Chicago City Neighborhood Crime Analysis

Shafiul Azam

INTRODUCTION

It is often important for Law Enforcement Agency to identify the crime rate and pattern in an area to identify proper strategies: enforcement, education, campaigning etc.

Chicago has violent crime about 49% higher than the average rate of crime in Illinois and 164% higher than the rest of the nation (https://www.areavibes.com/chicago-il/crime/). Therefore, it is often important for any potential home buyers or renters do some study and analyze the crime data before making decisions of a safe neighborhood.

In this study we will investigate a cluster analysis of the Chicago Neighborhood to find out the typical crime types, frequency etc. The analysis can also help someone a potential buyer to choose a relatively safe neighborhood.

DATA SOURCE AND DESCRIPTION

The following two primary data sources were used for this study.

1. Neighborhood Information was available from the following wiki page: (https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago)

2. Crime information was available from Chicago Data Portal

(https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6). The data was exported in csv format and imported into the Python data-frame for analysis. The file has all crimes for year January 5, 2019 until January 4, 2020

The crime report shows the following salient fields:

1. Case #
2. Date of Occurrence
3. Approximate Block Address.
4. Primary Description of the crime (i.e. Burglary, Battery, Theft etc.)
5. Secondary Description of the Crime
6. Location Description
7. Arrest made (Y/N)?
8. Domestic Violence (Y/N)?
9. GPS Point (Latitude and Longitude)

METHODOLOGY

In order to analyze the crime data, the following methodologies were considered:

1. Mapping the Neighbor - Mapping the neighborhood in Chicago was done using the follium library. The mapping shows spatial distribution of the neighborhoods.
2. Crime Data Import and Data Wrangling - The crime data was downloaded as csv file and imported into a Pandas data frame.
3. Exploratory Analysis of Crime Data: The crime data was then considered for exploratory analysis. The following analysis were considered for exploration:
    - i) Distribution of Primary Crime Types
    - ii) Distribution of Primary Crime Types by Arrest made
    - iii) Crime Distribution by Months of Year 2019.
    - iv) Distribution of Crimes by Days of Week
4. Clustering of Neighborhoods using K-means clustering

Mapping Chicago Neighborhood

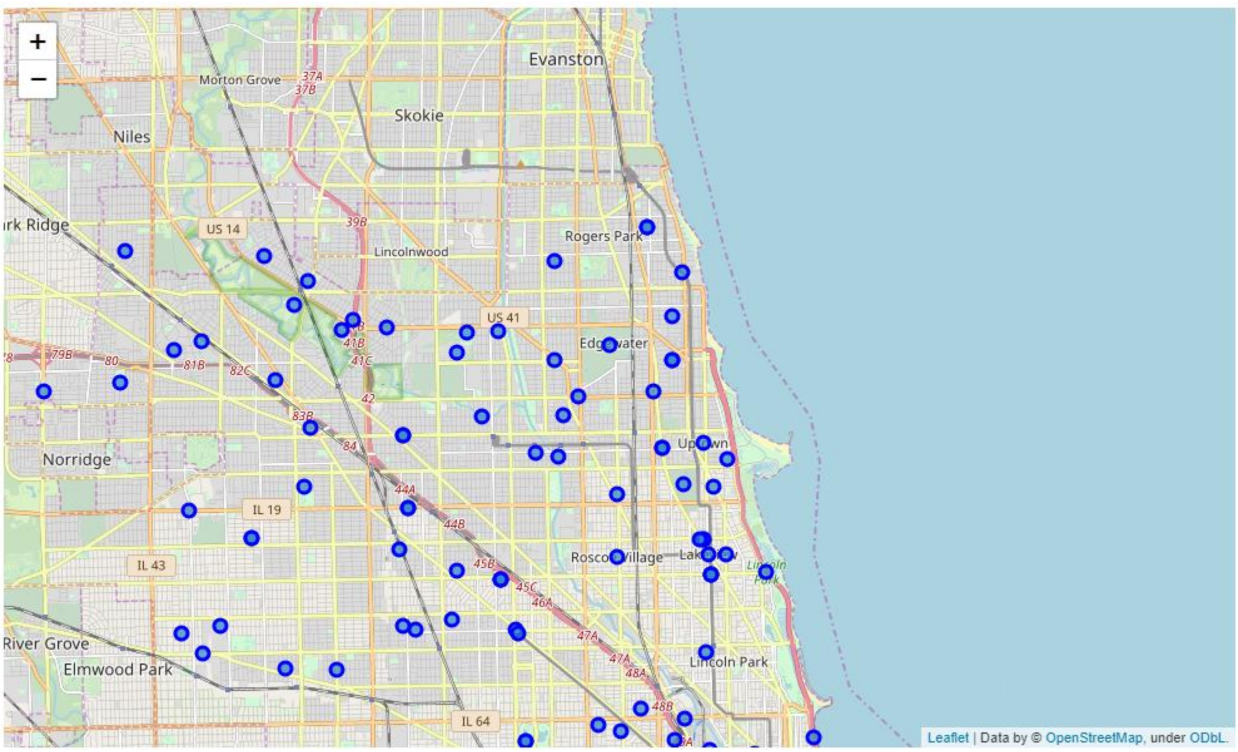

Figure 1 Map of Chicago Neighborhood Centers

Data Wrangling: Importing Crime Data and Exploratory Analysis

A discussed earlier the crime data was available from Chicago Data Portal

(https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6). The data was exported in csv format and imported into the Python data-frame for analysis. The file has all crimes for year January 5, 2019 until January 4, 2020. A snapshot of the data is shown below.

Table 1 Crime Data from City Portal

	CASE#	DATE OF OCCURRENCE	BLOCK	IUCR	PRIMARY DESCRIPTION	SECONDARY DESCRIPTION	LOCATION DESCRIPTION	ARREST	DOMESTIC	BEAT	WARD	FBI CD	C
256484	JC356571	07/19/2019 09:00:00 PM	031XX W HARRISON ST	5002	OTHER OFFENSE	OTHER VEHICLE OFFENSE	GOVERNMENT BUILDING/PROPERTY	N	N	1134	24.0	26	
256485	JC318635	06/23/2019 04:40:00 PM	005XX N CENTRAL AVE	0460	BATTERY	SIMPLE	RESIDENCE PORCH/HALLWAY	N	N	1523	37.0	08B	
256486	JC405144	08/23/2019 11:00:00 PM	019XX W SCHILLER ST	0810	THEFT	OVER $500	STREET	N	N	1424	1.0	06	
256487	JC379670	08/05/2019 12:30:00 PM	074XX S HARVARD AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	N	Y	731	6.0	08B	
256488	JC316226	06/18/2019 04:00:00 PM	034XX W 71ST PL	0610	BURGLARY	FORCIBLE ENTRY	RESIDENCE	N	N	831	17.0	05	

A preliminary data analysis shows the following distribution for primary crime type during the analysis period.

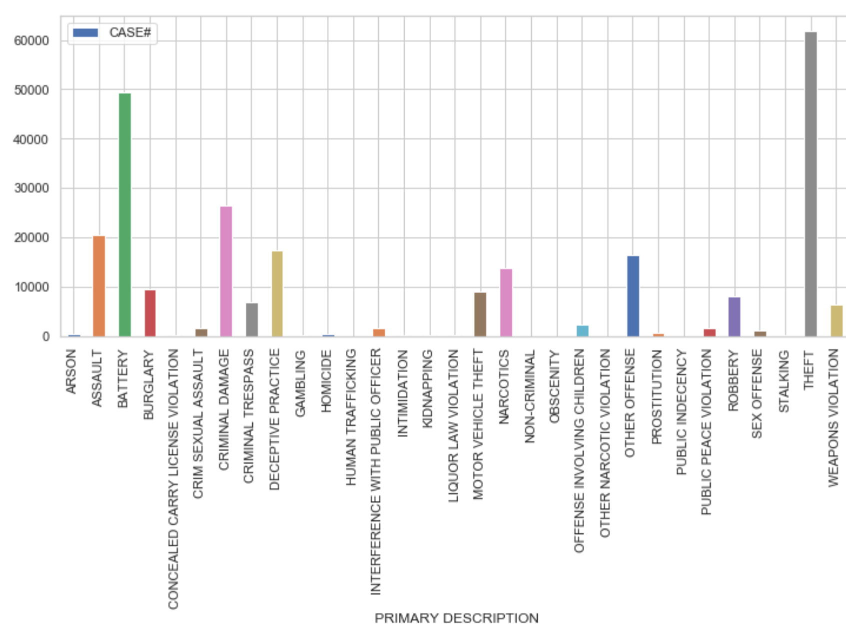

Figure 2 Distribution of Crime Types in Year 2019 in Chicago Neighborhoods

It is also important to know how many of these crimes have eventually end up with arrests done by law-enforcement agency. The following figure shows the distribution. It is evident that almost all Narcotics related crime ended up having arrest. However, for battery or theft the percentage of arrests being made are pretty low.

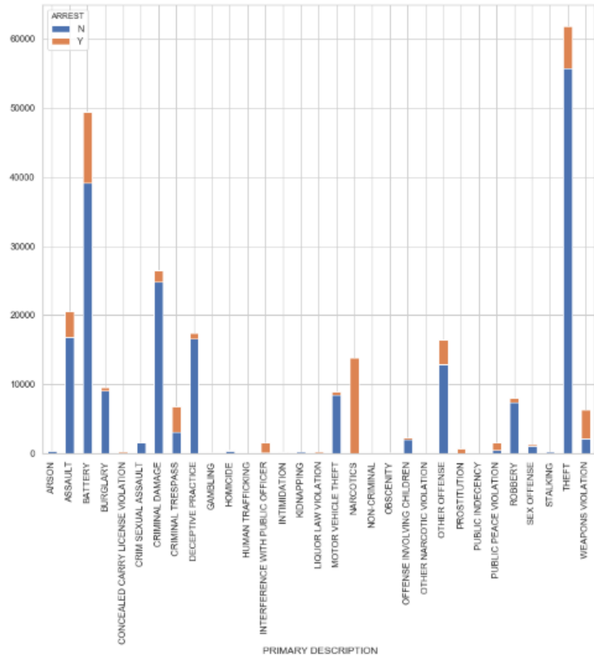

Figure 3 Distribution of Crime Types by Arrest Made

Crime records also showed variation of crimes across time. For example, the months of July and August have typically more crimes as shown by the figure below.

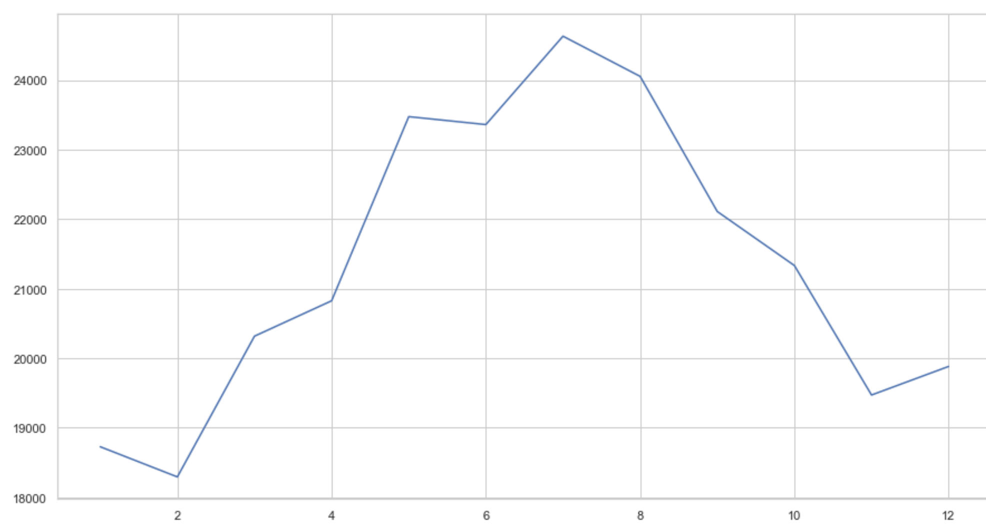

Figure 4 Number of Crimes in Chicago Neighborhoods by Months in 2019

However, the distribution of crime by day of weeks do not show any noticeable differences.

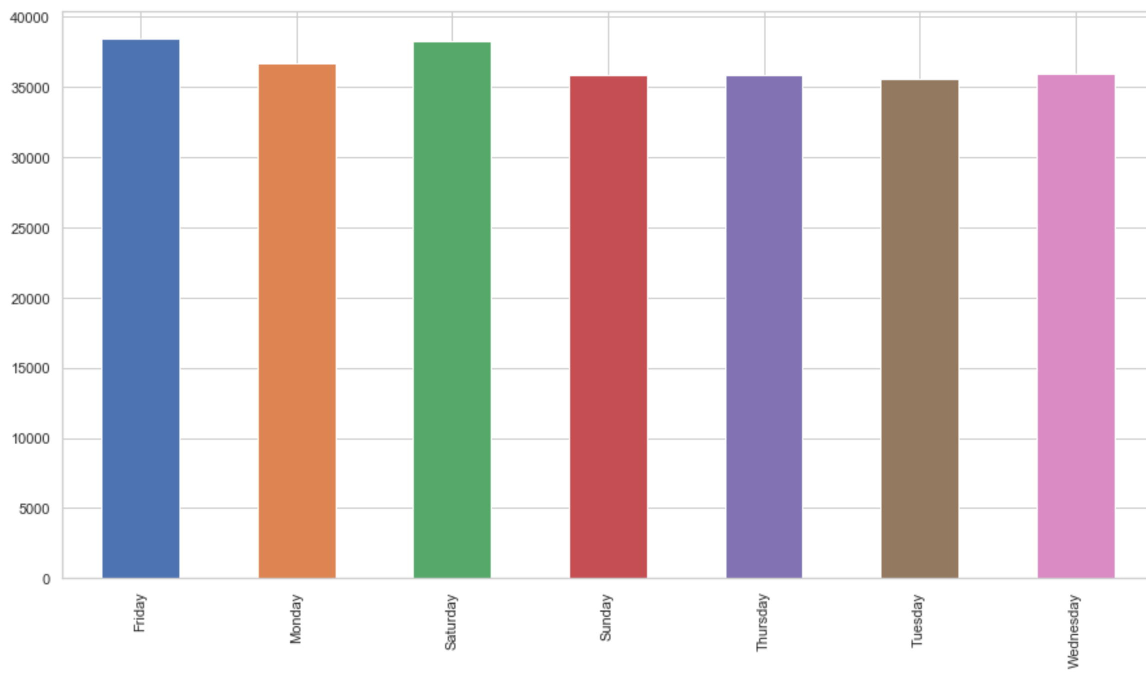

Figure 5 Number of Crimes by Days of Week

Linking Crime Data to Neighborhood

In order to link the crime data to neighborhood, the following steps were followed:

1. Each neighborhood address was used to find corresponding GPS point by using the geopy package. A snapshot of records after latitude and longitude are added is shown below:

	Neighborhood	Community Areas	Latitude	Longitude
0	Albany Park	Albany Park	41.971937	-87.716174
1	Altgeld Gardens	Riverdale	41.654864	-87.600446
2	Andersonville	Edgewater	41.977139	-87.669273
3	Archer Heights	Archer Heights	41.811422	-87.726165
4	Armour Square	Armour Square	41.840033	-87.633107
5	Ashburn	Ashburn	41.747533	-87.711163
7	Auburn Gresham	Auburn Gresham	41.743387	-87.656042

2. A function was created to calculate the distance from each crime site to the neighborhood centroid. If the distance from the crime site is less that 1 mile then the nearest neighborhood was assigned to the crime report. A snapshot of the function is shown below:

```
def cal_distance(lat, long):
    distance_dic ={}
    for index, row in df.iterrows():
        coords_1 = (lat, long)
        coords_2 = (row["Latitude"], row["Longitude"])
        distance = geopy.distance.vincenty(coords_1, coords_2).miles
        distance_dic.update ({index:distance})
        #print (row["Latitude"], row["Longitude"],distance)
    min_value = min(distance_dic.values())  # maximum value
    min_keys = [k for k, v in distance_dic.items() if v == min_value]
    if (min_value<=1):
        return str(df.loc[min_keys[0]]['Neighborhood']),min_value
        #return min_keys[0],min_value

    else:
        return -1, -1
```

3. Finally, the package "one hot coding" was used to create dummy variable for primary crime type. The snapshot of the new dataframe is shown below:

	Neighborhood	ARSON	ASSAULT	BATTERY	BURGLARY	CONCEALED CARRY LICENSE VIOLATION	CRIM SEXUAL ASSAULT	CRIMINAL DAMAGE	CRIMINAL TRESPASS	DECEPTIVE PRACTICE
46651	South Shore	0	0	0	0	0	0	0	0	0
175708	East Garfield Park	0	0	0	0	0	0	0	0	0
121870	Hamilton Park	0	0	0	1	0	0	0	0	0
188528	Palmer Square	0	0	0	0	0	0	0	0	0
156444	Chatham	0	1	0	0	0	0	0	0	0

Clustering Neighborhoods with Crime Data

The basis of the K-neighborhood analysis was to crime types so that neighborhood with similar crime trends can be clustered together. Before applying the K-neighbors classifier, the crime data was used to display the distribution of the crime across any neighborhood. For example, for Andersonville neighborhood the distribution of the crime is shown below.

```
----Andersonville----
           Neighborhood  freq
0                 THEFT  0.33
1         OTHER OFFENSE  0.13
2       CRIMINAL DAMAGE  0.13
3               BATTERY  0.10
4              BURGLARY  0.07
5     CRIMINAL TRESPASS  0.07
6   MOTOR VEHICLE THEFT  0.07
7               ASSAULT  0.07
8     DECEPTIVE PRACTICE  0.03
9                 ARSON  0.00
```

Then the dataframe was modified to list up to 10 common crime types for each neighborhood. A snapshot of the modified dataframe is shown below.

Table 2 Chicago Neighborhood by Most Common Crime Types

	Neighborhood	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
1	Albany Park	THEFT	BATTERY	CRIMINAL DAMAGE	ASSAULT	ROBBERY	WEAPONS VIOLATION	NARCOTICS	BURGLARY	DECEPTIVE PRACTICE	MOTOR VEHICLE THEFT
2	Altgeld Gardens	BATTERY	ASSAULT	CRIMINAL DAMAGE	OTHER OFFENSE	THEFT	MOTOR VEHICLE THEFT	WEAPONS VIOLATION	DECEPTIVE PRACTICE	NARCOTICS	CRIMINAL TRESPASS
3	Andersonville	THEFT	CRIMINAL DAMAGE	OTHER OFFENSE	BATTERY	ASSAULT	BURGLARY	CRIMINAL TRESPASS	MOTOR VEHICLE THEFT	DECEPTIVE PRACTICE	WEAPONS VIOLATION
4	Archer Heights	THEFT	CRIMINAL DAMAGE	ASSAULT	BATTERY	OTHER OFFENSE	BURGLARY	CRIMINAL TRESPASS	DECEPTIVE PRACTICE	MOTOR VEHICLE THEFT	WEAPONS VIOLATION
5	Armour Square	OTHER OFFENSE	THEFT	ASSAULT	DECEPTIVE PRACTICE	BATTERY	OFFENSE INVOLVING CHILDREN	WEAPONS VIOLATION	INTIMIDATION	BURGLARY	CONCEALED CARRY LICENSE VIOLATION

Finally, an initial K value of 10 was chosen to apply the K-neighborhood clustering technique. The map below displays the 10 identified clusters.

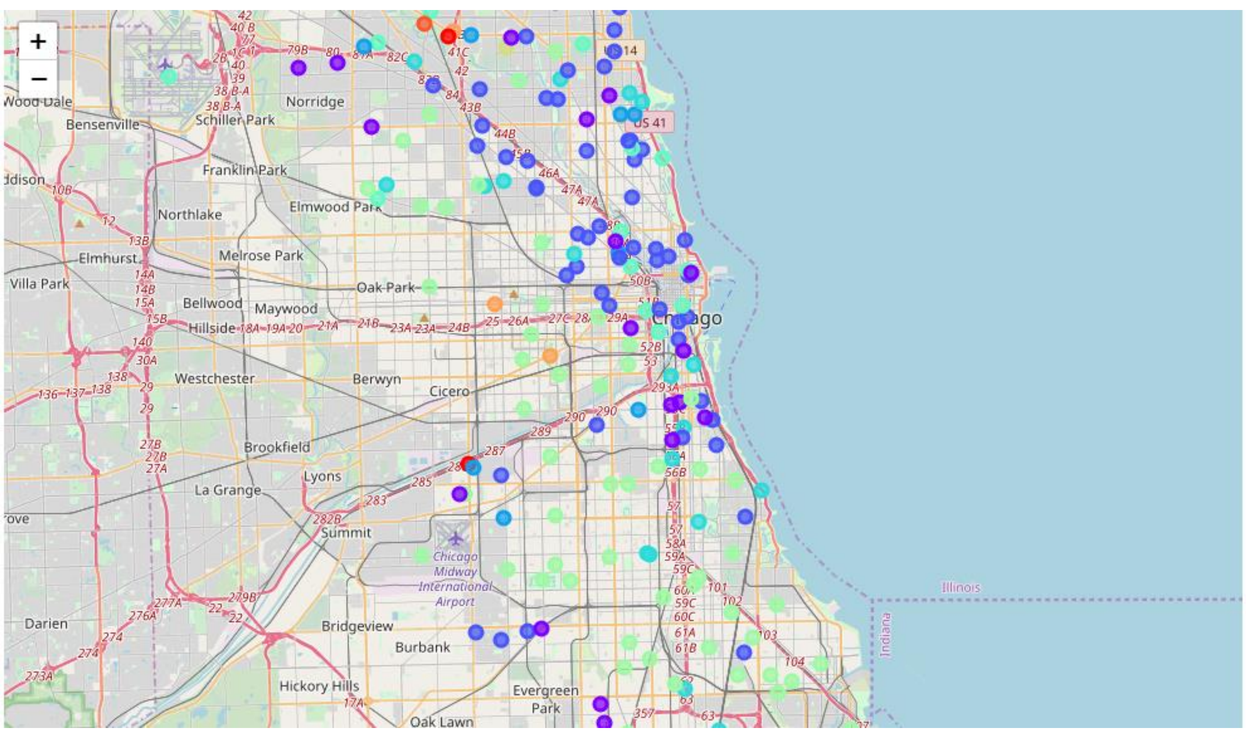

Figure 6 Distribution of K-neighborhood Clusters based on Crime Types

RESULTS

Based on the K-means cluster analysis the results can be summarized as below:

Individual Clusters are obtained with the Top 10 Most Common Crimes. For example, the cluster 1 and 3 has the following crime distributions.

Cluster 1

	Community Areas	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	8t
111	Garfield Ridge	WEAPONS VIOLATION	BURGLARY	CRIMINAL DAMAGE	OTHER OFFENSE	KIDNAPPING	ASSAULT	BATTERY	
204	Forest Glen	BURGLARY	OTHER OFFENSE	WEAPONS VIOLATION	KIDNAPPING	ASSAULT	BATTERY	CONCEALED CARRY LICENSE VIOLATION	

Cluster 3

	Community Areas	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	Co
2	Edgewater	THEFT	CRIMINAL DAMAGE	OTHER OFFENSE	BATTERY	ASSAULT	BURGLARY	CRIMINAL TRESPASS	MOTOR VEHICLE THEFT	DECEPTIVE PRACTICE	
3	Archer Heights	THEFT	CRIMINAL DAMAGE	ASSAULT	BATTERY	OTHER OFFENSE	BURGLARY	CRIMINAL TRESPASS	DECEPTIVE PRACTICE	MOTOR VEHICLE THEFT	
5	Ashburn	THEFT	OTHER OFFENSE	ASSAULT	BATTERY	MOTOR VEHICLE THEFT	INTERFERENCE WITH PUBLIC OFFICER	CRIMINAL DAMAGE	OFFENSE INVOLVING CHILDREN	WEAPONS VIOLATION	
9	Avondale	THEFT	BATTERY	DECEPTIVE PRACTICE	CRIMINAL DAMAGE	ASSAULT	OTHER OFFENSE	ROBBERY	MOTOR VEHICLE THEFT	CRIMINAL TRESPASS	
12	Belmont Cragin	THEFT	DECEPTIVE PRACTICE	CRIMINAL DAMAGE	BATTERY	BURGLARY	MOTOR VEHICLE THEFT	OTHER OFFENSE	ROBBERY	NARCOTICS	

Also, a distribution of clusters by 1st Most Common Crime is graphically displayed below.

Figure 7 Distribution of Most Common Crimes by Crime Clusters

DISCUSSION

The results of the K-nearest neighborhood for crime reports in Chicago area shows some interesting facts about crime.

- July is the month when there are most crimes, followed by August.
- There is not much difference in days of week distribution for crimes.
- First of all, there are diverse distribution of crime rates in the neighborhood. For example, cluster#2 has around 53 neighborhoods that has Theft pretty common, on the other hand cluster#6 has battery and cluster#3 has significant Criminal damages. The data can help law enforcement agency concentrate of particular strategy to consider depending on the pattern of crimes.
- Cluster #6 has the highest neighborhood where Battery is predominant, followed by Theft.

CONCLUSION

The study shows an application of web scraping, geo-coding and unsupervised machine learning technique (i.e. K-nearest neighborhood analysis) to analyze and cluster residential neighborhood in terms of types and extensity of crimes. The results can help a home buyer or renter choosing a safer neighborhood or cluster. It can also help law enforcement agencies consider strategic planning to mitigate the crimes.

Reference:

1. Chicago Neighborhood Information: https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago
2. Crime information – Chicago Data Portal https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6
3. Area Vibes (https://www.areavibes.com/)