Coursera Data Science CapStone Project

The following notebook is made for the description of the Capstone project. It is part of the IBM Data Science Professional Certification.

Introduction

This project is part of IBM's Data Science Professional Course.

Violent crimes, thefts, and robberies occur in majority of places. On daily basis, local businesses, houses etc encounter such crimes. Our job as data scientist is to analyze the pattern of crimes. study the distribution of types of crime that occur in various part of a particular place. It is important to know where the crimes are happening, usually areas with police stations farther in distance can result into increase crime count.

what's the use case for this? well, first of all, everyone is always wanting to look for areas that are very safe. It is very important even for a tourist to know, which part of the city they are traveling to, are dangerous, and to avoid them. keeping that idea in mind, I decided to perform some data science on the criminal activities in the chicago city area, and try to point out the distribution of types of crimes and their various corelation and numbers.

In this case, I decided to perform statistical analysis and data science on the crimes in the city of chicago.

We will also compare the location of local police station in the area, and their relation with the location of the crimes.

Data Understanding

Descrption of Data and how it will be used to solve the problem.

Here in this project, I will utilize the crime dataset & police stations provided by Chicago Data Portal.

Furthermore we will use the Foursquare API to fetch the venues to see in what specific type of environment or location does these crimes take place.

Crime Data Set

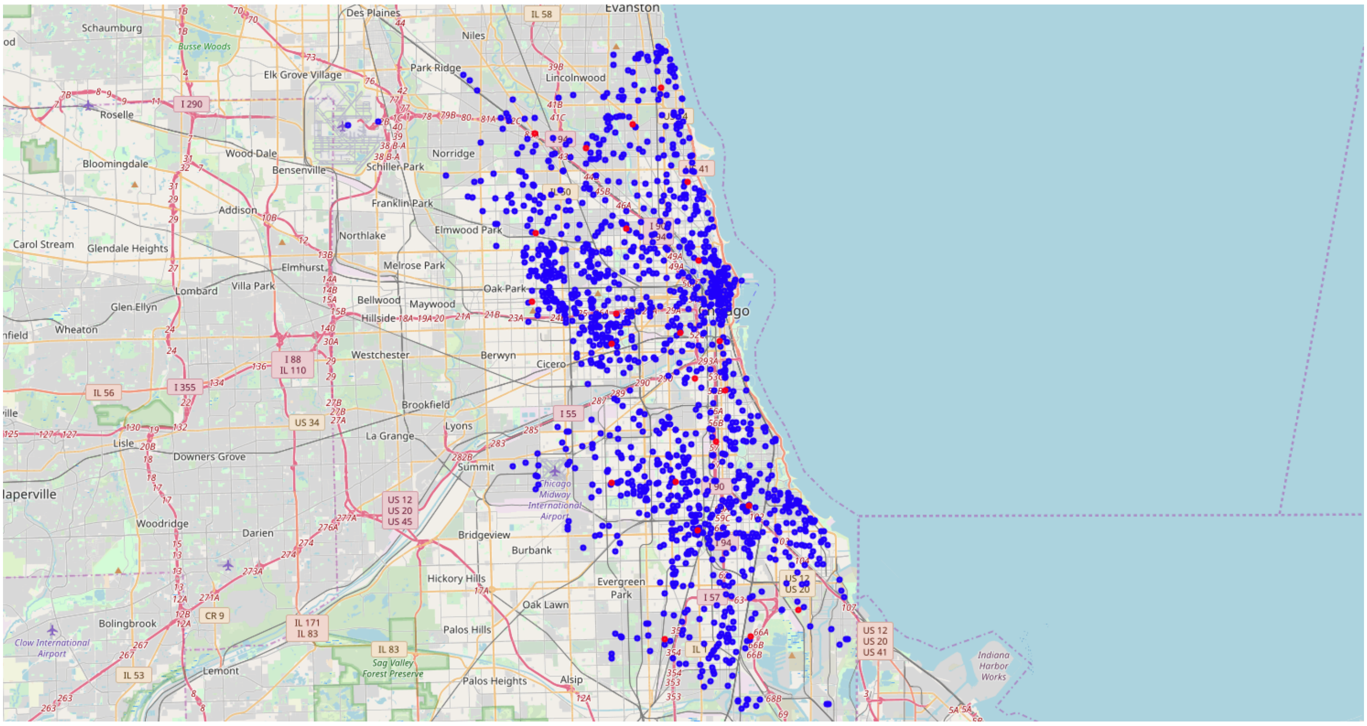
The crime data set consist of reported incidents of crime that occurred in the City of Chicago from 2018, The particular data set is subset of a bigger crime data with over 6.9 Million fields which would require greater computation power.. Data is extracted from the Chicago Police Department's CLEAR (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 dataset consist of **22 fields** and over **226K rows**. The fields include attributes related to crime such as *Case number, block, IUCR codes, Description, Type of Crime, location, Lattitude,* *Longitude, district.*

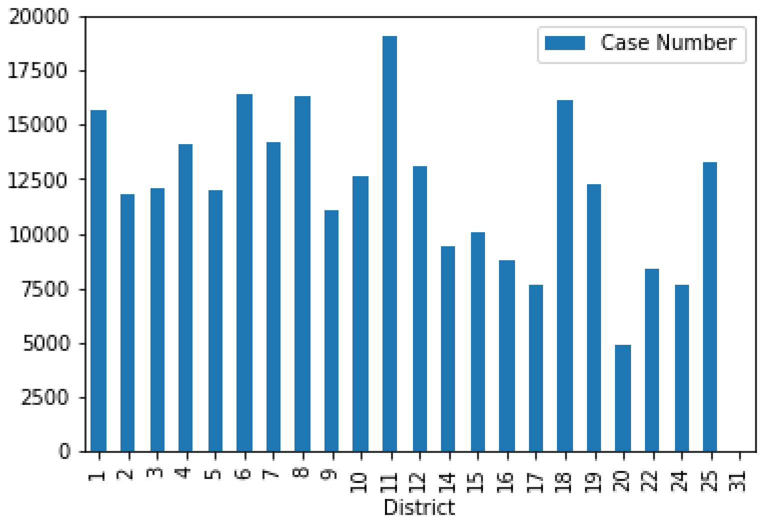
Since there are many fields that are not required in our analysis, we will require a lot of cleaning to do.

Methodology/Discussion

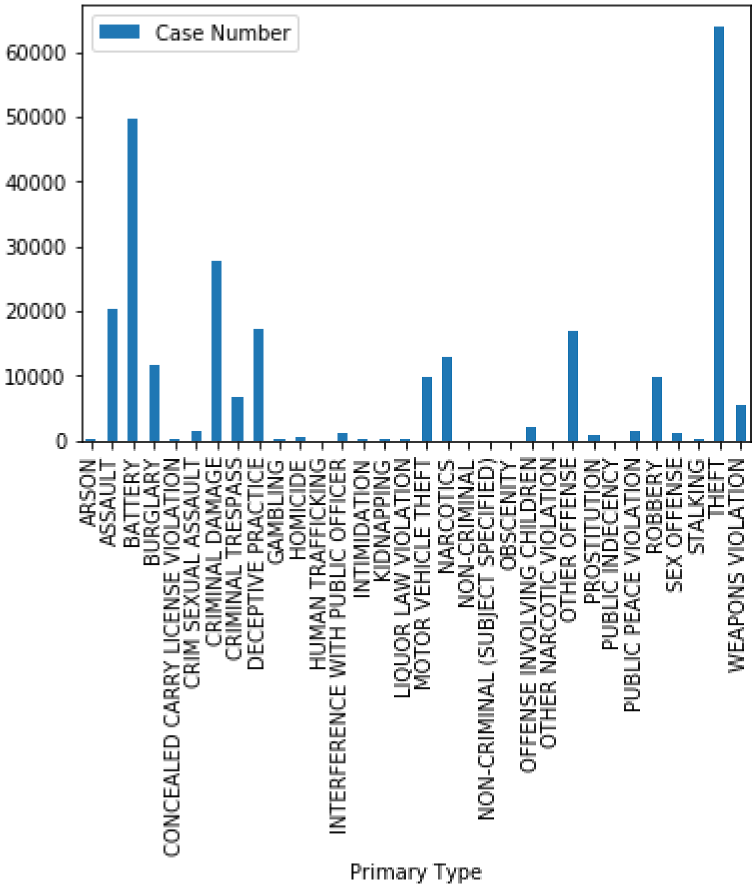
In this section we will discuss about the various operations that were performed on the dataset. Here in the image, we can see that the chicago area is distributed into 23 Districts, represented by the red dot, which are also the location of the police departments in each district. The blue dots scatter represents the location of registered crime. Be sure that our data set being very huge, we decided to plot sample of only 5% fraction of the data.



We start of the analysis by top level district v. num cases. A simple bar graph suffice the requirement. We can see in the graph that **District 11** has most registered cases followed by District 1 and District 18.

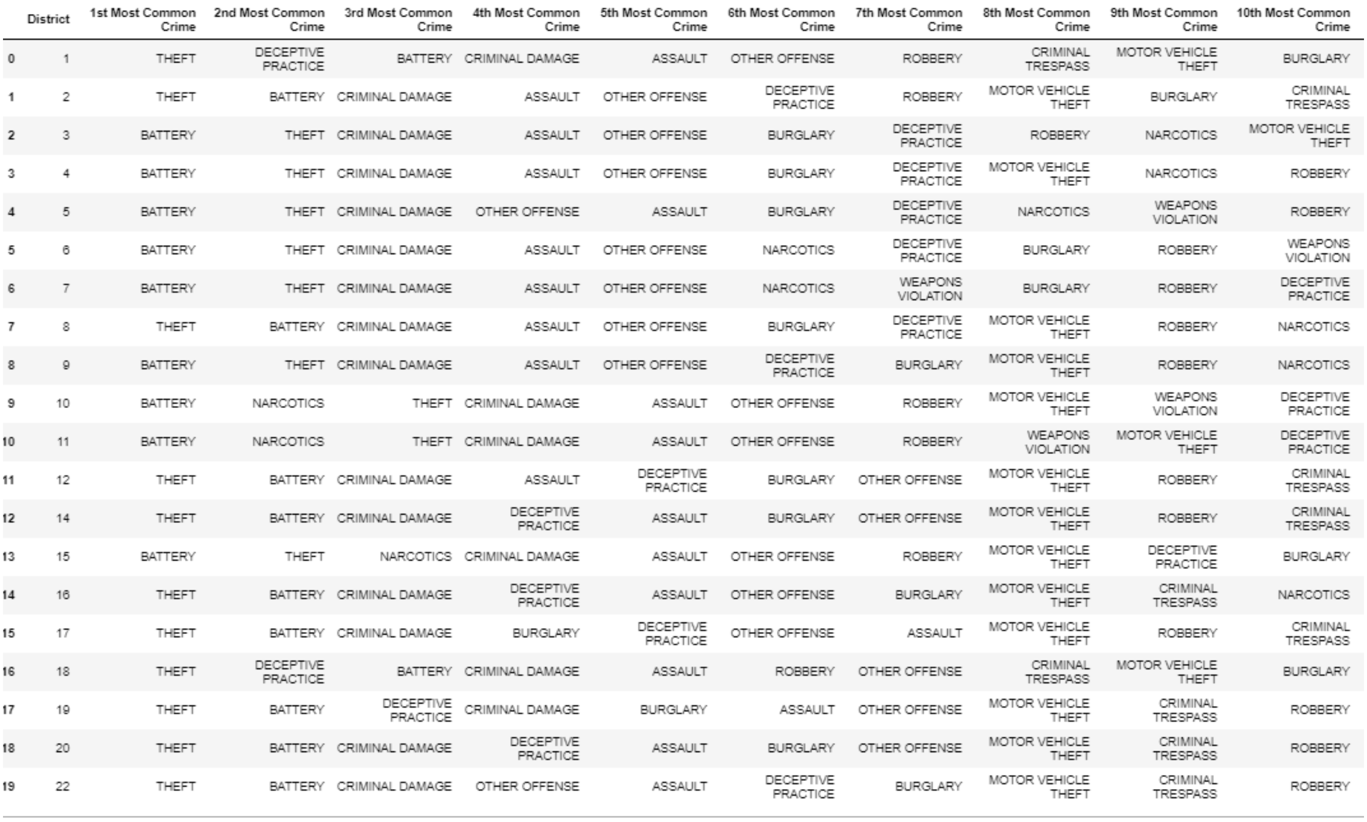


Another important thing to keep in mind, is the severity of crimes. Crimes like theft, Gambling on one side, and Homocide, rapes on other extreme end. This would provide a better inside, to which area to avoid. for the same, we developed a graph that shows number of registered cases for various type of crime



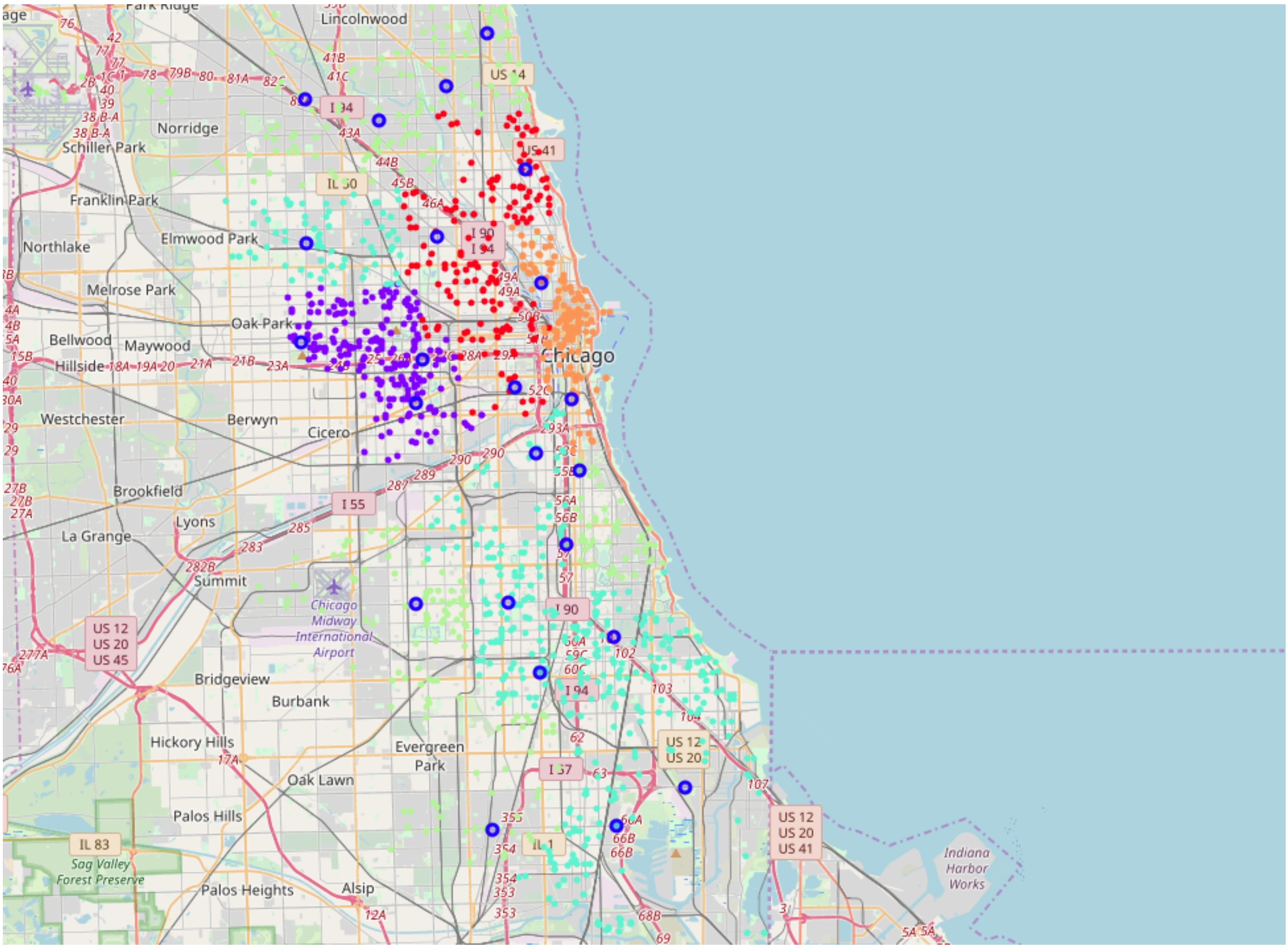
Moving on the same line. We start studying each type of crime individually for each district. This is where our main clustering model would help us from a group.

I used one hot encoding to aggregate a mean for every type of crime, in each individual district. Once that is done, I used sorting and filtering to get the top 10 most occuring type of crime in each district, and formed a new data frame that looks like this.



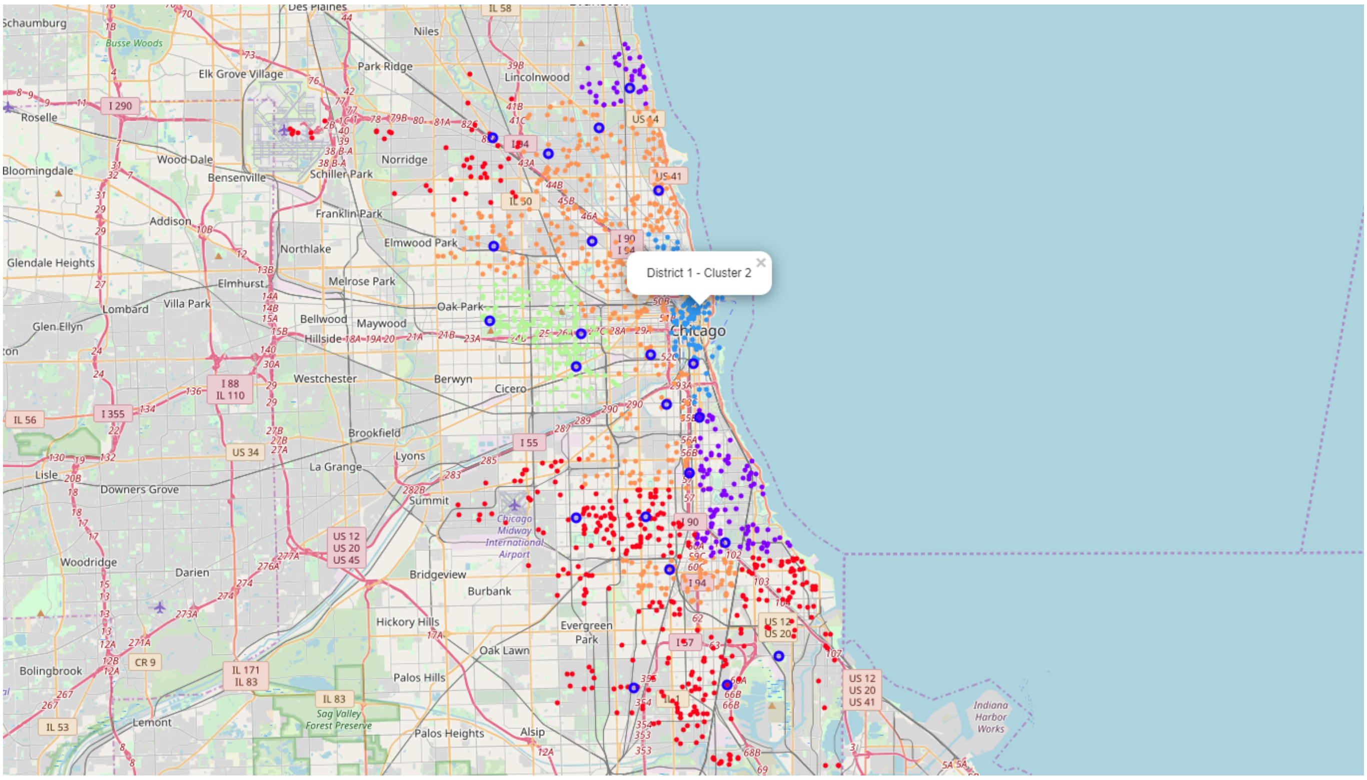
	District	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime	6th Most Common Crime	7th Most Common Crime	8th Most Common Crime	9th Most Common Crime	10th Most Common Crime
0	1	THEFT	DECEPTIVE PRACTICE	BATTERY	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	ROBBERY	CRIMINAL TRESPASS	MOTOR VEHICLE THEFT	BURGLARY
1	2	THEFT	BATTERY	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	DECEPTIVE PRACTICE	ROBBERY	MOTOR VEHICLE THEFT	BURGLARY	CRIMINAL TRESPASS
2	3	BATTERY	THEFT	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	BURGLARY	DECEPTIVE PRACTICE	ROBBERY	NARCOTICS	MOTOR VEHICLE THEFT
3	4	BATTERY	THEFT	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	BURGLARY	DECEPTIVE PRACTICE	MOTOR VEHICLE THEFT	NARCOTICS	ROBBERY
4	5	BATTERY	THEFT	CRIMINAL DAMAGE	OTHER OFFENSE	ASSAULT	BURGLARY	DECEPTIVE PRACTICE	NARCOTICS	WEAPONS VIOLATION	ROBBERY
5	6	BATTERY	THEFT	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	NARCOTICS	DECEPTIVE PRACTICE	BURGLARY	ROBBERY	WEAPONS VIOLATION
6	7	BATTERY	THEFT	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	NARCOTICS	WEAPONS VIOLATION	BURGLARY	ROBBERY	DECEPTIVE PRACTICE
7	8	THEFT	BATTERY	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	BURGLARY	DECEPTIVE PRACTICE	MOTOR VEHICLE THEFT	ROBBERY	NARCOTICS
8	9	BATTERY	THEFT	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	DECEPTIVE PRACTICE	BURGLARY	MOTOR VEHICLE THEFT	ROBBERY	NARCOTICS
9	10	BATTERY	NARCOTICS	THEFT	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	ROBBERY	MOTOR VEHICLE THEFT	WEAPONS VIOLATION	DECEPTIVE PRACTICE
10	11	BATTERY	NARCOTICS	THEFT	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	ROBBERY	WEAPONS VIOLATION	MOTOR VEHICLE THEFT	DECEPTIVE PRACTICE
11	12	THEFT	BATTERY	CRIMINAL DAMAGE	ASSAULT	DECEPTIVE PRACTICE	BURGLARY	OTHER OFFENSE	MOTOR VEHICLE THEFT	ROBBERY	CRIMINAL TRESPASS
12	14	THEFT	BATTERY	CRIMINAL DAMAGE	DECEPTIVE PRACTICE	ASSAULT	BURGLARY	OTHER OFFENSE	MOTOR VEHICLE THEFT	ROBBERY	CRIMINAL TRESPASS
13	15	BATTERY	THEFT	NARCOTICS	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	ROBBERY	MOTOR VEHICLE THEFT	DECEPTIVE PRACTICE	BURGLARY
14	16	THEFT	BATTERY	CRIMINAL DAMAGE	DECEPTIVE PRACTICE	ASSAULT	OTHER OFFENSE	BURGLARY	MOTOR VEHICLE THEFT	CRIMINAL TRESPASS	NARCOTICS
15	17	THEFT	BATTERY	CRIMINAL DAMAGE	BURGLARY	DECEPTIVE PRACTICE	OTHER OFFENSE	ASSAULT	MOTOR VEHICLE THEFT	ROBBERY	CRIMINAL TRESPASS
16	18	THEFT	DECEPTIVE PRACTICE	BATTERY	CRIMINAL DAMAGE	ASSAULT	ROBBERY	OTHER OFFENSE	CRIMINAL TRESPASS	MOTOR VEHICLE THEFT	BURGLARY
17	19	THEFT	BATTERY	DECEPTIVE PRACTICE	CRIMINAL DAMAGE	BURGLARY	ASSAULT	OTHER OFFENSE	MOTOR VEHICLE THEFT	CRIMINAL TRESPASS	ROBBERY
18	20	THEFT	BATTERY	CRIMINAL DAMAGE	DECEPTIVE PRACTICE	ASSAULT	BURGLARY	OTHER OFFENSE	MOTOR VEHICLE THEFT	CRIMINAL TRESPASS	ROBBERY
19	22	THEFT	BATTERY	CRIMINAL DAMAGE	OTHER OFFENSE	ASSAULT	DECEPTIVE PRACTICE	BURGLARY	MOTOR VEHICLE THEFT	CRIMINAL TRESPASS	ROBBERY

based on the normalized one hot encoding that I created, I performed the Kmeans clustering to fit the data. Since there were many attributes its hard to comprehend the dimensionality on a graph, so we assign the label to each cluster, and then map those districts back to the crime dataset. using the lattitude and longitude of each crime report, we generate a plot on folium map which distinct color based on the cluster.



As you can see, the clusters have formed according to their proximity, and similarities in the type of crime in each district. Important thing to note is, the orange cluster which is the "downtown" region of Chicaog doesnt have any police station in the center, and also that the crimes reported in that area are high in number for a small region which makes it dense.

Keeping the same concept in mind, we repeat the same process of clustering, but this time, we try to find the relationship between the location of the crime, ie Apartment, street, in car etc, and the district.



The following image shows the distribution of clusters on the map. Now this is different that the previous map. however the clusters seem to makes sense, since they are distrutied according the type of places each areas consist of. for instance, the light green colored cluster has low income middle class houses and apartment, which makes them vurnerable assualt, gambling etc.

Result

- There exist a strong relation between the type of crime, and the location.
- clusters as they are constructed show different group of district formed with which we can make a prediction of whether its a good area to be in or not.
- District 11 has the highest recorded cases in the year 2018.

Conslusion

- In this project we analyzed various distribution of crimes in the chicago city area. The city being divided into 23 districts became the key driver of the distribution.
- As we can see there is a strong co-relation of the clusters along with its primary type as well as location of the crime that is taking place