Title

Ben Horvath

December 9, 2018

Contents

Load libraries:

```
# library(corrplot)
library(dplyr)
library(ggplot2)
# library(lme4)

# source('../R/multiplot.R')
```

1 Introduction

The purpose of this document is to explore the relationship between 311 calls and housing violations in New York City. 311 is a phone number in New York City where citizens make civil, non-emergency calls, e.g., reports of graffiti or noise complaints. The expectation is that geographic areas with higher 311 calls should produce more housing violations – or vice versa. Either outcome is *a priori* plausible.

The end goal is to build a model to predict housing violations. If successful, potential applications include allowing the city government to plan and target housing inspections more efficiently.

The first section of this paper describes the data and how it is assembled. Logs of 311 calls and housing violations from 2014 are pulled from NYC's open data web site, https://opendata.cityofnewyork.us/. Additionally, I

use supplementary demographic data pulled from the U.S. Census Bureau (via a very nice R package called `tidycensus`).

NOTE: As it turns out, there's not a great relationship between 311 and housing violations. I still attempt to construct various models; my hope is that detailed data exploration and model diagnostics will make up for it.

2 Data Collection

2.1 311 Calls

The 311 call logs are updated every day on the NYC Open Data web site. The online records go back to 2010, and each year contains many calls. For the purposes of this project, I will limit my exploration to 2014.

I used this `curl` call to download the dataset directly from the web site, and then used the `grep` command line tool to filter to 2014:

```
curl https://data.cityofnewyork.us/resource/fhrw-4uyv.csv?%24limit=5000&%24%24app_toke
```

Note that this command requires a registration to get a token, and that I had to fiddle with various query parameters to get exaclty what I wanted.

Load the data: