Datathon Group 3

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

October 26, 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 caps mean trip distance have changed over time (rush/non-rush hours) as a result of Uber's trips growth?*
- **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 caps 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 Data Wrangling and Data Cleaning

The data cleaning process was done in two steps:

- For yellow and green cap trips, the rows that have distances equal to 0 we deleted. This, because we are aiming to take into account only the trips that traveled some distance.
- For yellow and green caps trips, the IQR 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.

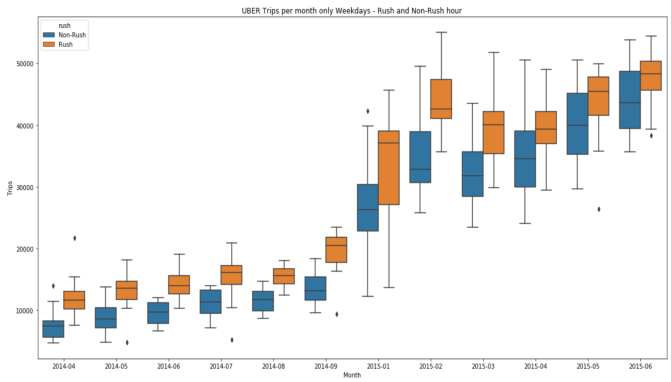
Feature engineering: We created a new variable that measures the ratio between the total amount of the trip and the distance it traveled. This feature was created for Yellow trips and Green trips and was used for the outlier cleansing.

2.1 Data Analysis

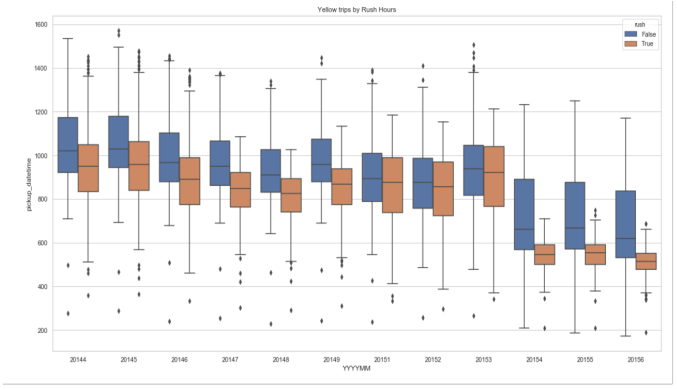
Different plots were created with the aim to understand the behaviour of the different transportation systems. The following plots summarize the important findings encountered, so far.

Figure 1 shows the graph plots of the monthly trips for the different transportation systems: Figure 1(a) for Uber's trips, Figure 1(b) for Yellow cap trips, Figure 1(c) for Green cap trips, and Figure 1(d) for MTA trips. The boxplots differentiate the trips between rush hours (orange boxes) and non-rush hours (blue boxes). From the figure it can be seen that the MTA is highly used in rush hours. Additionally, it is possible to see that there has been a significan 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 caps, especially in rush hours.

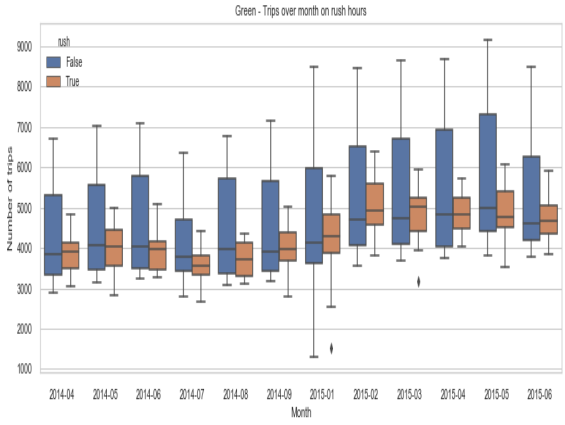
Figure 2 compares the monthly average travel distance covered by Uber and Yellow caps. From these plots, it is possible to see that from 2015, Uber is widely used for long distance trips, in contrast to Yellow caps. On the other hand, the average travel distance of Yellow caps experienced an increase in April 2015.



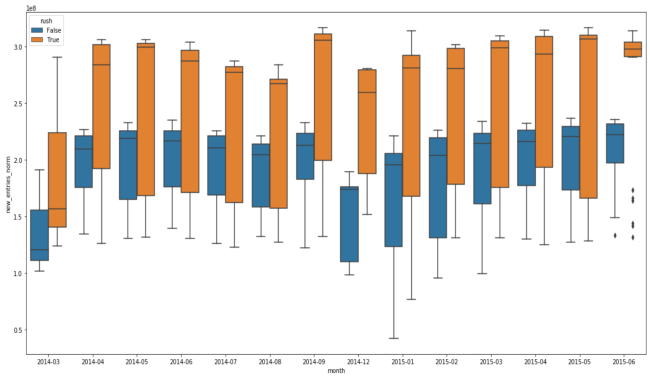
(a) Uber trips.



(b) Yellow cap trips

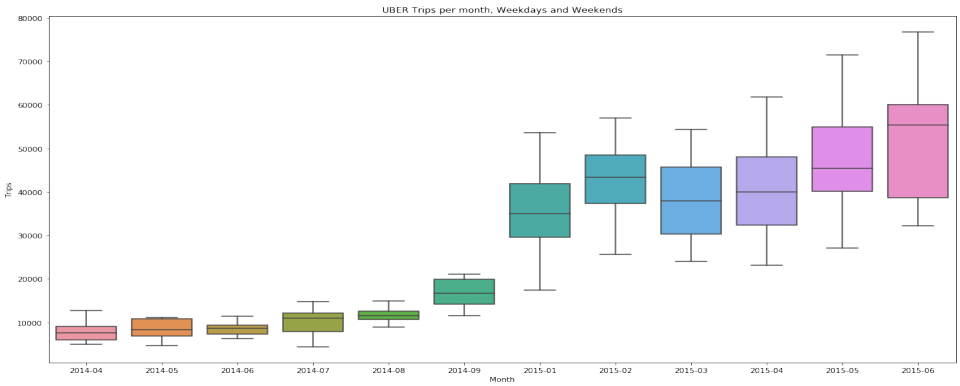


(c) Green cap trips

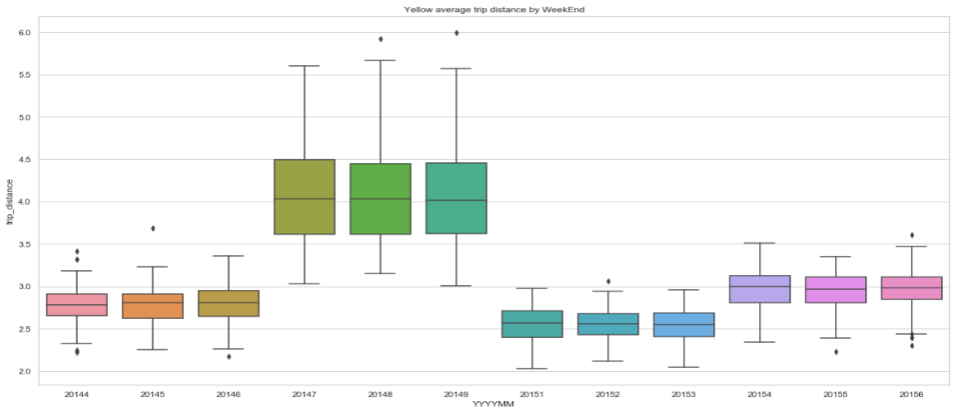


(d) MTA trips

Figure 1: Monthly behaviour of the number of trips. Orange boxes represent the number of trips in rush hours and blue ones correspond to non-rush hours.



(a) Uber



(b) Yellow caps

Figure 2: Comparison of the monthly average travel distance covered by Uber and Yellow caps. From the plot