NYU·poly

POLYTECHNIC INSTITUTE OF NEW YORK UNIVERSITY

Computer Science and Engineering

Massive Data Analysis
Taxi Data Analysis
Semester: Fall 2014

Professor: Juliana Freire

Team Name: Cygnets

Team Members:

Name	**Student ID**	**Email**
Savan Rupani	N10779202	spr297@nyu.edu
Aditya Nehulkar	N14427271	ann277@nyu.edu
Aniket Bezalwar	N12412473	asb637@nyu.edu

Table of Contents

Problem Area

For the project we have done analysis on various aspects of taxi service. We have done an analysis on taxi usage, taxi economics and taxi tip behaviors. Though we used these three aspects as our base of analysis we have mainly focused on the taxi economics and taxi tip behaviors.

Project Description

To analyze NYC taxi service we have used two taxi service related data, trip and fare data. While trip data had the information about the taxi rip like location and date etc. The fare data contain the payment information about the taxi ride like trip fare, tip, tax and all things.

We have also used census data to do some proper analysis about population and the taxi usage. To do all this analysis properly, we had to first convert all the location to some region so that it does make some sense. For that we have used Zillow shape files to convert taxi pickup and drop off location to NYC regions so that we can have some more broad analysis on data.

As the census data doesn't have region name and to map the census data to the given taxi data we had used a Tiger shape file to convert census area code to region name. To analyze taxi data with census data, we processed all the taxi data and mapped it to the region name after that we joined that data with census data. The final outcome of this process was we had region name, population and taxi usage =, economics connected together.

Technologies Used

To analyze data we have used a combination of different technologies. At the very first stage we had developed MapReduce to process data at a low level, then after that we used pig to do analysis on that that processed data.

We used amazon's S3 bucket and HPC's HDFS to store and retrieve data. To process all the data we used amazon's EMR service and HPS's Hadoop cluster's. As HPC was slow in processing very large file using MapReduce program, we processed all the data first using EMR then we switched HPC cluster.

After the first step of data processing, we used PIG on HPC cluster to analyze data. We used various charts on the d3. js library and Google Fusion Table to visualize our output.

Language: Python, Java (MapReduce), PIG

Storage: S3 bucket, HPC HDFS

Infrastructure: Amazon EMR, NYU HPC cluster, Cloudera

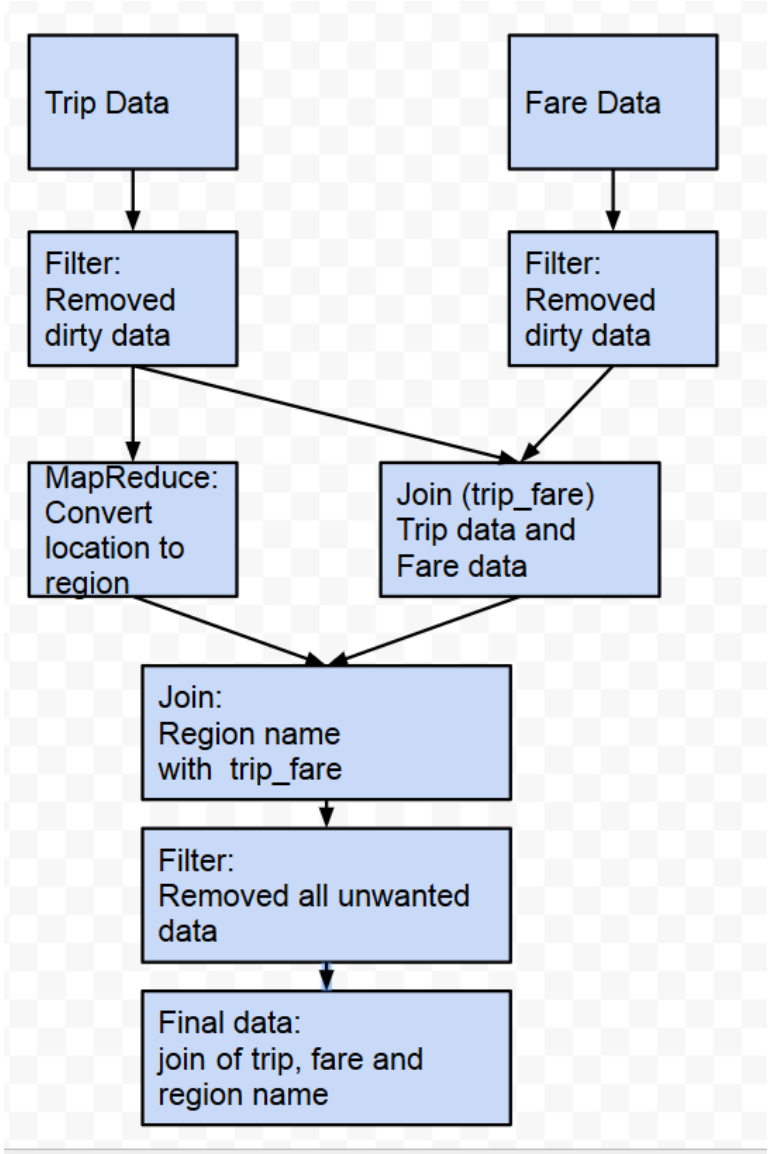
Data Used

We used to Zillow shape files to convert taxi location to NYC regions and Tiger shape file to convert census track code to regions.

- Taxi trip data
- Taxi fare data
- NYC census data
- Zillow shape files
- Tiger shape files

Data processing

1. Dumped all data to amazon s3 bucket
2. Processed location to region on trip data using amazon EMR and stored it on s3 bucket
3. Joined trip data and fare data using Amazon EMR (used pig to join data)
4. Filtered all the data
5. Transferred all the data to HPC HDFS
6. Used pig to analyze data on HPC cluster


Trip Data
Fare Data
Filter: Removed dirty data
Filter: Removed dirty data
MapReduce: Convert location to region
Join (trip_fare) Trip data and Fare data
Join: Region name with trip_fare
Filter: Removed all unwanted data
Final data: join of trip, fare and region name

- We coded a MapReduce program in Java to convert the pickup and drop off location of the taxi to the NYC regions.
- To map NYC population data to region name we used Tiger map files to convert census Track code to region name and then we combined this data with other data.

Analysis

1. *We calculated total amount, total fare and total tip of taxi trips for each hour of the day (0 -23), throughout the year*

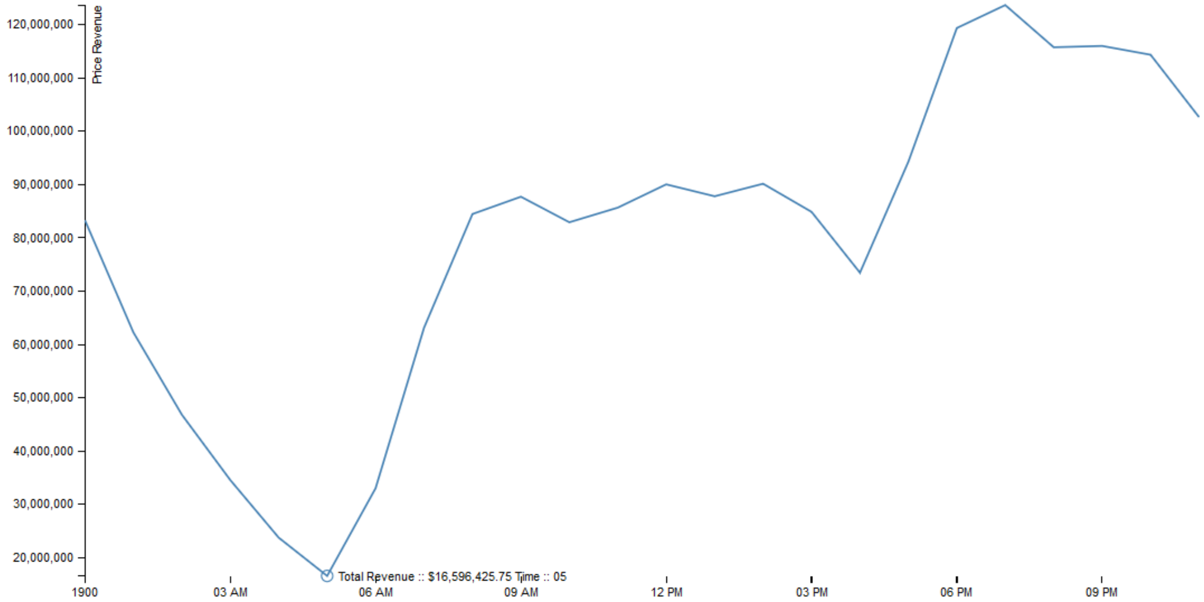


Fig 1. Total taxi revenue by hour

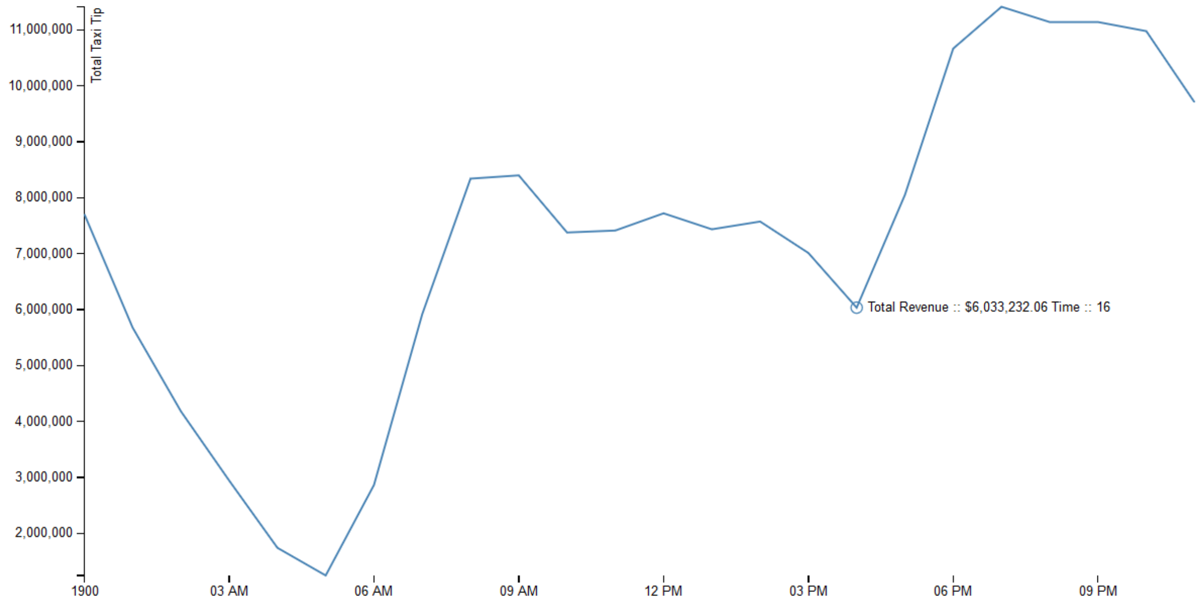


Fig 2. Total of taxi tip by hour for year 2013

Query Execution:

- Used to group by an hour filed on the date and calculated sum of total trip amount, fare and tip for every taxi trip

Query Conclusion:

- After analyzing the final outcome of the data we came to know that taxi usage and revenue is very less between 3 AM to 6 AM.
- And the total taxi tip is decreasing in comparing with taxi revenue between 9 AM to 1 AM.

2. *We calculated sum of total trip amount for each day for the year 2013.*

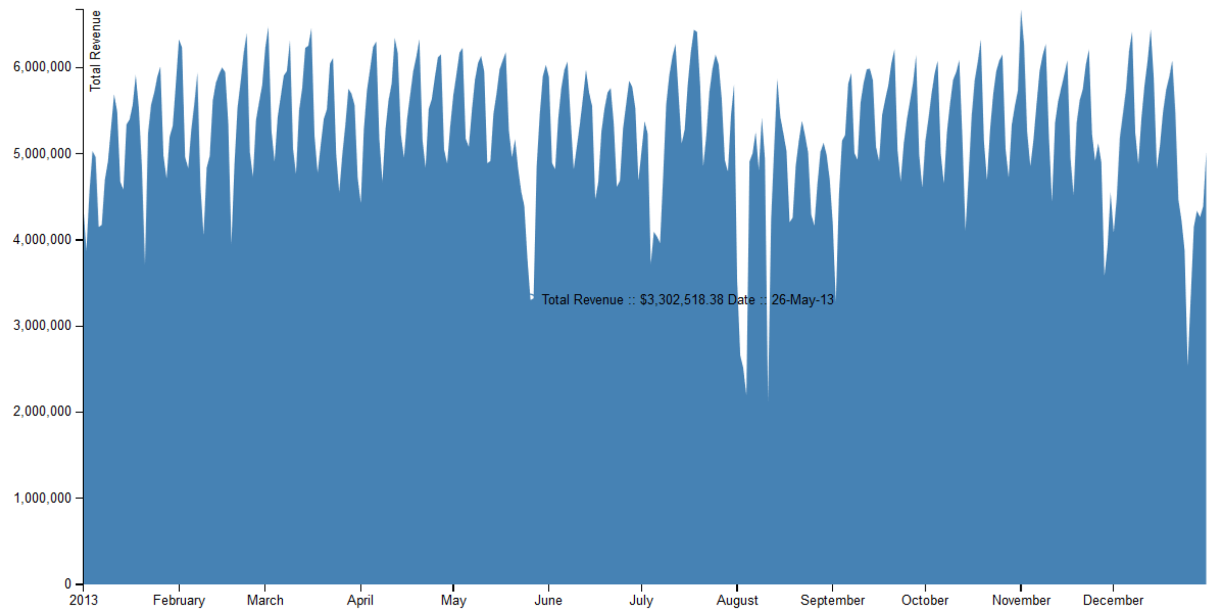


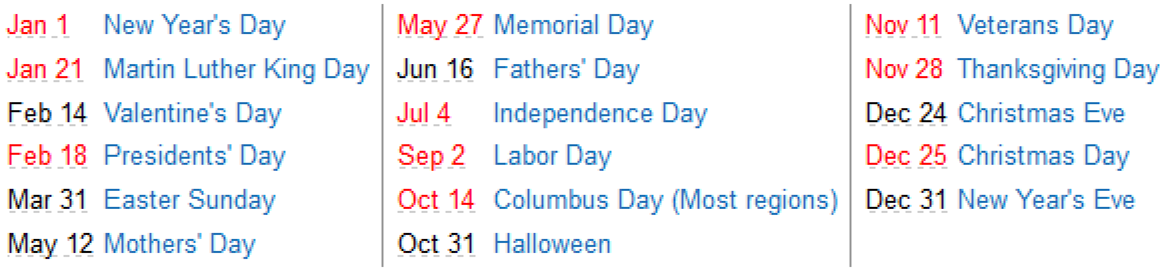
Fig 3. Total taxi revenue for each day (2013)

Query execution:

- We calculated sum of total_amount of each trip for each day of the year. In short group by on date and sum total_amount

Query conclusion:

- After visualizing the output we came to know that taxi revenue has a direct relationship with the public holidays
- List of holidays of 2013



Jan 1	New Year's Day	May 27	Memorial Day	Nov 11	Veterans Day
Jan 21	Martin Luther King Day	Jun 16	Fathers' Day	Nov 28	Thanksgiving Day
Feb 14	Valentine's Day	Jul 4	Independence Day	Dec 24	Christmas Eve
Feb 18	Presidents' Day	Sep 2	Labor Day	Dec 25	Christmas Day
Mar 31	Easter Sunday	Oct 14	Columbus Day (Most regions)	Dec 31	New Year's Eve
May 12	Mothers' Day	Oct 31	Halloween		

-
- As we can see that taxi revenue is very low compared to other days of the holidays like Jan 21, May 27, Sep 2, Dec 25…

3. *Popular (Top 10) pickup and drop off regions of the NYC*

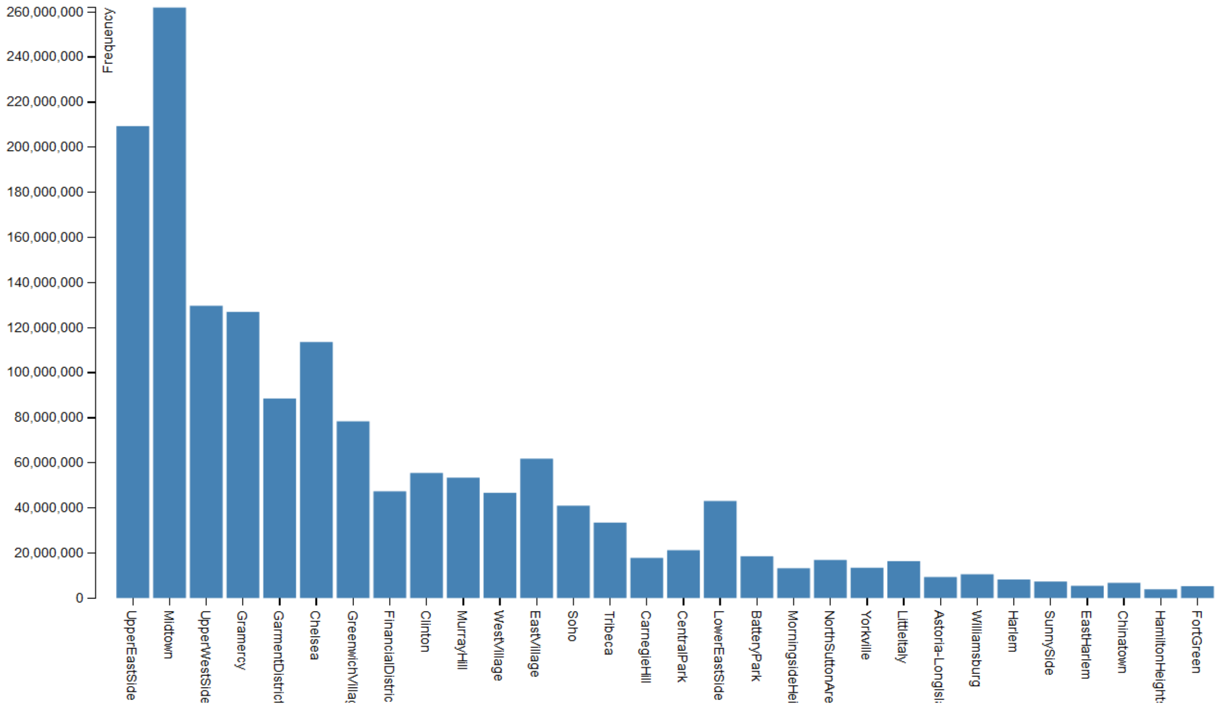


Fig 4. Total taxi pickup regions (Top 30)

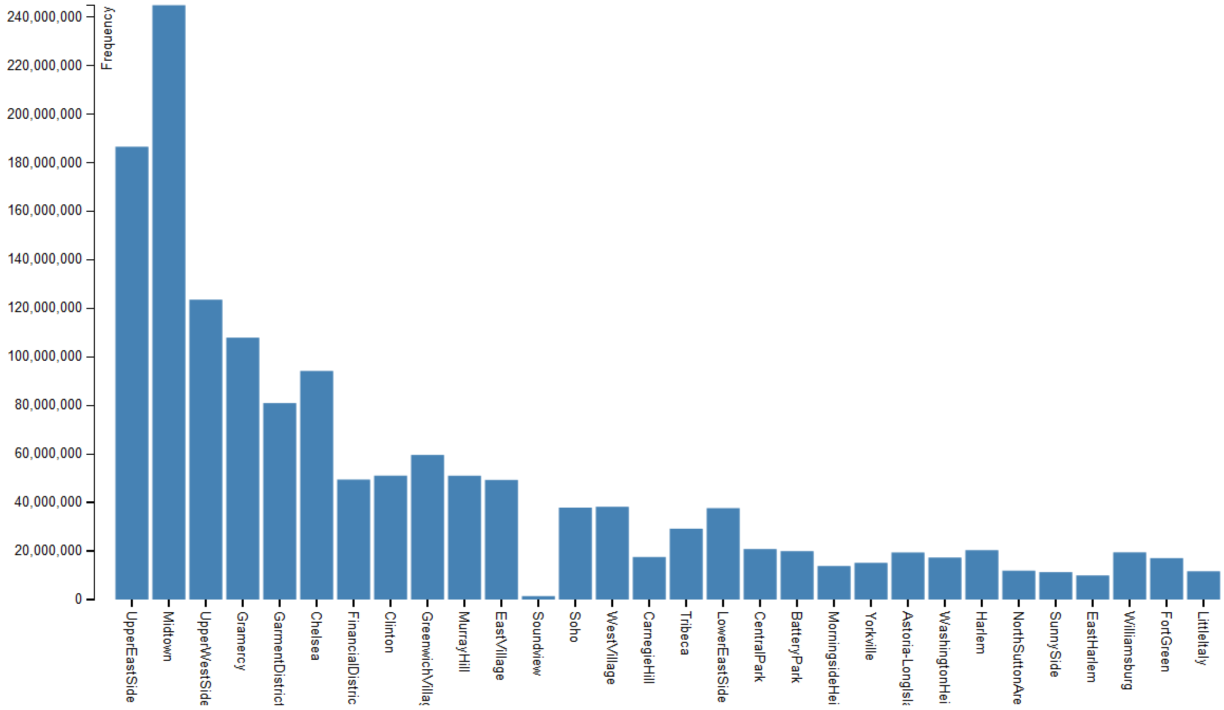


Fig 5. Total taxi drop off regions (Top 30)

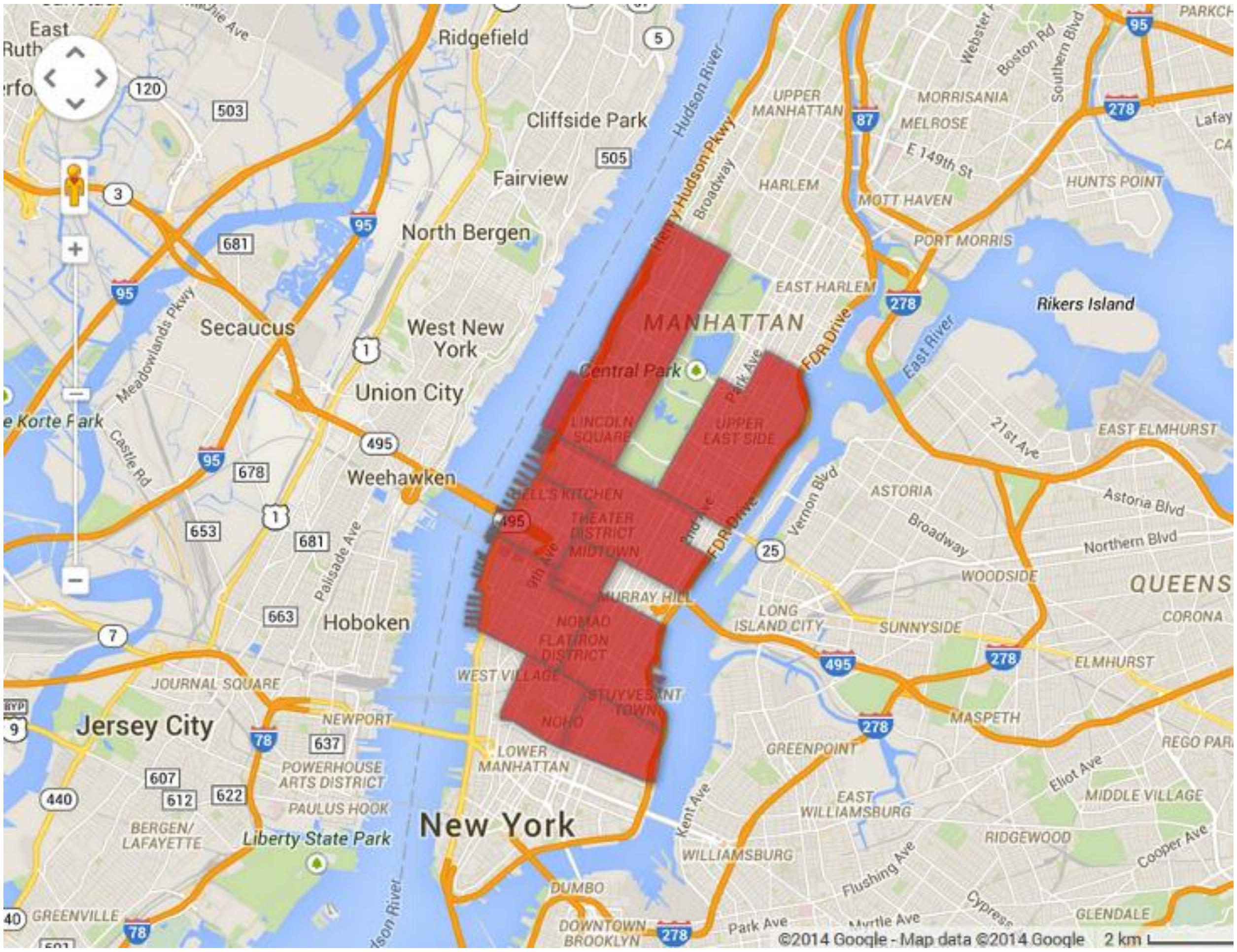


Fig 6. Total taxi picks up regions (Top 10)

Query execution:

- Group by on region and sum of total_amount and total_taxi_ride (year 2013)

Query conclusion:

- After analyzing the data we came to know that popular midtown and upper Manhattan is famous for taxi pickups and drop off.
- Another thing we analyzed that there is only one region financial district, which is in the top 10 of taxi drop off regions.

4. *We calculated popular regions on the basis of Avg tip paid by the taxi rider.*

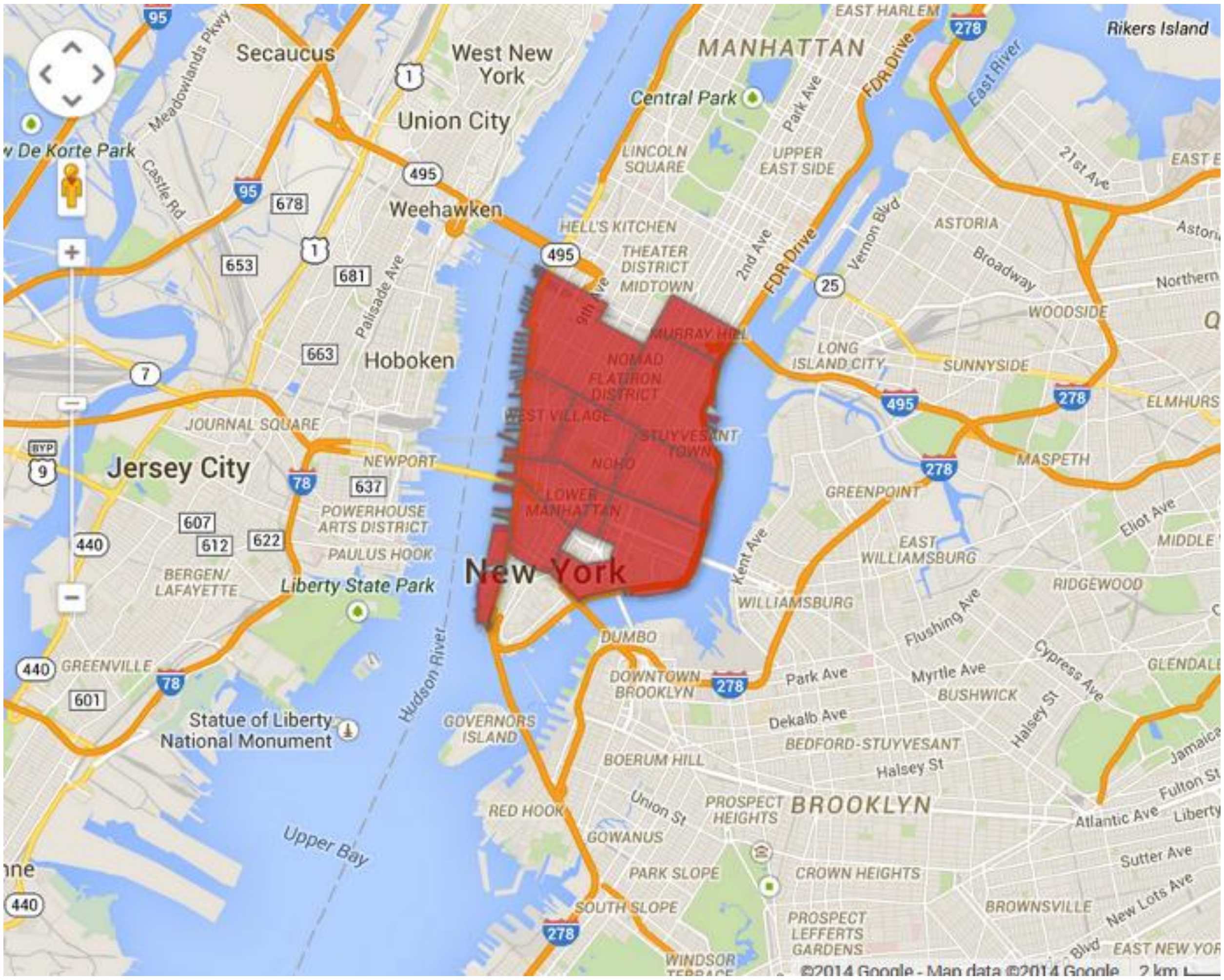


Fig 7. Popular regions for average taxi tips (top 10)

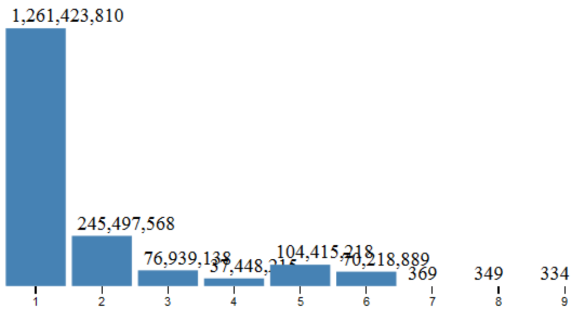
Query execution:

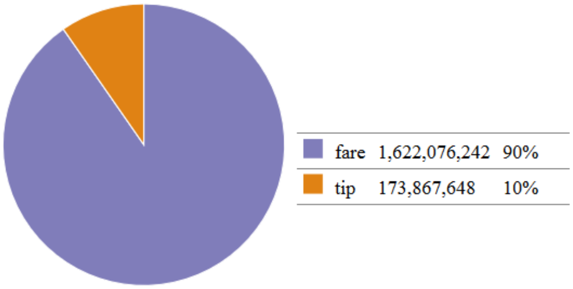
- We calculated percentage tip for each ride. After that we had done group by on the region and average at percentage tip.

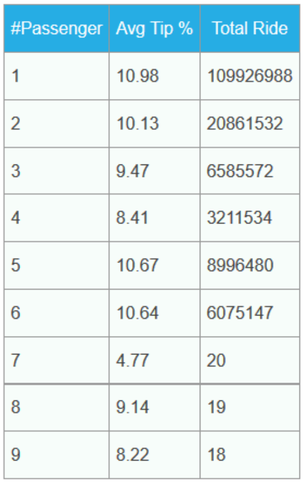
Query conclusion:

- After visualizing the data we came to know that though upper and middle Manhattan has more total revenue than lower Manhattan, regions in lower Manhattan have a high Avg tip per trip than other parts of the Manhattan

5. Analysis on passenger count and tip and total ride







#Passenger	Avg Tip %	Total Ride
1	10.98	109926988
2	10.13	20861532
3	9.47	6585572
4	8.41	3211534
5	10.67	8996480
6	10.64	6075147
7	4.77	20
8	9.14	19
9	8.22	18

Fig 8. Total fare,tip,avg tip and total ride for passenger count

Query execution

- First, we calculated the tip percentage for each trip
- Then we did group by on passenger count and sum on total fare and tip and Avg on tip percentage and count on total ride

Query conclusion

- 70% rides are of single passengers
- As we came to know that the number of passengers doesn't directly influence the Avg tip of the ride
- And for passenger count 1, Avg tip is consistent 10-11% all the month, while we may see some difference in the other but not that much.

6. Analysis of distance of the taxi ride

Query execution:

- Count total rides where a mile is less than 1

Query conclusion:

- We found that the total number of small rides are very large, so tried to find that the total riders who has ride distance less than 1 mile

Month	Rides < 1 mile	Total Rides
1	3247824	13323199
2	3085528	12645126
3	3355685	14150935
4	3161018	13539406
5	3042474	13132040
6	2910958	12802572
7	3090684	13478527
8	2309724	11242962
9	2834241	12609899
10	3055508	13437434
11	3021200	12846311
12	36136044	142515316

7. Population and taxi revenue NYC region wise

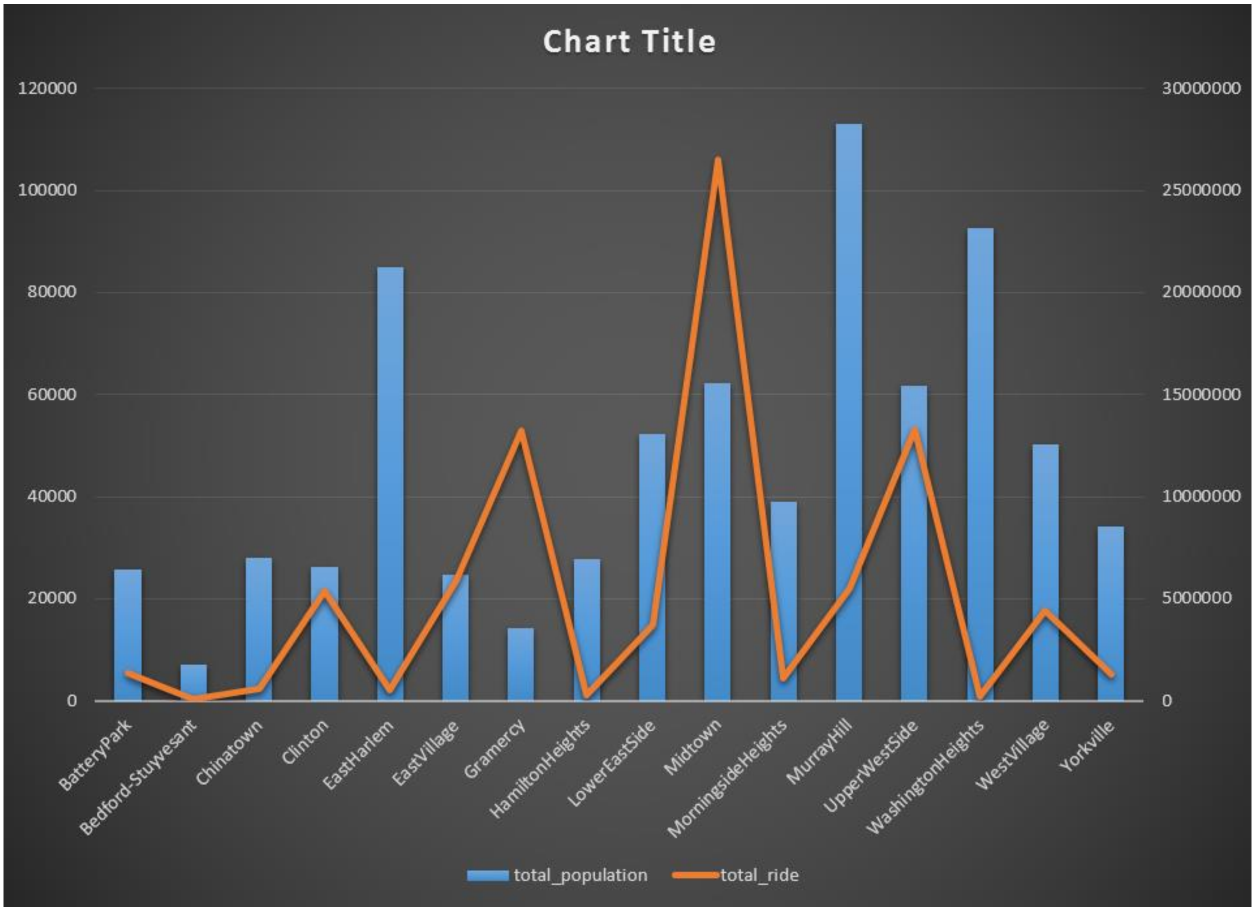


Fig 9. Chart for total population vs total_ride for some regions

Query execution:

- First we computed total ride, fare, tip fro each region
- Then we computed population of all the regions using census data
- Then we combined both the data

Query conclusion:

- We can see clearly that some region has more population than other but total ride count is less than other regions
- For example harlem has more population than Gramercy but less ride count
- So we concluded that population doesn't that much affect the taxi rides

8. Revenue and tip behaviors for top earning taxi drivers

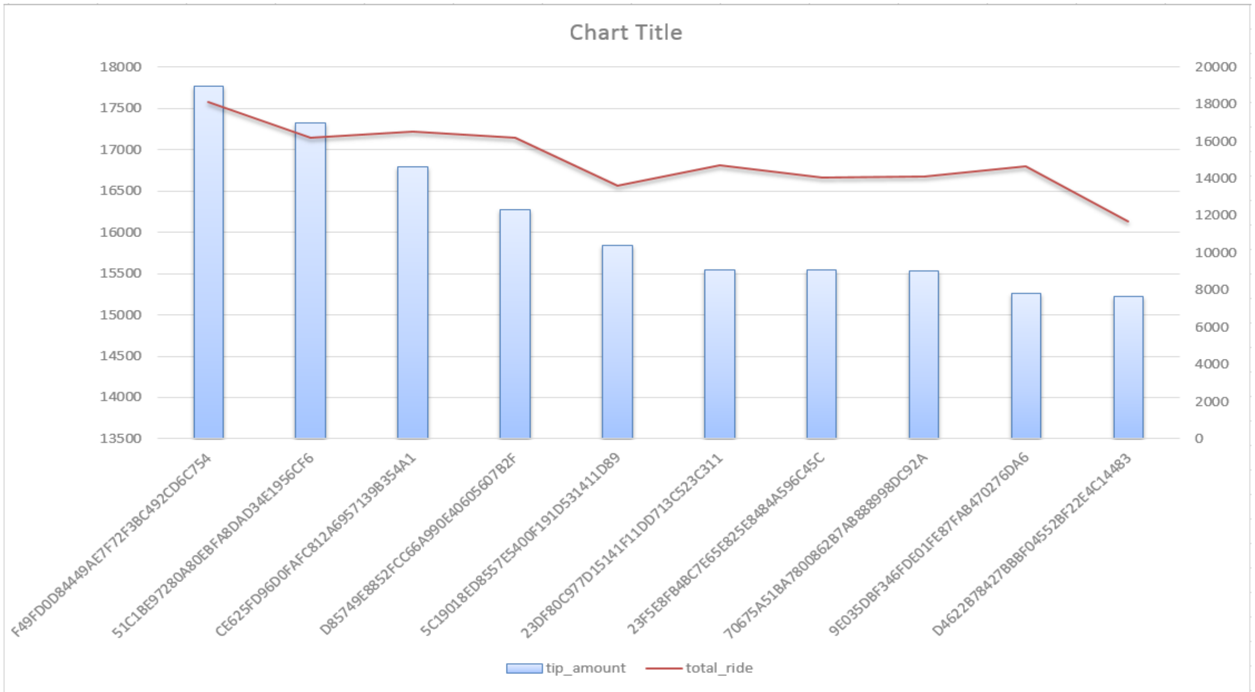


Fig 10. Chart for total_tip vs total ride for top tip earning drivers

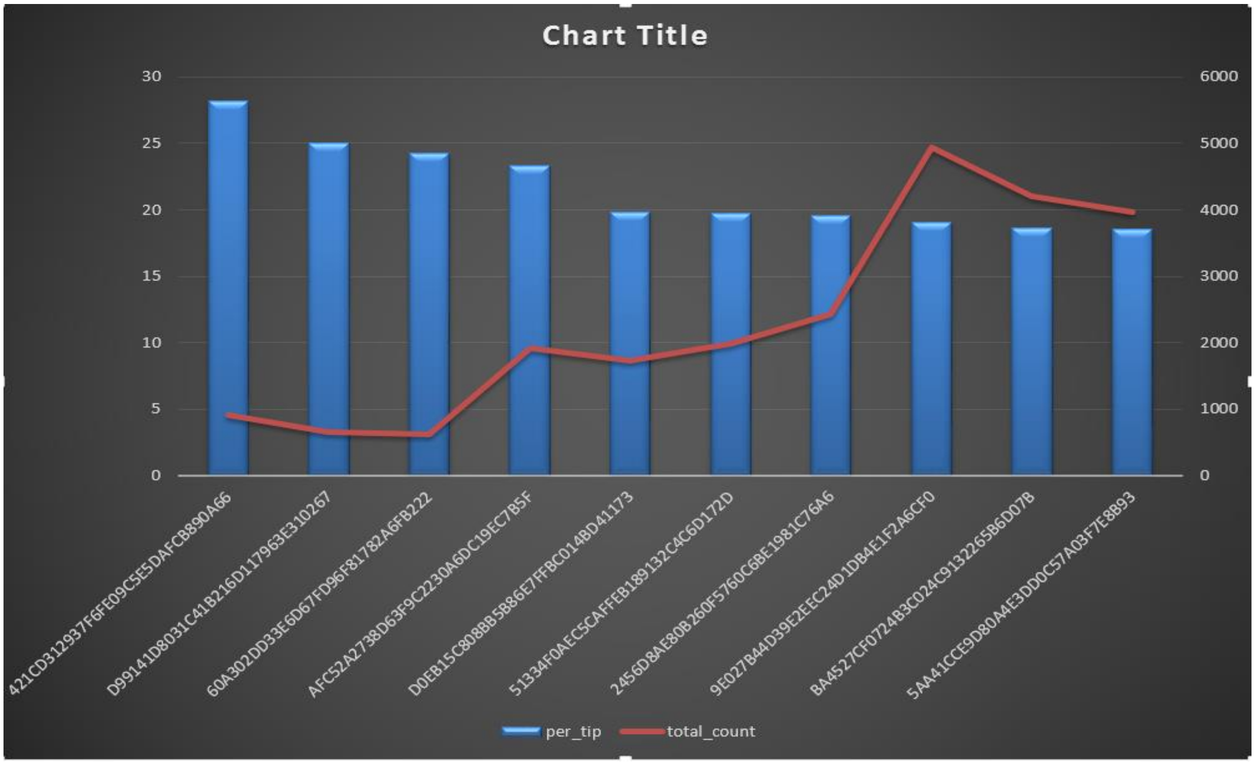


Fig 11. Chart for avg percentage tips and total ride for top avg tip drivers

Query execution:

- First we caculated percentage tip for each ride
- Then we have done group by on hack_license and sum on total fare,tip,ride
- Then we fecthed top 10 drivers who has earned most tips and other 10 drivers with most avg tip

Query conclusion:

- After calulating total tips of this drivers we queried data for where this drivers mostly has pickups and we found these regions UpperEastSide, UpperWestSide, Chelsea, Gramercy, Midtown
- So we are concluding that driver driving in this region are likely to earn more money

Technical difficulties

1. It was difficult to orient with the HPC as none of us was used to it. But later we found it very convenient to use the cluster.
2. We had problem in integration of the shape files with the trip and fare data files as the join was time consuming and the job would exit on HPC. We used AWS EMR to tackle this problem. Here we could easily add more machines and reduce the time take to run the job. Later we transferred the data/files to HPC and started the analysis part.
3. We found combining the census data with the trip data was not easy. As the shapes of the census are different than the shapes of the trip data. We split the regions in the census data and then joined the files.
4. Few of the data type in Pig are difficult to handle like Date. But eventually we managed to convert all the dates to one format which made it easy to group by date.

Contribution

Team members

Savan

- Conveting location to region usiing zillow shape file
- Pig scripts
- Report
- Analysis on taxi economics
- Visualization using d3.js

Aditya

- Pig scripts
- Cleaning the data
- Joining of different data sets
- Report
- Analysis on taxi tip
- Visualization using d3.js

Aniket

- Pig scripts
- Converting census track code to location
- Joining census data with taxi data
- Analysis on census data
- Visualization using d3.js
- Final presentation

Final conclusion

Analyzing such a big data needs a lot of resources. Each and every step needs to be done carefully as playing with the live data is too critical. There is a lot of dirty data which you have to exclude to get consistent result. Almost half of the efforts should be spent before starting the analysis. Analysis should be based on some facts and not on trial and error as the data is big and querying it needs a lot of time.

GitHub link : https://github.com/SavanRupani/BigData