1. **Data Science Problem**

Crime is still a big issue whether it is in big cities or counties. Police officers trying hard to decrease illegal activities and it is important for them to manage crime by predicting crime. Crime has patterns just likes everything else people do when a large group of people to do it. In this project, we plan to investigate crime patterns of Chicago and Montgomery County of Maryland. Using historical crime data to predict crime patterns, we could predict where and when the next crime will likely take place. This prediction will help police officers to implement prevention work, therefore, significantly reduce the incidence of crime and reduce the harm caused by crime.

For example, the LAPD used the vast data set to showcase which areas in Los Angeles are hotspots of crime and a mathematical model to predict where crime would take place. With success, as there has been a 33% reduction in burglaries, 21% reduction in violent crimes and 12% reduction in property crime in the areas Big Data mining techniques such as statistics, modeling, and machine learning are being used.

1. **Potential Analyzes that can be conducted using Collected Data**

The Crime data of both Chicago and Maryland is already well organized. Two enormous crime data set have a lots of attributes in common. Such as, locations where crime occurs, time occur and type of crime…etc. Although some are missing and miss document, still both of the data are very clean.

Predicting the crime for the upcoming crisis can be a good help for the police officer to accomplish their job, to make the community much more safer. For example we can draw out two attributes out of the categories, the start time of the crime and the district locations (longitude and latitude). First, we can make clusters of the locations. Then analyze when crime happened more frequently in that district, in order to let the police department to increase the patrol frequency around the area. The types of crime can even added in to the clusters, enable to assign expert in different of crime types to prevent from happening in advance. For example, if the shoplifting happen a lot in Rockville’s supermarket. The MCPD can assign more police car the patrol around all the supermarkets.

1. Data Issue

* Chicago Crime Data has 20 attributes, Maryland has 17 attributes.
* Fetch the data from 2013 to 2016.
* Some attributes are unnecessary, such as ID are FBICode. They are not useful for predicting crime patterns.
* Some attributes have missing values, such as ‘ARREST’.
* Some attributes have null values, such as some zipcode are ‘”null”.
* Almost no noise for both data set.

1. Attributes of Montgomery County

data[2]:{

"agency":"MCPD",

"beat":"2D3",

"case\_number":"16009862",

"city":"BETHESDA",

"date":"2016-02-28T00:22:01.000",

"district":"BETHESDA",

"incident\_id":"201067968",

"incident\_type":"2942",

"location":"CENTER DR",

"narrative":"MENTAL TRANSPORT",

"place":"Residence - Other",

"police\_district\_number":"2D",

"pra":"074",

"sector":"D",

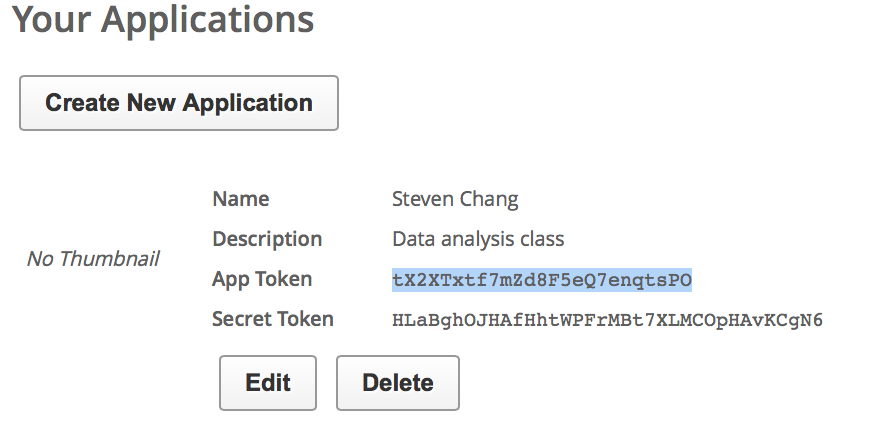
"start\_date":"2016-02-28T00:22:00.000",

"state":"MD",

"zip\_code":"20814"

}

1. Attributes of Chicago
2. **Collecting New Data**

We have collected over 150000 columns of crime sets (Montgomery County) approximately 115MB (Montgomery County) 

1. **Data Cleaning**

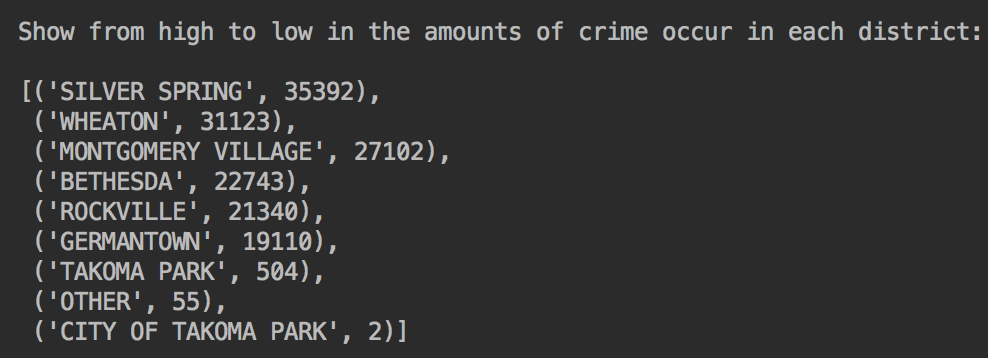
The data set of ours we can use each column to calculate the proportion of clearness of each attribute. Nearly all of the attributes are almost 100% clean with no noise or null. But in the data of Montgomery County the crime’s end date/time shows that almost 50 percent of the data are null. Although in this attribute of data is nearly half of the data did not document, still can analyze the execute efficiency between variety of crime.

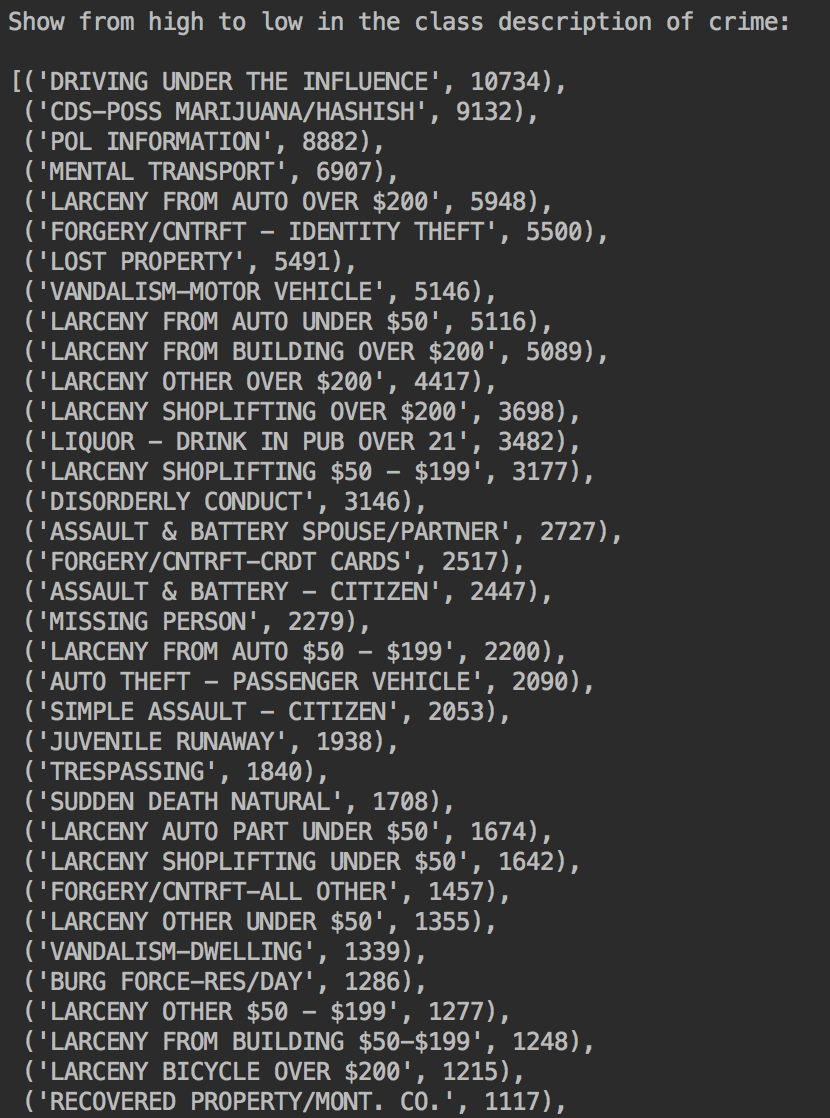
On the other hand we delete some of the unnecessary attributes that included some noise and null data.

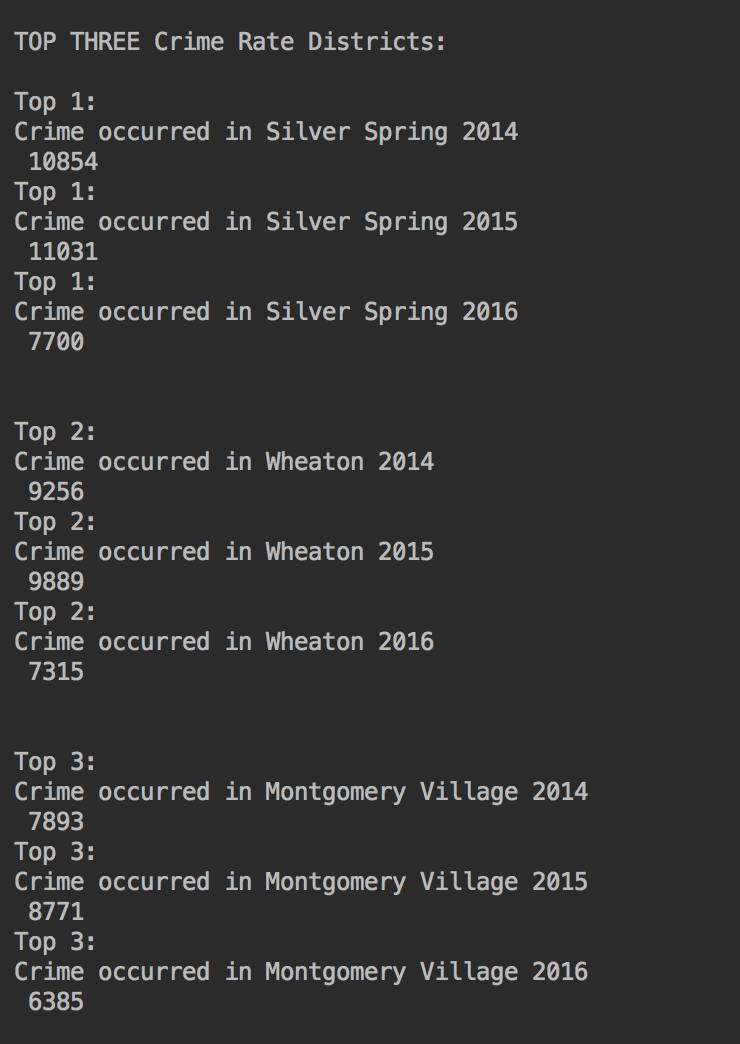
1. Feature Generation:

For Montgomery County we analyze:

* Which area owns the highest rate of crime occurrence.

Which places had the higher risk than other places.

* What kinds of crime take place more often.
* From analyzing which district had the most crime took place. Next add in the attribute of crime start date, separate into 2014, 2015 and 2016. Then can finalize in the Top 3 crime occurrence districts that crime rate slightly gain from 2014 to 2015 and drop significantly from 2015 to 2016.



* Analyze the data of Child Abuse in Single Family from 2013 to 2016
* The Number decrease significantly from 2014 to 2015

