Michael Fivis
May 21, 2013
GEOG 301 – Farmer

Who reports rats?

An exploration of the distribution of 311 complaints classified as "Rat Sighting" across ZIP codes of New York City.

Introduction

New York City has a notoriously entrenched rat infestation that has proven difficult to control, eradicate or even observe. Few credible estimates of the actual population exist, but the city's 311 feedback loop does collect and catalog all complaints pertaining to rat sightings (over 30,000 geocoded complaints in the three years between 1/1/2010 and 1/1/2013) and forward them to the Department of Sanitation, which oversees trapping and poisoning programming.

Many quality of life metrics like residential noise complaints are strongly associated with denser districts of nightlife business, which in turn is tightly connected with higher real estate values and, possibly, higher rates of complaining to 311 overall. Rat sightings could too have a similar relationship to wealth but it is not immediately evident in what way. It should seem like poorer neighborhoods with fewer resources for neighborhood cleanup should have worse problems. However, the ZIP codes that often top the rat-complaints list are the Upper West Side, Bushwick and the Lower East Side. These regions are disparate in their demographics, especially wealth but are similar in their high numbers of rat complaints.

Objectives and Goals

The goal of this project is to explore relationships between areas of rat complaints using available ZIP-level data and to build a minimal linear model to explain the variation.

Data and Study Area

The data to perform this analysis were all collected from NYC's Open Data portal and the American Fact Finder from the US Census. Census data was used solely for referential information such as the number of people residing in particular ZIP codes as well as their median incomes, as reported by 2011 American Community Survey estimates. The rest of the explanatory were sourced from NYC Open Data, including the 311 call-set which provided both the initial rat data as well as lots of other complaint data to furnish the large Department of Health dataset was used to determine the approximate number of registered restaurants and bars within each ZIP code. ZIPs that were eliminated and on what basis

Despite the prevalence of ZIP-level demographic and other metrics, there is a disparity in available information for New York City area ZIP codes. There are a handful of ZIP codes that lack all of the variables needed to make a predictive model. Some ZIP codes encompass a single large superstructure, such as the Empire State Building and Rockefeller Plaza, which famously have their own ZIP codes. While there are rat sightings in these ZIP codes there are no real residents to compare this information to.

The same applies to LaGuardia and JFK Airports, which have rat data inside of their ZIP codes but lack any residential demographics or metrics to examine. The airports were ultimately deleted from the map altogether. Altogether, 177 ZIP codes were kept which contained all of the necessary data. For the final, simplified counts for the incident data (restaurants or rats or complaints) a COUNTIF command was used in Excel to tease out the number of incidents or complaints or restaurants per ZIP. The rest of the measures were 2010 census data or 2011 ACS data, concerning population and demographics.

Methods

Because of the nearly infinite number of associated terms to search for with rat data, a small handful of demographic indicators are used. But even then the approach is exploratory. The *cor* command provided some interesting leads in providing a matrix of correlation rates between variables:

```
> cor(x = rat$rats, y = rat)
           zip rats     dohrb       pop        mhi noncomnoise
[1,] -0.135933    1 0.2066468 0.5932608 -0.3659837   0.2739447
      comnoise nypdnoise  FemaleHH PctBelowPo
[1,] 0.3753331  0.608326 0.6866062  0.5145494
     PctCollege       area blackpct   pctwhite   pctasian
[1,] -0.2005112 0.03382673 0.307539 -0.2421193 -0.4160742
       pcthisp
[1,] 0.3175325
```

The most interesting connections are related to race, gender (number female head-of-households provided a very strong correlation here) and other kinds of complaints (as shown with) *nypdnoise*. This information was used to guide trial and error attempts at building models using these terms in no certain order. Along the way, other terms were created, such as *borough*, a categorical term added to include another physical dimension to the data, but also to help explain connections to race and poverty.

The open-endedness of discovering relations between criteria for the rat data meant that *cor* would be used as a roadmap for piecing together attempts at models that explained the variation with significance. Linear models were constructed to explain the largest amount of data while retaining high levels of significance in each covariate used, checked against regression summaries and the *anova* function of R. Cross-comparisons across other terms were used to continually place new information and changes into context.

Results and Discussion

DOHRB

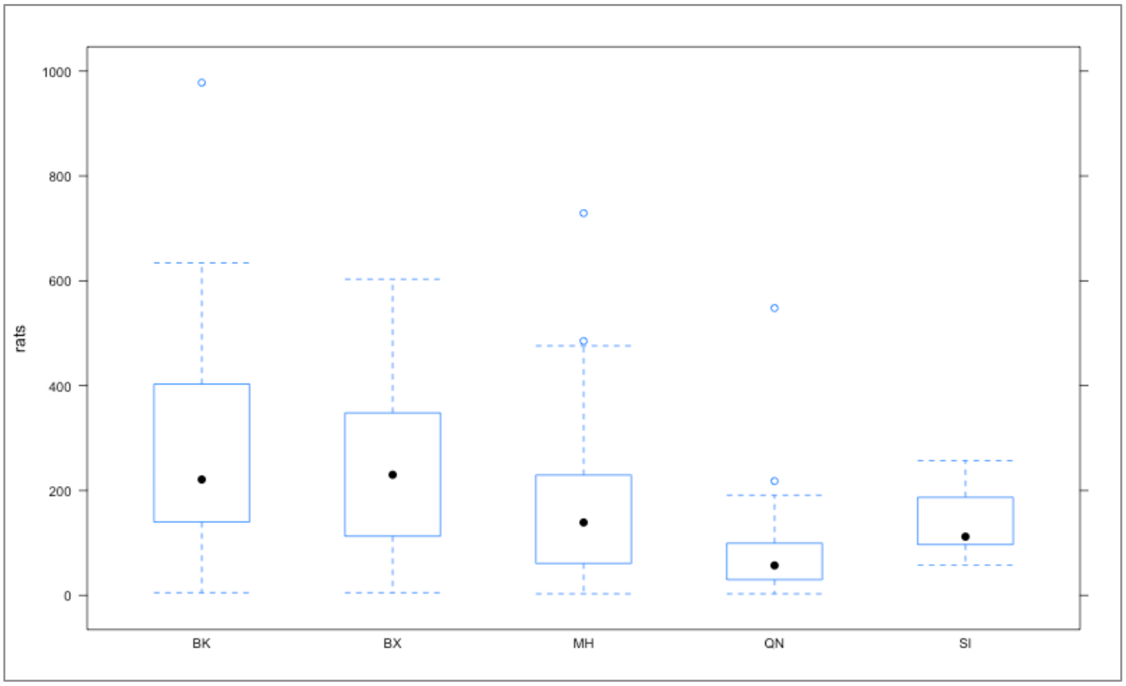
The term *dohrb*, Department of Health-registered Restaurants and Bars, which required the most massaging to integrate into the model turned out to be mediocre in predicting the number of rat complaints. While restaurants and bars do put out tremendous amount of attractive waste for rat populations on a nightly basis and usually on a local and dense scale (mountains of trash bags on minor sidewalks), and while they do predicate certain types of wealth and neighborhoods, the term did not explain too much of the variation ultimately. By itself, a model of rats as a function of *dohrb* produces an r-squared of less than .05.

However it is a significant additive to other models, holding on to its F-statistic strength and P-value significance no matter its position on a long chain of quantitative terms.

```
> mod9 = lm(rats ~ pop + noncomnoise + nypdnoise + dohrb, data = ratB)
> summary(mod9)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.8329487 18.6802891  -0.152   0.8796    
pop          0.0027211  0.0003754   7.248 1.37e-11 ***
noncomnoise  0.0621323  0.0304961   2.037   0.0431 *  
nypdnoise    0.1063635  0.0136345   7.801 5.68e-13 ***
dohrb       -0.5426743  0.1257160  -4.317 2.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 114.2 on 172 degrees of freedom
Multiple R-squared: 0.5545,      Adjusted R-squared: 0.5441
F-statistic: 53.51 on 4 and 172 DF,  p-value: < 2.2e-16
```

RACE AND BOROUGH



That blackpct as a covariate is able to explain anything about rat sightings has more to do with place than it does race. When a categorical term for borough was added to the table, most of the race-as-a-percent terms were difficult to use without some strong collinearity evident at one or more of the levels. For instance, adding borough to a model of *population* and *blackpct* resulted in not only the Bronx becoming insignificant as a factor, but also Staten Island (for its distinctly non-black population) and the base level of Bronx.

rats ~ blackpct + pop:

```
> summary(mod2)

Residuals:
    Min       1Q   Median       3Q      Max
-250.99  -78.63  -20.77   51.45  639.05
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.264e+01  2.161e+01  -1.510    0.133    
blackpct     1.545e+02  3.867e+01   3.994 9.58e-05 ***
pop          3.686e-03  3.856e-04   9.560  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 131.1 on 174 degrees of freedom
Multiple R-squared: 0.4064,                    Adjusted R-squared: 0.3996 
F-statistic: 59.56 on 2 and 174 DF,  p-value: < 2.2e-16
```

rats ~ blackpct + pop + borough:

```
> mod2 = lm(rats ~ blackpct + pop + borough, data = ratB)
> summary(mod2)

Residuals:
    Min      1Q  Median      3Q     Max 
-241.50  -67.60  -10.25   40.05  624.25 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.434e+00  3.674e+01  -0.148  0.88258    
blackpct     1.594e+02  3.810e+01   4.185 4.56e-05 ***
pop          3.496e-03  4.078e-04   8.573 5.97e-15 ***
boroughBX    2.938e+00  3.205e+01   0.092  0.92706    
boroughMH    3.995e+01  3.108e+01   1.285  0.20049    
boroughQN   -8.819e+01  2.853e+01  -3.091  0.00233 ** 
boroughSI   -8.500e+00  4.307e+01  -0.197  0.84379    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 122.1 on 170 degrees of freedom
Multiple R-squared: 0.4972,      Adjusted R-squared: 0.4795 
F-statistic: 28.02 on 6 and 170 DF,  p-value: < 2.2e-16
```

The overlap of borough and race is difficult to separate and it inflates the standard errors of the borough-levels when race-related terms are added. And just to confirm the borough-level differences in relation to *blackpct*, the mean values for each are accordingly very different.

```
> mean(ratB$blackpct ~ ratB$borough)
       BK        BX        MH        QN        SI 
0.3431454 0.3913858 0.1576212 0.2297070 0.1225464 
```

MEDIAN INCOME / POVERTY

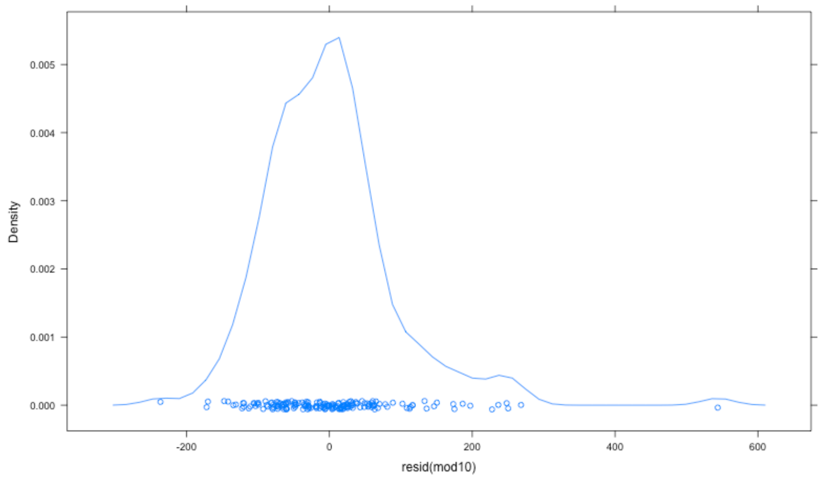
Despite the fairly strong report of negative correlation between the rats and income, many models incorporating *mhi* insignificant p-values. Of all the factors, median household income played a weak role in explaining ZIP to ZIP variation in the number of rats. (r-squared = .13). While some models it was able to add some explanatory value to, it was unable to contribute any value to the strongest model in the exploration (which incorporates *FemaleHH, nypdnoise and dohrb).* Income disparity was the inspiration for looking at this data to begin with – because some of the supposedly rattiest ZIPs were places of quite different levels of wealth – and it is actually apparent from statistical analysis that income had low relevance to the number of rats that might be found within a ZIP code. In models where it was able to add value, the P-values often toed the line of significance.

```
> summary(mod13)

Call:
lm(formula = rats ~ FemaleHH + nypdnoise + mhi, data = rat)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.605e+01  2.993e+01  -2.207   0.0286 *  
FemaleHH     2.189e-02  2.102e-03  10.417  < 2e-16 ***
nypdnoise    8.307e-02  9.138e-03   9.091 2.25e-16 ***
mhi          5.611e-04  3.057e-04   1.835   0.0682 .  
---
Residual standard error: 101.7 on 173 degrees of freedom
Multiple R-squared: 0.6447,      Adjusted R-squared: 0.6386 
F-statistic: 104.6 on 3 and 173 DF,  p-value: < 2.2e-16
```

The *PctBelowPo* term from American Community Survey data offered a clearer picture with stronger evidence than income, with higher significance, due to the nature of the data. As a rate which makes visible the proportion of people living below a threshold, it was perhaps less skewed by wealthier subsections of a ZIP code. However, it held a strong collinearity with *FemaleHH* that barred its use in *mhi*'s place, even though *PctBelowPo* often enhanced other models that made significant use of *mhi.*

FEMALE HEAD OF HOUSEHOLDS

The strongest/simplest model constructed in this study came from just four quantitative terms: *mod10 = lm(rats ~ FemaleHH + nypdnoise + noncomnoise + dohrb)*



```
> summary(mod10)

Call:
lm(formula = rats ~ FemaleHH + nypdnoise + noncomnoise + dohrb,
    data = rat)

Residuals:
    Min      1Q  Median      3Q     Max
-236.64  -60.58   -5.88   32.66  543.83

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.673562  14.683223  -2.021   0.0448 *
FemaleHH      0.022541   0.001798  12.533  < 2e-16 ***
nypdnoise     0.067868   0.012055   5.630 7.22e-08 ***
noncomnoise   0.149366   0.026636   5.608 8.03e-08 ***
dohrb        -0.490127   0.103447  -4.738 4.50e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.36 on 172 degrees of freedom
Multiple R-squared: 0.696, Adjusted R-squared: 0.6889
F-statistic: 98.45 on 4 and 172 DF,  p-value: < 2.2e-16


> anova(mod10)
Analysis of Variance Table

Response: rats
             Df  Sum Sq Mean Sq F value    Pr(>F)
FemaleHH      1 2374859 2374859 266.740 < 2.2e-16 ***
nypdnoise     1  838122  838122  94.136 < 2.2e-16 ***
noncomnoise   1   93374   93374  10.488  0.001442 **
dohrb         1  199863  199863  22.448 4.501e-06 ***
Residuals   172 1531366    8903
```

These four covariates all provide good proxy explanations for other things as well. *FemaleHH*, which stands particularly strong on its own as a model (r-squared = .46) has built into it a little racial difference and income difference, as well as a predisposition built into the gender of calling the city to complain about rats. *FemaleHH* demonstrates that it's strongly correlated with median household income and *PctBelowPo* (poverty line rates), while *nypdnoise* and *noncomnoise* have built into them not only perhaps an indirect measure of how populous a zip is but also of nightlife activity and the built environment.

Staten Island - whose ZIPs were consistently underestimated by all models is - in this model too, underestimated.

```
> ratB$resid10 <- resid(mod10)
> mean(ratB$resid10 ~ ratB$borough)
        BK         BX         MH         QN         SI 
 23.019775 -14.062281 -17.188729  -8.915906  65.180655 
```

The borough's lower densities and low population mean that one of the key components of the models, *nypdnoise*, vastly misses the mark. Residents of Staten Island have fewer complaints concerning other people and the noise they make than the rest of the city.

```
> mean(ratB$nypdnoise ~ ratB$borough)
       BK        BX        MH        QN        SI 
1272.3784  806.9200 1490.3182  423.9661  328.5833 
```

In this model, adding *noncomnoise* (or DEP noise complaints) furnished a small gain in r-squared while remaining statistically significant (p = 0.00261). The model stands to account for quite a lot of the variation even without it (> 0.60).

Comparing FemaleHH against other types of complaints demonstrates a unique affinity to rat sightings over [noise] complaints in general.

```
> cor(ratB$FemaleHH, ratB$nypdnoise)

[1] 0.3239613
```

```
> cor(ratB$FemaleHH, ratB$rats)

[1] 0.6866062
```

Modeling FemaleHH against noise complaints returns very weak explanations with significance.

```
> mod4 = lm(noncomnoise ~ FemaleHH, data = ratB)
> summary(mod4)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 611.594312  62.911160   9.722   <2e-16 ***
FemaleHH     -0.016523   0.008219  -2.010   0.0459 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 512.5 on 175 degrees of freedom
Multiple R-squared: 0.02257,     Adjusted R-squared: 0.01699

> mod5 = lm(nypdnoise ~ FemaleHH, data = ratB)
> summary(mod5)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 543.51159  103.45373   5.254 4.30e-07 ***
FemaleHH      0.06123    0.01352   4.530 1.09e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 842.8 on 175 degrees of freedom
Multiple R-squared: 0.105, Adjusted R-squared: 0.09984
```

And for comparison:

```
> mod6 = lm(rats ~ FemaleHH, data = ratB)
> summary(mod6)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.667845  15.142080   1.827   0.0694 .
FemaleHH     0.024715   0.001978  12.493   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 123.4 on 175 degrees of freedom
Multiple R-squared: 0.4714,      Adjusted R-squared: 0.4684
F-statistic: 156.1 on 1 and 175 DF,  p-value: < 2.2e-16
```

Conclusion and Suggestions for Future Work:

Incorporating a term that integrates information about the structural typology might be important. In many models, residuals largely underestimated the number of rats on Staten Island but its built environment is vastly different than the other four boroughs in density and housing stock. As well, the number of restaurants and its race makeup are also unlike the rest of the boroughs.

There are an infinite number of environmental factors at play with this kind of data and what would perhaps bind together the results tighter than any measure might be to begin mapping and modeling some of the data at their raw distances. The data is available to not only look at the *exact address* of a rat location but also to compute the distance between those and nearest restaurants and bars, which also have their exact addresses published. Working to analyze the data at a resolution finer than that of ZIP code would be the best way to enhance this information, and would likely introduce new discoveries altogether. Some ZIP codes may be just too large geographically for these comparisons to be relevant, especially in regard to income, when class differences can manifest themselves on a block-by-block basis.

The strength of the Female Head of Households term was a surprising but a rational answer, however more work could be done to understand the split of gender in 311 usage, as well as female predisposition in many other types of complaints using factors other than *FemaleHH*.

Table Codebook:

zip – zip codes of NYC, less the zipcodes that did not have all other requisite terms (including zipcodes that encompass only office buildings or airports).

dohrb – the number of Department of Health-registered restaurants and bars in a particular zip, deduced from the DOHMH's rating dataset.

mhi – median household income, as reported by 2011 ACS data at the zip level

noncomnoise – number of complaints forwarded from 311 to the DEP concerning noise. These complaints usually concern inanimate and environmental conditions (construction site noise, transportation noise, loud air conditioners, etc)

comnoise – a subset of **nypdnoise** specifically about commercial business noise (such as loud music coming from restaurants, or loitering patrons outside of a dance club)

nypdnoise – encompasses most noise complaints concerning other *people*-made noise. Loud music coming from parked cars, noisy neighbors, yelling, loitering, parties, dog barking are all part of these types of complaints.

FemaleHH – number of female-run households as determined by 2010 Census information

PctBelowPo – ACS ZIP level data concerning rates of people within a ZIP living below the federal poverty guidelines

Area – approximate area of ZIP in feet.

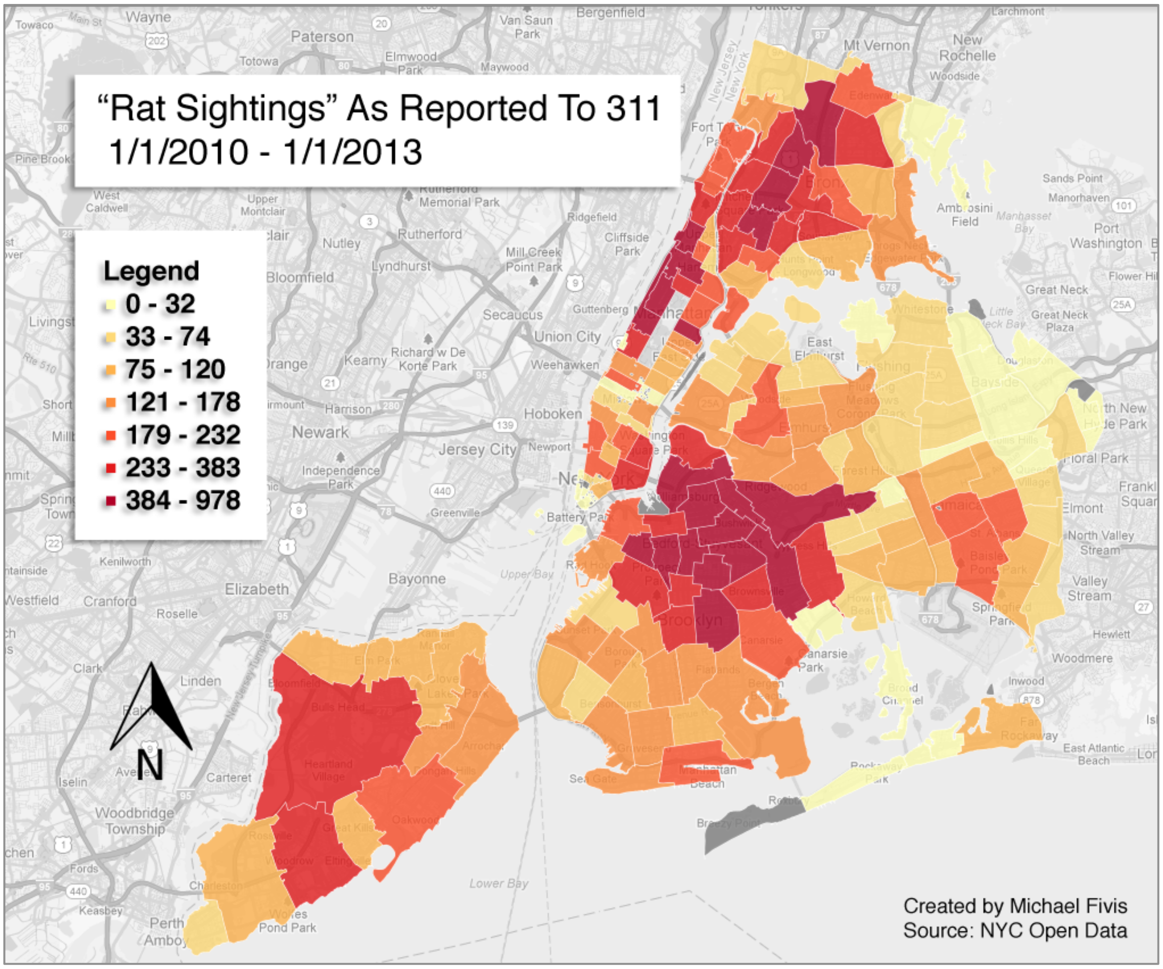
Note: Thank you for your help in calculating this but I never got it to return anything interesting. I think controlling for physical size may not work very well at the size of ZIPs or… something I didn't find is throwing it off. Perhaps some ZIP codes are so overbounded that it doesn't produce any rational measure of density.

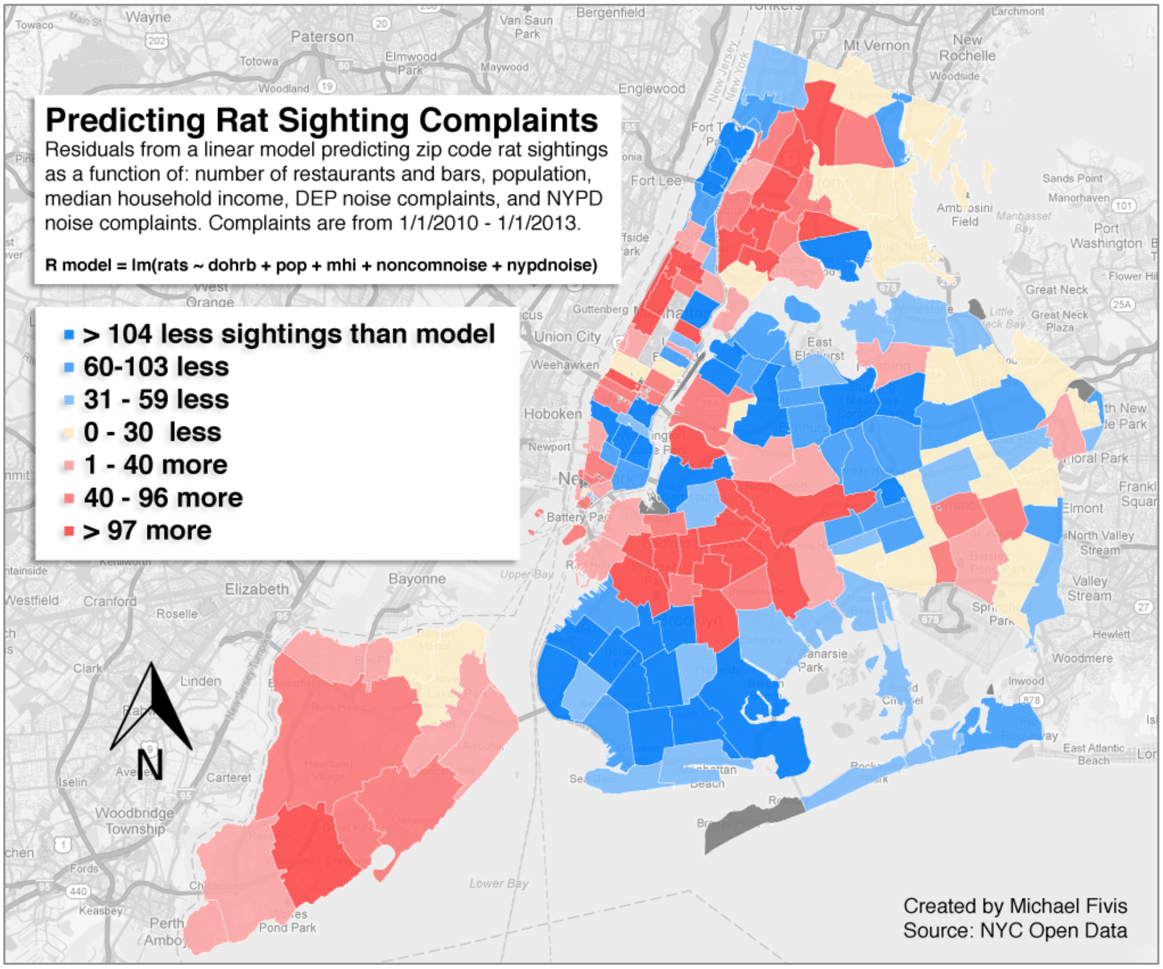
GIS Addendum

Submitted for interest only

For a GTECH 201 project, I submitted a simple visualization from early models tested for this project. By plotting out a simple model's residuals, I was able to get a fresh perspective on which ZIP codes threw an average model.

The hotspots are roughly similar – the model suggests that areas of Bushwick, the Upper West Side, and the Bronx *should* have far fewer rats given the values of the model terms. Some misses include the Lower East Side and sections of Williamsburg which are perhaps thrown by the sheer number of bars and restaurants (as well as the outsize number of associated noise issues) located within those zip codes. The aforementioned disparity between Staten Island and the rest of the city in the observed explanatory terms is also evident in its near uniformity of red residual "miss".





Predicting Rat Sighting Complaints

Residuals from a linear model predicting zip code rat sightings as a function of: number of restaurants and bars, population, median household income, DEP noise complaints, and NYPD noise complaints. Complaints are from 1/1/2010 - 1/1/2013.

R model = lm(rats ~ dohrb + pop + mhi + noncomnoise + nypdnoise)

Created by Michael Fivis
Source: NYC Open Data

Data Sources

- *United States Census American Fact Finder*
- *NYC Open Data (DOHMH dataset, 311 Complaints)*