Chicago Crime

An Analysis of Chicago Crime Trends

Graham Dominick
Computer Science
University of Colorado Boulder
Boulder, CO USA
Graham.Dominick@colorado.edu

Isaac Hammes
Computer Science
University of Colorado Boulder
Boulder, CO USA
Isaac.Hames@colorado.edu

Amanda Killeen
Computer Science
University of Colorado Boulder
Boulder, CO USA
Amanda.Killeen@colorado.edu

Daniel Kingsley
Computer Science
University of Colorado Boulder
Boulder, CO USA
Daniel.Kingsley@colorado.edu

ABSTRACT

In this data mining effort we attempted to answer a number of interesting questions about Chicago crime. One interesting question we set out to answer was “Is there a correlation between type of crime and location description or district?” We approached answering this question using a partial data cube and viewing a two dimensional cuboid. Our results showed a concentration of “Battery” accounts in districts 2 through 11. We also saw a concentration of “Theft” accounts in districts 1, 18 and 19. These results could be used by law enforcement or researchers to address this crime. In addition to identifying districts with higher rates of battery and theft, we also sought to develop an arrest prediction model using classification techniques and we were able to create an arrest prediction model with using a Decision Tree Model with ~85% accuracy Finally through the application of the association rule, we were able to determine the frequent itemset {Residential, Not Domestic, Property Crime} was the least likely to result in arrest.

INTRODUCTION

The United States is currently enjoying a period of all time low crime rates, according to the New York University Law School. The number of crimes per 100,000 people in the US’s 30 largest cities has dropped from ~9,300 in 1990 to less than 4,000 in 2016, and violent crime has dropped from ~730 to ~386 instances per 100,000.[5] Despite this, Chicago stands out as a city that has experienced a disturbing rise in crime, particularly violent crime, in recent years, with a rate of over 900 violent crimes per 100,000 people. A 2016 article in Time Magazine showed how Chicago responsible for for almost half the increase in the national murder rate in 2016 (13%).

Chicago’s spike in crime has received a lot of publicity, and has made the city somewhat of an example city for those interested in talking about or studying crime rates. In January 2017 President Trump tweeted about Chicago crime and suggested that he may “send in the feds” in response to a 24% increase in killings from 2016. In this project, we set out to answer the following questions:

- Is there a correlation between type of crime and location description or district?
- Do arrests occur more often in different locations for the same crime?
- What are the most common crimes for different areas?
- Are there periods when there were waves of certain types of crimes?
- Trends in crime types?
- Can an accurate model be generated for “arrest made” based on other information about a crime?

Our goal for this project was to use data mining techniques on a dataset of Chicago crime information in order to uncover novel trends about crime in the city such as crime waves, trends, and clusters, through patterns and correlations between crime, location, and time. We hoped that the trends we uncovered could potentially help understand the factors causing crime in Chicago, provide insight leading to possible solutions, and provide law enforcement with predictive tools.

These questions are important because crime trends in any city can be difficult to interpret, especially in such a large one like Chicago, and it’s unclear whether tougher policing would have more positive effects than other potential strategies, like investing in social services such as education, mental health, and drug rehabilitation. An article published by the Atlantic in 2017 discussed how the causes are essentially unknown, despite lots of research.

RELATED WORK

When we began work on this project we explored various articles, notebooks and repositories to understand the existing work that has been conducted on the topic of Chicago Crime using the Chicago Crime Portal data.

The first existing work we evaluated was from the City of Chicago itself and their Crimes - 2001- to Present Dashboard. This dashboard is a comprehensive resource of 13 visualizations covering the entire data set and has an emphasis on the locality of reported crimes, which is a topic we are planning to explore in our research. Their geographical visualizations however tended to implement a regional visualization technique and we saw an opportunity to explore other geographic visualization methods, such as a heatmap or point map to better express the volume of crimes and where the literal hotspots are.

In addition to the City of Chicago’s visualizations, we explored Kaggle**,** an online machine learning and data science community, where we found several projects on the use of the Chicago crime dataset that we are utilizing in our analysis. We mainly focused on the work conducted by Djona Fegmen, in his Chicago Crime Data Analysis Jupyter notebook. Fegmen uses Python along with the pandas, seaborn and matplotlib.pyplot libraries to create a number of visualizations on overall counts of primary crime type, arrest volume over time, top 5 crimes and trends over time, and volume and trends in domestic violence arrests.

On Github we found a repository by Tejal Behangle, where she conducted a similar project at the Missouri University of Science and Technology, but her project attempted to predict the location and time of day where crimes and when crimes were likely to be committed. Using both the Chicago crimes dataset and a second dataset of socioeconomic data to explore the impact of socioeconomic factors (such as literacy rate and employment) on the rate of crime occurrence.

Finally, we reviewed an article published on Medium.com by Sadaf Tafazoli. In his article, Tafazoli explains how following preprocessing, he used Spark SQL to explore the data and answer questions. This article also sought to answer multiple questions and created visualizations to support their questions, but did not have a cohesive or particularly interesting outcome or story. It was simply a lot of count queries that were later graphed in some manner.

While research has been conducted in the area of Chicago crime before, our work differs from the previous work surveyed in that our data mining focused on the type of crime, frequency, location, and whether or not arrest occurred to create a predictive arrest model. We also did not take into account any socioeconomic factors as previously been done.

DATA SET

The data set which we are using for our project comes from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2, which has been made available by the Chicago Police Department. It is updated daily and includes crimes that are reported in Chicago, IL from 2001 to present, minus the last 7 days. In the case that the crime was a murder, it includes data for each victim. There are 7.08 million instances and 22 attributes for the data.

Here is a list of all of the attributes:

- ID (Number) - Unique key for tuple
- Case Number (Text) - Unique number for the case
- Date (Floating Timestamp) - Date and time when the crime occured.
- Block (Text) - Partial address of crime location.
- IUCR (Text) - Illinois Uniform Crime Reporting Code which is based on crime type and description.
- Primary Type (Text) - Primary description from IUCR code.
- Description (Text) - Secondary description of the IUCR code.
- Location Description (Text) - Description of the location where the crime occured.
- Arrest (Boolean) - Whether an arrest was made for the crime or not.
- Domestic (Boolean) - Whether the crime was domestic or not.
- Beat (Text) - Beat where crime occured. A beat is the smallest geographic area that police coverage is divided into.
- District (Text) - Police district where crime occured.
- Ward (Number) - Ward where crime occured.
- Community Area (Text) - Community area where the crime occured.
- FBI Code (Text) - FBI's crime classification.
- X coordinate (Number) - X coordinate where crime occured.
- Y coordinate (Number) - Y coordinate where crime occured.
- Year (Number) - Year crime occured.
- Updated On (Floating Timestamp) - Date and time record was updated.
- Latitude (Number) - Latitude where crime occured.
- Longitude (Number) - Longitude where crime occured.
- Location (Location) - Coordinates of where crime occured.

TECHNIQUES APPLIED

1 **Data collection**

Data was collected by the Chicago Police Department and the data set was made available

by the city of Chicago. Once the data was cleaned, it added to a Github repository for team use.

2 **Preprocessing**

Fortunately, the data set was fairly clean to begin with, however we needed to remove the “N/A” or Null fields and narrow the scope of attributes to be implemented in our final dataset and we utilized Pandas and Numpy libraries for the cleaning and pre-processing.

The location data required some additional processing, as the original data set included 10 geographical attributes:

- Block
- Beat
- District
- Ward
- Community Area
- X coordinate
- Y coordinate
- Latitude
- Longitude
- Location Description
- Location

We narrowed the scope to Block, Beat, and Location, this was partially due to a large volume of Null values in the fields of Ward and Community Area. The Location attribute also contained the Latitudinal and Longitudinal coordinates, enabling us to remove the duplicate fields of X Coordinate, Y Coordinate, Latitude, and Longitude.

In addition to narrowing the scope of geographical attributes, there was normalization required on the location description attribute. For example, classifying “Apartment”, “Basement” and “Coach House” as “Residential”.

After the location and geographical attributes were cleaned, we reviewed the different Crime Codes applied to the data set. Each crime has an IUCR (Illinois Uniform Crime Reporting Code) and an FBI code attribute. In reviewing the IUCR and FBI codes, we saw an opportunity to further classify the Primary Crime Types using the FBI crime type definitions of “Violent Crime”, “Property Crime”, and “Less Serious Offense”. These classifications allowed us to focus on specific segments of interest, such as Violent Crime.

Finally, the “Identity Theft” Primary Type attribute was dropped due to null values and seemed less relevant to our project, since we are mainly interested in violent crime.

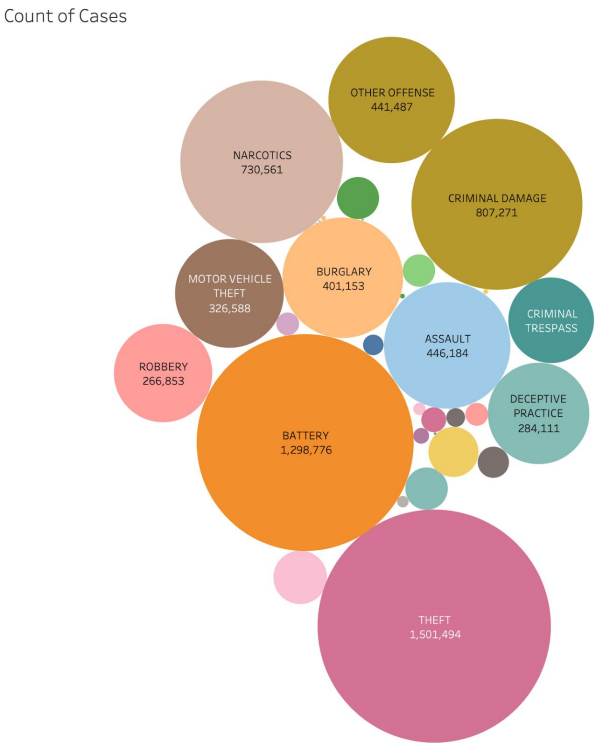
One area of obvious interest for this project was gang-related crime. While our dataset provided minimal information on the gang-affiliation of certain crimes, we hoped it could be inferred from the series in which crimes of a type occur in. For example, if a murder in a neighborhood known for gang activity is closely followed by another murder, it may likely be gang-related.

We looked for neighborhoods or “beats” with serial episodes of violent crimes, to identify areas with potentially higher violent gang activity. However, we were not able to derive a boolean “Likely gang related” or “Not” value with that process.

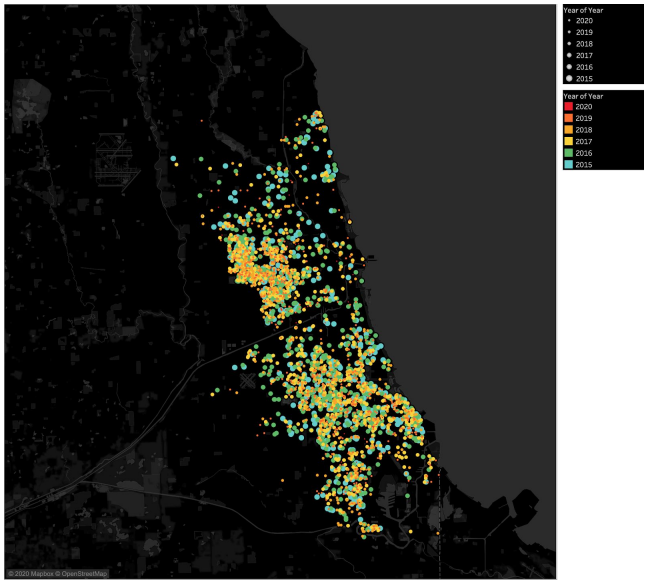
3 **Preliminary Results**

Following cleaning and preprocessing of data. We explored the data using data visualization tools, Tableau and MatPlotLib to identify any interesting visual patterns to explore furthering in our data analysis. While these are not necessarily reflective of our final results, they provide some interesting information that fed into our final analysis. Given the previously high rates of killings we started with a simple count of

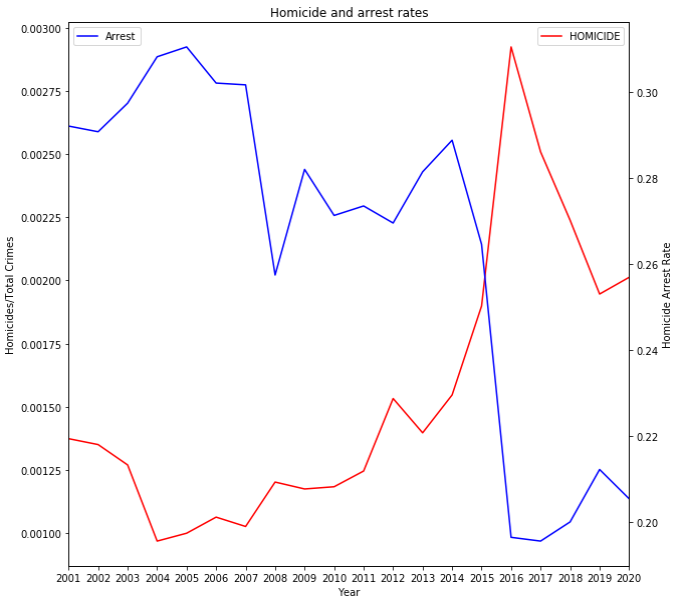
distinct cases and their Primary Type of crime. While we were expecting violent crime to have the highest volume, Theft actually was the top offender. As we dig into the data further, it will be interesting to see if there is any relationship between theft and violent crime.



Another interesting outcome of using visualization tools to do some initial data exploration was utilizing the geographical data to see if there were clusters of violent crime, specifically homicide, that were immediately obvious. Looking at the Primary Type - Homicide and the Years 2015-2020 inclusive, and stacking the cases on top of eachother, we can see there are not only geographical clusters where homicides occur, but areas where there are homicides appear year over year. The frequency of homicide can be derived from this visualization and provides a foundation at which researchers could explore further into, such as what are the characteristics of these geographical areas that would lead some to have a higher frequency of homicide than the others.



In our exploratory analysis we also noted that the rate of homicide, compared to the rate of other crimes has increased, noticeably over the 20 year period, while the rate of arrest for homicide has gone down significantly.



In our proposal, we discussed looking for ways to identify neighborhoods with gang activity by the pattern of homicides. We hypothesized that if homicides in a certain district tended to "clump" together, that might be a sign of revenge killings.

To measure the "clumpiness" of homicides, we borrowed the entropy measure from information theory. As was published in 2013 by a group of researchers from UPenn[7], entropy can be used to measure the expected amount of information in a time series of events. If there was a pattern of revenge killing in certain districts, the entropy metric might allow us to find it. It calculates the sum of the time intervals (between homicides) multiplied by the log of the time intervals.

We calculated this statistic for each of the 25 districts in the city, expecting to see higher entropy in e.g. district 11 which is known for gang activity. Unfortunately, this approach was ineffective because of the disproportionate amount of homicides in different districts, i.e. district 11 had over 1000 homicides in the 20 year period, compared to 68 in district 20. We thus determined it was not realistic to try to compare patterns of homicide between the districts.

4 **Data Mining Methods**

4.1 Association Rule Mining with the Apriori Algorithm: Utilized Pandas and Numpy libraries along with the mlxtend (Machine Learning extensions) library to search for interesting rules between the features. The features used were District, FBI Code (categories were reduced to lesser crimes, property crimes, and violent crimes), Domestic, Arrest, Location, and Year.

The pandas get_dummy function was used to convert the categorical data into indicator variables for every category. The Apriori algorithm was then run on the data and the rules were extracted into a table with the support, confidence, and lift for every rule found.
One of the common sense rules found is that if a crime is domestic, then it occured in a residential location. An interesting rule is that if a crime is a property crime, residential, and non domestic, there is a 96% chance that the perpetrator was not arrested. This seems to indicate that stealing or vandalizing something in a residential area (as long as the criminal is not stealing from his own family) is relatively safe and the lack of arrests could lead to an increase in property crime. This is a fairly significant increase in not being arrested when compared with the 89% of not being arrested for all property crimes. The results are shown in figure 1.2.

4.2 Decision Tree, Naive Bayes and Logistic Regression models: Utilized Sklearn and Pandas to predict the likelihood of arrest based on the District the crime was committed in, the FBI Code for crime, whether it was domestic or not, and the year it was committed to predict whether an arrest was made or not.

In order to process this data for modeling, we had to use the Pandas Get Dummies feature to convert each category in the model to a separate boolean feature. The size of the dataframe quickly became too large to use, so narrowed our scope to use only attributes mentioned above to help with predicting arrests.

Logistic Regression Results:

Prediction True	No Arrest	Arrest
No Arrest	1,517,066	11,885
Arrest	309,004	270,776

Accuracy: 84.7

Naive Bayes Results:

Prediction True	No Arrest	Arrest
No Arrest	1,519,913	9,038
Arrest	315,417	264,363

Accuracy: 84.6%

Decision Tree Results:

Prediction True	No Arrest	Arrest
No Arrest	1,496,610	33,721
Arrest	277,888	300,512

Accuracy: 85.2%

4.3 K-Nearest Neighbor: Attempted by unsuccessful due to volume of data and processing time required..

4.4 Partial data cube: The data cube method allows our data to be viewed in multiple dimensions. We used a partial data cube to reduce the drain on computing resources. The 2D cuboid helped us to look at slices of the data to inform further questions and we could make use of higher dimensions in the data to answer these second level questions.

KEY RESULTS

In the data visualization Figure 1.1 (Appendix) the x-axis is District and the y-axis is Primary Type of crime. The color represents the frequency of that type of crime in the corresponding district. A third variable, frequency of the type of crime per district, had to be derived for this result. The result facilitates answering the question, “Is there a correlation between type of crime and location description or district?”.

In this model we essentially drill down to a two dimensional cuboid and derive the frequency data value. This 2D cuboid allows us to build a visualization that could be used in discovery. As mentioned previously, the question we are trying to answer with this data mining method is “Is there a correlation between type of crime and location description or district?” The combination of these three attributes answers this question when visualized. We can see a concentration of “Battery” accounts in districts 2 through 11. We also see a concentration of “Theft” accounts in districts 1, 18 and 19.

Another key result that came out of our project was from the modeling of arrests. We modeled the arrests using Decision Tree, Naive Bayes, and Logistic Regression. The Decision Tree model was the most accurate, being able to predict arrests with an accuracy of 85.2%. The second best model was Logistic Regression with an accuracy of 84.7%. The third best model was the Naive Bayes model with an accuracy of 84.6%. It is possible that better models could be found using better methods for feature selection. The large numbers of categories in each of our features made memory a big issue when constructing the models.

APPLICATIONS

The knowledge that was mined from the dataset using the two dimensional cuboid that resulted in figure 1.1 in the appendix could assist law enforcement in tracking and taking action to counter these concentrations of crime types. It could also assist researchers in studying possible reasons for these trends and concentrations.

Not only does this 2D cuboid answer our initial question, “Is there a correlation between type of crime and location description or district?”, but it contributes to the development of new questions and targets for further mining efforts. One further question for example, “Is battery actually more prevalent in these areas per capita or is it a product of population?” In order to approach answering this secondary question we might attempt to move to the next level in the lattice and include population as another dimension. Then instead of deriving a frequency we could derive a ratio attribute that would give us frequency per capita. This would give us more detail and assist in the evolution of the mining process.

A possible application of an arrest model could be to look at what circumstances of a crime are likely to lead to an arrest and which are not likely to lead to an arrest. An investigation of why locations leave many crimes being committed with few arrests, while others do not and look into ways to reduce the number of crimes committed with the criminal escaping punishment. Another, almost opposite application, could be to improve the efficiency of police resources by spending less time on crimes that have a very low chance of ever leading to an arrest. After a crime is reported, the information from the crime could be put into the model, and based on the prediction of an arrest being likely or not, the police could allocate more resources to a crime that is likely to lead to an arrest and less resources on a crime that is most likely going to lead nowhere.

REFERENCES

[1] Djona Fegnem. 2017. Chicago Crime Analysis. Retrieved March 10, 2020 from https://www.kaggle.com/djonafegnem/chicago-crime-data-analysis

[2] Sadaf Tafazoli. 2018. My notes on Chicago Crime data analysis. Retrieved March 10, 2020 from https://medium.com/@stafa002/my-notes-on-chicago-crime-data-analysis-ed66915dbb20

[3] Chicago Crime Portal. 2020. Crimes - 2001 to present - Dashboard. Retrieved March 10, 2020 from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g

[4] Tejal Bhangale. 2017. Chicago Crime Analysis. Retrieved March 10, 2020 from https://github.com/Tbhangale/Chicago-Crime-Analysis

[5] FBI Crime in the United States Report 1995-2019 https://www.fbi.gov/services/cjis/ucr/

[6] Josh Sanburn. 2016. Chicago is responsible for almost half of the increase in U.S. homicides. Time Magazine online https://time.com/4497814/chicago-murder-rate-u-s-crime/

[7] Zhang, Y., Bradlow, E. T., & Small, D. S. (2013). New Measures of Clumpiness for Incidence Data. Journal of Applied Statistics, 40 (11), 2533-2548. http://dx.doi.org/10.1080/02664763.2013.818627

APPENDIX

Figure 1.1

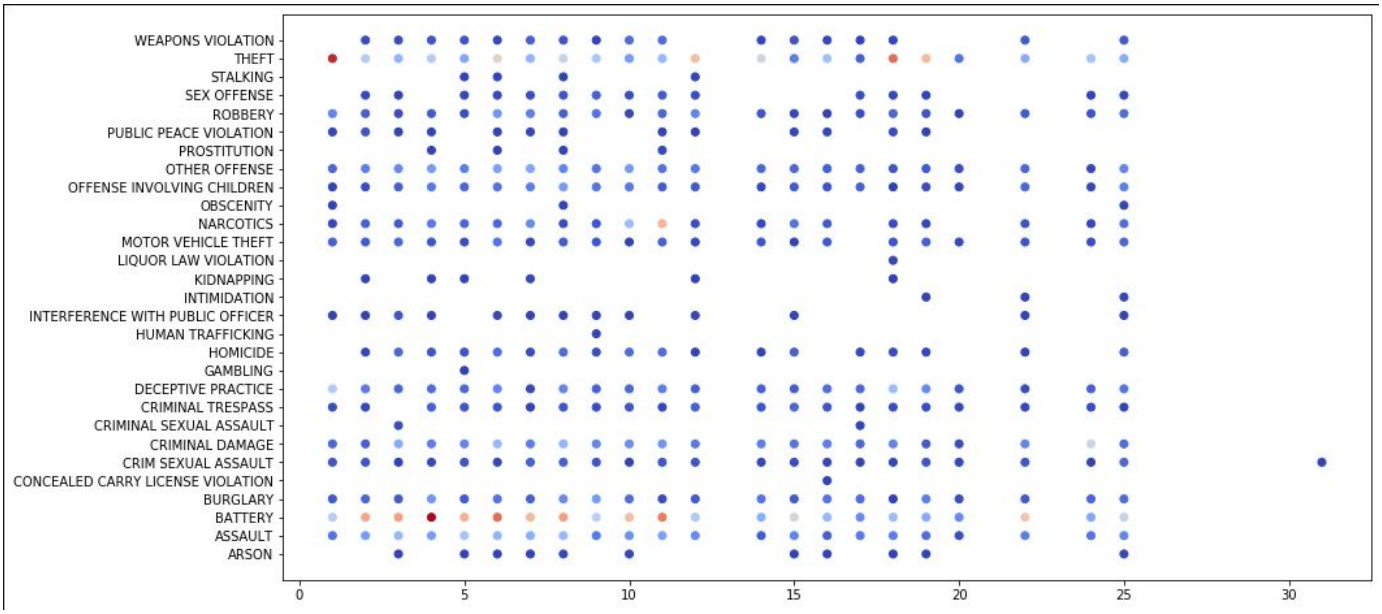


Figure 1.2

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(8, Property crime)	(No Arrest)	0.022534	0.725336	0.020001	0.887596	1.223703
(Not Domestic, Property crime)	(No Arrest)	0.309823	0.725336	0.277516	0.895725	1.234910
(Property crime)	(No Arrest)	0.315753	0.725336	0.282956	0.896133	1.235473
(OTHER)	(No Arrest, Not Domestic)	0.037795	0.618032	0.029156	0.771439	1.248218
(STREET, Property crime)	(No Arrest)	0.093959	0.725336	0.089434	0.951837	1.312271
(STREET, Not Domestic, Property crime)	(No Arrest)	0.093264	0.725336	0.088796	0.952102	1.312636
(RESIDENTIAL, Property crime)	(No Arrest)	0.091413	0.725336	0.087673	0.959085	1.322263
(RESIDENTIAL, Not Domestic, Property crime)	(No Arrest)	0.087016	0.725336	0.083629	0.961067	1.324996
(Property crime)	(No Arrest, Not Domestic)	0.315753	0.618032	0.277516	0.878904	1.422100
(RESIDENTIAL, Property crime)	(No Arrest, Not Domestic)	0.091413	0.618032	0.083629	0.914841	1.480248