COMPSCI 597: Big Data and NOSQL Databases

Term Project Final Report

Tanner, Reuben, Trisca, Gabriel

December 10, 2013

Contents

1 Introduction

1.1 Motivation

Cities and communities maintain databases of crime to be able to predict and prevent criminal activities from happening. A major motivation behind what is called "crime data mining" [5] is identifying patterns of criminal behavior[1].

Crime is a big burden on society: in the United States alone, 23 million criminal offenses were committed in 2007. The economic loss of victims was $15 billion and $179 billion were spent by the government on police protection, judicial and legal activities, and corrections [4]. Even if a small percentage of these crimes can be prevented, the impact would be extremely significant[3].

In addition, identifying what variables influence the likelihood of a certain type of crime happening can shed light on the complex dynamics behind all criminal behavior, helping understand the motivations behind these antisocial conducts, and how the environment affects people and the choices they make[2].

1.2 Dataset

We are working with four datasets:

Weather Daily weather summaries from the City of Chicago for the time period 2001 – 2013
Location: http://www.ncdc.noaa.gov/cdo-web/

Crime This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.
Location: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Socioeconomic Indicators Census Data - Selected socioeconomic indicators in Chicago, 2007 – 2011
Location: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

Public Health Statistics Selected public health indicators by Chicago community area.
Location: https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu

2 Problem

We want to predict crime by using socio-economical information, health indicators for the different areas in which crimes are committed and historical weather information collected in the city of Chicago.

3 Method

3.1 Code Structure

With our focus in mind of being able to predict crimes, we needed a glue that would allow us to connect our socio-economic indicators, public health indicators, weather and crime data. The "community area" ended up being the exact glue that we needed. Chicago is broken up into 77 community areas and, in each of our datasets (aside from weather) the community area is either explicitly or implicitly defined.

The first phase of our analysis was to map each crime to a community area based on its specified latitude and longitude. This involved finding a KML file which contained the boundaries for each of the 77 community areas and then generating polygons for each of the community areas and mapping each crime to its corresponding polygon. Once we had successfully mapped each crime to a community area, we remapped the entire dataset to a file, stripping the unnecessary fields of the crimes and

applying the community areas.

Once we had remapped the dataset to only contain the fields we needed and the one we extrapolated we were ready to begin focusing on the relationships. Because of the abundance of different crime types (384 specifically), we decided that attempting to derive correlations for each of them would be somewhat fruitless so we narrowed our focus to be only on the top 50 crimes committed over the entire city. This data was extracted by one simple Hadoop job.

After determining which crimes occurred frequently and which community areas they occurred in we had enough information to extract correlations from public health indicators and socio-economic indicators but we still needed a way to link crimes to our weather dataset which listed the weather daily. We decided this could best be accomplished by running a job to determine what crimes (of the top 50) occurred on any given day so the dataset was remapped to have a key of a day and a value of the frequencies of the top 50 crimes committed on that day.

At this point, we had enough pieces of data to link all of our datasets and now we needed to determine which of them were correlated with which others so we pulled out the big guns and wrote a manual full join against each of the files and the crime dataset. This was especially challenging considering there were 4 files being joined and one of the files (the weather dataset) required a different matching scheme than the public health indicators and the socio-economic indicators because weather was keyed by date. In the reduce phase of the correlation, x, y points were calculated for crime type and frequency vs. severity of public health indicator, crime type and frequency vs. severity of economic indicator and crime type and frequency vs. weather pattern. A file was produced containing the points for each of the features within the each of the files.

3.2 Challenges

One of our initial challenges that we faced was our choice of data set. When we first began this project we desired to perform analytics on a crime dataset for Washington DC to perform geospatial and temporal analysis. Once we started getting into the dataset that was on the opendata website for DC, we realized that the data did not have accurate locations provided through latitude and longitude or through any other means so we decided to go with a different dataset that we could more easily plot, specifically, the 311 requests dataset from Washington DC.

After we shifted our focus from crime to the 311 requests, we began performing analysis on it by following the pattern laid out in our Data Analysis section. After we generated scatter plots and correlations for the requests against weather patterns and begin critically examining them, much to our dismay, we found that there were very few interesting correlations. One of the most obvious and mundane correlations we found was that on cold days there are more requests for heating; ground breaking.

Once we reached the dead end with the 311 requests we decided to switch our focus to finding a new, crime dataset that would have the features which would make the analysis far more interesting. Our paper is written about that dataset, namely, the Chicago crimes dataset.

3.3 Data Analysis

As an initial step, we are calculating the correlation between crime frequency and different indicators. To achieve this, we performed a join operation between the weather data and the crime frequency data with date as the key. For the socio-economic and public health indicators we joined the datasets on the community area where the statistics came from and where the crime was committed, and the crime frequency was accumulated.

These scatter plots are the top most highly correlated pairs of indicator/crime type.

Indicator Name	Description
16_UNEMPLOYED	Percentage of persons 16 or older and unemployed
25_NO_HIGHSCHOOL	Percentage of persons 25 or older without a high school diploma
BELOW_POVERTY	Percentage of persons under the federal poverty line
BLOOD_LEAD	Micrograms of lead in blood
CANCER	Cancer rates per 10,000 people
CROWDED	Percentage of people living in crowded housing environments
INCOME	Average income level
LOW_BIRTH_WEIGHT	Percentage of babies being born with weight under normal level
NO_HIGHSCHOOL	Percentage of persons without a high school diploma
PRENATAL_CARE	Percentage of children that received prenatal care
PRETERM_BIRTH	Percentage of children born prematurely
TEEN_BIRTH	Rate of teen pregnancy and birth
UNEMPLOYMENT	Unemployment rate (percentage)
YOUNG_OR_OLDER	Percentage of poersons 16 or younger and 65 or older

Table 1: Indicators and their description

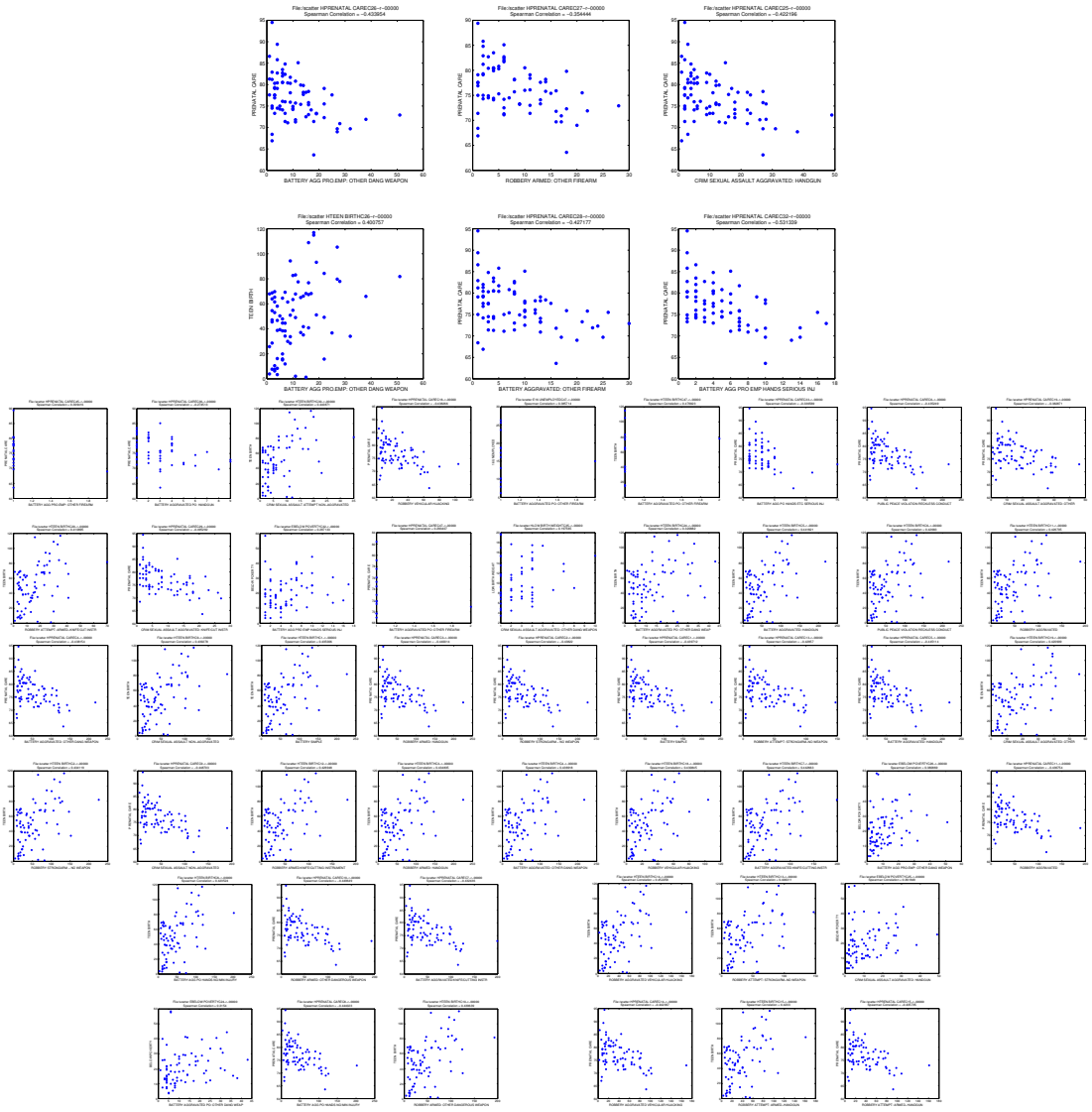

Figure 1: Scatter plots and Spearman Correlation

Out of the top 50 crimes, we chose to group them in three main groups: **Battery**, **Sexual Assault** and **Robberies**. Those crimes were the most frequent and we believed that they are different enough that a causal pattern was highly likely.

The distribution of the indicator values is described below.

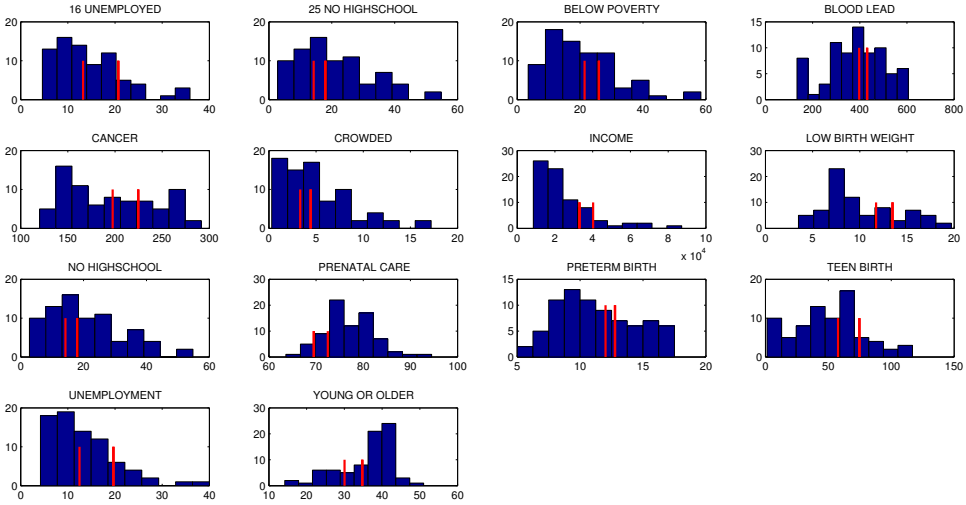

Figure 2: Distribution of predictor variables. Lines in red represent the mean of the top 5 community areas with the highest type of crime for the three categories (BATTERY, SEXUAL ASSAULT, ROBBERY).

Based on these observations, we proceeded to generate artificial samples from the mean values of the indicators for each crime type. The algorithm is described below:

1. Identify which group a given indicator belongs to (BATTERY, SEXUAL ASSAULT, ROBBERY).
2. Extract the mean of the top 5 values for said indicator
3. Create a sample with these values.

These steps are performed on every one of the top 50 crimes, and then the data is used as training for a classification tree.

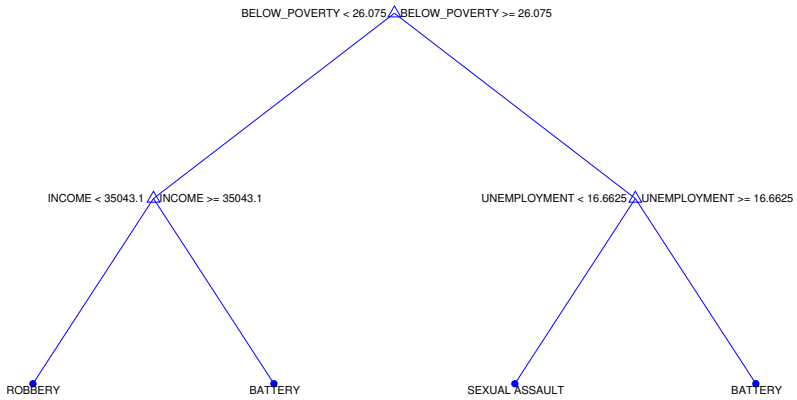

Figure 3: Classification tree based on three columns, INCOME, % BELOW POVERTY and UNEMPLOYMENT

4 Results

We can see that the there are different variables that interact and that there is no single indicator that can predict the likelihood of a type of crime. While the interpretation of these results are highly subjective and only reflect a statistical correlation, much less an actual causation, we can conclude that certain indicators related to poverty are highly correlated with violent crime.

Specifically, we see that looking at the percentage of people below the poverty line, income level and unemployment of a specific community area, we can characterize the type of crimes that are most likely to occur.

An alternative tree using all the indicators provides the following result:

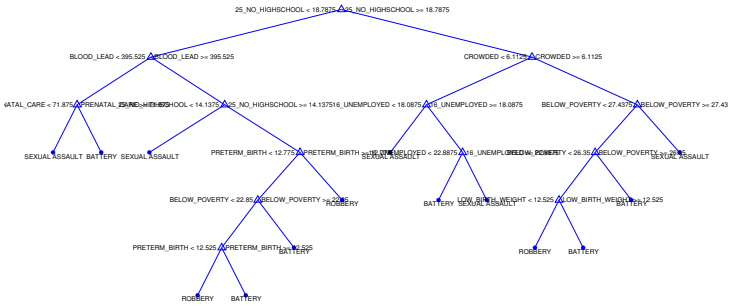

Figure 4: Tree using all socio-economic indicators (Tree has been pruned)

4.1 Future Work

These project didn't perform a class-weighted classification, that is, homicides for example represent less than 2% of all crime committed in Chicago, and therefore the classification technique used didn't capture the under-represented classes, missing them completely. Another future improvement would

include Montecarlo simulations to approximate the uncertainty in both the sample class and the indicators. These socio-economic and health indicators are sampled from a whole community area and local variability should be accounted for. Same for the sample classes: we are assuming that the crime classification for robbery is always accurate when different kinds of robberies related crimes could be analysed separately.

5 Conclusion

The process involved in putting this project together was very educational and we learned new techniques and familiarized ourselves even further with working on large datasets. We gained one of the more valuable gems of experience by reaching a dead end with our first attempt and rebuilding most things from scratch on a new dataset. There was also quite a bit of "peripheral data science" that needed to be done with various small programs that performed necessary operations, like the mapping of crimes to community area and other related activities which provided excellent learning opportunities.

In conclusion, we were able to use the MapReduce paradigm for data analytics like we had never done before. We truly enjoyed this class, as well as this project.

References

[1] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin, and Michael Chau. Crime data mining: A general framework and some examples. *Computer*, 37(4):50–56, 2004.

[2] Aurelio José Figueredo, Paul Robert Gladden, and Zachary Hohman. volume 1. University of Texas, 2012.

[3] Shyam Varan Nath. Crime pattern detection using data mining. In *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*, pages 41–44, 2006.

[4] Federal Bureau of Investigation U.S. Department of Justice. Crime in the united states: Uniform crime reports. http://www2.fbi.gov/ucr/cius2008/index.html, 2008. Accessed: 2013-12-5.

[5] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Learning to detect patterns of crime. 8190:515–530, 2013.