Implementing Big Data Methods to Analyze NYC Yellow Taxis

José M. Peña Marmolejos
Graduate School of Arts and Sciences, Fordham University
113 W 60th Street New York
jpenamarmolejos@fordham.edu

Luis A. Martínez-López
Graduate School of Arts and Sciences, Fordham University
113 W 60th Street New York
lmartinezlopez@fordham.edu

Sammy Ahmed
Graduate School of Arts and Sciences, Fordham University
113 W 60th Street New York
sahmed84@fordham.edu

ABSTRACT

New York City residents rely heavily on many different modes of transportation outside of personal vehicles. There is little time to waste in NYC, so no matter the mode of transportation, it needs to be quick and efficient. We will examine yellow NYC taxi data from 2017 to further understand the demand for taxis in NYC. Performing big data techniques on this dataset allows for future business strategies to be planned, which will not only help NYC taxis produce more revenue, but also help consumers receive faster and more efficient service. Our analysis of this dataset will be focused on a few key points of interest: time and location, region, month, and time and day of the week. Each of the following key points will entail multiple specific analyses, which result in potential business solutions. Ultimately, the analyses performed will be mapped to various types of visualization. This allows for easier interpretations of the results, rather than parsing through complex data files. The New York Taxicab and Limousine Commission dataset is a publicly available dataset offered by NYC.gov. Each record contains details of a single cab ride. We worked exclusively on yellow cab data.

Keywords

Big Data; Yellow Taxi; Spark; New York City

1. INTRODUCTION

The capacity of transportation contributes to the development and efficiency of a society. Generally, 10 to 20 percent of national economics is linked to transport. The taxicabs of New York City are widely recognized icons of the city and come in two varieties: yellow and green. Taxis painted canary yellow (medallion taxis) can pick up passengers anywhere in the five boroughs. Those painted apple green (street hail livery vehicles, commonly known as "boro taxis"), which began to appear in August 2013, are allowed to pick up passengers in Upper Manhattan, Bronx, Brooklyn, Queens (excluding LaGuardia Airport and John F. Kennedy International Airport), and Staten Island. Both types have the same fare structure. Taxicabs are operated by private companies and licensed by the New York City Taxi and Limousine Commission (TLC) [1]. In this study we take an in depth look at the yellow taxis and analyze the data to understand the behavior of the taxis as well as the customers. We have been able to achieve the proposed goals by utilizing the Spark technology in conjunction with Google's Cloud platform for processing almost 10GB of daily generated data from 2017. Due to the amount of data, it can be considered as an example of "Big Data". We explore the data in a "Drill Down" manner, analyzing as the city in general, as boroughs, and based on time.

The core objective of our project is to understand this public service that has been available for so long and offers an enormous public database of New Yorkers patterns regarding taxi transportation. Our analysis will be focus in three big pillars: 1) Analysis on trips overall, 2) Analysis based on time and location. And 3) Analysis based on fare/payment. The analysis on trips overall covers the number of trips in 2017, highlighting efficiency and the percentage of service providers to the taxi meters. It will also report on averages of passenger count, tips and total amount paid of all the trips. In addition, we calculate the total trips per month and hour. The analysis based on time and location focuses on common pickup and drop off

locations per borough and time. It is important to note that the taxi data for 2017 did not register the latitude and longitude for each trip, and instead uses a zone ID for a pre-established list of locations. As a result, we are not able to calculate the exact location of the pickup and drop off. However, the New York Taxicab and Limousine provides a dataset with the 265 location ID with their corresponding taxi zone, which we used to translated it into a map using a shapefile. Here we collect the most common location for pickup and drop off, percentage per borough, per day of week, day of month and the average of trip distance by hour and borough. For the analysis based on fare and payment, we had the goal to comprehend the best day and hour combination for tips, fare revenues and most common payment type. Furthermore, we collected common pickup and drop off location per payment type, and the average revenue per hour.

All our analysis was done with the help of Hadoop and Spark environment. Most of the code consisted of Spark SQL commands, which uses SQL queries. The rest of the paper is organized as follows. Section II introduces the related work. Section III presents our Preprocessing and Section IV proposes our Results. Section V presents Performance Evaluation and it is followed by Section VI with the conclusions.

2. RELATED WORK

There have been several analyses on New York City taxi transportation data. For example, on July 2016 a Data Science Boot Camp was held in NYC and reported an analysis of the 2014 yellow taxi data. Records included pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts [2]. However, data from the NYC TLC has changed from 2016, and since 2017 the latitude and longitude are no longer available. Therefore, those previous analysis are not possible to replicate anymore. At the present time, it provides a location ID feature that ranges from 1 to 265 for the specific service zone that a taxi cab can travel.

According to the NYC TLC, 13,587 Medallion Taxis are transporting New Yorkers in streets [3]. Only two (2) Taxicab Passenger Enhancements Project (TPEP) are licensed. TPEP focuses on the automated collection and submission of trip data, installation of a passenger information monitor, incorporating electronic message transmission capability into the taxicab and finally, the addition of equipment to enable the acceptance of credit/debit cards [4].

The complete analysis is performed with the help of Spark and Hadoop technology. Following is a brief definition of Big Data, Hadoop, Spark and Spark SQL.

A. Big Data

Big data is the term increasingly used to describe storage and analysis of large and or complex data sets using a series of techniques [5].

B. Hadoop

Apache Hadoop is an open source software project that enables distributed (parallel) processing of large amount of data sets across clusters of many computers. One server can scale up to thousands of machines, with a very high fault tolerance. Hadoop derives from Google's MapReduce and Google File System papers [6].

C. Spark

Apache Spark is a general-purpose cluster computing engine with APIs in Scala, Java and Python and libraries for streaming, graph processing and machine learning [7].

D. Spark SQL

Spark SQL runs as a library on top of Spark. It exposes SQL interfaces, which can be accessed through JDBC/ODBC or through a command-line console, as well as the DataFrame API integrated into Spark's supported programming languages [7].

3. PREPROCESSING

Before we could analyze the data, some preprocessing and storage was needed. All the

data was downloaded from the NYC TLC website, one file per month. An estimate of 10GB of size for all the files from 2017 was calculated. For processing the data, we implemented a Dataproc cluster from Google Cloud platform. We used the following set up: 1) One master with 4 CPUs, 15 GB memory and 500 GB of hard disk. 2) Two workers with 4 CPUs, 15 GB memory and 500 GB of hard disk each. Spark and Hadoop were pre-installed in the cluster. To run every command, we used the REPL (Read-Eval-Print Loop from Spark) with the Scala language. In addition, the 12 files produced a dataset of approximately 9.3GB.

To process the data after being combined with Spark SQL, we followed the subsequent steps. First, we filtered out all the "*null*" values from the datetime columns (pickup and drop off). We joined our Yellow Taxi cab dataset with the Taxi Zone dataset from NYC TLC, to obtain the Borough and Zone for pick up and drop off. We implemented a left join from the yellow taxi dataset with the taxi zone dataset through the column "PULocationID" and "LocationID" respectively. Further, it was also necessary to join with the drop off location "DOLocationID". The taxizone data set had four columns: LocationID, Borough, Zone and Service_Zone. The latter was drop, as it was not necessary for our processing.

Next, the column "tpep_pickup_datetime" and "tpep_dropoff_ datetime" allowed the creation of new columns to simplify the processing. These are the new columns added from the datetime values of the previous mentioned features: month, day of month, day of week, hour, minute and difference between drop off and pick up.

Thirdly, we only kept for the project all the data that contained the year 2017 in the "tpep_pickup_datetime" and drop off datetime column. The dataset contained 1896 rows with a different year for pickup and drop off than 2017, and some inconsistencies (e.g., a pickup in 2017 and drop off in 2016). Hence, our final preprocessed dataset contained 113,494,978 trips.

For calculating the time for every script, we encountered a problem with the Spark command "spark.time()" which can be used to see the time it took to a command sent as an argument. The problem consisted that it was not possible to obtain the value into a variable, therefore, making it unusable to us. Despite this, we implemented two variables to get the time from the system before and after the script, calculated the difference and then, saved it as a file. All the code here can be found within a repository in GitHub to freely download and run on a Spark environment [8].

Before the preprocessing, the original features of the dataset were the following:

- VendorID
- tpep_pickup_datetime
- tpep_dropoff_datetime
- passenger_count
- Trip_distance
- RatecodeID
- store_and_fwd_flag
- PULocationID
- DOLocationID
- payment_type
- fare_amount
- extra
- mta_tax
- tip_amount
- tolls_amount
- improvement_surcharge
- total_amount

After the preprocessing, more columns were added so the processing process would be simplified. Here is a list of those new features:

- PU_Borough/DO_Borough
- PU_Zone/DO_Zone
- PUMonth/DOMonth
- PUDay/DODay
- PUWeekday/DOWeekday
- PUHour/DOHour
- PUMinute/DOMinute
- trip_duration

4. RESULTS

During this section about the analysis of the NYC Taxi data set, the problem and solution are the principal focus. The analysis and its results are clearly stated one after the other, as follows. In order to understand some of the Spark SQL queries implemented in this analysis, we have set the name of our table view of the data set as “data”.

A. Analysis on trips overall

These analyses consist of understanding the dataset in general. As stated before, Spark SQL commands are used to run through the dataset of 113,494,978 rows.

1. Average distance travelled

This analysis was implemented in general, where we analyze and determine the average distance travelled in miles for 2017. For this year, the average travelled distance corresponds to 2.93 miles. Below is the code we implemented to obtain the result.

```
spark.sql("SELECT AVG(Trip_distance)
FROM data")
```

2. Average time travelled

This analysis was implemented in a global scope. It was calculated the average time travelled from the NYC Yellow Taxis of 2017. Therefore, the average time travelled for a person in New York was 995.69 seconds or 16 minutes and 35 seconds. Below is the code we implemented to obtain the result.

```
spark.sql("SELECT AVG(trip_duration)
FROM data")
```

3. Average efficiency based on distance over time

From the trip distance over the time in seconds, we calculated the average of efficiency for every trip. 0.3666% of an average efficiency was the result obtained from the dataset. Below is the code we implemented to obtain the result.

```
spark.sql("SELECT
AVG(Trip_distance/trip_duration) FROM
data")
```

4. Average passenger count per ride

This analysis took the average of all rides to determine the approximate number of passengers in a single cab ride, which was 1.625 passengers.

```
spark.sql("SELECT AVG(passenger_count)
FROM data")
```

5. Vendor percentage

Two TPEP providers exist in the dataset which provide the records. We determined the percentages of the records provided by each provider. Creative Mobile Technologies, LLC contributed 51 million records, or 45.08% of all records. Verifone Inc. contributed approximately 62 million records, or about 54.92% of all records in the dataset.

```
spark.sql("SELECT
VendorID,count(VendorID),COUNT(V
endorID)*100/(SELECT
COUNT(VendorID) FROM data)
FROM data GROUP BY VendorID")
```

6. Average of tips per trips

This analysis took all the trips to calculate the average amount of tips a passenger gives at the end of a ride. A taxi rider left a $1.83 tip on average per trip.

```
spark.sql("SELECT AVG(tip_amount)
FROM data")
```

7. Average of total paid per trips

We analyzed all the rides to determine the average of total amount paid per trip. A taxi rider paid $16.34 as an average per trip. This is a $2.94 increase from 2013, where the average trip cost $13.40 [9].

8. Total trips per hour

The processed data showed that the peak hour for taxi ride requests occurs at 6 PM, whereas the least taxi ride requests occurs at 4 A.M. This was determined after obtaining a total average of 7,087,806 and 1,131,708 respectively, for trips at each time of the day throughout the entire year. Figure 1 shows the distribution of hours and trips for the 2017 year.

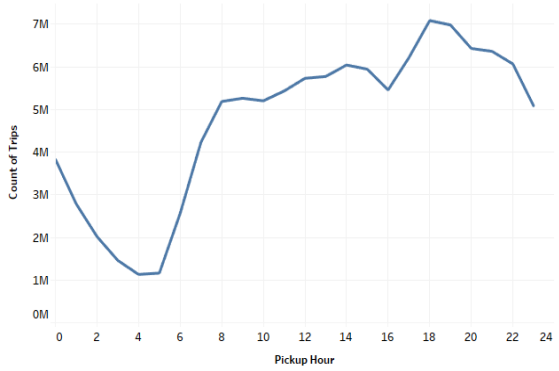

Figure 1: Total trips per hour

9. *Total trips per month*

The processed data showed the most active month as March; this was determined after obtaining a total count of 10,294,628 trips. The less active month was August with 8,422,196 pickups for the entire year. Figure 2 shows the distribution of hours and trips for the 2017 year.

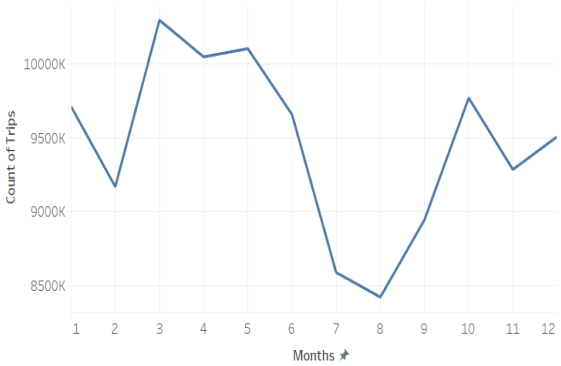

Figure 2: Total trips per month

10. *Common pickup locations*

The following map (Figure 3) shows overall highest pick up locations. To no surprise, Manhattan far surpasses any other borough. Other spots to note are popular airports.

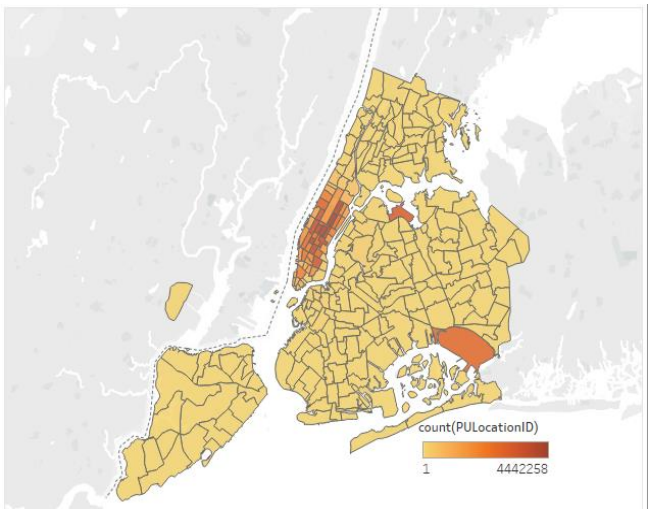

Figure 3: Common pickup locations

B. Analysis based on time and location

This analysis provides an in depth look at two main questions: when do customers commonly request for taxi rides and where do they commonly do it. We consider many different time intervals, such as hour, day of the week, day of month, and month.

1. *Average trips per day of the week*

For this analysis we considered the day of the week and the averaged count of trips per weekday. As a result, we discovered Fridays are the most active days for taxi ride requests, with a total average count of 2,483,952 passenger pickups throughout the year. Conversely Mondays were the least active days with 2,085,260 passenger pickups throughout 2017. Below, Figure 4 shows the averaged totals for every weekday.

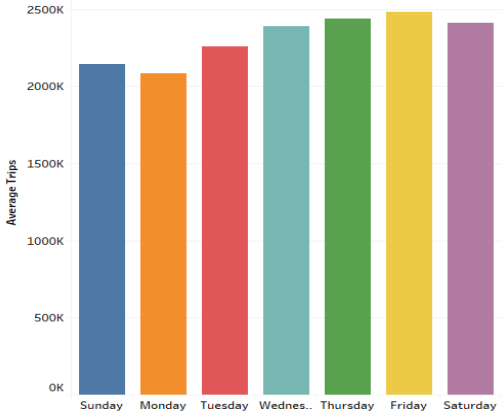

Figure 4: Average trips per day of the week

2. *Average distance travelled by time of day*

The following analysis demonstrates how far NYC taxis travel during the complete span of the day, an insight obtained due to this analysis was that the highest average traveled distance for 2017 was 17 miles at 2 P.M, on the other hand, the lowest average traveled distance was 13 miles at 2 A.M. The complete data is summarized in Figure 5 below.

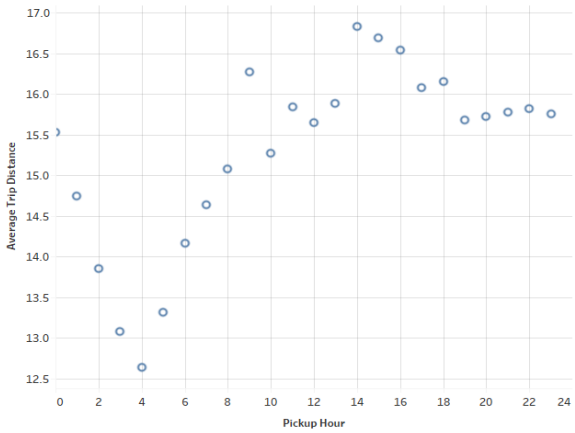

Figure 5: Average distance travelled by time of day

3. *Average distance travelled by time of day per borough*

A more detailed approach of the traveled distance is taken, by adding an extra feature to the previous analysis; this time we take into account the average traveled distance by time of the day by borough. Figure 6 demonstrates Queens travels the furthest on average by taxi. Also, the figure below illustrates the complete analysis.

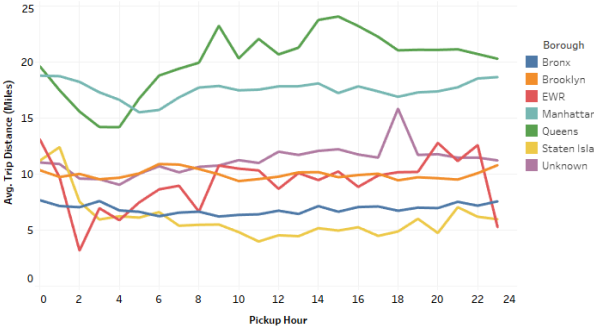

Figure 6: Average distance travelled by time of day per borough

4. *Average tips by time and day of the week*

Tip amount varies depending on pick up time and day. Interestingly, we found that the average tip amount peaks sharply around 5AM every day, then after a decline, slowly rises again around 4PM. These increases may be attributed to work commutes. We also found that tips are relatively low for both Saturday and Sunday. Figure 7 shows the fluctuation of tip amount over time.

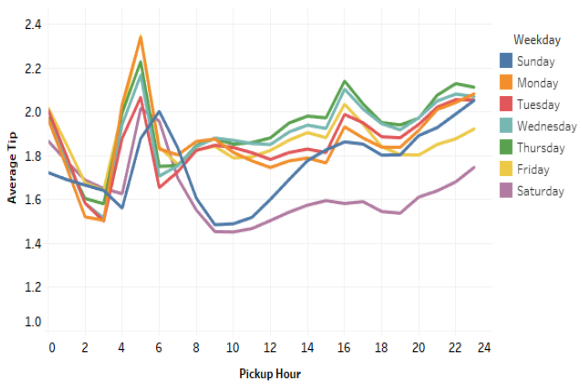

Figure 7: Average tips by time and day of the week

5. *Common nightlife drop-off locations*

New York City is known for being "the city that never sleeps". Analyzing which locations are most popular at night is very important. We analyzed drop-off times between 8 P.M. and 4 A.M. on Fridays and Saturdays. In Manhattan, Midtown Center is the most popular location overall. However, at night we see that East Village is the most popular drop-off zone, though it is not as popular during the day. In the following figure, the left map shows general Manhattan drop-offs, while the right shows nightlife drop-offs.

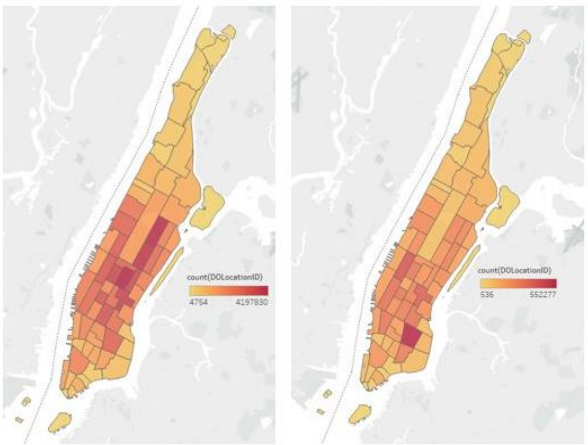

Figure 8: Common nightlife drop-off locations

6. *Pickup locations by season*

With this analysis, we attempt to find any interesting patterns in seasonal pick data. It's assumed that more people call for taxis in the winter, as it is colder and more difficult to walk; we found this to be the case with most boroughs. In Queens however, we found that more people called for cabs in the summer, shown in the following Figure 9.

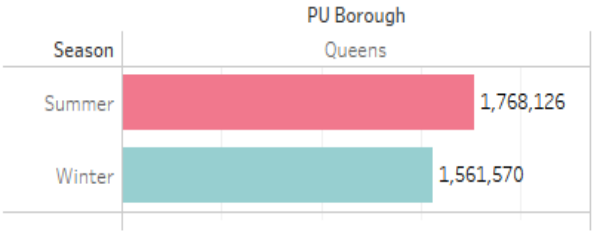

Figure 9: Pickup locations by season

7. *Percentage of pickup and drop off per borough*

The following analysis was implemented to understand how much of the pickups handle through each borough. Table 1 presents statistics on pick up and drop off percentages. Manhattan manages a 90.82% of pickups compared to the 6.15% from the second most, Queens. Moreover, Manhattan handles 87.98% of all drop offs with Queens and Brooklyn relative close to each other with 5.02% and 4.58% respectively.

TABLE 1
PERCENTAGE OF PICKUP AND DROP OFF PER BOROUGH

Borough	Drop off %	Pickup %
Bronx	0.60	0.08
Brooklyn	4.58	1.38
EWR	0.19	0.01
Manhattan	87.98	90.82
Queens	5.02	6.15
Staten Island	0.02	0.00
Unknown	1.60	1.56

C. *Analysis based on fare/payment*

1. *Most common payment type per borough*

For this analysis, we computed the most used payment type for the five boroughs. The most common form of payment Manhattan trips is credit card, followed by cash. Queens is the second with most transactions, having more credit card payment than cash as well. The figure below shows the results.

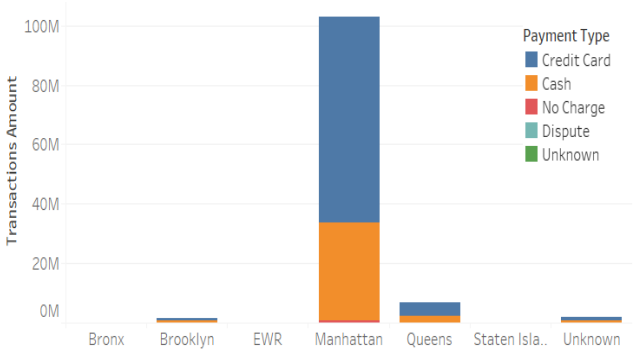

Figure 10: Most common payment type per borough

2. *Most common location for credit card transactions in Manhattan*

This analysis isolates Manhattan, the borough with the highest transactions per 2017. Credit card is the most used as presented above (See Figure 10). Therefore, Figure 11 presents the map of Manhattan with color representing the amount of transactions made with credit card.

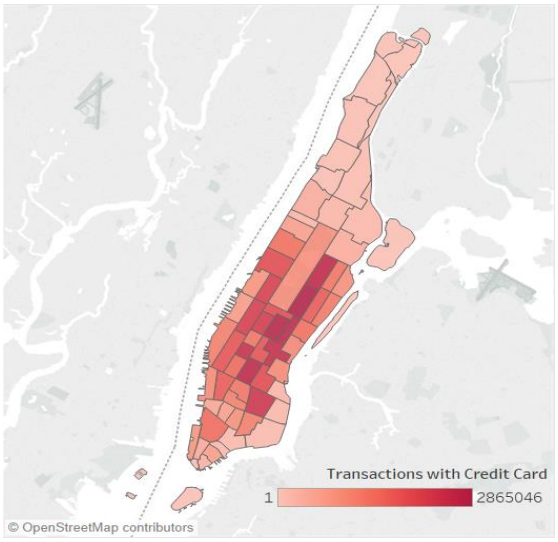

Figure 11: Most common location for credit card trans. in Manhattan

3. *Most common payment type according to hour*

The processed data showed how the trend for all the payment type goes through the hour of a day. In Figure 12, it is observed that the 6PM and 7PM hours contain the most credit card transactions. After 12 AM. there is a steep decline in credit card transactions, and a slower decline in cash payments. After 5AM however, credit card payments increase heavily relative to the slower increase in cash payments.

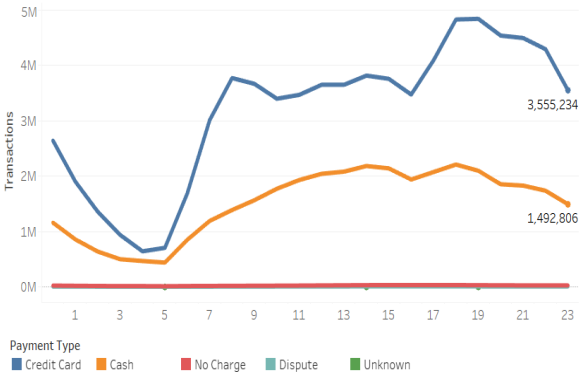

Figure 12: Most common payment type according to hour

4. Average gross revenue per hour

Understanding how much each taxi makes per hour allows us to see which hour is most profitable for the service and its drivers. Figure 13 lets us see the trend of gross revenue for a taxi driver. Gross revenue does not take into consideration the gas, or any fee charged to the taxi driver.

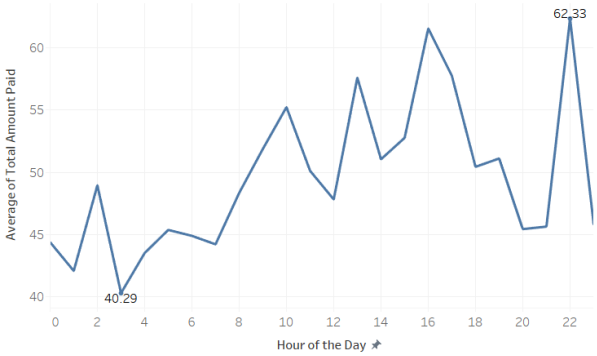

Figure 13: Average gross revenue per hour

5. PERFORMANCE EVALUATION

In this section, we evaluated performance of every Spark SQL statement implemented for our project. It is important to mention every script ran inside the REPL of Spark, with the computer specs stated in Section III. This data could be used to compare the results with other environments to find which tool is better to work with the NYC Taxi data. Table 2 presents the average time needed to accomplish those analysis in seconds and minutes.

TABLE 2
PERFORMANCE EVALUATION USING SPARK SQL

Analysis	Average Time	
	Seconds	Minutes
On trips in overall	424.83	7.08
Based on time and location	512.31	8.54
Based on fare/payment	484.44	8.07

6. CONCLUSION

Using the big data techniques taught in class this semester, we were able to successfully analyze 2017 NYC taxicab data. Being able to analyze a business' performance is very important for the growth of the business. The results found could be used to promote actionable business strategies for the future, which can yield great results for both NYC taxis and its customers. Location based analysis can be used to determine where is best to dispatch multiple taxis to maintain efficiency. Our time-based research also helps determine peak hours, allowing for a more timely, efficient service. Payment based analyses can help taxi drivers maintain a high average revenue, as it is known which locations generate the most revenue at which hour. With ride sharing apps such as Uber and Lyft surpassing traditional NYC taxis, hopefully further understanding the taxicab market with big data techniques can help them push to become more competitive [10]. Also, these models allow the creation and revision of a comprehensive vision of how travel patterns and use of transportation modes can be expected to respond.

REFERENCES

[1] NYC Taxi & Limousine Commission. http://www.nyc.gov/html/tlc/html/about/about.shtml.

[2] Analysis of NYC Yellow Taxi data, NYC Data Science Academy, Data Science Central, https://www.datasciencecentral.com/profiles/blogs/analysis-of-nyc-yellow-taxi-data

[3] NYC Taxi & Limousine Commission. 2017 Annual Report. http://www.nyc.gov/html/tlc/html/archive/annual.shtml

[4] NYC Taxi & Limousine Commission. Taxicab Passenger Enhancements Project (TPEP) Archive.

http://www.nyc.gov/html/tlc/html/industry/taxicab_serv_enh_archive.shtml

[5] J. Ward and A. Barker., “Undefined By Data: A Survey of Big Data Definitions”. Cornell University Library. arXiv:1309.5821v1. 2013. https://arxiv.org/abs/1309.5821

[6] P. Michalik, J. Štofa and I. Zolotová, "Concept definition for Big Data architecture in the education system," 2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, 2014, pp. 331-334.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6822433&isnumber=6822368

[7] Michael Armbrust, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi, and Matei Zaharia. 2015. Spark SQL: Relational Data Processing in Spark. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15). ACM, New York, NY, USA, 1383-1394. DOI: https://doi.org/10.1145/2723372.2742797

[8] NYC Taxi 2017 project in GitHub. https://github.com/luisadrianml/nyctaxi2017

[9] NYC Taxi & Limousine Commission. 2014 TLC Factbook.
http://www.nyc.gov/html/tlc/downloads/pdf/2014_tlc_factbook.pdf

[10] Todd W. Schneider, “Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance”.
http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/