Analysis and Visualization of Criminal Activity in Urban Illinois

A PROJECT REPORT

submitted by

Harshit Anand - 19BCB0071

Avish Aviraj Jha - 19BCE0812

Utpal Manishchandra Prajapati - 19BCE0759

Faculty: Prof. Jyotismita Chaki

CSE3020 - Data Visualization

School of Computer Science and Engineering

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

MAY 2022

Table of Contents

S. NO.	TITLE	PAGE NO.
	Acknowledgment	2
	Abstract	3
1.	Introduction	4
2.	Background Study	5
3.	Methodology Adapted	11
	3.1 Description of the Project	11
	3.2 Full Architecture	13
	3.3 Obtained Results and Analysis	15
	3.3.1 Visualization	15
	3.3.2 Prediction of arrest type using Decision Tree Classifier	24
4.	Conclusion	26
5.	References	27

ACKNOWLEDGEMENT

We would like to express our gratitude to all those who have helped us in the successful completion of this project. Without their support, we would not have been able to achieve the goal of completing our project successfully.

We would like to take this opportunity to thank our faculty, Dr. Jyotismita Chaki, for her constant support, guidance and mentorship without which it would have been extremely difficult to complete the project on time.

Finally, we would like to express our sincere gratitude to VIT University, which provided us with a platform to hone our skills over a period of three years.

Place : Vellore

Harshit Anand

Avish Aviraj Jha

Utpal Manishchandra Prajapati

Date : 25th April 2022

Abstract

Crime pattern analysis is related to public safety and helpful for police patrols. There are three important topics, which are crime type prediction, criminal behavior pattern analysis and hot-spot prediction, in the crime pattern analysis. Crime is an inseparable part of our society, either being a victim or an offender everybody has witnessed crimes. In our project, we analyzed the crimes data for which we selected the “Chicago Crime dataset report” which has the incidents of crimes for Chicago city that occurred from 2001 - present. We analyzed the trends of crime over the years, locations of the crimes where it happened the most, and hotspots of crimes over the years.

The Chicago Crime dataset with entries stretching back to 2001 is visualized and analyzed with multiple graphs and plots like barplot, line graph, scatterplot, 3D scatterplot and more for easy interpretation and trend analysis. Then the data is partitioned into train and test dataset and used to train DecisionTree Model to make predictions on whether arrests will be made or not. Its accuracy is measured and cross validated against the present data.

The prediction of arrest types is the problem in this project. In the data preprocessing, the features, which are hour, day of a week, business hour, and business day are extracted from date in the raw data. The column of output is descriptive words in the raw data and transformed into 4 categories. The 4 categories are home, public open space, public building, and others; however, the category of others is dropped before the data are fed into prediction models because this category provides no information about the crime location. Decision Tree is applied for the prediction of arrest types. Two important parameters, the maximum depth of each tree and the number of trees in a decision tree are optimized. Feature importance indicates the relationship between the output and these input features. Then, the decision tree is evaluated in terms of precision, recall, f1-score, and computation time.

1. Introduction

Safety is a highly concerned public topic, which is related to the crimes. Crime is an important problem in every city, especially in cities like Chicago. According to the crime data for Chicago in 2016, the total number of crimes reaches 264679, which means 725 crimes happen per day in Chicago approximately. Analysis of crime patterns in a city is beneficial for allocating patrol resources and residents improving their protection consciousness. In order to provide information about crime locations and alert police where to notice when they are patrolling, the prediction of crime location types is conducted in this project.

Some of the crimes are random and rarely happen in an area. It is hard to find latent patterns in the crime cases and predict future cases, which means the prediction could be not so accurate. However, it is great even if there is a minor help for decreasing the crimes since it may save or help some persons from danger. The field of crime prediction is similar to statistical inference, which is involved in a set of assumptions. However, incorrect assumptions could lead to incorrect conclusions and invalidate the inference such as incorrect assumptions in the Cox models. Traditional statistical methods analyze data for the population, race, surroundings, and education level in crime prevention, which is a time-consuming task. For the prediction problem in this project, it is a difficult task to provide a number of correct assumptions because of the complex relationship in the data set. On the contrary, machine learning is very powerful in solving such problems.

In this project, the problem about predicting arrest types is addressed. The crime location types are related to the primary crime types, for example, theft is more common in the residential areas than in public buildings. The community areas are important features for predicting crime location types because some community areas might be commercial areas with only public places. Community areas, primary crime type, and information related to time are selected as input features for the predictions. And the arrest types are separated into categories. Thus, the prediction of arrest types is a classification problem.

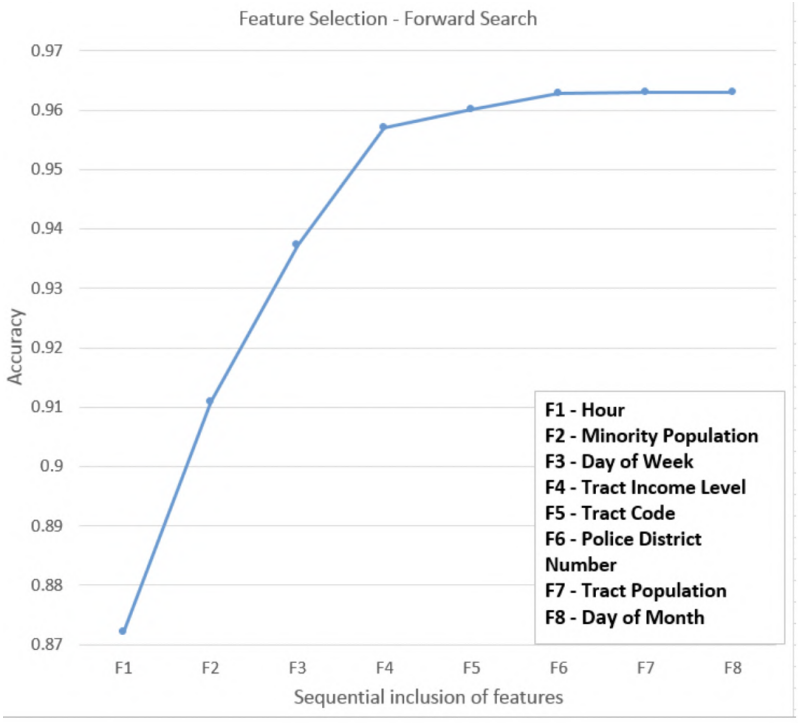
2. Background Study

2.1 Crime Prediction and Classification in San Francisco City

Author : Addarsh Chandrasekar, Abhilash Sunder Raj and Poorna Kumar

In this project, they analyze crime data from the city of San Francisco, drawn from a publicly available dataset. At the outset, the task is to predict which category of crime is most likely to occur given a time and place in San Francisco. To overcome the limitations imposed by our limited set of features, they enrich their data by adding information from the United States Census to it.

The initial problem of classifying 39 different crime categories was a challenging multi-class classification problem, and there was not enough predictability in their initial data-set to obtain very high accuracy on it. They found that a more meaningful approach was to collapse the crime categories into fewer, larger groups, in order to find structure in the data. They got high accuracy and precision on the blue-collar/white-collar crime classification problem using Gradient Boosted trees and Support Vector Machines.



Drawbacks: The classifier gave 30% accuracy on the training set and 25% accuracy on the cross-validation set. Hence, both training and cross-validation errors were very high. On performing cross-validation on the training data using 10 folds, they got a test error of 84%, which was huge compared to their training error, indicating high variance.

2.2 Machine Learning Applied To Crime Prediction

Author : Vaquero Barnadas, Miquel.

This project intends to provide an understandable explanation of what is it, what types are there and what it can be used for, as well as solve a real data classification problem (namely San Francisco crimes classification) using different algorithms, such as K-Nearest Neighbors, Parzen windows and Neural Networks, as an introduction to this field.

With the crime classification problem, it has been seen that the most accurate algorithm was the Artificial Neural Network. Although it is true that ANN was better than KNN, it did not outperform it. This might have been a model design problem, bad feature selection, poor training or bad adjustment, or a combination of all of them.

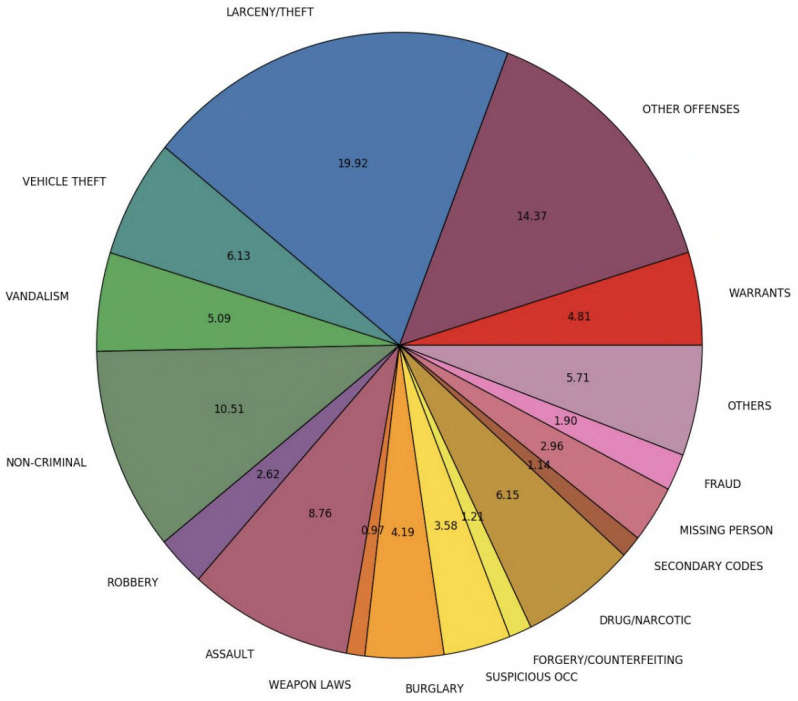


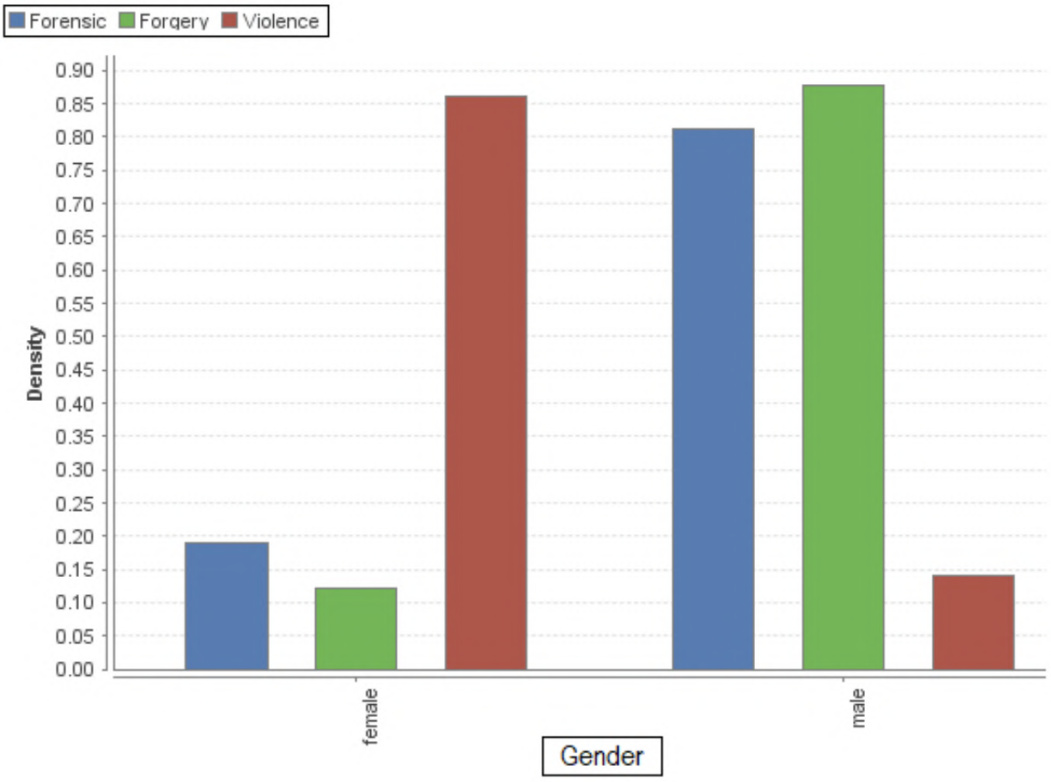
Figure 17. Prior probability of a sample to belong to each category

Drawbacks: With the crime classification problem, it has been seen that the most accurate algorithm was the Artificial Neural Network. Although it is true that ANN was better than KNN, it did not outperform it. This might have been a model design problem, bad feature selection, poor training or bad adjustment, or a combination of all of them. Probably, by adding more layers to the network or extending the training process, the results might have been better.

2.3 Mining Forensic Medicine Data for Crime Prediction

Author : A. Abdo , Hanan Fahmy , Amir Abobaker Shaker

In this research, a framework has been built to analyze forensic medicine data for crime prediction by applying data mining techniques (DMT). The proposed framework consists of six main phases. Data acquired from forensic medicine authority database (FMA DB) and flat files in Alexandria department, this department serves three governorates (Alexandria, Beheira and Mersa Matruh). This study aimed at giving recommendations to the Egyptian government to apply this work over forensic medicine authority in all Egyptian governorates. This work presents a new framework for crime prediction using data mining techniques based on real data from Egyptian forensic medicine authority data. Rapidminer software was used to analyze the collected data with acceptable accuracy (about 98%), the simulator provided an easy, simple use and real-time interface to crime prediction.



Drawbacks: Crime prediction accuracy is quite low. Need more data mining techniques to increase crime prediction accuracy. Need to use more datasets to train other classifiers, development of social networks to link crime with criminals, study their interrelationship, and development of online crime mapping to show crimes areas.

2.4 Designing efficient and balanced police patrol districts on an urban street network

Author : Huanfa Chen, T. Cheng, Xinyue Ye

They propose a street-network police districting problem (SNPDP) that explicitly uses streets as basic underlying units.This model defines the workload as a combination of different attributes and seeks an efficient and balanced design of districts. They also develop an efficient heuristic to generate high-quality districting plans in an acceptable time.

In this study, the SNPDP model is proposed. This is a novel approach to incorporating the street network structure and street-level predictive crime risk into the design of police districts. This model is multi-criteria-based, in that the objectives include theefficiency and balance of the district workload, and the workload is a combination of the crime risk, area size and district diameter.

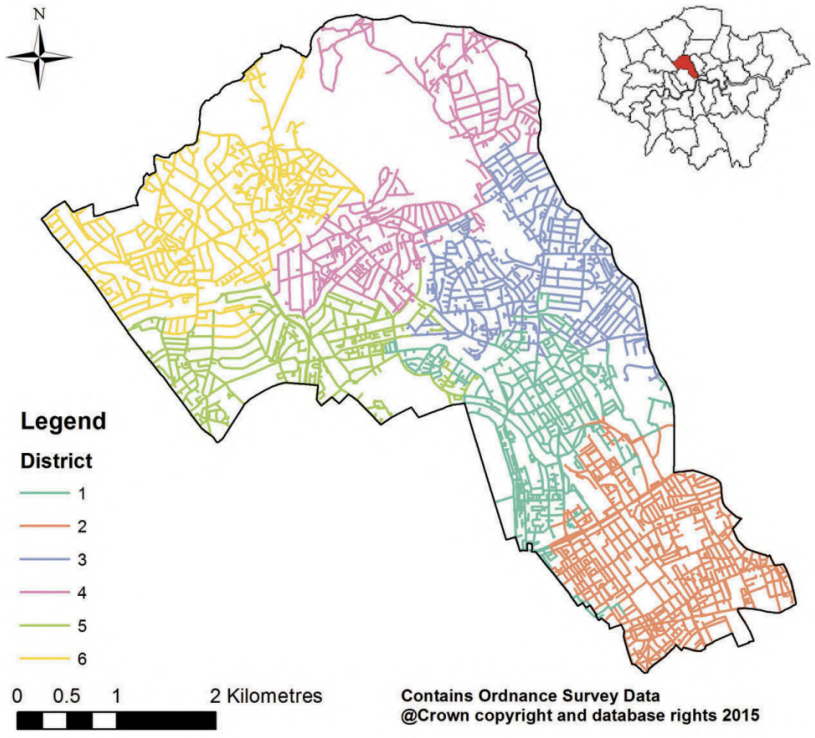


Figure. The best districting solution for Camden.

Table: A summary of the best solution using GP–TS

District	No. of streets	Area	Risk	Diameter	Workload
1	1121	0.192	0.152	0.394	0.246
2	1110	0.181	0.131	0.339	0.217
3	965	0.159	0.196	0.383	0.246
4	682	0.132	0.118	0.549	0.267
5	847	0.165	0.260	0.339	0.255
6	850	0.170	0.142	0.426	0.246

Drawbacks: As the street segments are incompatible with the current census units (e.g. census block, output area), it is difficult to carry out demographic research in a district consisting of streets. While most of the police activities take place on the street or near the street, there are situations in which police deviate from the street network, such as carrying out tasks in an area, such as a large green space, that has no streets.

2.5 Using Machine Learning Algorithms to Analyze Crime Data

Author : Lawrence McClendon, Natarajan Meghanathan

They observed the linear regression algorithm to be very effective and accurate in predicting the crime data based on the training set input for the three algorithms. The relatively poor performance of the Decision Stump algorithm could be attributed to a certain factor of randomness in the various crimes and the associated features (exhibits a low correlation coefficient among the three algorithms); the branches of the decision trees are more rigid and give accurate results only if the test set follows the pattern modeled. On the other hand, the linear regression algorithm could handle randomness in the test samples to a certain extent (without incurring too much prediction error).

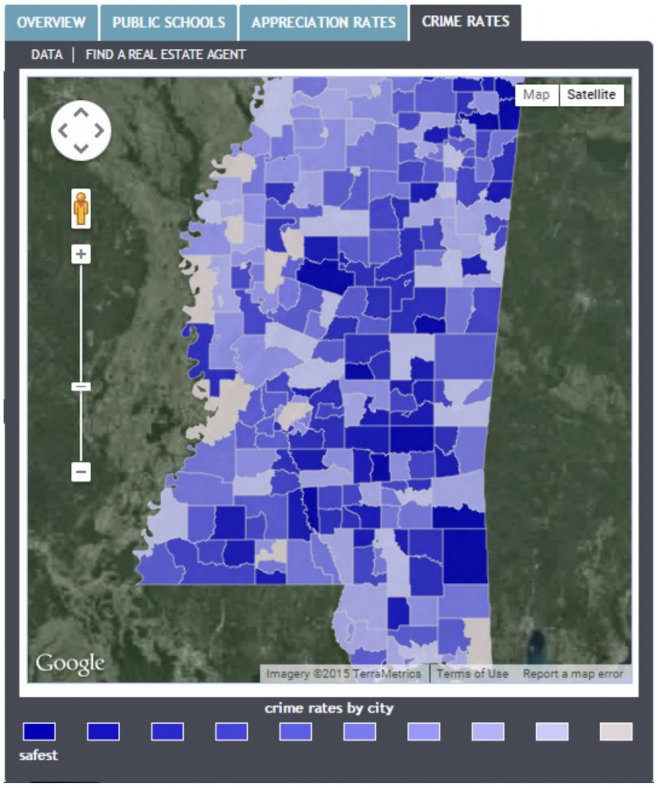


Figure 5: Crime Rates in Different Cities of Mississippi

Drawbacks : They observe the linear regression algorithm to be very effective and accurate in predicting the crime data based on the training set input for the three algorithms. The relatively poor performance of the Decision Stump algorithm could be attributed to a certain factor of randomness in the various crimes and the associated features (exhibits a low correlation coefficient among the three algorithms); the branches of the decision trees are more rigid and give accurate results only if the test set follows the pattern modeled.

3. Methodology Adapted

3.1 Description of the Project

The Chicago Crime Dataset is a collection of all crimes committed in Chicago and its districts from the year 2001 up until April 2022. This is maintained and updated regularly by the Chicago Police Department and is available to the public. 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. This contains over 7 million entries of crimes committed and recorded. This has record of every crime where each crime has been described by multiple attribute as follows:

- **ID -** Unique identifier for the record.
- **Case Number -** The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
- **Date -** Date when the incident occurred. This is sometimes a best estimate.
- **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 -** 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.

- **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. Chicago has 77 community areas.
- **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.
- **Year -** Year the incident occurred.
- **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.

Data is imported from the Chicago Crime Dataset and filtered for error lines. This is then indexed according to Date and used to plot multiple graphs depicting the trend in crime since 2001 based on day of the week, time, month, location and the trend each type of crime is following. Each plot gives an insight about crimes like at what time they happen the most and what location has witnessed multiple crimes. All of this information will give the law enforcement a better understanding of their demographic and they could modify their patrolling accordingly.

3.2 Full Architecture

We are using the Classification algorithm in this project to accomplish our task. As there are different types of crimes, we can classify them and perform operations accordingly to accomplish our objective of recognizing arrest types across the city based on geographical locations. Our first measure is dividing the entire city of Chicago into smaller units called cells, each district of Chicago is evaluated as a cell. In the meta-data obtained from the CLEAR system of Chicago Police Department, each criminal record is characterized by several attributes that includes crime description, location, longitudes and latitudes, etc as elaborated in Table 3.2.1.

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.

Table 3.2.1

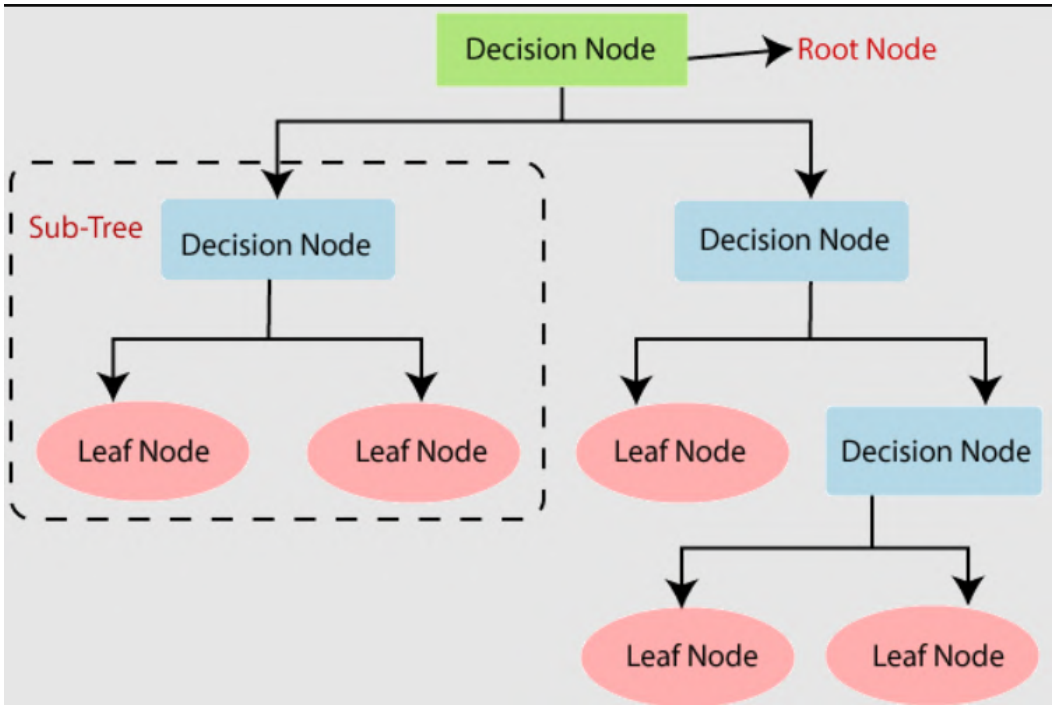
These attributes comprise the dataset of the system model adopted by our project and will be conducive while plotting the exact locations of the crimes. In addition, the CLEAR system classifies the crimes into 32 different categories as depicted in Table 3.2.2.

With all the attributes, we expect to depict the pattern of each crime-type across the City of Chicago for an entire year. For a sound prediction of the occurrence of a crime at any location and any hour of a day, it is required to consider the data that is consistent and out of exemptions. Therefore in order to abstain from false conjectures and guarantee a reliable prediction model, we plan on considering the criminal records of the past 14 years to train our algorithm. The derived prediction model is then tested against the records from recent years for validation and determining the accuracy rate of our model.

Theft	Battery
Robbery	Criminal Damage
Deceptive Practice	Narcotics
Domestic Violence	Non-Criminal (Subject Specified)
Assault	Criminal Trespass
Gambling	Arson
Burglary	Prostitution
Concealed Carry License Violation	Human Trafficking
Motor Vehicle Theft	Weapons Violation
Homicide	Offense involving Children
Crime Sexual Assault	Sex Offense
Obscenity	Non-Criminal
Liquor Law Violation	Interference with Public Officer
Kidnapping	Public Peace Violation
Intimidation	Stalking
Public Indecency	Ritualism

Table 3.2.2

The dataset is used to train a Decision Tree Classifier model with 9 tree depths for 'arrest' prediction and the same is cross validated to determine accuracy and precision.



Hardware Used:

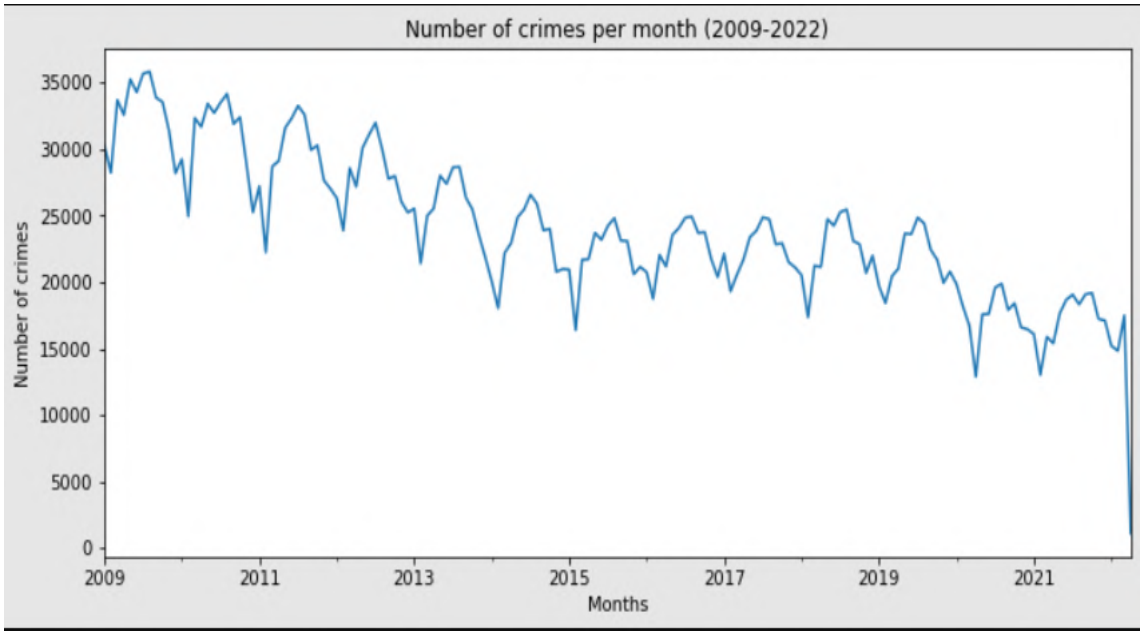
- ➢ Windows 11
- ➢ Ram : 16 GB
- ➢ Google Chrome

Software Used:

- ➢ Anaconda / Python 3.9
- ➢ Jupyter Notebook
- ➢ Libraries : pandas, numpy, matplotlib, seaborn, sklearn, wordcloud, pylab

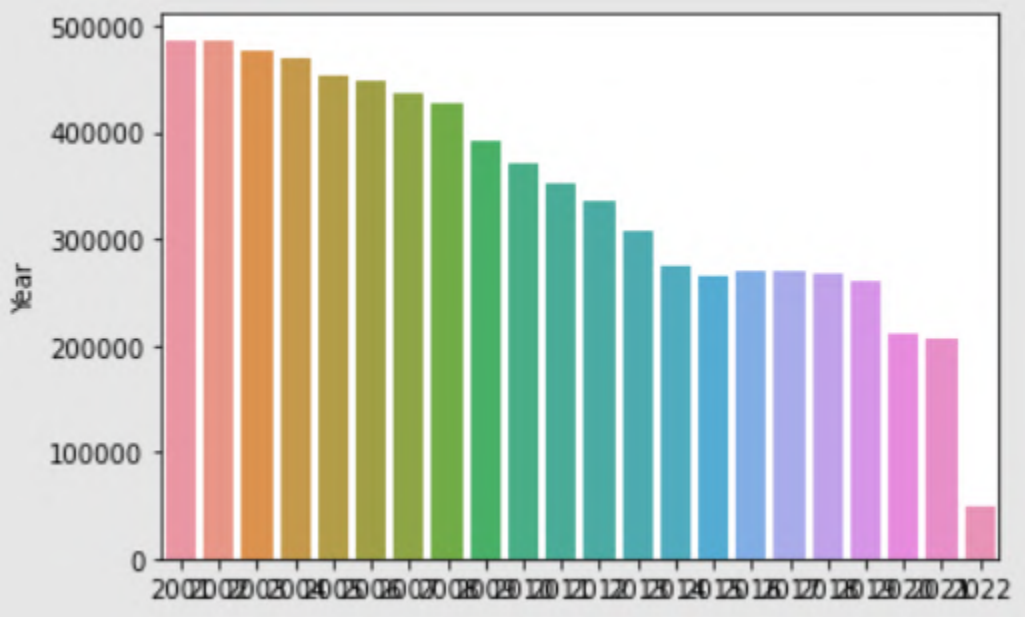
3.3 Obtained Results and Analysis

3.3.1 Visualizations

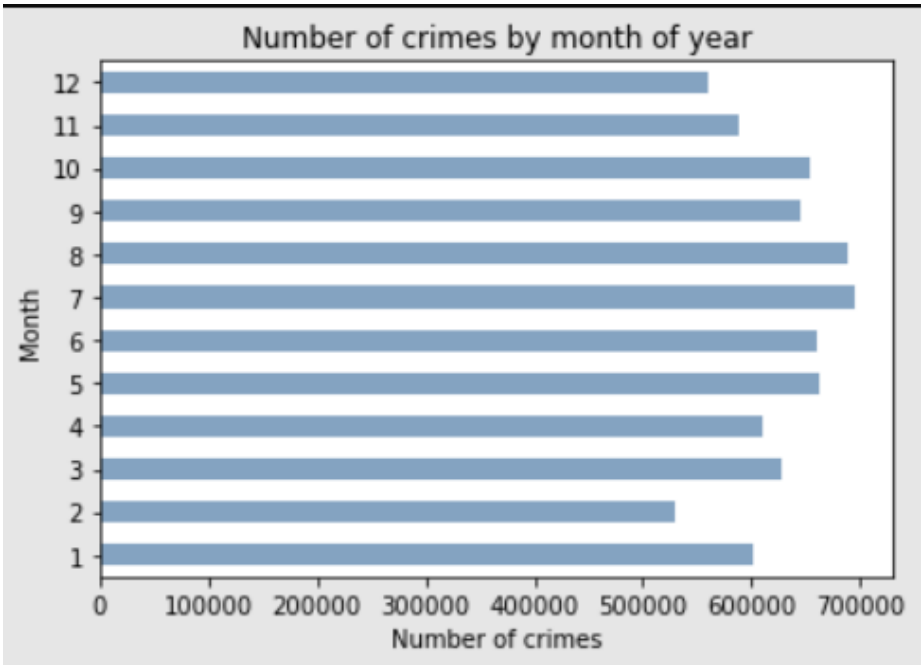


The number of crimes that have been committed each month has been visualized in this graph. It shows the overall pattern of a late summer / early fall peak in cases followed by a low period during the summer months. Number of crimes has not been steady over the years. The most

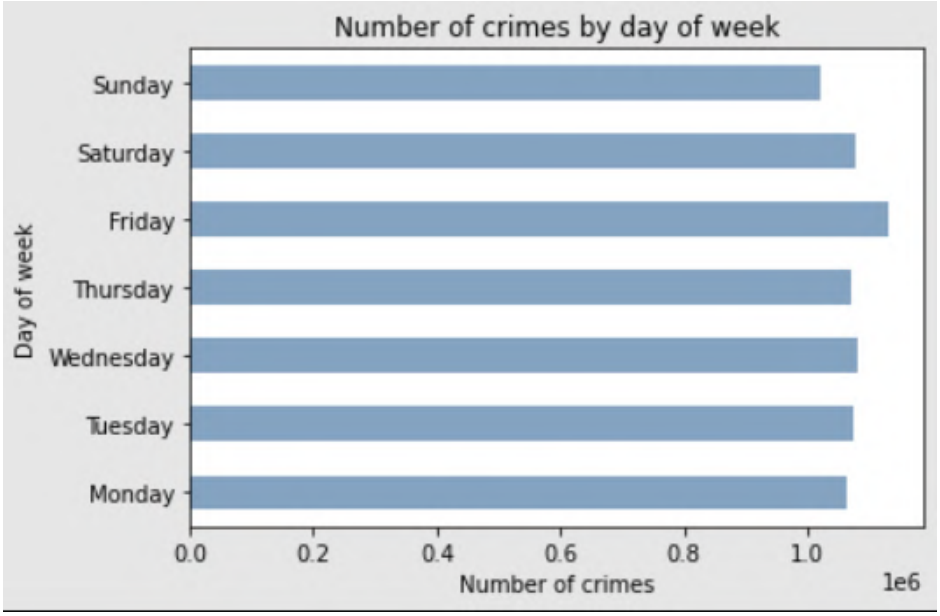
number of crimes were recorded between 2009-2010. There have been several peaks in the number of crimes in the last few years, but the number of crimes has been reducing since 2010.



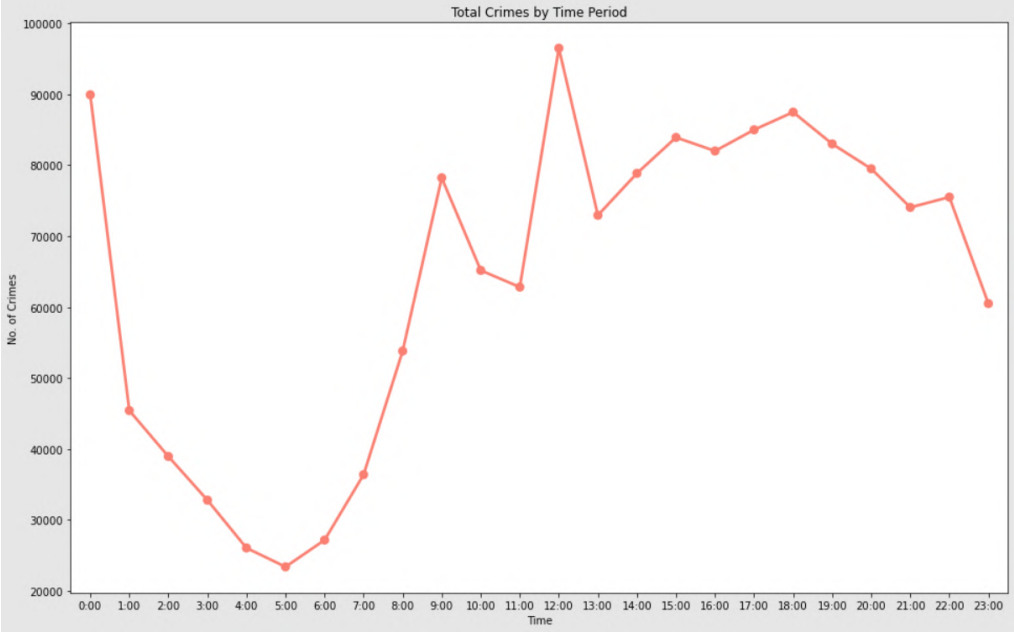
From over 450000 crimes in 2001 to less than 50000 crimes in 2022. The crimes over the last 21 years have been steadily decreasing. The number of crimes was constant between 2014-2016. This graph shows the overall progression of crimes across 22 years, live upto last week.



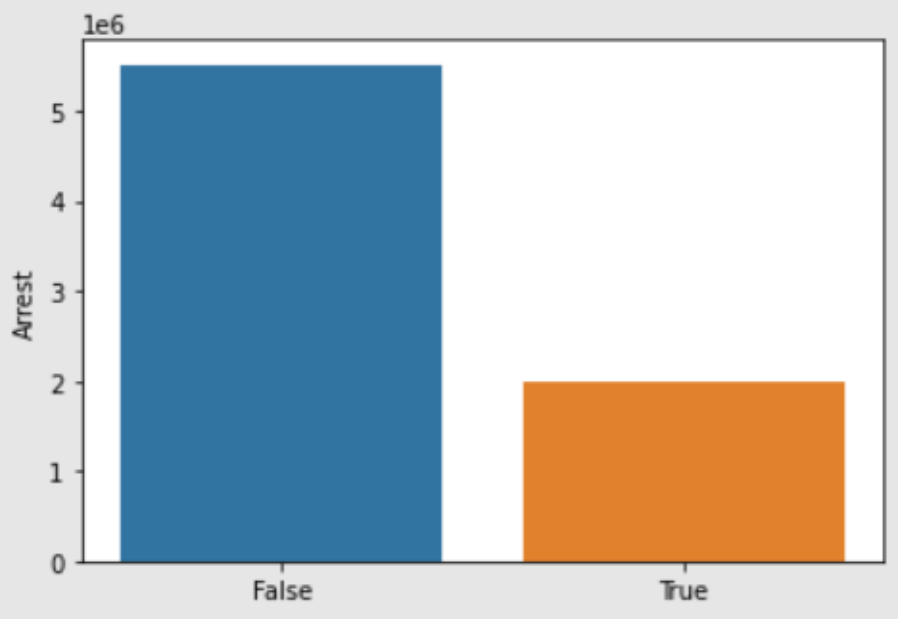
The number of crimes has been lowest in February over the years. The lowest number has been over 2000. The number of crimes has been the highest in September over the years. The highest number has been almost 3000. In conclusion, the number of crimes has been highest during May-October.



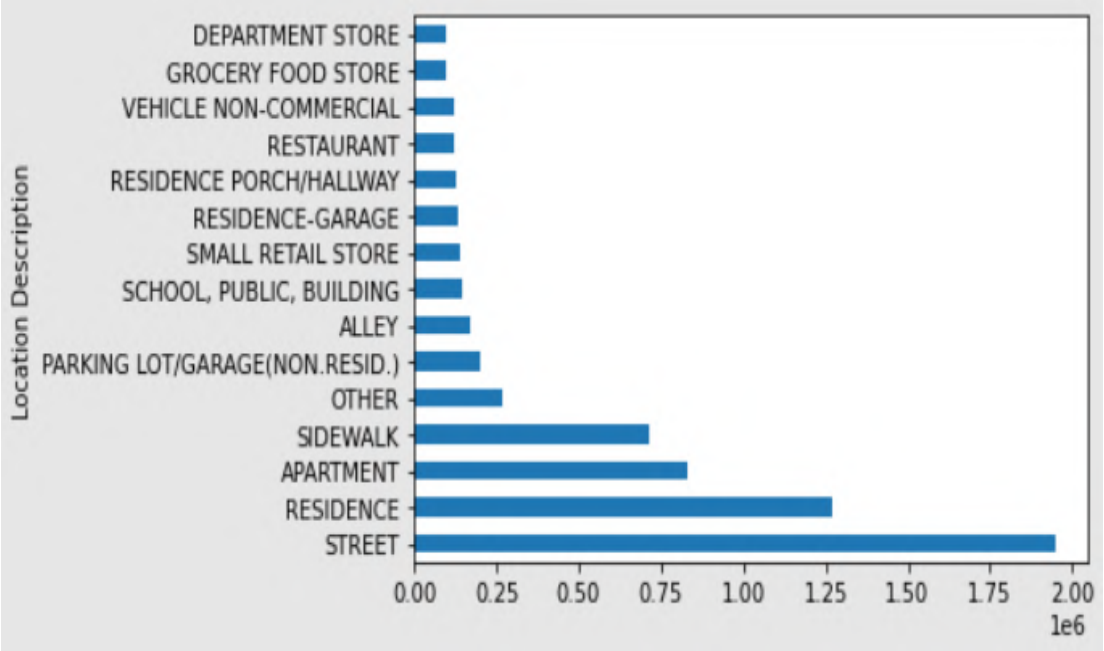
It has been seen that the number of crimes has been high throughout the week. The crimes have been well over 4000 crimes throughout the week.



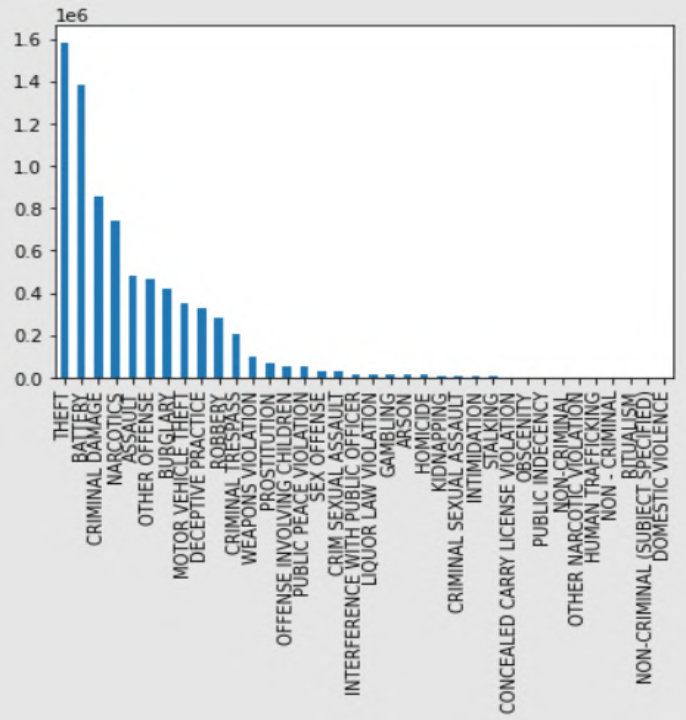
The number of crimes has been lowest in the early morning specifically during 4 am – 6 am. The number of crimes has been highest in the late-night, specifically from 7 pm – 11 pm. The crime density has been high throughout the day and till late at night.



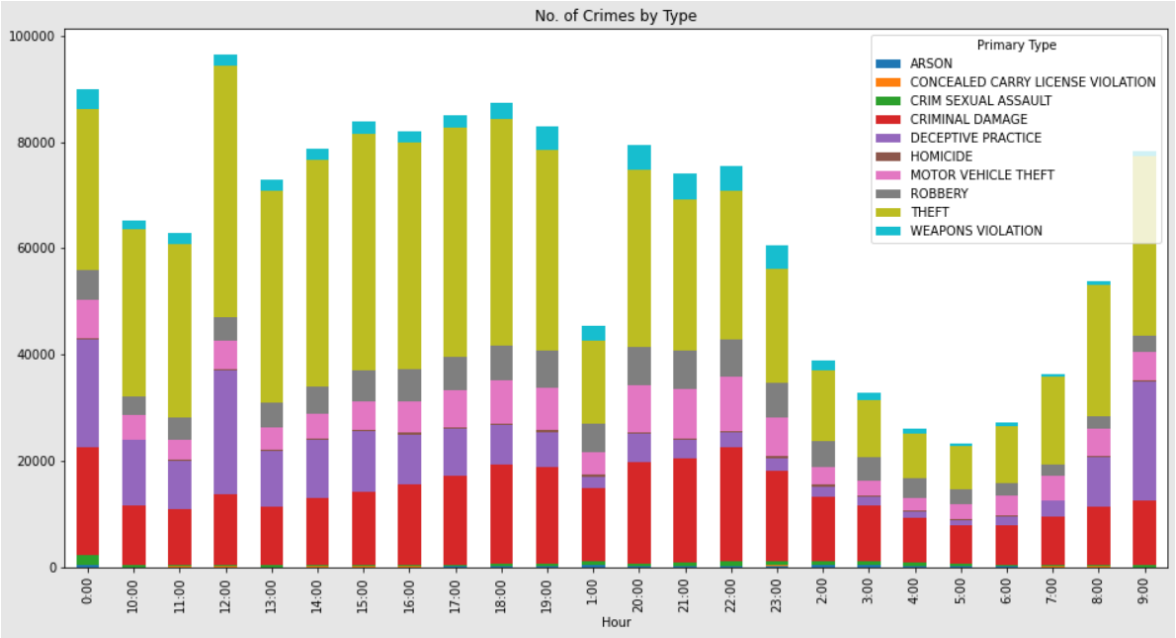
The number of suspects apprehended are way more than the one who got arrested. The number of false arrests made is almost thrice that of the true arrests.



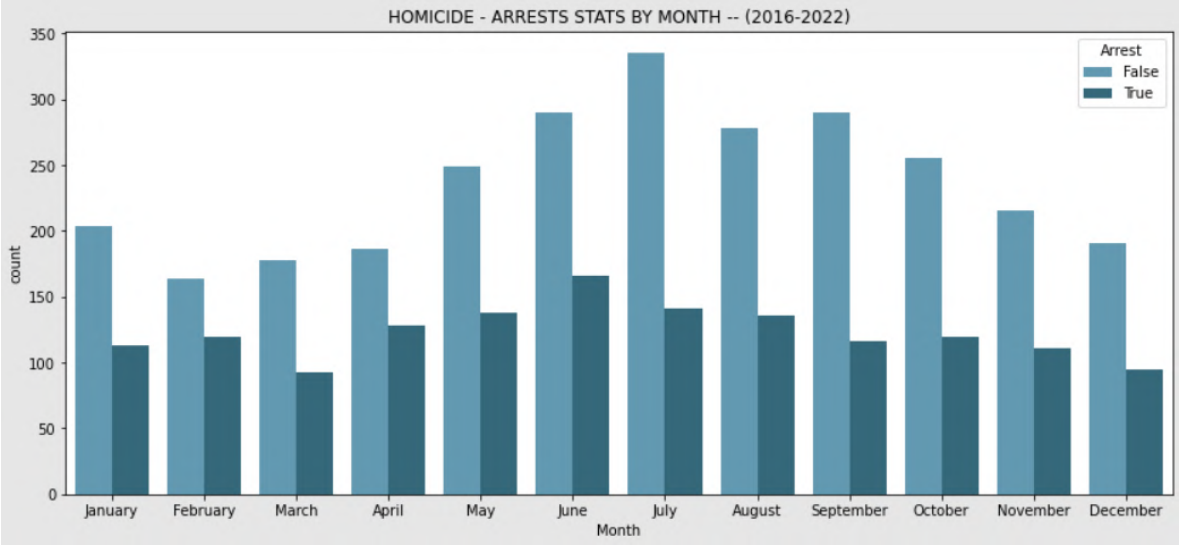
We identified locations that are more prone to crimes, the street being the scene with the highest crime rate. Also, we have pointed out the exact location (latitude and longitude) of that place.



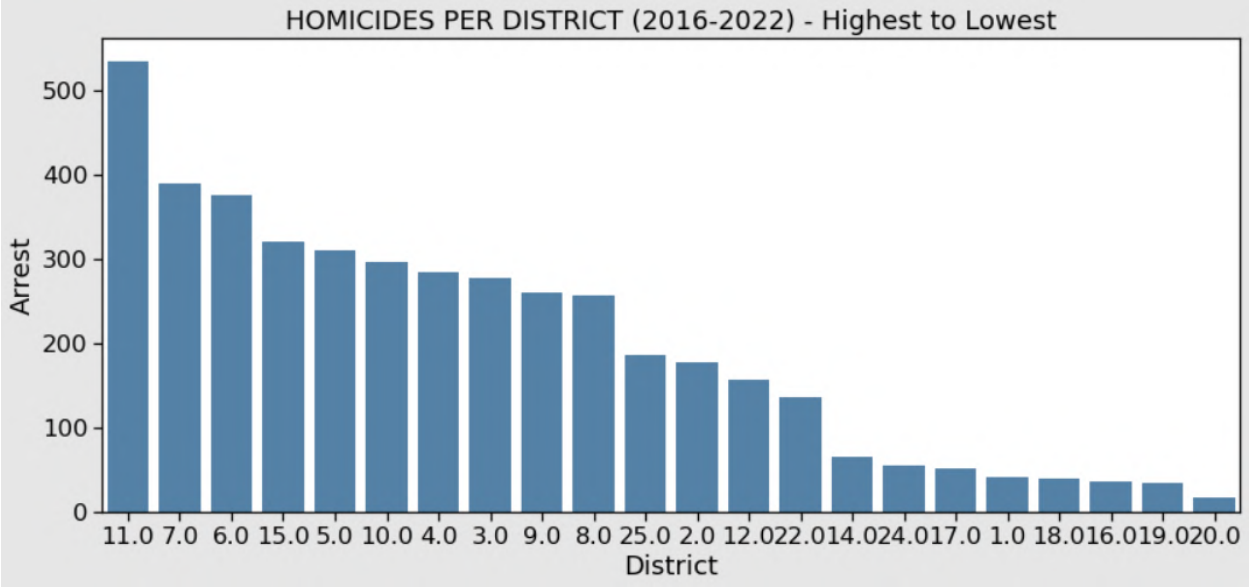
We interpret that theft has the highest percentage and is the crime type with the highest crime rate. Public Indecency and Domestic violence had least reporting.



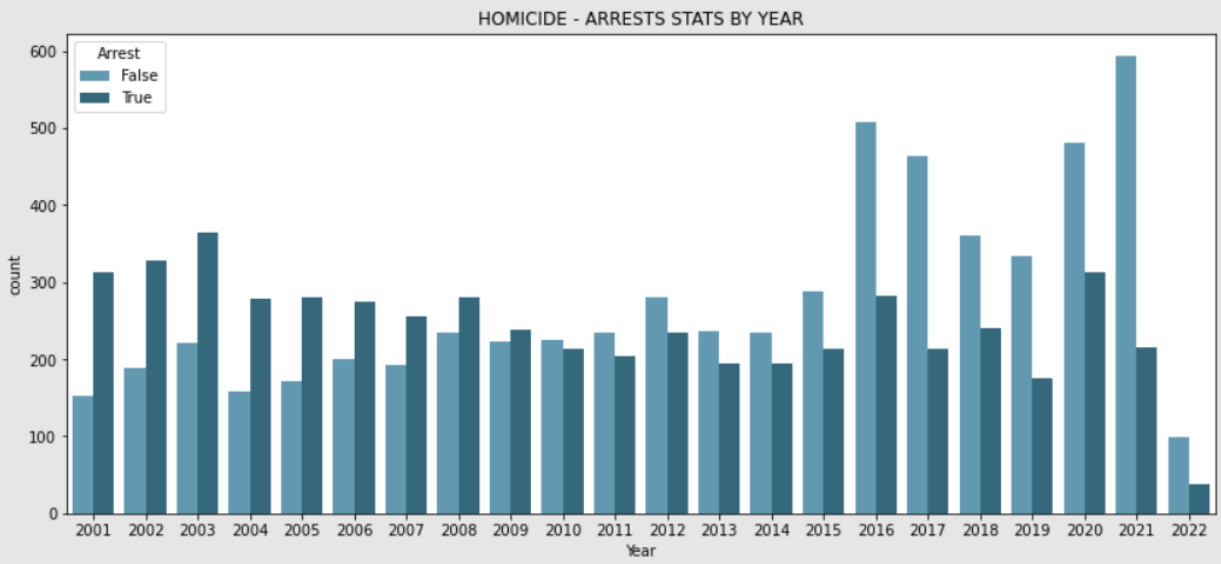
In general the number of crimes is less during the night and highest during the afternoon. Although a peek can be seen at 00:00 and 12:00 with later being the prominent one.



This shows that every month, there are more False arrests than True arrests. In July there is the highest number of False arrests, which is more than double of True arrests.



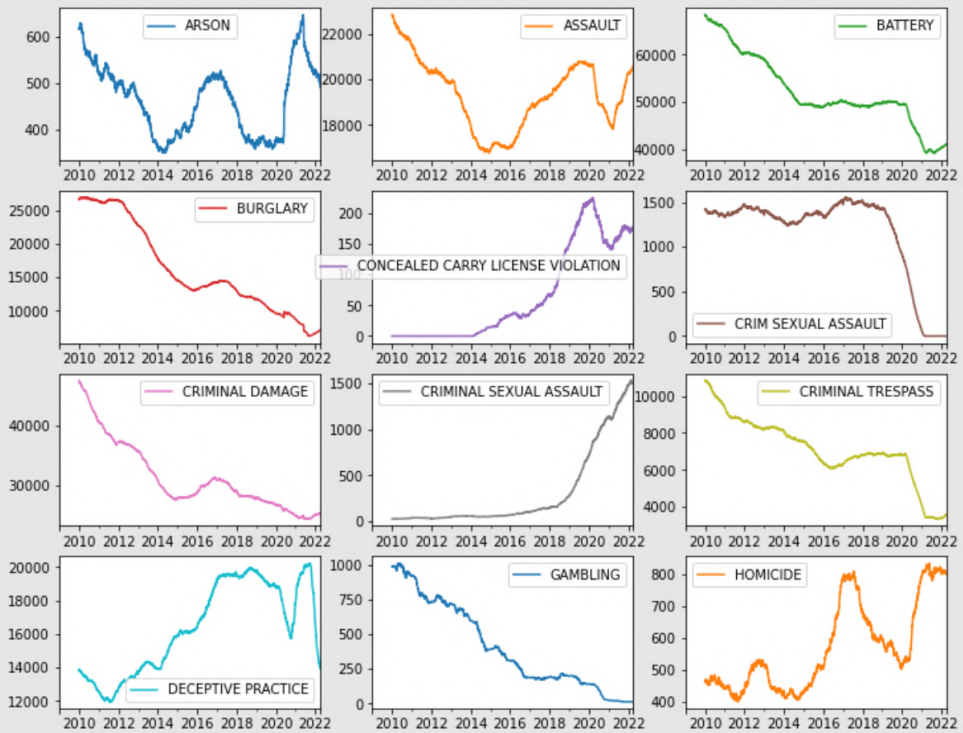
We visualize the crimes according to the districts, as it is easy for the department to coordinate and handle the crimes.



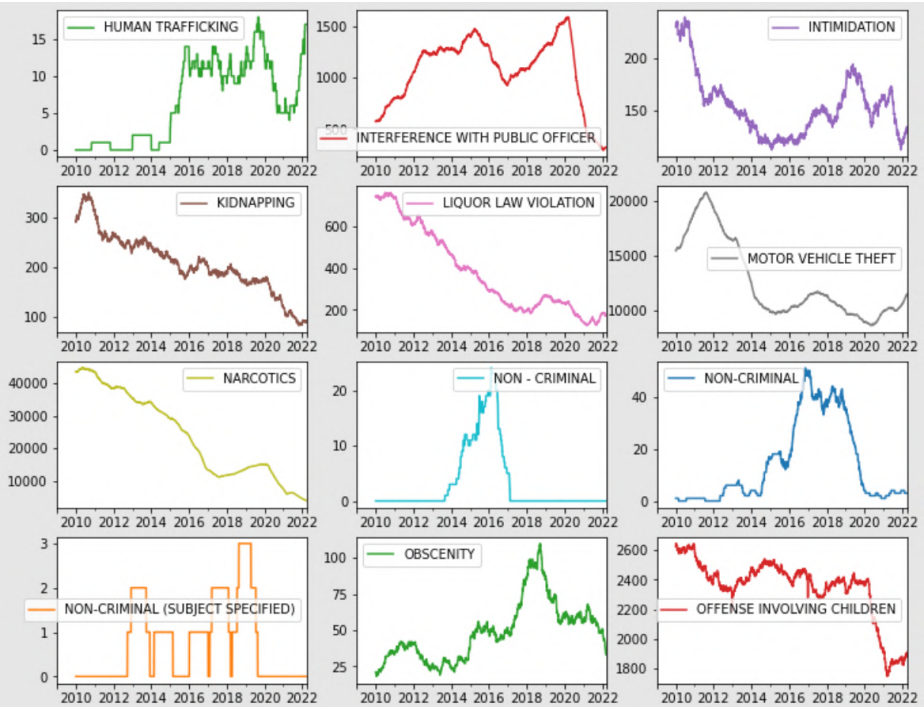
From this visualization, we interpret that before 2009 the number of True arrests were always higher than False arrests. After 2009 the number of False arrests is higher than True arrests.

This gives an insight on the change in investigation approach or indicates toward increasing complexity of crimes.

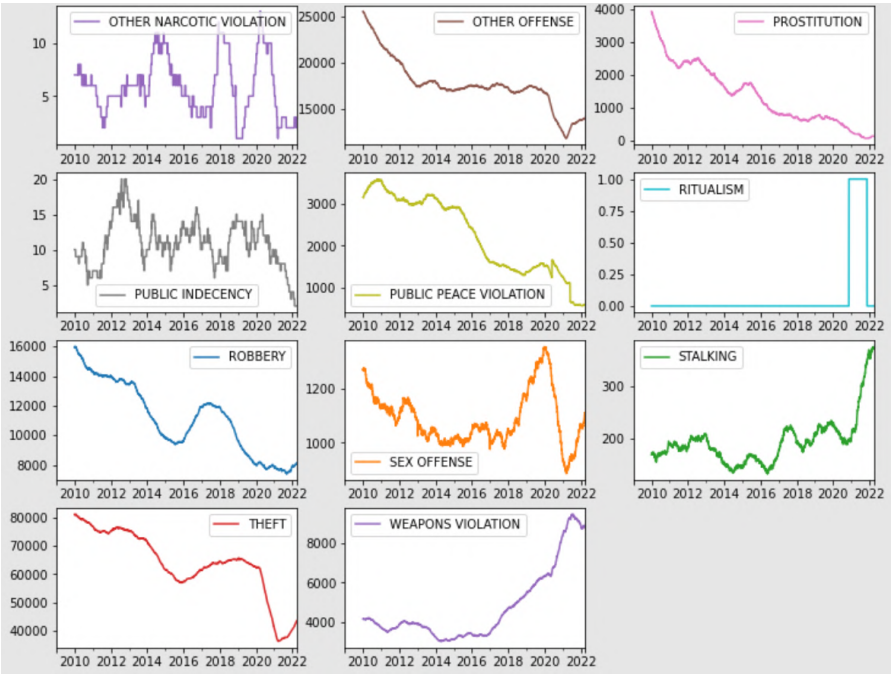
We can infer from the following visualization that the crime with crime type code “031A” was registered the most number of times. This was followed by 051A.



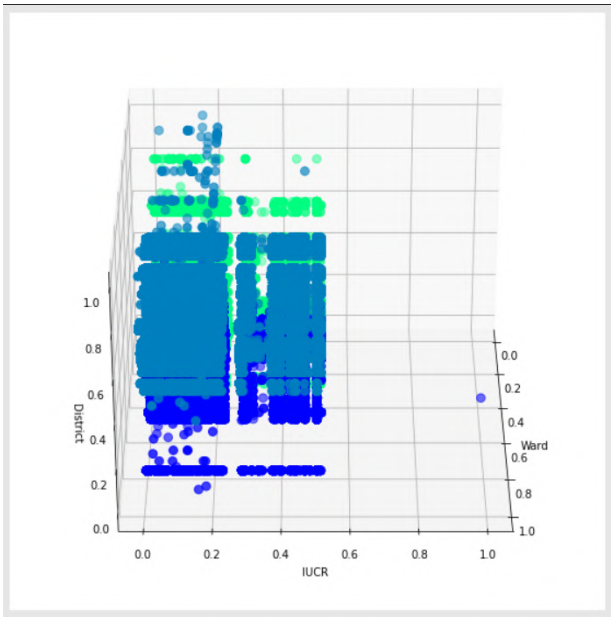
After 2020, there was a huge drop in sexual assaults and criminal trespasses. There was also a spike in homicide cases after 2020. Gambling, criminal damage, burglary, and battery are gradually dropping.



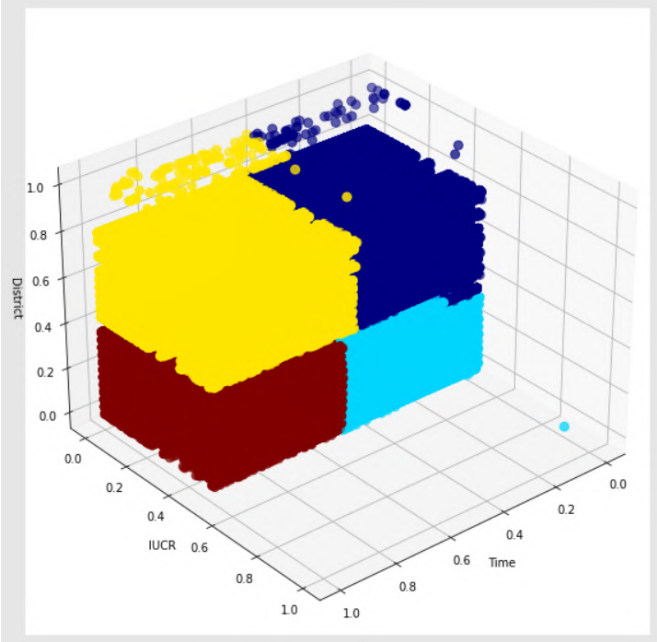
Narcotics, kidnapping, motor vehicle theft and liquor law violations are dropping gradually. Around 2015, there was a huge spike in human trafficking which dropped after 2020 and has started rising recently.



Theft, robbery, public peace violation, and prostitution is decreasing gradually. Crimes related to sex offense were highest in 2020. Ritualism cases were only in 2021.



First plot represents 3D Clustering with each point corresponding to their data values of District, IUCR and Ward.



Second plot represents 3D Clustering with each point corresponding to their data values of District, IUCR and Time.

3.3.2 Prediction of arrest type using Decision Tree Classifier

a) Percentage of True and False Arrests made according to Dataset :

```
Arrest
False    73.581252
True     26.418748
dtype: float64
------------------
Percentage Positive Instance = 26.418748369597775
Percentage Negative Instance = 73.58125163040222
```

b) Training accuracy, Precision and Recall for DecisionTree Classifier :

```
Training Accuracy = 1.0 Precision = 1.0 Recall = 1.0
```

c) Training accuracy for 9 depths in the Decision Tree :

```
Depth: 1 Accuracy: 0.736
Depth: 2 Accuracy: 0.779
Depth: 3 Accuracy: 0.839
Depth: 4 Accuracy: 0.828
Depth: 5 Accuracy: 0.817
Depth: 6 Accuracy: 0.821
Depth: 7 Accuracy: 0.802
Depth: 8 Accuracy: 0.792
Depth: 9 Accuracy: 0.771
```

d) Prediction accuracy of DecisionTree Classifier :

```
Accuracy for DT = 0.8387026326501109
Precision for DT = 0.964262110461366
Recall for DT = 0.964262110461366
```

e) Cross validation for prediction accuracy of DecisionTree Classifier :

```
Cross Validation Accuracy DT: 0.8387026369835059
Cross Validation Recall DT: 0.9636754268756554
Cross Validation Precision DT: 0.4044485755397509
```

4. Conclusion

This project offers visualizations for users to better interpolate data and understand the trend going on for the last 20 years. Law enforcement can gain insight on where to deploy extra force and in what capacity and at what time by studying the trend crimes have been showing for past years. These are not absolute rules but a calculated guess as to how an event may come to pass if it truly follows the past trends. It also tells about the change in investigation methodology or increase in complexity of crimes with variation in proportion of arrests.

The problem in this paper is predicting arrest types. The input features, hour, day of a week, business day, business hour, primary crime types, and community areas, are extracted from the raw dataset. The crime location and type descriptions in the raw dataset are preprocessed and classified into three prediction labels, which are home, public open space, and public buildings, before being fed into the prediction models. The Decision Tree model is implemented for the prediction. The machine learning decision tree classifier presents an estimated value as to whether the arrest made is True or False.

The prediction accuracy might be improved by reclassification. Thus, the accuracy might be much higher if crime location type is only classified into home and public places. Another strategy for improving the performance is that more datasets are applied in the prediction, such as demographic data, neighborhood appearance, and meteorological data. The crime occurrence is similar to our prediction. However, we focus on the prediction of arrest types, and it is helpful when the police patrols in the city.

Furthermore, this can help the police department to take the safety precautions for the civilians accordingly so that people can live safely and it will be easy for the department to handle all the situations.

5. References

[1] McClendon, Lawrence & Meghanathan, Natarajan. (2015). Using Machine Learning Algorithms to Analyze Crime Data. Machine Learning and Applications: An International Journal. 2. 1-12. 10.5121/mlaij.2015.2101.

[2] Chandrasekar, Addarsh, Abhilash Sunder Raj, and Poorna Kumar. "Crime Prediction and Classification in San Francisco City."

[3] Vaquero Barnadas, Miquel. "Machine learning applied to crime prediction." Bachelor's thesis, Universitat Politècnica de Catalunya, 2016

[4] Computer Science IJCSIS, Journal of. “Mining Forensic Medicine Data for Crime Prediction.” IJCSIS Vol 17 No 6 June Issue, 2019.

[5] Chen, Huanfa & Cheng, T. & ye, Xinyue. (2018). Designing efficient and balanced police patrol districts on an urban street network. International Journal of Geographical Information Science. 33. 1-22. 10.1080/13658816.2018.1525493.