Chicago Crime Analysis (Data Mining Semester Project)

[*]Updated 15/03/21

[redacted]

Abstract—**Most social phenomena are multi-faceted and complex problems that require years of experience to analyse. To this end there is an organised movement by government agencies to collect data that may be useful both for predicting and profoundly understanding these issues. One of the biggest and most common problems of societies is crime of all kinds. The vast volume of data available as well as the combinatorial nature of crimes make criminology appropriate for applying data mining techniques. In this paper societal, economical, temporal and spatial factors linked to delinquent behaviour were investigated to find possible emerging correlations between them.**

Index Terms—**Crime, Data-Mining, Chicago**

I. Introduction

In this effort of ours the main goal was to explore different factors and their relationship to criminal activity. The past couple years, data-driven techniques such as exploratory data analysis and data mining are proving to be useful tools in the field of criminology as they appear to shed some light on underlying, seemingly uncorrelated factors and improve both arrest rates and overall crime prevention. The introduction of such techniques has a twofold result; Firstly, it diminishes crime as seen in its declining trend and secondly, it led to effective management of police resources.

Knowledge discovery in databases (KDD) is the process of extracting meaningful information and finding patterns in a database [1]. Data Mining (DM) is a step in the KDD process which applies sophisticated algorithms to the database [1].

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

[2] The authors J. Agarwal, R. Nagpal, and R. Sehgal have conducted analysis on crime by employing K-Means clustering in order to extract useful information from a high volume dataset and interpret said information, assist police in identifying and analysing patterns and ultimately find ways to reduce said criminal activity. In this work, rapid miner tool was used which is an open source statistical and data mining suite written in Java.

[3] F. Mingchen, Z. Jiangbin, R. Jinchang, A. Hussain, L. Xiuxiu, X. Yue and L. Qiaoyuan analysed datasets on crime for Chicago, San Francisco and Philadelphia using the Big Data Analystics (BDA) approach. They also applied some deep learning models in order to make predictions.

Our approach employed combinatorial exploratory data analysis, time-series and data-mining techniques. Specifically, we provide insightful visualisations (graphs, kernel density estimation plots, animations) to help the reader quickly and easily understand the situation of crime as is. Also, with time-series decomposition techniques we managed to decouple its trend from the seasonal and the irregular component (which seems to be relatively small). Additionally, using CART [4] we successfully predicted whether an arrest was made given other attributes/features.

II. Data-Acquisition

Since our goal was to perform thorough research and the project is rather complex a combination of multiple sources of information was deemed necessary. These include multiple, rather heterogeneous, databases, spatial information in the form of maps and partitions of it (zip-code, community-area) and census data for demographic purposes.

The primary dataset is this of the Chicago Crime [5]. This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law

Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. The preliminary crime classifications may be changed at a later date based upon additional investigation and there is always the possibility of mechanical or human error.

The weather dataset was not straightforward to acquire since the concept of ”weather” is broad. For this analysis we decided to use only the temperature recorded in given area with an immediate result being that datasets containing other aspects of the weather phenomenon such as humidity, atmospheric pressure etc. were rendered useless. We ended up using data provided by National Oceanic and Atmospheric Administration and specifically from National Climatic Data Center [6] and the weather station which was taken into account was that of Chicago Midway airport. Since the datasets provided by NCDC are split by year we needed to compose a bash script to gather information from 2001-2020. Afterwards these separate datasets were combined using Pandas library to produce a file containing all temperature information.

Even though Chicago is one of the most diverse cities in the US an interesting phenomenon is observed as community areas tend to be racially homogeneous. The dataset containing information for the population in US by zip-code [7] will provide insight to this claim.

Another dataset which was explored to further enrich our research was this of public holiday data [8]. The results provided by this dataset were not significant so we decided not to present them in this paper.

A. Dataset Description

- Chicago Crime Dataset
  - **ID**: Unique identifier for the record.
  - **Case** Number: The Chicago Police Department Records Division Number,which is unique to the incident
  - **Date**: Date when the incident occurred.
  - **Block**: The partially redacted address where the incident occurred, placing it on the same block as the actual address.
  - **IUCR**: The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description
  - **Primary Type**: The primary description of the IUCR code
  - **Description**: The secondary description of the IUCR code, a subcategory of the primary description
  - **Location Description**: Description of the location where the incident occurred
  - **Arrest**: Indicates whether an arrest was made
  - **Domestic**: Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act
  - **Beat**: Shows the beat where the incident happened. It is the littlest police geographic territory – each beat has devoted police beat vehicle. Three to five whips make a police division, and three parts make up a police locale
  - **District**: Indicates the police district where the incident occurred
  - **Ward**: The ward (City Council district) where the incident occurred
  - **Community Area**: Indicates the community area where the incident occurred.
  - **FBI Code**: Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
  - **X Coordinate**: The x 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.
  - **Y Coordinate**: The 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.
  - **Updated On**: Date and time the record was last updated.
  - **Latitude**: The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block
  - **Longitude**: The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block
  - **Location**: The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block
- Weather Dataset
  - **Date**: Date when the record was created
  - **Temperature**: Average recorded temperature of this date.
- Public Holiday Dataset
  - **Date**: Date when the record was created
  - **normalizeHolidayName**: Normalized name of the holiday.
- US Population by ZIP Dataset
  - **zipCode**: Date when the record was created
  - **population**: Population of this segment.
  - **race**: Race category in Census data. If it's null, it's across all races.
  - **sex**: Male or female. If it's null, it's across both sexes.
  - **minAge**: Min of the age range. If it's null, it's across all ages.
  - **maxAge**: Max of the age range. If it's null, it's across all ages or the age range has no upper bound, e.g.

age ¿ 85.

III. Data Cleaning / Pre-processing

For the purposes of this project we excluded features that seemed to be too specific or did not convey extra information. Regrading the crime dataset we implemented the following steps:

1) Converted variables to categorical.
2) Removed missing values and other irregularities.
3) Date feature which previously contained date and time was split into Day,Month,Year,Weekday and Hour.
4) Constructed a new feature to express the Season which was extracted through the Month variable
5) Convert different types of " non – criminal ", "non-criminal ","non – criminal (subject specified)" to Non-Criminal as well as "criminal sexual assault" to "crim sex assault".
6) Kept "Description" and "Location Description" values that account for 90% of data and merged rare values to "Other".
7) Deleted records where location was either NaN or not in Chicago.

Regarding the weather dataset:

1) Changed Fahrenheit measurements to Celsius.
2) Deleted records with NaN values.

Regarding the public holiday dataset the only necessary pre-processing was to exclude records that were unrelated to the US. As for the dataset containing information for the population for each ZIP code what we needed to accomplish was to map zip codes to corresponding community areas, since this was the level of detail in our analysis. For this purpose we used a publicly available list providing this mapping thus transforming the dataset to our needs.

IV. Experimentation & Results

A. Exploratory Results

Through the process of Exploratory Data Analysis (EDA) we tried to thoroughly understand the data to figure out emerging patterns behaviours and possible irregularities.

As we implied before crime is trending downwards throughout the years. In the figure below we can further inspect this phenomenon through the time series decomposition. Each time series consists of four components.

- Trend: Trends represent the general tendency of the data to increase or decrease over time.
- Seasonality: Like seasons, the data patterns appear after regular intervals represent the seasonal component of time series. They repeat after specific intervals like day, week, month, year, etc.
- Cycle: They can be thought of like seasons with the only difference that cycles don't appear at regular intervals.
- Random Noise: The sudden changes in the time series data which do not fall under the above three categories and are also hard to explain otherwise are called random fluctuations or noise.

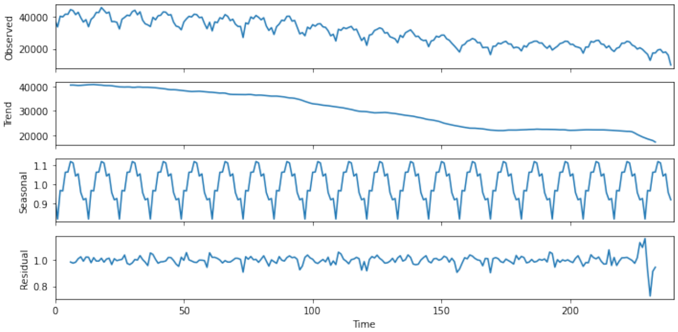

Figure 1. Time series decomposition (frequency 12 months)

As seen from the seasonal component of 1 there is strong seasonality which leads to the assumption that crime has a two-axis behaviour, one revolving around yearly trend and one around seasons. This will be investigated later on, but for now we can see that there are three important values in the seasonal component (0.9, 1.0, 1.1) meaning that there are months with steadily reduced or increased crime correspondingly.

To further analyse crime we plotted its behaviour into multiple sub-divisions of time to observe whether some interesting patterns emerge. This is shown in 2 where we can see that crime rate rapidly declines with a minimum at around 5 o'clock in the morning and then climbs back up with a maximum at around 11 o'clock in the morning. Regarding weekdays we can not distinguish any pattern as the behaviour seems to follow a uniform distribution. The same holds true for the third plot, with two irregularities; namely an unusual peak at the first day of the month and an unusual plunge in the last day of the month. We attribute this to the default value being the first day of the month when the day of the crime is unknown. As for the box-plot presented in the bottom right it demonstrates the monthly based crime statistics.The month of February seems the safest and is observed to have the least crime incidents followed by December and January whilst June-August are the least safe month with the maximum number of reported crime incidents. We found that monthly crime rate is very likely linked to local climates. Chicago has temperate continental climate with a cold winter and hot summer in June-August, significantly more outdoor activities can be found in summer than winter days thus the associated higher or lower crime incidents in different seasons.

This observation, that crime varies significantly with season, sparked our interest to enrich this hypothesis with weather data. It is rather obvious that temperature is not the only thing that changes between seasons but to keep this analysis simple and comprehensible we chose to solely rely on this. As seen in figure 3 there is a strong connection between the number of crimes recorded and the temperature. We notice a considerable rise between 20 and 25 degrees Celsius and a strong decline afterwards. Between 0 and 15 degrees the count of crimes is uniformly distributed.

In order to examine the connection between location and the types of crimes we made a heatmap 4 which densely presents

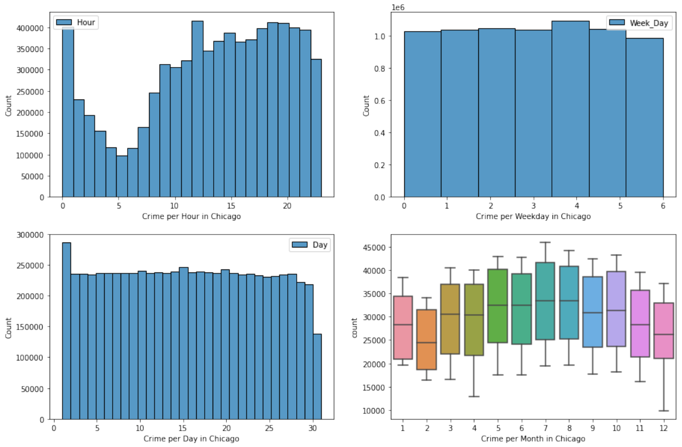

Figure 2. Crime behaviour by Hour, Weekday, Day, Month

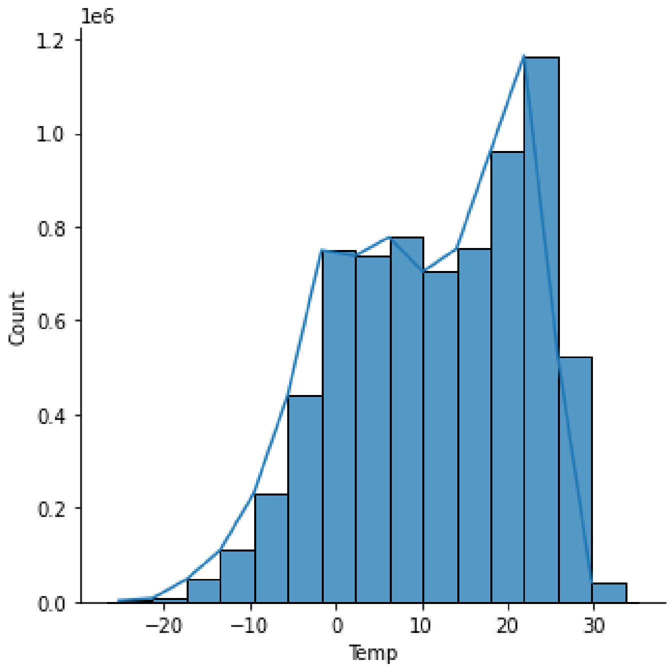

Figure 3. Histogram of temperature

how each one of the 10 most popular locations correlate to the 10 most popular types of crime. Looking at the heatmap one can make a distinction between two types of behaviour. The first being crimes that are uniformly distributed amongst all locations and vice versa and the second being crimes that show bigger concentrations in certain locations. In the case of assault for example it's clear that its numbers are somewhat uniformly distributed amongst all locations. Other crimes like theft and battery on the contrary have distinctly greater numbers in specific locations like street and apartment, residence, sidewalk and street respectively. The "other" category is an all encompassing group that includes many locations usually very specific places that didn't have enough entries in the dataset and thus weren't of sufficient statistical or interpretative value. The graph can also be looked at with a different perspective. When we focus on the columns instead of the lines we see that generally not many different crimes happen at each location and that only the opposite is true. One big exception is the "street" where all different types of crimes happen and which also happens to be the number one most frequent location. That is however not very surprising since the street is a very general term.

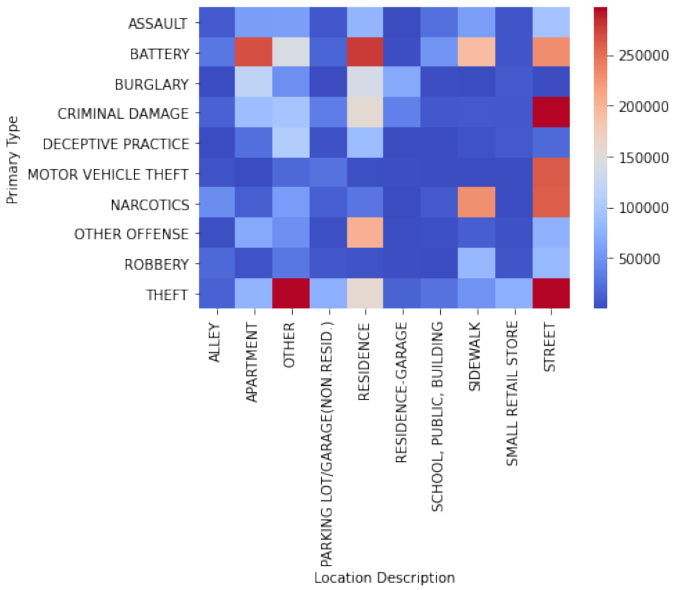

Figure 4. Heatmap of crime with regards to location

A rather demanding task is resource allocation for law enforcement. It would be rather inefficient to have uniformly distributed forces across the city since doesn't occur with the same density everywhere. To provide actionable information, we employed kernel density estimations for the most prominent crimes both in the form of a still image (mean densities per crime) 5 and GIF format crime densities per community area/hour with the extra dimension being the time of the day.

This geospatial data helps with the visualisation of crime hot-spots while also providing insightful information for strategic patrol placement. It is of utmost importance to note, that some areas seem to be more infamous regardless of the crime. These areas (in the detail of Community Areas) will be later on precisely identified and analysed in multi-faceted manner.

We used CART algorithm from Scikit [9] to predict whether there was an arrest for a specific crime given other attributes describing it. In order to reflect the natural relationships of features we utilised "cyclical feature engineering", in the sense that we transformed our features regarding day, hour, weekday and month to their sine, cosine counterparts. This is rather helpful as it helps encapsulate information as such: "Midnight (0'clock) is a close to 11 o'clock as it is to 1 o'clock".

We employed a 80-20% train-test split since the dataset is rather voluminous and the results are summarised in the table

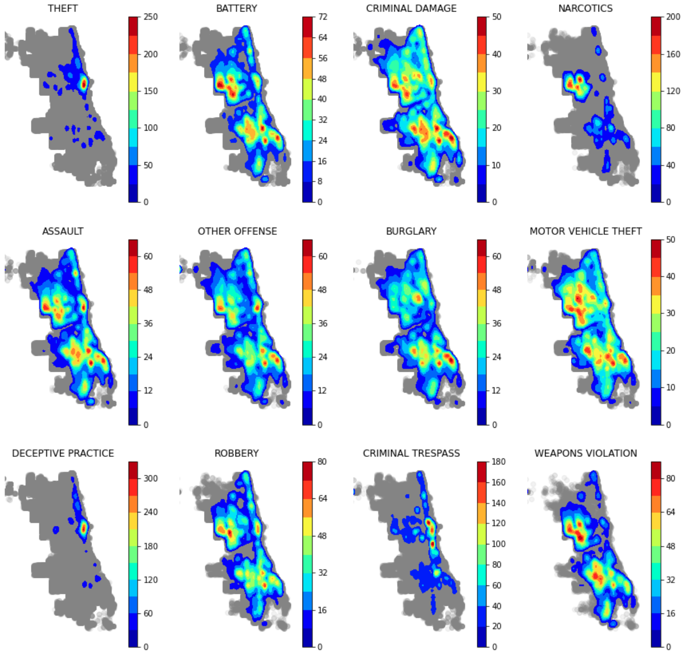

Figure 5. Crime densities per community area

below.

	precision	recall	f1-score	support
False	0.88	0.87	0.88	970884
True	0.66	0.69	0.67	356886
accuracy			0.82	1327770
macro avg	0.77	0.78	0.78	1327770
weighted avg	0.82	0.82	0.82	1327770

We observe that even a very simplistic classifier can achieve very good results (approx. 80% accuracy). A useful addition provided without extra cost through decision trees is feature importance. This will enlighten the reader with regards to how to the DT model is making its predictions. From the graph 6 we can see that the crime's primary type is the most importance feature to predict whether an arrest was made.

V. Social Analysis

In this section we will employ an analysis based on Census data which can be found here: Data regarding income, unemployment,education and demographics regarding populations in the area. Through our analysis we have deducted that the community areas mostly associated with crime are Austin (25), South Shore (43) and West Town (24). These areas' income can be seen in figure 7 where it clear that both Austin and South Shore are above average regrading population percentage with low income resulting in violent crimes such as assault, battery, theft, robbery and narcotics. On the other hand West Town seems to be a rather wealthy area as most of its population has above average income. Even though that is the case the crimes committed there are of different nature mostly vehicle thefts, burglaries and criminal damages.

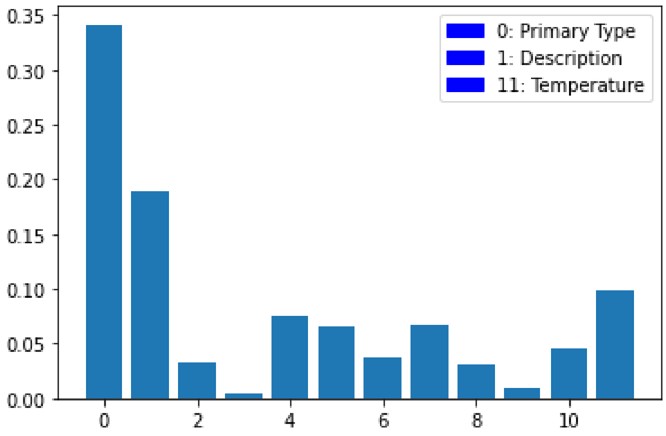

Figure 6. Decision tree entropy based feature importance

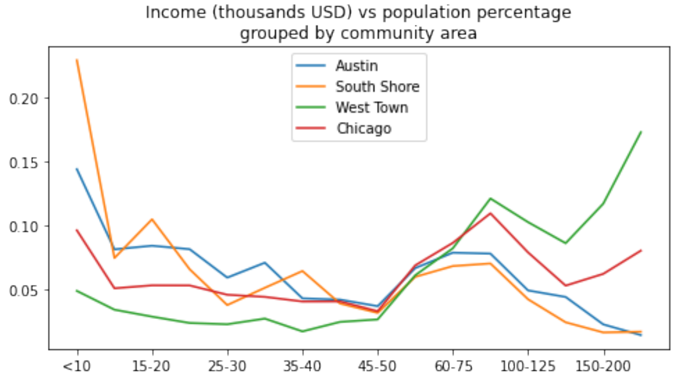

Figure 7. Chicago income per community area

So it is obvious that the wealth status of a community area can and will affect the type of crimes that take place. A similar picture can be depicted through figure 8 where the percentage of population was split by its education level for each community area. It has been suggested by sociological studies that the education level and crime rates are correlated to a significant degree [10]. Again, the picture for Austin and South Shore is troublesome as a large portion of their population have the minimum education type and also they are both below average in the percentage of people with BA or higher. In contrast, West Town, a community area where based on the education level of its population and their economic status we can only assume that is being targeted rather than that its population is the one that commits crimes.

By utilising the US-ZIP dataset we found the top-three race & ethnicity in each community area. For Austin, South Shore and West Town the populations are distributed as presented in Table I and Table II.

We can see that there are mostly black or African American populations in Austin and South Shore and white populations in West Town, which is the main argument weaponized by people with racist tendencies. It is rather important to take into consideration all factors and not jump to conclusions without

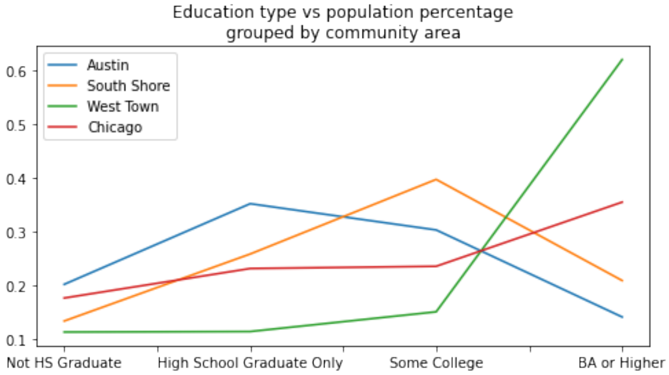

Figure 8. Education type per community area

	zip Code	Race/Ethnicity	Percentage
Austin	60624	BLACK OR AFRICAN AMERICAN	95.36
	60639	SOME OTHER RACE	40.32
	60644	BLACK OR AFRICAN AMERICAN	94.34
	60651	BLACK OR AFRICAN AMERICAN	64.01
	60707	WHITE	73.26
	60804	WHITE	52.01
South Shore	60619	BLACK OR AFRICAN AMERICAN	97.46
	60637	BLACK OR AFRICAN AMERICAN	77.95
	60649	BLACK OR AFRICAN AMERICAN	95.56
West Town	60610	WHITE	72.89
	60612	BLACK OR AFRICAN AMERICAN	61.36
	60622	WHITE	71.26
	60642	WHITE	67.68
	60647	WHITE	60.55
	60651	BLACK OR AFRICAN AMERICAN	64.01
	60654	WHITE	77.52
	60661	WHITE	63.28

Table I
CHICAGO PRIMARY RACE/ETHNICITY IN ZIP CODES FOR AUSTIN, SOUTH SHORE AND WEST TOWN

bearing all facts of the situation in mind.

A restriction posed by this dataset in general is that there is no information regarding the perpetrator. Were we to have this information available, then it would be rather interesting to visualise which community areas are the ones that are home to more criminals and which are the ones that are subjected to their crimes.

VI. ACKNOWLEDGEMENT

We would like to thank our professors for giving us the opportunity to work on such a stimulating project with real-life data which demonstrates the power of data mining.

REFERENCES

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth *et al.*, "Knowledge discovery and data mining: Towards a unifying framework." in *KDD*, vol. 96, 1996, pp. 82–88.

[2] J. Agarwal, R. Nagpal, and R. Sehgal, "Crime analysis using k-means clustering," *International Journal of Computer Applications*, vol. 83, no. 4, 2013.

[3] M. Feng, J. Zheng, J. Ren, A. Hussain, X. Li, Y. Xi, and Q. Liu, "Big data analytics and mining for effective visualization and trends forecasting of crime data," *IEEE Access*, vol. 7, pp. 106 111–106 123, 2019.

	zip Code	Race Ethnicity	Percentage
Austin	60624	WHITE	2.19
	60639	WHITE	37.15
	60644	WHITE	2.72
	60651	SOME OTHER RACE	17.19
	60707	SOME OTHER RACE	12.72
	60804	SOME OTHER RACE	39.3
South Shore	60619	TWO OR MORE RACES	1.26
	60637	WHITE	14.97
	60649	WHITE	1.18
West Town	60610	BLACK OR AFRICAN AMERICAN	17.44
	60612	WHITE	12.72
	60622	SOME OTHER RACE	13.27
	60642	SOME OTHER RACE	13.43
	60647	SOME OTHER RACE	24.67
	60651	SOME OTHER RACE	17.19
	60654	ASIAN	11.03
	60661	ASIAN	24.78

Table II
CHICAGO SECONDARY RACE/ETHNICITY IN ZIP CODES FOR AUSTIN, SOUTH SHORE AND WEST TOWN

[4] D. Steinberg and P. Colla, "Cart: classification and regression trees," *The top ten algorithms in data mining*, vol. 9, p. 179, 2009.

[5] C. P. Department, "Crimes - 2001 to present: City of chicago: Data portal," Mar 2021. [Online]. Available: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

[6] R. Rich Baldwin. [Online]. Available: https://www7.ncdc.noaa.gov/CDO/cdoselect.cmd?datasetabbv=GSOD

[7] "Us population by zip code." [Online]. Available: https://azure.microsoft.com/en-in/services/open-datasets/catalog/us-decennial-census-zip/

[8] "Public holidays." [Online]. Available: https://azure.microsoft.com/en-us/services/open-datasets/catalog/public-holidays/

[9] 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.

[10] L. Lochner and E. Moretti, "The effect of education on crime: Evidence from prison inmates, arrests, and self-reports," *American economic review*, vol. 94, no. 1, pp. 155–189, 2004.