ANLY 501 Final Project

# What Promotes Crime?

Dec. 10th, 2018

Jiaqi Tang

Jin Young Yang

Shiqi Ning

Xinyi Ye

Georgetown University

# Abstract

This report identifies the data science problem of “What Promotes Crime?”. Potential factors, both direct and indirect, including moon phase, weather, days of a week, regional education level and adult obesity/smoking rate are taken into consideration to see how different factors affect certain types of crime more closely. Crime types are categorized into eight groups: Assault, Burglary, Death, Drug, Fraud, Robbery, Theft and Sexual Crime. 2017 crime data for 14 US major cities are closely analyzed, which are Austin, Baton Rouge, Boston, Chicago, Denver, Detroit, Hartford, Las Vegas, Los Angeles, Philadelphia, New Orleans, New York, San Francisco, Washington DC. Different methodologies have been applied to analyze the relationship between different factors and crime type, such as correlation, clustering, association rule, hypothesis testing, PCA, spatial-temporal model, machine learning and network analysis. The results reveal that Theft and Burglary crime have a strong correlation between each other. Distinctive temporal and spatial patterns for different days of the week. Days of the week, Education level, Obesity rate, Smoking rate, New moon, Rainy Weather are found to be more related crime types compared to other factors being considered.

# Contents

# What Is the Issue?

According to FBI’s Uniform Crime Report (UCR), for the past 15 to 20 years, both the violent and nonviolent crime rates have been in decline.

However, a global study on homicide held by United Nations Office on Drugs and Crime (UNODC) had stated that overall the total crime rate of the United State compared to other developed countries, especially Europe, is higher, except for South American countries and Russia.

In the US, different types of crime happen every day and everywhere, the mass shooting in Las Vegas in 2017, countless school shootings, arson crime, and violent crime, etc. precious life have been taken. Students receive at least once a week the alerts sent by their schools regarding with the sexual assaults, burglary happened to their fellow classmates around the campus.

In order to better study crime and protect people from being harmed, various studies have been conducted to analyze the complex topic about crime. But there is no simple way to fully explain the subject matter, and it is impossible to find direct causation of one factor to crime. Thus, taking in consideration of various different factors appear to become one possible approach to better understand and analyze crime.

# Introduction

Crime has always been an intriguing topic for the majority of people, for it has a close relationship with the safety of each individual. Thus, in order to help society to develop a better understanding of the nation’s living standard and its law enforcement agencies, studying crime has become an important part in the criminal justice system. In addition, by closely examining the social and psychological factors that cause people to commit crimes, it also increase the possibility for individuals to better protect themselves against crime.

Various studies relate to the influences of ethnicity and demographics like age and sex on crimes have been analyzed, and the statistics have revealed that though different crime has various trends, victims of specific crimes could be determined by gender and ethnicity. For instance, males and people of black descent are more likely to be victims of aggravated assaults. On the other hand, females are more likely to be victims of kidnappings and grand theft property (Cung, 2013).

However, the reasons for crime occurrences can not be fully explained by only a single factor, behind the seemingly obvious statistics, there underlies countless co-factors such as education level, health conditions, historical conditions, discrimination, economic status, social and cultural perceptions towards sex, age, ethnicity, or environmental factors like weather and so on.

Weather, especially temperature, rainy and non-rainy weather, seem to have a direct effect on the physical environmental conditions for crime to occur. Additionally, there exists an interesting hypothesis which relates crime with moonphase. And series of studies on how different moonphase can affect human body and behavior have been conducted. But whether moonphase, like full moon actually promotes the different crime type is unknown and worthy analyzing. Exploring the relationships between different weather and moonphase conditions and different crime types, so that crime patterns can be foreseen.

Day of the week also seem to have an impact on different crime type, as different days, especially weekdays and weekends, have distinct characteristics that may lead to different crime patterns.

Education level of the criminals is another factor that has been extensively studied. But further questions like “do neighborhoods with higher average education level incline to have higher or lower crime occurrence?” and “does education level plays a role in different crime type?” are both valuable aspects that can be investigated.

The victims are also of interest as the question whether there is a tendency for certain type of people to be involved in different crime remains unknown. For example, do people with higher obesity rate and smoking rate more likely or less likely to be the victims of crime is also one possible perspective that can be analyzed.

Latest crime data in 2017 of the major cities, including Austin, Baton Rouge, Boston, Chicago, Denver, Detroit, Hartford, Las Vegas, Los Angeles, Philadelphia, New Orleans, New

York, San Francisco, Washington DC in the United States are collected, since major cities have more diverse and robust information that is useful for data analysis. Figuring out how different factors, traits or backgrounds mentioned above such as weather, moonphase, days of a week, education level and adult obesity/smoking rate that can potentially contribute to a propensity to promote crimes can be seen as a possible and worthwhile angle to better study crime and therefore prepare for it.

# Data Collection

## Crime Data

Crime data in 2017 for 14 different cities are collected from Data.gov, including Austin, Boston, Chicago, Los Angeles, New York, San Francisco, Washington DC, etc. The zip codes for each city are collected and the corresponding crime data is then scrapped. The crime data includes each crime record in 2017, which to be specific contains the date, location, crime type and other crime-related information.

## Weather and Moonphase Data

Moon phase datasets are retrieved via worldweatheronline.com using API, including each day's moon phase, moon illumination, sun hour and temperature, and weather in 2017. The moon phase attribute contains 8 categories: first quarter, full moon, last quarter, new moon, waning crescent, waning gibbous, waxing crescent and waxing gibbous. And the weather attribute contains 35 weather types including: blizzard, light snow, overcast, sunny, heavy rain and etc.

## Education Data

For education datasets, school data for elementary, secondary and postsecondary institutions are collected from the *Homeland Infrastructure Foundation-Level Data (HIFLD)*'s open data by specify the States of interest. Specifically, there are three datasets being downloaded: 1. private school information for elementary, secondary institutions; 2. public school information for elementary, secondary institutions; 3. Universities information. The data contains the zip code, the name of the institution, school population, types of the institution (regular, special education, vocational/technical, etc.), location of the institution, etc. for all the cities in the US.

## Health Data

The adult smoking rates, obesity rate of all counties of interest are collected from https://datausa.io/ by using API.

# Data Issues and Data Cleaning

## Crime Data

### 1. Data Issues

There are over 20 columns in crime datasets for each city, however, some of columns (features) are not necessary for analyzing crime, for example: weapon usage, weapon description. In addition, due to the lack of information from crime report, some columns contain plenty of missing values (over 60%). In this way, some of these columns are supposed to remove in order to reduce the dimensions of data.

In addition, for the types of crime, the description of crime types varying in different cities, based on the data from these cities, there 1139 unique values in the 'Type' columns, some value was describing crime so specifically. Nevertheless, in this project, the crime types tend to be the output feature in further predictive analysis, uniforming the inconsistent crime type data should be performed as data clean method.

### 2. Data Cleanliness and Clean Method

First, for missing data, all the records have been reported by police and victims, in this case, there is no reasonable way to replace or predict missing value. Thus, for crime data, first, the columns which contains missing value that are more than 30% have been removed, furthermore, after removing these columns, all the rows contains missing values have been removed too.

Second, for inconsistent data in crime dataset, all the crime types are renamed and categorized into the following 9 groups by the following grouping matrix:

1. Crime types that contain the following keywords will be grouped into "Assault":

- Assault
- Threat
- Firearm
- Kidnapping
- Trafficking
- Extortion
- Harassment
- Stalking
- Trespass
- Disturbance
- Child
- Violence
- Weapon
- Battery
- Offense
- Mississ

2. Crime types that contain the following keywords will be grouped into "Burglary":

- Burglary
- Computer
- Vandalism

3. Crime types that contain the following keywords will be grouped into "Death":

- Murder
- Suicide

- Death
- Homicide

4. Crime types that contain the following keywords will be grouped into “Drug”:

- Drug
- Marijuana
- Controlled
- Narcotics

5. Crime types that contain the following keywords will be grouped into “Fraud”:

- Fraud
- Identity
- Forgery
- Counterfeiting
- Deceptive
- Bribery
- Embezzlement
- Money
- Card

6. Crime types that contain the following keywords will be grouped into “Robbery”:

- Robbery
- Carjacking
- Property

7. Crime types that contain the following keywords will be grouped into “Sexual”:

- Rape
- Sex
- Prostitution
- Sexual

8. Crime types that contain the following keywords will be grouped into “Theft”:

- Theft
- Stolen
- Shoplifting
- Larceny

9. Crime types that do not contain the above keywords were grouped into “Other”. This group contains less common crime types and will not be used for future analysis.

## Weather and Moon Phase Data

The moon phase data has been retrieved through API in json format. After being checked missing data and noise data in python script, both numerical and categorical data in the weather and moon phase datasets are in appropriate ranges. Therefore, technically there are no data issues in those datasets. In other words, the datasets don’t have any missing values or noise.

# Education Data

## 1. Data Issues

For the three datasets that are collected separately for education data, there is no direct sign of empty values for each attribute, but there exist several other issues with the datasets:

- The inconsistency of the attributes: same column (Type, Status, Level) with different attribute directories
    - Even though the three datasets are downloaded from the same source, used similar column names such as “Type”, and categorized into groups of 1, 2, 3, 4, the groups actually represent inequivalent school types.
    - E.g. In public school dataset, 2 in column “Type” means Special Education, but in private school dataset, 2 in column “Type” means Montessori.
- Primary zip codes and secondary zip codes misses the starting 0:
    - There are primary zip codes that are not 5-digit, or secondary zip codes that are not 4-digit. They can be either wrong zip codes or just don’t have the leading 0.
    - E.g. Massachusetts zip codes can start with 0 and can be omitted by csv format file.
- Attribute “Population” has 0 and -999:
    - The “Population” column has negative values and 0, which could be deemed as missing values.

## 2. Data Cleanliness and Clean Method

For the education data, the unique values are first printed for categorical attributes, and no sign of abnormal data. Then, the numerical attributes mentioned in the previous section are checked, such as “zip codes” and “population”, etc. In addition, there are variable directory layouts available for each dataset, and several categorical attributes are already being grouped. By checking with the layout files, there are several attributes that are actually being labeled as missing values (such as 0 and -999). The detailed fraction of missing values for each attribute are shown as the following (Detailed directory layout can be found in the zip folder):

- For private elementary and secondary school data:
    - The fraction of missing values for attribute LEVEL is 0.0.
    - The fraction of missing values for attribute POPULATION is 0.0.
    - The fraction of abnormal values for attribute ZIP is 0.0159.

- For public elementary and secondary school data:
    - The fraction of missing values for attribute LEVEL is 0.0321.
    - The fraction of missing values for attribute POPULATION is 0.0377.
    - The fraction of abnormal values for attribute ZIP is 0.0164.
- For colleges and universities data:
    - The fraction of missing values for attribute TYPE is 0.009.
    - The fraction of missing values for attribute LEVEL is 0.009.
    - The fraction of missing values for attribute LOCALE is 0.0.
    - The fraction of missing values for attribute INST_SIZE is 0.0207.
    - The fraction of missing values for attribute SECTOR is 0.009.
    - The fraction of missing values for attribute POPULATION is 0.0869.
    - The fraction of ABNORMAL values for attribute ZIP is 0.0154.

The fraction of missing data is all below 0.1, thus on average, a data quality score of 90 out of 100 would be given. Because, the missing values may occur in the same row, the overall cleanliness of our data is relatively high.

To further clean and prepare the data, the categorical attributes for each dataset will be redefined. For instance, the “Type” attribute in the private and public school datasets is regrouped into regular school, alternative school, special education, and vocational/technical school, so that all the school information would be on the same scale. In addition, rows with missing values will be dropped. Then, the zip codes will be checked to further determine whether the zip codes are wrong or just lack the first 0.

## Health Data

Health data is collected from Datausa API with many interesting possible factors for crime rates. The issue of this dataset is its geographical information. The dataset has its unique geographical ID number for counties. In order to create a column designating each observation’s county name, another dataset containing the ID number and the name of each county from the same website is collected and merged with the previous dataset.

From the ID and name of county dataset, the unnecessary attributes such as “display_name”, “image_author”, “image_link”, “image_meta” and etc. are deleted. By deleting the trivial columns, a solid merged dataset of counties’ health status is formed.

When checking the cleanliness of the data, there exists neither missing values nor abnormal data for both obesity and smoking rates. Therefore, there are no data issues in the health dataset.

# Feature Generation

## New Attributes

- A new feature called “number of schools per zip code” is added to education data.
- A new feature called “Day” is added to crime data, which transform each date into days of a week and is represented as 0 ~ 6, with 0 = Monday, 6 = Sunday.
- A new feature called “Month” is added to crime data, which transform each date into its corresponding month and is represented as 1 ~ 12, with 1 = January, 12 = December.
- A new feature called “Weekend” is added to crime data, which categorizes each day into either weekdays or weekend, and is represented by 0 and 1, with 0 = Weekdays and 1 = Weekend.

## Data Merge and New Dataset

### 1. Crime Data vs. Weather and Moonphase Data

The crime data and the moon phase data are merged by date. Before merging the crime data with moonphase data, the formats of date column, which is the key column for both datasets for merging, is first unified.

### 2. Crime Data vs. Education Data

For education data, the zip codes for both data sets are used to merge crime data with school data.

3. Crime Data vs. Health Data

For health data, since the obesity rates and smoking rates in the health dataset is for each city, thus the city names for both data sets are used to merge health data with school data.

# Exploratory Analysis

## Basic Statistical Analysis and Data Cleaning Insight

### Data Cleaning

There are three different kinds of missing data issues in our dataset: inconsistent data, missing values, and outliers:

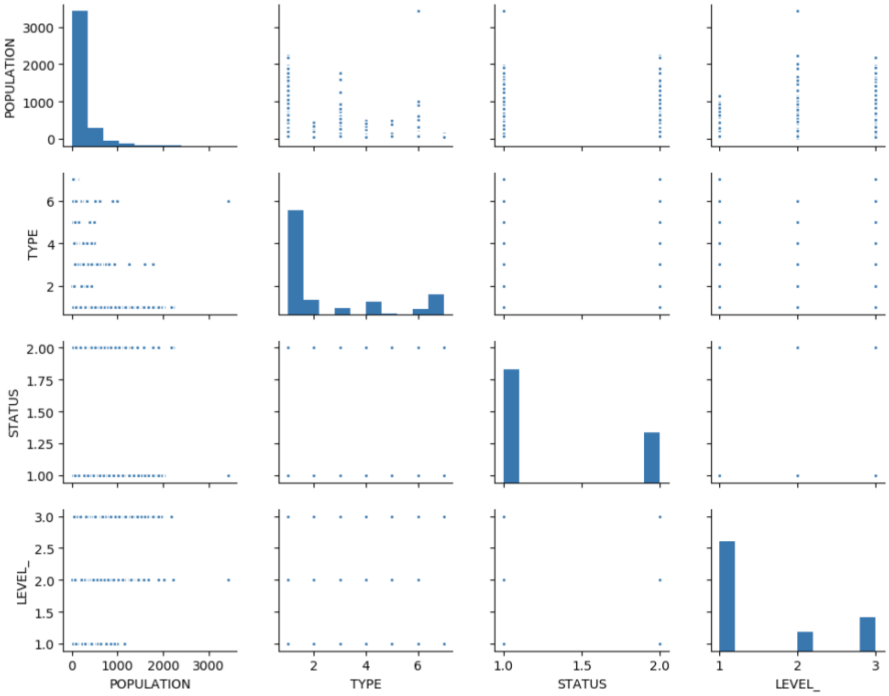
**Education Data:**

Since the education data that were collected contains all cities for each State, the three school datasets needed to be filtered by cities of interests. Thus, a list of city names and a list of county names were created to filter the raw datasets. The reason for creating two lists for the filtering process is that some cities may cover several counties or have city names represented by district names. So, filtering by two lists can greatly reduce the risk of losing cities of interest. However, this filtering method will contain cities other than the cities of interest. But it won't affect the analysis, because the education data will be merged with the crime data by matching the zip codes. Thus, those cities will be ruled out before the actual analysis part.

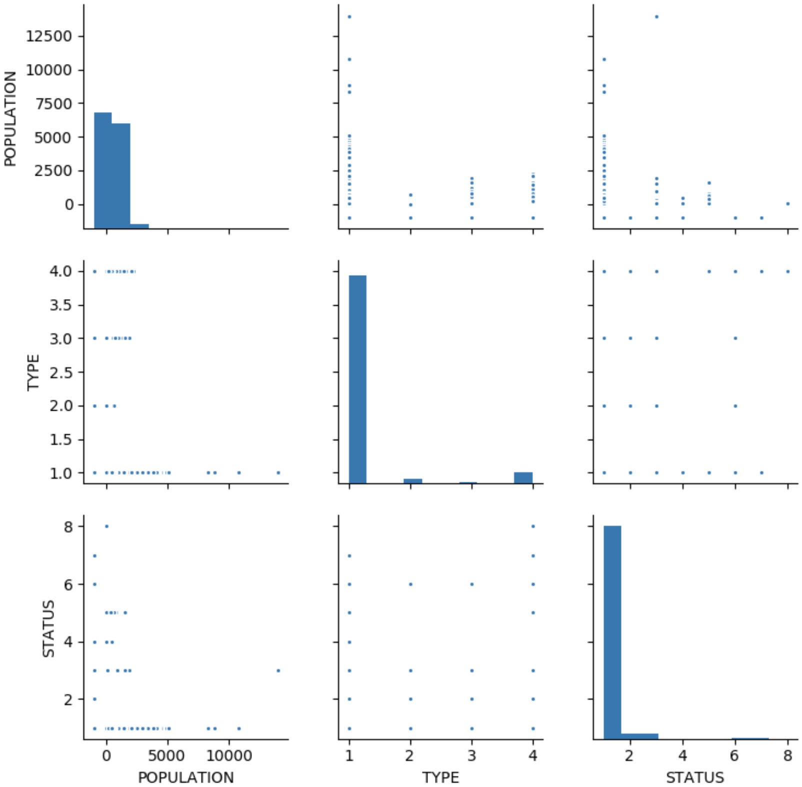
**Detect Outlier**

After the raw school datasets being filtered, a scatter plot for the attributes of interest were plotted, in order to see if there exists any outliers or abnormal data (see the following figures, upper: Private School Data, lower: Public School Data):

Graph 1. Outlier Detection for Education Data



*Private School Data Before Cleaning*



*Public School Data Before Cleaning*

From the private school dataset, at a glance, there is a potential outlier, which is POPULATION greater than 3000. However, after closely checking this point, this point is actually outside the range of cities of interest.

For the public school dataset, as mentioned above, there exists POPULATION smaller than and equal to 0, which according to the dataset's directory layout, are missing values and need to be dropped. And there were 3020 rows of them. When the POPULATION is negative or equal to 0, it means that the school is either temporarily inactive or closed, and in addition contains missing values in other columns. Including these rows won't add any valuable information to the analysis, and thus are dropped.

For attribute LEVEL, the missing values, which are represented by "N" also dropped. And most of the rows with LEVEL = N has POPULATION = -999, so they were dropped.

For attribute STATUS, according to the directory:

- 2 = School has closed since the time of the last report;
- 6 = School is temporarily closed and may reopen within three years;
- 8 = School was closed on a previous year's file but has reopened; were dropped

Since only active school information is desired, STATUS with the above representations was dropped. And there were 89 rows of them, which is not significant as well compare to the whole dataset.

For attribute ZIP in both private and public school datasets, there are some zip codes that have digits less than 5, which could be abnormal data. However, after closely examining the zip codes with only 4 digits, all of them are zip codes from Boston and Hartford, which should start with 0. Thus, they are not abnormal data.

In addition, the attributes LEVEL, TYPE in both datasets had different representations, and were reconciled as the following:

LEVEL:

- 1 = Elementary Public school;
- 2 = Secondary Public school;
- 3 = Combined Public school;

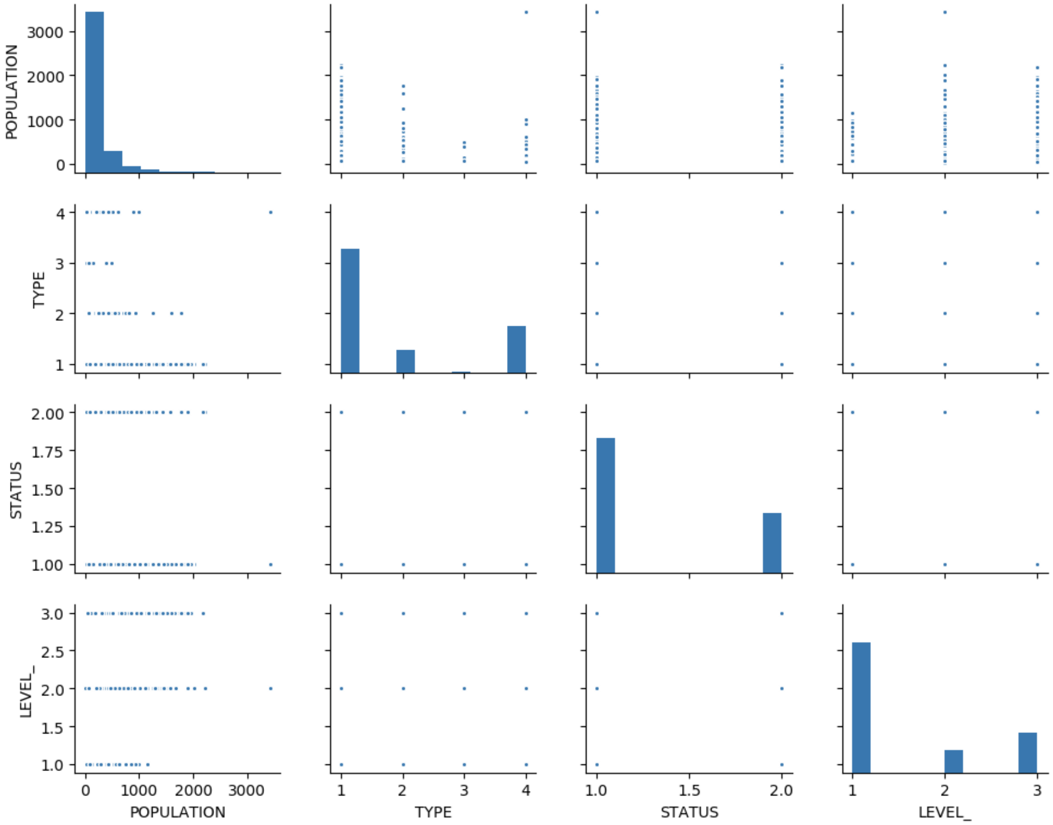
TYPE:

- 1 = Regular School;
- 2 = Special School;
- 3 = Vocational School;
- 4 = Other/alternative School;

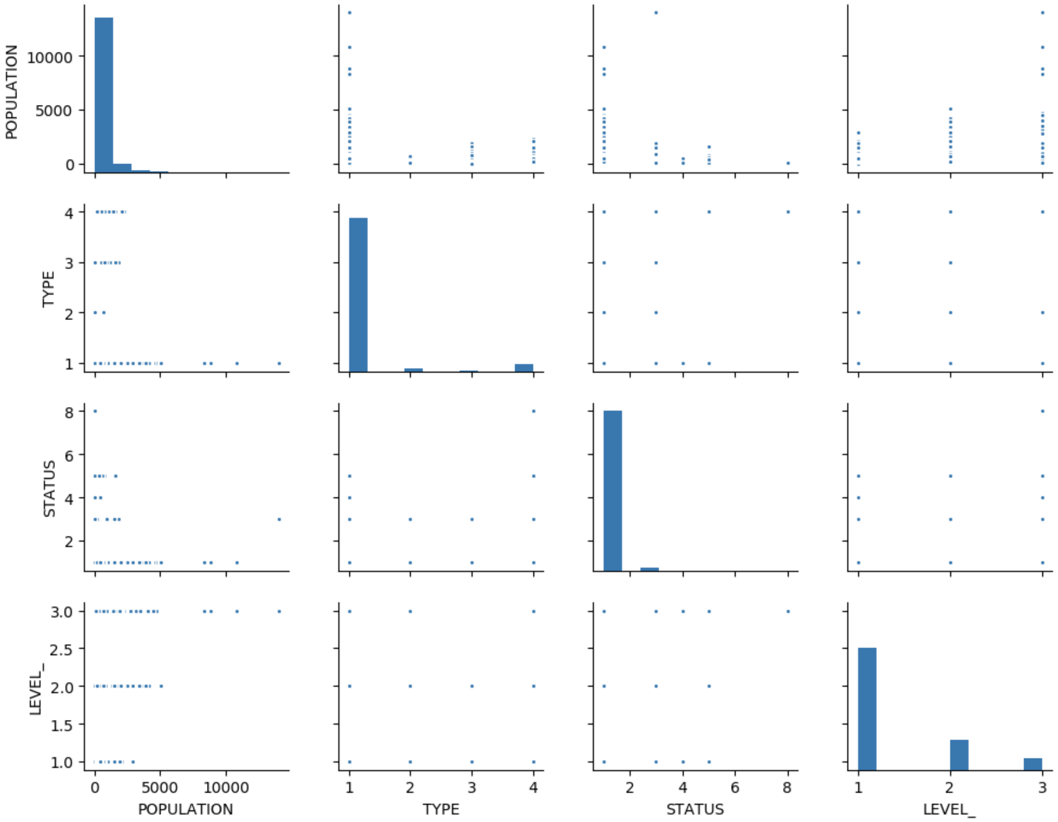
Furthermore, attributes LEVEL, TYPE, POPULATION has been converted into multiple columns by summing up its value by each zip code respectively. So, the columns containing the counts of each school TYPE, LEVEL, POPULATION for each specific zip code are generated.

The following figures are the plots after the two datasets are cleaned (upper: Private School Data, lower: Public School Data):

Graph2. Cleaned Education Data Summary



*Cleaned Private School Data*



*Cleaned Public School Data*

And the overall trend of each attribute did not change much.

For the universities and college datasets, the attribute STATUS is checked to exclude the non-active schools, and the rows with the following representations were dropped, since the rows with the following representations also contain negative values in attributes like POPULATION:

- M = an institution that closed in the current year;
- D = delete institution is out of business;
- C = combined with another active institution;
- G = not applicable; were dropped

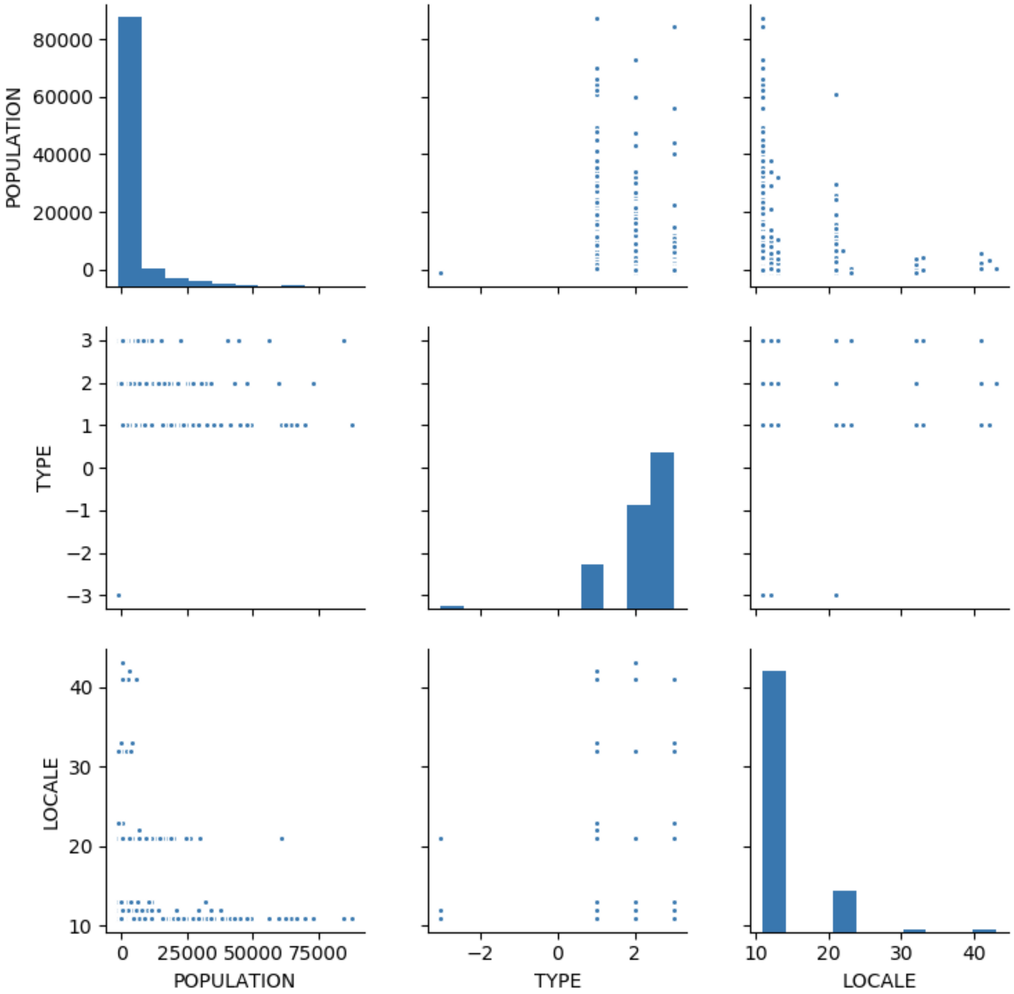
There were 18 rows out of 500 rows that contain the above representations.

The attribute ZIP was checked by the same way mentioned above, and all of the zip codes are correct, and POPULATION smaller than 0 is also check and dropped.

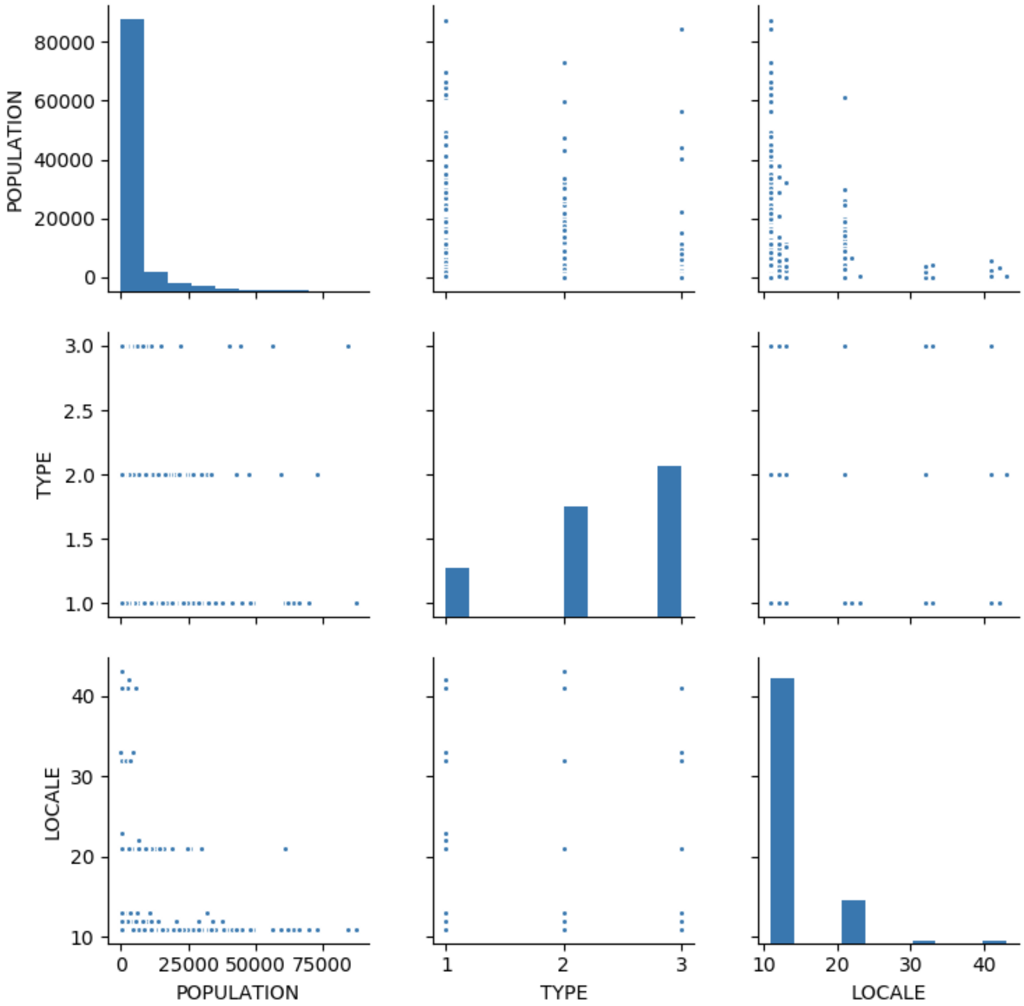
Also, attributes LOCALE, TYPE, POPULATION has been grouped by each zip code and converted into multiple columns by summing up the value in each row. So, new columns of each university type, location and population for each specific zip code are generated.

Following figures are the comparison of Universities and College dataset before and after cleaning (upper: University before cleaning, lower: University after cleaning):

Graph 3. Cleaned College Data Summary



*University Data Before Cleaning*



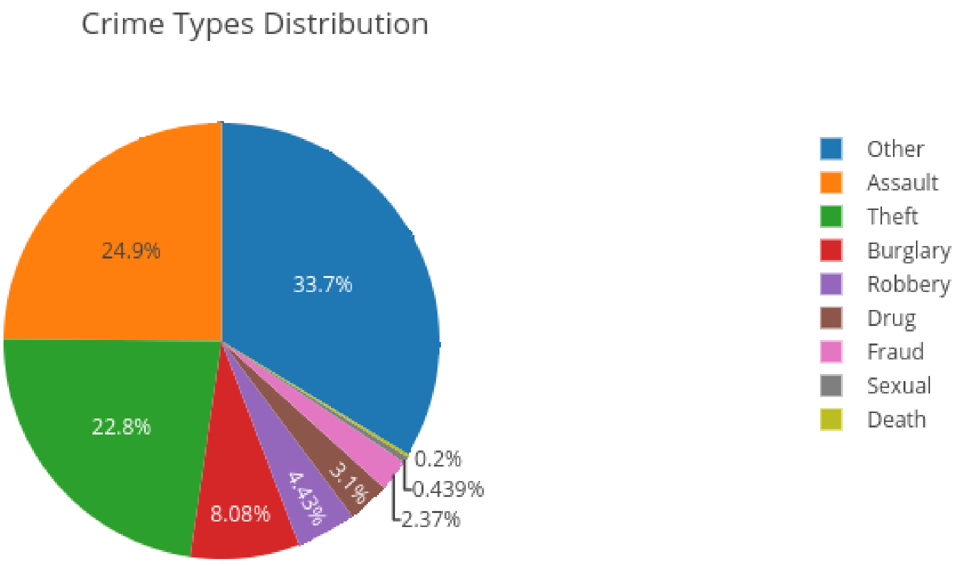
*Cleaned University*

Additionally, in order to perform association rule mining, the several attributes in school data were binned by equal-sized groups from 1 to 5, with the highest values are stored in group 5 for each attribute of interest, and the lowest values are stored in group 1 for each attribute of interest, specifically for school counts and school population for each school type.

**Crime Data:**

After data cleaning, the crime types in the dataset are distributed as the graph 4 below:

Graph4. Crime Types Distribution



As the graph shows, the crime types are not evenly distributed, in this way, in order to deal with imbalanced counts of crime types, the normalization and preprocessing have been performed in analysis part.

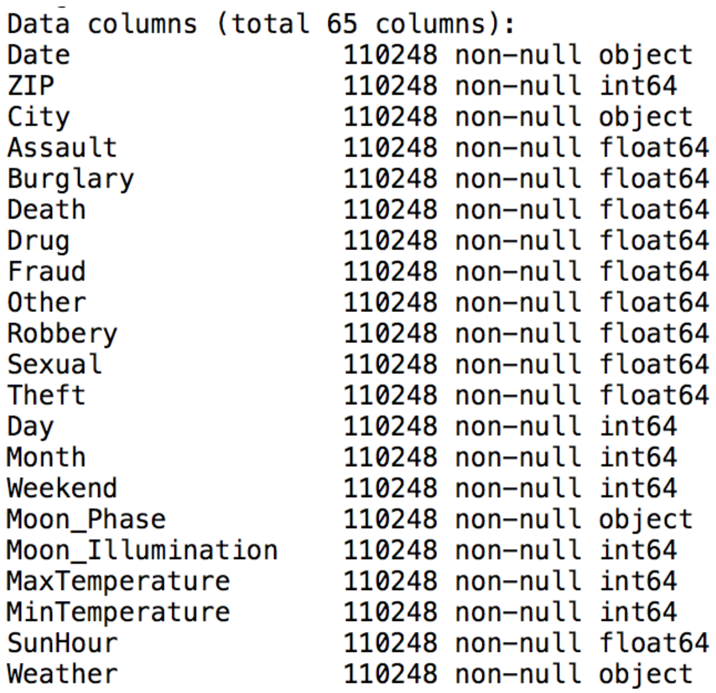
## Mean, Median, Mode, Standard Deviation:

The mean, median, and standard deviation of the following attributes are closely examined, the first four are collected from the Moon phase dataset, and the rest are collected from the Crime dataset:

1. Moon Illumination
2. Max Temperature
3. Min Temperature
4. Sun Hour
5. Assault
6. Burglary
7. Death
8. Drug
9. Fraud
10. Robbery
11. Sexual
12. Theft
13. Private School Count
14. Private School Population
15. Public School Count
16. Public School Population
17. University Count
18. University Population

Here is part of the results printed in the Python console by using the info() function in Python, and the data type are below:

Figure 1. Data Type Summary

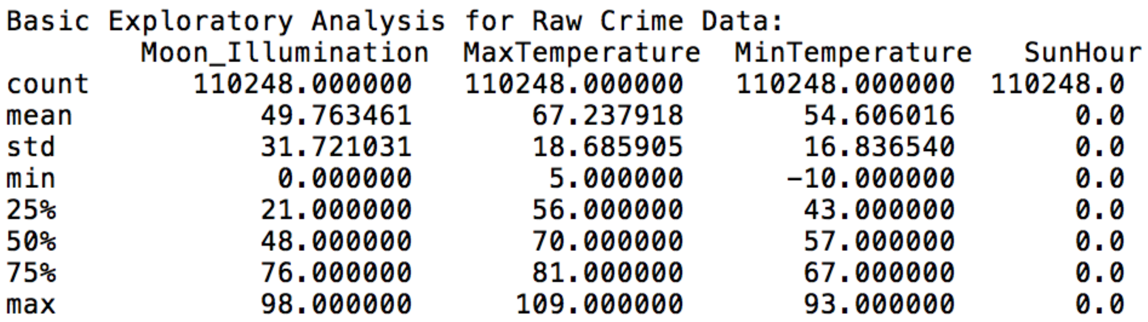


```
Data columns (total 65 columns):
Date                     110248 non-null object
ZIP                      110248 non-null int64
City                     110248 non-null object
Assault                  110248 non-null float64
Burglary                 110248 non-null float64
Death                    110248 non-null float64
Drug                     110248 non-null float64
Fraud                    110248 non-null float64
Other                    110248 non-null float64
Robbery                  110248 non-null float64
Sexual                   110248 non-null float64
Theft                    110248 non-null float64
Day                      110248 non-null int64
Month                    110248 non-null int64
Weekend                  110248 non-null int64
Moon_Phase               110248 non-null object
Moon_Illumination        110248 non-null int64
MaxTemperature           110248 non-null int64
MinTemperature           110248 non-null int64
SunHour                  110248 non-null float64
Weather                  110248 non-null object
```

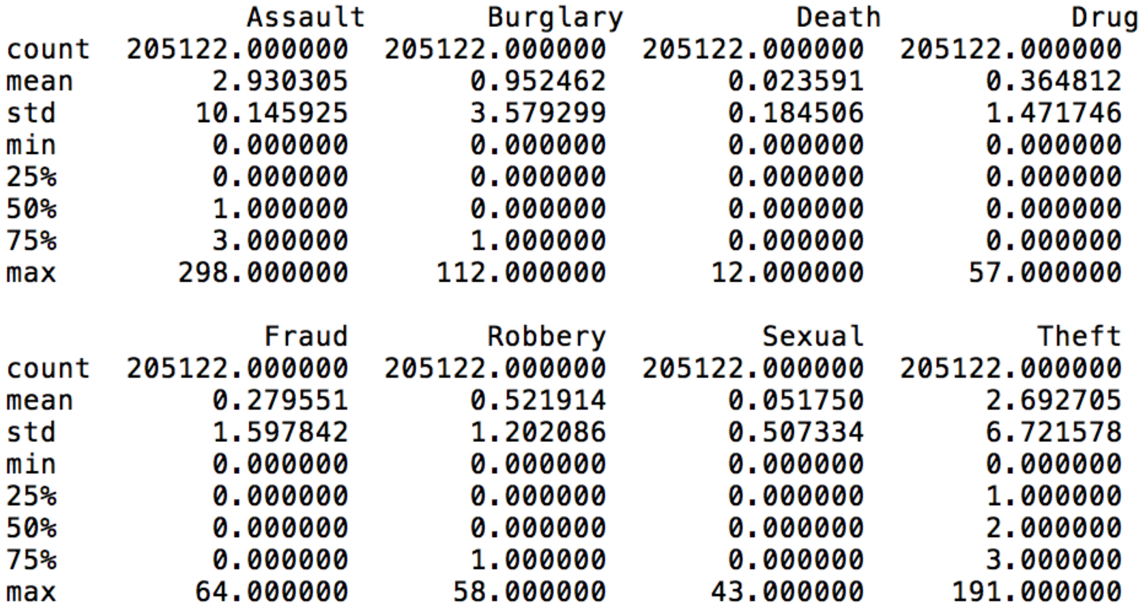
Both Date, City and Moon Phase attributes are object type. ZIP is in type int64, which were used a lot when merging the datasets; the rest of the attributes like Day, Month, Education (not shown), etc. are numerical data with either float64 or int64, since they are the counts for each attribute.

Here is part of the results exported for the attributes mentioned above by using the describe() function in Python:

Figure 2. Basic Exploratory Analysis for crime data



```
Basic Exploratory Analysis for Raw Crime Data:
        Moon_Illumination  MaxTemperature  MinTemperature   SunHour
count       110248.000000   110248.000000   110248.000000  110248.0
mean            49.763461       67.237918       54.606016       0.0
std             31.721031       18.685905       16.836540       0.0
min              0.000000        5.000000      -10.000000       0.0
25%             21.000000       56.000000       43.000000       0.0
50%             48.000000       70.000000       57.000000       0.0
75%             76.000000       81.000000       67.000000       0.0
max             98.000000      109.000000       93.000000       0.0
```



```
              Assault      Burglary         Death          Drug
count   205122.000000  205122.000000  205122.000000  205122.000000
mean         2.930305       0.952462       0.023591       0.364812
std         10.145925       3.579299       0.184506       1.471746
min          0.000000       0.000000       0.000000       0.000000
25%          0.000000       0.000000       0.000000       0.000000
50%          1.000000       0.000000       0.000000       0.000000
75%          3.000000       1.000000       0.000000       0.000000
max        298.000000     112.000000      12.000000      57.000000

                Fraud        Robbery         Sexual         Theft
count   205122.000000  205122.000000  205122.000000  205122.000000
mean         0.279551       0.521914       0.051750       2.692705
std          1.597842       1.202086       0.507334       6.721578
min          0.000000       0.000000       0.000000       0.000000
25%          0.000000       0.000000       0.000000       1.000000
50%          0.000000       0.000000       0.000000       2.000000
75%          0.000000       1.000000       0.000000       3.000000
max         64.000000      58.000000      43.000000     191.000000
```

The Moon Illumination, Max Temperature, Min Temperature, and Sun Hour are attributes relates to moon phase, which are used later for association rule mining and predictive analysis.

And min and max for most of the attributes for crime types are high since the datasets are discretely collected from different cities, or locations which may have various features, thus leads to high difference of each crime type.

In addition, there are lots of zeros in some of the crime type attributes, since some crime types have lower occurrence than other crime types, weighting method will be used to solve the problem.

Below is the frequency counts (mode) for the categorical attributes Moon Phase, Weather, and Day. The Waning Crescent has the highest frequency, while the New Moon has the lowest frequency. Which actually follows the overall changing pattern of the moon phase. For attribute Weather, Sunny has the most frequency of 31745. For attribute Day, Friday is the day with the most occurrence of crime, while Tuesday has the least occurrence of crime.

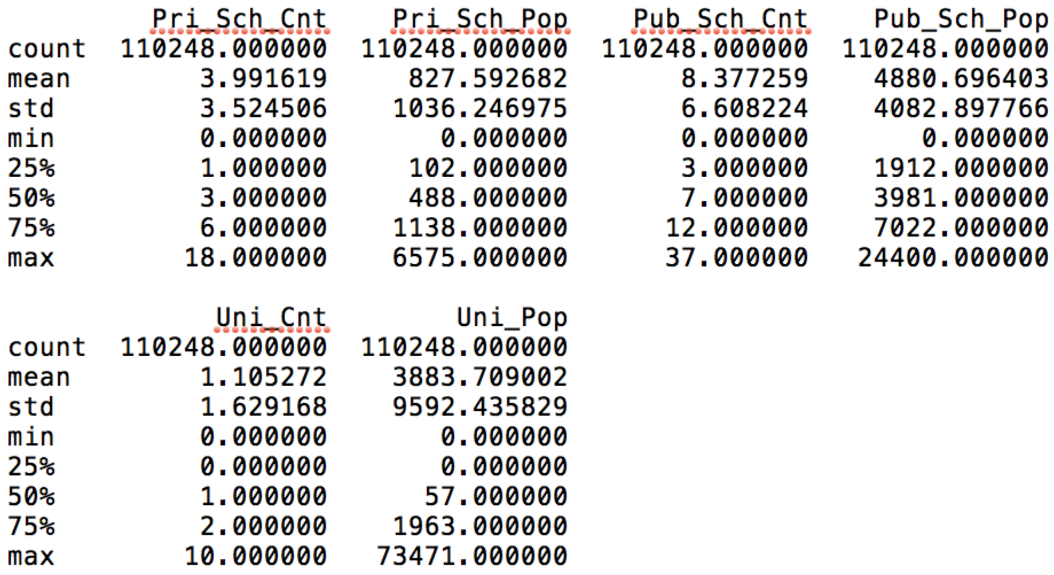
Figure 3. Basic Exploratory Analysis for crime data with moon phase data

```
Basic Exploratory Analysis for mode of Moon_Phase in Raw Crime Data:
 Waning Crescent    25395
First Quarter      15091
Waxing Crescent    14807
Waning Gibbous     14538
Full Moon          14470
Waxing Gibbous     11417
Last Quarter       10898
New Moon            3632
Name: Moon_Phase, dtype: int64
```

```
Basic Exploratory Analysis for mode of Weather in Raw Crime Data:
 Sunny                                     31745
Overcast                                  19962
Partly cloudy                             16889
Cloudy                                     7646
Patchy rain possible                       7260
Light rain shower                          6138
Moderate or heavy rain shower              3012
Mist                                       2689
Light drizzle                              2613
Light rain                                 2438
Moderate rain                              1443
Patchy light rain with thunder             1334
Moderate or heavy rain with thunder        1182
Thundery outbreaks possible                1142
Heavy rain                                 1043
Light snow                                  599
```

```
Basic Exploratory Analysis for mode of Day in Raw Crime Data:
 4    15832
6    15794
0    15744
2    15740
5    15723
3    15723
1    15692
Name: Day, dtype: int64
```

Below is the mean, median, standard deviation for Education data. The attributes are counts for different school types (Private, Public and University) and its corresponding population. For example, there are Private School Count, Public School Count, University Count, Private School Population, Public School Population and University Population. These data have high variance since there may have places with 0 schools, but other places with more than 10 schools. The high variance may tell interesting relationship between school data and other information.



```
          Pri_Sch_Cnt    Pri_Sch_Pop    Pub_Sch_Cnt    Pub_Sch_Pop
count   110248.000000  110248.000000  110248.000000  110248.000000
mean         3.991619     827.592682       8.377259    4880.696403
std          3.524506    1036.246975       6.608224    4082.897766
min          0.000000       0.000000       0.000000       0.000000
25%          1.000000     102.000000       3.000000    1912.000000
50%          3.000000     488.000000       7.000000    3981.000000
75%          6.000000    1138.000000      12.000000    7022.000000
max         18.000000    6575.000000      37.000000   24400.000000

              Uni_Cnt        Uni_Pop
count   110248.000000  110248.000000
mean         1.105272    3883.709002
std          1.629168    9592.435829
min          0.000000       0.000000
25%          0.000000       0.000000
50%          1.000000      57.000000
75%          2.000000    1963.000000
max         10.000000   73471.000000
```

## Preprocess Features and Create New Columns:

### Preprocessing, Normalizing, Binning crime types data:

Crime Type Data:

Preprocessing:

After dealing with this inconsistent data for crime types, the crime type data are grouped by zip code for merging with education data, and by date for merging with moon phase data. In addition, the crime type column ‘Type’ has been converted into multiple columns of different categories of crime types in order to show the counts of each crime type in specific zip code this year or in specific date in the city.

Normalizing:

Some crime types tend to occur more frequently than others (As the graph shown), such as Assault and Theft. In addition, crime occurrences are also related to cities, in order to avoid bias, the normalization has been utilized. The following method has been utilized to weigh the crime counts for each crime type.

Method 1:
X = Total counts of crime type i in zip i in date i
Y = Total counts of crime type i in city i (includes zip i) in date i

Weighted value = X / Y

Binning:

In order to find the most frequent crime type in each record, the largest weighted value in each row is supposed to be extracted. Therefore, a new column called “max” has been created for storing the crime type with the largest weighted value. The value of this new column means that this crime type happens the most frequent for each corresponding date and zip code. This column is the output for predictive analysis in this project.

### Weather Data

For the weather column, different weathers are categorized into either “rainy” and “not rainy” weather. For example, rainy weather contains weather such as light drizzle, light rain, moderate rain, heavy rain, etc.; and not rainy weather are like sunny, overcast, cloudy weathers.

### Education Data

Total population counts of the different school types are computed and transformed by the following weighting matrix, so the education level for each zip code can thus be calculated and used for further analysis:

Method:
X = Total school counts for each zip code
Y = Total population counts for each zip code

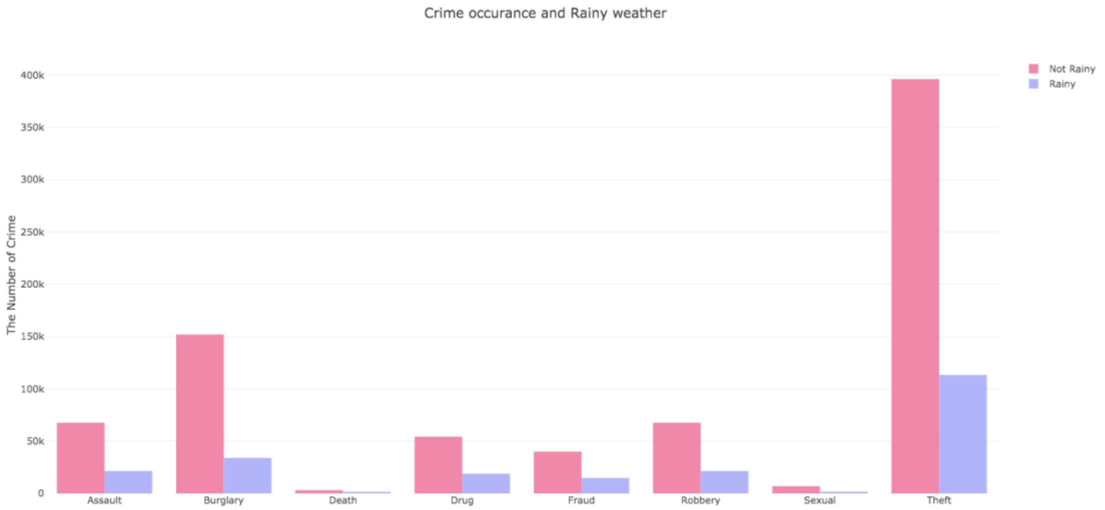
Weighted value = X / Y

And the weighted values are further grouped into 4 equal groups: Top 20%, 20 ~ 40%, 60 ~ 80% and Others, where “Top 20%” contains the highest 20% of weighted values, and “Others” contains the lowest rankings of the weighted values for each zip code.

## Exploratory Plots and Correlations

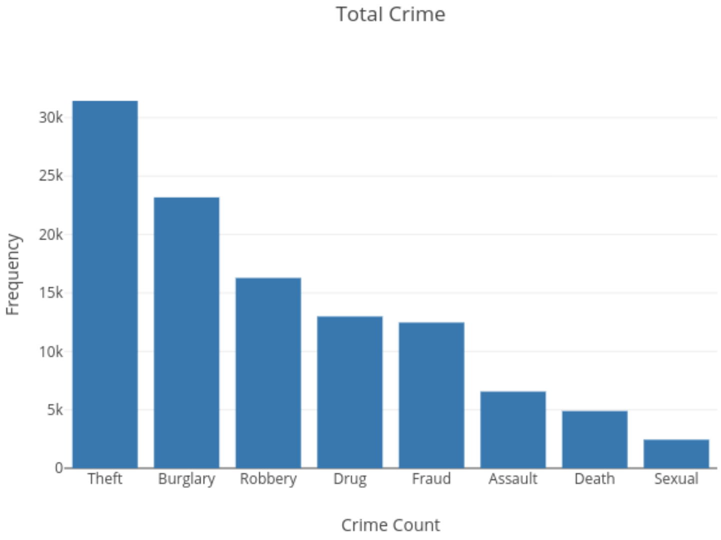
Graph 5. Distribution of Crime Type with Rainy Weather

Histogram



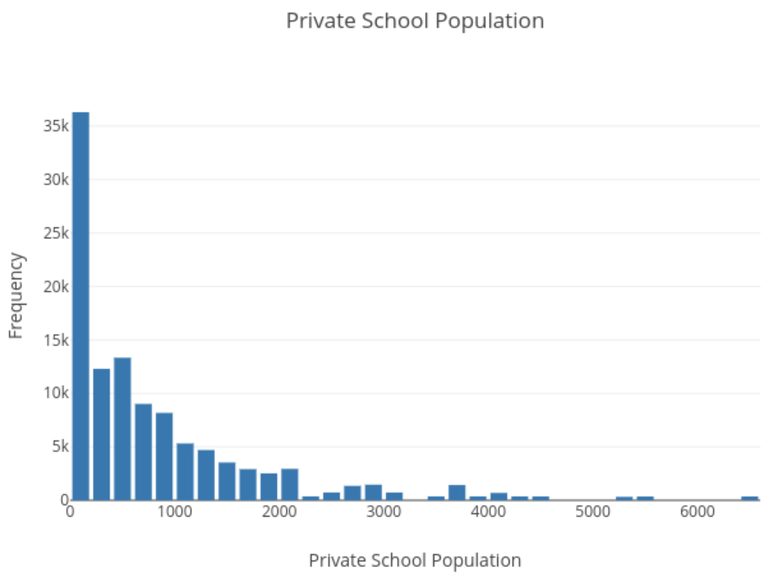
The above plot shows the number of different crime types happen in either rainy or non-rainy weather. And from the histogram, it is clear that overall, not rainy weather has higher crime occurrences than that in not rainy weather. Further analysis on how rainy weather would affect different crime types will be in the hypothesis testing section of the predictive analysis.

Graph 6. Distribution of Crime Types



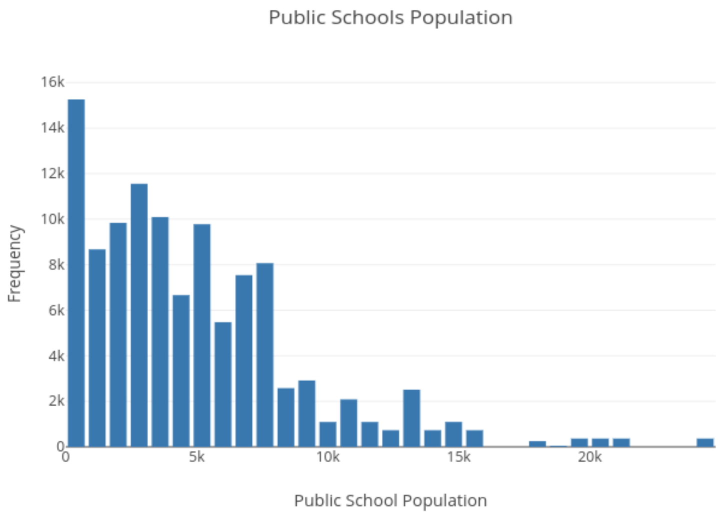
From the histogram above, it is clear that the two most commonly occurred crime type are theft and burglary, while death and sexual crime seem to occur less often, which is the same as common knowledge.

Graph 7. Distribution of Private School Population



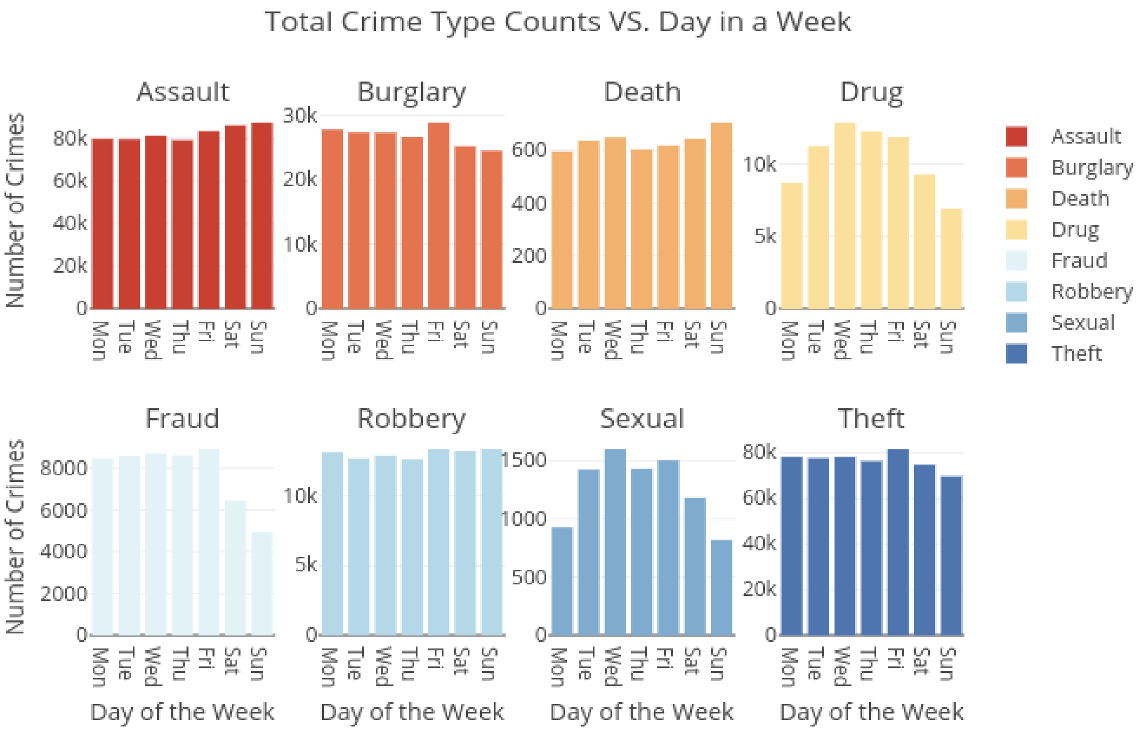
From the histogram, private schools overall have the most schools with student population smaller than 200.

Graph 8. Distribution of Public School Population



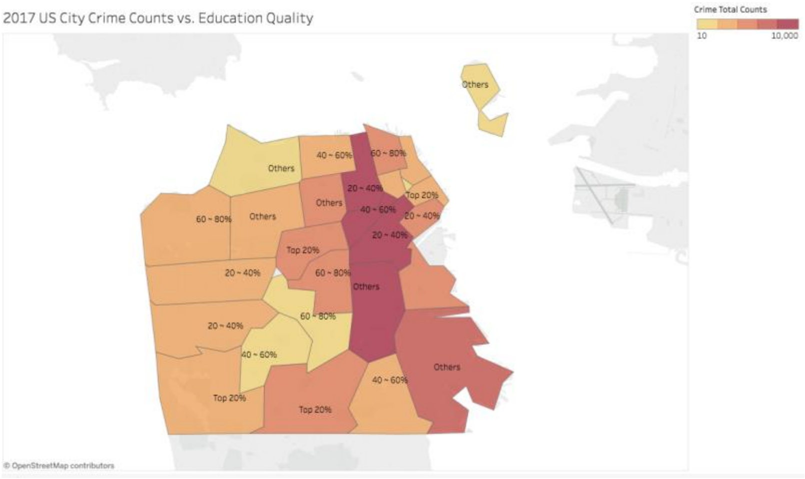
From the histogram, public schools overall have the most schools with student population smaller than 800. The relationship between private and public schools with crime types will be further examined by using association rule.

Graph 9. Distribution of Crime Types and Days in A week



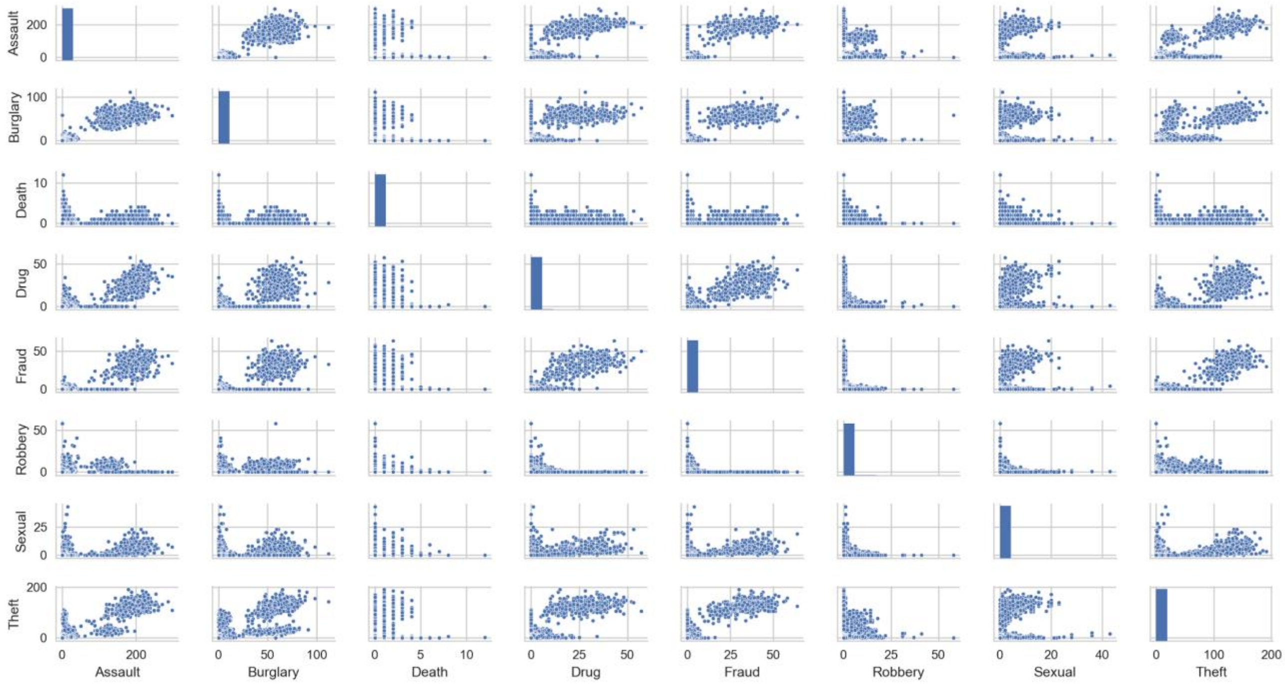
The above histograms are the total crime type counts vs. day in a week. Interesting patterns can be found in this plot. For instance, there seem to have a correlation between drug and sexual crime, as well as between burglary and theft. Further statistical test will be conducted in the additional analysis section. In addition, for certain crime type like assault and death, weekends tend to have higher crime occurrence, while for fraud weekdays seem to have higher occurrence. Furthermore, for drug crime, Wednesday seems to have the most crime occurrence, further analysis about day and crime types will also be conducted in the additional analysis section.

## Geographic Map


2017 US City Crime Counts vs. Education Quality
Crime Total Counts
10
10,000
Others
40 ~ 60%
60 ~ 80%
20 ~ 40%
Top 20%
© OpenStreetMap contributors

The geographic map above shows the crime occurrence in one of the major cities of interests (San Francisco) in 2017. The red color represents the locations where the crime happened the most frequent, and the lighter color represents the locations where crime happens less frequent. In most of the cases, the red areas are the downtown area for one city. The labels on the map are the rankings for education level per zip code. An interesting pattern is that in most downtown areas, the education level tends to be lower compared to other areas. Hypothesis testing will be made in order to check if there is statistical evidence to support such pattern.

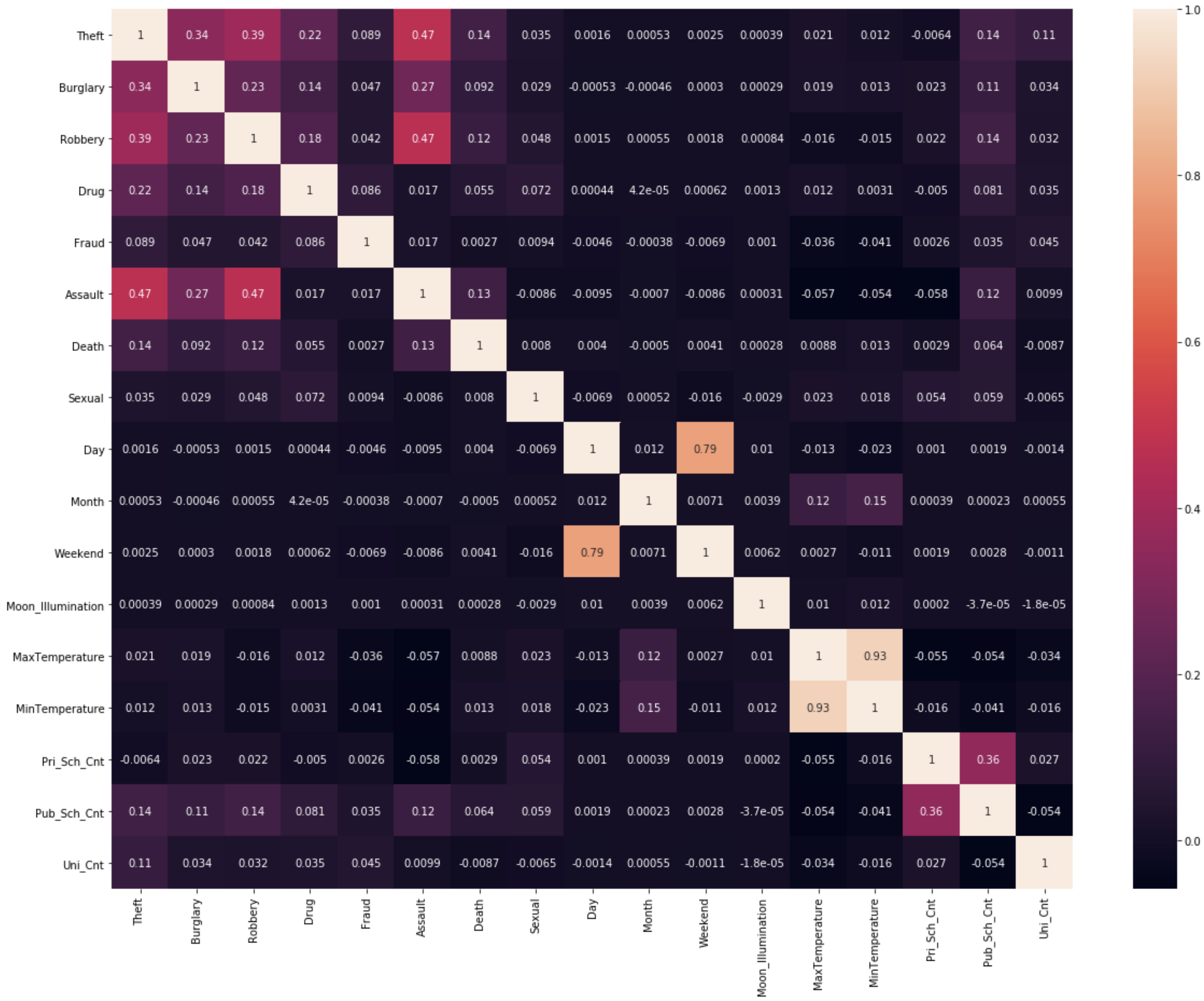
Scatter Plot



The above scatter plot shows the relationship between different crime types. Among the 8 crime types, there seems to exist some crime types that have a positive correlation between each other. For example, Assault and Burglary, Assault and Drug, Drug and Fraud.

## Correlation

Graph 10. Correlation Distribution between Crime Types

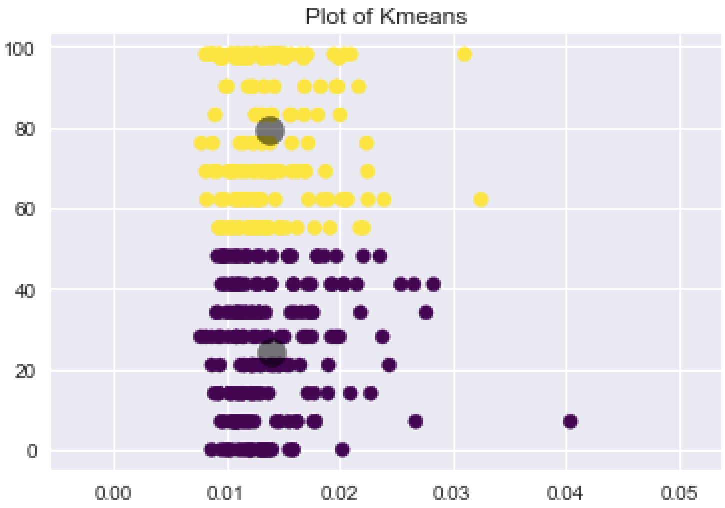


| | Theft | Burglary | Robbery | Drug | Fraud | Assault | Death | Sexual | Day | Month | Weekend | Moon_Illumination | MaxTemperature | MinTemperature | Pri_Sch_Cnt | Pub_Sch_Cnt | Uni_Cnt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Theft | 1 | 0.34 | 0.39 | 0.22 | 0.089 | 0.47 | 0.14 | 0.035 | 0.0016 | 0.00053 | 0.0025 | 0.00039 | 0.021 | 0.012 | -0.0064 | 0.14 | 0.11 |
| Burglary | 0.34 | 1 | 0.23 | 0.14 | 0.047 | 0.27 | 0.092 | 0.029 | -0.00053 | -0.00046 | 0.0003 | 0.00029 | 0.019 | 0.013 | 0.023 | 0.11 | 0.034 |
| Robbery | 0.39 | 0.23 | 1 | 0.18 | 0.042 | 0.47 | 0.12 | 0.048 | 0.0015 | 0.00055 | 0.0018 | 0.00084 | -0.016 | -0.015 | 0.022 | 0.14 | 0.032 |
| Drug | 0.22 | 0.14 | 0.18 | 1 | 0.086 | 0.017 | 0.055 | 0.072 | 0.00044 | 4.2e-05 | 0.00062 | 0.0013 | 0.012 | 0.0031 | -0.005 | 0.081 | 0.035 |
| Fraud | 0.089 | 0.047 | 0.042 | 0.086 | 1 | 0.017 | 0.0027 | 0.0094 | -0.0046 | -0.00038 | -0.0069 | 0.001 | -0.036 | -0.041 | 0.0026 | 0.035 | 0.045 |
| Assault | 0.47 | 0.27 | 0.47 | 0.017 | 0.017 | 1 | 0.13 | -0.0086 | -0.0095 | -0.0007 | -0.0086 | 0.00031 | -0.057 | -0.054 | -0.058 | 0.12 | 0.0099 |
| Death | 0.14 | 0.092 | 0.12 | 0.055 | 0.0027 | 0.13 | 1 | 0.008 | 0.004 | -0.0005 | 0.0041 | 0.00028 | 0.0088 | 0.013 | 0.0029 | 0.064 | -0.0087 |
| Sexual | 0.035 | 0.029 | 0.048 | 0.072 | 0.0094 | -0.0086 | 0.008 | 1 | -0.0069 | 0.00052 | -0.016 | -0.0029 | 0.023 | 0.018 | 0.054 | 0.059 | -0.0065 |
| Day | 0.0016 | -0.00053 | 0.0015 | 0.00044 | -0.0046 | -0.0095 | 0.004 | -0.0069 | 1 | 0.012 | 0.79 | 0.01 | -0.013 | -0.023 | 0.001 | 0.0019 | -0.0014 |
| Month | 0.00053 | -0.00046 | 0.00055 | 4.2e-05 | -0.00038 | -0.0007 | -0.0005 | 0.00052 | 0.012 | 1 | 0.0071 | 0.0039 | 0.12 | 0.15 | 0.00039 | 0.00023 | 0.00055 |
| Weekend | 0.0025 | 0.0003 | 0.0018 | 0.00062 | -0.0069 | -0.0086 | 0.0041 | -0.016 | 0.79 | 0.0071 | 1 | 0.0062 | 0.0027 | -0.011 | 0.0019 | 0.0028 | -0.0011 |
| Moon_Illumination | 0.00039 | 0.00029 | 0.00084 | 0.0013 | 0.001 | 0.00031 | 0.00028 | -0.0029 | 0.01 | 0.0039 | 0.0062 | 1 | 0.01 | 0.012 | 0.0002 | -3.7e-05 | -1.8e-05 |
| MaxTemperature | 0.021 | 0.019 | -0.016 | 0.012 | -0.036 | -0.057 | 0.0088 | 0.023 | -0.013 | 0.12 | 0.0027 | 0.01 | 1 | 0.93 | -0.055 | -0.054 | -0.034 |
| MinTemperature | 0.012 | 0.013 | -0.015 | 0.0031 | -0.041 | -0.054 | 0.013 | 0.018 | -0.023 | 0.15 | -0.011 | 0.012 | 0.93 | 1 | -0.016 | -0.041 | -0.016 |
| Pri_Sch_Cnt | -0.0064 | 0.023 | 0.022 | -0.005 | 0.0026 | -0.058 | 0.0029 | 0.054 | 0.001 | 0.00039 | 0.0019 | 0.0002 | -0.055 | -0.016 | 1 | 0.36 | 0.027 |
| Pub_Sch_Cnt | 0.14 | 0.11 | 0.14 | 0.081 | 0.035 | 0.12 | 0.064 | 0.059 | 0.0019 | 0.00023 | 0.0028 | -3.7e-05 | -0.054 | -0.041 | 0.36 | 1 | -0.054 |
| Uni_Cnt | 0.11 | 0.034 | 0.032 | 0.035 | 0.045 | 0.0099 | -0.0087 | -0.0065 | -0.0014 | 0.00055 | -0.0011 | -1.8e-05 | -0.034 | -0.016 | 0.027 | -0.054 | 1 |

There are several attributes showing correlation to some extent, for example, Burglary and Theft has a correlation of 0.34, Robbery and Theft has a correlation of 0.39, and Theft and Assault have a correlation of 0.47. Other correlations are relatively close to zero and thus are less interested.

## Cluster Analysis

### K-means: Moon Illumination vs. Assault

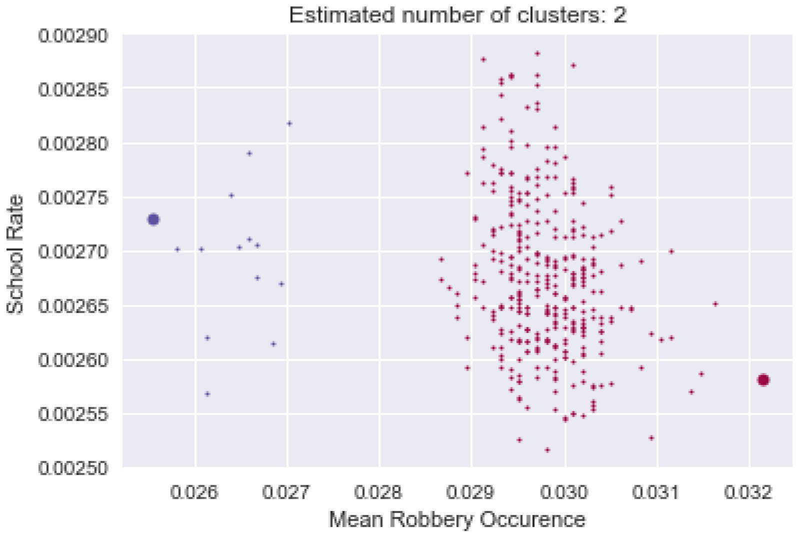


The x-axis is the mean Assault occurrence every day, and the y-axis is the Moon Illumination. There are two clusters in the plot, one with a higher level of moon illumination, and one with a relatively lower moon illumination. The assault occurrence range of most points are between 0.008 to 0.02.

The score of k-means method is 0.6248513603390273 which is acceptable.

DBScan: Robbery vs. School Rate

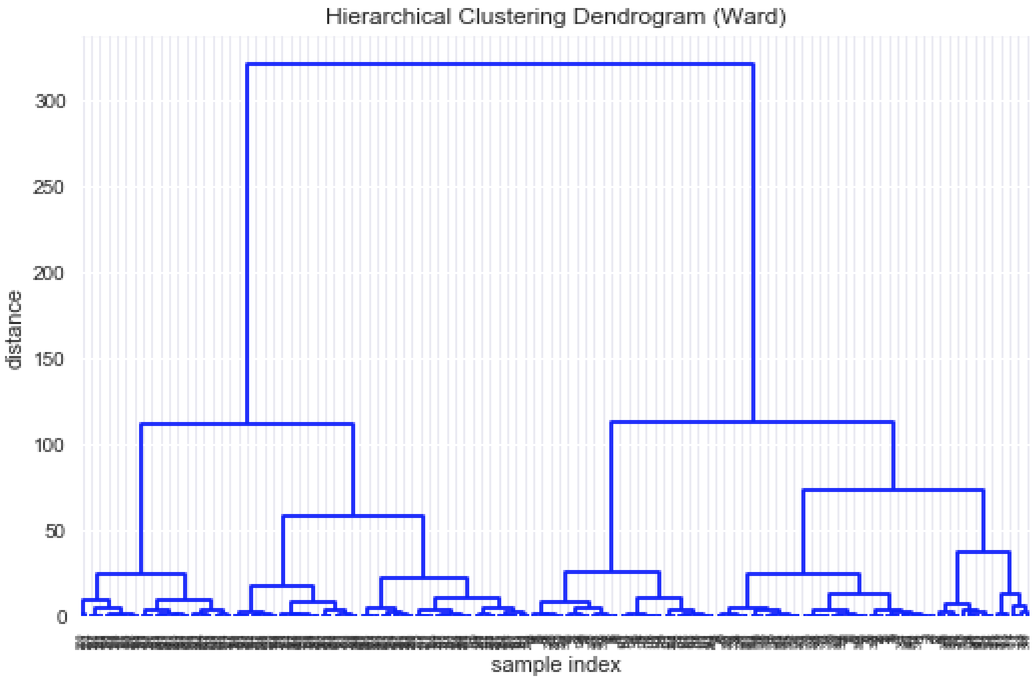
School Rate is a new variable representing the ratio of people and school number in an area.



From the centroid of two estimated clusters, the information is quite clear. That is, the higher the School Rate is, the lower the Robbery Occurrence is. Also, the cluster on the right is much denser than that on the left, which indicates that the group on the right represents common situations, and the group on the left represents some unusual condition (where in some areas the robbery rate is extremely low). The School Rate ranges of two groups do not have much difference.

The score of DBScan is 0.8400999793698287 which is relatively high.

Hierarchical: Drug & Max Temperature



The x-axis is Drug crime, and the y-axis is the Max Temperature. There are roughly 6 small clusters and they can be further grouped into two bigger clusters, where each bigger cluster has three subgroups.

# Association Rules

The cleaned final dataset is used to generate association rules, all detailed results are stored as text files in the zip folder. For Max temperature and minimum temperature, the average temperature is computed. The data used for association rules is first binned by equal-sized group of 1 to 5 for each attribute, so the highest values are stored in group 5, and the lowest values are stored in group 1 for each attribute of interest. The selected attributes formed a new dataframe and it is transformed into the matrix format. And the Apyori package is used for generating the rules. And different support level, confidence level and lift level are chosen based on the overall performance of the dataset.

## Most Frequent Crime Type vs. Moon Phase

Association Rules consist of the most frequent crime, different moon phase, sun hour level and average temperature level are generated individually with a minimum support of [0.03, 0.05, 0.08, 0.1, 0.3]. The minimum confidence is set at 0.3 and minimum lift at 1. If the lift is greater and 1, it may indicate that the items involved may have an underlying relationship.

The pattern that is the most frequent for this case is:

- {SunHour_3, AvgTemp_4}, with support = 0.433532

This rule indicates that 43.35% of the time, when the Sun hour is in the middle level, it will associate with a relatively higher average temperature.

Other association rules with minimum support of 0.2 are as the following:

- {Theft} -> {SunHour_3}, support = 0.285148, confidence = 1.0, lift = 1.0
- {Waning Crescent} -> {SunHour_3}, support = 0.230344, confidence = 1.0, lift = 1.0
- {Burglary} -> {SunHour_3}, support = 0.210244, confidence = 1.0, lift = 1.0

These rules indicate that more than 20% of the times, when the most frequent crime types are Theft and Burglary, the sun hour level of that day is more likely to be in the middle; and when the moon phase is Waning Crescent, the sun hour will also likely to be in the middle level.

## Different Crime Types vs. Moon Phase

Association Rules with minimum support of [0.05, 0.08, 0.1, 0.2, 0.5] are generated individually with minimum confidence 0.3 and minimum lift of 1. Rules associate different crime types with different moon phase.

The pattern that is the most frequent for this case is:

- {Theft_1, Assault_1}, with support = 0.987583

This rule indicates that 98.76% of the time, when the Theft crime occurrence is low, the occurrence of Assault crime will also be low. This rule intuitively makes sense since when the occurrence of one crime is low, it may indicate that this place/location overall is safer, and thus the likelihood of the occurrence of another crime will also be low.

Other association rules with minimum support of 0.9 are as the following:

- {Robbery_1} -> {Theft_1}, support = 0.965197, confidence = 0.993418, lift = 1.004894
- {Drug_1} -> {Fraud_1}, support = 0.953677, confidence = 0.984042, lift = 1.001289
- {Fraud_1, Assault_1} -> {Drug_1} ,  support = 0.951419 ,  confidence = 0.984097, lift = 1.001228

The above rules also imply that when an area has a overall low level crime occurrence of one type, it will also have low levels occurrence of other crime type. And certain crime types seem to associate with each other for more than 95% of the time, such as Assault and violence crime.

Association rules relate moon phase and crime types are as the following:

- {Waning Crescent} -> {Fraud_1}, support = 0.226453, confidence = 0.983107, lift = 1.000338
- {Waning Gibbous} -> {Robbery_1}, support = 0.128202, confidence = 0.972211, lift = 1.000638

This may infer that there may exist certain relationship between moonphase and certain types of crime.

## Most Frequent Crime Type vs. School Type

This association rule considers the relationship between the most frequent crime type and different school types. With minimum support of [0.05, 0.08, 0.1, 0.2, 0.3], association rules are generated individually with minimum confidence 0.3 and minimum lift of 1.

The pattern that is the most frequent for this case is:

- {Total_Pub_1} -> {Total_Pri_1}, with the support = 0.713591

This rule indicates that 71.34% of the time, lower number of public schools in one place/location may associate with lower number of public schools in that area. So, for places with lower education level, low level of one type of school should associate with a lower number of another type of school.

Other association rules with minimum support of 0.2 are as the following:

- {Theft} -> {Total_Pub_1}, support = 0.211224, confidence = 0.740751, lift = 1.010172
- {Burglary} -> {Total_Pri_1}, support = 0.20072, confidence = 0.9547, lift = 1.004445

These rules indicate that when the most frequent crime types are Theft and Burglary, more than 20% of the times, that location would have lower numbers of public and private schools respectively.

# Predictive Analysis

## Hypothesis Testing

Rainy Weather vs. Crime Type:
Hypothesis tests are utilized to select features for further investigation. Here is result of running t-test:
$H_0$: Crime rates of varied types are not affected by rainy weather.
$H_A$: The crime rates are higher when it’s rainy than not rainy.
The confidence level is set to 0.95.

Table1. T test Result for Rainy Weather vs. Crime Type

Rainy Weather vs. Crime Type

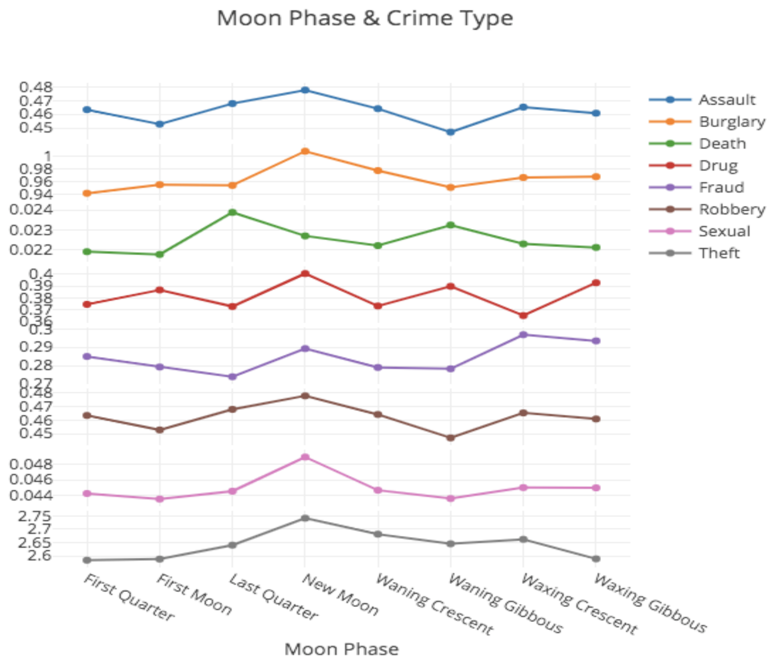
| | Assault | Burglary | Death | Drug | Fraud | Robbery | Sexual | Theft |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| p-value | 1 | < 2.2e-16 | 1 | 1 | 1 | 1 | 0.000112 | 0.01834 |

Therefore, the conclusion is:

For Assault, Death, Drug, Fraud and Robbery, the crime rates are not affected by rainy weather. However, the crime rates of Burglary, Sexual and Theft is lower when it’s rainy.

New Moon vs. Crime Rate:



This plot contains the occurrences of different crime type during different moon phases. Crime types such as Assault, Burglary, Drug, Robbery, Sexual and Theft can be roughly visualized the most occurrence during New Moon. Death crime happens the most during Last Quarter. And Fraud happens the most during Waxing Crescent. And further analysis on how different moon phase would affect crime occurrence will be examined later.

This t-test is performed to see if the crime rate is higher when it’s new moon than when it’s other moon phases. The crime rate data is grouped by moon phase:

$H_0$: Crime rates of varied types are not affected no matter what the moon phases is.

$H_A$: The crime rates are higher when it’s new moon.

The confidence level is set to 0.90.

Table2. T test New Moon vs. Crime Type

New Moon vs. Crime Rate

| | Assault | Burglary | Death | Drug | Fraud | Robbery | Sexual | Theft |
|---|---|---|---|---|---|---|---|---|
| p-value | 0.09825 | 0.1796 | 0.4481 | 0.1429 | 0.3873 | 0.09825 | 0.2631 | 0.1001 |

From the table above, only the crime rates of Assault, Robbery are affected by New Moon and are higher than usual, other crime types are unaffected. Compared to the line plot above, though they appeared to be affected by New Moon, in fact Burglary, Drug, Sexual and Theft are proved to not have statistical significance.

## Chi-square Test of Independence

The chi-square test of independence is also used to check whether education level, adult obesity and smoking rate would affect different crime types. The chi-square test statistic is computed and compare to a χ2 distribution so that the P-value can be given for two factor variables involved. All the crime types are categorized into 5 groups, with group 1 has the lowest values and group 5 has the highest crime occurrence. And education level is categorized into 2 groups, with group 1 has lower rate and group 2 has higher rates. The adult obesity rate and smoking rate are categorized into 2 groups by the national average obesity and smoking rate. The obesity group 1 is the obesity rate smaller than 0.3, and obesity group 2 is the obesity rate greater than 0.3. The smoking group 1 is the smoking rate smaller than 0.15, and smoking group 2 is the obesity rate greater than 0.15.

The hypothesis and results are shown as the following, with a significant level set at α = 0.05:

### Education

H0: Education level is independent of crime type (there is no association between education level and crime).

HA: Education level is not independent of crime type (there is an association between education level and crime).

Table3. Chi_Square test For Education level with crime type

| | Assault | Burglary | Death | Drug | Fraud | Robbery | Sexual | Theft |
|---|---|---|---|---|---|---|---|---|
| X-squared | 0.86569 | 7.4082 | 6.6929 | 11.66 | 4.617 | 10.708 | 5.9616 | 4.2303 |
| df | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| p-value | 0.9294 | 0.1158 | 0.153 | 0.02006 | 0.3289 | 0.03005 | 0.202 | 0.3757 |

From the above results, only Drug and Robbery crime have p-value below 0.05, which means there is enough evidence to reject the null hypothesis and accept the alternative hypothesis. Thus, education level is not independent with Drug and Robbery.

## Health

H0: Adult obesity rate is independent of crime type (there is no association between obesity rate and crime).

HA: Adult obesity rate is not independent of crime type (there is an association between obesity rate and crime).

Table4. Chi_Square test For Adult obesity rate with crime type

| | Assault | Burglary | Death | Drug | Fraud | Robbery | Sexual | Theft |
|---|---|---|---|---|---|---|---|---|
| X-squared | 55.085 | 898.86 | 1949.7 | 1946.5 | 119.81 | 452.01 | 24.397 | 668.49 |
| df | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| p-value | 3.118e-11 | 2.2e-16 | 2.2e-16 | 2.2e-16 | 2.2e-16 | 2.2e-16 | 6.65e-05 | 2.2e-16 |

From the above results, all the crime types have a p-value below 0.05, which means there is enough evidence to reject the null hypothesis and accept the alternative hypothesis. Thus, adult obesity rate is dependent with different crime types.

H0: Adult smoking rate is independent of crime type (there is no association between smoking rate and crime).

HA: Adult smoking rate is not independent of crime type (there is an association between smoking rate and crime).

Table5. Chi_Square test For Adult smoking rate with crime type

| | Assault | Burglary | Death | Drug | Fraud | Robbery | Sexual | Theft |
|---|---|---|---|---|---|---|---|---|
| X-squared | 96.125 | 842.79 | 907 | 1847.5 | 218.4 | 1295.8 | 86.564 | 301.06 |
| df | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| p-value | 2.2e-16 | 2.2e-16 | 2.2e-16 | 2.2e-16 | 2.2e-16 | 2.2e-16 | 2.2e-16 | 2.2e-16 |

From the above results, all the crime types have a p-value below 0.05, which means there is enough evidence to reject the null hypothesis and accept the alternative hypothesis. Thus, adult smoking rate is dependent with different crime types.

## Principal Component Analysis

Moreover, PCA is conducted to select two more most correlated features from moon illumination, max temperature, min temperature, day, month, week day/weekend:

Table 6. Two Component PCA analysis

| | Day | Month | Weekend | Moon_Illumination | Max Temperature | Min Temperature |
|---|---|---|---|---|---|---|
| PC-1 | -0.0006409 | -0.0006878 | -8.73E-05 | -0.9997474 | -0.0161386 | -0.0156138 |
| PC-2 | 0.0014325 | -0.0188789 | 7.09E-05 | 0.02244588 | -0.7452932 | -0.6660899 |

Comparing the absolute value of each components, the two most important components are moon illumination and max temperature.

## Spatial-Temporal Model (Monte Carlo Spatial-Temporal Algorithm):

1) Background and Introduction:

Theoretically, routine activity theory has proven most instructive for understanding temporal patterns in crime. However, the most prominent of the temporal crime patterns investigated is seasonality: crime (most often assault) increases during the summer months and decreases once routine activities are less often outside.

In this case, there is very little research investigating temporal patterns of crime at shorter time intervals such as within the week or even within the day. Also, most of research selected Block group as spatial units. Nevertheless, people are not familiar with the block groups units, instead, most of residents in the  US are more likely to use zip code as space unit.

Therefore, in order to detect and investigate spatial temporal pattern. Days in week are chose as time unit, in the same time, zip code is used as space unit in this project. Moreover, for visualizing as map for specific city with zip code. The LA city is the example city for showing the interactive map for spatial temporal analysis.

2) Algorithm and Method:

1. Choose a base dataset which is already normalized in earlier. From dataset without normalized, randomly sample 85 % of the points, with replacement. And normalized crime types as earlier in EDA part. This is effectively a bootstrap created by sampling from the test dataset. Simulate 300 times.

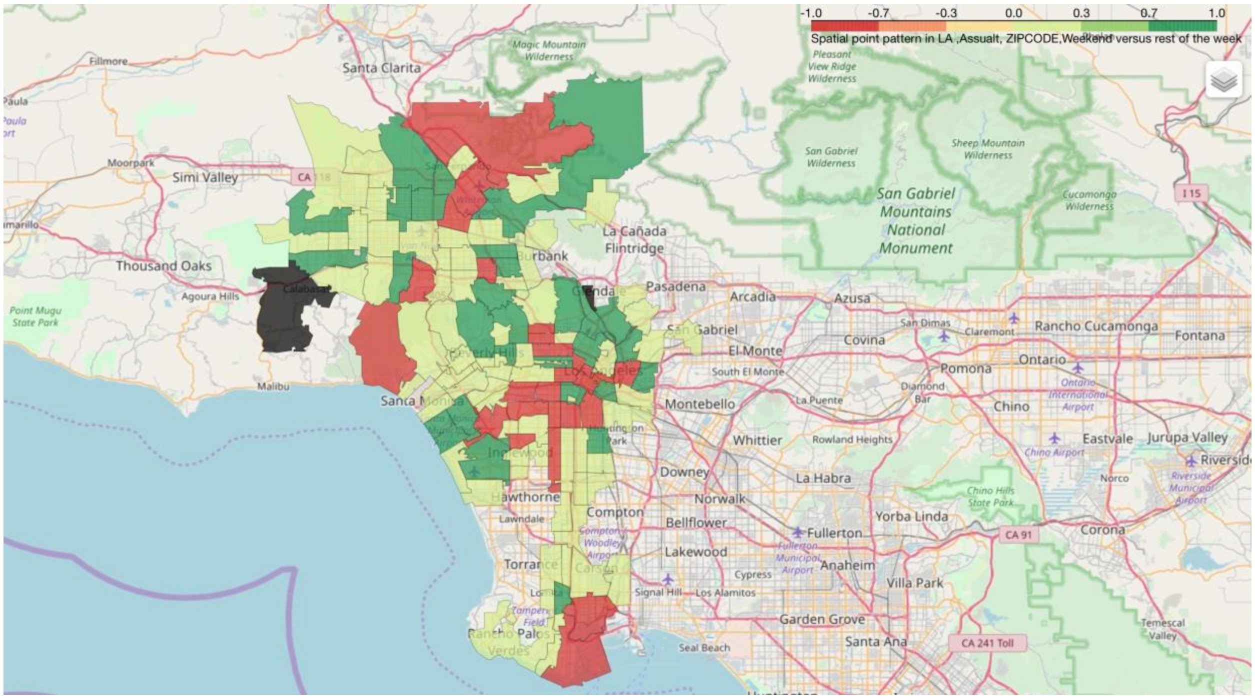
2. For spatial unit in the test data set, calculate the percentage of crime that has occurred in the area. Use these percentages to generate a 95 % nonparametric confidence interval by removing the top and bottom 0.25 % of all counts.

3. Calculate the percentage of points within each area for the base dataset and compare this to the confidence interval generated from the test dataset. If the base percentage falls within the confidence interval, then the two datasets exhibit a similar proportion of points in the given area (use 0 to indicates). Otherwise they are significantly different (Andresen and Malleson) (use 1 to indicates) As such, statistically significant changes/differences can be identified at the local level.

The output of the test is a global parameter with three values -1 (Significantly Lower) to 1 (Significantly higher), 0(no significantly difference)

3) Result:

Based on Spatial Temporal Analysis, the results clearly show that for the crime types under analysis there are distinctive temporal and spatial patterns for different days of the week. In this case, the predictive modeling in this project was performed with spatial temporal pattern, the spatial unit is zip code, and the time unit is day.



## Machine Learning

Predictive analysis:

From the analysis in earlier parts, there is a trend that the relationship and correlation between those features (Moon phase, weather, temperature, day, month, education level, obesity rate, smoking rate) and crime rates vary in different crime types. For examples, according to table 2, based on 90% confidence interval and the p value of t test, new moon and Assault crime and Robbery crime has statistical relationship. However, for other crime types, there is no significant relationship with new moon.

Thus, in predictive analysis, the predictive modeling was utilized for predicting the level of accuracy of different types of crime, (when the normalized occurrence higher than median, using 1 to represents , it means the high frequency. On the other hand, when the normalized occurrence lower than median, using 0 to represents, it means the low frequency. in this way, new 8 binary output features have been create, for example, 'Robbery' feature which contains value is [0,1], 1 represents the robbery crime tends to occurs more in this day and this area, and also, 0 represents tends to occurs more in this day and this area.

1. EDA result and Hypothesis test:

From the results in EDA part, first, from graph 5 and graph 6 , these two graphs show that the day in weeks and rain weather tend to be related with crime rates in different crime types. Consequently, these two features (day and rainy weather) are included for predictive analysis.

Moreover, From hypothesis test, Moon phase, especially new moon has significant relationship with two of crime types, and the table3 convinced that there is the significant relationship between education level and all different crime types.

Furthermore, based on the Table 3, which is the Two Component PCA analysis, the features Max Temperature and moon illumination attained high scores for both component, and 'Month' value obtain relatively higher results.

In this case, based on all these analysis, there are features have been selected for predictive analysis:
Day, Moon phase, Moon illumination, Max temperature, Month, Education level, Rainy weather .

2. Predictive results:

In this part, five predictive models have been used to predict if the specific crime type will be highly frequent type in the specific time-space pattern. Time unit is the day, Spatial unit is the zip code.

### 1) Robbery crime

For predicting Robbery crime type, the output 0 represents that the robbery crime is not the most- frequent- occurred crime type in this day and zip code, 1 represents that robbery crime plays a role as the most frequent crime type in the space time pattern.
the result of these four predictive modeling is that turned out below after 10-fold cross-validation:

KNN: 0.665366 (0.004904)

CART: 0.700742 (0.004369) - decision tree
NB: 0.696013 - naive bayes
SVM: 0.696

## 1. Decision tree:

The result of the decision tree is that:
CART:0.700742 (0.004369)

The theory of Decision Tree Classification is a supervised learning algorithm with a predefined target variable, which in this case, is “max”, which represents the most frequent crime types. It works for both categorical and continuous variables. Based on the result with 0.7 as accuracy and 0.0043 as standard deviation, it means 75.61% of data in the cross-validation test part attained accurate prediction.

## 2. Naive Bayes:

The result of the naive Bayes is that:
NB: 0.696013

It means the 69.6% of data in cross-validation test part attained accurate prediction.

To be specific Naive Bayes is a classification algorithm that combines prior knowledge based on information obtained from data to make predictions and assign class labels. And the conditional probability of the class given certain conditions is calculated using the Bayes theorem based on the datasets in order to give predictions.

In this way, the result of Naive Bayes shows that it is the optimal model.

## 3.SVM:

The result of the SVM is that:

SVM:0.696

Support Vector Machine is a supervised classifier. In SVM, a hyperplane is selected to best separate the points by their classes, usually two classes. The goal of an SVM is to determine the “best” way to separate the points in 2D (or hyperplane in higher dimension). The kernel selection will have significant influence on the result of this method.

In this predictive analysis, based on the performance showed above, it means the 69.6% of data in the cross-validation test part attained accurate prediction,  without high dimensional data, the SVM doesn't not to be used for this project.

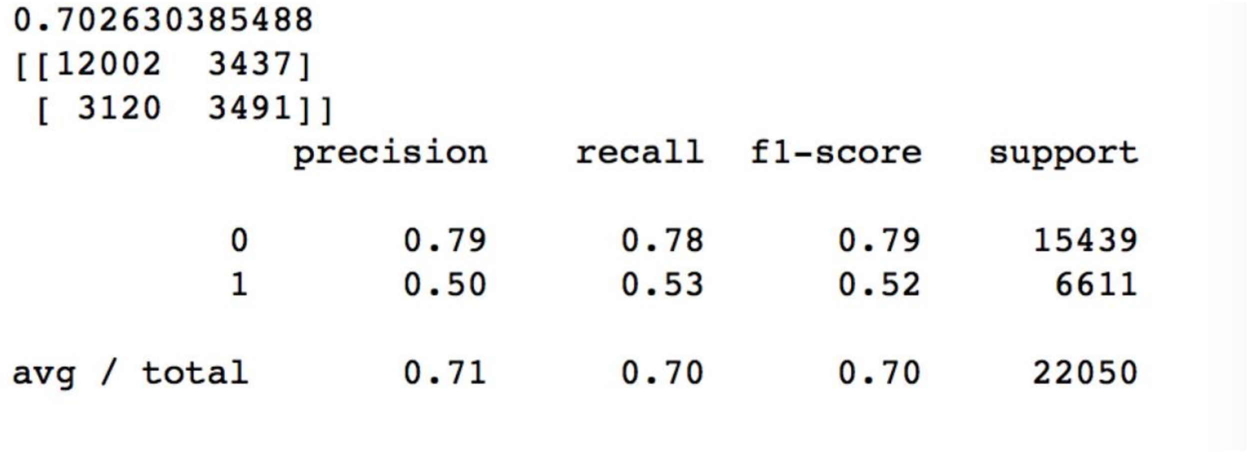
## 4.KNN:

The result of kNN is that:
KNN: 0.665366 (0.004904)

It means the 66.54 % of data in the cross-validation test part attained accurate prediction.
K-Nearest Neighbor Classification is to find a predefined number k of the training samples with the closest in distance to a new sample. The label of the new sample will be defined from these neighbors. When the kth nearest neighbors were calculated, the majority class of the kth nearest neighbors will be in same the class with the new sample.

## Choosing Optimal Model:

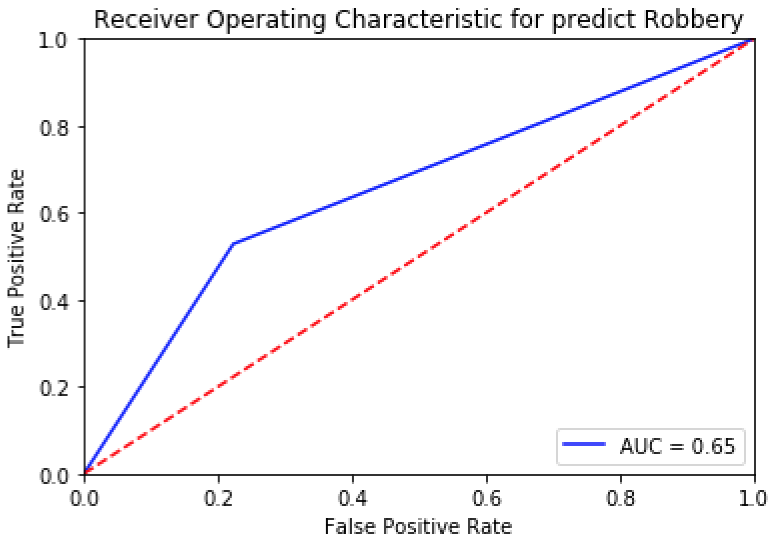
According to these four models which have been performed for predict, the optimal model in this project is Decision Tree. After testing the Decision Tree modeling, the result is below:
Accuracy is: 0.7, and the result shows below:



```
0.702630385488
[[12002  3437]
 [ 3120  3491]]
             precision    recall  f1-score   support

          0       0.79      0.78      0.79     15439
          1       0.50      0.53      0.52      6611

avg / total       0.71      0.70      0.70     22050
```

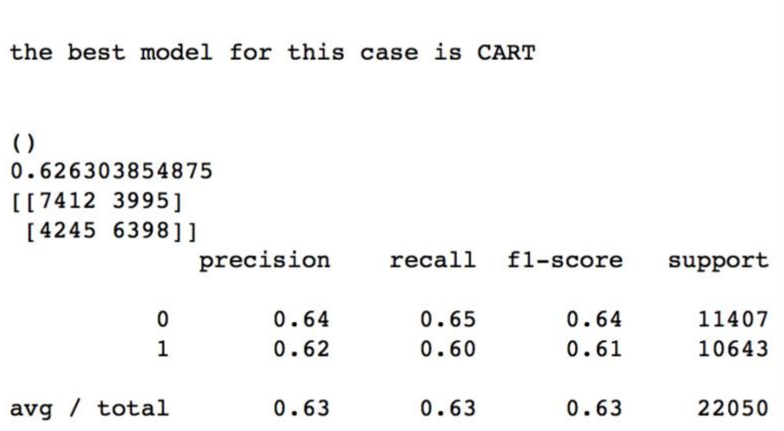


According to these four models which have been performed for predict, the optimal model in this project is Decision Tree. After testing the Naive Bayes modeling, the result is below: Accuracy is: 0.7, and the result shows below:

Based on the precision table, it shows that even though overall accuracy is 0.7, However, for '1' class, naive bayes model obtained 50 % accuracy, however, for the '0' class, it attained 79% accuracy. In this way, this modeling tends to perform relatively better for low frequency prediction, compared to predict high frequency in each spatial temporal pattern.

In the rest of crime types, the similar model selection method and analysis has been performed. The optimal model and ROC for each crime type prediction as the results shown below:

### 2) Burglary

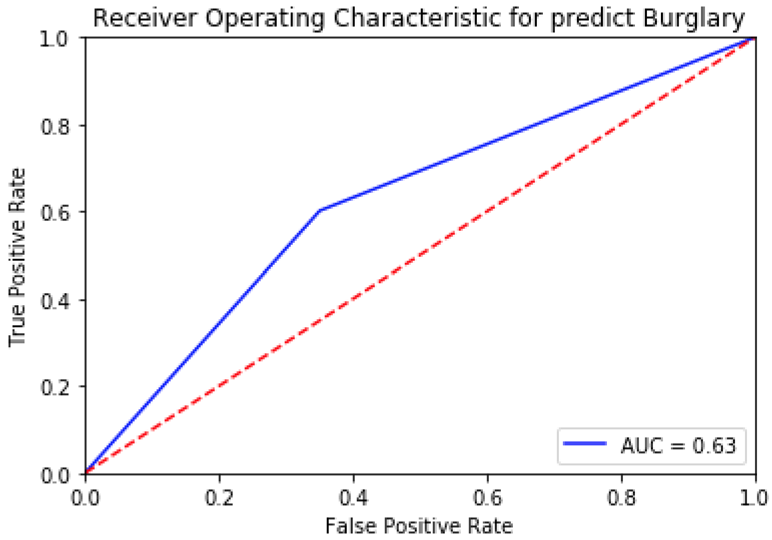


```
the best model for this case is CART


()
0.626303854875
[[7412 3995]
 [4245 6398]]
             precision    recall  f1-score   support

          0       0.64      0.65      0.64     11407
          1       0.62      0.60      0.61     10643

avg / total       0.63      0.63      0.63     22050
```



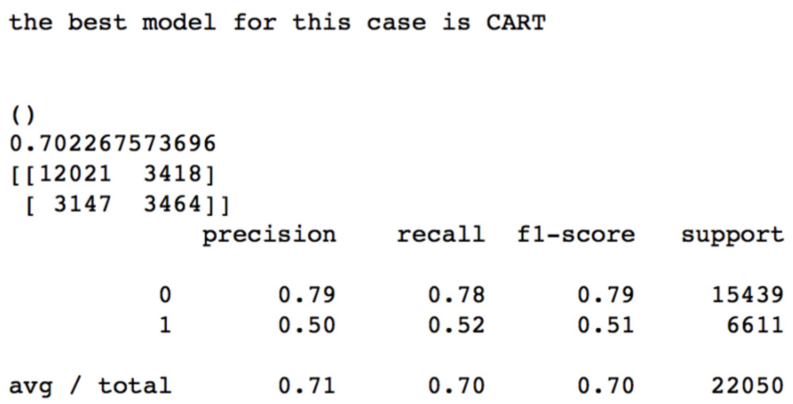
According to these four models which have been performed for predict, the optimal model in this project is Decision Tree. After testing the Naive Bayes modeling, the result is below: Accuracy is: 0.62, and the result shows below:

Based on the precision table, it shows that even though overall accuracy is 0.63, However, for '1' class, naive bayes model obtained 62 % accuracy, however, for the '0' class, it attained 64% accuracy. In this way, this modeling tends to perform relatively well for both high frequency prediction and low frequency prediction in each spatial temporal pattern.

The ROC curve for this optimal Model:

The AUC score is 0.63 and the ROC curve shows that it is relatively have a curve of the ROC space, it means the accuracy would not be lower when considering the performance as false positive rate and true positive rate.

### 3) Assault

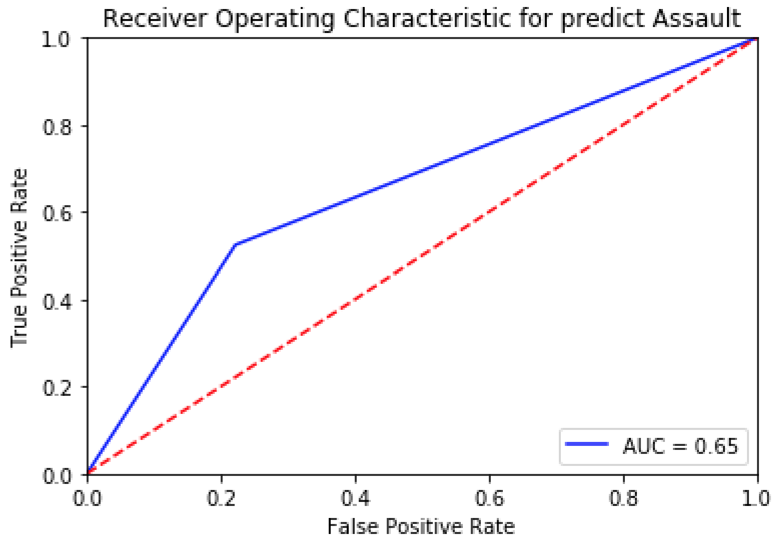


```
the best model for this case is CART


()
0.702267573696
[[12021  3418]
 [ 3147  3464]]
             precision    recall  f1-score   support

          0       0.79      0.78      0.79     15439
          1       0.50      0.52      0.51      6611

avg / total       0.71      0.70      0.70     22050
```



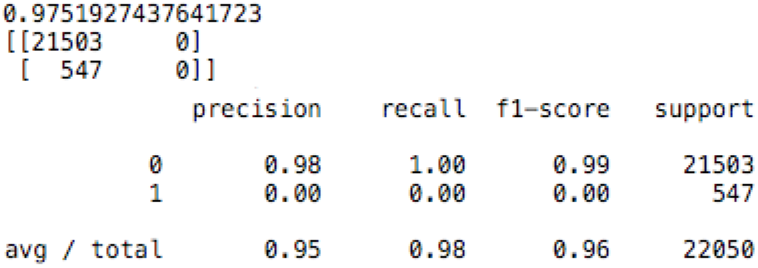
According to these four models which have been performed for predict, the optimal model in this project is Decision Tree. After testing the Naive Bayes modeling, the result is below:
Accuracy is: 0.70, and the result shows below:

Based on the precision table, it shows that even though overall accuracy is 0.70, However, for '0' class, naive bayes model obtained 79 % accuracy, however, for the '1' class, it attained 50% accuracy. In this way, this modeling tends to perform relatively better for low frequency prediction, compared to predict high frequency in each spatial temporal pattern.

The ROC curve for this optimal Model:

The AUC score is 0.65 and the ROC curve shows that it is relatively have a curve of the ROC space, it means the accuracy would still be relatively high when considering the performance as false positive rate and true positive rate.

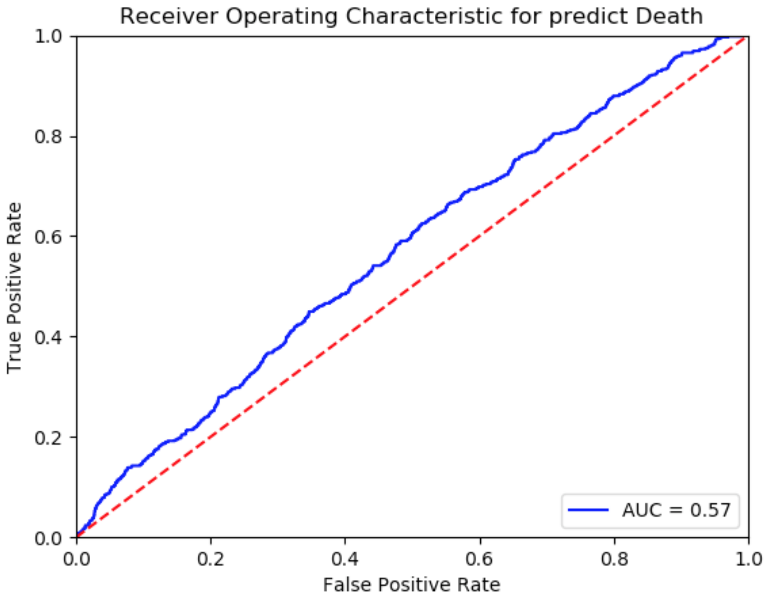
## 4) Death



```
0.9751927437641723
[[21503     0]
 [  547     0]]
             precision    recall  f1-score   support

          0       0.98      1.00      0.99     21503
          1       0.00      0.00      0.00       547

avg / total       0.95      0.98      0.96     22050
```



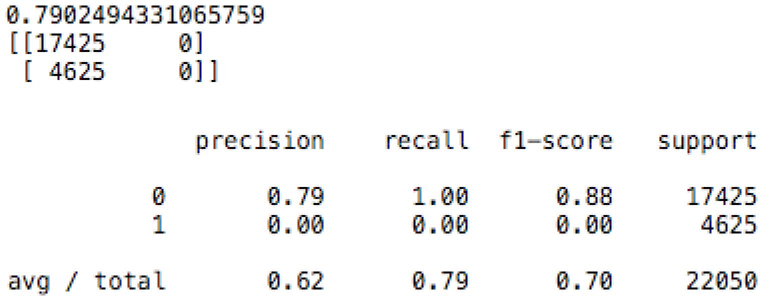
According to these four models which have been performed for predict, the optimal model in this project is Navie Bayes. After testing the Naive Bayes modeling, the result is below:
Accuracy is: 0.98, and the result shows below:

Based on the precision table, it shows that even though overall accuracy is 0.98, However, for '0' class, naive bayes model obtained 98 % accuracy, however, for the '1' class, it attained 0% accuracy. In this way, this modeling tends to perform well only for high frequency prediction.

The ROC curve for this optimal Model:

The AUC score is 0.57 and the ROC curve shows that it is closer to the 45-degree diagonal of the ROC space, it means the less accurate the test. Even though the average accuracy (0.98) is relatively high, due to the imbalanced data (most of the crime values in Death crime column in normalized dataset are 0) , and also the low correlation between the factors with death crime type, when considering the performance as false positive rate and true positive rate, it is less accurate.

### 5) Drug

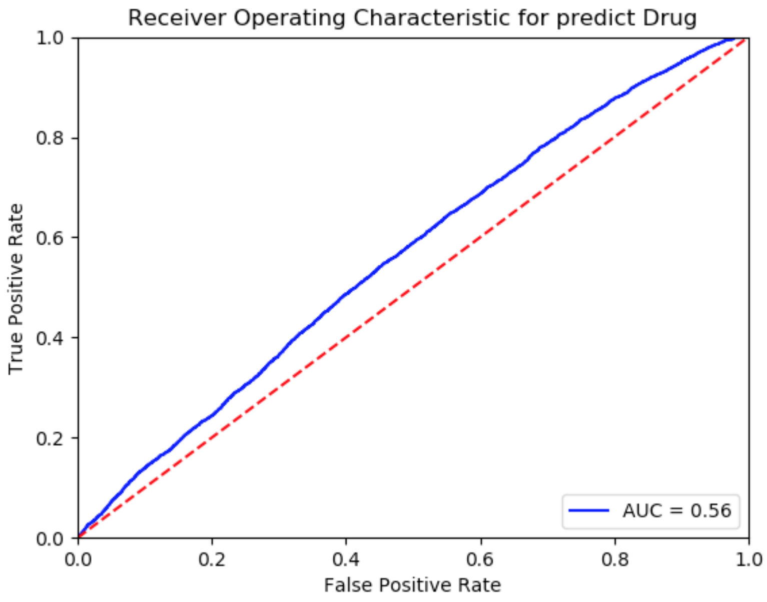


```
0.7902494331065759
[[17425     0]
 [ 4625     0]]

             precision    recall  f1-score   support

          0       0.79      1.00      0.88     17425
          1       0.00      0.00      0.00      4625

avg / total       0.62      0.79      0.70     22050
```



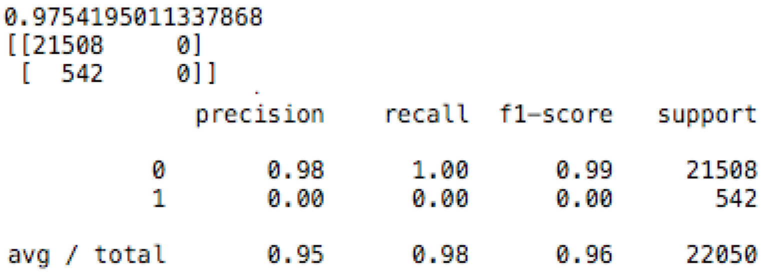
According to these four models which have been performed for predict, the optimal model in this project is Naive Bayes. After testing the Naive Bayes modeling, the result is below:
Accuracy is: 0.79, and the result shows below:

Based on the precision table, it shows that even though overall accuracy is 0.79, However, for '0' class, naive bayes model obtained 79 % accuracy, however, for the '1' class, it attained 0% accuracy. In this way, this modeling tends to perform well only for high frequency prediction.

The ROC curve for this optimal Model:

The AUC score is 0.57 and the ROC curve shows that it is closer to the 45-degree diagonal of the ROC space, it means the less accurate the test. Even though the average accuracy (0.98) is relatively high, due to the imbalanced data (most of the crime values in Drug crime column in normalized dataset are 0), when considering the performance as false positive rate and true positive rate, it is less accurate.

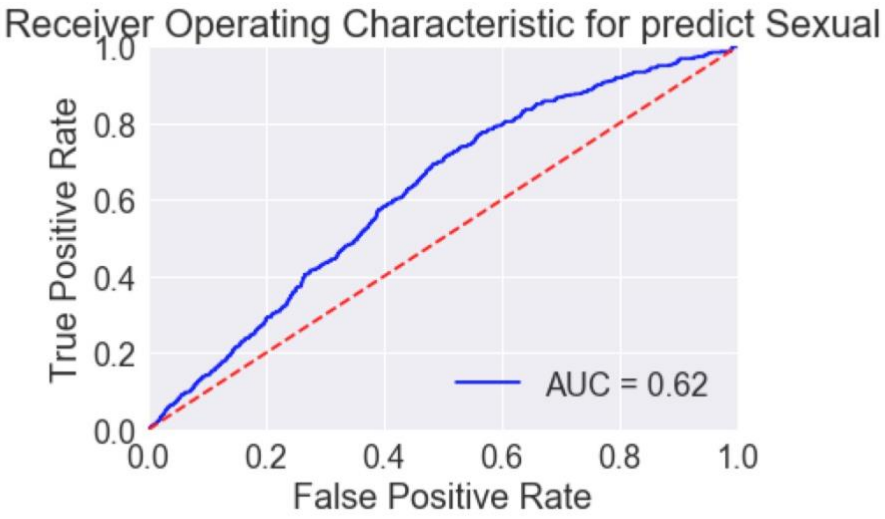
### 6) Sexual



```
0.9754195011337868
[[21508     0]
 [  542     0]]
             precision    recall  f1-score   support

          0       0.98      1.00      0.99     21508
          1       0.00      0.00      0.00       542

avg / total       0.95      0.98      0.96     22050
```



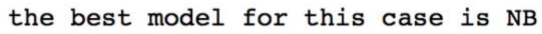
According to these four models which have been performed for predict, the optimal model in this project is Naive Bayes. After testing the Naive Bayes modeling, the result is below:
Accuracy is: 0.98, and the result shows below:

Based on the precision table, it shows that even though overall accuracy is 0.98, However, for '0' class, naive bayes model obtained 98 % accuracy, however, for the '1' class, it attained 0% accuracy. In this way, this modeling tends to perform well only for high frequency prediction.

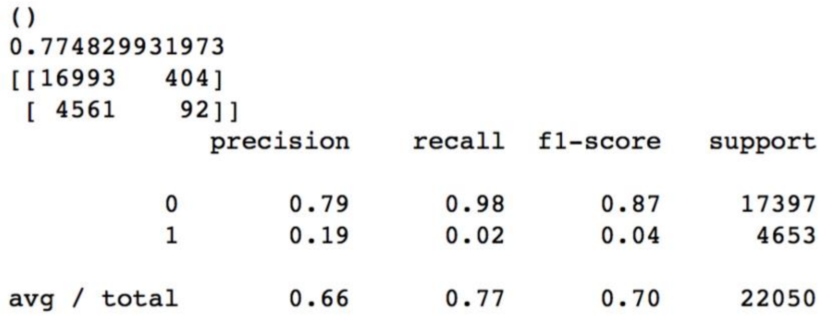
The ROC curve for this optimal Model:

The AUC score is 0.57 and the ROC curve shows that it is closer to the 45-degree diagonal of the ROC space, it means the less accurate the test. Even though the average accuracy (0.98) is relatively high, due to the imbalanced data (most of the crime values in Sexual crime column in normalized dataset are 0), and also the low correlation between the factors with sexual crime type, when considering the performance as false positive rate and true positive rate, it is less accurate.

### 7) Fraud



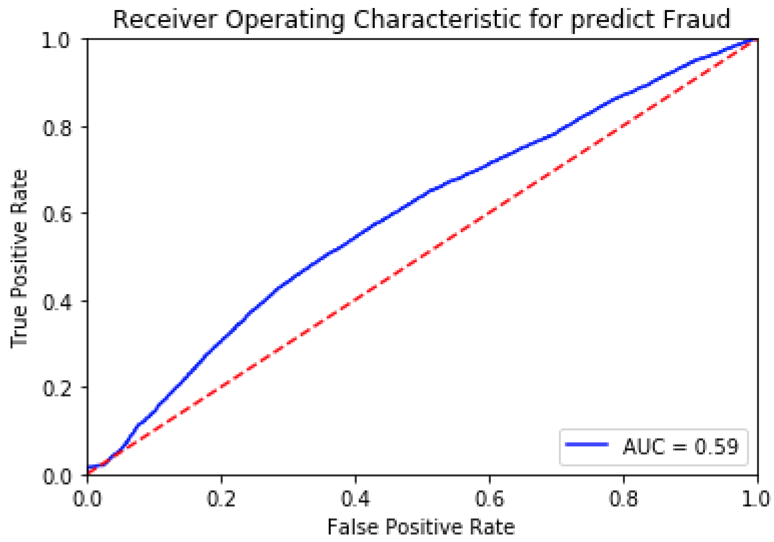
```
the best model for this case is NB
```



```
()
0.774829931973
[[16993   404]
 [ 4561    92]]
             precision    recall  f1-score   support

          0       0.79      0.98      0.87     17397
          1       0.19      0.02      0.04      4653

avg / total       0.66      0.77      0.70     22050
```



According to these four models which have been performed for predict frequency for Fraud crime, the optimal model in this project is Navie Bayes. After testing the Naive Bayes modeling, the result is below:

Accuracy is: 0.77, and the result shows below:

Based on the precision table, it shows that even though overall accuracy is 0.77, However, for ‘0’ class, naive bayes model obtained 79 % accuracy, however, for the ‘1’ class, it attained 19% accuracy. In this way, this modeling tends to perform well only for high frequency prediction.

The ROC curve for this optimal Model:

The AUC score is 0.59 and the ROC curve shows that it is closer to the 45-degree diagonal of the ROC space, it means the less accurate the test. Even though the average accuracy (0.77) is relatively high, when considering the performance as false positive rate and true positive rate, it is less accurate.

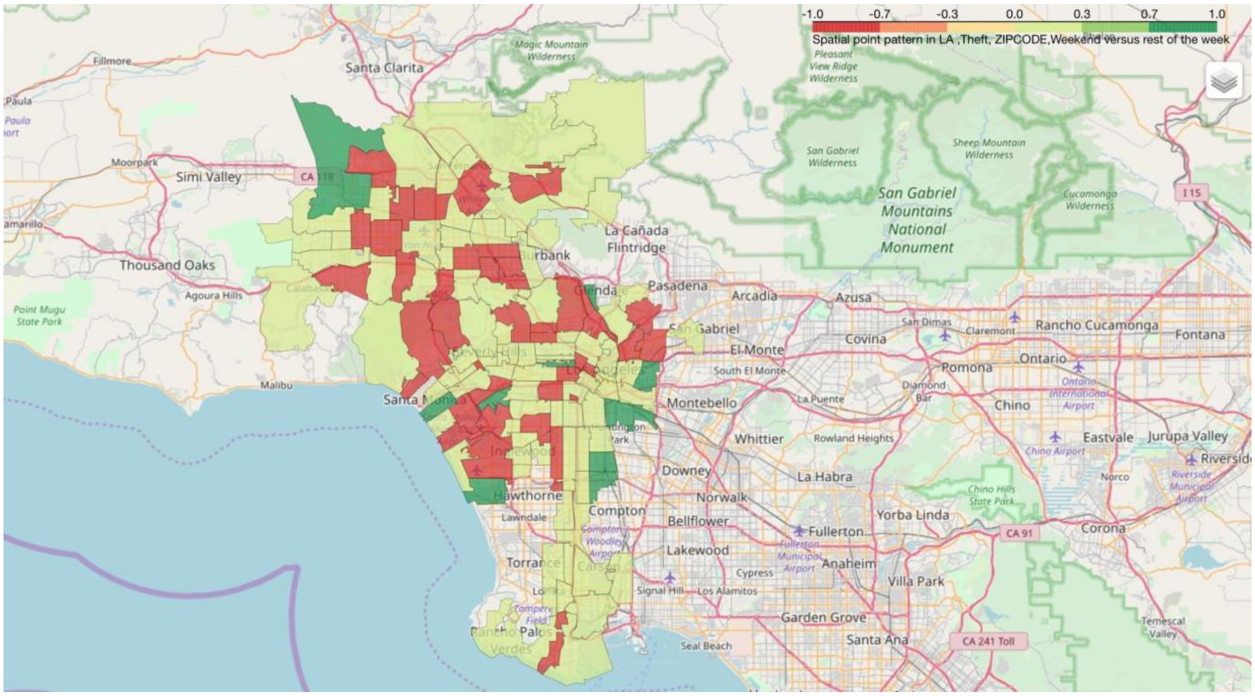
In conclusion, considering about performance both of high frequency and low frequency, this predictive modeling with these 7 features tend to perform well only for Robbery crime, Assault crime and Burglary crime which are three most frequent crime types. For other crime types which occurs more rarely, the performance of this predictive modeling only can predict low frequency.

# Additional Analysis

1. Spatial temporal analysis with Monte Carlo Approach (This analysis process was in Hypothesis test)
Result visualized in the map:

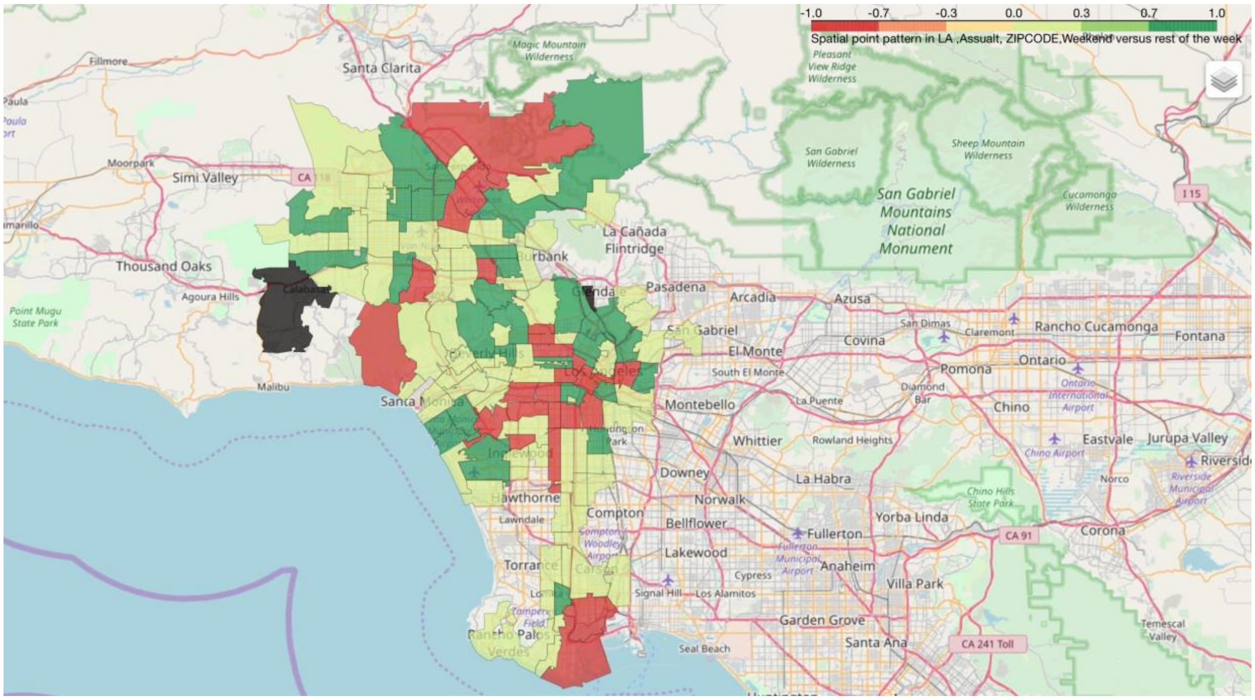
From this map, the dark green shows that the occurrence of Theft crime in Weekend is significantly higher than the corresponding in the rest of week. The red shows that the occurrence of Theft crime in Weekend is significantly higher than the corresponding in the rest of week. the light green shows that there is no significant difference between weekend and the rest of days in the week. In this way, from this graph, compared to the graph 9, even though the distribution of theft type in different days shows that the theft crime rates in the weekday is generally higher than weekday, the results is different when considering spatial pattern. For example, statistically, in the spatial pattern around Santa Monica area which is the recreation parking area, the theft crime rate is significantly higher on weekend than the crime on weekday.

Graph Theft crime spatial-temporal pattern in LA (Weekend versus rest of the week)



From this map, the dark green shows that the occurrence of Assault crime in Weekend is significantly higher than the corresponding in the rest of week. The red shows that the occurrence of Assault crime in Weekend is significantly higher  than the corresponding in the rest of week. the light green shows that there is no significant difference between weekend and the rest of days in the week. In this way, from this graph, compared to the graph 9, even though the distribution of Assault type in different days shows that the Assault crime rates in the weekday is generally lower than weekday, the results is different when considering spatial pattern. For example, statistically, even though the numbers of dark green units are more than numbers of red green units, compared to the graph 9, there are still some large area with the opposite results that is red area, it means some specific area with the lower Assault crime on weekend.

Graph: Assault Crime Spatial-temporal Pattern in LA (Weekend versus rest of the week)

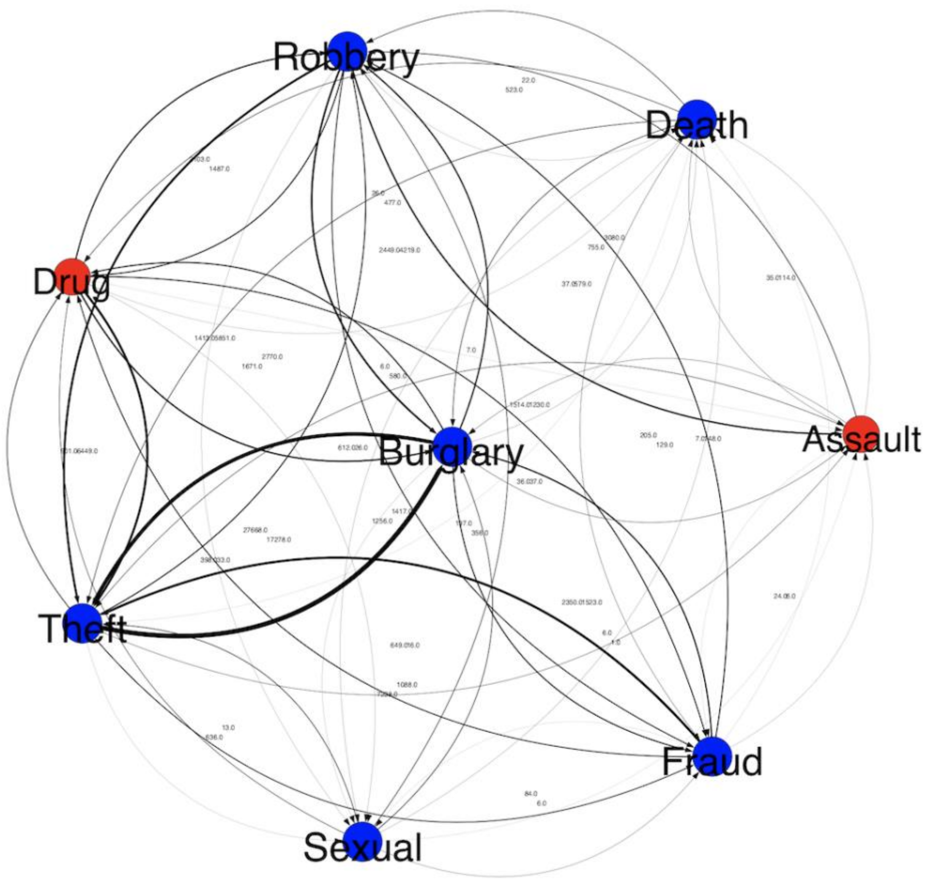


Conclusion:
Based on Spatial Temporal Analysis, the results clearly show that for the crime types under analysis there are distinctive temporal and spatial patterns for different days of the week. In this case, the predictive modeling in this project was performed with spatial temporal pattern, the spatial unit is zip code, and the time unit is day.

2. Network Analysis:

In the exploratory analysis, the scatterplot of different crime types are plotted, and there seem to have correlations between certain crime types. Thus, a network analysis of different crime types is conducted to further investigate the relationship among different crime types.

The two most crime types happened for each day and each zip code are selected, and the frequency of the same pattern for each pair are computed and treated as weight, which will be used when plotting the network graph. Both a 2D and 3D network graph are plotted, the 2D graph is like the following:



The following are the summarizations of this network:

- Nodes Labels:
  - Assault, Burglary, Death, Fraud, Robbery, Sexual, Theft, Drug
- Number of Nodes: 8
- Number of Edges: 55
- Density: 0.982143
- Degree of Each Node:
  - 13, 14, 14, 14, 14, 14, 14, 13
- KNN:
  - 76.935707, 4.612677, 152.915874, 20.746856, 14.095713, 157.406189, 3.840974, 19.166044
- Eigenvector Centralities:
  - 0.036921, 1.0, 0.002108, 0.115819, 0.168534, 0.001483, 0.824035, 0.116584
- Betweenness:
  - 25.0, 0.0, 6.0, 8.0, 0.0, 25.0, 0.0, 0.0
- The Local Transitivity (Clustering Coefficient):
  - 0.733503, 1.050825, 0.636321, 0.817849, 0.873051, 0.995601, 0.971484, 0.989918

The density is the fraction of actual connections (edges) divided by all possible connections (edges) in the network (graph). Since, the density for this network is really, it means that almost all different types of crime have been paired up and be the two most frequent crime type in a specific day.

The degree is the number of neighbors an individual node has. In this case, only Assault and Drug has degree of 13, other crimes all have degree of 14, which means the co-occurrence of Assault and Drug compared to other crime types is lower.

The edge weights are the strength of the relationship between two nodes, which are the numbers one the lines in the network graph. The thicker the line, the heavier the weights are. Thus, from the above graph, crime like Burglary and Theft are linked with heavy lines, which indicate that there exists a strong relationship between the two crimes. It may also infer that the two crime type are similar in some way.

The KNN will take weights into consideration and calculate the average degree of the neighbors for each vertex. And Death and Sexual crime have the highest KNN, which means that the two crimes are less weighted and do not occur as often as other crimes. Thus, the network can be clustered into two groups, with Death and Sexual as one group and the rest crime are in the other group.

The eigenvector centralities calculate the centralities of the vertices in a graph. From the results above, Burglary and Theft, which has the highest centralities, could have a considerable influence within the network by virtue of their control over information passing between others.

The local transitivity or clustering coefficient calculates the probability that two neighbors of a vertex are connected; in this case, this probability is calculated separately for each vertex. From the output above, it is more likely that Burglary, Robbery, Theft and Drug are more likely to be clustered together.

# Conclusion and Limitation:

## Conclusion:

Under EDA analysis, Hypothesis testing as well as Association Rule Mining, there are some significant relationships: Days and crime types, Education level and crime types, Obesity rate and crime types, Smoking rate and crime types, New moon and crime types, Rainy Weather and crime types. However, for rainy weather, education levels, new moon, these two factors are not significantly correlated with all crime types. For example, for New moon, it is only correlated with Assault and Robbery crime types.

In order to better investigate the potential relationship among different crime types. A network analysis is conducted, which considers the top two crimes happened in one specific day for each zip code. The results show that there seem to exist two clusters for the 8 crime types. Specifically, the crimes can be clustered into: Assault and Drug, which have relatively lower degree; and the rest of the crime types, which have relatively higher degree. This implies that the chance for Assault and Drug crime to be the two most frequent crime is lower than other combinations, which could also indicate that Assault and Drug will associate with other crime types more often individually. And from the network graph, Theft and Burglary also have a strong correlation between each other, which are connected by thicker edges. The thick edges connecting the two crime means that weight of these network is much heavier and the possibility for the two crime to happen together is much higher than other crime combinations. This is because Theft and Burglary are the two most common types of crime being seen in daily life, require similar conditions for committing the crime, and they share similar traits and have similar patterns during the week. Both tend to happen less frequently during the weekends, and more during weekdays.

Additionally, spatial temporal pattern results are crucial in the context of situational crime prevention. This project applied the Monte Carlo simulation method to investigate spatial temporal pattern for different types in days in the week. The results show that there are distinctive temporal and spatial patterns for different days of the week. As such, for the predictive modeling, it was based on zip code units and considered day and month as features in order to spatially predict different types of crime in day.
In this case, the predictive analysis modeling are with the features: Day, Month, Moon phase, Moon illumination, Education level, Rainy Weather, Max Temperature, and the predictive modeling in this project was performed for different types of crime to predict the frequency level of specific crime type (If this specific crime type would occur as the most frequent crime type in this space time pattern or not). After model selection and evaluation, the predictive models for all different crime types are with more than 60% accuracy. However, considering about performance both of high frequency and low frequency, this predictive modeling with these 7 features tend to perform well only for Robbery crime, Assault crime and Burglary crime which are three most frequent crime types. For other crime types which occurs more rarely, the performance of this predictive modeling only can predict low frequency.

In this case, these features tend to predict well for future prediction in specific crime types. To know if the specific crime type would be highly frequent events in specific day and particular location would facilitate crime department to make decision for crime prevention. Furthermore, for people who wants to travel to some particular places, this comprehensive prediction modeling is able to help them to know which type it would most likely to occur in these days. In a word, for crime, which is related to a variety of factors, and also it varies in different crime types, time and location. Therefore, it is really important to take space-temporal and different crime types into consideration. Consequently, accurate prediction based on location, time and crime type, as the predictive modeling in this project, is really critical for the success of any crime prevention initiative.

## Limitation and future work:

Data limitation:

In this project, there are more than 10 big cities in the US which have been considered for crime analysis. Unlike demographics and transportation data, most of the features used for predictive modeling in this project are utilized in this project are not city-related and policy-related. In this way, this project has limited ability to be a reference for policy-making department to see which policy promotes the crime or improves safety. In this case, other datasets are related to crime, such as demographics data, transportation data, are supposed to be utilized in future analysis.

Model limitation:

From the predictive modeling, there is lack of analysis for investigating the relationship between these features, in other words, the correlation between each feature should be considered in future for improving the accuracy. Furthermore, the problem of imbalanced labels for output feature has not been solved in this project, in this case, the method for dealing with imbalanced labels is supposed to apply for avoiding overfitting.

Moreover, the spatial temporal pattern analysis was performed only for LA city, for both of location and time are extremely crucial for crime, the spatial temporal analysis need to be extended to various cities.  Also, some other factors distributed by space and location are supposed to be considered for spatial temporal analysis.

For the relationship between education/health and crime, it is hard to conclude an overall pattern which could perfectly fit for all cities. Since each city may have its own characteristics, and in order to come up with a proper model, all factors that could make an effect should be taken into consideration. But, at the same time it is hard to consider and unify all the factors. Thus, a more comprehensive method should be considered.

# Link to Website:

https://lemonning0713.wixsite.com/website

# References

Cung, B. (n.d.). Crime and Demographics: An Analysis of LAPD Crime Data. Retrieved December 5, 2018, from https://escholarship.org/uc/item/2v76v571

Fischer, C. (2010, June 15). A crime puzzle: Violent crime declines in America. Retrieved December 5, 2018, from http://blogs.berkeley.edu/2010/06/16/a-crime-puzzle-violent-crime-declines-in-america/

Global STUDY on Homicide 2013. (n.d.). Retrieved December 5, 2018, from https://www.unodc.org/documents/gsh/pdfs/2014_GLOBAL_HOMICIDE_BOOK_web.pdf

Intra-week spatial-temporal patterns of crime. https://crimesciencejournal.springeropen.com/articles/10.1186/s40163-015-0024-7