Real Estate evaluation Based on Crime & Complaints

Pratik Sunil Rane
NYU Tandon SOE
Brooklyn
psr280@nyu.edu

Glancy Cajitan Rodrigues
NYU Tandon SOE Brooklyn,
NY
gcr253@nyu.edu

Shivraj Patil
NYU Tandon SOE
Brooklyn, NY
srp468@nyu.edu

ABSTRACT— The paper examines the study on how real estate prices are affected by crimes (Emergency) & complaints (Non-Emergency) in New York City neighborhoods. Crime & complaints in the neighborhood serve as an important parameter for change in the socio-economic composition of communities. While such change occurs over a long period of time, it is difficult for a naive user to find best neighborhood to live. We aim to do so, by using big data technologies.

Keywords- Big Data, Hadoop, HDFS, Map Reduce, Impala, Hive, Heat Map, Visualization in R.

I. INTRODUCTION

The goal of the paper is to evaluate the neighborhoods / precincts in New York City based on their crime & complaint records & property values. To achieve this, we are using the major crime dataset (emergency) & 311 dataset (non-emergency complaints) and real estate property data. We have used big data technologies to profile these datasets, clean noise in data and create an analytics which will list best precincts to live in New York City. We had identified the common and unique attributes in each of these datasets. After successful mapping of these common attributes, we used Hadoop - Map Reduce to generate useful information with respect of all 3 datasets. With the results of Map Reduce we used hadoop programming technology - Impala to build our final analytics.

II. MOTIVATION

Thousands of people move to New York City frequently, unaware of the property/rent values & crime rate in the neighborhood. With this project, we aim to solve this problem also the results that we get with the analytics will be helpful to find an area which will be both safe and within their budget of a user in search of apartment. In addition, realtors & property websites can use this information to analyze and quote the property rates.

Furthermore, analyzing and mapping various datasets would help us understand various challenges and big data technologies in current use.

III. RELATED WORK

A. Crime and Residential Choice: A Neighborhood Level

The data that is used is the housing, crime, and demographic data at the census tract level for the city of Columbus, Ohio, a Midwestern city. The results of the experiments they ran is said to misleading i.e., the analytic that the housing price decreases with increase in the crime rate was not proved from their results They use Hedonic modeling which is a revealed preference method of estimating demand or value of housing transactions over time to examine the impact of changes in crime in addition to levels of crime. They measure changes in behavior by examining whether changes in crime levels affect housing values when holding the crime rate constant.

Finally, they also examine whether the impact of crime varies across different types of neighborhoods and show that local business owners capitalize the cost of changes in violence differently based upon pre-existing neighborhood conditions. They offer several competing hypotheses regarding how and why crime will differentially impact housing prices depending upon levels of per capita income in the neighborhood, one of them being changes in violent crime rates will have a larger impact in higher-income neighborhoods in which violent crime is typically a less frequent event. Also they say that crime is under- reported in certain types of neighborhoods, particularly those comprised of poor or largely immigrant residents.

B. Crime Pattern Detection Using Data Mining

This paper discusses clustering techniques since the dataset is categorical data & output response is undecided. The contribution here was to formulate crime pattern detection as machine learning task and to thereby use data mining to support police detectives in solving crimes. Some of the significant attributes are identified using expert based semi-supervised learning method and developed the scheme for weighting the significant attributes. Our modeling technique was able to identify the crime patterns from a large number of crimes making the job for crime detectives easier. Also data mining is sensitive to quality of input data that may be inaccurate, have missing information, be data entry error prone etc.

C. Big Data: New Tricks for Econometrics

If you have several gigabytes of data or several million observations, standard relational databases become unwieldy. Databases to manage data of this size are generically known as "NoSQL" databases. Once a dataset has been extracted, it is often necessary to do some exploratory data analysis along with consistency and data-cleaning tasks. Data analysis in statistics and econometrics can be broken down into four categories: 1) prediction, 2) summarization, 3) estimation, and 4) hypothesis testing. For Prediction first step is penalize models machine learning experts have come up with various ways to penalize models for excessive complexity. In the machine learning world, this is known as or excessive complexity, this is known as "regularization". Second, it is conventional to divide the data into separate sets for the purpose second, it is conventional to divide the data into separate sets for the purpose of training, testing, and validation. You use the training data to estimate a model, f training, testing, and validation. You use the training data to estimate a model, the validation data to choose your

model, and the testing data to evaluate how well he validation data to choose your model, and the testing data to evaluate how well your chosen model performs." Third, if we have an explicit numeric measure of model complexity, we can third, if we have an explicit numeric measure of model complexity, we can view it as a parameter that can be "tuned" to produce the best out sample predictions. Then build a classifier is to use a decision tree.

Most economists are familiar with decision trees that describe a sequence of decisions that results in some outcome. After this return to the familiar world of linear regression and consider the et us return to the familiar world of linear regression and consider the problem of variable selection

D. Cassandra - A Decentralized Structured Storage System

Cassandra is for large amounts of structured data spread out across many commodity servers, while providing highly available service with no single point of failure. Cassandra does not support a full relational data model; instead, it provides clients with a simple data model that supports dynamic control over data layout and format. There are strict operational requirements in terms of performance, reliability and efficiency, and to support continuous growth the platform needs to be highly scalable. Cassandra uses a synthesis of well-known techniques to achieve scalability and availability. A table in Cassandra is a distributed multi-dimensional map indexed by a key. The value is an object which is highly structured. Cassandra exposes two kinds of columns families, Simple and Super column families. Super column families can be visualized as a column family within a column family.

The system allows columns to be sorted either by time or by name. The Cassandra API consists of insert, get and delete method. It partitions data across the cluster using consistent hashing. It uses replication to achieve high availability and durability. Cluster membership in Cassandra is based on Scuttlebutt, a very efficient anti-entropy Gossip based mechanism. The Cassandra bootstrap algorithm is initiated from any other node in the system by an operator using either a command line utility or the Cassandra web dashboard. This system relies on the local file system for data persistence. The data is represented on disk using a format that lends itself to efficient data retrieval. Thus Cassandra can support a very high update throughput while delivering low latency.

E. Can Poor Neighborhoods be Correlated with Crime? Evidence from Urban Ghana

Using the household survey data and a qualitative study conducted in different socio-economic neighborhoods in four cities Accra, Kumasi, Sekondi-Takoradi, and Tamale. It was concluded from the paper that the relative safety of low-class neighborhoods compared with middle-class neighborhoods is attributed to strong social cohesion and the presence of guardianship at all times of the day in poor neighborhoods. Various others theories like Hipp and Yates theory on analytics poor neighborhoods being correlated with crime are explained. The study is done based on a survey of 2,745 households, conducted in 13 selected low-, middle-, and high-class neighborhoods in Ghana's four key cities. It is also observed that although individually the selected study sites of low-, middle-, and high-class neighborhoods possess unique identities in terms of their location, age and history, ethnic composition, and socio-economic development, each group possesses distinct characteristics which distinguish it from the rest.

The paper is concluded saying the study has revealed that while low-class neighborhoods have high poverty characteristics, they tend to have high social cohesion and strong community bonding, both of which help in positively impacting crime—and hence the assessment of these neighborhoods as relatively safe compared with middle-class neighborhoods. However, the peculiar situation of low-class neighborhoods, especially poor housing and congestion as well as low income, tends to impact negatively on sexual crime incidence. In other words, the household and neighborhood characteristics of low-class neighborhoods facilitate sexual crimes.

F. Bigtable: A Distributed Storage System for Structured Data

Bigtable has successfully provided a flexible, high-performance solution for all of these Google products. In this paper we describe the simple data model provided by Bigtable, which gives clients dynamic control over data layout and format, and we describe the design and implementation of Bigtable. Large distributed systems are vulnerable to many types of failures, not just the standard network partitions and fail-stop failures assumed in many distributed protocols. We understood that it is important to delay adding new features until it is clear how the new features will be used.

A practical lesson that we learned from supporting Bigtable is the importance of proper system-level monitoring (i.e., monitoring both Bigtable itself, as well as the client processes using Bigtable).

IV. DESIGN

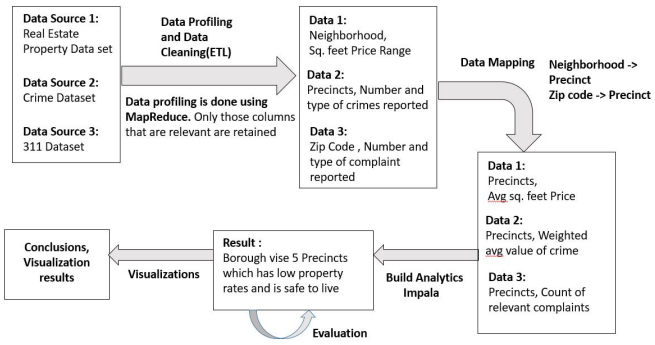

Fig 1. Design diagram

All the three data sets where available at NYC open data and where found separated with respect to year. Following years where considered for each data set:

1. Real Estate data set: 2008-09, 2009-10, 2010-11, 2011-12
2. Crime data set: 2009, 2010, 2011, 2012
3. 311 complaints data set: 2008, 2009, 2010, 2012

In the stage of data profiling only the columns that are necessary to the analytics are chosen. The data box after data profiling shows the only columns that were present after data profiling in case of each data set.

In the stage of data cleaning checks were incorporated with respect to all column data to filter out any errors that may be present in the data sets. For instance, the check for the valid zip code of the NYC was performed in case of 311 data to filter out all the unwanted data whose Zip code doesn't belong to New York city.

In the stage of mapping the data, deciding on a common field with which to perform the future analytics, we had choices on to select either precincts, zip code or neighborhood data. We decided to continue with precincts as there was easier mapping available to convert zip code and neighborhood data into corresponding precincts. The result of this stage contained the columns as shown in the design diagram after the data mapping process.

For building analytics part with the data that was available, we considered the average prices of the real estate in each of the precinct areas. In the case of crime, we calculated weighted average of crime (Weights are given based on severity of crime type) for each precinct. In case of 311 data only that departments (under which complaint is logged) is considered which may affect the living conditions of a specific area instead of considering those complaint type which may be specific only to building or apartment. For instance, the complaints like heat problems that a tenant is logged against his/her apartment owner is not considered for our analytics as it will be a problem which is specific to an apartment not the area in which the apartment is present.

All these steps till now where performed on each year's data that were available separately. The analytics was built using Impala. The results from the previous steps were used as the input to the Impala. We experimented with various combinations to generate a good analytics varying priorities to different datasets. We have given the least priority to 311 results since crime results & property results would relate more to a good or a bad neighborhood. Following are the two best analytics we have come up with:

Approach 1:

Priority 1: Crime; Priority 2: Property & Priority 3: 311 Results.

Here we are retrieving top 20 precincts for each borough from 311 results with least complaints, which are further filtered using property results to get top 10 precincts which have least property value. Finally, these 10 precincts are filtered using Crime results which have least crime rate(weighted average).

Approach 2:

Priority 1: Property; Priority 2: Crime & Priority 3: 311 Results.

The explanation is similar to the first approach only the priority of property and crime is interchanged.

V. RESULTS

Our analytics evaluates the top 5 precincts for each borough in New York City from 2009 to 2012. Following figure visualizes the output for year 2010 from our analytics (using approach 2 as mentioned in previous section):

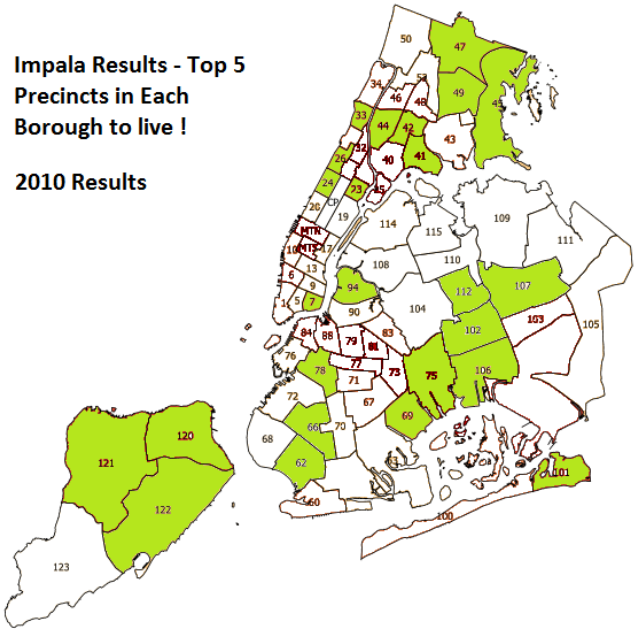

Fig 2. Analytics Results for year 2010

Most of our results are the same using approach 1 & approach 2. Following represents top 5 precincts for Brooklyn Borough using both approaches:

	Approach 1				Approach 2			
Year	2009	2010	2011	2012	2009	2010	2011	2012
Top 5 Precincts	94	94	78	69	77	69	69	69
	78	69	69	72	69	62	62	66
	88	62	88	66	78	94	66	62
	69	66	62	62	94	66	88	72
	77	60	66	60	88	78	78	88

Table: Analytics Results for Brooklyn Borough from 2009 to 2012 using approach 1 & 2.

Goodness of Analytics:

We used many of our analytics results to check its goodness with respect to actual dataset as follows:

Experiment 1:

Precinct 7 is evaluated as good precinct for 2009, 2010 & 2011 and is not included in top 5 precincts in Manhattan for year 2012. Below figure 3. shows how property values, crime & 311 is varied for precinct 7 from 2009 to 2012.

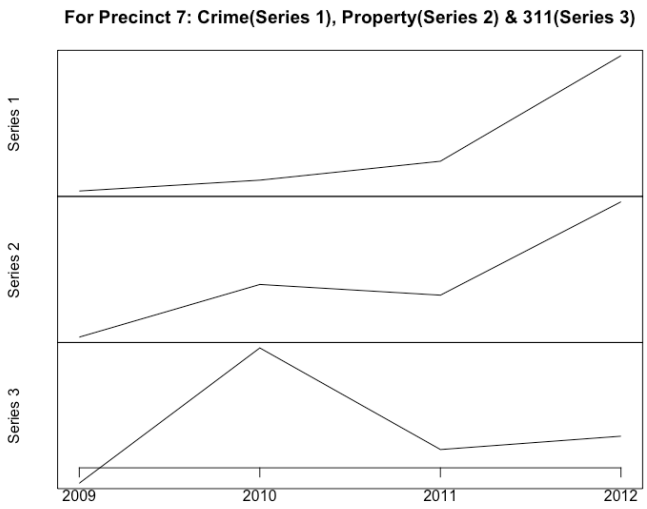

Fig 3. Crime, property & 311 plot from 2009 to 2012 for precinct 7

It is clearly seen that for year 2012, crime & property values have increased drastically. Hence our analytics is correct.

Experiment 2:

Precinct 26is evaluated from our analytics for all the years from 2009 to 2012. Following Figure 4. shows how property, crime & 311 varies for this precinct from 2009 to 2012, which proves that precinct 26 is good:

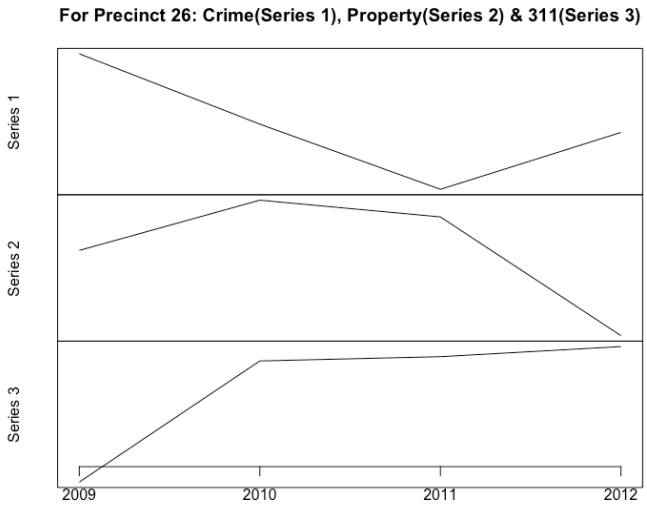

Fig 4. Crime, property & 311 plot from 2009 to 2012 for precinct 26

VI. FUTURE WORK

The results from our analytics can be used by users looking for a house to buy/rent, real estate companies. We can give priorities to different dataset and filter the results according to user requirements. In addition, the analytics can be further improvised by adding more external factors which affect the standard of living in neighborhoods.

VII. CONCLUSIONS

We have studied how real estate prices are affected by crimes (Emergency) & complaints (Non-Emergency) in New York City neighborhoods. We have successfully evaluated the best neighborhoods / precincts based on our datasets from 2009 to 2012. We have done this using big data technologies and proved the goodness of our analytics.

ACKNOWLEDGEMENT

We wish to thank Prof. McIntosh and for providing all the required technical guidance, valuable comments and suggestions and thank our TAs Shashank Pavan Segu and Zhengxin Cai for their constant support in guiding us throughout the project.

REFERENCES

[1] George E. Tita, Tricia L. Petras, Robert T. Greenbaum, “Crime and Residential Choice: A Neighborhood Level”.

[2] Shyam Varan Nath, “Crime Pattern Detection Using Data Mining”.

[3] Varian, Hal R, “Big Data: New Tricks for Econometrics”.

[4] Avinash Lakshman, Prashant Malik, “Cassandra - A Decentralized Structured Storage System”.

[5] George Owusu, Martin Oteng-Ababio, Adobea Y Owusu, Charlotte Wrigley-Asant, “Can Poor Neighborhoods be Correlated with Crime? Evidence from Urban Ghana”.

[6] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber, “Bigtable: A Distributed Storage System for Structured Data”.