Quantifying Peace of Mind

Analyzing 311 Call Data to Deduce Economic Decision Making

Sam Blumenthal

Tom Hervieu

Marvin Lee

Suneet Mishra

Ronak Mukhopadhyay

Jen Newell

Introduction

The aim of our project is to estimate a dollar figure for the price people are willing to pay in order to avoid living in areas sporting a high incidence rate of 311 calls. 311 is a hotline people can call in order to report non- emergency complaints, such as potholes, streetlight issues, noise complaints, etc, and call data is generally publically available in large comma- separated data sets for free. Our approach involves filtering these datasets by their type of call (maintenance, noise complaint, etc), volume (total calls in the year), and frequency (number of calls per month) by year, from 2012 to 2015, and the zip code in which the call was made. We analyzed zip codes in New York City, and Los Angeles, in order to have data points in major metropolitan areas across the country and take into account different housing cultures and attitudes (suburban homes vs urban apartments, etc.) Originally our analysis intended to cover Chicago and Philadelphia, but both publicly available datasets did not have the zip code information required for our analysis, and additionally was presented by type of 311 call, which required significant manual labor in joining the many datasets together. After this data is filtered and ready for use, we matched the zip codes with the real estate prices in the same area code, and analyze trends.

Problem Definition

311 data has been available to the public since the late 1990's, and our literature details some of the studies that utilize this data in their research. For example, 311 incidence rates were analyzed in order to predict civic engagement and voter turnout in municipal elections in the Zha and Veloso study. Throughout our research and literature studies, which mostly concern either analysis done on 311 data or real estate pricing algorithms, no study has actually attempted to analyze the link between 311 incidence rates and real estate prices. We believe that identifying this correlation, if it exists, would be an interesting and important data point in decision making for many audiences, such as policy makers, those looking to move, business owners, and many others.

Survey

The literature from our project comes from a variety of sources. The majority describe methodologies for analyzing 311 or house pricing data, whereas other discuss various algorithms such as neural networks or machine learning that can be used to analyze enormous datasets. The sources that devote analyses to 311 data analyze trends such as voter turnout, civic engagement, interaction with the local government, how to improve upon 311 service times, most likely types of 311 incidences in certain areas, but nothing attempts to correlate real estate values with 311 data. We also found papers that describe the regressors we used and their various benefits.

Method

Our project is distinct from most other studies because we are one of the firsts to correlate 311 incidences with real estate pricing; the majority of other studies have used other factors such as crime rates or demographics. By analyzing different factors we hope to be ahead of the curve on existing research.

We used a wide variety of different methods to analyze our data. Our first task was to get the 311 data. As each city makes this information public, all we had to do was download the data for our target years. Our next task was to clean up the data. To do this, we used a mixture of Python, Pandas, and Seaborn to visualize complaints in various zip codes by type. Our code grouped each complaint by type filed, then counted each occurrence and graphed it on a log scale. Once we had our data we conducted our analysis.

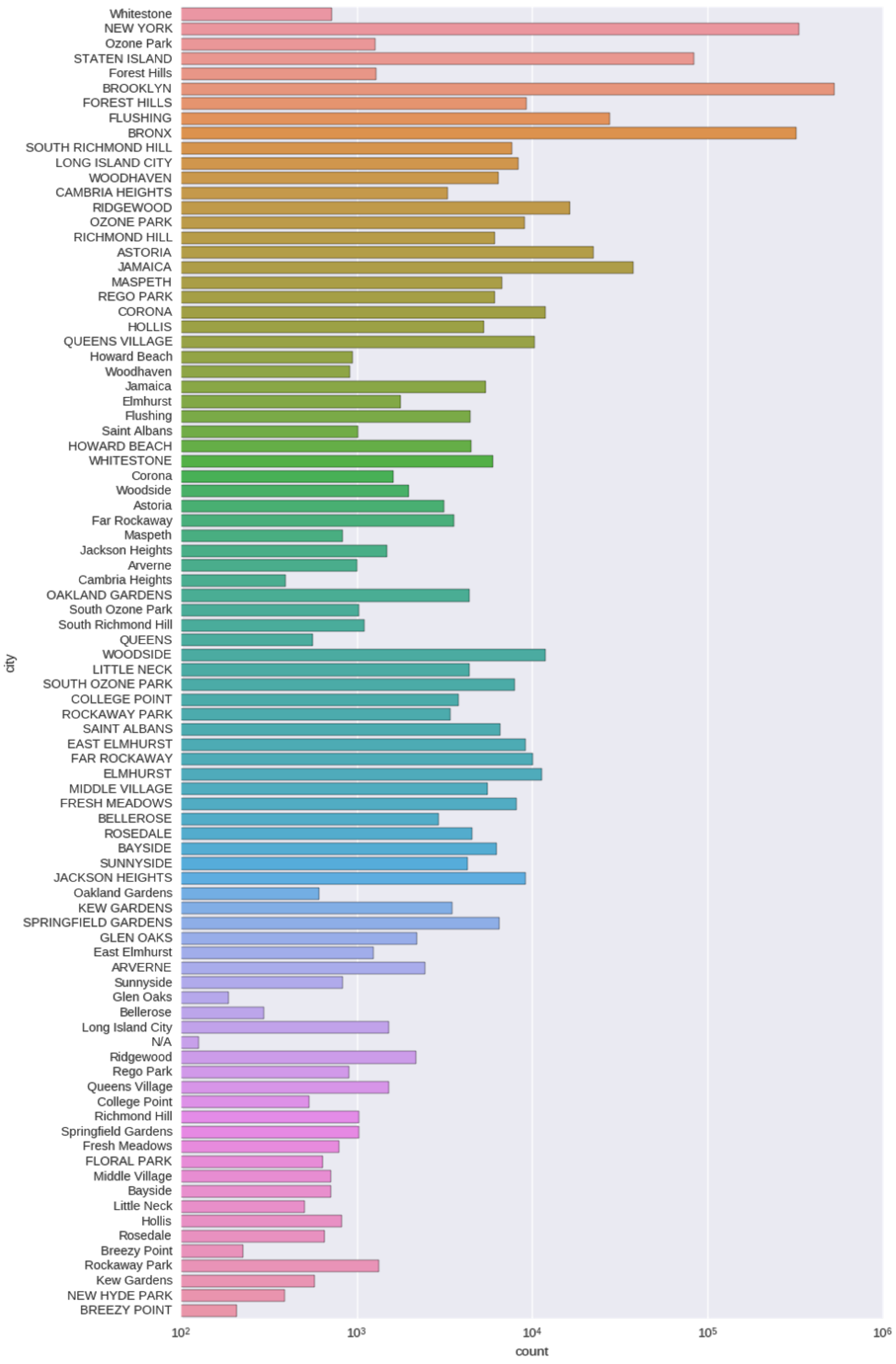
Experiments

Our experiments are designed to answer one question: how much will a person pay to avoid living in an area with a high rate of 311 calls? In addition, we want to see if a high number of certain 311 calls cause more “anxiety” than others. For example, graffiti pushes homeowners away more than potholes.

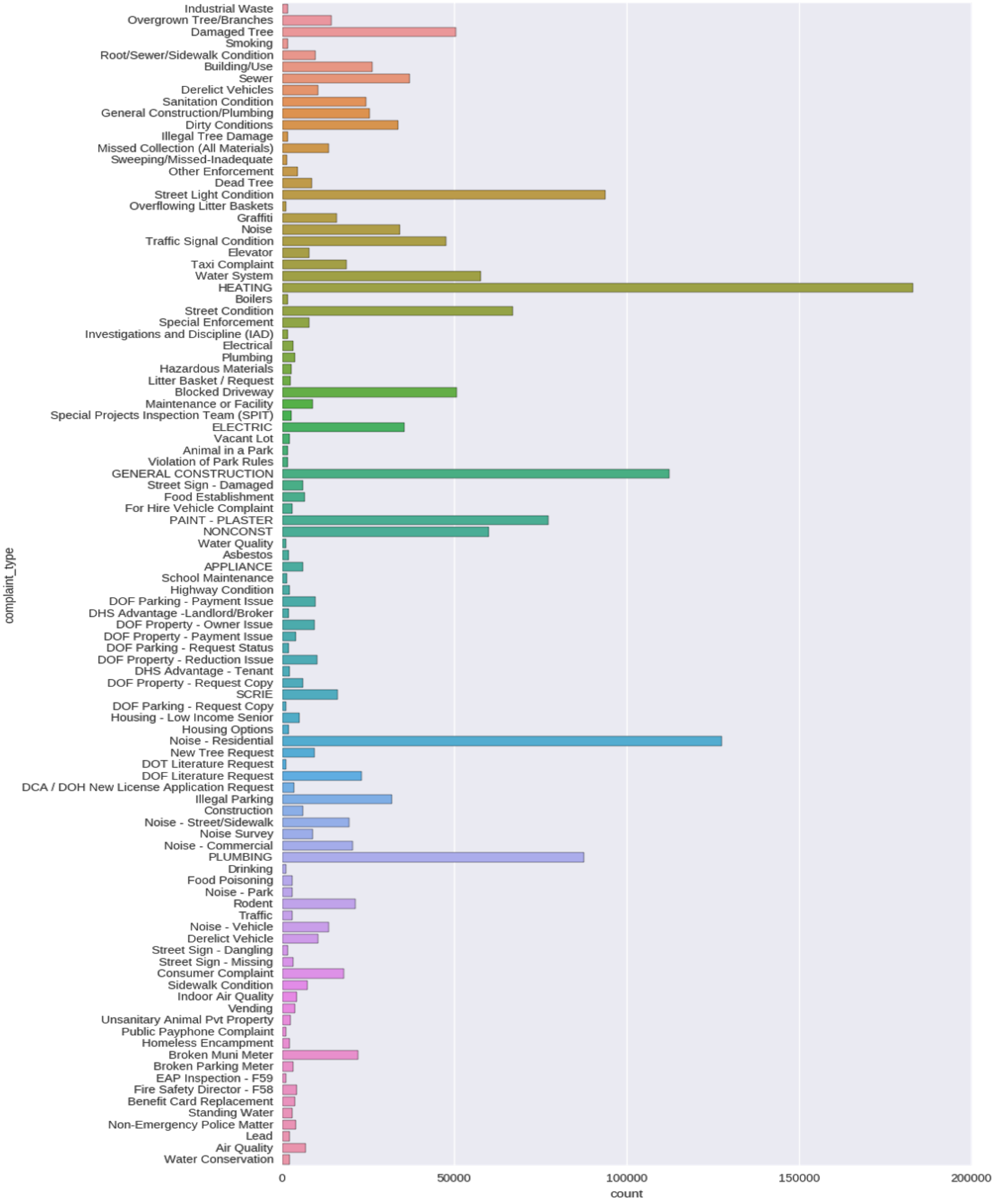
Our experiments will first find trends regarding type, volume, and frequency of 311 data by zip code. For example, we might find that for zip code 10036 (Times Square) there were approximately 50 calls regarding potholes every month during 2014. We did this for each year for each of our target cities, and excluded data that is irrelevant or incomplete. Then, we will find the median price of living in each of the zip codes for which we have data, and find a correlation between the type, frequency, and volume of 311 calls and the cost of living. We also considered other factors that drive up real estate prices that are not necessarily tied to 311 information and take these factors into account when analyzing real estate prices. To our advantage, we are analyzing data starting from 2012, which is three years following the house-market crash- when prices have started to stabilize.

Implementation of Experiments

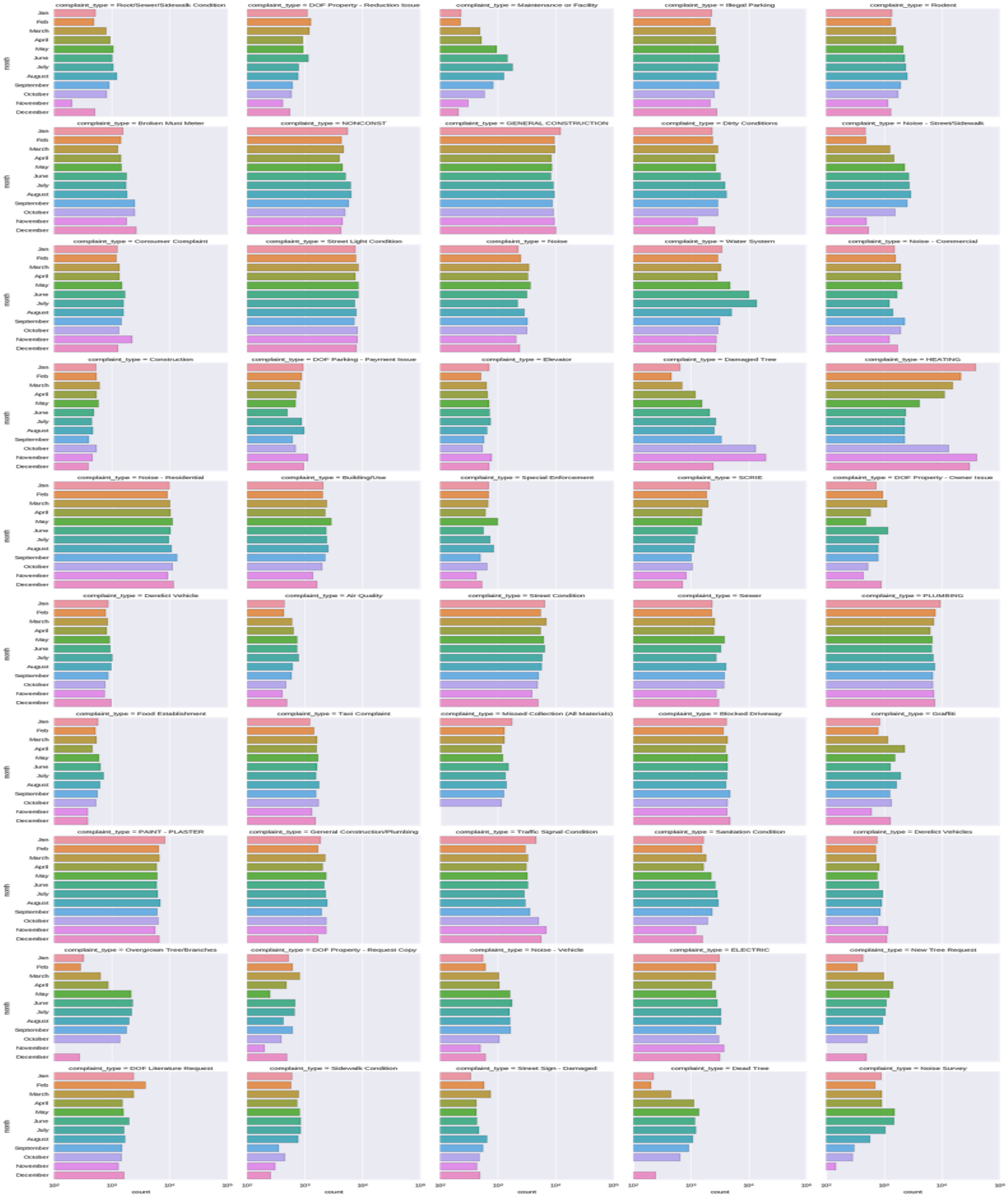
We began by downloading datasets from publically available csv files from the city websites of New York City and Los Angeles. We then visualized the total number of requests against each metropolis area, as shown below:


Whitestone
NEW YORK
Ozone Park
STATEN ISLAND
Forest Hills
BROOKLYN
FOREST HILLS
FLUSHING
BRONX
SOUTH RICHMOND HILL
LONG ISLAND CITY
WOODHAVEN
CAMBRIA HEIGHTS
RIDGEWOOD
OZONE PARK
RICHMOND HILL
ASTORIA
JAMAICA
MASPETH
REGO PARK
CORONA
HOLLIS
QUEENS VILLAGE
Howard Beach
Woodhaven
Jamaica
Elmhurst
Flushing
Saint Albans
HOWARD BEACH
WHITESTONE
Corona
Woodside
Astoria
Far Rockaway
Maspeth
Jackson Heights
Arverne
Cambria Heights
OAKLAND GARDENS
South Ozone Park
South Richmond Hill
QUEENS
WOODSIDE
LITTLE NECK
SOUTH OZONE PARK
COLLEGE POINT
ROCKAWAY PARK
SAINT ALBANS
EAST ELMHURST
FAR ROCKAWAY
ELMHURST
MIDDLE VILLAGE
FRESH MEADOWS
BELLEROSE
ROSEDALE
BAYSIDE
SUNNYSIDE
JACKSON HEIGHTS
Oakland Gardens
KEW GARDENS
SPRINGFIELD GARDENS
GLEN OAKS
East Elmhurst
ARVERNE
Sunnyside
Glen Oaks
Bellerose
Long Island City
N/A
Ridgewood
Rego Park
Queens Village
College Point
Richmond Hill
Springfield Gardens
Fresh Meadows
FLORAL PARK
Middle Village
Bayside
Little Neck
Hollis
Rosedale
Breezy Point
Rockaway Park
Kew Gardens
NEW HYDE PARK
BREEZY POINT
city
10^2
10^3
10^4
10^5
10^6
count

After plotting the volume of 311 data in each area, we plotted the number of 311 calls by type, as show below:



This gave us an insight into what kind of 311 requests were most common. With this knowledge, we filtered the 311 data by type and plotted frequency over the course of the year.



After visualizing this data, we chose regressors to run against our data. We set each parameter to be types of 311 data, and since there were so many to choose from, we narrowed our potential regressors to those that penalize models and deal with having a large number of parameters. The regressors we chose to ran were Ridge, Lasso, Random Forest, and Gradient Boosted.

Each of these regressors is designed to deal with large amounts of data and parameters. Ridge and Lasso are “penalization” regressors in that they seek to minimize the effects of irrelevant factors; Ridge regression aims to compensate for biased estimators and factors by setting those effects to less than a fixed value whereas Lasso regression does the same by setting those effects to zero. Random Forest regression also reduces bias by aggregating prediction models of ensemble trees; this produces a highly accurate model with low bias, which has made Random Forest one of the most popular machine learning prediction algorithms on the market. Finally, Gradient Boosted regression uses boosting that incorporates randomness to further improve on accuracy.

Innovations

Moreover, after identifying trends in data, we can conduct research on significant events happening in certain zip codes or to the entire city in order to confirm trend observations or eliminate or distorted data.

Our approach involves several innovations- as stated previously, there has not been significant undertakings to find relationships between 311 call incidence rates and real estate prices; in the literature we have explored, this correlation has not been mentioned at all.

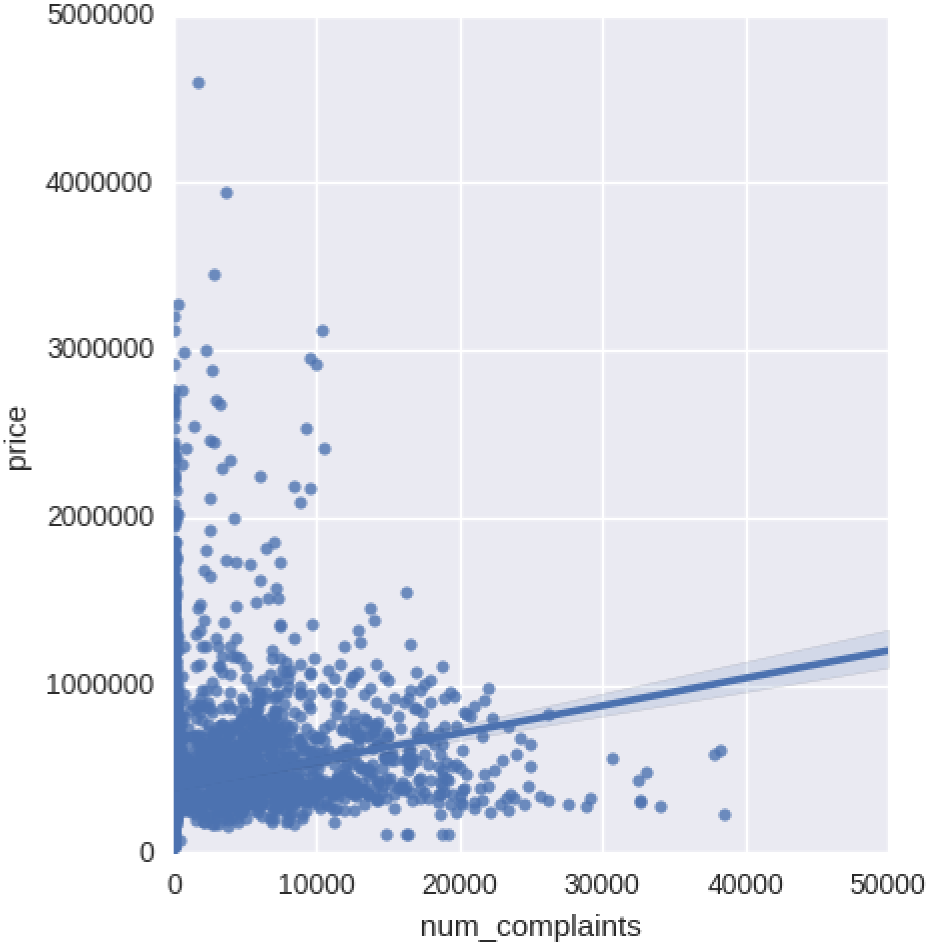
Secondly, our clustering of 311 data by type, frequency, and volume is unique and produces patterns and visualizations that did not exist previously- the Zha and Veloso study has a similar approach to categorizing 311 data but analyzes more factors which produces different groupings altogether.

Conclusion

Our approach was unable to find a significant correlation between 311 incidence and housing prices. Our mean absolute error for NYC was around $120,000 while in LA the mean absolute error was about $190,000. There are hundreds if not thousands of factors that can affect real estate pricing and it is highly unlikely that 311 volume and frequency alone can cause fluctuations in the housing market, although combined with other factors we may be able to find a larger correlation.

Our best performing regressors were two decision-tree based ensemble regressors: Random Forest and Gradient Boosted Regressors. For each city we printed out the top 10 most significant variables. Both cities had similar results although the order of importance was different for both. In New York City noise, taxi complaints, and electric issues were the most important features while in Los Angeles asbestos, construction, and safety issues were more significant.

We then used a linear model based on the total number of complaints for both cities to determine the correlation between the number of yearly complaints and price. We also created a scatter plot with a fit curve to visualize the correlation (below).



The scatter plot had many outliers and the R-squared score was only 0.03. This shows that our model- i.e., using 311 incidence as our only factor- has high levels of bias. However the measured effect size due to the number of complaints was 16.55 with a p-value of 3.59e-54, which is statistically significant. This shows that our model was able to correlate increased 311 calls with increased housing prices by $16.44. This supports our hypothesis that people who live in more expensive areas will be more likely to call 311.

Thus, we are unable to conclude whether or not people will pay more to avoid living in an area with fewer 311 calls, or if it is even possible to "escape" 311-worthy grievances by living in more expensive areas. There are several factors that could explain this; people who own more expensive homes may be more invested in the quality of their neighborhood. Furthermore people living in higher-quality neighborhoods may feel a higher sense of civic pride that inspires them to do more to clean their areas up.

Group Contributions

All group members contributed equally

Literature Survey:

1. http://pracademics.com/attachments/article/607/Ar_3_Schwester_Carrizales_Holzer.pdf

    a. Schwester, Richard W., Tony Carrizales, and Marc Holzer. "An examination of municipal 311 system." *International Journal of Organization Theory and Behavior* 12.2 (2009): 218-236.

2. http://scholar.harvard.edu/files/arwhite/files/noteson311data_july2015.pdf

    a. Trump, Kris-Stella, and Ariel White. "Research Note: The Promises and Pitfalls of 311 Data." (2015).

3. https://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/download/8787/8264

    a. Nalchigar,, Soroosh, and Mark S. Fox. "An Ontology for Open 311 Data." (n.d.): n. pag. Web. 17 Oct. 2016.

4. http://datasmart.ash.harvard.edu/assets/content/cr_87.pdf

    a. Goldsmith, Stephen. *Digital Transformation: Wiring the Responsive City*. Rep. Vol. 87. New York City: Manhattan Institute, 2014. Civic Report.*Manhattan-institute*. Web. 17 Oct. 2016.

5. http://discoveryjournals.com/discovery/current_issue/v44/n205/A2.pdf

    a. Anil Kumar KM, Anil B, Anand CU, Aniruddha S, Rajath Kumar U.Machine Learning Approach to Predict Real Estate Prices. Discovery, 2015, 44(205), 173-178

6. http://yann.lecun.com/exdb/publis/pdf/caplin-ssrn-08.pdf

   a. Caplin, Andrew, et al. "Machine learning and the spatial structure of house prices and housing returns." *Available at SSRN 1316046* (2008).

7. https://link.springer.com/chapter/10.1007/11760191_173
   a. Liu, Jian-Guo, Xiao-Li Zhang, and Wei-Ping Wu. "Application of fuzzy neural network for real estate prediction." *International Symposium on Neural Networks*. Springer Berlin Heidelberg, 2006.
8. https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/viewFile/10206/10261
   a. Chohlas-Wood, Alex, et al. "Mining 911 calls in New York City: Temporal Patterns, Detection, and Forecasting." (2015): 4-10.
9. https://icma.org/Documents/Document/Document/4599
   a. Howlett, Kristin and Tyler P. Reinagel. DeKalb County Process Improvement. *Call Routing for 311: The Issues and Solutions*. October 2008. Print.
10. https://www.ctg.albany.edu/publications/journals/hicss_2013_philly-nyc311/hicss_2013_philly-nyc311.pdf
    a. Nam, Taewoo and Theresa A. Pardo. "Identifying success factors and challenges of 311-driven service integration: a comparative case study of NYC311 and Philly311." *Proceedings of the 46th Hawaii international conference on system sciences*. 2013.
11. http://www.inquiriesjournal.com/articles/1404/2/effect-of-vacant-buildings-on-crime
    a. Chandar, Bharat, and Oliver Dean. "The Effect of a 311 Vacant Building Call on Crime Rates." *The Developing Economist* 3.1 (2016).
12. *http://www.tandfonline.com/doi/abs/10.1080/10630732.2014.887933*
    a. Nam, Taewoo, and Theresa A. Pardo. "Understanding Municipal Service Integration- An Exploratory Study of 311 Contact Centers." *Journal of Urban Technology* 21.1 (2014): 57-78.
13. https://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/viewFile/8834/8266

    a. Yilong Zha and Manuela Veloso. "Profiling and Prediction of Non-Emergency Calls in New York City." *New York University* (2014): 41-47.

14. https://www.researchgate.net/profile/Alexander_Schellong/publication/231168668_The_evolution_of_311_A_review_of_the_City_of_New_York/links/0912f50657afb7c473000000.pdf

    a. Charis E. Idicheria, Alexander Schellong, and Jobst Fiedler. "A Review of 311 in New York City." *ISPRAT* (2014): 11-30.

15. http://knowledgecenter.csg.org/kc/system/files/Holzer_Article_1.pdf

    a. Marc Holzer, Richard Schwester, Angie McGuire and Kathryn Kloby. "State-level 311 Systems: Leveraging Service Enhancement and Performance Measurement at the State Level." *Graduate Department of Public Administration* (2006): 409-413.

16. http://uar.sagepub.com/content/52/2/211.short

    a. Minkoff, Scott L. "NYC 311: A Tract-Level Analysis of Citizen–Government Contacting in New York City." *Urban Affairs Review* 52 (2016): 211-46. Web. 17 Oct. 2016.

17. http://www.ejst.tuiasi.ro/Files/48/15_Trnka.pdf

    a. Trnka, Adrej. “Big Data Analysis.” European Journal of Science and Theology 10.1 (1014): 143-148.

18. http://www.rosebt.com/uploads/8/1/8/1/8181762/big_data_the_management_revolution.pdf

    a. McAfee, Andrew, et al. "Big data." *The management revolution. Harvard Bus Rev* 90.10 (2012): 61-67.

19. http://www.jarad.me/stat615/papers/Ridge_Regression_in_Practice.pdf

    a. Marquardt, Donald W., and Ronald D. Snee. "Ridge regression in practice." *The American Statistician* 29.1 (1975): 3-20.

20. http://www.jstor.org/stable/2346178?seq=1#page_scan_tab_contents

    a. Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* (1996): 267-288.

21. http://pubs.acs.org/doi/pdf/10.1021/ci034160g

    a. Svetnik, Vladimir, et al. "Random forest: a classification and regression tool for compound classification and QSAR modeling." *Journal of chemical information and computer sciences* 43.6 (2003): 1947-1958.

22. http://ac.els-cdn.com/S0167947301000652/1-s2.0-S0167947301000652-main.pdf?_tid=64426b14-bd8e-11e6-b5c3-00000aab0f6b&acdnat=1481233174_3d89fd5462e03ff6b105ac43fd530807

    a. Friedman, Jerome H. "Stochastic gradient boosting." *Computational Statistics & Data Analysis* 38.4 (2002): 367-378.