Big Data Analytics Project

Spring 2017 Master

NYC Yellow Cab Analysis during Nights under the Impact of Uber

Team: F4

Team Member: Chuanze Cai, Di Zhu, Shaoying Tang, Zhe Zhang

Contents

[Abstract 3](#_Toc480920550)

[Introduction 4](#_Toc480920551)

[Data 5](#_Toc480920552)

[Conceptional Model and Analytics&Methodology 7](#_Toc480920553)

[Result 8](#_Toc480920554)

[Managerial Implications and Future Research 15](#_Toc480920555)

[Sample data for analyzing 17](#_Toc480920556)

[References 17](#_Toc480920557)

# 

# Abstract

The main subject of this project is to discover the future of the taxicabs under the critical impact of shared economic represented by Uber and Lyft in transportation services in most recent years. Regarding the fact that New York city subway is the largest rapid transit system in the world by the number of stations which delivered over 5.7 million rides daily and 5.9 million rides over the weekend in 2015, the study of yellow cabs will be limited to midnights when subways are running on late-nights schedule and fewer people are using subways as the first choice. After restricting the time frame of the study within the midnights, Uber is expected to be the major competitor to the yellow cabs.

Based on the yellow cabs data collected from April to September in 2014 including information such as pickup and drop-off locations, date, and fare, the team will be able to find the busiest pickup and drop-off locations during nights and compare the service area and reviews of yellow cabs and Uber within the same period. According to the result, recommendations will be made to Yellow cab companies to ensure the profit from service areas with absolute advantages, maximize the business opportunities within the competition and improve the service quality based on the reviews in order to differentiate the business from Uber.

Introduction

Taxicab, one of the most frequently used transportation services, provides a convenient, fast, private however expensive way for individuals to travel around the city, which can critically reduce the use of private vehicles and thus eliminate the traffic and pollution to economize energy resources and benefit the environment. Compared to mass public transit like metros and buses, taxicab allows customers to customize pickup and drop-off locations without unnecessary stops based on their own requests. Modern taxicabs are originated from horse-drawn for-hire hackney carriage services in Paris and London in 17th century at the time when people were traveling using horse-drawn carriages. In the history, the first electric battery-powered taxi appeared in 1897 in London and later in the same year, 12 electric hansom cabs started running in New York City by Samuel’s Electric Carriage and Wagon Company. Nowadays, taxicabs are recognized as an icon of New York City with two distinct varieties: yellow cabs, sometimes recognized as medallion taxis and street hail livery vehicles in green. According to the regulation of the authorities, yellow cabs are allowed in all 5 boroughs in New York City while green cabs are excluded from lower Manhattan or either airport.

However, the rise of the transportation network companies especially the competition from rideshare services such as Uber and Lyft from 2009 dramatically affected the business of Taxicabs. The new types of transportation methods generated from the prevalence of the smartphones and internet services in the most recent years provides the public relatively cheaper and more economic ways to travel around the city. According to New York City Taxi and Limousine Commission, by July 2016 the estimated total number of taxicab has dropped to 13,587 from 13,605 in March 2014 and the meter revenue has declined by over 18% from 2013 to 2015.

# Data

Firstly, the dataset is collected from Google BigQuery public datasets. The dataset is called nyc- tlc: yellow trip (<https://bigquery.cloud.google.com/dataset/nyc-tlc:yellow>). It includes 19 variables: vendor\_id, pickup\_datetime, dropoff\_datetime, pickup\_longtitude, pickup\_latitude, dropoff\_longtitude, dropoff\_latitude, rate\_code, passenger\_count, trip\_distance, payment\_type, fare\_amount, extra, mta\_tax, imp\_surcharge, tip\_amount, tolls\_amount, total\_amount and store\_and\_fwd\_flag and more than 1 trillion rows. So as you can see, the dataset is super big, which actually costs 130 GB storage size. The dataset records NYC Taxis picking up and dropping off passengers from the first day of 2009 to June 30, 2015. The trend of taxis business can also be described from the time span of 6.5 years.

Secondly, data preprocessing has to be adopted. In order to detect abnormal data, queries with descending and ascending order were implemented. Few extremely abnormal records of tips were detected as below. The suggestion is to delete the rows with tips which were above 100 percent of the fare amount and which were negative. Feature selection will also be implemented. After observing the dataset, all variables are supposed to be valuable if put into analysis except the vendor\_id. So probably no variable will be abandoned, and different selection of variables will depend on different sub-analysis. The negative tips and abnormally high tips were deleted from the dataset. Then we use SQL to select the records we are interested in. For this project, we aimed to make night analysis.

SELECT pickup\_datetime, pickup\_longitude, pickup\_latitude  
FROM [nyc-tlc:yellow.trips]  
WHERE TIME(pickup\_datetime) > '00:00:00' AND TIME(pickup\_datetime) < '06:00:00'   
ORDER BY pickup\_datetime

So far, we get our completed dataset.

Thirdly, hypothesis should be listed. By now only four hypotheses are agreed by the team based on experience in real life.

|  |  |  |
| --- | --- | --- |
|  | **Hypothesis** | **Rationale** |
| 1 | The business for taxis were going down a little bit. | Uber, the most popular transportation mobile app emerged since 2009, offering more economic and sometimes more convenient picking up customer’s service. |
| 2 | Passengers are picked up more frequently in the mid-town than in the upper town. | There are more skyscrapers and white collars in mid-town than in the upper town. The white collars usually needs Taxis more for emergency in work. |
| 3 | Passengers picked up before 6 a.m. have a greater chance of demanding long-distance rides. | More active people during midnights are expected on business trip or travel to the airports. |

Table 1 Hypothesis and Rationale

# Conceptional Model and Analytics & Methodology

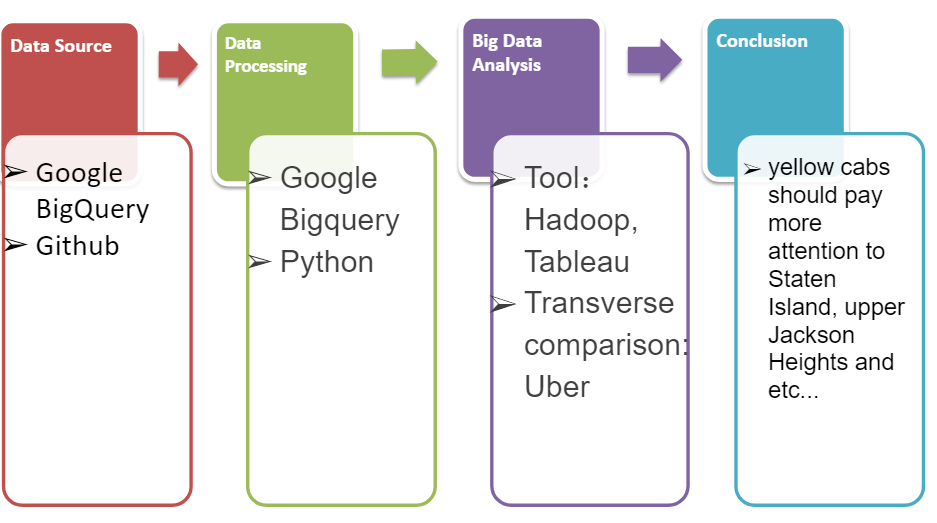


Figure 1:Work Flow

Then the descriptive analysis assisted by visualization will be implemented. After this, the team is going to dig deeper into the dataset. We will try predictive analysis. The planned models for the analysis will be explained in the following context.

Considering the data size, the major tools for the project are Python and Google Bigquery. For the visualization function, Tableau, which is able to display very beautiful charts, was also leveraged. The domain of the project is basically the combination of business and data analytics. Since the project is set to get a conclusion that is easy to comprehend and execute, we need to make the conclusion jargon free and imply more visualized elements such as charts and tables.

Based on the dataset, two kinds of analysis are primarily applied. The first one is descriptive analysis, the pickup locations can be mapped around New York City. By this way the first two hypotheses are easily to be examined. Predictive analysis can find the important factors of the customers bringing in most revenue to the taxi company and predict whether the designated types of passengers are valuable to the business.

For the result part, visualization charts are generated with different layers including gross income per capita to convey a subtle correlation between pickups and local economy developments. Plus, in order to learn more detailed situations of pickups of yellow cabs and Uber separately, we leverage python to organize and categorize pick up coordinators and visualized the ‘sweet spots’ for each company. With digging deeper into the dataset, we can figure out some trends including popular pickup location as well as frequency which can be told from the map created. Considering the goal of the research, reviews about both companies were also scraped and analyzed.

Result

From the analysis process before we can get some descriptive conclusions and visualizing results that are listed below.

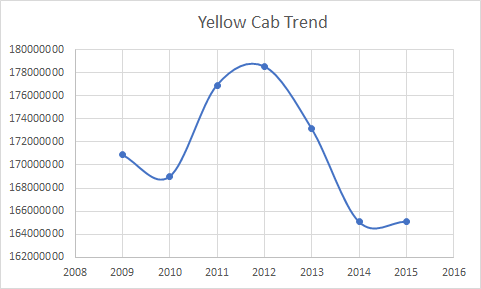


Figure 2:Yellow Cab Performance Trendline

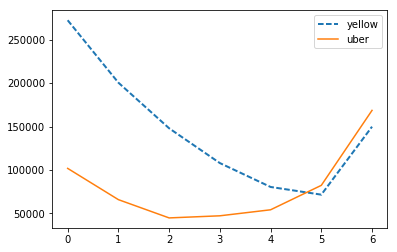


Figure 3:Comparison of Yellow Cab and Uber pick-up times per month

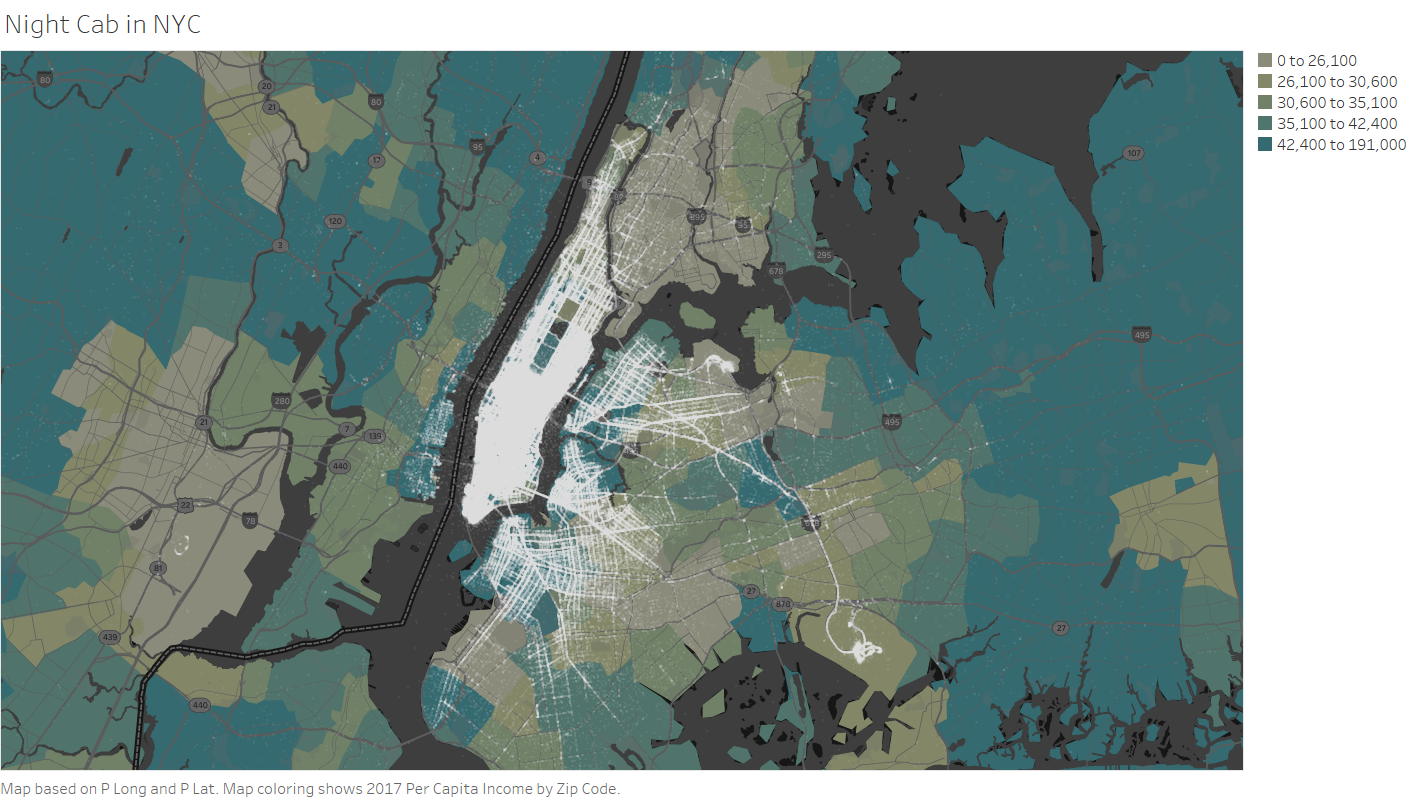


Figure 4:Yellow Cab Service area by picking up location

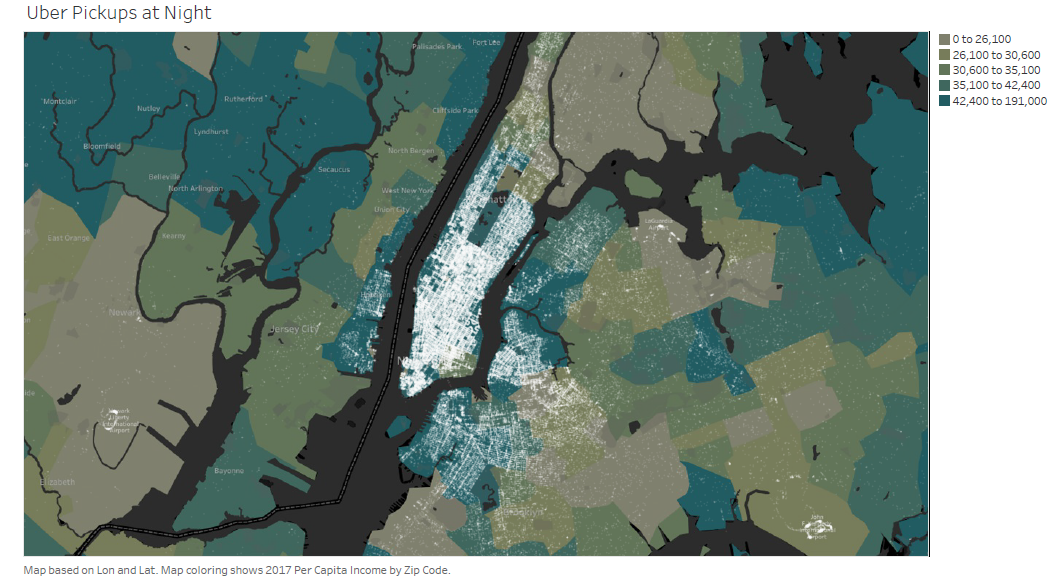


Figure 5:Uber Service area by picking up location

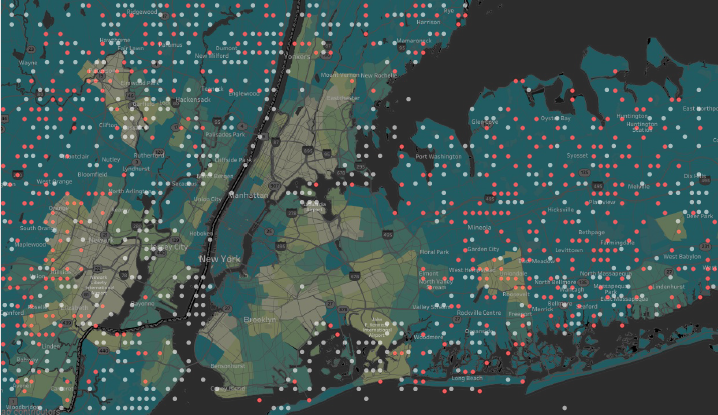


Figure 6:Unique services areas (light blue for Yellow and red for Uber)

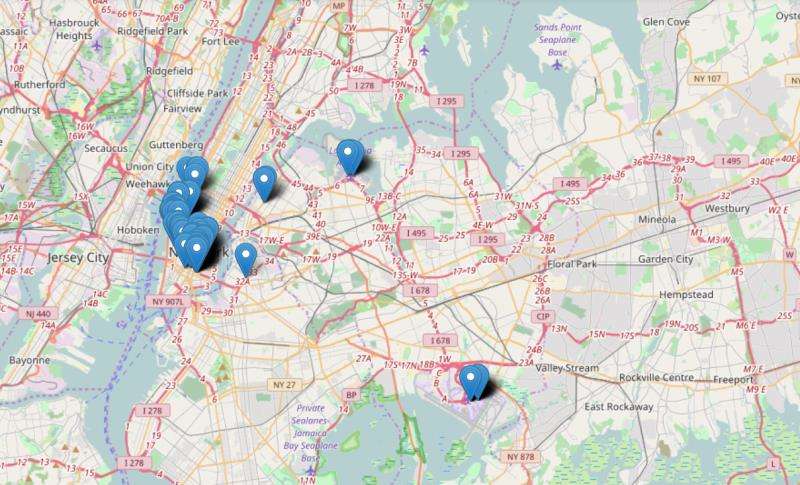


Figure 7:Yellow Cab Apr 2014 - Sept 2014 Top 200 Pickup Locations in NYC

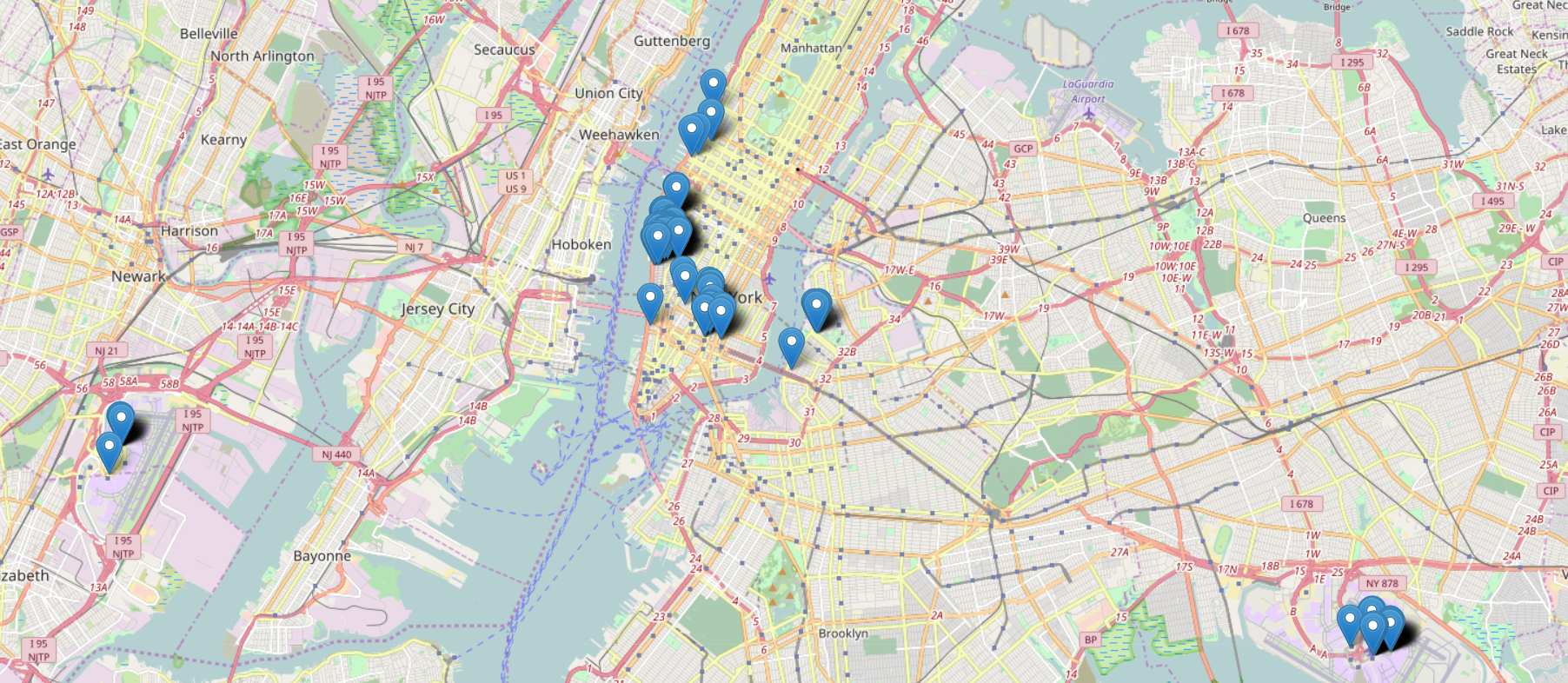


Figure 8:Uber Apr 2014 - Sept 2014 Top 200 Pickup Locations in NYC

By counting the number of rides for yellow cabs during nights, the time series plots shows that the business dramatically decreased from over 178 million to less than 166 million in 2012-2014 with the rise of Uber usage. Overall, rides occur more frequently in high income areas and both yellow cabs and Ubers are busier around the area with more restaurants, bars, and clubs in addition to the airports during nights. According to the overview maps, the white dots density representing pick up locations for Uber are lower than the yellow cabs resulted from fewer business. And Newark airport is a significant place for Uber drivers instead of LaGuardia Airport for yellow cabs. Moreover, the overview for yellow cabs shows that the pickup locations are more spread out and the pickup can happens anywhere along the streets while Uber tends to pick up customer at certain locations. By removing the last digit of the longitude and latitude, the pickup locations will be grouped by 0.001 degrees. The comparison of the pickup locations will illustrate the unique business for yellow cabs where rides never made by the Uber customers. Ubers are not generally used in Staten island, Bayonne, upper Jackson Heights, Hunts point, and sheepshead bay by the customers as many as the yellow cabs.

Also, with using Semantria, we can get the theme volume by the sentiment of yellow cab together with the theme volume by sentiment of Uber. Based on the result gained from below, we can learn the interesting fact that both customers of Uber and yellow cabs mentions yellow cabs most. However, customers of Uber comment better than those who take yellow cab for a ride.

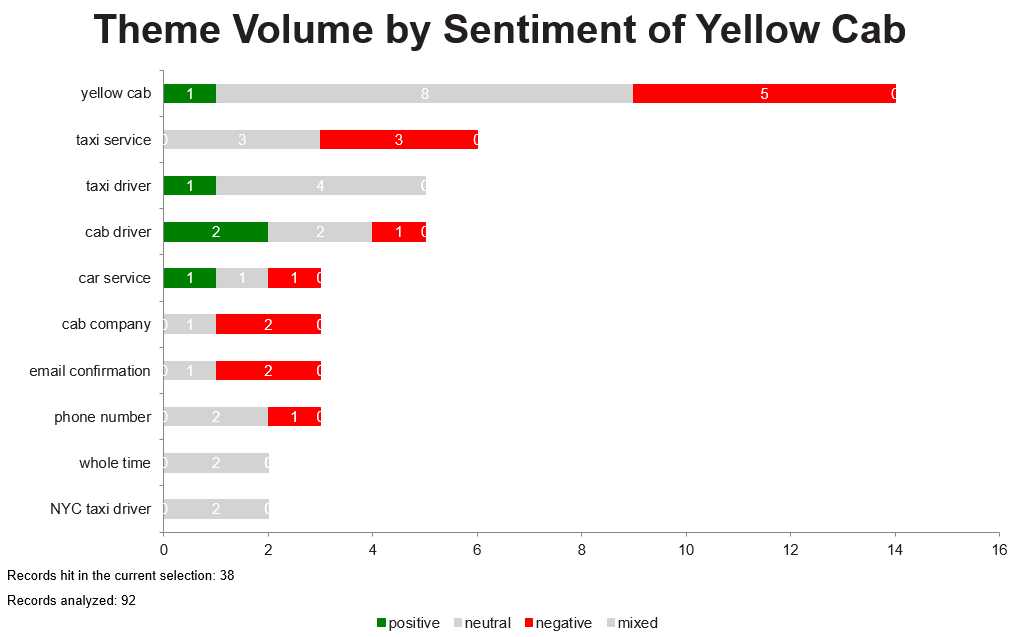


Figure 9:The me volume by sentiment for yellow cab

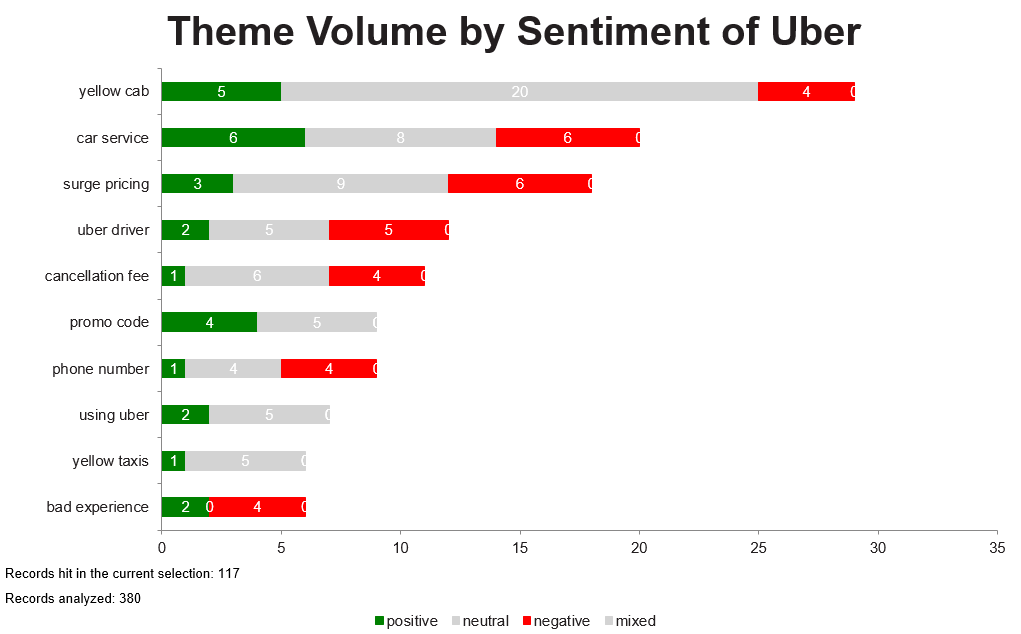


Figure 10:Theme volume by sentiment for Uber

Also, we analyzed the count of different themes according to positive, negative and neutral sentiment. Based on the table we can learn that the total general count of those three categories can gain a inner insight about the satisfaction of customers.

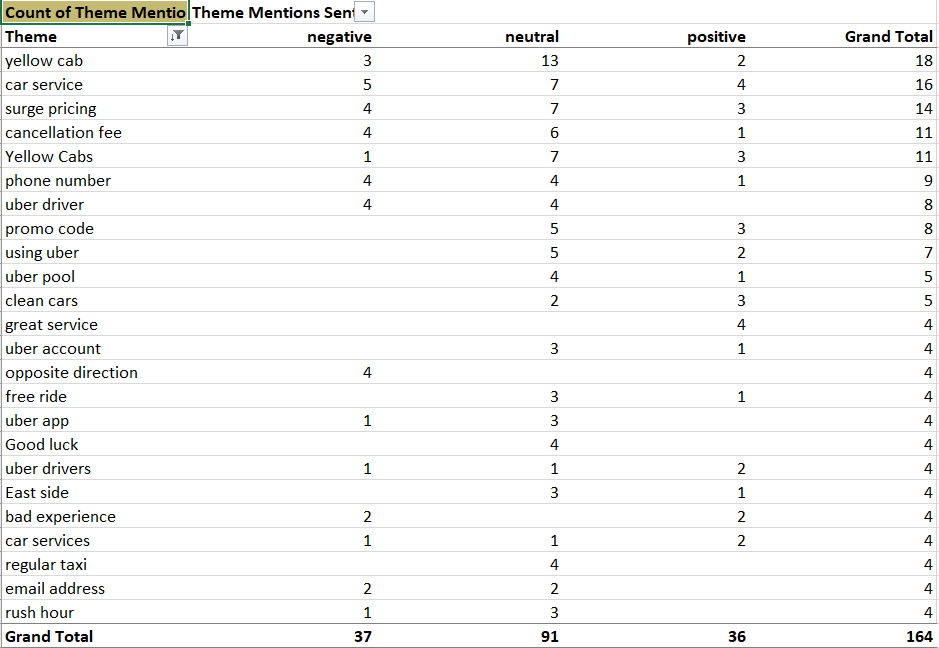


Figure 11:Count of Theme mention in yellow cab reviews

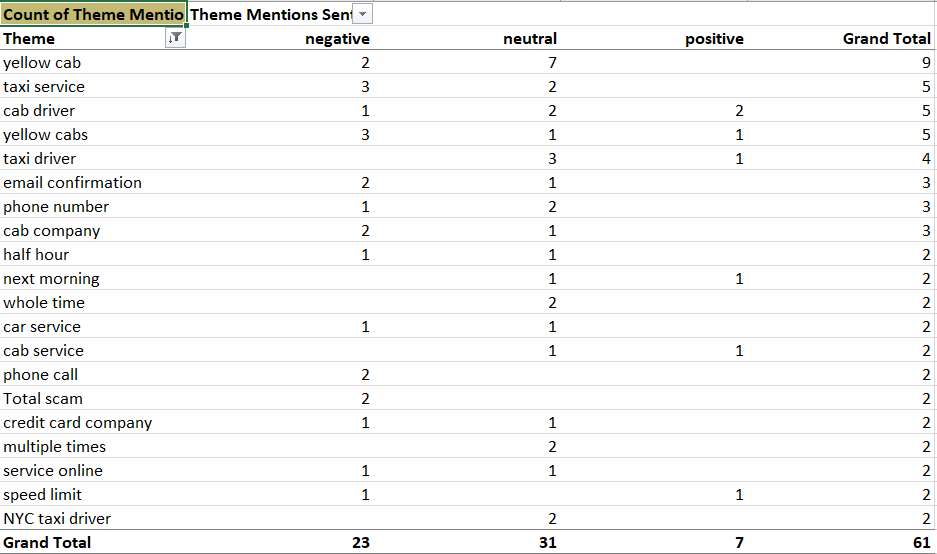


Figure 12:Count of Theme mention in Uber reviews

We use Hadoop to count the word frequency from the reviews gotten from Yelp. From the output, it shows that the customers mention a lot about the locations. The locations which appear a lot are popular spots from the perspective of customers. Hadoop is very quick and useful to process large dataset.



Figure 13:Word Count of yellow cab reviews

We also do word count on the reviews of Uber. In those reviews, customers mention a lot about price and ride type (for example, Uber pool, Uber X, etc.). It seems that price and ride type are the things customers mostly care about. Uber is more novel so the customers’ focus point is more diversified.

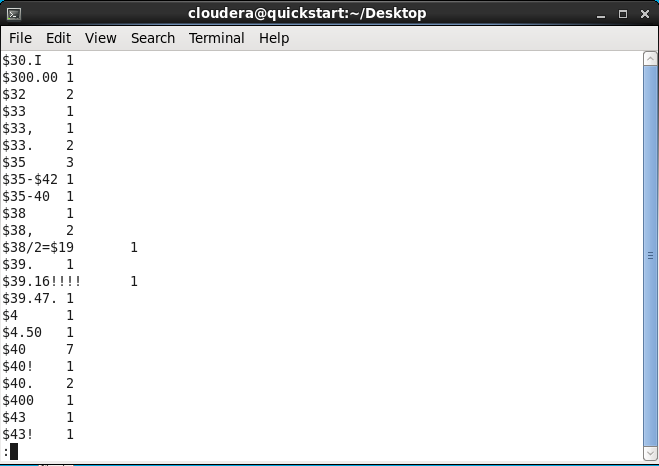


Figure 14:Word Count of Uber reviews



Figure 15:Word Count of Uber reviews (continued)

# Managerial Implications and Future Research

Based on the previous analysis, in grand conclusion Uber does impact the business of yellow cab in New York City, not only seen in the significantly declining trend of the amount of pick-ups since Uber emerged, but also implied in the frequently mentioned topic in the reviews of Uber. By analyzing the performance of Yellow cab and Uber in ways of comparing the popularity, service areas, sentiment analysis of customer reviews on yelp.com, there are conclusions as well as suggestions for Yellow cab.

1. Yellow cab still has significant advantage over Uber in late night’s transportation service even though in some news Uber has outnumbered Yellow cab business (http://ny.curbed.com/2017/1/17/14296892/yellow-taxi-nyc-uber-lyft-via-numbers).
2. Yellow cab should maintain its advantage on the slightly more percentage of positive sentiment in feedback of drivers, while it should improve its confirmation service and find ways to compete with the price advantage of Uber like promo code.
3. In order to enlarge the market share of yellow cabs, Yellow cab company should pay more attention to places where Uber cannot be easily accessed such as Staten Island, and upper Jackson Heights.

In future research, since Uber’s pick up locations are set by the customer before the ride and cab’s pick up the place customers generally determine locations notice an available cab, latitude and longitude can be evaluated along with the locations of major neighborhoods, hotels or shopping malls where Uber rides are expected with a higher probability to be requested from. In addition, with more detailed information about the profitability of the rides such as fare rate and drop-off locations for both Uber and yellow cabs, we can possibly find the trend of industry and figure out the pros and cons of taking cabs for citizens, better decision may be made to help yellow cab differentiate the business from Uber and take an advantage of the services.

# Sample data for analyzing

## References

[**https://en.wikipedia.org/wiki/Taxicabs\_of\_New\_York\_City**](https://en.wikipedia.org/wiki/Taxicabs_of_New_York_City)

[**http://www.yellowcabnyctaxi.com/**](https://www.yellowcabnyctaxi.com/)

[**http://ny.curbed.com/2017/1/17/14296892/yellow-taxi-nyc-uber-lyft-via-numbers**](http://ny.curbed.com/2017/1/17/14296892/yellow-taxi-nyc-uber-lyft-via-numbers)