**Chicago Violent and Non-Violent Crime Rates Per 1000 People Per Year: Northern Neighborhoods May Not Be as Safe as People Think**

ISyE 602 Final Report

Chicago Crime PreDictor (CCPD): Thiyagarajan Annamalai Ramalingam, Sanyam Kakkar, Yaoshihui Wang, Angad Vir Singh, Kathleen Buckingham (100/5)

**Abstract**

The number of violent and non-violent crimes per 1000 people per year in Chicago shows that northern neighborhoods may not be as safe as people think. In particular, this report uses two choropleth maps to isolate the number of violent and non-violent crimes per 1000 people per year for each neighborhood. In addition, it uses a 24-hour time series graph to show the number of violent and non-violent crimes per 1000 people per year for each neighborhood. We generated choropleths as it’s the best way to show the spatial pattern of these crime rates, not for supporting judgment. In addition, we added the time-series graph to make up for the choropleths’ inability to convey temporal patterns. Our target audience is Chicago tourists and/or new Chicago residents who may not be as familiar with Chicago neighborhoods and their crime levels. Our app may be used as an explanatory graph for our audience to understand historical crime rates, and an exploratory graph to target areas they may be traveling through to be able to predict their risk for crime in these neighborhoods overall as well as a particular hour of the day. We used ggplot2, dplyr, shiny, and other related packages to construct the plots in R and publish them online in a Shiny application. In conclusion, we warn the audience that their risk for both violent and non-violent crime is highest in Fuller Park. Risk for violent crime is lowest in Forest Glen and risk for a non-violent crime is lowest in Edison Park. We highlight that there is always less crime of both types from midnight to noon and more crimes from noon to midnight. Finally, we note to our audience that areas such as the loop are higher in non-violent crime, which tends to be overlooked due to the attention that south side violent crimes receive.

**Background and Purpose**

We were motivated to create a visualization of Chicago because we all have some ties to the city. Some of us currently live in the city, others will soon be moving to the city, and most of us fly to the city when coming and going from Madison. In addition, Chicago has always been at the center of attention when it comes to crime rates in certain areas and is notoriously an even more dangerous city than New York. For this reason, we thought it would be important to address the issue of crime for people who are not as familiar with the Chicago area, such as those traveling through Chicago or those looking to move into the city.

It is generally common knowledge that violent crime rates are higher in the south side, but there is also both violent and non-violent crime all over the city where tourists, visitors, and residents may not expect it. Oftentimes, violent crimes such as homicide, sexual assault, and battery get more attention from the media. Non-violent crimes such as theft, criminal narcotics, and burglary are more common than violent crimes, but most often do not get as much attention as violent crimes. We believe it’s important for people to be informed of both types of crime, especially for neighborhoods that they may not expect. In addition, we further inform the audience by showing temporal data of violent and non-violent crimes by the hour for each neighborhood. We have been inspired by many maps online, such as a dashboard posted by the City of Chicago and an interactive graph made with excel posted on Open Data Bits [1, 2]. However, we have not seen a visualization that isolates violent and non-violent crime per 1000 people per year over the 20 years. In addition, we have also not seen a visualization detailing the temporal crime data by the hour for each neighborhood.

**Basis and Approach**

In our shiny application, we decided to present three graphs. The first two are choropleth maps showing the risk of violent and non-violent crimes, and the third is a time-series graph showing violent and non-violent crime counts per hour for the neighborhood in question. We believe it would be a compelling story to show. First, the audience can look between the choropleth maps and compare their risk of encountering a violent or non-violent crime in the same neighborhood based on historical data. The audience may be interested in a particular neighborhood or a group of neighborhoods depending on their reasons for being in Chicago. Consequently, the audience can look at the time series graph. Not only will this combination of graphs give them a better understanding of their risk for both violent and non-violent crimes in their neighborhoods of interest, but they will also be able to see their risk of encountering both types of crimes down to the specific hour. For example, say a person knows they will be traveling through Lincoln Park at 5am. By looking at the choropleths, they could see that Lincoln Park has a lower risk for violent crimes and a higher risk for non-violent crimes from the choropleths. Next, they can see that at 5am, the risk for encountering both types of crime is much lower than at other times of the day.

We decided to use a choropleth because people who are not familiar with Chicago may not recognize neighborhood names. If we were to use a bar or a line graph that only referenced neighborhood names, it may be meaningless to someone who doesn’t know where they are. We also believe it is very helpful to visualize the location of neighborhoods because someone may be familiar with one neighborhood, but not be aware of what neighborhoods surround that area. It’s possible that a safer neighborhood could be neighboring a dangerous neighborhood. If someone not familiar with Chicago is planning on visiting a friend in a safe neighborhood, they may be able to see that a neighborhood with a higher risk for crime is right next to the neighborhood in question. This could be useful as it helps them to be cautious of accidentally traveling into a more dangerous area. Another reason why we use a choropleth is that we wanted to provide information on every neighborhood. Including 77 bars in a bar chart or 77 data points in another form of a graph would make the graph too cluttered. Thus, it would be a great and engaging way to represent many points of information without feeling like it is too cluttered. Finally, we included a time series graph to make up for the lack of temporal analysis included in a spatial choropleth.

**Bringing Meaning to Data through Abstraction and Aggregation**

The dataset that we used is from Kaggle and contains reported incidents of crime that have occurred in Chicago from 2001 to the current year [3]. The dataset is pulled directly from the Chicago Police Department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system. The dataset includes 22 columns and 7 million rows containing both geographic location data of where the crimes were reported to occur, the type and descriptions of the crimes.

A particular challenge we faced was the large size of our dataset, which caused the code to run very slow. To mitigate this and transform our raw data to support our intended visualization, we filtered for the columns that we intended to work with. Notable columns that we selected include date (date when the incident occurred), primary type (crimes were classified into 37 primary types including homicide, assault, battery, etc), description (crimes were further described with over 500 descriptions), and community area (defined areas in Chicago numbered 1-77).

We abstracted the data in two ways. The first was by classifying the data as violent or non-violent. To do so, we manually went through each of the 37 primary types of crime and leveraged the National Justice Institute guidance on violent crime classification [4]. Next, we used our best judgment to classify each of the primary types to be violent or non-violent. When we were unsure, we looked further into the descriptions of the type of crime to get more information on whether violence may have been used or not. Once we classified violent and non-violent crimes, we then created a new binary column called “violent” containing a “1” if the crime was violent, and “0” if it was non-violent. We did the same with a “non-violent” column. This abstraction is then used to total the number of violent and non-violent crimes per neighborhood to be used in the per-capita calculation. Another way we abstracted the data is by separating date and time, and then pulling out the hour from the time column to be used in our hourly graph and mutating it to 24-hour format.

Next, we combined a few datasets using left join to make our data more meaningful. We discussed that using community area numbers in our map would not be as effective to the audience as using neighborhood names. So, we joined a dataset containing neighborhood names such as “Loop” or “Lincoln Park” to be the most recognizable by the audience. In addition, we also combined the dataset with a population per neighborhood to use in our per capita calculation [5]. Finally, we joined our data with spatial day to create the map of Chicago in a special features map [6].

With all the information in a single data frame, we used the summarize function to aggregate violent and non-violent crimes per 1000 people per year for each neighborhood. This was done by taking the sum of violent or non-violent crimes calculated from our previous abstraction. This number was divided by the population of the neighborhood, which was then divided by 20 years. Finally, rather than having a very small decimal, we decided it would be more useful to the audience to multiply the crimes per capita per year by 1000 to get the final crimes per 1000 people per year data.

Finally, we also wanted to briefly discuss the concept of construct validity. The number of people getting arrested, and the number of crimes reported does not necessarily reflect the rate of crimes. In more predominately white areas, people might be less likely to get arrested for crimes due to racial bias, so they aren’t accounted for. It is important to think about how the dataset provided may not necessarily be an accurate reflection of actual crime rates.

**Visualization Techniques**

To make our choropleth maps visually attractive and capture the attention of the audience, we used the proximity compatibility principle and interaction. The proximity compatibility principle is used by putting the choropleth maps side by side, rather than only having one map with an input variable that toggles between violent and non-violent crime. Two maps side by side make it easier for the audience to compare the two crime types by looking back and forth between the two, instead of only viewing one map at a time and having to remember what one map looks like before toggling to the next. In addition, we were conservative with our interaction and were careful not to include too much so we could better control what information the audience takes in. Users can hover over neighborhoods and only see the neighborhood name and the number of crimes per 1000 people per year. Another interesting element of our interaction is if a neighborhood on one map is selected, the corresponding neighborhood is selected on the other map to allow for easier comparison. We also made sure that rather than the entire fill of the neighborhood being highlighted, we only highlighted the border so the audience can still see the shading corresponding to crime rates. Finally, we changed the scale of the map so that neighborhoods with more crimes are darker to align with the audience’s perception of darker equating to more of something.

In our time-series map, we used figure-ground separation, reference lines, and interaction. We added a reference line of the average number of crimes per 1000 people for the entire city of Chicago as a reference to be able to understand how a certain neighborhood compares to the average. In addition, we made this line gray and transparent, so it appears to fade into the background as we don’t want this piece of information to distract from our main points. In the shiny application, we allow users to select from a drop-down list to view the graph of their neighborhood of choice. Because there are 77 neighborhoods, the user can use a search bar to speed up their search if they proactively know what neighborhood they are interested in.

Even though we had ample data, we still lacked some key information. Initially, we didn’t have the name of the neighborhoods - they were just numbered from 1 through 77. In addition, our initial discussions considered the integration of housing prices of the neighborhoods to see if there was a relationship between that and crime. However, we did not pursue this idea because we weren’t able to find a credible data source. To overcome these challenges, we manually googled the neighborhood area codes to create a data frame with all the names of the neighborhoods.

**Overview of the Visualization**

We want the audience to observe the historical trends of crime data to understand their risk for both violent and non-violent crime in certain neighborhoods by the hour. See a screenshot of the choropleth maps in Figure 1, and the time-series graph in Figure 2 below.

Diagram

Description automatically generated

***Figure 1:*** *Interactive Choropleth Maps: Violent and Nonviolent Crimes per 1000 People per Year*

Chart, line chart, histogram

Description automatically generated

***Figure 2:*** *Interactive Time Series Plot: Time of the day Vs. Crime per 1000 People*

To immediately capture attention, we used a Choropleth plot to set a stage for the audience and allow them to not only see violent and non-violent crime data for any neighborhood, but also allow them to make comparisons between the two. We think an innovative aspect of our map is the ability to select a neighborhood on the map which remains highlighted until it is unselected. This may be particularly useful to the audience if they know they will be traveling through, say, 4 neighborhoods, and want to select the path they are traveling through (see Figure 3). Next, we intend of the audience to move from interacting with the choropleth maps, to interacting with the time-series map to understand crimes by hour.

Chart

Description automatically generated

The Loop

South Lawndale

***Figure 3:*** *Example of how a user can interact with the graph and identify the neighborhoods that are more safe to travel through while entering in South Lawndale Chicago and ending in the Loop.*

We hope that the primary elements of our story include which neighborhoods have the highest and lowest risks of encountering both types of crimes and comparisons between the two types of crime. Fuller Park leads both types of crimes with 10.42 violent crimes and 29.13 non-violent crimes per 1000 people. It’s interesting to note that even in a neighborhood with the highest risk of violent crime, you’re around 3 times more likely to encounter non-violent crime. Compare this to Forest Glen with 0.53 violent crimes and 2.83 non-violent crimes per 1000 people per year. In this neighborhood with one of the lowest rates of both types of crime, you are around 5 times more likely to encounter non-violent crime than violent crime. Additional elements to note are that the Loop has the third highest rate of non-violent crimes with 24.39 non-violent crimes per 1000 people per year. Compare this to 3.83 violent crimes per 1000 people - in the Loop you are 6 times more likely to encounter non-violent crime than violent crime. We believe this difference is due to the fact that the Loop is one of the most vibrant neighborhoods including many shops, businesses, parks, museums, high rises, and attractions. Another element we’d like to draw attention to is that Hyde Park has a lower number of both types of crimes, but is surrounded by neighborhoods with high numbers of crime, especially violent crime. This is likely due to the presence of campus security at the University of Chicago (located in Hyde Park). We think it would be very useful for an incoming student at the University to be aware of the violent and non-violent crime data in surrounding neighborhoods, especially at what hour of the day. Finally, the main takeaway from our time series is that for all neighborhoods, there is less crime from midnight to noon, and more crime from noon to midnight. We have noticed common spikes of crime at the hours of noon and 3pm, which is interesting.

One secondary message we thought would be interesting to note is that there isn’t necessarily a direct relationship between the number of violent crimes per 1000 people per year, and the number of non-violent crimes per 1000 people per year. Through our exploration of the data, we discovered that by just looking at crime count per neighborhood, multiple neighborhoods have high per capita non-violent crime but low per capita violent crime, are with the highest crime count. Without investigating more and representing just the total crime count would have been very misleading in informing our audience. If this was our primary focus, then a scatterplot of violent vs non-violent crime would have been more appropriate than a timeline.

There are benefits to using R for our visualization than Excel.  If we were to use Excel sheet for the same purpose the complexity of manipulating data would have been comparatively higher, also excel doesn’t have interactive features which make the plot visually appealing to the viewers. The major limitation in Excel is that if we were to use a color gradient to differentiate the neighborhoods, we have to manually use the color pallets and adjust the gradients, whereas R has predefined pallets that can be used off the shelf. Overall, R creates a better audience experience and a more straightforward interactive feature.

The time, data set size and the computing power we had were the major constraints, had we more time we could have done rather a detailed study involving the whole data, integrating with the cost of living in a neighborhood to create detailed visuals, moreover initially we thought of doing a time-series wherein audience can zoom in, zoom out and various other criteria they can compare, this has to dropped to achieve the deadline. In addition, with more time we would have liked to label notable landmarks on our choropleth maps (Such as University of Chicago or Navy Pier) to orient the audience who may be unfamiliar with Chicago, as well as provide reasons that people may be visiting certain neighborhoods. Finally, we also thought it would be very interesting to drill down into primary types of crime such as homicide, theft, burglary, to understand which specific primary crime types occur in each neighborhood.

**Conclusion**

Our team used an iterative approach when doing the project as we had to start over a few times. The critical response process in the class provided us with great insight as we realized where we could improve upon and make the final changes. We had to reflect on which graphs would be optimal to show the information to our intended audience. We kept the graph simple and only used interactive data visualization when it provided value for our audience for enhanced user experience in interpreting the graphs. The result was an amalgamation of all such deliberations.

The final shiny app was able to give us many important insights. By looking at the time series plots, we concluded that most of the neighborhoods are much safer in the mornings. Moreover, the south has more violent and non-violent crimes than the north. However, in the case of loop neighborhood, we found that it had more non-violent crimes per capita than most neighborhoods. In conclusion, this is a shiny app which would be useful to us, and would be a very insightful guide to go around Chicago.

**Link To Shiny Web Application:** <https://shihuiwang.shinyapps.io/Projectapp/>

**References**

[1] Crimes - 2001 to present - Dashboard | City of Chicago | Data Portal [WWW Document], n.d. URL https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g (accessed 12.12.21).

[2] Bits, O.D., n.d. City of Chicago Crime Data Now Featured at Open Data Bits – Open Data   
 Bits. URL http://opendatabits.com/city-of-chicago-crime-data-now-featured-at-open- data-bits/ (accessed 12.12.21).

[3] Crimes - 2001 to present - Dashboard | City of Chicago | Data Portal [WWW Document], n.d. URL https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g (accessed 12.12.21).

**Appendix**

Comment 1

*"We generated choropleths as the best way to show the per capita crime in the Chicago neighborhoods." Generally, qualify any statement that involves "best" or "optimal" with a mention of for what. In the case of choropleth plots they are good for showing patterns not for supporting judgment. Add "the spatial pattern..." Add a sentence linking this to the sentence about the time series graph. This might say something about the choropleth not showing temporal patterns well.*

**Addressed in Abstract**: “We generated choropleths as it’s the best way to show the spatial pattern of these crime rates, not for supporting judgment. In addition, we added the time-series graph to make up for the choropleths’ inability to convey temporal patterns.”

Comment 2

*“Your presentation included a bunch of striking crime statistics and you included references. Make sure you use APA citation style to link them to appropriate sentences in the main part of the document. Zotero makes this easy.”*

**See updated format of citations in Reference section and sources cited throughout the report.**

Comment 3

*Make sure you specify the units of the crime rate. Are those per person per year or per person per duodecade? Maybe convert to crimes per 1000 people per year?*

**Instead of crimes per 100 people, we changed it to crimes per 1000 people to be more useful to the audience. In addition, we went through the report and made sure to specify units of crime by saying “crimes per 1000 people per year” instead of a more general “crimes per capita”.**

Comment 4

*Unless there is a good reason don't mix typeface. Not sure why there is serif font for the map.*

**All fonts are now the same.**

Comment 5

*The grey background for the timeline seems a bit odd relative to the white for the maps.*

**Gray background removed.**

Comment 6

*Including an "Overall Chicago" reference lines in the timeline graph could be useful, but make sure you push it into the background so that it doesn't compete with the neighborhood data. Not sure you need the points on the lines: are they clutter? Make sure the axis limits are constant when you change neighborhoods.*

**Reference line added and faded into the background. Axis labels remain constant when changing between neighborhoods.**

Comment 7

*“One message we want the audience to take away is that a neighborhood that has a higher per capita non-violent crime does not necessarily have a high per capita violent crime, and vice versa. " Mention how this is a secondary focus because if this was the primary then a scatterplot of violent vs non-violent crime would have been more appropriate than a timeline. It all depends on your primary aims. Add a few sentences to discuss this.*

**Addressed in the Overview of the Visualization section: “**One secondary message we thought would be interesting to note is that there isn’t necessarily a direct relationship between the number of violent crimes per 1000 people per year, and the number of non-violent crimes per 1000 people per year. Through our exploration of the data, we discovered that by just looking at crime count per neighborhood, multiple neighborhoods have high per capita non-violent crime but low per capita violent crime, are with the highest crime count. Without investigating more and representing just the total crime count would have been very misleading in informing our audience. If this was our primary focus, then a scatterplot of violent vs non-violent crime would have been more appropriate than a timeline.”

Comment 8

*It would be great to select the neighborhood for the timeline from the map. Also I think it is easy to change the selected neighborhood highlight from the fill to the color of the boundary.*

Minor Suggestions

*1) Avoid underlining for emphasis (main title). Typography nerds consider that to be appropriate only when you are using a typewriter and need a way to indicate bold to the typesetter of the manuscript. It really bothers a very small part of the population. Same with centered text.*

**Underlines removed from title.**

*2) Add page numbers*

**Page numbers added.**

*3) Use grammarly.com to do a final check of your writing. I am always amazed by what it catches.*

**Spell check complete!**