Title: Detecting Crime Patterns from FBI Crime Records from 1979-2014

Author: Blake Conrad

Email: [bmconrad@iupui.edu](mailto:bmconrad@iupui.edu)

Project Advisor: Dr. Murat Dundar

My project is the analysis of K-means clustering on *FBI Crime Data from 1979-2014*. This data set has been provided by [*Murat Dundar*](http://cadkd.cs.iupui.edu/). This data set contains over 300,000 observations with 14 features (10 numeric and 4 categorical). In the analysis, my hope is to collectively sort and accrete short term goals to achieve my overall desired goal of predicting crime patterns in various cities over time and to detect healthy or unhealthy population growth or decay.

By applying K-means clustering to my data set, I want to make predictions of crime patterns based on the model. These predictions could include: finding unexpected patterns in the data, predicting healthy and unhealthy growth and decay, and grouping problematic cities and their potential future. Each of these individual goals leans on the necessary understanding of the data and a quality clustering algorithm to be applied to it. By manually assessing whether the K-means clustering algorithm makes sense, this will allow an understanding of both the data and the model being used. In the manual assessment of the clusters, I plan to take time to do exploratory analysis and potentially apply feature engineering on the data to increase feature importance by transforming a collection of less meaningful features individually into a more meaningful features. To confirm that the data clusters make sense, I need appropriate visualizations for a technical understanding of what the clusters are explaining to me and how accurate they are. This fine-tuned model will allow for potential cluster ranking, cluster analysis, or cluster predictions on cities. The analysis of what the data is explaining is a critical piece of this project. My hope is to spend enough time understanding the clusters, their meaning, and their purpose to find unintuitive patterns in the data. After obtaining reasonable clusters, I will apply that knowledge to the concept of ranking. With an understanding of cluster ranking (like a heat map), I will be able to model cities over time. With a clear visualization of a city over time, I will have the opportunity to see the trend and direction of any particular city and the analysis of the entire population. My end goal is to have a presentation of my analysis of the data, how I determined my K-means clustering algorithm and its effectiveness on my data set, how the ranking system was built, and ultimately how well the ranking system predicted the growth and decay of particular cities over time based on the ranking system I hope to create. This will be a power point presentation with slides showing my progress along each step of the way. I will also be presenting my interesting relevant and irrelevant findings from the research in the data set applied by the listed techniques and process above. If the ranking system fails, a backup plan will consist of understanding the limitations and expectations of K-means clustering, such that the clusters make sense on the data.

A timeline to approach this issue will be as follows: five weeks will be dedicated to understanding the data, hypothesis testing, and visualization to create a persuasive K-means that clusters well and makes sense, the next four week I plan on implementing the ranking system that will effectively determine favorable and less favorable crime patterns, the following week I will investigate particular instances of cities and temporal crime patterns amongst them with visualizations to explain the findings, and the last two weeks I will dedicate to preparing meaningful and analysis driven presentation via PowerPoint and my final thesis paper to present at the end of the semester. The timeline leaves the last week of the semester as a buffer in case something happens along the way where I will need the extra time, which will be used to help prepare my formal presentation, paper, or plots if unutilized throughout the semester. This project is meaningful in that it is both scalable and impactful. This project could scale to a large back end data driven engine to support the police force in crime stationing and detection, which would have a direct impact in each of our lives on a daily basis.