Rodent Baiting in NYC

Sanchit Khandelwal

Master's in CS Student, NYU

New York, USA

sak649@nyu.edu

Rohit Shankar

Master's in CS Student, NYU

New York, USA

rs4737@nyu.edu

Simran Kaur

Master's in CS Students, NYU

New York, USA

sk5340@nyu.edu

*Abstract*-The aim of this analytic project is to find the factor which can be best used to predict the occurrence of Rodents in a particular area in New York City (NYC). Using few of "311 NYC complaint" databases we have tried to find which factor (complaint) causes an increase in the number of rodent complaints in each area of NYC .We have also analyzed the frequency of rodent complaints made in the city with respect to the temperature. Furthermore, we have used the 311 rodent complaints database to try and get a rough estimate of the rodent population of the city.

Keywords—Analytics, Big data, Hadoop, New York, Rodent

# Introduction

Our aim of this project was to find out the main factors that are causing the infamous rodent problem in New York City (NYC). We have used NYC’s official 311 complaints database for this purpose. 311 is a special [telephone number](http://en.wikipedia.org/wiki/Telephone_number) supported in many communities in [Canada](http://en.wikipedia.org/wiki/Canada) and the [United States](http://en.wikipedia.org/wiki/United_States) which provides access to non-emergency [municipal](http://en.wikipedia.org/wiki/Municipality) services. The number 311 is intended in part to divert routine inquiries and non-urgent community concerns from the [911](http://en.wikipedia.org/wiki/9-1-1) number which is reserved for emergency service only. Examples of calls intended for 311 are: dead animal removal, debris in roadway, illegal burning, noise complaints, potholes, sinkholes and utility holes in streets etc. In our case, we used the 311 complaints database pertaining to rodent complaints and analyzed its correlation to database containing sanitation complaints and water leak complaints respectively. We checked to find for each zip code, if for every garbage or water leakage complaint there was a rodent complaint within the next week or not. That is, we wanted to find out which are the factors that lead to an increase in the number of rodent complaints.

Along with this, we also correlated the rodent complaints database with the weather database of NYC to check whether any relation can be found between the frequency of rodent complaints and various ranges of temperature.

Using the rodent complaints database, we also made a rough estimation of the total rodent population of the city. Rather than trying to get an exact number our aim was to disprove the often quoted statement ‘Eight million rats for eight million New Yorkers.’ This was done by tying the rodent complaints data with general information about rodent habitats and their life expectancy in the city.

# Motivation

Studies indicate that within the United States, the city is particularly well-suited for rats, taking into account such variables as (human) population patterns, public sanitation practices, climate, housing construction standards, etc. Rats primarily find food and shelter at human places and therefore interact with humans in various ways. More often than not, rats are found in corner stores in New York. In particular, the city's rats adapt to practices and habits among New Yorkers for disposing of food waste. Curbside overnight disposal from residences, stores, subway and restaurants, as well as littering, contribute to the sustenance of the city's rats. Rats have shown the ability to adapt to efforts to control them, and rodent infestations have increased as a result of budget reductions, more wasteful disposal of food, etc. Rats in New York have been known to overrun restaurants after hours, crawl up sewer pipes and enter apartments through toilets. They have also attacked homeless people, eaten cadavers in the city morgue, and bitten infants to get food off their faces. In 2003, a fire station in [Queens](http://en.wikipedia.org/wiki/Queens) was condemned and demolished after rats had taken over the building. Despite this the New York Department of Health and Mental Hygiene (DHMH) has taken very few preemptive actions to control this rodent problem.

Furthermore, when a number of complaints are received they are taken care of on a first come first serve basis. Rather, by analyzing which factor caused the rodent call, action can be taken on a priority basis.

A similar predictive approach has been used in Chicago. In partnership with the [Event and Pattern Detection Laboratory](http://www.cs.cmu.edu/~neill/EPD/index.html) (EPD Lab) at Carnegie Mellon University, Chicago’s Department of Innovation and Technology (DoIT) has managed to tackle the problem with a degree of success. It has been reported that resident requests for rodent control services have dropped by 15% in 2013. Hence, we felt that this approach could be replicated in NYC as well.

The New York City Department of Health and Mental Hygiene has never made an actual calculation of the rodent population, rather they identify the areas that are conducive for rodents to habitat. Hence, we realized that there was scope for us to get rough estimate of the rodent population

# Related Work

The combination of the geographic location information and the new 311 system has caused DSNY and its associated researchers Sarah Williams and Nick Klein [1] to review how they evaluate their complaint data and helps look at the spatial relationship between high complaint areas and socio-demographic factors. Like all 311 calls, complaints are logged, and inspected for service and review.

This study looks at the spatial patterns of DSNY missed pick-up complaints and analyzes if there is a relationship between spatially clustered complaints and neighborhoods with low density housing, people living in poverty, and minority status.

We found: 1] High Complaint Rates and Lower Density Housing areas with very high numbers of one and two family homes (almost anomalies) appear to have some correlation with high complaint rates.

2] Complaints and Neighborhood Demographics: Race & Income it appears that high poverty neighborhoods complain less than areas that with low numbers of minorities and higher income levels.

3] Complaint Hot Spots: Places to Explore Finally we implement “hot spot” analysis to help determine areas with statistically significant numbers of complaints

The research is similar to our analytics that involves deriving relationship among complaints and other factors that might lead to such complaints and also defining hot spots will help us recognize areas with higher probability of rodent complaints.

The goal of the research work done by Scott.L.Minkoff [2] is to develop an ultra-local measure of public goods quality and the number of 311 calls about goods quality in a defined space is reflective of neighborhood conditions where more calls is an indicator of poorer conditions.

The most important thing to note is the seasonality of complaints. Both the seasonality and the trend in city-wide complaints suggest the need to take modeling precautions when estimating the tract-level causes.

This is a spatial problem and the data are also spatial auto correlated—census tracts, not spatially independent from one another—such that nearby tracts have similar numbers of complaints.

This spatial problem is addressed three ways. First, the inclusion of community attributes and crime rates should ensure that estimates for council representation experience are not merely stand-ins for community attributes.

Second, a time invariant spatial lag is included in the model.

The spatial lag is the monthly average number of 311 calls for all surrounding tracts for all months in the time series.

This analytics provides us with methods and strategies to analyze the factors in each area based on 311 databases to derive a correlation among those factors and rodent complaints and enables us predict the rodent population in a defined space

The authors of the research paper Bruce Colvin, A.Daniel AShton,Wellard McCartney, William Jackson [3] aim at analyzing the rodent control program involving standard IPM techniques of surveys, public education, sanitation, rodent proofing, and baiting. These actions will be distributed among three zones adequately to match environmental conditions and proposed construction activities. The surveys will identify sanitation deficiencies (exposed garbage, unapproved refuse storage) and signs of rodent activity on premises. Furthermore, comparisons of survey data collected before and during construction will be used to evaluate any claim of community impact during construction.

Results of community and utility surveys; records and evaluations of any complaint calls from the public; and the locations, types, and amounts of bait placed will be recorded in the field on standardized data sheets and subsequently transferred to computer file. Statistical analysis of data sheets that detail the baiting activities, inspections for rodent activity and sanitation deficiencies, and any complaint calls investigated.

The paper relates to our analytics since it involves evaluating complaint calls in accordance with the location, types and amounts and taking into account correlation among factors that lead to rodent complaints.

According to Zha’s and Veloso‘s [4] general analysis of NYC’s 311 database there are large amounts of complaints received every day on an extensive range of issues. They studied the complaints on certain broad fields like time, location, type, agency of complaint etc. They tied this data with calendar and weather databases to find any trends based on which they made a prediction. They used a ‘Random Forest; method for training the first 3 years’ dataset and the 4th year as test dataset for this prediction. They found that complaints like heating and water system increase with even slight variation in weather. Similarly, they found that most complaints decrease during the weekends except for noise related complaints which they found increased. Several of the methods used by them for analyzing the 311 data have helped us in our analysis as well, as we too will be analyzing the rodent complaints with weather data.

The following research takes a look at online restaurant

Recommendation communities and reviews to study what endogenous and exogenous factors influence people’s participation in the communities. The authors Saeideh Bakhshi, Partha Kanuparthy, Eric Gilbert [5] analyze a corpus of 840K restaurants and their 1.1M associated reviews spread across every U.S. state. They construct models for number of reviews and ratings by community members, based on several dimensions of endogenous and exogenous factors. They find that while endogenous factors such as restaurant attributes (e.g., meal, price, service) affect recommendations, exogenous

factors such as demographics (e.g., neighborhood diversity, education) and weather (e.g., temperature, rain, snow, and season) also exert a significant effect on reviews.

Certain attributes of restaurant are taken into account such as Atmosphere ambience, monetary, service delivery and so on and all these factors are analyzed to find the ones that influence the most in calculating performance of a restaurant and the reviews.

A distribution of quantitative variables is then prepared that takes into account the following factors:

Restaurant Rating, Review, polarity, Demographics Weather

The next step is to model review and ratings

Effect of endogenous attributes as plotted such that the conditional analysis depicts impact of such attributes on review and ratings

The research done by Yu Zheng,Tong Liu1, Yilun Wang, Yanmin Zhu[6] presents a noise map of New York City, based on data sources: 311 complaint data, social media, road networks, and Point of Interests (POIs). The noise situation of any location in the city, consisting of a noise pollution indicator and a noise composition is derived.

The output of our model is a fine-grained noise pollution indicator which can diagnose citywide noise situation in various settings.

The noise indicator of a region is denoted by the number of 311 complaints. The noise map is overlaid on Bing Map which supports real-time scaling and moving. The darker a region’s color is the heavier noise pollution in it. It’s composed of the components

a) Location Ranking based on inferred noise indicators in various settings in a time slot and in a noise category.

b) Noise Composition Analysis. It is used to dig in the noise composition in a region and understand the main noise categories and their distribution changing over time. c) Revealing the distribution of noises over different noise categories in a location.

The Paper is useful in our research work since it involves study and analysis of the factors that contribute most towards noise pollution in the city of New York. We can learn how it works and apply to the factors for the rodent problem.

# Databases.

1) *311 Rodent Complaint Database*

Contains rodent complaints with details like timestamp of complaint, zip code, location type etc. for year 2010- Nov ’14.

Size: 38MB; Format: ‘.CSV’

2) *311 Sanitation Complaint Database*

Contains sanitation complaints having fields similar to rodent database for 2010-Nov’14.

Size: 41MB; Format: ‘.CSV’

3) *311 Water Leak Database*

Contains several water complaints like water leaking, standing water, hydrant overflow along with timestamp, zip code etc. for 2010-Nov’14.

Size: 30MB; Format: ‘.CSV’

4) *NCDC Weather Database*

The National Climate Data Center (NCDC) weather database for NYC contains fields like max, min temp, rainfall, wind speeds for each day for years 2012-Nov’2014.

Size: 1MB; Format: ‘.CSV’

# Design

Analytic 1: Sanitation, Water leak factors.

In this analytic, we have tried to find that whether for particular garbage or water leak complaint is there is an increase in the number of rodent complaints in the next seven days for a particular area in NYC. The method applied to both the 311 sanitation (garbage) complaints and water leakage complaints is the same. Figures 1, 2 explain the steps that are taken by us in the analytic. The 311 sanitation complaint and 311 rodent complaint databases are considered for calculating the effect of sanitation on rodent complaints.

Out first step involves cleaning up the database and to extract the relevant fields which we have used for our analytic. After extracting the date and zipcode fields for both these complaints we have used PIG to perform a join operation to get for each sanitation date all the rodent dates in a single row, along with the respective zipcode for that sanitation complaint in another column.

We then feed this huge matrix of data to a MapReduce (MR1). MR1calculates, for each sanitation date, the count of no. of rodent complaints, 1 week prior (negative) and 1 week (positive) after the sanitation date, along with zipcodes (area). This result has same zipcodes values scattered in different rows. So to club these same zipcodes into a single row and to get the Average of no of negative and positive complaints for a particular zipcode we use another MapReduce (MR2). The final output is that we have for a particular zipcode the average of number of rodent complaints that the zipcode (area) gets after a complaint for a sanitation problem has been made. This average is done on a data of 5 years, so we have very accurate measure of the average positive and negative values.

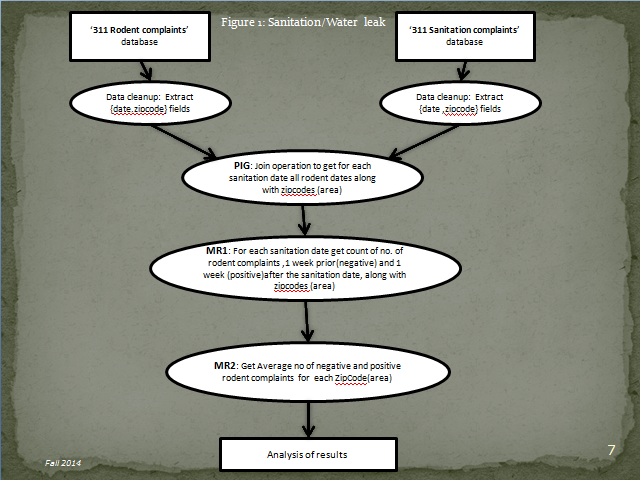


Figure 1: Sanitation/Water leak

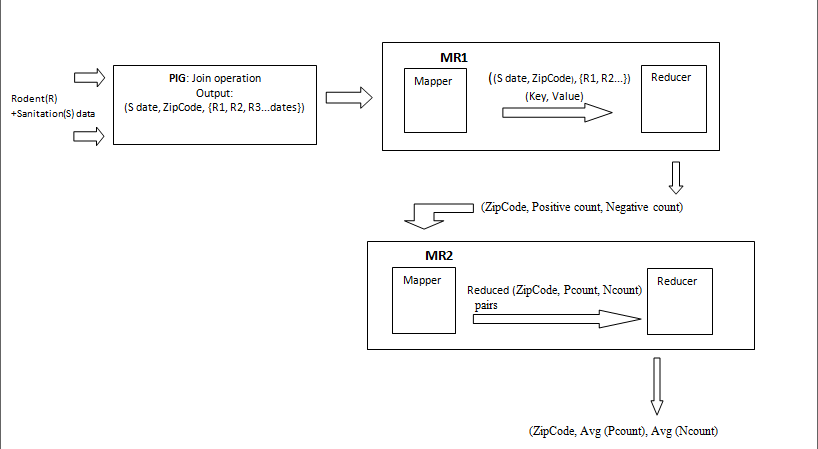


Figure 2: Input and Outputs in each Stage using Cloudera VMware.

We then calculate the difference in the positive and negative values and try to check for the places where the positive value is more. We then analyze our results.

The same method is applied to the water leakage and rodent databases. The results for the same are given below in the results section.

Analytic 2: Weather affecting Rodent Complaints.

Figure 3. explains the design and the steps of this analytic. In this we are trying to find the relation between the temperature and the rodent complaints. We are trying to figure out for which temperature range the rodent complaints is maximum i.e. to also say that for which temperature range the rodents population is the most. This is assuming that more the number of complaints more is the rodent population in that season.

In this analytic our first step is to normalize the databases. Formatting the date in both these data sets is also done. For that part we have used MapReduce for the rodent database. Then we have replaced the individual temperatures in the weather set by 5 ⁰C temperature intervals. So now for example a temperature 6⁰C lies in 5-10 temperature range.

Now the next step is to use PIG to join these datasets so that for each date in rodent complaints we have its respective temperature range value. Using MapReduce we then aggregate the rodent complaints based on the temperature ranges. Finally we get the number of rodent complaints for each temperature range. We have analyzed these findings in the result section.

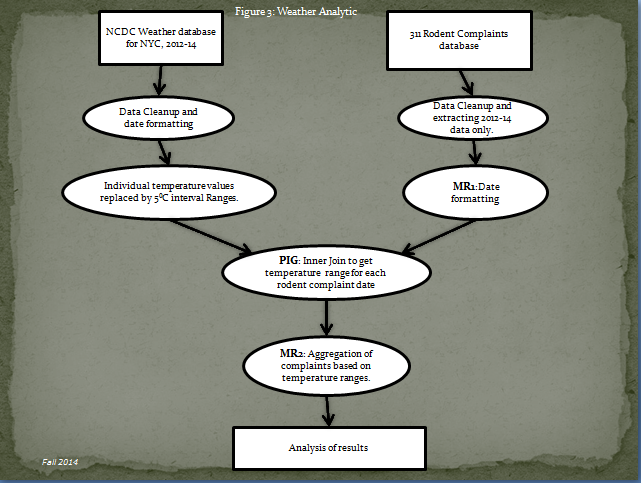


Figure 3: Weather Analytic

Analytic 3: Estimation of Rodent Population:

Figure 4: depicts the steps of this analytic. The diagram is simple and self explanatory. Scientific studies have shown that life expectancy of a rodent is 1 year in a city. Hence we have the found Average number of rodent complaints for 1 year. Taking the big overestimation-each rodent call represents each entire colony (on an avg. rodents live in a colony of 40-50) we have multiplied the result of the code with 50. Please check the results section for analysis.

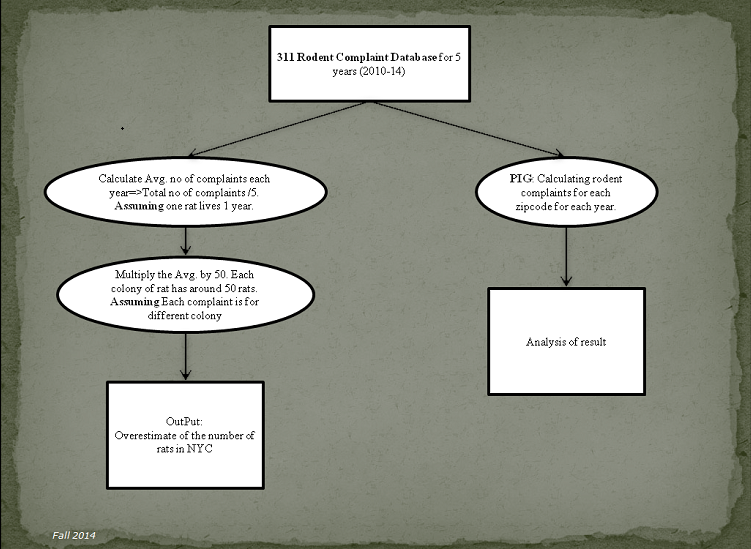


Figure4: Rodent Population Estimation

# Results

Analytic 1: Sanitation, Water leak factors Results.

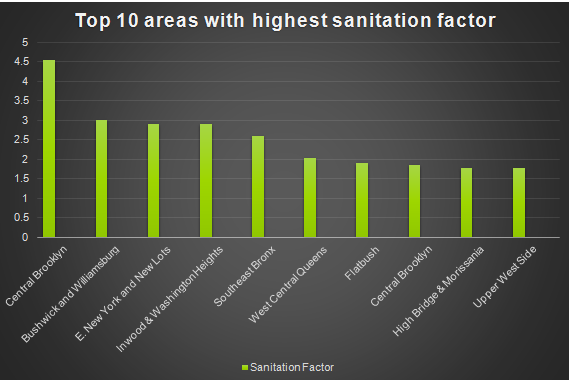


Figure 5. Sanitation

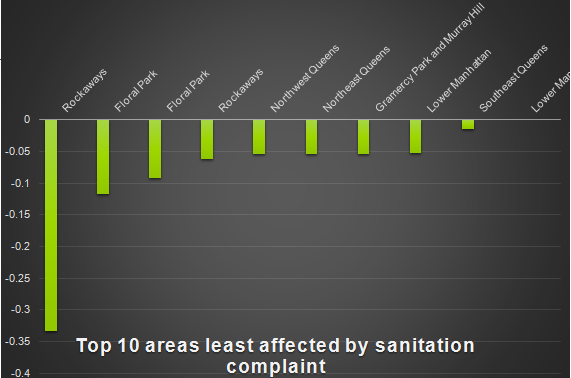


Figure 6. Sanitation

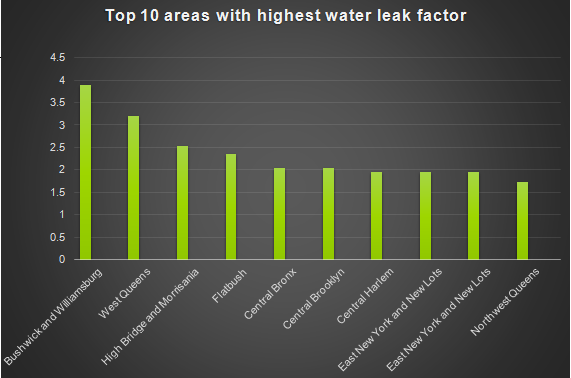


Figure 7. Water leak

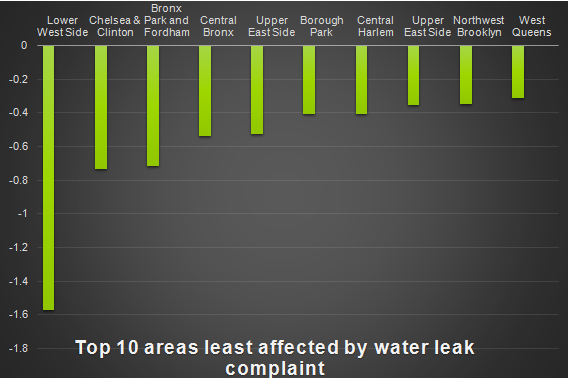


Figure 8. Water leak

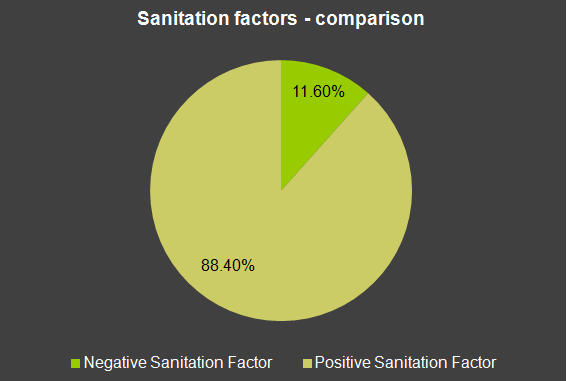


Figure 9 Sanitation factor.

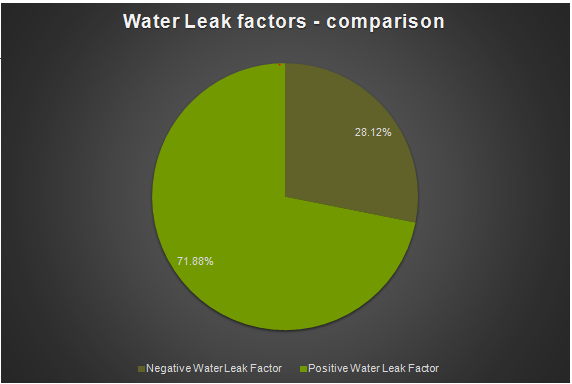


Figure 10. Water leak factor.

1) Figure 5, 7depicts the places where when a sanitation/water leak complaint is received there is an increase in the number of rodent complaints in the next seven days, meaning the rodent count increases. Hence at these sites sanitation/water leakage is a very big factor for rodents. Therefore, the health department should take preemptive rodent control measures as soon as they get a sanitation/water leakage complaint from any of these locations.

2) Figure 6, 8 depicts the places where the rodent complaints decrease after a sanitation/water leakage complaint is made. We can conclude that at these places sanitation/water leakage is not the main cause of rodent complaints. Hence sanitation is a low factor here.

3) What we can conclude is that when the health department gets a call for sanitation (garbage)/ water leakage they should check out how big the garbage/water factor for rodents for that area is. Based on that they can take preemptive actions and can also determine the degree (intensity) of their operation.

4) Also the health department can use our analytic to prioritize the rodent complaint calls. Suppose the department gets 10 rodent calls from 10 different places in the city. They can use our results to rank the places by predicting the number or rodents that they might find at the place. For the prediction they can check the last 7 days record for these places and see if they have received any other complaints like garbage or water leakage from these places. They can use our findings to see where there is a high chance of the rodent population to rise based on the complaint types. This helps them in prioritizing the calls and not just solving the cases on a first come first serve basis.

5) They should also take special care to keep the place clean for the top 10 areas most affected by the factors.

6) Figure 9, 10 depict that in almost all cases the number of rodent complaints a week after a sanitation/water leakage complaint is more than the rodent complaints a week before.

Analytic 2: Weather affecting rodent complaints Results

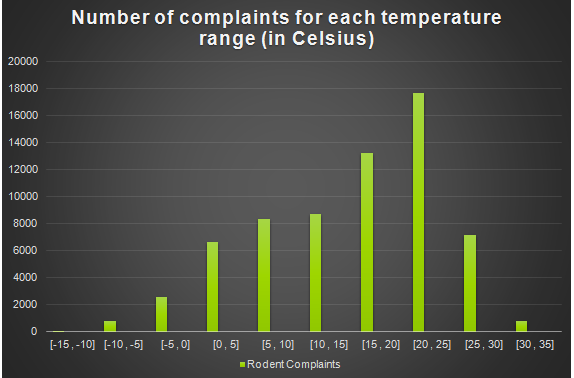


Figure 11. Temperature

1) WE can conclude from the results that as NYC experiences moderate temperature [15 – 25 C] the number of rodent complaints increase.

2) These results are analogous to the scientific fact that rodents prefer living in temperatures from 16-30⁰C. Hence our results validate this scientific fact.

3)Also we notice that when we move from summer to winter ((30-25) -> (10-5)) Rodent complaints increase. This is because rodents move indoors during the winter season. Hence Preemptive measures should be taken by the health department in all the lots and properties when fall ends and winter starts. This will help in reducing the rate population and also the diseases that are caused by this animal.

Analytic 3: Estimation of Rodent Population Results.

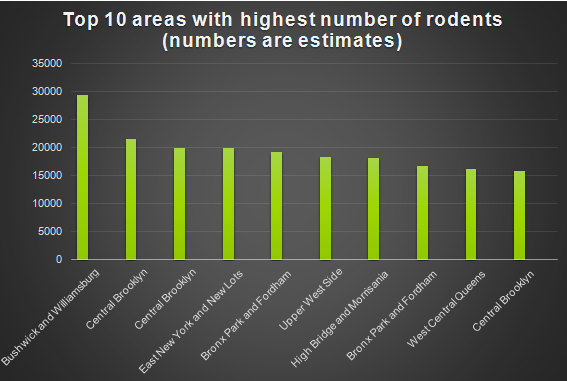


Figure 12. Areas with highest rodent count.

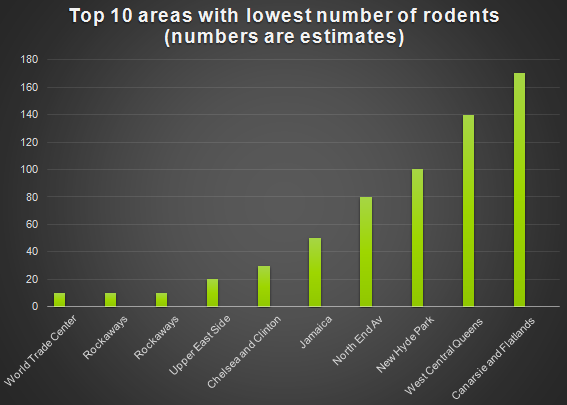


Figure 13: Areas with lowest rodent count

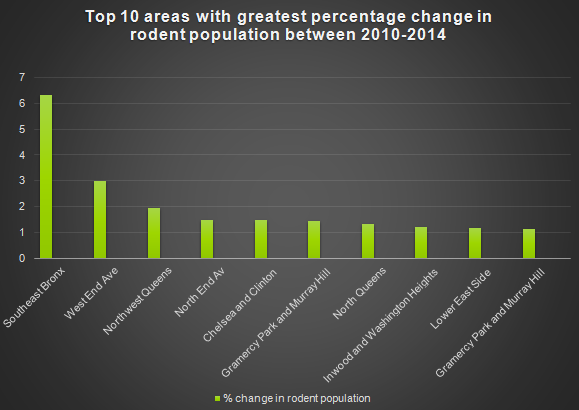


Figure 14: Areas with highest increase in rodent (2010-14)

1) The final value that we got after multiplying the results with 50(on an avg. rodents live in a colony of 40-50) was approx 1.2 million. Considering that we didn't take into account the sewer population and also the finding that the rodent sewer density is less than that is found on the surface above, we approximated the rodent population to 2Million, which is still less than 8million.

2) The fact that we overestimated and still fell way short the 8million mark, clearly suggests that the rodent population is nowhere close to 8 million. Hence we debunk the urban myth "8 million rats for 8 million people".

3) The bar charts are self explanatory. The most important result to focus on is that Brooklyn had the most number of rodents and Manhattan had the least if we club the areas borough wise.

# Future Work

In our first analytic we did not take into account the fact that a particular rodent complaint could be categorized in the ‘one week after’ window for a particular sanitation complaint and in the ‘one week before’ window for the next sanitation complaint date. This can be removed by considering only those sanitation complaints that are more than a week apart.

Further accurate analysis can be made on the data and our findings by using the aid of several statistical models. Chicago’s DOIT, in their analysis, used around 31 factors in predicting a rodent problem. Hence, in future analysis, the number of factors can be increased.

In our project we have used zip code as the default for location. We can further increase the granularity of the location by using GPS coordinates of the location of the complaints made.

# Conclusion

We have conclusively shown that the number of rodent complaints increases 88% of the times when a sanitation complaint is made and 78% of the time when a water leaking complaint is made.

Also, the number of complaints increases between 65F – 90F which matches with the ideal temperatures in which rodents thrive.

Lastly, despite making a large overestimate, we have concluded that the number of rodents in the city cannot exceed two million disproving the notion that there are eight million rodents in NYC.

##### Acknowledgment

##### We would like to thank the New York City’s Public Department’s initiative to put up all the 311 complaints databases online for free access. We would also like to thank National Climate Data Center (NCDC) for providing us the weather database for New York City for the past two years. Lastly, we would like to thank Professor Suzanne Macintosh for her guidance and support throughout the course of this project.

##### References

[1] New York Department of sanitation: Spatial Analysis of Complaints. Sarah Williams, Nick Klien

[2] Gotham 311: Equitability, Representation, and Public Goods Quality in New York City

Scott L. Minkoff

[3] Planning Rodent Control For Boston’s Central Artery/Tunnel Project. Bruce Colvin, A.Daniel AShton,Wellard McCartney, William Jackson

[4] Profiling and Prediction of Non-Emergency Calls in New York City. Yilong Zha, Manuela Veloso

[5] Demographics, Weather and Online Reviews: A Study of Restaurant Recommendations. Saeideh Bakhshi, Partha Kanuparthy, Eric Gilbert

[6] Diagnosing New York City’s Noises with Ubiquitous Data. Yu Zheng,Tong Liu1, Yilun Wang, Yanmin Zhu