PREDICTIVE CRIME DATA ANALYSIS

A study on associating crime type with the days in Chicago

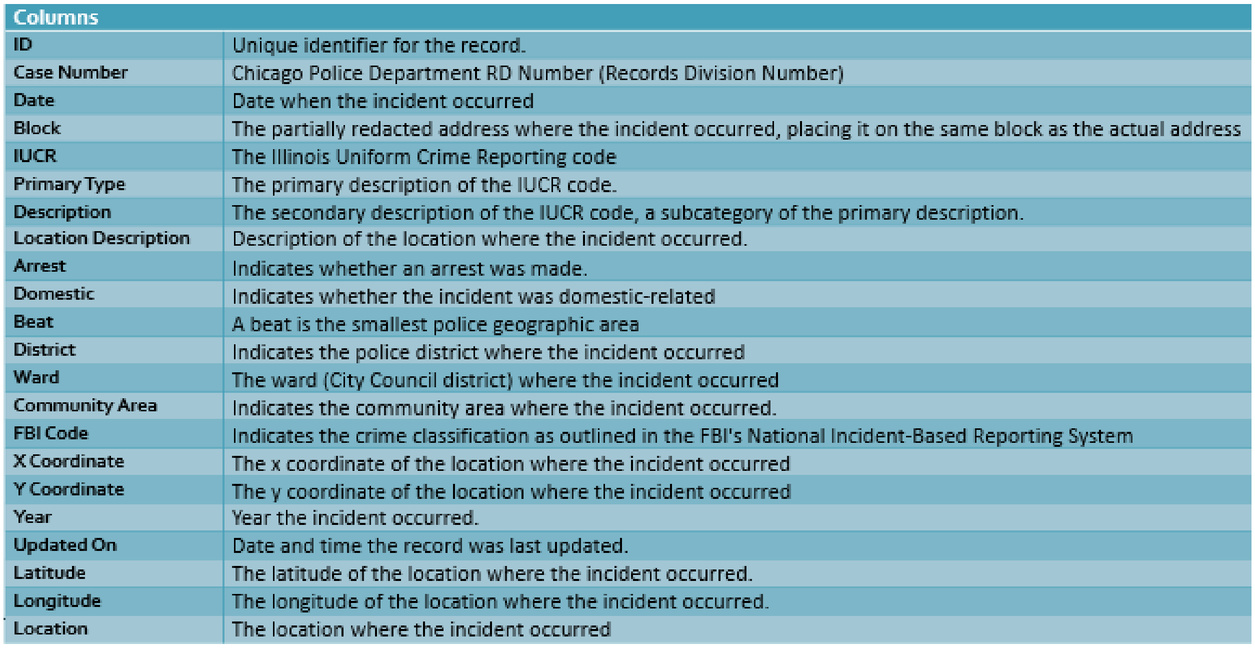
GERALD WIJAYA

SPRINGBOARD CAPSTONE 1

The Chicago Data Portal

Around the world, there are open source datasets provided by the government on each sector a crime has been reported. In the United States, cities like Chicago, San Francisco and New York, have their datasets on the city data portal to share these data for individuals to analyze them for their needs and to create a public awareness. The primary idea behind this project is to achieve the predictive analysis of a large crime dataset from the Chicago Data Portal[1]. The dataset, exported to csv format, includes every reported crime archived in case number from the year 2001 to present. Further specification on the file will be derived using iPython Notebook.

When identifying on the features on the columns of the dataset, it can be observed that each report explains the type and description of the crime with the year and time it occurred. Furthermore, each crime report also explores into the detail of the geo location for each crime. The crime dataset generally already has all the respective variables to build a predictive model for future crime. However, no association has been put into detail for which crime type is the most on which day.



Columns	
ID	Unique identifier for the record.
Case Number	Chicago Police Department RD Number (Records Division Number)
Date	Date when the incident occurred
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address
IUCR	The Illinois Uniform Crime Reporting code
Primary Type	The primary description of the IUCR code.
Description	The secondary description of the IUCR code, a subcategory of the primary description.
Location Description	Description of the location where the incident occurred.
Arrest	Indicates whether an arrest was made.
Domestic	Indicates whether the incident was domestic-related
Beat	A beat is the smallest police geographic area
District	Indicates the police district where the incident occurred
Ward	The ward (City Council district) where the incident occurred
Community Area	Indicates the community area where the incident occurred.
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System
X Coordinate	The x coordinate of the location where the incident occurred
Y Coordinate	The y coordinate of the location where the incident occurred
Year	Year the incident occurred.
Updated On	Date and time the record was last updated.
Latitude	The latitude of the location where the incident occurred.
Longitude	The longitude of the location where the incident occurred.
Location	The location where the incident occurred

[1] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g

The Problem Statement

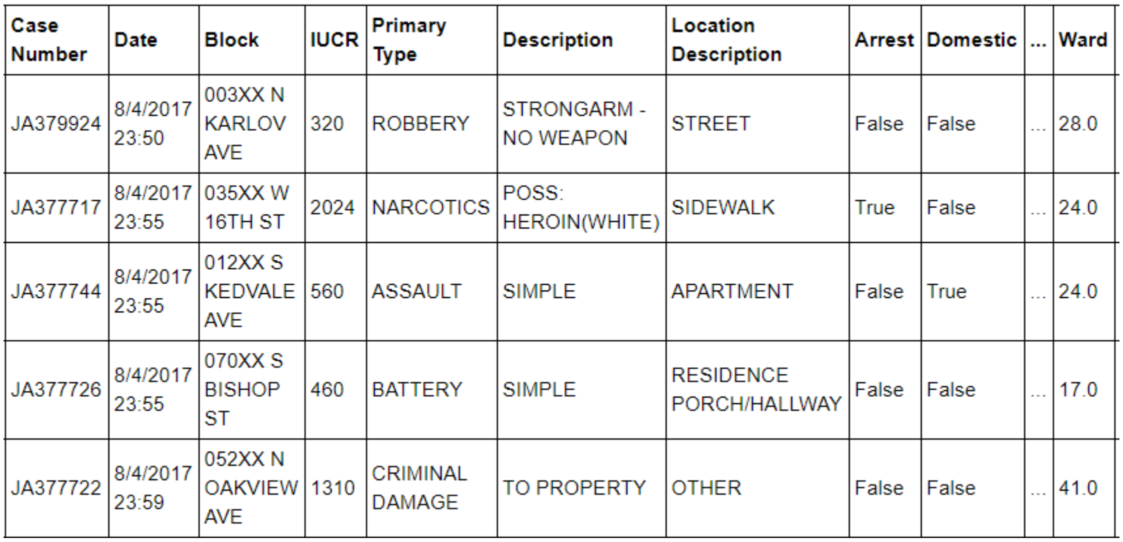
A Predictive Crime Analysis is what an ideal world would need, and since a human is unable to handle the multiple variable that could cause crime from happening, this technology as portrayed in the movie *Minority Report,* is in the brink of becoming a reality. Given the collection of data over the years, the police department can build a crime models based on the proximity of the crime, the location and the type of crime. Based on this analysis, the system could potentially round down the upcoming crime and its threat level, which gives a preemptive measure police could react to in any situation. The main objective of the project will be to create associate properties between the days of the week to the crime type for five years of dataset provided from the Chicago Data Portal. The overall purpose of this analysis is to be able to give residents pre-emptive alerts and advices on areas they can avoid or preventive measures they can act on to avoid the crime from happening. This statement usually goes by saying that crime cannot be eliminated, but it can be prevented. Thus, the conjecture at the end of this assignment can be used to assist policing strategies and forecast unusual movement in crime before the time of occurrence.

For this data story, a walkthrough will be made on how to tackle the major concerns of this system; based on how statistically consistent and accurate the predictive model is. Is there a chance that the model targets the wrong and potentially innocent people as a threat? Will the system be a biased profiling to the public based on the provided information? Lastly, how can we in turn use this technology properly and not have the tragedy like in *Minority Report* from happening? With these data information, models can be charted to visualize a trend. The problem with the open source crime data is the lack of community involvement in taking advantage of the possible analysis that can be done with it. Perhaps the data provided is not enough to create a predictive model, perhaps it is enough and yet probably sensitive approach to the public for a sort of social experiment. Here in this project, I would like to approach this problem differently; not in a way to predict a crime based and assumed the happenings like in Minority Report, but rather I would like to create an associate rule within the past data from the Chicago Data Portal. A suitable client and obvious one would be the police department, however the association variables found in this analysis would be more useful if it is used for rehabilitation than assumption that a pre-crime analysis is targeted on a person prior to the happenings of the crime. My goal of this project is to be able to find the correlation between the occurrence of the crime and the crime type in a specific location. For example, on a Saturday night, there is highly likely that Battery Theft is going to occur. Hence, data will be sent to keep people vigilant and preventive measures will be taken on people with any thoughts of stealing way before the crime occurs. The client would care because relatives of whomever will be involved in crime would be conscious of the behavior of the individual they are related to and will act to send them to rehabilitation before any action of the crime happening. Let's call this hypothetical client Rehab Centre A. Given the trend produced from the predictive model, we can hypothesize what crime type is best needed to be treated as an awareness to the public so that it will reduce the happenings. The recommendation can be based on personal rehabilitation and a town hall meeting to make the public in each area more vigilant and a stricter approach to policing based on imposing new laws on the penalty.

Data Descriptions and features

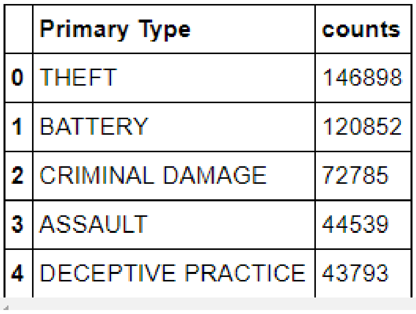
The dataset has existing models plotted on their website based on their data collection from 2001 up to 2 weeks before the current date. Although graphs are plotted to represent crimes and types of crimes in each area, no statistical analysis is done to depict correlation to a future prediction model. So, ideally to start visualizing the trend, the required features of the dataset must be explored.

Table 1.1 Output from the data acquisition of crimechicago.csv



Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward
JA379924	8/4/2017 23:50	003XX N KARLOV AVE	320	ROBBERY	STRONGARM - NO WEAPON	STREET	False	False	...	28.0
JA377717	8/4/2017 23:55	035XX W 16TH ST	2024	NARCOTICS	POSS: HEROIN(WHITE)	SIDEWALK	True	False	...	24.0
JA377744	8/4/2017 23:55	012XX S KEDVALE AVE	560	ASSAULT	SIMPLE	APARTMENT	False	True	...	24.0
JA377726	8/4/2017 23:55	070XX S BISHOP ST	460	BATTERY	SIMPLE	RESIDENCE PORCH/HALLWAY	False	False	...	17.0
JA377722	8/4/2017 23:59	052XX N OAKVIEW AVE	1310	CRIMINAL DAMAGE	TO PROPERTY	OTHER	False	False	...	41.0

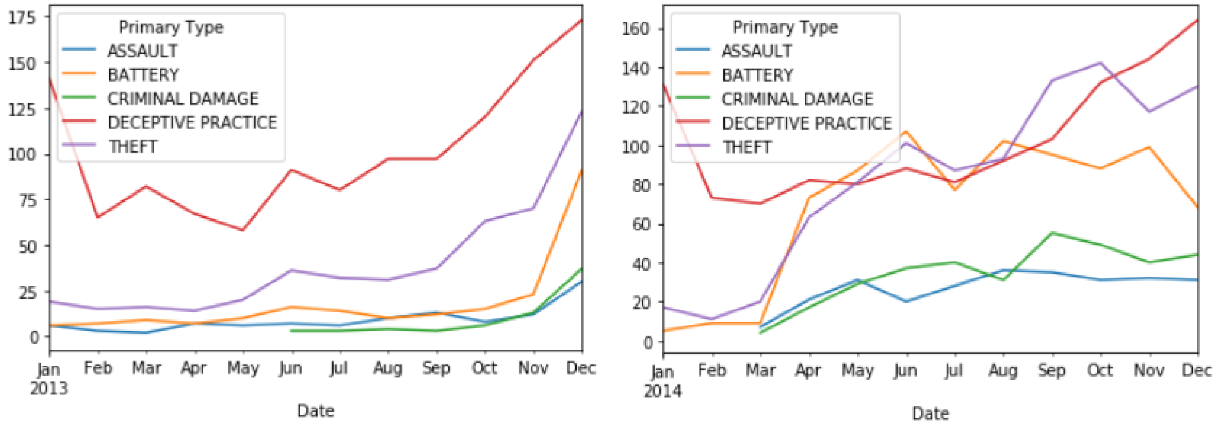
Generally, the dataset contains reported crime with date and its crime information such as type, description, location and ward. The problem can be simplified by focusing on just the column ‘Date’ and ‘Primary Type’. Doing so will lead the problem closer to finding the correlation between the crime and time of crime being reported.



	Primary Type	counts
0	THEFT	146898
1	BATTERY	120852
2	CRIMINAL DAMAGE	72785
3	ASSAULT	44539
4	DECEPTIVE PRACTICE	43793

Table 1.2 Determining the top 5 Crime type over the 5 years period.

From Table 1.2, the top five crimes have been determined and technically the focus on these variables should be a head start on determining the trend in crimes over the months, weeks and days. Notice that Date feature includes both the date and exact time of the crime when reported. This feature requires the conversion of ‘Date’ to datetime data type to conserve the difference and further clarity for analysis in iPython notebook.

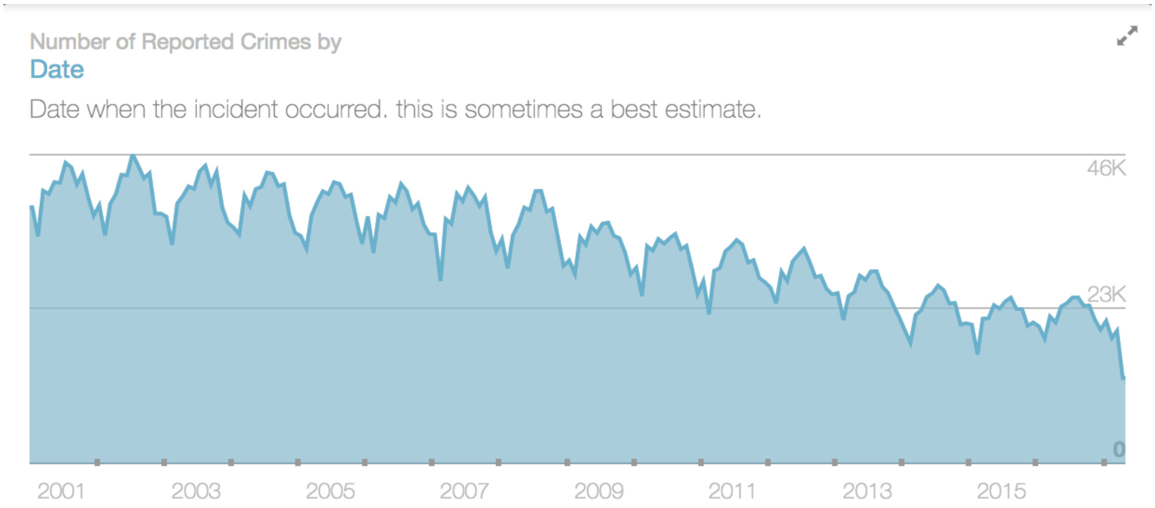


Graph 1.1 Graph plot of Date vs Crime Type for the year 2013 and 2014

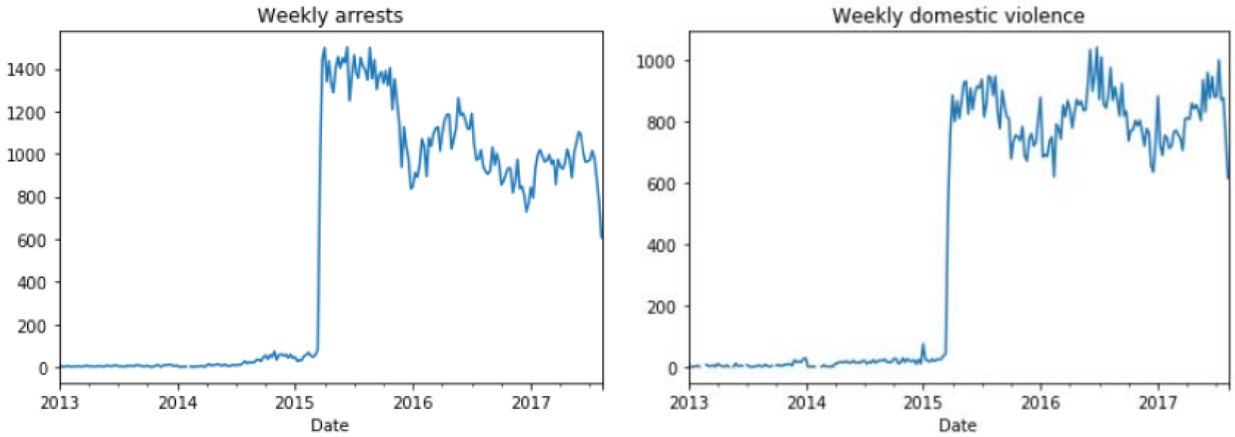
The key variables to plot all the crimes occurred vs. time are CRIME_FREQ and CRIME_DATE. From this plot, a rough estimate of how the overall crime pattern can be determined. The two highest peaks as seen from the plot is during winter and summer season. Winter season being the end of December towards February while summer season started from May to end of August. From this analysis, a deduction can be made that there seems to be a relation between crime frequency and the frequency of the overall crime. Thus, there must be a reason to the following occurrence to be guilty of crime. Now when looking at crime types from the two years of plot, Deceptive Practice is unusually high and always occurring, further analysis on the association between crime types and the day of week will be shown below. Although, the deceptive practice crime frequency is high, the other types of crime are high towards the end of each year. Therefore, police must be warned by the fact deduced from the data above that during winter, the policing should be extra vigilant. Have the police been doing a good job patrolling and taking preventive measure?

A simple data modelling can be done for the number of arrest based on domestic violence and outdoor arrests should be put into comparison.

Initial Data Acquisition and Analysis



We can hypothesize that from 2001 to the current year, the reported crimes or crime occurrence has reduced and should keep decreasing in the coming years. This trend makes sense because over years, technological advancement of tactical and urban planning improved the safety of the city and maintaining it towards the needs of the people. Therefore, the client for this project is the Chicago Police Department, specifically the division that takes charge of increasing the patrol planning within the city.

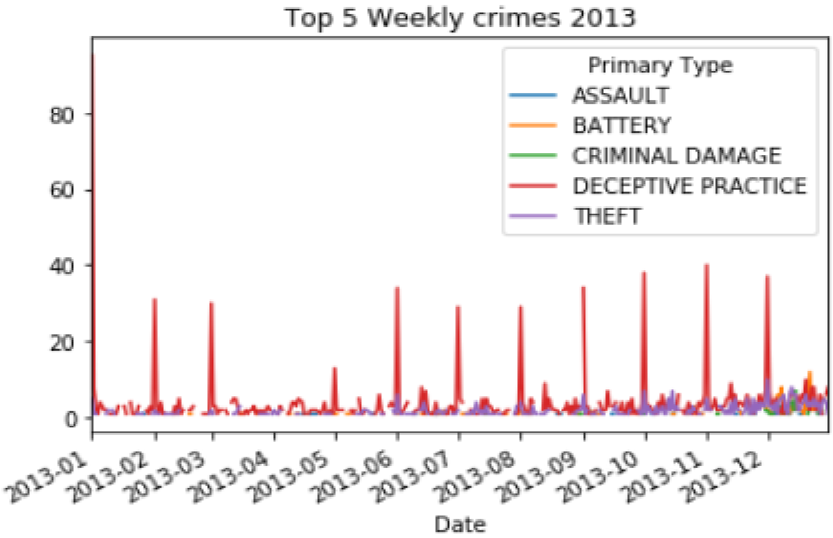


Graphs 1.2 Comparison between weekly arrests and domestic violence

Note that the data acquisition shown above is based on arrests that are only true. Before the year 2015, a lot of the reported data includes false arrests and domestic violence. Due to that nature, the data seemed like there is low crime rate, but this is not the case for the above plots. Some implementation must because of the major turnover "bad cop- good cop" event[2] in Chicago shooting during 2015 and 2016 that significantly increase the enforcement of laws on the policing tasks. From the above plot, we can see the eventual decline in the arrest post of the years of 2015. While this assumption might be true, it is not for the domestic violence. As city of Chicago supports the prevalence[3] of domestic violence that over 30% of women, seniors, LGBT and children were physically abused by a related or some by stalkers or former

[2] https://www.newyorker.com/magazine/2015/12/21/bad-cops-good-cops
[3] https://www.cityofchicago.org/city/en/depts/fss/supp_info/general_facts_aboutdomesticviolence.html

members of family. An in-depth analysis on whether these facts really exists in the dataset will be shown below.

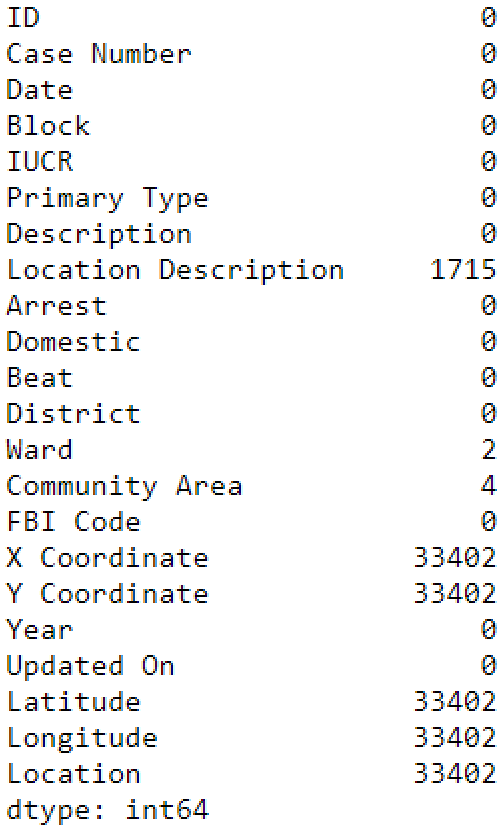


Graph 1.3 Days view of Crime Data over the year 2013

Another impact coming from the data is the legitimacy of the data acquisition, just how dependable is the data from the Chicago Data Portal? Looking from the plot above, it is hard to get into detail, the crime when just plainly looking at the view throughout the years. Inevitably, it can be said that the dataset, when checked using the '.nunique()' on iPython, it shows that not all data are one unique data. Thereby, the contents of some crime reports may be duplicated submission from a police clerk who logged the data in the system. The next data wrangling step below will introduce further analysis on a micro scale and will go about the method in which variable can be used to analyzed while minimizing inaccuracy in the fact drawn from the analysis.

Outlining the approach to solving this problem

Data Wrangling step is taken into consideration based on a second crime analysis workbook. The dataset from l is a huge complex one. The website includes already many existing analyses such as graphs and descriptive plots on the highest crime type aggregated from the year 2001 to current. Being able to sort the data according to date and time is important. First, a decision has to be made on whether variables with missing data entries can be omitted before modelling begins. The command df.isnull().sum() is used to add all the missing data entries as shown in the result below.

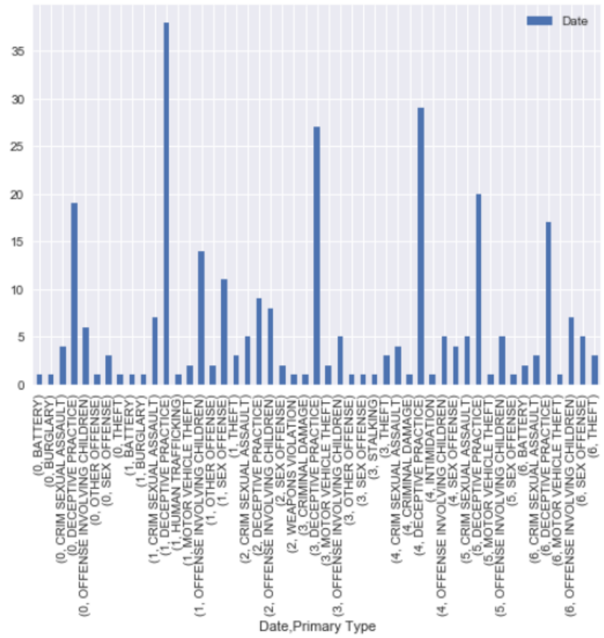


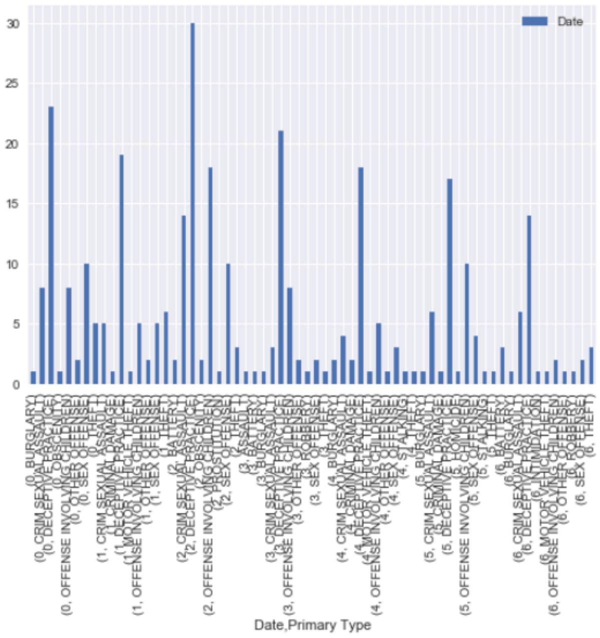
```
ID                          0
Case Number                 0
Date                        0
Block                       0
IUCR                        0
Primary Type                0
Description                 0
Location Description     1715
Arrest                      0
Domestic                    0
Beat                        0
District                    0
Ward                        2
Community Area              4
FBI Code                    0
X Coordinate            33402
Y Coordinate            33402
Year                        0
Updated On                  0
Latitude                33402
Longitude               33402
Location                33402
dtype: int64
```

Table 1.3 Results from running null aggregation on the dataset.

From table 1.3, it is apparent that the missing values are location features. Since location featured that is going to be used for analysis is not coordinate and description related, we can choose to ignore that further action is required. Whereas, for ward and community area, these two features may be required, but the amount of missing data is too small to make any significant difference in the result.

The essential step before modelling the analysis is to index the dates so that it can be sorted accordingly. This method can be done using pd.to_datetime(df.index). We want to be able to simplify the findings based on the top 5 crimes of the total crime types. To do this, we account for the total years of crime reports and count all the crime types using grouby method. It can be found through this step that the five primary types of crime are theft, battery, criminal damage, assault and deceptive practice with the order from biggest count to smallest count. Further, a plot is depicted using seaborn to present this data visually.



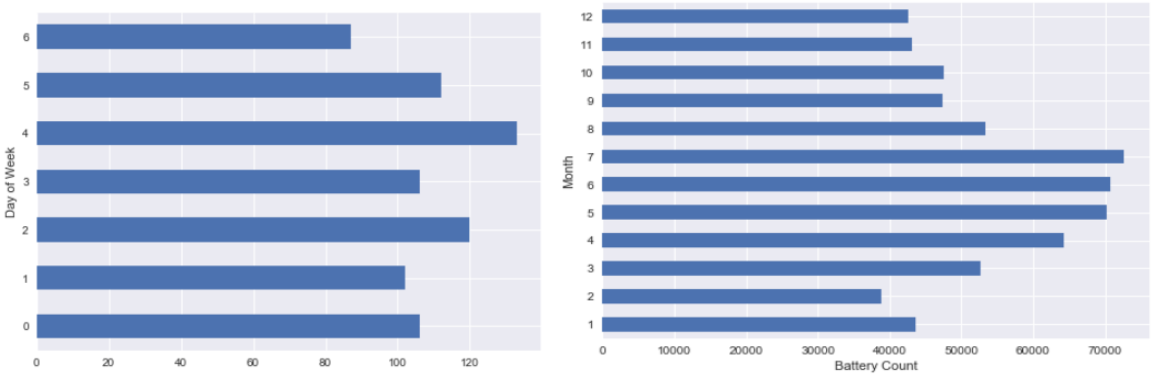


Using groupby method we can achieve classifying Crime types based on the day of the week when the data type is converted to datetime; day 0 = Monday to day 6 = Sunday. From table 2.1 in Appendix, the data for the frequency of each crime can be seen associated with the respective day. While it is deceptive practice that has the top crime reported, like the previous analysis above, crime type like ‘offense involving children’ or otherwise known as the domestic violence, is also a crime highly corelated beginning of the week. This crime type is unusually high and also particularly not visible when analyzing in terms of year. Thus, crime types like domestic abuse, sexual assault and sexual abuse are the unseen crime features that Chicago policing should be let informed about.

Location Analysis

Many of the terminology used in the dataset such as Ward and Beat are what the police department used to classify reported crimes. Since some of the crimes filed and reported does not escalate to an arrest, it can be considered a false crime. We use a Boolean command for which the arrest is 'True' only to filter the false report crimes and focus on the last 5 years (2013 -2017) for analyses of data based on the years. Beat specifically is based on territory policing and patrols that "can create more interaction between police and community members".[4]

Subsequently, Ward variable plays a stable role in determining which crime were required for police to be informed for the policing to improved and hence provide a better rehabilitation support for the victims. For the modelling, a random beat number of 1621 (16th District) is taken into analysis in conjunction with Ward 1.



Graph 2.2 Crime Count based on the Beat 1621 Ward 1.

Similarly, in this analysis, the days of the week that is significant would be the Day 4= Friday and Month 6,7 that are June and July respectively. Hence, this supporting data shows that there might be an overall reason behind the increase in crime and confirmed the trend in the data mentioned above.

[4] https://en.wikipedia.org/wiki/Beat_(police)

Using Association rules and Apriori to analyze the different variables

Association rules method is often used in the analysis of transaction of sales. It is widely known that this technique can find how closely correlated are a transaction with item A and item B. Although we all know tea and sugar, or milk are complementary shopping items. With association rules, we might find a third variable to the habits of tea drinkers. For instance, tea drinkers also purchase biscuit brand C when shopping for tea. Thus, stocks of biscuit brand C will be kept in stock proportionately with tea or discounts can be applied to one or the other to boost performance of sales. This is just one example of how association rules work for data analysis. Now imagine the transactional data are crime data. The variables involve are dates, crime types, ward, beat and location. With the right association of the variables, the dataset can output the 'habit' or a pattern for the crime that is likely to occur.

Further, an algorithm exists within this method called the Apriori principle. This algorithm reduces the number of item sets that needs to be analyzed. The two main indicators of the algorithm output are support and confidence. Support defines the proportion of the transaction belonging to an item type. Given that there is for instance theft happening 4 out of 7 days of the week, then the Support value for theft is going to big (>50%). Consequently, confidence is a derivative of support of two variables. The larger the value for confidence, the higher associated item A and B are. For example, theft mostly happen on Thursday. Then Thursday and Theft are associated. While there are other indicators to show correlation between variables, Apriori focuses on Support and Confidence to find out the associative properties between dates, crime types, ward, beat and location. On the next page, results are computed on a table to show the summary of associated variables. With these steps taken, a crime forecast based on weekly predictions for each police district can be interactively plotted on Chicago maps using random forests for detailed crime report visualization

Date vs. Primary Type

Year	Most frequent day of Occurrence	Most Frequent Crime Type	Other Frequent Crime Type
2013	Monday	Deceptive Practice	Offense involving Children
2014	Tuesday	Deceptive Practice	Offense involving Children, Sex Offense
2015	Wednesday	Deceptive Practice	Theft, Criminal Damage
2016	Monday, Friday	Deceptive Practice	Theft, Criminal Damage
2017	Sunday	Deceptive Practice	Theft, Criminal Damage

Location vs. Primary Type

From the analysis with apriori for the variables location and primary type, the top 3 locations associated with the most frequent crimes are Battery in Apartment, Criminal Damage around Residence and Theft on the street. From the two analyses ran for location and date, we could add one more variable (Beat, Ward, Time) to the algorithm to find out what else are those two variables associated with.

Ward vs. Primary Type

Ward Number	Most Frequent Crime Type
2	Theft
27	Theft
28	Battery
42	Theft

Beat vs. Primary Type (Min Support of 0.004)

Beat Number	Most Frequent Crime Type
111	Theft
112	Theft
1834	Theft

Beat and Ward vs. Primary Type

Ward Number	Beat Number	Most Frequent Crime Type
42	111	Theft
42	112	Theft
42	1834	Theft

Time vs. Primary Type

Timestamp is too intricate and need to be converted to category; morning, noon, night which defeats the purpose of the analysis

Time category	Crime Type
Morning	Associative support < 0.05, so no relation
Noon	Battery, Theft
Night	Battery, Theft
Late Night	Associative support < 0.05, so no relation

Most equally frequent and associative variables

Week of Day	Time of Day	Beat	Ward	Location	Crime Type
Sunday, Monday	Noon- Night	111, 112, 1834	42	Street	Theft

Open Questions and Conclusions

At this point, the forecast that was done for this project still has room for further analysis. Given the access to R, one would be able to visualize further detail on geolocation like Tableau, a current trending product that uses longitude and longitude to visualize data in a map. Consequently, the address of each crime occurrence can be accurately mapped as coordinates and each point can be pinpoint for predictive analysis.

Further, census data can also be used for secondary add-on to the current analysis. For instance, dataset for nightlife and other events may spur more crime that is happening during night time. Before deducing any further, one should also account for the fluctuation trend in the plot above. What causes the peak of crime incident being reported annually? Is it because of any recent political event or any economic crisis that occurred, which brought down many jobs, thus bringing people into the desperate choice of doing crime? One should analyze these reports in terms of their types and their related news to find out the answer to the spike in the trends. Some require data outside of this source would be to track down the news on politics, stock market (S&P) over certain years to determine the trend.

Grid location analysis could provide a better deduction between location, time and crime types; therefore, being able to find some highly correlated variables for association. While Apriori method has been used to find out the association between the core variables, there are some limitations to the method. Association used for analysis of large datasets would consists more item set tweaking. The support value must have a minimum threshold to determine whether the proportion is considered big or small. The method to determine this threshold could be difficult as the data needs to go through training and testing before it can be generalized as a certain degree of association. Hence, the summary above could be a spurious result. On the other hand, apriori algorithm on large dataset can be considered computationally expensive[5] as it requires many sorting algorithms for the many variables involved.

[5] https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html/2

Appendix

Table 2.1 Crime Aggregated data vs. Day of week for Year 2013

	Date	
Date	**Primary Type**	
0	**BATTERY**	1
	BURGLARY	1
	CRIM SEXUAL ASSAULT	4
	DECEPTIVE PRACTICE	19
	OFFENSE INVOLVING CHILDREN	6
	OTHER OFFENSE	1
	SEX OFFENSE	3
	THEFT	1
1	**BATTERY**	1
	BURGLARY	1

	Date	
Date	Primary Type	
	CRIM SEXUAL ASSAULT	7
	DECEPTIVE PRACTICE	38
	HUMAN TRAFFICKING	1
	MOTOR VEHICLE THEFT	2
	OFFENSE INVOLVING CHILDREN	14
	OTHER OFFENSE	2
	SEX OFFENSE	11
	THEFT	3
2	CRIM SEXUAL ASSAULT	5
	DECEPTIVE PRACTICE	9
	OFFENSE INVOLVING CHILDREN	8

	Date	
Date	**Primary Type**	
	SEX OFFENSE	2
	WEAPONS VIOLATION	1
3	**CRIMINAL DAMAGE**	1
	DECEPTIVE PRACTICE	27
	MOTOR VEHICLE THEFT	2
	OFFENSE INVOLVING CHILDREN	5
	OTHER OFFENSE	1
	SEX OFFENSE	1
	STALKING	1
	THEFT	3
4	**CRIM SEXUAL ASSAULT**	4
	CRIMINAL DAMAGE	1

	Date	
Date	Primary Type	
	DECEPTIVE PRACTICE	29
	INTIMIDATION	1
	OFFENSE INVOLVING CHILDREN	5
	SEX OFFENSE	4
5	CRIM SEXUAL ASSAULT	5
	DECEPTIVE PRACTICE	20
	MOTOR VEHICLE THEFT	1
	OFFENSE INVOLVING CHILDREN	5
	SEX OFFENSE	1
6	BATTERY	2
	CRIM SEXUAL ASSAULT	3

	Date	
Date	Primary Type	
	DECEPTIVE PRACTICE	17
	MOTOR VEHICLE THEFT	1
	OFFENSE INVOLVING CHILDREN	7
	SEX OFFENSE	5
	THEFT	3