K-Means Clustering and Its Use in Determining Crime Hotspots

David D Berberena

Bellevue University

DSC 500 Introduction to Data Science

Nasheb Ismaily

October 8, 2023

K-Means Clustering and Its Use in Determining Crime Hotspots

K-means clustering methods at their most basic definition provide the user with the ability to group data points that are similar to each other within a given dataset into clusters, the number of which can be specified by the variable *k*. There are many use cases for K-means clustering in society today, one of them being its use in determining local crime hotspots. Crime happens virtually everywhere in people’s minds, yet data can suggest that there are certain areas within a city, a county, a state, or even a country that are considered “hotspots” for crime. K-means clustering can be used to evaluate a dataset outlining the exact location of a crime as well as the type of crime committed to see what areas are prone to high levels of crimes, and even more so see what types of crime are rampant in a certain area. Data collected by local police departments can be compiled and uploaded to a central database that can cluster similar crimes and crimes committed in similar areas.

Potential analysis results could be that k-means clustering models show that a burglary hotspot exists in Long Island, NY (this is pure speculation, no data has been collected for the validity of this statement or any further statements) or that human trafficking cases are extremely high in Waukesha, WI. Depending on the data collected and the scale of k-means clustering, varied results will be displayed via models and graphs inferred from the data. It is difficult to identify city-sized crime hotspots in Utah if the models generated from k-means clustering show a large crime cluster covering the entirety of the Midwestern United States. Keeping data points relevant to the scope of where the user is trying to identify crime hotspots is a key part of validating the use of k-means clustering. If a dataset includes four cities listed as having data on multiple categories of criminal incidents, organizing the clusters into types of crime would organize the data, then breaking those clusters up even more by the four cities indicated would create multiple clusters that could identify the city with the most data points and the city with the highest ratio of crime-specific data points. The clusters created would be able to tell someone which city has the most crime relative to the other cities, and which city has the highest instances of a specific crime relative to the other categories seen in that city.

***Conclusion:*** K-means clustering has the ability to see further into data that may at first seem all over the place. By grouping the data points into clusters and modeling those data points graphically, we can make inferences that could potentially warn people of the severity of crime in an area, allowing the data viewer to make a decision on whether they wish to move to the area as someone looking to move, or if an established resident may decide to move away from that area.