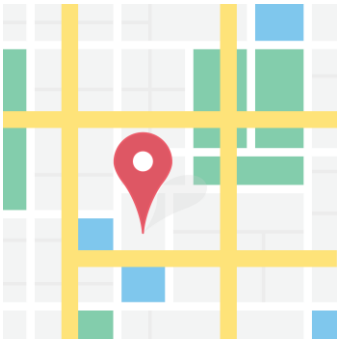
UID: 204256656
CS 170A
Winter 2016

Los Angeles Urban Crime Patterns



Kang (Frank) Chen

Table of Contents

I. Dataset Description

Criminology has always been an essential area of behavioral and social research, especially since crime is, and has always been, omnipresent since the earliest formation of society. Many of the justice laws passed by our government today relies on our understanding of criminal behavior, and the course of actions we should take to decrease crime frequency.

Unfortunately, despite new laws being passed year after year, crime continues to be a part of our everyday lives; the toll it has taken on us as citizens have transformed our sociological behavior. We have a growing distrust towards certain groups of individuals based on news/social media biases; as a result of this prejudice, the gap between citizens and law enforcements widens, sometimes turning into violent protests and movements. In addition, crime is given high screen time in movies and television; it is inserted in almost every plot sequence, and we, the jaded citizens, are constantly reminded that crime simply 'happens'.

How do we prevent crime? Do we teach criminals a lesson by incarcerating them (The United States has the most incarceration rate out of all counties)? Or should we focus on building better neighborhoods and strengthening the relationship between citizens and the police? One method I'm particularly in is using the Broken Windows Theory[1], which states that maintaining and monitoring urban environments to prevent small crimes such as vandalism ultimately creates an atmosphere of order, therefore preventing more serious crimes from occurring. How are criminals and their environments connected? How does criminal activity spread in a neighborhood, and how do we control and decrease criminal behavior?

To explore these questions and more, I found a group of data sets called "LA Historical Crime Data" from LA County Sheriff's Department's jurisdiction[2]. These series of huge datasets (0.5 GB total) contains crime entries dating from 2005 – 2013, with more than 250,000 crime entries each year. Furthermore, I used another dataset[3] showing all the building and safety inspections done in the greater LA area for an analysis I will go into detail later in my report.

Tuple #	Incident Date	Crime Category	...	Address	...
1	7/12/2013	NARCOTICS	...	500 W ARROW HWY, SAN DIMAS, CA	...
2	7/12/2013	LARCENY THEFT	...	HOLLYWOOD BLVD & WESTERN AVE. HOLLYWOOD, CA	...
3	8/6/2013	VANDALISM	...	100 W HARRIET ST, ALTADENA, CA	...
...	...	...	...	...	...

Figure 1: Sample Table showing some of the attributes of the data; we are primarily concerned with Incident Date, Crime Category, and Address; the '...' entries in the cells are symbolic for other data as part of the dataset that we are not concerned with, such as Incident Reporting Number etc.

II. Hypothesis & Analysis Plan

My primary goal in this project is to answer the following question:

"Does the Broken Window Theory Exist in the Greater LA Area?"

In other words, I want to use my obtained dataset to show neighborhoods that are consistently being monitored for small issues, such as building infrastructure and safety, have less crime frequencies over the past couple of years.

My approach is to begin by first understanding the crime dataset by understanding its distribution, spread, and correlation. Using histogram, PCA, and correlation analysis, I should be able to have a better understanding. In addition, I plan to visualize the dataset by finding the longitude and latitude given the address from my parsed attribute, then plotting the result on a map of the Greater Los Angeles. From here, I can see the geographical locality of the crimes, thereby understanding its relationship with the LA neighborhood and environment.

Furthermore, I will use the dataset obtained from *data.gov* on the different building safety inspection logs to build a map visualizations of all the places in LA where the building infrastructure are kept in check by the government. This is an example of maintaining a good environment, thereby preventing the phenomenon described by the Broken Window Theory. My hypothesis is that the locations where there are high concentrations of building safety inspections will have low concentrations of crime in general.

III. Preparing the Dataset for Analysis

I am primarily using MATLAB to analyze my dataset in this project, and one of the first issues I came across was reading in the dataset for MATLAB. All of my datasets are in `.csv` format, so a simple `import` function on MATLAB can load the files into my workspace. However, I realized that the size of the dataset plays an important role in the MATLAB runtime. Loading all 593 MB of data into the workspace took a very long time, almost crashing my MATLAB program; furthermore, this is without doing calling any computation or analysis on the dataset, which will surely take even longer. Foreseeing this issue ahead of time, I wrote a script to extract only the attributes I need to do my analysis from each year. This significantly decreased my loading time. I loaded only the attributes `ADDRESS`, `CATEGORY`, and `INCIDENT_DATE` into my MATLAB workspace.

In addition, I had to sample my dataset as well, since loading in all 2005 – 2013 attributes resulted in around 3.5 million tuples per attribute. When any analysis is performed on these variables, MATLAB crashes or takes a significantly long time (> 400 seconds, or around 7 minutes). To avoid that, I sampled every 100 tuple in my dataset, and have a resulting 35,000 tuples for each attribute.

```
% load first dataset
ADDRESS       = load('C:\Frank_Chen\UCLA\Winter 2016\CS
170A\project_proposal\data\ADDRESS_2013.mat');
CATEGORY      = load('C:\Frank_Chen\UCLA\Winter 2016\CS
170A\project_proposal\data\CATEGORY_2013.mat');
INCIDENT_DATE = load('C:\Frank_Chen\UCLA\Winter 2016\CS
170A\project_proposal\data\INCIDENT_DATE_2013.mat');

% sample every 100th tuple
ADDRESS = ADDRESS.ADDRESS(1:100:length(ADDRESS.ADDRESS));
CATEGORY = CATEGORY.CATEGORY(1:100:length(CATEGORY.CATEGORY));
INCIDENT_DATE = INCIDENT_DATE.INCIDENT_DATE(1:100:length(INCIDENT_DATE.INCIDENT_DATE));

% load second dataset
tmp1 = load('C:\Frank_Chen\UCLA\Winter 2016\CS 170A\project_proposal\data\ADDRESS_2012.mat');
tmp2 = load('C:\Frank_Chen\UCLA\Winter 2016\CS 170A\project_proposal\data\CATEGORY_2012.mat');
tmp3 = load('C:\Frank_Chen\UCLA\Winter 2016\CS
170A\project_proposal\data\INCIDENT_DATE_2012.mat');

% load
tmp1 = tmp1.ADDRESS(1:100:length(tmp1.ADDRESS));
tmp2 = tmp2.CATEGORY(1:100:length(tmp2.CATEGORY));
tmp3 = tmp3.INCIDENT_DATE(1:100:length(tmp3.INCIDENT_DATE));

% concatenate tuples together per attribute
ADDRESS = [ADDRESS;tmp1];
CATEGORY = [CATEGORY;tmp2];
INCIDENT_DATE = [INCIDENT_DATE;tmp3];
```

IV. Analysis Log

Histogram

A useful visualization is generating histogram representing the distribution of data.
I generated four types of histograms:

- Type 1: Distribution of crime per year from 2005 – 2013
- Type 2: Distribution of crime per crime type for the years 2005 – 2013

Histogram Type 1 is shown in Figure 2.

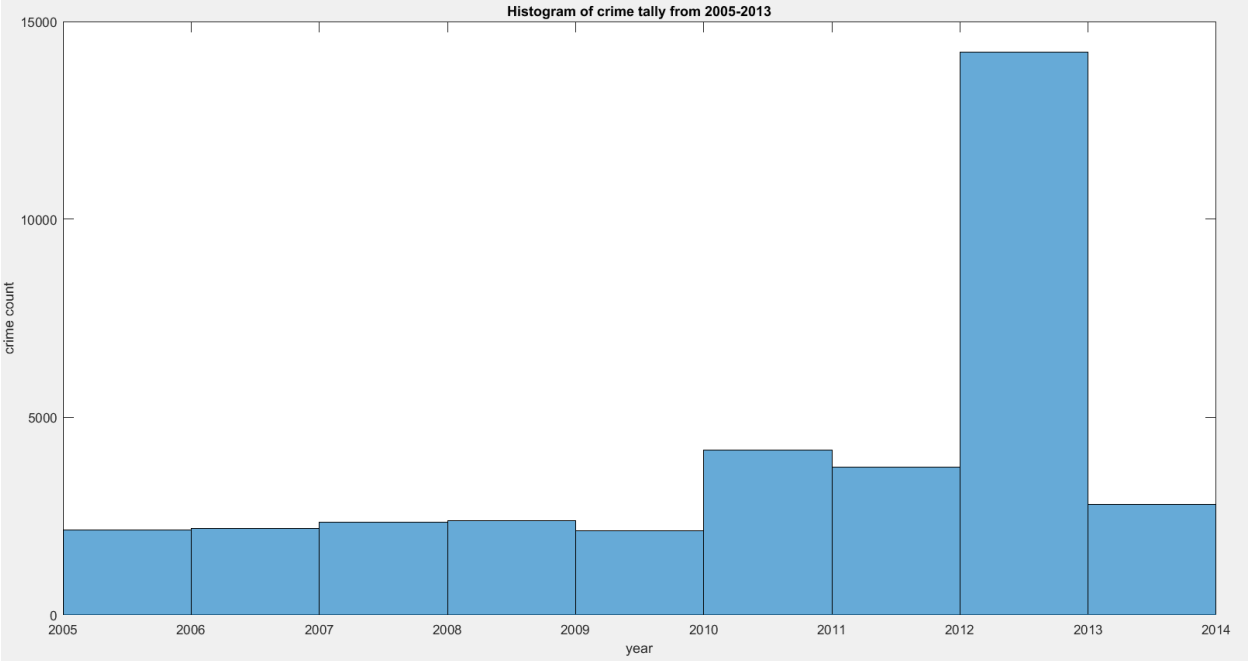


Figure 2: Histogram of crime frequency per year, from 2005 – 2013; 2014 index at the end can be ignored, as the index corresponds to the bar right after it.

The anomaly in the year 2012 was confusing at first; after checking the dataset, I realized it was simply a limitation of the given dataset; the number of tuples in the 2012 dataset was significantly larger than the rest of the given datasets. Two possible explanations: 1). that the sheriff department happened to receive more crime entries that year; 2). The dataset does not account for all the crimes. Given the recent steady decrease in crime rates in the Los Angeles area, I have made the assumption that the dataset entry for the year 2012 is due to dataset error; this, however, will not affect our analysis of the locations of these crime occurrences.

Histogram Type 2 is shown in Figure 3:

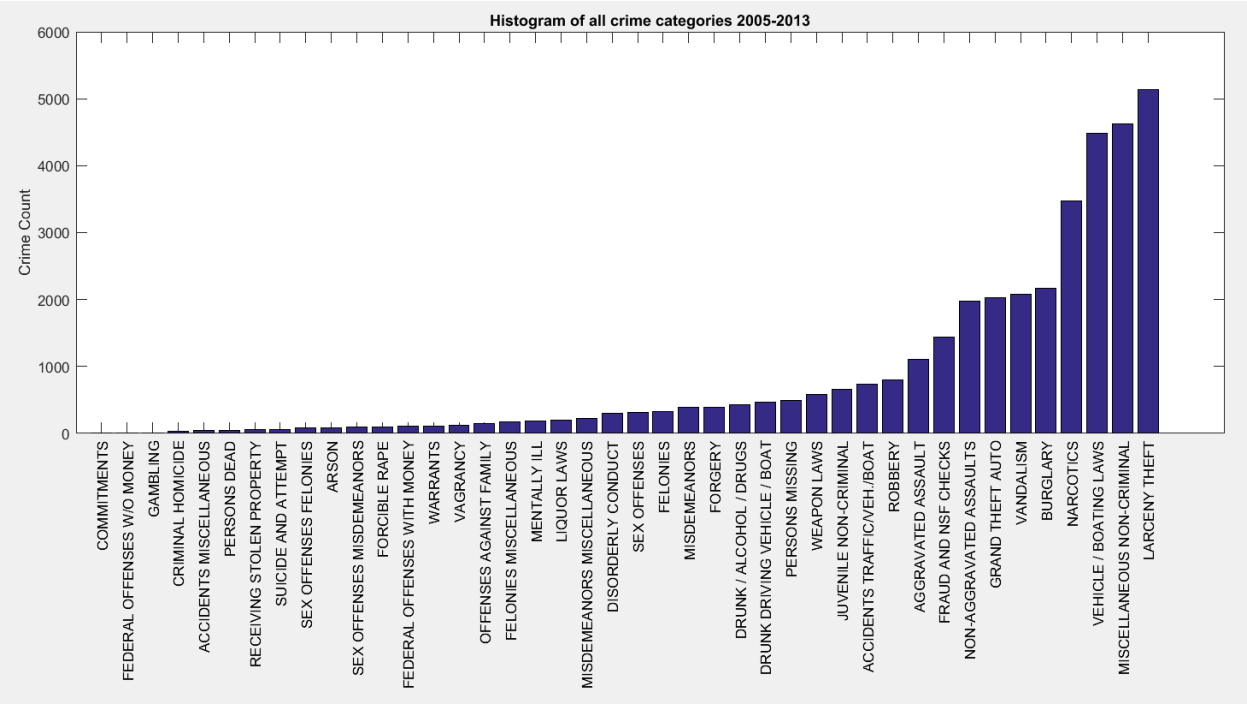


Figure 3: Histogram of all crime categories from 2005 – 2013; the x-axis shows all crime categories; the y-axis shows the crime frequency (sampled every 100 time for MATLAB performance issues).

MATLAB Script:

```
CATEGORY = CATEGORY.CATEGORY(1:100:length(CATEGORY.CATEGORY));
        INCIDENT_DATE =
INCIDENT_DATE.INCIDENT_DATE(1:100:length(INCIDENT_DATE.INCIDENT_DATE));
        month = month(INCIDENT_DATE);
        edges = [1:13];
        figure, histogram(month, edges);
        title('Histogram of crime tally in 2005');
        xlabel('month');
        ylabel('crime count');
        crimes = CATEGORY;
        [n, cellout]=cellhist(crimes);
```

In order to generate the histogram with the x-axis displaying all the crime categories, I used a special function, named `cellhist:`

```
function [n, cellout]=cellhist(CELL)
% this function plots a histogram based on char cell array.
% Input: CELL - a cell string array (Nx1)
%
%Output: n - alements in a bin
%        cellout - the bin value( a char)

%Example

if size(CELL,2)>1
```

```
    error('CELL need to be a vector of Nx1')
end

if sum(cellfun(@ischar,CELL))~=size(CELL,1)
    error('CELL must be a cell string array')
end

[cellout, mm, nn] = unique(CELL);
for i=1:length(cellout)
    n(i,1)=sum(nn==i);
end
[n,IX] = sort(n);
cellout=cellout(IX);
figure, bar(1:length(n),n);
1;
set(gca,'XTick',1:length(n))
set(gca,'XTickLabel',cellout)
set(gca,'Position',[0.2 0.4 0.7 0.5])
title('Histogram of all crime categories 2005')
ylabel('Crime Count')
rotateticklabel(gca,90)
```

Figure 3 shows the top four highest frequency crimes are:
Larceny, Miscellaneous Non-Criminal (trivial crimes), Vehicle / Boating Laws, & Narcotics
Larceny and Vehicle / Boating Laws are expected, since they are considered small crimes, but Miscellaneous showing up at 2nd place is a little misleading (we'll find out later why), and narcotics showing up so high indicates the prevalence of drug-related crime in LA.

2D Projection

I began by seeing the spread of the data by performing a random-2D projection. First, I formatted the data to find the frequency of each crime category given the year from 2005 – 2013 (similar to the *Hall of Fame* dataset we used in Homework 0). I then created a matrix variable called `Stats` to keep track of the frequency of crime given a category and a year. From there, I can do my analysis, starting with 2D projection.

MATLAB script to generate 2D random projection:

```
result = proj_2d(Stats, 2);
        figure
        plotmatrix(result(:,1), result(:,2));
        title('Graph of Crime Data using 2D-projection');
        outlier = findOutlier(Stats, result);


function [AP] = proj_2d(A, k)
[n, p] = size(A);
s = sum(A);
for idx = 1:p
A(:,idx) = A(:,idx)/s(idx);
end
P = rand(p, k);
AP = A*P;
```

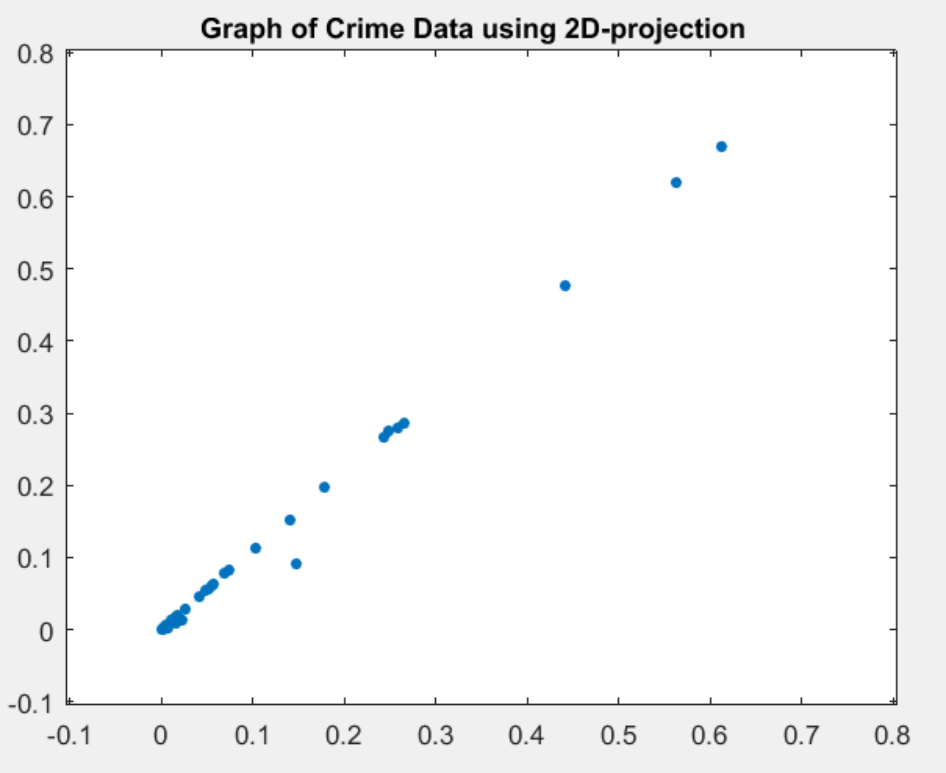


Figure 4: Random 2D projection of the data

The data seem to have a large spread, but also cluster at the bottom left. I will continue to analyze this variance through PCA.

Principle Component Analysis (PCA)

I performed three kinds of PCA:

- Type 1: 1st and 2nd PCA graph
- Type 2: Correlation PCA graph
- Type 3: 1st, 2nd, and 3rd PCA graph

Type 1 PCA is shown in Figure 4 (on the next page):

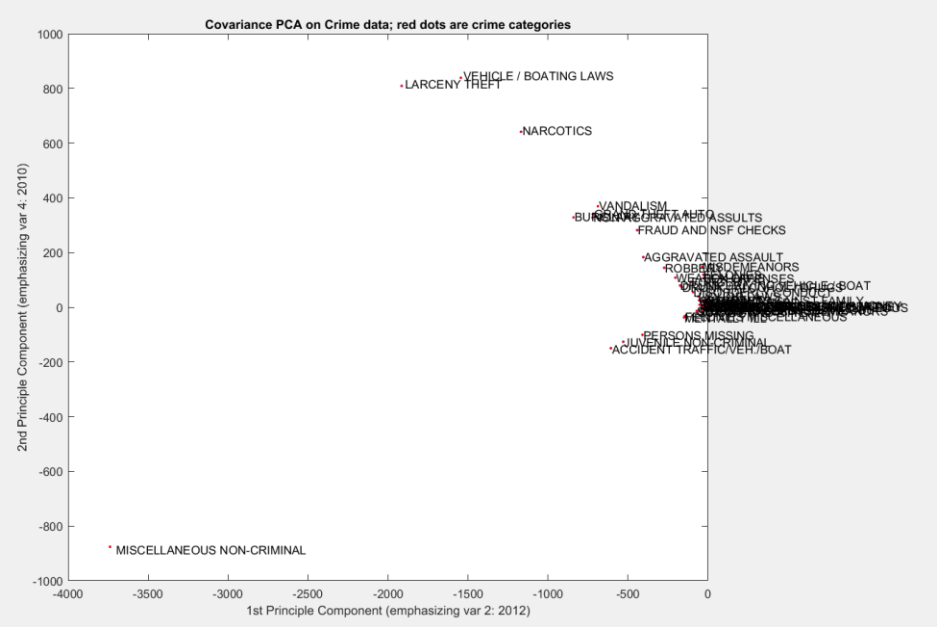


Figure 5: Taking the SVD of the covariance matrix, we see how each crime category is grouped.

Type 2 PCA is shown in Figure 5:

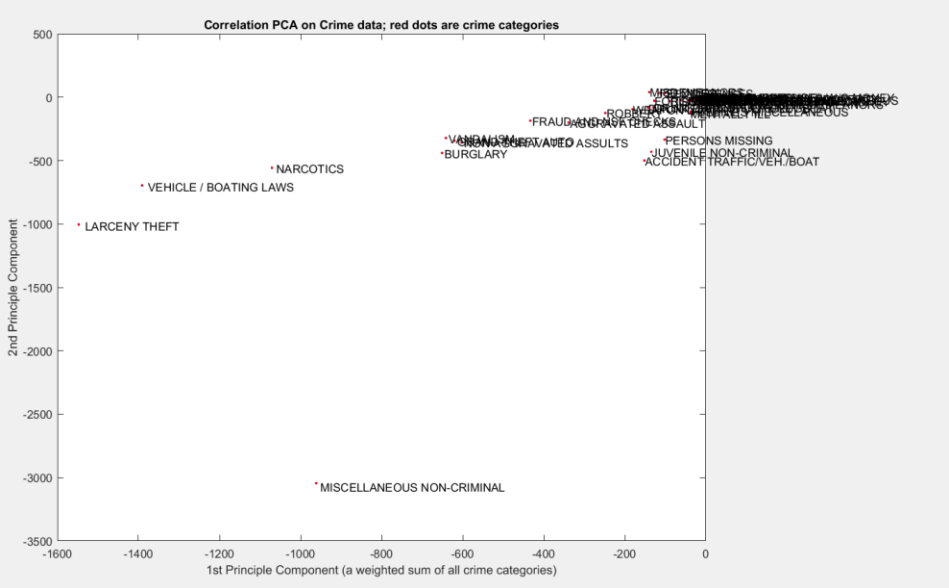


Figure 6: PCA Taken on the Correlation matrix.

Figure 4 and 5 both show that the crime category 'Miscellaneous non-criminal' has a lot of variation over the years, and therefore does not group well with the other crimes. After checking the actual data, I saw that its appearance from 2013 going back has exponentially decreased. There are no explanations in the dataset website about this, so I will not take this category into consideration. However, the other categories show consistency in in the PCA analysis, indicating that their frequency is relatively constant throughout the years.

Furthermore, we see that PC 1 puts weight on the year 2012 and PC2 emphasizes the year 2010. From our knowledge of PCA, we know that this means our data spreads are largely determined by those two years.

PCA Type 3 is shown in Figure 6.

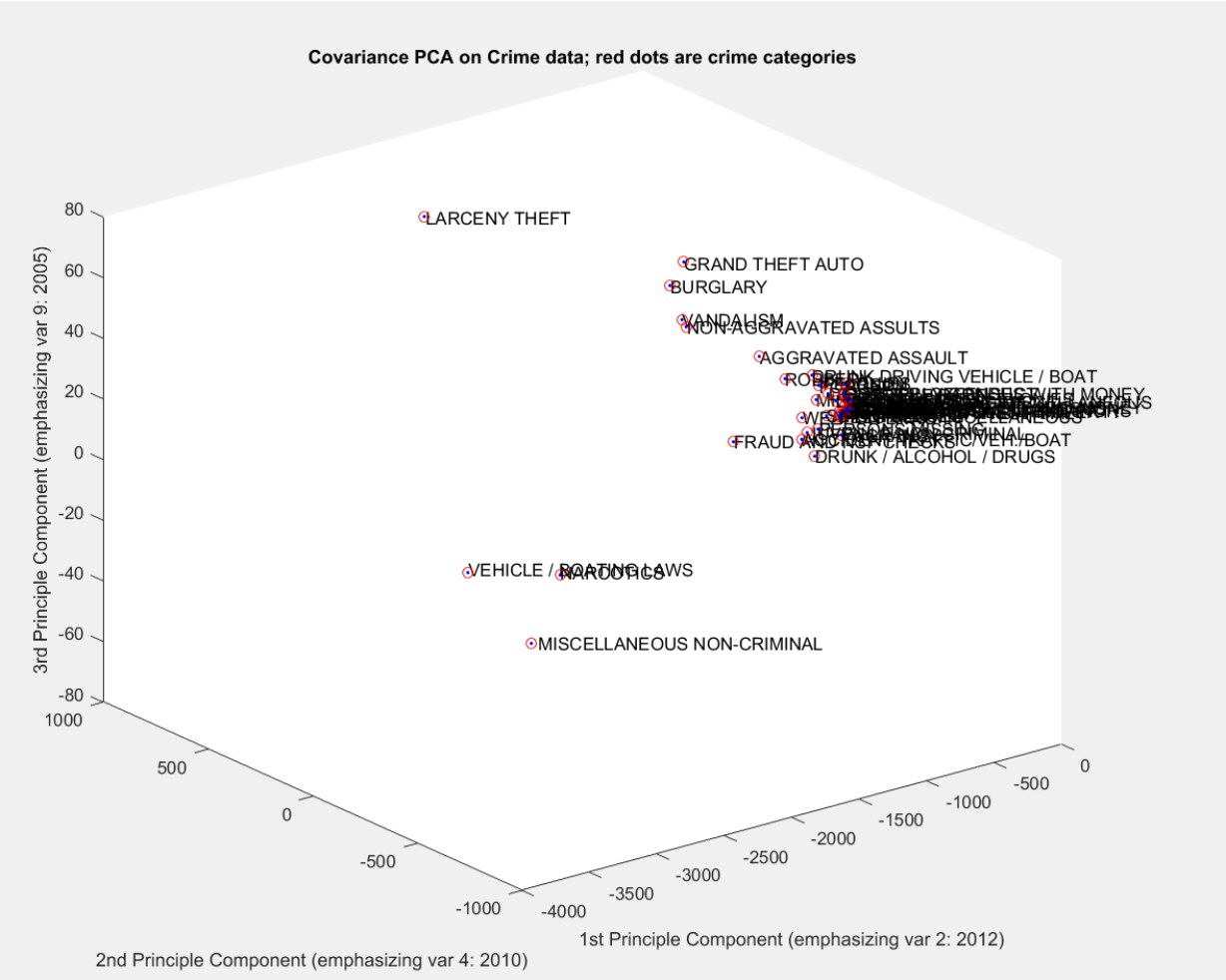


Figure 7: data plotted on the first three Principle Components

Figure 7 is an extension of our PCA analysis on the first two principle components, this shows us that the spread of crime categories such as *Larceny* or *Vehicle / Boating Laws* are largely governed by these three years: 2012, 2010, & 2005. We have already determined that *Miscellaneous Non-Criminal* is an outlier, so it is no surprise that the data point is once again below the general correlation.

V. Map Visualization

The next step in this project is to visualize the crime data as a heat map on the greater LA area. I am using a map visualization tool called MapsData[4]. I formatted my crime data into three attributes:

- Crime category
- Latitude
- Longitude

First, I had to generate longitude and latitude points from the `ADDRESS` attribute from my crime data. I used a function called `geocode`[5], which uses OpenStreetMap as an API call to convert addresses to longitude and latitude. I was able to finish converting my address data to the appropriate coordinates, but OpenStreetMap limits the number of API calls a user can make in a given amount of time, and since I created a loop to continuously make calls to convert addresses to coordinates, I was temporarily denied access to using the service. However, I had gathered all my data for the visualization.

Using the Excel file I generated from MATLAB through the command `xlswrite`, I was able to create a heat map visualization of my crime data, shown in Figure 8.

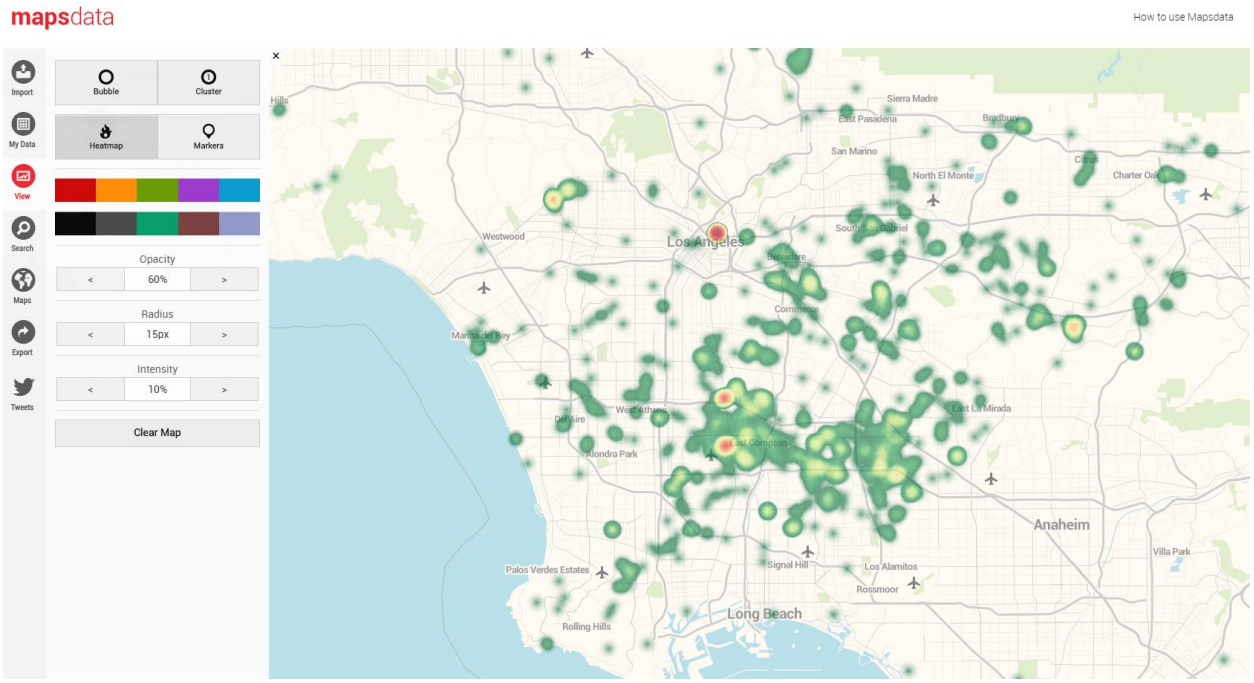


Figure 8: Heat map of crime frequency in the LA area, from 2005 - 2013

Immediately, it is clear that some neighborhoods do get a greater *spread* of crime frequency. By *spread*, I mean the area covered, not necessarily the concentration. East Compton, for example, lies in the middle of that spread of green, indicating that while there's no intense concentration of crime in a single location, there are crimes occurring within short distances of one another.

Another visualization is generating the cluster of crimes occurring at close radiuses of each other, as shown in Figure 9.

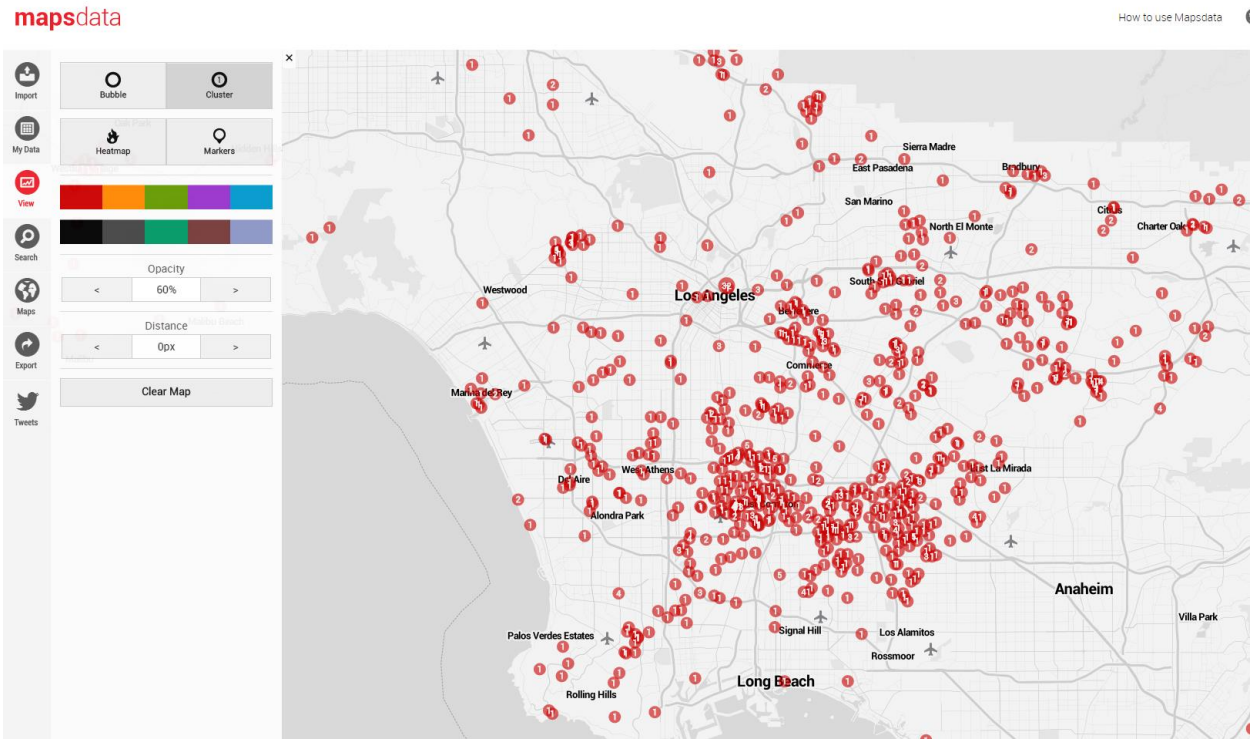


Figure 9: each red circle has a number that represents the number of crimes that occurred near that location

From here, it is clear to observer that there are definitely clusters of areas where crime frequency covers the map; the relative distance among each occurrence is very short. Furthermore, there are areas where multiple crimes happen at very similar locations.

More visualizations can be made by isolating certain conditions, as the next figure shows.

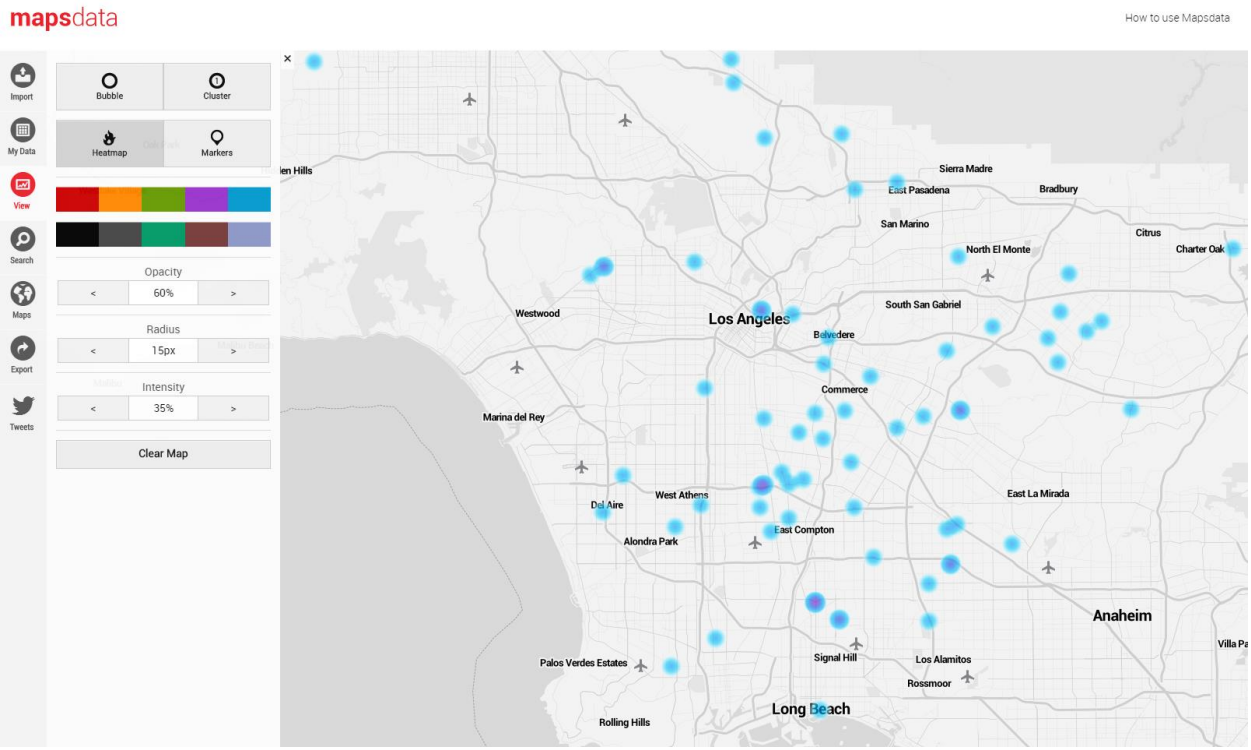


Figure 10: heat map of crimes under the category "Narcotics".

Now, it is time to visualize the building inspection dataset. My original hypothesis was that, due to Broken Windows Theory, areas where building inspections happen are generally areas where the neighborhood is maintained and watched over, and therefore has less crime frequency. The building inspection dataset has the following data organization:

Tuple #	Latitude	Longitude	Address	...
1	33.9658	-118.4241	11444 Olympic Blvd	...
2	33.9741	-118.4622	11479 1/2 W Tiara St	...
3	33.9692	-118.4584	1 S LMU Dr	...
...	...	...	...	...

Figure 11: data organization of the building safety inspection dataset

Figure 12 shows the locations where building inspections were conducted on MapsData:

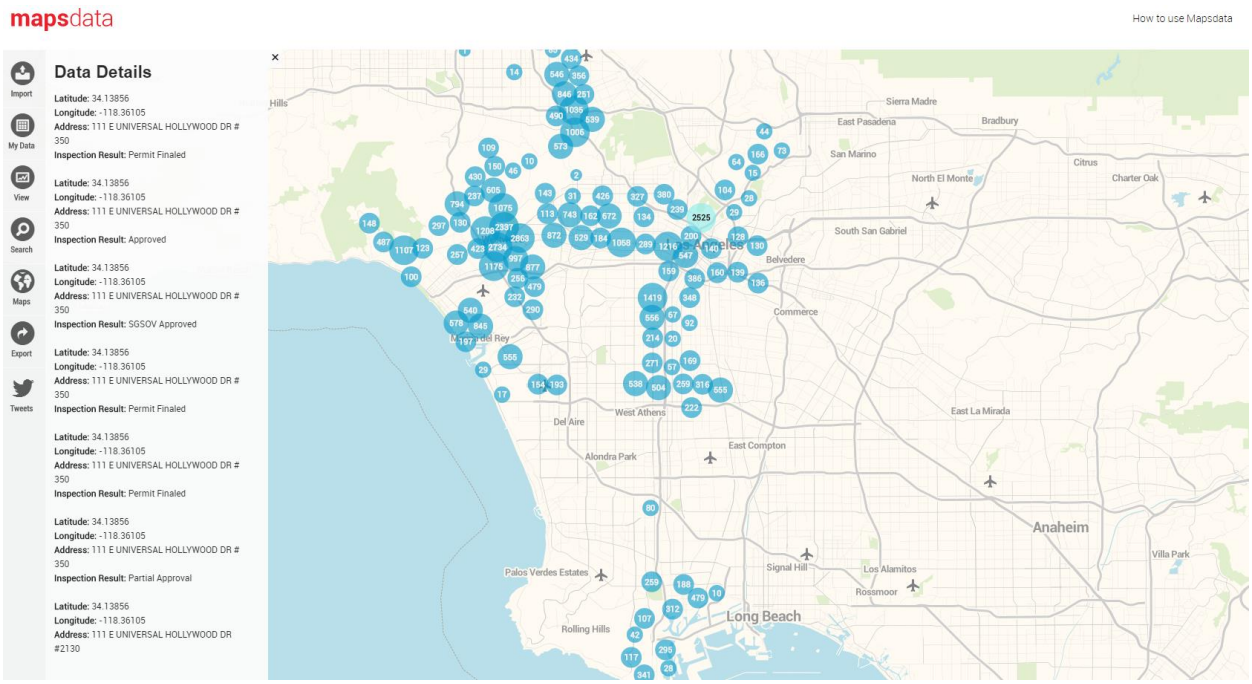


Figure 12: cluster map showing the locations where building safety inspections were done. Clicking on a cluster circle shows all the building inspections that have happened in that area (left)

Now, I will overlay the locations where building safety inspections are done with the heat map of the crime locations:

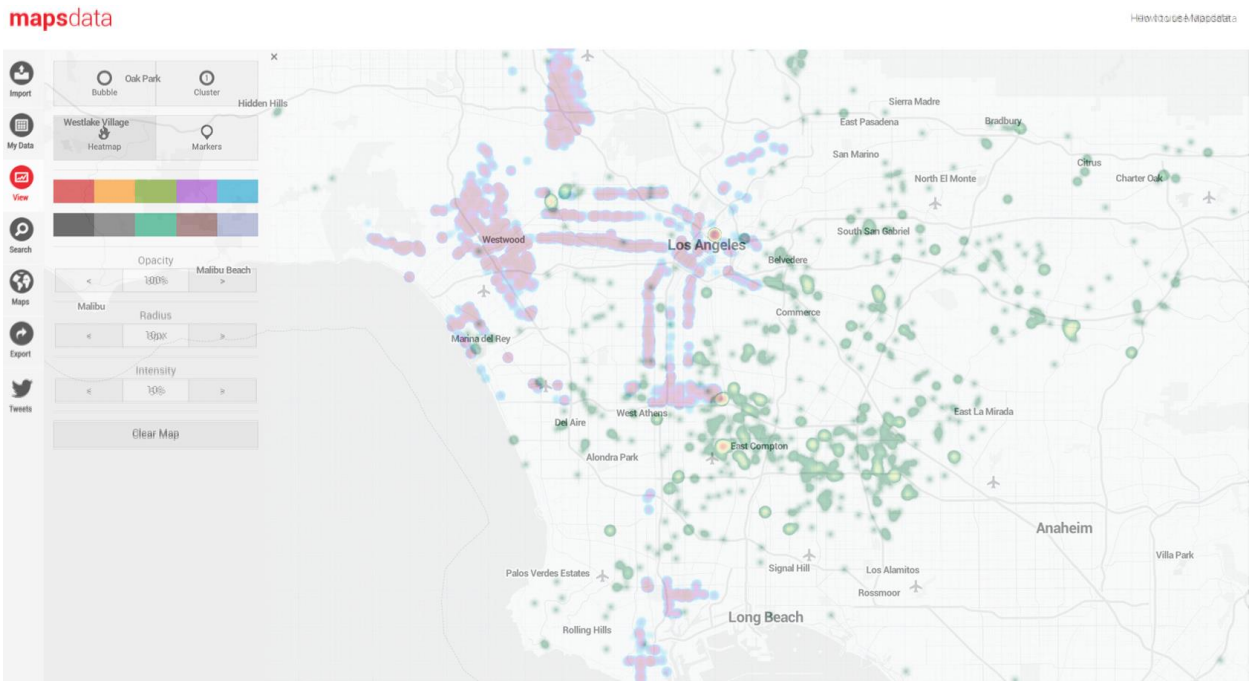


Figure 13: building inspections location (in purple) overlayed with crime locations (in green and red heat map). Notice how separate the two datasets are on the map.

VI. Results & Analysis

Figure 13 illustrates the Broken Windows Theory very well. Areas in LA such as Bel Air, Marina Del Rey, and Westwood have a high concentration of building safety inspections, while at the same time, the crime data heat map is not concentrated in that area. In fact, almost all the areas where the building inspection heat map shows up, the crime data does not, and vice versa. Perhaps this is obvious, but the pattern continues to show for specific crimes, such as narcotics, shown in Figure 10. Undeniably, there exists an inverse relationship between areas with high building safety inspections and crime frequency.

My hypothesis has been proven correct, and while it may seem like a straightforward conclusion, it is ironic that many of our actions towards reducing crime is to place more people in prison and increase police/citizen hostility instead of focusing on education for the younger generation and spending money on maintaining neighborhood infrastructure. A key idea of the Broken Window Theory is that a sense of civility is greatly reduced in neighborhoods environments that suggest apathy; likewise, the sense of civility is greatly increased in neighborhoods that appear to be 'civilized' or ordered.

VII. Conclusion & Future Works

This project has been a great experience for me, mainly because I was able to use a variety of tools, many of which I learned in CS 170A, to analyze and make conclusions from large datasets. However, what made me really excited to pursue this topic was the potential to relate social analysis into big data. I mentioned in my project proposal that I learned about the Broken Window Theory from my GE class freshmen year, Geography 3, which is the study of how we shape our environments, and how our environment sometimes shape us and our behaviors. My professor for that class mentioned the potential of using data to visualize human and environment interactions. Now, two years later, I was able to use my knowledge from that class in Mathematical Modeling.

I believe the results from this project clearly indicates that a major factor in decreasing crime in urban cities is to tackle the roots of the problems: it's about educating the younger generations; it's about supporting laws that focuses on providing a budget to reform city infrastructure; it's about replacing gentrification with renovations. Ultimately, it's about letting everyone understand that we shape our environment, and each of us have a responsibility to maintain and take care of our neighborhood.

VIII. References

[1] Wilson, James Q., and George L. Kelling. "Broken Windows." Manhattan Institute (n.d.): n. pag. Manhattan Institute. Web. 18 Mar. 2016, Retrieved from *https://www.manhattan-institute.org/pdf/_atlantic_monthly-broken_windows.pdf*

[2] "Current Crime Data." Los Angeles County Sheriff's Department. Los Angeles County Sheriff's Department's Jurisdiction, n.d. Web. 18 Mar. 2016, Retrieved from *http://shq.lasdnews.net/CrimeStats/CAASS/desc.html*

[3] "Data Catalog." Building and Safety Inspections. Data.lacity.org, 3 Feb. 2015. Web. 18 Mar. 2016, Retrieved from *http://catalog.data.gov/dataset/building-and-safety-inspections-f6000*

[4] MapsData Tool, Retrieved from *http://www.mapsdata.co.uk/*

[5] geocode MATLAB function, Retrieved from *http://www.mathworks.com/matlabcentral/fileexchange/37860-geocode*