Blake Conrad

CSCI-49500 – Capstone Research

Project Advisor: Murat Dundar

FBI City Crime Prediction Analysis Via Unsupervised Learning

**Introduction**

Data analysis has played an important role for centuries and alongside the ushering in of the digital age, this demand for data and comprehension thereof has only increased. Machine learning, data mining, and applied statistics have become a hot area of research due to the efforts of leveraging cutting edge techniques to develop unique and innovative solutions to making optimal sense of data. Scientific principles have also been meshed with this hybrid data driven field of so many disciplines uniting to make some amazing discoveries in the area of predictive analysis. My personal project in this exciting field of research has been to look at an *FBI Crime* dataset provided generously by [Dr. Murat Dundar](http://mdundar@iupui.edu), my faculty advisor. By definition, data mining is the technique of finding unintuitive patterns in data through diverse algorithms and techniques. This was precisely my aim for this project and is my hope that the approaches taken may shed some light on the uniqueness of the problem, the data set, and the future of crime for cities in the United States. This projected piqued my interest when I saw its potential for scalability. By developing some type of technique that can classify cities into crime patterns, diagnose how healthy those patterns are, and determine if new cities are trending in a good or bad direction based on a live feed of new data, I immediately saw the value in exploring such an engine to produce these results. The milestones for a project like this include many important questions that must be answered before any results can be examined credibly. The types of questions that must be addressed include:

* How can we distinguish are good, bad, and neutral cities?
* How many crime patterns exist in the data set given?
* What is a good way to group cities with similar tendencies (population increasing/decreasing/stagnant)?
* How do certain groups of cities behave as their population changes?
  + Is there a way to determine more favorable crime from less favorable?

Each of the listed questions and more will be explored in the analysis of this project report. A detailed project overview can be found in *Figure 1*.

../THE_QUEST___Progress_Folder/images/uml_outline_bmconrad.pdf

*Figure 1:*

A high level overview of the projects goals and initiative.

**What are good and bad cities?**

At first glance of the problem pertaining to assessing what determines a good or a bad city, was to reach out to what others think. Social media has been a great resource for many to examine how a population views particular issues and topics of a diversity of interests. In the U.S, one of many reliable new media sources is <http://www.usatoday.com>, which has a specific section for best and worst cities to live. After spending far too much time assessing these findings, I found my specific sources to be population skewed. My faculty advisor helped me in directing me toward resources that are not skewed by population, because good and bad crime is independent of population. The sources for the best and worst cities to live based on crime may be found at:

* <http://time.com/money/collection/best-places-to-live/>
* <http://247wallst.com/special-report/2016/06/28/the-worst-cities-to-live-in/11/>

These links each represent a disjoint set containing 50 unique city state observations that have been known with reliability to be either good or bad based on crime. By writing a small script, I was able to scrape each web page and pipe the unique city states into my program to do more rigorous data mining techniques and data analysis.

**How many crime patterns are in the data?**

When looking at any type of large scale problem, the hope is to find general tendencies or patterns. Once a general tendency is found, we can find a general solution to help aid or fix the problem. When looking at crime data, our hope is the same. Because of the tabular structure of the data set, I was able to apply the *K-means Clustering* algorithm on the data to determine groups for each of the cities. These groups are often interchangeably referenced as clusters or patterns throughout the data mining community. The only glaring issue this technique is the hyper parameter associated with the algorithm; K. My initial approach at this involved an exhaustive approach from. After structuring a data file with my 50 best cities and worst cities from my data set over time from 1979 to 2014, I could apply purity-scores to each cluster attempt, and measure:

1. What is the purity score of this attempts K-value?
2. Which clusters yielded the highest purity in the best cities?
3. Which clusters yielded the highest purity in the worst cities?

Thanks to support from my faculty advisor, I was able to construct a purity matrix to represent these data:

The matrix A is a *purity matrix* which tells us each column represents a cluster, the first row represents the number of the best cities to live in who landed in that cluster, and the second row represents the number of the worst cities to live in who landed in that cluster. From this purity matrix I was able to assess the highest purity score, with my final purity scores results:

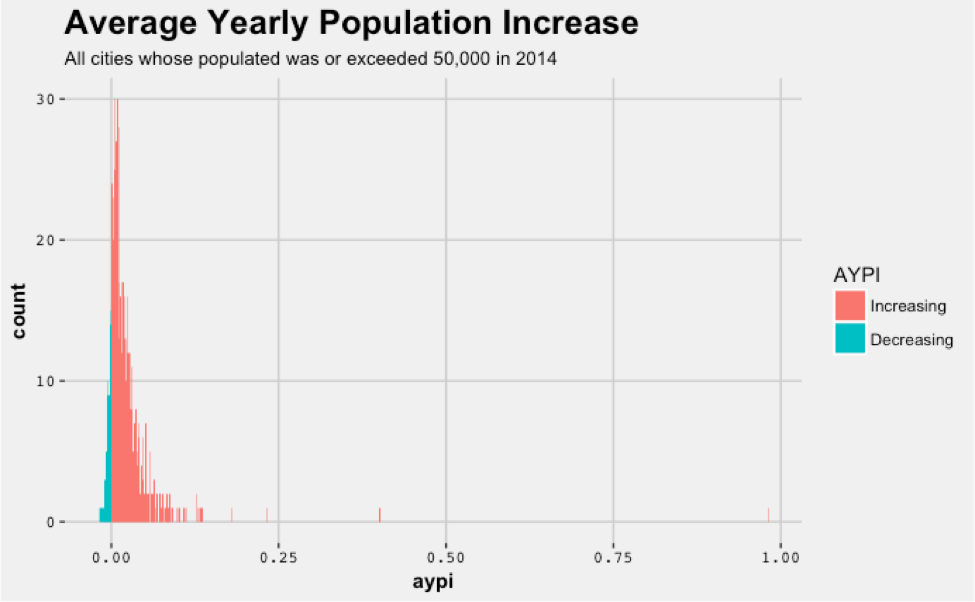
* K = 5 🡪 purity-index value = 0.8453
* K = 10 🡪 purity-index value = 0.8905
* K = 15 🡪 purity-index value = 0.8786
* K = 20 🡪 purity-index value = 0.8419
* K = 25 🡪 purity-index value = 0.8688

With the analysis from above, one can conclude that the best hyper parameter K for this data set is at 10. To assess our original question, we can with reasonable confidence say that there exist about 10 crime patterns in the data. We can also infer from looking at our clusters which clusters contained the most purity toward the best cities to live in and worst cities to live in by looking at the majority who ruled across each column between each row. For example, by looking at the first column of matrix A, we can see the best cities only had 2 observations in this cluster, while the worst cities contained 34. After continuing this process through all columns, we can form two new sets to represent the better crime patterns and the worse crime patterns:

By knowing the best and the worst cities to live in it, would only make sense to find some sort of way to see which types of cities in general had these less favorable attributes associated to the best and worst cities. By separating the cities into different groups, we can see how those groups changed over time with respect to one another, to see what types of crime made the worse ones worse, and the best ones best.

**How can we group cities into increasing, decreasing, or stagnant (from 1979 to 2014)?**

A primary focus and personal interest of the project was to examine large cities. By examining large cities, one can group them appropriately as they ascend or descend in population. By setting the qualifying condition that makes a city large at its population being greater than 50,000, we can classify cities that meet this set requirement. My initial approach simply looked at raw numbers, for example: Look at all cities whose population in 2014 was above 50,000 and whose population in 1979 was below 10,000. This approach had several holes, and was quickly revamped to looking at a cities *Average Yearly Population Increase (AYPI)*. This approach introduced a good metric for all cities to be compared and grouped by. By applying the *AYPI* algorithm starting from 2014 on my subset of data whose population at 2014 was 50,000 or greater, I was able to accurately split this subset into two distinct groups. The curve that represents the findings from this is in *Figure 2*:

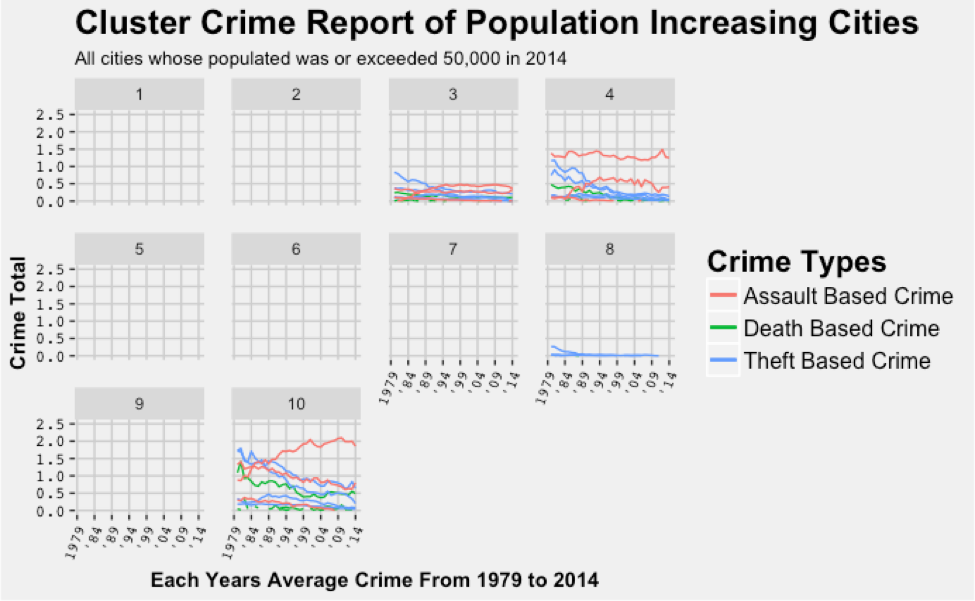


*Figure 2:*

A curve representing the decreasing cities and increasing cities from our subset of data.

After successfully categorizing the groups into those which increased and decreased on average, I am now able to fine tune my analysis by seeing how they trended together in view of their crime correlations.

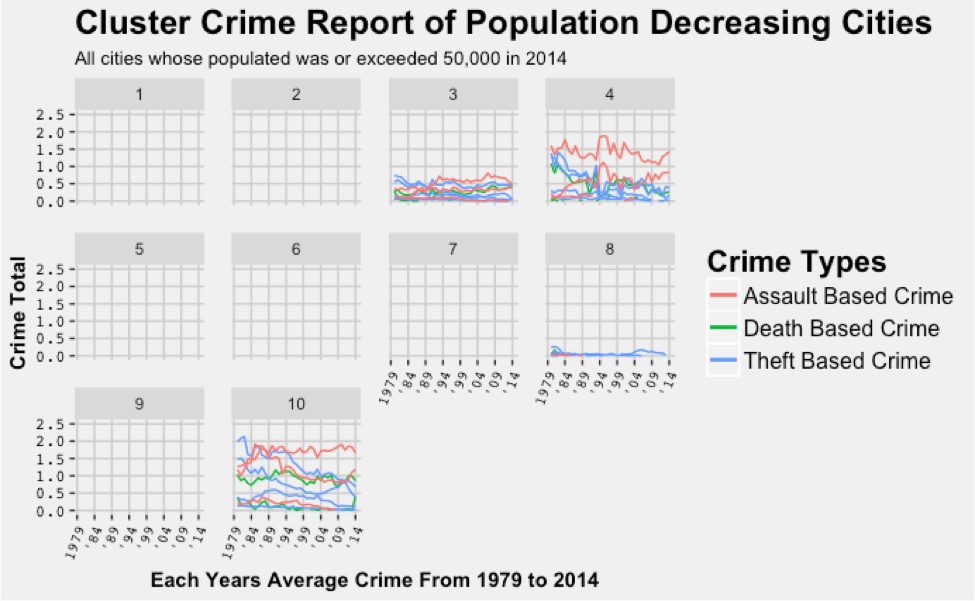
**How does crime change in a group of cities which increase/decrease together?**

After categorizing our subset of data into two distinct groups (increasing or decreasing) we can start to look at correlations between different clusters, groups, and cities within those groups for interesting findings. Specifically, I began by looking at my population increasing cities. First, I wanted to see how and why the clusters were acting the way they were and what the *K-means* was actually finding in the data. To do this, I look at each city’s cluster assignment over each year, then by finding every occurrence of a city in clusters 1 through 10, I can average them up for each year’s occurrences, then make a distinct plot for the 10 different crimes in the data set on average over time per cluster. This lead to the plot generated in *Figure 3*. The plot in *Figure 3* allows us to see inside each cluster over time. I was also able to better categorize my crime data fields into three groups, *Assault Based Crime* (Total Assaults, Officer Assaults, and Total Rape), *Death Based Crime* (Murder, Man Slaughter, Officers Killed), and *Theft Based Crime* (Total Robbery, Total Burglary. Total Larceny, and Auto Theft). The first interesting thing that popped out from the plot was the worst cities highest crime was the assault based crime, in addition to its general trend as most dominant. I could also see that the clusters 4 and 10 had very spread out amounts of crime, with appearances of high volumes. Clearly, high volumes of crime are never good, but *spread out* volumes of crime may also be a general problem. *The* *K-means* algorithm was able to find this unique pattern amongst the worst crime clusters. It is not that we do not see the appearance of assault based crime in our good crime clusters, but it is not dominant with a high amount or containing a heavy spread of crime in different volumes. After generating a second plot for the decreasing cities, I came up with *Figure 4*. By looking at

*Figure 3:*

A plot to represent each cluster’s average crime reported over time for increasing cities.

the decreasing cities next, there is little distinction aside from some increased fluctuation in theft based crime for cluster 4 and cluster 10. This leads me to believe that whether a city is increasing or decreasing in population, high amounts of assault based crime trends and high spread of different crimes, tend to be the driving factors when classifying a city from our model. I found this discovery relatively odd due to my intuition that death and murder based crime would be the most undesirable trait for humans when deciding to live in a city. Without a doubt, death based crime plays a role in our model, but it is not the most significant distinction between the groups. Not only is high assault the main factor in determining if a city is becoming worse over time, but the theft crime as well. When there exist high amounts of theft based crime, we can see that it is only slightly behind the assault based crime as a driving factor. This also explains cluster 8 as being a healthy transition or generally healthy cluster from our model as well.



*Figure 4:*

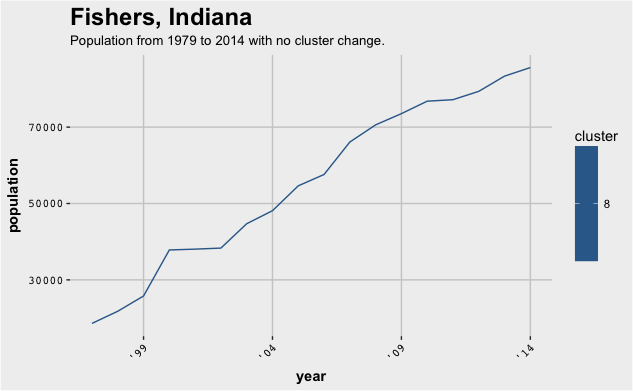
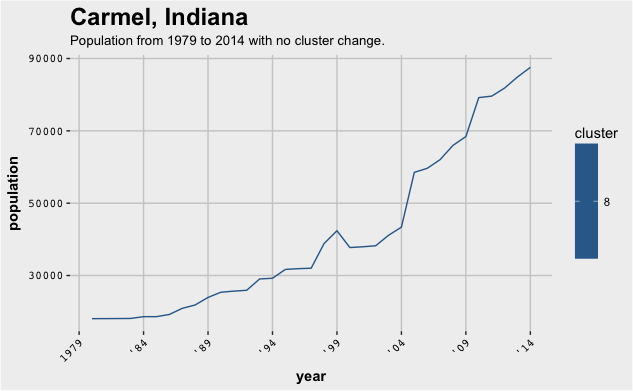
A plot to represent each cluster’s average crime reported over time for decreasing cities.

**Case Study:**

In addition to finding the necessary classifications of crime patterns, I took a look at a couple different case studies of states. By looking at a couple of different states, our previous knowledge about good crime and bad crime will be put to practice. The first case study is a look at Indiana crime. We begin this study by taking a look at cities whose *AYPI* was strictly bounded between 5% and 10% growth. The following Indiana cities found themselves in this fold:

* Carmel, Indiana
* Fishers, Indiana

We can see how each of these cities looked visually in *Figure 5* by looking at the population vs. time graph colored by crime pattern.



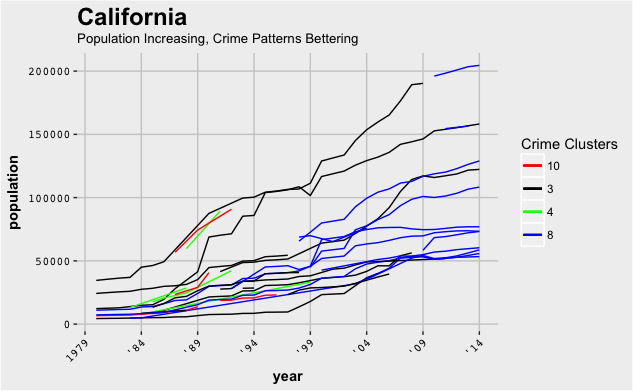
*Figure 5:*

Case study images showing the cities Carmel and Fishers as their population changes over time with respect to their cluster assignments.

As one can see, the crime in Indiana for these large cities experiencing population growth is very healthy. As the population for both cities increases over time, we see no change in their crime pattern. Also, we can note from *Figure 3* that this is only theft crime in low volumes. This case study was an example of a healthy population increase with little complications. Another case study endeavored was that of 12 different cities in California, the cities consist of:

* Brentwood
* Perris
* Rocklin
* Folsom
* Palmdale
* Cathedral City
* Victorville
* Lake Elsinore
* Roseville
* Temecula

By looking at *Figure 6*, one can doily note that there is several cities whose crime changed from better to worse. The clearest reality from this visualization is the change of clusters from around 1999 to 2014. Though the figure is not entirely complete due to the data set provided, we can see that the crime pattern got way better after 1999 to minimal volumes of crime and better crime clusters. This is a great feature to see in a city as the population grows.



**Conclusion**

With many different interesting points and conclusions that can be drawn from a project involving time series crime data, the one of many I was able to explore was detecting why a city was classified the way it was by K-means. The project at large has been valuable in that the techniques implemented were on real data, real problems, and provided real solutions that could actually be used by the *FBI* or crime organization. This project has not only given me a great deal of insight into crime data specifically, but also what data analysis, data mining, and predictive modeling looks like when the project is truly applicable to a specific domain. I can now see how helpful scientific and statistical techniques are used in this type of problem solving. Additionally, by having an extremely knowledgeable faculty advisor to help answer questions and steer me in the right direction along the way, I was able to reach much farther than I could have on my own. Having such direct contact with a data scientist for my questions and concerns on this project, I have been able to also see how a data scientist thinks about problems and just how creative and robust the solutions can be. I have nothing but the utmost appreciation for my faculty advisor and the complexity of this type of data science driven project. In final remarks, the most valuable things I have taken away from this project include but are not limited to:

* Structure a data science project
* Think more like a scientist
* Designing a problem to meet interesting and beneficial conclusions
* Creating unique solutions to solve abstract problems
* Building up logical evidence to support analysis and conclusions
* Learning how to structure and ask meaningful scientific questions