

Chicago Crime Data Analysis

Group: Jason Becker, Ankith Gunapal, and Eric Yang

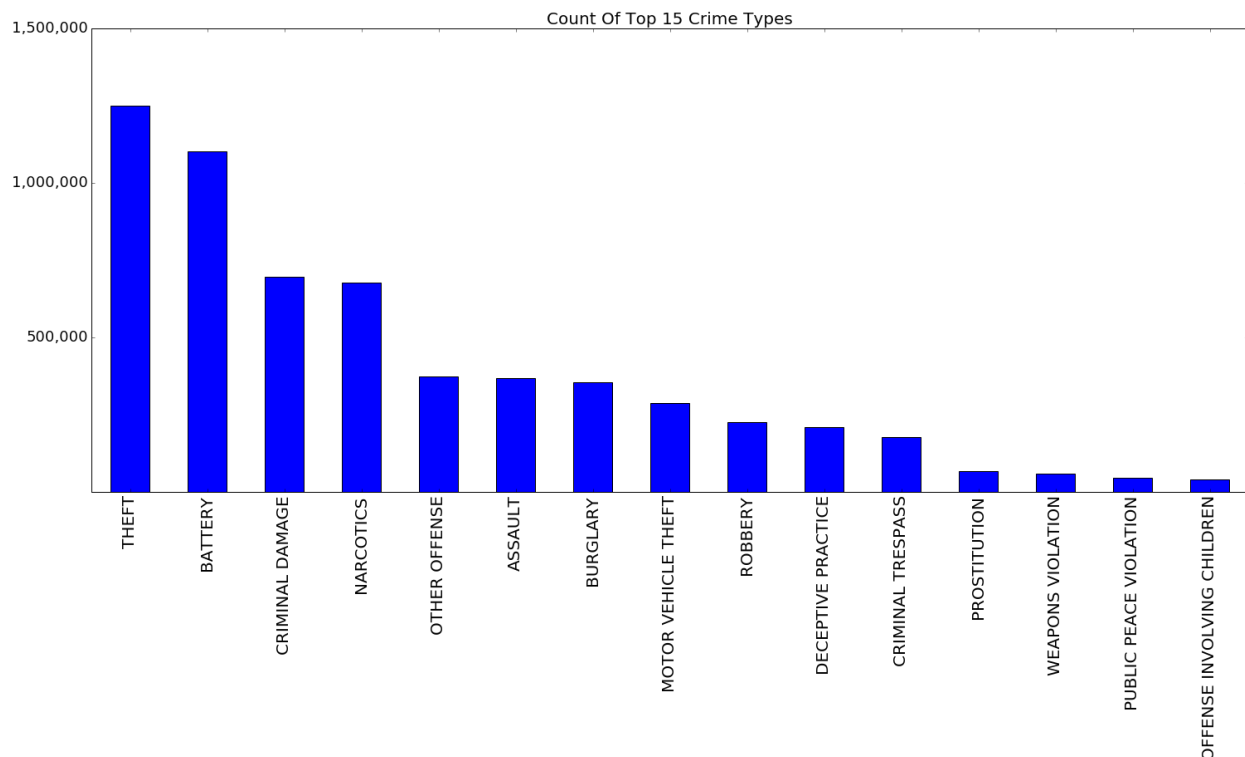
Project Outline

The City of Chicago provides data on all reported incidents of crime in the city of Chicago from 2001 to present (minus seven days). At the time of download that included 6.03 million lines of data, or approximately 400,000 crimes per year, or almost 8,000 crimes per week.

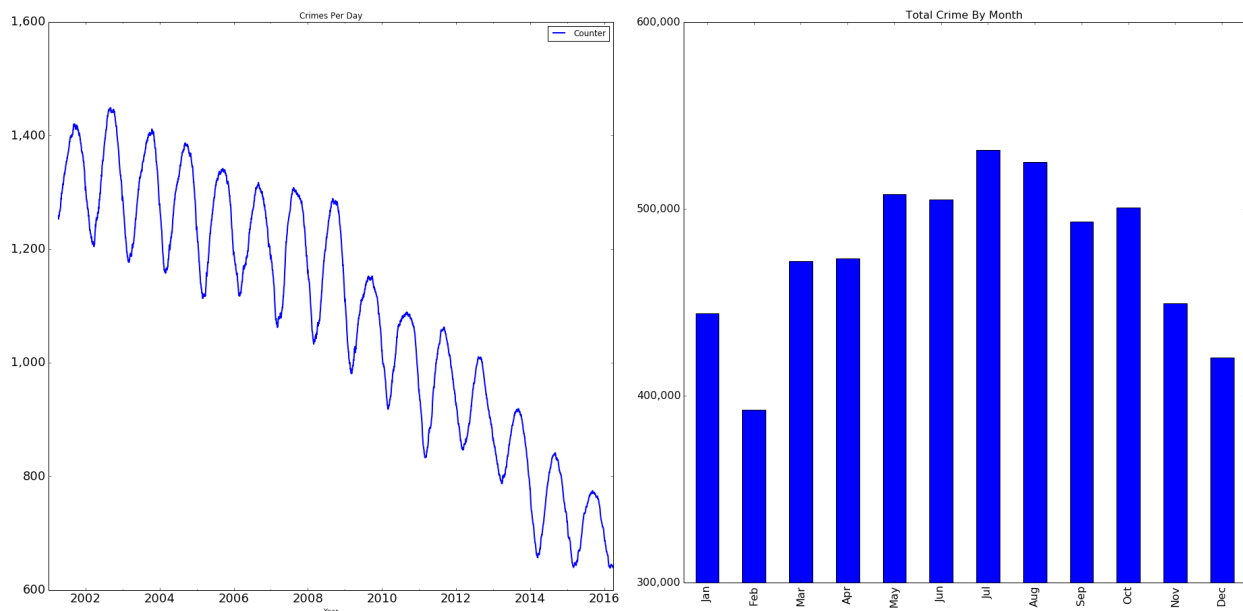
The key data points include types of crime committed, location of the crime committed, and when the crime was committed. Broadly, we plan on using this data to establish a clear picture of what types of crime are being committed in Chicago, how this has changed over time, and what if any common factors are associated with certain types of crimes. Aside from the initial dataset we brought in two more datasets in order to investigate the correlation between types of crimes committed and socioeconomic factors as well as crime and weather patterns.

Data Overview

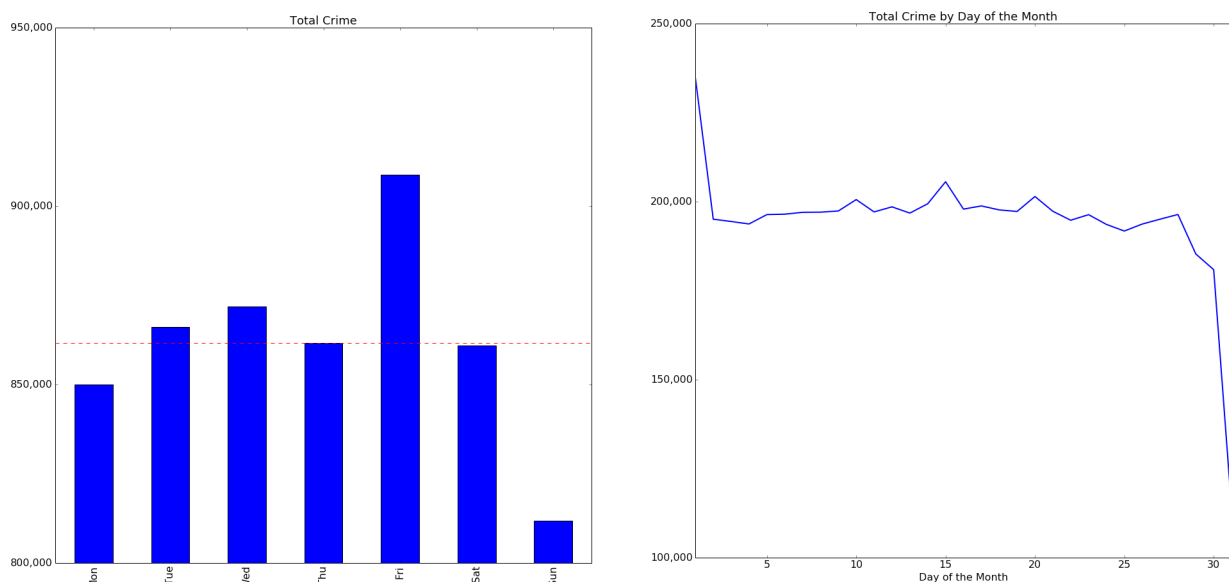
Each crime has 22 columns of information including the date and time, case number, type of crime, district, ward, a location description as well as geospatial information. Each crime is categorized into 35 groups ranging from homicide to non-criminal activities, below are the 10 most common categories of crime.



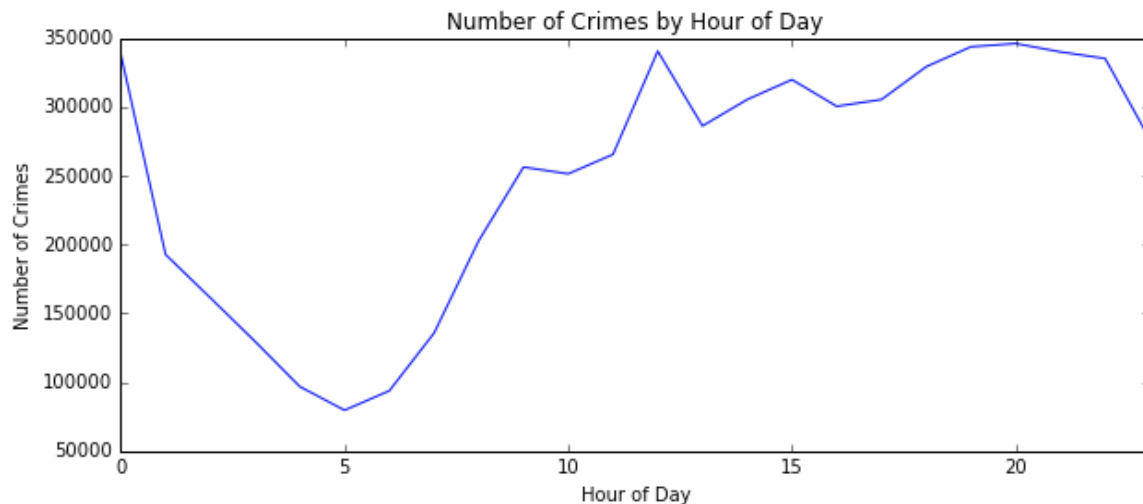
The dataset spans 15 years of crime information, so it is also important to understand how it changes over time, as well as the patterns that emerge over time. Over the entire data set two patterns emerge. Every year shows a cyclical pattern of increasing and decreasing crime, as well as a strong decline in the overall occurrence of crime. shows a 100 day rolling average of crimes per day over the entire time frame. The frequency of the cyclical pattern appears to be annual, and this is borne out by breaking down all crimes into the month in which they occurred after excluding the partial results of 2016. Summer months have had up to 100,000 more crimes than the low of February. An exception to this seasonality trend is the month of January, which has had more crime than adjacent months.



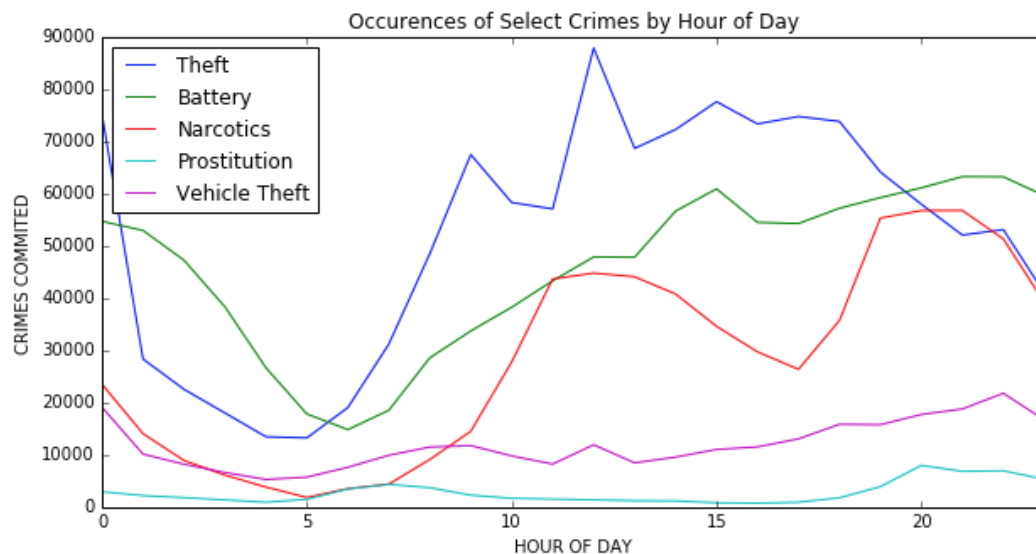
Breaking down to the day of the week shows a an increase on Fridays and a decrease on Sundays, however the variance is only 11% (whereas monthly crime varies by 29%). The day of the month looks relatively consistent with the exception of the first and the 31st. Only 7 months out of the year have a 31 days, which explains the nearly 50% reduced incidence of crime, however no such explanation exists for the first of the month.



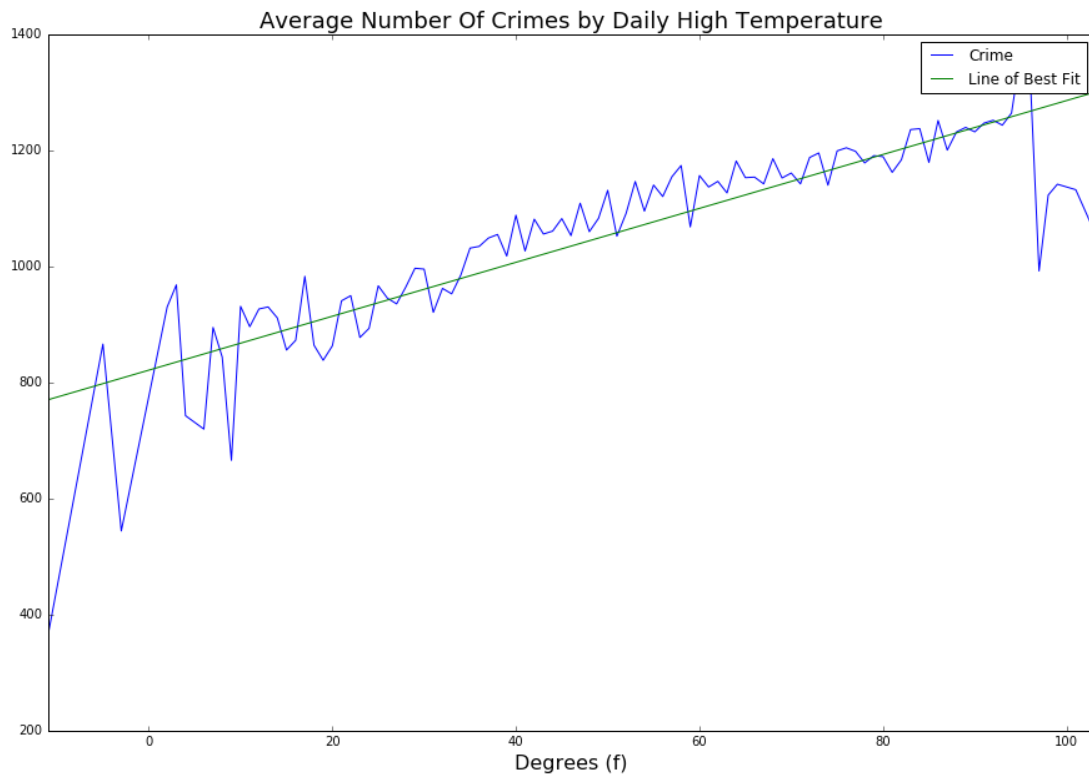
Examining number of crimes by hour of day shows that frequency of crime is at its lowest at 5AM in the morning. This is likely because in the early morning hours most people are sleeping or at home. Crime steadily increases from the low point at 5AM and reaches it maximum level at 8PM. There's an interesting spike in crime at 12PM, which is largely driven by a high incident of theft during this period. Theft at 12PM accounted for 25% of crimes during this specific hour, whereas theft only comprised 20% of crimes during all other hours.



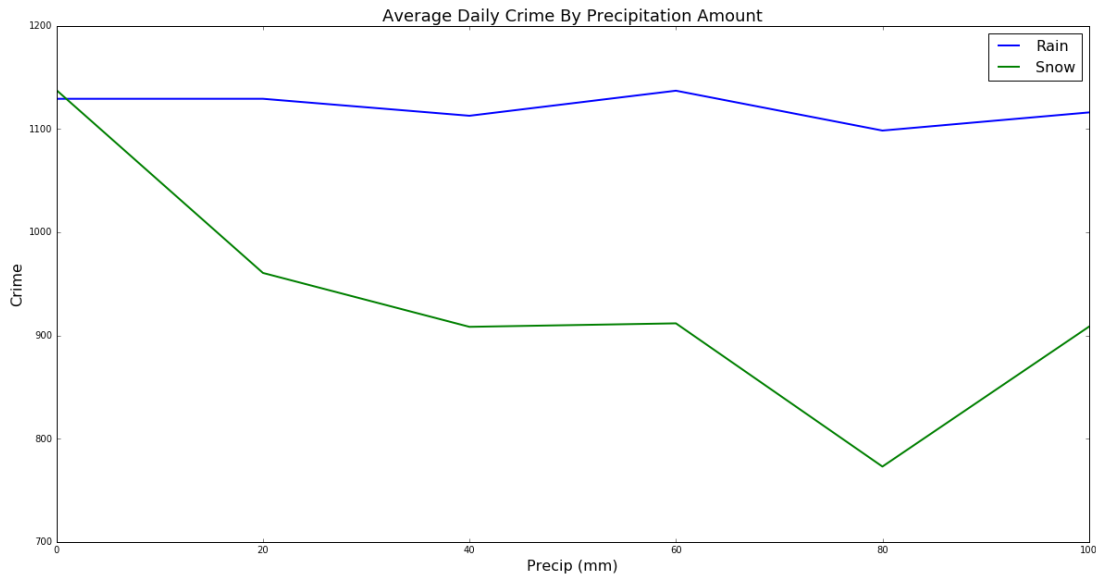
Looking at the frequency of specific crimes by hour of day also reveals some interesting findings. We see that theft does indeed peak at 12PM and in general is at it's highest in the early afternoon period. This is unlike battery and vehicle theft which increase steadily throughout the day and peak in the late evening. Narcotics crime shows an interesting pattern with two spikes throughout the day in the late morning and early evening. Prostitution, as expected is at it's peak during the late evening hours.



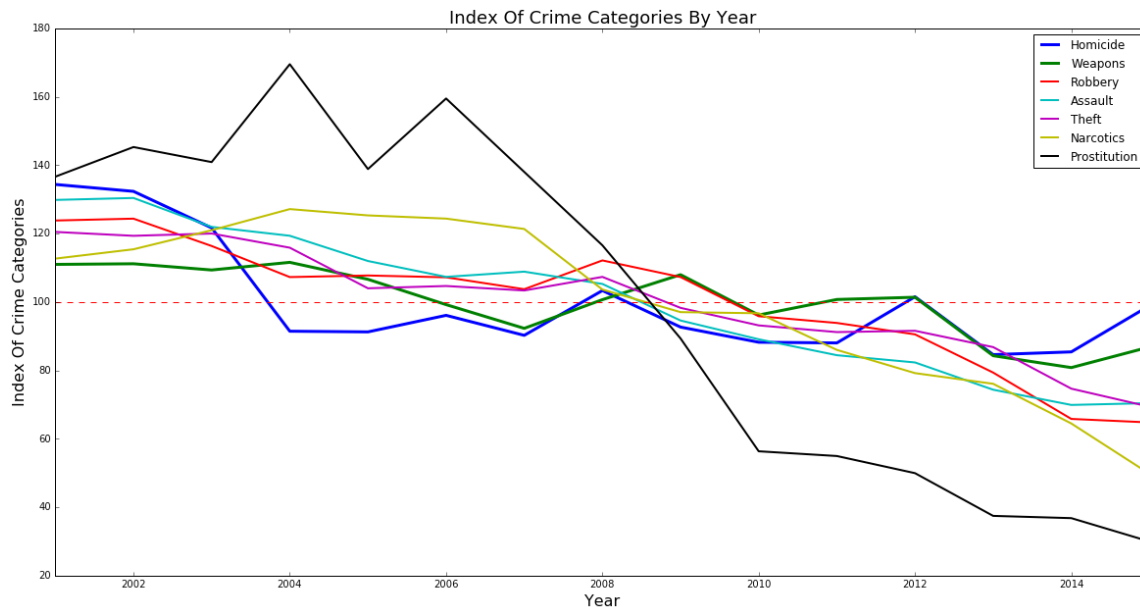
Every year crime increases and decreases on an annual frequency; crime is higher in the summer and lower in the winter. To see if this holds true, weather records for every day in the data set (up to 2014) were merged together. By plotting the average number of crimes committed by the daily high temperature, the pattern holds up. Comparing the results to a line of best fit, we see increasing volatility at both extremes of temperature.



It would also be expected that weather events like rain and snow would have a negative impact in the daily crime rate. Rain and snowfall amounts were bucketed into 20 mm groups and then plotted versus their average count of daily crimes. Increased rain appears to have little impact on crime, however snow is associated to significantly lower occurrences of crime.



The data show a strong trend of decreasing crime, however that conflicts with many current perceptions of crime in the city. Whether it's news articles [[Washington Post](#)] or public response to Spike Lee's recent film *Chi-Raq*, there is a feeling of increasing crime and violence in recent years. Perhaps it stems from increased media coverage of crime, or is a side effect of ongoing arguments between the city, state, and police departments. One possible explanation for this perception is that not all types of crime are falling. In the below graph each type of crime was indexed to allow for comparison, meaning that values below 100 are below their 15 year average.



Most categories are showing consistent decreases over the data set, however homicides and weapons violations (green and blue lines) are currently at or near their 15 year average. If this

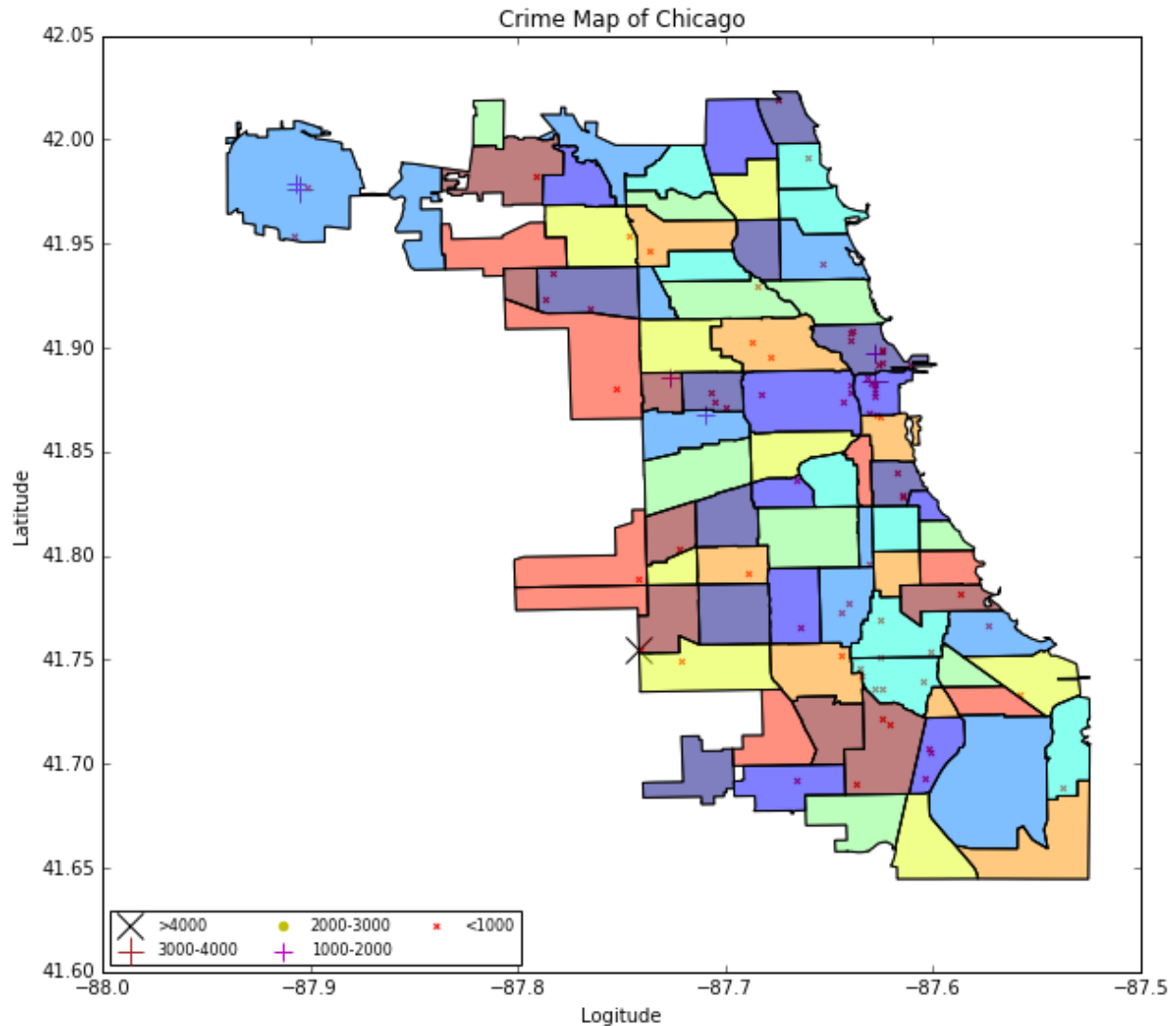
study were performed on only the last 10 years, homicide is indexing above average at 104. Recent studies are showing 2016 murders are currently up 72% and shootings are up 88% compared to the first three months of 2015 ([USA Today](#)).

Analysis with location information

The dataset comes with location information (latitude & longitude) for each reported crime. Using this, we can sort the locations based on the number of crimes committed in a given location. In the below table, Count shows the total number of crimes committed in a given location.

	Count	Location
Index		
0	4078	(41.754592961, -87.741528537)
1	3517	(41.976290414, -87.905227221)
2	3215	(41.883500187, -87.627876698)
3	2290	(41.909664252, -87.742728815)
4	2168	(41.897895128, -87.624096605)

We used the geopandas library to read in the Chicago map and plot a scatter plot of the top 100 locations . The resulting map is shown below



We want to look at domestic crimes and check where they occur the most.

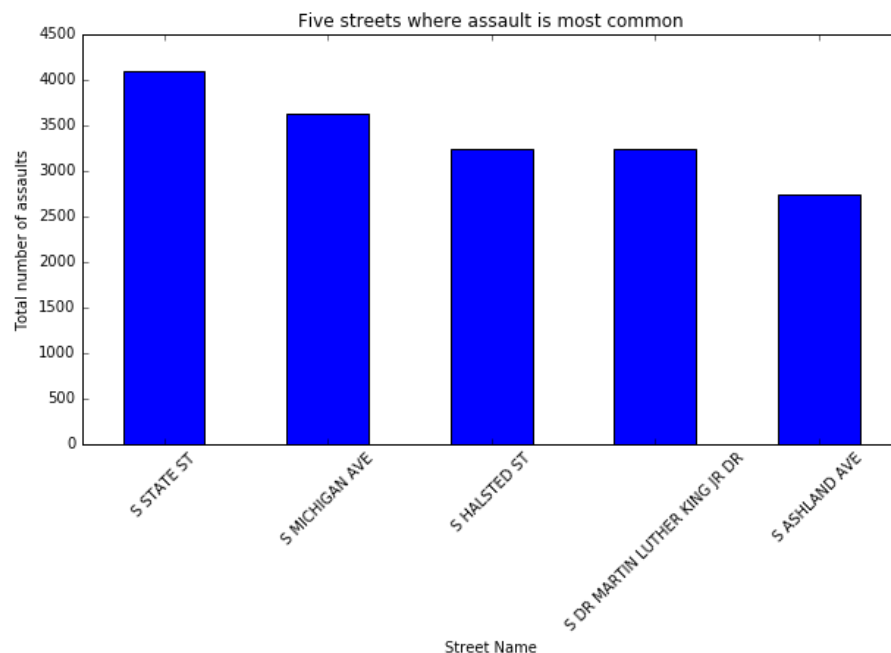
- One surprising observation below is that Street is the third most likely place where a domestic crime occurs

RESIDENCE	298006
APARTMENT	229256
STREET	95779
SIDEWALK	49467
RESIDENCE PORCH/HALLWAY	15112

- As expected, the most common crime locations are not a place for domestic crimes

Location	Domestic	Count
(41.754592961, -87.741528537)	False	4013
(41.976290414, -87.905227221)	False	3470
(41.883500187, -87.627876698)	False	3209
(41.909664252, -87.742728815)	False	2266
(41.897895128, -87.624096605)	False	2156

- Create a new column 'Street Name' by stripping the building number from the column named 'Block' . This new column can be used to find on which street , any kind of crime is more likely to happen
- This column can be used to identify on which street 'Assault' is common. We obtain the following graph



We can also obtain the Top 5 number of crimes that have occurred in any street



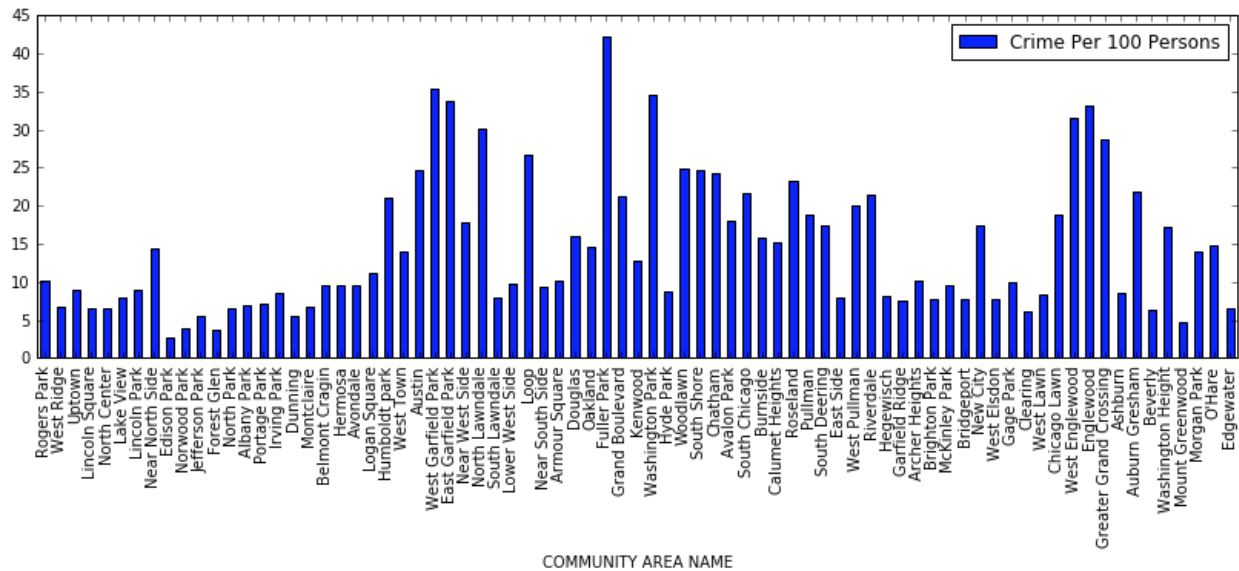
Crime and Socioeconomic Variables

We were also interested in evaluating crime against several socioeconomic factors such as unemployment, per capita income, high school education, etc. We brought in two additional datasets that included socioeconomic indicators by neighborhood and population by neighborhood and joined this information to our original crime dataset on neighborhood areas. The datasets were both from the City of Chicago and are located here:

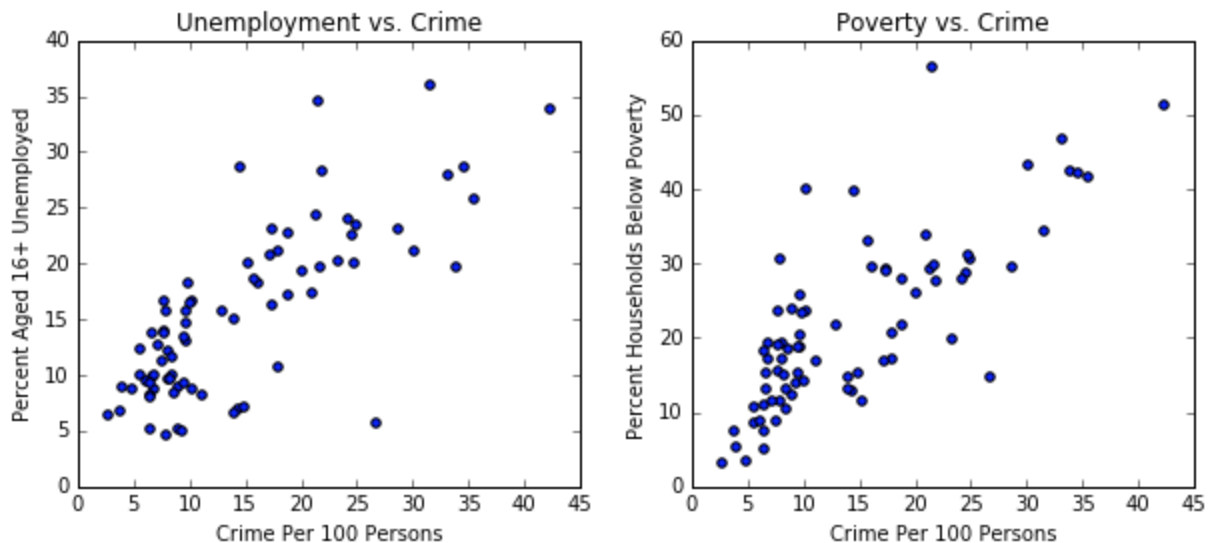
[City of Chicago - Socioeconomic Indicators](#)

[City of Chicago - Population by Neighborhood Area](#)

The City of Chicago population data allowed us to measure incidents of crime as a portion of population; segmented by neighborhood area. This showed that the top three neighborhoods in terms of crime rate were Fuller Park, Washington Park, and Edgewood. Fuller Park showed a very high 42.2 crimes per 100 people living in the area.



Our hypothesis was that the crime rate by neighborhood would have a strong positive correlation with negative socioeconomic factors such as high unemployment rate and poverty rate. Both these socioeconomic indicators do show a high correlation with crime, with r values of 0.76 and 0.77 respectively. Potentially neighborhoods with high unemployment and poverty reflect the limited economic opportunities in the area leading to higher incidents of crime.



Other negative socioeconomic factors did not show a strong positive correlation with crime. Neighborhoods with a high percent of crowded housing or those with a high percent of without a high school diploma show do not necessarily have higher incidents of crime. This was somewhat surprising, especially for crowded housing since one might think that crowded housing could be an indicator of poverty. Crime and per capita income showed a slightly

negative correlation. The negative correlation with income makes sense since ~34% of the crimes in our dataset are related to theft.

R Values	% Aged 16+ Unemployed	% Households Below Poverty	% Housing Crowded	% 25+ Without High School Diploma	Per Capita Income	Hardship Index
Crime Per 100 Persons	0.76	0.77	0.01	0.08	-0.34	0.55