What Is New York Taxi Data Telling Us

Viet-Nhat Nguyen
DISI, University of Trento, Italy
vietnhat.nguyen@studenti.unitn.it

ABSTRACT

By utilizing PySpark with Spark SQL—the tool connecting line between RDD and relational table—we analyse a big volume of data thereby providing a comprehensive view of taxi trips in New York City in over 2 years, from 2016 to 2018. In addition, our contribution includes using different kinds of visualization techniques to present information in a visual manner. We are able to reveal hidden pattern behind the data. Our analysis suggests that the New York traditional Green taxi have been facing a decline in the number of passengers. New York City Taxi and Limousine Commission (TLC) could use our result as a ground to understand more what is happening with the taxi industry thereby introducing adjustments to improve this kind of transport efficiently.

Keywords

Big Data, Spark, New York taxi, clustering, visualization

1. INTRODUCTION

NYC Taxi and Limousine Commission has collected and published the data about trip records including fields capturing 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. It is a huge dataset consisting of hundered Gigabytes storage of 3 types of taxi in New York: Yellow, Green and For-Hire Vehicle. Due to the limitation in time and computational capacity, we choose to focus only on Green taxi over the time of 2 years: from July 2016 to Jun 2018. By doing this, although we cannot compare between different kinds of taxi, we are able to mining insightful knowledge and understand how the habits of passengers and the quality of taxi had changed during that time. For example by just looking at figure 1, we could easily grasp the information about flow of taxi trips in New York.

This paper is structured as following: we will first start with preprocessing data in which we will present an overview of data features and how to clean it in section 2. Then we perform the clustering task in section 3. In section 4, we will explore the data by querying top trips according to some criterion. In particular, top 5 best trips, worst trips about distance, payment, speed and top 5 best locations, worst locations about number of trips, trip distance, trip speed, tip amount will be presented in section 4.1 and 4.2 respectively. We will illustrate how quantities of trips changed over hours of day, days of week and months of 2 years in section 4.3. Taking a further step, in section 5, we demonstrate some visualizations to discover facts underlying the data such as trend of payment method, what makes passengers give more tip amount, how trips were distributed geographically, by using various of visualization techniques. Finally, we summarize our study and discuss more about difficulties and future works in the section 6.

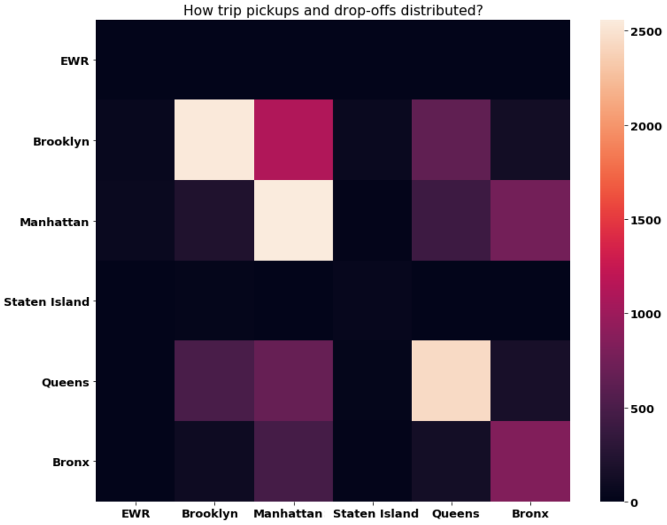

Figure 1: Heat map of from pickup locations to drop-off locations of all trips. The numbers in this heat map have been rooted (take the sqrt) to make it more visible.

2. DATA PREPROCESSING

This raw data was recorded by receiving information from drivers, which includes both automatically generated and manually created data. For this reason, apparently these data are not perfect so we have to preprocessing it before exploring some insights. By visualizing and plotting some statistics, we will have an overview of the big picture and be

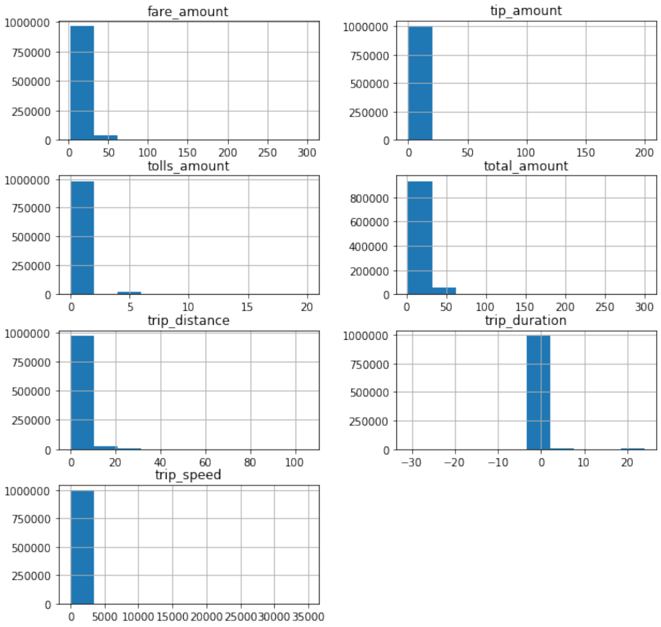

Figure 2: Histogram number of trips by different values of real-value features

able to observe anomalies, irrelevant parts for cleaning the coarse data.

We divide columns of the dataset into 2 types: discrete values and continuous values. With column having discrete values, we can easily check whether that column has outliers. Given that its valid values are well predefined, an instance is outlier if it contain value which is not in valid list.

For those features having real number values, it is a bit more complicated to detect wrong values. Take trip distance for example, how could we determine that a taxi trip moving 50 kilometers was abnormal or not? Let us start by visualizing the histogram of fields having float number values in figure 2. We can see that all of them have outliers. For example, $Total_amount$ has negative values which are impossible in reality.

In order to set the thresholds at that we can decide one row is inaccurately recorded, we have to plot some statistics over the data. Table 1 shows statistics of average, minimum values, maximum values and percentiles of different columns over first 1 million rows in the dataset.

Base on the above observation and the data dictionary describing Green taxi data [1], we clean the dataset according to following criterion:

- **VendorID**: value is either 1 or 2. This is a code indicating the LPEP provider that provided the record. 1 = Creative Mobile Technologies, LLC; 2 = VeriFone Inc.
- **fare_amount**: greater or equal to 2.5 which is at 1% percentile. We also found that this is the minimum fare for a taxi trip in New York.
- **trip_distance**: greater than 0.1. There are a number of trips (at least 0.1%) having 0 distance and also many trips having distance less than 100 meters. We found that the trip_fare for these trip also varies, not only be $2.5. These trips will not be consider and will be removed.
- **extra**: value is in $\{0, 0.5, 1.0\}$. This is extras and surcharges applied for rush hour and overnight charges.
- **tip_amount**: non negative values. This number in the dataset could be up to $250. However we do not set the upper bound for this column since there are some passengers are very generous in reality.
- **tolls_amount**: non negative values and not greater than 20. We heuristically set maximum tolls amount to $20 as it is reasonable and it statistically covers at least 99.99% trips in the dataset.
- **improvement_surcharge**: value is in $\{0, 0.3\}$. If this number is -0.3, then convert it to 0.3 as it is likely a mistake when redundantly pressing the minus button.
- **total_amount**: non negative values and not greater than 300. Similar to tolls amount, maximal at rate 300 is enough to cover at least 99.99% number of trips in the dataset. Besides, it is the dataset about trips in a city, so it is unreasonable that the amount of a trip excesses $300.
- **trip_duration**: positive values and not greater than 5. There are some trips recorded which has this trip duration up to 24 hours. However we set the maximum to 5 hours as it is possibly realistic.
- **trip_speed**: this is the average speed of a trip, calculated by the fraction of trip_distance and trip_duration. This value is must be positive and not greater than 60, considering that New York is a highly congestive city and the speed limit On New York City streets is 25 miles per hour.

After filtering the dataset with the above conditions, the remain contains more than 23 million trip records. For feature engineering, we also add several columns indicating different hierarchical levels in time of trips. All the queries in this paper will base on this cleaned data.

3. TRIP CLUSTERING

In this first task we will perform clustering algorithm on the dataset to find some patterns and popular profiles of trips. The results of this task is a list of centroids, each of them is a representative bearing characteristics of trips belong to that cluster. We will do it in unsupervised manners to group all the trips and assign each single trip to each cluster which has a profile. The features we chose are number of passengers, trip distance, trip duration, trip speed, total amount and tip amount. The algorithm we use is K-means. Before feeding data into clustering algorithms, we have to normalize the data since each feature has it own range and distribution. If we do not do so, the clusters will be biased to several particular features because clustering algorithm use Euclidean distance for the measurement.

We run the K-mean algorithm with number of clusters is 2, 3 and 4 sequentially then report the result in table 2. We evaluate the clustering performance with Silhouette score

[1] `https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf`

Table 1: Statistics of values of different features over 1 million rows in the dataset.

	mean	std	min	0.01%	0.1%	1%	median	99%	99.9%	99.99%	max
Fare_amount ($)	12.45	10.78	-475.00	-52.00	-3.50	2.50	9.50	51.00	95.00	250.00	1250.00
Tip_amount ($)	1.27	2.58	-5.16	0.00	0.00	0.00	0.00	3.70	22.22	75.00	250.00
Tolls_amount ($)	0.12	1.39	-30.00	0.00	0.00	0.00	0.00	0.00	10.50	18.04	900.00
Total_amount ($)	15.00	12.28	-475.00	-52.80	-4.80	6.30	11.44	27.80	104.30	300.00	1202.30
distance (km)	2.91	3.02	0.00	0.00	0.00	0.66	1.92	6.43	24.55	42.86	148.72
duration (hour)	0.38	1.87	-30.77	0.00	0.00	0.06	0.17	0.44	23.90	23.98	24.00
speed (km/h)	15.52	138.46	-0.08	0.00	0.00	7.31	11.77	19.87	468.00	5760.0	34800.00

which is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). We can see that number of passengers does not contribute to the clustering decision. The centroids are mainly affected by the change in trip distance, trip duration, total amount and also tip amount. Although the silhouette score is maximized when $k = 2$, cluster centres are still separable with $k = 3$, but become similar when $k = 4$. Therefore, based on this observation, we suggest to divide trips into 3 groups corresponding to 3 types of trips: short, medium and long.

We also tried to apply Gaussian mixture model, an algorithm based on probability. However, even with a single value number of clusters, the program did not finish in 5 hours so we had to terminate.

4. DATA EXPLORATION

4.1 Trip exploration

After cleaning and clustering the data, we would like to see some of the top statistic of trips. Particularly, we will use Spark SQL to identify top-5 best trips and top-5 worst trips in term of distance, speed and total amount.

Table 3 shows top-5 trips which have the longest travel distances and top 5 trips having shortest distances. Although we have filtered the dataset, there are still some outliers in the query result. Take top-5 longest distance trips for example, the fare amounts range from $70 to $300 while the distances are around 120 kilometers. It is very hard to tell which ones of these trips were wrong recorded. Top 5 shortest trips all are 0.1 kilometer since it is the lower bound of our filtering. It is feasible that the fare amounts of these trips are only around $3 and the pickup places and drop-off places are the same. Except for the trip that started from location number 80 and ended at location number 112. We found that these locations are neighbouring on Brooklyn map (figure 3).

Table 4 indicates top-5 fastest trips and top-5 slowest trips as well, according to average speed. Owing to the fact that we have discarded all the trips with average speeds greater than a threshold, top-5 fastest strips are at 60 km/h. An interesting observation is that most of the trips appearing in this table have the same starting and destination location ID. The slowest trips were in areas of Manhattan and Bronx borough and the fastest trips were in areas of Queens and Brooklyn borough. In term of time aspect, top slowest trips happened around 7 P.M and 8 P.M which are usually rush hours. By contrast, fastest trips were in the afternoon and at midnight. Further insights about average speed of trips will be analysed in section 4.2 and 5.

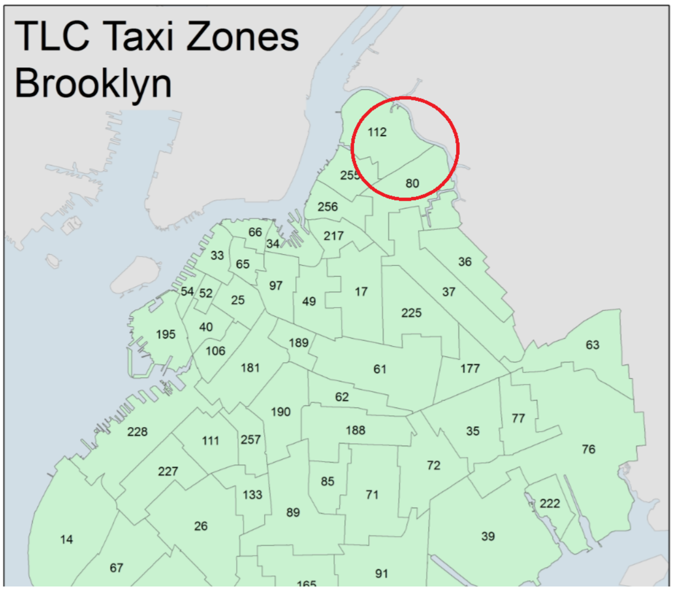

Figure 3: A part of taxi zone map in Brooklyn. There is one shortest trip that went from location number 80 to number 112. We can see that they are adjacent areas.

The top highest paid trips and lowest paid trips are presented in table 5. It seems that most of top highest paid trips are outliers because they all have total amount at $300 but their distances are less than 1 kilometer. Similarly, in top-5 lowest paid trips, there are two trips having distances greater than 6 kilomters but costed only $2.5—the minimum amount that passengers have to pay for a taxi trip in New York. We observe this result because it is very intractable to determine threshold for filtering noise. However, although remain outliers could affect top-k identifications, they are not problem for trend analysis which requires computations over large number of trips in next sections.

4.2 Location exploration

Table 6 and table 7 convey information about best locations and worst locations for pickup and drop-off respectively. It is clear that top appealing locations for taxi were in Manhattan and Queens, where is the largest residential and central business district in the United States, while most of locations that had the least trips were in Staten Island. We will see this distribution in more visual way in next section.

Table 8 shows top locations from there passengers took longest trips and shortest trips, mean while top locations

Table 2: Statistics of values of different features over 1 million rows in the dataset.

K	centroids						counting	Silhouette score
	passenger count	distance	duration	speed	tip amount	total amount		
2	2.72	4.64	0.39	23.64	1.90	25.57	19401764	0.7811
	2.74	10.71	0.72	30.02	4.31	47.38	3769868	
3	2.72	4.29	0.37	23.12	1.77	22.27	16455128	0.6632
	2.74	14.65	0.92	31.81	6.25	61.97	1126599	
	2.74	7.77	0.59	27.76	3.05	36.84	5589905	
4	2.49	7.70	0.58	27.66	3.03	36.62	5338341	0.4856
	2.70	14.58	0.92	31.84	6.20	61.65	1139616	
	2.48	4.28	0.37	23.13	1.76	24.22	15394485	
	6.56	5.21	0.42	24.39	2.13	27.62	1299190	

Table 3: Top 5 trips having the longest distances (top half) and shortest distances (bottom half)

pickup_time	pickup location	drop-off location	distance	fare amount
2017-05-02 16:07:02	205	23	128.81	70.00
2017-06-19 14:24:13	242	23	126.98	90.00
2016-08-22 13:29:34	92	23	122.82	100.00
2016-08-24 16:59:14	197	19	114.17	200.00
2018-06-24 16:44:32	130	265	112.2	300.00
2016-07-01 06:31:33	80	112	0.10	3.50
2016-07-01 08:55:49	97	97	0.10	2.50
2016-07-01 06:24:26	243	243	0.10	3.00
2016-07-01 02:28:45	179	179	0.10	3.00
2016-07-01 07:53:48	17	17	0.10	3.80

Table 4: Top 5 fastest trips (top half) and top 5 slowest trips (bottom half) by trip_speed

pickup time	pickup location	drop-off location	distance	trip speed
2016-07-09 21:56:13	152	152	0.40	60.00
2016-07-15 23:22:47	17	17	0.50	60.00
2016-07-13 22:34:43	91	91	0.10	60.00
2016-07-02 07:59:46	28	28	0.10	60.00
2016-07-15 16:48:31	258	258	1.70	60.00
2016-09-07 19:25:57	116	116	0.10	0.022
2017-02-24 19:09:56	254	254	0.11	0.023
2017-12-14 20:34:32	182	212	0.10	0.029
2016-07-29 19:39:49	74	74	0.13	0.030
2017-09-06 19:24:52	196	56	0.14	0.031

for average speed aspect are represented in table 9. We can see that from many areas of Manhattan, passengers usually went for quite short distances, typically around 2 kilometers. In these areas, taxi were not able to go fast due to the high traffic jam while in some zones of Staten Island, taxi could travel much more faster.

Table 10 exhibits top pickup locations respect to tip amount. We observe the big disparity because there were very few trips in several locations. For instance, there is only 1 trip starting from Great Kill Park of State Island in the dataset and the driver in this trip was not received any tip money. We will present a more comprehensive view for these locations in the visualization section.

4.3 Time-related exploration

One of the most interesting part is examining how trips changed by time. The advantage of our analysis is that we use the data recording trips for over 2 years, therefore we can inspect how the trips had changed, what was the trend of passengers' habit.

Firstly, we group the dataset by the hours when trips started and compute the average quantities then plot the result in figure 4. Predictably, passengers' demand on taxi reached the pick at round 6 and 7 PM which are rush hours when most people left from works, and constantly decreased from that rush hours to 5 AM of the following day. Contradiction to number of trips were the average speed of trips and distance of trips. From 4 PM to 5 AM, passengers had to pay averagely much more twice times of tax amount than the other time of day, seeing that trip in rush hours and overnight were charged an extra tax. Passengers in trips happening in the morning time tended to be more generous since they tipped more money than those on trips in the afternoon. Our hypothesis is that people have better mood in the morning than in the afternoon when they are in the middle of works. Trips in the early of morning had highest average number of passengers but the dissimilarity was not remarkable.

Secondly, we illustrates how trips changed on different days of a week in 5. Not surprisingly, Saturday was the day having the highest number of trips, followed by Friday and Sunday. Weekends are the time in that trips had higher average speed. Passengers also tended to went on longer trips with more companies on the weekends. Drivers was received lowest tip amount on Monday as it is the starting day of a week when people start to work again after weekend holidays. This graph reflects facts of life.

Finally, we plot the changing in the number of trips, average trip distance, average trip speed and average tip amount each month changed over 2 years in figure 6. We can see that number of trips had gone down from more than 1.2 million in 2016 to under 0.8 million in 2018. Passengers had a tendency to went longer in each time they called a taxicab. This is because of the emergence of peer-to-peer ride sharing like Uber. Besides, the average speed of taxi and the tip amount drivers was received also tended to declined. Conjecturally, the degeneration of quality of taxi services and the com-

Table 5: Top 5 highest paid trips (top half) and top 5 lowest paid trips (bottom half)

pickup time	pickup location	drop-off location	distance	total amount
2016-07-11 19:40:13	225	61	0.86	300.00
2016-07-09 15:52:21	37	37	0.12	300.00
2016-07-10 12:14:40	89	89	0.55	300.00
2016-07-11 10:00:52	85	85	0.44	300.00
2016-07-23 18:58:05	243	265	15.49	300.00
2016-07-17 16:34:18	152	235	12.55	2.50
2016-07-30 14:58:47	28	28	0.10	2.50
2016-07-09 12:58:16	166	151	1.03	2.50
2016-07-26 14:48:01	74	247	6.84	2.50
2016-07-30 12:17:30	89	22	1.91	2.50

Table 6: Top 5 locations with most pickups (top half) and top 5 locations with least pickups (bottom half)

pickup location	borough	zone	service zone	num of trips
74	Manhattan	East Harlem North	Boro Zone	1401230
41	Manhattan	Central Harlem	Boro Zone	1306566
75	Manhattan	East Harlem South	Boro Zone	1223346
7	Queens	Astoria	Boro Zone	1161720
82	Queens	Elmhurst	Boro Zone	1003063
110	Staten Island	Great Kills Park	Boro Zone	1
12	Manhattan	Battery Park	Yellow Zone	1
99	Staten Island	Freshkills Park	Boro Zone	6
204	Staten Island	Rossville/Woodrow	Boro Zone	8
2	Queens	Jamaica Bay	Boro Zone	9

Table 7: Top 5 locations with most drop-offs (top half) and top 5 locations with least drop-offs (bottom half)

pickup location	borough	zone	service zone	num of trips
74	Manhattan	East Harlem North	Boro Zone	771911
42	Manhattan	Central Harlem North	Boro Zone	764769
7	Queens	Astoria	Boro Zone	721274
41	Manhattan	Central Harlem	Boro Zone	685612
129	Queens	Jackson Heights	Boro Zone	641134
110	Staten Island	Great Kills Park	Boro Zone	1
105	Manhattan	Governor's Island...	Yellow Zone	2
99	Staten Island	Freshkills Park	Boro Zone	15
2	Queens	Jamaica Bay	Boro Zone	38
204	Staten Island	Rossville/Woodrow	Boro Zone	74

Table 8: Top 5 pickup locations that from there passengers averagely took longest trips (top half) and shortest trips (bottom half)

pickup location	borough	zone	service zone	distance (km)
117	Queens	Hammels/Arverne	Boro Zone	16.67
86	Queens	Far Rockaway	Boro Zone	14.79
201	Queens	Rockaway Park	Boro Zone	14.60
109	Staten Island	Great Kills	Boro Zone	12.83
5	Staten Island	Arden Heights	Boro Zone	11.48
237	Manhattan	Upper East Side S...	Yellow Zone	1.95
234	Manhattan	Union Sq	Yellow Zone	1.97
142	Manhattan	Lincoln Square East	Yellow Zone	2.07
41	Manhattan	Central Harlem	Boro Zone	2.11
141	Manhattan	Lenox Hill West	Yellow Zone	2.12

Table 9: Top 5 pickup locations that from there trips had highest average speed (top half) and lowest average speed (bottom half)

pickup location	borough	zone	service zone	speed (km/h)
99	Staten Island	Freshkills Park	Boro Zone	31.15
44	Staten Island	Charleston/Totten...	Boro Zone	28.71
204	Staten Island	Rossville/Woodrow	Boro Zone	25.69
1	EWR	Newark Airport	EWR	24.56
132	Queens	JFK Airport	Airports	24.34
12	Manhattan	Battery Park	Yellow Zone	6.60
234	Manhattan	Union Sq	Yellow Zone	9.06
186	Manhattan	Penn Station/Madi...	Yellow Zone	9.20
161	Manhattan	Midtown Center	Yellow Zone	9.29
107	Manhattan	Gramercy	Yellow Zone	9.55

Table 10: Top 5 pickup locations that starting from there trips had highest average tip amount (top half) and lowest average tip amount (bottom half)

pickup location	borough	zone	service zone	tip ($)
109	Staten Island	Great Kills	Boro Zone	7.23
23	Staten Island	Bloomfield/Emerso...	Boro Zone	6.81
1	EWR	Newark Airport	EWR	4.77
204	Staten Island	Rossville/Woodrow	Boro Zone	4.27
115	Staten Island	Grymes Hill/Clifton	Boro Zone	4.18
110	Staten Island	Great Kills Park	Boro Zone	0.0
187	Staten Island	Port Richmond	Boro Zone	0.08
248	Bronx	West Farms/Bronx ...	Boro Zone	0.30
94	Bronx	Fordham South	Boro Zone	0.32
173	Queens	North Corona	Boro Zone	0.34

petition from carpooling applications made the passengers hesitate to give gratuity to taxi drivers.

5. FURTHER VISUALIZATION

In this section, we want to dive deep into the data and visual techniques is the most effective way to tell the story. Figure 7 shows the percentages of different types of payment methods and how the ratio between paying by cash and paying by credit card changed over time. Generally, most of passengers pay by cash or credit card as they accounted for 99.6% of trips. Not surprisingly, the ratio of passengers choosing pay by credit card gradually increased over time. This trend have happened in over the world and in all aspects of life where cash have been steady replaced by electronic payment, and taxi paying in New York is not an exception.

We want to discover the relationship between tip amount and the other factors of a trip. Figure 8 is the trip distribution that represents how tip changed when the other factors changed by plotting randomly 3,000 trips that passengers paid by cash (since there was no tip when passengers paid by credit card). Surprisingly, we found that passengers tend to tip more when they went alone and gave less tip in four-passengers trips. It is quite obvious that tip amount was likely to increase linearly when fare amount rose. Although the dependence is not so clear but we can see that gratuities also tend to get bigger for faster trips. Finally, there is no correlation between tax amount and tip amount.

Figure 1 shows the flow of taxicab. Most of the trips

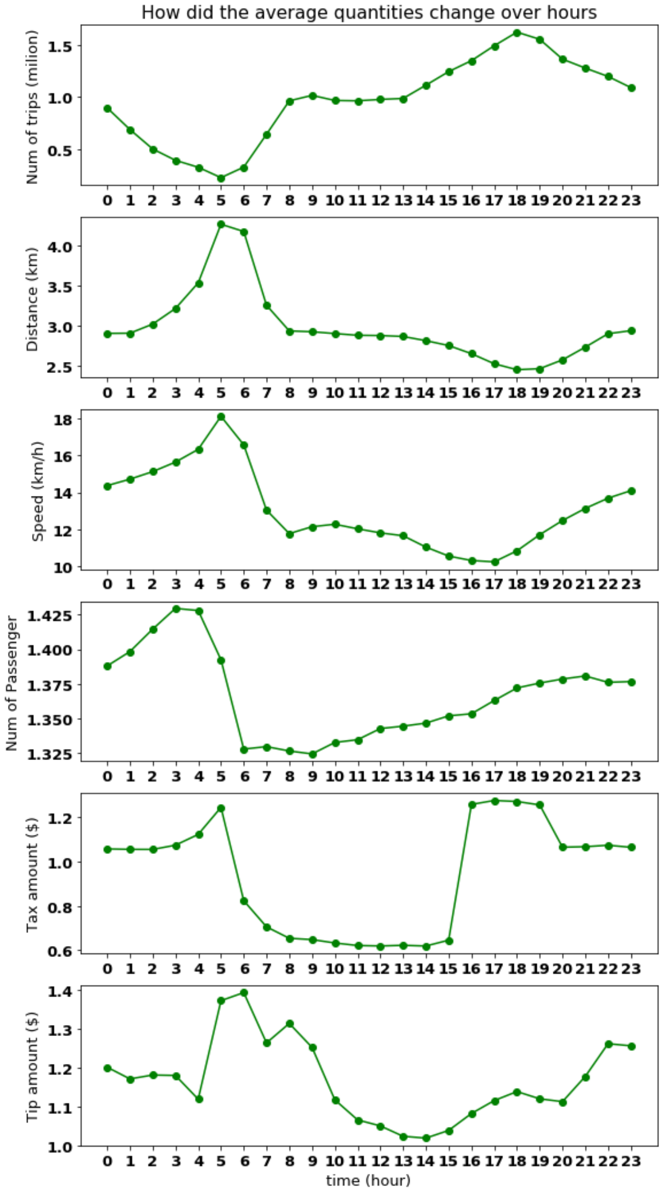

Figure 4: Average numbers of various quantities changed over hours each day.

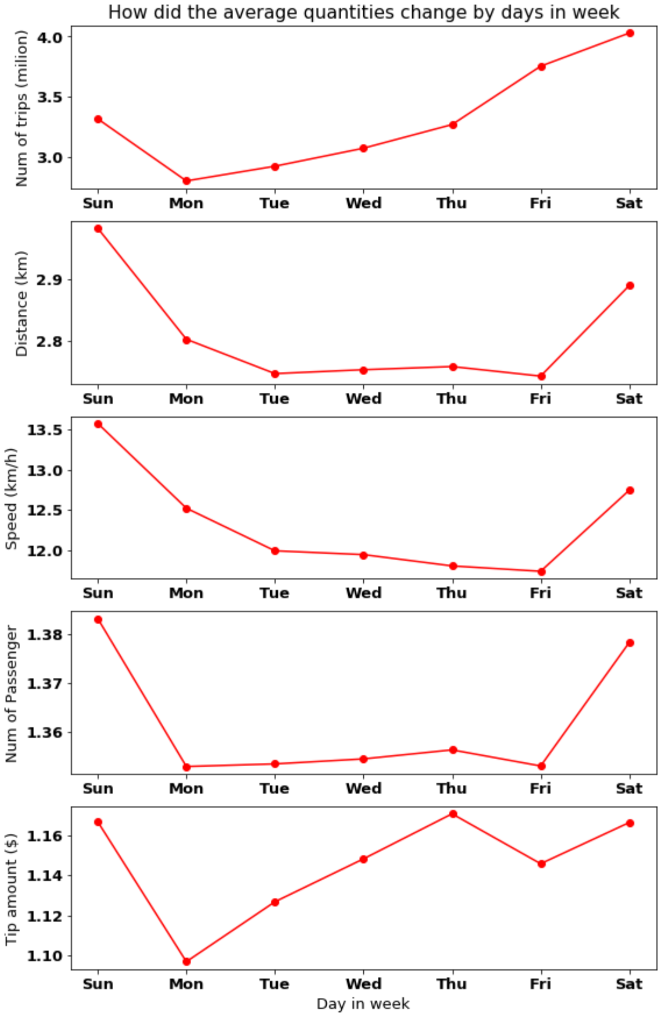

Figure 5: Average numbers of various quantities changed over days in week.

started and ended within the same borough as the diagonal cells are brighter than the other part of the matrix. Besides, trips from Manhattan to Bronx, from Queens to Manhattan and trips from Brooklyn to Manhattan dominated other routines.

The last, but probably the most fascinating part, is the visualization on the map of New York City. By using the shapefile provided by TLC, we are able to plot heat maps to see how quantities changed over locations in figure 9. We can see that most of the trips concentrated on Manhattan and vicinity areas. This is apparently because that area has high density and also the airport. In contrast, trips from Staten Island and south east areas of New York could archive speed three times higher than trips from Manhattan. In addition, drivers on trips from these locations got higher tip, as we analysed the relationship between trip speed and tip amount. It is quite interesting that passengers on trips starting from Staten Island had to pay more taxes than the other areas.

6. CONCLUSION

In this paper, by combining the strength of Big Data tools–Spark SQL and MLLib–and visualization libraries, we dived deep into the data of Green taxi trips in New York City in 2 years from July 2016 to Jun 2018. Our investigation reflects the behaviours of passengers and reveals some underlying relationships between factors of trips. We performed clustering algorithm on the dataset, queried top-5 trips, locations in term of various criterion, drew charts to show how trips changed by time, looked at the data using geo-spatial level. In our opinion, cleaning data is the most important job when working with data analysis since all the other experiments are based on this process.

Due to the limitation of computational capacity, a shortage of our work is that we only are able to run our experiments on 1 type of taxi and perform only K-means al-

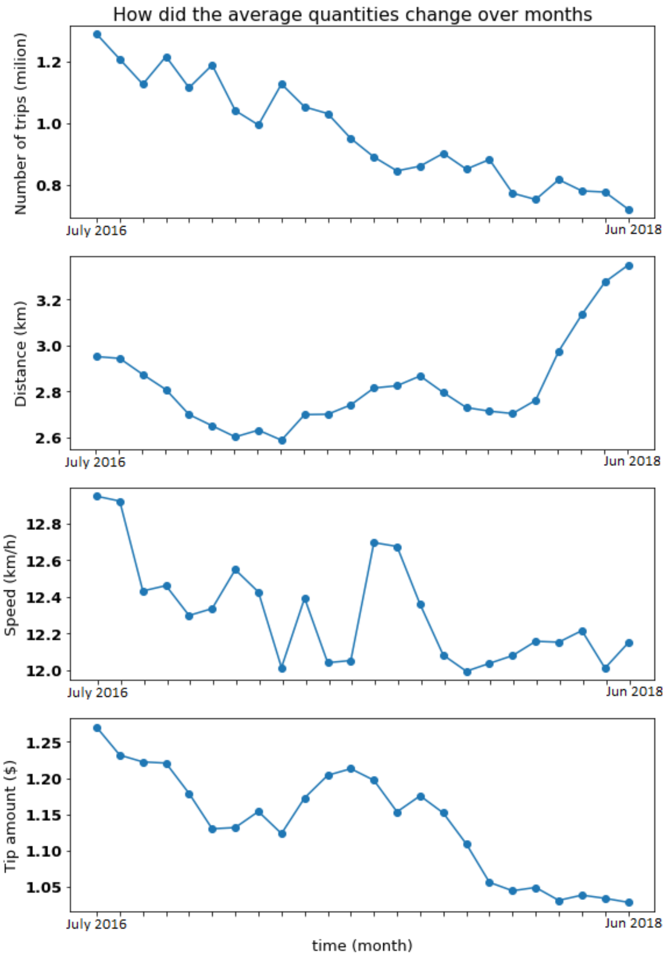

Figure 6: Number of trips, average trip distance, average trip speed and average tip amount each month changed over 2 years.

gorithm for clustering. In further works, it is worth to use more data, compare behaviours of different kinds of taxi and try to cluster trips with other algorithms such as Gaussian Mixture model. Besides, it is also necessary to apply domain knowledge to data cleaning, given that our results still contain some outliers when querying and visualizing since we heuristically set some thresholds to remove abnormalities.

7. ACKNOWLEDGMENTS

This is the final project of Big Data course at University of Trento, A.Y. 2018-2019. The data used in this project was downloaded from TLC website in January 2019.

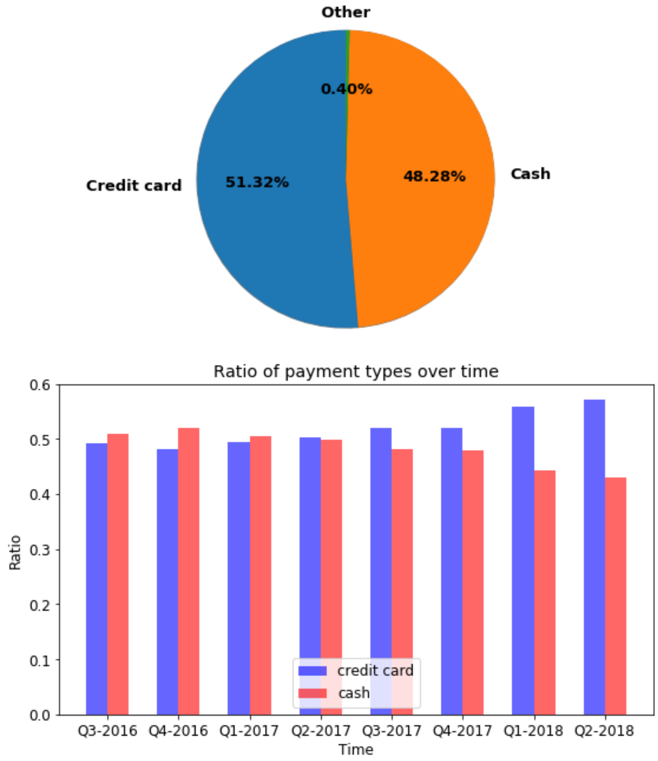

Figure 7: Percentages of different types of payment (top) and trend of two major payment methods: by cash and by credit card (bottom).

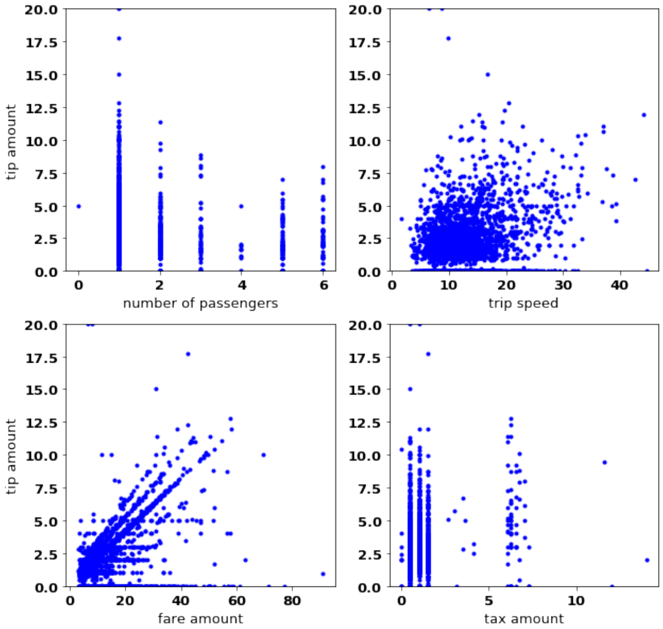

Figure 8: Relation between tip amount drivers received with number of passengers, fare amount, total tax amount and average speed of trips.

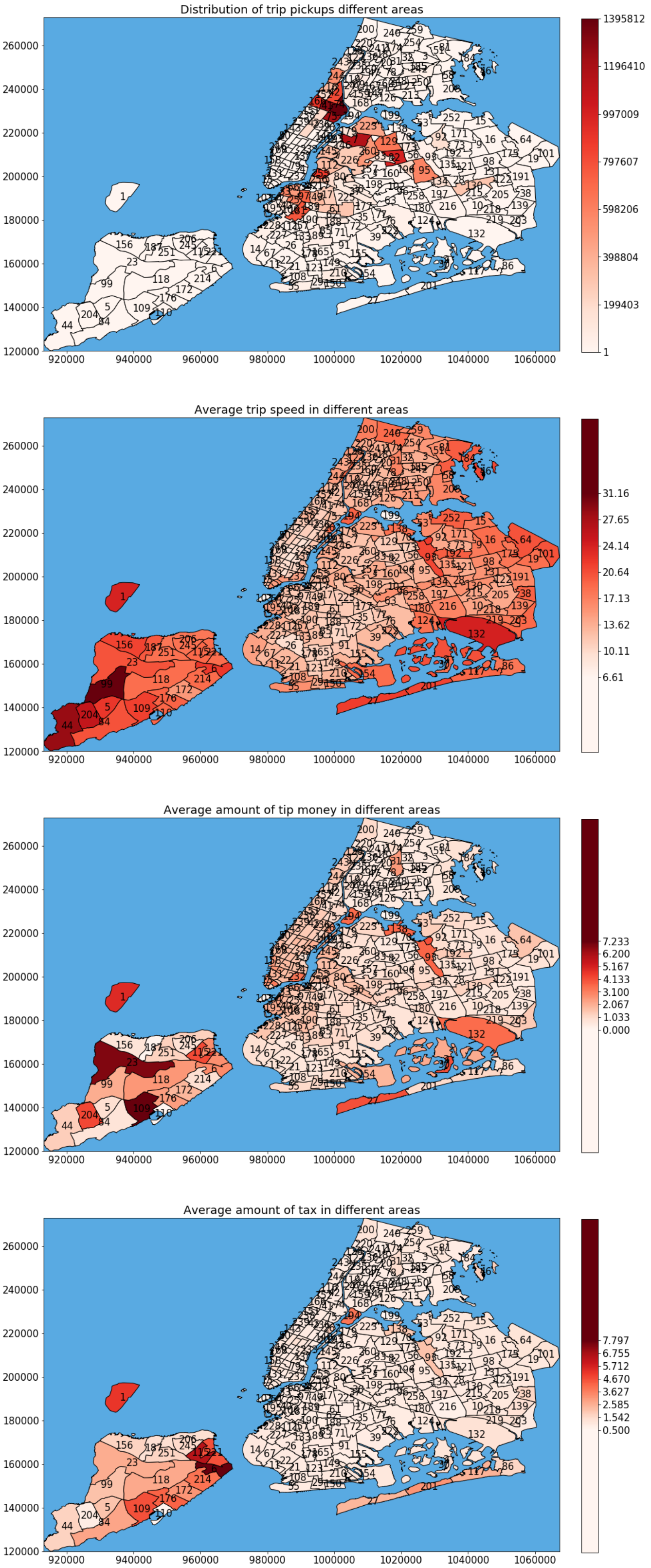

Figure 9: Relation between tip amount drivers received with number of passengers, fare amount, total tax amount and average speed of trips.