Final Report: Chicago Crime Data Analysis

MIDS -INFO- S18 Python Fundamental

Project 2: Data Analysis

Angshuman Paul, Anu Sankar, Yue Hu, Kuangwei Huang

Summary:

The analysis of Crimes in Chicago has been an area of interest in most studies related to crimes, especially those in the United States of America. Since the beginning of the 20th century, the crime rate in Chicago has been above the US average[1]. We will be analyzing the crime data made available by the city of chicago and try to uncover some insights related to the crimes that were committed in the city between 2015 and 2017. We will also be using weather data as a supplementary dataset to understand any bearing that weather may have had on these crimes.

[1] - https://en.wikipedia.org/wiki/Crime_in_Chicago

Research Questions:

Our research surrounds the following questions:

1. How does crime vary across the years from 2015 through 2017?
2. What type of crime is the most frequently committed (and in which season)?
3. How does theft vary during the holiday season (Thanksgiving to New Year) compared to the rest of the year?
4. How does crime in Chicago relate to the weather?
5. Are there any high incidence areas whereby the police should increase patrols during different weather, times of the year?
6. Are domestic crimes more or less violent than the non-domestic ones in general? Are they more or less frequent?
7. Is any part of the city more susceptible to crimes involving firearms?
8. Is there any correlation between arrests and the type, location or time of crime?
9. Is there any correlation between the type of crime and the type of location where they are committed? For example, are thefts more common in retail outlets and gas stations as opposed to apartment or residences?

Datasets:

1. **Chicago Crime Data -** This dataset reflects reported incidents of crime that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. We will slice this dataset to focus on a more manageable subset spanning 3 years from January 1, 2015 through to December 30, 2017.

   Dataset Link: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
   Data Definition / Description: http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html

Shapefile for Geospatial Plots:

https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

Structure:

- Full data set total lines (Since 2001): 6,563,183,
- 3-year subset total lines(2015 - 2017): 798,748
- Number of columns: 17

	date	block	iucr	primary_ type	description	location_ description	arrest	domestic
count	798748	798748	798748	798748	798748	798748	798748	798748
unique	342113	31884	353	33	329	144	2	2
top	2015-01-01 00:01:00	001XX N STATE ST	820	THEFT	SIMPLE	STREET	FALSE	FALSE
freq	150	2536	73379	183051	84790	181658	624787	671585

	beat	district	ward	community_ area	fbi_code	year	latitude	longitude
count	798748	798748	798748	798748	798748	798748	798748	798748
unique	274	24	51	78	26	3	264443	264353
top	1834	11	42	25	6	2016	nan	nan
freq	7720	55925	44548	49356	183051	268428	8570	8570

2. Weather Data - In order to study the crime rate with weather, the weather data in Chicago from January 1, 2015 through to December 30, 2017 are downloaded from Weather Underground website and parsed into a flat CSV file.

Python Code Link: https://github.com/fivethirtyeight/data/tree/master/us-weather-history
(wunderground_scraper.py and wunderground_parser.py are used to download and store weather data in CSV file)

Dataset link: https://drive.google.com/open?id=1aqqyxW77aafZINpV4fHIXVN-SKag2sACzT67DY66odY

Structure:

- Data set total lines (January 1, 2015 - December 30, 2017): 1,096
- Data set total columns: 13

	actual_mean_temp	actual_min_temp	actual_max_temp	actual_precipitation
count	1095	1095	1095	1095
mean	52.005479	43.099543	60.395434	0.10853
std	19.690428	18.913837	21.019921	0.295517
min	-2	-13	4	0
25%	37.5	30	43	0
50%	54	44	63	0
75%	69	59.5	79	0.05
max	84	76	95	4.19

Sanity Check:

A majority of the attributes in the crime dataset are categorical and mostly discrete in nature. The supplemental weather dataset mostly consists of quantitative data. We did some basic sanity checks by looking at the descriptive stats and value counts. The data-types for and values for the variables were as expected and within the expected ranges wherever applicable. Null checks for the variables showed that location_description have a couple thousand nulls while district and ward have 1 and 3 null values respectively. Variables for longitude and latitude have close to 8500 null records which is roughly 1% of the overall dataset.

Data Cleaning and Preparation:

The original crime dataset contains data since 2001 and contains about 6.6 million records. We focussed on the data from 2015 to 2017 for our analysis which contains about 800 K records. We dropped the 'ID', 'Case Number', 'Updated On', 'Location', 'X Coordinate', 'Y Coordinate' columns from this dataset as we were not planning to use these columns. The original dataset was then sliced to retain only 3 years worth of data. Then this dataset was broken down into 12 subsets each containing data for one quarter so that they were more easily manageable. For the weather dataset, we imported only 'date','actual_mean_temp', 'actual_min_temp', 'actual_max_temp', 'actual_precipitation' columns as these were the only ones of interest for our analysis. These datasets were then joined based on the date column.

As some of the key variables have codes instead of the actual values, we added functions to create new columns with more descriptive data so that the reports become more meaningful. 'Chicago sides' data was derived based on community area, generic street names based on block, district names based on district, firearm boolean based on description, location type based on location_description and fbi category, fbi crime type, & boolean columns indicating if the crime was relate to property or crime were derived using the fbi code variables. We also added a couple of bins for mean temperature and time of day.

Data Analysis

1. Trends in Crime Over Time

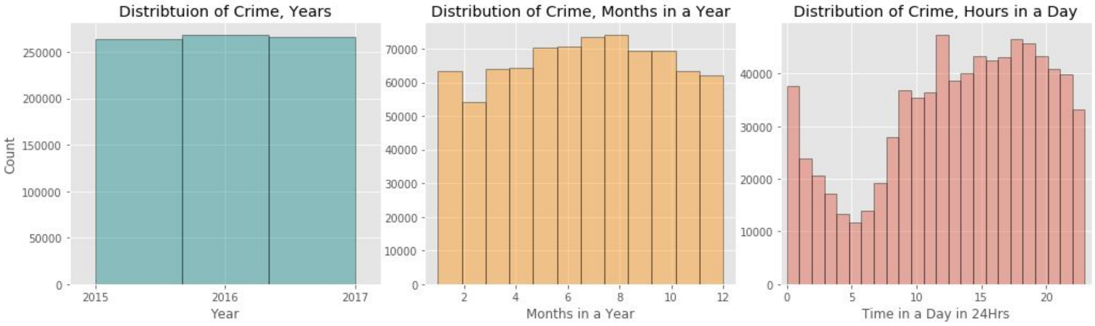


Figure 1.1: Crime Distribution over Years, Months, and Hours in a Day.

We started exploring how crime varies across the time period between 2015 through 2017 by looking at the distribution plots in Figure 1.1. The total number of crimes in Chicago has remained relatively constant over the years, at a mean of 266249 crimes a year. Number of crimes do appear to increase in the summer months from June through August and crimes are more often committed from the late morning 10am onwards to 10pm at night. There is an unusual spike at 12 midnight as well as 12 noon, perhaps more criminals find it convenient to plan and coordinate crimes at these hours.

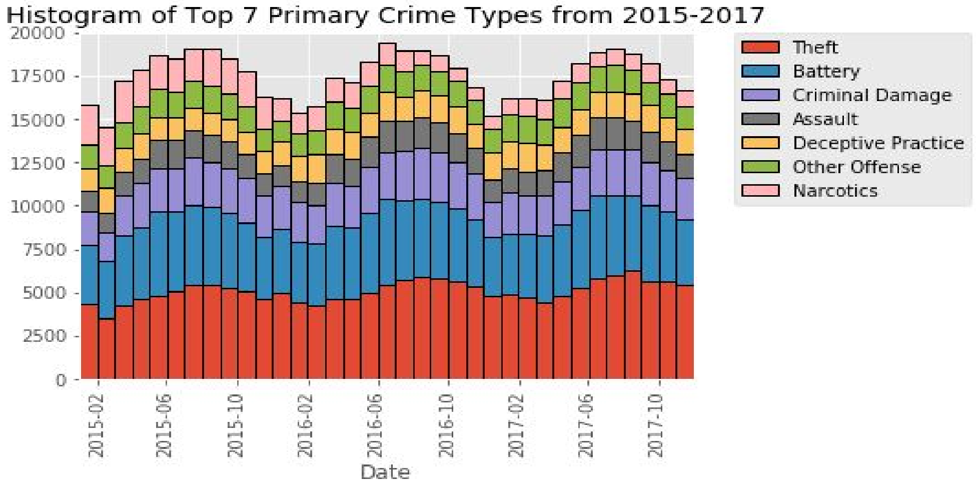


Figure 1.2: Three Year Histogram of Top 7 Primary Crime Types

Upon investigation how the top 7 crimes vary across the months 2015 to 2017 in Figure 1.2, we do see a repeated pattern that crimes are higher in the summer months from June to August. There could be a relationship that during the warmer months, there is more criminal activity in Chicago. There also seems to be slight increase in some types of crime, and decrease in other crime types over this 2015 to 2017 period.

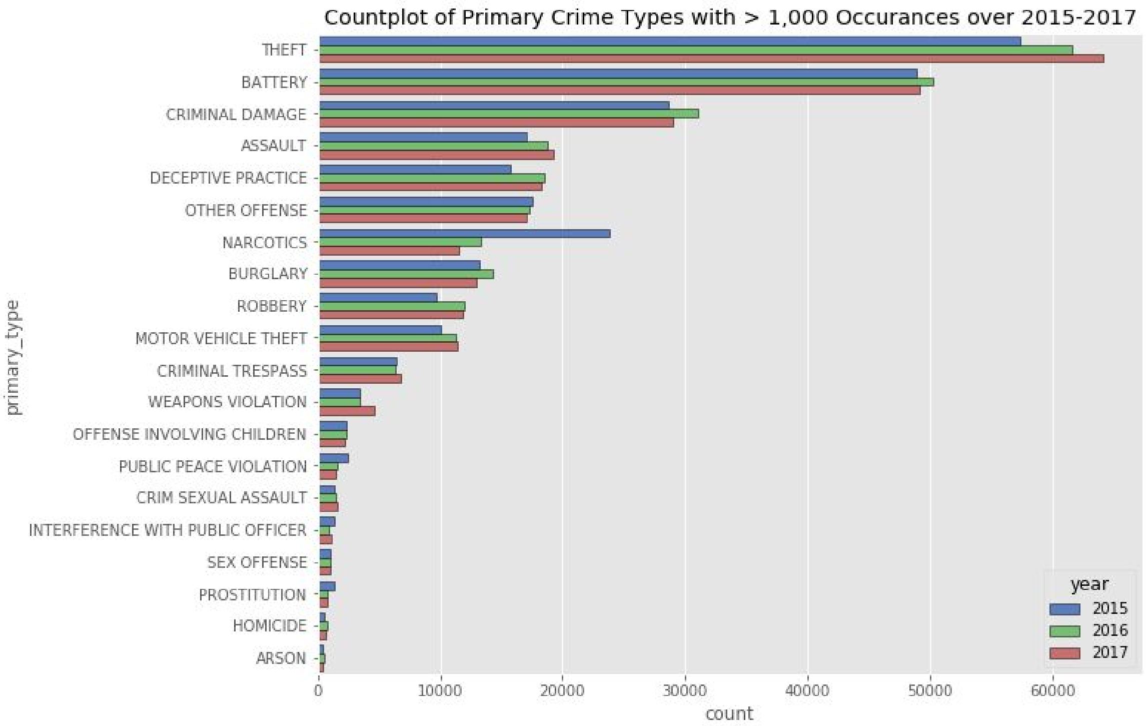


Figure 1.3: Countplot of Primary Crime Types with > 1,000 occurrences over 2015-2017

Plotting the countplot of the primary crime types with more than 1,000 occurrences over the 3 years in Figure 1.3, we see that Theft is by far the most frequently committed crime, followed by Battery, Criminal Damage and Assault. In looking at the individual trends of Theft crimes and Narcotics crimes, we do see an increasing trend for Theft crimes and a starkly decreasing trend for Narcotics, as shown in Figure 1.4. We also noticed that Theft activity increases in the summer months, with the highest activity in the month of August.

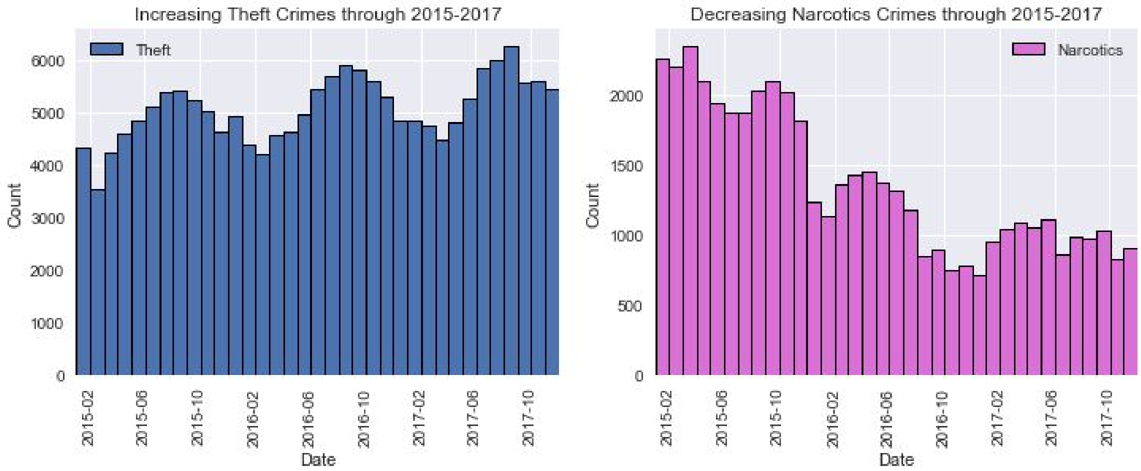


Figure 1.4: Trend of Theft and Narcotics Crimes over 2015-2017

Diving further into different sub-descriptions for Theft crimes in Figure 1.5, we can see that the highest proportion of Theft crimes are petty in nature with less than $500 value. We also see an increase in four types of Theft crimes from 2015 to 2017: Over $500, From Buildings, Retail Theft as well as Pocket-Picking.of the crimes that have the most observed changes from 2015 to 2017, Theft and Narcotics stood out.

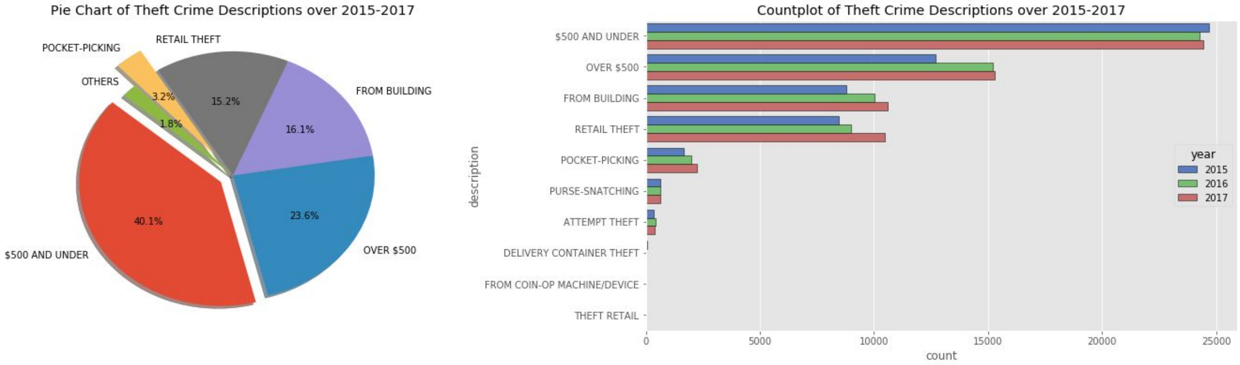


Figure 1.5, Pie Chart and Countplot for Subcategory Descriptions of Theft Crimes

Our next analysis was to see if some days of the week were more prone to crime. We analyzed the data for each of the top 5 crime types and for Narcotics.

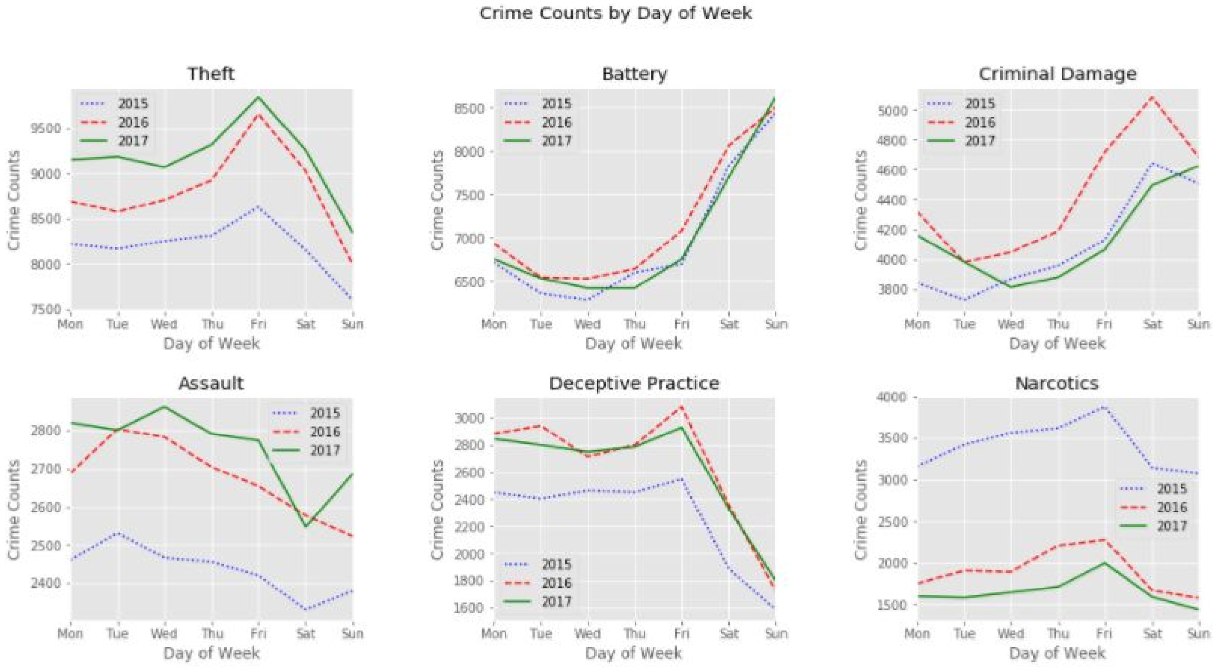


Figure 1.6: Crime Counts by Day of Week

The analysis shows a clear pattern between the type of crime and the day of the week for all three years of data. There is a consistent pattern of high incidence of Theft on Fridays. We have seen in a previous analysis that theft of $500 or less is the most common type of theft crime. As many people go out on Fridays and stay out late, there is ample opportunity for such petty theft. Theft on Saturdays is still higher than the other days of the week while it is lowest on Sundays when most people stay home. It should also be noted that Theft crimes have gradually increased year-over-year for the period 2015 to 2017.

Similarly, Battery type of crime peaks on Sundays while Criminal Damage is higher on Saturdays. Assault can be high on Tuesdays or Wednesdays. Both Deceptive Practices and Narcotics show a rise on Friday along with Theft. Again, more people being out and about on Fridays could be a key factor influencing the spike in these categories of crime.

In a previous analysis, we have seen that crimes involving Narcotics have declined over the three years from 2015 - 2017. The same conclusion is apparent in the Narcotics y-o-y chart.

2. Impact of Weather on Crime

Does weather affect the crime rate? Are more crimes committed with temperature or precipitation?

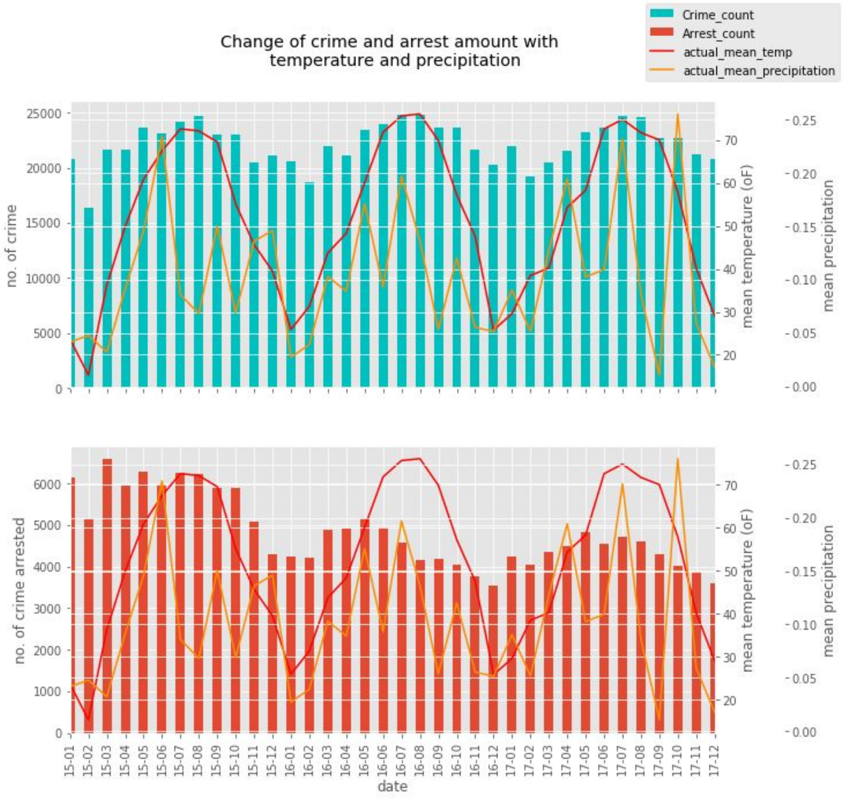


Figure 2.1a, 2.1b: Total Crime Count and Arrest Count against Temperature and Precipitation

Based on the trends in crimes over time, it shows crimes are higher in the summer months from June to August, which could result from the hot weather. Therefore, it is necessary to study the impact of weather on crime. Figure 2.1a and 2.1b shows variation of total number of crimes and number of crimes of which arrests were made with time, respectively, along with the change of mean temperature and precipitation over month. It is obvious that number of crimes are closely related to mean temperature in Figure 2.1a, because they share the same trend over the month. However, number of crimes of which arrest was made doesn't show close relationship with weather (temperature and precipitation) in Figure 2.1b.

In order to verify the close relationship between crime occurrences and temperature, heatmap for mean temperature and total crime distribution over time are plotted in Figure 2.2. These heatmaps confirm that the occurrence of crime is high during May to September, when the mean temperature is higher than 60 ^{o}F . Usually during warm weather, people head outside for activities and park is crowded with parades and music festival, which brings more opportunities for crime occurrence.

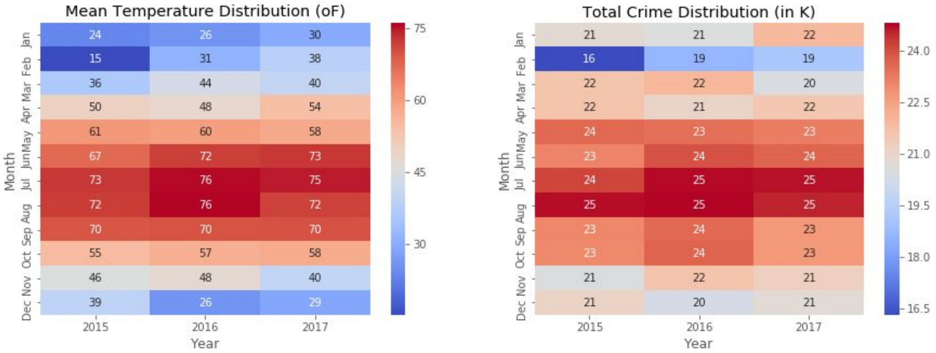


Figure 2.2: Heatmap for Mean Temperature and Total Crime Distribution over Time

In Figure 2.1b, it indicates that the trend of total crimes and crimes of which arrests were made over time are different. In order to study the relationship between these two variables, the new term arrest ratio is introduced and defined by:

$$arrest\ ratio \ = \ \frac{number\ of\ crimes\ of\ which\ arrests\ were\ made}{total\ number\ of\ crimes}$$

Figure 2.3a demonstrates that total number of crimes has remained relatively constant over years, while the number of crimes of which arrests were made is continuously decreasing, and as such, the arrest ratio is declining over year as shown in Figure 2.3b. It could cause by the worse police performance or the increase of crimes that police is not bothered with arrests.

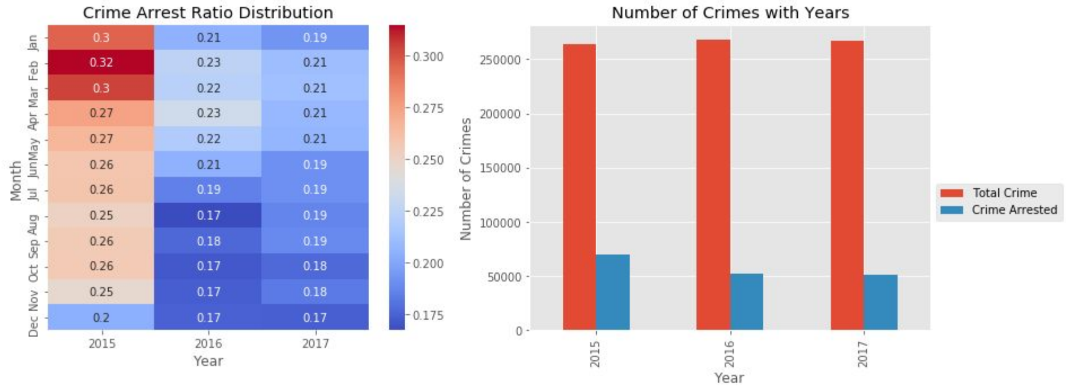


Figure 2.3a: Heatmap for Crime Arrest Ratio; Figure 2.3b: Crime Arrest Trend over the Years

In this study, the impact of weather on arrest ratio is investigated in Figure 2.4 based on domestic and non-domestic crimes. Two category variables are created based on temperature and precipitation as shown in Table 2.1. In Figure 2.4a the arrest ratio of domestic crime is highest during the day without precipitation, while arrest ratio of non-domestic crime is lowest during no precipitation day. According to temperature, both non-domestic and domestic crimes obtain the highest arrest ratio during cold day with temperature range from 15 to 50 °F.

Table 2.1: Category Variables for Temperature and Precipitation

Temp Range (°F)	-10 - 0	0 - 15	15 - 50	50 - 65	65 - 75	75 - 85	85 - 95	95 - 100
	Freezing	Very Cold	Cold	Pleasant	Warm	Very Warm	Hot	Very Hot

Precipitation (inch)	0		0 - 3		3 - 6	
Temp Range(°F)	< 32	>= 32	< 32	>= 32	< 32	>= 32
	no precipitation		light snow	light rain	heavy snow	heavy rain

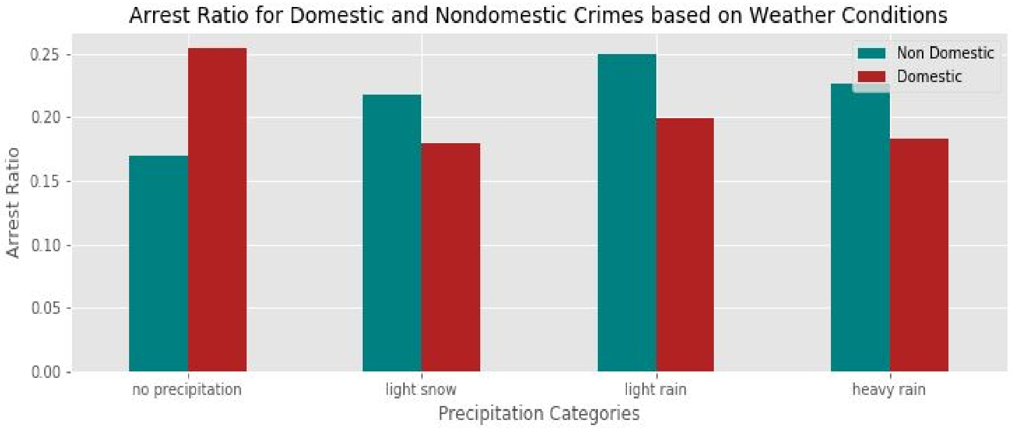


Figure 2.4a: Arrest Ratio for Domestic and Non-Domestic Crime across Precipitation Categories

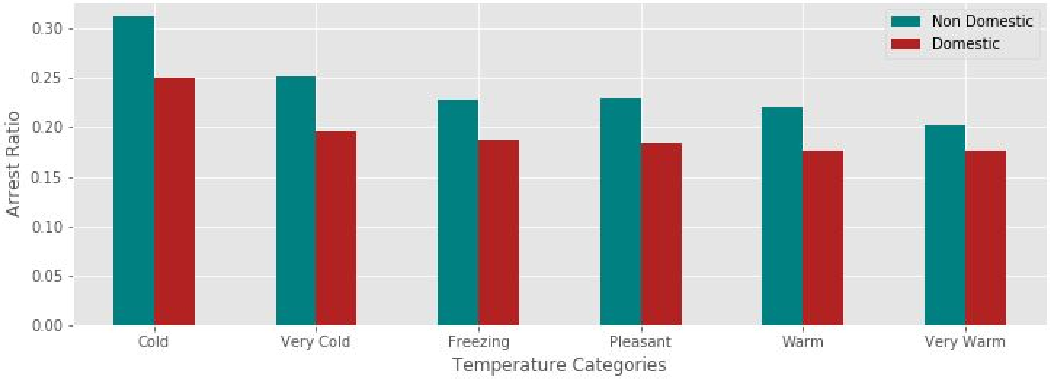


Figure 2.4b: Arrest Ratio for Domestic and Non-Domestic Crime across Temperature Categories

3. Crime Correlation with Location

Using community areas variable in the data set, and aggregating in the various Chicago 'Sides', we can see in Figure 3.1 overall tendency for most crimes to be committed in the Chicago 'West Side', and the police would definitely need to focus their resources there, especially during the summer months. For Theft cases, 'Central' location is second highest, and this is likely attributed to it being the prime commercial area.

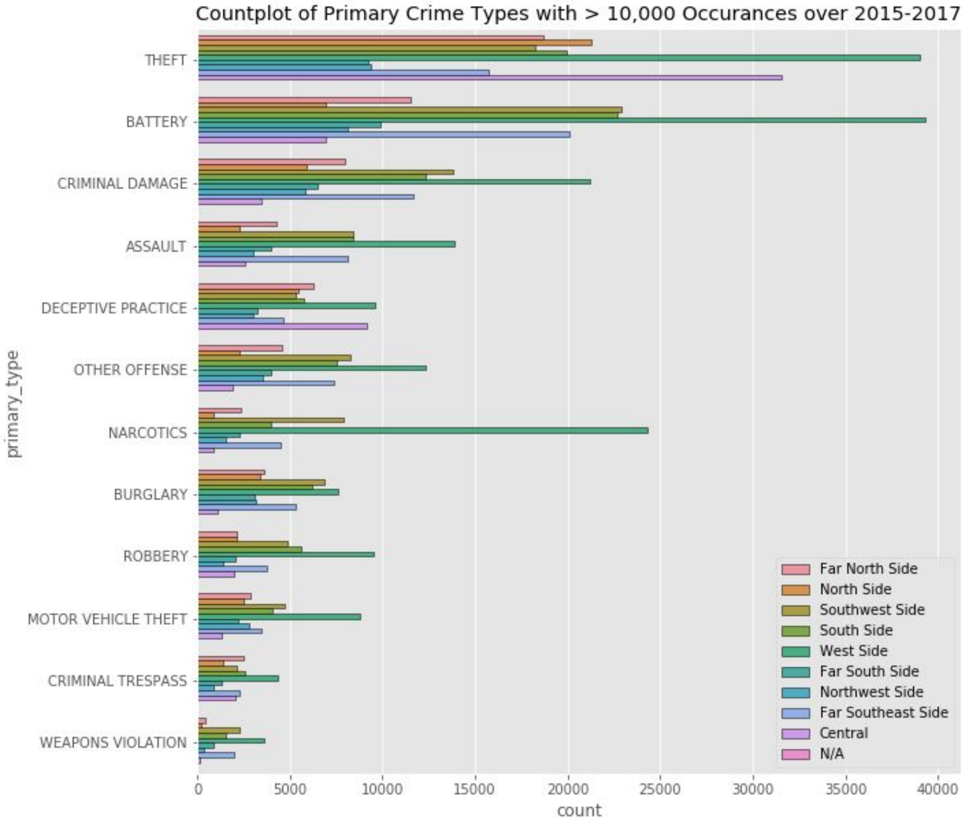


Figure 3.1: Countplot of Primary Crime Types across different 'Sides' of Chicago

Over the 3 year period from 2015 through 2017, from Figure 3.2, we see that Central region and to some extent the Far North Side have experienced year-on-year increase in crime, and the Southwest Side does see a slight year-on-year decrease in crime.

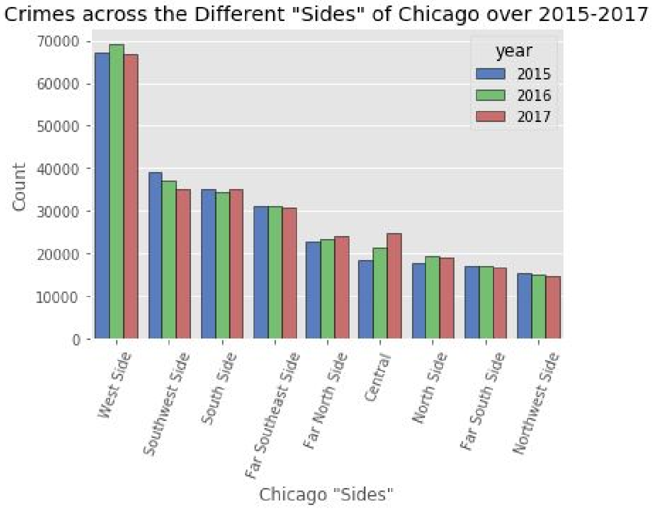


Figure 3.2: Crimes across Different ‘Sides’ of Chicago over 2015-2017

We then investigated the top crimes by location, type of location and FBI crime categorization. Referring to Figure 3.3, the plots showed us that crimes are more prevalent in streets and residential areas as compared to others. Crimes in the west side of the city seemed to be far more than those in the other parts of the city. Also, crimes related to Larceny far outnumbers all other crime categories.

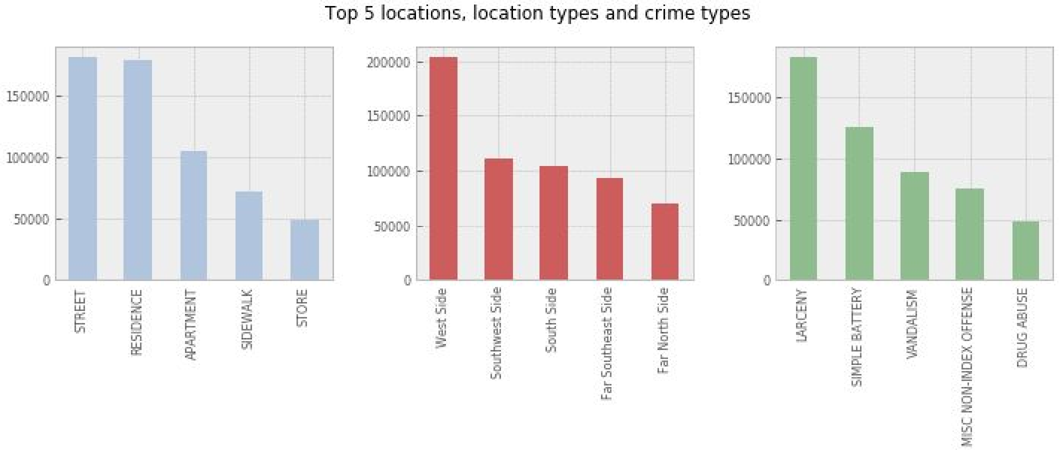


Figure 3.3: Top 5 Locations, Location Types and Crime Types

Next we tried to gather insights on any correlations that may exist between the top 5 locations, location types and crime categories and if the temperature played a role in increase or decrease in the number of crimes that were commited. We noticed that irrespective of the location, type and category, the number of crimes increase with temperature and peak when the mean temperatures are between 60 and 80 degrees after which they start to decline steadily.

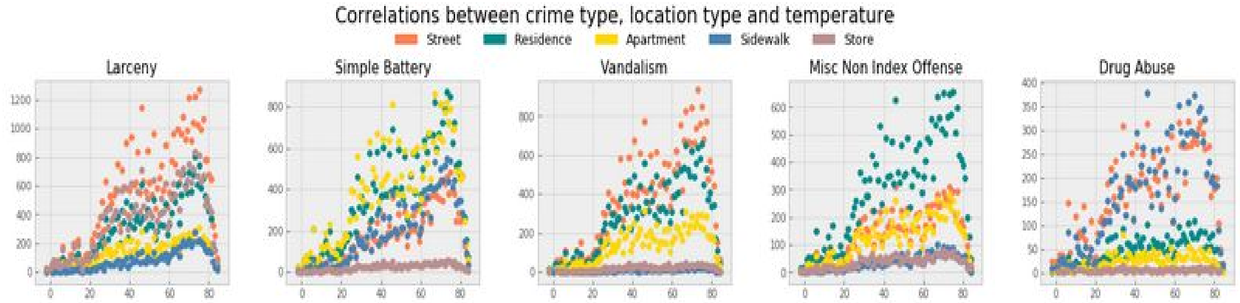


Figure 3.4: Correlations between crime type, location type and temperature

The study of the correlation within location types and categories however, brought some interesting patterns to light. Larceny related crimes are more common in streets and least in sidewalks. Crimes related to simple battery occur more in residences and apartments and the least in stores. Vandalism is more common in streets and the least in stores and sidewalks. Drug Abuse is more common in sidewalks and streets and least in stores while misc non index (non serious) offenses are more common in residences and least in sidewalks and stores. Overall, streets seem to be most prone to crimes while stores are the least.

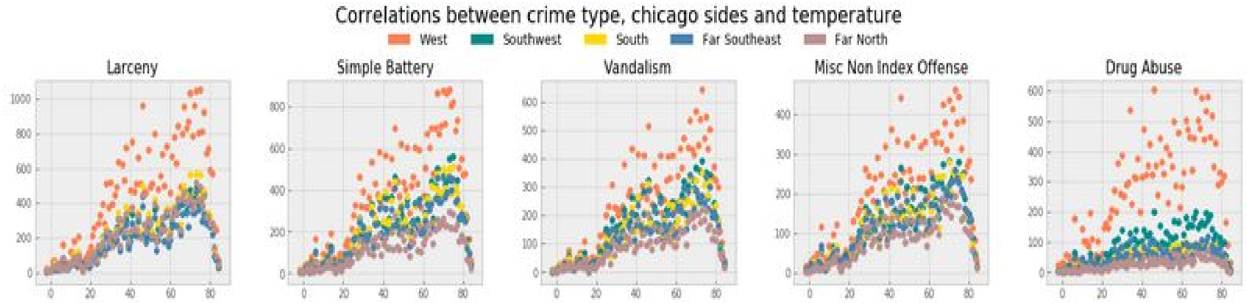


Figure 3.5: Correlations between crime type, Chicago Sides and temperature

When it comes to locations though, West Side stands out no matter what category of crime it is. However there are variations in the number of crimes for the top categories and the other 4 locations. Larceny is equally prevalent in Southwest, South, Far Southeast and Far North sides. But Battery, Vandalism and Misc non serious offenses seem to be more in Southwest and South as compared to the other two while Drug Abuse is more in Southwest side as compared to South, Far Southeast and Far North.

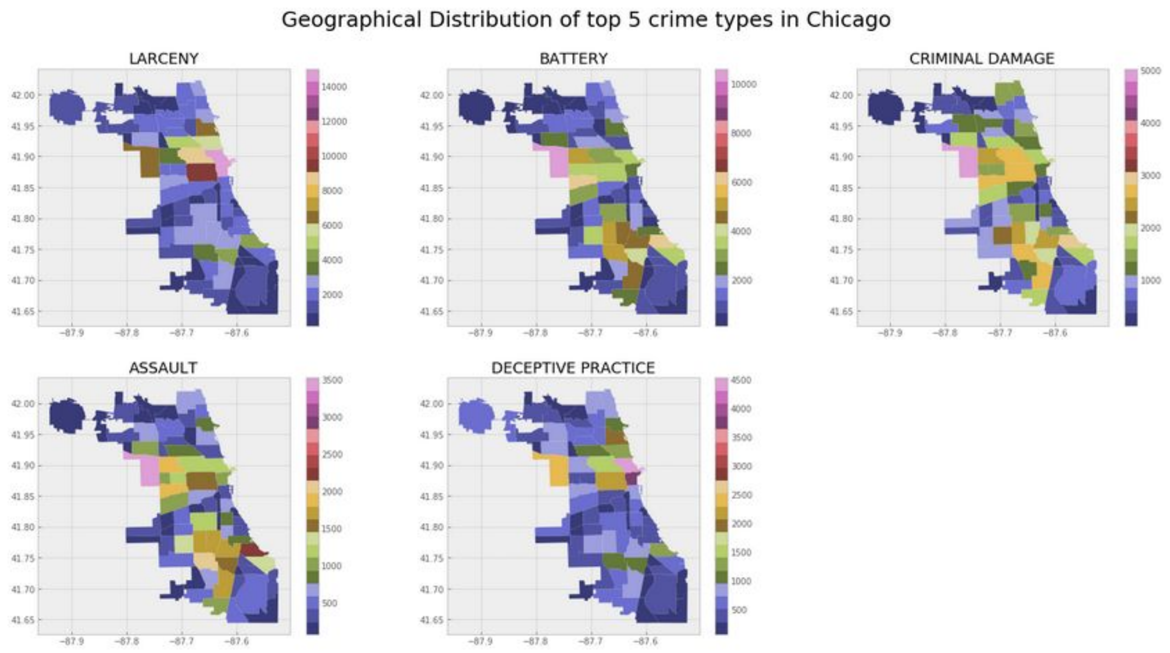


Figure 3.6: Geographical Distribution of top 5 crime types in Chicago

An analysis of the top 5 crimes by geography shows that the majority of these crimes occur in Central and West side communities. Larceny related crimes are mostly occurring in all communities in the Central side. Battery, Criminal Damage and Assault crimes are mostly in the community area 25 in the West side while Deceptive Practice is mostly happening in community area 8 of the Central side.

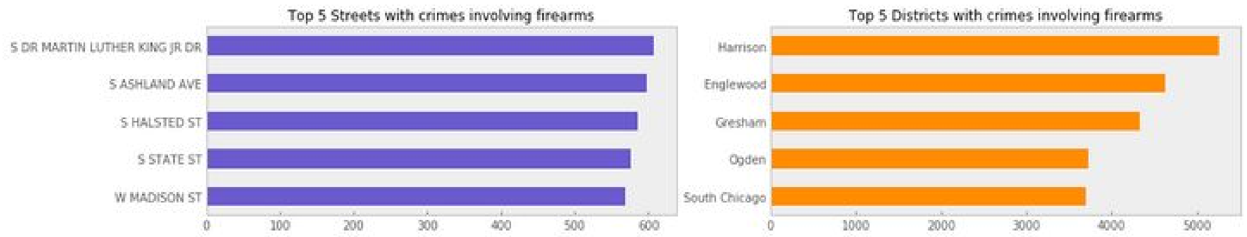


Figure 3.7

An analysis of the crimes involving firearms and other dangerous weapons indicates that S Dr Martin Luther King Jr Dr. and S Ashland Ave are the two most dangerous streets while Harrison and Englewood are the two most dangerous districts in Chicago.

The investigation of crime rate and arrest ratio based on police district would be important for a measure of police performance. The figure on the left shows the occurrence of crimes based on police district area. It shows that Lincoln, Rogers Park, Albany Park and Morgan Park are the most safest district area to live. An interesting discovery is that the arrest ratio in Harrison is as high as 0.38, though Harrison obtains the highest crimes. Therefore, it would be reasonable to evaluate the police performance based on not only crimes rate but also the arrest ratio.

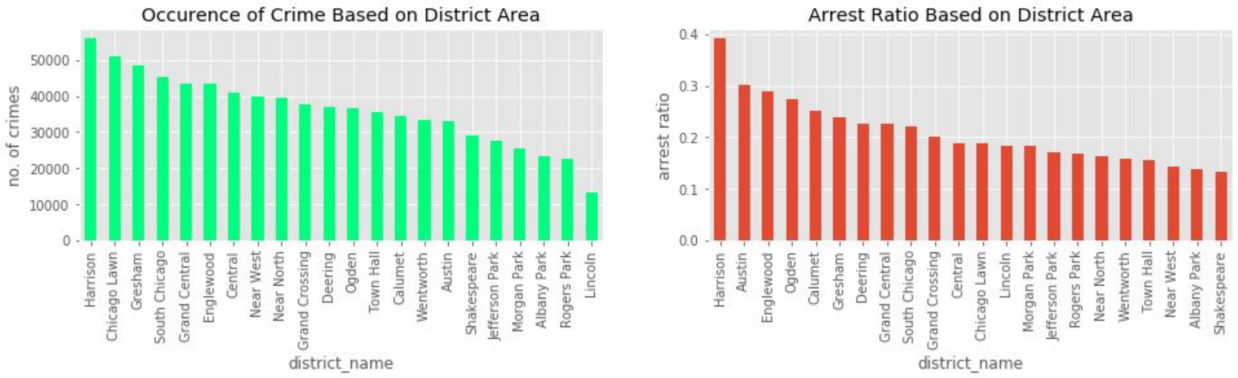


Figure 3.8: Crimes occurrences and arrest ratio based on police district area

In order to check the worst police district area with high crime occurrence and low arrest ratio, heatmaps of both crime and arrest ratio distribution over the time and police district area are plotted. Even though the arrest ratio is not directly related with police performance. Somehow, it could cause worries of low arrest ratio for crimes of which arrests should be made. It is obvious that police district areas, Gresham, Grand Central, Chicago Lawn amd Central, obtain low arrest ratio with high crime occurrence, indicating better police strength and more patrols are needed in this areas.

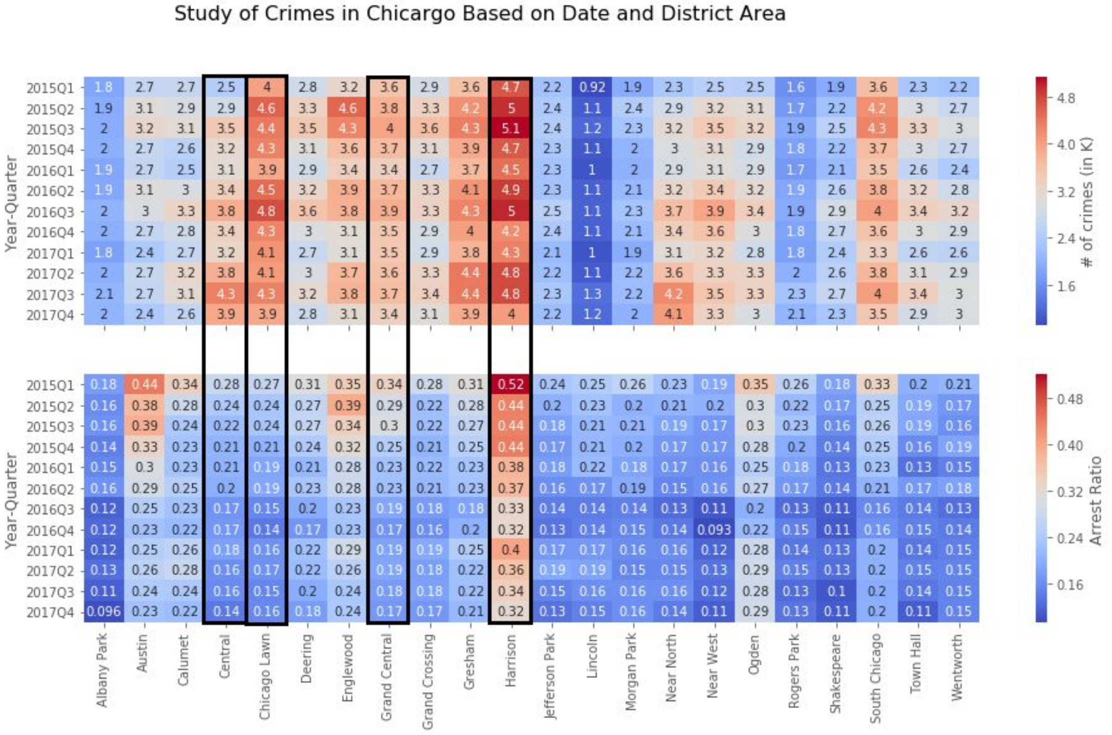


Figure 3.9: Crimes occurrence and arrest ratio distribution over time and police district areas

According to the study, Gresham, Grand Central, Chicago Lawn and Central police area needs better police performance. However, it comes with the questions when more patrols are needed in these areas. Therefore, the investigation of crime occurrences based on quarter, day of week, time and weather conditions are performed and illustrated in Figure 3.10. It shows that these four district areas share some common findings that high crime occurrence happens in Quarter 2 and 3, Friday, early afternoon and the day without precipitation. However, in Gresham, Grand Central, and Chicago Lawn, the crime rate is also high during evening. Therefore, the patrols in these district areas are necessary to increase during these findings.

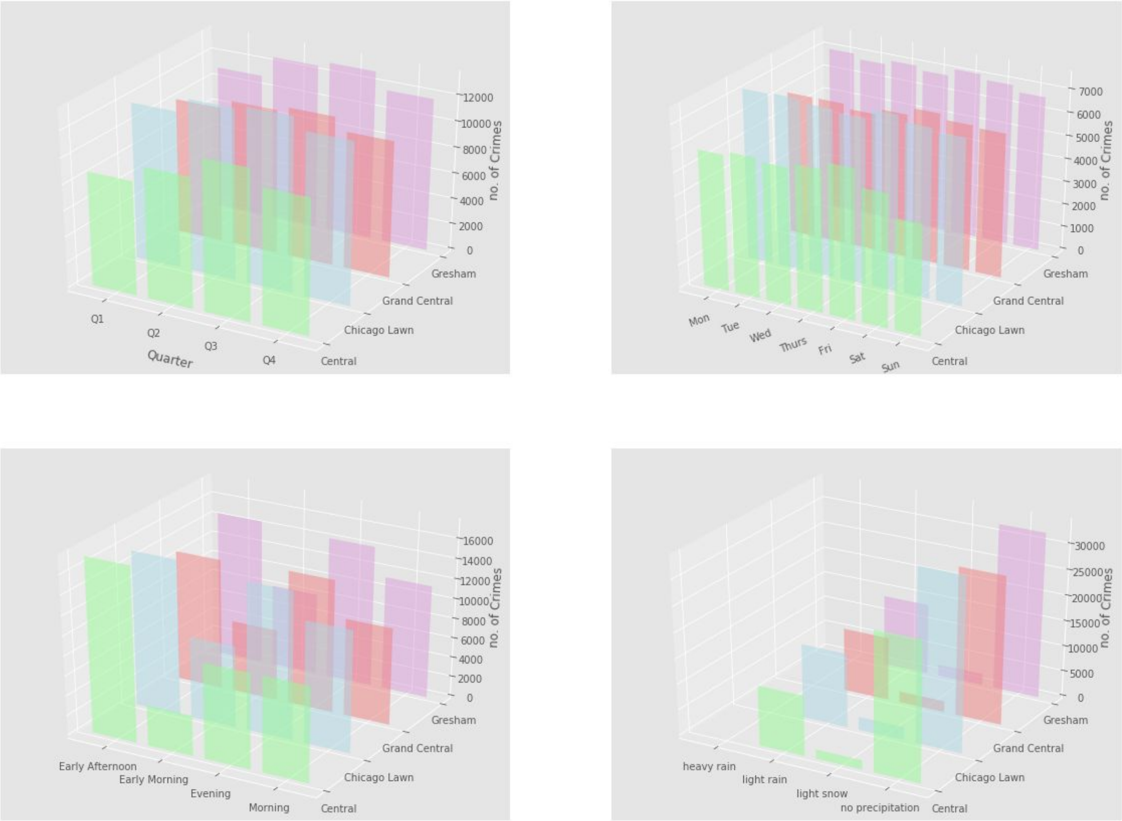


Figure 3.10:Crime occurrences in four district areas (Central, Chicago Lawn, Grand Central and Gresham) based on quarter, day of week, time and weather conditions.

4. Variation from the Mean

To understand the seasonality of crime we plotted the change of crime statistics for given month from a typical average month of the year.

Specifically, the plot is designed as follows

1. The monthly average number of thefts is calculated for each year as an average over all 12 months

2. Every month's theft statistics is then normalized as a percentage change from the above monthly average for a given year. For example, if in Feb of 2016, there were 120 crime count and the monthly average in 2016 was 100, then February’s normalized statistic would be 20%. This helps in studying monthly patterns over different years on the same graph.

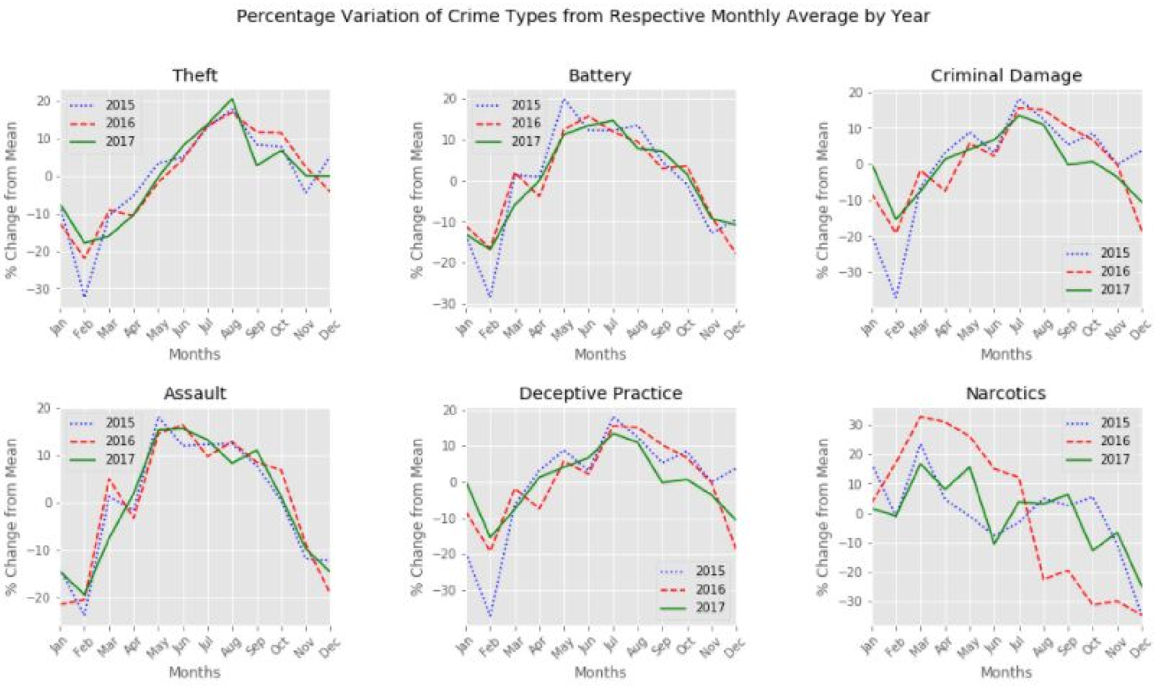


Figure 4.1:Percentage Variation of Crime Types from Respective Monthly Average by Year

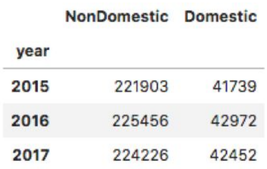
The crime pattern is consistent across all three years for the top 5 crimes. Narcotics in 2016 was abnormally high from March to June.

Various crimes peak at different months. Theft, criminal damage and deceptive practices peak in late summer whereas battery, assault and narcotics peak in late winter/early spring months. The correlation between the seasons (temperatures) and when the crimes peak seem to be strong.

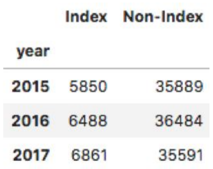
We had a hypothesis that crimes and especially thefts could increase during holiday seasons (November, December) when families travel , but the data doesn't bear that relationship. So, it can be concluded that theft does not increase substantially during the holiday season.

5. Domestic Crimes

Finally, we analyzed the trend of domestic crimes in Chicago. A comparison between domestic and non-domestic crimes shows that number of domestic crimes is about 1/5th of the other types of crimes. Within domestic crimes the number of serious (index) crimes is much lesser than the non serious (non index) crimes. However the number of serious crimes seems to be increasing over the years.



	NonDomestic	Domestic
year		
2015	221903	41739
2016	225456	42972
2017	224226	42452



	Index	Non-Index
year		
2015	5850	35889
2016	6488	36484
2017	6861	35591

The trend for Domestic crimes doesn't appear to be any different than the overall trend. The frequency is lesser earlier in the year and towards the end of the year and goes up during the summer months with February recording the lowest number of crimes every year. Interestingly, when comparing data across different years, the trends seem to be similar for alternate years. The domestic crime distribution across the various months of 2015 seems to resemble the ones for 2017 more than the ones for 2016.

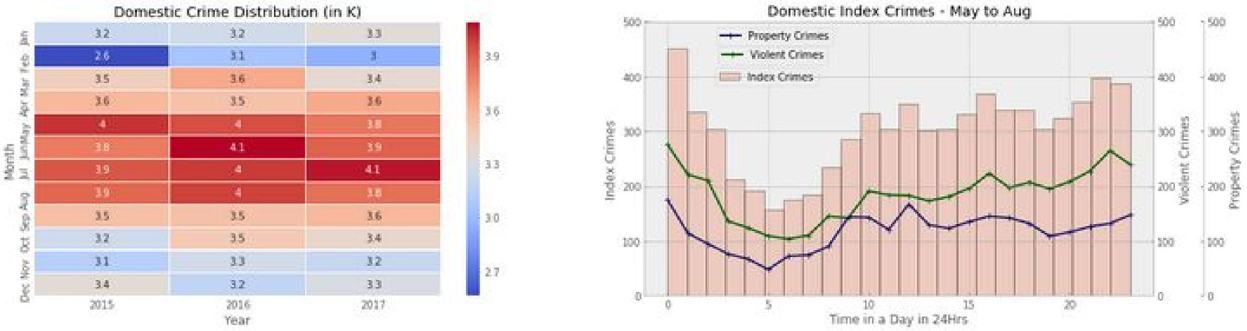


Figure 5.1: Domestic Crime Distribution

An analysis of the index crime patterns for the peak months (May to August) shows that these crimes are at their peak between midnight and 1AM after which the numbers go down considerably. The numbers are the lowest in the morning between 3 and 8 AM. Serious crimes are broadly categorized into property related and violent crimes. On analyzing the pattern of violent and property crimes, it is seen that the violent crimes are always more than property related crimes. However, at certain times of the day, their numbers are very close to each other.

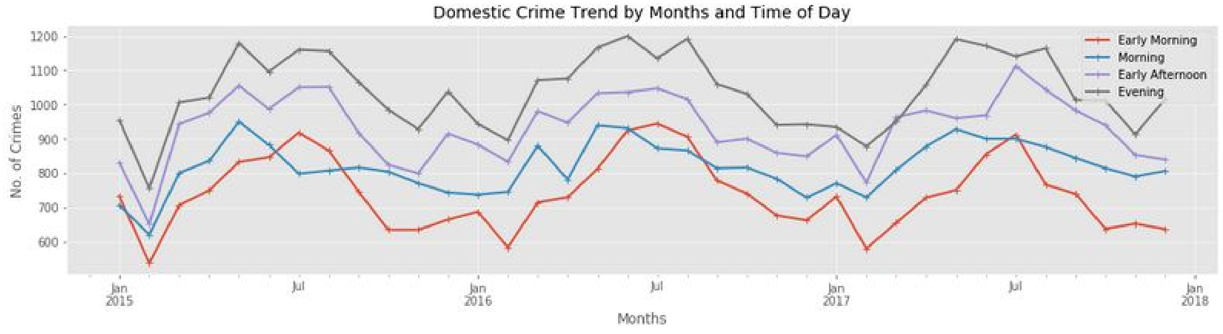


Figure 5.2: Domestic Crime Trend by Months and Time of Day

Domestic crimes are at their peak during evenings and are the least during early mornings. Usually the number of crimes goes up as the day progresses. However, during the summer months, the crimes committed during early mornings exceed the number of crimes committed before early afternoon.
It also appears that the rate decline in domestic crimes in February every year before picking up again in March.

Conclusions

We believe that we have answered all our research questions set forth in our project proposal, and managed to derive some unique observations on crime in Chicago from 2015 through 2017. Our answers and additional insights to our research questions are as follows:

1. Crimes rate remain quite constant from 2015 to 2017. However, through the investigation of crime occurrences over time, it is found number of crimes do appear to increase in the summer months from June through August and crimes are more often committed from the late morning 10am onwards to 10pm at night.

2. The study demonstrates that Theft is by far the most frequently committed crime, followed by Battery, Criminal Damage and Assault. Moreover, most of it is petty theft. Theft activity increases in the summer months, with the highest activity in the month of August. there is a decreasing amount of Narcotics crimes over the years, but tends to peak in March every year.

3. We had a hypothesis that crimes and especially thefts could increase during holiday seasons (November, December) when families travel , but the data doesn't bear that relationship. The analysis of theft crimes show that the % change of thefts varied very little from the mean during the month of December. So, it can be concluded that theft does not increase substantially during the holiday season.

4. In order to better study the crime rate in Chicago, Chicago weather dataset are combined with crime data to investigate the impact of weather on crime rate. It is found that crimes are highly related to mean temperature and are higher when the mean temperature is above 60 ^{o}F , which is consistent with the finding that crimes are more committed in Summer. Usually during warm weather, people head outside for activities and park is crowded with parades and music festival, which brings more opportunities for crime occurrence.

5. Using community areas variable in the data set, overall tendency for most crimes to be committed in the Chicago 'West Side'. 'Central' location is second highest, and this is likely attributed to it being the prime commercial area. Moreover, based on the analysis with police district area, Lincoln, Rogers Park, Albany Park and Morgan Park are the most safest district area to live, while police district areas, Gresham, including Grand Central, Chicago Lawn and Central, obtain high crime occurrence, indicating better police strength and more patrols are needed in this areas.

6. A comparison between domestic and non-domestic crimes shows that number of domestic crimes is about 1/5th of the other types of crimes. However Within domestic crimes the number of serious crimes is increasing over the years. Domestic crimes peak between 12 midnight and 1AM.

7. According to the study of firearm related crimes, it is better to stay away from streets: S Dr Martin Luther King Jr Dr. and S Ashland Ave ; as well as districts: Harrison and Englewood to keep safe.

8. In order to study the total crime and crimes of which arrests are made, the new term arrest ration is introduced. Unfortunately, the arrest ratio is continuously decrease from 2015 to 2017, indicating better police performance is needed.

9. Correlation between the type of crime and the type of location where they are committed is also investigated in this study. Larceny related crimes are mostly occurring in all communities in the Central side. Battery, Criminal Damage and Assault crimes are mostly in the community area 25 in the West side while Deceptive Practice is mostly happening in community area 8 of the Central side.

10. Fridays are the most crime ridden day of the week, with a consistently high incidence of Theft, Deceptive Practices and Narcotics. Many people are out and about on Fridays, and that could be a key factor influencing the spike in these 3 categories. Other crime types show peaks on different days of the week as well. These observations can help the Police to be vigilant for specific crimes on certain days of the week.

We believe that these findings provide some insights to our readers, and hope that it helps them stay a little safer while enjoying this wonderful windy city!