Big Data Analytics Symposium - Fall 2019

Analytics Project: Analysing New York Taxi Data to Determine Affluence and Road Safety of New York Neighbourhoods

Team:

Praveen Oak

Alankrith Krishnan

Abstract: This analytic aims to use taxi data, monthly rent data and camera and parking violations data as an indicative measure to analyse the relationship between taxi cab usage for a neighbourhood and its median rent, and the taxi cab usage with violations in the neighbourhood. The aim is to analyse and look up interesting insights that can be learnt about neighbourhood rents and safety of taxi drivers based on the taxi data.

Motivation

Who are the users of this analytic?

Home buyers, Taxi drivers, City planners, Traffic police, Business owners

Who will benefit from this analytic?

Home buyers, Taxi drivers, City planners, Traffic police, Business owners

Why is this analytic important?

Taxi data is a vast, constantly updating real time statistic that is open and free for the public. Hence, any analytic derived from taxi data is future proof, especially for fields where data loses value with time. Eg Real estate, traffic violations. This is applicable for monthly rent data and traffic violations - and hence they can be extended for the future to check for results in the future.

Goodness

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

For the tip percent vs neighbourhood analysis, we assessed the results based on 2 metrics.

1. The average tip percent derived from data seemed to match tipping trends in NYC(~15%)

2. A similar study done by GrubHub for tips for home deliveries tend to match our analytic in many cases.

https://firstwefeast.com/eat/2014/06/nycs-best-and-worst-tippers-by-neighborhood

3. The data for correlations was normalised and plotted to see if the trends were visible on the graphs - and they seem to fit the general trend.

Analysing New York City Taxi Data

Data Sources

Name: NYC TLC Open dataset.

Description: Dataset contains the yellow cab taxi ride data for every ride since Jan 2009 upto 2019. Contains information like pickup and dropoff area, fare amount tip etc.

Size of data: 10-20GB

Name: Streeteasy Median Rental Open Dataset

Description: Contains the median rent for each neighbourhood in NYC on a monthly basis as a time series. Data available since 2009.

Size of data: < 1MB

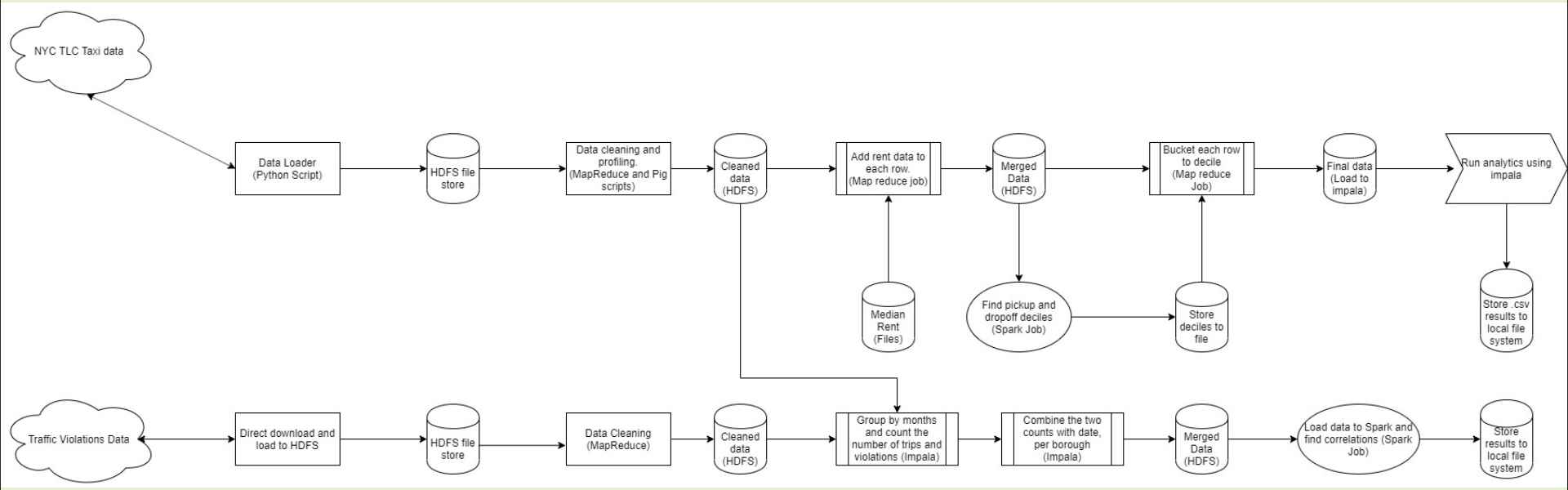
…

Name: Open Parking and Camera Violations

Description: Contains every single traffic violation made in New York on a daily basis from 2016.

Size of data: 12GB

Design Diagram



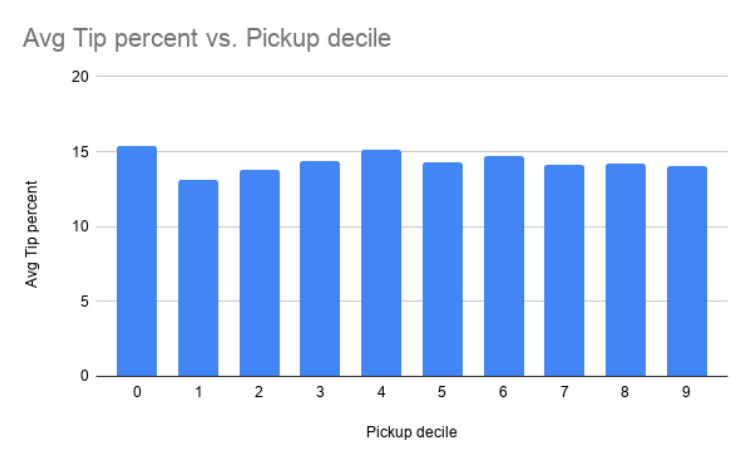
Trained on NYU HPC - Dumbo from start to finish

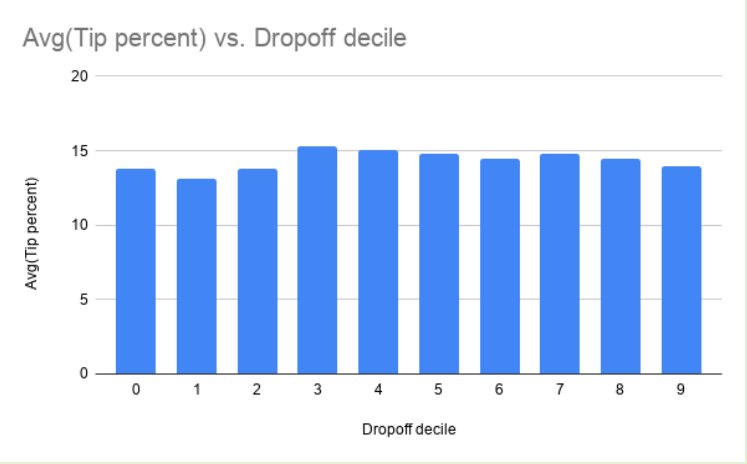
Results

1. Average tip percent across different rent category neighbourhoods does not vary a lot(because of averaging effects across large areas)
2. However, affluent neighbourhood dominate best ranked neighbourhoods and poorer neighbourhoods dominate worst ranked neighbourhoods, based on average tip percent.
3. Correlations between number of taxi trips and number of traffic violations shows that there is a negative correlation for Manhattan, positive correlation for Queens and Staten Island, and no correlations for Bronx. This means that taxi drivers are safer than self drivers in Manhattan, and the inverse is true for Queens and Staten Island, while Brooklyn and Bronx have equal driving problems.

Analysing New York City Taxi Data

Average Tip Percent vs Pickup and Dropoff Decile





Analysing New York City Taxi Data

Best and Worst Tips and their Localities

Best pickup tip	Locality ID	Avg(Pickup rent)	Locality Percentile	Locality Name
17.3	138	1745	0	La Guardia(Queens)
15.9	125	No available data	0	Hudson Sq(Manhattan)
15.7	234	No available data	0	Union Sq(Manhattan)
15.6	249	3739	6	West Village(Manhattan)
15.6	52	3068	4	Cobble Hill(Brooklyn)
15.5	90	4624	8	Flatiron(Manhattan)
15.5	107	3512	6	Gramercy(Manhattan)
15	13	4376	8	Battery Park(Manhattan)
15.3	113	3704	8	Greenwich Village North(Manhattan)
15.2	158	3737	6	Meatpacking district(Manhattan)

Worst pickup tip	Locality	Avg(Pickup rent)	Decile	Locality Name
3	35	1843	0	Brownsville(Brooklyn)
3.4	173	1939	0	North Corona(Queens)
3.5	39	2085	0	Canarsie(Brooklyn)
4	69	1699	0	East Concourse(Bronx)
4.1	26	1798	0	Borough Park(Brooklyn)
4.4	177	3162	4	Ocean Hill(Brooklyn)
4.4	76	2122	0	East New York(Brooklyn)
4.5	168	2087	0	Mott Haven(Bronx)
4.7	83	1946	0	Elmhurst(Queens)
5	92	2000	0	Flushing(Queens)

Analysing New York City Taxi Data

Manhattan (Correlation = -0.224009)

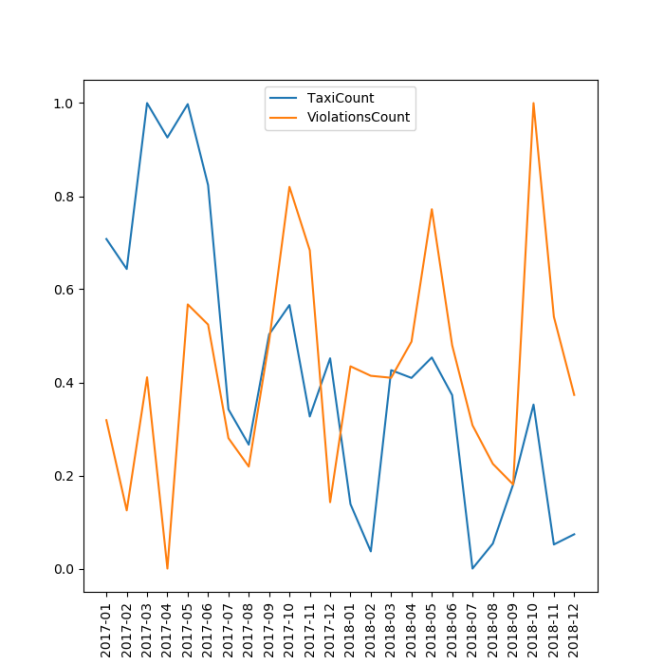
Analysing New York City Taxi Data

Queens (0.223480)

Staten Island (0.275796)

Analysing New York City Taxi Data

Brooklyn (0.000756)



Bronx (0.089859)

Obstacles

1. Lot of bad rows in taxi dataset. Which meant results were not making sense in the beginning. Also, the data format changed in one of the years, which meant that data processing had to account for this.
2. The boroughs were mapped with different ID's on the traffic violations dataset - Manhattan was referenced with MN and NY, Brooklyn was referenced with K and BK, and Queens was referenced with Q and QN.

Analysing New York City Taxi Data

Summary

New York Taxi data is robust and real time source of data. While the single data point alone will probably not be able to predict complex variables like rent, it will serve as one of the data points that can be used to predict the prices of rent.

The correlations between the number of taxi trips and the number of traffic violations can help with understanding which boroughs you should be driving in, and which boroughs you should be taking taxis in, based on the relative safety of taxi drivers in the neighbourhood.

Acknowledgements

NYU HPC
NYC Open Data and StreetEasy
Prof.McIntosh
Graders Omkar Patinge and Srishti Grover

References

1. Stoyanovich, Gilbride, Moffitt. Zooming in on NYC taxi data with Portal, 2017
2. Nadai and Nepri. The economic value of neighborhoods: Predicting real estate prices from the urban environment, 2018
3. Zamani and Schwartz.Using Twitter Language to Predict the Real Estate Market, 2017.
4. Yellow taxis have fewer accidents than blue taxis because yellow is more visible than blue (Teck-Hua Ho, Juin Kuan Chong, and Xiaoyu Xia), 2017
5. NYC TLC Open dataset. Link: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
6. Streeteasy Median Rental Open dataset. Link: https://streeteasy-market-data-download.s3.amazonaws.com/rentals/All/medianAskingRent_All.zip
7. Open Parking and Camera Violations Dataset. Link: https://data.cityofnewyork.us/City-Government/Open-Parking-and-Camera-Violations/nc67-uf89

Thank you!