Analytics Project Presentation - Spring 2016

Analytics Project: Predictive analysis and geolocation mining of taxi data in New York City

Abstract: With the advent of variety of many open datasets being made available in public domain, different analytics can be drawn from them. These datasets are often huge and require use of big data technologies like MapReduce, HDFS for data processing and storage. We analyze the New York City Taxi & Limousine Commission datasets (for the year 2015) for different analytics that can be useful for the passengers as well as the taxi drivers.

Predictive analysis and geolocation mining of taxi data in New York City

Motivation

Who are the users of this analytic?

- People who are *currently driving* taxis.
- People who *want to drive taxis*.

Who will benefit from this analytic?

The primary beneficiaries are the people who will be using this analysis (taxi drivers and people wanting to drive taxis). Other than them, using the results of our analytics, many other firms can benefit. For example, looking at our analysis related to the payment method by taxi commuters, credit card companies can come with special discounts or schemes that will benefit not only their customers but also increase their profit.

Why is this analytic important?

- These analysis will help the above users to decide in which area a cab is most likely to be hired; pickups from which area gets more tips and so on.
- It will also help people wanting to be taxi drivers and are confused whether to drive a yellow cab or green. Looking at these analytics, he/she will make a much more informed decision.

Predictive analysis and geolocation mining of taxi data in New York City

Goodness

What steps were taken to assess the 'goodness' of the analytic?

We looked at the results of each and every analytics and found that they perfectly align with the general perception of people and common logic. For example,

- As we all know that people living in areas like Lower East side, Upper East side and Upper West side belong to a higher income group. The result to one of the analysis showed that these people tend to be more generous than the rest.
- In time intervals which fall under either office hours or the time when people go out and hang out with friends or colleagues, more taxis are hired. This was confirmed by our results.
- Since most of the corporations/companies, shopping areas, Central Park, tourist places, etc. are located in Manhattan, people travel to these areas daily. People from areas like Port Richmond and South Shore which are located very far from Manhattan, cover more distance resulting into more average fare. This was also confirmed by our results.
- Since people don't generally carry a lot of cash with them, trips with high fare are paid through credit cards rather than cash.
- We all know that many tourists come during summer. Our results show that during these months, the no. of cabs hired are more than the rest. Similar result has come for peak winter months when people don't prefer to walk or take public transport.

Predictive analysis and geolocation mining of taxi data in New York City

Data Sources

Name: The official TLC trip record dataset – Yellow Cabs.
(http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)

Description: NYC TLC provides taxi data for different years. Some key fields of interest for this project are:

- Payment type
- Total amount
- Tip amount
- Toll amount
- Trip Distance
- Pickup date & time
- Drop off date & time
- Pickup longitude
- Pickup latitude

etc…

Size of data: In GBs. (~ 1 Million entries per month, and our project is based on full year 2015)

Name: The official TLC trip record dataset – Green Cabs.
(http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)

Description: Same data source as described above.

Size of data: In GBs. (~ 1 Million entries per month, and our project is based on full year 2015)

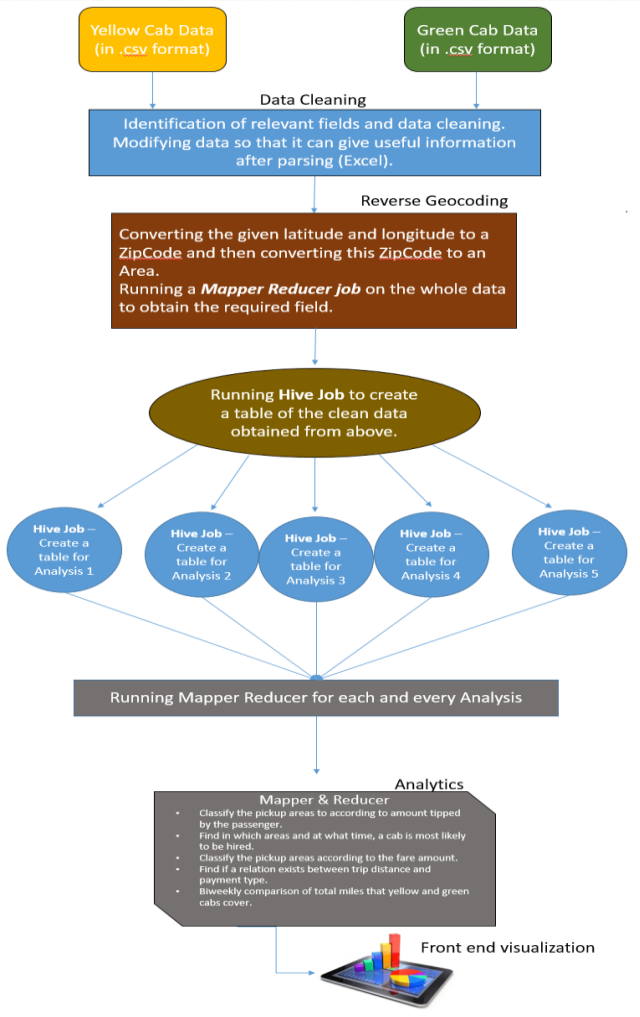
Predictive analysis and geolocation mining of taxi data in New York City

Design Diagram

The data from the two sources (NYC yellow and green taxi data) is collected and then pickup latitude and longitude is converted to neighborhood.

Now the resulting data has the field neighborhood instead of latitude and longitude. Then from this dataset, columns specific to each analytic is extracted using Hive and MapReduce and sent to corresponding Mappers.

Once Mappers/Reducers for each analytic finish their computation, the results are collected. The results are then plotted appropriately using clustered column charts, stacked column charts and scatter plots.



Platform(s) on which the analytic ran:

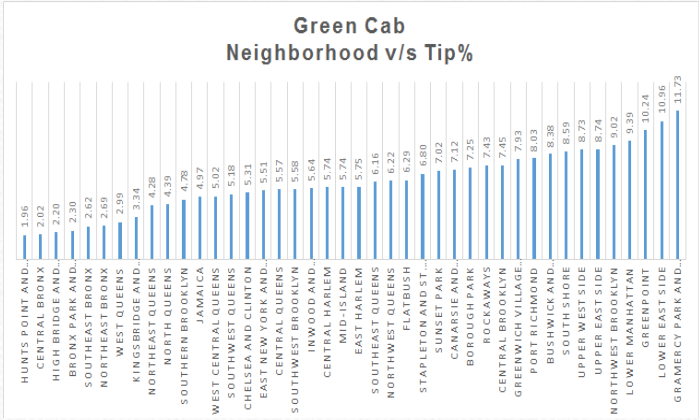
We ran our whole analysis on Quickstart VM.

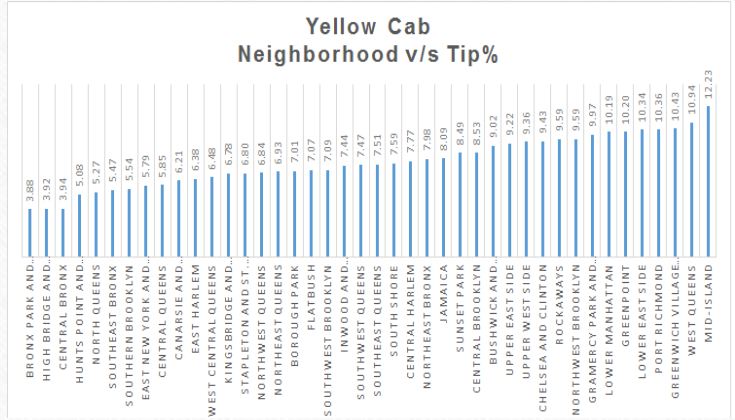
Results

Predictive analysis and geolocation mining of taxi data in New York City

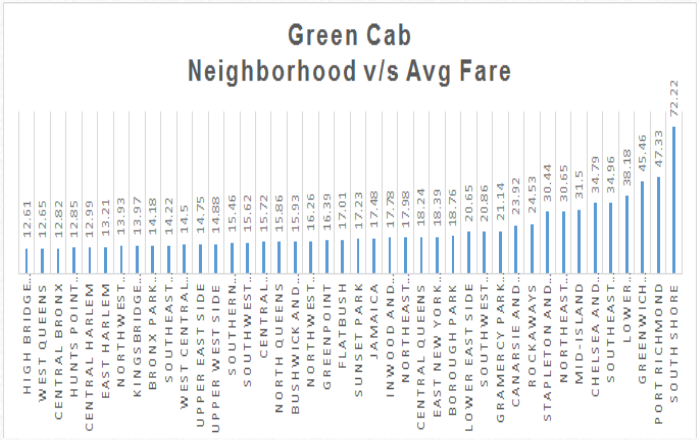
Result 1.

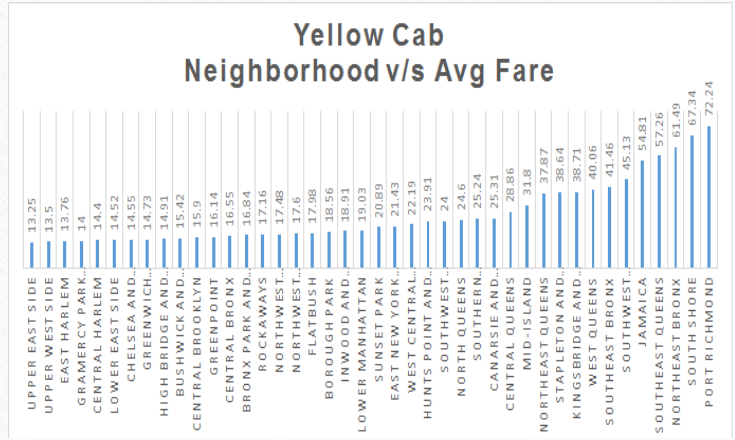
- In areas like Lower East side, Greenpoint, Upper East side and Upper West side, people tend to be more generous and give more tips. Whereas in areas like Bronx (All parts), Highbridge and Morrisania, Bronx park and Fordham and certain areas of Queens, people tend not to give much tip.





- It can be seen that areas like Port Richmond and South Shore have very high average fare. Whereas areas like Upper East Side, Upper West Side, Lower East Side, etc. have very low average fare. This trend could be mainly because most of the corporations/companies, shopping areas, Central Park, tourist places, etc. are located in Manhattan. Hence people travel to these areas daily. Port Richmond and South Shore are located very far from Manhattan, hence cabs cover more distance resulting into more average fare.

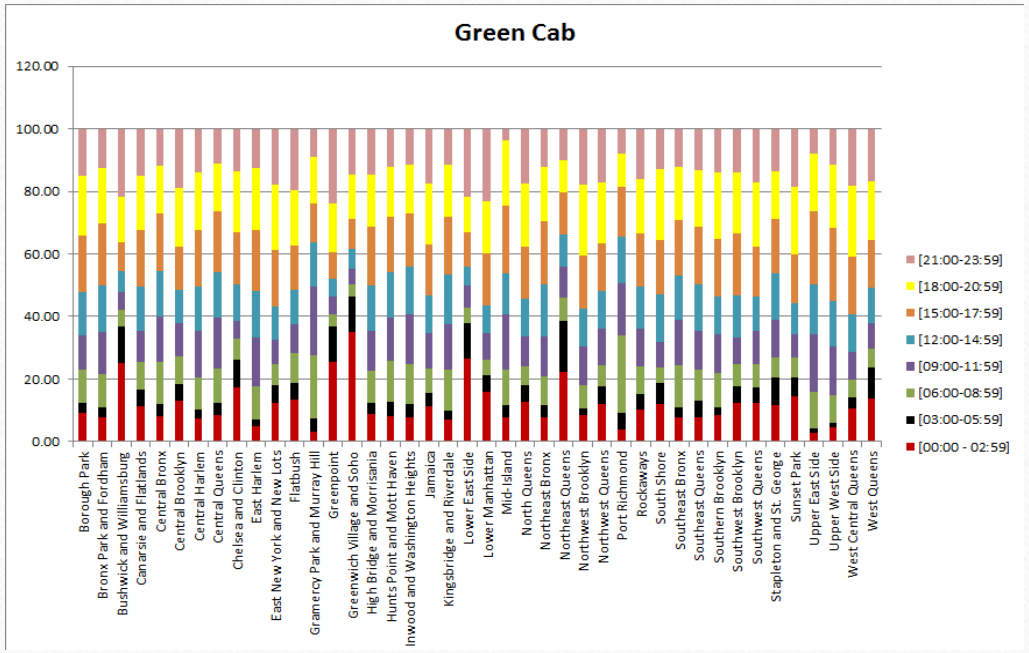


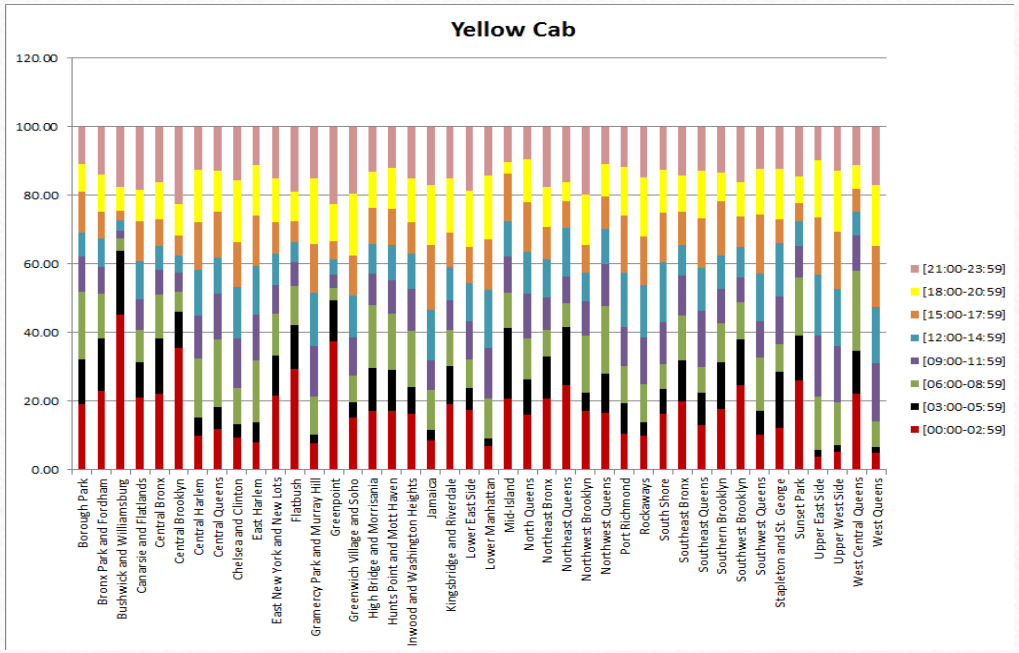


Predictive analysis and geolocation mining of taxi data in New York City

Result 2.

- As can be seen from the graphs, high number of cabs are hired in the time intervals 15:00 – 21:00, 06:00 – 09:00 across all neighborhoods. This could be mainly because these intervals fall under either office hours or the time when people go out and hang out with friends or colleagues.
- In areas like Greenwich Village and Soho, Lower East side and Bushwick and Williamsburg, no. of taxis hired during the time interval 00:00-3:00 is very high. This can be related to the fact that these areas consist of bars, pubs and restaurants where people hangout during the night.

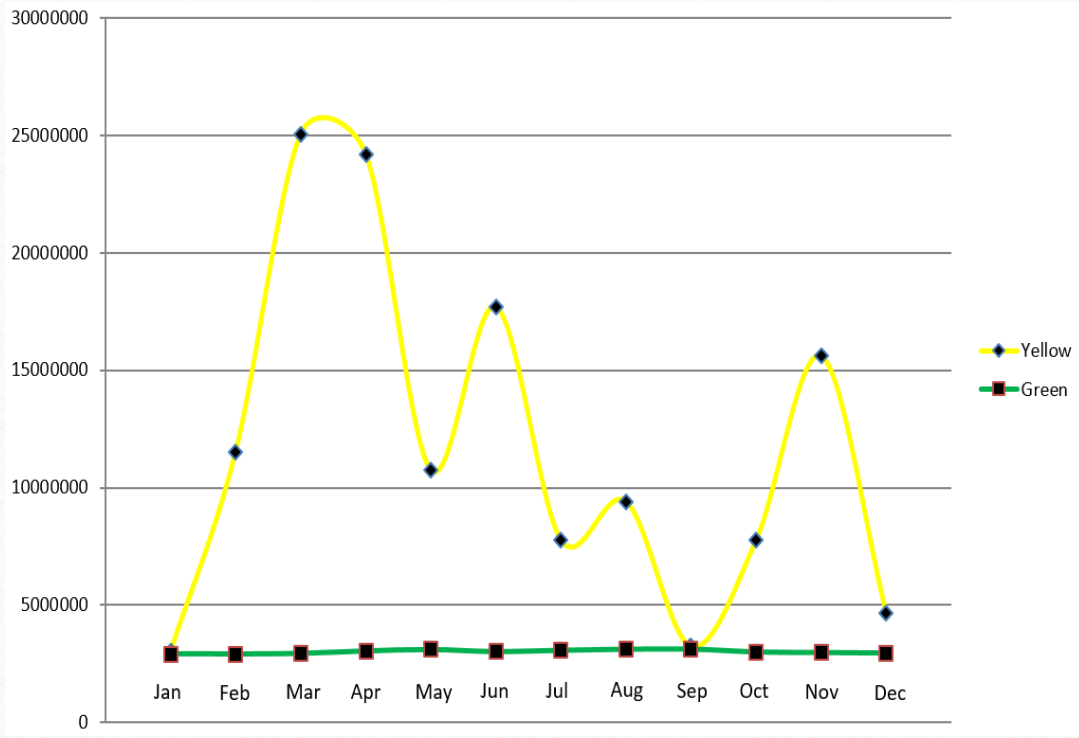




Predictive analysis and geolocation mining of taxi data in New York City

Result 3.

- From this graph, after comparing the total miles covered by yellow cabs and green cabs, it can be observed that total miles covered by yellow cabs are way more than that covered by green cabs throughout the year. This could be because yellow cabs are hired more than green cabs.
- Moreover, it can be seen that yellow cabs have a peak during peak winter months and summer months. This could possibly be because during winters, people tend to travel more through cabs more than walking or taking public transport. During summers, NYC is filled with tourists and they prefer taking cabs than public transport.



Predictive analysis and geolocation mining of taxi data in New York City

Obstacles

1. For our analysis, the first step was to convert the given latitude and longitude into an area / neighborhood. We initially used a Google API in which we sent the longitude and latitude and it used to return the area. But when we ran our mapper reducer on the whole data, we learnt that there is a limit to the number of requests that can be sent to this API from one IP address per day. That limit was very small (50,000). We are dealing with around 24 million entries, hence we couldn't use this.

2. Like in any team project, there were a lot of clash of ideas. Sometimes, we both had different opinions and methodology to do a certain task. We both had a different style of coding and hence it took time to get accustomed to working in a team.

Predictive analysis and geolocation mining of taxi data in New York City

Summary

On the basis of these analytics,

- We are able to classify areas as generous or not on the basis of amount of tip given.
- We are able to find in which areas and at what time, a cab is most likely to be hired.
- We are able to classify the pickup areas according to the fare amount.
- We are able to find a relation between trip distance and payment type.
- We are able to find interesting inferences after comparing the total miles that yellow and green taxis cover.

Acknowledgements

We would like to thank NYC TLC for providing us with the data without which we couldn't have had run the analytics to get the desired results. We would like to thank Suzanne McIntosh for giving us the opportunity to explore these new technologies in the industry and equipping us with proper skillset to use these technologies. We would also like to thank Cloudera for providing us with the QuickStart VM so that we were able to run everything on our hardware. We would also like to extend thanks towards Google for providing us with the Open Source API for Maps.

Predictive analysis and geolocation mining of taxi data in New York City

References

- https://en.wikipedia.org/wiki/Taxicabs_of_New_York_City
- Timothy H. Savage, Huy T. Vo. Yellow Cabs as Red Corpuscles.
- Xianyuan Zhan, Xinwu Qian, Satish V. Ukkusuri. Measuring the Efficiency of Urban Taxi Service System.
- M. Anil Yazici, Camille Kamga , Abhishek Singhal. A Big Data Driven Model for Taxi Drivers' Airport Pick-up Decisions in New York City.
- J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implemenation, 2004.
- http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
- T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
- https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm
- https://boutell.com/zipcodes/

Predictive analysis and geolocation mining of taxi data in New York City

Thank you!