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

Praveen Oak
Department of Computer Science
New York University
ppo208@nyu.edu

Alankrith Krishnan
Department of Computer Science
New York University
ak7380@nyu.edu

***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.

I. Introduction

Cabs are an essential form of transportation in the city of New York. They operate in every nook and corner of the city and are used for varying sections of society. This makes them one of the important metrics of spending pattern of the cities residents. The Taxi and Limousine Commission of New York makes available all of the commuting data regarding yellow(and green) cabs from 2009 upto present freely available on the internet. In addition, we also have the median rent data for all of New York's neighbourhoods for every month back upto 2009. We decided to use these two datasets to analyse if there is a co relation between spending patterns of commuters in each neighbourhood and the median rent of that neighbourhood. The safety of a neighbourhood also depends on the number of traffic violations happening in the neighbourhood (since traffic violations could lead to accidents), and hence we analyse whether the taxi drivers in each neighbourhood contribute to the violations happening in the neighbourhood as well.

The end goal of the project was to analyse the data to see if there are any interesting insights we can gather from the taxi usage and fare patterns for a neighbourhood and the median rent for the neighbourhood. The taxi dataset had a lot of information for each taxi ride, but four datapoint were of particular interest to us. They were pickup neighbourhood, drop-off neighbourhood, fare amount and tip amount. To start, we decided to look to test the hypothesis which most of us have about taxi riders, that more affluent riders generally tip better than less affluent ones. For this, we used the median neighbourhood rent of the pickup and drop-off location as an indicator of how affluent the taxi rider is. Using this information we tried to analyse how the tip percent varies based on the median rent of the neighbourhood.

The dataset also contained information of what payment type was used to pay for the ride eg : credit card, cash or pre booked. With this information we wanted to test out another hypothesis, that more affluent riders generally tend to pay for rides electronically while less affluent ones prefer to pay in cash.

Using the taxi trips, we can also count the raw number of trips taken per month in each neighbourhood, and we compare this with the number of traffic violations happening in the neighbourhood per month to see if there is a correlation between the number of trips being taken and the traffic violations. The amount of correlation will tell us how safe the taxi drivers in the neighbourhood are - and by extension - how safe normal drivers are in the neighbourhood, depending on whether the correlation is positive or negative.

II. Motivation

Predicting rent prices especially in a competitive real estate market like the city of New York can be a valuable tool. Such a model will be helpful in pricing real estate valuations more accurately, which can be useful for developers. It can also help out landlords prince their properties as per the market. Unlike other fields, say retail, the real estate industry is largely unorganised and decentralised. This makes it hard to collect large amounts of data regarding current rent prices and their movement over time. It also makes it hard to predict which neighbourhoods are currently popular or which ones may become so in the near future. However, taxi data is a real time metric that is updated constantly. If we can use this data to predict real estate prices, it can be used well into the future as the data will always remain current.

The safety of the neighbourhood is also taken into account when prospective buyers go looking for houses, and this includes road safety as well. If we can quantify how safe the taxi drivers and normal drivers are in a neighbourhood - we can determine whether taking a cab or driving on your own in that specific neighbourhood is a better option if you are buying a house.

III. Related Work

In working on this project, we were interested in looking into other research work focused on using taxi data and especially ones that used Big Data technologies to do so. The research paper by Stoyanovich, Gilbride and Moffitt[1] which discusses an effort to analyse the taxi dataset used in the paper as well to look for patterns and interesting insights.

The paper makes use of the publicly available taxi data from NYC Taxi & Limousine Commission for the year 2016. This paper uses the same big data environment(Hadoop) and also analyses the same set of data points(taxi data) to extract insights and hence the techniques and tools used here are helpful in our own project.The initial part of the paper talks about the inspiration and goals of the paper. It broadly defines3 results, namely, hotspots, popular routes and data visualization of the results. The next part of the paper talks more deeply about the data used, the format, the cleaning process. Section 2then covers the softwares and the execution environment used. It goes into the details of why the Hadoop ecosystem and particularly Portal was chosen for this task. It also talks about howPortal uniquely fits into the requirements of this analysis. The remaining parts of section 2 go into the implementation and design detail of how the analysis was carried out in Portal.Section 3 talks about the results of the analysis. The paper uses a whole lot of graphs to explain the main findings of the paper. The data is analysed at different levels of granularity in terms of distance and time and at each level the main findings are discussed. Some of the surprising findings are that Laguardia is a more frequent destination/source than JFK even though JFK is a much bigger airport. It also talks about how many of the trips along the same route and time are often with just 1-2 passengers, a clear opportunity where ride sharing is beneficial.

The work done by Nadai and Lepri[2] is also interesting and gives insights into what are the different factors that influence housing prices for a neighbourhood. By analysing close to 70000 online listing in Italy, the authors discuss how different factors like building age, proximity to airport, property taxes and many others influence buyers decisions.

Researchers have also used other data sources to try and determine housing prices. In this paper by Zamani and Schwartz[3], the author look to predict housing prices using twitter data. The authors investigate how social media activity and language of a community can capture economic activity and hence be an indicator of real estate prices in the area.

The paper by Teck-Hua Ho, Juin Kuan Chong, and Xiaoyu Xia[4] tries to analyse different factors of taxis that lead to accidents in Singapore. The authors consider multiple factors like driver characteristics, past accident records, their contracts, daily driving records along with the colour of the taxis they drive. When comparing these factors with the accident data for the same time period, the conclusion was that the difference in drivers (age, demographics, education, experience) do not affect the accident trend at all, and the kind of training and recruitment for all the drivers were the same without any special treatment. With all of the other factors being the same, and the number of accidents with yellow cab taxis being higher than blue cab taxis by over 9%, they concluded that the colour should have an effect on the number of accidents.

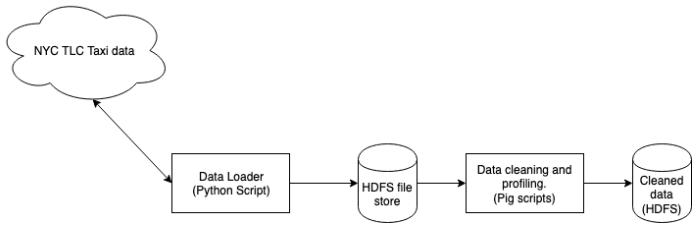
IV. Design and Implementation

The entire project was implemented using open source tools from the Hadoop and HDFS family of softwares.

The design is divided broadly into three phases. The ETL phase, which is responsible for loading the raw data from the datasource into the internal HDFS file system, the processing phase, which is responsible for add additional information to the dataset and finally the analytic phase, where the data is crunched to extract interesting insights.

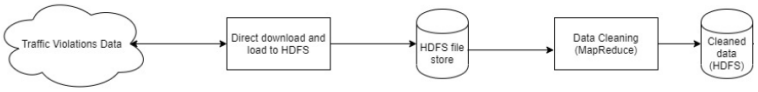
The NYC TLC taxi data is a freely available open dataset. The dataset is available for each month and each year from 2009 to 2019. For the purposes of this paper, we considered only the data from 2016 onwards to ensure the recency of data and hence remove any bias resulting from stateless on information. The data was loaded into the NYU Dumbo HDFS cluster using python scripts. The data was first downloaded in chunks and stored temporarily in local files before being loaded into HDFS.

Once the data was present in HDFS, it was then cleaned to remove any rows that contained obviously incorrect information such as negative fare amounts, or incorrect location IDs. Once this was done, the relevant features were extracted using pig scripts and MapReduce and re-loaded into HDFS as clean data.



Above is a diagrammatic representation of the ETL process.

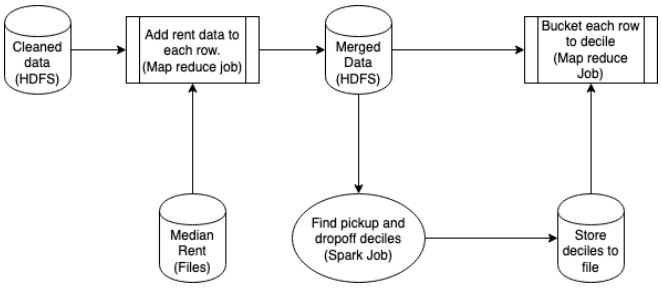
A similar approach was taken for the cleaning of data for the Traffic Violations Dataset (removing invalid rows with invalid dates, empty rows) and formatting the date to yyyy-mm and the results were stored into HDFS.



The processing phase for matching taxi trips with rent involved multiple map reduce jobs to append rent information to each row of the taxi dataset. To do this, we used an auxiliary dataset which contained the rent information for each neighbourhood in a time series format for each month for 2009. Using this information we first mapped the location name in the rent dataset with the corresponding location id in the taxi dataset. Since the names were not an exact match, we used a simple python script to calculate Levenshtein distances between each pair of locations from the two dataset to map

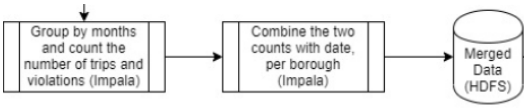
each location in the taxi dataset to its closest matching location on the rent dataset if a perfect match was not available. With this mapping ready, the next step was a map reduce job which took each row of the taxi dataset and appended the pickup and drop-off median rents to each row.

The second map reduce job in this phase added the rent decile information to each row of the taxi dataset. This information was required in doing aggregate analysis of how tipping is related to rent prices in the neighbourhood. The decile were calculated using a spark job and these decile values were fed into a second map reduce job which added the information to the taxi dataset.

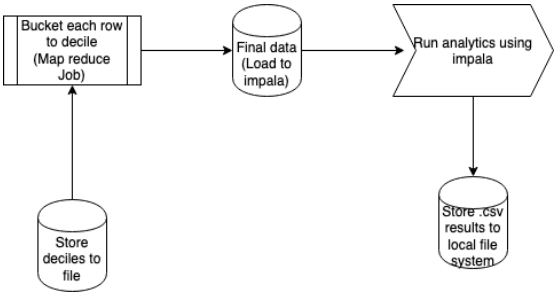


Above is a representation of the process.

The processing for taxi trips with violations involved exporting the cleaned data of taxi trips and violations from HDFS to Impala by creating separate tables for them. Once the tables were created, the values were grouped by borough and date (which was already cleaned to have a yyyy-mm format) and storing a new column of counts for each group (i.e, each month. Once this was done for both datasets, the datasets were combined with a join and the counts were merged. The final step involved separating the datasets by borough to make neighbourhood based analysis easier, and was sorted by date and exported to a csv file on the system.



The last phase was to run analytics on the processed dataset. For this, we wanted to use only big data tools which can scale as the size of the data increases. Therefore, the data was then loaded into Impala using an sql script. The final part of the process involved running .sql scripts on the impala database to get analytics regarding tip percent distribution across different localities and across rent deciles.



In the last phase of analysis for taxi data and violations, we exported the combined data from Impala to Spark to run a correlation analysis to check the results of the correlations for each borough.

V. DATASETS

A. Name of dataset #1

NYC TLC(Taxi and Limousine Commission) Open Taxi Dataset. This dataset is freely available on the Taxi website. The data is divided by month and year and is available for each year 2009 onwards. Each data block(for each month and year, say Aug 2013), is around 300 - 500MB. For this project data from 2016 and onwards was considered. The total size of the aggregated dataset was 7GB in size.

The dataset is in .csv format and contains a total of 17 fields. Each row of the dataset describes a taxi ride. The columns present include the pickup locality(in the form of an id), drop-off location, fare amount, tip, distance, etc. The data and detailed description of the dataset is available in the following location mentioned in the references page. [5]

B. Name of dataset #2

Streeteasy monthly rent data describes the median rent of each neighbourhood for the city of New York starting from 2009 onwards. The data is freely available on the street easy website. The data is in the .csv format and is a single file in the KB size range. Each row of the dataset represents the time series change of rent for a neighbourhood over the last ten years, recoded monthly. The data is available in the following location mentioned in the references part of this paper. [6]

C. Name of dataset #3

Open Parking and Camera Violations Dataset (from NYC Open Data). This dataset is freely available on the NYC Open Data site, and has multiple columns for each violations that was recorded - the ones of interest to us were mainly the date of the violation and the borough. The dataset was roughly 12GB in size. The data is available in the location mentioned in the references part of this paper.[7]

VI. RESULTS

There were two hypothesis with respect to rent that we wanted to test out in this paper. The first was : that more affluent riders generally tip better than less affluent ones. In this experiment we used the rent of a particular location as a proxy for affluence. It turns out that this is not in general true across all location in New York. Table 1 shows the average tip percent for each of the 10 deciles of the pickup rent.

Dropoff Decile	Avg Tip percent	Avg Dropoff Rent
9	13.99	7749
8	14.45	4353
7	14.78	3850
6	14.42	3649
5	14.79	3294
4	15.05	3000
3	15.32	2998
2	13.79	2705
1	13.11	2200
0	13.80	1839

Table 1

Avg Tip percent	Pickup decile	Avg Pickup rent
14.03	9	8446
14.19	8	4787
14.13	7	3978
14.75	6	3946
14.33	5	3751
15.14	4	3500
14.41	3	3436
13.80	2	2931
13.09	1	2200
15.36	0	1803

Table 2

From the tables we can see that there is no strong co relation between affluent neighbourhoods and a higher tip. In fact as the data shows, that there is actually no great different in the tip value. The difference is minuscule and we cannot make any generalisations across it. This is unexpected and surprising, but our theory regarding this is that while there may be islands or high/low tipping across neighbourhoods that are affluent/not affluent, when generalising across the entire city, the differences get averaged out.

Pickup Tip Percent	Locality ID	Avg Pickup Rent	Locality Percentile	Locality Name
17.3	138	1745	0	La Guardia(Queens)
15.9	125	No available data	NA	Hudson Sq(Manhattan)
15.7	234	No available data	NA	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)

Table 3

To test out this new hypothesis, we then checked if such neighbourhoods do exists, and whether in such neighbourhoods, the hypothesis that affluent riders pay more tips holds. So next, we tried to get the best ten and worst ten neighbourhoods for tips and with the average pickup and drop-off rents of those neighbourhoods. Here, it looks like the hypothesis does hold. The data in tables 3, 4, and 5 describe the results.

As you can see that among the top ten neighbourhoods for tips, we see that almost all of them are in Manhattan and also areas with a high median rent. The one outlier here is LaGuardia airport, which expectedly does not follow the trend

Best dropoff tip	Locality	Avg Drop off Rent	Rent Decile	Locality Name
18.2	138	1743	0	La guardia
17.4	40	No data	NA	Carroll Gardens(Brooklyn)
17.1	52	3069	2	Cobble Hill(Brooklyn)
16.6	257	2531	1	Windsor Terrace(Brooklyn)
16.3	249	3739	5	West Village(Manhattan)
16.2	181	2825	1	Park Slope(Brooklyn)
16.1	112	2794	1	Greenpoint(Brooklyn)
16.1	189	2900	1	Prospect Heights(Brooklyn)
16.1	224	1822	0	Stuvesant Heights(Brooklyn)
16	1	2831	1	EWR(Newark Airport)

Table 4

because most of the riders going to and from the neighbourhood do not stay there. And similarly, when analysing the data for neighbourhoods with the lowest average tips, we see that none are in Manhattan, and all of them have really median rents.

Also another interesting fact observed here is that the there is a significant difference in the median rent for best locations

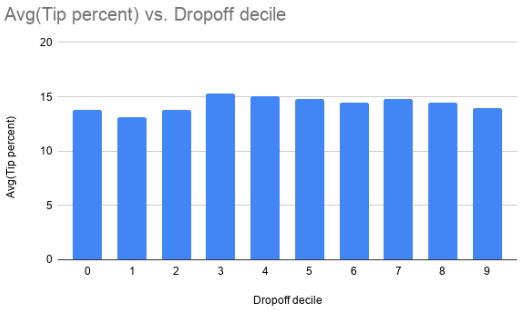
for pickup and drop-off. On further analysis, the reason for this becomes clear as the drop-off locations are mostly in the suburbs(Brooklyn, Queens, Bronx), where the average rent is in general lower for the same desirability.

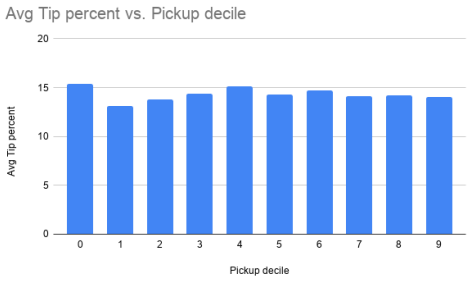
Pickup Tip Percent	Locality	Avg Pickup Rent	Pickup Rent Decile	Locality
3	35	1843	0	Brownsville(Brooklyn)
3.4	173	1939	0	North Corona(Queens)
3.5	39	2085	0	Canarsie(Brooklyn)
4	69	1699	0	East Concourse(Bronx)
4.1	26	1798	0	Borough Park(Brooklyn)
4.4	177	3162	4	Ocean Hill(Brooklyn)
4.4	76	2122	0	East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

Table 5

And similarly we can see that the locations where the tip percent is low are mostly areas where the rent is also low and almost all of them fall in the lowest decile of the rent scale even though they are not in the prime area(Manhattan) of New York.

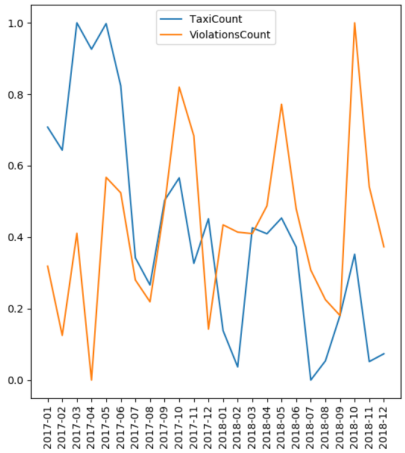
Below are some graphical depictions of tables 1 and 2 for a more visual representation of the data. From here we can see that the decile has very little effect on the tip percentage in general.



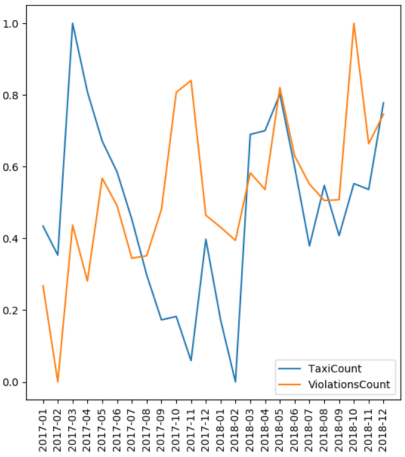


When looking at the number of taxi trips and the number of violations across the boroughs, these are the plots of the taxi trips and violations, plot across each borough.

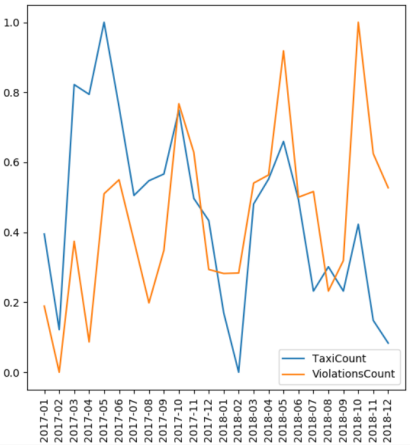
Manhattan (Correlation = -0.224009)



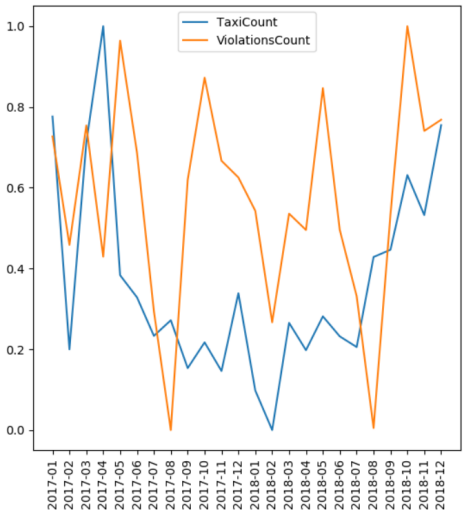
Brooklyn (Correlation = 0.000756)



Bronx (Correlation = 0.089859)



Queens (Correlation = 0.223480)



Staten Island (Correlation = 0.275796)

With these correlation values, we can see that Manhattan has a negative correlation of 22%, which shows that more regular drivers are involved in violations across Manhattan. This means that self driving in Manhattan is not preferable, taking a taxi is more likely to keep you safe.

In the case of Brooklyn and Bronx, the correlations values are close to 0%, which means that the chances of violations and accidents are almost equal, irrespective of self driving or taking a taxi.

Queens and Staten Island show a positive correlation (22% and 27% respectively), and this means that taxi drivers are involved in violations and accidents that happen across Queens and Staten Island, and hence it is preferable to drive yourself over taking taxis in these neighbourhoods.

VII. Future Work

This work was meant to be a hypothesis testing experiment. The hypothesis has been partly proven and partly disproven. Taxi spending pattern do correlate to median housing prices in a neighbourhood but the correlation is highly localised and cannot be generalised over large areas. Predicting housing prices is a complex problem, and is highly unlikely to be dependent on just a single variable like taxi prices.

We believe that the current work is a step forward in building a prediction model. The learnings from the experiments done in this paper can be taken forward and the taxi dataset can be used as one datapoint in a more complex model which also takes into account other traditional variables like schools, commercial establishments, office districts and many other such variables to be a better prediction engine.

Much like the work done by Zamani and Schwartz[3] where twitter data alone proves to be not very efficient but in conjunction with other traditional house price prediction models, increases the accuracy of the model. In the same way we hope that the learnings of this paper will lead to taxi data being an input in building more accurate housing models.

VIII. Conclusion

In this paper, we have studied the cause and effect relation between median rent prices and the taxi usage for the neighbourhood. Predicting housing prices is a complex problem to solve and will probably require far more datapoint to predict with greater accuracy. Our work is fulfils just one facet of the many variables which can be used for the purpose. Our work suggests that while there is a correlation between the taxi data and median house data for a given neighbourhood, it cannot be generalised over a large area and hence if such a hypothesis is used in building a model to predict housing prices, it will need to be aware of the localised nature of the co relation and adapt accordingly. We have also studied the relationships between the number of taxi trips and the number of violations across each borough in New York, and this can help decide whether taking a taxi or self-driving is a better option in these boroughs.

Acknowledgment

We would like to thank Prof. Suzanne McIntosh under whose guidance this project and course was completed. We would also like to extend our gratitude to the the graduate teaching assistants Omkar Patinge and Srishti Grover for their feedback. We would like to thank NYC Open Data and the relevant departments, and StreetEasy for making the data openly available and accessible for analysis. We would also like to that the High Performance Computing group at New York University for allowing us to run our experiments on their cluster.

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