An Analysis of Violent Crimes in the City of Chicago from the years 2001 to 2023

> Zak Kastl, Mario Ruoff, Julissa Hutchison-Ybarra

Abstract

Introduction

In every society throughout history, criminal behavior has hindered the normal operation of that society. Webster defines *crime* as 'an illegal act for which someone can be punished by the government'[1]. When a crime is committed, all of society suffers, and the individual cost can vary wildly with the type of crime committed. An 2010 analysis of the various costs of crime by McCollister, French, and Fang[2] estimates the cost of a single murder, including tangible and non-tangible costs, at around 9 million dollars. In the City of Chicago, there were 494 murders in 2023 as of October 14. [3] This would put the total cost for citizens of Chicago at around $4.4 billion dollars. Similarly, the cost of a rape/sexual assault this year is around three million dollars.

While the costs due to crime can be very large, often police budgets are reduced in times of austerity. The has been reported in various news articles[[X]] with the ... It is also known that crime affects a city unevenly, with incidents of crime being more likely in particular areas of a city over other parts. It makes sense that police presence should reflect the distribution of crimes in the city.

In Machine Learning, the technique of clustering divides points of data into different clusters or categories. Additional points of data can then be classified as being part of one of those categories. However, We believe there is more that we can visualize in regards to the clustering of crimes. To this end, we have developed a Python-based web application to examine and analyze the distribution of crime in the city of Chicago. The application uses multiple forms of clustering algorithms to examine the distribution of different type of crime and to provide clusters in order to examine the distribution of crimes in relation to police stations.

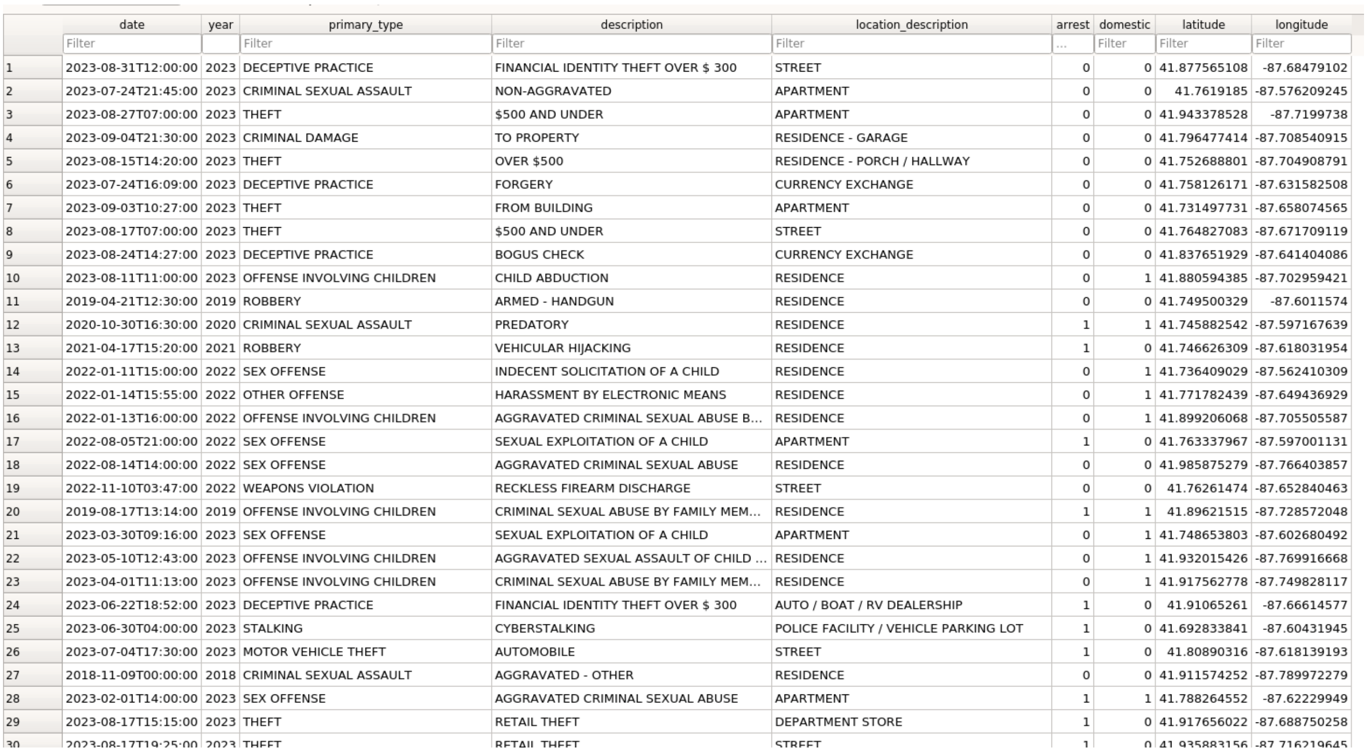
This paper will examine three different clustering algorithms that are currently available and plot the cluster centers as a Google Maps overlay. It will allow an examination of clustering as a function of a particular crime type and will compare those clustering with the true locations of police stations within the city.

Prior Work

Put prior work stuff here

Chicago Crime Dataset

For a thorough examination of crime, we are utilizing the Chicago Crime Dataset[3]. This dataset, provided by the city of Chicago, is a record of all reported crimes committed in the city from the year 2001 until the present day. The dataset provides anonymized crime statistics, including the primary crime type, location, and the latitude and longitude (partially anonymized) of the crime. It also provides whether the crime resulted in an arrest or if it is classified as a domestic. We decided to use this dataset due to its exceptional quality and robustness.



	date	year	primary_type	description	location_description	arrest	domestic	latitude	longitude
	Filter		Filter	Filter	Filter	...	Filter	Filter	Filter
1	2023-08-31T12:00:00	2023	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	STREET	0	0	41.877565108	-87.68479102
2	2023-07-24T21:45:00	2023	CRIMINAL SEXUAL ASSAULT	NON-AGGRAVATED	APARTMENT	0	0	41.7619185	-87.576209245
3	2023-08-27T07:00:00	2023	THEFT	$500 AND UNDER	APARTMENT	0	0	41.943378528	-87.7199738
4	2023-09-04T21:30:00	2023	CRIMINAL DAMAGE	TO PROPERTY	RESIDENCE - GARAGE	0	0	41.796477414	-87.708540915
5	2023-08-15T14:20:00	2023	THEFT	OVER $500	RESIDENCE - PORCH / HALLWAY	0	0	41.752688801	-87.704908791
6	2023-07-24T16:09:00	2023	DECEPTIVE PRACTICE	FORGERY	CURRENCY EXCHANGE	0	0	41.758126171	-87.631582508
7	2023-09-03T10:27:00	2023	THEFT	FROM BUILDING	APARTMENT	0	0	41.731497731	-87.658074565
8	2023-08-17T07:00:00	2023	THEFT	$500 AND UNDER	STREET	0	0	41.764827083	-87.671709119
9	2023-08-24T14:27:00	2023	DECEPTIVE PRACTICE	BOGUS CHECK	CURRENCY EXCHANGE	0	0	41.837651929	-87.641404086
10	2023-08-11T11:00:00	2023	OFFENSE INVOLVING CHILDREN	CHILD ABDUCTION	RESIDENCE	0	1	41.880594385	-87.702959421
11	2019-04-21T12:30:00	2019	ROBBERY	ARMED - HANDGUN	RESIDENCE	0	0	41.749500329	-87.6011574
12	2020-10-30T16:30:00	2020	CRIMINAL SEXUAL ASSAULT	PREDATORY	RESIDENCE	1	1	41.745882542	-87.597167639
13	2021-04-17T15:20:00	2021	ROBBERY	VEHICULAR HIJACKING	RESIDENCE	1	0	41.746626309	-87.618031954
14	2022-01-11T15:00:00	2022	SEX OFFENSE	INDECENT SOLICITATION OF A CHILD	RESIDENCE	0	1	41.736409029	-87.562410309
15	2022-01-14T15:55:00	2022	OTHER OFFENSE	HARASSMENT BY ELECTRONIC MEANS	RESIDENCE	0	1	41.771782439	-87.649436929
16	2022-01-13T16:00:00	2022	OFFENSE INVOLVING CHILDREN	AGGRAVATED CRIMINAL SEXUAL ABUSE B...	RESIDENCE	0	1	41.899206068	-87.705505587
17	2022-08-05T21:00:00	2022	SEX OFFENSE	SEXUAL EXPLOITATION OF A CHILD	APARTMENT	1	0	41.763337967	-87.597001131
18	2022-08-14T14:00:00	2022	SEX OFFENSE	AGGRAVATED CRIMINAL SEXUAL ABUSE	RESIDENCE	0	0	41.985875279	-87.766403857
19	2022-11-10T03:47:00	2022	WEAPONS VIOLATION	RECKLESS FIREARM DISCHARGE	STREET	0	0	41.76261474	-87.652840463
20	2019-08-17T13:14:00	2019	OFFENSE INVOLVING CHILDREN	CRIMINAL SEXUAL ABUSE BY FAMILY MEM...	RESIDENCE	1	1	41.89621515	-87.728572048
21	2023-03-30T09:16:00	2023	SEX OFFENSE	SEXUAL EXPLOITATION OF A CHILD	APARTMENT	0	1	41.748653803	-87.602680492
22	2023-05-10T12:43:00	2023	OFFENSE INVOLVING CHILDREN	AGGRAVATED SEXUAL ASSAULT OF CHILD ...	RESIDENCE	0	1	41.932015426	-87.769916668
23	2023-04-01T11:13:00	2023	OFFENSE INVOLVING CHILDREN	CRIMINAL SEXUAL ABUSE BY FAMILY MEM...	RESIDENCE	0	1	41.917562778	-87.749828117
24	2023-06-22T18:52:00	2023	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	AUTO / BOAT / RV DEALERSHIP	1	0	41.91065261	-87.66614577
25	2023-06-30T04:00:00	2023	STALKING	CYBERSTALKING	POLICE FACILITY / VEHICLE PARKING LOT	1	0	41.692833841	-87.60431945
26	2023-07-04T17:30:00	2023	MOTOR VEHICLE THEFT	AUTOMOBILE	STREET	1	0	41.80890316	-87.618139193
27	2018-11-09T00:00:00	2018	CRIMINAL SEXUAL ASSAULT	AGGRAVATED - OTHER	RESIDENCE	0	0	41.911574252	-87.789972279
28	2023-02-01T14:00:00	2023	SEX OFFENSE	AGGRAVATED CRIMINAL SEXUAL ABUSE	APARTMENT	1	1	41.788264552	-87.62229949
29	2023-08-17T15:15:00	2023	THEFT	RETAIL THEFT	DEPARTMENT STORE	1	0	41.917656022	-87.688750258
30	2023-08-17T19:25:00	2023	THEFT	RETAIL THEFT	STREET	1	0	41.935883156	-87.716219645

The only preprocessing work we had to perform on this dataset was to reduce certain dimensions that were unnecessary or redundant. Additionally, we chose to only examine crime statistics on a single-year basis. This is partially due to the sheer size of the record list on the application, but also to isolate the changing trends in demographics, affluence, and structure that may affect results beyond simple police station location. In the end, we reduced the dataset down to the following dimensions:

- Datetime of the crime committed
- Year crime was committed (seemingly redundant, but allows for easier retrieval of the data by year).
- Crime's Primary Type - One of 33 different primary crime types that the user can filter on.
- Description - A description of the crime committed.
- Location Description - A short description of the location where the crime was committed: e.g. 'Street' or 'Residence'.
- Was an arrest made?
- Is this considered a domestic crime?
- Latitude/Longitude where the crime was committed.

Clustering Algorithms

For this application, we have implemented three separate clustering algorithms: **KMeans Clustering**, **DBSCAN**, and **Spectral Clustering**.

KMeans

This clustering algorithm, first formally proposed by Lloyd in 1982 5, is the defacto standard for clustering algorithms in machine learning. K-Means is easy to implement, although computationally hard through a technique known as Lloyd's algorithm. In Lloyd's algorithm, a number of entries in the dataset, K, are selected at random from the entire set. Then, a distance metric, often the Euclidean or Manhattan distance metrics are applied to each point in the data to each of the cluster centers selected earlier. The point is "assigned" to the cluster with the shortest distance. After each point is assigned, the centers of each cluster are moved to the mean of the points in the cluster. Then the process is repeated until either the centers no longer move or a specified number of iterations have passed.

LLoyd's Algorithm (in a Python pseudo-code style)

```
def LloydsAlgorithm(data, k, max_iterations):

    # Select k rows from the data
    centers = [select random k rows for row in data]
    clusters = []

    while i < max_iterations:
        # Assign a row to the closest center
        for row in data:
            clusters[center].add(row if euclidian_distance(row) < all_other_centers)

        # Update the centers to the mean of the points in the cluster
        for cluster in clusters:
            centers[cluster] = mean(cluster[cluster])

        # Repeat until the centers no longer move
        if centers not Move():
            break
```

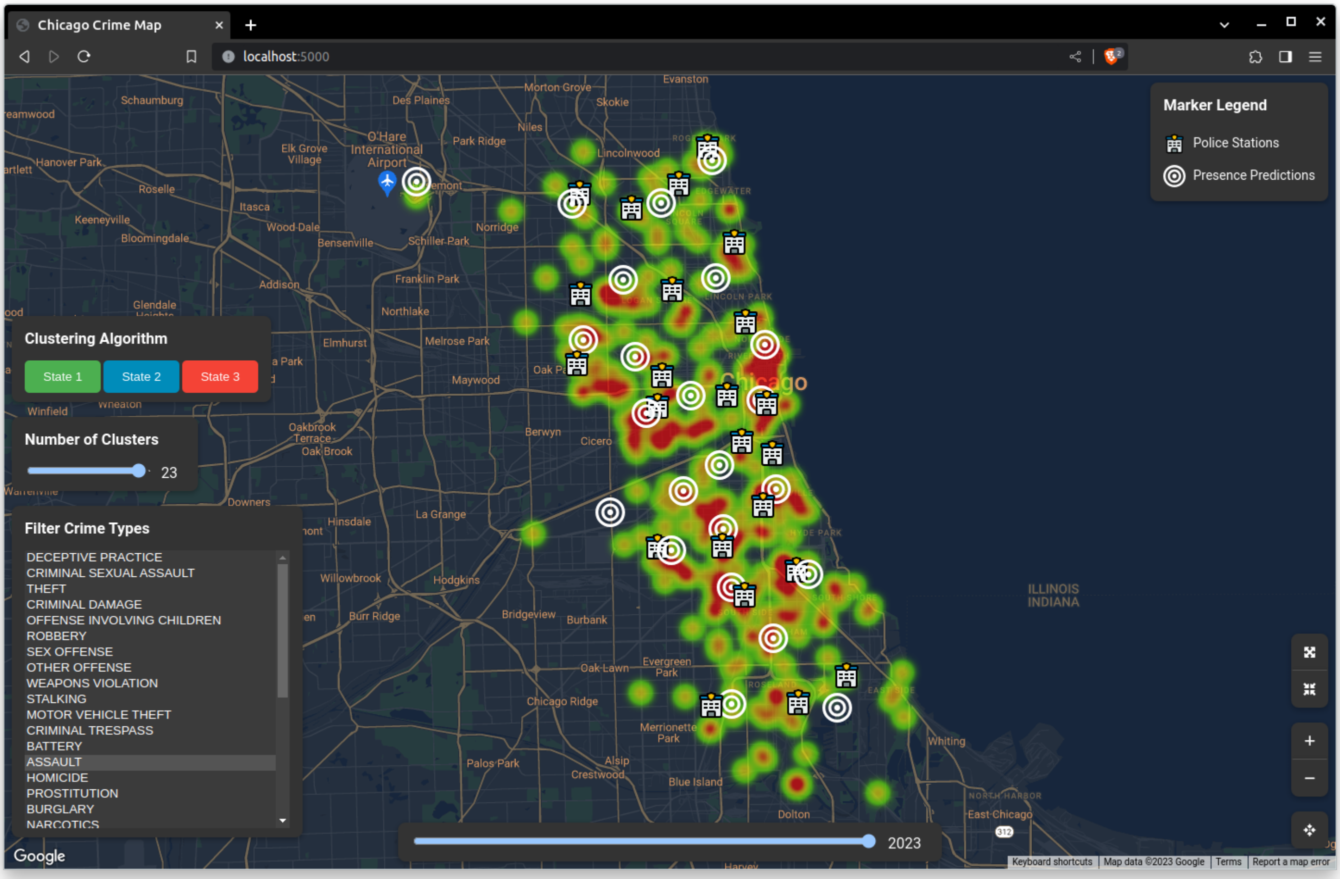
[Figure 1: Pseudocode for Lloyd's algorithm]

The primary advantage with this algorithm is its simplicity of implementation along with its wide application. However, it is an NP-hard problem in higher dimensions, and other clustering algorithms have outpaced it in performance and it is not guaranteed to find the optimum distribution. Furthermore, the initial cluster centers and the value of **K** used affect it greatly. Still, it is an exceptionally effective algorithm and worthy of consideration of clustering technique.

Density-based Spatial Clustering of Applications with Noise (DBSCAN)

Spectral Clustering

Application



[Figure 2 - A screenshot of the application]

Figure 2 shows an overview of our application. It is based on the Flask web framework written in the Python programming language. The application is designed to run locally on the user's computer and connects to a SQLite3 database. This database was converted from the raw Tab-Separated Value (TSV) file provided by the City of Chicago [3]. The database was not reduced in any way from its original format. We chose the TSV version of the data due to issues importing the data from the CSV file provided by the City of Chicago website. In order to reduce the sheer quantity of data that could be pulled from the database to the application, our application only queries the database for latitudes and longitudes of crimes that fit the criteria requested. This data is pulled from a separate view of the database that only contains the dimensions mentioned in section X.

The colored heatmap represents the quantity of crimes committed in an area. Red values indicate a greater number of crimes committed in that area over orange, yellow, and green areas.

Clustering Techniques

On the left side of the application, the user has the option to run one of the three clustering techniques on the data.

Analysis

Write something here about how the distributions are different based on the clustering technique used.

Conclusions and Future Work

References

1. https://www.merriam-webster.com/dictionary/crime
2. https://doi.org/10.1016/j.drugalcdep.2009.12.002
3. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2
4. https://heinonline.org/HOL/P?h=hein.journals/policejl58&i=122
5. https://cs.nyu.edu/~roweis/csc2515-2006/readings/lloyd57.pdf