Data Mining Project: Chicago Crime

Abstract—**In order to assist the Chicago Police Department lower crime rate in the most violent areas of Chicago, we employ supervised and unsupervised learning techniques to analyze crime events and possible arrests that may have followed. We also aim to suggest measures of crime prevention and prediction, and maybe shed some light on why Chicago has such high crime rates, compared to the rest of the US by analyzing certain socioeconomic factors.**

I. Introduction

Crime in Chicago has been a major topic of study in criminological studies and data science, as the city's overall crime rate, especially the violent crime rate, is higher than the US average [1]. The reasons for the higher numbers in Chicago remain unclear. Our main source of data is the Chicago Crime Dataset of the Google Cloud Platform [2], which reflects reported incidents of crime that occurred in the City of Chicago from 2001 to present. Detailed description of the dataset is presented in VI.

We were hired as data mining specialists by the Chicago Police Department, to extract unusual knowledge and hidden patterns about crimes in the city of Chicago, in order to suggest new measures and tactics that the force can apply, so that crime rate can become lower.

Crime theory suggests that crime is not random, it is either planned or opportunistic, crime happens when the activity space of a victim or target intersects with the activity space of an offender. So, by analyzing crimes, arrest rates and their respective locations, we can find which areas are underprotected (and thus which are overprotected) so that the Police force can aim its resources more intelligently.

First subject of our analysis is an overview of the crime situation in Chicago. Chicago is divided into 77 different community areas and 22 police districts. We will attempt to give a detailed description of crime and arrest rates depending on the different community areas and districts, as this will give the police a general idea of the crime taking place throughout the state.

Next, is is organized retail crime, and what are the arrest rates for each store type. Organized retail crime (ORC) is a serious subject, ORC costs retailers $777,877 per $1 billion in sales — an all-time survey high while Chicago landed 4th in the top cities where ORC occurs, on a 2018 survey by the National Retail Federation [3].

Next, correlating weapons used per crime with arrest rates, may give us some insights on how the police handles the effects of the gun policy in Chicago. This led us to a really useful analysis into Chicago's homicides.

Consequently, we will attempt to give an explanation of our findings from socioeconomic perspective. We will try to explain why crime happens in certain districts, by understanding the population living there. We aim to achieve that by combining data for some socioeconomic indicators [4] and population [5] for the city of Chicago.

Last but not least, we will use clustering methods to group crime incidents from a spatial perspective. We strongly believe that a real-time implementation of this approach will prove to be very useful for the Police, because clusters of crime incidents are due to happen, and its spatial characteristics help the Police to make better judgement calls.

We aim to achieve the subjects mentioned above by employing supervised learning (e.g. Decision Trees), unsupervised learning (e.g. k-means clustering) and basic exploratory analysis techniques. Finding location and/or primary types of patterns of whether an arrest happened or not, we can lead the force and fight crime more effectively.

II. Methodology

In this paper, the methodology part includes all knowledge discovery and data mining stages from the beginning of data extraction from GCP's database, data processing, transformation, and data mining, to the analysis and evaluation stages.

The first part of every data mining project, is data cleaning, in order to bring the data to a form appropriate for analysing. After that, an exploratory data analysis is appropriate in order to summarize the main characteristics of the data. Primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. This is a crucial task as it tells us a lot on how to proceed. After that, we decided that in order to explain our data by classifying them we need an algorithm that is rule based, and thus chose the Decision Trees Classifier by the scikit-learn library [6].

A decision tree is a flowchart-like structure in which each internal node represents a ”test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. This helped us find hidden patterns on our data, that we proceeded to extract using basic data analysis techniques and enabled us to present them with useful figures and plots. After every useful piece of information we extract, we try to support our argument by searching related work and referencing when appropriate.

Last part of our analysis consists of spatial clustering of the crime data. We achieved this by using the KMeans algorithm (Scikit-learn library again). Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). Clustering is an unsupervized learning technique. More specific, in the k-means algorithm, the sum of the squares of the Euclidean distances of data points to their closest representative, is is used to quantify the objective function of the clustering, so k-means is a representative-based clustering algorithm.

An important hyperparameter of the k-means algorithm that needs to be determined is the number of clusters. As there is no straightforward way to determine that, we used the ’elbow method’, which is merely a heuristic. The method consists of plotting the SSE (or explained variation, distortion etc.) as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

On figure 1 it is evident that the elbow is located where the number of clusters is 3. Because we found 3 clusters to be too few, we settled for 4 clusters for our dataset.

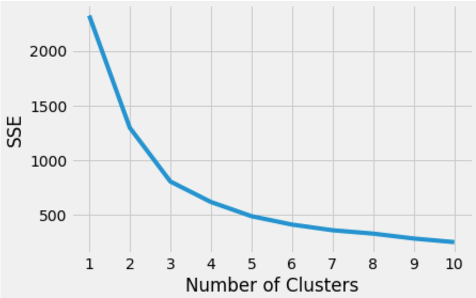

Figure 1. SSE vs number of clusters.

III. Related Work

Though it is difficult to find relevant work that fits our paper completely as it is a combination of different research subjects, this study by *Papachristos et. al* [7] examines the spatial patterning of violent crime in Chicago to determine whether or not all neighborhoods experienced decreases in violence. Also, this work [8] by *De Nadai et al.* is a study between different cities which explores how crime is related not only to socio-economic factors but also to the built environmental (e.g. land use) and mobility characteristics of neighbourhoods.

This paper [9] examines the effect of street lights on crime, by estimating the effect of nearly 300,000 street light outages in Chicago neighborhoods on crime.

Other related work: Robberies in Chicago [10], evidence on relationships among urban green space, violence, and crime in the US [11].

IV. Results

We will begin our analysis with an overview of crime on various districts of Chicago. On plot 2 we have a barplot showing the arrest rates per police district of Chicago, along with the mean arrest rate represented by the dashed line. This will prove useful later on our analysis, because we divide our locations by police districts, so we need to have an idea of the arrest rates distribution.

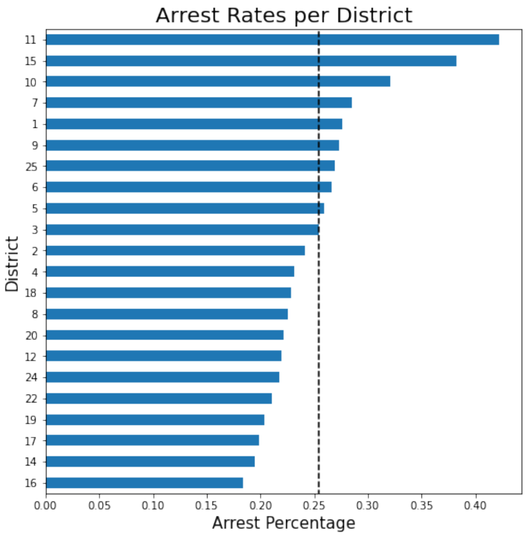

Figure 2. Arrest Rates per District.

As it is obvious from the plot, Police District 11 leads the way with the highest percentage of arrests. For a person living in Chicago this comes as no suprise, as the 11^{th} district is notorious for its high crime and violence incidents, as is evident from the local news, *”The 11th District on the West Side has had more fatal shootings this year (2020) than in all of that and other big cities.”* [12]. But for a non-US person reading this article, it helps build a perspective for the situation in Chicago. Throughout the years, crime in Chicago has shown a decline, as seen on plot 3 characterized as *The Great Crime Decline* by *Papachristos et. al* [7].

Despite this substantial decline, the number of murders occurring in neighborhoods at the top of the homicide distribution relative to the rest of the city increased. We refer to homicides because, a homicide, which is a statistically

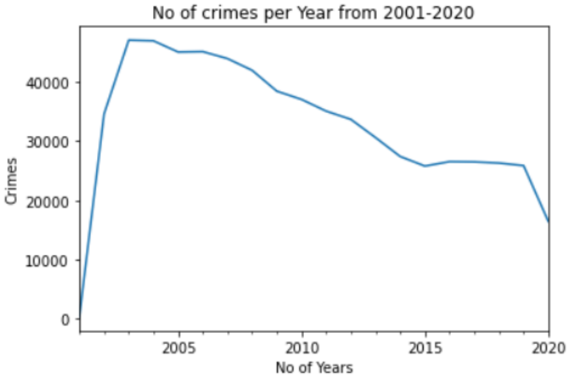

Figure 3. Crime in Chicago on 2001-2020.

rare event, tends to be the benchmark for determining crime trends in the United States because of the accuracy of its measurement over time and across jurisdictions. This overall pattern of decline, however, applies to every other major category of crime.

We will focus on 4 specific districts, two of relative arrest rates (and thus crime rate), districts 11 and 10, and two of relative low arrest rates, 14 and 17.

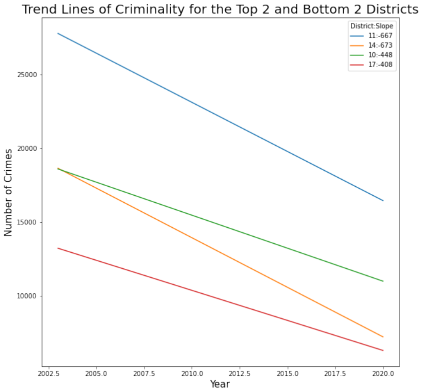

Figure 4. Trendline for crime on districts: 10,11,14,17.

As seen on plot 4, all of the four districts have a declining trendline throughout the years. On plot 5 it is evident that districts with high crime rates, have higher homicide rates, *despite the overall crime decline*. The outcome of this analysis is that while the drop in violent crime is shared between low and high crime areas alike, there remain areas of the city where violent crime rates are high, thus implying that underlying socioeconomic factors of the population, such as poverty, education etc., may be the reason for the wide crime gap between the safest and highest-crime neighborhoods.

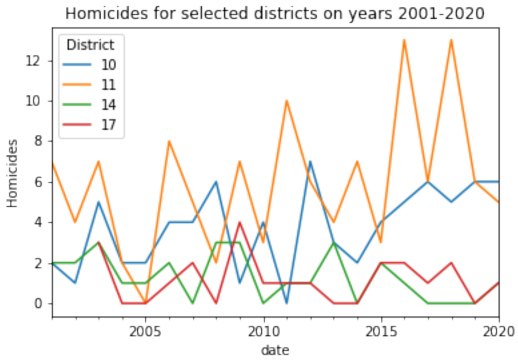

Figure 5. Homicides on districts: 10,11,14,17.

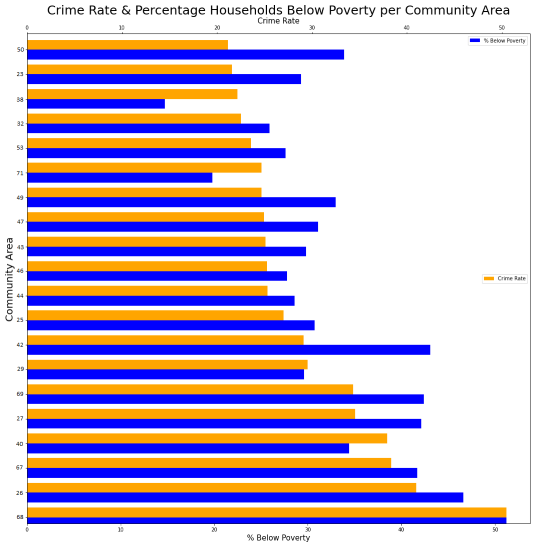

Figure 6. Crime Rate - Households below poverty per com. area.

On figure 6 we have a barplot of crime rate and % percentage of households below poverty, for every community area of Chicago[1]. There is a strong correlation between poverty and crime rate. Crime rates were calculated as the ratio of number of incidents to the population of every community area. The community areas with the highest crime rate and relatively high percentage of households below poverty, are: 68, 26 and 67. Community areas 67 and 68 belong to the 7^{th} Police District while community area 26 belongs to the 11^{th} Police District. These districts have the highest crime rates in Chicago. That's why it's important to show a some socioeconomic indicators about these three community areas, to compare with Chicago's corresponding average.

Community Area Number	Community Area Name	Households Below Poverty(%)	Aged 16+ Unemployed(%)	Aged 25+ Without High School Diploma	Per Capita Income	Hardship Index
68	Englewood	46.6	28.0	28.5	11888	94.0
26	West Garfield Park	41.7	25.8	24.5	10934	92.0
67	West Englewood	34.4	35.9	26.3	11317	89.0
Mean	-	21.8	15.4	20.3	25563	49.5

Table I
SOCIOECONOMIC INDICATORS FOR CHICAGO'S SELECTED COMMUNITY AREAS.

[1] Community Areas are different than Districts, we will mention their corresponding districts.

As someone can easily deduce from table I, community stricken community areas tend to be poorer than average, as indicated by socioeconomic indicators such as percent of households below poverty, per capita income. They also have a higher hardship index. The hardship index provides a more complete, multidimensional measure of community socioeconomic conditions than individual measures such as income or employment alone. A community with a high hardship index score has worse social and/or economic conditions than a community with a low or medium hardship score.

The 6 Community areas with the lowest hardship index, share borders, and are all located to the Northeast side of Chicago. The areas with the highest hardship index scores are mostly clustered together in the South and Southwest Chicago region [13].

Beyond the division of the city into Districts and Community Areas, there is another traditional division, instigated by the Chicago River. The river, which runs through the whole city, splits naturally Chicago into the North, West and South sides.

Side of Chicago	Mean Hardship Index	Mean Crime Rate(%)	Mean Arrest Rate(%)
North	27.64	8.4	20.5
West	69.62	18.2	28.9
South	55.23	18.6	23.5

Table II
Socioeconomic indicators for North, South and West Chicago

It is evident from table II, that this is not merely a geographical division. The North Chicago Side has a lower mean hardship index as well as a lower crime rate compared to the other 2 sides. In fact, the West and South sides have more than double the crime rate and mean hardship index. A staggering statistic, is the 30 year difference in life expectancy a New York University School of Medicine analysis found between 2 Community Areas, that lay opposite the Chicago River, with mean life expectancy in the Englewood community being just 60 years old [14]. Locals of the South and West Chicago areas, seem to recognize this disparity, and point out that North Chicago residents actively avoid the rest of the city, fearing the high crime rate, and generally worse living conditions [15].

Another facet of the forenamed disparity, is that racial segregation also follows that same geographical division. Until the 1950s, when the civil rights movement started, african american residents of Chicago were segregated to the South and West sides, with the North side Community Areas beign at least 90% white. Seventy years on, even though country wide advances in racial equality were definitely made, the segregation problem of Chicago has not been adequately addressed. This in turn, only amplifies the difference in hardship, job security and crime rate between the different sides of Chicago [16].

As it has already been hinted at, poverty and crime seem to be strongly interconnected. Undoubtedly, where crime rates are high it is natural for arrest rates to be accordingly high as a measure to decrease the amount of criminal activity. Having said that, this doesn't address the underlying problem, which is poverty and general hardship, driven by, at least in part, racial segregation. Furthermore, it is found that investors in Chicago tend to be interested mainly in low crime North Side areas, only furthering the vicious cycle of poverty and crime in the rest of the city [17].

Now, we will shift our focus on arrest regarding different locations. By harnessing the classification power of decision trees, using the Scikit learn library [6], we found an imbalance on arrest rates between different types of stores. As seen on image 7, a crime commited at a small retail store leads to an arrest at 25.5% of the times, while at a department store, it leads to an arrest at 55.1% of the times.

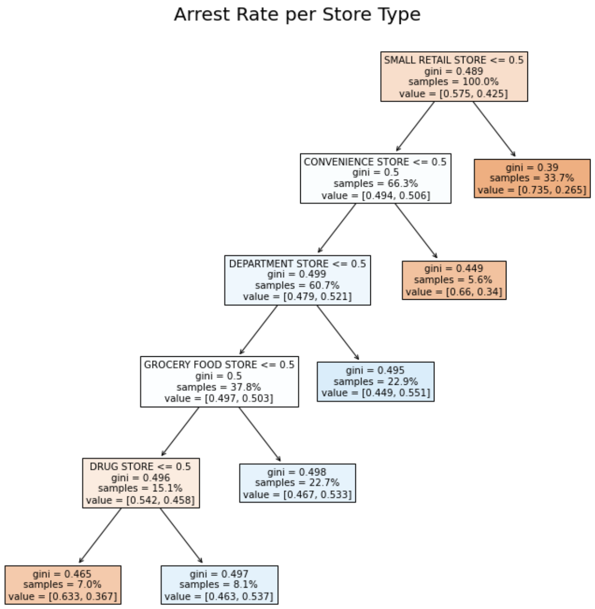

Figure 7. Arrest Rates per Store Type.

This is a perfect example of underprotection in small department stores, and can be explained because that kind of stores are not always part of big shopping centres, thus their security is more limited. Another fact that reinforces this argument is that convenience stores have a lower arrest rate of 34%, relative to big department stores. Convenience stores are often popular targets for a variety of crimes, most notably shoplifting and robbery, and it's not unsual for some areas of the US, for clerks to be working behind bulletproof glass windows, even during daylight hours. The reasons for the higher rate of crime at convenience stores may be attributable to various factors:

- including the small number of employees per store makes it difficult to stop or deter criminals,
- the extended hours of many convenience stores present more opportunities when few customers and/or witnesses are present,
- the smaller size of the stores makes it easy for criminals to quickly navigate the floor plan and enter and exit close to the cash registers and the majority of purchases are in cash as opposed to electronic transactions,
- leading to a relatively large amount of cash (often minimally secured) at any point.

A major problem in Chicago and the US in general, is gun violence. According to [18], US landed first in gun-related homicide and suicide rates in high-income OECD[2] countries, in a 2010 study. The fact that US has more than 6 deaths per 100.000 population, stems from its permissive firearm guiding policy, while the ownership and regulation of guns are among the most widely debated issues in the country.

Another finding that originated by the rules of the decision trees, was a short tree showing arrest rates regarding the weapon type that accompanied an assault or battery incident, as shown on figure 8. As one can infer from the decision tree rules, gun related crimes lead to an arrest only 18.9% of the times. This is a tremendously low arrest rate for a crime involving guns, that invited us to research furthermore chicago's arrest rate regarding serious crimes, such as homicides.

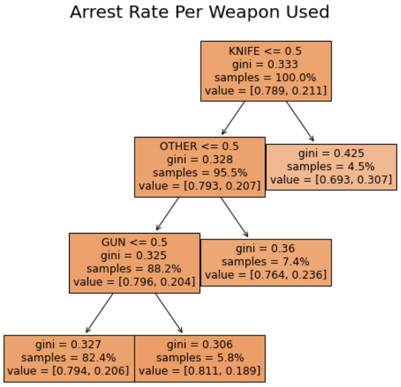

Figure 8. Arrest Rates by weapon type.

What we found was that, while on figure 5 homicides tend to increase throughout the years in specific districts, on figure 9 we see that the overall homicides have an increasing trend, with a 20-year mean of 7.35%, while arrest rates, as seen on figure 10, tend to decrease.

These numbers are alarming to the Police of Chicago, most homicide acts remain unpunished, meaning the Police don't get to catch the perpetrator or don't have enough incriminating evidence to lead to an arrest.

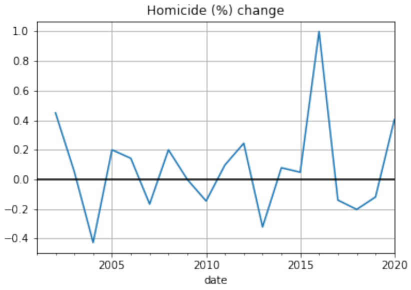

Figure 9. Homicide per cent change per year.

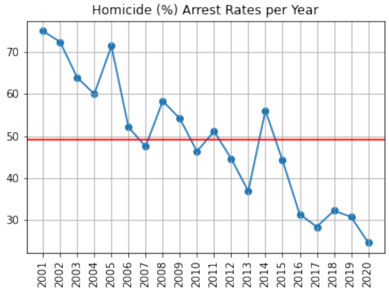

Figure 10. Homicides Arrest Rates. Horizontal Line is the average arrest rate.

The last subject we will present, is the use of unsupervised learning in crime analysis. The use of spatial clustering could prove to be really useful in future endeavors of the Police if applied in real time, as the distribution of crime is know to have a spatial dimension; that is, a map of point locations of crime often reveals spatial patterns, or clusters.

This phenomenon is partly explained by the fact that population is not homogenously distributed over space (i.e., neither the density of population, or characteristics like age, income, etc.) – it is expected that crime, too, will not be homogeneously distributed. Other explanations, like proximity to bars or night clubs, may also have a spatial aspect.

The presence of clusters, or hot-spots in this case, will provide the Chicago Police Department with useful information that will lead to better resources coordination, meaning, send reinforcements where appropriate.

As mentioned in the Methodology section, we used the k-means algorithm with a number of clusters of 4. As the use of this technique aims to apply at real time, we only used a subset of our data, the crimes that happened in the range between the dates October 1^{st},2020 and December $3rd$, 2020. The findings of our cluster analysis are shown in figure 11. Obviously, the clusters are overlapping at times, but one would expect that, as a crime is characterized by multiple features.

[2] Organisation for Economic Co-operation and Development

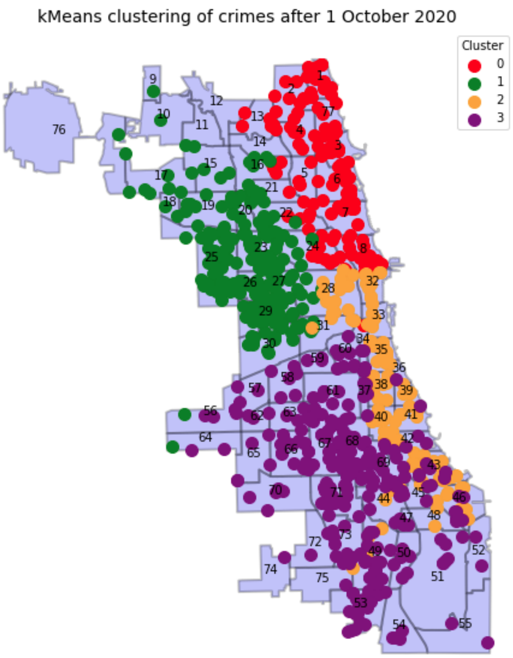

Figure 11. Clustering of crimes that took place in 1/10/2020 - 3/12/2020.

V. Conclusion

We will now summarize our findings in a more compact way, in order to provide useful feedback and suggestions to the Police of Chicago. Some of our conclusions were anticipated, while some others were not, but both confirming something we would think it may be true by combining different data sources and finding new and unprecedented knowledge are useful.

- Chicago Districts 11,15,10,7 and 1 are the top five districts with highest arrest rates, with district 11 leading the way both in arrest and crime rates, as supported by local newspapers and further investigation.
- In the last 20 years, the city of Chicago underwent a great decline in crime rates, as investigated by previous studies, that affecter nearly all community areas in Chicago, but the high-crime communities had the most to gain from the so-called *Great Crime Decline*.
- While nearly all of Chicago's neighbourhoods experienced tremendous declines in crime in that 20-year period, it also saw an average rise on homicides of 7.35%, while the homicide arrest percentage is in a declining trend throughout the same time period. This is an unusual outcome, as someone would expect a rise in crime to lead in an analogous rise in arrests.
- We found that crime-stricken communities, are highly associated with socioeconomic factors, such as the percentage of households below poverty, or the per capita income. This comes as no surprise but it's really important to correlate the wide gap between low and high crime areas with the wealth inequality between these communities.
- We realized that organized retail crime causes a huge loss in the retail industry, and that the bigger the type of department store, the higher the chances that the crime will need to an arrest, thus causing an underprotection problem in smaller stores such as convenience stores and drug stores.
- With gun violence being a big issue in the US, we found that gun related crimes lead to an arrest only 18.9% of the times, while knife related crimes for example have a 30.7% arrest rate.
- Finally, we found that real time clustering of crimes can prove to be really useful to the police force, as clusters are prone to emerge due to the non homogeneity of the population.

We will recommend the city of Chicago with some hopefully useful measures, with the sole purpose of fighting crime more effectively thus leading to more safety in the streets of Chicago.

- Gun Violence is a major problem in the US, the government should really reconsider their gun policy, and pass gun laws that actually reduce gun violence. Investments in smart gun technology could be useful too, we can set up our phones so you can't unlock it unless you've got the right fingerprint, so why not do the same thing for guns?
- Increase police patrols in underprotected areas and areas where organized retail crime takes place. It is necessary that we learn from the data.
- Address the overarching problem of racial segregation and economic despair certain regions of Chicago face. That could be done by encouraging the investment in such areas, providing more employment opportunities and job security for residents, which will in turn improve living conditions and further decrease crime rates.
- Implementation of real time clustering of the reported crime events will help the Police to better command and direct its resources.
- Ensure that all residents can pursue opportunities to thrive without fear of violence by providing prevention and intervention supports that are available to individuals of all ages and levels of risk.

References

[1] "Chicago is responsible for almost half of the increase in u.s. homicides," *Time magazine*. [Online]. Available: https://time.com/4497814/chicago-murder-rate-u-s-crime/

[2] "Chicago crime data - google cloud platform." [Online]. Available: https://console.cloud.google.com/marketplace/details/city-of-chicago-public-data/chicago-crime/

[3] "2018-organized retail crime survey." [Online]. Available: https://cdn.nrf.com/sites/default/files/2018-11/NRF_ORCS_IndustryResearch_2018_FINAL.pdf

[4] "Socioeconomic indicators." [Online]. Available: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

[5] "Community areas in chicago." [Online]. Available: https://en.wikipedia.org/wiki/Community_areas_in_Chicago

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] A. Papachristos, N. Brazil, and T. Cheng, "Understanding the crime gap: Violence and inequality in an american city: Understanding the crime gap," *City Community*, vol. 17, 12 2018.

[8] M. D. Nadai, Y. Xu, E. Letouzé, M. C. González, and B. Lepri, "Socio-economic, built environment, and mobility conditions associated with crime: A study of multiple cities," 2020.

[9] A. Chalfin, J. Kaplan, and M. LaForest, "Street light outages, public safety and crime displacement: Evidence from chicago," *Public Safety and Crime Displacement: Evidence from Chicago (January 27, 2020)*, 2020.

[10] W. Bernasco and R. Block, "Robberies in chicago: A block-level analysis of the influence of crime generators, crime attractors, and offender anchor points," *Journal of Research in Crime and Delinquency*, vol. 48, no. 1, pp. 33–57, 2011. [Online]. Available: https://doi.org/10.1177/0022427810384135

[11] B. K. Bogar S, "Green space, violence, and crime: A systematic review. trauma violence abuse."

[12] "Violence soars in minneapolis after floyd killing, but one chicago police district is even worse." [Online]. Available: https://tinyurl.com/ferhyxud

[13] "Economic hardship index shows stark inequality across chicago." [Online]. Available: https://greatcities.uic.edu/2016/09/19/economic-hardship-index-shows-stark-inequality-across-chicago/

[14] "Study: 30-year life expectancy gap in 2 chicago communities." [Online]. Available: https://tinyurl.com/422csbju

[15] "Dear north siders: Stop acting like chicago doesn't have a south side." [Online]. Available: https://tinyurl.com/3r3h3k6j

[16] "Separate, unequal, and ignored." [Online]. Available: https://tinyurl.com/t3c3ujnr

[17] "Neighborhood disparities in investment flows in chicago." [Online]. Available: https://www.urban.org/research/publication/neighborhood-disparities-investment-flows-chicago

[18] H. D. Grinshteyn E, "Violent death rates: The us compared with other high-income oecd countries," *Am J Med. 2016 Mar;129(3):266-73.*

VI. Appendix A: Dataset Description

The dataset is characterized by the following columns:

- **unique_key** (integer): Unique identifier for the record.
- **case_number** (string): The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
- **date**(timestamp): Date when the incident occurred (Best estimate).
- **block**(string): The partially redacted address where the incident occurred, placing it on the same block as the actual address.
- **IUCR** (string): The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description. See list on this link.
- **primary_type** (string): The primary description of the IUCR code.
- **description** (string): The secondary description of the IUCR code, a subcategory of the primary description.
- **location_description** (string): Description of the location where the incident occured.
- **arrest** (boolean): Indicates whether an arrest was made.
- **domestic** (boolean): Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
- **beat** (integer): Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at this link.
- **district** (integer): Indicates the police district where the incident occurred. See the districts at this link.
- **ward** (integer): The ward (City Council district) where the incident occurred. See the wards at this link.
- **x(or y)_coordinate** (float): The x(or y) coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
- **latitude (or longitude)** (float): The latitude (or longitude) of the location where the incident occured.
- **location** (string): The location of an event in a single format.
- **year** (integer): Year the incident occured.
- **updated_on** (timestamp): Date and time the record was last updated.
- **community_area** (integer): Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at this link.