The New York City Taxi and Limousine commission has provided with the data of the trips by their taxis. We will be using the dataset for September, 2015.

The data consists of:

- Total Columns : 21
- Total Rows: 1,494,926 observations

Each observation has 20 features and the details can be found at
http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf

Since the number of observations is more than a million, it is very expensive to store it on memory and analyze it. For this reason, I have used the MapReduce framework which generates the RDD, which is easier and cheaper to analyze.

Python Packages

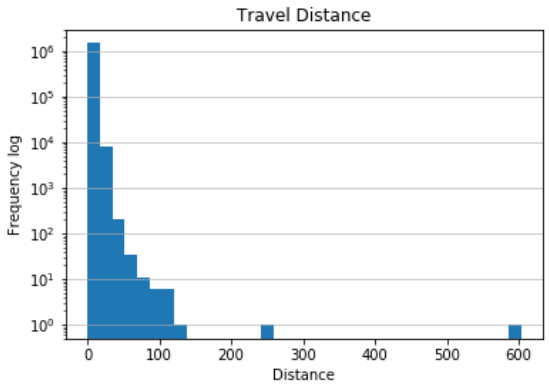
The python packages used for the analysis are:

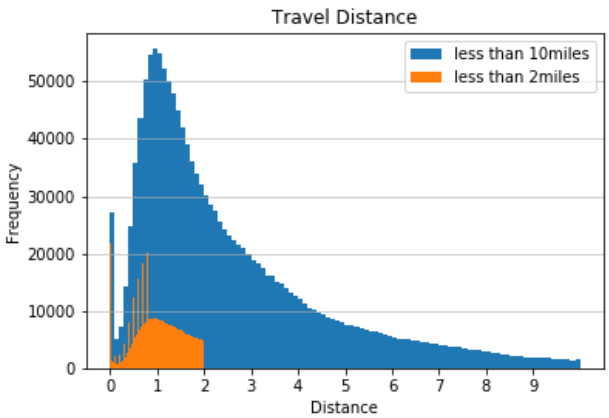
- Pyspark → mllib →Kmeans
- Numpy for data wrangling
- Matplotlib for data plotting
- Pandas for data analysis

Analysis of the Dataset

Distance Travelled

Since the data set is for taxi service, the provider has recorded the distance travelled by each taxi during the ride. Due to high variability, the plot is not interpretable. To solve this, I have logged scaled the Y axis. Number of Bins used are 35.





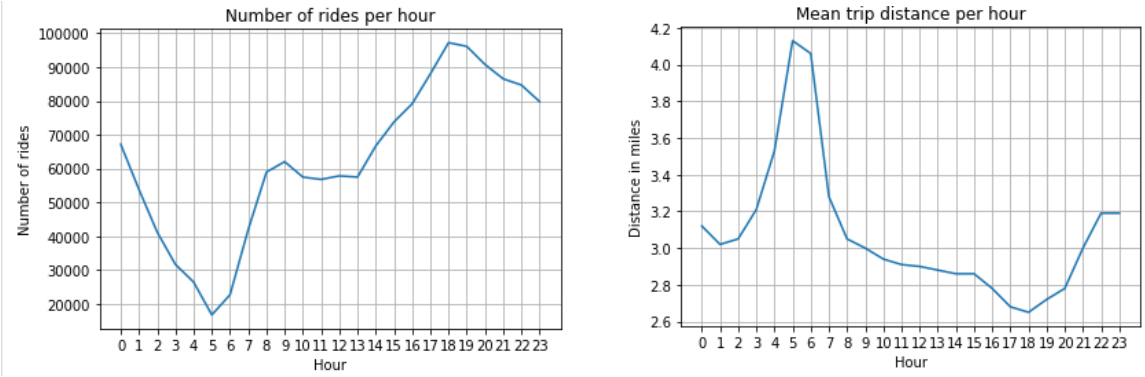
It can be observed that the distance below 10miles has the highest frequency. To analyze this, I plotted another histogram with less than 10miles travelled and less than 2 miles travelled.

- 96% of the total trips are within 10 miles ride distances
- 50% of the total trips are within 2 miles ride distances

Analysis of daily trends

Another metric useful for operations analysis is the daily trip distances. This can be used to understand the demand of the customers over the day. I have divided this into two parts:

- Total number of rides based on hour of the day
- Mean trip distance during each hour of the day



Observing the two graphs, it can be seen that during the low peak hours, the drivers tend to get long distance trips or is it vice versa, because the drivers take long trip, they don't get enough rides? It seems logical that during the late night hors they might take long trips and during the evenings the drivers get more trips due to short and quick trips.

Airport Trips

It is observed that the taxi provides special services to JFK and Newark airports, since they have assigned codes for the same. I have tried to answer the following questions.

1)Trips to airports : Which airport is more used?

Total Number of trips to JFK or Newark Airports is 5,552
Rides to Newark: 1,117
Rides to JFK: 4,435

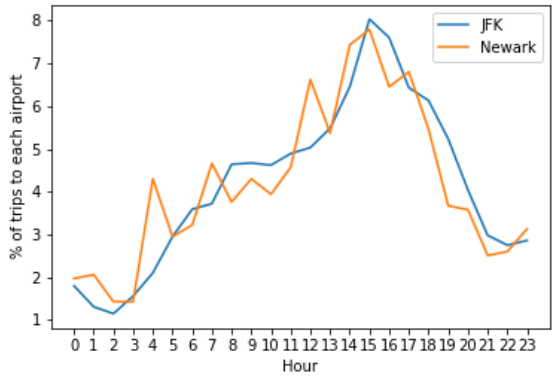
2)Average Fare : Which airport is expensive to reach?

Average all ride fare: $ 15.03
Average Fare for both airports: $ 57.21
Average Fare for JFK: $ 56.53
Average Fare for Newark: $ 59.9

3)Average Tips: Does a driver get better tips because of the luggage passenger is carrying?

Average all ride tip: $ 1.24
Average tip for both airports: $ 4.35
Average tip for JFK: $ 4.08
Average tip for Newark: $ 5.44

4)Daily trends of rides to airport



It can be observed that the highest percentage of rides to both the airports is during the evening**. People do prefer flying after office hours.**

5) Type of Payment : Do passenger carry enough cash for such expensive trips?

Airport/Payment Method	1 (Credit Card)	2 (Cash)	3 (No Charge)	4 (Dispute)	5 (Unknown)
Newark	460	499	110	47	1
JFK	1740	2498	138	58	1

From the numbers, two specific observations can be made: The highest number of transactions are via cash, and the number of disputes for Newark is quite high as compared to JFK. (Since total trip are 1/3rd of JFK, but number of disputes are almost the same)

Analysis of localities

Considering that the taxi service provides service in both NJ and NY, it is advisable to look at how the ridership is divided between each state. For the purpose of analysis, I have give New Jersey as code = 1 and NY as code = 2

The rides are bifurcated as follows:

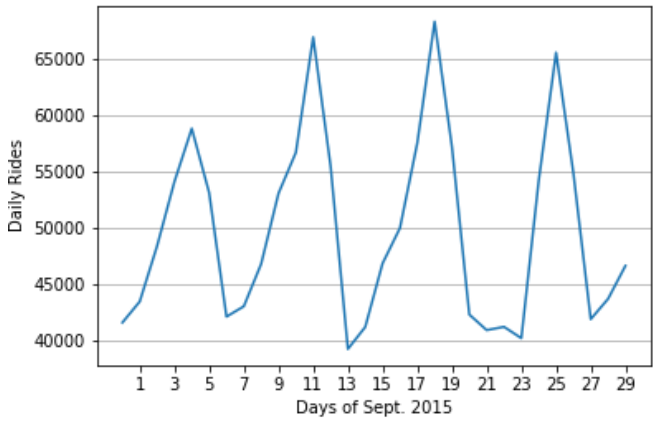
1. NJ to NJ: 449,130 or 30.0 %
2. NJ to NY: 37,133 or 2.0 %
3. NY to NY: 864,553 or 57.0 %
4. NY to NJ: 144,110 or 9.0 %

Within New York Total Rides: 864553

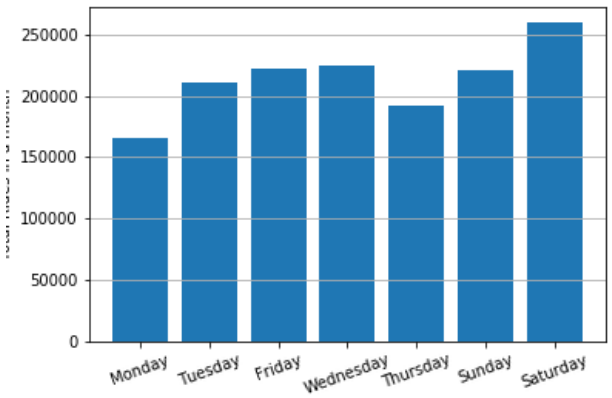
1. Uptown to Downtown Rides: 398,428
2. Downtown to Uptown Rides: 46,6125

Weekly and Monthly trends in Ridership

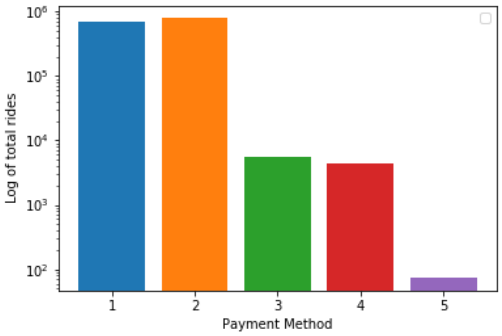
I further analyzed the total monthly data to identify any trends in the demand. There is a clear trend. The demand fluctuates every couple of days and is cyclical.



The peaks are usually on Saturday. Out of the seven days, the highest travelled days are as shown



Further Recommendations

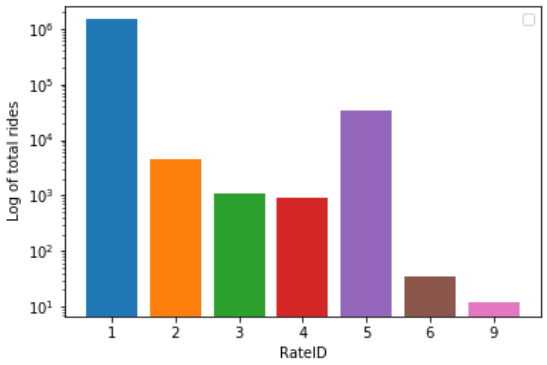


Payments Method:

Using cash is mostly preferred, followed by credit card.

Vendors

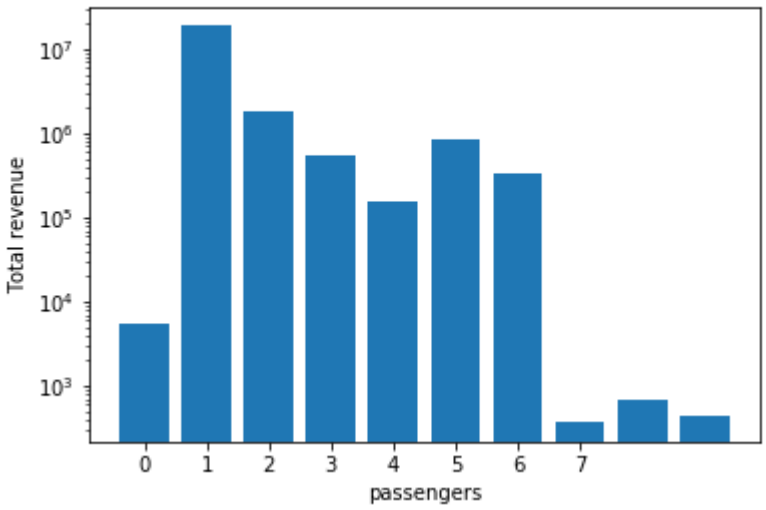
Even though there are two vendors, the VeriFone Inc. is dominant in market. They account for 78% of the total rides. ***What can be done to increase Creative mobile's share?***



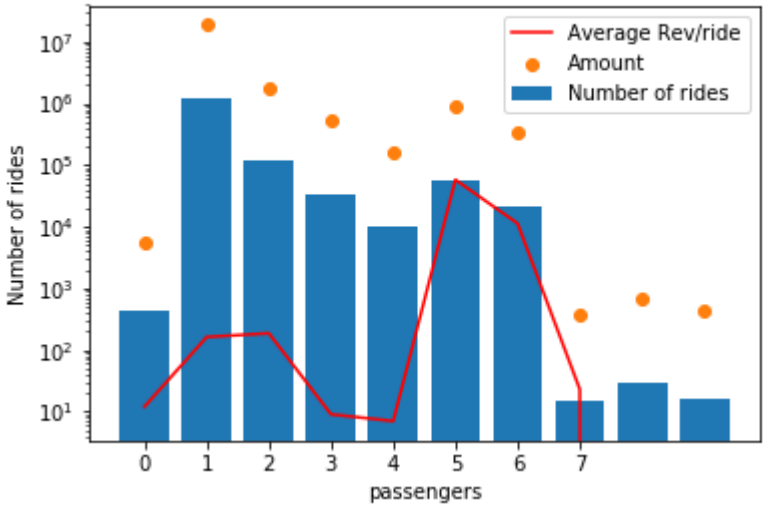
Rate ID distribution

Rate ID 1 is for standard fare, which accounts for most of the rides. Following that is negotiated fare (ID 5). ***Are the driver negotiating themselves to avoid commission?***

Analyzing which kind of passenger ride generates highest revenue?



The above graph shows the distribution of how the rides are divided with number of passengers. It is observed that the solo passengers account for maximum number of rides. ***Will carpooling attract more solo passengers?***



While the solo passenger rides the maximum, ***the trips with five passengers contribute in highest average revenue/ ride.***