What's Your Problem: An Analysis of Urban Conditions in New York

Chelsea Chandler, Cassidy Haas, Annika Muehlbradt and Bryce Wilson
Department of Computer Science
University of Colorado Boulder
Boulder, CO 80309 USA
{chelsea.chandler, cassidy.haas, annika.muehlbradt, bryce.d.wilson}@colorado.edu

ABSTRACT
This project explores the urban condition of New York through the analysis of spatial and temporal distributions of 311 service call data and NYPD complaint data. We provide a brief overview of the datasets and our approach to data preprocessing and sampling. We then discuss our methodology and share insights into interesting patterns of complaint types by year and location. Our findings include distributions of complaints by global coordinates, borough, month, and year, explanations of outliers and anomalies, and an in depth analysis of the differences between neighborhoods by zip code. We conclude our discussion with a short evaluation and project reflection.

Author Keywords
New York City; Urban Condition; Data Mining.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

1. INTRODUCTION

New York is one of the largest metropolitan areas in the United States; it is home to more than 8 million people of diverse heritage, age, identity, and interests, and more than 2 million businesses. As a melting pot, the city frequently experiences friction between its inhabitants. New York City (NYC) officials receive thousands of 311 service requests a day, and hundreds of complaints are filed with the New York City Police Department (NYPD). By reporting these complaints, inhabitants are providing a detailed account of the conditions of urban life. So what is urban life like in New York City?

To explore this question, local government, newspapers, and journals have investigated New York City's urban condition. The New York Times has looked at the city's noise distribution, I Quant NY investigated NYC property taxes, and the New York City Department of Health and Mental Hygiene has tracked air quality over several decades. While these are interesting factors to explore, the analysis of these individual conditions reveals little about urban life in the city. 311 service call data and NYPD complaint data offer insights about the city's daily hustle and bustle by exposing residents' accounts of the city's affairs. Analyzing this data can inform us about the urban condition, about urban trends and unique events.

This project seeks to explore the urban condition of New York City through an in depth analysis of 311 service call data and NYPD complaint data. We plan to analyze and map the spatial and temporal distribution of residents' complaints to assess which NYC neighborhoods offer the best living conditions. We also want to provide strong association rules on resident behavior. By understanding normal patterns of behavior, we can make inferences about anomalies and assess if these anomalies correspond with specific events.

2. RELATED WORK

While no general exploratory work has been done on NYC 311 service requests and NYPD complaint data, we have found quite a few articles and papers that touch on the topics we are interested in. In a more general study, Coburn et al. utilizes Geographical Information Science and spatial distribution analysis to categorize areas into neighbourhoods [6]. This is an approach we plan to utilize in our project. Bonaiuto et al shows the relevance of predictors from four areas in predicting attachment to neighborhoods, namely architectural and town-planning features, social

relations features, punctual and in-network services, and context features [2]. We believe that this paper will shed some light on the different reasons people choose to live where they do. This knowledge will come in handy when it is time for us to draw some conclusions from our results. Neckerman et al. analyzes individual block faces in NYC and their aesthetic, commercial, and criminal tendencies [1]. This work will be particularly useful in studying the intricacies of life in NYC. We will be able to compare our findings with this study in order to draw better conclusions and refine our search. Finally, Trump et al. cautions against interpreting 311 service requests data as a measure of mass civic engagement akin to voter turnout, but argue that it is a potentially useful measure of aggregate demands for public service [7]. While this work does not touch on the same types of questions we will be asking, it will be extremely important when it comes time to draw conclusions. Namely, we need to remember that this dataset is not the end-all be-all source of information on social conditions in NYC and we must take these cautions seriously in our analysis.

Our work is very different from the related works that we have come across; these works are either more general in location or more specific in goals. Our study will be able to harness the 311 service requests and NYPD complaint data to draw some conclusions on the populace of NYC. We will want to answer important questions drawn from the dataset as well as uncover important insights in human behaviour.

3. DATA DESCRIPTION

Our work looks at two primary datasets obtained from the NYC Open Data project: 311 service requests data and NYPD complaint data. Both datasets are freely available on the project's website and can be downloaded as a XML spreadsheet. One supplementary dataset, the New York census data, provides insight into trends and patterns found in the primary datasets.

3.1 311 Service Requests Data

The 311 service requests dataset contains information about the daily service requests received by the city of New York from 2010 through 2017. The objects are the individual service requests and the attributes provide information such as the date the service call was received, the type of complaint, the address at which services are being requested, etc. Most of the attributes are nominal and represented by text entries. The dataset contains roughly 16.4M records with 53 attributes each. The dataset is sparse as some of the attributes are mutually exclusive -- the events described by the attributes cannot occur simultaneously. For example, a service request pertaining to the removal of a dead tree will not have an entry for the attribute "vehicle type."

3.2 NYPD Complaint Data

The NYPD complaint dataset contains information about criminal and non-criminal offenses that occurred within the city of New York between 2006 and 2016. Each record describes a single offense that was reported by a New York city police precinct. The attributes provide information such as the type of offense, the date on which the offense occurred, the address at which the offense occurred, etc. Most of the attributes are nominal and represented by text entries. The dataset contains roughly 5.58M records with 24 attributes each. The dataset only has two sparse attributes, park name and development name.

3.3 Other Data

3.3.1 Census Data

The census data was obtained from the New York City Department of City Planning and contains selected demographic information by zip code such as gender and ethnicity, and by borough such as employment status and income for the years 2011 through 2015. These data are estimates and a margin of error is provided for each data point.

3.3 Data Limitations

Our primary datasets have one limitation: the data has been collected and recorded manually by New York City employees. This makes the data error prone and subject to missing values. For example, the date on which a 311 service request was completed is missing more than 50% of the time. It is probable that some service requests were never resolved, though the high percentage of missing values indicates that employees were likely to forget to fill in this value.

As mentioned above, the limitation of the New York census data is that these data are estimates with margins of error as high as 10%.

3.4 Data Management

To facilitate data mining, the datasets are stored on GitLab, a web-based repository management system.

For this project, GitLab serves as a data warehouse in that it stores data from different sources and facilitates reporting of analytics. It allows us to quickly share analysis tools and new discoveries. From GitLab, the datasets are imported to local database instances for data processing. The local databases are powered by MongoDB, a flexible system for storing and processing records. MongoDB offers native native analytics tools for data search and data aggregation, fast indexing, and a rich query language. Using MongoDB, we are able to perform data mining tasks such as classification and outlier analysis in parallel.

4. PROBLEM FORMULATION

We propose to mine the NYC 311 service requests data set and the NYPD complaint data set for patterns and anomalies such that we can make inferences about the urban conditions of New York City. To break the analysis of the datasets into smaller, more clearly defined subtasks, we devised the following questions:

1. What is the spatial and temporal distribution of service requests and complaints?
2. Can we define clear neighborhood boundaries based on the spatial distribution of service requests and complaints?
3. What is the frequency of service requests and complaints?
4. Are there any correlations between service requests and complaints?
5. How do service requests and complaints change over time?

We plan to create a number of high fidelity visualizations for each of the subtasks to aid in the interpretation of the results. We hope that the convergence of these results will provide a clear picture of the urban trends and will allow us to make predictions about future events in NYC.

5. METHODS

5.1 Data Preprocessing

Our data preprocessing focused on three areas: attributes with missing values, attributes with redundant values, and dimensionality reduction. In preprocessing the missing values, we were able to generate entries for attributes based on information gained from others. For example, a common missing value in the 311 dataset is “Closed Date” for when the 311 service request was properly dealt with and the issue was closed. To amend this, we calculated the mean duration of an open ticket for a given complaint type (e.g. Noise, Construction, Mold, etc.) and used that to fill in missing “Closed Date” values. The 311 service request data contains redundant information. Namely, it has both attributes for intersection streets and cross streets. When an entry had one but not the other, we were able to use one to fill the other. Once they were all filled we were able to do some attribute reduction. Now we have redundant information like intersection and state planar coordinates since we already have cross street and latitude/longitude. There was also irrelevant data, such as the city the complaints were made in. Since all of these entries stated “NYC”, we were safe to cut out that attribute completely. Reducing the number of attributes allows us to make faster searches and frequent pattern finding. Once the attributes were reduced, we also reduced the dimensionality. If we had a sparse entry, we would cut it from the dataset because it wouldn’t help us with our analysis later on. All in all, we were able to reduce our dataset by 30%

5.2 Sampling

In order to further reduce the size of our data and to prepare the data for classification with multiple models, we performed two kinds of sampling. First, we randomly sampled each dataset (without replacement) to compose 10 samples each containing roughly 1M records. These samples provide a basis for performing k-fold cross-validation in which each of the samples represents the test set in one of the iterations of testing. The second sampling method we used was stratified sampling in which each strata was composed of randomly selected records for a specific year (e.g. 2008). The number of samples per dataset varied as the 311 service requests dataset contains records for 8 years (2010 through 2017) and the NYPD complaint dataset contains records for 11 years (2006 through 2016).

Each sampling method has its advantages. Sampling for cross-validation provides a simple way to split the data into training and test sets. With multiple random samples, many data mining tasks can be performed in parallel. In comparison, stratified sampling ensures that training and test sets are not skewed as a result of oversampling records from a particular year. Stratified

sampling also provides a means of comparing patterns year-to-year.

5.3 Basic Statistical Descriptors

We started our analysis by looking at some basic distributions of the number of complaints and complaint types. These distributions did not reveal anything out of the ordinary, but they did guide further investigation. For example, looking at the distribution of the 277 different complaint types we found "residential noise", "heating", "street conditions", and "street light condition" to be among the top complaints. These types of complaints are often frequent in larger cities while a myriad of other complaints happen more sporadically as seen in Figure 1.

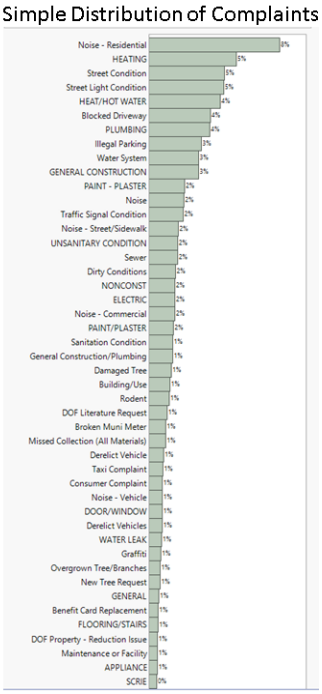

Figure 1: Distribution of complaint type.

Similarly, looking at the distribution of the number of complaints by borough in Figure 2 revealed that Brooklyn has the most complaints (nearly 10% more complaints than any other borough). This is also not surprising as Brooklyn has the largest population. In fact, a visualization of complaints by latitude and longitude can be overlaid onto a map of New York to see that the number of complaints roughly correlates with population size.

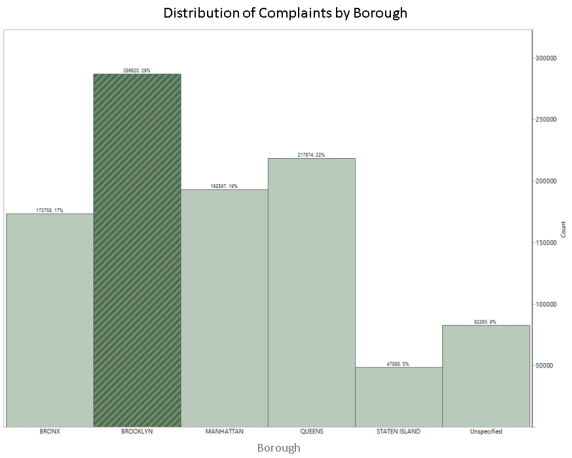

Figure 2: Distribution of the number of complaints by borough.

After analyzing the basic distribution of complaints and complaint types, we turned to spatial distributions of complaints by latitude and longitude. We wanted to find out in which areas certain kinds of complaints were most common. Using geographic coordinates, we plotted all noise complaints for the year 2011 (see Figure 3).

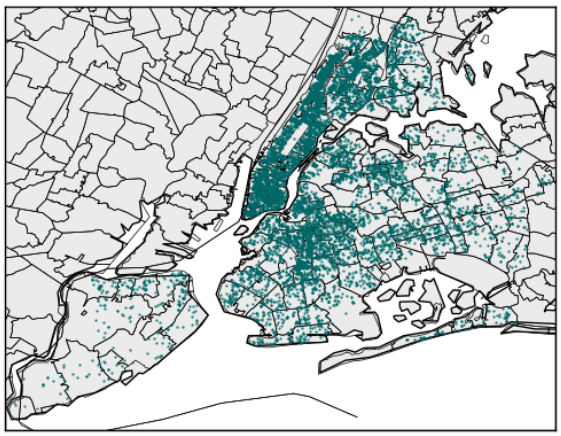

Figure 3: Spatial distribution of noise complaints in 2011.

Looking back at the distribution of complaints by borough, one might expect Brooklyn to have the highest number of noise complaints. However, we are specifically looking at the spatial distribution, or concentration, of complaints. Figure 3 shows that Manhattan actually has the highest number of noise complaints. One possible explanation is that Manhattan has the highest density of people and thus is more sensitive to noise issues.

Lastly, we created a temporal distribution to probe for trends in the number of complaints by month, season, and year. Figure 4 revealed that there are significantly more complaints in the winter months than in the summer months. While there was also an increase in complaints year over year (2010 through 2017), the highest number of complaints reported was in January 2014.

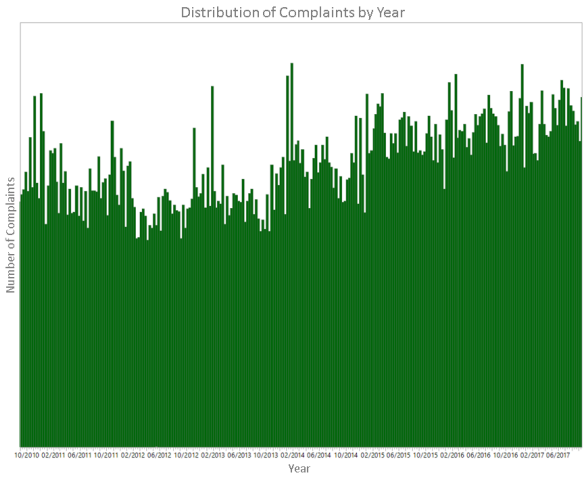

Figure 4: Distribution of complaints by year.

6. FINDINGS

6.1 Classification

In a more in-depth look at potentially associated variables and individual complaint types, we took each complaint type as a proportion of the complaints made that week and averaged that over the years, giving us relative frequencies per week. Figure 5 shows a frequency count of all complaints combined in this format as a baseline. Of note here is that there is a drop in calls around the holidays. However, the final data point (which is around 2500) is not representative of a full week; a year has 52 groups of 7 days and 1 day remaining, so week 53 on these weekly graphs is always just December 31st. A significant drop in calls is still present in a daily frequency graph, however.

As for individual complaint types, it's common to find that certain types of calls are made more often in warm months or cold months. For example, calls about heating and hot water are (unsurprisingly) very common in winter and nearly nonexistent in summer.

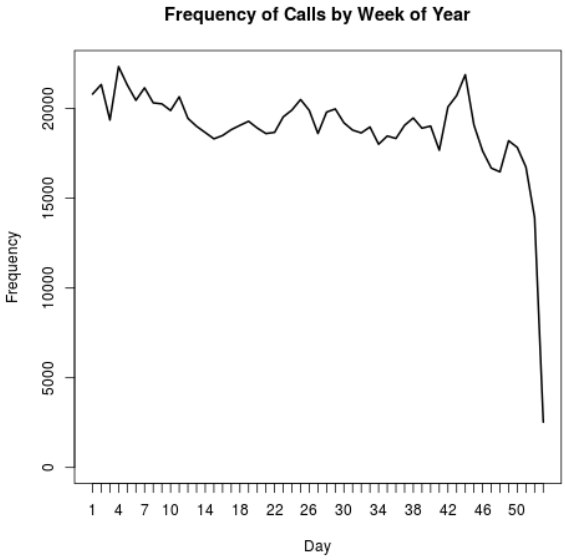

Figure 5: Average weekly frequency of total calls.

Potentially contrary to what one might guess on intuition, calls about parking violations come largely during warm months. As such, the relative frequencies of calls about heat and calls about parking have a relatively strong negative correlation (coefficient of determination equal to 0.6228).

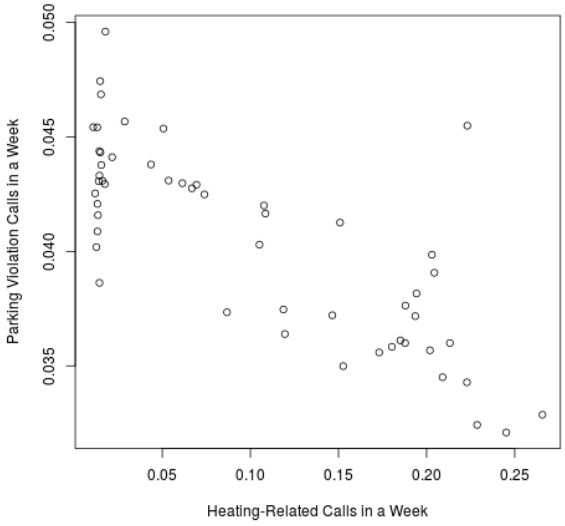

Figure 6: Proportion of heat-related calls plotted against proportion of parking-related calls in a week.

One of the stronger linear relationships we found was between noise complaints and calls concerning rodents (Figure 7).

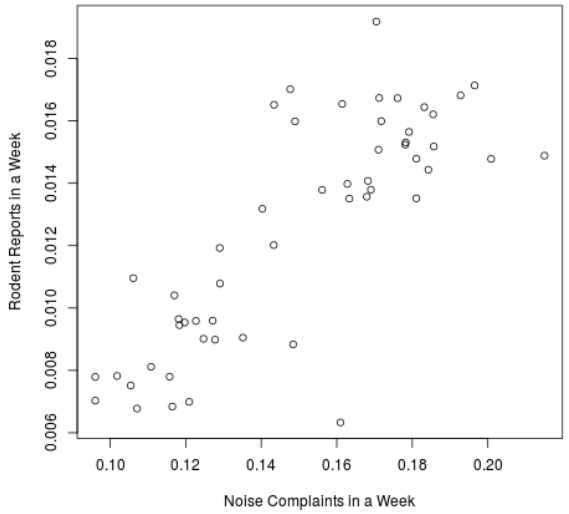

Figure 7: Proportion of noise complaints plotted against proportion of rodent-related calls in a week.

There is a similarly strong relationship between rodent reports and food-related calls (reports of food poisoning which may be explainable separately from seasonal changes.

Many of these relationships might be explained more by season than anything else, but incorporating weather data would be necessarily complex and worth a study in and of itself.

One type of complaint that is unexpectedly seemingly not related to season is construction (Figure 8). It is a common notion that construction happens more during summer months, so one might expect the number of complaints to be higher during the summer months. However, there appears to be no significant change based on season. There is, however, a very large amount of variability from week to week, which is especially curious given the weekly data is an average over 6 years, which should reduce the variance and expose patterns. Data spanning more years would help to better identify weeks that should be investigated (or it would reveal that there aren't outliers as there appear to be in this data.

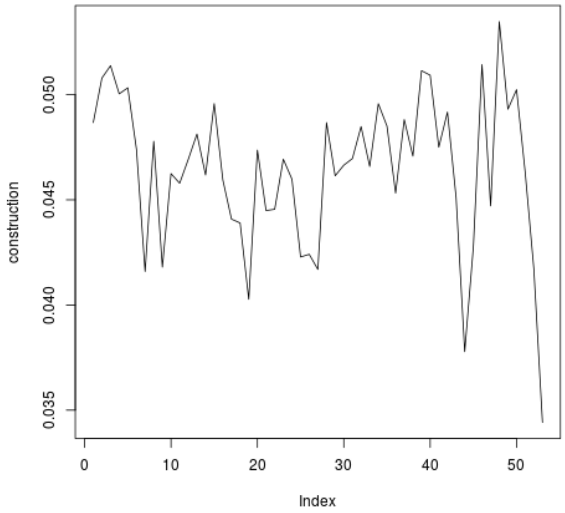

Figure 8: Proportion of calls made over time that had to do with construction.

6.2 Outlier Analysis

In our outlier analysis, we focused on finding both stray data points as well as anomalies caused by unexpected events. To start out, we explored the data using k-Nearest Neighbor clustering where $k = 1, 2, 3, ..., k$ skipping values by the Fibonacci sequence to avoid extraneous computation. We plotted each k-Nearest Neighbor computation on a multivariate scatter plot. As noise was among the top complaints in our basic statistical analysis, we performed k-Nearest Neighbor clustering on latitude and longitude of the subset containing all noise complaints from the years 2010 through 2017 (see Figure 9).

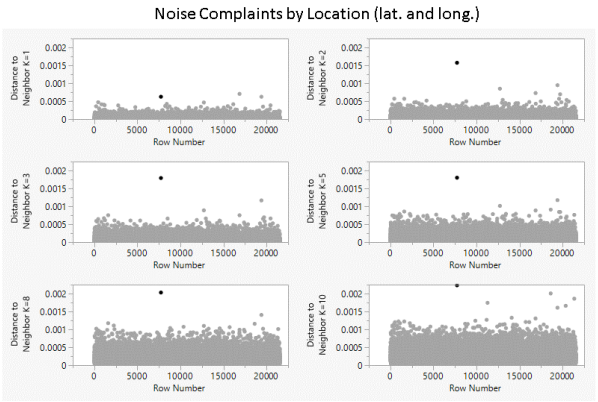

Figure 9: Multivariate k-Nearest Neighbor plot.

The scatterplot shows few random, global outliers. We were unable to determine if these outliers were significant (or simply errors caused by misreporting data) by simple inspection of the data points. Thus, we compared the data points to a visualization of the spatial distribution of all noise complaints shown in Figure 10. Inspecting the visualization we can see that these outliers exist in isolated, but populated areas opposed to a park or a inhabitat isle. We can assume that these were likely not misreported and are in fact outliers.

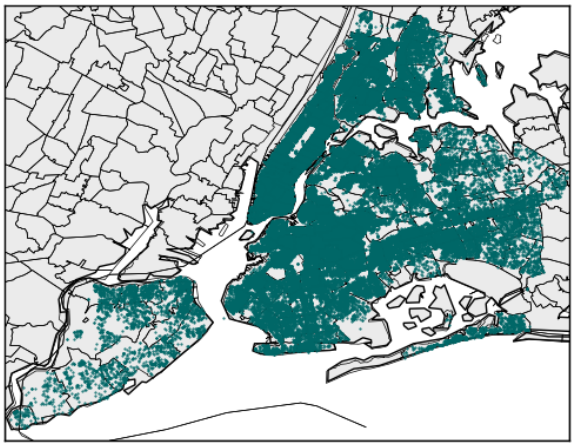

Figure 10: Spatial distribution of all noise complaints.

In addition to noise, we looked at a number of other complaint types including “heating”, “water system”, “street/sidewalk noise”, and “street condition”. Most k-Nearest Neighbor plots showed few outliers similar to the cluster of noise complaints seen in Figure 9 above. Some of the aforementioned complaint types, however, also revealed interesting contextual and collective outliers. For example, the complaint type “heating” shows an up and down, seasonal trend for the years 2012 through 2014. There are no complaints reported for the year 2015. When we expand the analysis to include the distribution data of heating complaints for all years, we can see that there are no complaints reported after 2014. We can surmise that the city decided to stop tracking this complaint type and that these are collective outliers (see Figure 11).

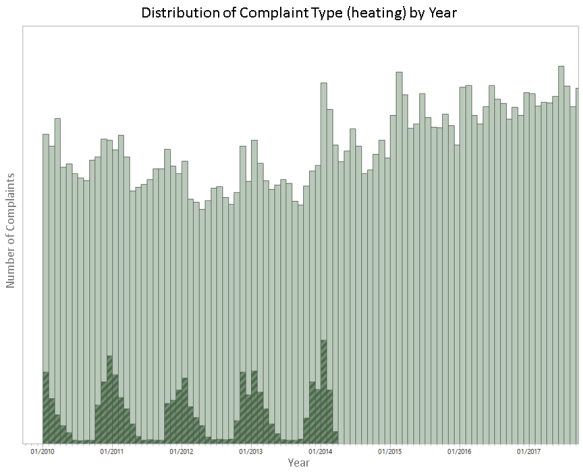

Figure 11: Distribution of heating complaints overlaid on top of the distribution of all complaints by year.

An example of an interesting contextual outlier uncovered by our analysis was that of an increase in street condition complaints in the month of April 2015. As shown by Figure 12, there is a clear ramp and spike in street condition complaints which then taper off over a couple of months. A Google search revealed that the demolition of a plot of land next to Grand Central Station was underway during April of 2015 to make way for the new One Vanderbilt building. A subset analysis of latitude and longitude confirmed that the spike in complaints largely originated in areas near Grand Central Station. We can conclude that there is a strong correlation between the demolition and the increase in complaints, and that this is a contextual outlier.

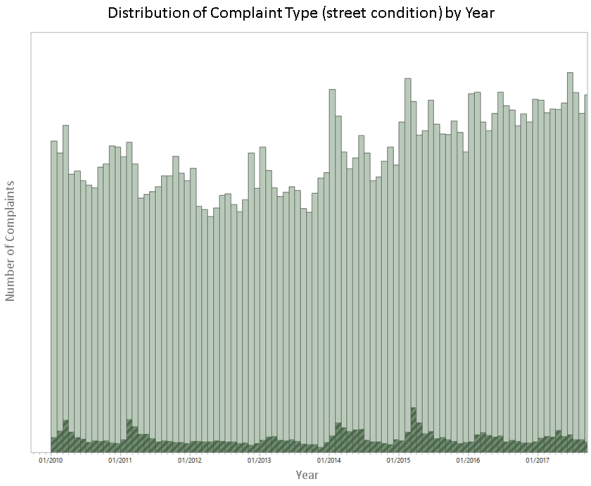

Figure 12: Distribution of street conditions overlaid on top of the distribution of all complaints by year.

6.3 Subset Analysis

In order to do a thorough analysis of conditions in New York City, we split the data into subsets corresponding to zip codes and did a general analysis on similarities and differences between neighborhoods. To start off, we downloaded shapefiles from the internet to make visualizations. The shapefiles contain zip code data so it was easy to extract information from the 311 data and intersect it with the geometries from the shapefile in order to visualize different properties and relationships.

The first area we explored was whether there were any zip codes that were outliers with respect to the number of complaints received in a year normalized by the population of the area. Using both a percentile-based outlier test (using the 95th percentile) and a median-absolute-deviation test (with a z-score threshold of 3.5), we were able to identify multiple neighborhoods whose complaint counts differed from the rest with statistical significance. The zipcodes 11040 (New Hyde Park), 10119 (a small block of Manhattan near Korea Town), and 10162 (a few blocks of Upper East Side in Manhattan) all were below the normal amount, and 10018 (Garment District), 10004 (Governors Island), and 10040 (Washington Heights) were all above the normal amount according to the percentile-based test. Only 10018 was considered an outlier according to the median-absolute-deviation test. The Garment District clearly had more complaints per capita than any other zip code. Figure 9 shows just how drastic this normalized complaints count outlier is. The graph displays each zip codes' normalized complaint counts for the entire year with the zip codes removed from the x axis for clarity.

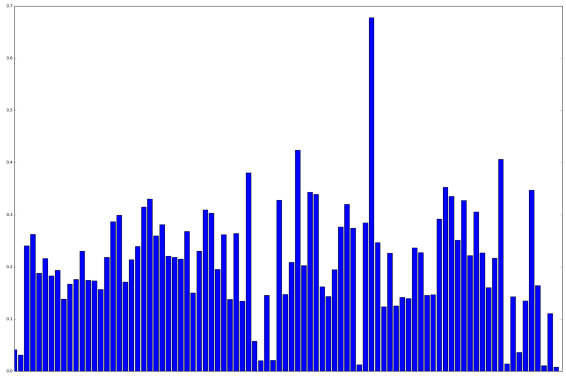

Figure 9: Normalized complaint count by zip code.

Next we explored the most popular complaint type in each zip code. The most popular complaint types overall are Heating and Street Condition. When looking at the most popular complaint among entire boroughs of NYC, Staten Island was overwhelmingly Street Condition, the Bronx was overwhelmingly Heating, and Manhattan was a mix of Heating and Street Condition, but with an unsurprising addition of Noise and Taxi Complaints. Figure 10 shows just how many heating complaints were made in the Bronx. This is extremely telling in the analysis of conditions in NYC as a whole. We can see which boroughs struggle the most with different issues.

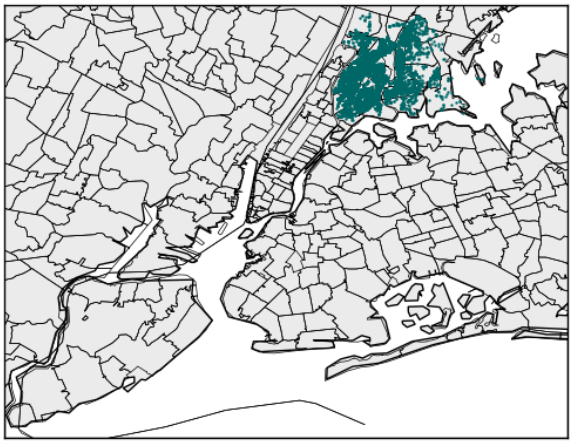

Figure 10: Heating complaints in the Bronx.

Finally, as a time series analysis, we analyzed how the complaints differed from year to year in specific zip codes. While this analysis could be done on every neighborhood in the dataset, we chose to refine the exploration to just one area. We chose to investigate the complaint trends in Williamsburg. This neighborhood is known to have gone through a gentrification and quickly became one of the top places to live in NYC within the past 10 years. Because of this, we hypothesized the trends of this area would be the most interesting to explore. While the gain in population in most neighborhoods in NYC have risen in amounts proportional to the rise in population of the city as a whole, Williamsburg has not. This neighborhood started becoming gentrified in the early 2000s and became a hub for so-called 'hipsters' soon after. In the mid 2010s, it became less hipster and more of an extension of Manhattan as far as demographics go.

Figure 11 shows the drastic rise in 311 calls starting in the year 2015.

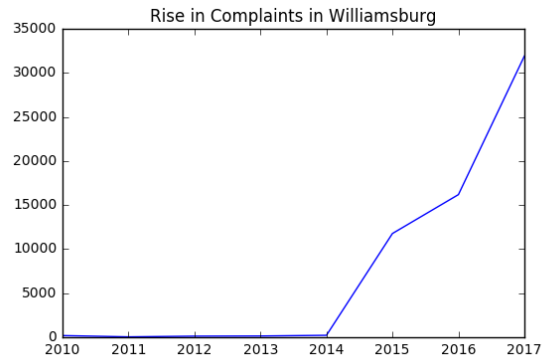

Figure 11: Complaints made in Williamsburg over the past 7 years.

This rise in the year 2015 is not surprising. Although 311 calls do not contain demographic information of the callers, research by Dr. Joscha Legewie, a sociologist at New York University, showed that most complaints are made by middle-aged white people [8]. It wasn't until the mid 2010s that Williamsburg saw the flocking in of upper class business people and the flocking out of the so-called hipster.

The most popular type of complaint also saw a shift in 2015. The years 2010 through 2014 all shared the same most frequently made complaint: Traffic Signal Condition. This was no longer the case after 2014. The most frequent complaints made between 2015 and 2017 were concerning Mold and Site Safety. Does this mean that Williamsburg invested money into their infrastructure around this time and improved traffic conditions, or that with the influx of upper class people, living and working conditions took precedence? One possible answer could be that with gentrification comes a rise in construction and the corresponding uncovering of mold and more opportunities for job safety concerns. It has become quite obvious that the demographics and changes in population in a neighborhood can play a huge role in the types and quantity of 311 complaints received.

7. EXPERIMENTAL EVALUATION

In an exploratory experiment such as this one, the evaluation can be done empirically. In the results we obtained through clustering, classification, rule finding, outlier analysis, etc., we were able to uncover trends and facts about the state of the world. For example, when outliers arose in the data, we were able to verify that there was in fact an event that sparked a change in trends. Additionally, when we say trends in the type of complaints people tend to make in certain areas, we can understand why; road conditions in Staten Island and noise and taxi complaints in Manhattan both make perfect sense. Being able to back up the insights found in the data with real knowledge from the world helped us to verify that what we were uncovering in the data was legitimate.

In addition to using real world data, we were also able to compare our findings to other relevant works (see introduction and related works). For example, our distributions of noise data matched distributions and time series analyses compiled by Wired Magazine. Our interpretations of these distributions were similar -- though our investigation was far more extensive -- and we can therefore surmise that these conclusions are accurate.

8. KNOWLEDGE DISCOVERY

The knowledge discovered in this project was mostly general. We were able to back up many things that people assume to be true in the world. For example, when a huge construction project is underway, there is no doubt that the number of complaints will skyrocket; when a neighborhood is being gentrified and the demographics are rapidly changing, the behavior of the constituents will change as well. Not only do different types of people tend to view certain problems as nuances when others wouldn't, but different types of people also tend to call in to 311 to complain more frequently than others.

There are types of complaints that are ubiquitous across the nation: troubles with heating and road conditions to name a couple, but there are also complaints that are specific to a certain type of area. The types of complaints that are made on a regular basis in different areas are quite telling of the living conditions.

Finding relationships between types of complaints was unsurprising, but again, neat to see backed by data. Correlations such as the one between food quality and rodents or the one between season of the year and number of heating complaints were strongly backed by this dataset.

The knowledge gained from our analysis of resident behavior can be used to predict when and why residents will issue 311 service requests. For example, if certain attributes are present or events occur, this could indicate a possible onslaught of 311 complaints.

9. PROJECT REFLECTION

With such a large dataset, this analysis made obvious the need for sampling and data reduction. Had the data not been cleaned and reduced before our exploration, we would have encountered more noise and slower performance on our tests. Data preprocessing is necessary not only for accuracy but for allowing more exploratory work to be accomplished with less time and computer power.

Visualizations of the data is one of the biggest keys to understanding distributions and changes over time. Without making graphs of our data, it would have been extremely difficult to analyze the data and draw conclusions. Some visualizations even sparked ideas of different routes to investigate next and that is a very important factor in an exploratory project such as this.
In the analysis of the 311 data, the importance of bringing in knowledge about the status of the world became overwhelmingly clear. What may at first look like abnormal and strange behavior both in changes over time and between different geographic areas can be easily explained by commonly-known facts about the world that are not explicitly in the data. Interpreting the data would be impossible without domain knowledge.

9. CONCLUSION

This project successfully explored the living conditions in different neighborhoods of New York City. Through in depth analysis of trends and associations of attributes in 311 service call data, we were able to gain a thorough understanding of the quality of life for the general population of NYC. In this paper we uncovered what were consistent, as well as anomalous, urban conditions and behaviors across the different boroughs. We were able to capture resident behavior in a way that could be used to better predict the frequency and types of 311 service requests, as well as the different types of instigators for time periods with frequent calls.

REFERENCES

1. Neckerman, K. M., Lovasi, G. S., Davies, S., Purciel, M., Quinn, J., Feder, E., ... & Rundle, A. (2009). Disparities in urban neighborhood conditions: evidence from GIS measures and field observation in New York City. *Journal of public health policy*, *30*(1), S264-S285.
2. Bonaiuto, M., Aiello, A., Perugini, M., Bonnes, M., & Ercolani, A. P. (1999). Multidimensional perception of residential environment quality and neighbourhood attachment in the urban environment. *Journal of environmental psychology*, *19*(4), 331-352.
3. Rundle, A., Roux, A. V. D., Freeman, L. M., Miller, D., Neckerman, K. M., & Weiss, C. C. (2007). The urban built environment and obesity in New York City: a multilevel analysis. *American Journal of Health Promotion*, *21*(4_suppl), 326-334.
4. Mielke, H. W., Gonzales, C. R., Smith, M. K., & Mielke, P. W. (1999). The urban environment and children's health: soils as an integrator of lead, zinc, and cadmium in New Orleans, Louisiana, USA. *Environmental research*, *81*(2), 117-129.
5. Krupat, E. (1985). *People in cities: The urban environment and its effects* (No. 6). Cambridge University Press.
6. Corburn, J., Osleeb, J., & Porter, M. (2006). Urban asthma and the neighbourhood environment in New York City. *Health & place*, *12*(2), 167-179.
7. Trump, K. & White, A. (2015) The Promises and Pitfalls of 311 Data. *Forthcoming, Urban Affairs Review.*
8. Ryna, B (2015) What 311 Calls Can Tell Us About Gentrification. *The Cut*