A Study on the Relationship Between Housing Prices and the Air Quality, Metro Stations and Crimes in New York City

1st Yi Zhang
Department of Computer Science
New York University
New York, United States
ian.zhang@nyu.edu

2nd Taikun Guo
Department of Computer Science
New York University
New York, United States
tg1539@nyu.edu

***Abstract*—How the housing prices are affected by their surrounding environments? Though it is a tough work to figure out all the related factors. In this study, we mainly focus on the factors of the air quality, metro stations and crimes, and exploit how they influenced the housing prices in New York City (NYC). To do this, we profiled and compiled historical information on changes in housing prices, air quality, metro stations and crimes from 2010 till now in all the boroughs and blocks in NYC. With the refined data, we calculated the scores of each factor in each block with different measurements, then trained and tested some distinct multiple linear regression models which implied a large and statistically significant association between housing prices and these three factors. And the result will tell us the factor of metro stations and crimes play an important role on the housing prices. And there are still some important factors that we need to consider making the model better and get a more precise suggestion.**

***Index Terms*—Housing prices, air quality, metro, crime, analytics, Big Data**

I. Introduction

It is an interesting task to discover the law how the housing prices increased till now. However, it is also a complicated and tough work to figure out all the factors that affect the housing prices. For example, the location of your house, the nearby metro lines and your working place determine how long will you take to reach your destinations, the environment quality and noise environment affect the degree of comfort of your house, and the criminal situations, the locations of nearby police department and fire department influence you sense of security. The most important thing to be considered should be the price of your house or apartment.

In this study, we chose three main factors, air quality, metro stations and crime, to simulate their relationships with the housing prices, which implied there is a statistically significant relationship between housing and these factors. And we built a multiple linear regression model with considering the influence of the imbalance of development for each borough in NYC (As mentioned by Glaeser (2005) [4], the cost of land and construction played an important role in the housing prices), to approach a dwelling prices function which can estimate the housing price. With the predicted housing prices and the historical housing prices, we can make some analytics on the prices of houses in NYC and give some recommendations for consumers. Besides, we can also analyze those blocks with a high error with the predicted to find why they are so different.

The content of this paper is organized as follows. First, we will introduce the motivation of our study and describe why we think this application is important. Then, some related work will be introduced and compared with our work, which underscores the creativity of our study. In the fourth part, the paper will show how we design the application, including how the application was built and how the data flew. Then in the experiments part, we will describe our data source, experimental setup and discuss our experiments results. Then we will reach our conclusion in the next section.

II. Motivation

Housing prices has been one of the most important things that need to be considered by people in the United States, especially in those megacities like New York City, the Bay Area, etc. As Glaeser (2005) [1] mentioned, the housing price has been kept increasing in the past 60 years. And this is always a heavy burden for those new young graduates or poor. Besides the prices of houses, it is also important to consider the housing environments, such as air quality, metro stations, crime, etc. Thus, to help consumers giving a vision on how the housing prices related to the crimes, metro and air quality, and choosing a favored house with lower price and good environments is urgent by analyzing the blocks with abnormal scores. The aim of our study it to build a Big Data application to help consumers and investors making wise decisions on where and how to buy their desirable houses.

III. Related Work

We barely found a paper that considered multiple factors affecting the housing prices, and most of the papers only focused on one factor, such as air quality, crime, subway line or locations. For the factor of air quality, Vincenza (2014) [2] provided the relationship between housing prices and air quality by building some hedonistic multiple linear regression

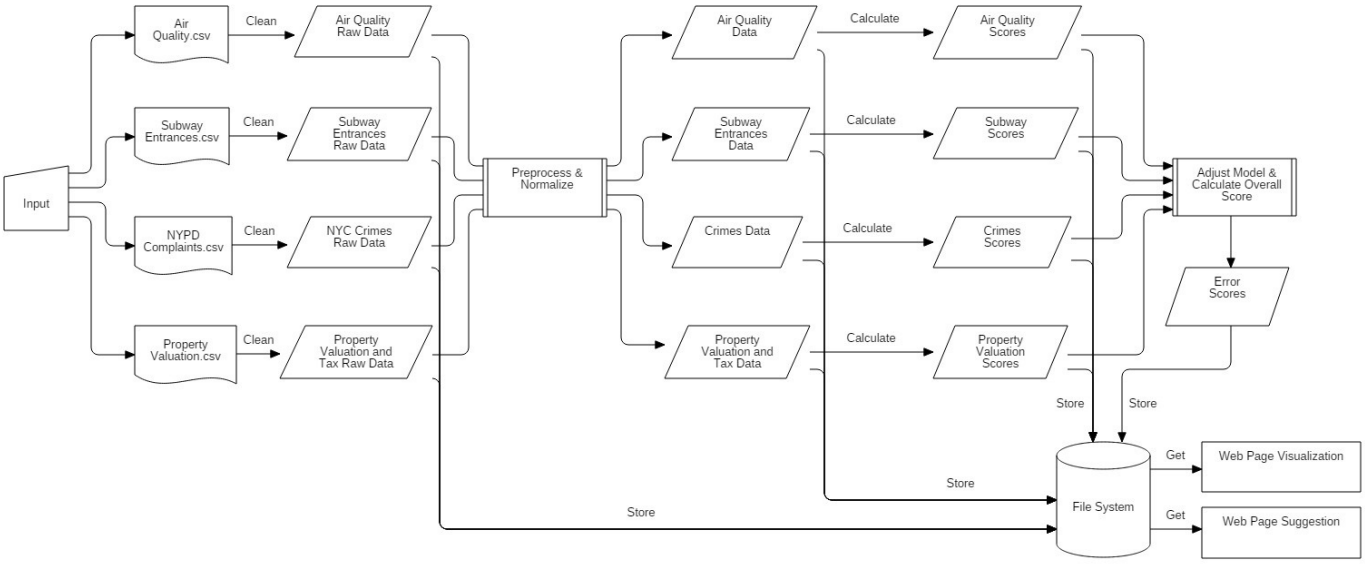


Fig. 1: The data flow graph in our program. It depicts how the raw data was cleaned, preprocessed and scored, and then the program trained the linear model and calculated the goodness of fit.

models to estimate the data in the province of Taranto, and his model also considered the house types, metro lines, roads and the sources and types of pollution. Though this is good experiments, his data source and size were not so perfect. His data was from the survey of more than 1000 observations, and such little data cannot ensure the precision and universality of the experiment data.

For the aspect of crime, Devin (2012) [3] gave a good experiment on the association between crime and property values by focusing on the 1990s crime drop with enormous and persuasive data, and reached a conclusion that the reduction in crime surely has an immediate benefit that should be reflected in housing values. However, the 1990s crime drop was irreproducible and peculiar, so that it cannot be considered as a normal simulation for our study. So, in our study, we should more focus on the relationship between the housing prices and crime under normal circumstances.

In addition, Agostini (2008) [5] presented his conclusion that building new metro lines will prompt the surrounding housing prices to rise in his paper. He also considered various factors including the distance to clinic, school and metro stations, which is a good and useful estimation for us to estimate the scores of distances of houses to nearest metro stations in our study.

IV. Design

In this section, we will introduce how the data was cleaned, preprocessed and scored to the data that can be trained in the linear model, and then we will discuss how the linear model was trained and how to score the errors. Finally, how to visualize these data was included.

A. Profile and Clean

The program codes were written in Scala and will be run on the Spark nodes. As shown in Fig. 1, when these raw data imported into the program, it will profile the raw data, for example, gather the type, range, and average value of each data, to help us to choose the data columns we need, and then we can decide how the raw data should be cleaned. After getting the cleaned data, then program will store them for backup. The detailed profiled information will be introduced in the experiment section.

B. Preprocess

The next step is to preprocess these cleaned data. The previous step is to choose the data we need, and now the program will standardize and transfer the data to the format which is needed for calculating the model and visualizing on the webpages. And in this step, using the BBL data from NYC government is very important, because the program will finally visualize the data for each borough and block in NYC. For instance, the raw crime data was composed by enormous distinct crime records from NYPD, where each record contains the crime date, position, severity, etc. To convert these data to what we want and to show how the crime situations of each block, the program will count the crime for a block as:

$$CrimeCount_i = \sum_{\forall r \in S} In(r, i) \tag{1}$$

$$In(r,i) = \begin{cases} Score(r) & \text{if } r \text{ was in 0.5 mile around block } i \\ 0 & \text{otherwise} \end{cases}$$

$$Score(r) = \begin{cases} 1 & \text{if } r \text{ is a Misdemeanor} \\ 2 & \text{if } r \text{ is a Violation} \\ 3 & \text{if } r \text{ is a Felony} \end{cases}$$

where S is the set of all crime records. And we also gave different scores for different severities of crime to depict the crime situation more detailed. For the air quality data, because the Air Quality Index (AQI) data was collected for each zip code area, then the program will simply assign each block with the data in its zip code area. And how to preprocess the metro entrances data is that, the program will calculate the distance of the block with its nearest metro entrance, by calculating their Euclidean distance and converting them to miles, as shown as (2):

$$MetroDistance_i = toMile(i.geom, entrance.geom) \quad (2)$$

where the *geom* represents their latitudes and longitudes. And the *toMile* function is defined as:

$$toMile(geom1, geom2) = \sqrt{d_{lt}^2 + d_{lg}^2} \quad (3)$$

$$d_{lt} = 68.6864 \cdot (geom1.lt - geom2.lt) \quad (4)$$

$$d_{lg} = 69.1710 \cdot \cos(geom.lt) \cdot (geom1.lg - geom2.lg) \quad (5)$$

where the *lt* is the abbreviation for latitude and *lg* for longitude. And for the property value, the program will calculate the average price per square feet for each block. Importantly, for those blocks which lacks data, the program will assign them the average value for AQI, crime counts and metro distance, and 0 for the property value.

C. Score

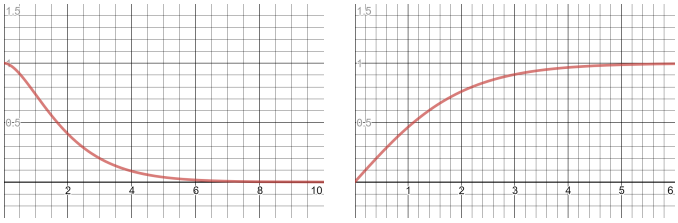
These refined data will be counted for each year separately and then stored in the file system so that the developers and users can easily view and download the data. After that, because these different types of data have different ranges and distributions, we need to normalize these data to (0,1) so that they can be trained in the same model. For the data of crime counts, metro distances and AQIs, as we know that the lower their values are, the higher their scores should be. Thus, the normalized function we used for these three types of data is:

$$Score(v) = e^{-\alpha v}(\alpha v + 1) \quad (6)$$

$$\alpha = \frac{\beta}{Average(v)} \quad (7)$$

where the α is a threshold value which differs from each type of data, and how their values were chosen will discussed in the experiment section in detail. In our experiment, we used (7) to determine the value of α, which is related to the average value and another threshold value β. It is important to choose the value of α, and we need to ensure the uniform distribution of each type of scores so that the image of the scores could be more precise and clearer. Its function graph is shown in Fig. 2a.

And for the property values, we used an updated sigmoid function to score its score, and unlike the data of crimes, metro and AQI, we do not need to change its growth property. The sigmoid function we chose is the (8), and its graph is shown in Fig. 2b.



(a) The normalize function for crime, metro and AQI data.

(b) The sigmoid function for scoring the property value data and the errors.

Fig. 2: The graphs of normalization functions.

$$Score_{property}(v) = \frac{2}{1 + e^{-\alpha v}} - 1 \quad (8)$$

As the same, these scores will be stored in the file system, so that we can easily use them to do analysis and show them on the web pages.

Finally, to count the data for each block and show them on the map correctly, the program also needs the Borough Block Lot (BBL) and zip codes data, and each record in the BBL data contains a polygon denoted by multiple geographic points, which represents the geographical area for that lot. When calculating the crime scores and metro scores for each block, the program first needed to calculate the contour points of this block by merging all the lots in this block. The algorithm is shown in Algorithm 1, and these contour points can depict the area for each block on the map. Besides, to calculate the crime counts and metro distance for a block, the program needs to find the center point, and the method is in Algorithm 2.

Algorithm 1 Algorithm for extracting the contour points of a block

Input: $points$: List[(Double, Double)]
Output: $result$: List[(Double, Double)]
Initialization:
$start$ = Find p in $points$ with least p._2
$result$ = List()
$current$ = $start$
$vector$ = (-1, 0)
Start Loop:
while $next$!= $start$ **do**
  $next$ = Find p in $points$ with least $\cos(vector, p - current)$
  $result$.add($next$)
  $vector$ = $current$ - $next$
  $current$ = $next$
end while
return $result$

Algorithm 2 Algorithm for finding the center point of a block

Input: $borders$: List[(Double, Double)]
Output: $center$: (Double, Double)
Initialization:
$totalArea$, $totalX$, $totalY$ = 0.0
Start Loop:
for i = 0 to $borders.length$ - 1 **do**
  a = $borders((i + 1) \bmod borders.length)$
  b = $borders(i)$
  $area$ = 0.5 $\times$ ($a.\_1 \times b.\_2$ - $b.\_1 \times a.\_2$)
  $totalArea$ += $area$
  $totalX$ += $\frac{1}{3}$ $area \times (a.\_1 + b.\_1)$
  $totalY$ += $\frac{1}{3}$ $area \times (a.\_2 + b.\_2)$
end for
return $(\frac{totalX}{totalArea}, \frac{totalY}{totalArea})$

D. Train Model

With the scores of each factor and of property values, we can start to train the linear model to simulate the relationship among them. Especially, to consider the geographical factor and the imbalance development among different boroughs, we added another feature that represented the borough of each block in the model. When in the experiment, we used the *MLLib* package to build the ordinary least squares linear model. When training this model, we marked the property values data as the labels and the crime, metro and air scores as the features. After that, we calculate the Chi-Squared goodness of fit for each label point *i*, which is:

$$Error_{Chi squared}(i) = \frac{(observed(i) - prediction(i))^2}{prediction(i)} \quad (9)$$

Then the errors were still needed to be normalized, so that it can be shown on the web pages averagely. And with the error scores, we can easily find which block is abnormal and figure out why, then we can give some suggestions to our users.

E. Visualization

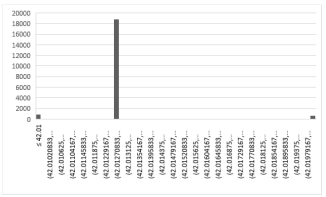
To visualize the data that we processed in the previous steps, we use the *Carto* web sites to help us show our data on the NYC map, and the results are stunning. Our program will automatically generate the *KML* file, which contains the geographical information for each block, crime scores, metro scores, AQI scores, property value scores and error scores. Then we only needed to upload it on the *Carto* datasets, and view them on the map.

V. Experiment

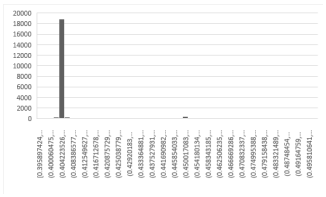
In this section, we will introduce our experiments data, the set of some threshold values, the experiment processes and the results in detail.

A. Experiment Data

In this study, we totally needed various types of data, including the air quality data, the crime records data, the metro entrances data, the property values data and the Borough Block Lot (BBL) data for NYC. So, next we will introduce where and how we got these data.

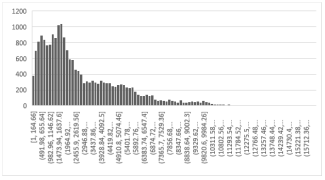


(a) The distribution of the AQI raw data.

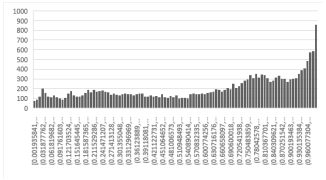


(b) The distribution of the AQI scores.

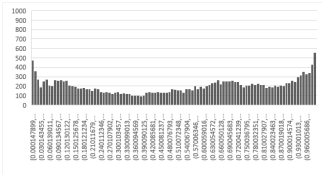
Fig. 3: The distributions of the AQI data.



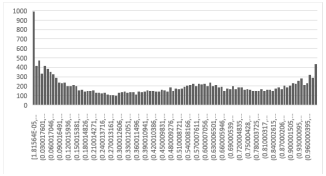
(a) The distribution of the crime counts data.



(b) The distribution of the crime score when β = 1.5.



(c) The distribution of the crime score when β = 2.0.



(d) The distribution of the crime score when β = 2.4.

Fig. 4: The distributions of the crime data.

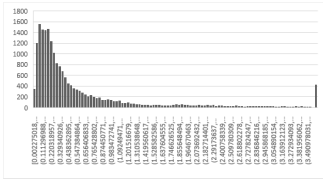
1) AQI Data: Our AQI data was fetched from the *AirNow API* (https://docs.airnowapi.org/). It is a query API that can enable users to query the historical air quality in a specified zip code area or a geographical point. However, at the very beginning we used a historical air quality data from *NYC OpenData*, and the reason why we did not use it was because its data was ambiguous and incomplete. There are many types of measurements for the AQI, and we finally chose the PM2.5 as the standard because of its widespread uses. To extract the AQI data, we wrote a fetching program to get the data for the first day of each month from 2010 to 2017 for each zip code area. And the profiled information for our AQI data is shown in Table I. After preprocessing the AQI data, the distribution of the data for each block is shown in Fig. 3a. The interesting things we can find is that the air quality in NYC is very average, and theres nearly no difference among each block.

To score the AQI, the β was set to 2, which means that α was 0.04769, and the distribution of AQI scores is shown in Fig. 3b. The scores of AQI preserve the property of raw data perfectly.

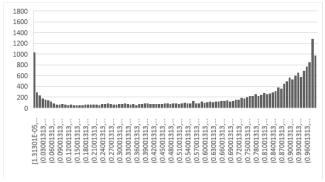
2) Crime Data: The crime data was fetched from NYPD Crime Statistics, which contains all the crime records in NYC during 2010 to 2017. And there were 351510 records in our raw data. After preprocessing, the profiled information is

TABLE I: The profiled information for the AQI data.

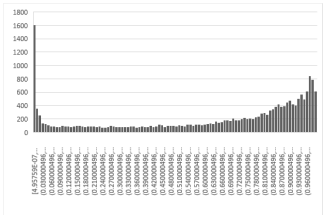
Count	22394
Maximum	156
Minimum	1
Average	41.934
Standard Deviation	0.545



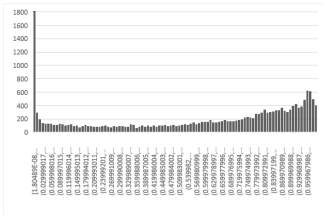
(a) The distribution of the metro distances data.



(b) The distribution of the metro score when β = 1.6.



(c) The distribution of the metro score when β = 2.0.



(d) The distribution of the metro score when β = 2.4.

Fig. 5: The distributions of the metro data.

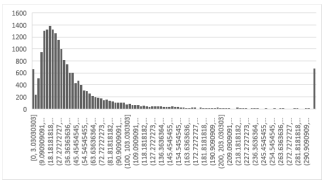
shown in Table II and its distribution is displayed in Fig. 4a. From the graph we can notice that, the distribution of crime counts is not uniform. Thus, to find the proper value of α, we have tried three different values of β, which were shown in Fig. 4b, 4c and 4d. Among these three values, when β was equal to 2.0, there is maldistribution either in the head or the tail, comparing with 1.5 and 2.4. And we also did some tests that, when the β went smaller, the distribution would have a fatter tail, and when it went larger, the head of distribution would expand. Thus, when β is set to 2.0, the model would have the best performance.

TABLE II: The profiled information for the crime data.

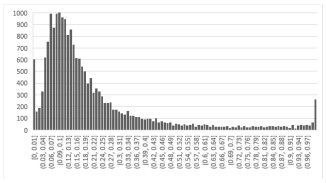
Count	20429
Maximum	16367
Minimum	1
Average	2888.862
Standard Deviation	2447.948

TABLE III: The profiled information for the metro data.

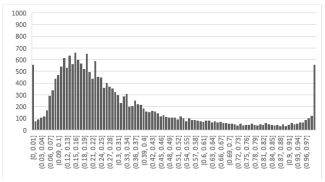
Count	20429
Maximum	5.811
Minimum	0.002
Average	0.667
Standard Deviation	0.880



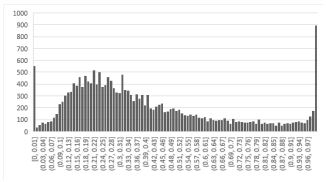
(a) The distribution of the property values data.



(b) The distribution of the property value score when β = 0.6.



(c) The distribution of the property value score when β = 1.0.

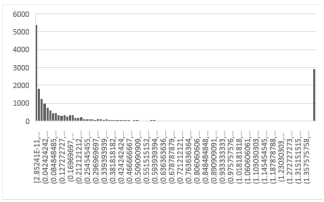


(d) The distribution of the property value score when β = 1.4.

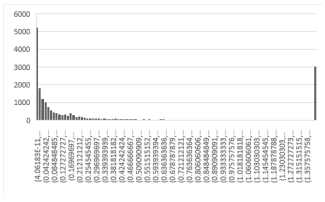
Fig. 6: The distributions of the property value data.

3) Metro Data: As for the metro data, because the unchanged of metro system in NYC has been kept for nearly 80 years, fortunately we only needed to find the metro entrances data, and we got it from the MTA datasets. After preprocessing the metro entrances data, as introduced in the design section, our program could use the data for each block to calculate the metro scores for each block, and the profiled information for the metro distance data is shown in Table III. However, it is a bit tougher to choose the threshold than to choose the α values for other types of data, and we have tried some values for β, which were shown in Fig. 5. Let us set β to 2.0 as the middle point, when it went smaller, 1.6 for example in the figure, the tail of the distribution would be fatter, and we did not regard this as a good choice because too many data points were gathered in the tail, which would make the scores unbalanced and uneven. And when β grew to 2.4, the head, which is the duration from 0 to 0.03, was extremely large, and the number of points in this range was over 2500. Though the tail of the distribution is more even than the distribution when β was 2.0, we did not choose it because it ignored too many information. When β was 2.0, its distribution was a little larger than of 2.4, and the number of scores in the range (0, 0.03) was also acceptable, which was about 7.8%, because for those blocks far away from the metro stations, it makes no difference for the distance from a block and its nearest metro station is 2 miles, 3 miles, or more. So, 2.0 could be the best values for β after our experiments.

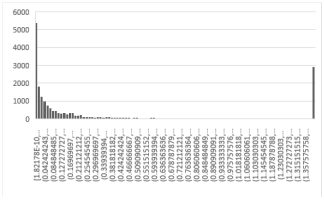
4) Property Valuation Data: The property valuation data was fetched from *Department of Finance* (DOF), which contains the property valuation, assessment data and the related BBL information from 2016 to 2017. And we chose some of the fields we needed, such as property values, area, BBL, etc., then the program calculated the average housing price for each block, and its profiled information was shown in Table IV. From its distribution graph Fig. 6a, we could find that this is a Gaussian-like distribution, and its tail and head are really



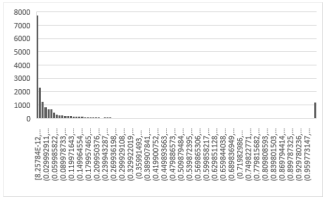
(a) The distribution of the errors of ordinary linear regression model.



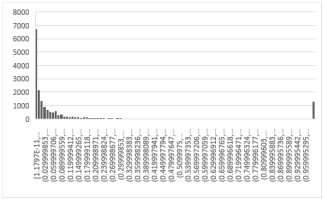
(b) The distribution of the errors of lasso regression model.



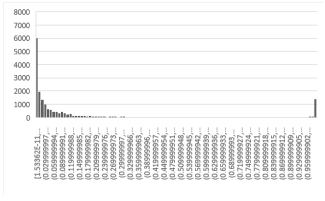
(c) The distribution of the errors of ridge regression model.



(d) The distribution of the error scores when β = 0.7.



(e) The distribution of the error scores when β = 1.0.



(f) The distribution of the error scores when β = 1.3.

Fig. 7: The distributions of the error of models.

fat. We believe that those data points in the tail should be in the most prosperous districts on the Manhattan island. In our experiment, when β was set to 0.6, the distribution of scores became very unbalanced, and when it was set to 1.4, the tail of distribution became much larger than the distribution of the raw data, which distorted the property of the raw data. We finally set β to 1.0 in our experiment, because it kept a balance between the uniform distribution and the reservation for properties of the data points in the tail.

TABLE IV: The profiled information for the property values data.

Count	20429
Maximum	3591.301
Minimum	0.054
Average	64.467
Standard Deviation	141.664

5) BBL Data: The BBL and zip codes data was fetched from the DOF and the *CartoDB* data library, and the program merged and preprocessed them together to get the geographical data we need in the next steps.

B. Experiment Results

After gathering all the required scores, the program started to train the linear model. In our experiments, we used three

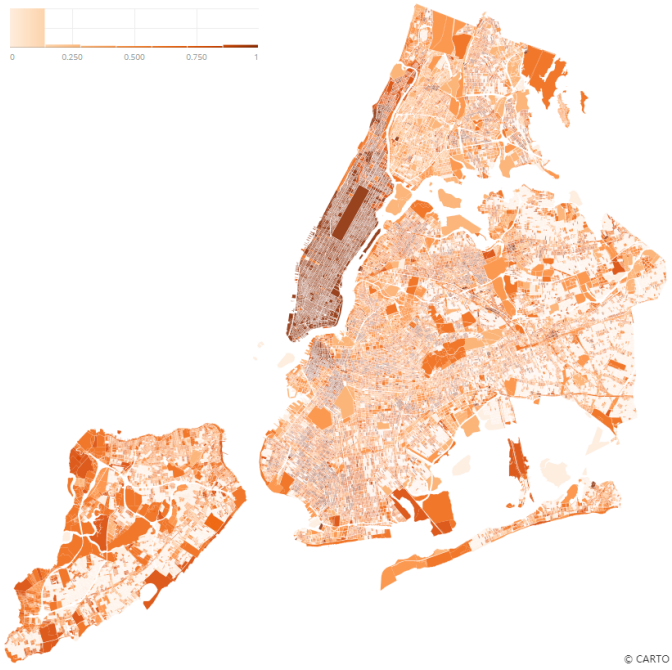


Fig. 8: The map of the error scores distribution on *Carto*.

types of linear regression models, ordinary, lasso and ridge, to try to simulate the relationship between the factors and the property values. The iteration number was set to 10 times the number of total data points, which was 204290, and the pace and regularization parameter were set to 0.01. Then the program calculated the Chi-Squared errors for these three models to test their goodness of fit, and the distributions were shown in Fig. 7a, 7b and 7c. From the results, we can find that there was nearly no difference among them. And their tails, which was larger than 1.4, were well-marked, we believed this was because some of the blocks cannot be well fitted by the linear models with the factors we gave, and we will analyze why it happened in the next section. However, for most of the blocks, the model could fit the observed data well, and the average Chi-Squared error was 1.209.

Additionally, showing the error to users is very important, because this can give users a deep vision on those abnormal blocks which may be the chosen place to invest or buy houses. Thus, we also need to normalize the error to (0, 1) with a sigmoid function, and the results were shown in Fig. 7d, 7e and 7f. All the three distributions preserved the marked tails in the raw error distributions, and that was the blocks that we wanted to recommend to our users.

After getting all the data and scores we needed, the program generated a KML file that could be read by the *Carto* and then shown on the web pages. And the error distribution on the map was shown in Fig. 8. From the map, we could find that most of the country side in NYC fitted the linear model well. However, for those busy districts, such as the Manhattan Island and the downtown in Brooklyn, the performance of our linear regression model was not very good, and we believe that there

must be some other factors we need to consider, for example, the prosperity, surrounding facilities, cultural environment and residential class of the blocks.

VI. Conclusion

Though it is hard to figure out how the housing prices is influenced, and there are so many factors we need to consider, in this study, we chose some of the most important factors that can also be quantized, the crime situation, metro stations and air quality, to simulate whether and how these factors affect the housing prices in NYC. From the experiment results, we reached a conclusion that, in most of the blocks in NYC, especially for those in the outskirts, the relationship between the factors and the property values can be simulated by the linear regression models perfectly. But for those downtown districts, especially for the Manhattan Island, the model did not perform well. This was because there were many unquantized factors affecting the housing prices on Manhattan, for example, the development level, the economic status, the cultural factors and the advantaged position of Manhattan. Even if our models considered the influence of different districts, its prediction still cannot make up the lack of these crucial factors. Maybe in the further study, we can first quantize these affecting factors and then train a new model considering all the possible factors.

VII. Acknowledge

This study was supported by our professor Suzanne McIntosh (mcintosh@cs.nyu.edu). We also thank the *High Performance Computing* department in NYU for supporting the computing resources, and most parts of our program was run on *Dumbo*.

We thank *CartoDB* for providing a visualization web page to show our processed data and give a direct vision for our users to help them and give them advises.

We also thank *AirNow API* for providing an API to enable us to fetch the historical AQI data in NYC, though the air quality in NYC is really good and there is hardly difference among all the historical records.

References

[1] Glaeser, Edward L., Joseph Gyourko, and Raven Saks. Why have housing prices gone up?. No. w11129. National Bureau of Economic Research, 2005.

[2] Chiarazzo, Vincenza, et al. The effects of environmental quality on residential choice location. Procedia-Social and Behavioral Sciences 162 (2014): 178-187.

[3] Pope, Devin G., and Jaren C. Pope. Crime and property values: Evidence from the 1990s crime drop. Regional Science and Urban Economics 42.1 (2012): 177-188.

[4] Glaeser, Edward L., Joseph Gyourko, and Raven Saks. Why is Manhattan so expensive? Regulation and the rise in housing prices. The Journal of Law and Economics 48.2 (2005): 331-369.

[5] Agostini, Claudio A., and Gastn A. Palmucci. The anticipated capitalisation effect of a new metro line on housing prices. Fiscal studies 29.2 (2008): 233-256.