Chicago Crime Landscape

Nádia Soares
Departamento de de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
Porto, Portugal
up201804296@fc.up.pt

Rafael Santos
Departamento de de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
Porto, Portugal
up201804298@fc.up.pt

***Abstract*—This project explores the Chicago Crime dataset, extended with information about population density, socioeconomic indicators, languages spoken and the shape of community areas. We created a interactive tool, which includes interactive and animated map visualizations, and a dashboard, allowing to answer on the changes of crime over time, across the city and with respect to socioeconomic factors.**

***Index Terms*—data visualization, crime, time, maps, dashboard**

I. Introduction

This work intends to create data visualizations which explores the Chicago crime dataset.

For that, we started by doing an exploratory data analysis (EDA) and creating some graphics to help understand the data at hand.

We propose some questions based on the EDA performed and prepared the data to allow us to answer them. This involved manipulating the data to remove duplicate entries and rows with missing values and joining multiple datasets. Finally, the data was ready to be used.

The final result, is set of visualizations, in an interactive Shiny application, which allows the viewer to answer the questions we proposed and confirming the hypothesis we posed. It contains an interactive map, which facilitates the analyse of the correlation between the count of different types of crimes per capita and several socioeconomic indicators, for each community.

The application also shows choropleth maps of the most common primary type of crime, absolute counts of crimes, arrests, and domestic crimes with respect to the community areas.

Finally, the shiny application provides a dashboard for exploring other dimensions of the dataset, namely, time, the primary types of crime, arrests and domestic crimes.

II. Data Manipulation

A. Data cleaning

After importing and uniting the four files, we removed 1.770.469 duplicated entries. From the remaining dataset, 692.486 cases had missing values in at least one column. We also removed them and obtained a dataset of 5.478.330 rows, and 22 columns.

The first crime was recorded in January 1st, 2001 and the last one on January 18th, 2017. To avoid having one year with only 18 days, and one month, January, with more values than the others, we left 2017 out. The year of 2001 was also left out, because it contained a very low number of entries.

B. Datasets

To answer the questions we proposed, we needed to use other datasets, and join them with the crime data. The following datasets were used:

- **Boundaries - Community Areas** [1] - geographic information about community areas, needed to draw maps.
- **Population per Community Area per the Census 2008-2012** [2] - Population of each community, used to calculate the population density needed to answer question 3.
- **Selected socioeconomic indicators from Census 2008-2012** [3] - socioeconomic indicators used to answer question 3.
- **Languages spoken in Chicago per the Census 2008-2012** [4] -

Each of these datasets has *Community Area* column, which maps directly to the one with the same name in our crime dataset. Those were used to join all the datasets.

C. Feature Engineering

New features were added to the dataset, namely, by breaking down the date into:

- **Date** - containing only the date in the format yyyy-mm-dd.
- **Hour** - containing an integer from 00 to 23, representing the hour of the day.
- **Weekday** - a factor containing the day of the week.
- **Month** - a factor containing the month the crime occurred.
- **ALL** - a new aggregation of crime primary type that is the sum of the number of occurrences in a community area.
- **Population density** - obtained by dividing the population of a community area by its shape area.
- **Per capita crimes** - obtained by dividing the population of a community area by each number of occurrences of a crime's primary type.
- **Latitude and Longitude of centroid** - obtained by applying geospatial functions to the geometries of each community area to find their centroid's coordinates.

III. Questions

We did some preliminary data analysis of the several variables in the dataset and started by posing the following questions:

- **Question 1:** How did crime change over time?
- **Question 2:** How did crime change over geographically?
- **Question 3:** Which socioeconomic indicators have more correlation?

We will go through the four of them, detailing the visualizations created to answer them and how we explored the follow up questions that arose.

A. *Question 1: How did crime change over time?*

This is a very broad question, but we started by decomposing it into the following:

Does the total number of crimes vary over the years, months, days of the week and hours?

*a) **Visualizations:*** To answer this question we started by creating a barplot of the counts of crimes per year, to show the tendency. To see how each year diverges from the mean, we added an annotation with the mean of all the years.

A barplot of the months was used to complement this one, by showing the cycles over the months.

We then grouped the counts by year, and created boxplots to the cycles over the months, weekdays and hours.

*b) **Answer:*** Crime has had a decreasing tendency over the years, as can be seen in "Fig. 1".

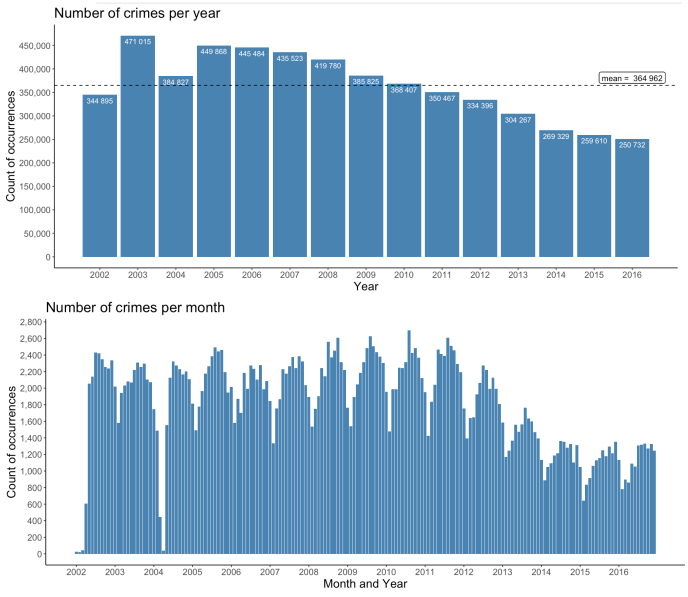

Fig. 1. Distribution of all crimes over the years, on the top, and over the months on the bottom.

Two months, March and April, have very low crime counts, creating a deep in the total number of crimes in 2004. This may have been caused by some temporary issue that let to many crimes to not be recorded. For that reason, and without more information, we cannot know for sure if 2004 really had a decrease in the number of crimes.

It is periodic with respect to the months and hours, being lower in cold months, and higher from around May to October. It is also much lower between 1am and 8am, and the hours between 8pm and midnight have a higher variance across the years. In terms of the days of the week, crime is slightly lower during the weekend.

*1) **Question 1.1**: Do all crimes have this same patterns?:* We then wondered if some types of crimes happened at different times.

*a) **Visualizations**:* To answer that, we created filters for the primary type of crime on the plots mentioned above.

*b) **Answer**:* There are types of crimes with different patterns compared to the overall trend. For example, *burglary* has a more or less constant trend until 2011, and then rapidly decreased to half the count of crimes per year, by 2015. On the other hand, *arson* appears to be increasing since 2013, and has higher counts during the weekends and from 22pm to 4am. *homicides* picked in 2016, an increase of 34.7 with respect to the mean, as shown in "Fig. 2" and "Fig. 3".

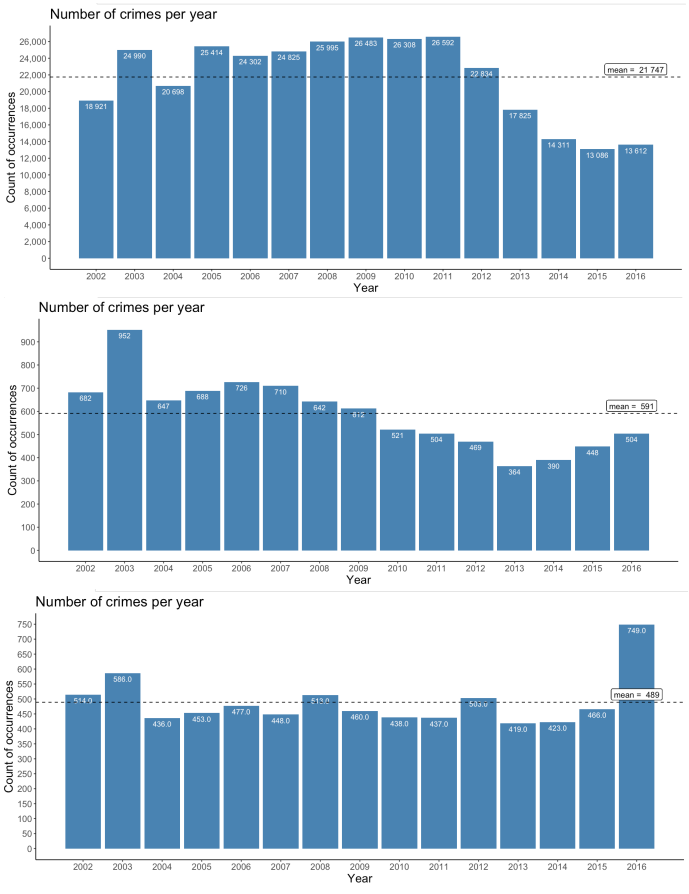

Fig. 2. From top to bottom, the distribution of *burglary*, *arson* and *homicides* over the years.

*2) **Question 1.1.1**: What is the tendency of the top 5 more common crimes?:*

*a) **Visualizations**:* For this, filtered the dataset by the five primary types with higher overall count, and we created a lineplot with with five lines, one per type of crime.

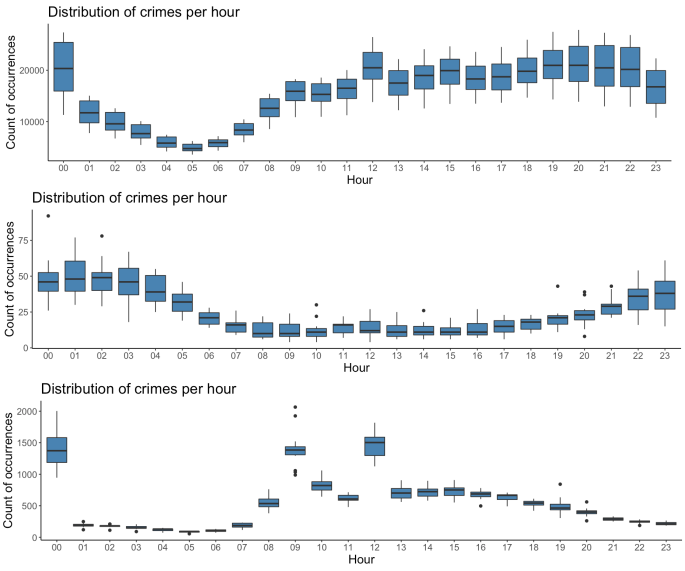

Fig. 3. Variations of different types of crimes over the hours. From top to bottom, the aggregated distribution of all crimes, *arson* and *deceptive practice*.

*b) **Answer**:* The most common crimes, from higher to lower, are *theft* , *battery* , *criminal damage*, *narcotics* and *assault*, as shown in "Fig. 4".

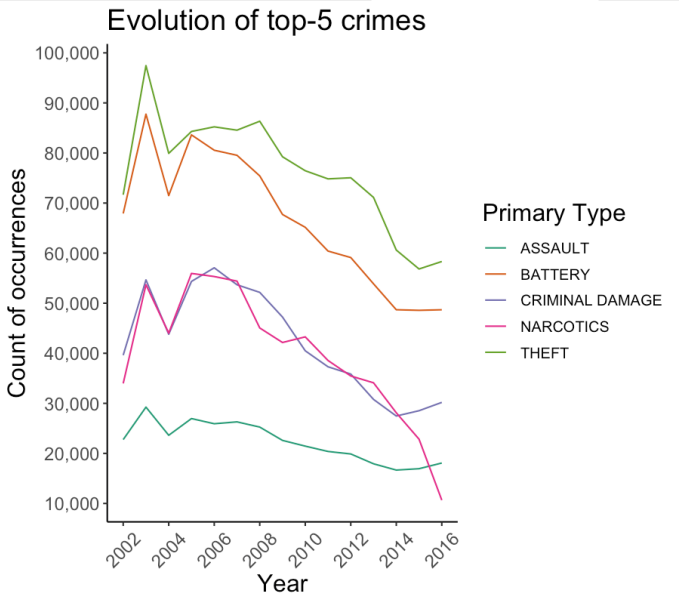

Fig. 4. Top 5 most common primary types of crime and their tendency over the year.

All five are decreasing overall, but *assault* decreased more slowly. Narcotics is decreasing much faster than the others and presents a steep decrease in the last years, while the other for had a slight increase in the last year.

B. *Question 2: How did crime change over geographically?*

We want to understand if, over the course of time, crime crime has been more or less constant across the communities, or if it has significantly increased or decreased in some regions.

*a) **Visualizations**:* For that purpose, we created a choropleth map, using leaflet, using the color to encode a discretized variable of the absolute counts of crime. We added a slide and a play button to animate the passage of the years.

*b) **Answer**:* It is possible to see in "Fig. 5" that, over the course of the years, some communities have become safer, in the sense that they have lower number of crime.

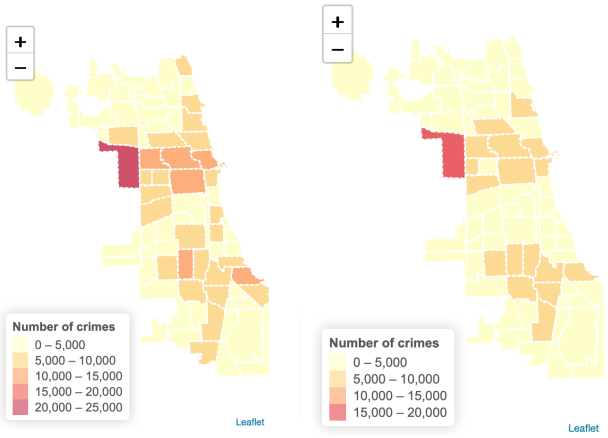

Fig. 5. Difference in the number of crimes from the year 2002, on the left, to 2016, on the right.

*1) **Question 2.1**: Did the type of crime change geographically?:* We also want to know if the most common primary type of crime of each region changed over the years. For that, we used the same type of visualization as before, but encoded the most common type of crime with color.

Four crimes were always the top crime of at least one community: *battery*, *criminal damage narcotics* and *theft*.

Some regions did sometimes change to another of the four crimes, but no major shifts were sawn over the years. "Fig. 6" shows the difference between the years 2006 and 2012.

*2) **Question 2.2**: Do some communities tend to have more arrests? Does that depend on the most common types of crimes?:* We could also ask if some communities tend to have more crime that lead to arrests. And, if so, does that depend on the types of crimes that happen there?

"Fig. 7" shows a choropleth map of the number of arrests per community.

Also, "Fig. 8", shows that crimes like *narcotics*, *prostitution*, *gambling*, *liquor law violation* and *public indecency* have a rate of arrest close to 100%.

We hypothesized that regions with high levels of arrests also have high numbers of the types of crime mentioned above.

We looked into Austin in particular, because it is the community with a higher number of arrests, and found that indeed, *narcotics* appears as the most common crime, as shown in "Fig. 9" corroborating our hypothesis.

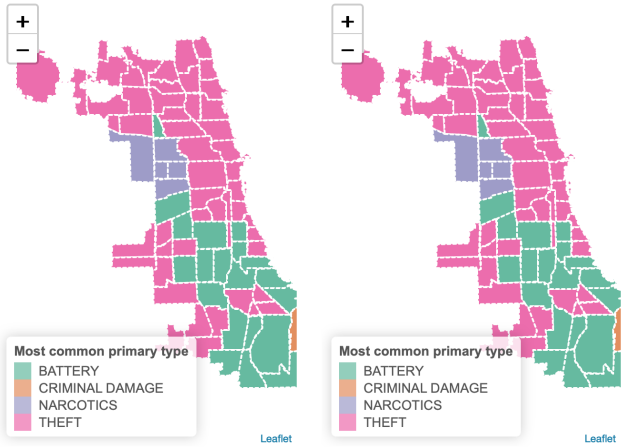

Fig. 6. Choropleths displaying the most common primary type of crime for each community, between the years of 2006, on the left, and 2012, on the right.

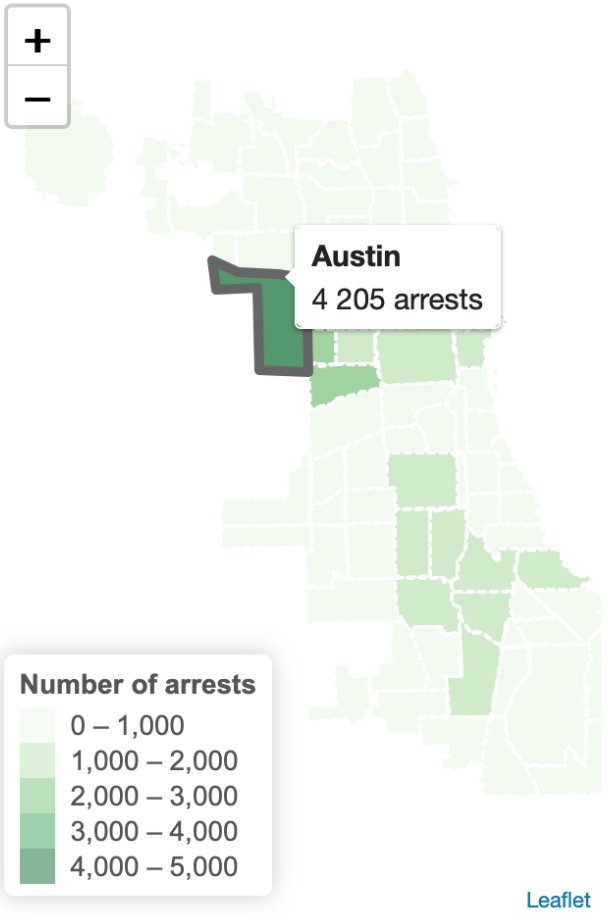

Fig. 7. Choropleth map of the number of arrests per community, in 2016. Austin is the community with more arrests.

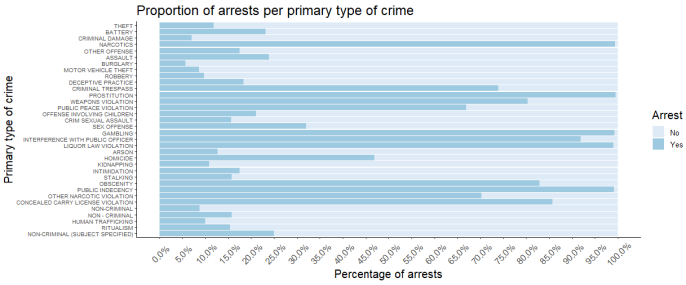

Fig. 8. Proportion of arrests per crime by primary type of crime.

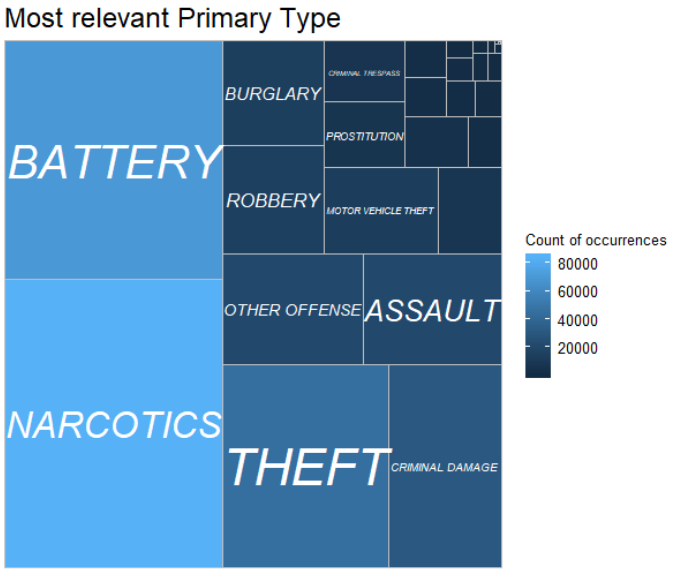

Fig. 9. Proportion of arrests per crime by primary type of crime.

C. *Question 3*: *Which socioeconomic indicators have more correlation to the amount of crime in a community area?*

To answer this question we tested the following indicators: *population density*, *per capita income*, *percentage of households below the poverty level*, *percentage of housing crowded*, meaning percentage of occupied housing units with more than one person per room, *percentage of unemployed citizens with age above 16*, *percentage of citizens over 25 years old without a high school diploma*, *number of Hispanic people per capita*.

*a) **Visualizations**:* Since communities may have different population density, we use the crime per capita in this plot. Also, since census data, from which in social indicators were obtained, only covers between 2008 and 2012, the crime dataset was also filtered to only keep those years.

To assess the correlation, we created a scatter plot using the library [5], where each point represents a community area, and x and y are the amount of crimes and the value of the indicator, respectively. The visualization also shows the Pearson correlation and draw a regression line on the points.

To visualize the location of the communities areas, we created an interactive map, using leaflet, with circle marks representing their center. The number of crime per capita,

which may be filter by type of crime, is shown as the size of the circle. The different indicators can be selected in an input selector and were discretized to be shown as the color of the circle.

b) ***Answer****:* The *percentage of households below the poverty level*, shown in "Fig. 10", and *percentage of unemployed citizens with age above 16* have high correlation to the amount of crime per capita in a community, 0.77 and 0.76 respectively. The remaining indicators did not show significant correlation.

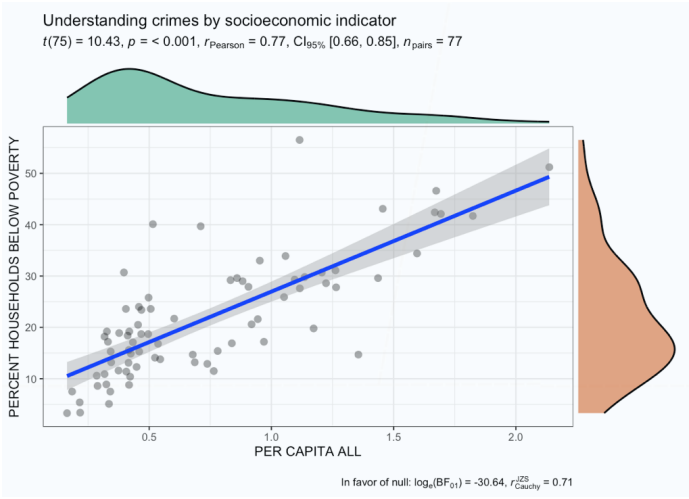

Fig. 10. Strong correlation between the percentage of households below the poverty level of a community area and the amount of crimes per capita.

The most prevalent crimes in when *percentage of households below the poverty level* is high are battery, shown in "Fig. 11", assault and criminal sex assault, with a correlation of 0.84, 0.82 and 0.79, respectively.

Percentage of unemployed citizens with age above 16 also presents similar correlations, of 0.85, 0.86 and 0.83, but also 0.86 for criminal damage.

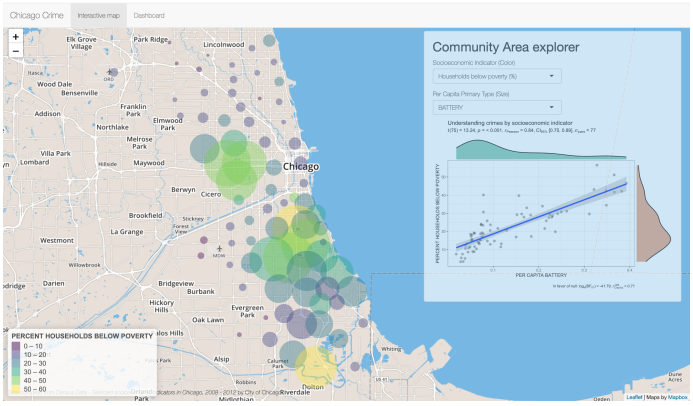

Fig. 11. Strong correlation between the percentage of households below the poverty level of a community area and the amount of battery crimes per capita.

1) ***Question 3.1****: Which crimes are more correlated to low and high income places?:* We now know that there is no relevant correlation between income and the amount of crime per capita. But could it be that, some types of crimes happen more often in richer areas and others in poorer regions? If so, which crimes are more correlated to low and high income places?

a) ***Answer****:* The primary type of crime with higher correlation is *arson*, with and inverse correlation of 0.60. Other than that, no significant correlations were found.

IV. Conclusions

We concluded that crime has been decreasing in the city of Chicago, and that different types of crime present very different patterns across time.

Crime varies slowly over the communities. Some regions did sometimes swap between the most common crimes, but no major shifts were sawn over the years.

Some socioeconomic factors, like the percentage of households below poverty level and percentage of unemployed citizens older that 16 years-old, are highly correlated with crime per capita.

Other factors like population density and income do not show a significant correlation to crime.

In the end, we were able to develop an engine for data analysis of the Chicago Crime dataset, containing dashboards and maps, as "Fig. 12" shows. It allows several degrees of freedom to explore multiple variables and dimensions.

A. Future work

We used the number of Hispanic people per capita as an example, but from the *Languages spoken in Chicago* dataset, other social could be extracted. There would need to be significant manual preprocessing of the dataset in order to aggregate the many different languages the dataset provides into less and more significant groups of people that tend to inhabit in close quarters, forming social communities based on ethnic groups. In that sense, an interesting next step would be to do that manual aggregation and provide more one more map plot that allow to understand which social communities are exposed to higher crime rates.

References

[1] City of Chicago Portal, Boundaries - Community Areas (current) , https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6.

[2] City of Chicago, Community Area 2000 and 2010 Census Population Comparisons, https://www.chicago.gov/content/dam/city/depts/zlup/Zoning_Main_Page/Publications/Census_2010_Community_Area_Profiles/ Census_2010_and_2000_CA_Populations.pdf.

[3] City of Chicago Portal, Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012, https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2.

[4] City of Chicago Portal, Census Data - Languages spoken in Chicago, 2008 – 2012, https://data.cityofchicago.org/Health-Human-Services/Census-Data-Languages-spoken-in-Chicago-2008-2012/a2fk-ec6q.

[5] https://github.com/IndrajeetPatil/ggstatsplot

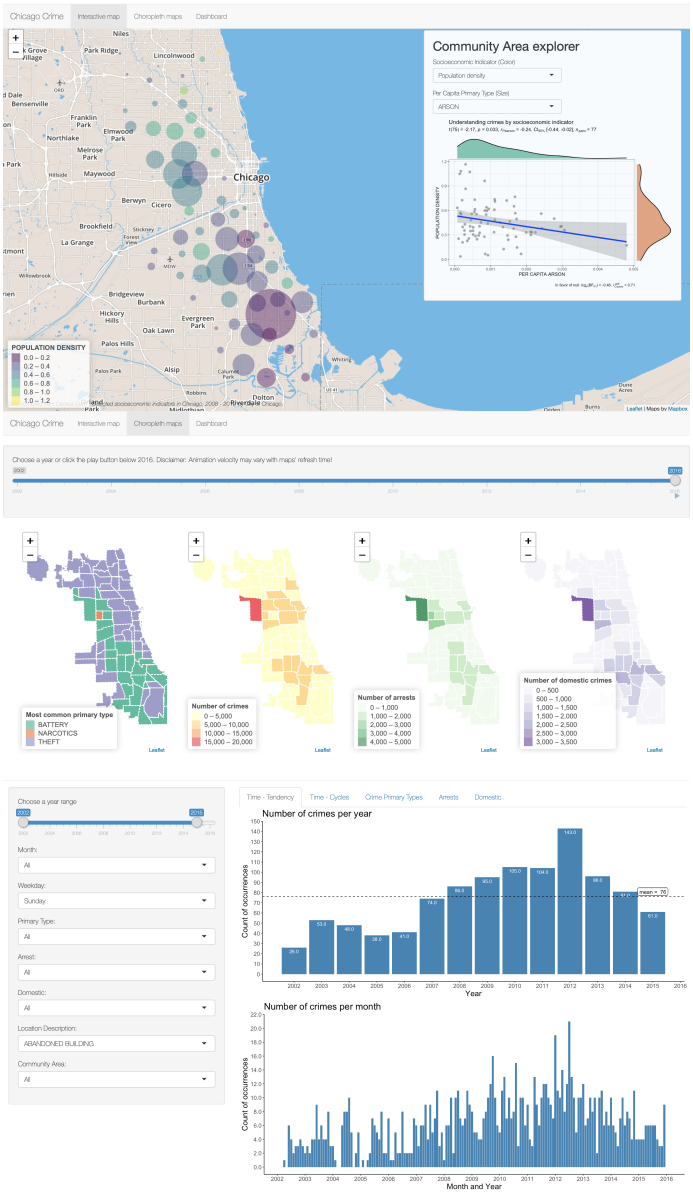

Fig. 12. Components of the resulting application. From top to bottom, the interactive map component, the choropleth maps and the dashboard for interactive exploration.