Chicago Crime Analysis

Abdullah Zaher, Anh Pham, Ziting Tang, Xueqi Lu

December 6, 2022

1 Introduction

Regardless of city, county, state, or country, crime is a negative aspect of society that affects a vast majority of people whether directly or indirectly. There are multiple datasets available that analyze different types of crimes in different regions of the world that may draw different conclusions. These types of crimes may differ in popularity depending on many factors of the local societies such as geographical location, socioeconomic status, technological advancements, strength of media coverage, and many more. For these reasons, data can be beneficial in tracking and analyzing the trends of each crime in each region for the sake of public information for increased precautions and potentially to help combat these trends in the correct manner depending on the respective trends. In this report, the preliminary objective is to gather crime data in a specific region, analyze and visualize it, and potentially relate it to other data that may seem very unrelated at first glance.

2 Data Description (Primary Dataset)

The primary dataset being analyzed is titled "Crimes - 2001 to Present | City of Chicago". This dataset consists of well over one-million crime reports (2001 - Present) with 22 features. Each report contains information on crime type and sub-type, date and time, address and coordinates, and result of the report (arrest or non-arrest). It is important to note that the crime types included in this dataset are all arrestable offenses in the specific year that they occurred (some laws may have changed over the years). Chicago is notoriously and statistically known for its high crime rates which is a local's nightmare but a data-scientist's gold mine because of the plethora of data available. The abundance of features/variables provides us with many routes and trends to analyze while utilizing different approaches and visualizations to draw up some questions and have the data answer them.

In our project, the following variables in the dataset have been frequently used.

Crimes Parameters			
Variables	Interpretation	Variables.1	Interpretation.1
Date	Date and time	Beat	Geographic Area of the City Broken Down for Patrol
Primary.Type	Crime Type	District	District Number
Description	Crime Description	Community.Area	Community Area Number
Location.Description	Crime Location	X.Coordinate	Geographic Coordinates of the crime
Arrest	Whether arrested	Y.Coordinate	Geographic Coordinates of the crime
Domestic	Whether domestic	Year	Year

Table 1: Crimes Parameters

For visualization purposes, we have included a few rows of data below:

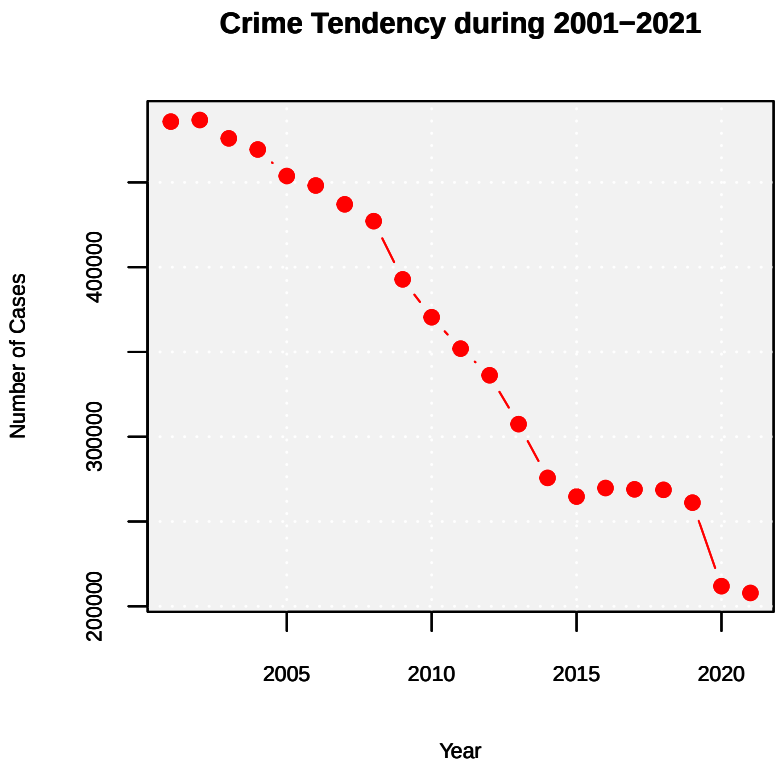
Few Rows of Dataset											
Date	Primary.Type	Description	Location.Description	Arrest	Domestic	Beat	District	Community.Area	X.Coordinate	Y.Coordinate	Year
9/5/15 13:30	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	FALSE	TRUE	924	9	61	1165074	1875917	2015
9/4/15 11:30	THEFT	POCKET-PICKING	CTA BUS	FALSE	FALSE	1511	15	52	1138875	1904869	2015
09/01/18 12:01	THEFT	OVER $500	RESIDENCE	FALSE	TRUE	631	6	44	1152037	1920384	2018
09/05/15 12:45	NARCOTICS	POSS: HEROIN(BRN/TAN)	SIDEWALK	TRUE	FALSE	1412	14	21	1141706	1900086	2015
09/05/15 13:00	ASSAULT	SIMPLE	APARTMENT	FALSE	TRUE	1522	15	25	1168430	1850165	2015

Table 2: Few Rows of Dataset

3 Questions/Hypotheses and Preliminary Conclusions

1. After test driving the dataset for the first time, the natural initial question is "Is overall crime trending upwards or downwards over the years?" Due to the increased reach of the media in the modern day, our expectation/assumption is that crime will be trending upwards with high confidence.

   (Method: Base R)

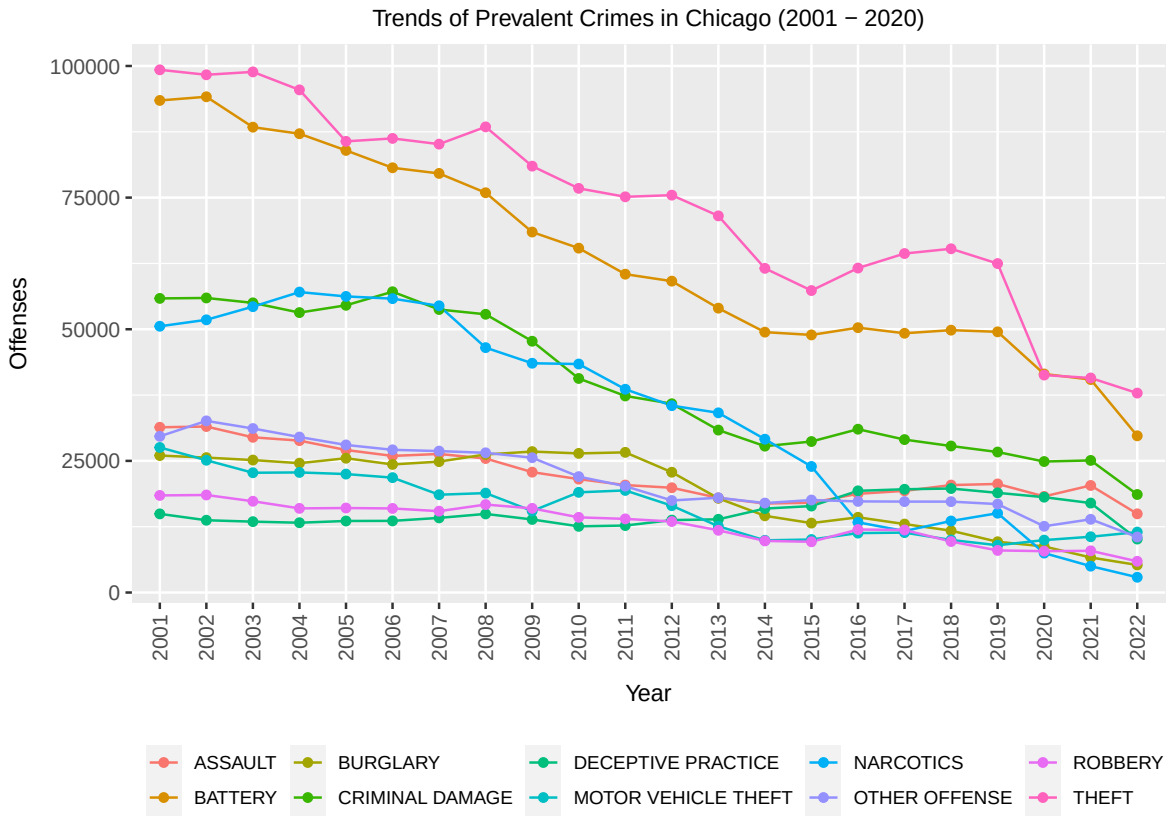


   To our surprise, the end result was not what we expected. Except for a slight increase from 2021 to 2002, the overall trend shows a downward trend year by year. The crime rate had a slow decline that started in 2002 until the first dip in 2009, and the dip continued into 2015. The crime rate leveled off from 2015 to 2019, then began to plummet in 2020, reaching the lowest crime rate in Chicago in 2021.

2. Naturally, our next instinct was to figure out what crimes are the most common in Chicago over the years. We now know that overall crime has been decreasing over the years in Chicago but knowing what

crimes are the most common could help us develop more questions to ask. Initially, our assumption was that Robbery and Theft-Related crimes were the most common since we all know friends or family that have experienced robberies.

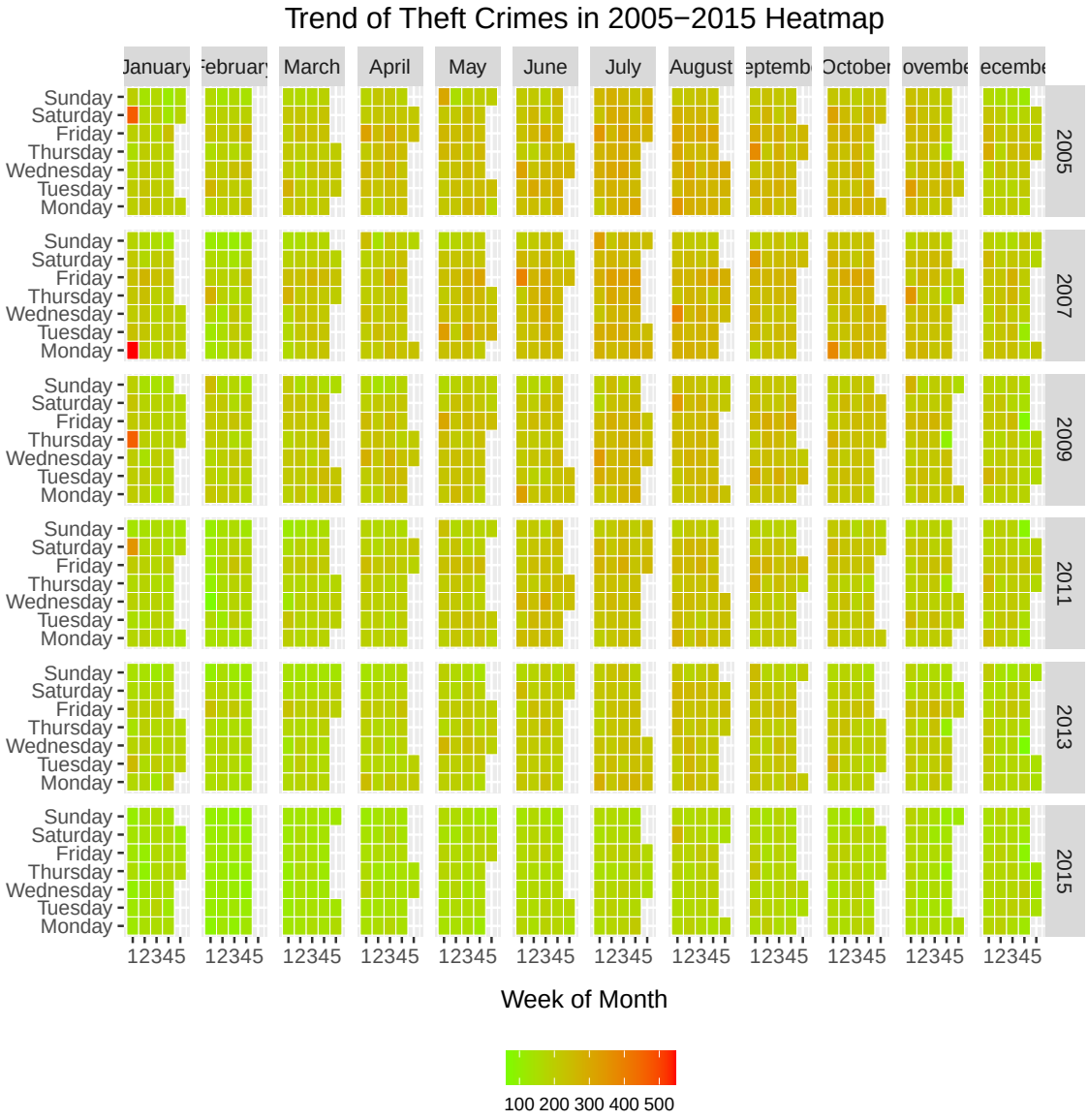
(Method: DPLYR)



The results supported our initial assumptions that theft-related crimes were the most prevalent. It can be seen from the graph that although the number of thefts has declined, it is still the most prevalent crime of all crime types. Even as the overall crime rate leveled off from 2015 to 2019, the crime rate for theft increased.

3. Our next question was geared towards date and time. Since we now know that theft is the most common crime, we wanted to know at what times of the year are theft crimes most common and whether or not time of year was correlated with the number of crimes committed. Our initial assumption was that crime was more likely to occur during holidays when people were distracted and away from their homes. The holidays that apply to this category are Valentines, Summer Vacation, Spring Break, and New Years Eve.

   (Method: DPLYR and GGPLOT)



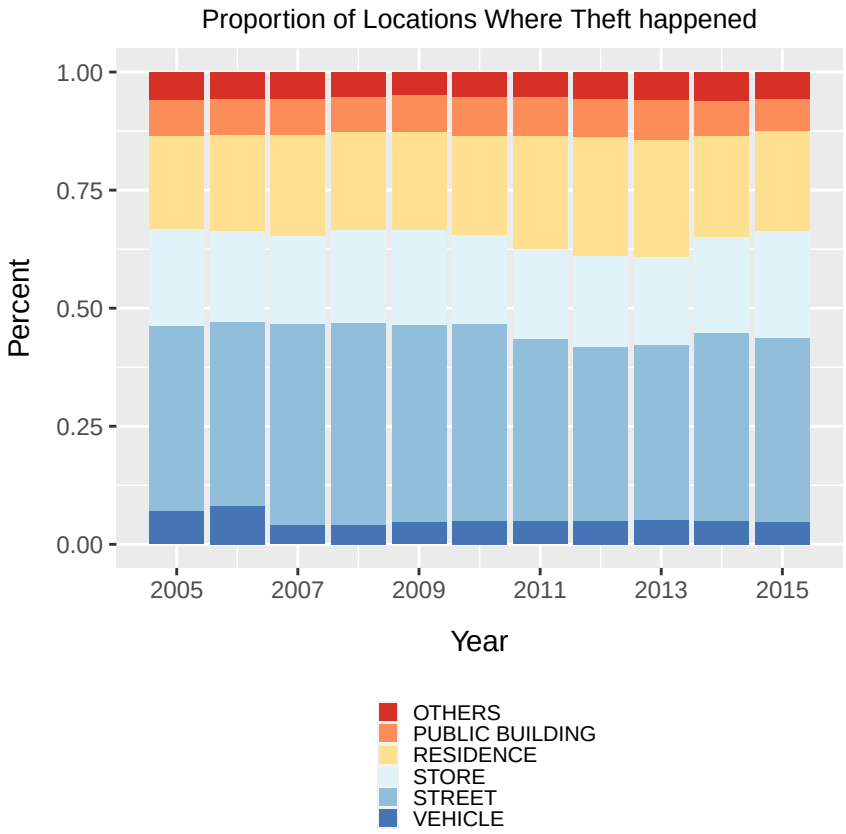
After creating the plot above, the result that stands out (dark red) is that theft is most prevalent around the New Years holiday, which is somewhat consistent with our assumption. Other than that day, the spread of theft crimes across the years did not seem to have a consistency but it was more common around the Spring - Summer timeframe. It seems as though thieves do not like to commit crimes during cold weather.

4. We have now dealt with the date and time aspect of theft. Next, we will discuss the location aspect and answer the following question: In which locations are theft-related crimes the most common. The results of the date and time analysis did not create a solid and concrete conclusion and that is what drove us towards the location aspect next. Due to the fact that there were over 200+ location categories in our dataset, we decided to create a few larger categories and group the sub-categories together to provide a cleaner result. Below is the split that we used for grouping for reference purposes (used RegEx for location replacement):

   - Residence :Residence, Apartment, Driveway, Condo, etc.,
   - Street: Street, Sidewalk, Alley, Park, Parking lot, Construction Site, Land, Highway, Bridge, Forest, Cemetery, Gas, Railroad etc.

- Store: Grocery, Retail, Restaurant, Store, Gas station, Bar, Barbershop, Shop etc.
- Vehicles: Bus, Taxi, Vehicle, CTA, Train, Car, Boat, Truck etc.
- Public Building: Hospital, School, Bank, Stadium, Fire Station, Police, Credit Union, Building, Theater, Hotel, Church/Synagogue/Place of Worship, College, University, Motel, Office, Factory, Airport, Garage, Warehouse, Pool, etc.
- Other: Other and Abandon.

(Method: SQL, GGPLOT, RegEx)



From the plot we can see the proportion of theft on the street is the highest which close to 40%. The proportion of theft from the residence building and store is relatively close both about 12.5%. There are fewer thefts in public buildings, followed by other and thefts from vehicles.

5. One particularly interesting feature in this data was the Arrest/No Arrest feature. This feature provides information on whether or not an arrest was carried out in that specific crime. Looking at the same location split as the previous plot, we decided to figure out the ratio of arrests-non arrests for thefts in those specific locations across the years. With the evident increase in security technology such as CCTV, home-intruder alarm systems, and sophisticated locking systems, our natural assumption was that arrests were taking place more often than not.

   (Method: SQL, GGPLOT, RegEx)

The results of the plot above were very shocking to us. The amount of non-arrests completely outweighed the amount of arrests in all but one location type (Stores). Even in the store category, it seems as though the number of arrests in the last 7-8 years has been declining, which is the overall conclusion. As years go by, the number of arrests has been declining and the number of non-arrests has been also declining in most cases but still outweighing the number of arrests by a very large margin. We would be very concerned if we were living in Chicago after realizing this.

4 Analysis with more datasets

1. Data Description (Secondary Dataset)

   It was difficult to find a Chicago-specific dataset and so we took a bit of a different approach. Instead, we searched for a secondary dataset that would show us national data in lieu of Chicago-specific data and our plan was to potentially filter out this data to get us what we wanted. Our secondary dataset is titled "Daily Temperature of Major Cities". The dataset description is very much stated in the name. The main columns are the City, Day, Month, Year, and Temperature. We used this dataset and filtered by the city of Chicago and the Years 2005 - 2015 in order to grasp the data that was going to be helpful in our analysis.

The following table shows the variables included in the secondary dataset and the explanation of each variable.

Temperature Dataset Parameters			
Variables	Interpretation	Variables.1	Interpretation.1
region	Continent name	Month	Month
Country	Country name	Day	Day
State	State name	Year	Year
City	City name	AvgTemperature	Average Temperature of the day

Table 3: Temperature Dataset Parameters

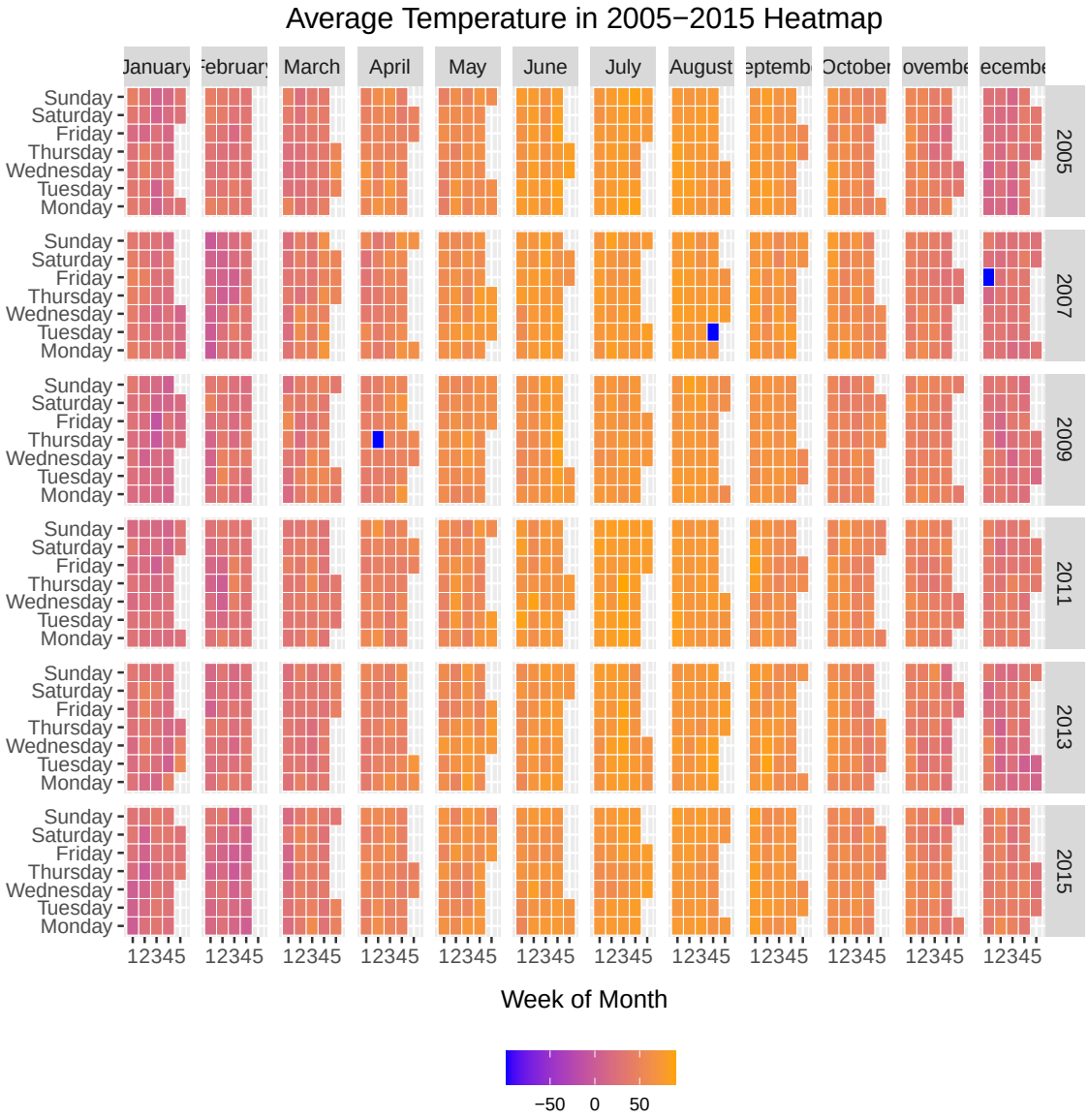
For visualization purposes, we have included a few rows of data below:

Few Rows of Dataset							
Region	Country	State	City	Month	Day	Year	AvgTemperature
Africa	Algeria	NA	Algiers	1	1	1995	64.2
Africa	Algeria	NA	Algiers	1	2	1995	49.4
Africa	Algeria	NA	Algiers	1	3	1995	48.8
Africa	Algeria	NA	Algiers	1	4	1995	46.4
Africa	Algeria	NA	Algiers	1	5	1995	47.9

Table 4: Few Rows of Dataset

2. Secondary Dataset Analysis (Questions/Hypotheses)

   Although we all have a general idea on which months are the warm months and which months are the cold months, we did not want to make any generalizations. We have the data and wanted the data to speak for itself. Similar to what we did with our primary crime dataset, we decided to build a calendar heat map across the years being analyzed to help visualize which months are the cold months and which are the hot months to possibly help connect crime with temperature. Though this sounds simple, it could help us tremendously to build a connection between the two datasets and draw up some initial conclusions.



As you can tell, the results were just as expected. The warmer temperatures were between the months of May and October and the colder months were between the months of November and April. Though it slightly varies year to year, this can probably be taken as the general consensus.

3. For our Killer plot, we decided to mix and match a few different variables together. The variables we chose were Crime Type (in this case, we are looking at theft), Number of Cases, Arrests/Non-Arrests, Temperatures, Months, Years, and variation of temperature from the median temperature of that specific year. We want to see from this graph whether the crime rate has a linear relationship with temperature.

   Firstly, we decided to split the plot into two halves. One half was for the analysis of the colder months and One half was for the analysis of the warmer months. The classification of the months (warm and cold) was solely dependent on the median temperature of that specific year (unlike the average temperature, choosing the median temperature can avoid the influence of individual extreme weather in the month or year). For a better visualization, we decided to draw a line to represent the year's median temperature and arrows next to each month visualizing the variation of that month's median temperature from the year's median temperature.

Next, we wanted to visualize the trend of theft cases across the cold months and across the warm months and to do that, we plotted the trend in theft cases on each half of our plot to better show the numbers. The scale was placed dead center between the two halves.

Lastly, we figured it would be interesting to analyze the percentage of Arrests/Non-Arrests during the cold months and the warm months. Instead of plotting it, we decided that in order for our plot to look nice and clean, we would display these percentages on the top of the plots. We later found out that the percentages year by year for Arrests/Non-Arrests were quite similar. Our Killer Plot of year 2010 and 2015 can be found below: (Method: Grid)

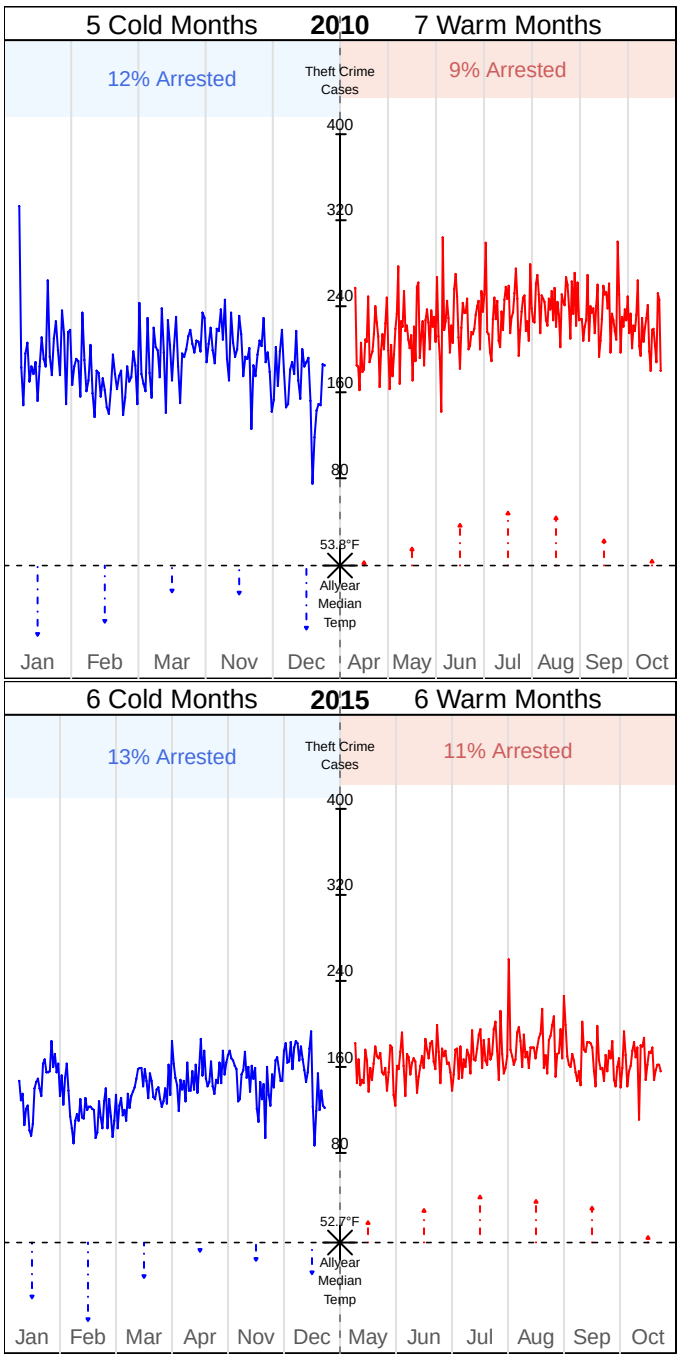


Figure 1: Temperature-Theft Crime: Cold VS Warm Month

From the plot we can draw two obvious conclusions:

1. Each of these killer plots shows that the overall temperature trend and the crime rate trend are basically consistent. We cannot clearly see the relationship between temperature and crime rate in 2015 in the previous heat map of time and crime rate but the killer plot shows it clearly. We can assume there is a linear relationship between temperature and crime rate and we will do a linear regression analysis on temperature and crime rate.

2. The arrest rate in the warm month is generally lower than that in the cold month. This may be because temperature has little effect on arrest efficiency but has an obvious effect on crime rate.

4. After discovering that temperature may have an effect on theft, we hope to conduct a more detailed analysis. We previously divided the 200+ location types in the data set into six categories: Residence, Street, Store, Vehicles, Public Building and other. We all know that the temperature in the public building in summer must be different from the temperature in the street, and the temperature in the day and night differ greatly. So we added more variables to analyze the relationship between the number of crimes and locations, time, temperature, etc.

   (Method: Grid)

   Basic Plot Explanation:

   Line: Each line represents a location.

   Center of circle: Since the number of cold months may not equal to the warm months so we first calculated monthly average crime cases for cold month and warm month respectively, and the center of the circle represents the ratio of the monthly average crime cases.

   Size of circle: Theft cases number of that location.

   Color of circle: Using average temp when theft happens V.S. average temp of this year. If hotter than average temp then it is red. Cooler then blue.

   Parameter: show-tieline

   ***Dotted line:** Located at the point that the ratio of the monthly average crime cases in cold month and warm month is 1:1.

   Parameter: show-locationAvg

   ***Dashed Circle (Unit Circle):** Average cases (all case/6 locations) and average ratio (cases happened cold month to warm month).

   Compare the solid circle with dash circle:

   If the center of solid circle not closed to the center of dashed circle, The influence of the time might be varied on the theft crime to the location.

   If solid circle is bigger, then a greater proportion of thefts occur at this location.

   Parameter: show-dayandnight

   Rate on the left of line: The percentage above the line represents the proportion of theft occurring during the daytime. The percentage under the line represents the proportion of theft occurring during night.

   When we use shinny to present this killer plot, we can see the different effects of temperature, location and time on the crime rate in different years.

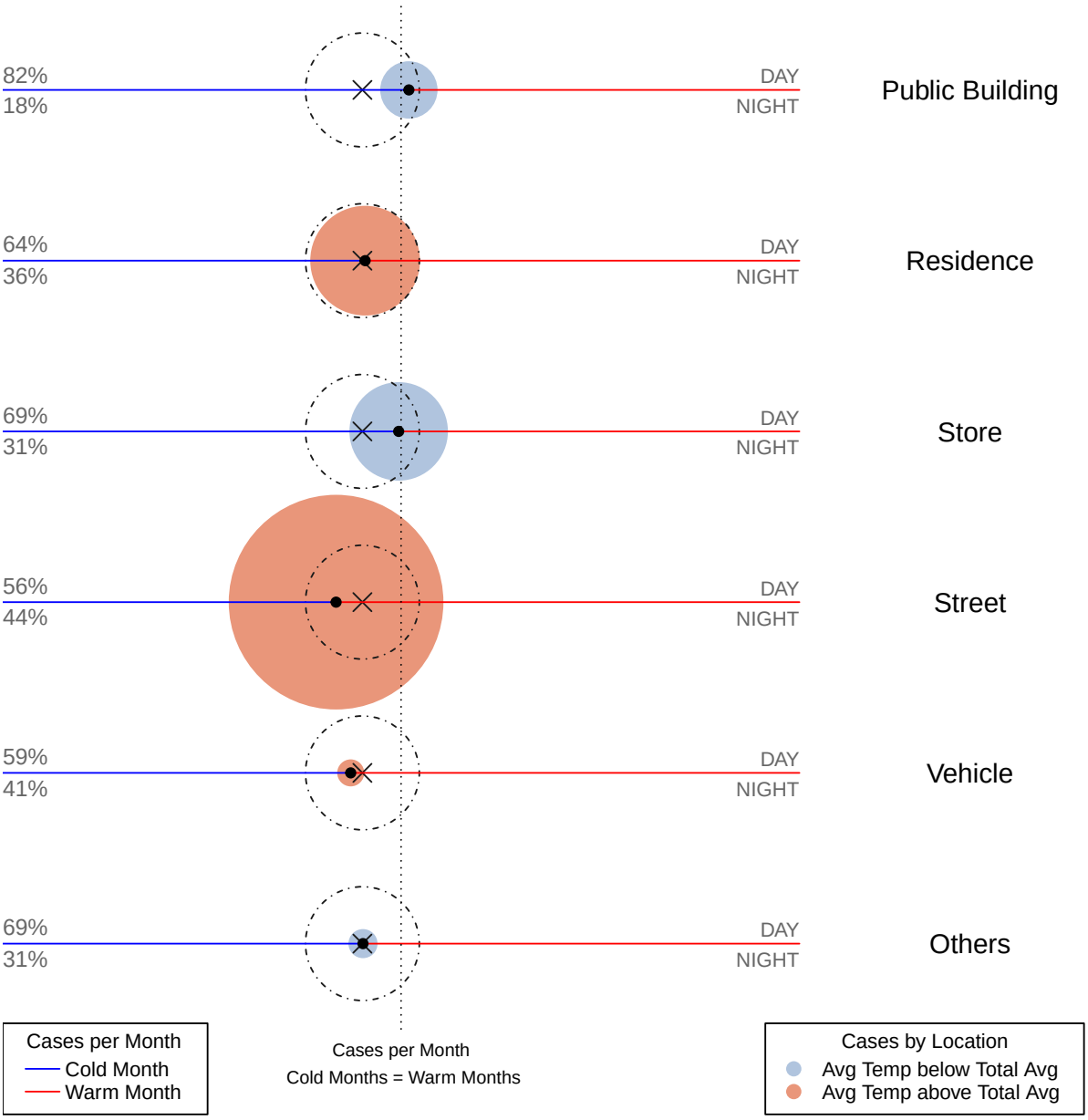


Figure 2: Theft and Related Factors in Six Location Groups

5. By observing the first two graphs, we really want to confirm what kind of linear relationship exists between temperature and the number of theft crimes, and whether we can predict the number of crimes through temperature. So we use linear regression to fit the data and first used the model

$$y = \beta_0 + \beta_1 x_1$$

where y represents daily theft crime cases, x_1 represents daily avgerage temperature. The following table shows the simple linear regression model result and it shows that there exist a strong relationship between the theft crime cases and temperature of that day.

Variables	Estimate	Std..Error	t.value	p.value
Intercept	153.09390	1.61271	94.93	$< 2e-16$
AvgTemperature	1.08130	0.02953	36.62	$< 2e-16$

Table 5: SLR Model Result

Since we introduced the concept of cold and warm month from the temperature-theft plot, we further design a dummy variable to analyze group difference of the temperature influence on cases happened in cold or warm month.Therefore we use the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where y represents theft crime cases, x_1 represents daily avgerage temperature and x_2 is a categorical variable which indicates whether this month belongs to hot month or not.

$x_2 = 1$ if the month belongs to hot month;

$x_2 = 0$ if the month belongs to cold month

The table shows all statistic values of the coefficients and we can clearly see that all parameters are significant.

Variables	Estimate	Std..Error	t.value	p.value
Intercept	155.12166	1.89758	81.747	$< 2e-16$
AvgTemperature	1.00202	0.04902	20.442	$< 2e-16$
hotmonthTrue	3.97317	1.96134	2.026	0.0429

Table 6: MLR Model Result

The diagnostic plots show that our model fit the data perfectly and our data follow the normal assumption.

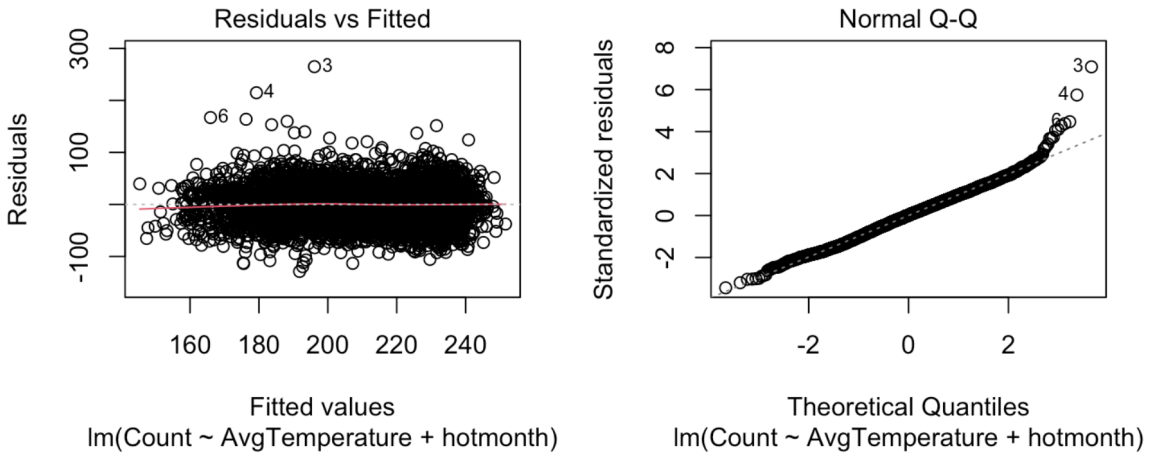
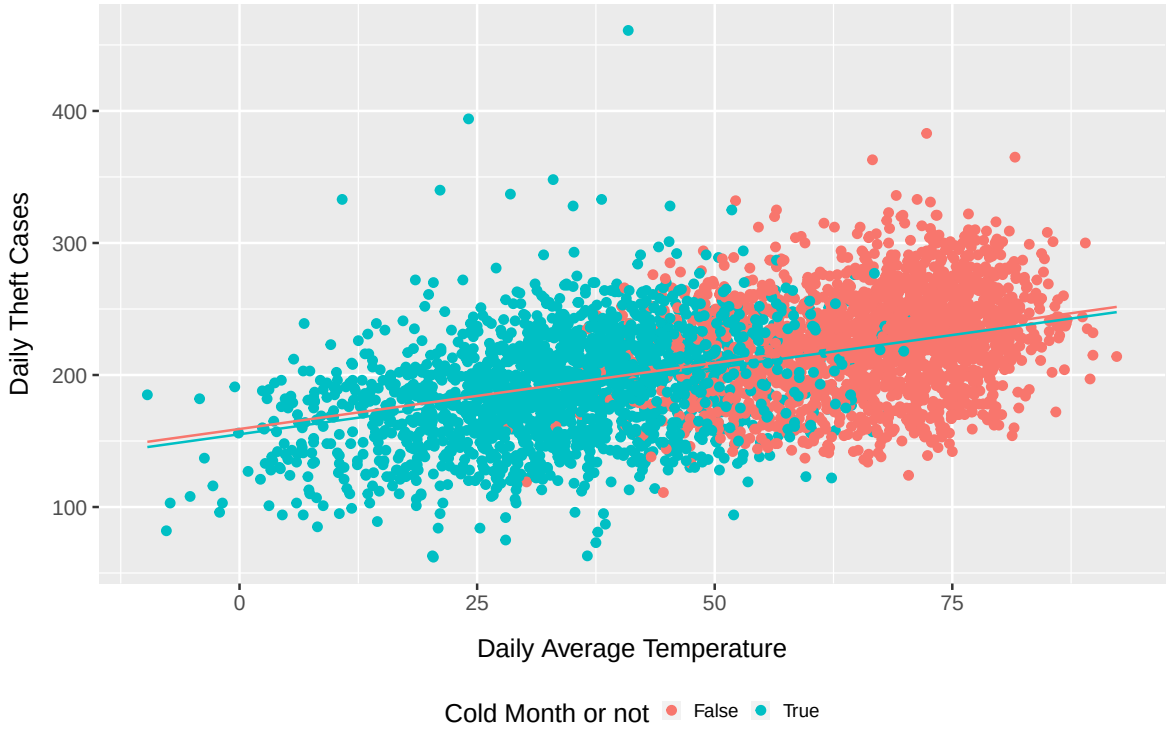


Figure 3: Diagnostic Plot



From the linear regression plot we can see that temperature has a great influence on theft crimes, the higher the temperature, the more crimes. The slope of the model is slightly different between cold and warm months, which means that the effect of temperature on crime is somewhat different between cold and warm months.

5 Conclusion

Through this entire analysis journey, we were able to draw up a few conclusions to the main question we have asked. Is theft crime rate directly correlated with temperature in Chicago? The short answer would be

yes. In the data we were given to analyze, the theft rates and the temperature showed somewhat of a linear relationship. As temperature increases, theft rates as a whole increase as a result. We have also found that more arrests are made in the colder months than in the warmer months.This tells us that law-enforcement is more equipped to handle the cold weather of Chicago than the criminals. Realistically thinking, any type of theft would seem to be more difficult in the winter months than in the summer months. In fact, the weather in the Windy City shows no mercy to anyone walking the streets during the cold months and that could very well be a contributing factor. One surprising detail that our killer plot uncovered was that theft occurred mostly during the day rather than at night which went against our preconceived notions. Based on the analysis we have made, the advice we would give to the general public of Chicago is to be on the lookout for theft especially during the warmer months when the weather is not a roadblock or contributing factor that could make theft significantly more difficult.

6 Reference

The primary dataset of Chicago Crimes (2001-Present) was extracted from the Chicago Data Portal provided by Chicago Police Department which is a government work not subject to copyright. The secondary dataset of Temperature Data Files for 157 U.S. and 167 International Cities is sourced from the National Climatic Data Center and collected by University of Dayton in their Environmental Protection Agency Average Daily Temperature Archive.

[1] Chicago Data Portal - Crimes (2001 to Present), https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

[2] University of Dayton - Environmental Protection Agency Average Daily Temperature Archive, http://academic.udayton.edu/kissock/http/Weather/default.htm