INTRODUCTION

The grand and lustrous city of Chicago is one that wears many masks. To some it is an illustrious metropolis with an inner jungle of steel and glass illuminating its surroundings. However, to others it is a food desert plagued by segregation and violence which precipitated from the roots.

This report looks to analyze crime statistics taken from the City of Chicago Data Portal as well as Socioeconomic data from the 2010 census. The goal of the report is to analyze crime trends in the city of Chicago in comparison to community areas; as well as to compare crime with socioeconomic data. A correlation was ran between socioeconomic indicators and crime. This allowed for us to run a k-Nearest Neighbor (KNN) algorithm in order to pinpoint similar neighborhoods and their trends. The main objectives to be accomplished through this analysis is to have a standardize method supported by machine learning to measure socioeconomic improvements in neighborhoods throughout the city.

The first analysis is on crimes from 2001 to present. The second analysis is a combination of 2010 census data for the city of Chicago combined with average crimes by Community Area for 2010 so we can run a correlation model between socioeconomic factors and crime. The analysis looks primarily into the data of the city as a whole, as well as the neighborhoods of Pilsen, Back of the Yards (BOTY), and Little Village. The data was modeled in Python in order to facilitate the replication of the analysis once 2020 census data is published.

To begin, our Chicago crime dataset contains 7,079,493 rows and 29 columns of which, we will be using the following:
Primary Type: Type of crime committed (Homicide, Narcotics, Assault, Burglary, ect.)

- Location Description: Where the crime was committed (Street, Home, Restaurant, ect.)
- Arrest: Whether or not an arrest was done
- Domestic: Whether or not it was a domestic crime
- Community Area: In which one of the 77 Chicago community areas did the crime occur
- Year: The year in which the crime occurred
- Month: The month in which the crime occurred
- Day: The day in which the crime occurred
- Crime_hour: the hour in which the crime occurred

Our second dataset is a combination of census data and crime data. It contains 77 rows (one for each Community Area) and 15 columns. The following fields in it were the focus:

- Community Area: One of the 77 Chicago Community Areas
- Num: One of the 77 Chicago Community Areas in numerical form
- Percent of Housing Crowded: Percent of households in the community area in which houses are considered crowded per Census guidelines
- Percent of Households Below Poverty: Percent of households in Community Area living below the poverty line

- Percent aged 16+ Unemployed: Percent of persons over the age of 16 that are unemployed per Community Area
- Percent Aged 25+ Without a High School Diploma: Percent of persons over the age of 25 without a high school education per Community Area
- Percent Aged under 18 or over 64: Percent of people per Community Area under the age of 18 or over the age of 64
- Crime Rate: Crime rate by Community Area

MODELS

As stated above two statistical models were applied on the data sets. The first, a correlation model to asses which fields seemed to have some sort of statistical interaction. Secondly a KNN model was ran in order cluster similar neighborhoods based on their socioeconomic relation.

Correlation: The Correlation model looks to measure if two variables have some sort of statistical relationship. Scores range between -1 and 1. Generally speaking a score about .4 indicates a strong correlation. A score between .2 - .4 can be considered a moderate relationship. Any score below a .2 can be considered a weak correlation.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

k-Nearest neighbor: The KNN model is a supervised learning algorithm. It is a form of clustering which, groups together the most similar points in a dataset. Each cluster is given its own centroid (mean).

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Euclidean distance was then used to measure the difference between clusters. This is simply the distance from point A to point B. It is better known as the Pythagorean metric.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

RESULTS

This section will be broken up into sections. Section 1 is a simple analysis/exploration of the two datasets. This will help us see interesting trends and areas of interest in the city as a whole and in specific neighborhood. Section 2 will show the results of our models.

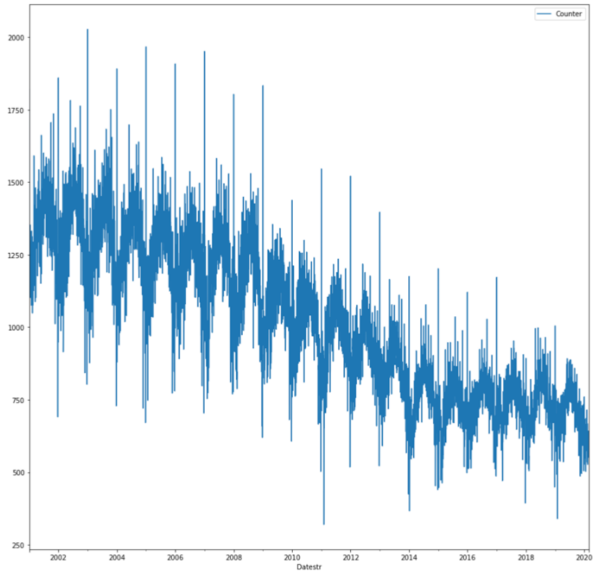
1.0

Crime was analyzed in the city of Chicago as a whole as well as in individual Community Areas. Crime was broken down by time and crime type. Right away we can see that crime has reduced over the past 20 years.

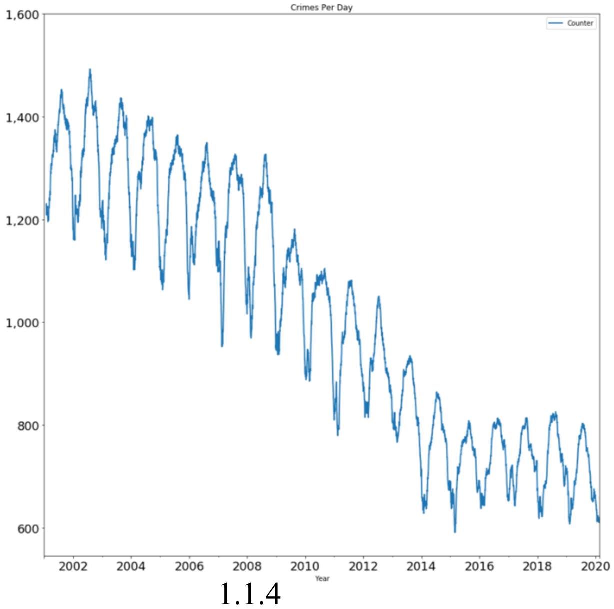
1.1 Time

Over the past 20 years, crime has dropped rather significantly. Figure 1.1.1 below counts the amount of crimes that occurred each day for the last 20 years. While this is a very noisy graph, it shows the decrease in the count of crimes. For more clarity figure 1.1.2 takes the moving average by month. As we can see, Chicago goes from 1200 - 1400 crimes a day in 2001 to 600 – 800 crimes a day. Figure 1.1.3 zooms in and looks solely at 2019. The seasonality of crime is easily distinguished in the patterns each figure shows with less crime in the winter and more crime in the summer. If we zoom out and look at crime trends per month in the last 20 years, the same trend persists. This is shown in Figure 1.1.4.

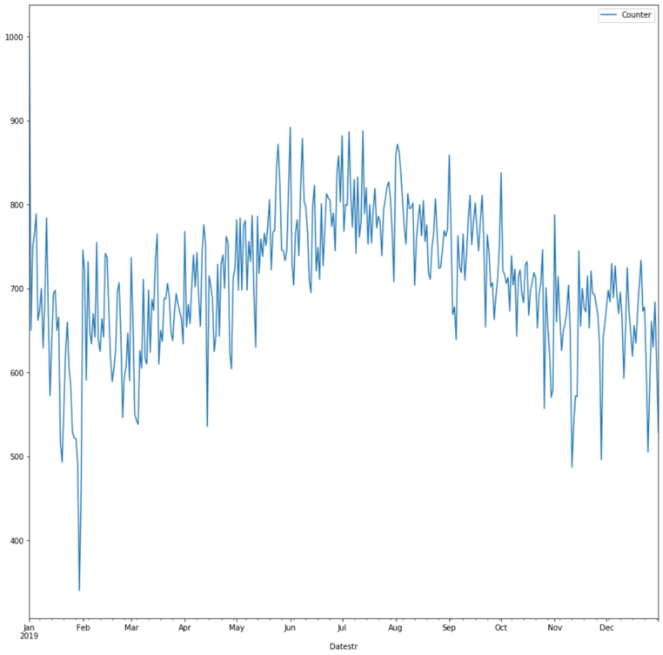
1.1.1



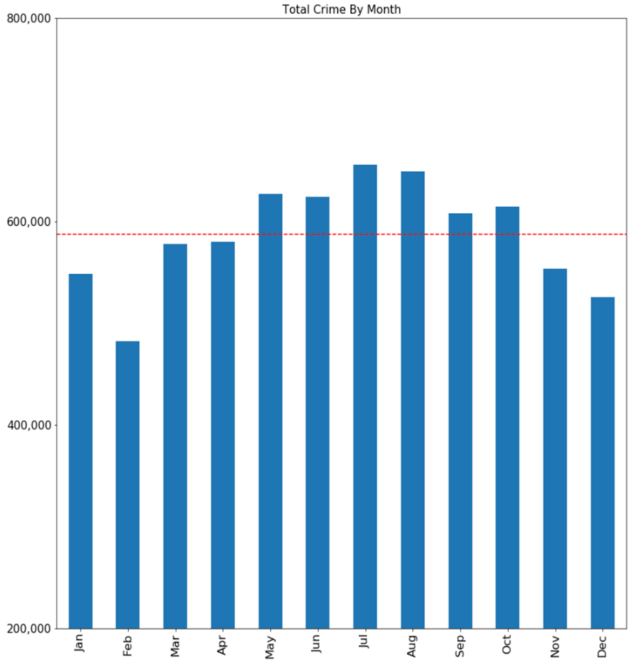
1.1.2



1.1.3

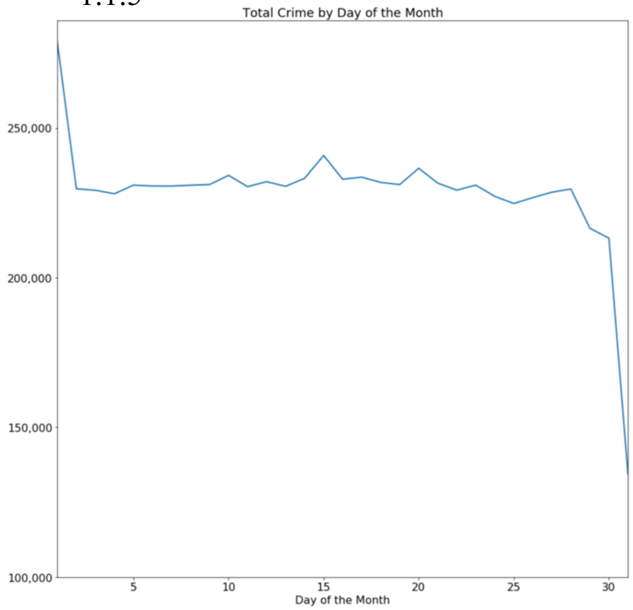


1.1.4

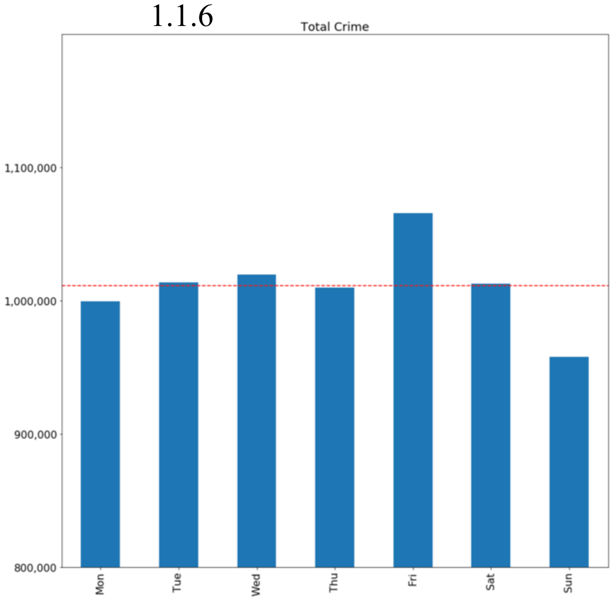


Interestingly as we can see in figure 1.1.5, crime peaks at the beginning of the month. Other than that, when crime is broken down by day of the month it seems to stay consistent. The drop at the end of the month can be attributed to not all months having 30 or 31 days. Lastly, crime was looked at by day of the week in figure 1.1.6. In the past 20 years Friday is easily the day where most crimes are committed while, Sunday is the day where the least amount of crimes occur. All other days deviate very closely to the average.

1.1.5



1.1.6



We then looked into what time crime is being committed at. Sub data frames were created in order to perform the same analysis on community areas of interest for the organization (Pilsen, Little Village, and BOTY). Interestingly for the city of Chicago crime peaks at noon and is followed closely with large amounts of crime occurring between 7pm – 9pm. Figure 1.1.7 illustrates this while Table 2 shows the aggregate amount of crimes per hour over the last 20 years. This of course sparked the interest of what the top crimes were during the peak hour. Figure 3 pulls the 5 crimes at the peak hour. In summary the most likely crime to occur in the city of Chicago is a theft at noon.

Figure 1.1.7

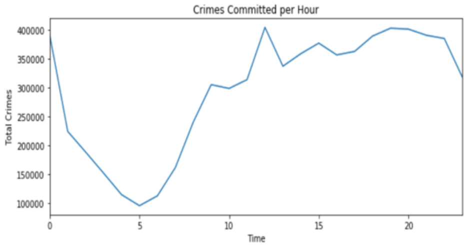
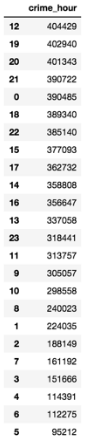
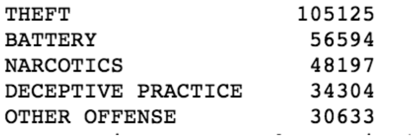


Table 2



	crime_hour
12	404429
19	402940
20	401343
21	390722
0	390485
18	389340
22	385140
15	377093
17	362732
14	358808
16	356647
13	337058
23	318441
11	313757
9	305057
10	298558
8	240023
1	224035
2	188149
7	161192
3	151666
4	114391
6	112275
5	95212

Table 3



```
THEFT                     105125
BATTERY                    56594
NARCOTICS                  48197
DECEPTIVE PRACTICE         34304
OTHER OFFENSE              30633
```

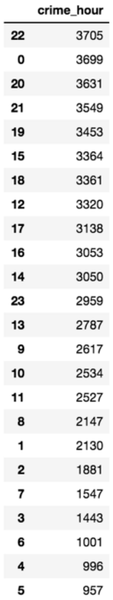
Our process was then replicated on to the neighborhoods of interest. Results were as follows:

Pilsen

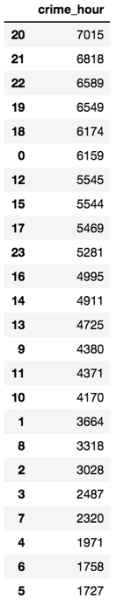
Little Village

Back of the Yards

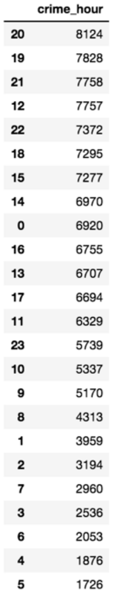
Crime by Hour Table



	crime_hour
22	3705
0	3699
20	3631
21	3549
19	3453
15	3364
18	3361
12	3320
17	3138
16	3053
14	3050
23	2959
13	2787
9	2617
10	2534
11	2527
8	2147
1	2130
2	1881
7	1547
3	1443
6	1001
4	996
5	957

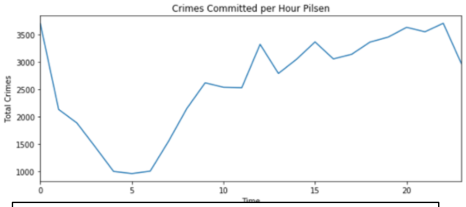


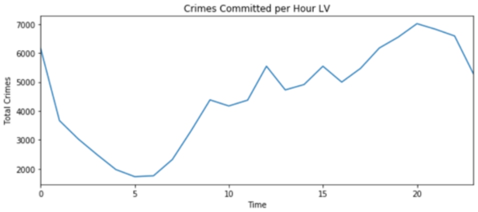
	crime_hour
20	7015
21	6818
22	6589
19	6549
18	6174
0	6159
12	5545
15	5544
17	5469
23	5281
16	4995
14	4911
13	4725
9	4380
11	4371
10	4170
1	3664
8	3318
2	3028
3	2487
7	2320
4	1971
6	1758
5	1727

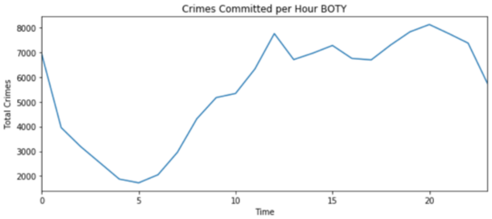


	crime_hour
20	8124
19	7828
21	7758
12	7757
22	7372
18	7295
15	7277
14	6970
0	6920
16	6755
13	6707
17	6694
11	6329
23	5739
10	5337
9	5170
8	4313
1	3959
2	3194
7	2960
3	2536
6	2053
4	1876
5	1726

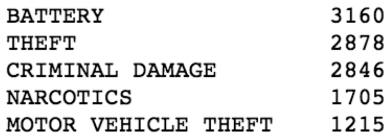
Crime by Hour



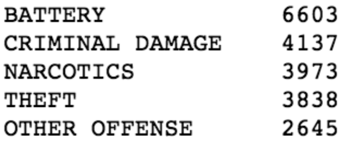




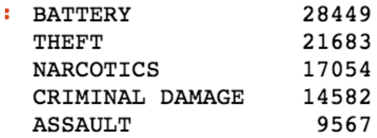
Crime Committed at Peak Hours Table



```
BATTERY              3160
THEFT                2878
CRIMINAL DAMAGE      2846
NARCOTICS            1705
MOTOR VEHICLE THEFT  1215
```



```
BATTERY           6603
CRIMINAL DAMAGE   4137
NARCOTICS         3973
THEFT             3838
OTHER OFFENSE     2645
```



```
: BATTERY            28449
  THEFT              21683
  NARCOTICS          17054
  CRIMINAL DAMAGE    14582
  ASSAULT             9567
```

For all 3 neighborhoods crime peaks into the late evening. The “Crime Committed at Peak Hours Table” pulls the top 5 hours in the evening when the most crimes are committed in each neighborhood. As we can see, they all share very similar crimes however, BOTY by far has the most crime being committed. They exceed the amount of battery’s in Little Village by 430%.

1.2 Crime Type
We next looked into the types of crimes being committed the most in the city of Chicago. The dataset records 35 different types of crimes. The frequency at which each occurred is recorded in figure 1.2.1. Table 4 besides it gives the breakdown of each crimes occurrence in figure 1.2.1

Theft is the type of crime that occurs the most with almost 1,500,000 occurrences. Interestingly while homicide is the crime heard about the most on the news, it falls into 23rd place with a still staggering 10,113 occurrences since 2001.

Figure 1.2.1

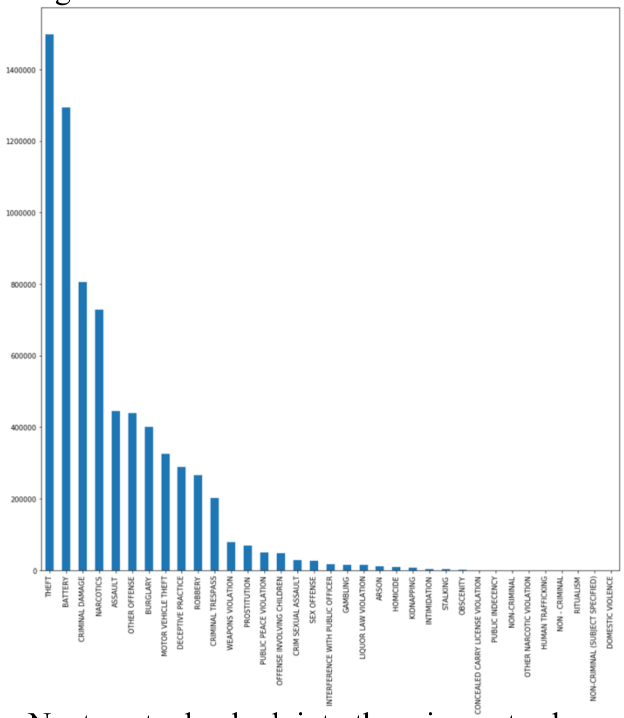
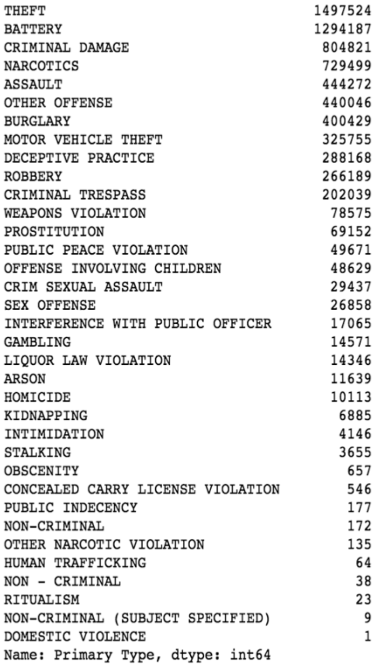


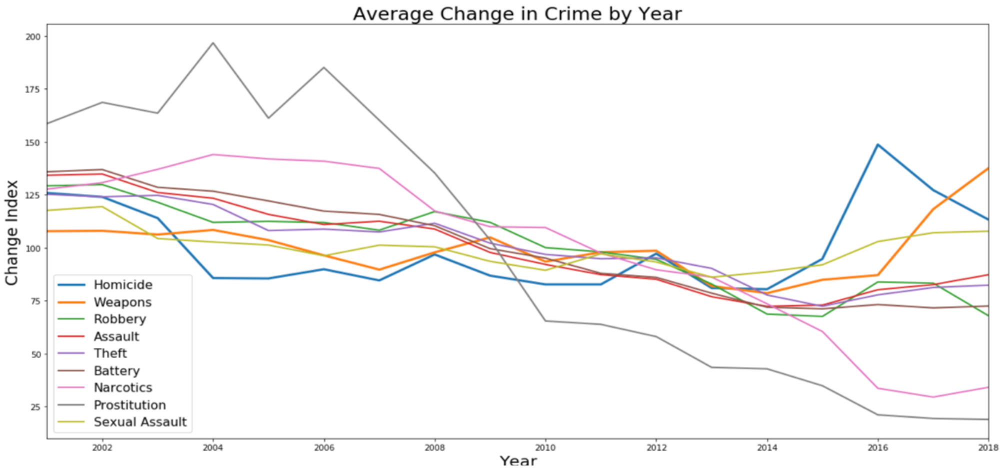
Table 4



```
THEFT                                  1497524
BATTERY                                1294187
CRIMINAL DAMAGE                         804821
NARCOTICS                               729499
ASSAULT                                 444272
OTHER OFFENSE                           440046
BURGLARY                                400429
MOTOR VEHICLE THEFT                     325755
DECEPTIVE PRACTICE                      288168
ROBBERY                                 266189
CRIMINAL TRESPASS                       202039
WEAPONS VIOLATION                        78575
PROSTITUTION                             69152
PUBLIC PEACE VIOLATION                   49671
OFFENSE INVOLVING CHILDREN               48629
CRIM SEXUAL ASSAULT                      29437
SEX OFFENSE                              26858
INTERFERENCE WITH PUBLIC OFFICER         17065
GAMBLING                                 14571
LIQUOR LAW VIOLATION                     14346
ARSON                                    11639
HOMICIDE                                 10113
KIDNAPPING                                6885
INTIMIDATION                              4146
STALKING                                  3655
OBSCENITY                                  657
CONCEALED CARRY LICENSE VIOLATION          546
PUBLIC INDECENCY                           177
NON-CRIMINAL                               172
OTHER NARCOTIC VIOLATION                   135
HUMAN TRAFFICKING                           64
NON - CRIMINAL                              38
RITUALISM                                   23
NON-CRIMINAL (SUBJECT SPECIFIED)             9
DOMESTIC VIOLENCE                            1
Name: Primary Type, dtype: int64
```

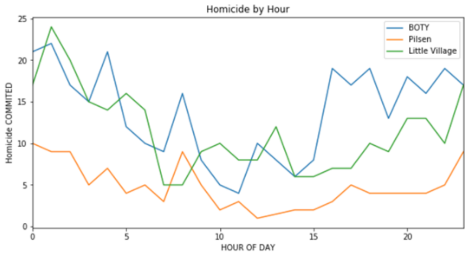
Next we took a look into the crime rate changes over the years in figure 1.2.2. For these the top 9 crimes of interest were looked into (Homicide, Weapons, Robbery, Assault, Theft, Battery, Narcotics, Prostitution, and Sexual Assault). As we can see, there is a downward trend in crime between 2001 – 2019 most notably in prostitution and Narcotics. However, after the year 2014 there are major jumps in homicides, weapon crimes, and sexual assault. Furthermore, assaults, theft, and robberies also increase.

Figure 1.2.2

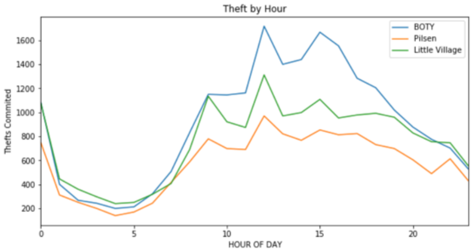


We can look at this in detail in our community areas of interest in order to see how crime dynamics have changed there. This will allow for us to compare different community areas with each other as well as with the city. For figures 1.2.3 – 1.2.6 the occurrence of specific crimes was compared through the time of day in the BOTY, Pilsen, and Little Village areas. The common trend we see here is that in all 4 figures BOTY has the most crime, followed by Little Village. Furthermore, there are very similar trends and peaks in the neighborhoods for Narcotic, Theft, and Battery crimes. For these 3 crimes it is clear that intervention during the daytime would be effective seeing as this is when the upward trends begin.

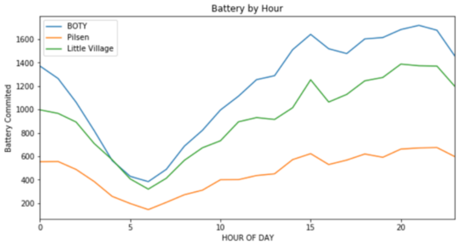
1.2.3



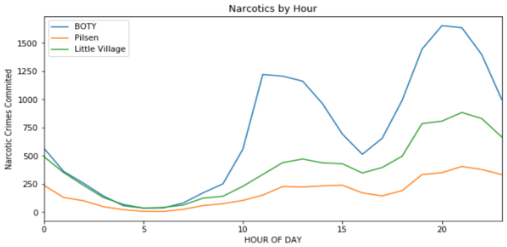
1.2.4



1.2.5



1.2.6

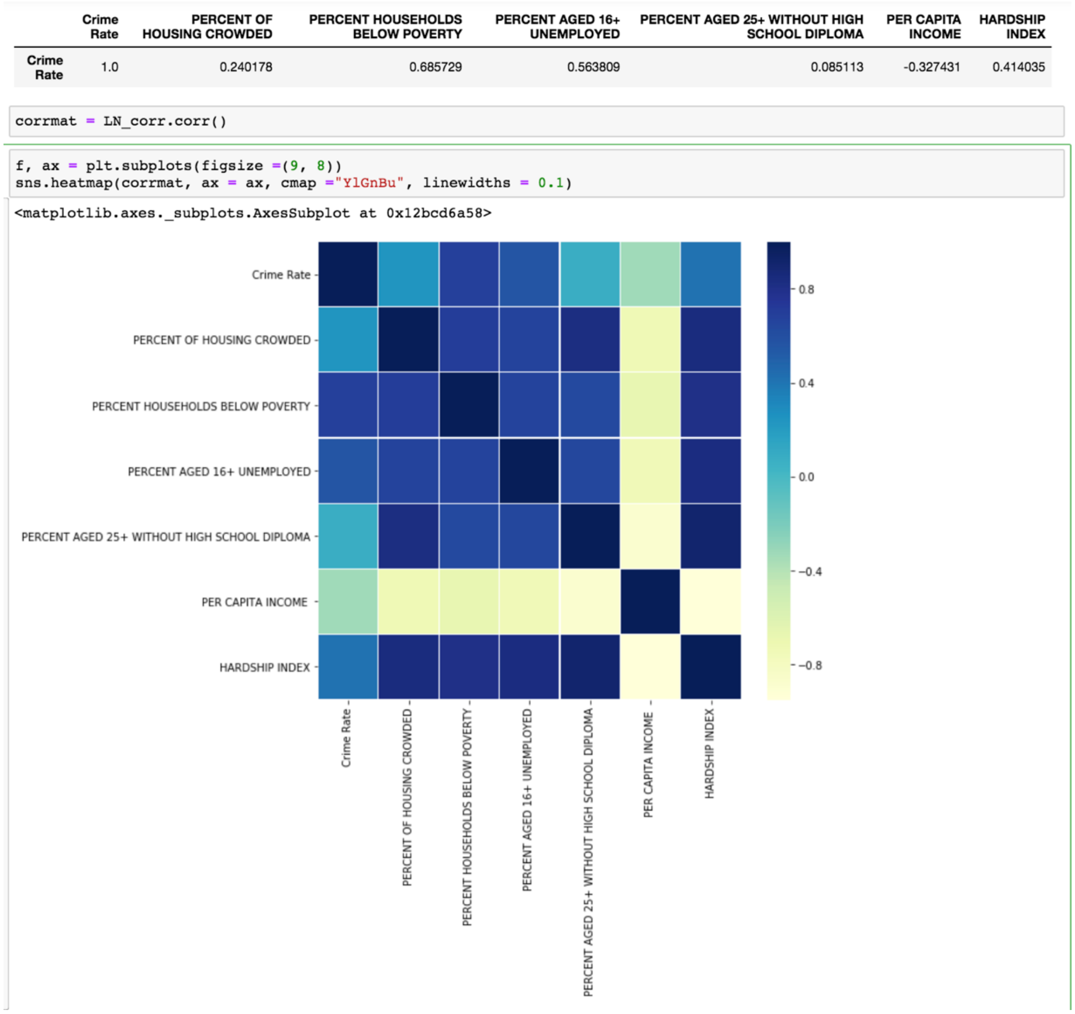


2.1 Correlation

The final piece of this analysis is to integrate socioeconomic data in order to view its correlations with crime. For this, data was pulled from the 2010 census. The census data was then combined with the Chicago crime data by taking the average crime rate in each community area and binding the datasets through the community areas. This new dataset was ran through a correlation matrix. The regressor variable chosen was the Crime Rate. This variable was ran against Percent House Hold Crowded, Percent Households Below Poverty, Percent Aged 16+ Unemployed, Percent Aged 25+ without High School Diploma, Per Capita Income, and Hardship Index. Hardship index is a combination of the above-mentioned metrics. Therefore, while included in the correlation for a cumulative analysis purpose, it will not be further analyzed.

Results obtained were a pretty strong correlation between Crime Rate and Households Below Poverty with a correlation coefficient of .685. This is followed by a Percent Aged 16+ Unemployed which got a correlation coefficient of .56. Percent Capita Income score the lowest with a correlation coefficient of -.32. This could be due to it being the only metric that was not broken down into percentages. For future analysis this metric should be converted into a percentage. A good alternative to this would be to replace it with the AMI. Table 5 shows the full results, followed by a heat map indicating the strength of the correlations through color.

Table 5



	Crime Rate	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PER CAPITA INCOME	HARDSHIP INDEX
Crime Rate	1.0	0.240178	0.685729	0.563809	0.085113	-0.327431	0.414035

```
corrmat = LN_corr.corr()
```

```
f, ax = plt.subplots(figsize =(9, 8))
sns.heatmap(corrmat, ax = ax, cmap ="YlGnBu", linewidths = 0.1)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x12bcd6a58>
```

Crime Rate
PERCENT OF HOUSING CROWDED
PERCENT HOUSEHOLDS BELOW POVERTY
PERCENT AGED 16+ UNEMPLOYED
PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA
PER CAPITA INCOME
HARDSHIP INDEX
0.8
0.4
0.0
−0.4
−0.8

2.2 k-Nearest Neighbor

The final model we ran our data through was a KNN. Once again, a KNN model looks to find similarity between data points and calculate the means based off the clusters and not the entire dataset. For this particular analysis, this will allow for us to understand which neighborhoods are more similar based off of the Crime Rate and Percent Households Below Poverty.

In order to determine the optimal number of clusters, both the distortion and inertia elbow methods were used. The distortion method iterates through different numbers of clusters and uses the average of the squared distance from the points to the cluster center of the respective clusters. On the other hand, the inertia method uses the sum of squares to calculate the distance from each iterated cluster center to the data points pertaining them. This method is called the Elbow Method because the optimal number of clusters is determined by where your graph turns and makes an elbow shape. Through this method it was determined that the optimal number of clusters would be 4.

Figure 2.2.1

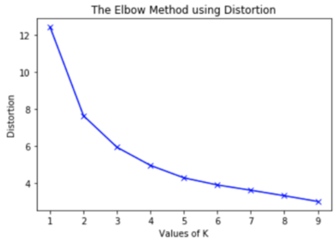
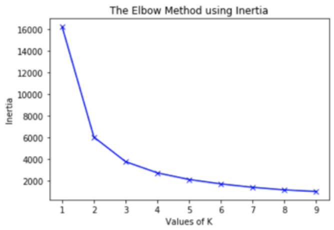


Figure 2.2.2

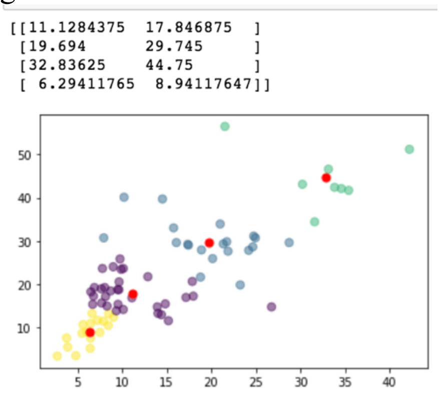


Finally, we were able to run our KNN algorithm. Each cluster has a centroid represented by a red dot. The centroid is simply the mean of each cluster. Figure 2.2.3 below contains a table that shows the coordinates of each centroid. The X axis represents the Crime Rate as a percentage while the Y axis represents a Percent Households Below Poverty. Cluster 1 in yellow has an average Crime Rate of 6.29% and an average Household Below Poverty Rate of 8.94%. Cluster 2 in purple has an average Crime Rate of 11.12% and a Household Below Poverty Rate of 17.84%. Cluster 3 in the blue has an Average Crime Rate of 19.69% and a Household Below Poverty Rate of 29.74%. Cluster 4 in the green has an Average Crime Rate of 32.83% and a Household Below Poverty Rate of 44.75%. Please look at table 6 for reference. The KNN Cluster output was replicated in Tableau so we can more easily view which neighborhoods were clustered together. This can be seen in Figure2.2.4.

Table 6

Cluster Number	Color	Crime Rate	Household Below Poverty Rate
1	Yellow	6.29%	8.94%
2	Purple	11.12%	17.84%
3	Blue	19.69%	29.74%
4	Green	32.83%	44.75%

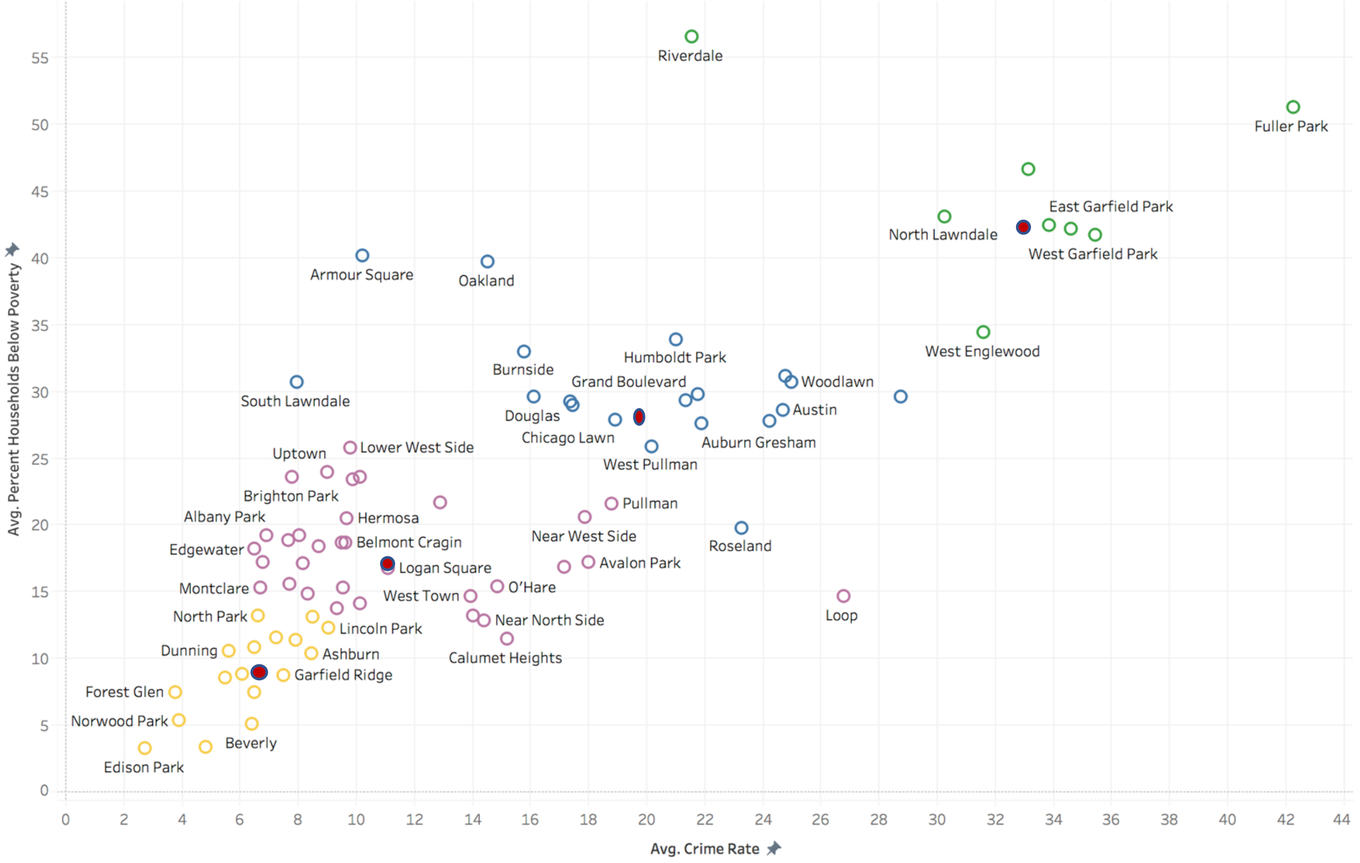
Figure 2.2.3



```
[[11.1284375   17.846875   ]
 [19.694       29.745      ]
 [32.83625     44.75      ]
 [ 6.29411765  8.94117647]]
```

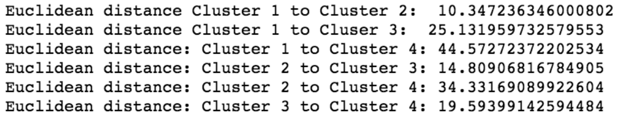
Figure 2.2.4

KNN Cluster



The Euclidean distance was then taken from each cluster. Euclidean distance in this case tells us on a 2D plain the distance from one cluster to another. This metric will allow us to measure how similar or different clusters are. Table 7 below calculated the Euclidean Distance between all clusters. The Euclidean distance from Cluster 1 to Cluster 2 is 10.34. This means that on average, for a neighborhood in Cluster 2 to move up to cluster 1 a combination of the Crime Rate and Households Below Poverty would have to improve by a combined total of 10 points.

Table 7



```
Euclidean distance Cluster 1 to Cluster 2:  10.347236346000802
Euclidean distance Cluster 1 to Cluser 3:  25.131959732579553
Euclidean distance: Cluster 1 to Cluster 4: 44.57272372202534
Euclidean distance: Cluster 2 to Cluster 3: 14.80906816784905
Euclidean distance: Cluster 2 to Cluster 4: 34.33169089922604
Euclidean distance: Cluster 3 to Cluster 4: 19.59399142594484
```

CONCLUSION

Various different components go into the socioeconomic development of a community. Currently, due to the core of this report depending on census data there is a huge limitation of only being able to run it once per decade. Luckily, we are on a census year so we will very soon be able to view the changes that have occurred over the past 10 years. Furthermore, as the world becomes more data driven hope for augmentation that datasets are high. This will allow for deeper looks and more precise results.

We have learned that crime has reduced significantly since the early 2000s. There is also a seasonality in crime with it peaking in the summer months and declining over with winter months. The beginning of the months and 15th of the month seem to have the most crimes and Fridays are the peak for the week. If we look at it by hour noon interestingly is when a crime is most likely to happen, followed by the hours of the evening. Overall theft and battery are the most common crimes for the City.

When looking into Pilsen, Little Village, and BOTY we learn very quickly that BOTY has the most crimes going on. Interestingly the trends by hour are very similar as are the types of crimes committed. These can be used to narrow and target for example when Community Ownership leaders should hold interventions or when Financial Wellness counselors should hold classes; with the hopes of positively impacting the crime during the peak moments. Furthermore, new neighborhoods of interest can be pinpointed for expansion and dissected in order to impact the communities as quickly and efficiently as possible.

In terms of the KNN results, graduating clusters into Cluster 1 should be the objective since, it has the lowest crime rate and households below poverty. Of course, taking a neighborhood from Cluster 4 and aiding it into Cluster 1 will be very difficult. However, through this metric we will be able to monitor the improvement over time. Furthermore, this analysis will be easily replicated once 2020 census data is released to the public. We will now have a means by which to compare changes from 2010 to 2020. In relation to crime data only, we will have the ability to monitor yearly changes in the neighborhoods based off of the city's public crime dataset.

Reference:

Acosta-Cordova, Jose M. *The Latino Neighborhood Report: Issues and Prospects for Chicago*. October, 2017.

Hunt, Bradford. “Redlining.” *Encyclopedia of Chicago*, 2004, www.encyclopedia.chicagohistory.org/pages/1050.html.

Wilson, Reid. “FBI: Chicago Passes New York as Murder Capital of U.S.” *The Washington Post*, WP Company, 18 Sept. 2013, www.washingtonpost.com/blogs/govbeat/wp/2013/09/18/fbi-chicago-passes-new-york-as-murder-capital-of-u-s/.