Fight Crimes, Live Safe

A crime analysis report of Chicago by Xiangyu Chen/Springboard Data Scientist Career Track 2017-18

Table of contents:

1. Motivation of this project and how the analysis would be useful to many people.

Crime is one of the major problems people have to deal with living in a society. There is no way to stop all crimes from happening, but there are certainly ways to reduce the rate of crimes and prevent certain types of crimes. Also, what people think about crime rates and what the actual crime rates look like do not always agree (Figure 1).

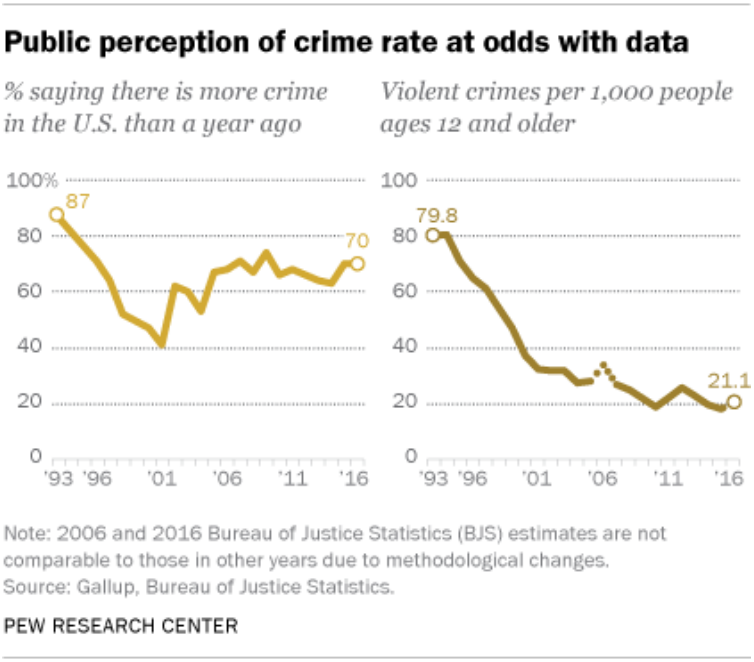

Figure 1: Perception of crime rate and actual crime rates

Source: Pew Research Center by John Gramlich, http://www.pewresearch.org/fact-tank/2018/01/30/5-facts-about-crime-in-the-u-s/ft_18-01-26_crimetrends_perception/

A thorough and in-depth crime analysis is therefore valuable in that it will reveal patterns that could help in the detection and prevention of crimes. Moreover, machine learning techniques could be used to build a model and predict the type of crimes based on location, time of day, neighborhood, and/or even temperature/weather data.

Chicago is a large city with a readily available crime data set from 2001 to 2016. This data set contains detailed description for each reported crime including time, block, community, crime type, etc. The total number of rows is about 6.5 million, so a lot of work could be done to extract useful information. This is also why I chose this data set to do my analysis. When different parts of the city are broken down into neighborhoods/communities (defined by the US Census and these two terms are used interchangeably in this report from now on), crime hotspots are observed (Figure 2: https://commons.wikimedia.org/wiki/File:Chicago_violent_crime_map.svg). It would be useful for law enforcement to know whether there are places where certain types of crimes happen more often than others.

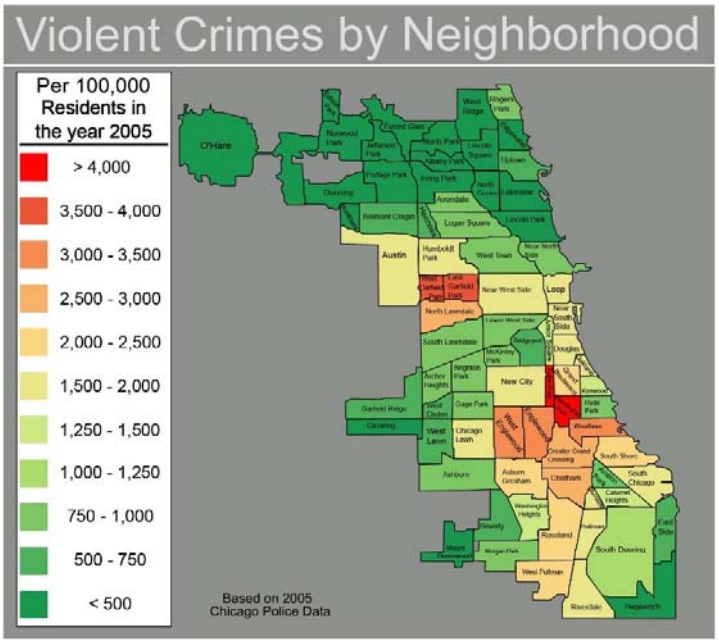

Figure 2: Crime rates in different neighborhood of Chicago in 2005

Source: Wikepedia commons

In a big city like Chicago, it's not possible to cover all the areas with police patrols so sending officers to 'crime hotspots' more frequently would be more efficient than patrolling randomly in the city. For example, neighborhoods which historically have high frequencies of homicides, are areas more policing could reduce the likelihood of homicides. Another factor that might be relative to crime patterns is the time of day. Crimes may more frequently occur at night since it limits visibility, making it more difficult

for crimes to be detected while they are occurring. Do crimes peak in late hours? If so, is this true for all crimes, or just some crimes? The answer lies within the temporal analysis of different type of crimes. To maximize the use of police force, both time and location are critical.

Social economics data will be obtained from the US Census to study the relationship between different statistics (such as income, poverty, and education levels) and crime rates. Do higher poverty levels correlate with higher crime rates? If so, is it true for all communities in Chicago? Communities with high poverty levels but low crime rates or vice versa would be interesting targets for further investigation.

By comparing neighborhood crime rates and poverty levels or education levels, it may be possible to infer the underlying causes for different crimes. Among the above correlations, education plays an important role. For example, do school closings affect crime rates in the vicinity? Is there a relationship between the percent of a neighborhood's population with bachelor's degree and crime rates? These problems are relevant to the local government because they can point out the areas which the officials or legislatures need to improve to reduce crimes in the communities of Chicago.

The last goal of my analysis is use machine learning algorithms to predict crime types. From the EDA we will learn a lot about the patterns of different crimes and the relationships between the patterns and other factors such as location, time of day, community, and geographic coordinates. These factors can be useful in predicting crime types.

And of course, there is always more work to be done to get more insights into the crimes in the city. This report will pave the road for more complete crime analysis and more informative answers to the above questions.

2. **Data acquisition and wrangling**

The main crime data set is available on Chicago Data Portal website as a csv file for download. This data set includes complete data of crimes reported during 2001 and 2017 (note that 2017 data was not complete by the time the analysis was conducted). The link to the data set is below:

Crime Data Set

To get weather data, I used National Oceanic and Atmospheric Administration (NOAA) website and downloaded weather data from 2001 to 2016 for Chicago. Social economics data were downloaded from the Chicago Data Guy. Transportation/transit location file was acquired from Chicago Data Portal. School closing data were obtained from WBEZ news. Colleges and universities coordinates were processed by gathering their addresses from their websites and retrieved using Google Maps API.

Crime data were read in using Pandas package for Python (Pandas website) as a data frame. It contained many missing values in the 'Ward' (legislative district) and 'Community Areas' (Census tracts) columns. This analysis focused on the Neighborhood community level so that Ward column was dropped. I then filled the missing values in the 'Community Areas' using 0 so that we keep most of the crime data. The 'Date' column was parsed using to_datetime method of pandas to datetime object and set as index for time series analysis. The first three rows were displayed as below.

Date	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	District	FBI Code
2001-01-01	3272413	HK299011	2001-01-01	076XX S EMERALD AVE	0842	THEFT	AGG: FINANCIAL ID THEFT	RESIDENCE	False	False	621	6.0	06
2001-01-01	5508664	HN320275	2001-01-01	009XX N HARDING AVE	0266	CRIM SEXUAL ASSAULT	PREDATORY	RESIDENCE	False	True	1112	11.0	02
2001-01-01	2743135	HJ380339	2001-01-01	039XX N MOZART ST	0266	CRIM SEXUAL ASSAULT	PREDATORY	APARTMENT	True	False	1733	17.0	02

Figure 3: Crime data frame representative rows

The weather data were downloaded as two files since the NOAA website has a limit on the size per request. The two csv files were also read in using Pandas and appended together to form one data frame. Since the weather types were labeled by numbers, the corresponding actual weather types were mapped into the new data frame (Figure 4).

	index	LATITUDE	LONGITUDE	ELEVATION	DATE	PRCP	SNOW	SNWD	TAVG	TMAX	TMIN	weather_type	date
0	0	41.995	-87.9336	201.8	2001-01-01	0.0	0.0	17.0	14.5	24.0	5.0	Unknown	2001-01-01
1	1	41.995	-87.9336	201.8	2001-01-02	0.0	0.0	15.0	12.0	19.0	5.0	Unknown	2001-01-02
2	2	41.995	-87.9336	201.8	2001-01-03	0.0	0.0	14.0	17.5	28.0	7.0	[Fog, ice fog, or freezing fog, Mist, Snow, sn...	2001-01-03

Figure 4: Weather data representative rows

I then merged the two data frames as a left join using the merge method in Pandas to add weather information to the crime data using the 'Date' column.

After reading in the education, income, and poverty data, I found they needed quite a bit of wrangling to make them into the right shape for analysis. They contained more information than needed. To extract just the data between 2011 to 2015 (most recent data), irrelevant columns were dropped. A multiindex approach was used before the unstack method in Pandas to align community area numbers to their names and the education, income, and poverty data (Figure 5).

	Community Name	Community Area	(2011-2015, BA or Higher)	(2011-2015, High School Graduate Only)	(2011-2015, Not HS Graduate)	(2011-2015, Percent HS Grad or Higher)	(2011-2015, Percent with a BA or Higher)	(2011-2015, Some College)	(2011-2015, Total)
0	Rogers Park	1	15,502	6,998	5,862	84.0%	42.0%	8,667	37,029
1	West Ridge	2	19,004	10,665	8,318	83.0%	39.0%	10,346	48,333
2	Uptown	3	24,740	6,844	4,784	89.0%	55.0%	8,502	44,870

Figure 5: Representative rows from the education data table after wrangling

The numbers were still in string format and they were converted to integers by first deleting the comma or % sign by using str attribute of the columns followed by astype method. The same approach was used on the income and poverty data sets to convert numbers from strings to integers.

3. **Exploratory Data Analysis (EDA)**

Once the data tables were cleaned and wrangled, the data were analyzed to try to find trends and patterns. Since the time stamp of each crime is available, a time series was performed to discover how crime rates have changed over time. Over the course of 16 years, crime rates decreased in Chicago (Figure 6). (also true at the whole country level according to the Brennan Center for Justice website).

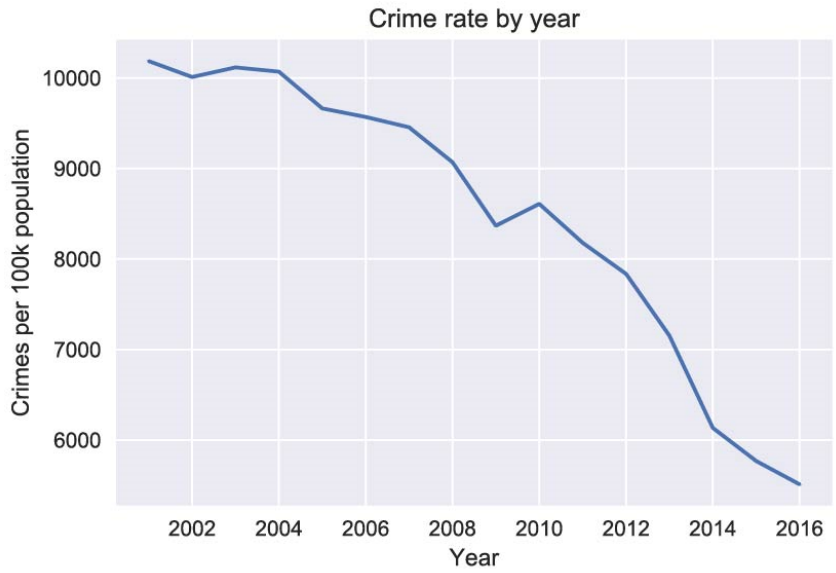

Figure 6: Annual crime rates in Chicago between 2001 and 2016

These data include all different types of crimes and do not indicate what is happening with specific types of crime. Therefore, the trends of eight different felonies: 'Theft', 'Battery', 'Narcotics', 'Burglary', 'Robbery', 'Weapon violation', 'Sex offense', 'Homicide' were examined between 2001 and 2016 (Figure 7).

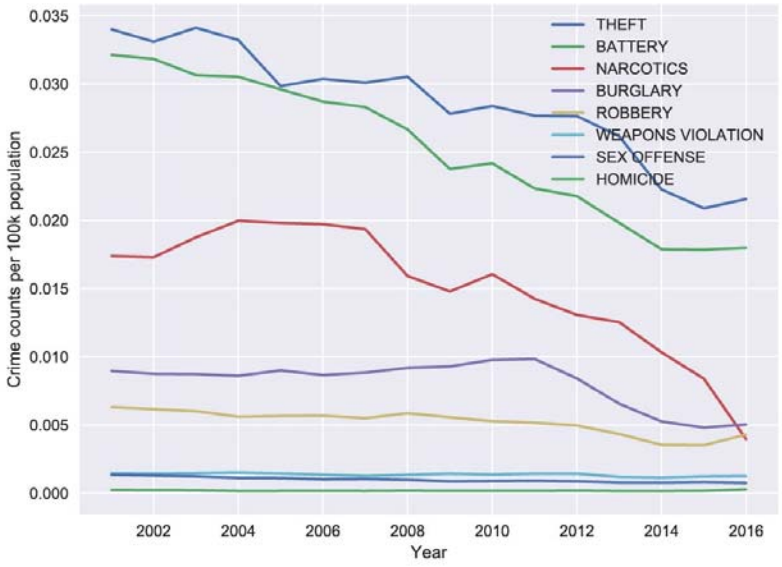

Figure 7: Annual rates of specific crimes in Chicago between 2001 and 2016

This figure shows except for ‘Narcotics’, all other crimes seem to have a little increase after 2014. Due to the large counts of some crimes, ‘Weapons violation’, ‘Sex offense’, and ‘Homicide’ trends are flattened. Another plot with just the above 3 crimes is shown below.

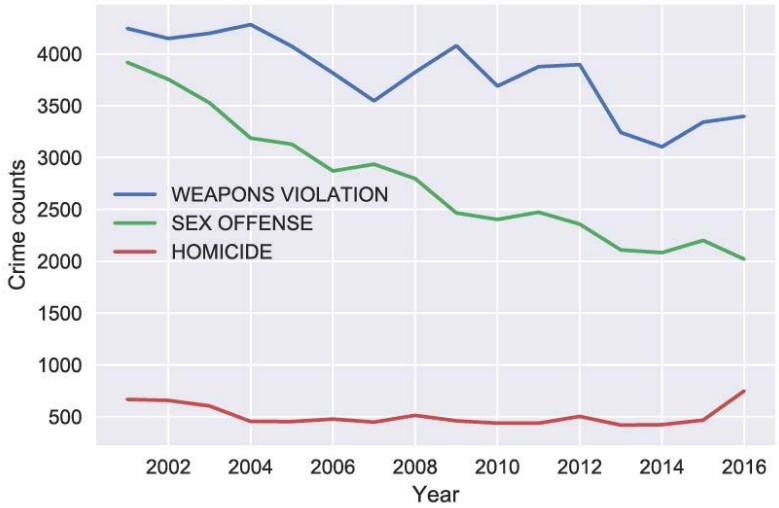

Figure 8: Crime rates of ‘Weapons violation’, ‘Sex offense’, and ‘Homicides’

Both ‘Weapons violations’ and ‘Homicides’ increased after 2014. Are those causatives? Without the details of all the homicides cases, we can’t draw the conclusion yet. The homicide counts are drastically elevated in 2016. What about the rates then? By plotting the homicide rates by year, we can see how bad the situation was.

Figure 9: Annual murder rates between 2001 and 2016

Murder rate skyrocketed after 2015. Chicago made news headlines in early 2017 for the unusually high murder incidents (see this CNN article). Gun problems were the major culprit for these murders (according to this Chicago Tribune article). But where did all these new murders emerge? Did some neighborhoods get worse over the year or were the new cases spread out in different communities? To answer this question, homicides in 2015 and 2016 were broken down by communities (Figure 10). There are 77 communities in Chicago as defined by the US Census.

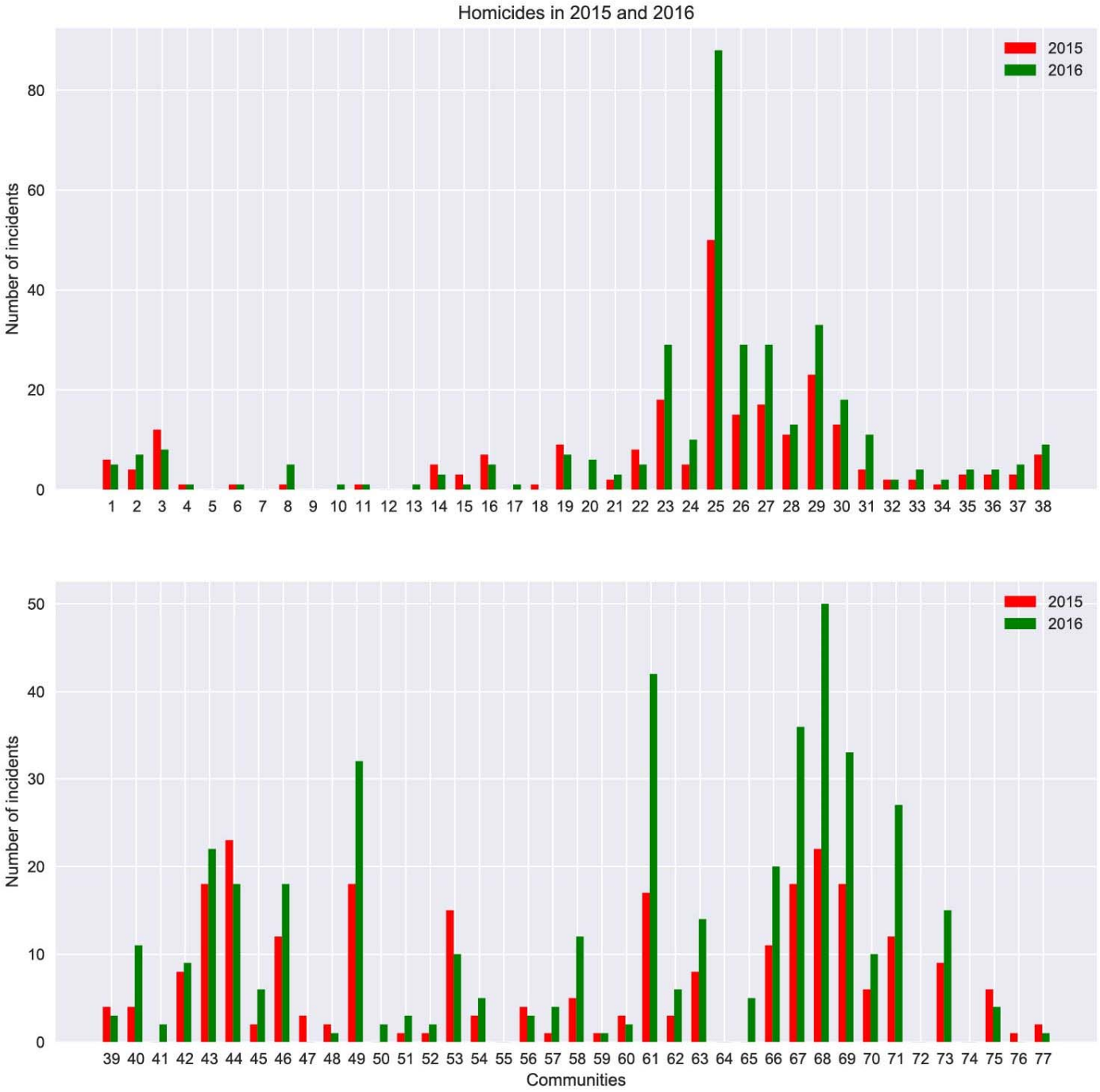

Figure 10: Comparing the number of homicides during 2015 and 2016 in different Chicago communities

Murders increased in these certain communities as two clusters: 23 to 29 (cluster 1) and 66 to 71 (cluster 2). These communities are geographically clustered as well (Figure 11, Social Work, Loyola University, Chicago).

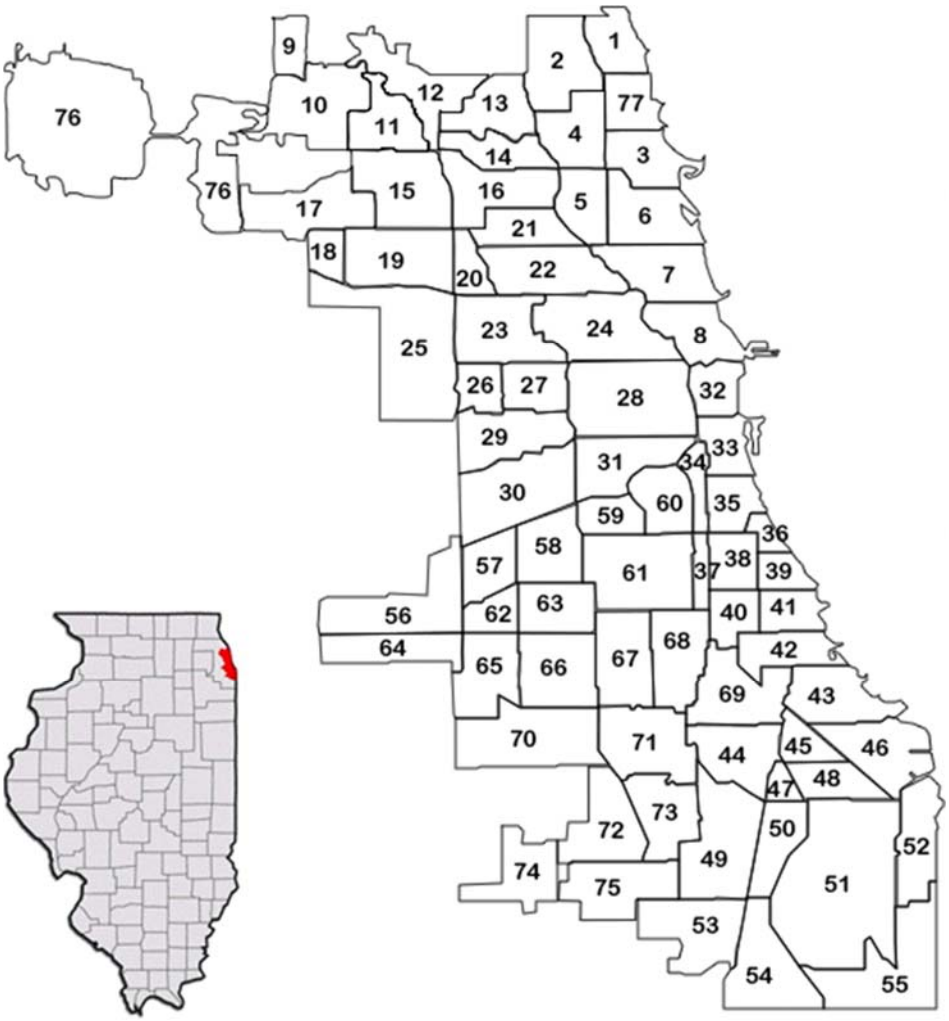

Figure 11: Chicago communities as defined by the US Census

I then visualized murders in different communities as choropleth map. This plot used the gmaps package for Python (gmaps github repo).

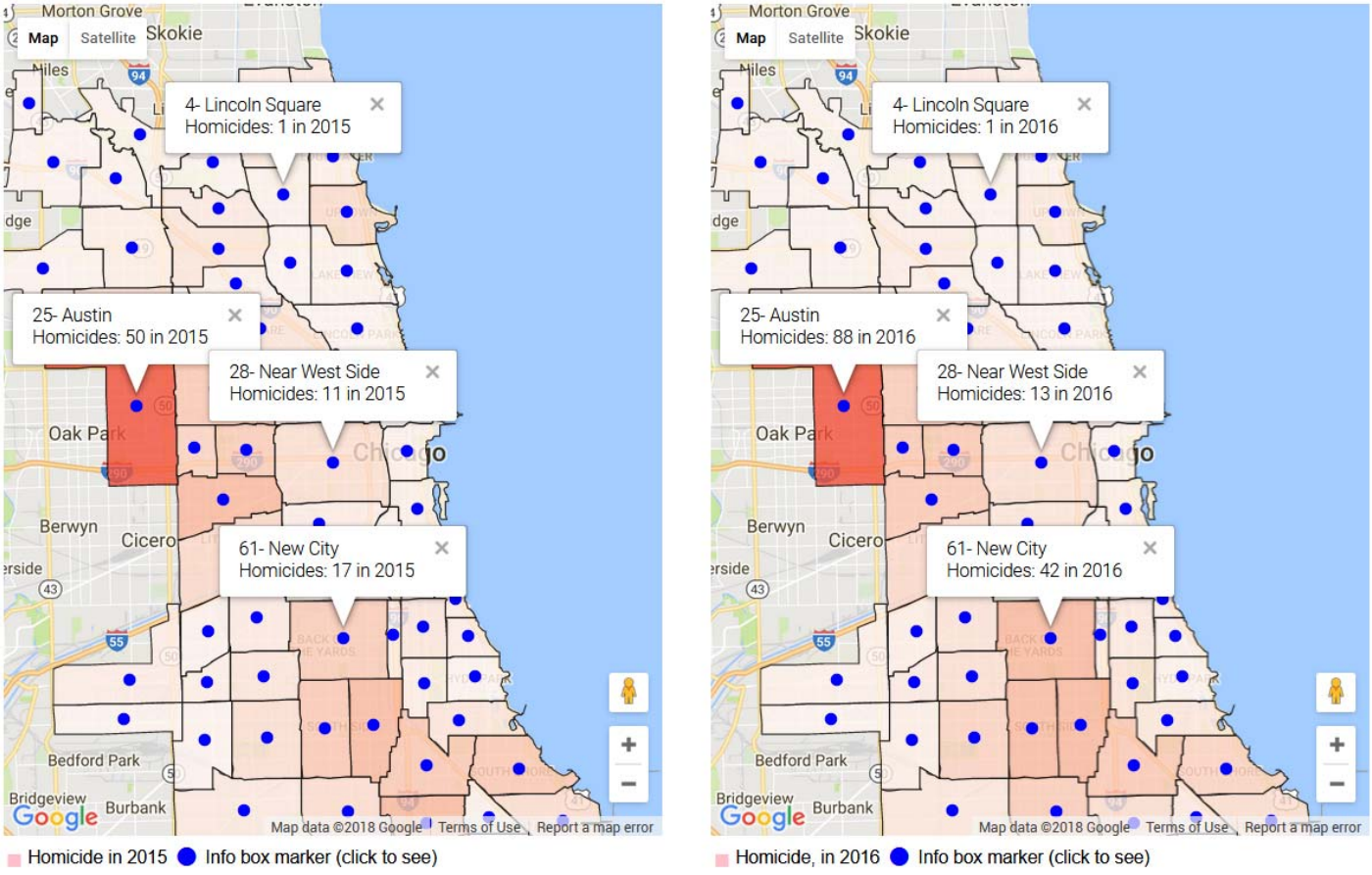

Figure 12: Murders are high only in certain communities in 2016 compared to 2015 (this is a screenshot from an interactive map)

This shows some neighborhoods got more murders compared to others. The map visualization is a great way to see the changes. Austin and New City had substantial increase in murders in 2016 (see the numbers in the pop-up info box).

This approach was used to visualize different types of crimes in different neighborhoods. On the other hand, I also used bar graphs to see if different types of crimes have different distributions in various neighborhoods. The potential patterns would serve as features for my prediction models. First all crimes counts in different neighborhoods were plotted betweeen 2001 to 2016 to show the ones with the highest number of crimes (Figure 12).

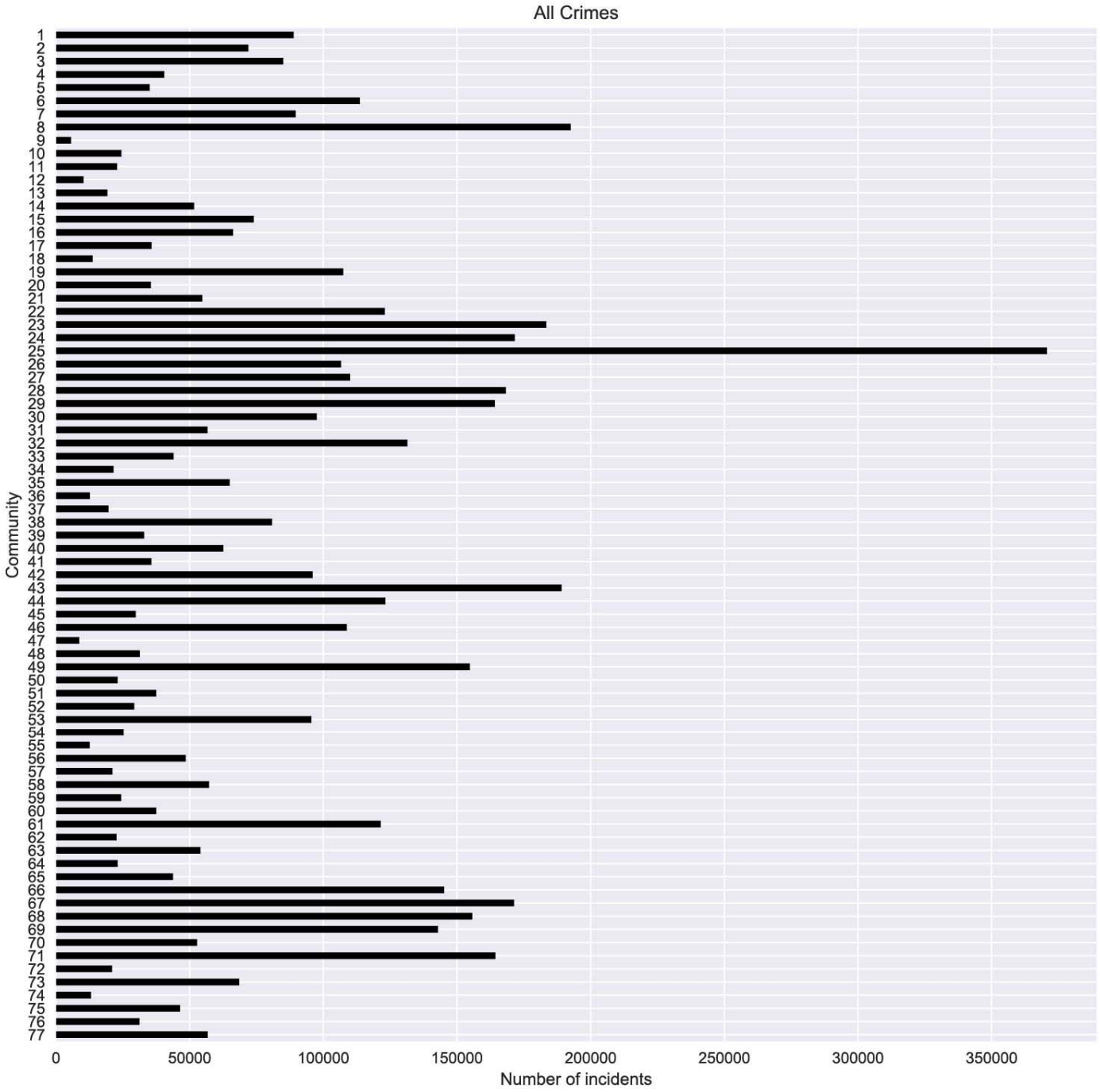

Figure 13: Community 25 (Austin) had the highest crimes in all communities in 16 years

Community 25 (Austin) has had by far the most crimes committed, however, Austin doesn't have the lowest median income or the highest poverty rate. Maybe using other statistics could reveal why it is the most dangerous neighborhood in Chicago.

Next, I examined the distribution of sex offenses in all communities in 16 years as well which is a concern to many people.

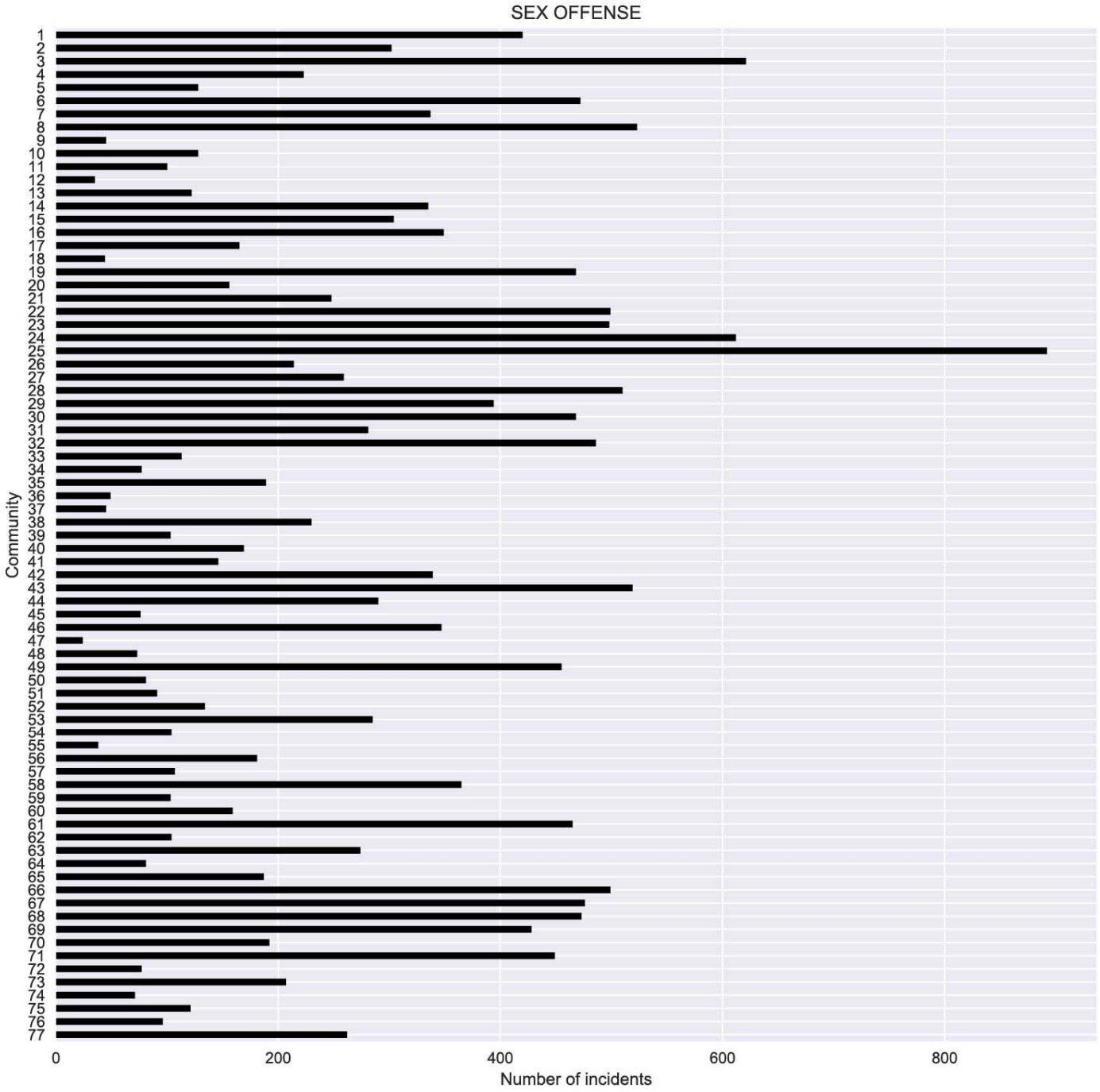

Figure 14: Community 25 (Austin) again had highest number of sex offenses in all communities

Austin still topped the list, but we also notice that Uptown (community 3) came in close second. If we check all crimes, Uptown is not even close compared to Austin. This is an interesting pattern we can use for our machine learning models: different neighborhoods have different crime profiles.

Are there any correlations between different crime pairs? We can plot a heatmap matrix to show the pairwise relationships.

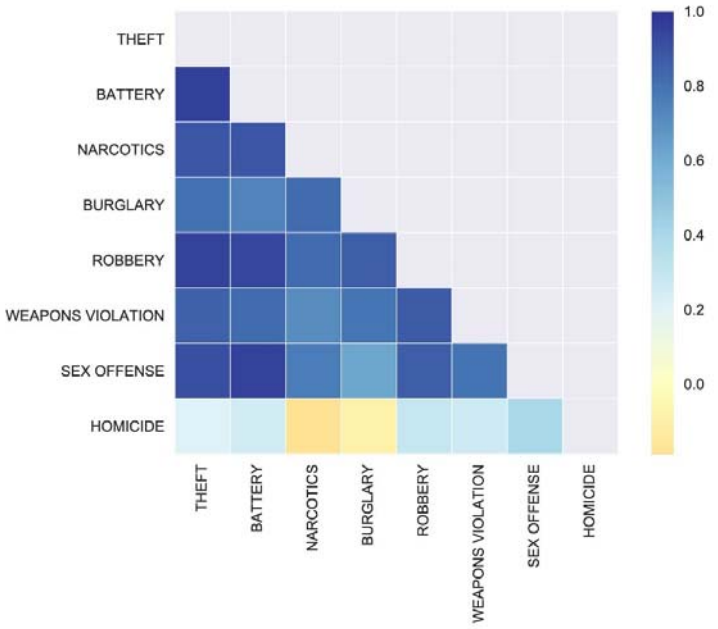

Figure 15: Pearson correlation map between the eight felonies, pairwise

Many felonies are correlated. For example, battery is very likely to be associated with robbery.

Next, I examined the distribution of these felonies over time. What month(s)/day(s) had the most crimes and what type of crimes (Figure 15)?

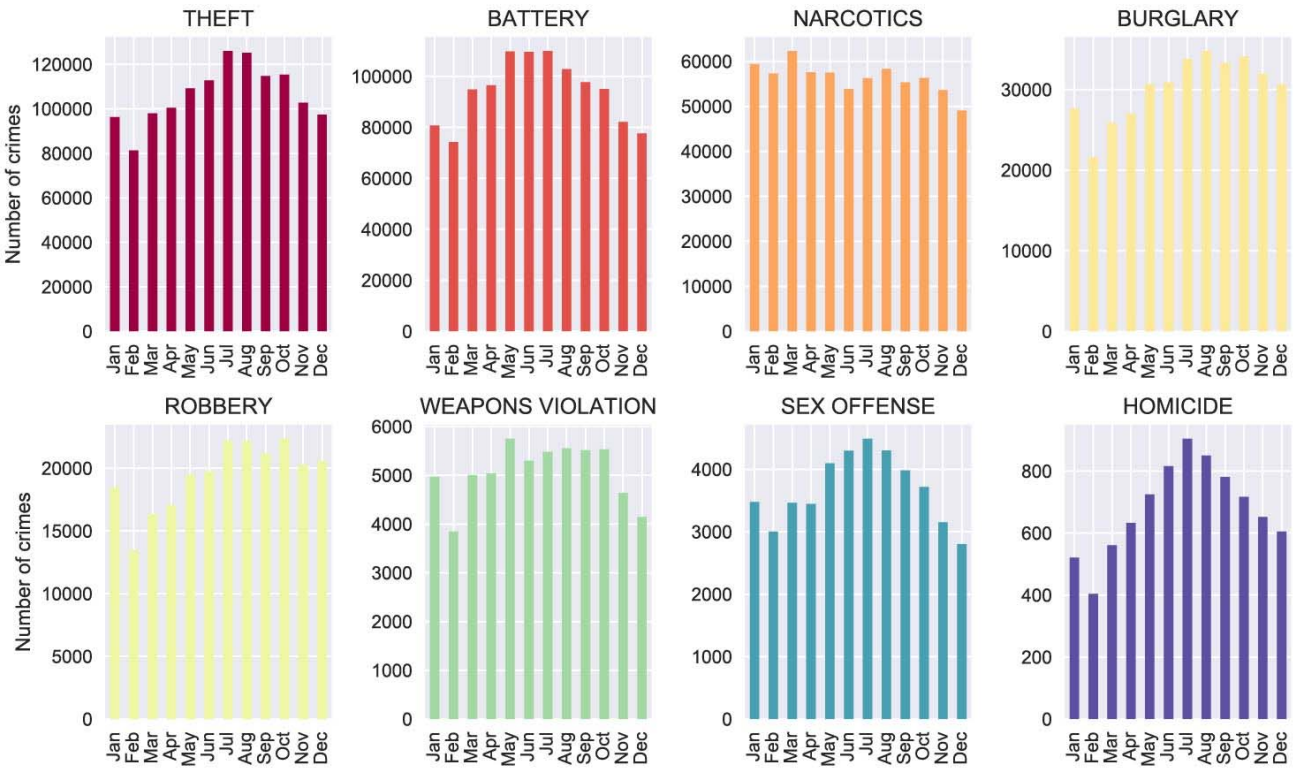

Figure 16: Monthly crimes between 2001 and 2016

Out of the eight felonies, seven peaked in warmer months (May-August) except Narcotics which shows more incidents in Jan-Feb-Mar when temperatures are supposed to be low. To confirm that most crimes occur in higher temperatures, a histogram of all crimes was plotted.

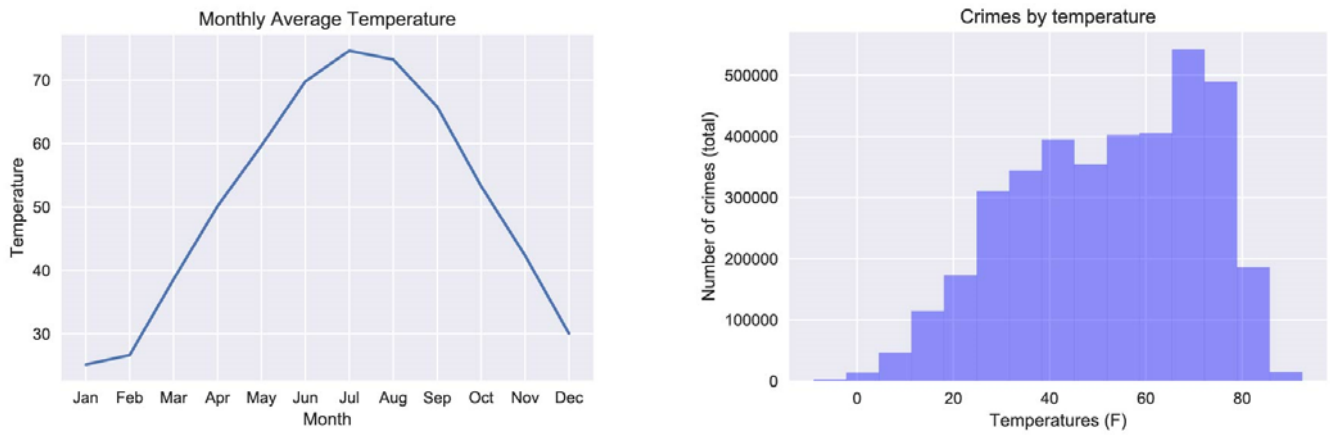

Figure 17: Distribution of all crimes at different temperature

While most crimes did occur in higher temperatures, 'Narcotics' seems to be higher in colder months. After plotting individual crimes by temperature, I could see this abnormal trend.

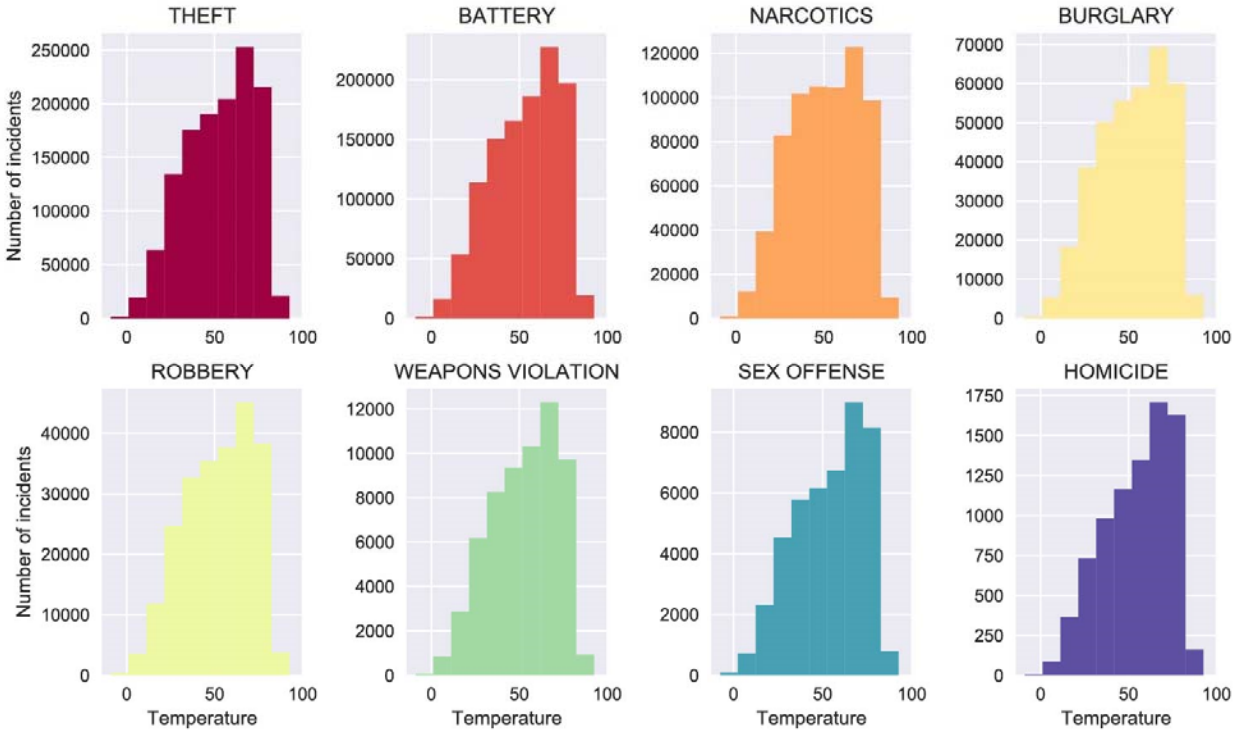

Figure 18: Distribution of individual crimes at different temperatures

The unusual shoulder around 40-50 ºF in ‘Narcotics’ plot tells us that they indeed have more incidents in lower temperatures. I then compared ‘Narcotics’ in March (highest) and December (lowest) (Figure 18).

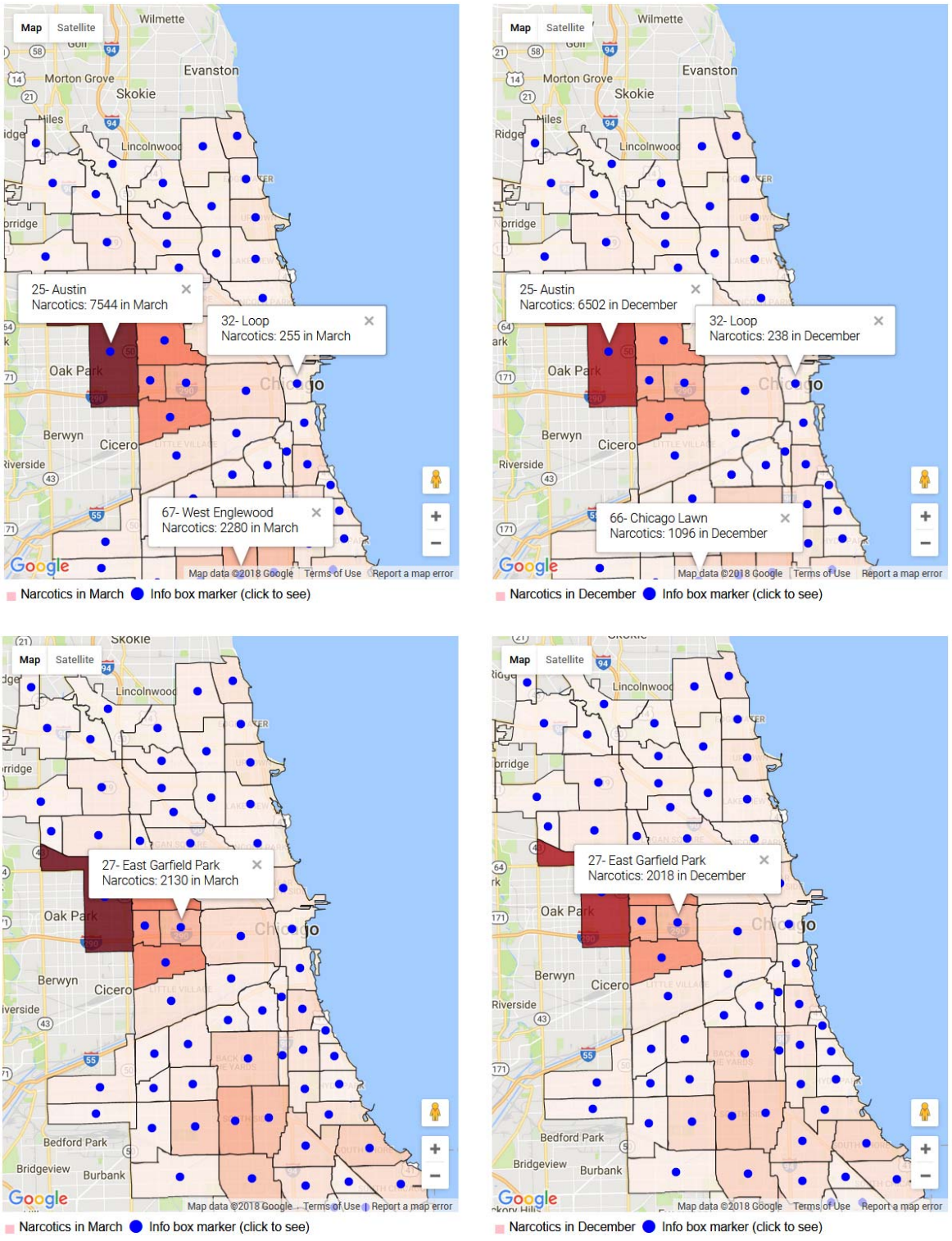

Figure 19: All Narcotics in March and December between 2001 and 2016

In Austin there is substantial decrease down by 1000 cases but in East Garfield Park there is not much difference. I also tested for significance (t-test) for Austin and East Garfield Park to compare March and December narcotics incidents between 2001 and 2016 to test if the difference for Austin was significant and there was no difference for East Garfield Park.

I took all narcotics that happened in March and December from 2001 to 2016 for either Austin or East Garfield Park. I then tested significance of comparing March numbers to December numbers for the two communities after discarding NaN values and outliers (2003-2016 kept) using *statsmodels ttest* (statsmodels package website) method in Python. Here are the results for p values (Figure 19).

	Austin	East Garfield Park
Compare March to December		
p value	0.048766	0.389275

Figure 20: T test for comparing narcotics between 2003 and 2016

For Austin, I used one tailed test ('larger') to show that in March there is more narcotics in average compared to December whereas for East Garfield Park, I used 'two-sided' test to show that there is no apparent difference between March and December. The p value for Austin is <0.05 and the p value for East Garfield Park is >0.05. Even though the 0.05 cutoff is arbitrary but from examining the actual data for these two communities, I could see the tests are real and significant (Figure 20).

Date	Austin March	Austin December	East Garfield March	East Garfield December
2003-12-31	565	545.0	197	174.0
2004-12-31	782	532.0	178	194.0
2005-12-31	737	680.0	185	163.0
2006-12-31	580	588.0	145	161.0
2007-12-31	724	547.0	209	160.0
2008-12-31	576	443.0	124	126.0
2009-12-31	596	485.0	123	132.0
2010-12-31	603	389.0	139	120.0
2011-12-31	459	444.0	119	110.0
2012-12-31	482	329.0	148	70.0
2013-12-31	431	383.0	98	156.0
2014-12-31	357	246.0	164	160.0
2015-12-31	316	159.0	158	55.0
2016-12-31	193	140.0	45	45.0

Figure 21: Austin had more Narcotics in March than in December, but East Garfield had no differences

Next, I broke down the crimes by weekdays and by hours of the day to reveal some trends (Figure 21).

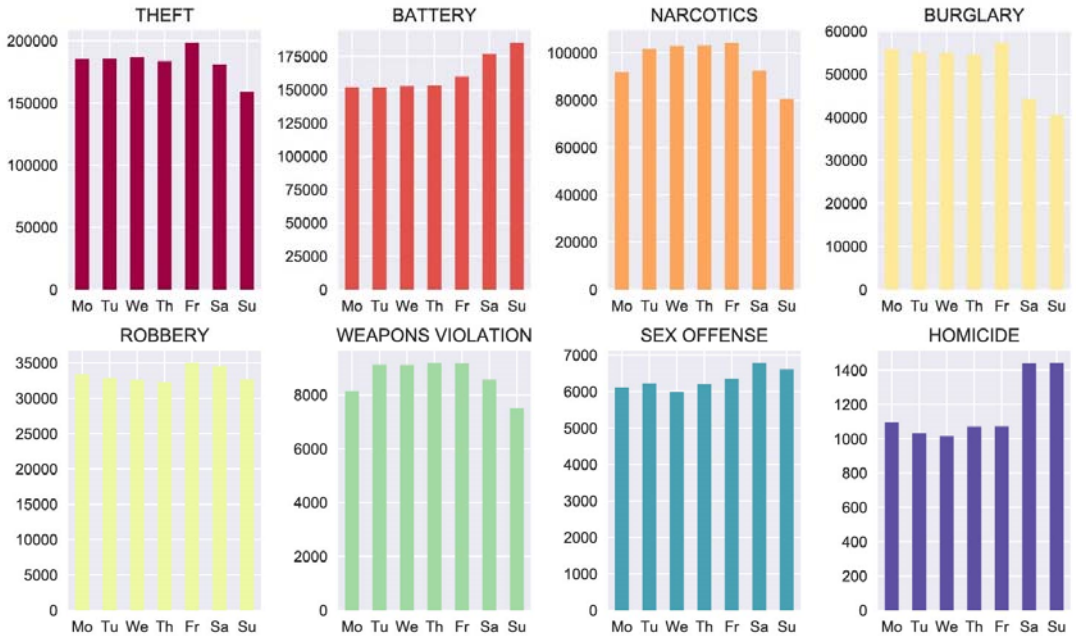

Figure 22: Daily crimes between 2001 and 2016

These crimes show two patterns, either they peaked during weekdays or on weekends. For burglaries, weekend cases are low, but homicides are a lot higher on weekends. This pattern can be used in our machine learning models to train the algorithm. Then hourly crimes were plotted (Figure 22).

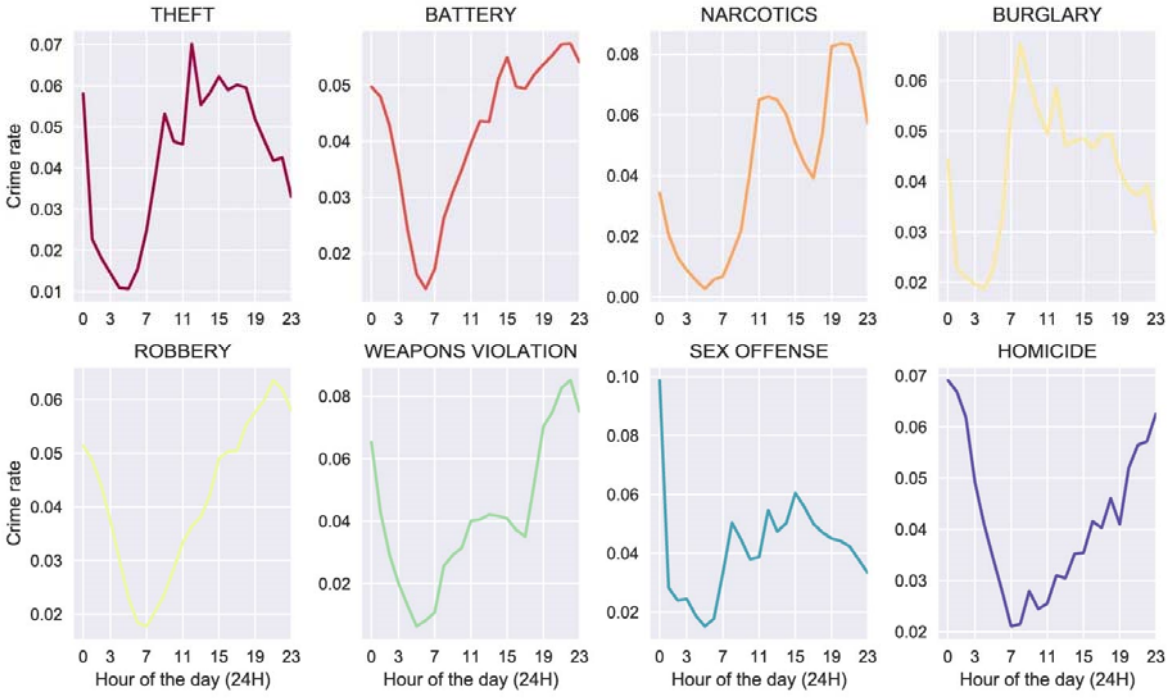

Figure 23: Hourly crime distribution

This interesting plot shows how different crimes peaked at different hours during a 24-hour period. For ‘Narcotics’, there are two peaks at 12pm and 8pm. ‘Battery’ peaked at 3pm and 10pm. ‘Burglary’ peaked at 8am in the morning and that’s when most people leave for work. Let’s compare ‘Narcotics’ of 12pm and 8pm in all communities (Figure 23).

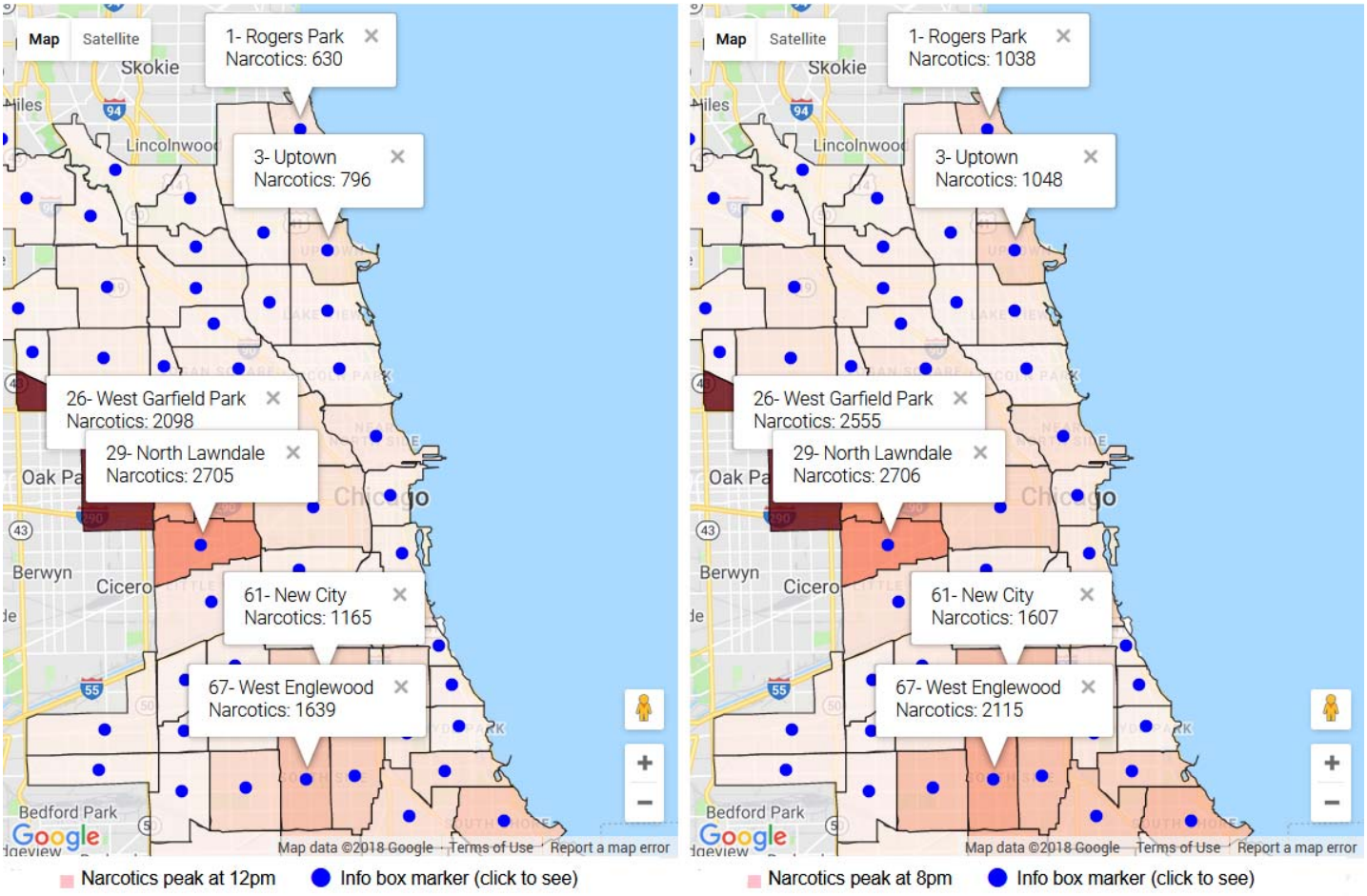

Figure 24: Narcotics are higher in certain communities at 8pm than 12pm

In North Lawndale, there is no difference between 12pm and 8pm but in other places such as Uptown and West Garfield Park, there are a lot more incidents at 8pm. A similar t-test was done to show significance like the one before. The difference for New City was significant by the p value I got.

	New City	North Lawndale
Compare 8pm to 12pm		
p value	0.03505	1.0

Figure 25: Statistical test for narcotics at 8pm and 12pm between 2002 and 2016

I then used bar graphs to compare narcotics at 8pm and 12 pm in all neighborhoods.

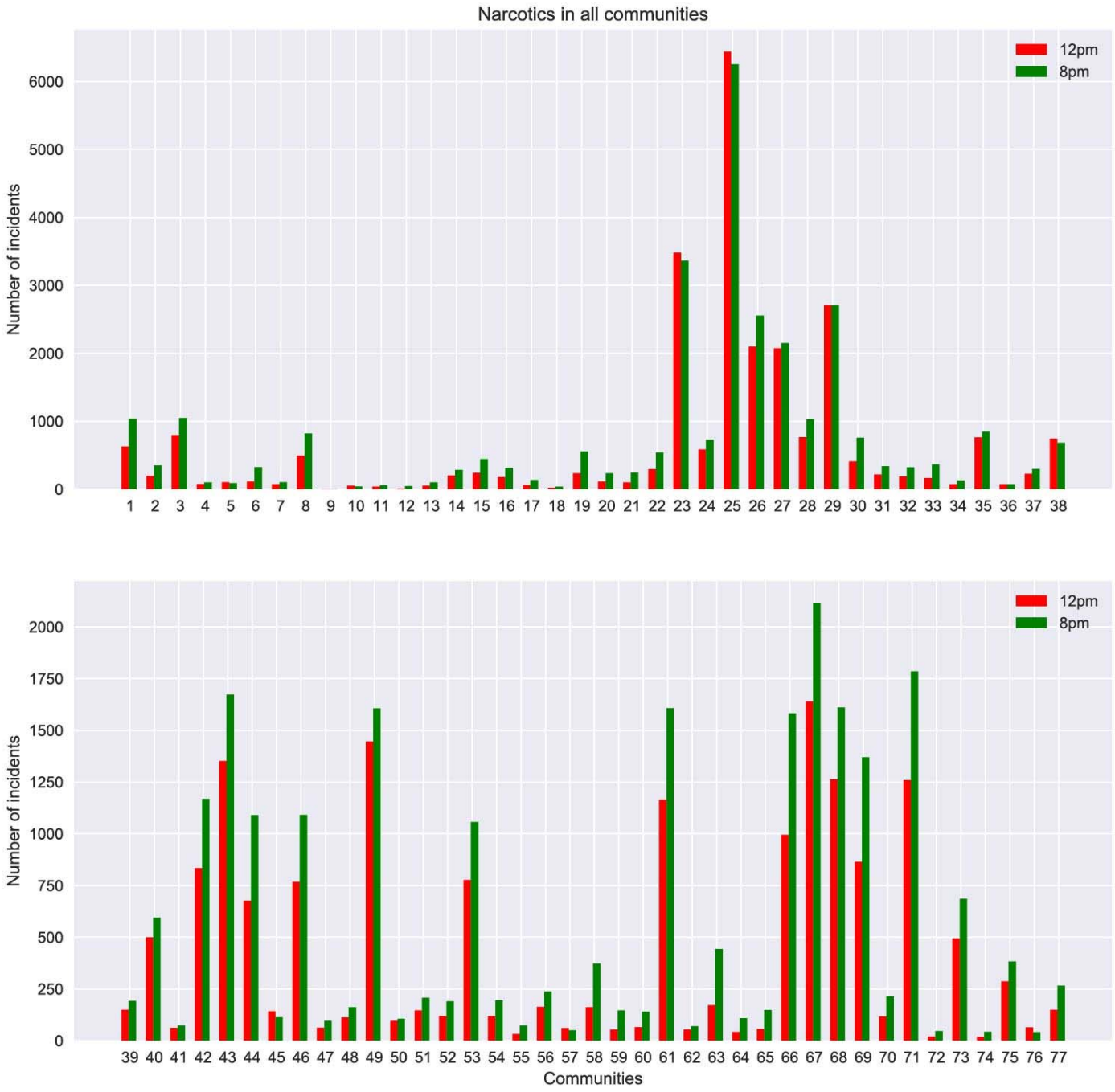

Figure 26: 'Narcotics' are higher in certain communities at 8pm than 8pm

This plot shows two clusters of communities 42 to 46 as cluster 1 and 66 to 71 as cluster 2 (these are also geographically clustered as shown in Figure 11 above) that had substantial increase at 8pm. From the above analysis, I decided to include the hour of the day as another feature to use in subsequent machine learning models for crime prediction.

To get the whole picture of Chicago crime distribution, I plotted all crimes as choropleth maps.

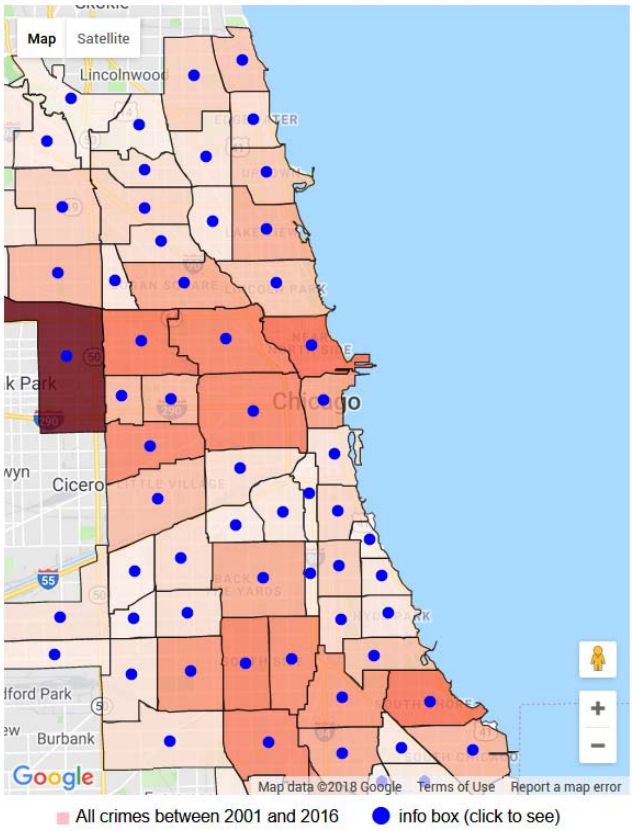

Figure 27: All crimes as choropleth between 2001 and 2016

In the West and the South, there are two major clusters of crimes. They happen to be areas where income levels are low and poverty levels are high. I then used poverty and education data to correlate with eight felonies in all communities. Eight crime data were broken down to different communities together with poverty and education averaged between 2011 and 2015 (here I used % of population with a Bachelor's degree or higher). I tested for Pearson Correlation and significance (Figure 28).

	Correlation with education level	Correlation with poverty	Crime type	Education correlation p value	Poverty correlation p value
0	0.181593	0.285939	THEFT	0.11397	0.0117
1	-0.473203	0.829419	BATTERY	0.00001	0.0000
2	-0.417031	0.647396	NARCOTICS	0.00016	0.0000
3	-0.466991	0.618249	BURGLARY	0.00002	0.0000
4	-0.362280	0.666467	ROBBERY	0.00120	0.0000
5	-0.522311	0.723058	WEAPONS VIOLATION	0.00000	0.0000
6	-0.299172	0.738853	SEX OFFENSE	0.00821	0.0000
7	-0.528851	0.794385	HOMICIDE	0.00000	0.0000

Figure 28: Eight crimes correlate positively with poverty levels and negatively with education levels

All eight felonies except theft show positive correlations with poverty levels and negative correlations with education levels for all the communities. However, when I plotted the crime rate against poverty levels for each crime type, some communities stood out as outliers (Figure 29).

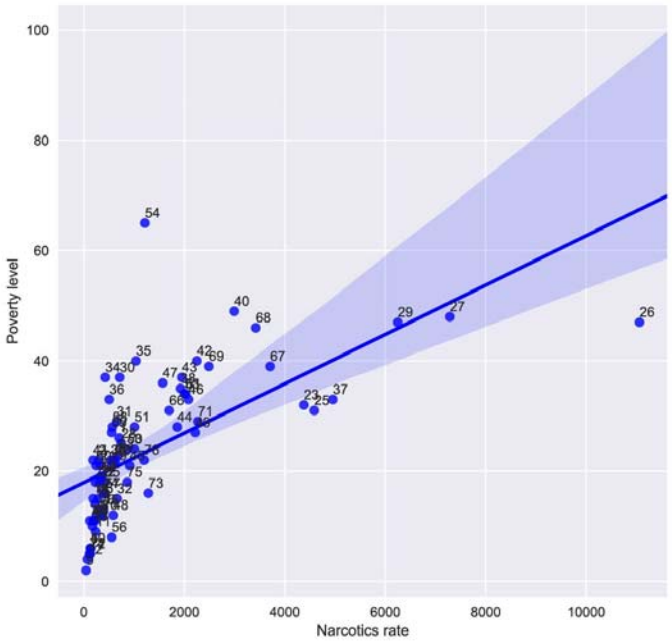

Figure 29: Community 54 (Riverdale) has the highest poverty level but very low narcotics crime rate (per 100k population)

As shown above, Riverdale is the poorest community in Chicago, but its narcotics rate is very low. However, this is not true for all other crimes. Battery for example, is very high in Riverdale (Figure 30).

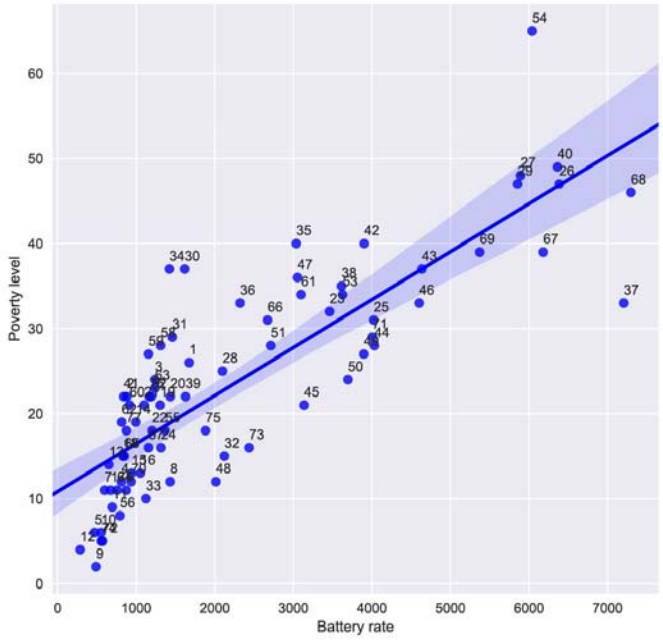

Figure 30: Riverdale has high battery crime rate (per 100k population)

Another community that showed this abnormal behavior is community 32: Loop. This is a rich community with low poverty rate. However, its sex offense crime rate is unusually high (Figure 31).

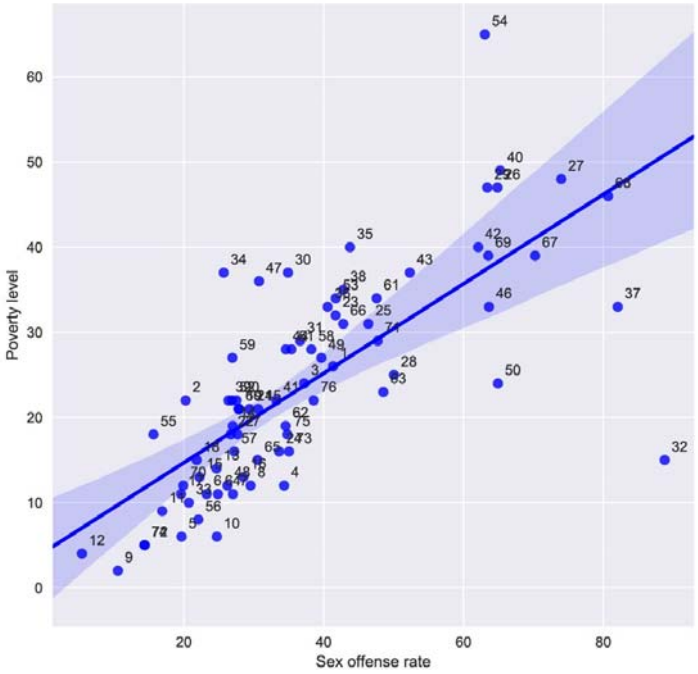

Figure 31: Community 32 (Loop) had highest sex offense crime rate despite being a rich community (per 100k population)

Loop showed the highest sex offense crime rate and yet its poverty level is in the low range. For all other crimes, Loop showed low crime rates such as robbery (Figure 32).

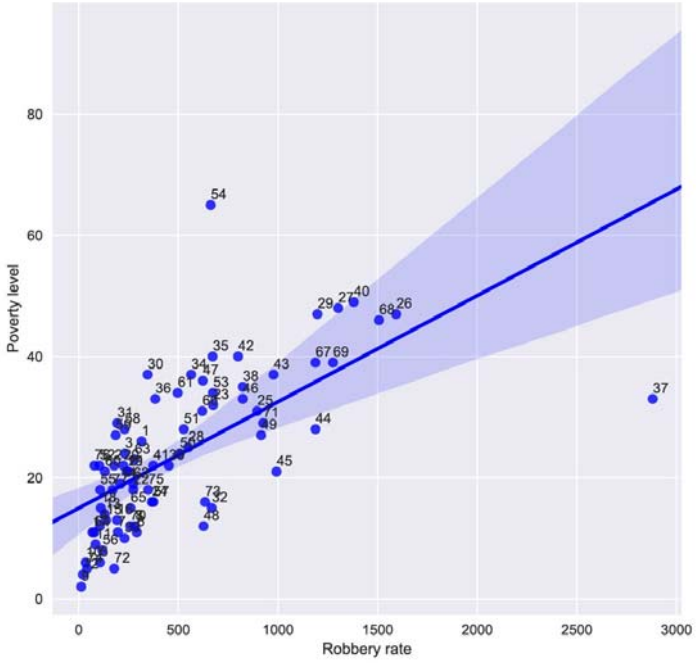

Figure 32: Loop had normal low robbery crime rate conforming to its low poverty level (per 100k population)

The abnormally high sex offense rate was also visualized on a map (Figure 33).

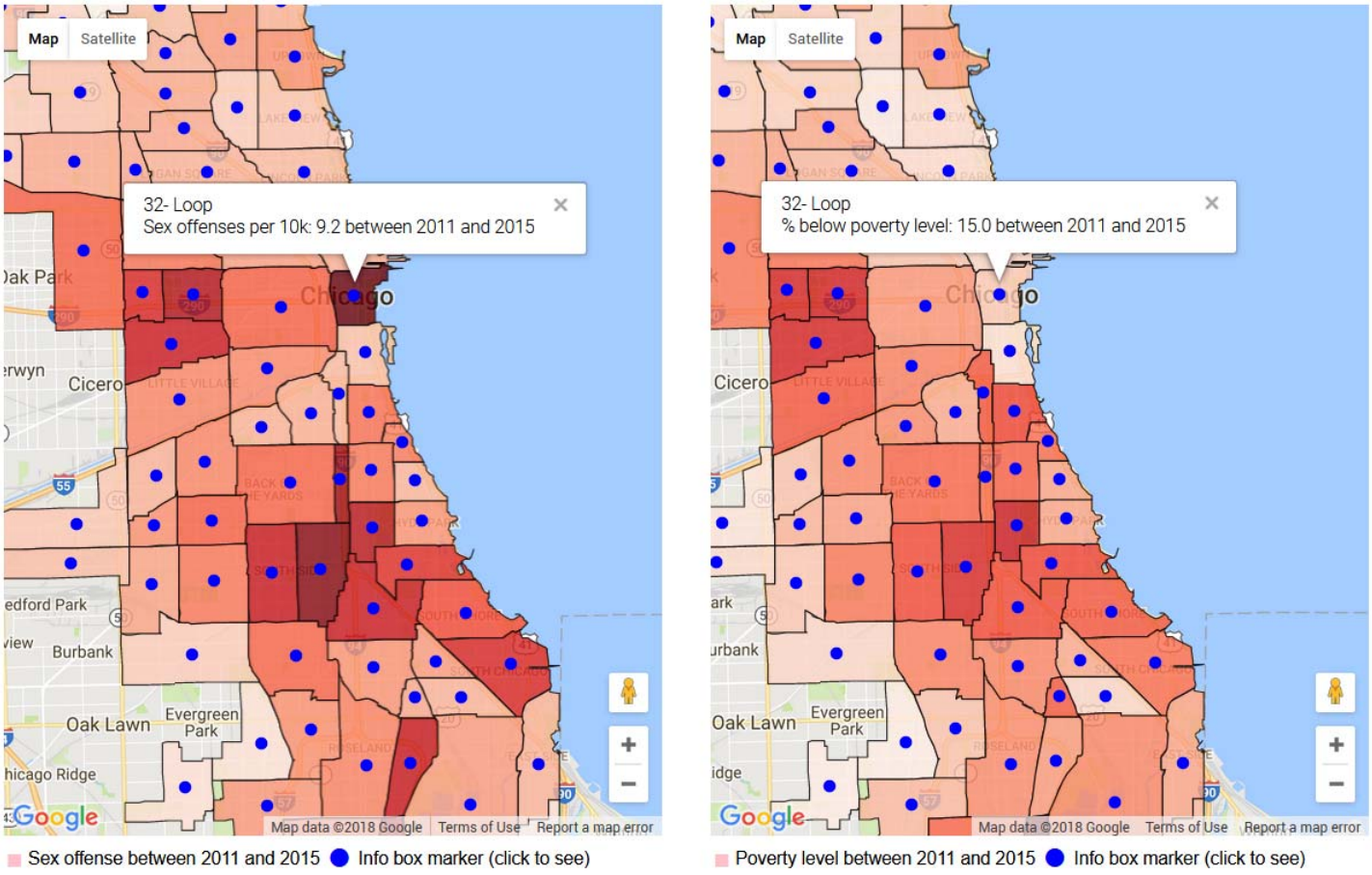

Figure 33: Community 32, Loop had highest sex offense crime rate while having low poverty level

These two communities: Riverdale and Loop are interesting targets according to the analysis above. Riverdale showed mostly low crime rates even though it's the poorest neighborhood in Chicago. Maybe the police department in Riverdale was very effective or there were few gang activities. On the other hand, Loop is not a poor community and it had unusually high sex offense crime rate. A possible explanation would be, prostitution was high in Loop and the police department did not do a good job. Maybe there was corruption in the law enforcement in Loop. If I had the details of all the sex offenses (e.g. categories), I'd be able to draw more meaningful conclusions.

I felt there was still substantial EDA work to be done to get a more complete sense of all the crimes, however, all analysis must stop at some point. I then switched gear and applied machine learning algorithm (logistic regression) to this data set using all the relevant features I noticed from my EDA. Crime prediction is hard but can be very informative in some situations. The complexities in the data set may not yield very robust models and maybe more information is needed to build a decent model.

In the following section, I will demonstrate the approach I took and the results I got with the current data and features.

4. **Build a classifier to predict crime types**

From the EDA analysis above, I have had a good sense of what features I should use in order to build my model. Because there are too many classes (crime types) for this data set, I tried the one vs rest model using LogisticRegression. I transformed the target column (crime type) to either 0 for the crime type I wanted to predict and all the rest as 1. I did this for all crime types and the example I am going to show here is for Narcotics (meaning Narcotics is class 0 and all others are class 1).

The features I used for the machine learning model were:

'Domestic, Location Description, Community Area, Month, Day, Hour, Latitude, Longitude, Temperature'

Among those, 'Latitude', 'Longitude', and 'Temperature' are numerical features and the rest are categorical. For 'Month' and 'Day', I grouped them into seasons and weekdays or weekends so for 'Month' there are four categories: Spring, Summer, Fall and Winter. And for 'Day' feature, there are two categories: weekday or weekend. I did this because according to my EDA, seasons and weekdays had more impact than individual month or day of the week. I then used sklearn LabelEncoder to transform those into integers. The DataFrame for machine learning was as follows:

	Domestic	Location Description	Community Area	Month	Day	Hour	Latitude	Longitude	Temperature	Crime Type
0	0	145	0	3	0	12	42.002478	-87.669297	34.0	0
1	0	122	65	1	0	19	41.780595	-87.683676	29.5	1
2	0	152	67	3	0	1	41.787955	-87.634037	24.0	0
3	0	47	22	1	0	16	41.901774	-87.709415	29.5	1
4	0	152	43	1	0	22	41.748675	-87.599049	29.5	1

Figure 34: Representative rows from the DataFrame for machine learning

I took a random 500000 sample from this data frame to build my LogisticRegression model. A train-test split was done to use 80% for training and 20% for testing. I then used StandardScaler to rescale my numerical features ('Latitude', 'Longitude', 'Temperature'). Pandas get_dummies was used to transform all categorical features to dummy columns (like a sparse array) for machine learning. Cross validation was used on the training data and the C parameter was tuned to pick the best C value for the model.

Since the 0 and 1 classes were not equally distributed, the model was set as class_weight='balanced' to reflect this weigh imbalance. After cross validation, the model was tested on the test data set and the confusion matrix was plotted as follows:

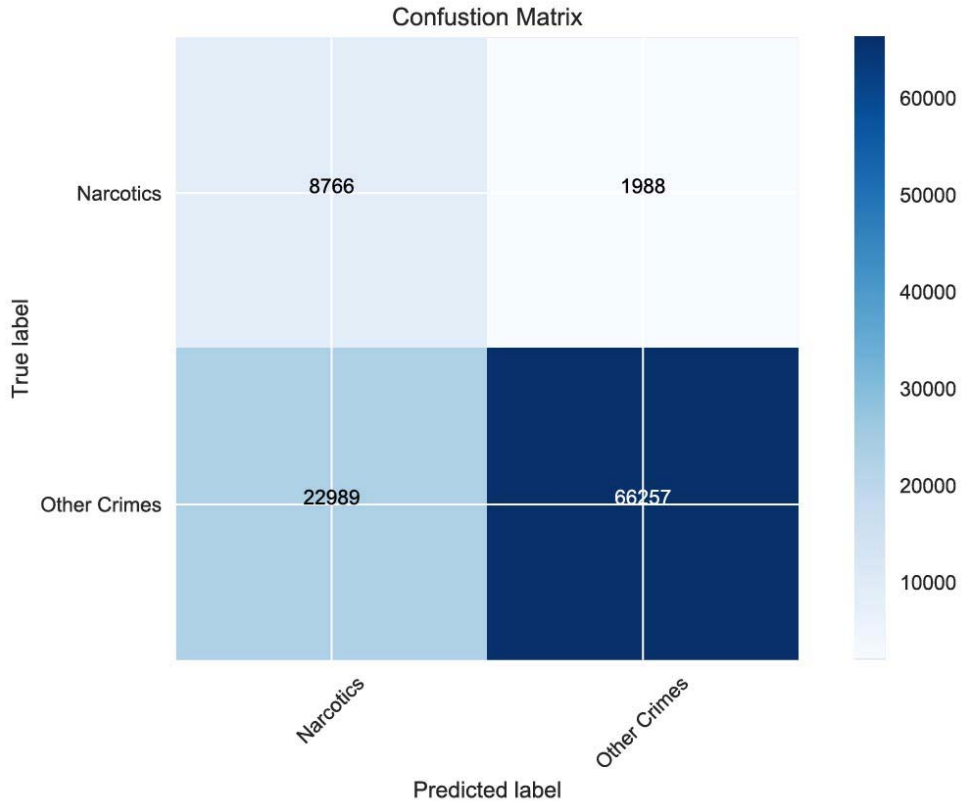

Figure 35: Confusion Matrix of LogisticRegression model on crime prediction

Note that for Narcotics the recall was pretty good but the precision was not ideal mostly due to the class imbalance in the original data set. On the other hand the other class (rest of the crimes) had both good precision and recall. The ROC curve was also plotted as below and AUC value was calculated (see figure legend):

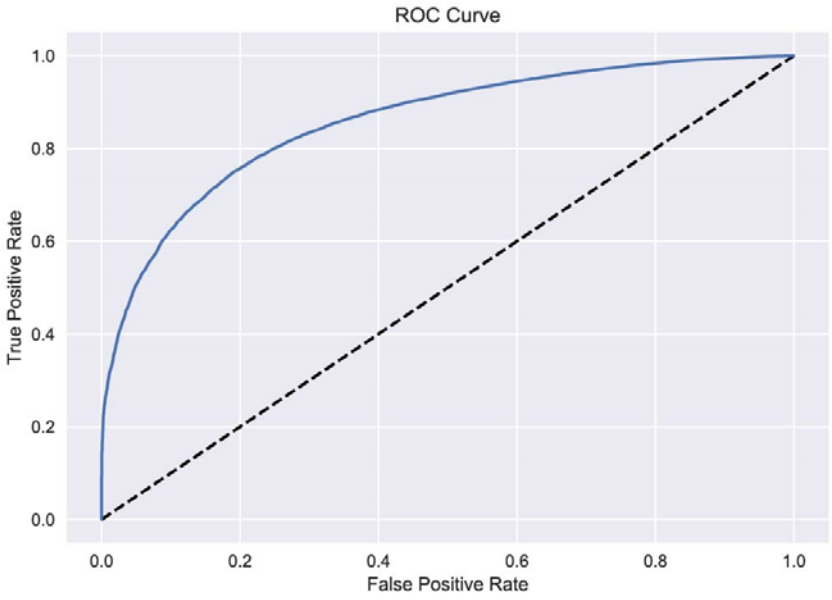

Figure 36: ROC curve of the LogisticRegression model. Area Under Curve (**AUC**) is: **0.85835622162**

The AUC was greater than 0.86 which was a pretty good value for logistic regression model. This same approach could be used for all other crime types (the one vs rest model) and then when applied to a new piece of data, all models could be used to compute probabilities for all crime types and then we could take a majority vote on them.

5. **Conclusion and future work**

Based on my analysis, several conclusions can be drawn about this data set and they are listed below.

Conclusions:

- Overall crime rates decreased in Chicago, however, weapons violations and homicides were on the rise. Especially for homicides, 2016 was a bad year. Future data after 2016 would be able to test whether homicides would keep rising. If true, it would be Chicago's top priority to work on because murder rate is still the no.1 metric when judging whether a city is suitable for living. Gun problems may be the culprit for the unusually high murder rate in this city and worth working on.
- Even though most crimes tend to happen in warm weather, Narcotics were found to happen more in colder month (March). Austin (community area 25) seemed to have more Narcotics than other communities in this regard. When plotted by temperatures, Narcotics seemed to happen more often in temperatures around 50 ^{o}F. Maybe all the drug dealers and transactions were done when fewer people were outside in the cold. I would suggest law enforcement in Chicago pay more attention to drug violations in colder months.

- Different crimes show different patterns during a week and the hours of the day. Some crimes happened more often during weekdays while other peaked on the weekend. For instance, burglaries happened most during weekdays around 8am in the morning and lowest on the weekend. On the other hand, homicides peaked on the weekend and around 11pm while lowest during the week in day time. These time patterns could serve as a guide for police officers to consider when patrolling in the city.
- Social economic status is usually indicative of crime rates and there is a correlation between them. Low income level and high poverty rate most likely would lead to more crimes. But it is not always true in Chicago. On example I found during my EDA was in Community 32, the Loop. The Loop had one of the lowest poverty levels (<15%) in Chicago but there was an extremely high sex offense rate (92 cases per 100000 population between 2011 and 2015) when compared to the entire city. This is worth investigating and maybe a brothel was located in Loop or an active sex offender lived there and needed to be brought to justice.
- When visualized on google map using choropleth, the West and South of Chicago showed up as bad areas (clusters also revealed by bar plots). The North area was far better in terms of total crimes, but they also showed some high crime rates for certain types of crimes (e.g. Narcotics and Sex offense). Areas around Chicago University were much better (map not shown in this document but in jupyter notebooks) even though it is located in the South (bad areas in general). Public safety in the university's perimeter is better than other areas.

There is still more to work on this data set and I will bring up some ideas I had but did not have the time to accomplish for this project.

Future work:

- The causation between weapons violation and homicides is worth investigating. By gathering data on the cause of each homicide to see if firearms were involved could shed light on this relationship.
- Look into the strange behavior of Narcotics that peaked in colder months unlike all other crimes. Maybe find the specific neighborhoods that showed this pattern and send more police officers in colder weather to these neighborhoods.
- One idea was to look into association of certain crimes with public transportation to see if some crimes tend to happen more often around transit lines.

- Another fact that someone could investigate was whether school closings affected crimes rates in the vicinity. Chicago has had a lot of school closings over these years and a time series analysis combined with map visualization could certainly gain some insights.
- Regarding the crime type prediction, other classifiers can be tested and tuned to see if they would provide an even better model. For example, random forest classifier offers more flexible parameter set and tuning. It also handles multiclass prediction better.

Acknowledgement

I'd like to thank Springboard and my career track mentor Danny Wells for your help along the way. The weekly Skype meeting and the community on Springboard website are great resources for me to learn and get insights. A special thank you to my current postdoc mentor Nancy Hollingsworth for her warm support. And always thank my family for their caring and understanding.