THE CYPRUS INSTITUTE
RESEARCH•TECHNOLOGY•INNOVATION

New York Taxis Database

SDS403 - Final Assignment

Charalambos Tsioutis

January 22, 2023

1 Introduction

For this report we will use a New York taxis database in order to derive insigns , patterns and trends on the characteristics of taxi rides in New York. The database contains information on the date and time for the pickup and dropoff of the ride, the distance of the ride, the number of passengers, the fare, tip and tolls amount, the color of the taxi, the payment method and also information of the pickup and dropoff location. The database gives information for a total of 6500 rides out of which we have to clean in order to get the best database possible.

Description of the dataset

The *taxis* dataset contains 6433 rows and 14 columns, out of which two are datetype, six are numerical and six are in text form. In order to better analyse the taxis characteristics we create additional columns from the existing data. We created the cost column that gives as the price of the ride without a tip, the duration column that gives us the time of each ride in minutes and the speed of each ride in km per h. We found useful to calculate the day of the week for the ride and thus we created another column day using the pickup time.

Just by looking through the values of our dataset we can already derive some information that can be useful for our analysis. New Yorkers appear to use taxi , according to our dataset, for distances as short as 90 meters and as long as 25 km. Although, passenger do not always tip the driver we can see that tips can reach up to 18 dollars per ride. The maximum duration for the taxi rides was 88 minutes and the most expensive ride was found at 82 dollars.

2 Statistical Analysis

In this section, we will apply some statistical tools to gain further insight into the data, providing a more comprehensive description of the dataset.

We found out that 79 % of the customers used taxis alone and also 65 % of the customers tipped the drivers. The most preferred method of payment appears to be via credit card and also the borough that had the most rides is Manhattan. The day that the taxis had the most rides was Friday while Monday had the least rides.

Additionally we checked if the color of the taxi can affect the amount that customers tip the drivers. First we checked if the average tip for the people that gave a tip had any significant difference for the two colors. We found that the mean value of the tip for each color was 2.58 +/- 0.10 $ for the green taxi and 3.09 +/- 0.04 $ for the yellow taxi. To have a complete picture for the color - tip correlation we checked if customers do not tip a specific color more than the other.

We found out that only 29 % of the people that used yellow taxis did not tip and the percentage for the green taxis was 67 % .

3 Data Visualization

In this section, we will explore the dataset through visualization to gain insights and understand patterns in the New York City taxi dataset that they couldn't be identified through statistical analysis.

To begin, we created a visualization of the cross correlation for our dataset in order to recognize relationships between the attributes (Figure 1). Through the correlation graph we found some strong correlations within our data, distance appears to be correlated with fare, total cost and duration of the ride and also total cost is correlated with the duration of the ride. Tip seems to be correlated with the distance, duration and total cost but only for the passengers that gave tip. The last correlation is not as strong as the other ones.

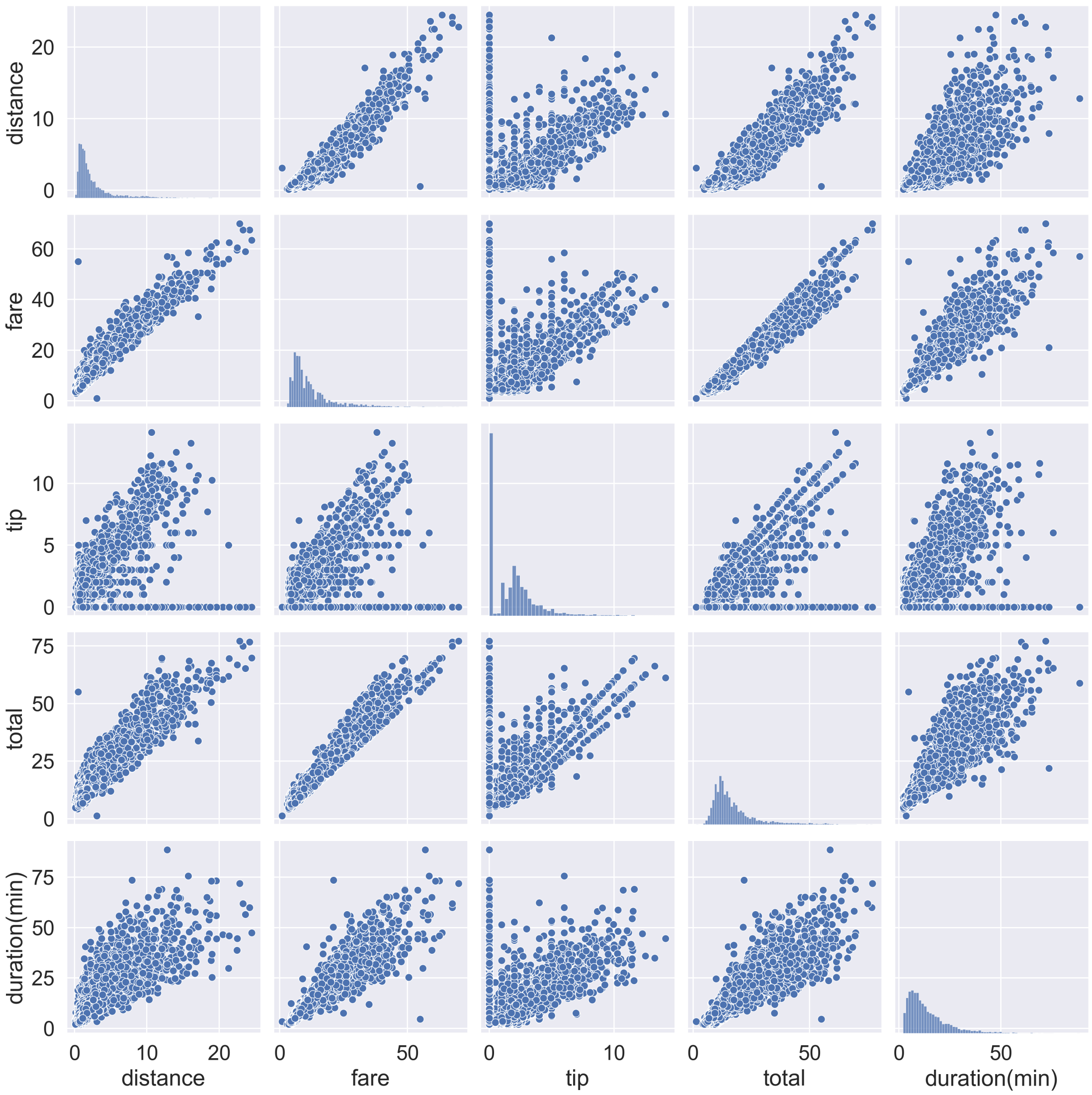

Figure 1: Cross correlation visualization.

Next, we will examine if the day and hour can affect the number of customers for taxis in New York. To achieve that we create histograms that show the number of rides for each day of the week and then, the number of rides per hour for each hour of the day (Figure 2 , Figure 3).

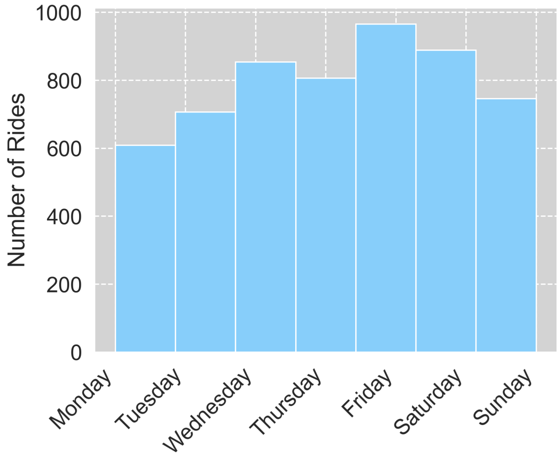

Figure 2: Number of rides for each day of the week.

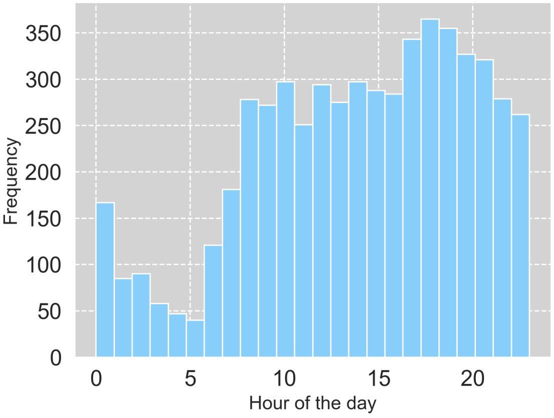

Figure 3: Frequency of rides per hour of the day.

Continuing the report we will examine how tipping can be influenced by different factors. In figure 4 we plot a histogram that shows the frequency of the tips given by borough. For this figure we used only the rides that gave a tip and excluded the non-tip data, showing the amount of the tip that was given more frequently for each borough.

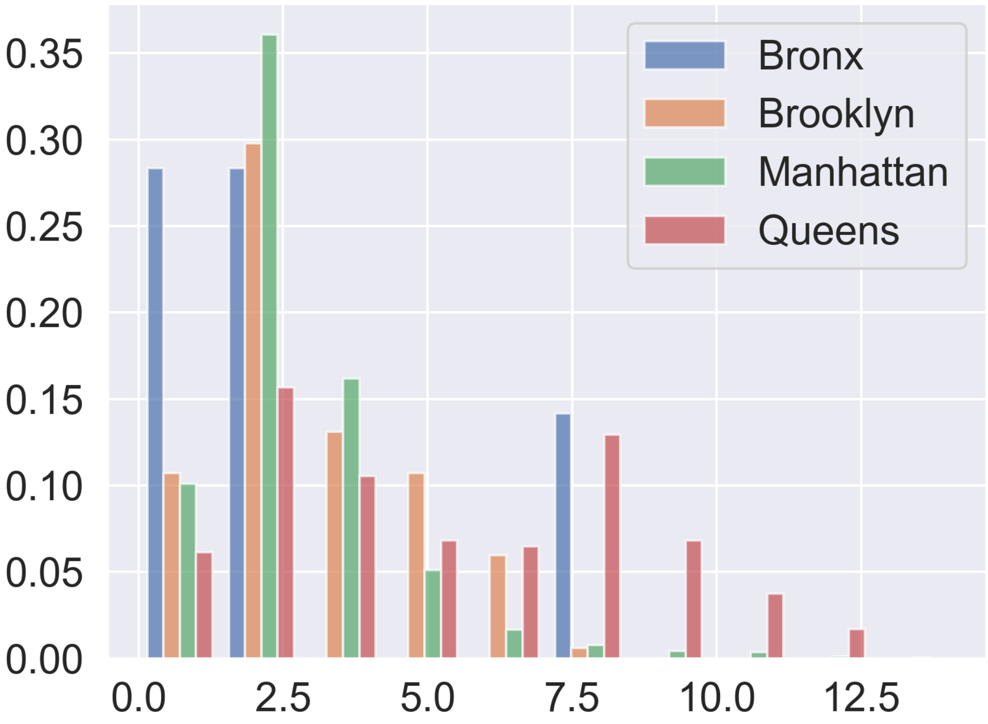

Figure 4: Histogram of the normalized tip amount for each borough.

We also created a scatter plot that shows how the speed of each ride can impact the amount of tipping. We also use only the rides tip was given (Figure 5).

From our data it was evident that many of the customers did not gave any tip. We examined if

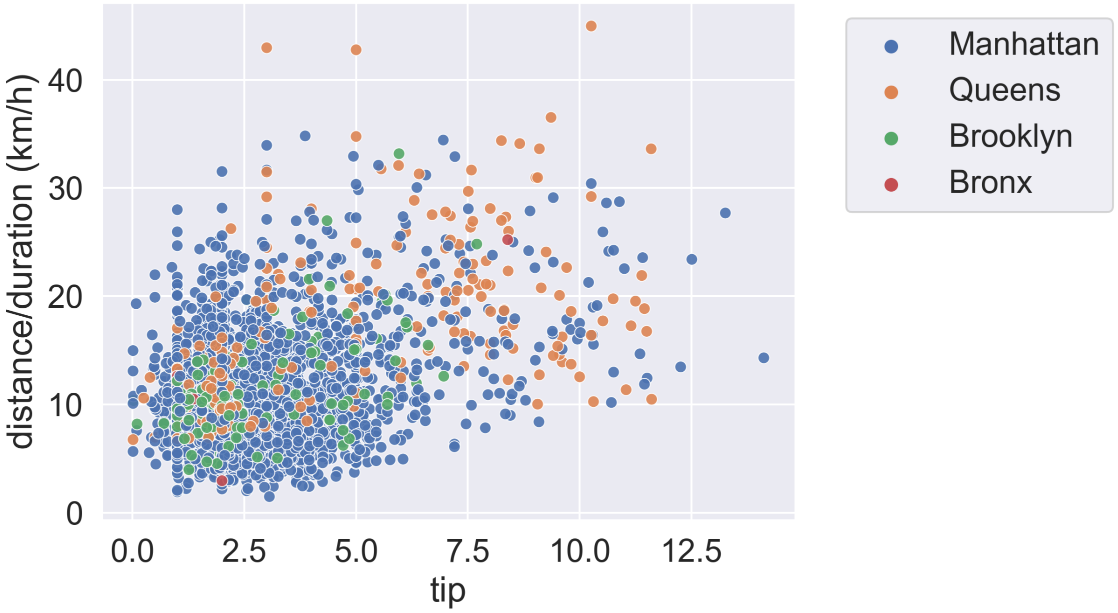

Figure 5: Scatter plot for tip per speed for each ride.

there are specific boroughs that don't give tips more that other by creating a histogram with the percentage of the non-tipped rides (Figure 6).

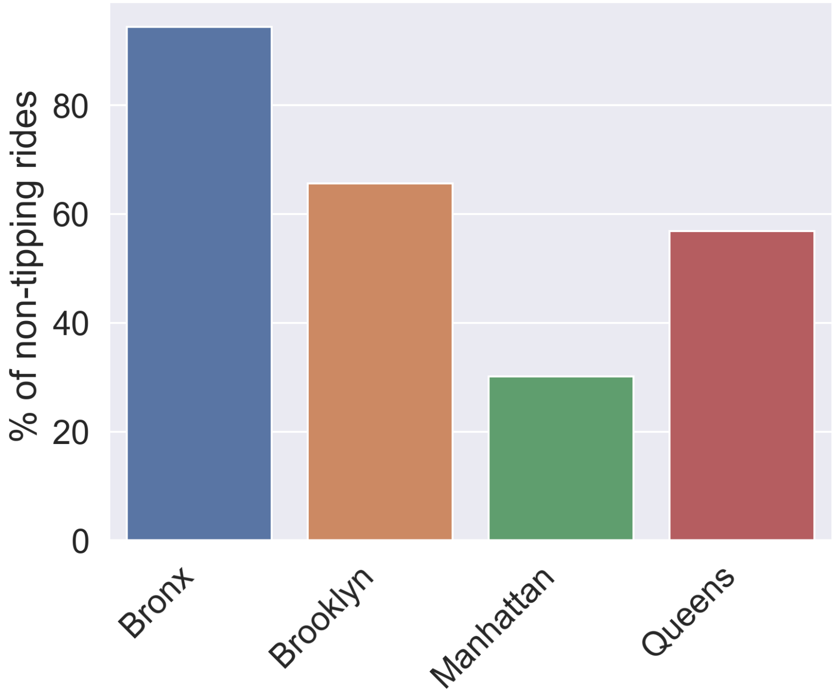

Figure 6: Percentage of the people that do not tip in each borough.

4 Derived observations

Through the analysis of this dataset, we have gained insights into the characteristics of taxi rides in New York city. Firstly, it is clear that the number of rides is highest during the evening hours while the number of rides is lowest during the early morning hours.

Additionally, it is apparent that in the borough of Bronx it's more likely that the taxi driver will not get any tip , while Manhattan is tipping more frequently. Furthermore, the data shows a positive correlation between duration and tipping. This suggests that as the duration of a ride increases, so does the likelihood of a tip. Lastly, the figure 4 indicates that Queens is the most likely borough to give a larger tip, while Bronx gives the least.

In conclusion, this report has provided an overview of the patterns and trends of taxi rides in New York. We can use the information to help businesses make the appropriate planning regarding the number of taxis per day and hour that should be hired but also give recommendations to the drivers, that their outcome relies on tipping, to make the best decisions.