# Clustering Crime in Chicago

Steven Hoang

January 26, 2020

## 1 Introduction

Today's law enforcement officers are equipped with various technologies that enable faster response times to criminal incidents. Modern advancements in weaponry provide officers with more optimized combative capabilities to negate threats in these incidents. The technologies and weaponry utilized, while effective, were accompanied by a total cost of 1.5 billion USD from 2010-2017. Preventive measures to forecast criminal incidents provide agencies with pattern recognition capabilities in environments that crime is most likely to occur. In an experiment conduced by Jain et al (2017), a k-means clustering algorithm is used with crime data obtained from an online database to identify geospatial areas with high probability of criminal incidents occurring. In this study, I utilize a k-means algorithm with a dataset obtained from the Chicago Data Portal and venue data Foursquare to create clusters segmenting Chicago communities by venues present. Knowledge which combinations of venues conducive to crime could help city government officials take proactive action in regulating these infrastructure combinations to reduce environments where crime is most likely to occur.

## 2 Materials and Methods

### 2.1 K-Means Clustering

K-means clustering is an unsupervised learning method used to partition n number of clusters into groups with individuals exhibiting similar characteristics. Data is divided into non-overlapping subsets, minimizing intra-cluster distances and maximizing inter-cluster distances. Venue data scraped from Foursquare is added to a k-means clustering algorithm to create 5 clusters segmenting the different communities of Chicago.

### 2.2 Data

The Crimes dataset recording incidents from 2001-present (minus the most recent seven days) downloaded from the Chicago Data Portal is visualized to show which clusters have the most criminal incidents.

# 3 Results

The data was partitioned into 5 groups: Cluster 0, Cluster 1, Cluster 2, Cluster 3, and Cluster 4. After executing the clustering algorithm, we saw that Cluster 1 had the most incidents out of the the rest of the clusters with 393 incidents. 90 incidents were grouped into Cluster 0, 1 into Cluster 2, 1 into Cluster 3, and 2 into Cluster 4. Upon further analysis, the 1st Most Common Venue with the highest frequency was Fast Food Restaurants.

# 4 Discussion

There are multiple opportunities to extend this experiment. Other variables could be analyzed including venue customer demographics. Venue ratings could also be analyzed to see if there is any correlation with number of crimes.

# 5 Conclusion

Based on the findings of this experiment, communities with high numbers of fast food venues are more susceptible to crimes. While correlation does not guarantee causation, city leadership could attempt to reduce the likelihood of crime occurring by limiting the amount of fast food restaurants allowed to do business in each community.