Datathon Group 3

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

November 9, 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?*
- **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 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.

3 Exploratory Data Analysis

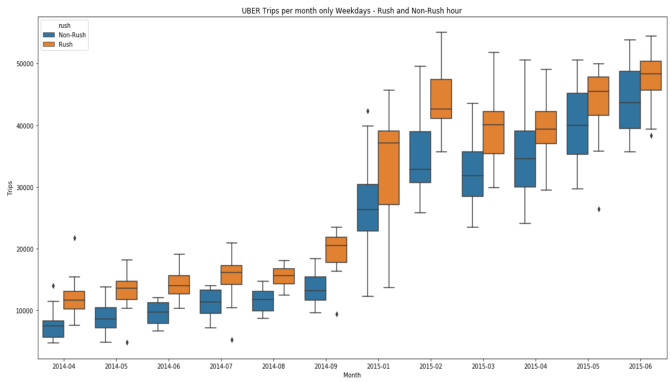
What hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?

3.1 Data Analysis

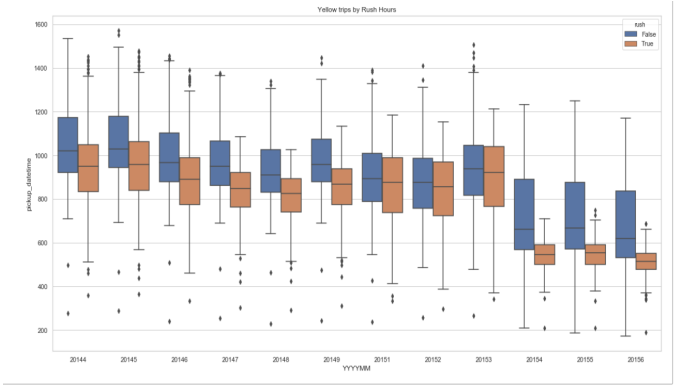
Different plots were created with the aim of understanding the behaviour of the different transportation systems. The following figures show some of the comparisons made, so far.

Figure 6 shows the boxplots of the monthly number of trips for the different transportation systems. Figure 6(a) for Uber's trips, Figure 6(b) for Yellow cab trips, Figure 6(c) for Green cab trips, and Figure 4(d) 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. On the other hand

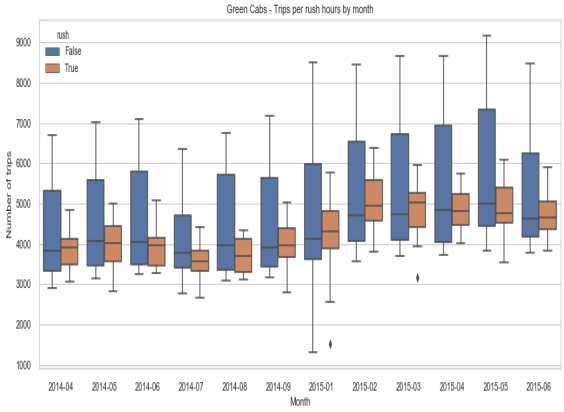
Figure **??** compares the number of trips made by Uber with the monthly average travel distance covered by Yellow cabs. The aim of this comparison was to analyze if the increase of Uber trips affected the average travel distance of plots, it is possible to see that from 2015, Uber is widely used for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



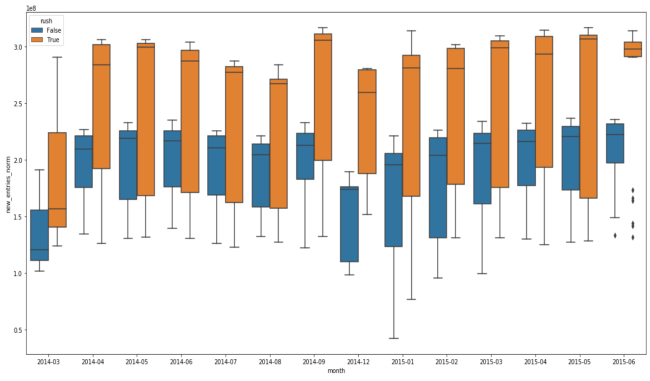
(a) Uber trips.



(b) Yellow cab trips

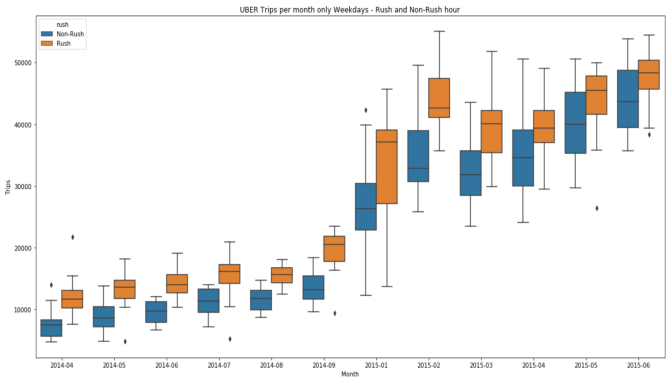


(c) Green cab trips

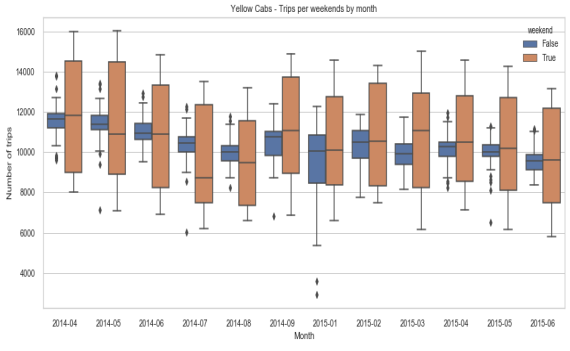


(d) MTA trips

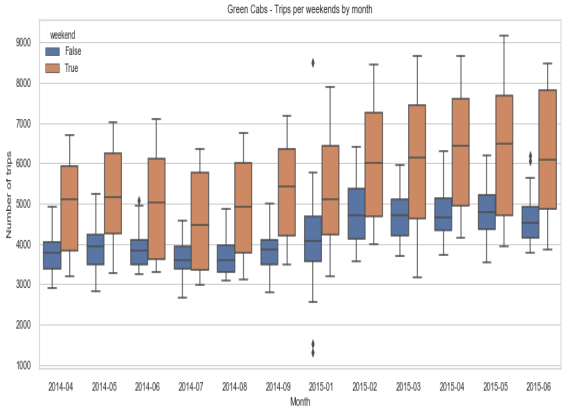
Figure 1: Has the increase of Uber trips affected the number of trips of Yellow cab, Green cab, and MTA trips? Orange boxes represent the number of trips in rush hours and blue ones correspond to non-rush hours.



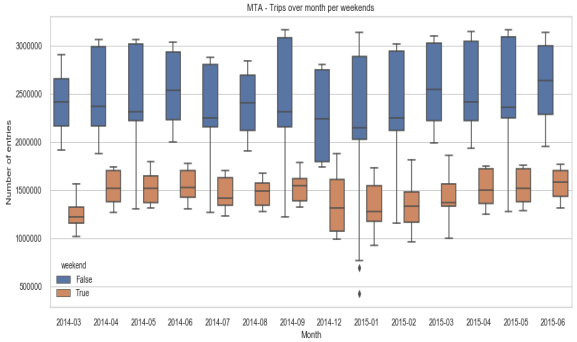
(a) Uber trips.



(b) Yellow cab trips

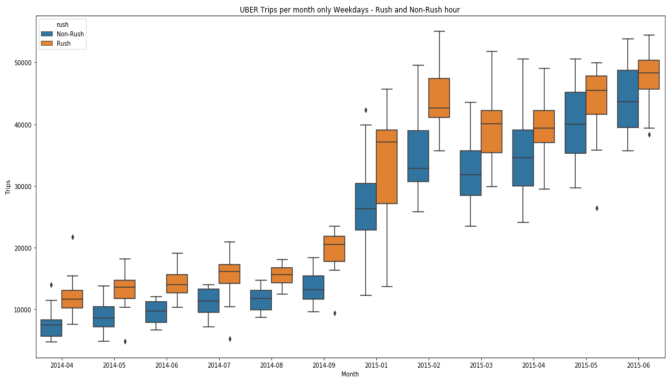


(c) Green cab trips

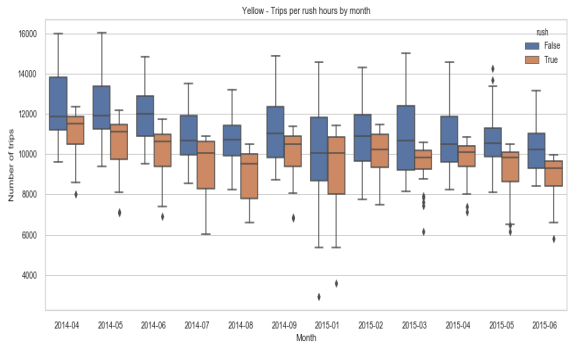


(d) MTA trips

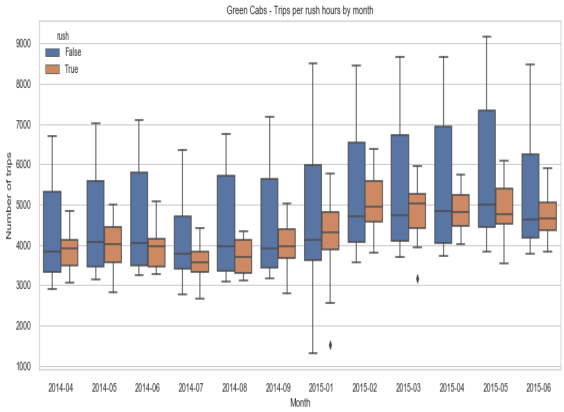
Figure 2: trips weekend by month



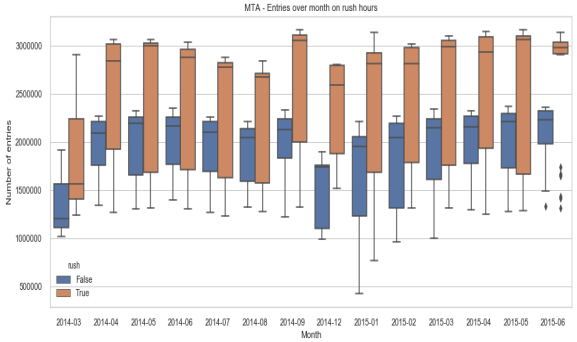
(a) Uber trips.



(b) Yellow cab trips

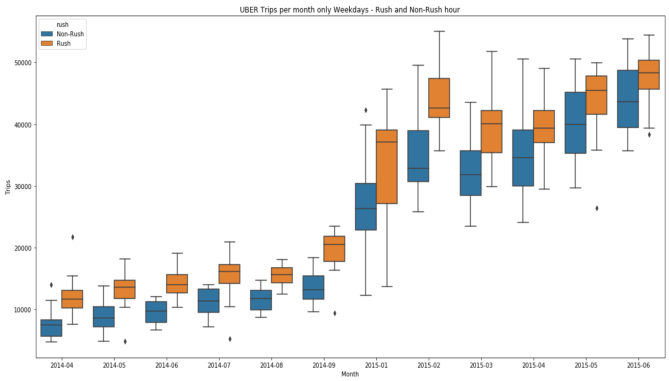


(c) Green cab trips

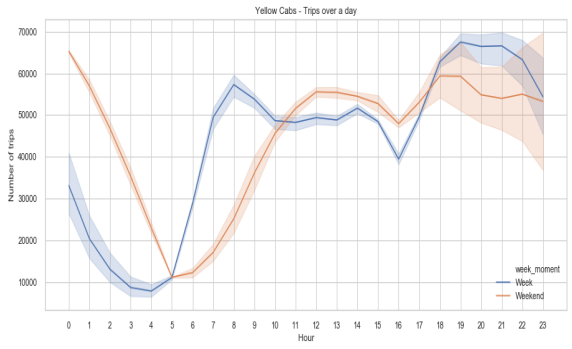


(d) MTA trips

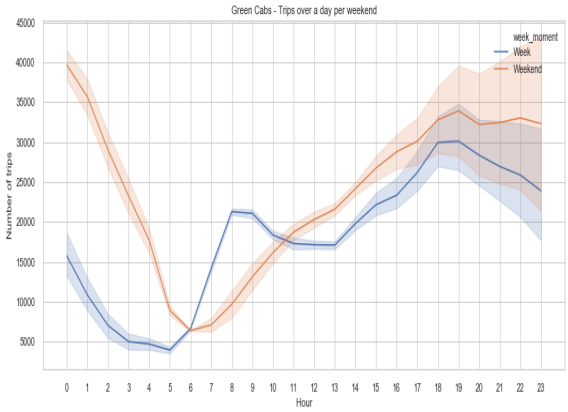
Figure 3: trips month rush hours



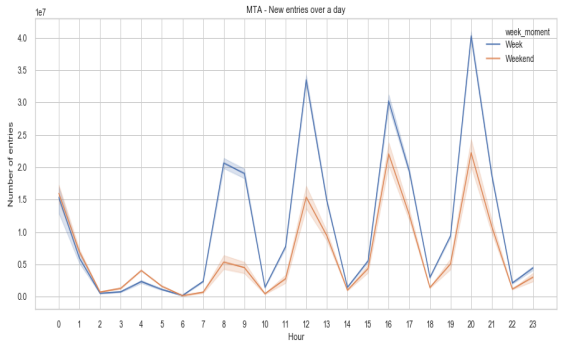
(a) Uber trips.



(b) Yellow cab trips

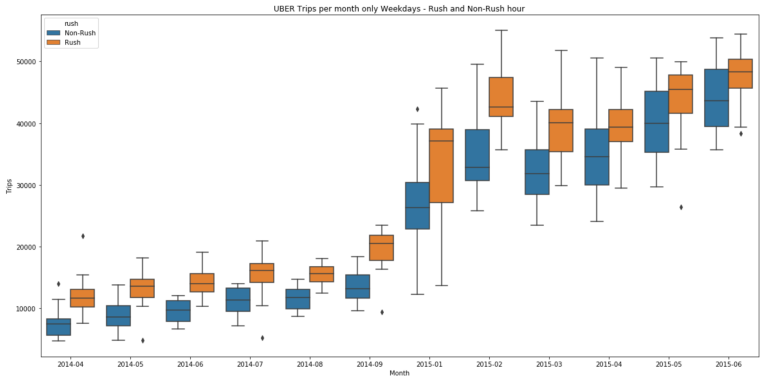


(c) Green cab trips

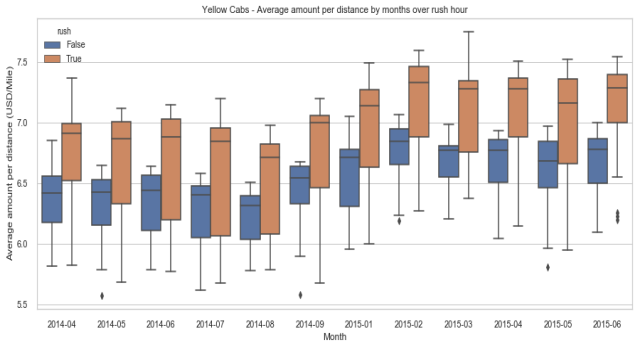


(d) MTA trips

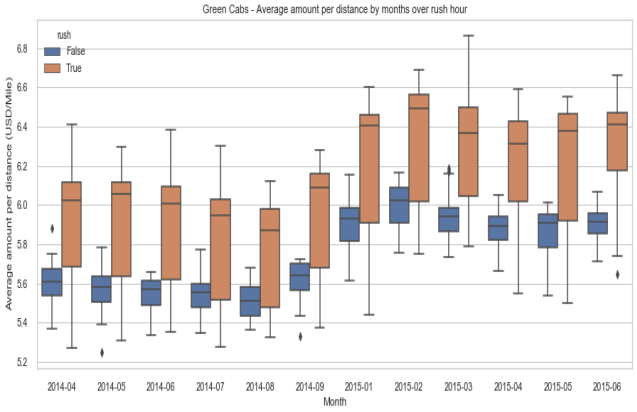
Figure 4: Hour week



(a) Uber trips.

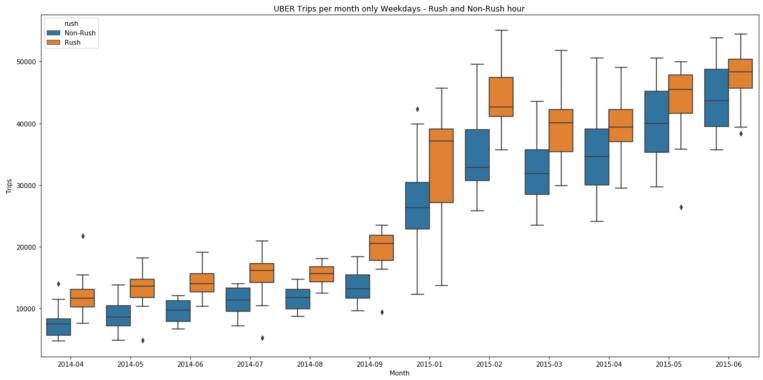


(b) Yellow cab trips

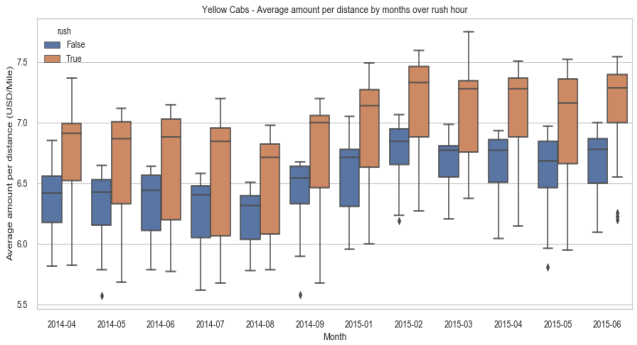


(c) Green cab trips

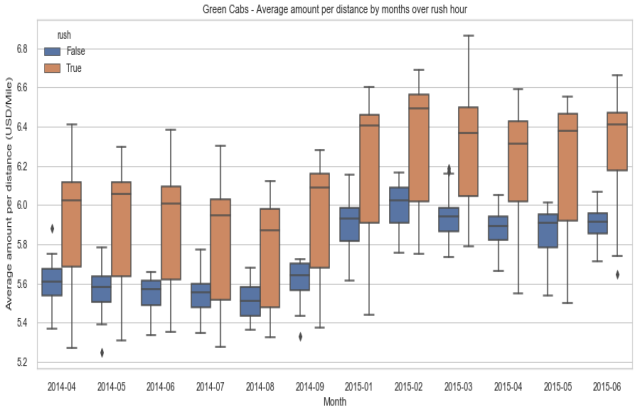
Figure 5: price per distance



(a) Uber trips.



(b) Yellow cab trips



(c) Green cab trips

Figure 6: avrg distance