9. AIRPORT TRAFFIC

Given the relevance of the airport traffic emerged by the previous analysis step, we decide to perform a more thorough analysis of the traffic originated or directed to the three airports. The first remark is that, due to Newark airport not being in a valid taxi zone, the data cleaning phase removed many of the trips to the Newark airport and so data regarding it is not reliable.

We first notice from fig. 40 that airport trips make a relatively small 5.92% of the total number of taxi trips. Despite that, as fig. 41 shows, they are a major stream of revenue for taxis due to the very high cost per trip, making 18.7% of their profits. In fact, by looking at the fare rate distribution in fig. 42, and comparing it with fig. 17 we can see that fare amounts for LaGuardia averages around 30$, fare amounts towards JFK have a fixed fare rate of 52$ (45$ before fare rate revision operated in 2012), while the in majority of common taxi trips the fare amounts are below 10$. It it thus not surprising that yellow taxi owners fought for forbidding green taxis to pick up passengers at airports.

By looking at the distribution of payment types in fig. 43

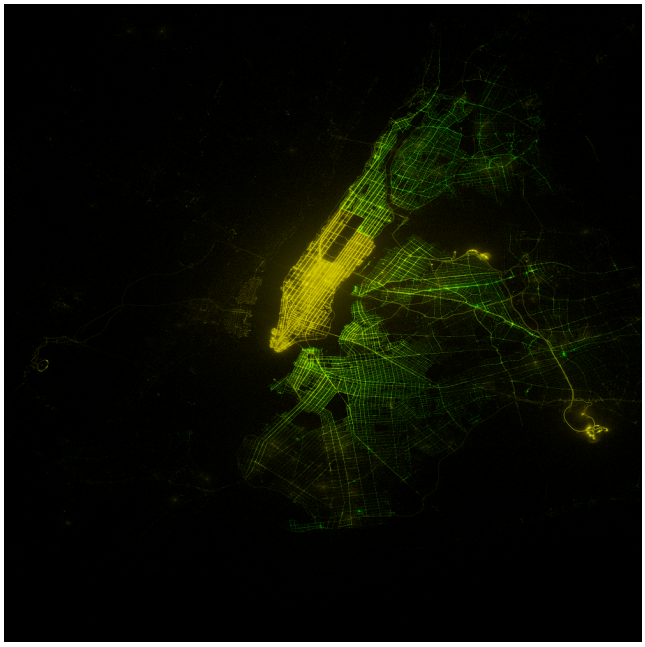


Figure 39: Pickup points of yellow and green taxis in their respective color. Notice how Manhattan and the airports feature yellow prevalence, while the other zones feature a green prevalence. For the high resolution version see appendix A

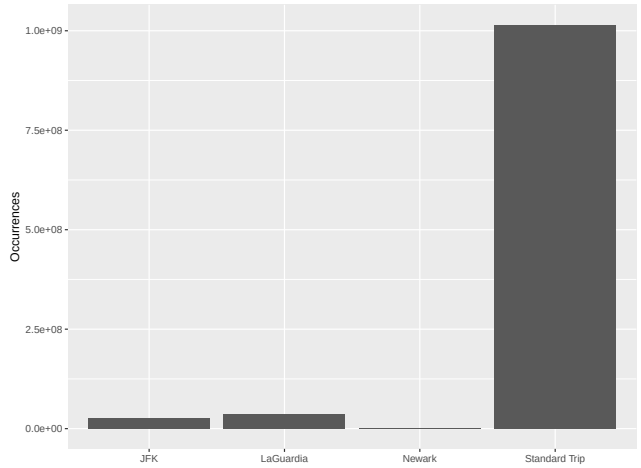


Figure 40: Total number of trips performed from January 2010 to June 2018 divided by airport. Airport trips make 5.92% of total taxi trips.

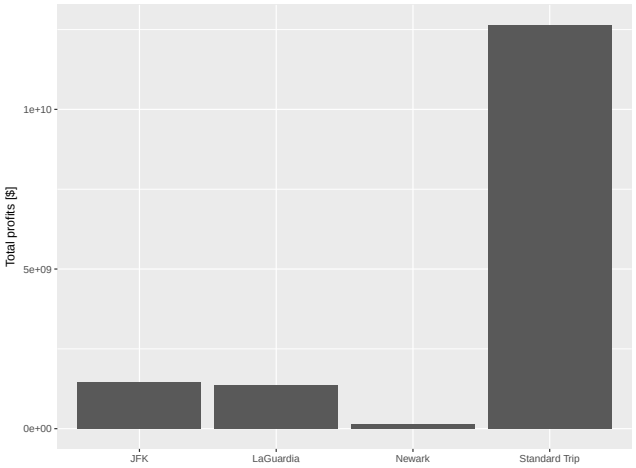


Figure 41: Total revenue from January 2010 to June 2018 divided by airport. Airport trips make 18.7% of total taxi revenue.

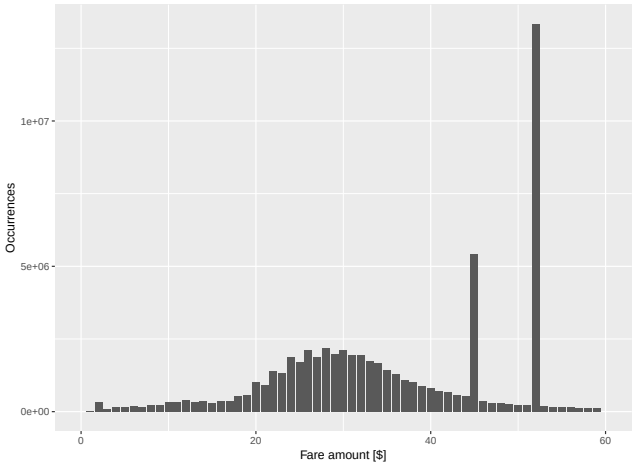


Figure 42: Distribution of fare amount. Notice the mode at 30$ corresponding to trips to LaGuardia airport and the 45$ and 52$ flat rate amounts corresponding to trips to JFK from Manhattan. The right heavy tail is given by trips to the JFK from outside Manhattan that are not subject to the flat rate.

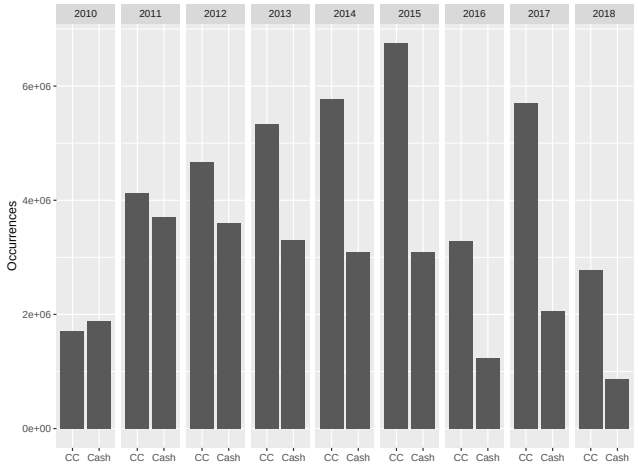


Figure 43: Distribution of payment methods from 2010 to 2018. Notice how already from 2011 credit card payments make for the majority of payments. The higher percentage of credit card payments is probably given by the high cost of airport trips.

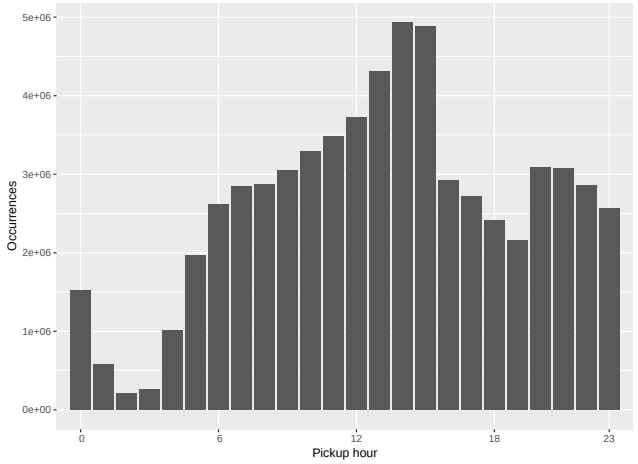


Figure 44: Number of pickups divided by hour. The peak activity is reached at 14 and is almost absent from 24 to 5.

it is also interesting to notice that the percentage of people paying with credit cards is higher than its percentage in common taxi trips shown in fig. 16. This is probably due to the higher total amounts.

From fig. 44 we can look at the number of trips as a function of pickup hour. This can be seen as a rough indicator of airport activities through the day. We notice that activity starts at about 5, then peaks at 14 and slowly decreases until 24, after which activity is enormously reduced. By comparing it with standard taxi activity shown in fig. 5 we can see that airport traffic completely lacks the peak experimented at 20 and reinforces the hypothesis that the peak at 20 in normal traffic is caused by nightlife.

As a last remark, by looking at the distribution of tolls amount in fig. 45, we notice that the number of trips subject to tolls is the majority with respect to the total number of

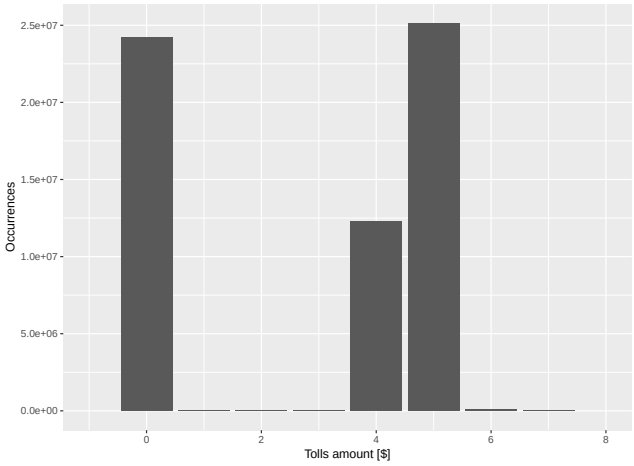


Figure 45: Distribution of tolls amount. Notice how the number of trips subject to tolls are the majority.

trips. By comparing it with fig. 21 we see that airport trips are responsible for the majority of trips that were subject to tolls.

10. CONCLUSIONS

In this paper we presented an overview of many aspects of the NYC taxi service, starting from an high level description and later proceeding more in depth in some selected topics. We highlighted some of the service's problems such as the concentration of yellow taxis in Manhattan and the airports which creates a difficulty for passengers of outer boroughs to find hails despite the introduction of green taxis, the increasing trend of fare amounts and the problematic level of city traffic which causes very low average moving speeds. It is our opinion that these and other factors such as the historically inadequate number of yellow taxi Medallions and problems in the public transport infrastructure play an important role in the success of ride hailing applications such as Uber which are experiencing huge gains in popularity, threatening the position of the taxi service.

Further topics of interest which have not been analyzed at the moment include a study of weekend traffic which seems to have peculiar characteristics, study of the effects of weather events on the taxi service, studies regarding the popularity of local businesses such as bars, clubs and pubs in the different zones of the city or a series of studies regarding the movement of employees of certain companies such as Accenture, the Goldman Sachs Group or American Express which could give a picture of where the wealthiest part of the population lives in the city.

11. REFERENCES

APPENDIX

A. FILES AND SCRIPT EXECUTION

This appendix describes the contents of the shared Google Drive directory and describes what is needed in order to run the scripts. The directory is composed of the following parts

- `src` folder, containing the Python and R sources
- `report` folder, containing the files used to produce the report
- `data` folder, containing the dataset and some helper files for faster dataset construction
- `plots` folder, containing all the obtained plots and the high resolution version of images shown in the report
- `spark_results` folder, containing all the `.csv` results produced by Spark

The src folder is composed of the following parts:

- Spark Python files in the root directory
- R scripts for the production of plots and additional data analysis in the R folder
- Docs and additional material in the docs directory

The dataset folder contains the following elements:

- `tar_gz_dataset` folder, containing the compressed dataset from 2010 to 2018 for yellow and green taxis
- `clustering_model.model` precomputed clustering model
- `clustered_dataset.parquet` precomputed dataset, ready for the analysis
- `daily_trips_by_geography.csv` public archive from toddwschneider containing aggregate information about for hire vehicle operations in New York City

In order to obtain from scratch the results presented in the report, the following steps must be carried out in this order, but you may use intermediate results to speed up the process if you wish to, as described later in this section:

- Install Python packages `matplotlib`, `numpy`, `pandas`, `networkx`, `fiona`, `shapely`.
- Import the compressed .tar.gz dataset in your cluster infrastructure and convert it in parquet format executing the gz_to_parquet_main.py script. The script needs to be configured as described in its header.
- Convert the different dataset years into a dataset with common schema using the `common_schema_conversion_main.py` script. The script needs to be configured as described in its header. In particular you need to ensure that the shapefile in the docs folder is available to the main script. Having the `lookup_matrix_1000_3.npy` precomputed lookup matrix in the same folder as the python script avoids its recalculation.
- Perform data cleaning by executing `data_cleaning_main.py` in order to produce the cleaned dataset. Follow the instruction in the script header to configure it before execution.

- Perform clustering. The full clustering procedure is the longest and most expensive part of the processing. For this reason we provide the precomputed `clustering_model.model` precomputed clustering model along with the dataset. If you insert it in the dataset folder you can comment the lines in the code performing clustering and model saving and direct apply the clustering model to your dataset, producing the clustered dataset. You may also skip the passages until this point by obtaining the `clustered_dataset.parquet` dataset from the Google Drive folder. Please refer to the file header to correctly configure your script. If you wish to execute the complete clustering procedure, you may execute the script directly after the required parameters are inserted without commenting out portions of the code.
- Obtain the main data statistics by executing `data_statistics_main.py`. Please refer to the header file for correct script configuration. The results in form of .csv files can be subsequently used by the R scripts.
- Obtain the clustered data statistics by executing the same script, but setting `clustered_analysis` to True. Remember to save the results in a different folder in order to be able to import them in R more easily in later steps.
- Obtain the airport statistics by executing `airport_data_statistics_main.py`. Please refer to the header file for correct script configuration.
- (Optional) Execute `data_imaging_main.py` in order to obtain plots such as the one of fig. 39.
- (Optional) Execute `graph_building_main.py` in order to obtain an interactive graph visualization of the taxi zones.
- Install R packages `ggplot2`, `RColorBrewer`,`rgdal`, `ggmap`
- Configure and execute `R/main.R` in order to obtain the first set of plots. In particular the script needs to be given the location of the .csv files obtained during the main analysis step.
- Configure and execute `R/airport_main.R` in order to obtain the first set of plots. In particular the script needs to be given the location of the .csv files obtained during the airport analysis step.
- Configure and execute `R/clustering_main.R` in order to obtain the first set of plots. In particular the script needs to be given the location of the .csv files obtained during the clustered analysis step.

It may be possible that some other minor steps or slight modification to the code (such as specifying file access protocols `hdfs:\\` instead of `file:\\`) are required in order to completely run the code on the system of the reader.

If you are interested only in seeing the result `.csv` files or in running the R scripts to produce the plots you can avoid running Spark entirely and use the `stats.zip`, `clustered_stats.zip` and `airport_stats.zip` archives available in the Google Drive folder to obtain the results of Spark elaborations. If you are interested only in seeing the plots you can download the corresponding zip archives from the Google Drive folder.

If you are interested in visualizing the interactive graph, download the `traffic_graph.zip` contained in the `plots` directory of the Google Drive folder, extract the content of the zip in an arbitrary directory, open the `traffic_graph` directory, open the `web_session` directory and open the `index.html` file. A web page displaying the interactive version of the graph will be opened. If you are interested in visualizing the graph legend, open the file `legend.png` in the same folder.