Chicago Crime Data Analysis

Francisca Conrado Silva
Group D
FCUP

Sofia Pereira
Group D
FCUP

Abstract

Criminality is a horizontal problem throughout the world. Studying this topic is crucial for making the world a better place. A data science approach to this matter can reveal a lot of interesting and useful information. In this article, a data set regarding criminality in Chicago was analyzed under this perspective, focusing on answering a set of carefully defined questions in a visual manner.

Keywords—Visualization, Chicago, crimes, R, shiny.

Introduction

The City of Chicago has an open database with information about crime occurrences from 2001 to 2017. Analyzing this data set allows for a better understanding of Chicago's criminal scenario. Using data processing tools and combining additional data sets crucial information can be obtained, such as identifying the most problematic districts and blocks, understanding what the most frequent type of crime is or even if some cultural or atmospheric events interfere with criminality. The aim of this project was to create visualizations to answer carefully thought questions. The answers to these questions might be interesting to several entities (such as Police Departments) to ultimately develop solutions to reduce criminality in the city.

Using R programming language, a detailed analysis of this data was carried out. Exploratory analysis allowed the authors to establish a set of interesting questions that once answered would yield useful information. These questions were thought of in a way that answering them sequentially ends up "telling a story" about the data. Each of them was carefully analyzed and answered graphically. Some of those visualizations are present in this report but some are not (due to the reduced available space).

I. PREPROCESSING

Four input files containing data from the Chicago Police Department were merged into one single file containing a total of 22 variables and 7941285 instances, ranging from 2001 to 2017. The resulting file will be referred to as the main data set. Due to the size of the main data set, some measures had to be taken in order to reduce its size, making it easier to process. Early exploratory analysis revealed the existence of some redundant variables such as two separate coordinate systems and different ways to spatially divide the city of Chicago like Beat, Districts, Ward, Community Areas, and Blocks. Therefore, some of these redundant columns were dropped to downsize the data. Thirteen columns were kept: ID, Date, Arrest, Domestic, District, Community Area, Year, Latitude, Longitude, Block, Primary Type, Description and Location Description. It was decided to carry out the analysis using Community Area (CA) as the selected "area variable". In some cases (namely regarding homicides) two or more data entries can correspond to different crimes in the same occurrence. It was decided to ignore this by dropping the Case Number column and to treat each line of data as a single occurrence.

Some measures were also taken to handle missing values (NAs). Most NAs were present in the columns containing information about the coordinate systems. The first approach was to remove all instances containing NAs and proceed to the next step (sampling). However, removing NAs leads to an imbalance in the input ratio of entries per year, i.e., some years (like 2017) have a lot more entries with NAs and therefore would have far more instances removed and fewer instances left on the data set than other years. In addition to the fact that years with fewer instances would be poorly represented, the distribution of the number of occurrences per year would not reflect the distribution of the original data. Thus, NAs were maintained at this stage of the project. Through the previously described analysis, some uncertainty about the range of time the data from each year was covering emerged. To verify this question, a function was developed to assess the range of time each year present in the data was built The result obtained showed that the year 2017 only has 23 records as there is only data regarding the first 20 days of the year. Thus, it is clear that this year should be excluded to make the year's ranges uniform therefore allowing a fair comparison between years. By this time, an outlier in the year column was found (an entry with value 41 in the Year column) and removed.

The data in the main data set was then sampled. The process of selecting a subset of data is crucial since it is through this subset that statistical inferences and deductions of characteristics of the total data population are made. In this project, sampling is necessary to be able to process and handle such a large data set. Stratified Random sampling was implemented using the function stratified from the splitstackshape R package to sample 70% of the data. Using this methodology, data is sampled without overlapping (there are no repeated instances) and without losing the representativeness of the original population. This sampling technique is the one that best adapts to this project because (once again) it guarantees that the distribution of the occurrences in the sample is as close as possible to the one observed in the data before sampling. If the distribution of the number of occurrences per year would not reflect the distribution of the pre-sampling data, the validity of some data analysis tasks such as studying the evolution of the number of crimes over time would be compromised. The resulting data set contains 5550950 rows and 13 columns. A comparison between the distribution of occurrences per year in the sample and the pre-sampling data is shown in **Figure 1**. Both distributions look quite similar which indicates that the sampling technique and all the other precautions took to ensure uniformity was successful. As expected, the number of occurrences in the sampled data set is 30% smaller than that of the pre-sampling data set.

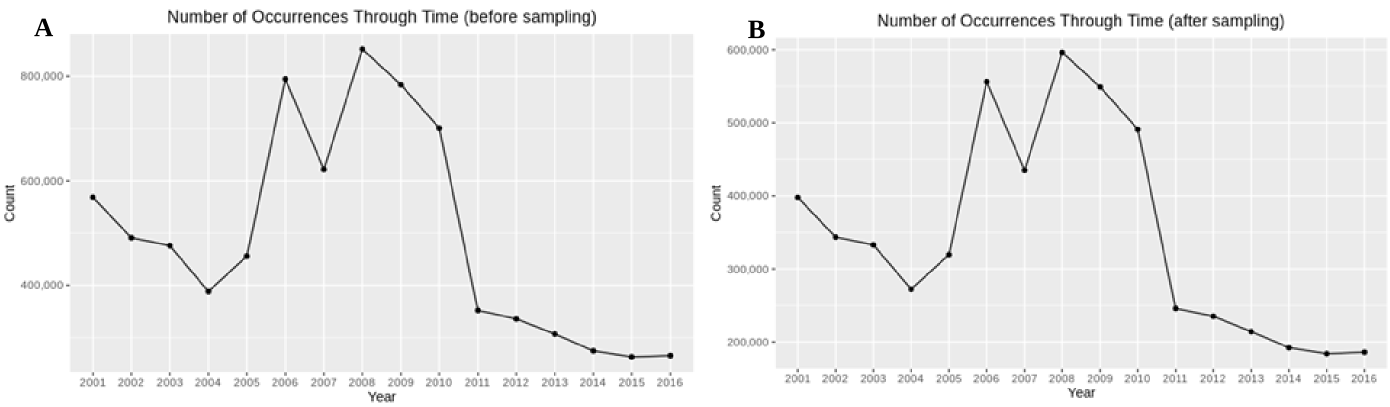

Figure 1: Number of crime occurrences per year before sampling (A) and after sampling (B)

Three external datasets were additionally used. The first contains data regarding socio-economic indicators over Chicago's community area between 2008 and 2012 (available at:https://catalog.data.gov/dataset/census-data-selectedsocioeconomic-indicators-in-chicago-2008-2012-36e55). This dataset will be referred to as the socio-economic dataset. The second external dataset contains information about the sunrise and sunset time per year in Chicago. This data set has been made through web scraping of information available at https://sunrise-and-sunset.com/en/sun/united-states/chicago, using the *octoparse* website's free trial. Using the described tools, data between 2008 and 2012 was scraped. The last external dataset (the NBA data set) contains information regarding the NBA league (https://www.kaggle.com/pablote/nba-enhanced-stats/download/BIs25MT3QSudfYu9HEJC%2Fversions%2F7Qk9SAql8EnCplVErzCK%2Ffiles%2F2012-18_officialBoxScore.csv?datasetVersionNumber=1). Instances with dates ranging from 2012 and 2016 regarding games where the Chicago Bulls played in Chicago were kept. Further information about the purpose of each external dataset and their processing steps will be supplied in the section containing the discussion of the questions they relate to.

Upon trying to figure out how to merge the socioeconomic data with the main dataset, a large number of NAs in the CA column of the main dataset was found. Additionally, some entries had a value of 0 in the same column, although there are only 77 CAs (values range from 1 to 77). As most of the instances in either one of the scenarios had data on the columns regarding latitude and longitude the authors were able to correct the CA column: In conjunction with a shapefile representing Chicago's CAs, the function *point.in.polygon* from the *sf* package was used to check if a given pair of coordinates falls inside any of the 77 polygons representing the CAs and if so, which one. The missing values in the CA column were then imputed using the information above. Some of the records did not fall inside any of the CAs and those were kept as NA. Also, some of the records did not have any data in the columns regarding the coordinates and therefore those were obviously also kept as NA. The number of NAs before and after this correction is shown **Supplementary Table 1**. This corrected version of the main data set will be the one used in all the analysis described below this point.

II. QUESTIONS DEFINED BY THE AUTHORS

1. How has crime evolved in Chicago over time?

2. What are the most frequent categories of documented crime?

3. What is the trend of the most frequent crimes over time?

4. Is there a preferred location and time of day for crimes to occur?

5. What is the pattern of arrests according to crime category, geographical area and over time?

6. What is the pattern of domestic crime according to crime category, geographical area and over time?

7. Is crime associated with areas of the city with a specific socioeconomic level?

8. Does the presence or absence of sunlight affect the number of crime occurrences?

9. Is Chicago's criminal scenario affected when the Chicago Bulls are playing?

III. VISUALIZATIONS AND DISCUSSION

III.I How has crime evolved in Chicago over time?

To answer this question, a contingency table containing the number of crimes per year was created using the main data set. **Figure 2** shows that 2008 was the year with the greatest criminality, followed by 2006 and 2009. One must be careful when trying to justify certain behaviors based on the data, as this behavior might just reflect the way data was entered in the database (i.e., it is not known if there was a uniform methodology for registering crime occurrences). However, if this information were to be extrapolated into the real world, it is worth mentioning that the year with more registered criminality coincides with the year of one of the greatest (and most recent) stock market crashes ever, which might explain the high criminality observed in 2008 and the following years. A trend line is the most simple and effective way to visualize this sort of information. An interactive version of figure 2 allows the user to hover over the dots of each year to know the exact number of occurrences.

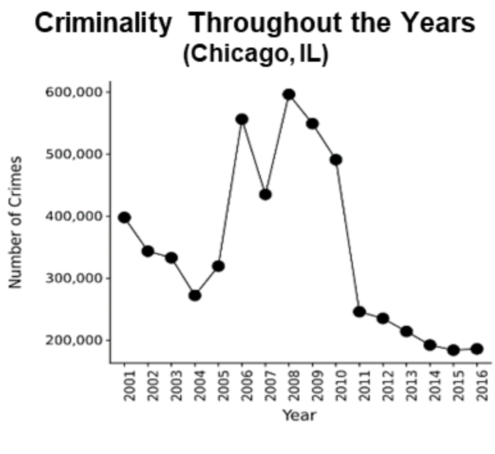

Figure 2: Number of Crimes over the year (from 2001 to 2016)

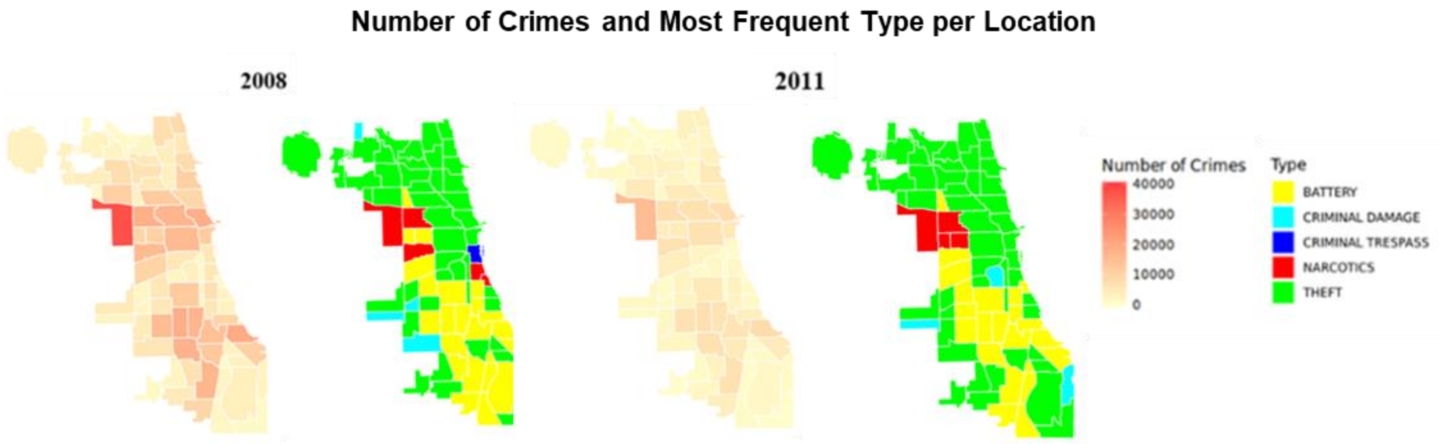

Figure 3- Comparison of the number of crimes and most common crime category per community area in 2008 and 2011

Some additional questions arise with the analysis of Figure2: what happened between 2008 and 2011? Was the observed drop in crime rate throughout Chicago or in a specific CA? Did the most frequent types of crime change? **Figure 3** shows that the drop in criminality between 2008 and 2011 happened thorough the city and was not due to a change in a specific CA. One can also conclude there are not a lot of changes between the two years when it comes to the most common crime category per CA.

III.II What are the most frequent categories of documented crime?

For this type of problem, a word cloud can be a very intuitive approach. To achieve such a visualization, a contingency table containing the number of occurrences of each crime was built. The word cloud represented on **Figure 4** was then created based on that information. In total, there are 35 types of crime. Due to the enormous difference between frequencies of each type of crime, it is not possible to represent all types of crimes in the same visualization maintaining the right proportionality between the words. Therefore, an interactive version of this figure was created. In this version, is possible to choose the maximum number of words and select a range of specific frequencies. This allows the user to zoom in, for instance, the crime types that are not so common. Results show that are the most frequent types of crime in the city are battery, theft, criminal damage and narcotics and "other offense".

Figure 4-Word cloud representing the top 10 most common crime categories. The size of each word is proportional to its frequency.

III.III What is the trend of the most frequent crimes over time?

Besides knowing what the most common types of crime are, it is interesting to see how these behave through time. **Figure 5** shows how the five most frequent types of crime behave throughout time. A contingency table with the number of crimes per each for each of the five categories was built in order to create Figure 5. Once again, and interactive version is available where the user can select which lines to visualize and can also hover over the points and to know the exact number of crime occurrences.

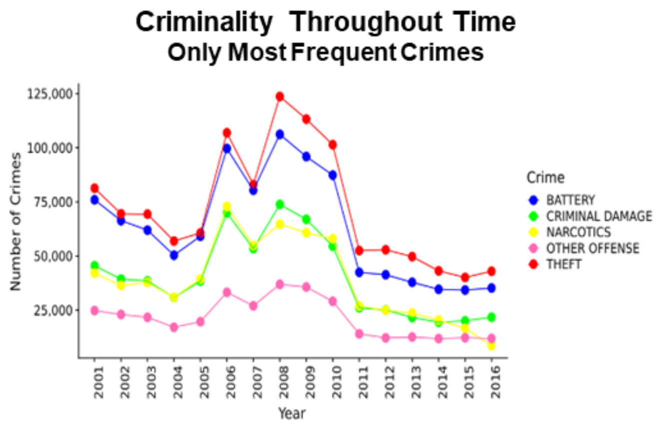

Figure 5: Number of crimes over time per most common crime types

Instead of looking how the top-five crime types evolve over time, another way to look at this question is checking if the most common types of crime change between years. To explore this, the slider input available in the interactive version of Figure 4 can be used to navigate through the years and see how the word cloud adapts.

II.IV Is there a preferred location and time of day for crimes to occur?

A map of Chicago such as the ones depicted on Figure 3 is a very useful way of visualizing an answer to this question. To obtain such a visualization, several steps were necessary. First, a new column containing the hour of each occurrence was derived from the date column of the main dataset. Then a contingency table containing the number of crimes for each hour in the selected range was built. This contingency table was merged (by CA) with the "shapefile data frame" (i.e., the data frame containing the data necessary to build the polygons representing the CAs.) In a way, it is like adding a new feature

Criminality per location and time of the day

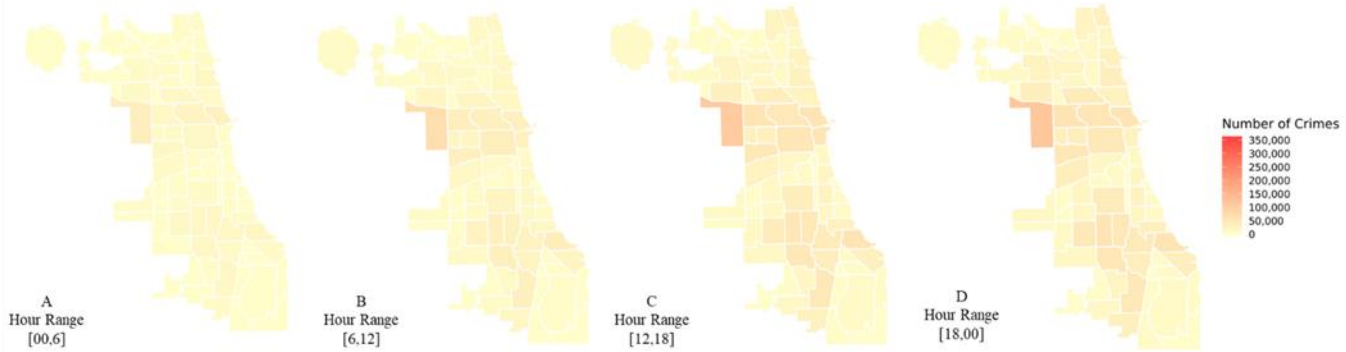

Figure 4: Number of Crimes per community area time of the day.

(besides the ones already there) to each polygon in the shapefile and use that same feature as a scale to fill in the polygons.

This was the way all questions handling map visualizations were handled. Depending on the desired time interval, the corresponding contingency table is built, added to the "shapefile data frame" and a map depicting the criminality per CA is generated. **Figure 6** contains four facets of the interactive version of this graph, each corresponding to a different time interval. Unfortunately, unlike in the interactive version of this graph, in the static version presented here it is not possible to hover over the CAs to check their name or number of crimes. Among other things, by looking at the facets one can clearly see that the CA with higher criminality (Austin) stands out, and also that the CA with higher criminality remains constant throughout the hours. Regarding a preferred time of the day, there are more crimes during late afternoon/early nighttime (between six pm and midnight) than during the night or early in the morning. In the interactive view, besides the hovering functionality it is also possible to select and visualize any desired time intervals as the data used to fill in the map will be adjusted according to the user's input.

II.V What is the pattern of arrests according to crime category, geographical area and over time?

The steps required to get an answer to this question are quite similar to the ones in question four. First, the desired set of crime types is selected. Then, based on that subset, a contingency table with the number of times an arrest took place (TRUE) and the number of times an arrest did not take place (FALSE) per CA is built. Based on this information, a new data frame containing the percentage of arrest per CA (=TRUE/(TRUE+FALSE) *100) is created. This data frame is then merged with the "shapefile data frame" and a map is generated using the information of the added feature (% arrest) to fill in the CAs (in the same manner as described for figure 3). Three facets of the interactive version of this plot are depicted in **Figure 7**. In the interactive version of this figure, the user can select the set of crimes take into account. Additionally, it can hover over each CA and see its name, percentage of arrest and number of crimes for the selected set of crime types.

The comparison between Figure 7B and 7C clearly shows that different types of crimes result in different patterns of arrest. Thus, the percentage of arrest of a given CA is most likely linked to the type of crime that go on in a CA. One can see that the CA with the highest crime rate (as seen in fig. 5) is also the CA with the highest percentage of arrest. This might be due to one of two things: either there is a bias for police to arrest more in that CA due to the significantly higher criminality rate or (as mentioned above) the types of crimes occurring in that CA are more prone for an arrest to take place. To get an answer for this, the overall percentage of arrest for each crime type was calculated (with a methodology analogous to the one used in the previous graph). **Figure 8** displays those percentages for the top-5 most frequent crimes in Austin. As expected, the most frequent type of crime in Austin has an extremely high arrest rate for in Chicago. Thus, the high percentage of arrest in this area is in fact due to the types of crimes that happen there.

III.VI What is the pattern of domestic crime according to crime category, geographical area and over time?

Domestic crimes represent all types of crimes that occur inside a house. In a way analogous to what was done for Figure 5, to assess whether the percentage of arrest in domestic and non domestic crimes is different, a contingency table with the

Percentage of arrest per location and crime type

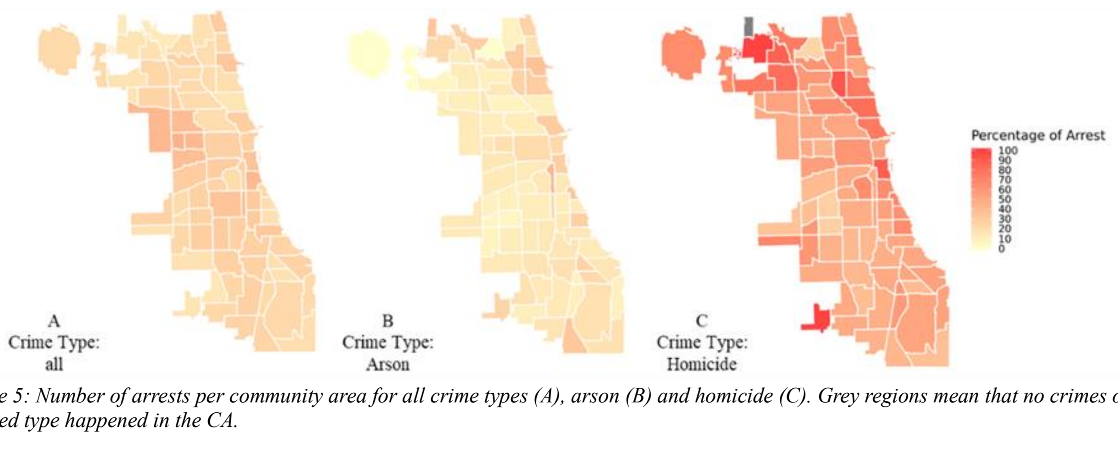

Figure 5: Number of arrests per community area for all crime types (A), arson (B) and homicide (C). Grey regions mean that no crimes of the selected type happened in the CA.

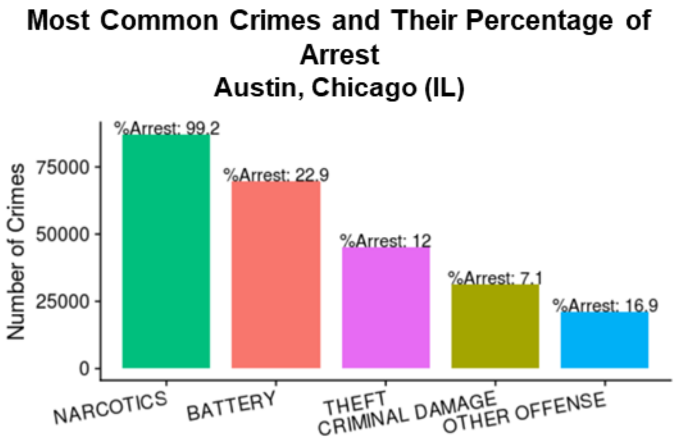

Figure 8- Percentage of Arrest of the five most frequent crimes in Austin

number of with the number of times an arrest took place (TRUE) and the number of times an arrest did not take place (FALSE) when the crime was domestic and when the crime was non domestic was used to build a data frame with the percentage of arrest in general, the percentage of arrest when a crime was domestic and when it was not. The results are shown in **Figure 9** in the form of a bar plot. The percentage of arrest in domestic crimes is smaller when compared to percentage of arrest in non-domestic crimes. However, even though the results are in percentual form, it is important to mention there are fewer domestic crimes than non-domestic crimes.

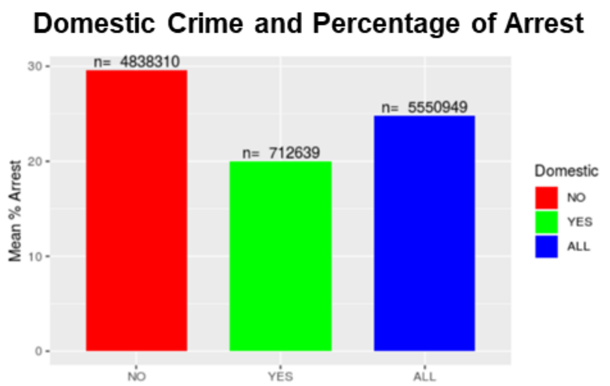

Figure 9-Mean Percentage of Arrest in Domestic crimes

III.VII Is crime associated with areas of the city with a specific socioeconomic level?

For this question, the socio-economic external dataset was used (Supplementary Material Table2), namely the column regarding the Per Capita Income (PCI) per CA. As the external dataset only comprises data from 2008 to 2012, the main data set was subset to the same period of time in order to fit this constraint. Thus, the answer to this question only considers data from 2008 to 2012. A contingency table with the number of crimes per CA was generated and merged (by CA) with the "shapefile data frame". Then, the information regarding the PCI was added to the same data frame (also merging by CA). A bivariate scale and map were created with guidance of https://timogrossenbacher.ch/2019/04/bivariate-maps-with-ggplot2-and-sf/), using the package. The resulting map is represented on **Figure 10**. It shows that there is a large number of CAs where there are both low PCI and high criminality, which is what the authors suspected. The link between low criminality and high PCI is not as evident but it's still visible in a few CAs. Comparing Figure 6 with Figure10 leads to very interesting observations: the CAs with higher criminality (shown in a darker tone in Figure 6) roughly coincide with the CAs in the high criminality/low PCI

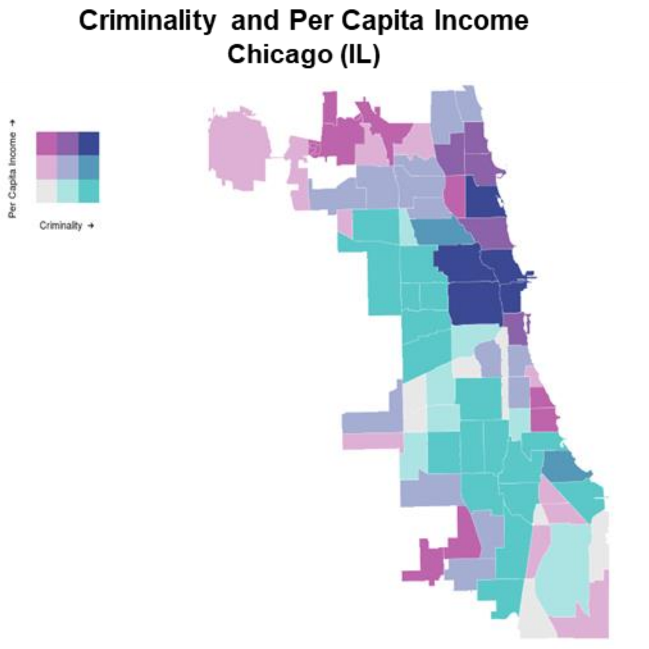

Figure 10: Relationship between criminality and PCI per community area. A bivariate scale was implemented in order to facilitate the visualization

situation in Figure10. However, a few CAs with high criminality rates do not have such a low PCI.

III.VIII Does the presence or absence of sunlight affect the number of crime occurrences?

For this question, the external sunlight (Supplementary Material Table3) dataset was used. As the external dataset only comprises data from 2008 to 2012, the main data set was subset to the same period of time in order to fit this constraint. Thus, the answer to this question only considers data from 2008 to 2012. It was merged (by column "date") with the main data set. Then, a binary column where 1 encodes for sunlight and 0 encodes for absence of sunlight was generated. To make it more realistic, a buffer was considered: to sunrise and sunset times half an hour was subtracted and added (respectively), since there is already a bit of light before sunrise and still a bit of light after sunset. Also, sunrise and sunset times are not exactly the same throughout the entire city, so this data is an approximation. A contingency table regarding the type of crime, the presence/absence of sunlight and number of crimes occurring in

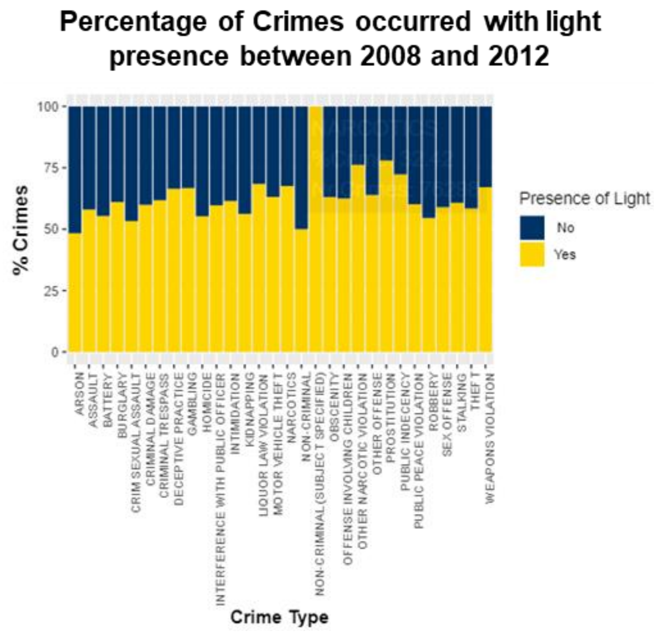

Figure 11: Percent of Crimes occurred in the presence or absence of light

each of those scenarios was built. This information was then represented in the stacked bar plot in Figure 9. This type of visualizations allows for the comparison between the realities of different crime types. In general, that are more crimes when there is daylight than when there is not. A possible explanation for this will be approached in the following section. There is not a significant difference between the patterns observed for different types of crime when it comes to this question.

II.IX Is Chicago's criminal scenario affected when the Chicago Bulls are playing?

For this question, the external NBA (Supplementary Material, NBA dataset information) dataset was used. This data set required additional preprocessing steps: first, only the games were the Chicago Bulls played at home were considered. The main reason for this is that the authors considered that when the Chicago Bulls were playing away in another city the level of disturbance in the criminal scenario would not be as significant as when the games occur in Chicago (were people from the city can go and watch). Also, the United States have several different timezones. Therefore, games in different cities would be happening at different timezones. Then, a correction due to time zone discrepancies was needed: the NBA dataset is in Eastern Time (ET), which Chicago's timezone only between March and November. During the other months Chicago's timezone is one hour early than ET. Starting from October 2016, the NBA external data set has a data input error: the column referring to the time the game has started start being written in the 12-hour format, without considering that the previous records were using the 24-hour format. This leads to errors as R assumes the time format is still 24h and thus that games occur in the morning, which is not the case (checked and corrected based on: https://www.basketball-reference.com/leagues/NBA_2017_games.html). As the external dataset only comprises data from 2012 to 2018, both the external and the main data set were subset to the common period of time between them (2012 to 2016).

The first approach to this question was in an exploratory perspective: using a graph like the one in Figure2 as a "base", an interactive figure was built, where the user could define the time range to look at. On top of that "base graph", a layer highlighting the time ranges where games were occurring was added. By looking at the color of the rectangles in that layer the information on the outcome of the game is also obtained. This interactive approach serves exploratory purposes. However, it does not allow the user to come to a general conclusion when it comes to finding out if Chicago's criminal scenario is affected by the Chicago Bulls' home games.

In order to make users able to take proper conclusions, a different (and static) of the previous representation was created. It works like a "summary" of the exploratory analysis one can perform with the interactive approach. The graph in figure 10 compares the mean number of crimes when NBA games are happening and when they are not. To allow for a fair comparison, only hours where games actually occur were considered. The average duration for an NBA game is considered to be 138 minutes (https://www.reference.com/world-view/average-length-nba-game-7df1bcf6ed0aa740). Since 94% of games occurred approximately between 19 and 22:18 pm that was the time interval to consider. After sub setting the main data set as explained above, a new column containing the hour of each occurrence was derived from the date column of the main dataset. Then, a binary column was added (also to the main dataset) through the crossing of information of the two datasets (main and NBA): for a given crime occurrence, 1 means that an NBA game was happening at the moment of the crime, while 0 means that no game was happening. Figure 10 makes it pretty clear that the fact that Chicago Bulls are playing does not make much of a difference when it comes to criminality in Chicago.

This might seem inconsistent with what one can conclude by exploring the interactive version of this topic. However, if the user selects a wide enough time range it can verify that criminality follows a very well-defined time series. This point relates to the previous question: the fact that there are more crimes when there is sunlight is most likely a realization of this cyclic behavior observed in the data, and it is not actually related to the presence or absence of light (just like criminality is most likely not related to the happening of an NBA game). Due to its seasonal behavior, a SARIMA model would probably be good fit to model its behavior.

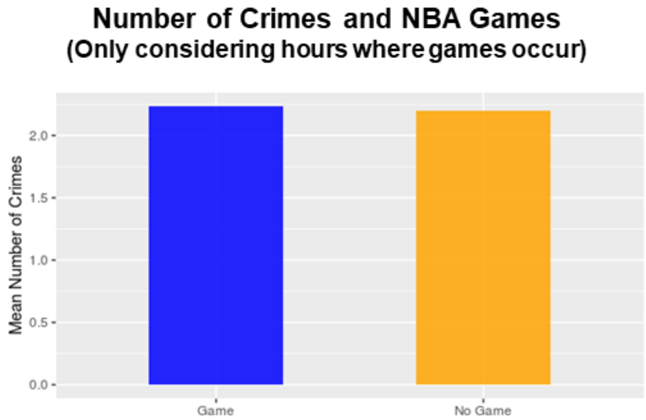

Figure 12: Comparison of the number of crimes occurred (during the time period where 94% of Chicago Bulls' home game occur) when there the Bulls are playing and when they are not.

IV. CONCLUSIONS

With this study it was possible to identify some parameters that may contribute to crime occurrences and can be used by Local Parties and Police Departments to develop solutions and tools to downsize the criminality levels in the Chicago City. The aim of the project was achieved by making curious and relevant visualizations about Chicago Crimes data set.

V. LIBRARIES

To preprocess all the data, were used the following R libraries: dplyr, splitstackshape, ggplot2, scales, lubridate, cowplot, maptools, rgdal, ggmap, biscale and sp.

To answer all visualizations the following R libraries were used: ggplot2, scales, lubridate, cowplot, ggiraph, shiny, rsconnect, rgdal, dplyr, plotly, wordcloud, wordcloud2, maptools, sp, reshape2, plyr, biscale and ggmap.

VI. REFERENCES

All external data sets are referenced upon the first time they are mentioned in the report.

Supplementary Materials

Table1: Chicago Crime dataset description

Column Name	Description
ID	Unique identifier for the record
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident
Date	Date when the incident occurred. this is sometimes a best estimate
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address
IUCR	The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e
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 as defined by the Illinois Domestic Violence Act
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74
District	Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r
Ward	The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block
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. This location is shifted from the actual location for partial redaction but falls on the same block
Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

Table2: Socioeconomic dataset description

Column Name	Description
Community Area Number	Number of Community Area
Community Area Name	Name of Community Area

Percent of Housing Crowed	Percent of occupied housing units with more than one person per room
Percent Households below poverty	the percent of households living below the federal poverty level
Percentage Aged 16+ unemployed	the percent of persons in the labor force over the age of 16 years that are unemployed
Percent Aged 25+ without high school diploma	the percent of persons over the age of 25 years without a high school diploma
Percent aged under 18 or over 64	the percent of the population under 18 or over 64 years of age (i.e., dependency)
Per Capita Income	Income per capita
Hardship index	Index

Table3: Sunlight dataset Description

Column Name	Description
Date	Record of the date of the year
Sunrise	Sunrise time
Sunset	Sunset time
Solar Noon	Time when the Sun passes a location's meridian and reaches its highest position in the sky
Daylength	Number of hours with light

Table4: Correction of missing values in the Community Area variable

Year	count	NA_count (Before Correction)	NA_count (After Correction)
2001	397830	394305	2137
2002	343544	95732	7369
2003	332897	34	8
2004	272193	36	6
2005	319435	36	6
2006	556265	63	10
2007	434944	155	12
2008	596211	326	62
2009	549050	310	34
2010	490923	250	10
2011	245918	123	12
2012	235154	18	6
2013	214211	7	8
2014	192283	2	2
2015	184012	1	2
2016	186095	1	1

NBA dataset Information

This dataset presents 100 variables related with Game Percentages. Among all variables only game date and hour, home team and opponent team, also the result of the game was kept (i.e., if the home team win or lose the game).