**Project Assignment 1**

1. **Data Science Problem**

Crime is still a big issue whether it is in big cities or counties. Police officers work 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**

We plan to collect the data which could reflect incidents of crime that occurred in the City of Chicago and Montgomery County of Maryland in last a few years. Specifically, the dataset will include the location where the incident occurred, the date when the incident occurred, type of crime and description of the incident, etc. When we have a large group of these data, we use them to build a model and then we could get the pattern of the crime.

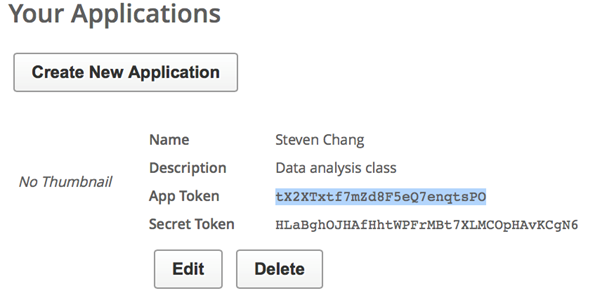
We can draw out two attributes out of the categories, the start time of the crime and the district locations (longitude and latitude). First, we can make clusters of the locations. Then analyze when crime happened more frequently in that district, in order to let the police department to increase the 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**

* 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”.

1. **Collecting New Data**

There are over 150000 records of crime data of Montgomery County. The dataset is approximately 115MB.



We also collect 150000 records of crime data of Chicago. Url is <https://data.cityofchicago.org/resource/6zsd-86xi.json?$limit=150000>.

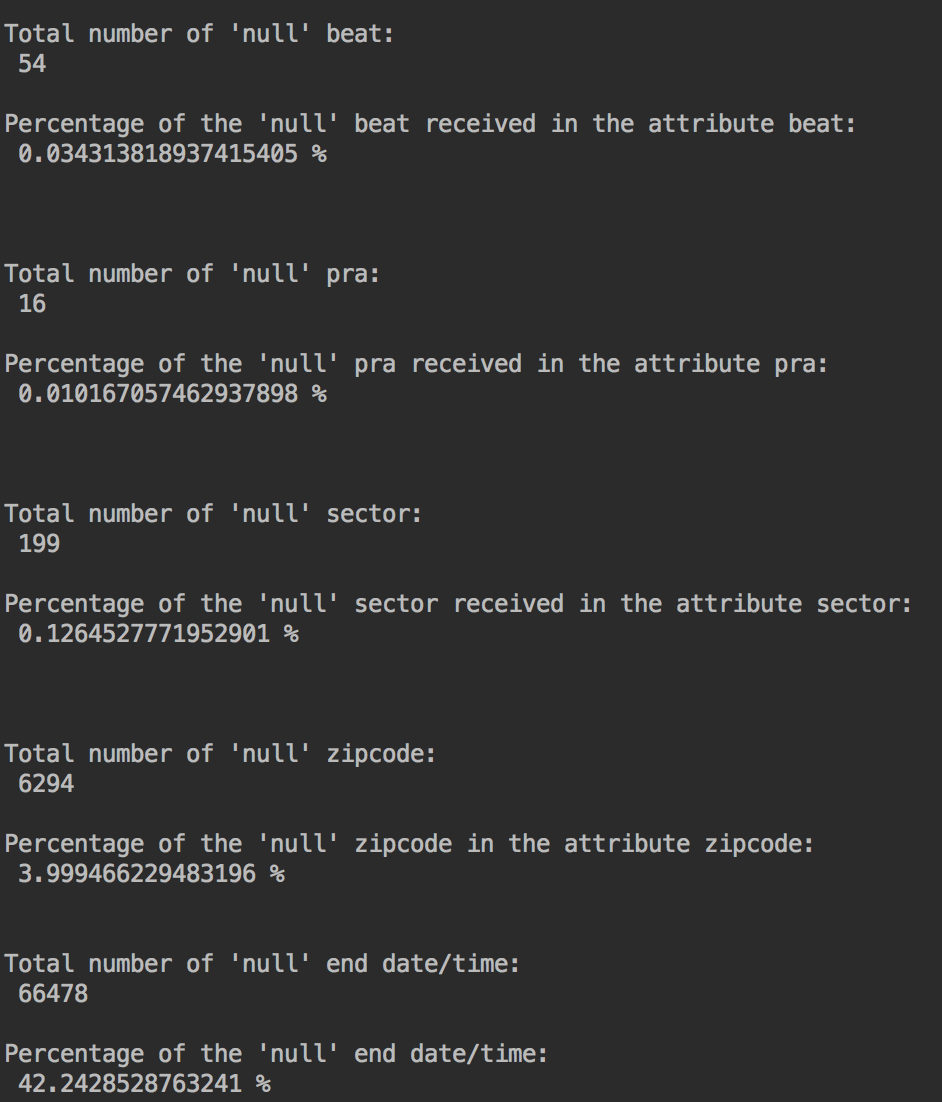
1. **Data Cleaning**

The Crime data of both Chicago and Montgomery County of Maryland is already well organized. Although some are missing and miss document, still both of the data are very clean.

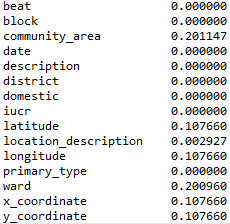
We use the fraction of missing values for each attribute and fraction of noise values to quantify how ‘clean’ the attribute is. In the dataset of Montgomery County of Maryland, most attributes are almost 100% clean with no noise or null. But in the data of Montgomery County the crime’s end date/time 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. In the dataset of Chicago, we remove some unnecessary attributes which are not useful for investigating the problem.

Fraction of missing values of Montgomery crime data:

5 out of 17 of the attributes have noise, all the other are all very clean (0% null value)



Fraction of missing values of Chicago crime data:

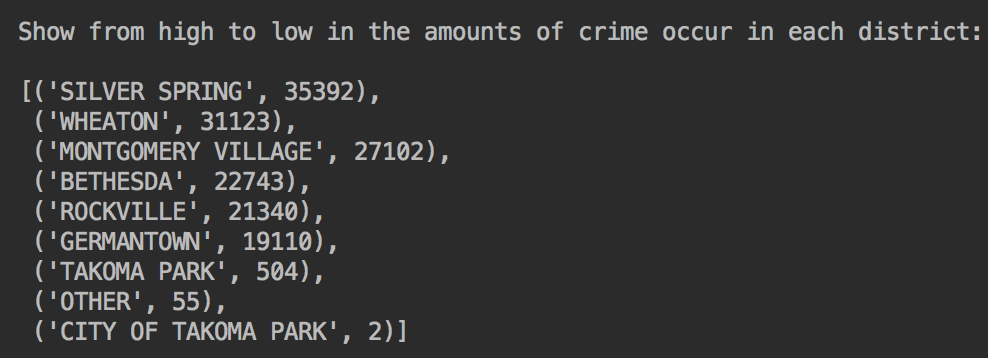


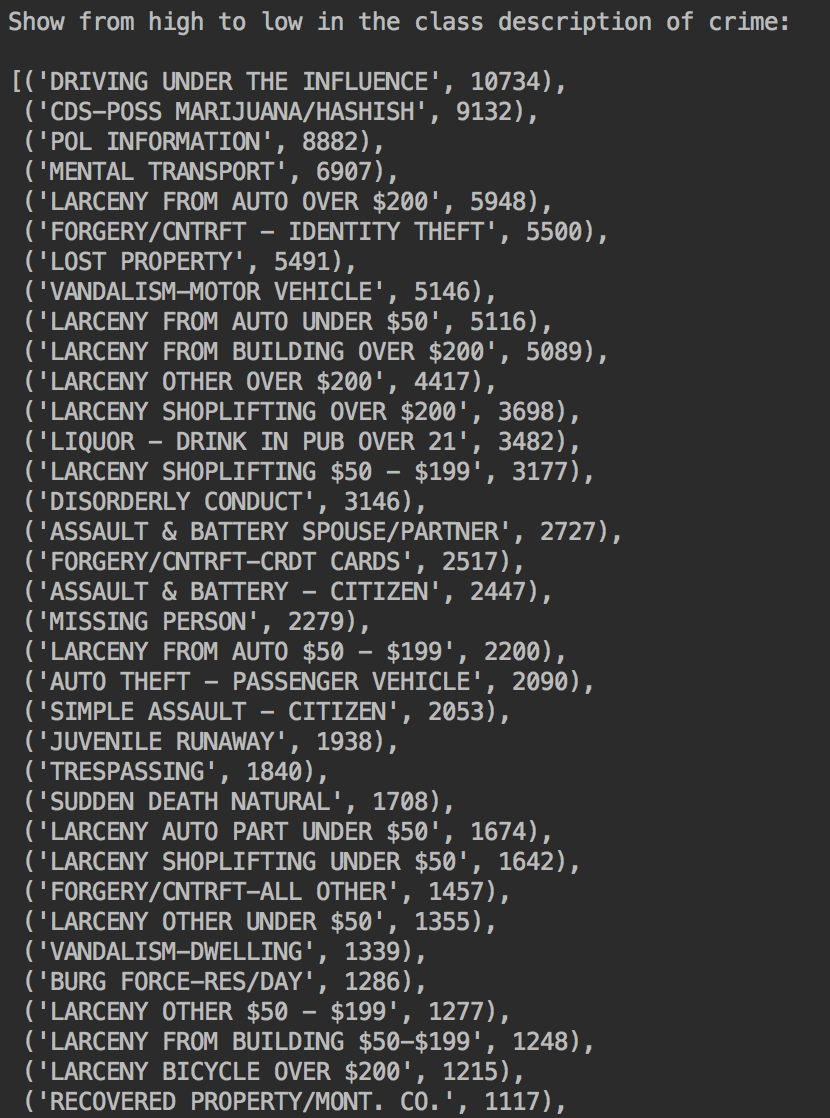
1. **Feature Generation**

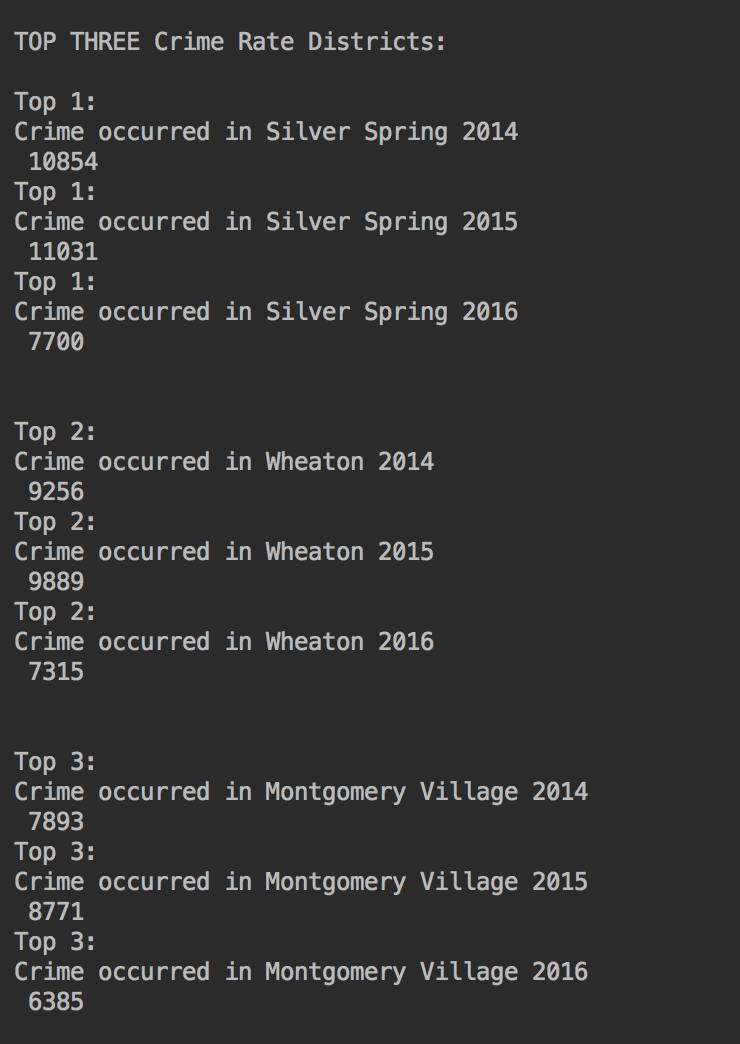
In dataset of Chicago, we use latitude and longitude to generate location. We also use the attribute of date to generate day of week. Beside, we create the new attribute (crimetype, location).

In dataset of Montgomery:

* 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

