Analysis And Visualization Of Crimes In Urban Illinois

SUBMITTED BY:

HARSHIT ANAND (19BCB0071)
AVISH AVIRAJ JHA (19BCE0812)
UTPAL MANISHCHANDRA PRAJAPATI (19BCE0759)

SUBMITTED TO:
Prof. Jyotismita Chaki

Group - 1

Abstract:

Crime pattern analysis is related to public safety and helpful for police patrols. There are three important topics, which are crime visualisation, criminal behavior pattern analysis and hot-spot prediction, in the crime pattern analysis.

The Chicago Crime dataset with entries stretching back to 2001 is visualized and analysed 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. It's accuracy is measured and cross validated against the present data.

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 pattern in a city is beneficial for allocating patrol resources and residents improving their protection consciousness. In our to provide clear view of crime pattern to police, data analysis and visualization is done in our project.

Traditional statistical methods analyze data for the population, race, surroundings, and education level in the crime prevention, which is a time-consuming task [4]. 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.

Literature Survey

S. No.	Author Name	Research Paper	Summary
1	Addarsh Chandrasekar, Abhilash Sunder Raj and Poorna Kumar	Crime Prediction and Classification in San Francisco City	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 they 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
2	Vaquero Barnadas, Miquel.	Machine Learning Applied To Crime Prediction	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 Neighbours, 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.

S. No.	Author Name	Research Paper	Summary
3	A. Abdo , Hanan Fahmy , Amir Abobaker Shaker	Mining Forensic Medicine Data for Crime Prediction	In this research, a framework has been built to analyze forensic medicine data for crimes 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 technique 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.
4	Huanfa Chen, T. Cheng, Xinyue ye	Designing efficient and balanced police patrol districts on an urban street network	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

S. No.	Author Name	Research Paper	Summary
5	Lawrence McClendon, Natarajan Meghanathan	Using Machine Learning Algorithms to Analyze Crime Data	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 modelled. On the other hand, the linear regression algorithm could handle randomness in the test samples to a certain extent (without incurring too much of prediction error).

Motivation

The city of Chicago publishes an up-to-date list of all reported crimes. The records span from 2001 to the modern date, allowing a few days of delay to catalog the crimes and publish them. According to official FBI data, Chicago is one of the leading cities in homicides, having more than quadruple the amount of crimes in New York City, and more than double the amount of crimes in Los Angeles. Being able to parse the data presented by these huge data sets is a problem central to understanding the crimes in the city of Chicago.

By analyzing the data in a mathematically rigorous way, researchers may be able to glean insight into the underlying causes of crimes, and also may be able to figure out indicators of future crimes to occur.

Proposed Approach

- We are using the Classification algorithm in this project to accomplish our task.
- The task will be carried out in two steps : Visualization and Prediction.
- In the data obtained from the system of the City Police Department, each criminal record is characterized by several attributes that include crime description, location, longitudes, and latitudes, etc.

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.

- In addition, the system classifies the crimes into 32 different categories.
- With all the attributes, we expect to depict the pattern of each crime type across the city.

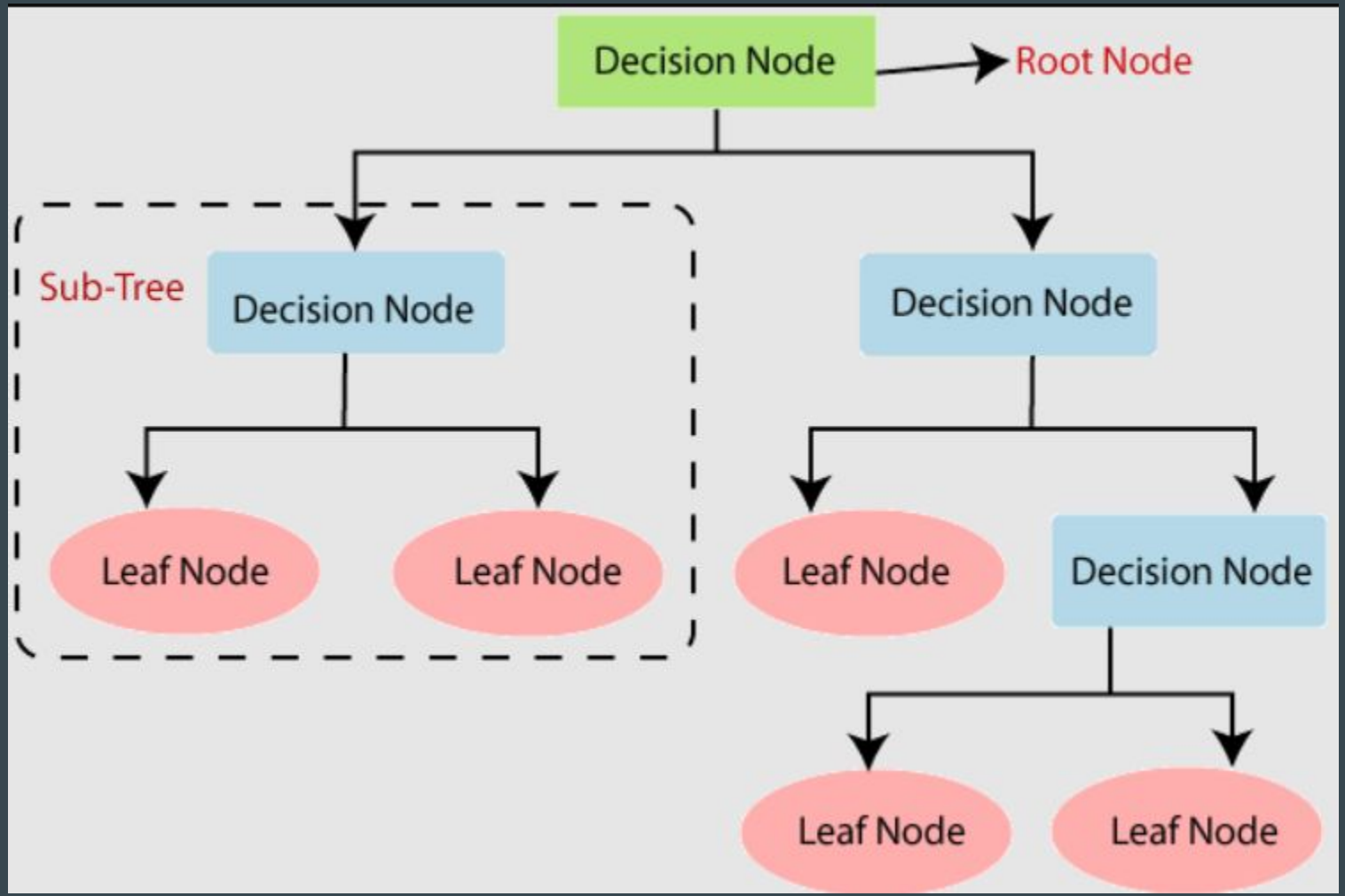
Theft	Battery
Robbery	Criminal Damage
Deceptive Practice	Narcotics
Domestic Violence	Non-Criminal (Subject Specified)
Assault	Criminal Trespass
Gambling	Arson
Burglary	Prostitution
Concealed Carry License Violation	Human Trafficking
Motor Vehicle Theft	Weapons Violation
Homicide	Offense involving Children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

Visualization :

- 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 graph 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.

Prediction :

- 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.

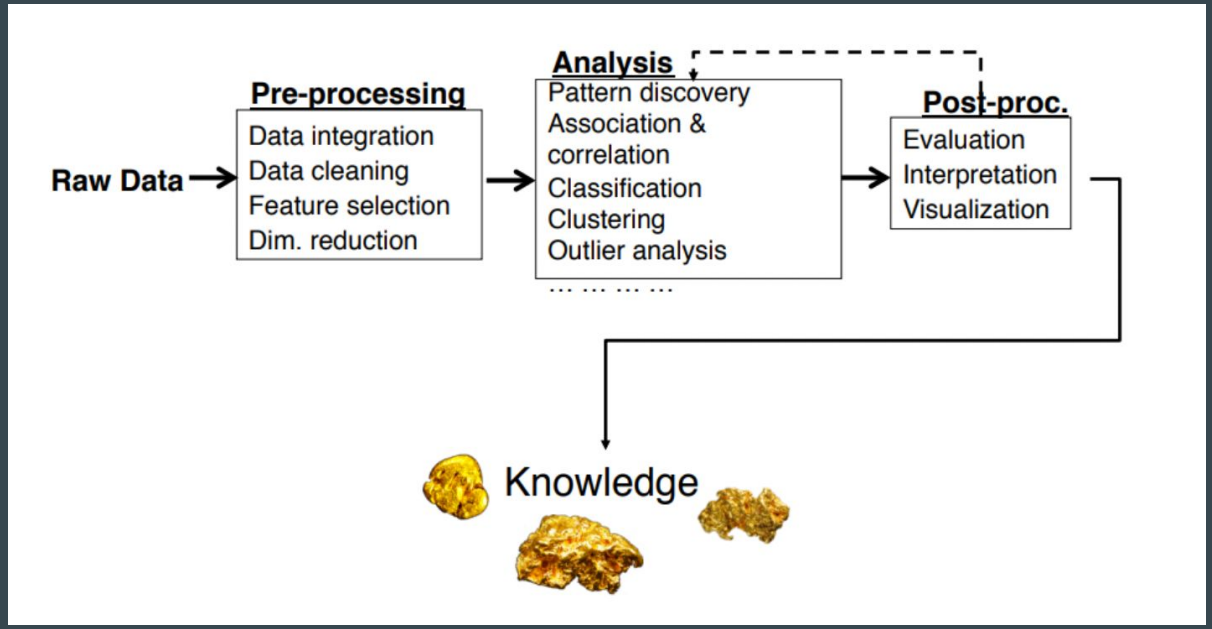

Decision Node
Root Node
Sub-Tree
Decision Node
Decision Node
Leaf Node
Leaf Node
Leaf Node
Decision Node
Leaf Node
Leaf Node

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.

IMPLEMENTATION

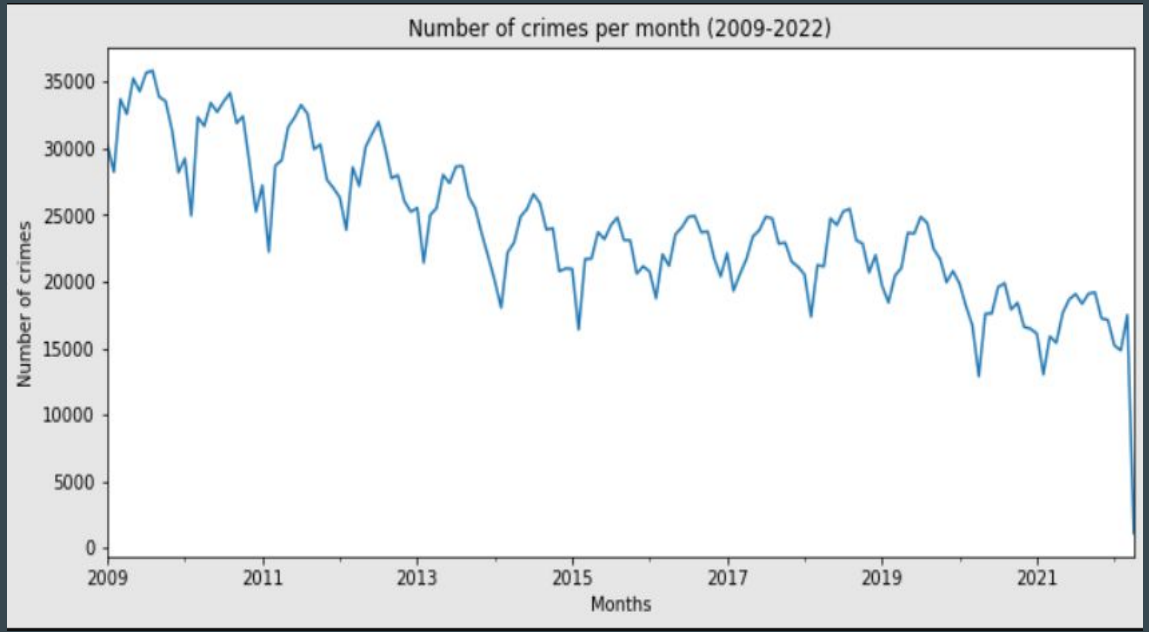
Flowchart



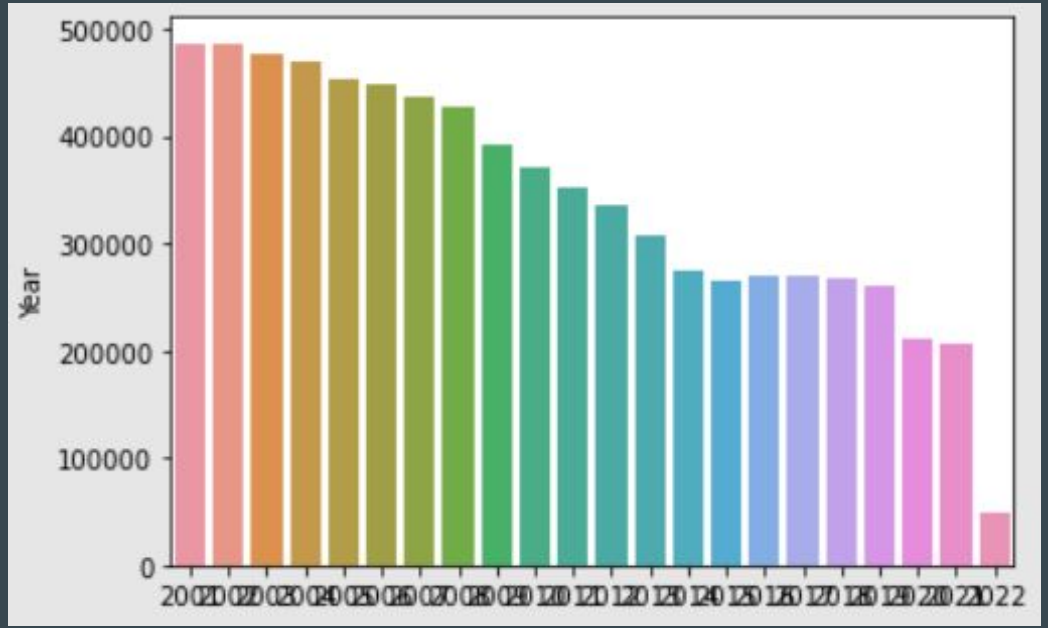
Results Obtained and Analysis

- We visualized the data using different ways. We analyzed the trend of crime occurrence for each year.
- Then, we plotted crime occurrence rates of the following: crime type, scene of the crime, hour day-month of crime.
- This gives us a better understanding of the major crimes that occur.

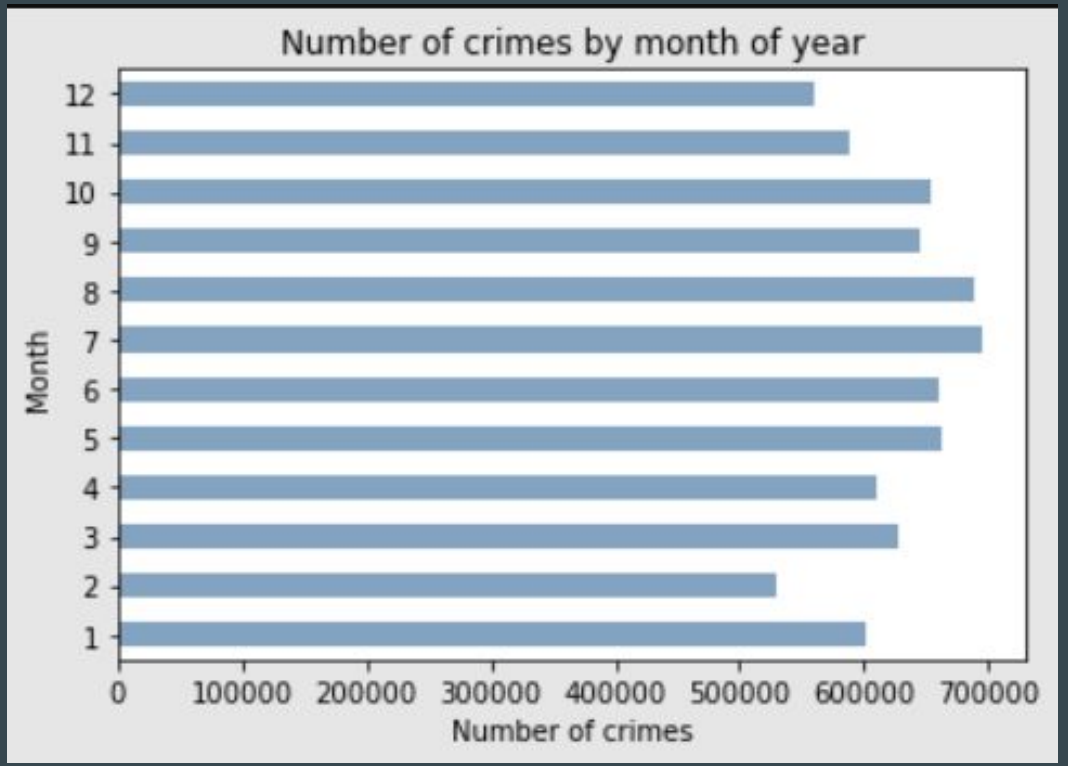
- Number of crimes have not been steady over the years.
- The most number of crimes were recorded between 2009-2010.
- There have been several peaks in the number of crimes in the last few years, but the number of crimes has been reducing since 2010.



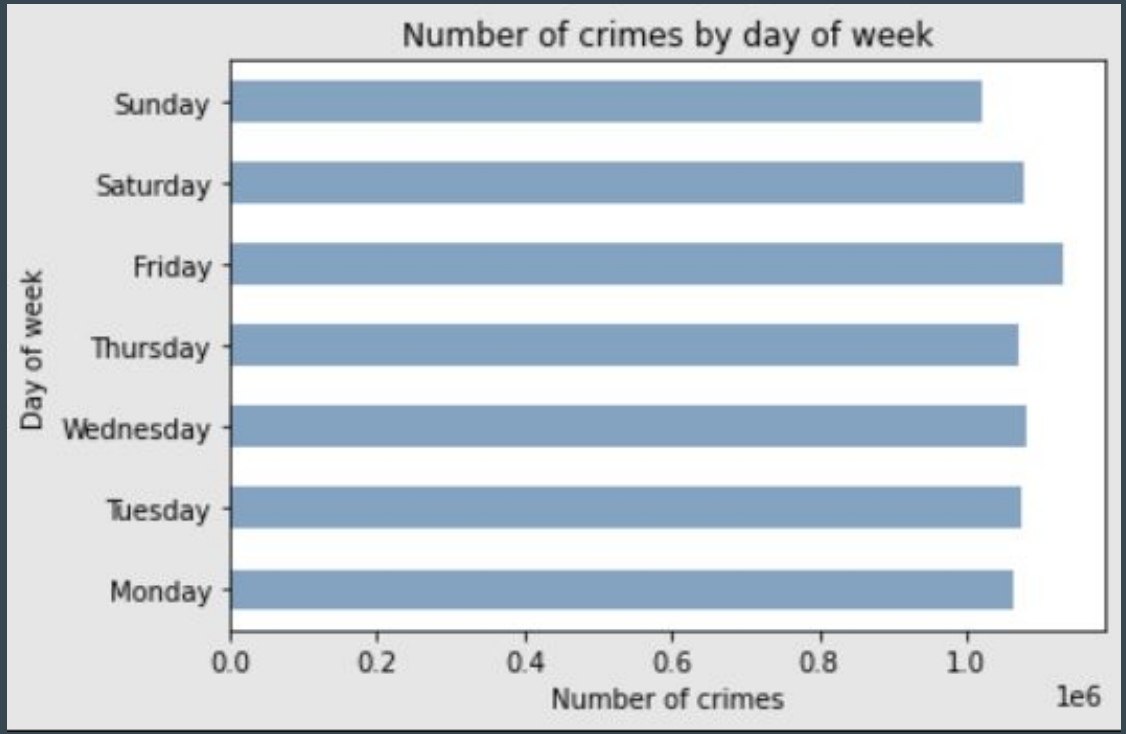
- From over 450000 crimes in 2001 to less than 50000 crimes in 2022.
- The crimes over the last 21 years have been steadily decreasing.
- The number of crimes was constant between 2014-2016.



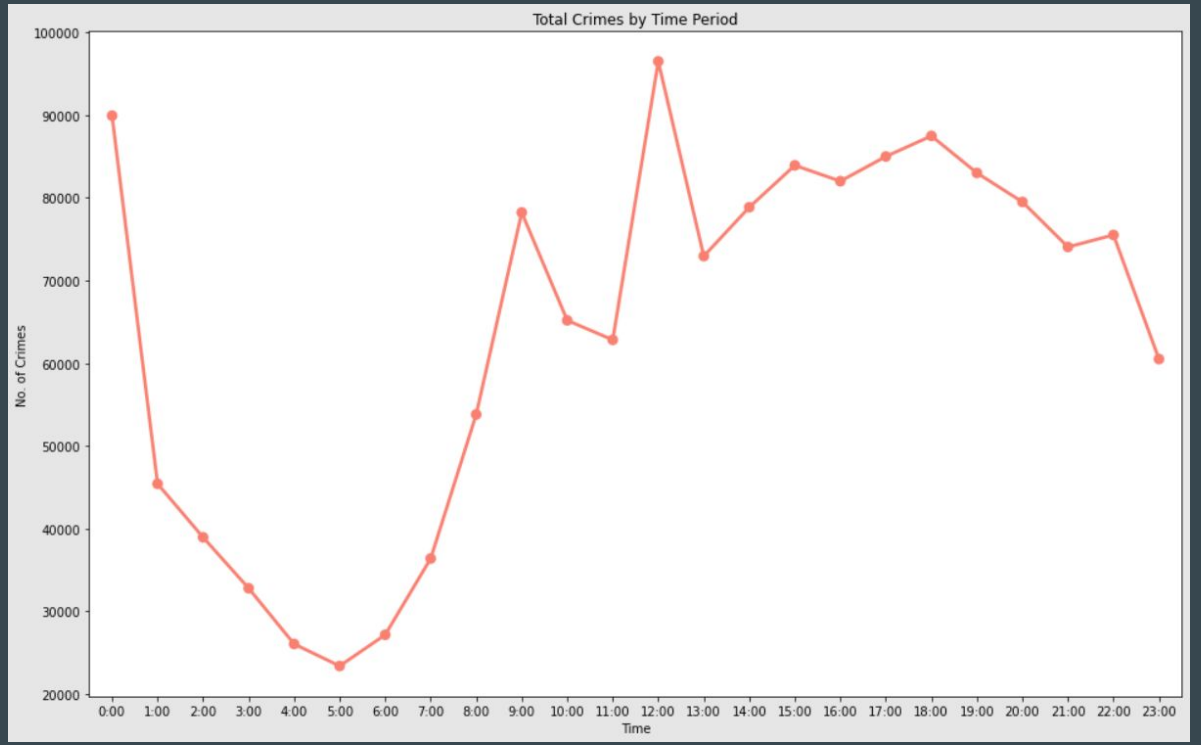
- The number of crimes has been lowest in February over the years. The lowest number has been over 2000
- The number of crimes has been the highest in September over the years. The highest number has been almost 3000.
- In conclusion, the number of crimes has been highest during May-October.

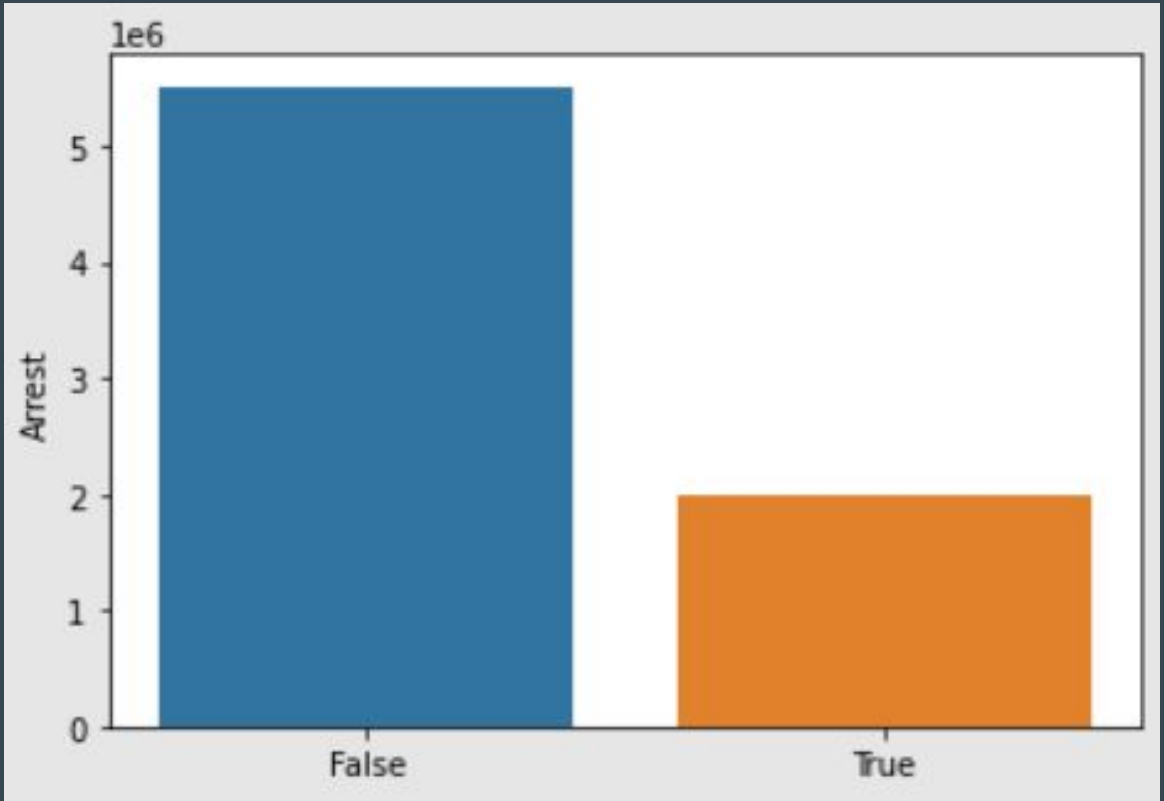


- It has been seen that the number of crimes has been high throughout the week.
- The crimes have been well over 4000 crimes throughout the week.



- 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.



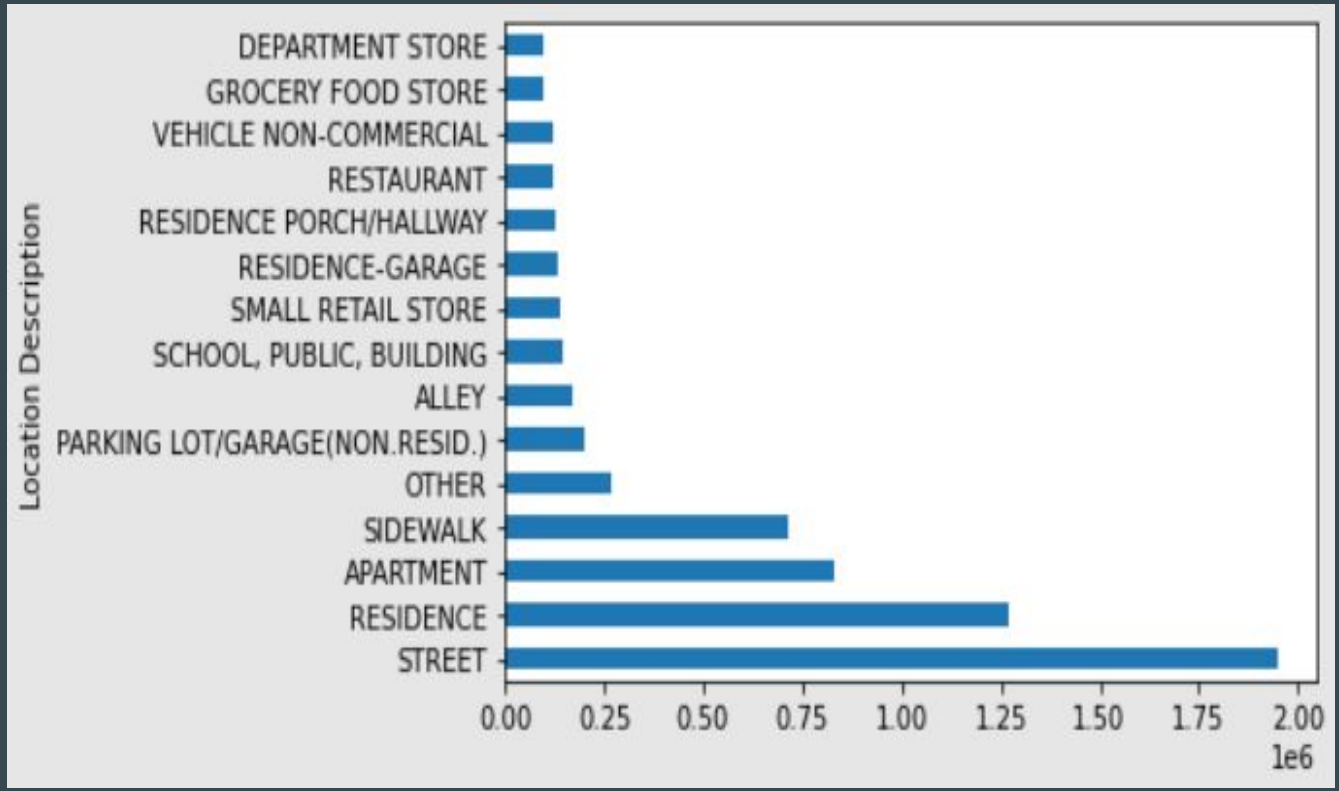


The number of suspects apprehended are way more than the one who got arrested.

The number of false arrests made is almost thrice that of the true arrests.

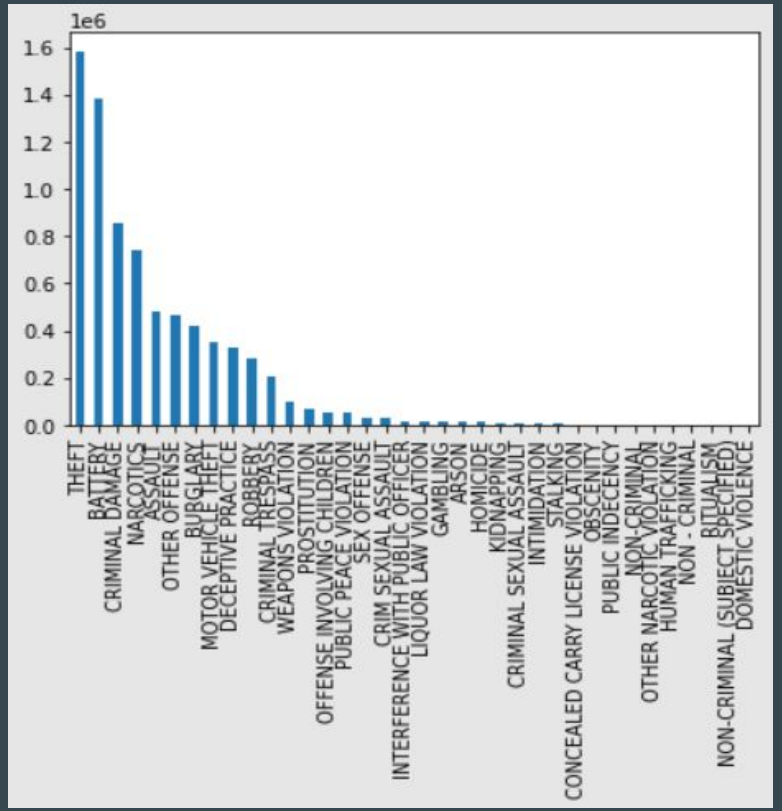
• We identified locations that are more prone to crimes, the street being the scene with the highest crime rate.

•Also, we have pointed out the exact location (latitude and longitude) of that place.



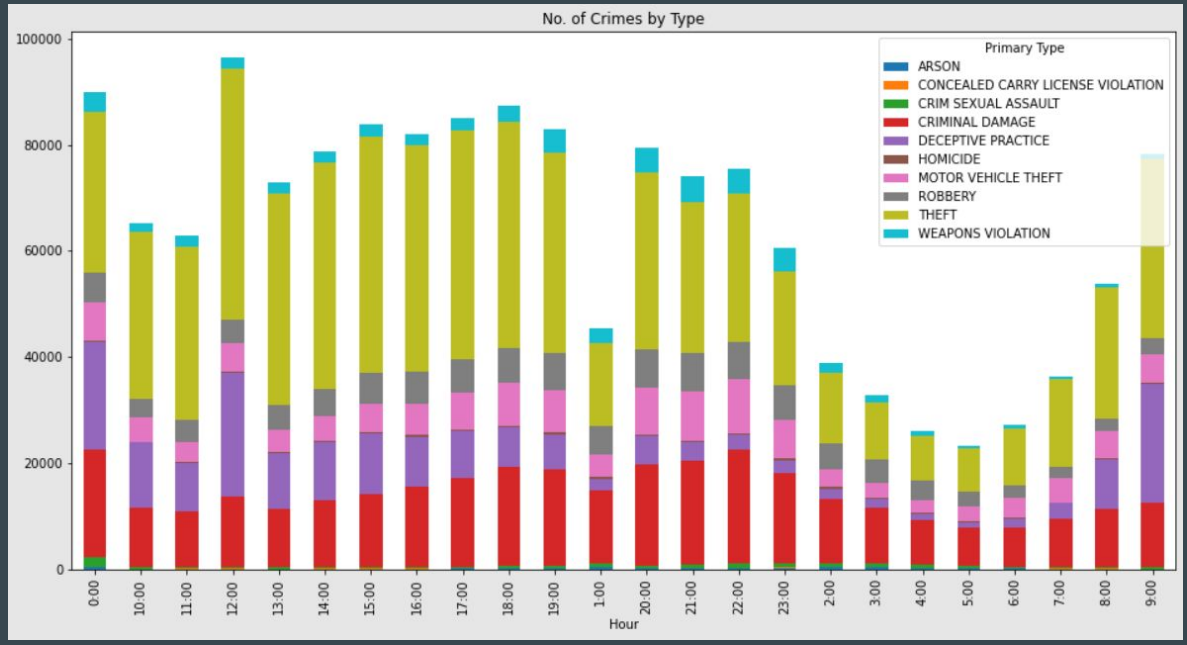
- We interpret that theft has the highest percentage and is the crime type with the highest crime rate.

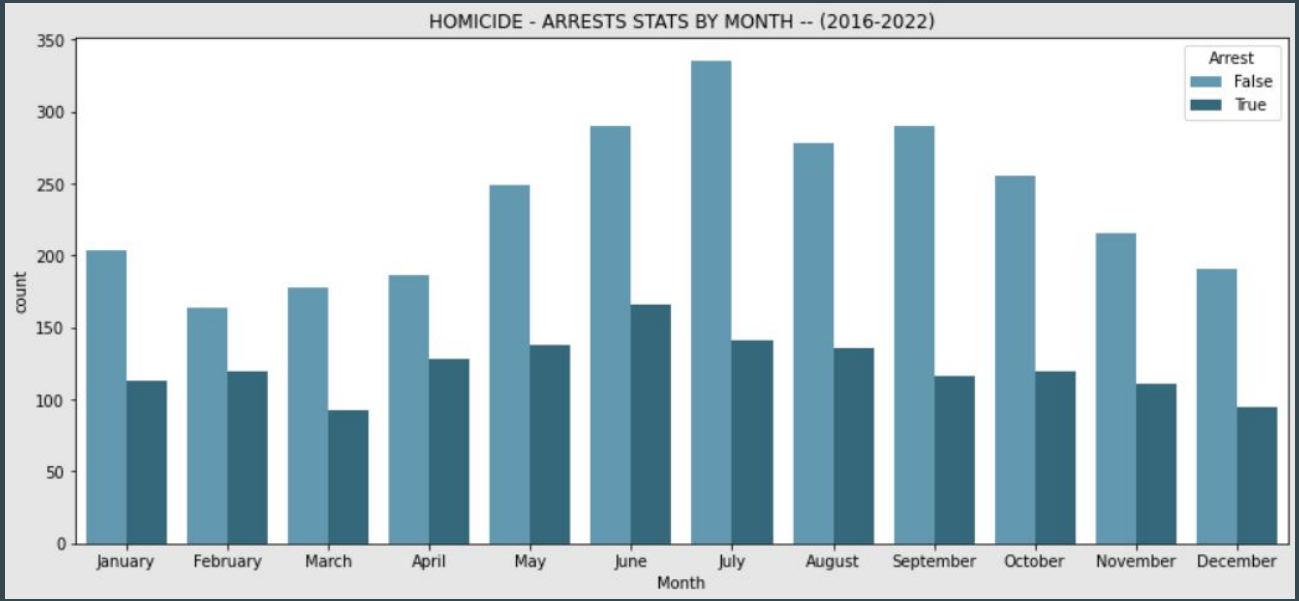
- Public Indecency and Domestic violence had least reporting.



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.

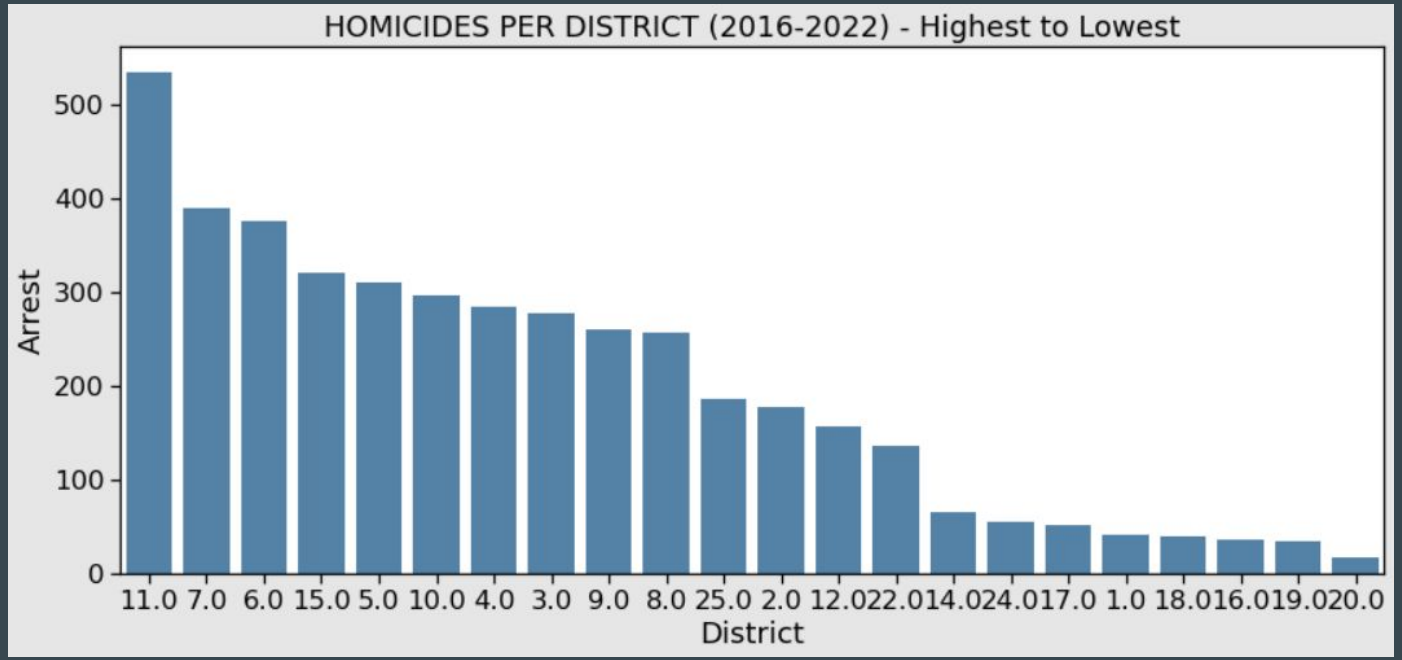




This shows that every months, there are more False arrests than True arrests.

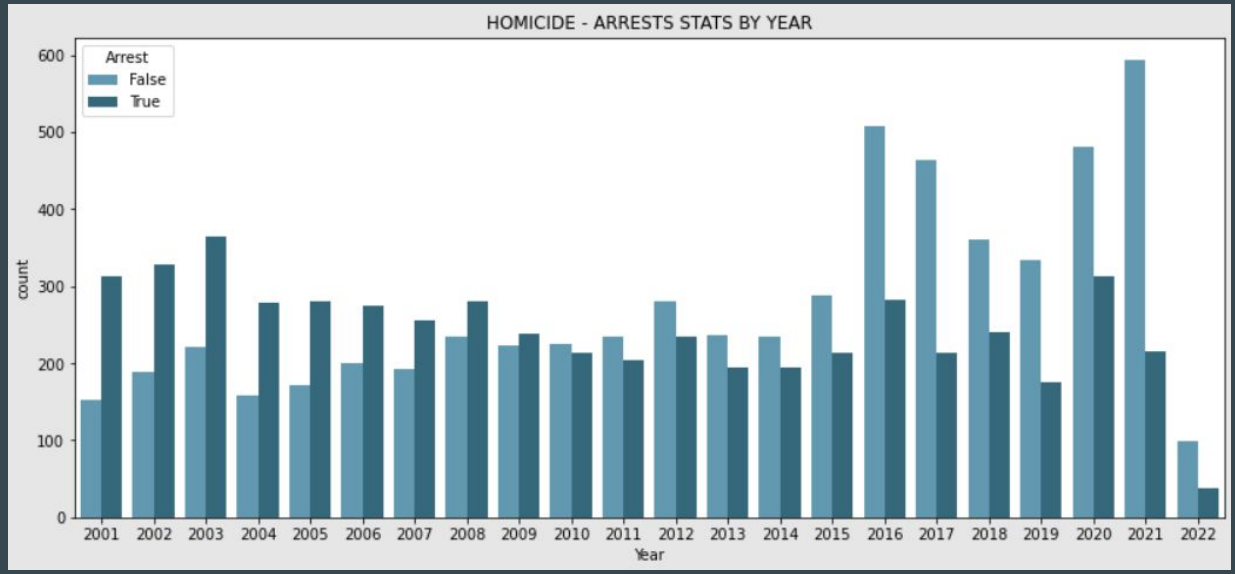
In July there is the highest number of False arrests, which is more than double of True arrests.

- 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 number of True arrests are always higher than False arrests. after 2009 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.



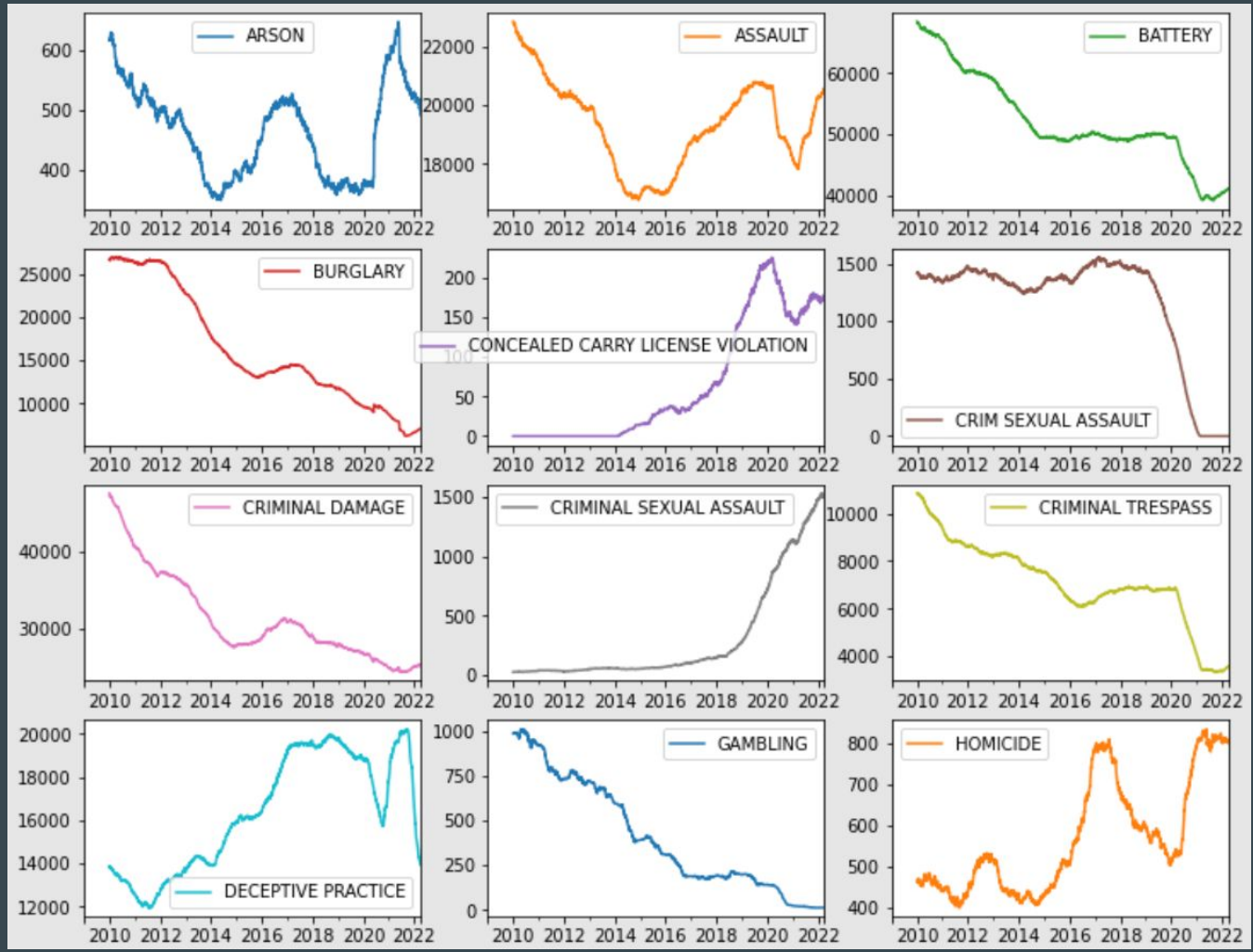
- We can infer from the following visualization that the crime with crime type code “**031A**” was registered most number of times.

- This was followed by 051A.

After 2020, there was a huge drop in sexual assaults and criminal trespasses.

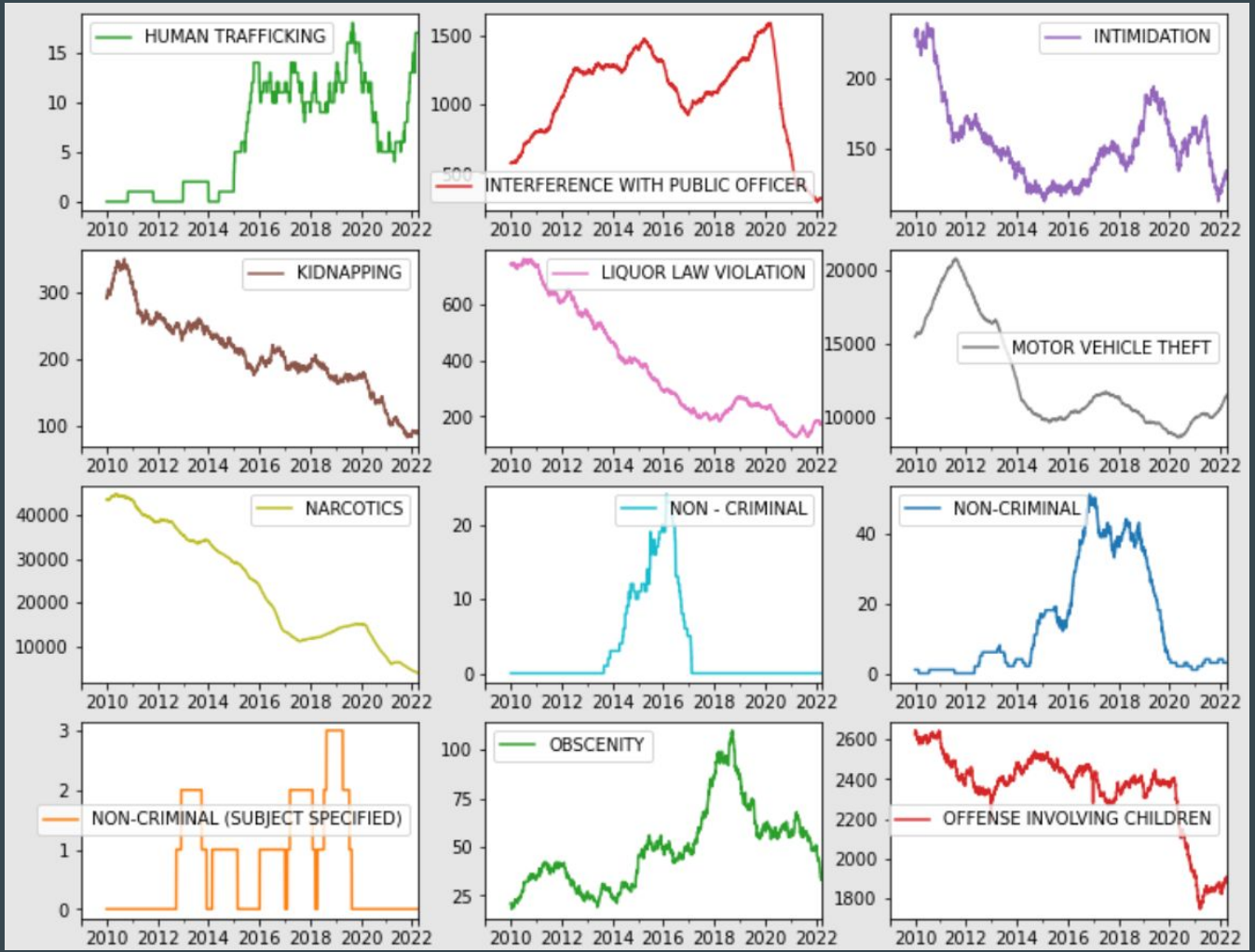
There was also a spike in homicide cases after 2020.

Gambling, criminal damage, burglary, and battery are gradually dropping.



Narcotics, kidnapping, motor vehicle theft and liquor law violations are dropping gradually.

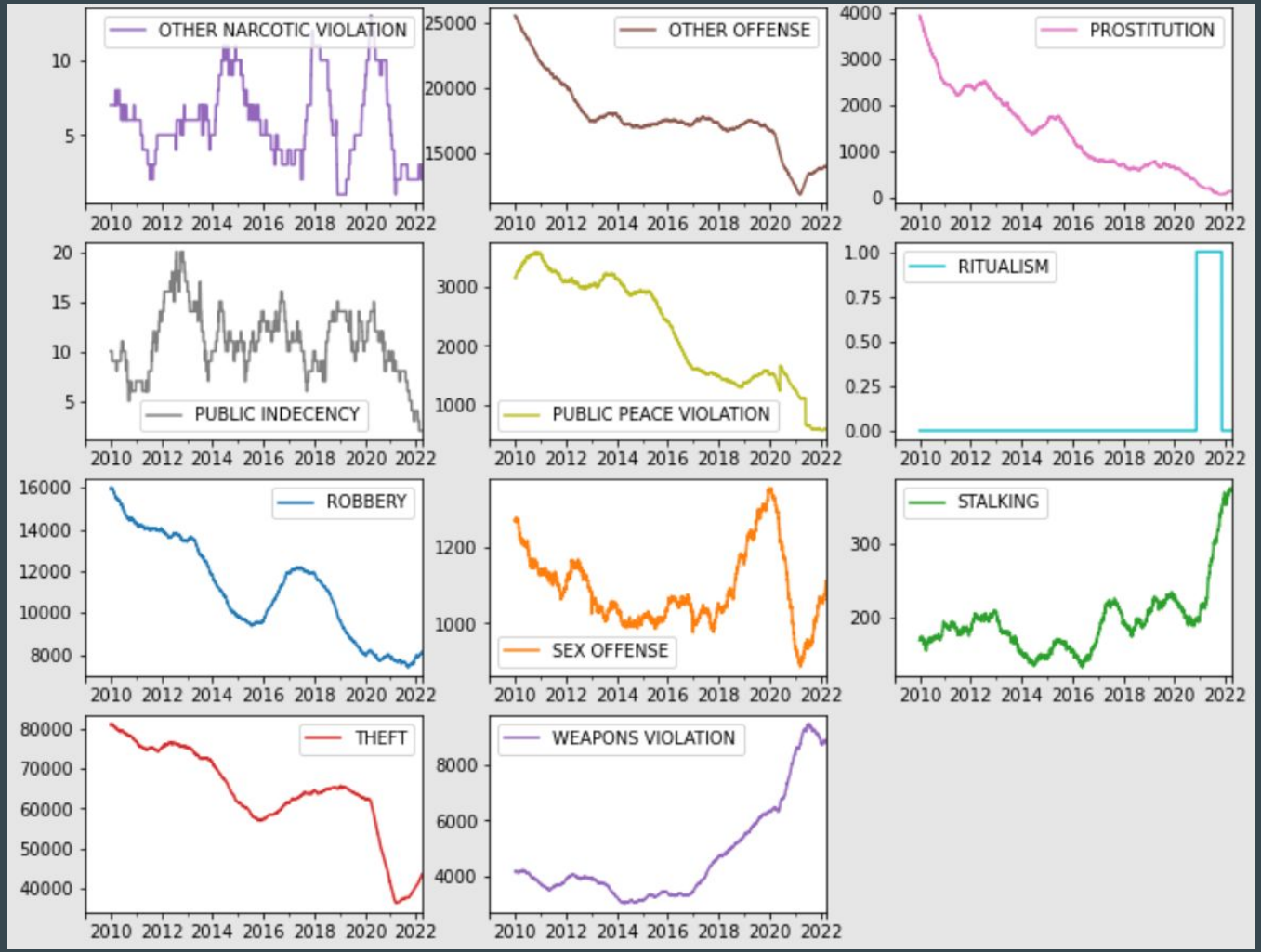
Around 2015, there was a huge spike in human trafficking which dropped after 2020 and has started rising recently.



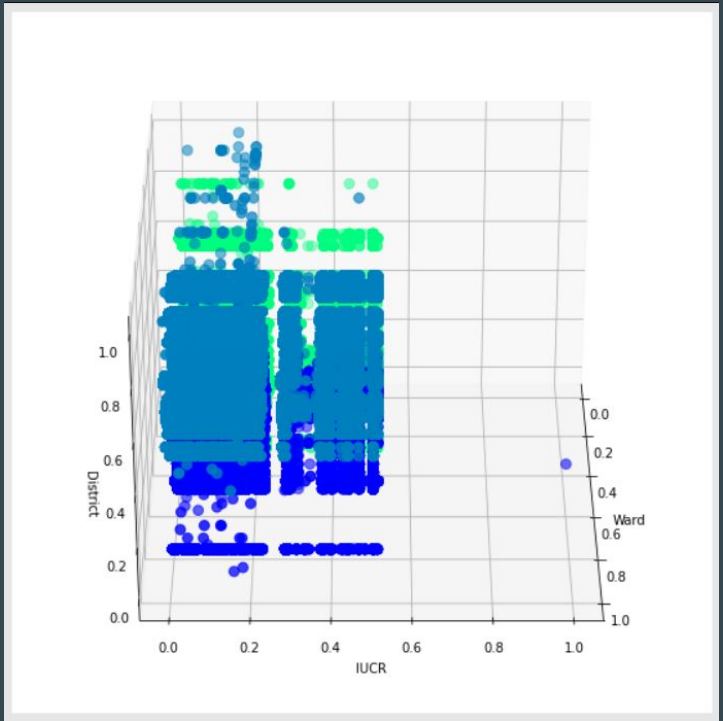
Theft, robbery, public peace violation, and prostitution is decreasing gradually.

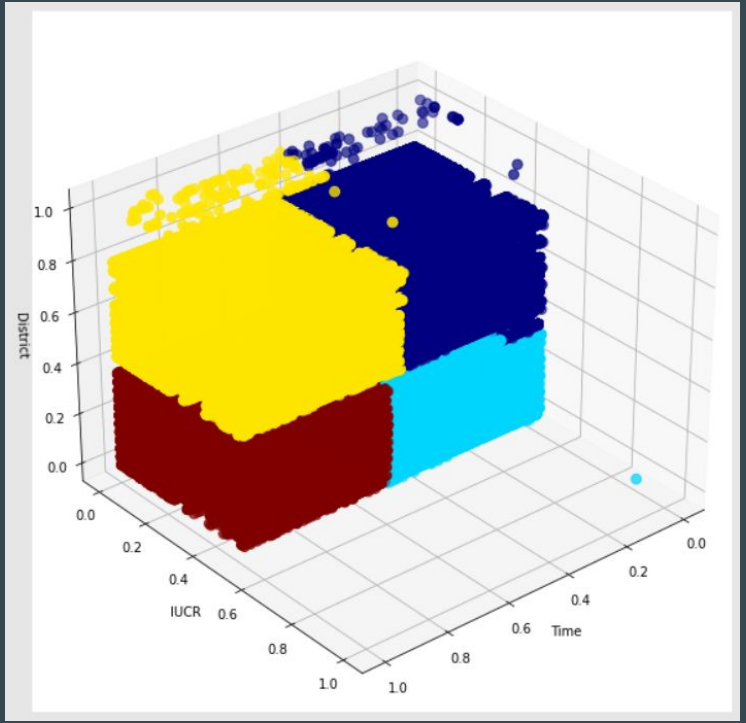
Crimes related to sex offense were highest in 2020.

Ritualism cases were only in 2021.

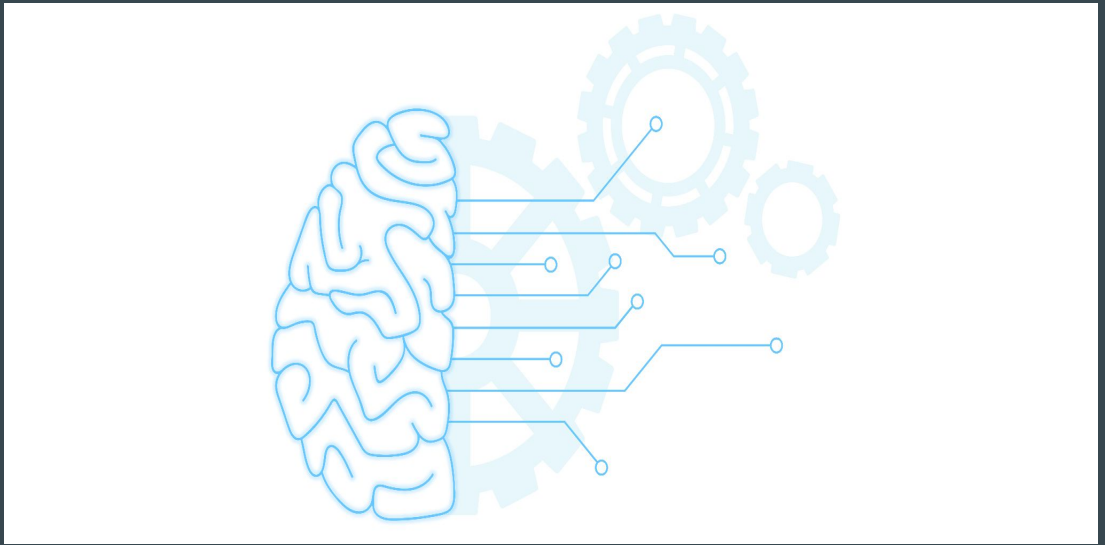


- First plot represents 3D Clustering with each point corresponding to their data values if District, IUCR and Ward.
- Second plot represents 3D Clustering with each point corresponding to their data values if District, IUCR and Time.





Machine Learning Model to Predict Arrest



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
```

Training accuracy for DecisionTree Classifier :

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

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
```

Prediction accuracy of DecisionTree Classifier :

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

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
```

Conclusion

This project offers visualizations for users to better interpolate data and understand the trend going on for 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 has been showing for past years. These are not absolute rule 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 machine learning decision tree classifier presents an estimated value as to whether the arrest made is True or False. This is reached by creating a decision tree based on the principal attributes in the dataset.