Exploratory Data Analysis on Crimes in Chicago Dataset 2001-2017

Submitted by: Mustafa Habeeb (mhabee2), Ibrahim Ayoob (iayoob2), Vidhyasagar Udayakumar (vudaya2)
The University of Illinois at Chicago

December 2nd, 2019

Our team will be exploring the Crimes in the Chicago data set while analyzing the crimes that are committed in the community areas of Chicago throughout various times of the day. We will be using classification and clustering techniques to produce our outputs and create visualizations of what crimes are trending throughout the years.

The datasets we will be using are from the Crimes in Chicago dataset from 2001- 2017. The data is available at https://www.kaggle.com/currie32/crimes-in-chicago.

To prepare the dataset we began to merge the csv files we had of the multiple years of crime in Chicago. Then, we came to the understanding that in order for the data to be more meaningful, the data must be altered. For starters, instead of simply having the time stamp in one column, we separated it out so then we can create a new column called 'TimePeriod'. The reason we created this column is to see in what time of day the most crimes are being committed. Based on a source found online, http://learnersdictionary.com/qa/parts-of-the-day-early-morning-late-morning-etc, we divided up into four parts. Morning (5:00 am to 11:59 am), Noon (12:00 pm to 4:59 pm), Evening (5:00 pm to 8:59 pm), and Night (9:00 pm to 4:59 am).

Next, we saw to it to create a new 'Primary Type' column called 'Primary Crime Type'. This column generalizes all the crimes (36 different crimes) into smaller groups (nine new groups) which can help us gain a better visual on the severity of this problem.

Furthermore, we grouped 'Community Area', 77 communities, into a new column called 'Major Sections'. This generalizes where the crime took place based on the nine major Chicago sides. This information can be found based on this link, https://upload.wikimedia.org/wikipedia/commons/2/24/Map_of_the_Community_Areas_and_%27Sides%27_of_the_City_of_Chicago.svg. By also decreasing the number of areas where we will be plotting our data, it will make it easier to read the data and determine the major crimes in that area such as theft, homicide, etc.

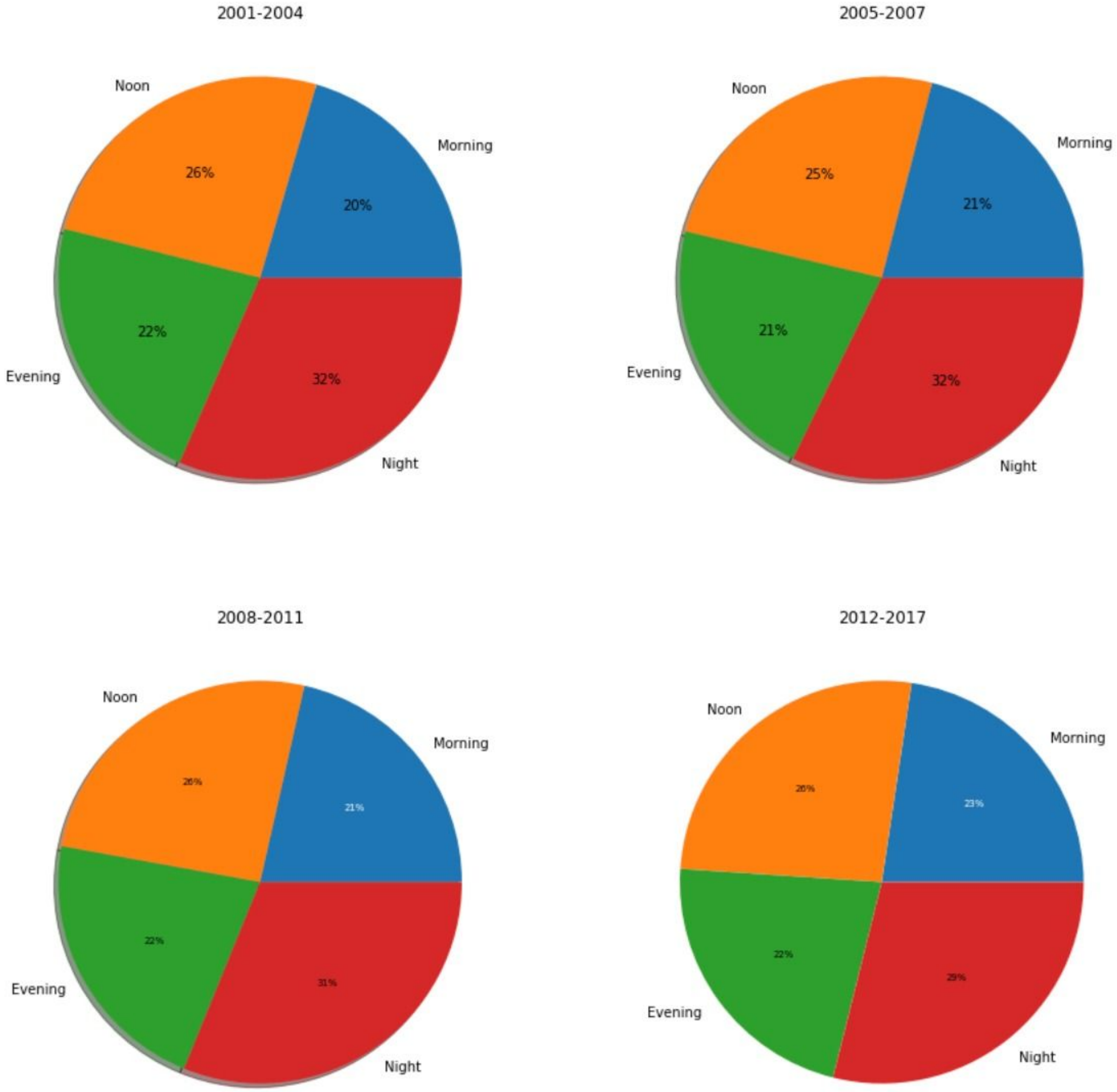
We finally got rid of unnecessary columns that we saw would not be useful based on the analysis were conduction such as, '#', 'ID', 'Case Number', 'Block', 'IUCR', 'Description', 'Domestic', 'Beat', 'Ward', 'FBI', 'X-Coordinate', 'Y-Coordinate', and

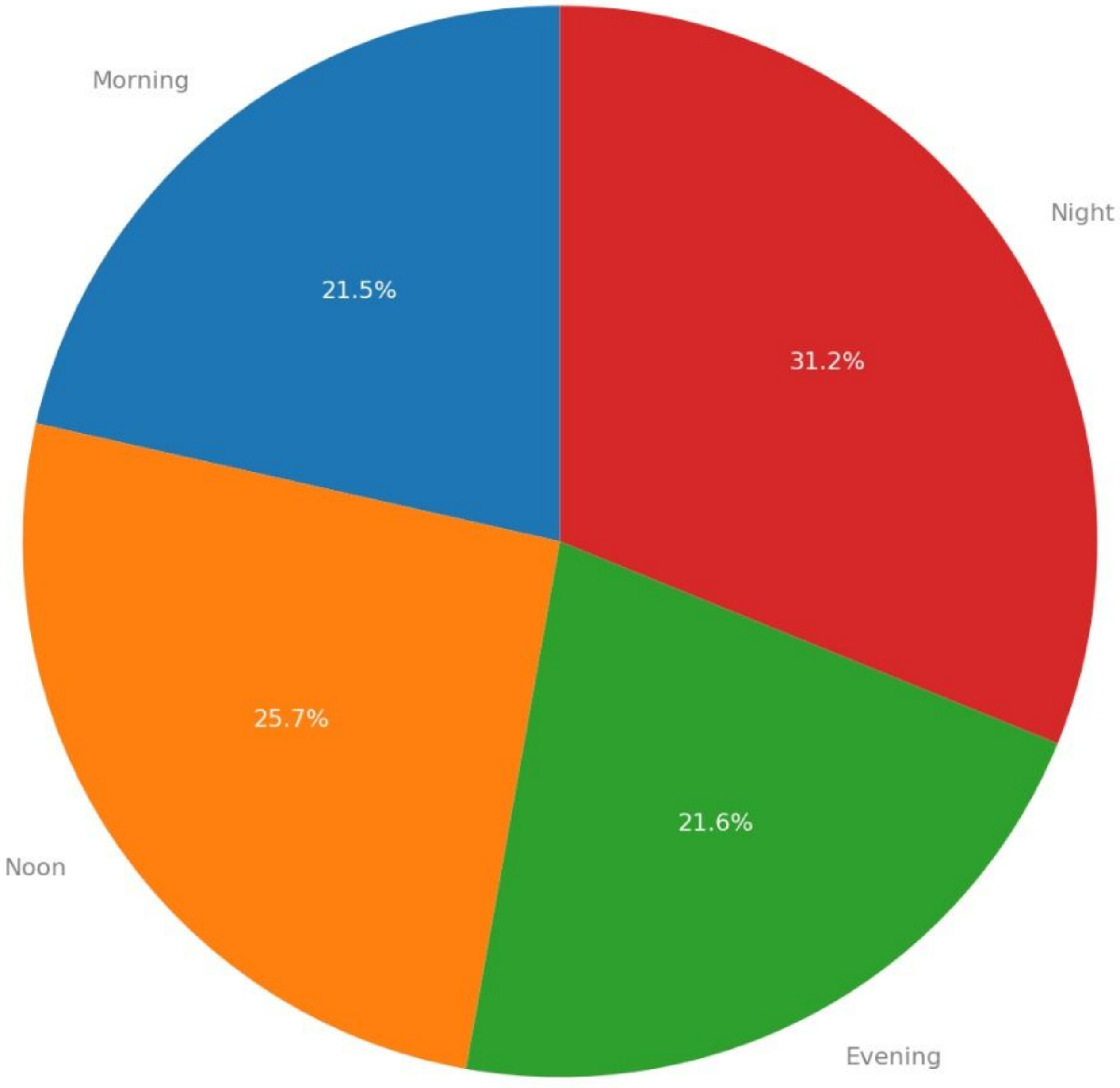
‘Update On’. These columns are not valued with our analysis because it does not help us classify or cluster the data.

Because of how large our data is, we will be decreasing the size ~~of our data~~ by taking every other 4th row of the CSV file so that we get 25% of the original dataset. We will also be ignoring NaN and missing values because if we estimate it, it will produce wrong values for where crimes are being produced in Districts, if Arrests occurred, etc. Essentially, estimating the values would hinder our results for this project.

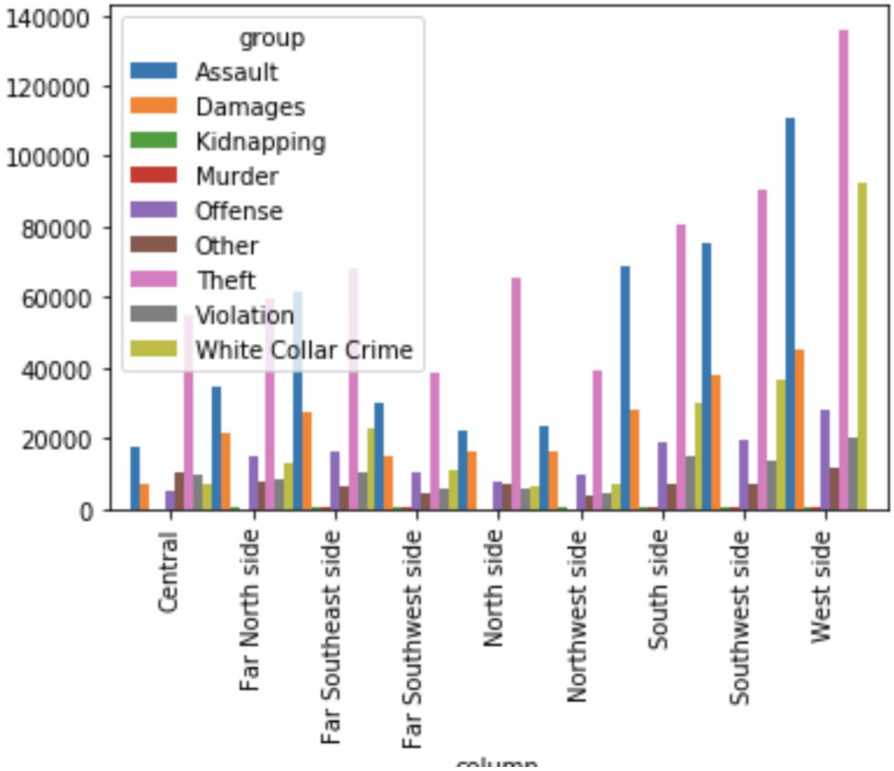
For our exploratory analysis, after considering what columns we will be using to properly show the trends of crimes throughout the years, we have decided to create some graphs to visualize crimes happening in different manners using the variables that we believe are most important.

a. We developed pie charts that would represent ‘Time of Day’ and Total crimes committed. In the first picture below, we can see the averages of the four CSV files provided and the statistics of the crimes being committed at certain times in a day. In the second picture following, we do a similar computation but this time of all the years totaled together from 2001 to 2017. As you can see, the data is very consistent with the averages showing little difference in the averages of when crimes are being committed throughout the day.

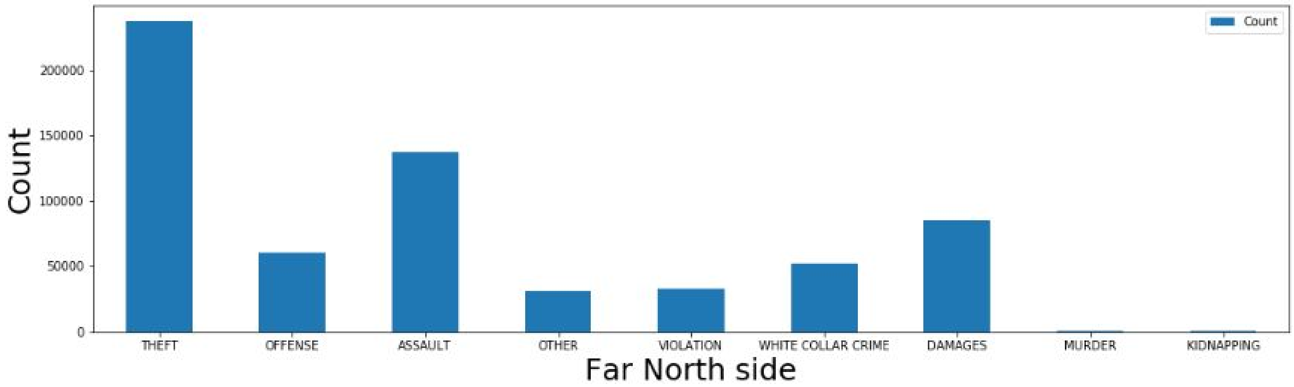

2001-2004
Noon
Morning
26%
20%
22%
32%
Evening
Night
2005-2007
Noon
Morning
25%
21%
21%
32%
Evening
Night
2008-2011
Noon
Morning
26%
21%
22%
31%
Evening
Night
2012-2017
Noon
Morning
26%
23%
22%
29%
Evening
Night


Morning
21.5%
Night
31.2%
25.7%
21.6%
Noon
Evening

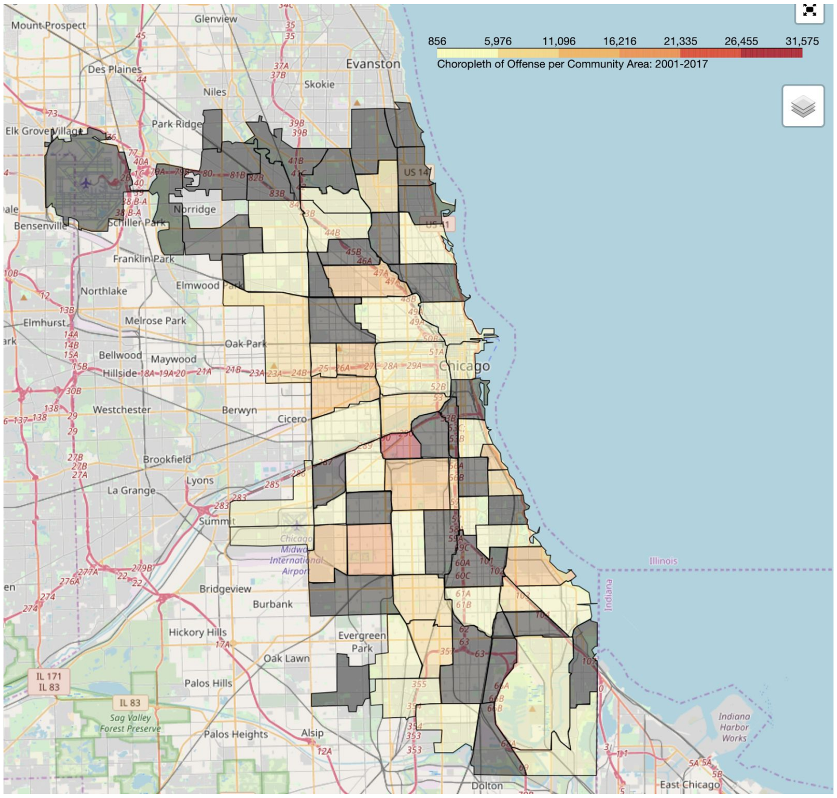
b. Another visualization that we created is the bar graphs of all the average crimes committed from 2001 to 2017. The ‘Primary Type’ column originally had 36 unique crimes listed but after regrouping the data and creating the new column, ‘Primary Crime Type’, it now has 9 unique crime values. Also, because there are many community areas in Chicago, 77 communities, we also generalize the communities into the major ‘sides’ of Chicago as explained above. As you can see below, bar graphs for each Chicago sideshow the average amount throughout the years of 2001 to 2017.



The same results are also shown for each individual Chicago ‘Sides’ are as well below. This is one of the Chicago Sides, specifically the Far North Side, that shows the average statistics of the crimes committed in that area from 2001 to 2017 in representation as a bar graph. (There are eight more graphs as so in our code).



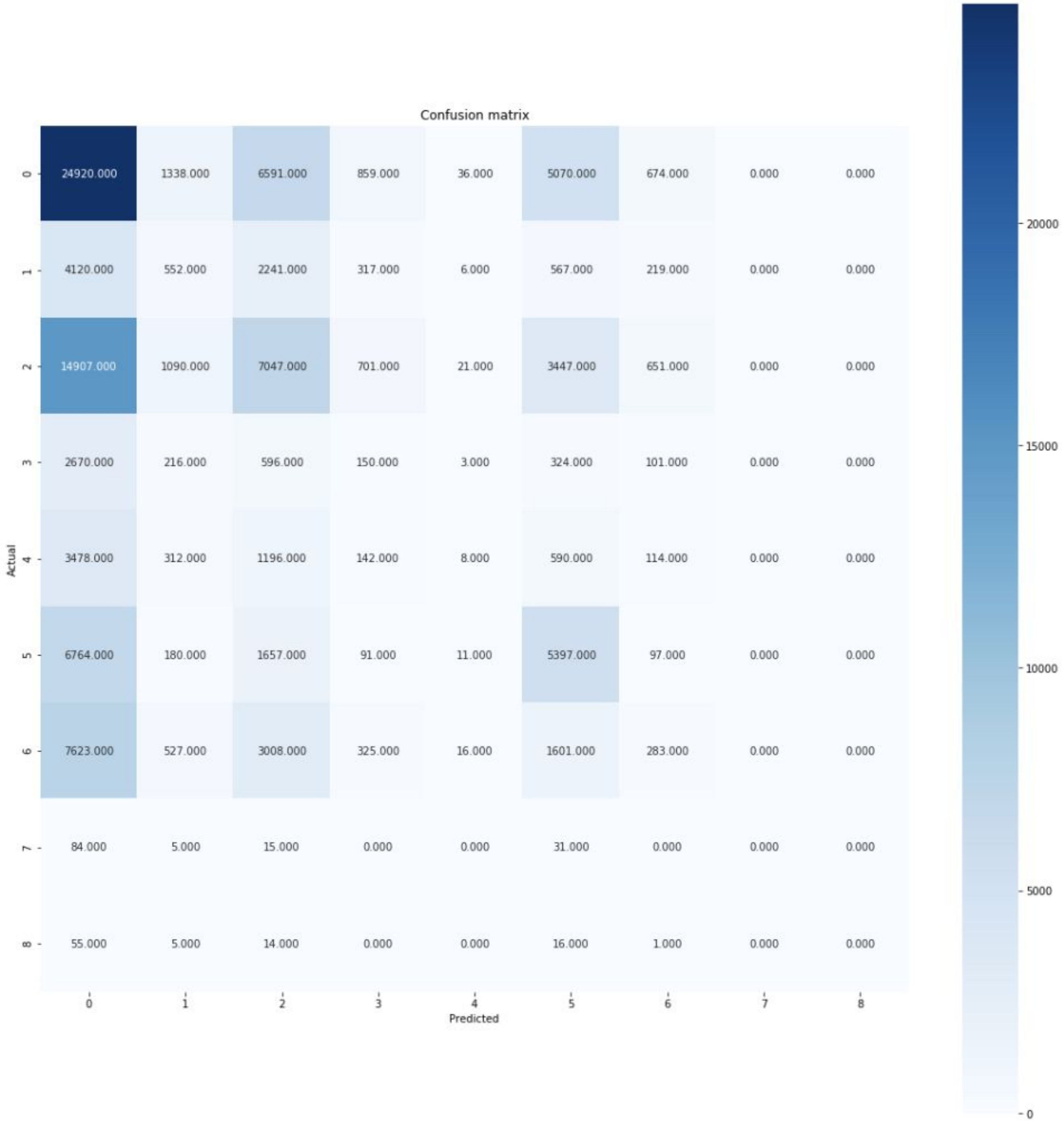
c. Lastly, we explored the different crimes committed in the 77 community areas and created a heat map for when those crimes that also included an arrest. This is just one map based on the number of Assault that follows with an arrest. (There are eight more graphs as so in our code).



5.

Classification:

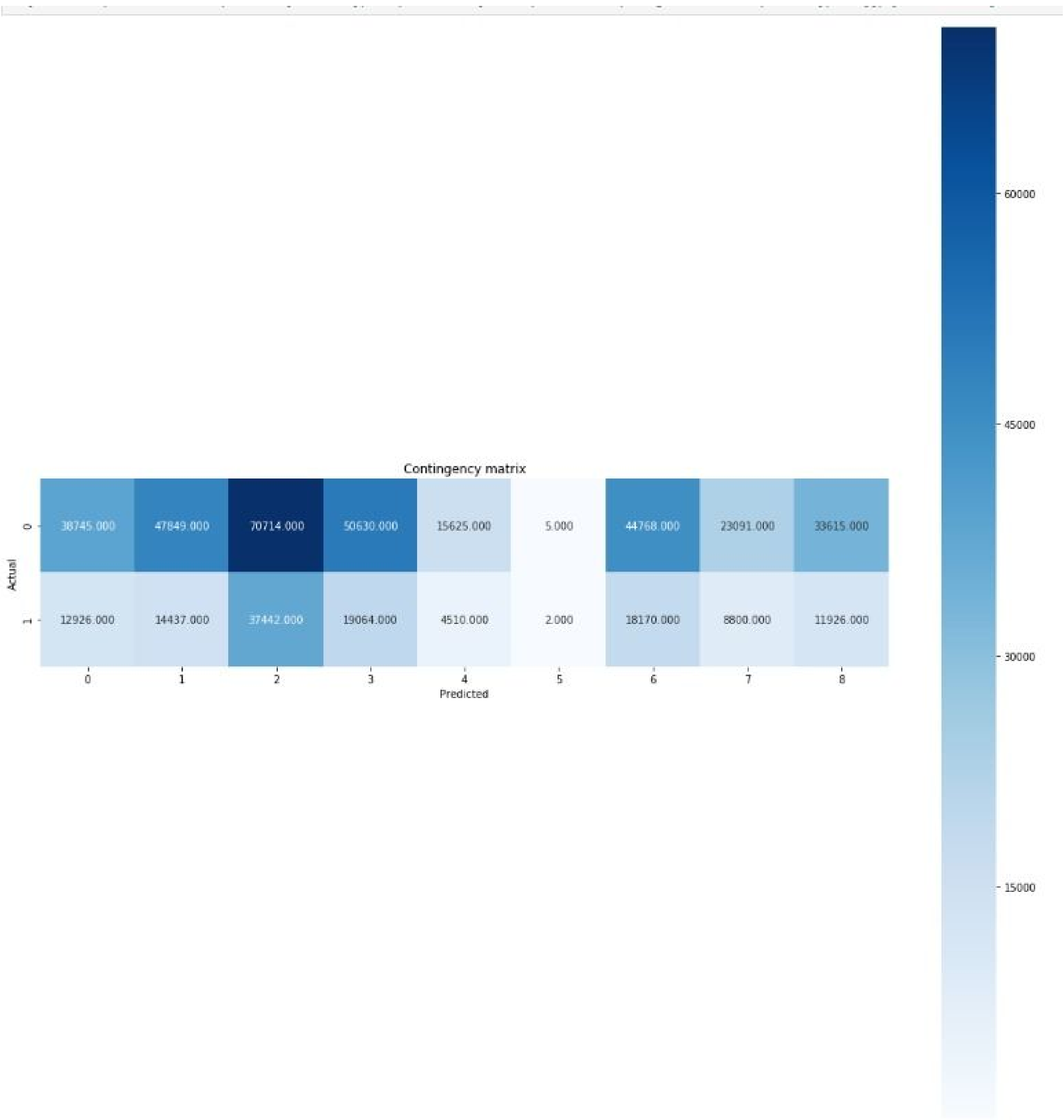
We used the Hold Out Method to split the data set. This leads to our test_size =.25 and Train_size = .75. We used the K-Neighbors classifier to classify Primary Crime Type. The classification results came out to be an F1 score of .0792 and an Accuracy score of 0.339. Below is the confusion matrix which describes the performance of the classification model.



Actual \ Predicted	0	1	2	3	4	5	6	7	8
0	24920.000	1338.000	6591.000	859.000	36.000	5070.000	674.000	0.000	0.000
1	4120.000	552.000	2241.000	317.000	6.000	567.000	219.000	0.000	0.000
2	14907.000	1090.000	7047.000	701.000	21.000	3447.000	651.000	0.000	0.000
3	2670.000	216.000	596.000	150.000	3.000	324.000	101.000	0.000	0.000
4	3478.000	312.000	1196.000	142.000	8.000	590.000	114.000	0.000	0.000
5	6764.000	180.000	1657.000	91.000	11.000	5397.000	97.000	0.000	0.000
6	7623.000	527.000	3008.000	325.000	16.000	1601.000	283.000	0.000	0.000
7	84.000	5.000	15.000	0.000	0.000	31.000	0.000	0.000	0.000
8	55.000	5.000	14.000	0.000	0.000	16.000	1.000	0.000	0.000

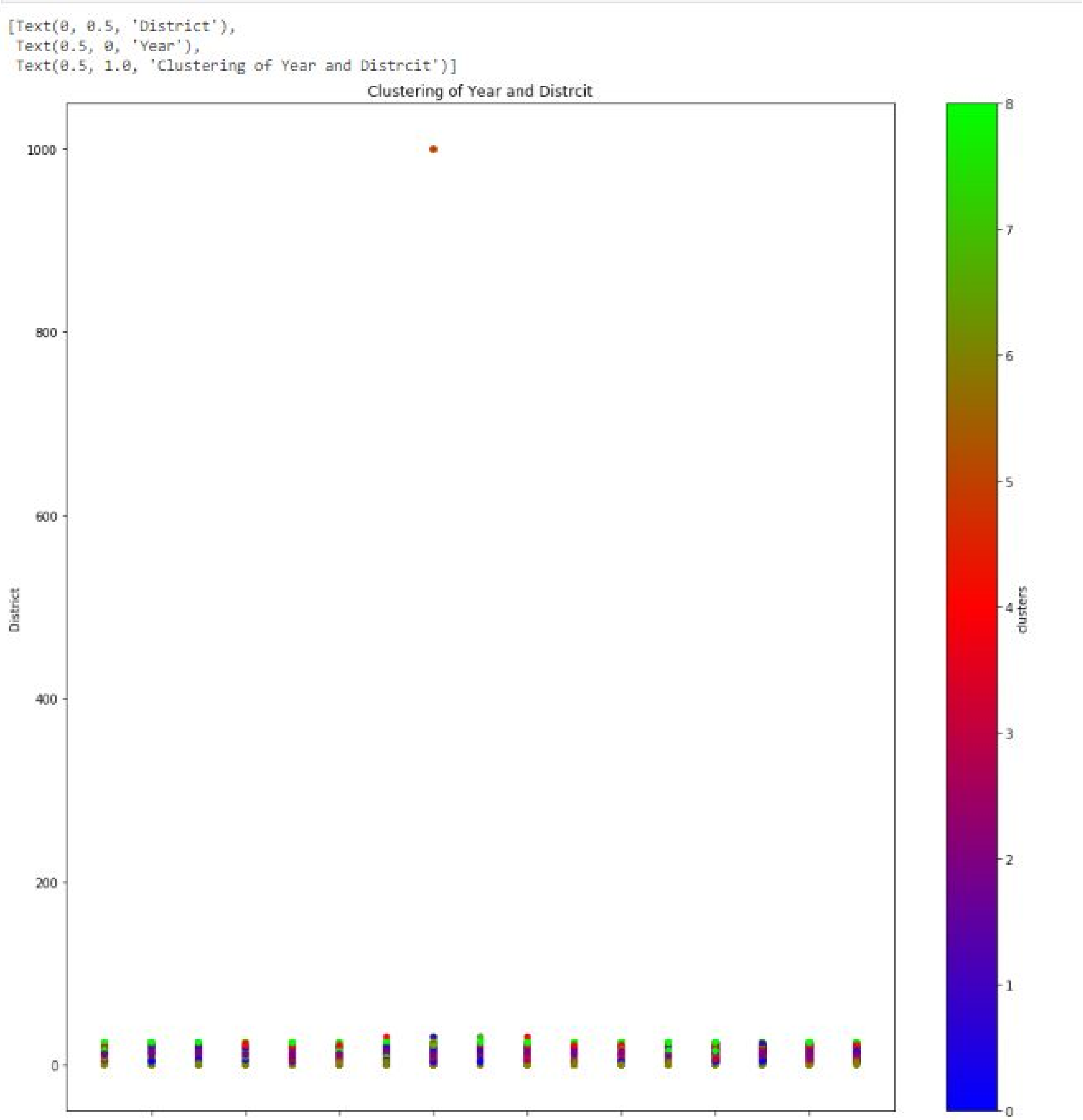
Clustering:

We clustered the data by having the centroid equal to Arrest and clustered the specific columns: Year, TimePeriod N, Arrest_N, Primary Crime Type N, Community Area, District, Major Section N, and Loc Type N. This gave us the output of the Contingency matrix below. This provides a basic picture of the interrelation between two variables and can help find interactions between them. We also received a Silhouette Coefficient of .38 which is not close to 1. This tells us that our clusters are not the most accurate. Below is also a Kmeans clustering.



Contingency matrix

	0	1	2	3	4	5	6	7	8
0	38745.000	47849.000	70714.000	50630.000	15625.000	5.000	44768.000	23091.000	33615.000
1	12926.000	14437.000	37442.000	19064.000	4510.000	2.000	18170.000	8800.000	11926.000



```
[Text(0, 0.5, 'District'),
 Text(0.5, 0, 'Year'),
 Text(0.5, 1.0, 'Clustering of Year and Distrcit')]
```