

# Datathon Project Group 3 DS4A

Johnathan Salamanca, Mario Cerón,  
Carol Martinez, Javier Cocunubo, Jairo Nino, Alvaro Munoz

December 9, 2019

## 1 Introduction

This document describes the process followed to define, analyze and answer the questions of the Datathon project. The data provided correspond to datasets from 2014 and 2015 of four different transportation means: Yellow cabs, Green cabs, UBER, and MTA. Additionally, one dataset provides weather information, and another one provides demographic information of the boroughs.

### 1.1 Background Information

New York city is the most populous city in the United States [5] with around 8.6 million inhabitants. It is expected to reach 9 million by 2040 [3], with Bronx being the borough with the highest increase in population with 14% (between 2010 and 2040).

Average Travel speeds in New York City is 10 mph, 4 pmh or less in Brooklyn and Queens, and over 3 mph NYC Midtown.

In terms of transportation, according to [1], New York City is the second most congested city in the USA and number 42 in the world. It has one of the largest and oldest (1904) subway systems in the world. Due to its congestion, New York City inhabitants prefer to use public transportation (67.2% of workers commuting to work by this means in 2006 [5])

However, MTA the worst commutes in the world [4] commute time of 35.6 minutes on average commuting has been linked to obesity, stress, anxiety, depression, higher blood pressure, higher rates of divorce, neck and back pain and shorter lifespans.

The following lists summarize key points of the different transportation means used in the Datathon.

### Yellow Cabs

- Mostly located in Manhattan
- It is very difficult to get a taxi out of Manhattan, especially in rush hours.
- There are not many yellow cabs 130000, which are not enough.
- They can operate midtown and lower Manhattan and airports
- Rarely pick up outside manhattan

### Green Cabs

- Where created to standardize street hails outside of NYC. They operate in Brooklyn, Queens, The Bronx, and upper Manhattan.
- Airports: they can drop-off but not pick-up unless sent by a dispatcher
- They are not allowed to stop in the South of the upper west and upper east sides.
- Rides cheaper than Yellow cab rides

- For drivers green cabs was a way to get money without the pressure that yellow drivers have.
- Green drivers are also drivers or UBER
- 1/3 pick ups from Brooklyn, 1/3 Northern Manhattan, 1/3 Queens, a few in Bronx and Staten Island

## UBER

- Started in 2011 in Manhattan but expanded at the same time green.
- Uber has made yellow cabs steady but has impacted green that were just started
- Uber has made yellow cabs steady but has impacted green that were just started
- Connects drivers with more rides
- May 2015 busiest month on record.

According to the mobility report of 2016 [2], New York City is growing in jobs, residents and visitors. Its transportsations modes have extend to mass transit, walking, and cycling. New York advances focus on applying technology to real-time traffic management, Select Bus Service routes, reducing travel times, Expanding the city's bicycle lane network, improving pedestrian access to transit. It is still required to invest in ways to keep the city moving.

Despite the incursion of new transportation means e.g. Green Cabs, UBER, among others, it is still difficult to catch a taxi from outer Boroughs to Manhattan. Lower access to legal taxi rides for people in outer Boroughs.

## 1.2 Topic Question

**General question:**

*Is public transportation coverage in New York City  
equally attending all areas?*

**Exploratory questions:**

- What are the patterns related to unattended areas of public service?
- Is there are relationship between demographic information and peoples' choices of public transportation?

## 2 Data Wrangling and Data Cleaning

### 2.1 Data Cleaning

The data cleaning process was done in two steps:

- For yellow and green cab trips, the rows that have distances equal to 0 were deleted. This, because we are aiming to take into account only the trips that traveled some distance.
- For yellow and green cab trips, the IQR (Inter Quartile Range) methodology was used to clean the outliers from the data. A variable called "amount\_per\_distance" was created. It was calculated as the ratio between "total\_amount" and "trip\_distance". With this new variable, the values that did not show a common relationship between distance and values were deleted.
- When analyzing the data, we encountered that the columns precipitation, snowfall and snow\_depth had missing values in the form of a "? ?" character. For each column, we found 237 (10.82%), 91 (4.15%), 24 (1.09%) empty values respectively. Considering that these variables are highly correlated with the average temperature, we decided to apply an iterative imputation with a decision tree regressor estimator to them.

Dataset	Initial	Deleted	Final
Uber trips	18676106	0	ss
Yellow cab trips	7926168	337998	7588770
Green cab trips	3537586	186494	3351092
MTA trips	7554197	0	ss
Weather	2190	0	2190

Table 1: Summary of the main information available to develop the project.

## 2.2 Feature Engineering

The following explains the different new features that were created.

- A new variable that measures the ratio between the total amount of the trip and the distance it traveled. This feature was created for Yellow trips and Green trips and was used for the outlier cleaning.
- “trips\_pickup”: this variable contains the total number of pick ups of all the transportation means (green cabs, yellow cabs, and UBER, NTA), for every NTA.
- “log\_trips\_pickup”: the previous variable “trips\_pickup” was transformed to log space for better visualization of the values.
- “yellow\_indicator”: A threshold was applied to the variable “log\_trips\_pickup”. If the number of trips was  $\zeta$  threshold, then “yellow\_indicator” is 0, otherwise 1. The threshold was defined as the 0.4 quantile of the variable “log\_trips\_pickup”. This variable allow us to know in which NTAs, public transportation has not an adequate coverage, or in which ones public transportation is not commonly used. Both analyzed with the total number of trips per NTA.

## 3 Exploratory Data Analysis

What hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?

### 3.1 Hypothesis

During the data analysis we stated different hypothesis:

- **H1:** Due to UBER incursion in the city, the **travel distances** of yellow cabs trips increased.
- **H2:** Due to UBER incursion in the city, the **number of trips** of yellow cabs was reduced.
- **H3:** The areas of the city where there is **no public transportation** coverage, correspond to areas with **low income families**.
- **H4:** Due to UBER incursion in the city, the price of yellow cabs trips have decreased.

### 3.2 Data Exploration

Different plots were created with the aim of understanding the behaviour of the different transportation systems. The following figures show some of the comparisons that were done, and shows the relationship of the plots with the previously mentioned hypothesis.

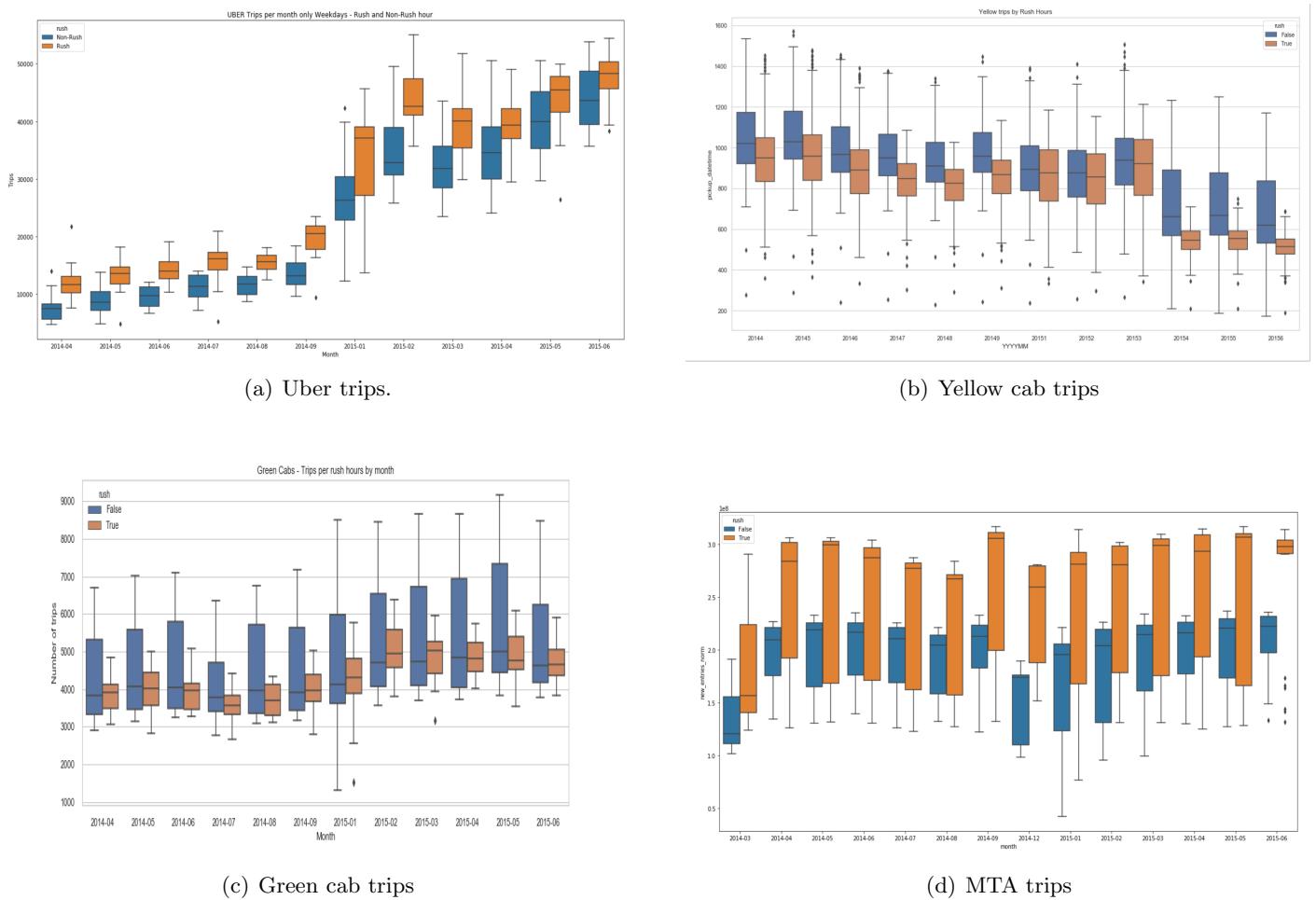


Figure 1: Has the increase of Uber trips affected the number of trips of Yellow cab, Green cab, and MTA? Orange boxes represent the number of trips in rush hours and blue ones correspond to non-rush hours.

### 3.2.1 General Plots

Figure 1 shows the boxplots of the monthly number of trips for the different transportation systems. Figure 5(a) for Uber's trips, Figure 6(a) for Yellow cab trips, Figure 6(b) for Green cab trips, and Figure 6(c) for MTA trips. The boxplots differentiate the trips between rush hours (orange boxes) and non-rush hours (blue boxes). From the figures, it can be seen that the MTA is busier in rush hours than in non-rush hours. Additionally, it is possible to see that there has been a significant increase of the number of trips taken by Uber from 2015 both in rush and non-rush hours; and a decrease on the number of trips taken by Yellow cabs, especially in rush hours. The latter corresponds to the ideas stated in hypothesis **H3**.

Figure 2 shows the same plot, but separating weekdays (blue boxes) from weekends. In the figure we can also see the previously mentioned behaviour: UBER trips increase and yellow trips tend to decrease.

Additionally, Figure 3 was created to analyze if users prefer to use a specific service at a specific hour. We can see that pick ups are high when people are moving to work, i.e. 7 am (yellow, UBER, and green cabs have a high number of pick ups during that hour). Nevertheless, the highest number of pick ups occur between 6 and 7 pm, the time when many people go home or go out for fun. MTA, on the other hand, shows a different behaviour during the day, with different busy moments.

Figure ?? compares the number of trips made by Uber with the monthly average travel distance covered by Yellow cabs. The aim of this comparison was to analyze if the increase of Uber trips affected the average travel distance of plots, it is possible to see that from 2015, Uber is widely used for long distance trips, in contrast to Yellow cabs. On the other hand, the average travel distance of Yellow cabs experienced an increase in April 2015.

On the other hand to analyze Hypothesis **H4**, the plots of Figure 4 were created. Figure 4(a) and 4(c) proof what was mentioned in in Section 1.1, the price of a trip by Green cab is cheaper than by Yellow cabs. However,

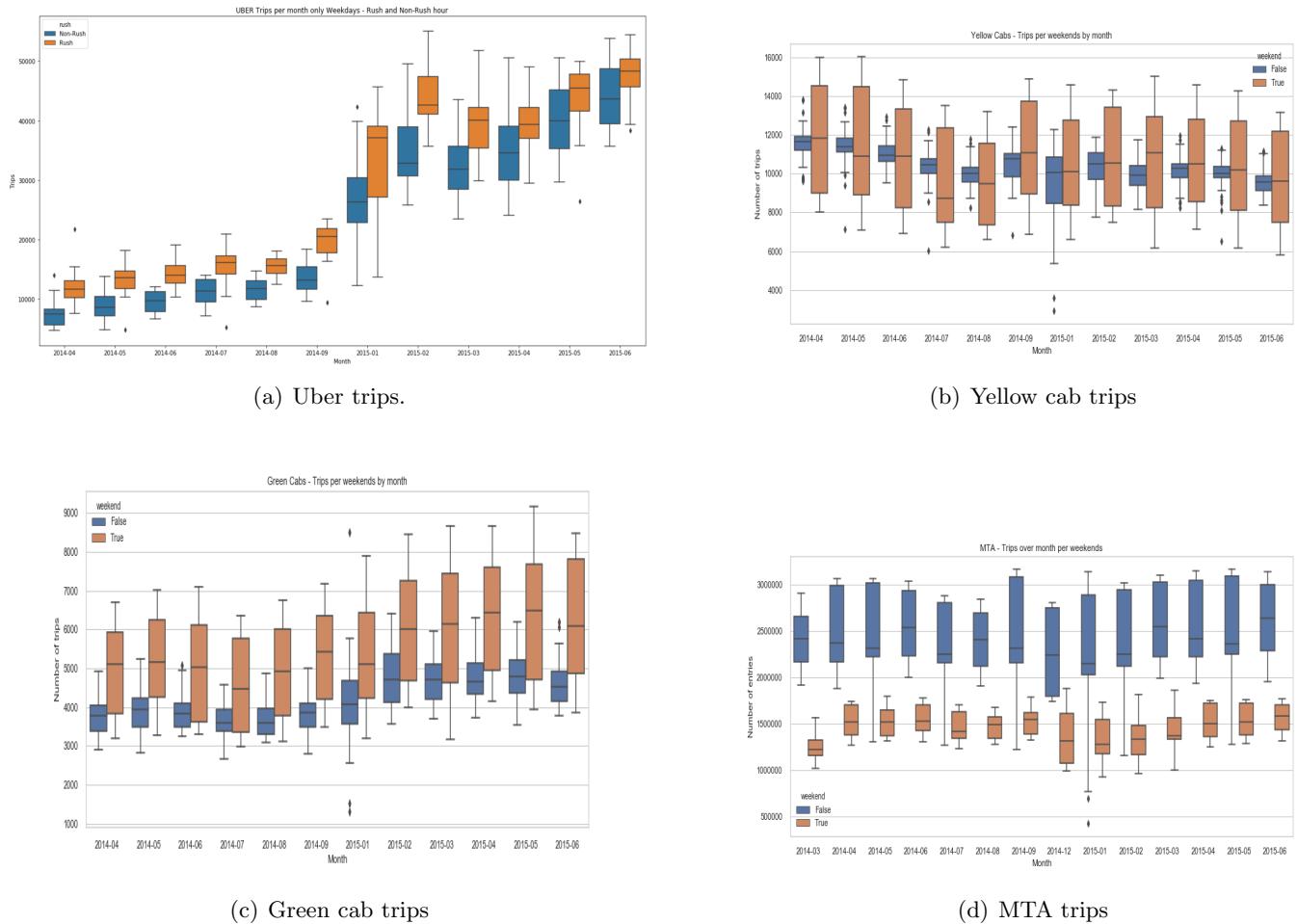
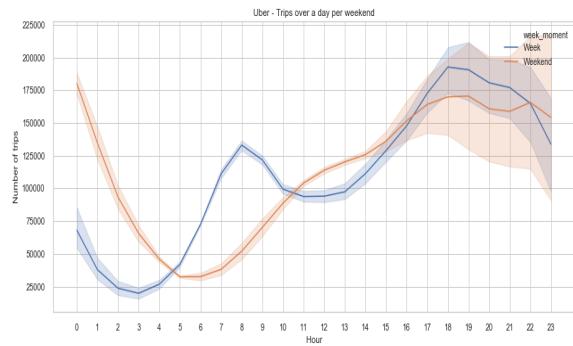
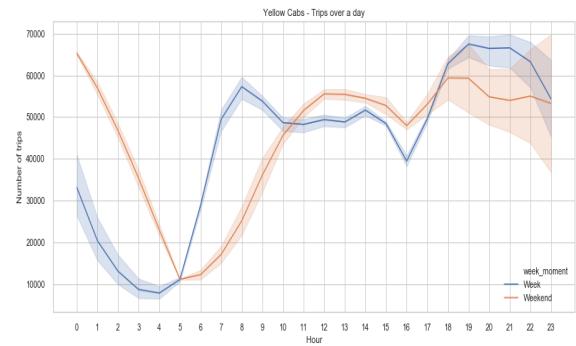


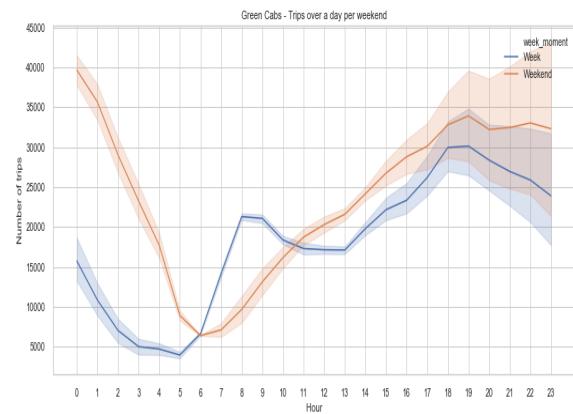
Figure 2: Has the increase of Uber trips affected the number of trips of Yellow cab, Green cab, and MTA? Orange boxes represent the number of trips in weekends and blue ones correspond to weekdays.



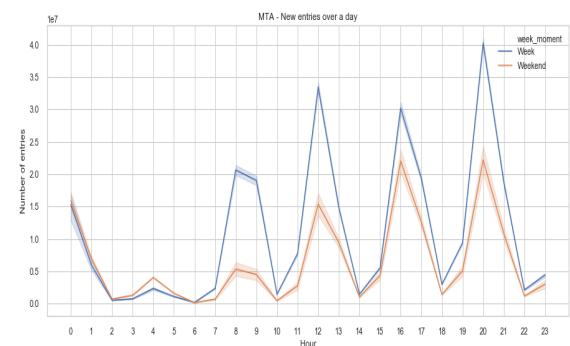
(a) Uber trips.



(b) Yellow cab trips



(c) Green cab trips



(d) MTA trips

Figure 3: Behaviour of the number of trips per hour.

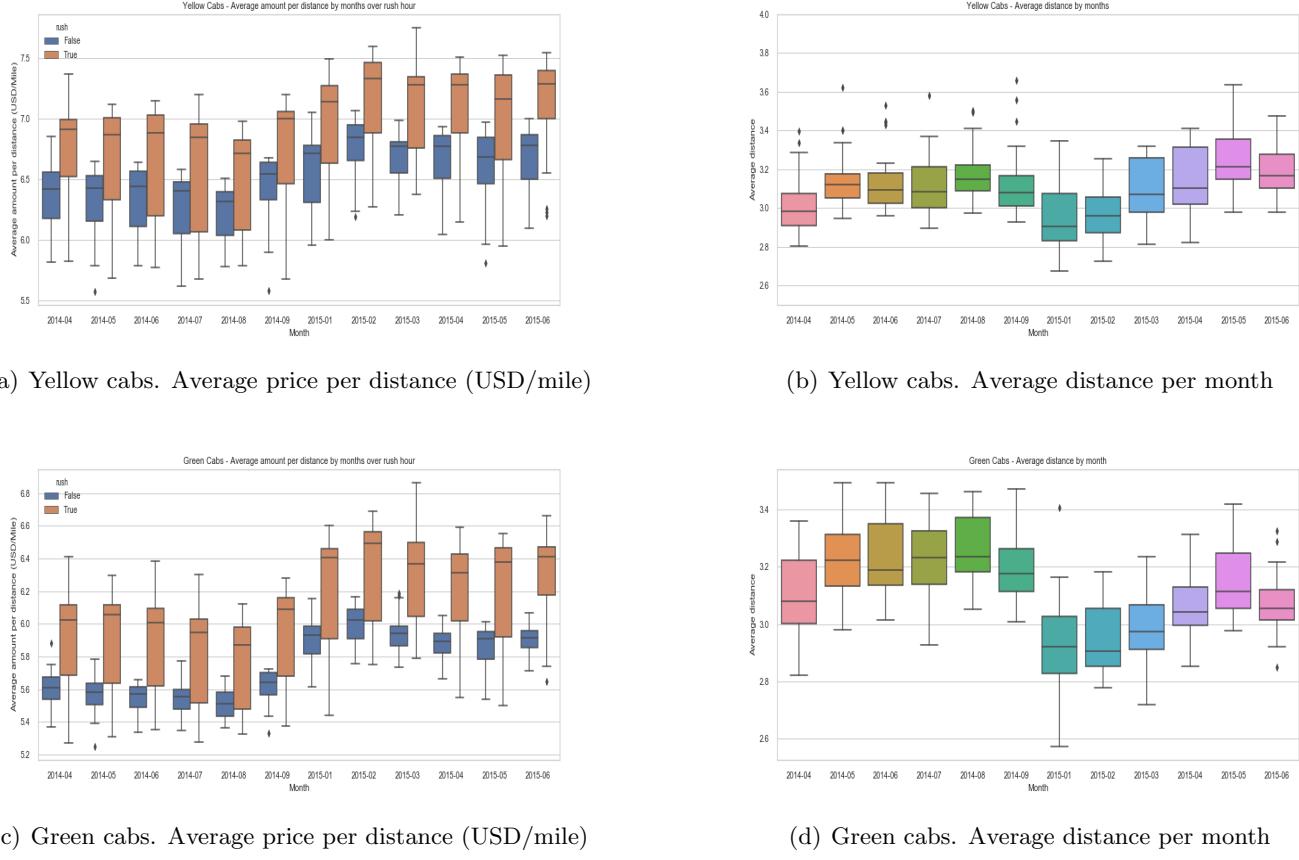


Figure 4: Has the price per distance decreased in Green and Yellow cabs?

the plots show that the price of Green and Yellow cab trips have not reduced as we thought.

Figures 4(b) and 4(d) analyze the traveled distances of yellow and green cabs. The plots show that they have not vary significatively, as we stated in Hypothesis **H1**.

### 3.2.2 Heat Maps

In this section different maps were created to analyze which areas of the city, in terms of NTAs, are the ones with less coverage; or what are the most common transportation choices of people, per NTA. Two different maps were created. The first one, Figure 5, shows the number of pick-ups and and the pickup zone, of the different transportation options. On the other hand, Figure 6, shows the number of drop-offs and the drop-off zones.

Table 3.2.2 shows the Links to access the interactive Heat Maps. The maps show the number of trips per NTA, in the different transportation options.

Figure Name	Link to Map
Trips NTA Green Dropoff Map	<a href="#">Link</a>
Trips NTA MTA Dropoff Map	<a href="#">Link</a>
Trips NTA Yellow Dropoff Map	<a href="#">Link</a>
Trips Population NTA Green Map	<a href="#">Link</a>
Trips Population NTA MTA Map	<a href="#">Link</a>
Trips Population NTA Uber Map	<a href="#">Link</a>
Trips Population NTA Yellow Map	<a href="#">Link</a>

Table 2: Links to access interactive Heat Maps. The maps show the number of trips per NTA, in the different transportation options.

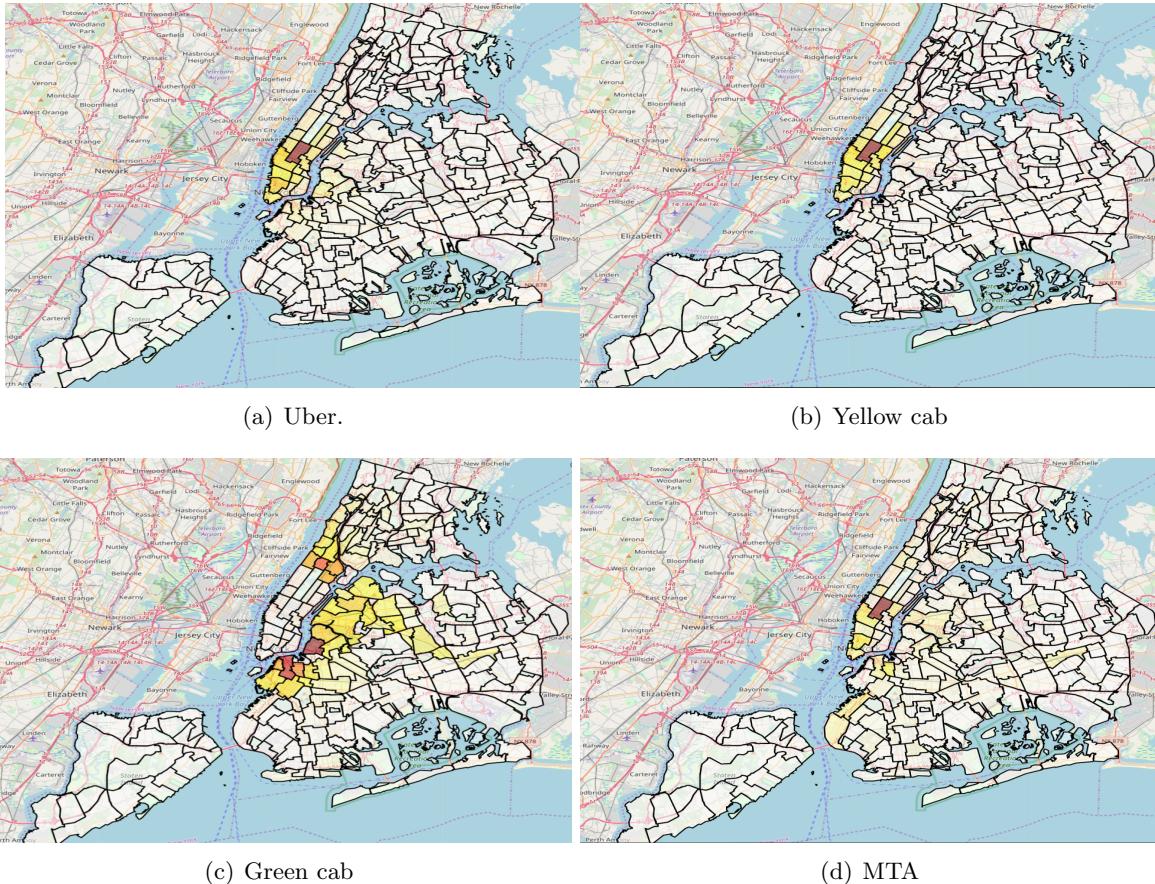


Figure 5: Heat maps of Pick Ups. The maps show the number of trips per NTA, in the different transportation options. The data has been normalized by the NTA population.

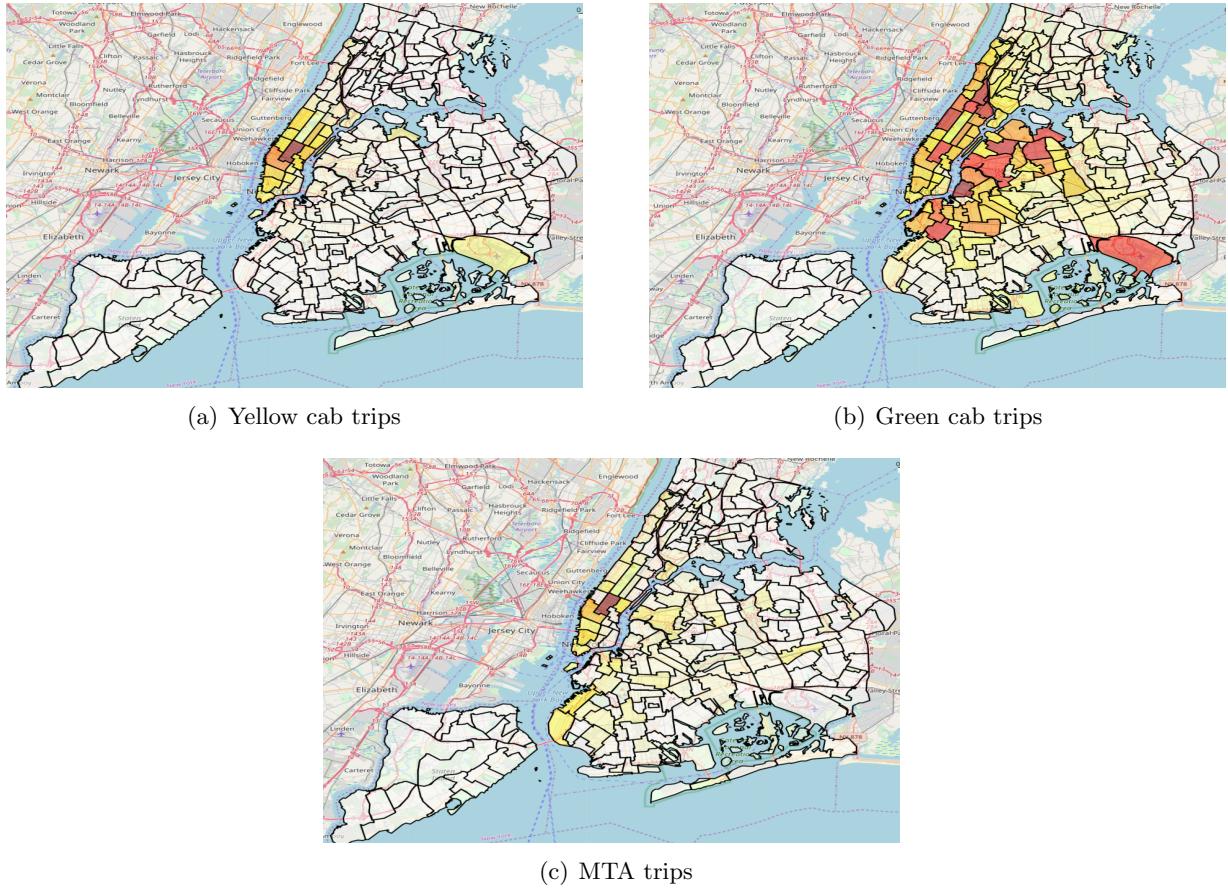
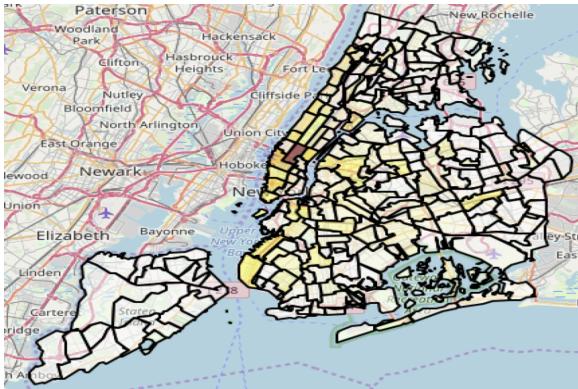
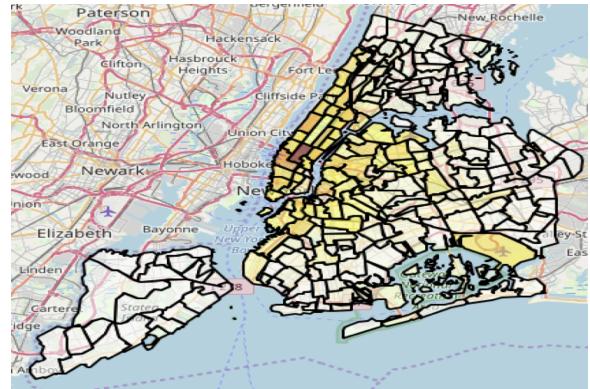


Figure 6: Heat maps of Drop offs. The maps show the number of trips per NTA, in the different transportation options.



(a) Total Pick Ups per NTA



(b) Total Drop Offs

Figure 7: Analyzing the total number of Pick Ups and Drop Off of all the transportation options.

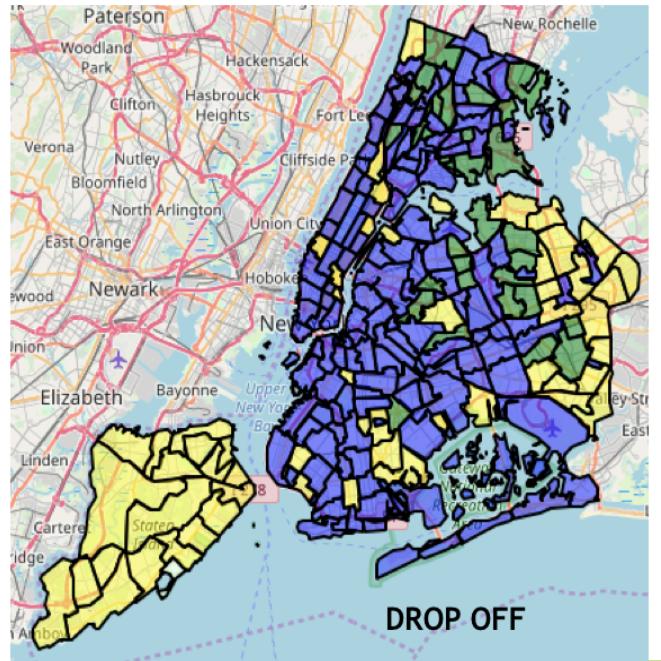
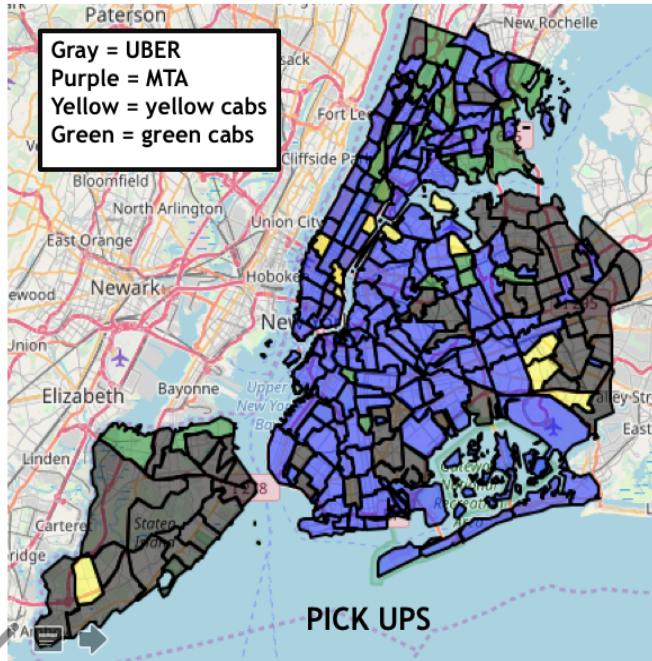


Figure 8: Most frequently transportation mean per NTA. The maps show the number of trips per NTA, in the different transportation options.

Figure 7 shows the location and the total number of pick ups and drop offs, of Green, Yellow, UBER (only pick up information was available) and MTA transportation means. The data is normalized by the population of each MTA. These maps start showing areas where public transportation is not widely used.

On the other hand, Figure ?? shows per NTA, which is the transportation option more used in the area. Analyzing the map that appears on the left (Pick ups), it can be seen that in the outer areas of the city, UBER is widely used. However, MTA is the prefer transportation option in most of the city. The latter coincides with what was found in Section 1.1: NYC is famous because people does not like to own a car, they prefer to use public transportation.

The right figure shows the drop off map. It is important to remember that there is no drop off information of UBER, and this is why in the outer areas, it is possible see that yellow and green cabs options are widely used. This map is different from what we expected. The reason is that according to the literature, Yellow cars do not like to move far from Manhattan. However, the map shows that they are moving far from it.

## References

- [1]
- [2]
- [3] Ten worst commutes in the world. Date retrieved 7-12-2019.
- [4] Ten worst commutes in the world. Date retrieved 7-12-2019.
- [5] Transportation in new york city. Date retrieved 7-12-2019.