Chicago Crime Reformatter

André Duarte

September 1, 2017

1 Introduction

It is often the case that an analysis company has a specific layout for their internal databases. Therefore, the first step when trying to incorporate new data (online or elsewhere) usually resides in reformatting the data into the needed layout. In this exercise, I will show how we can achieve this process in ECL, using as an example the Chicago Crime Data (link) from 2001 to the present. In the subsequent paragraphs, I will show how to import the data without loss of information, how to reformat the data into atomic fields, and how to adapt it into the three specific layouts provided. In addition, I will also show a few examples of analytics that can be performed in ECL on this data in order to answer some insightful questions about crime in Chicago in the 21st century.

The dataset we download from the above website is managed and provided by the Chicago Police Department, and contains clean information about incidents of crime in the city. There are more than six million rows, where each one represents a crime that occurred. Several fields give us anonymized (to an extent) information about the particular crime, such as time and date, type, description, location (anonymized block, district, beat, and ward), whether the crime was considered domestic and whether the criminal was arrested.

Recently, Chicago has been under media scrutiny for its increasing violence. In particular, president Trump has used this information in several of his speeches during his campaign, and continues to do so. It has been reported that 2016 was the bloodiest year in two decades in the city, with a massive increase in murders (link). As such, the second objective of this exercise is to provide some data-driven insights into this issue, and investigate if such statements are correct or manipulated.

2 Reformatter Exercise

On the day I downloaded the dataset, there were 6,412,313 records, which I downloaded as a CSV file. After uploading the file to the landing zone of HPCC and spraying it appropriately, we can switch on over to ECL in order to load it and start our reformatting exercise.

2.1 Layouts

I decided to create a module called `modLayouts` to include each of the major layout used in this exercise. These include the raw layout (`lRaw`), the clean layout (`lClean`), and each of the three provided layouts for crime (`lCrime`), address (`lAddress`), and the relation between the two (`lCrimeAddressRelation`).

The raw layout defines each field as a variable sized string, so that no information is lost. For the clean layout, the value types are defined in the exercise proposal. Inside the module, this corresponds to:

```
EXPORT lRaw := RECORD
  STRING ID;
  STRING Case_Number;
  STRING Date;
  STRING Block;
  STRING IUCR;
  STRING Primary_Type;
  STRING Description;
  STRING Location_Description;
  STRING Arrest;
  STRING Domestic;
  STRING Beat;
  STRING District;
  STRING Ward;
  STRING Community_Area;
  STRING FBI_Code;
  STRING X_Coordinate;
  STRING Y_Coordinate;
  STRING Year;
  STRING Updated_On;
  STRING Latitude;
  STRING Longitude;
  STRING Location;
END;
```

and

```
EXPORT lClean := RECORD
  UNSIGNED     Row_ID;
  UNSIGNED     ID;
  STRING       Case_Number;
  UNSIGNED4    Date_Crime;
  UNSIGNED4    Time_Crime;
  UNSIGNED     Address_Number;
  STRING       Address_Street;
  STRING       Address_City;
  STRING2      Address_State;
  STRING       Address_Country;
  UNSIGNED     Zip_Code;
  UNSIGNED     IUCR;
  STRING       Primary_Type;
  STRING       Description;
  STRING       Location_Description;
  BOOLEAN      Flag_Arrested;
  BOOLEAN      Flag_Domestic;
  UNSIGNED     Beat;
  UNSIGNED     District;
  UNSIGNED     Ward;
  UNSIGNED     Community_Area;
```

```
    UNSIGNED      FBI_Code;
    UNSIGNED      X_Coordinate;
    UNSIGNED      Y_Coordinate;
    UNSIGNED      Date_Crime_Year;
    UNSIGNED      Date_Crime_Month;
    UNSIGNED      Date_Crime_Day;
    UNSIGNED      Time_Crime_Hour;
    STRING        Date_Crime_Season;
    UNSIGNED4     Date_Updated;
    UNSIGNED4     Time_Updated;
    DECIMAL15_12  Latitude;
    DECIMAL15_12  Longitude;
    STRING        Location;
  END;
```

2.2 Transforming the data

To perform this reformatting, I wrote some transform actions in the file modTransformFunctions and some macros to parse the date and address in the file modMacros. Here, I will go into a little detail about a few interesting things I did in ECL.

First, in order to include a field with the rowID, I chose to use the COUNTER option in the transform action. This attributes a new integer value for each row that is processed. Doing this for each of the datasets allows us later to identify the originating row, the crimeID and the AddressID to relate the tables if necessary.

The macros parseAddress and parseTime are worth mentioning and transcribing here (below). Inside each macro, instead of returning a single value, I chose to return a module with several fields instead. As such, I bypass the need for several small functions to get the year, month, day, etc and only have one, which I can then "dive into" to retrieve more granular data. While we can extract the date as an UNSIGNED using the Std library, the fact that the time is in a 12h format (with AM or PM disambiguation) does not let us use the equivalent function in this library for the time, so I had to manually create one. Ideally, regular expressions would be more suited for this task, but the data set is well formatted already, so we can hardcode these numbers.

```
EXPORT parseDate(infield) := FUNCTIONMACRO
  uDate := STD.Date.FromStringToDate(infield[1..10], '%m/%d/%Y');
  uTime := (UNSIGNED4) (IF(infield[21..22] = 'AM',
                           infield[12..13],
                           (STRING) (((UNSIGNED) infield[12..13]) + 12)
                          ) + infield[15..16] + infield[18..19]);
  uYear     := (UNSIGNED) infield[7..10];
  uMonth    := (UNSIGNED) infield[1..2];
  uDay      := (UNSIGNED) infield[4..5];
  uHour     := (UNSIGNED) IF(infield[21..22] = 'AM',
                             infield[12..13],
                             (STRING) (((UNSIGNED) infield[12..13]) + 12));

  mod := MODULE
    EXPORT FullDate := uDate;
    EXPORT FullTime := uTime;
    EXPORT Year     := uYear;
    EXPORT Month    := uMonth;
    EXPORT Day      := uDay;
    EXPORT Hour     := uHour;
    EXPORT Season   := (STRING) IF(uMonth BETWEEN 3 AND 5, 'Spring',
                                   IF(uMonth BETWEEN 6 AND 8, 'Summer',
                                      IF(uMonth BETWEEN 9 AND 11, 'Fall',
                                         'Winter')));
  END;

  RETURN mod;
ENDMACRO;

EXPORT parseAddress(infield) := FUNCTIONMACRO
  mod := MODULE
    EXPORT Number  := (UNSIGNED) infield[1..5];
    EXPORT Street  := infield[7..];
    EXPORT ZipCode := (UNSIGNED) 0;
  END;

  RETURN mod;
ENDMACRO;
```

Since all the information in the dataset is now clean, it is a simple matter of projecting into the given layouts for Crime, Address, and the relation between the two (while adding an identifier row for each, using the COUNTER attribute). This is accomplished in the BWR_02_MakeTables file.

2.3 Inconsistencies, missing fields, wrong formats

While performing this data cleansing exercise, I noticed some inconsistencies in the formats provided. First, the blocks in the dataset are anonymized, meaning that the address numbers are not shown in their full extent. Indeed, the last two digits of the five provided are always replaced by 'X'. Therefore, when converting to an UNSIGNED, we lose and mangle that information.

Similarly, the IUCR codes are converted to UNSIGNED, but some have a letter at the end (such as 141A, 141B, and 141C), which have the same primary type, but different descriptions. This information is lost when converting the IUCR to simple integers. The same thing happends with the FBI codes (for example 08B).

Dates and times are to be conserved in yyyymmdd and hhmmss formats as UNSIGNED4 fields. Although this works for the dates, the time format becomes hmmss if there is a leading 0 (i.e., if the time is before 12PM). It could be preferable to store these in strings.

Finally, some required fields in the final layouts are missing from the raw data. While we can complete the city, state, and country to 'Chicago', 'IL', and 'USA' respectively, it is not trivial to input the zip codes for each crime. We would either need a map of address/beat/ward/district to zip codes or extra data to complete the forms. Similarly, we currently do not have information about the victims, such as gender and name. These fields therefore all remain blank.

2.4 Conclusion

Reformatting and cleaning data is an important part of any data retrieval project. Sometimes the layout that is currently used does not fully satisfy the needs of the new, incoming data, and several methods of retaining the maximum amount of data are possible. In other cases, it is simply non viable. In the end, for this assignment, it was possible to keep the overall information, albeit some data could not be imputed.

3 Data Exercise

Now that we have imported, cleaned, and reformatted the Chicago Crime Data, we can focus on performing some data analysis on it. Mainly, I have focused on answering the proposed questions using ECL queries. I also used the output tables to create plots and graphs in R in order to better visualize these answers.

3.1 BWR Files

I tried to keep my code organized in such a way that individual Builder Window Runnable (BWR) files should be run in sequence. Therefore, I adopted the naming convention `BWR_XX_Purpose`, where 'XX' is a two-integer number that defines the order in which the BWR files should be run, and 'Purpose' is a simple string that defines what the BWR file does or answers. In such fashion, if someone else gets my code, they will intuitively understand that they must run `BWR_01_CleanData` first, then `BWR_02_MakeTables`.

BWR files 03 to 13 create the output to answer the provided questions. BWR file 99 outputs all the tables (raw, clean, crime, address, relation) for easy visualizing of the formats and contents.

3.2 Data Analysis and Answers

Here, I will try to answer the questions posed in the document using the available data at our disposal. I grouped several questions together in some answers to give a more complete look at the issue.

Inside the `modMacros` file, I wrote three macro functions to help perform a grouping and counting analysis. `macCountSimple` performs a count by grouping a single field, `macCountDouble` does the same with two fields, and `macCountTriple` with three. I believe this task could be simplified into a single macro, where the user can specify a comma-separated list of fields, and the macro will automatically create the output record layout dynamically based on how many fields were passed. Unfortunately, I could not find a way to do this, hence why the three functions are very similar but separate.

3.2.1 Evolution of crime in Chicago since 2001

Using the macros described above, I created a table counting the total number of crimes per year and month. Also, I did the same analysis for arrests, using the `Flag_Arrested` field. These two tables can easily be joined, either in ECL or in R (I eventually did it in R for plotting), using the year and month combination to join records. The resulting plot is shown in Figure 1.

This figure summarizes many different pieces of information. First, we can see that the total number of crimes and arrests both show a decreasing trend since 2001, maybe stagnating after 2015. While there were between 35 and 45 thousand crimes per month in 2001, this number is between 20 and 25 thousand in 2016. For arrests, the numbers went from approximately 12 thousand to 4

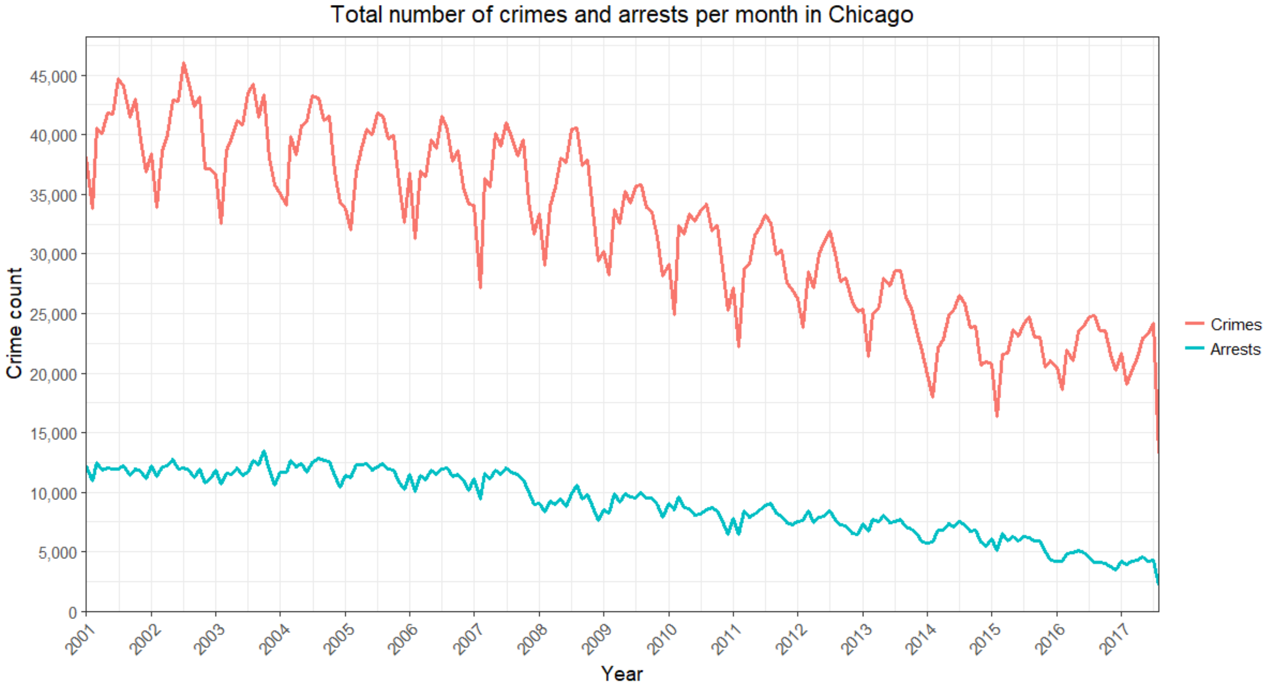

Figure 1: Monthly-aggregated crimes and arrests in Chicago from 2001 to present

thousand in the same period. At first glance, this seems very promising for Chicago, since less crime is happening in the city year after year. But this makes us wonder why president Trump and the media keep saying that Chicago is an increasingly violent city. We will break this down later.

Another information that this plot presents is that crime in Chicago presents a clear seasonality. Indeed, there is a peak in crime in the summer every single year, while the lowest point yearly is reached in the winter.

3.2.2 Breakdown of most common crimes

Something that is interesting to investigate is how each type of crime has varied through time. Figure 2 presents a histogram of the 20 most popular types of crime in the city in this period. We can see that theft, battery, criminal damage, and narcotics top the list. Indeed, these four types of crime comprise over 60% of the total through the years. Similarly, Figure 3 shows the yearly contribution to total crime, in percentile points, of each of these top 20 most prevalent crimes.

We can look at the evolution of each of these through the years, in order to see whether some crimes are increasing in detriment of others. Figure 4 shows the overall monthly variation of each type of crime. It can be seen that most types of popular crimes have steadily decreased with time, such as theft, battery, and criminal damage. This drives the overall crime rates of the city to decrease year over year. However, it seems that most of these reach a plateau in recent years, or even show some indications that they may be increasing again. In addition, less prevalent crimes are increasing, and some alarmingly fast. Arson, homicide, interference with a public officer, obscenity, and weapons violation all show upward trends. In particular, homicides have increased threefold since 2014, and weapons violations doubled in a few years.

To corroborate this last point, Figure 5 presents the year over year change for each primary type of crime. This plot allows us to see whether each crime has increased or reduced compared to itself the previous year. What we can see is that many crimes (18 out of the 33 total) present an increase over themselves in 2016 relative to 2015, including many of the popular crimes.

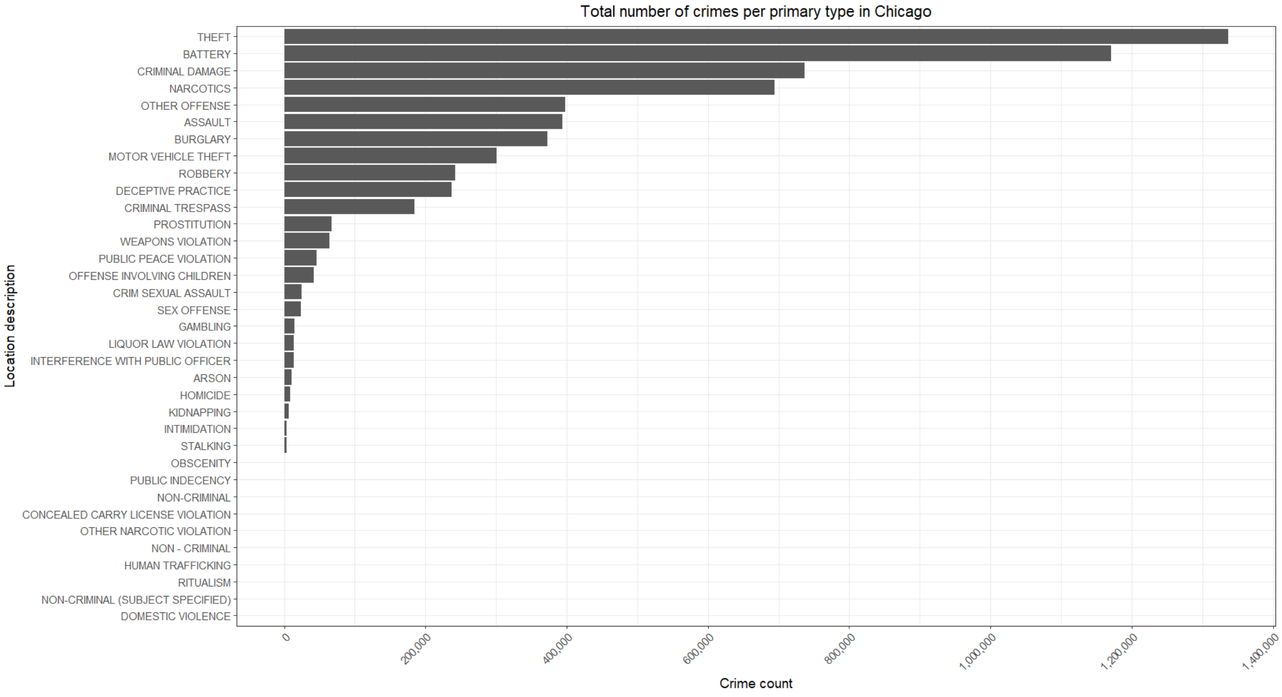

Figure 2: Most popular types of crime in Chicago since 2001

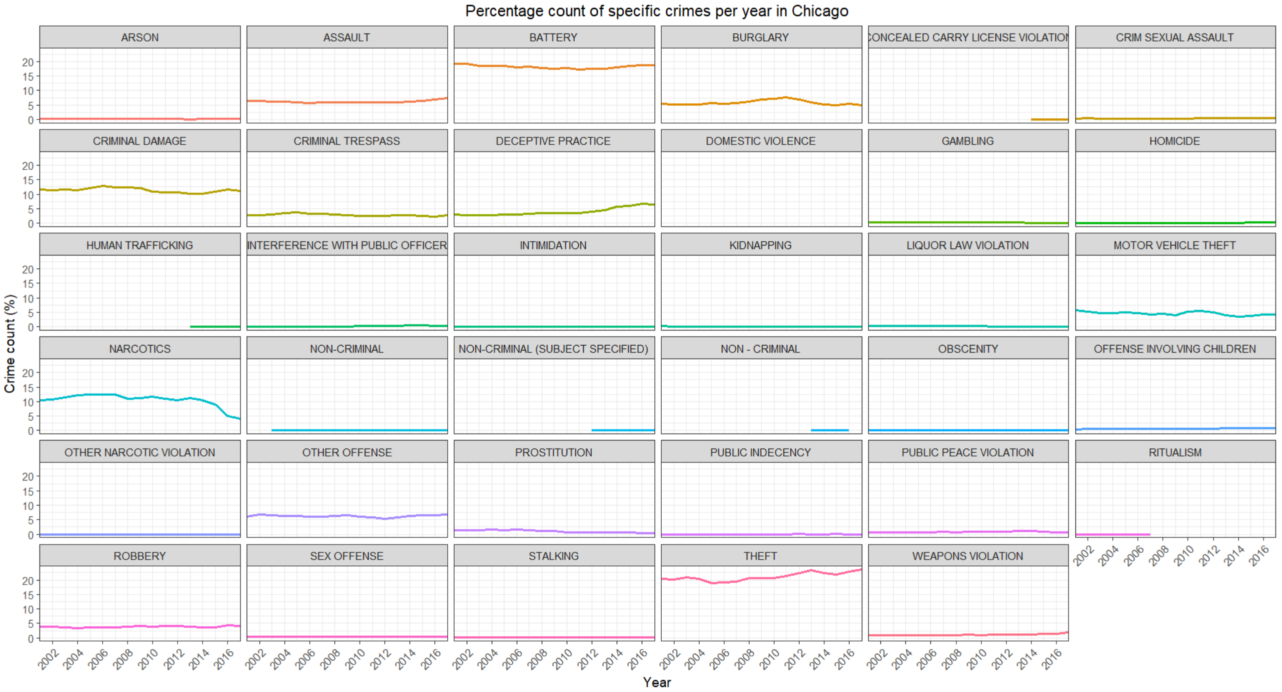

Figure 3: Yearly percentual contribution of each crime type (top20) on the total in Chicago since 2001

So what conclusion can be reached from analyzing these figures? Although crime has decreased in Chicago since 2001, it has seemed to reach a plateau around 2015 onwards. While some crimes have decreased year over year by big margins from 2015 to 2016 (around 40% for narcotics, gambling, and prostitution for example), others have increased in the last two years, including the most popular crimes. Additionally, important crimes like homicides have seen a jump of almost 60% from 2015 to 2016, which is very serious.

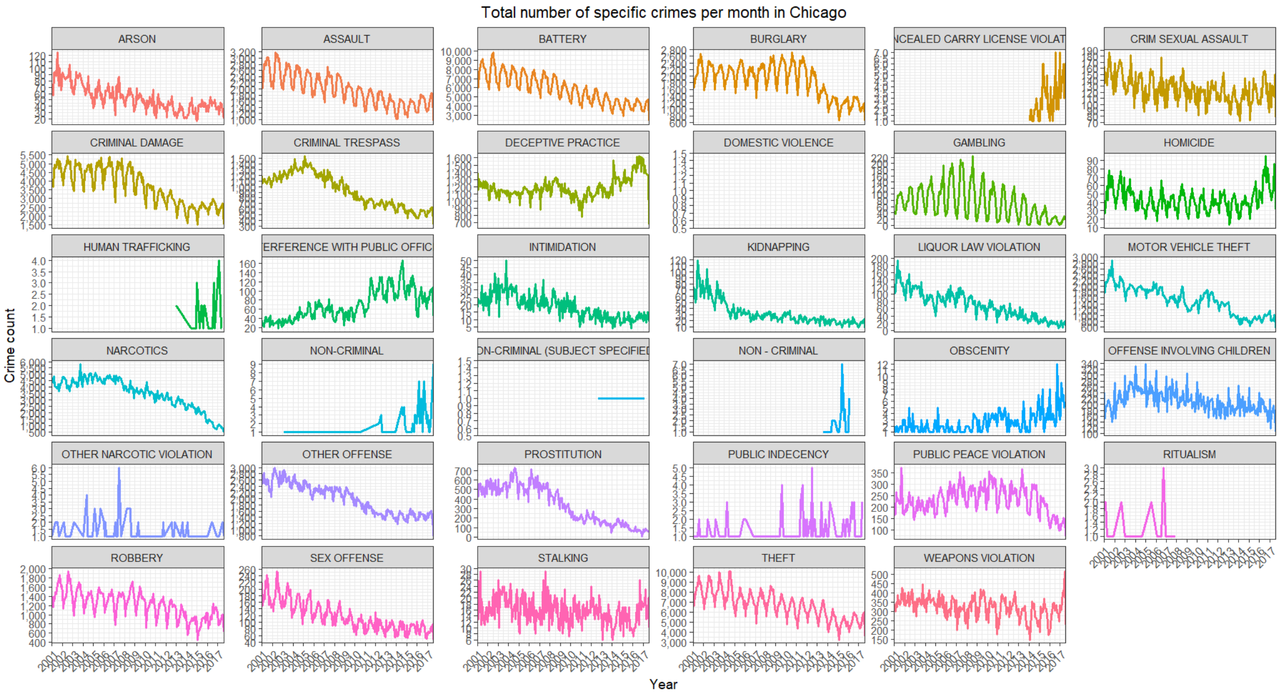

Figure 4: Evolution of each primary crime type in Chicago since 2001

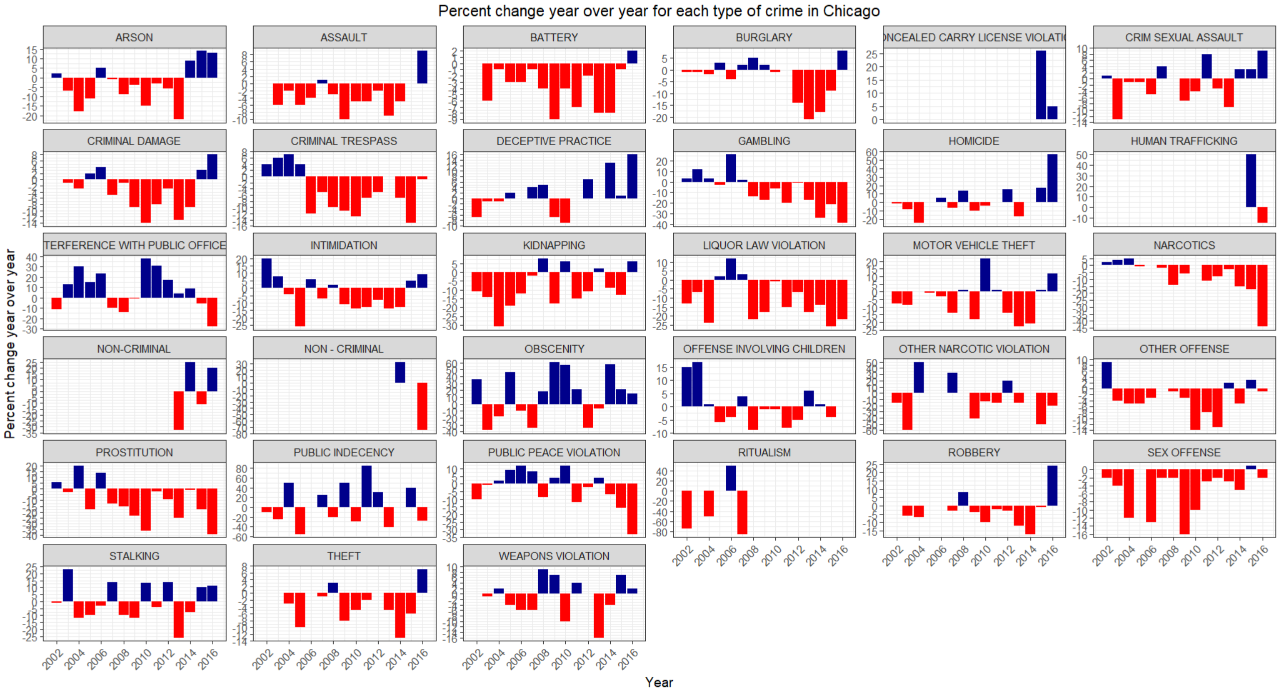

Figure 5: Percent change year over year for each type of crime in Chicago since 2001

3.2.3 The particular case of homicides

The case of homicides is important to look at, since it is a very serious crime. Figures 6 and 7 are taken from the previous plots to show more detail. It can be seen that, whereas homicides stayed fairly constant from 2005 to 2014, an incredible increase happened in 2015 and 2016, reaching a 16-year high in the summer of 2016, when more than 90 homicides were registered in July alone. Clearly, this is a problem in Chicago in recent years. According to the police department, this is largely due to an increase in gang presence and violence in the city in recent years.

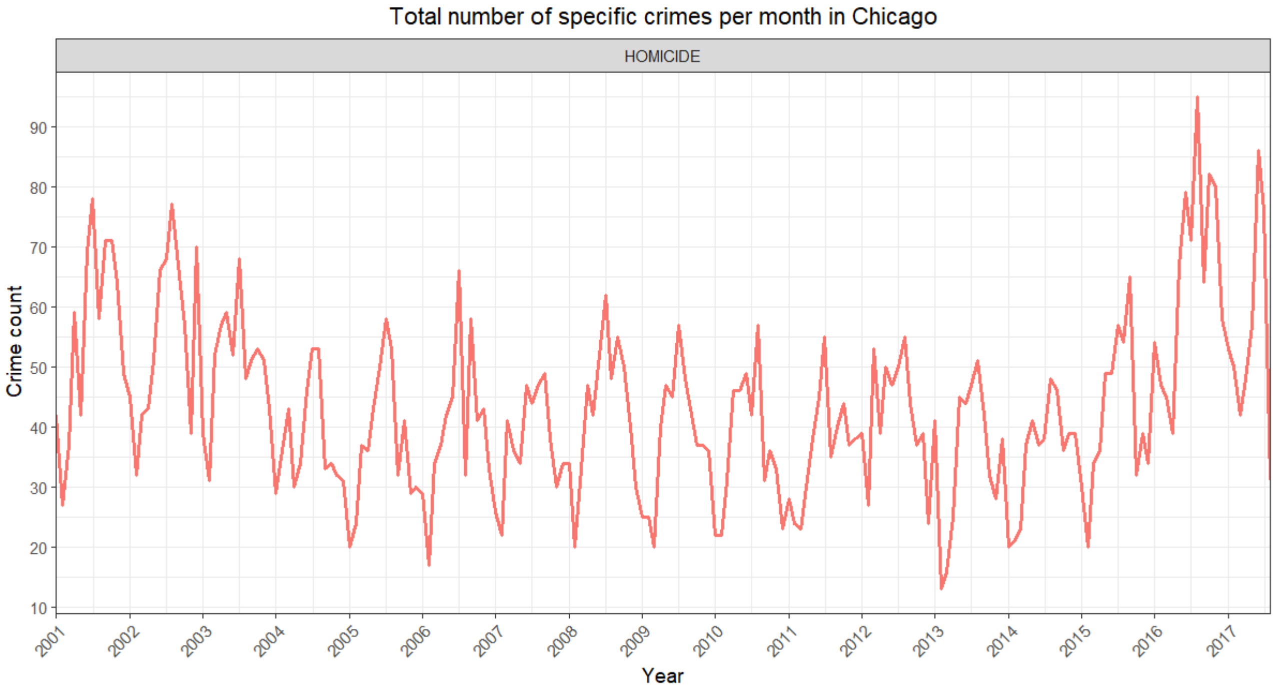

Figure 6: Number of homicides in Chicago since 2001

3.2.4 Crime on a day-to-day basis

While looking at crime month by month is important to assess the overall trend of crime in the city, we can also take a look at crime throughout the day. Indeed, the data set is granular enough to provide a fairly precise time when the crime happened (it is often an approximation). Figure 8 plots the total number of crime by hour of the day. It can be seen that this follows a wave pattern, from a minimum around 5AM reaching a maximum at around 7-8PM. The exact time of a crime often being unclear, this could explain the peak observed at 12PM, and the decrease at 11PM (i.e., rounding the time of crime to 12PM and 12AM). It seems that crime follows the same regular schedule as a working person.

So it seems that it is safer to stay at home during the end of the day and early night. Is it possible to say whether it is more likely to find a burglar at home though? Figure 9 presents the locations of crime by hour of day. Again, it is somewhat unclear what the distinction between residence and apartment is, or even between street and sidewalk. It seems to me that it would vastly depend on the person who is inputing the crime in the database. However, this plot does not show any massive discrepancy with the previous graph, albeit we could potentially note an increase in crime in residence at 12AM. This would potentially make it more likely to find a burglar at home at that time. Indeed, Figure 10 shows a spike in burglaries at this time. However, it could be due to time rounding, as mentioned previously.

3.2.5 Hotspots of crime

As I mentioned before, the data in this set is anonymized to a certain extent: the last two digits of the address number is masked, and we don't have the corresponding zip code without resorting to external help. This makes it harder to draw hotspots of crime in Chicago, where crime occurs more

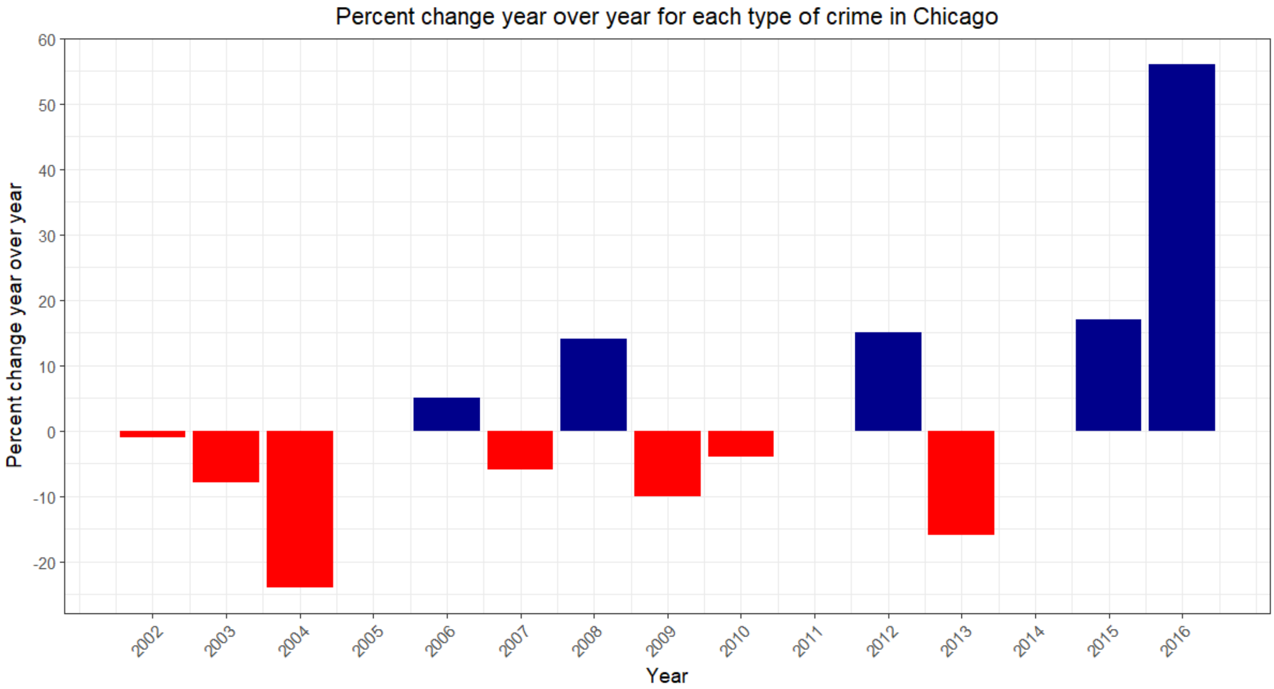

Figure 7: Percent change year over year for homicides in Chicago since 2001

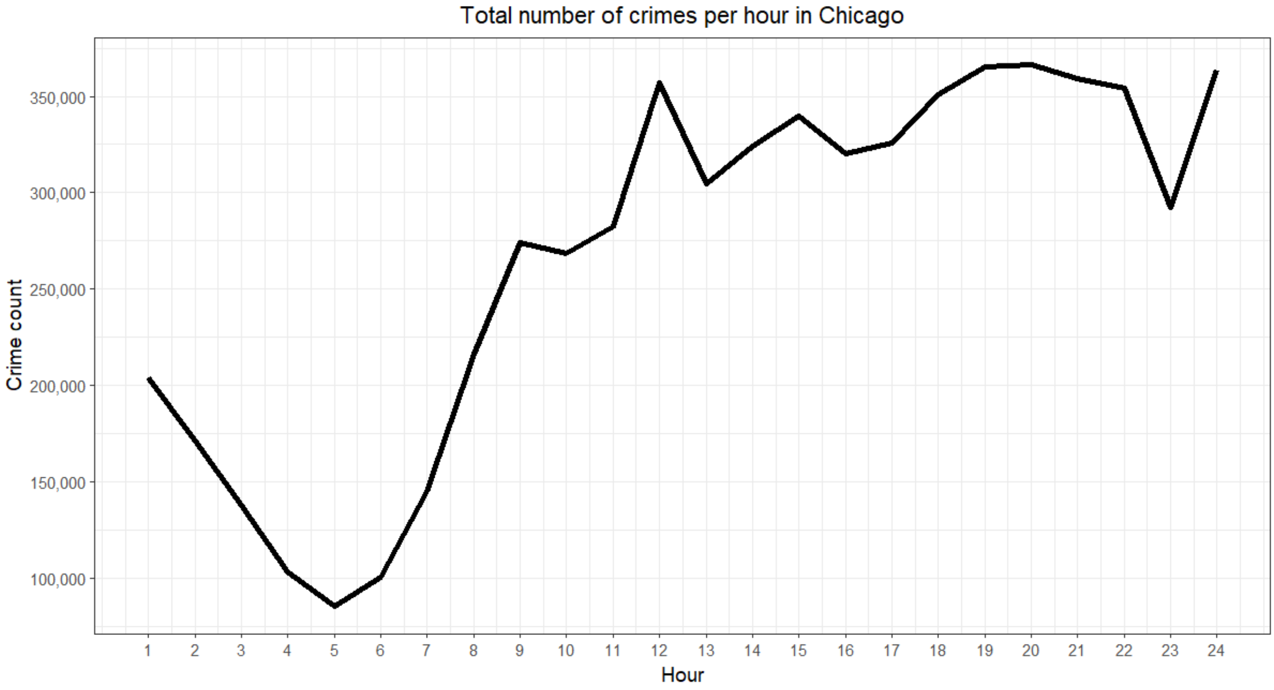

Figure 8: Crime by hour of day in Chicago

often. What we can do, however, is use the district, ward, beat, and community area information to draw approximate results.

In particular, I will limit this search to districts for ease of plotting. Chicago has 22 districts, but 27 show up in the data. Districts 31, 0, 21, 13, and 23 have 123, 49, 4, 1, and 1 crimes reported respectively, which make them obvious typing mistakes. Indeed, the valid district with lowest reported crimes is number 20, with 111,442 crimes, 3 orders of magnitude higher than 123.

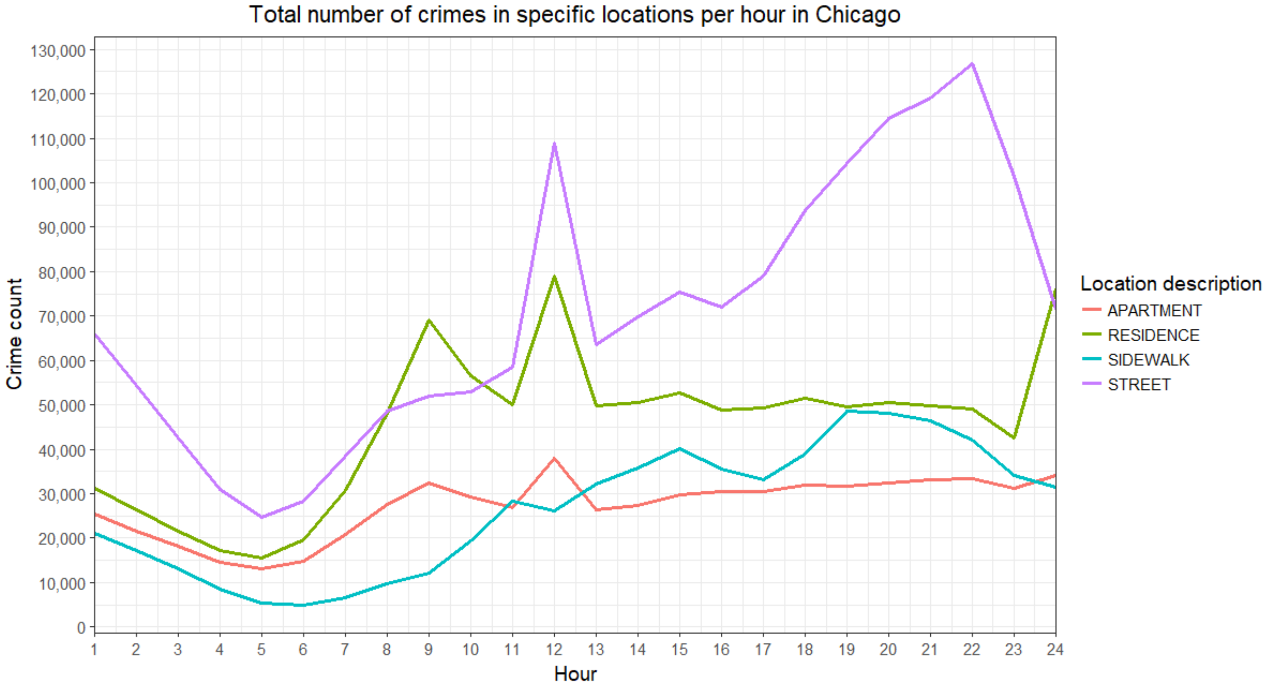

Figure 9: Crime by hour of day at apartment, residence, sidewalk, and street in Chicago

Figure 11 shows the number of crime per district. Districts, 8, 11, 7, 25, 6, and 4 top the list, while district 20 is the safest by a considerable margin. In general, the northern districts are safer than the southern and western areas of the city.

For community areas, numbers 0 and 25 have more than double the amount of crimes than the third highest area (number 23). 0 is a mistake, and most likely means that the information was not input into the table. However, this still leaves us with community area 25 with more than double the crime of number 23. The top three (25, 23, and 29) are all part of Chicago's West Side, an area whose residents are largely low-income families, and immigrants. In counterpart, the north/northwest areas present less crime overall (9, 12, 18).

These results shine a light into a growing problem in Chicago: a division of the city based on wealth. Poorer neighborhood are being segregated against and distanced from the economic center. These neighborhoods often consist of mainly immigrant families that have more difficulty in providing for their well-being. Coincidentally, these are a hotbed for the rise of gangs and criminal activity, which are undoubtedly linked one another.

3.2.6 Arrests and narcotics

While some US cities have already legalized the use of marijuana, Chicago has not. As such, possession of cannabis is the leading cause of arrest in narcotics, with 273,395 arrests since 2001. Following the list, possession of crack and heroin complete the top 3, with 116,777 and 88,571 arrests respectively. Cocaine comes in at number 5 with 23,834 arrests. These numbers are fairly high, and reflect an important drug problem in the city. Chicago are considering legalizing the use of marijuana, so that should highly impact the number of crimes and arrests if it comes to fruition.

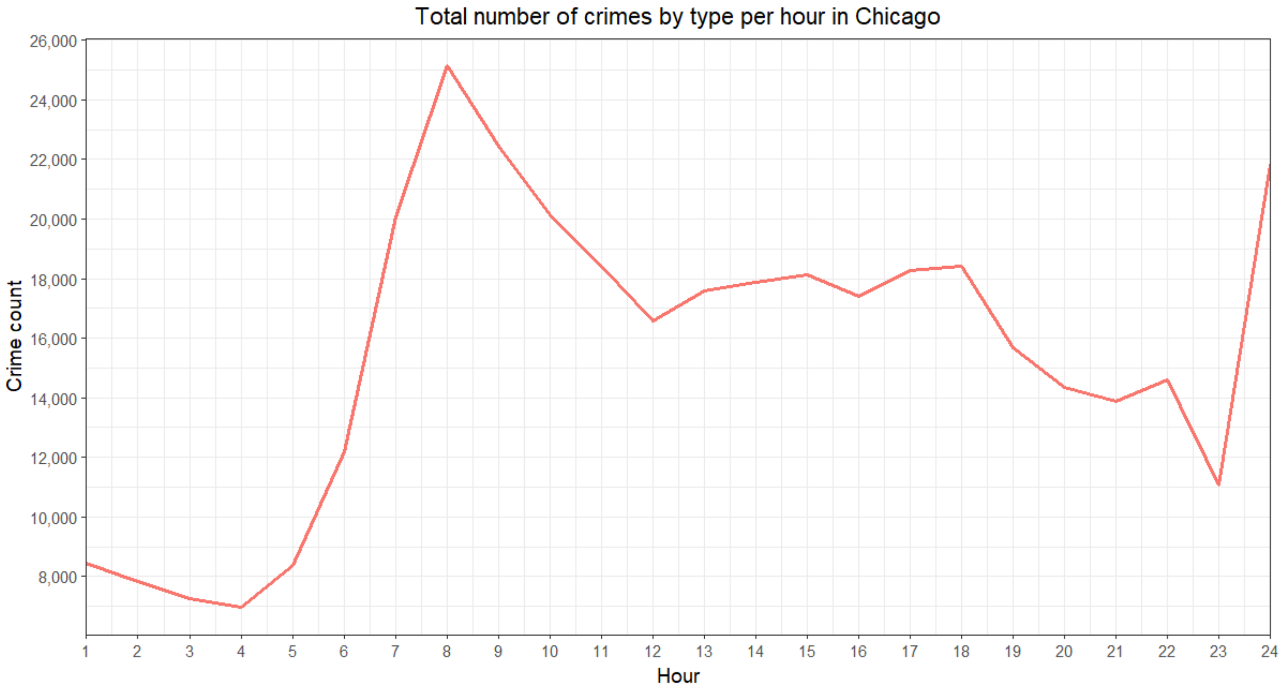

Figure 10: Burglary crimes by hour in Chicago

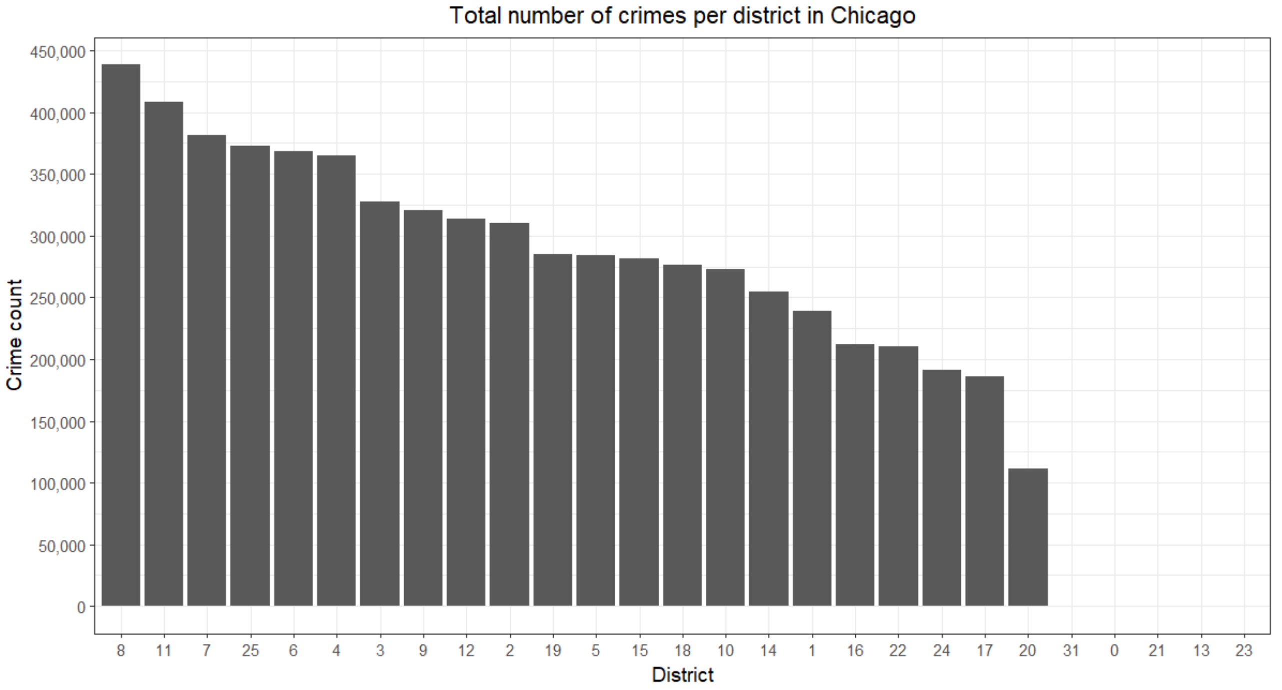

Figure 11: Crime by district in Chicago since 2001

3.2.7 Gun-related crimes and shootings

It has been widely reported that Chicago has a growing problem with shootings. President Trump cited over 4 thousand people were shot in the city in 2016. In this data set, we do not have direct information concerning these types of crimes. Indeed, it is very hard to know how to define whether a person was shot using the fields available to us. The Chicago Tribune keeps an up-to-date counter of these events on their website (link). It shows that 4,355 people were shot by the end of 2016 in Chicago, which corroborates with the president's statement.

For gun-related crimes, I defined a set of IUCR codes that relate to guns in ECL, and filtered the data to account only for these types of crimes (which is what I would have done for the shootings, but I don't believe such IUCR codes exist). In general, gun-related crimes are decreasing, and show the same periodicity as overall crime throughout the seasons (a high peak in the summer, and a low point in the winter). In the summer of 2016, there were 3,231 of these crimes, compared to 4,151 in the same period in 2002 for example.

3.2.8 Masquerading and downgrading of crime

There have been reports that Chicago has been masquerading their numbers downwards and shoving crime into lesser categories. From this analysis and the plots from the above sections, I cannot conclusively discern whether this is the case.

3.3 Final thoughts

In this exercise, I have used the data gathered and provided by the Chicago Police Department in order to build an analysis of crime in Chicago since 2001. We have seen that, overall, crime has been decreasing steadily since 2001, although a plateau seems to have been reached since 2015. Theft, battery, and criminal damage continue to be the major types of crimes committed, and narcotics are still a struggling point for the city. Many types of crimes have seen an increase year over year since 2015 after many years of continuous improvement. More alarmingly, homicides have increased 20% in 2015 and 60% in 2016, which is a major cause for concern. Finally, we also saw that there is a division happening in the city, where more crime tends to occur in poorer regions, while the wealthier areas in general are safer, such as the northwest district. Overall, it does seem that Chicago is experiencing a rough time since 2015 and that the government needs to take action to control the situation. Increased gang activity is surely a big contributor to this increase, and much attention is needed to make sure that this problem does not worsen in the coming months and years. The media may exaggerate certain points and bend the truth to some extent, but they are right that homicides have increased greatly in the past couple of years and it is definitely cause for concern, and even impacts the national average numbers. I do not think that the city is masquerading crime or reducing it to lesser types, but further investigation may provide additional information into this conclusion.

4 Conclusion

In this report, I have completed two exercises that showcase how ECL can be used for data analysis.

First, a reformatting example was implemented to explain how to clean an incoming data set and store it into pre-defined formats or layouts. We have learned that this task is straight-forward when the data is nicely organized, but can be troublesome if data is missing and imputation needs to happen. I have tried to follow the guidelines of ECL and code formatting as well as possible, but I am sure that there is still much that I need to learn to make this process faster and more robust.

Second, I used ECL to create tables and outputs based on several questions that were asked and other that I wanted to investigate. Here, the use of macros greatly simplifies the task, since the same process can be repeated for different fields. As such, I tried to minimize code repetition as best as I could. I then used R to import these tables in order to create plots that help visualize the results obtained.

I look forward to talk about this report some more, and even take a look at what other people did, and how they implemented code differently.