**Basic Statistical Analysis and data cleaning insight**

1. The results of basic statistical analysis (by running statistical\_analysis.py):

For some categorical attributes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SHIFT | METHOD | OFFENSE | NEIGHBORHOOD\_CLUSTER |
| mode | EVENING | OTHERS | THEFT/OTHER | Cluster 2 |
| unique | 3 | 3 | 9 | 39 |
| frequency | 0.4265 | 0.9304 | 0.4361 | 0.0816 |

|  |  |  |
| --- | --- | --- |
|  | BLOCK | VOTING\_PRECINCT |
| mode | 3100 - 3299 BLOCK OF 14TH STREET NW | Precinct 129 |
| unique | 447 | 143 |
| frequency | 0.0265 | 0.0427 |

For some numeric attributes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PSA | CENSUS\_TRACT | LATITUDE | LONGITUDE |
| mean | 373.48 | 6259.70 | 38.906884 | -77.007081 |
| median | 401 | 3600 | 38.906430 | -77.011367 |
| std | 194.17 | 3126.13 | 0.0298 | 0.036469 |

“SHIFT” is when the crime happens (day, night or midnight). We can see that crimes are more likely happen in the evening than in the daytime

“METHOD” is how the crime is committed (with a gun/knife or other way). From the above results we can conclude that most of reported crimes do not involve with weapons.

“OFFENSE” refers to crime offenses which includes many types of crime. But from the results most of crimes belong to a broad inclusion of Theft offenses including embezzlement, theft of services and fraud/false pretenses

The mode of attributes “NEIGHBORHOOD\_CLUSTER”, “BLOCK” and “VOTING\_PRECINCT” shows that where happens the most crimes.

The numeric attributes “LATITUDE”, “LONGITUDE” also shows where a crime happens. The mean or median may shows the center of the DC since crime happens everywhere.

Though PSA is a numeric attribute, the number is only a code. The mean, median or std doesn’t make much sense.

1. The strategy for handling missing values

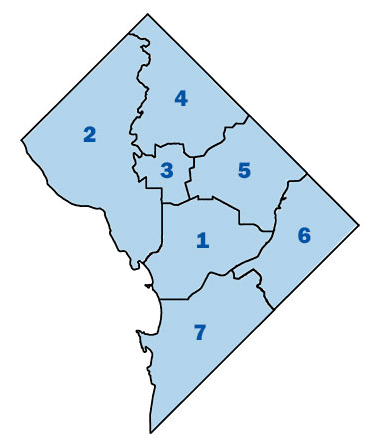
The attributes “DISTRICT” and “PSA” have a few missing values

Though they are numeric attributes, the number is only a code. So we use it’s mode to replace the missing values.

1. The binning strategy:

Since we may explore whether the location has a relationship with crime rate and PSA represents the locations well, we decided to bin the attribute “PSA”

For the attribute “PSA”, we use pre-binning strategy: {100, 200, 300, 400, 500, 600, 700, 800}. This is because the code between 100 and 200 belongs to the first district, the code between 200 and 300 belongs to the second district… and there are 7 districts in total:



1. Other strategies for data pre-processing:  
   a) drop the attributes which is not helpful for our further questions

Since the object\_id is a unique code for each observation and it is not helpful for us to do the association rule mining or machine learning. We decided to drop the attribute.

b) encode some nominal attributes or convert them to numeric attributes

For attribute “OFFENSE”, we encode this using number 1-9, where the bigger number indicates the higher level of crime.

And for some other attributes such as “SHIFT” or “METHOD”, we map them to do the following machine learnings.

**Association Rules/ Frequent Itemset Mining Analysis**

running association\_rules.py on dataset crime2017

Let the minimum confidence be 0.5 and set different minimum support

1. Minimum support = 0.2

|  |  |  |
| --- | --- | --- |
| patterns | support | confidence |
| {'DAY'}→{'OTHERS'} | 0.353 | 0.965 |
| {'EVENING'}→{'OTHERS'} | 0.400 | 0.939 |
| {'THEFT F/AUTO'}→{'OTHERS'} | 0.317 | 1.0 |
| {'THEFT/OTHER'}→{'OTHERS'} | 0.436 | 0.999 |
| {'THEFT/OTHER', 'EVENING'}→{'OTHERS'} | 0.221 | 1.0 |
| {'EVENING'}→{'THEFT/OTHER'} | 0.221 | 0.506 |

1. Minimum support = 0.3

|  |  |  |
| --- | --- | --- |
| patterns | support | confidence |
| {'DAY'}→{'OTHERS'} | 0.353 | 0.965 |
| {'EVENING'}→{'OTHERS'} | 0.400 | 0.939 |
| {'THEFT F/AUTO'}→{'OTHERS'} | 0.317 | 1.0 |
| {'THEFT/OTHER'}→{'OTHERS'} | 0.436 | 0.999 |

1. Minimum support = 0.4

|  |  |  |
| --- | --- | --- |
| patterns | support | confidence |
| {'EVENING'}→{'OTHERS'} | 0.400 | 0.939 |
| {'THEFT/OTHER'}→{'OTHERS'} | 0.436 | 0.999 |

The most frequent patterns are {'EVENING'}→{'OTHERS'} and {'THEFT/OTHER'}→{'OTHERS'}.

This is not surprising at all. Because through the basic statistical analysis, we can see that the frequency of value “OTHERS” is up to 0.93 for the attribute “METHOD”, which means it has a high chance of being included in the most frequent patterns.

These patterns do make sense: One shows us that most of crimes that happens in the evening do not involve with dangerous weapons like guns or knives. And the other one shows us that almost every theft, embezzlement, theft of services or fraud/false pretenses doesn’t involve with weapons.