Datathon Group 3

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

November 16, 2019

Abstract

In this document the project scope and plan for the Datathon are presented. The document provides information of the data cleaning process and some plots with preliminar results of the data wrangling process.

1 Project Scoping & Plan

1.1 Scope

- **Project Objective:**
  - **Main question**: *How do yellow cabs mean trip distance have changed over time (rush/non-rush hours) as a result of Uber's trips growth?*
    What are the patterns related to unattended areas of public service?
- **Main stakeholders**: the NYC citizen and government, and transportation industry (at all levels).
- **Boundaries of the project**:
  - We will show metrics of the impact of Uber incursion in NYC over the other transportation means.
  - The analysis will be made only on the information of the NYC Boroughs.
- **Risks:**
  - Data quality issues in the datasets.
  - The data might be not sufficient to answer the proposed question.

1.2 Plan

- **Summary**: *How do yellow cabs mean trip distance have changed over time (rush/non-rush hours) as a result of Uber's trips growth?* From this one we can analyze the mean income of the zones where yellow cabs drop-off zones changed.
- **Expected Deliverable**: A report with the topic question, Data wrangling and Cleaning process, Exploratory Data Analysis EDA, Statistical Analysis and Modeling, Results Interpretation and Conclusions.
- **How to get there**:
  - Clean, wrangled and analyze the dataset.
  - Conduct exploratory data analysis.
  - Conduct Analysis & modeling.
  - Conclusions and final report (source code and power point presentation).

2 Introduction

2.1 Background Information

2.1.1 Yellow cabs

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

2.1.2 Green cabs

- Where created to standardise street hails outside of NYC. Brooklyn, Queens, The Bronx, upper Manhattan.
- Created to provide more access to metered taxis. Less expensive than livery cars.
- They operate Manhattan Bello 110th St on the west side and below 96th street on the east side or at either la guardia or JFK airports.
- Airports: they can drop-off but not pick-up unless sent by a dispatcher
- They can be on call by dispatcher
- They are not allow to stop in the South of the upper west and upper east sides.
- The permits of Green cabs are cheaper than yellow cabs and easier to acquire. Cab license affordable for drivers.
- Rides cheaper than yellow
- 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

2.1.3 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.

2.1.4 NYC Passengers

- Difficult to catch a taxi from outer Boroughs to Manhattan. Especially for green cabs where it is not worthy for drivers because they cannot pick up a fare on the way back out of Manhattan.
- Lower access to legal taxi rides for people in outer Boroughs
- Green cabs try but never fulfilled their promise

3 Data Wrangling and 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.

Feature engineering: We 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 cleansing.

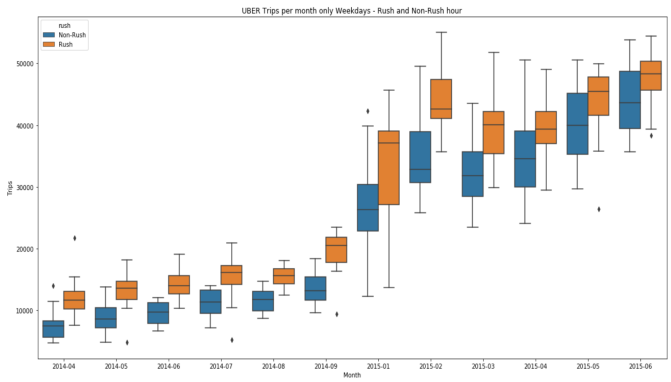
4 Exploratory Data Analysis

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?

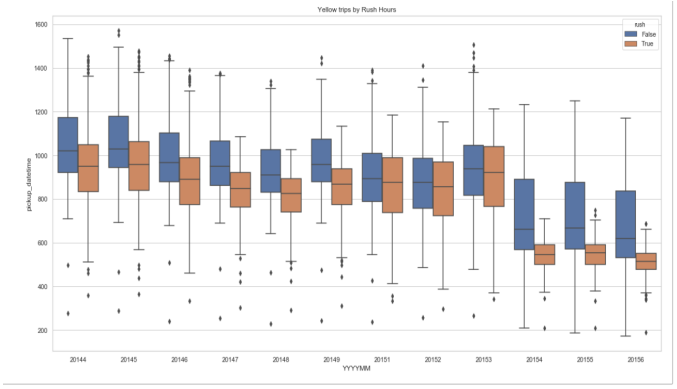
4.1 Data Analysis

Different plots were created with the aim of understanding the behaviour of the different transportation systems. The following figures show some of the comparisons made, so far.

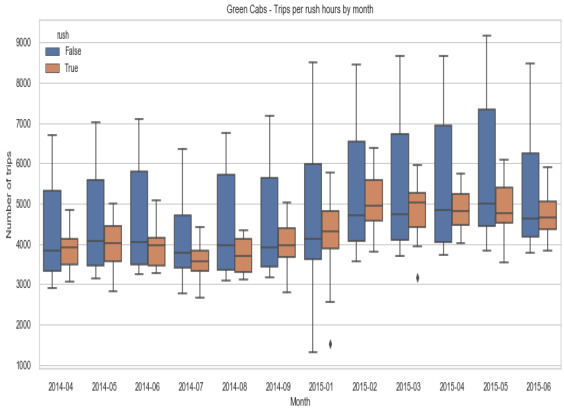
Figure 6 shows the boxplots of the monthly number of trips for the different transportation systems. Figure 7(a) for Uber's trips, Figure 8(a) for Yellow cab trips, Figure 8(b) for Green cab trips, and Figure 8(c) for MTA trips. The boxplots differentiate the trips between rush hours (orange boxes) and 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



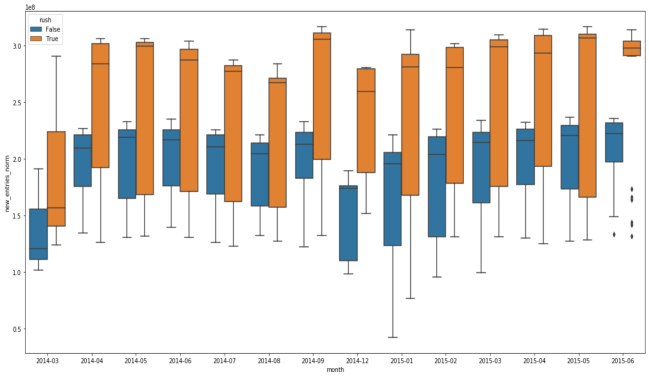
(a) Uber trips.



(b) Yellow cab trips

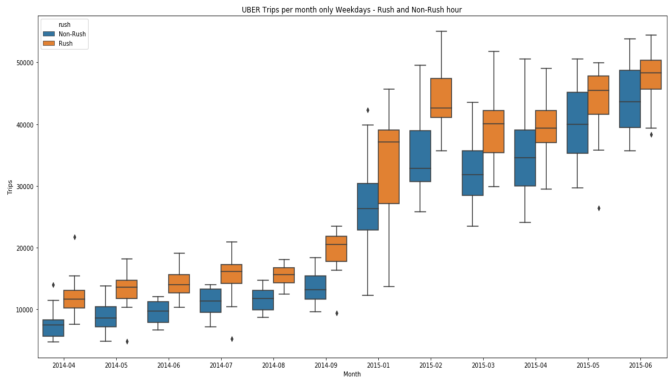


(c) Green cab trips

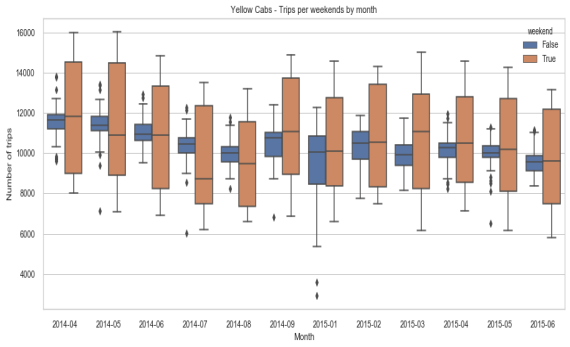


(d) MTA trips

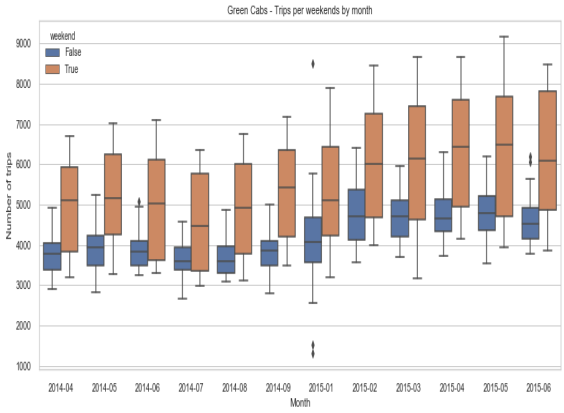
Figure 1: Has the increase of Uber trips affected the number of trips of Yellow cab, Green cab, and MTA trips? Orange boxes represent the number of trips in rush hours and blue ones correspond to non-rush hours.



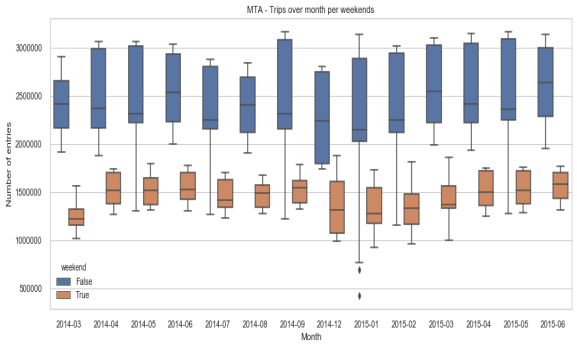
(a) Uber trips.



(b) Yellow cab trips



(c) Green cab trips



(d) MTA trips

Figure 2: trips weekend by month

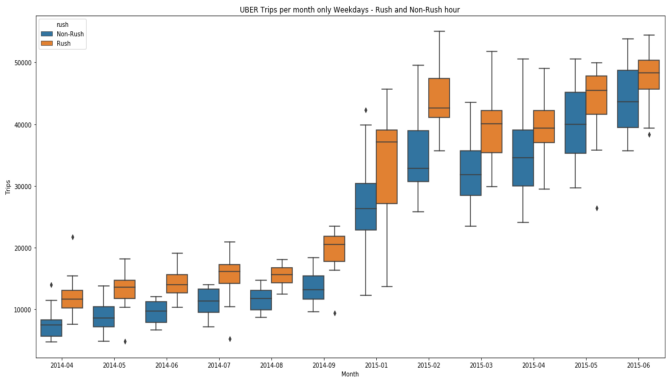
rush and non-rush hours; and a decrease on the number of trips taken by Yellow cabs, especially in rush hours. On the other hand

Figure 5 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.

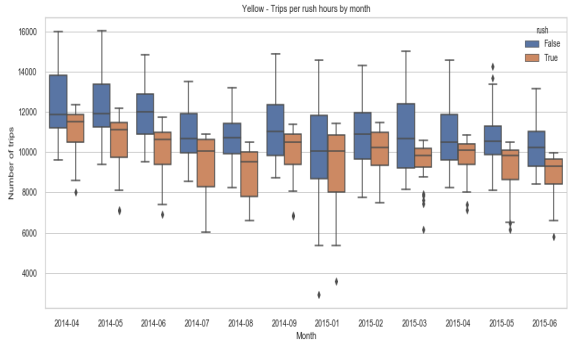
4.1.1 Heat Maps

In this section different maps were created in order to analyze which areas of the city, in terms of NTAs, are the ones with less coverage. Two different maps were created. The first one, Figure 7, shows the number of pick-ups and and the pickup zone, of the different transportation options. On the other hand, Figure 8, shows the number of drop-offs and the zones.

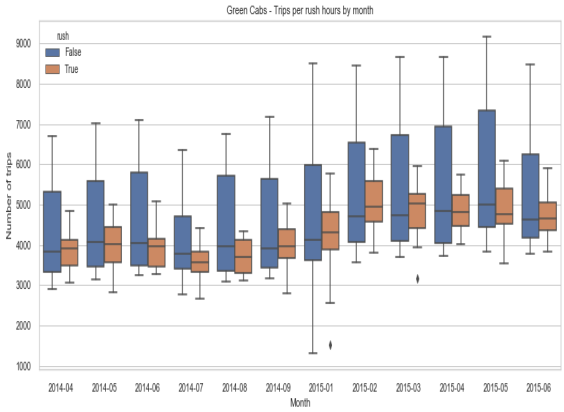
Table 4.1.1 shows the Links to access the interactive Heat Maps. The maps show the number of trips per NTA, in the different transportation options.



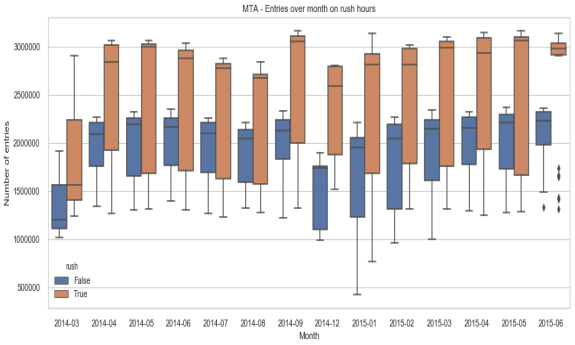
(a) Uber trips.



(b) Yellow cab trips

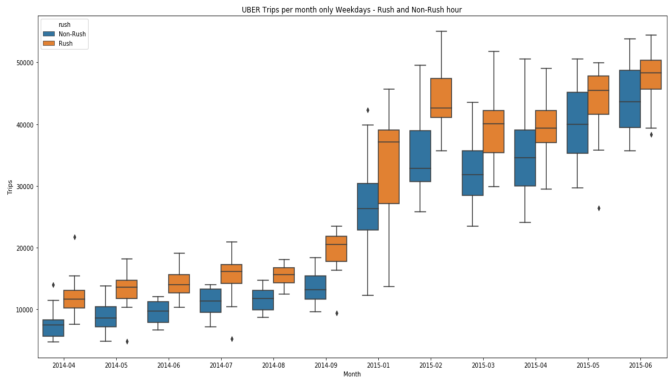


(c) Green cab trips

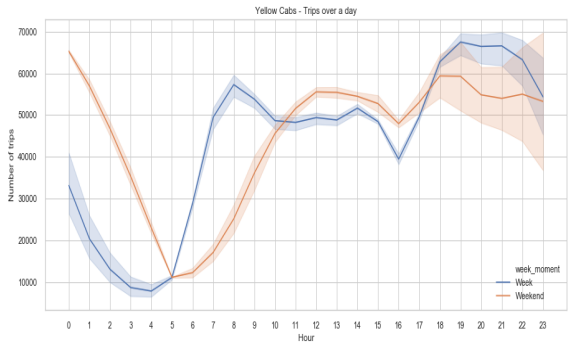


(d) MTA trips

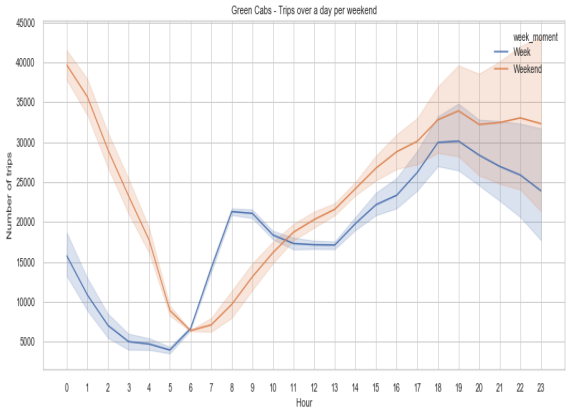
Figure 3: trips month rush hours



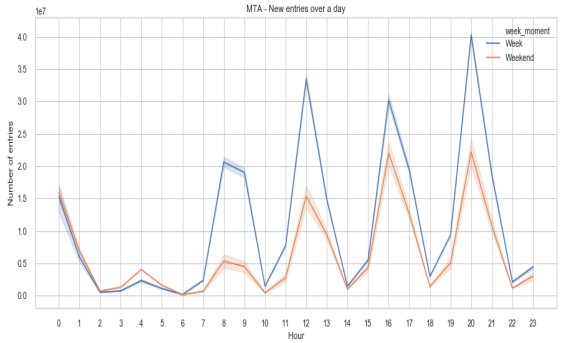
(a) Uber trips.



(b) Yellow cab trips

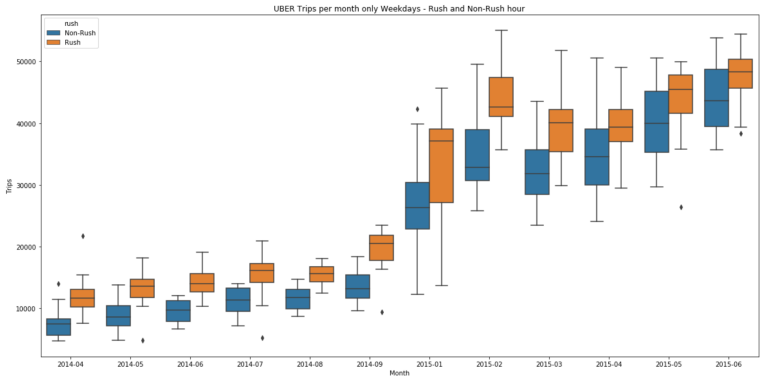


(c) Green cab trips

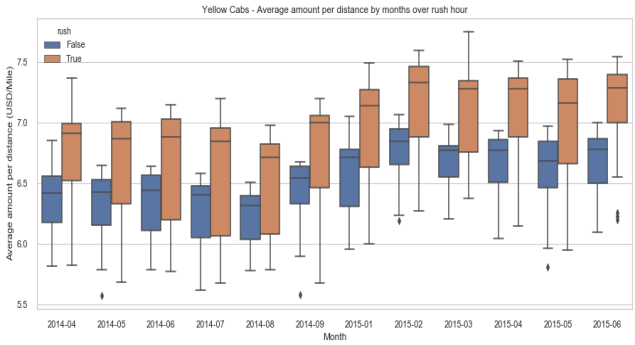


(d) MTA trips

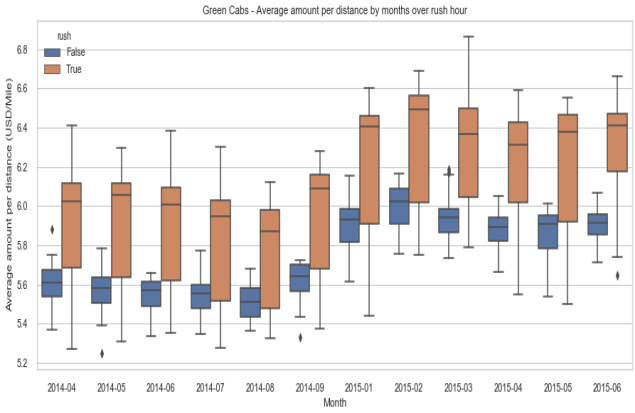
Figure 4: Hour week



(a) Uber trips.

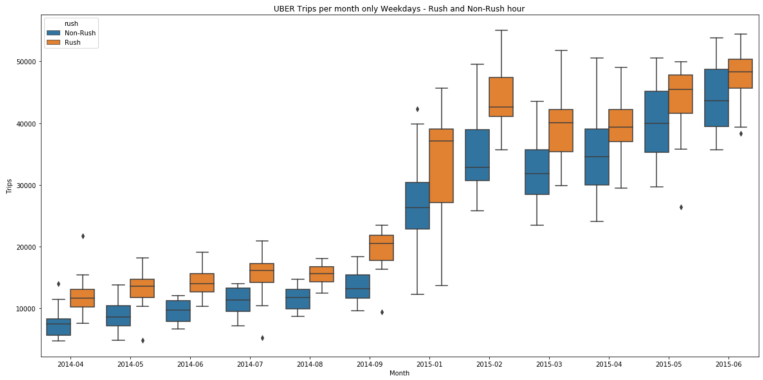


(b) Yellow cab trips

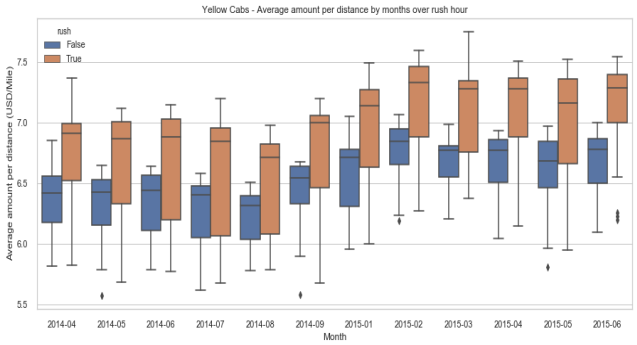


(c) Green cab trips

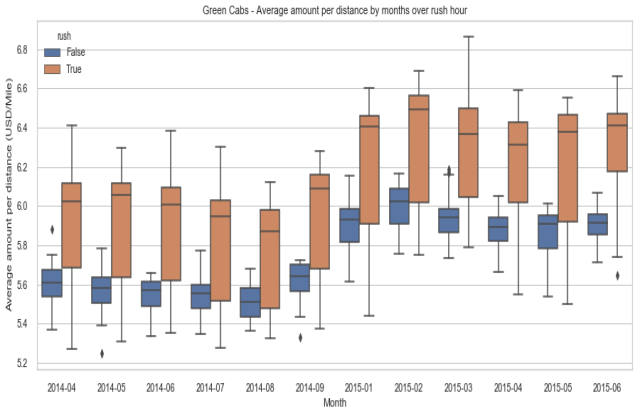
Figure 5: price per distance



(a) Uber trips.

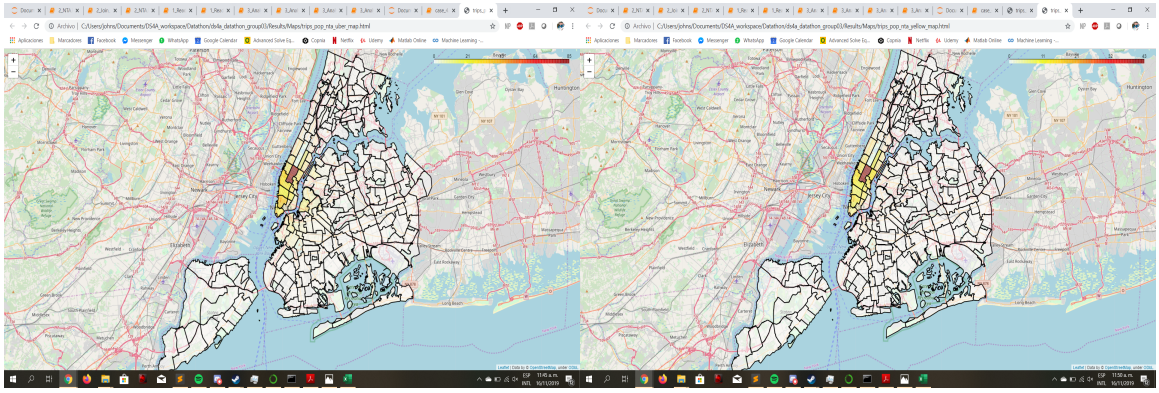


(b) Yellow cab trips

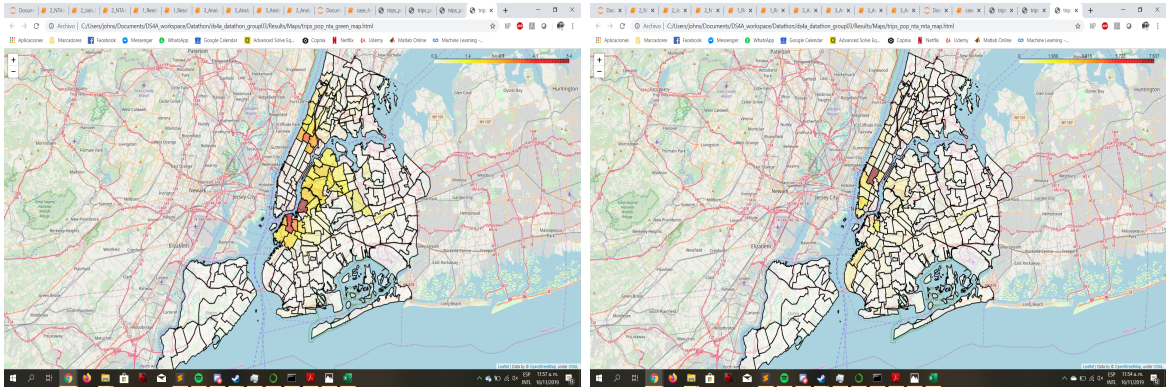


(c) Green cab trips

Figure 6: avrg distance

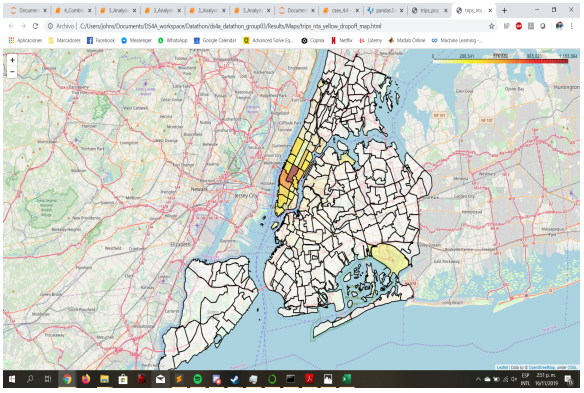


(a) Uber. (b) Yellow cab

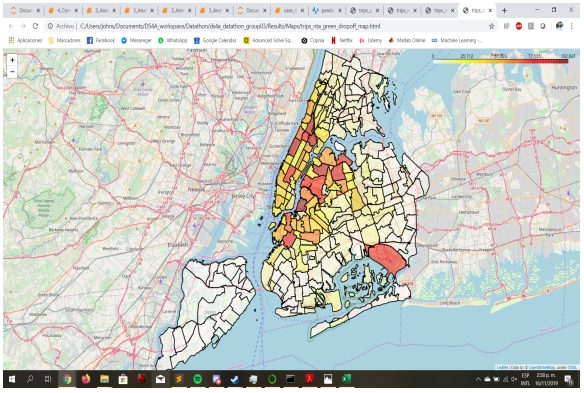


(c) Green cab (d) MTA

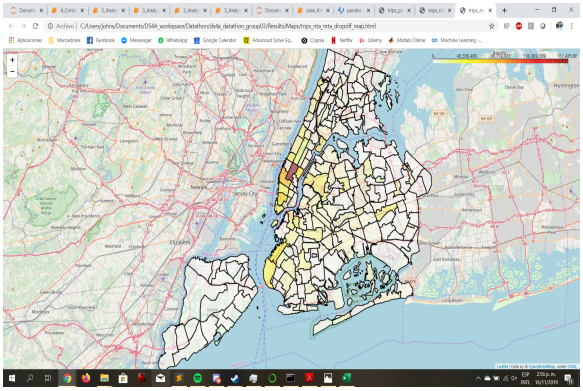
Figure 7: Heat maps of Pick ups. The maps show the number of trips per NTA, in the different transportation options.



(a) Yellow cab trips



(b) Green cab trips



(c) MTA trips

Figure 8: Heat maps of Drop offs. The maps show the number of trips per NTA, in the different transportation options.

Figure Name	Link to Map
Trips NTA Green Dropoff Map	Link
Trips NTA MTA Dropoff Map	Link
Trips NTA Yellow Dropoff Map	Link
Trips Population NTA Green Map	Link
Trips Population NTA MTA Map	Link
Trips Population NTA Uber Map	Link
Trips Population NTA Yellow Map	Link

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