Data Analysis of NYC Cab Services

Abhinandan Dubey
Stony Brook University
New York, USA

Raju Khanal
Stony Brook University
New York, USA

Teja Madiraju
Stony Brook University
New York, USA

[adubey, rkhanal, lmadiraju]@cs.stonybrook.edu

Abstract

Our project aimed at performing an analysis of commuting patterns, neighborhoods, traffic, tipping patterns, taxi fares (and more) in urban communities. The purpose is to extract useful insights so that any solutions that we derive can be mapped to other large cities.

1 Introduction

Analyzing commuting patterns in urban communities has become a key research area lately. Such analysis reveals critical results about a citys transport services. The NYC Taxi and Limousine Commission[1] has been releasing these staggering datasets since 2009, and these datasets hold information about millions of rides in the city every month. The datasets have a huge value and provide a fertile ground for fetching useful insights that may help policy makers and city administration to make effective policies towards curbing traffic and reducing the number of disputes in public transport. We have made an attempt to perform experiments that reveal such insights. More specifically, we have tackled the following questions:

1. Do disputed payments indicate something about the neighbourhood?

2. How does the tipping rate change according to location and/or peak timings?

3. What are the Frequent Pick-Up and Drop-Off points and how they relate to specific times of a year?

4. Do people travel in groups during festive seasons?

5. Do people in affluent neighbourhoods really tip higher?

6. How much should you tip, given a ride, location?

7. How much do you end up paying more for the fare because of traffic, and does it have a specific pattern of increasing over the years?

2 Background

The NYC Taxi and Limousine Commission releases the datasets and also drafts a report annually[2] which includes similar insights as our data but the experiments which we have performed in this project are completely different from those done previously. Todd W. Schneider has also performed a rigorous analysis of the dataset which provides a holistic study of the rise of Uber, and how it has affected the conventional YellowCabs system in NYC, how travel times to JFK and La Guardia have changed over time, and what are the peak hours for the same. Our analysis uses the same dataset, but the experiments we have performed in the project involve a critical investigation of data. Previously, the analysis were to be restricted to a month of data. In our experiments, we have stretched it to years, dealing with upto 72 GB of data to extract insights and find patterns.

3 Data

The New York City Taxi and Limousine Commission has released an astonishing 156 Gigabytes of detailed dataset[3] covering over 1.5 billion taxi trips in the city from January 2009 through June 2016. The dataset holds a variety of features critical to our analysis. The dataset itself is divided into files of around two gigabytes for each month for a single type of vehicle (YellowCab / GreenCab / FHV). The datasets have 19 features which include VendorID, PickupDateTime, DropOffDateTime, PassengerCount, TripDistance, PickupLongitude, PickupLatitude, RateCodeID, Store_FWD_Flag, DropOffLongitude, DropOffLatitude, PaymentType, FareAmount, Extra, MTA_Tax, ImprovementSurcharge, TipAmount, TollsAmount, TotalAmount.

Our observations included extracting insights that help us answer the questions proposed in the Introduction.

4 Methods Used

Our experiments involved using different methods for each task and the same explained below:

4.1 Do disputed payments indicate something about the neighborhood?

Our main goal here is to find if the drop-off location where the disputed payment happened is more frequent in one borough compared to others. The approach involves computing the following parameter over the huge dataset.

$$\frac{\textit{(No of disputed payments in that borough)}}{\textit{(No.of rides/payments made in that borough)}}$$

4.2 How does the tipping rate change according to location and/or peak timings?

Our main goal here is to analyze whether the tipping rates would increase in festive season (December) compared to other months. We calculate the average tip per mile as:

$$\frac{\textit{(Total Tip across all rides in a month)}}{\textit{(Total distance covered across all rides in a month)}}$$

4.3 What are the Frequent Pick-Up and Drop-Off points and how they relate to specific times of a year?

This problem is about finding Frequent Itemsets from the huge dataset. We used the classical apriori approach for finding the Frequent Itemsets ($size = 2$). Here we are considering only pickup and drop-off points as our items. The support of the itemsets is not mentioned currently in the code so as to visualize the entire boroughs thoroughly. However we can easily add support to remove the boroughs which have less frequency of pickup and drops(like the Staten Island pairs).

4.4 Do people travel in groups during festive seasons?

This experiments involves performing analysis which would allow us to extract information from `passenger_count` feature in our dataset. For festive season, were considering the entire week of Thanksgiving and Christmas. We followed two approaches to get the same.

1. *Get the average number of passengers in the taxi rides over a sequence of time during festive season.*
A naive approach to answering this method is perhaps taking the average of the number of passengers in the rides during the festive week, and contrasting it to the average number of passengers in the rides during a typical non-festive week. However, as the results below show, this technique fails to provide any useful insights about the same.

2. *Get the percentage of* `k-passenger` *($k > 1$) rides during a festive season and contrasting it with the percentage of* `1-passenger` *rides.*
This approach reveals quite meaningful results which have been discussed in the results section.

4.5 Do people in affluent neighbourhoods really tip higher?

For finding the affluent neighbors we applied a filter to find the tips with amount greater than 150 dollars per ride. After filtering, we calculated the tip per miles, by grouping all the rides with same dropping-location[1]. Since the data for Staten Island was much less, it might not be present in the final results (pie charts). We are considering affluency in this case by considering the `dropoff-location`. Another study could be to find the affluency by the pickup location and/or taking the average across all data without applying any filteration.

4.6 How much should you tip, given a ride, location?

In this experiment, we built an analogy between cab rides and user-user collaborative filtering techniques to suggest as a recommendation system's output, the amount of tip a passenger may pay for a ride.

To do this, the pickup and drop points were reverse-mapped to different boroughs of New York and a tuple of `(pickup, drop)` locations is treated as a user. The total fare is treated as an item for that user and the rating was analogized to the amount of tip for each of these rides. We considered 5 boroughs, which in turn generated 30 different possible users. All fares were rounded to the nearest integer and duplicate records were deleted. We then performed user-user collaborative filtering on this data.

To calculate the similarities between users, cosine distance was used. Considering the large size of the input data set and the small size of the user set, items were filtered on the basis of more than 1000 occurences.

The missing tip amounts were calculated using the formula

$$tip_{user} = \frac{\sum_{i \in U} sim(user, i) + tip_a mount(i)}{\sum_{i \in U} sim(user, i)}$$

4.7 How much do you end up paying more for the fare because of traffic, and does it have a specific pattern of increasing over the years?

This experiment involved using the latest libraries and some state-of-the-art techniques to perform time-series analysis. Taking into account the features like `pickup-datetime`, `dropoff-datetime`, `trip_distance` and `fare_amount`, we used the power of advanced data structures provided by **`spark-ts`**, a library developed by Clouderas Data Science team (and in use by customers) [5] that enables analysis of data sets comprising millions of time series, each with millions of measurements. The package runs atop Apache Spark, and exposes Scala and Python APIs. We had to build the spark-ts package over Maven to get started. We transformed our data into the time-series DataFrame. A time series is generally a sequence of floating-point values, each linked to a particular point in time. The size of the data doesnt allow us to use multiple features at once along with the timeseries, and hence we had to hit a bound of using a univariate model. In Python, a 1-D NumPy array denotes timeseries for a particular index, and has a **`DateTimeIndex`** somewhere nearby to link its values to points in time. An instant is the vector of values in a collection of time series corresponding to a single point in time. In spark-ts, each time series is generally labeled with a unique identification key to help us to identify it uniquely as a tuple. An observation is a tuple of **`(timestamp, key, value)`**, i.e. a single value in a time series or instant. As per spark-ts parlance, we created an observations dataframe layout and used it as out data structure for dealing with the series. We create a **`timeseriesRDD`** as required to perform map or filter operations. After getting the per mile fare rate macro-averaged over a month, we use the pandas library to perform time-series analysis. We assumed an *autoregressive* **`AR-1`** model.

$$y_t = \rho y_{t-1} + u_t$$

This involves performing the famous ***Dickey-Fuller Test***, and finding out stationarity, mean, deviations of the series from 2013 to 2015. We generate curves which show how much people end up paying than the average estimated per mile rate.

The notation **`AR(p)`** indicates an autoregressive model of order p. The AR(p) model is defined as

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

where $\varphi_1, \ldots, \varphi_p$ are the parameters of the model, c is a constant, and ε_t is white noise. In our case, the we had to assume a single parameter, `average_per_mile_rate`.

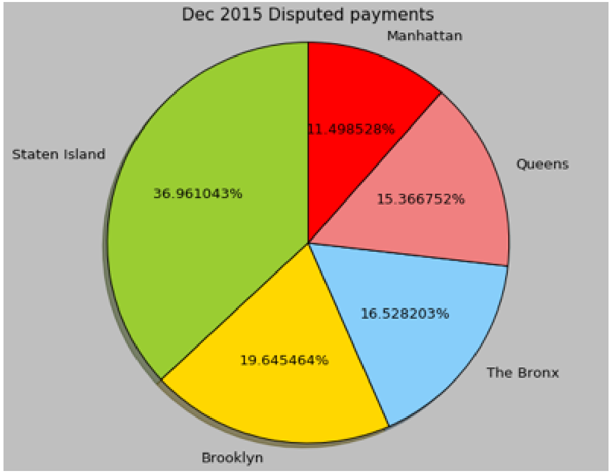
[1] By same dropping location, we mean the same borough. For example, Manhattan or Brooklyn

5 Results

The following are the brief results of our experiments.

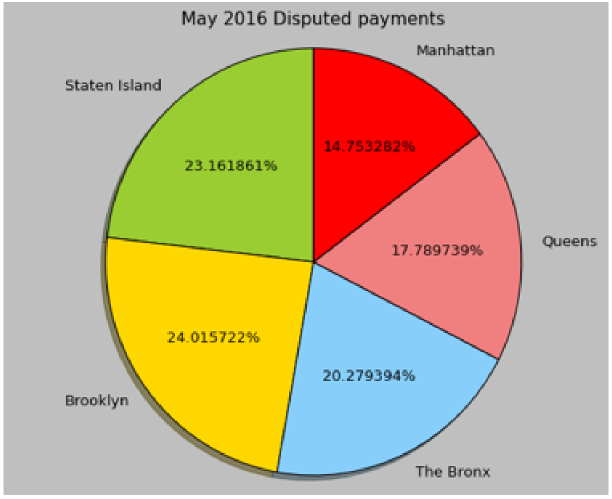
5.1 Do disputed payments indicate something about the neighbourhood?

Figure 1: Disputed Payments by Boroughs in Dec 2015



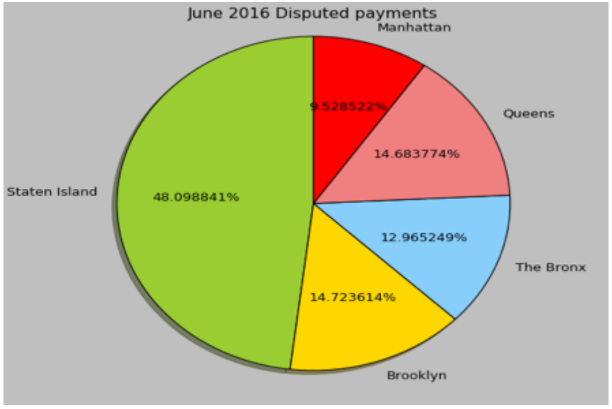
As can be seen in Figure 1, 2, and 3, all 3 datasets seem to suggest that the number of disputed payments in Staten Island are more than most(all). Also Manhattan has the least number of disputed payments compared to any other borough.

Figure 2: Disputed Payments by Boroughs in Dec 2015



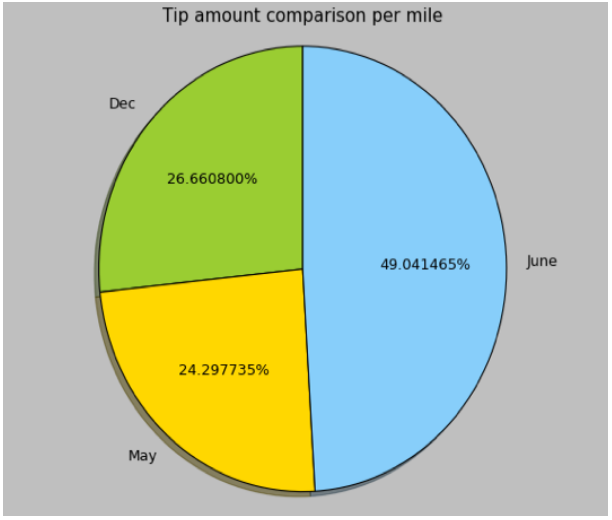
Figures 2 and 3 show the same statistics for May and June 2016. *Question : Is Manhattan a more safer borough to live than Staten Island or Brooklyn?*

Figure 3: Disputed Payments by Boroughs in Dec 2015



5.2 How does the tipping rate change according to location and/or peak timings?

Figure 4: Tipping Rates



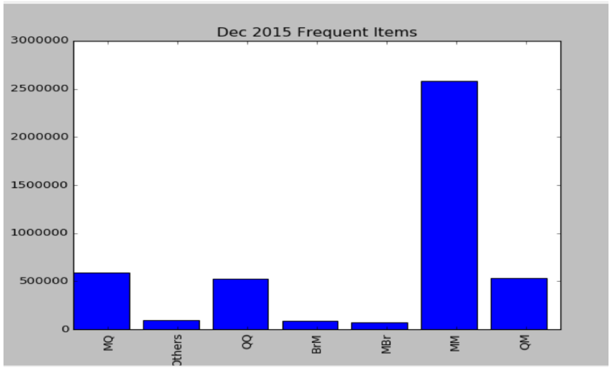
Although the tip amount for December does indicate that to be true compared to May, the statistics show contrasting results for the month of June.

Question: Was our assumption wrong? Is there no particular relation between months and the tipping rates? Why does June have such a big difference? Was there some musical events or other shows in that month that might have prompted people to give more tips?

5.3 What are the Frequent Pick-Up and Drop-Off points and how they relate to specific times of a year?

As can be seen from the bar graph, Manhattan to Manhattan(indicated by MM) has the most frequent

Figure 5: Disputed Payments by Boroughs in Dec 2015



pickup and drop-off locations followed by Manhattan to Queens. As can be seen, the Others category are the rides from boroughs other than the one mentioned above. So it can be concluded that the yellow cabs rides from Staten Island to Manhattan or any other borough is quite less compared to MM.

5.4 Do people travel in groups during festive seasons ?

We analyzed two different American holidays: Christmas and Thanksgiving.
Average number of people in a cab was 0.49% higher than the normal during the festive time than the normal. A closer analysis, following approach 2 as mentioned in the Methods section, however reveals that the number of rides with 2 or more passengers increased by 3.88 % during these holidays.

5.5 Do people in affluent neighbourhoods really tip higher?

We analyzed the tipping patterns as mentioned in Section 4. The results are shown in Figure 6,7, and 8.

5.6 How much should you tip, given a ride, location?

Our recommendation system was able to achieve pretty good results for different ”test rides”. Below is a table (Figure 9) of a few tip amounts that were calculated using item-item collaborative filtering described in the previous section.

Figure 6: Tip Affluence projected in May 2016

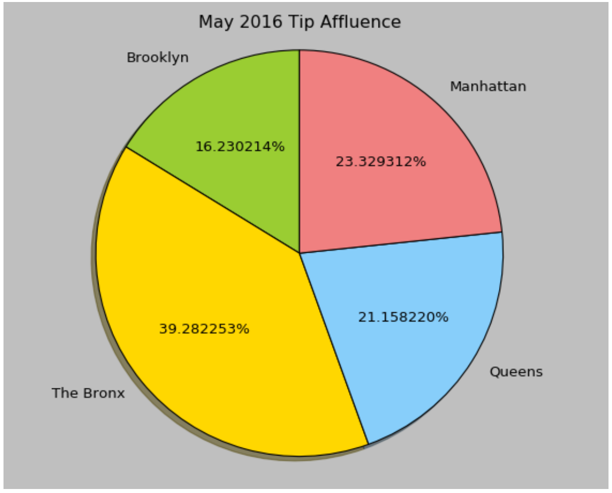
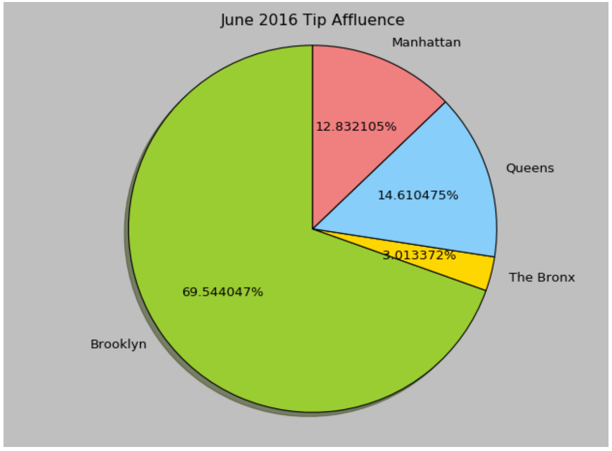


Figure 7: Tip Affluence projected in June 2016



5.7 How much do you end up paying more for the fare because of traffic, and does it have a specific pattern of increasing over the years?

We achieved quite significant results which show an increasing graph of the per-mile fare that passengers actually end-up paying because of traffic. This is shown in the figure.

Also, the Dickey-Fuller test statistic is more than the 10% critical value, thus the TS is not stationary with 90% confidence[2] As evident from the graph too, the fares show a fairly increasing series over the years. This shows the increasing traffic causing passengers to pay more for a ride than the usual.

[2] The Dickey-Fuller test we perform here assumes an AR-1 model, results might vary slightly for other models

Figure 8: Tip Affluence projected in Dec 2015

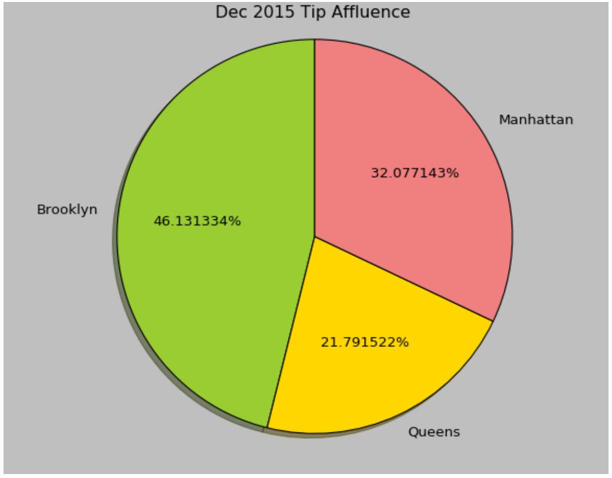
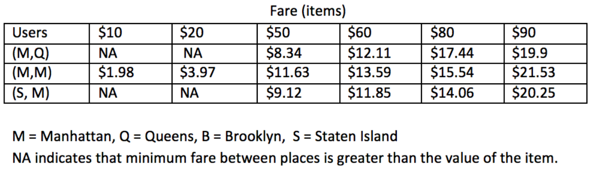


Figure 9: Results for Item-Item Collaborative Filtering



Fare (items)

Users	$10	$20	$50	$60	$80	$90
(M,Q)	NA	NA	$8.34	$12.11	$17.44	$19.9
(M,M)	$1.98	$3.97	$11.63	$13.59	$15.54	$21.53
(S, M)	NA	NA	$9.12	$11.85	$14.06	$20.25

M = Manhattan, Q = Queens, B = Brooklyn, S = Staten Island
NA indicates that minimum fare between places is greater than the value of the item.

6 Discussion

Our experiments revealed some interesting facts about the urban communities and their commuting patterns. The experiments show that with increasing population, the average trip times will increase and hence people will end up paying more for a ride and will end up wasting more time on commute than usual. The insights which map affluency of locations also indicate that tipping patterns vary according to areas.

From the results, we also conclude that at lower fare people tip lower than the average tip (expected of 20%). As the fare keeps increasing the tipped amount crosses widely above 20%. This can be attributed to the fact that fare amount encompasses both distance and time spent in the cab. This leaning of the larger tips to the right of the 20% curve is indicative of the tip's dependency on the time spent and less on the distance.

Figure 10: Time-series analysis for per-mile fare

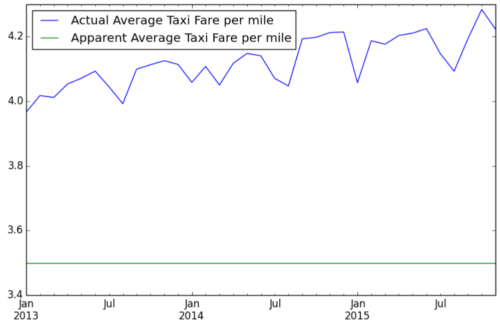
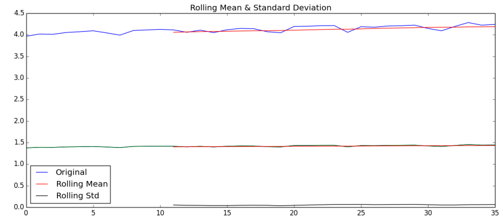


Figure 11: Rolling mean and standard deviations



Overall, our analysis extends the previous work to a new level. Some results conform to our assumptions but others differ or deviate.

Metric	Value
Test Statistic	-0.927128
p-value	0.778907
#Lags Used	6.000000
Critical Value (5%)	-2.967882
Critical Value (1%)	-3.679060
Critical Value (10%)	-2.623158

Table 1: Results of The Dickey-Fuller Test: