Homework 1: Diagnostic

Ye Ji

April 4, 2017

**Problem 1: Data Acquisition and Analysis**

* 1. ***Process***

After I acquired the city of Chicago 311 requests data, I used Python to clean, study and graph the data of graffiti removal, vacant and abandoned buildings, potholes, and sanitation code complaints. All the numbers and visuals in the following paragraphs are generated in Python. My work can be found with the GitHub repository link I submitted.

* 1. ***Summary Statistics***

In the year of 2016, The total number of 311 requests is 178831. That is 490 requests per day and 14903 per month.

- The majority of the requests are for graffiti removal, taking up **64.9%** of the total;

**21.2%** is potholes;

**11.3%** is sanitation code complaints;

**0.02%**. is vacant/abandoned building report.

- On average, city of Chicago responds to request in:

**0.6** days for graffiti removal request;

**9.4** days for sanitation code complaints;

**21.6** days for potholes.

For vacant/abandoned building reports, Chicago simply records them.

- By month, **August** has the most pf graffiti removal requests.

**May** has the most vacant/abandoned building reports.

**January** has the most pothole reports.

**June** has the most sanitation code complaints.

- By neighborhood top-five neighborhoods:

**Most graffiti** request (number of requests, neighborhood name):

(4885, 'LOWER WEST SIDE'),

(5339, 'AVONDALE'),

(5984, 'BRIGHTON PARK'),

(7881, 'LOGAN SQUARE'),

**(9149, 'WEST TOWN')**

Least graffiti request:

(12, 'OAKLAND'),

(12, 'RIVERDALE'),

(31, 'PULLMAN'),

(38, 'AVALON PARK'),

(41, 'BURNSIDE'),

**Most sanitation** request:

(617, 'AUBURN GRESHAM'),

(716, 'LOGAN SQUARE'),

(747, 'WEST TOWN'),

(748, 'AUSTIN'),

**(769, 'LINCOLN PARK')**

Least sanitation request:

(7, 'RIVERDALE'),

(17, 'OHARE'),

(20, 'EDISON PARK'),

(24, 'OAKLAND'),

(32, 'MONTCLARE'),

**Most pothole** request:

(1006, 'NORWOOD PARK'),

(1015, 'AUSTIN'),

(1099, 'NEAR WEST SIDE'),

(1114, 'NEAR NORTH SIDE'),

**(1194, 'WEST TOWN')],**

Least pothole request:

(63, 'OAKLAND'),

(75, 'FULLER PARK'),

(80, 'BURNSIDE'),

(102, 'RIVERDALE'),

(125, 'WEST GARFIELD PARK'),

**Most vacant building** request:

(249, 'NEW CITY'),

(280, 'AUBURN GRESHAM'),

(297, 'ROSELAND'),

(316, 'ENGLEWOOD'),

**(453, 'WEST ENGLEWOOD')**

Least vacant building request:

(0, 'NEAR SOUTH SIDE'),

(0, 'OHARE'),

(1, 'EDISON PARK'),

(1, 'LOOP'),

(1, 'ROGERS PARK'),

* 1. ***Five Interesting Findings***

**Finding #1 #2 #3:** Subtype information are fun.

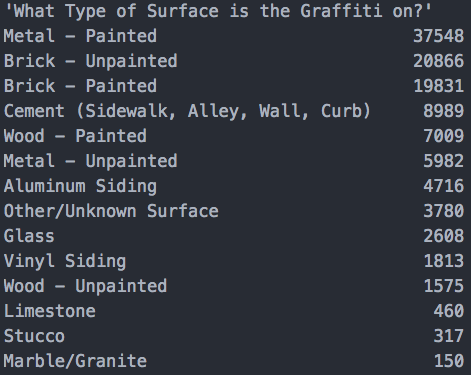
With a simple count, subtype information in these

datasets reveal some fun facts about the dataset. For

example, you can learn a list of common architecture

façade materials from the cheapest to the most

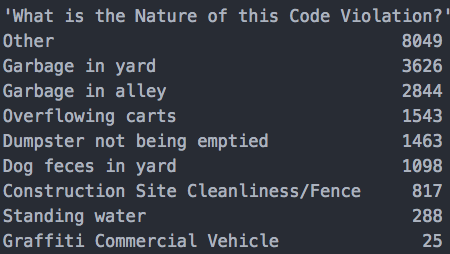
expensive from this table:



Subtype table 1 - Graffiti

- For Sanitation code complaints, it seems the nature

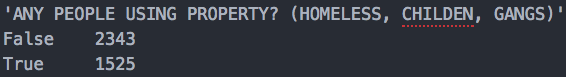
of code violation is usually hard to describe.



Subtype table 2 – Sanitation Code Complaint

- Good amount of the vacant/abandoned buildings are

de-facto occupied.



Subtype table 3 – Vacant/Abandoned Building

**Finding #4: Correlations**

Between the requests, graffiti is positively related to sanitation,

negatively related to vacant building, and has no significant relationship

with pothole. Pothole has a positive correlation with sanitation. And

sanitation has significant correlations with all other requests.

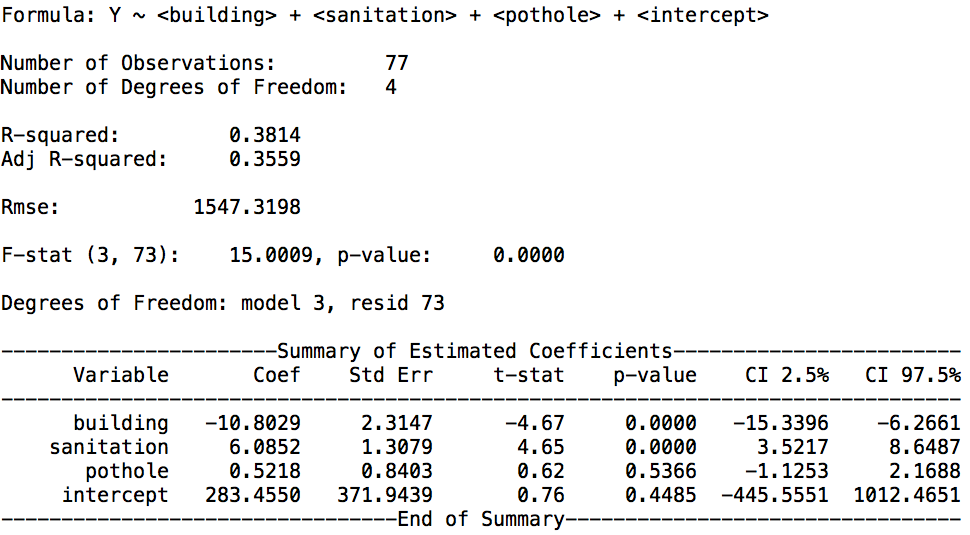


fig. Sample regression table – as equivalent to "reg building sanitation pothole" in Stata.

**Finding #5: Implications**

If we are choosing neighborhood to live in, without more detailed information,

the cleanness of a neighborhood could be used as a key indicator to eyeball.

If a neighborhood is not clean, chances are it also has other issues.

**Problem 2: Data Augmentation and APIs**

For the American Community Survey data, I collected "mean income", "median age", "average household size" characteristics variables on block-group level. Block group is a census-defined geographic unit that is much smaller than a neighborhood. Chicago has 77 community areas defined as neighborhood, and the city has 800 census blocks. Out of personal interest, I also added 'population' and 'Asian population' for each Census block, just to give me an idea for "how minor in population percentages in Chicago?".

***2.1 What types of blocks get "Vacant and Abandoned Buildings Reported"?***

Low mean income about $980, average household size about 3.

***2.2 What types of blocks get "Sanitation Code Complaints"?***

Average mean income about $3750, average household size about 2.8.

***2.3 Does that change over time in the data you collected?***

Result unavailable.

***2.4 What is the difference in blocks that get "Vacant and Abandoned Buildings Reported" vs "Sanitation Code Complaints"?***

Compared to blocks that gets 'Sanitation code', the blocks that get 'vacant/abandoned buildings' has much lower mean income, slightly higher household size, fewer Asians, and is younger.

**Problem 3: 311 Call Center Scenario Questions**

(Below are the answers to some problem set style questions.)

***3.A Of the four types of requests you have data for, which request type is the most likely given the call came from 7500 S Wolcott Ave? What are the probabilities for each type of request?***

This address is located in the Auburn Gresham neighborhood (community are code 71). Using the data from problem 1, This neighborhood has a total request of 2047, out of which:

**13.67%** is vacant building report;

**11.14%** is graffiti removal;

**45.04%** is pothole;

**30.14%** is sanitation code complaint.

That would be the probabilities distribution for which type of 311 request the call could be.

***3.B Let's now assume that a call comes in about Graffiti Removal. Which is more likely – that the call came from Lawndale or Uptown? How much more or less likely is it to be from Lawndale versus Uptown?***

Based on the 2016 data, graffiti requests from North Lawndale and South Lawndale combined is 5105, where graffiti requests from Uptown is 1788. On the condition that call is about graffiti, it is more likely to come from Lawndale than Uptown. The probabilities are:

**4.39%** that it comes from Lawndale;

**1.54%** that it comes from Uptown.

Dividing Lawndale probability by Uptown probability, it is 2.8 times more likely that the call comes from Lawndale.

***3.C Now assume that you don’t have access to all the raw data and you know the following things: There are a total of 1000 calls, 600 from Englewood and 400 from Uptown. Of the 600 calls from Englewood, 100 of them are about Graffiti Removal. Of the 400 calls from Uptown, 160 are about Graffiti Removal. If a call comes about Graffiti Removal, how much more/less likely is it that the call came from Englewood versus Uptown?***

The odds between Englewood vs. Uptown is roughly 38: 62. It is 23% more likely that the call is from Uptown.

Using Bayes' theorem:

P ( Englewood | Graffiti ) = [P (Graffiti) \* P (Englewood) ] / P (Graffiti)

= [(100 / 600) \* (600 / 1000)] / [(100 + 160) / 1000]

= 0.3846

P ( Uptown | Graffiti ) = [P (Graffiti) \* P (Uptown) ] / P (Graffiti)

= [(160 / 400) \* (400 / 1000)] / [(100 + 160) / 1000]

= 0.6154

P ( Uptown | Graffiti ) - P ( Englewood | Graffiti ) = 0.2308