Valuing NYC Neighborhoods

Daffney Deepa Viswanath
New York University
ddv246@nyu.edu

Aneri Dalal
New York University
and301@nyu.edu

Rong Feng
New York University
rf1316@nyu.edu

***Abstract*—**

This paper describes an application that gives users access to hard to acquire information on the services and characteristics of neighborhoods in New York City in order to aid them in the process of purchasing a home. The application uses a metric that combines school ratings, 311 service requests, and the median price of single-family homes in a given neighborhood, and produces a neighborhood value for every neighborhood in each of the five boroughs of New York City. The outcome of the application is an interactive map of New York City that gives us insights on the various aspects of a home buyer/seller. It also helps in determining if houses in a neighborhood are overpriced or underpriced based on the important features of the neighborhood. It is our hope that, upon using this application, a potential home buyer or real estate agent would be able to get an extensive picture of the neighborhood under consideration before making a final decision.

Keywords—analytics, spark, home buyers, real estate, 311, school ratings, house price, big data, folium map

I. Introduction

Finding the perfect neighborhood in which to buy a house can be exhausting especially if a family is moving to a new area and cannot distinguish one enclave from another. Rising home prices are often seen as one of the biggest markers of a good neighborhood. However, there are also other factors involved, such as school ratings and good city services, that have a greater impact on the lifestyle of a neighborhood. In service to homebuyers as well as the real estate agents of the New York City, we have come up with an application that makes use of publicly available datasets such as GreatSchools school ratings, 311 service request, and Annualized Property Sales in order to provide them with highly relevant, but hard to access, analysis on the neighborhoods. Each neighborhood is given a value/score that is calculated using a metric that tells the user how much value per sale price is available. The analysis presents users with a map of New York City, which enables users to view the value/score of each neighborhood based on each of the factors mentioned above. It not only gives an insight on the neighborhood but also has a direct impact on the value for the price of a house.

II. Motivation

When a family is in the process of purchasing a home, they are not just buying a house but also a part of that neighborhood. And most of the time, home buyers only get cursory insight of the neighborhood unless they have previously lived in the neighborhood for a couple of years. Our motivation through this analysis is to provide detailed information on commonly considered factors of a neighborhood, that one may experience by practically living in the neighborhood.

III. Related Work

Problem/Concerns regarding school ratings, crime analysis and influence of a neighborhood has been researched a lot in the past and several analyses have been proposed on the same.

[1] (A. Chiodo et al, 2010) compared the prices of homes near school zone boundaries to better control for the neighborhood effects on home pricing. If two houses are directly across the street and in different school zones, it would be a much more controlled comparison

Three arguments for non-linear pricing:

1. Concentrated buyer preference in the very-highly rated schools and limited supply of housing in these zones
2. Alternative schooling (privates) can be used in lower rated zones, further reducing the price of homes since cost of privates are high.
3. Good schooling is a luxury good in economic terms (to the chagrin of advents of American equality).

Education quality is proxied by standardized Math scores The Author finds that on the downside, for below-average schools, homes are priced on physical characteristics alone. On the other hand, there is a convex relationship between price and school quality on the upside for above-average schools.

[2] (L. Wang et al, 2017) shows one way to use 311 service request data. Their goal is to use 311 data from three cities, as well as 2014 US Census data and Zillow data to determine if 311 data can be used to classify the socioeconomic characteristics of a neighborhood and, going even further, possibly predict changes in housing prices. First, they use the 311 data to create "neighborhood signatures". The signature is a vector of relative frequencies of requests of different types. They then use k-means clustering to group the neighborhoods by signatures so they can see if the similarities in 311 signatures would mean similarities in the census data. They were able to find that each cluster had certain features in common such as ethnicity and income levels. They then used the 311 signatures and the Zillow data to see if they could predict changes in the annual average sale price of housing per square foot over time relative to the NYC mean. Their goal was not necessarily to predict a year to year change but rather to create a 311 based model that can predict future fluctuations. They tested on data from 2014 and their results were able to predict the trend of housing prices in direction and magnitude.

[3] (P. Visser et al, 2006) talks about the influence of the location and characteristics of a neighborhood on house prices and how this information can be used to improve the neighborhood for its betterment. The sales price of a house depends on various environmental attributes which are grouped into 4,

1. Physical housing characteristics

2.Physical characteristics of the residential environment

3.Social-economic characteristics of the residential environment

4.Functional characteristics of the residential environment

The paper not just analyzes the different characteristics but also the intensity of the influence of each of these characteristics. This analysis along with how it affects the price of houses in urban and rural areas, expands the scope of the analysis to other cities in the world as well. Finally the paper addresses one of the complex subjects to measure, housing market pressure and the influence of these characteristics on it.

IV. Datasets

3 public datasets were used in the analytics.

A. GreatSchools School Ratings

The dataset contains school ratings of all public schools in NYC. The dataset was < 1 MB and provided valuable insights on various public schools in each neighborhood. The data was scraped from the website using a bot we wrote.

Reference link where the dataset can be found: https://www.greatschools.org

Schema of GreatSchools School Ratings

NAME: String	Name of the NYC school
TYPE: String	Type of school (public, private, district, charter, etc)
GRADES: String	Grades offered at the school (ex: K-9)
ADDRESS: String	Address of school
ZIP: String	Zip code of school
RATING: Integer	Rating between 1 and 10
TOTAL_NUMBER_OF_STUDENTS_ENROLLED: Integer	Number of students enrolled at the school
STUDENTS_PER_TEACHER: Integer	Number of students to 1 teacher
REVIEWS: Integer	Number of reviews by parents
USER_REVIEW_STARS: Integer	Rating by parents between 1 and 5 stars
DISTRICT: String	School district
LINKS: String	Link to the GreatSchools.org page for the school

B. 311 Service Requests from 2010 to Present

The 311 dataset contains information about all the 311 service requests made in NYC since 2010. We chose to use only the service requests created in 2019 since they would be most relevant to the current state of the neighborhood. The dataset of just the 2019 service requests was approximately 1.25 GB. It contains various details about the complaint or requested service including date, locations, city government agency involved, and whether or not the service request was resolved.

Reference link where the dataset can be found: https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/7ahn-ypff

Schema of 311 Service Requests from 2010 to Present

DATE_CREATED: String	Date service request was created
COMPLAINT_TYPE: String	Category of complaint
DESCRIPTOR: String	(optional) Details of complaint
INCIDENT_ZIP: Integer	Zip code where incident took place
BOROUGH: String	Borough where incident took place

The original dataset contains a total of 41 columns but these five were the ones used in our application.

C. NYC Citywide Annualized Calendar Sales

This dataset contains sales information on properties sold in New York city between 2016 to 2020. The dataset was around ~70 MB. The dataset has information not only on houses sold in NYC but also on commercial buildings, religious buildings and warehouses. Our analysis will be concerned only with the price of one family dwellings.

Reference link where the dataset can be found: https://data.cityofnewyork.us/City-Government/NYC-Citywide-Annualized-Calendar-Sales-Update/w2pb-icbu

Schema of NYC Citywide Annualized Calendar Sales

BOROUGH: String	The name of the borough in which the property is located
NEIGHBORHOOD: String	The neighborhood name in the course of valuing properties.
BUILDING_CLASS_CATEGORY: String	This identifies properties broad usage
TAX_CLASS_AS_OF_FINAL_ROLL: Integer	Present tax class
BLOCK: String	A Tax Block is a subdivision of the borough on which real properties are located
LOT: String	A tax Lot is a subdivision of a tax Block and represents the property unique location
BUILDING_CLASS_AS_OF_FINAL_ROLL : String	The building classification is used to describe a property's constructive use
ADDRESS: String	The street address of the property
APARTMENT_NUMBER : String	The apartment number of the property,if any
ZIP_CODE : Integer	The property's postal code
RESIDENTIAL_UNITS : Integer	The number of residential units at the listed property
COMMERCIAL_UNITS : Integer	The number of commercial units at the listed property
TOTAL_UNITS : Integer	The total number of units at the listed property
LAND_SQUARE_FEET : Integer	The land are of the property listed in square feet
GROSS_SQUARE_FEET : Integer	The total area of all the floors of a building as measured from the exterior surfaces of the outside walls of the building , including the land area and space within any building structure on the property
YEAR_BUILT: Integer	Year the structure on the property was built.
TAX_CLASS_AT_TIME_OF_SALE: Integer	Tax Class at time of sale
BUILDING_CLASS_AT_T	The building classification at the time of sale

TIME_OF_SALE : String	Time of the property Sale
SALE_PRICE : Integer	Price paid for the property. A $0 sale price indicates that there was a transfer of ownership without a cash consideration.
SALE_DATE : java.sql.Date	Date the property was sold
COMMUNITY_BOARD : Integer	The Community Board field indicates the New York City Community District where the building is located.
CENSUS_TRACT : Integer	The Census Tract field indicates the U.S. Census Tract where the building is located.

V. Description of Analytic

Four different types of analytics were done using the datasets - one per dataset and one on the combined dataset.

A. GreatSchools - Analytics

Before averaging the ratings based on zip code, we applied a nonlinear function to the ratings. While the ratings provided by GreatSchools are already in the range of 1 to 10, we believed that the relationship between school ratings and attractiveness was nonlinear. This is also seen in [1]. One argument they provide is that “as school quality increases, competition from other buyers creates an increasingly tight housing market” [1]. This means that schools with lower ratings would not impact the housing prices and attractiveness of the neighborhood as much as schools with higher ratings. We applied an exponential function to the ratings in order for them to linearly relate to housing prices and attractiveness.

$$GSScore = \frac{e^{rating}}{2000} \div 2$$

Once each school had an adjusted score, we took the average for each zip code. The resulting dataset consisted of the zip code and average adjusted score.

B. 311 Service Requests - Analytics

Due to possible typos, the original dataset contained rows with zip codes that were not in New York City and some were not even in the United States. All rows with those types of incorrect zip codes were removed from the dataset during the cleaning process. Once that was completed, we totalled the number of service requests/complaints by zip code. We then applied a nonlinear function to the counts in order to create a score.

We chose a nonlinear function rather than a linear function because we believed that the relationship between service request/complaints counts and the attractiveness of a neighborhood is not a linear relationship. We believe that when there are very few service requests/complaints (<5,000), the neighborhood can still be considered very attractive whether there are 2,000 or 5,000 service requests/complaints. A similar argument can be made for when there are a lot of service requests/complaints (>20,000). Once the numbers are that high, whether it is 20,000 or 40,000, the neighborhood would be seen as very unattractive. The biggest difference in attractiveness would be seen in the middle values (5,000 < count < 20,000). In other words, there would be a large difference in attractiveness with a count of 8,000 versus a count of 18,000. Due to this reasoning, we chose to use a sigmoid function or S-curve to create the attractiveness scores.

$$311Score = \frac{10.8}{1 + e^{2.5 + (-0.0002 * count)}} + -0.8$$

The resulting data frame contained the zip codes and their respective attractiveness scores.

C. NYC Citywide Annualized Calendar Sales- Analytics

The original dataset has information on several building categories including 1,2 and 3 family house buildings, commercial buildings, religious buildings, rental buildings as well as warehouses. But, our analytic focuses only on one family dwellings in the city of New York. We filtered out data for one family dwellings category only. We filtered data with building categories relevant to One family’s requirement, as below.

1. One Family Dwellings
2. Coops - Walkup Apartments
3. Coops - Elevator Apartments
4. Condos - Walkup Apartments
5. Condos - Elevator Apartments
6. Condos - 2-10 Unit Residential

However, there were different building classes for each building category. We researched on how the buildings are classified in New York City and came up with a list of building classes that would fit into our criteria. Given a single family’s requirement,, we have taken into consideration the following building classes only.

A0	Cape Cod
A1	Two Story Detached Buildings
A2	One Story
A3	Large Suburb Residential Building
A4	City Residential One Family
A5	One Family attached or semi detached.
A6	Summer Cottages/ Mobile Homes
A7	Mansion Types
A8	Bungalows
A9	Miscellaneous One Family
S0	Primary One family with two stores or office

S1- Primary One family with one store or office

A new Dataframe was created with data pertaining to one family building category and building class. A median price was calculated for each building category, all building classes combined.

Median Price Calculation for each Building Category:

The Median Sale price was calculated using Approximate Percentile method.

1. The data frame was grouped by each building category for each zip code and the price was sorted in ascending order.
2. Total number of entries in each category per zip code was calculated and the median number was identified
3. The entry with the median number was picked as the median sale price for that building category of the zipcode.

The resulting data frame contained the zip code, the building category and the median sale price for the building category.

D. Combined Dataset - Analytics

The three datasets are then joined together by zip code. The GreatSchools dataset has the fewest distinct zip codes so some rows were dropped from the 311 and NYC Property Sales datasets in the process of joining. The resulting dataset contained zip codes, median housing prices, the adjusted school ratings, and the attractiveness score calculated from the 311 dataset.

Using Python, we generated summary graphs to visualize the data in terms of boroughs rather than zip codes just to get an overall look of the datasets.

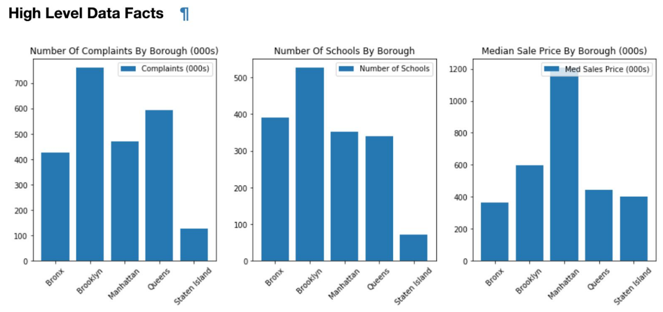


Fig.1 (Left to right) Number of complaints by Borough (000s), Number of Schools by Borough, Median Sale Price by Borough (000s)

Next, we calculated the relative attractiveness score for each neighborhood using the values/scores gained from the individual processing of the three datasets. The scores for each zip code is first mapped to a range of 1 to 5.

$$GSMappedScore = \frac{GSScore}{max(GSScore)} * 4 + 1$$

$$311MappedScore = \frac{max(311Score) - 311Score}{max(311Score)} * 4 + 1$$

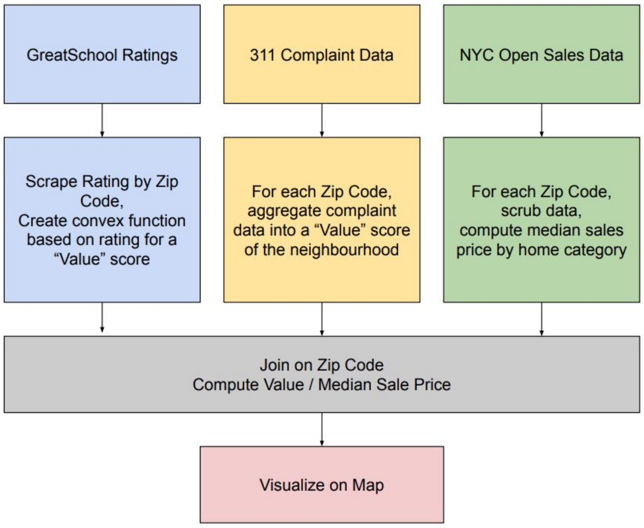
$$MedianSalesMappedScore = \frac{MedianSales}{max(MedianSales)} * 4 + 1$$

$UserWeight$ = user specified weight of GSMappedScore between 0.0 and 1.0

Then the final score for each zip code is calculated using the mapped scores and the weight provided by the user.

$$FinalScore = \frac{UserWeight*GSMappedScore+(1-UserWeight)*311MappedScore}{MedianSalesMappedScore}$$

VI. Application Design


GreatSchool Ratings

311 Complaint Data

NYC Open Sales Data

Scrape Rating by Zip Code,
Create convex function based on rating for a "Value" score

For each Zip Code, aggregate complaint data into a "Value" score of the neighbourhood

For each Zip Code, scrub data, compute median sales price by home category

Join on Zip Code
Compute Value / Median Sale Price

Visualize on Map

Fig.2 Project Design Diagram

The three datasets are used as big data for our application. The datasets were scraped or downloaded from the public domain and transferred to HPC Dumbo Hadoop cluster at NYU. Cleaning, profiling, and processing the datasets was done using Spark and Spark SQL. We joined the cleaned datasets based on zip code and transferred them to local drive. A value for each neighborhood was computed using the Median Sale Price of a one family dwelling, a score computed using 311 service requests and a school rating score. Finally, data visualizations are done using Python. We also used Folium, a third party library, that was used to render usable maps of New York City.

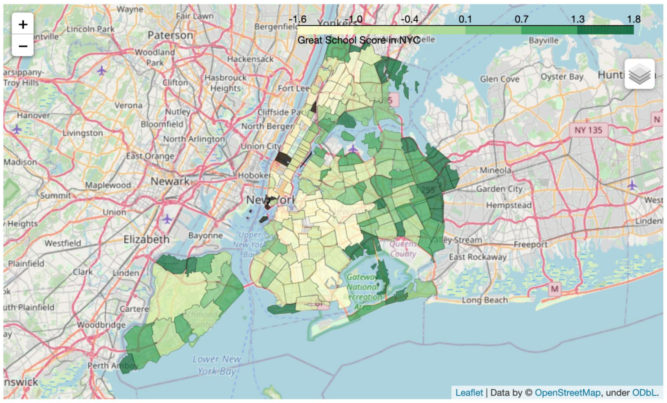


Fig.3 Map of NYC - UserWeight = 0.0 (FinalScore of zip codes based entirely on 311MappedScores)

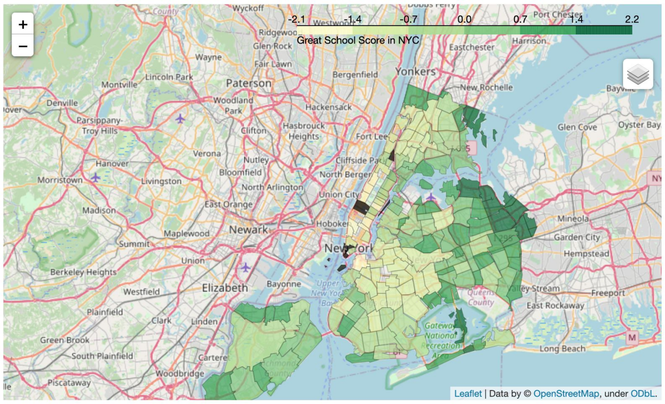

-2.1
-1.4
-0.7
0.0
0.7
1.4
2.2
Great School Score in NYC
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig.4 Map of NYC - UserWeight = 0.5 (FinalScore of zip codes based equally on GSMappedScores and 311MappedScores)

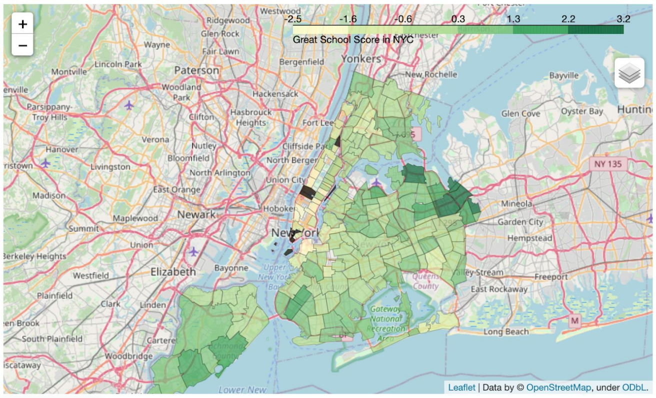

-2.5
-1.6
-0.6
0.3
1.3
2.2
3.2
Great School Score in NYC
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig.5 Map of NYC - UserWeight = 1.0 (FinalScore of zip codes based entirely on GSMappedScores)

VII. Analysis

From the analytic, with UserWeight = 0.5, we have concluded with the Best and Worst Neighborhoods of New York City.

Top 5 Neighborhoods:

1. Flushing
2. Glen Oaks
3. Little Neck
4. Bellerose Manor
5. Rockaway Beach

Bottom 5 Neighborhoods:

1. Lincoln Square
2. Gramercy
3. Chelsea
4. Downtown Brooklyn
5. Upper West Side

It is to be noted that all the 5 Top Neighborhoods were from Queens Borough. However, we cannot term Queens borough as one with best neighborhoods, most of them are in Manhattan and some parts of Brooklyn.

Manhattan was viewed as the worst value in this analytic. However we know that this doesn't take into consideration the life-style and the easy commute into a Manhattan job. Since the analytic rated them so poorly we can conclude the market is pricing these benefits into Manhattan

If spending time in Manhattan is not of high priority,then one should look in some of the fringe neighbourhoods on the outer edge of the city: Bayside, Van Cortlandt Park, Staten Island, where one can get the best value for school and services.

If school rating is one's primary importance, then Bayside is the best bet in the 5 boroughs.

If one prioritizes clean streets, less noise and functional city services, It would be best to avoid Central Bronx and some parts of Northern Brooklyn.

VIII. Conclusion

We hope that our analytic application is filled with information useful for Home Buyers, Real Estate Agents and Property Investors alike. Source of this analysis is entirely available on the public domain. But, they lie dormant in a format that is not easily accessible. More such applications are in need to weaponize the publicly available information to enlighten prospective investors and real estate professionals.

IX. Future Work

Given the current situation of COVID-19, all attributes of the analysis are impacted to a greater extent. With the evolving public data, we can perform the same analysis to show the impact of COVID-19 on Property prices.

We have focussed only on one family property. The same analysis can be performed on commercial buildings as well, to give better insight for commercial investors as well.

We could also expand the analysis into other cities and states. Most major cities have openly available 311 datasets and property sales datasets. GreatSchools also has ratings for schools all over the country. It would be possible to use the same analytic application framework and look into other major cities and compare them.

Acknowledgment

We thank HPC Dumbo for its service. We also thank the government of New York City for providing 311 service request data and NYC Property Sales Data. We thank GreatSchools for providing the school ratings data.

References

1. A. Chiodo, R. Hernández-Murillo and M. Owyang, "Nonlinear Effects of School Quality on House Prices", Review, vol. 92, no. 3, 2010. Available: 10.20955/r.92.185-204.

2. L. Wang, C. Qian, P. Kats, C. Kontokosta and S. Sobolevsky, "Structure of 311 service requests as a signature of urban location", PLOS ONE, vol. 12, no. 10, p. e0186314, 2017.

   Available: 10.1371/journal.pone.0186314.

3. P. Visser and F. van Dam, "The price of the spot. Neighbourhood characteristics and house prices in the Netherlands", European Network for Housing Research, 2006. Available: https://www.enhr.net/documents/2006%20Slovenia/W02_Visser_vanDam.pdf.

4. S. Islam, "Impact of Neighbourhood Characteristics on House Prices", Proceedings of ASBBS, vol. 19, no. 1, 2012. Available: http://asbbs.org/files/ASBBS2012V1/PDF/I/IslamS.pdf.

5. L. Chun Chang and H. Lin, "The Impact of Neighborhood Characteristics on Housing Prices-An Application of Hierarchical Linear Modeling", International Journal of Management and Sustainability, vol. 1, no. 2, 2012. Available: https://www.researchgate.net/publication/304597534_THE_IMPACT_OF_NEIGHBORHOOD_CHARACTERISTICS_ON_HOUSING_PRICESAN_APPLICATION_OF_HIERARCHICAL_LINEAR_MODELING.

6. G. Caetano, "Neighborhood sorting and the value of public school quality", Journal of Urban Economics, vol. 114, 2019. Available: 10.1016/j.jue.2019.103193.

7. G. Tita, T. Petras and R. Greenbaum, "Crime and Residential Choice: A Neighborhood Level Analysis of the Impact of Crime on Housing Prices", Journal of Quantitative Criminology, vol. 22, no. 4, pp. 299-317, 2006. Available: 10.1007/s10940-006-9013-z.