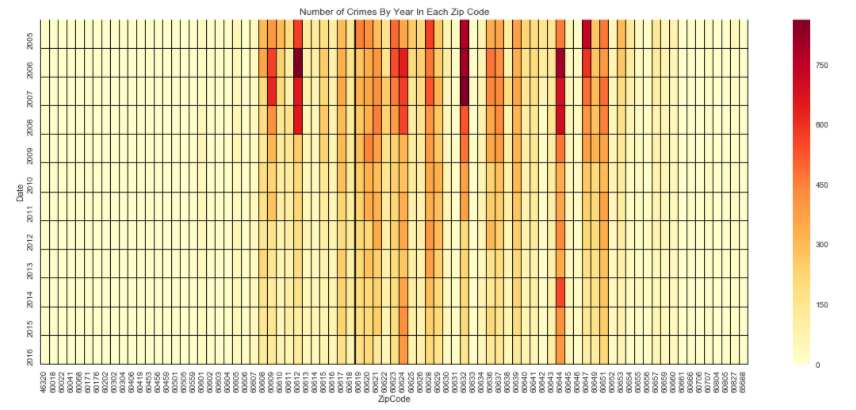
Chicago is the 3rd largest city in America with a population over 2,700,000 and also one of America's most historic. It laid the foundation for the first skyscraper in the world in 1885, the first ever television recorded presidential candidates' debate happened there between JFK and Nixon, and the word "Jazz" was invented there. However, in recent times Chicago become known primarily for its crime. While crime in Chicago has drastically decreased in modern times, and has also been widely studied already, it's still very interesting to look at.

One of the things that boggles my mind is how data can be used to make real predictions, and in this post I'll be focusing on that. For a person, your life is incredibly unique: you pick the movies that you like, you drive the cars you like, you eat the food you like. It's all personalized and unique to you, and what we like to do seems like it's been cultivated over the years we've been alive. For a thief, how could a bunch of numbers predict whether he will commit crime? His motivations are entirely his own, but somehow this machine picked him out of a line up accurately. It's not to say it's a perfect world; the data does go wrong. But then again, so do people.

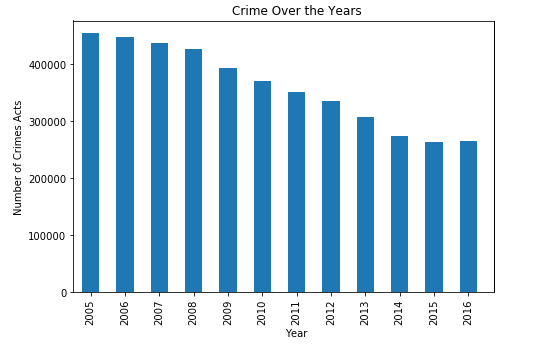
In this analysis, I'll be looking into crime as it happened in Chicago between the years 2005 to 2016, specifically the sex crimes. I primarily picked this because, in terms of volume of crime, it doesn't compare to many of the other facets of crime that lace the city. My purview ([seen here](https://github.com/ajitkoduri/Code-for-Blog-Posts/blob/master/Chicago%2BSex%2BCrimes%2BAnalysis.py)) was primarily to see if I could make predictions about whether a criminal would be found or not. As for the reasons I chose to do a crime analysis, it was because this project had 1) a large data set to play with, 2) a simple goal, and 3) a host of literature that I could read to base my interpretations off of.

My data was based off the Chicago Police Department's CLEAR system, where they transcribe a crime, location data, time of report, and whether an arrest was made for the report. The data contained over 5 million instances of crime, but when it was separated into sex crimes (classified as a crime that was reported as a offense against a child, human trafficking, prostitution, sex assault, sexual offenses, or stalking), it was only around 105,000 data pieces. I used a geotagging service to connect location to zip code, and connected population data about the zip code to the crime committed to the actual crime using U.S. Census data in 2010 on Chicago.

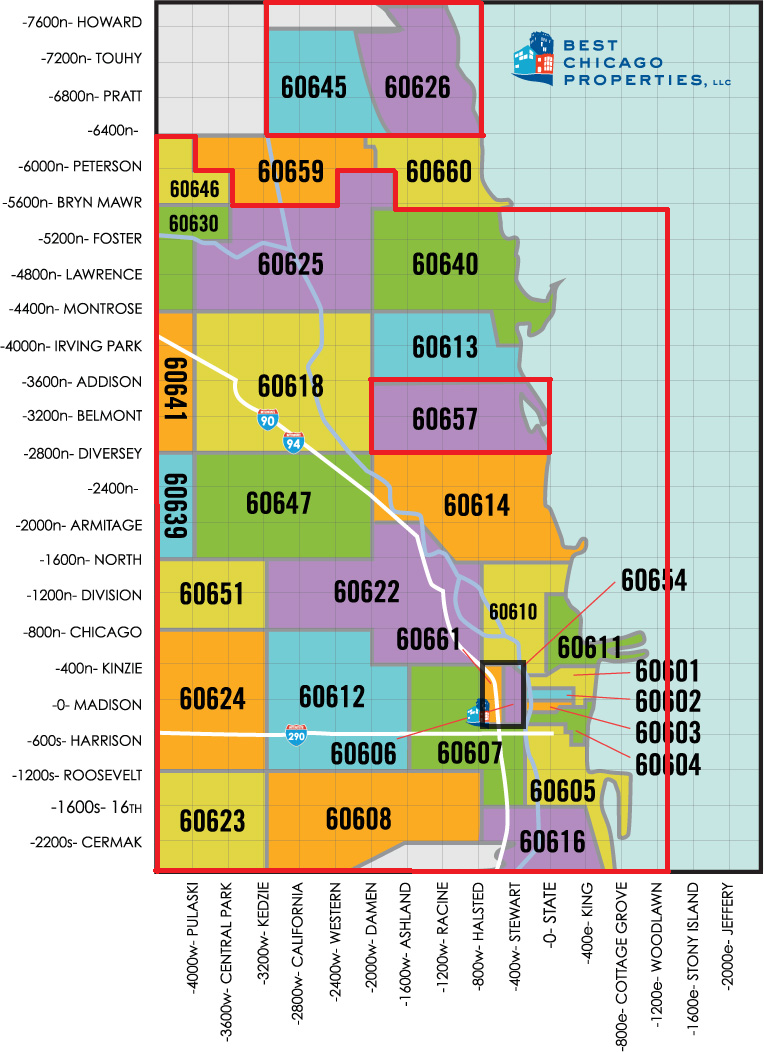
**Crime by Zip Code**

[](https://ajitdoes.files.wordpress.com/2017/12/a.png)

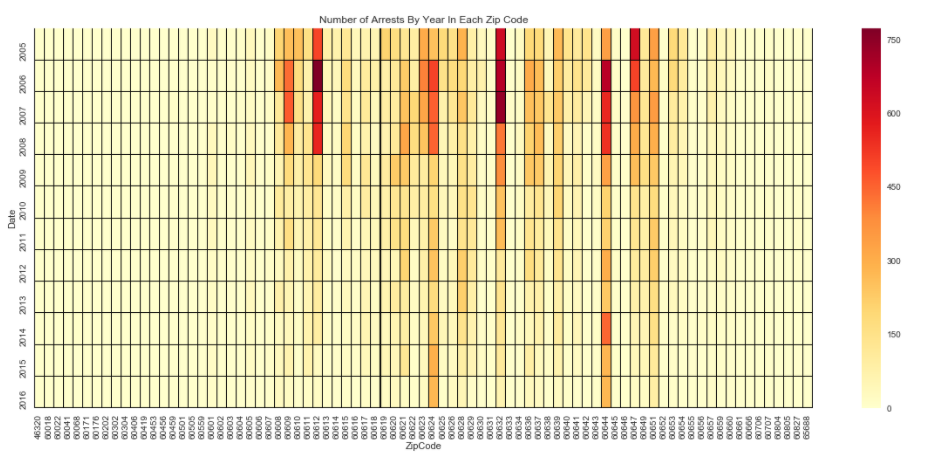
Looking at the above heat map of the amount of sex-related crimes over the 2005 to 2016 period, we can see a few general trends. Firstly, crime has tended to go down in recent times - the periods of 2005 to 2007 had the highest incidences, but recent times has seen an overall decrease in crime. It's important to remember that crime in general, as well as in Chicago, has always fallen over the years, as can be seen in the below graph ([generated here](https://github.com/ajitkoduri/Chicago-Crime-Analysis/blob/master/Chicago%20Crime%20Data.ipynb)).

[](https://ajitdoes.files.wordpress.com/2017/12/b.png)Looking more specifically at the heat map, it also becomes very clear that certain districts have a significantly higher level of sex-related crimes than others, those particularly between the zip codes 60608 and 60651.

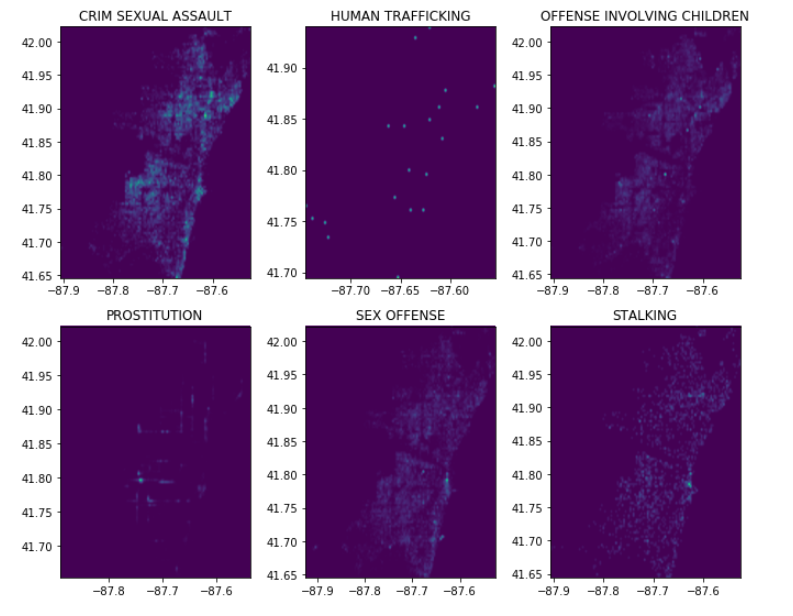
Overlaying the graph of the zip codes onto the actual map of Chicago (the areas surrounded in red are the zip codes of interest, while

[](https://ajitdoes.files.wordpress.com/2017/12/c1.png)the areas surrounded by black are not), we can see that most of the crime takes place in the city limits of Chicago as opposed to the suburbs. This intuitively makes sense, as most of the inner cities of any major city has the highest density of crime.

What's also there is how overpopulated these areas are compared to the rest of the greater Chicago area. Additionally, these areas tend to be more female dominated, suffer from large unemployment rates, have more numbers of previous sex offenders, be less educated than other parts of Chicago, and also tend to have a higher proportion of African Americans.

[](https://ajitdoes.files.wordpress.com/2017/12/d.png)In a general crime sense, not a lot of this is new. The highest proportion of crime is generally by the more poorly educated, the unemployed, and under-served minorities as their populations tend to have these characteristics more than Asian Americans and Whites. A heat map of the arrests by zip code shows that no one area has a significantly different arrest rate than its sex crime rate, showing that, in this specific category, **there is not a particular leaning to one area or another for a crime to be pursued.**

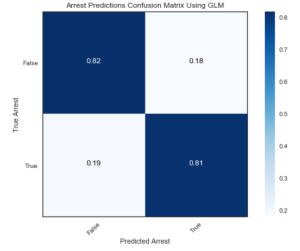
**Crime by Type Committed**

Sex crimes in general do not get a good follow up in Chicago. The average arrest rate for a specific type of crime in Chicago is around 30%, meaning that every third crime committed leads to an arrest. In terms of sex crimes, that list is much lower, closer to about 10%, except prostitution which is around 99% ([as shown at the bottom](https://github.com/ajitkoduri/Chicago-Crime-Analysis/blob/master/Chicago%20Crime%20Data.ipynb)) and children's sex offense which is around the normal rate. Why is there such a drastic difference between every other type of sex crime and prostitution? In terms of volume, prostitution isn't even the most common sex crime reported.[](https://ajitdoes.files.wordpress.com/2017/12/e.png)

Above is several heat maps of sex crimes as they were committed across Chicago. Stalking and sex offenses take place primarily closer to Lake Michigan, they still sporadically occur through the rest of the city. Sexual assault and child offenses typically occur throughout with no real central point, but they do come at a higher rate than prostitution does. Significantly, human trafficking is rarely reported - people that are spirited away tend to not talk about it. Even with that, only one-in-ten people that commit this type of crime are ever caught. What you'll notice as you look at the prostitution map, is that the data doesn't resemble any of the other maps - it's not being so severely under-reported as human trafficking, but also not occurring to the same extent that the rest of the crimes are.

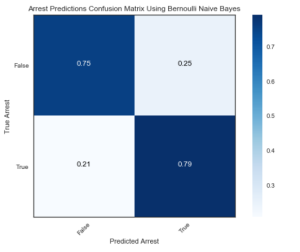
But this is strange; prostitution is the easiest to commit and the victims are more willing that someone who suffered from a sex offense or assault. When looking at the prostitution data, you can clearly see that the data points are lying literally on lines - *the reported crimes are on streets.* What you can probably find, if you delve a little deeper than I did, is what streets they occur on. Also, there's a small bright spot that is Vittum Park, but no spots in other parks in Chicago. **The most likely reason for this data to look this way is that the reported crimes are from the officers themselves from streets or areas they know where to arrest people who are in the act of prostitution.**

**Making Predictions**

[](https://ajitdoes.files.wordpress.com/2017/12/f.png)So with the data at hand, how can we turn it into something useful? The thing that stood out to me the most was the arrest rates, so I figured why not look for an algorithm that could help determine whether an arrest will be made. This is one of my first machine learning projects, so I tried to make it a simple classification problem - will there be an arrest made or not. I made three models to create these predictions, splitting a portion of the data into teaching my machine to make the predictions and another portion to test it on. I wasn't fine-tuning any particular aspects of the algorithm, so I didn't create a validation section (I'll be working on a post that's entirely about getting the most accurate model possible, so there will be one there). For this, I split it into 80% of the data going towards training it and 20% of the data going to testing.

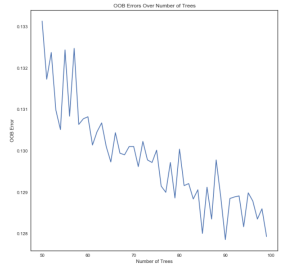
The first model was a simple Generalized Linear Model (GLM). It's a simple approach; first, see if each of your predictors linearly correlate to the chance of an arrest being made. Then, convert this chance of an arrest to arrest predictions by a simple formula: if the chance of an arrest occurring was greater than or equal to 50%, it would count as an 'arrest' prediction, else it was a 'no arrest' prediction.

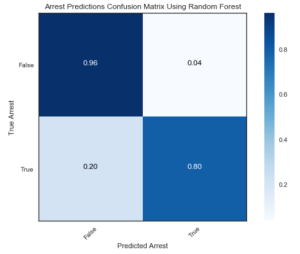
The accuracy of this model was 81.52%. You can see that there's not really a preference for getting the prediction correct or incorrect, so the model is fairly robust. While the predictions are not necessarily perfect, it's quite good given that the model only has details about the location of the crime, the population data of the zip code the crime was committed in, and what the type of crime was; it has none of the knowledge that a police report would contain and certainly no idea how much evidence the police have. I was initially worried that the data might have been skewed heavily towards a 'no arrest' prediction, given that for a majority of the crimes, there was no arrest. Luckily, you can see that that is not the end result as the algorithm does predict arrests made correctly as well (otherwise the bottom row would consist of 100% false negatives and 0% true positives).

The next model I did was a Naïve Bayes Model (NBM). This model is based on Bayesian Statistics, which bases its predictions on previous instances. This is different from the normal Statistics we use, called Inferential Statistics, because normal statistics bases its predictions off assuming that the population is normally distributed. Bayesian Statistics lacks a presumption of a general function that presides over it and as such is the basis for quite a lot of Machine Learning algorithms today. I chose a variation of the typical NBM by using a Bernoulli Bayesian Model, which can create predictions from the data using the discrete predictors inside. Normally, in a Machine Learning protocol, you make your classification predictions based off splitting the data into partitions that each of your classifications own. In the case of a binary classification problem, that results in a cutting the data in half, like what I did with the GLM. However, that would make predictions supplemented by the type of sex crime occurring incorrect since that is a varied group where the specific type affects the arrest rate. That's why I chose the Bernoulli [](https://ajitdoes.files.wordpress.com/2017/12/g.png)Model, because it could overcome that dilemma.

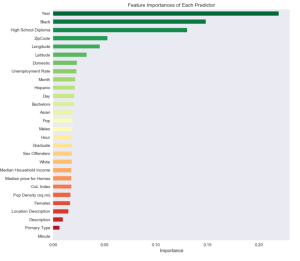
The model was actually only 77.4% accurate though, and it's clearly weaker than the GLM in every respect. Again, it is still robust, and the model did not fall into the trap of making every prediction a 'no arrest'. But from what we've seen, the linear model still surpasses this model and that may be because getting an arrest for these types of crimes is truly random or based on a factor that my data doesn't include about the crime itself. Either way, re-analyzing the data post hoc doesn't make for very good analysis and I certainly cannot get any details that the CPD keeps as confidential.

The last model I chose was a Random Forest. This model by using several Decision Trees and then aggregates their decision. The Decision Trees themselves operate by using binary logic, splitting that data through several leafs in order to partition the data into the categories it needs to predict from. What normally happens is that each Decision Tree makes a prediction and the Random Forest just spits out the the most popular decision. In that case, we can actually watch the evolution of the Random Forest as the number of Decision Trees it has increases in the below graph:

[](https://ajitdoes.files.wordpress.com/2017/12/h.png)

[](https://ajitdoes.files.wordpress.com/2017/12/i.png)What we can clearly see is the rapid error decrease as the number of trees increases (with some inconsistencies in between) in the training data set. In my model, I chose a Random Forest with 90 Decision Trees, as it had the lowest error rate.

As seen in the model, it quite good at predicting the 'no arrest' feature compared to previous models. The predictions for the 'arrest' category are worse, however they are around the same degree as the other models. This indicates to me that there is surely a variable that is missing that can account for these 'arrests' that I didn't include in the data, probably something that is privy to the individual case. Regardless, the accuracy of this model is 87.5%, the best of all the models.

[](https://ajitdoes.files.wordpress.com/2017/12/j.png)Also, the Random Forest lets us visualize how much weight each variable has towards the overall prediction. The first thing on the list is the 'Year', and taking into account what the diminishing criminal acts that occur over the years, it probably means that these types of crimes usually take a while to solve. The next most important is what percentage of African Americans makes up the area the crime was committed in. Does this show a bias for the CPD to arrest African Americans? I don't know without more individualized data of the crime, but it could also be that African Americans in these areas commit more or are victims of these sex crimes more than the other races.

As the third most important feature, the average education of the population has to meet a minimum standard of High School Diploma. This is bizarre to me since it doesn't increase or decrease significantly with increasing education. Perhaps those that are caught tend to be under-educated, or perhaps those that are under-educated commit these crimes more often (again, it is tough to say which with just the data I have). Following that is the location of the crime, but that, from the previous graphs, is because specific zip codes simply have more sexually related crimes committed than others. If the crime was domestic is next most important, but the arrests for that are probably due to having information that you wouldn't have if the crime was committed outside.

After that is unemployment rate of the zip code, which fits into common knowledge of what you expect crime to be. Then is the month, which there is already sufficient literature about how summer times tend to have the most crime committed. The Hispanic population comes afterwards, but also has the same question of whether the CPD discriminates or if they simply are connected to these cases more often than their counterparts, not including African Americans. The factors that follow from there are rarely helpful, so I will not continue discussing them.

In any case, I hope you guys thought this was interesting. I personally enjoyed making it, and there are plenty of factors I have not really explored that you can look into further if you want. What's really cool to me is how much of the current literature in criminal analysis we could see as we went through this data. I realize it's a little cheating, since hindsight is 20-20, but it's absolutely amazing to me how much knowledge of the past we could synthesize just by analyzing a large volume of data. Future things to look into might be the various other types of crime that are committed in Chicago, and see if there's any trends to look into or explore further. I'm personally satisfied for this project, but it could certainly be pushed to further heights for those that are interested in criminology