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.

2. 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.

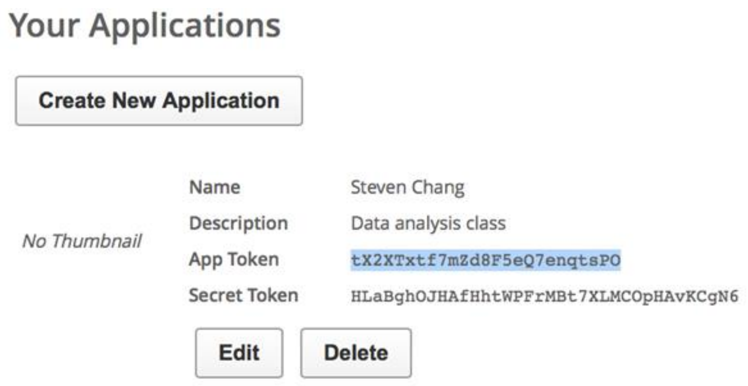
3. Data Issue

- 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".

4. Collecting New Data

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



Name Steven Chang
Description Data analysis class
App Token tX2XTxtf7mZd8F5eQ7enqtsPO
Secret Token HLaBghOJHAfHhtWPFrMBt7XLMCOpHAvKCgN6

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

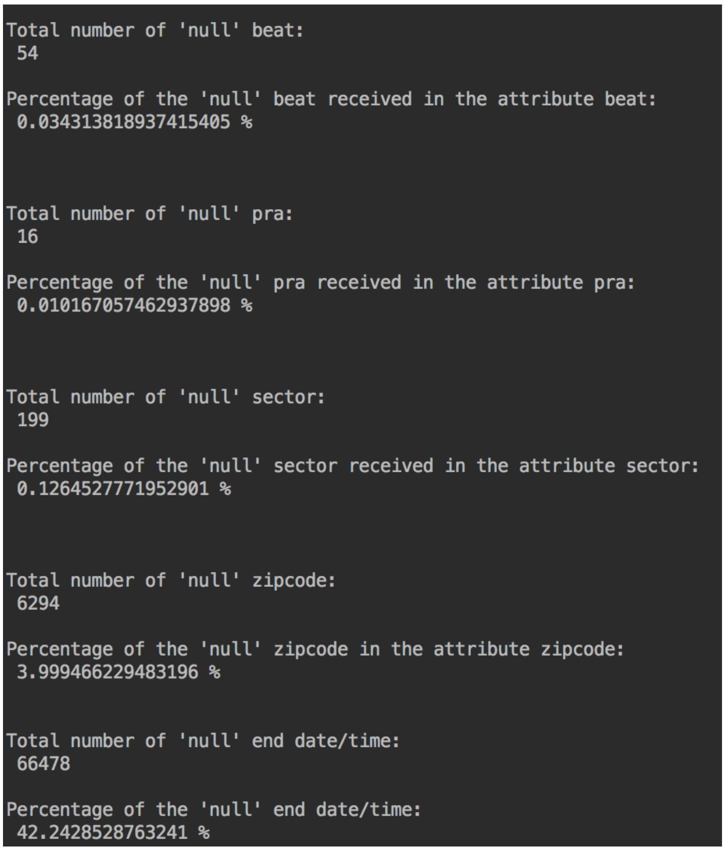
5. Data Cleaning

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.

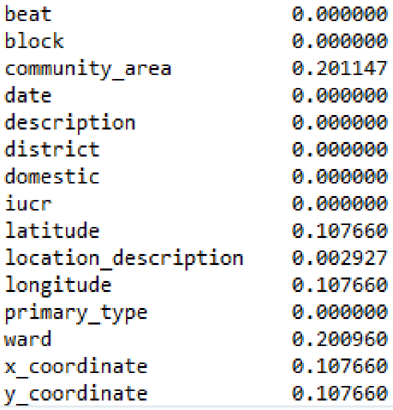
Fraction of missing values of Montgomery crime data:

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



```
Total number of 'null' beat:
 54

Percentage of the 'null' beat received in the attribute beat:
 0.034313818937415405 %


Total number of 'null' pra:
 16

Percentage of the 'null' pra received in the attribute pra:
 0.010167057462937898 %


Total number of 'null' sector:
 199

Percentage of the 'null' sector received in the attribute sector:
 0.1264527771952901 %


Total number of 'null' zipcode:
 6294

Percentage of the 'null' zipcode in the attribute zipcode:
 3.999466229483196 %


Total number of 'null' end date/time:
 66478

Percentage of the 'null' end date/time:
 42.2428528763241 %
```

Fraction of missing values of Chicago crime data:



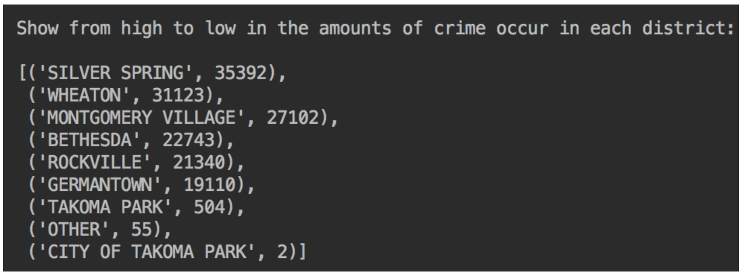
```
beat                    0.000000
block                   0.000000
community_area          0.201147
date                    0.000000
description             0.000000
district                0.000000
domestic                0.000000
iucr                    0.000000
latitude                0.107660
location_description    0.002927
longitude               0.107660
primary_type            0.000000
ward                    0.200960
x_coordinate            0.107660
y_coordinate            0.107660
```

6. 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 (crime type, location).

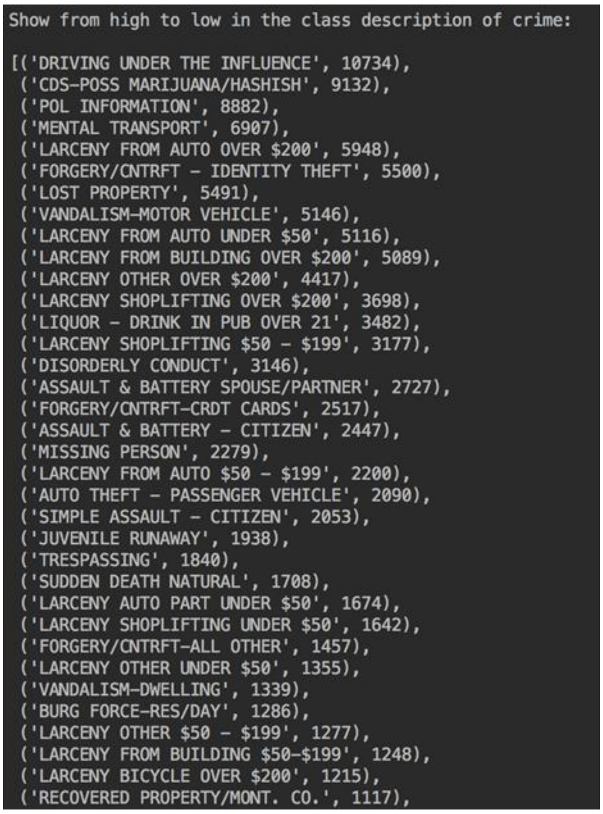
In dataset of Montgomery, we use start_date and district to generate (start_date, district). We also use place, narrative and district to generate (place, narrative, district). These new features help us to do the following analysis.

The area which has the highest rate of crime occurrence:



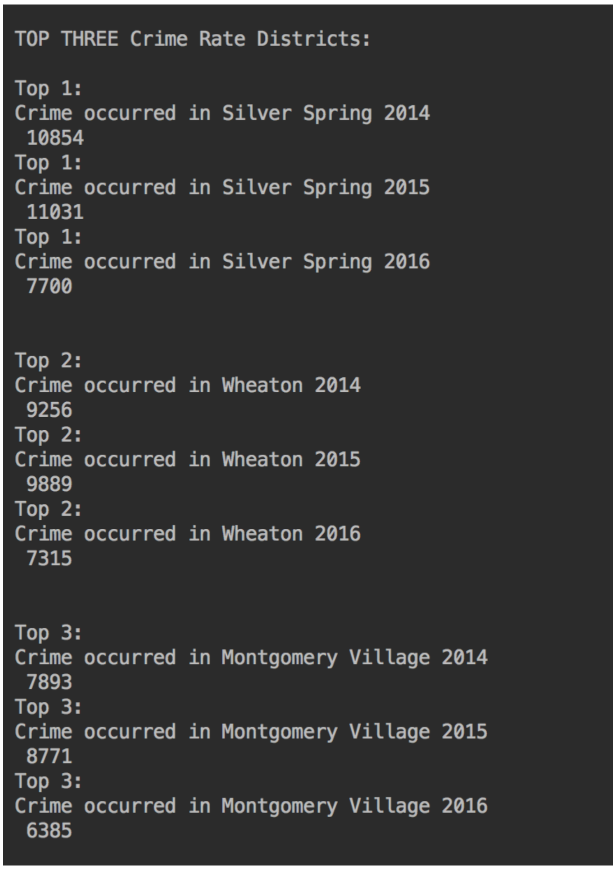
```
Show from high to low in the amounts of crime occur in each district:

[('SILVER SPRING', 35392),
 ('WHEATON', 31123),
 ('MONTGOMERY VILLAGE', 27102),
 ('BETHESDA', 22743),
 ('ROCKVILLE', 21340),
 ('GERMANTOWN', 19110),
 ('TAKOMA PARK', 504),
 ('OTHER', 55),
 ('CITY OF TAKOMA PARK', 2)]
```

The type of crime that takes place more times:



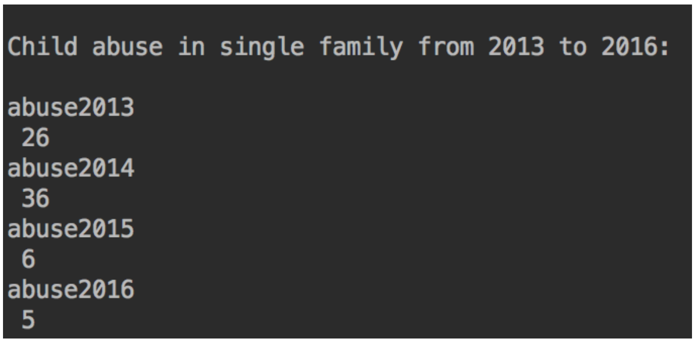
```
Show from high to low in the class description of crime:

[('DRIVING UNDER THE INFLUENCE', 10734),
 ('CDS-POSS MARIJUANA/HASHISH', 9132),
 ('POL INFORMATION', 8882),
 ('MENTAL TRANSPORT', 6907),
 ('LARCENY FROM AUTO OVER $200', 5948),
 ('FORGERY/CNTRFT - IDENTITY THEFT', 5500),
 ('LOST PROPERTY', 5491),
 ('VANDALISM-MOTOR VEHICLE', 5146),
 ('LARCENY FROM AUTO UNDER $50', 5116),
 ('LARCENY FROM BUILDING OVER $200', 5089),
 ('LARCENY OTHER OVER $200', 4417),
 ('LARCENY SHOPLIFTING OVER $200', 3698),
 ('LIQUOR - DRINK IN PUB OVER 21', 3482),
 ('LARCENY SHOPLIFTING $50 - $199', 3177),
 ('DISORDERLY CONDUCT', 3146),
 ('ASSAULT & BATTERY SPOUSE/PARTNER', 2727),
 ('FORGERY/CNTRFT-CRDT CARDS', 2517),
 ('ASSAULT & BATTERY - CITIZEN', 2447),
 ('MISSING PERSON', 2279),
 ('LARCENY FROM AUTO $50 - $199', 2200),
 ('AUTO THEFT - PASSENGER VEHICLE', 2090),
 ('SIMPLE ASSAULT - CITIZEN', 2053),
 ('JUVENILE RUNAWAY', 1938),
 ('TRESPASSING', 1840),
 ('SUDDEN DEATH NATURAL', 1708),
 ('LARCENY AUTO PART UNDER $50', 1674),
 ('LARCENY SHOPLIFTING UNDER $50', 1642),
 ('FORGERY/CNTRFT-ALL OTHER', 1457),
 ('LARCENY OTHER UNDER $50', 1355),
 ('VANDALISM-DWELLING', 1339),
 ('BURG FORCE-RES/DAY', 1286),
 ('LARCENY OTHER $50 - $199', 1277),
 ('LARCENY FROM BUILDING $50-$199', 1248),
 ('LARCENY BICYCLE OVER $200', 1215),
 ('RECOVERED PROPERTY/MONT. CO.', 1117),
```

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.



```
TOP THREE Crime Rate Districts:

Top 1:
Crime occurred in Silver Spring 2014
 10854
Top 1:
Crime occurred in Silver Spring 2015
 11031
Top 1:
Crime occurred in Silver Spring 2016
 7700


Top 2:
Crime occurred in Wheaton 2014
 9256
Top 2:
Crime occurred in Wheaton 2015
 9889
Top 2:
Crime occurred in Wheaton 2016
 7315


Top 3:
Crime occurred in Montgomery Village 2014
 7893
Top 3:
Crime occurred in Montgomery Village 2015
 8771
Top 3:
Crime occurred in Montgomery Village 2016
 6385
```

Analyzing the data of Child Abuse in Single Family from 2013 to 2016, we find that the Number decrease significantly from 2014 to 2015.



```
Child abuse in single family from 2013 to 2016:

abuse2013
 26
abuse2014
 36
abuse2015
 6
abuse2016
 5
```