Chicago Crime Analysis: Theft

Kevin Chea, Tony Lieu

California State University

Los Angeles

E-mail : kchea2@caltstatela.edu, tlieu@calstatela.edu

**Abstract:** This paper offers a rough breakdown and overview of theft crime in Chicago, ranging from 2001 to the present day. Using data extracted from the Chicago Police Department's CLEAR system, an analysis is performed in hive to ascertain which theft crimes are the most common, what times they are committed, and where they are committed.

**1. Introduction**

The focus of this analysis will be the details surrounding theft crime in the city of Chicago. We will use HIVE to export the necessary data we need. Then, we will conduct a variety of different analyses on the data using different tools, such as a temporal-spatial analysis using Excel, and a geospatial analysis using Google’s Fusion Tables. We will use the information that we gather to attempt to answer questions such as what types of thefts are being committed, and when and where they’re being committed. A tutorial, which includes the codes used in this analysis, as well as other documentation can be accessed at this link: https://github.com/Tonylieu833/CIS-4560

**Data:** The data used in this analysis is sourced from the Chicago Police Department’s CLEAR system.[[1]](#footnote-1) The database comes in the form of a CSV file and contains information on crimes committed in Chicago from 2001 to the present day. The data is made up of 22 columns with approximately 6.76 million rows as of this writing. The data is updated daily, and contains information such as crime types, time, and location.[[2]](#footnote-2)

**Hardware:** The Hadoop cluster we used contains 5 nodes, with 10 OCPUs. The CPU speed is 2.20GHz. We have access to 150 GB of memory, with 678 GB of storage. The capacity of HDFS is 147 GB.

**2. Related Works**

Similar studies have been conducted in relation to crimes in Chicago, some of which were carried out using the same dataset. While these studies are similar in that they discuss criminal activity in Chicago, they are conducted differently, and offer new and interesting insights into this topic.

**2.1** **“Analysis of Crimes in Chicago 2001 – 2017”**

One example is a general analysis of crimes in Chicago by Vivek Mangipudi. Using the same dataset that we used, Vivek executes a broad analysis of criminal activity in Chicago from 2001 – 2017. Unlike our analysis, he doesn’t focus on a any single type of crime, but instead attempts to gather information on the entire spectrum of criminal of criminal activity within the city. Similar to our own analysis, the goal of Vivek’s analysis is to uncover key points of information, such as crime trends, most common crimes, and crime locations. His analysis incorporates a variety of different charts, graphs, and maps, all of which are interactable. He also incorporates a variety of different analysis types, such as geospatial and temporal. His findings revealed several interesting points of information. He discovered that 2016 saw a massive spike in homicide crimes. His geospatial analysis revealed that the airport and harbor areas experienced noticeably higher crime rates. One peculiar point that he noted was that there were a considerably higher number of crimes committed than there were arrests made [1].

**2.2** **“An Analysis of Crime in Chicago”**

Another example of Chicago-oriented crime analysis is one performed by Ry Storey-Fisher, Elana Feldman and Lisa Schroer. The first goal of this analysis was to gather information on the types and location of crimes within Chicago. The second and primary goal of the study was to use the initial information to find relationships between economic factors and crime patterns. The first part of the study begins with a broad analysis of all crimes in Chicago. The timeframe for the data in this analysis is limited to 2012. From there, the analysis shifts its focus to a specific type of crime, specifically homicides. Results are visualized using charts, graphs, and a map. The second part of the study investigates how crime patterns are possibly linked to economic factors. Data regarding drug and stealing crimes is compared against income data on a scatter chart. And indeed, the study found that there are various correlations between crime patterns and income across Chicago’s communities. Generally, most of these crimes were more likely to be committed in lower income communities. The study concludes that crimes are not evenly distributed across the city of Chicago, and are in fact correlated with the economic conditions of each area [2].

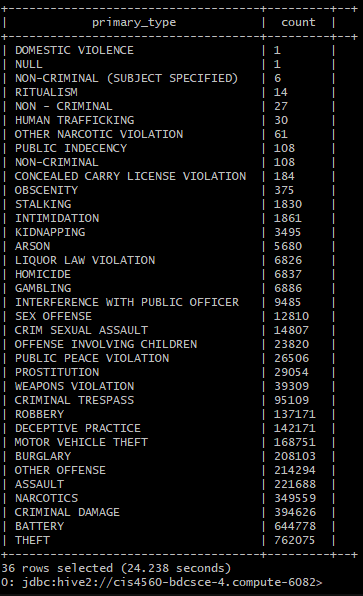
3. Queries

We began our analysis by downloading the data into Hadoop. From there, we created a table in HIVE and populated it with the data. We then proceeded to query the data to obtain information for our analysis. Finally, we exported our findings and visualized them using charts, graphs, and maps.

**3.1 Crime Breakdown**

**Approach:** Our first query is a simple count of the various types of crimes contained within the data. The crime type is contained within a column named “primary\_type”. Table 1 shows the results of performing a COUNT query on this column.

Table 1. Count of different crimes committed  
in Chicago from 2001 to present day.



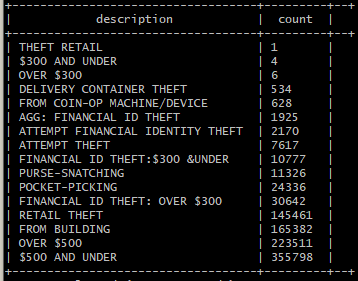
**Analysis:** According to the query, at 762,075, theft crimes are the most prevelant in Chicago, making up about 20% of overall crime since 2001, followed by battery at 644,778, and criminal damage at 394,626. It is important to note that the data categorizes vehicle thefts in a separate category. For the rest of our analysis, we will be focusing on non-vehicular thefts.

**3.2 Theft Types**

**Approach:** Next, we will attempt to discern what types of thefts are being committed. The dataset includes a brief description for each type of crime, which essentially act as subcategories. These descriptions are listed under a column titled “description”. Performing a COUNT query of this column for all crimes classified as thefts yields the results in Table 2.

**Analysis:** Based on this query, we can see that with 355,798 results, thefts valued at $500 and under are the most common. This is followed by thefts over $500 at 223,511, thefts from buildings at 165,382, and retail thefts at 145,461. The numbers drop off quite drastically after thefts from buildings. Note that among the top 4 results, retail theft is the least common at 145,461 records. This fact will be referenced later in the analysis.

Table 2. Count of 'description for theft crimes.



**3.3 Thefts by the Hour**

**Approach:** In the next analysis, we attempt to learn which hours theft crimes are most likely to occur by using the timestamps from the data. Within the data there is a column titled “date” which not only contains the date that each crime was committed on, but also the time. By performing a COUNT query that counts each theft crime for every hour, we were able able see how many thefts occurred within each hour. Figure 1 illustrates our results in the form of a graph generated using Excel.

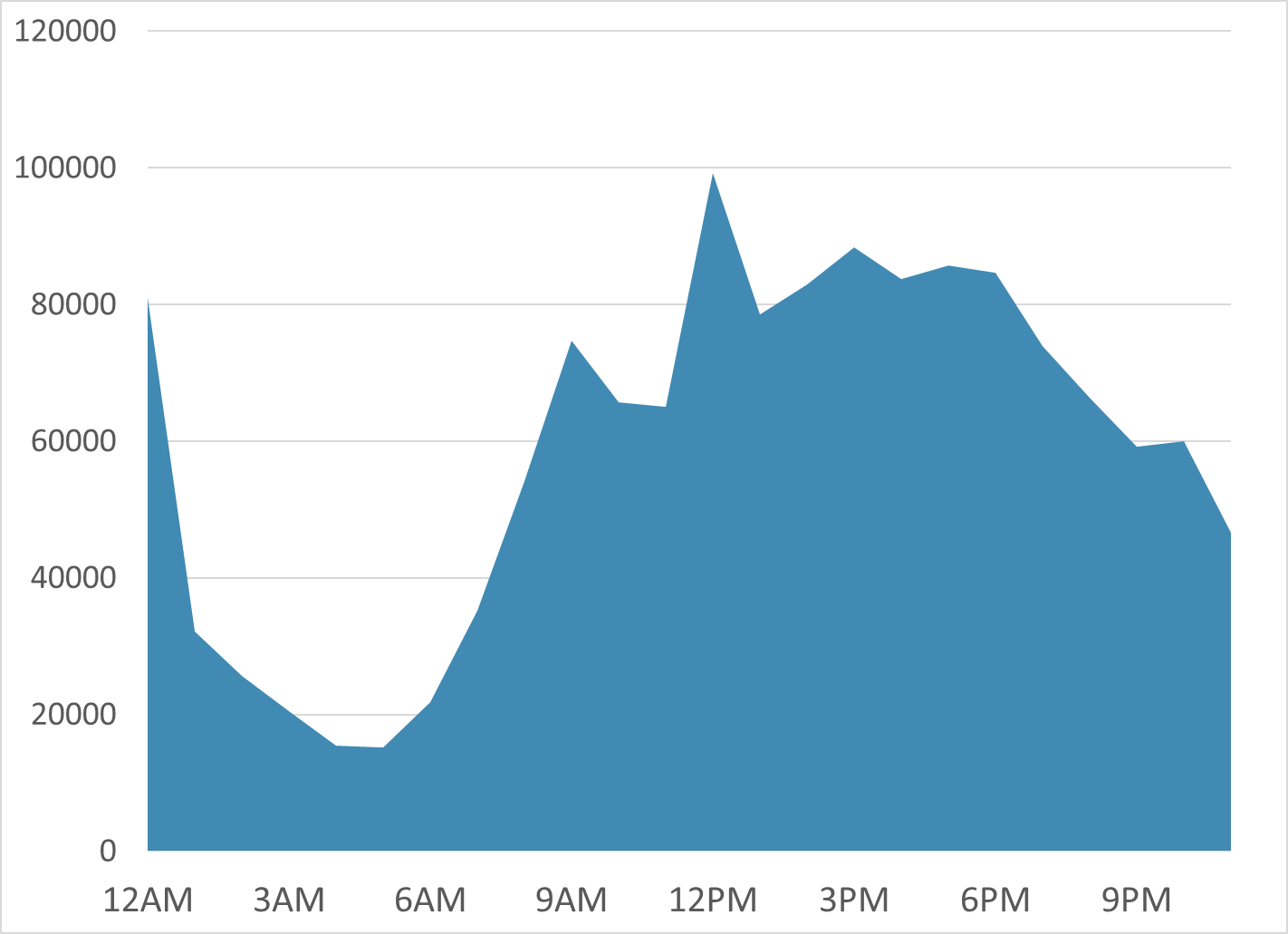


Figure 1. Chicago thefts by the hour, 2001 - present.

**Analysis:** By looking at the graph, we can see a couple of clear patterns. Most notable is that the bulk of thefts appear to occur between the hours of 9AM and 9PM. There also appears to be a spike in thefts around noon and midnight.

**4. Spatial Analyses**

Using the data we’ve gathered from our queries so far, we now proceed to our spatial analyses. The purpose of these analyses will be to obtain geographical information regarding theft crimes within the city. This will help us complete the “where” portion of our objectives.

**4.1 Temporal-spatial Analysis**

**Approach:** We decided to conduct a temporal-spatial analysis to see if we could spot any patterns in theft activity throughout the day. In the interest of gathering more useful information, we decided to narrow this analysis down a specific day and timeframe. We selected January 24th, 2004 because we found that it had the most crimes committed on a single day. We also limited the time frame from 9AM to 9PM because as we learned earlier, a majority of thefts occurred within this timeframe. We used Excel to visualize the results.

**Analysis:** The results didn’t turn out to be as interesting as we had hoped. Generally, the crimes occurred at a relatively even pace throughout the entire city. However, one interesting detail that caught our eye was how the downtown area seemed to have a much higher concentration of incidents than anywhere else in the city.

**4.2 Geospatial Analysis**

**Approach:** Once we noticed the concentration of crimes in the downtown area, we decided to generate a heatmap using a larger pool of data, in hopes of extracting more meaningful information. Like our temporal-spatial analysis, we decided once again to limit the timeframe from 9AM to 9PM. However, instead of using data from a single day, we decided to generate a map using all theft records in the dataset that fell within this time frame. Figure 2 shows the resulting heatmap that was generated using Google’s Fusion Tables.[[3]](#footnote-3) In addition to the heatmap, we also decided to create a map showcasing the downtown area of Chicago, with each theft incident individually plotted on the map. This map can be seen in Figure 3, and was also created in Fusion Tables using the same data.

**Analysis:** Again, in Fig. 2 we can see a strong concentratin of thefts happening in the downtown area of the city. However, we can also see a few other notable points of interest. Towards the northwest, the O’Hare International Airport shows up as a hotspot. The stronger hotspot towards the southwest portion of the map is a shopping mall. Along with the downtown area, one thing all of these hotspots have in common is that they are place where large numbers of people tend to congregate. Looking closer at the downtown area in Fig. 3, we can see high concentrations of thefts occurring near the bay area, as well as along the major streets. We know that these can’t be vehicle thefts, as we have already filtered the vehicle theft category out of our data. Our next analysis will focus on uncovering the reason for these concentrated patterns of thefts.

**4.3 Zoning Map Analysis**

**Approach:** To explain these theft patterns, we will be using a city zoning map for the city of Chicago. The map was created by DataMade, and was created using publically available zoning data from the city’s official data portal [3].

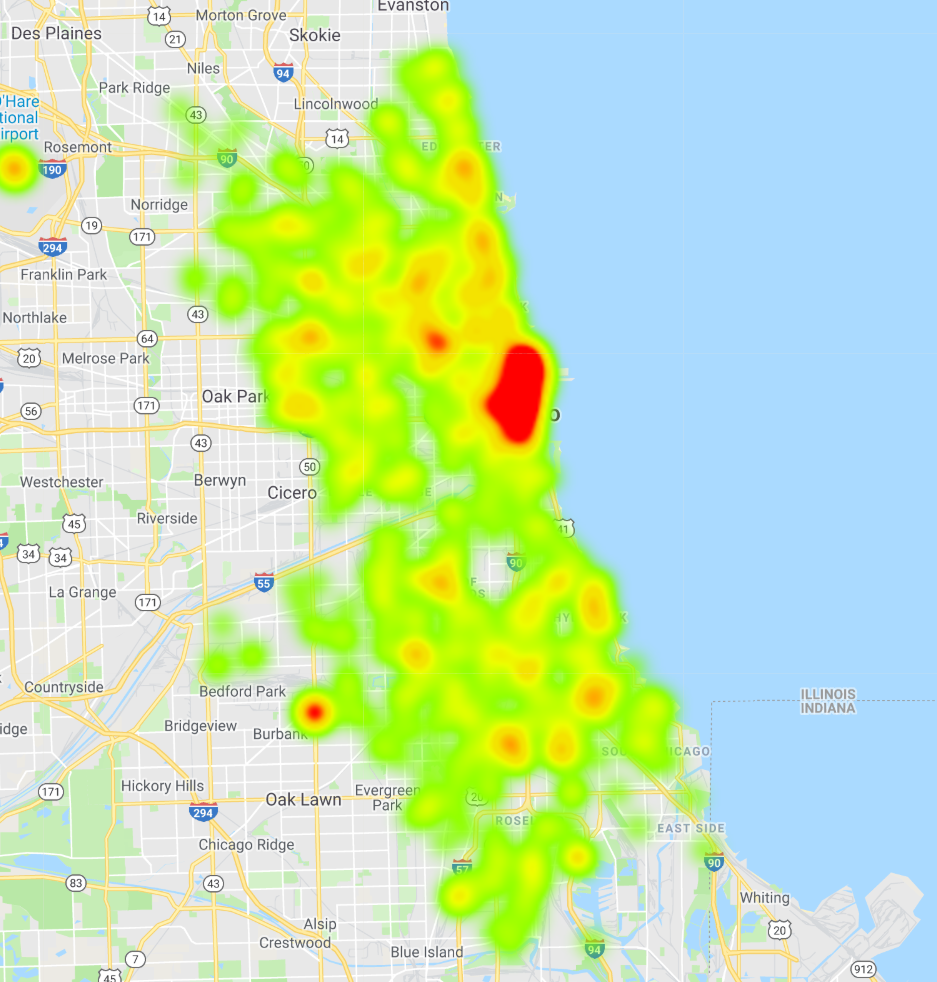


Figure 2. Heatmap of theft crimes in Chicago, 2001 - present.

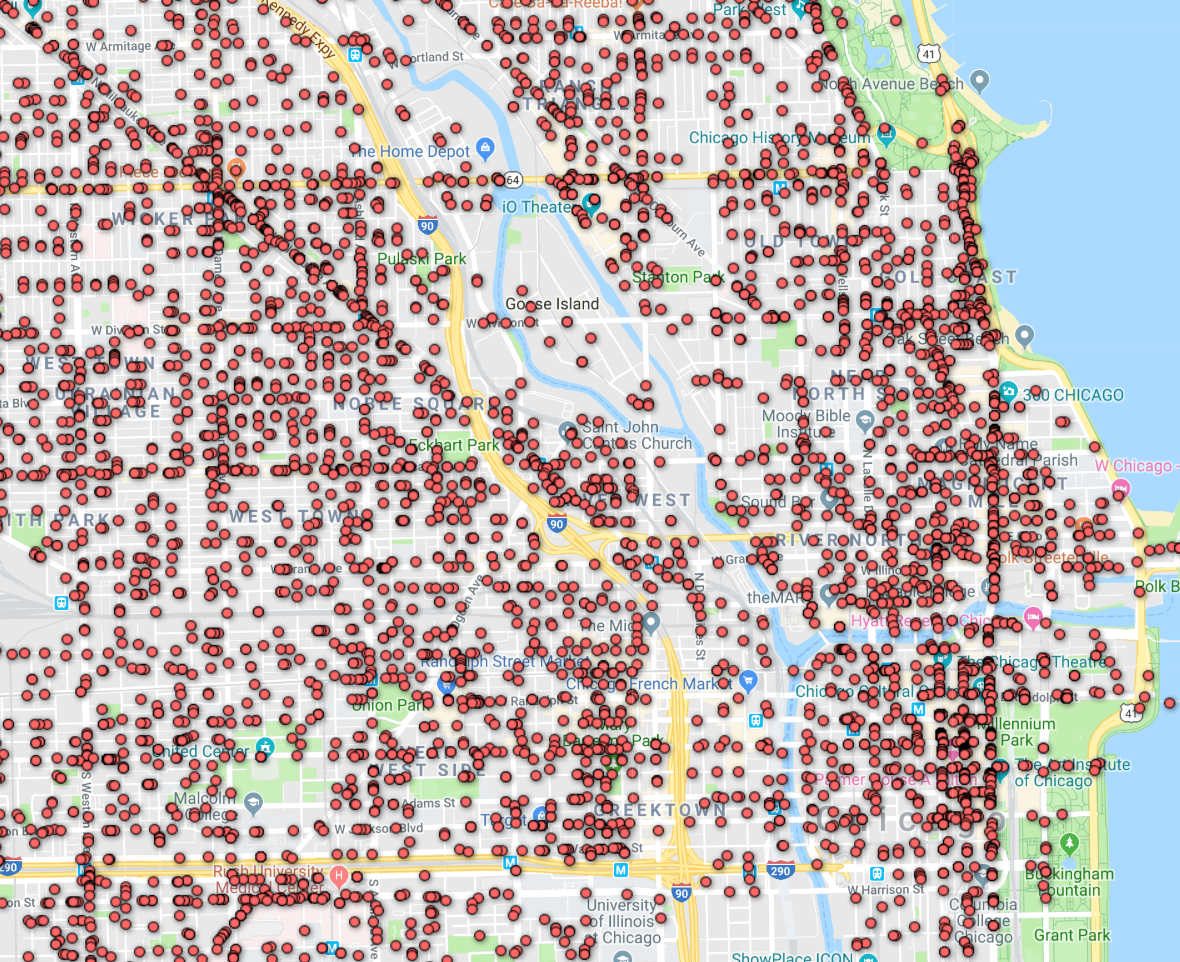


Figure 3. Downtown Chicago, which each theft  
incident individually plotted, 2001 - present.

Figure 4 gives us an overview of how the entire city is zoned, while Figure 5 highlights the zoning in the downtown area. The information is accurate as of 2017. For our analysis, we will be comparing the zoning map side-by-side with the maps we generated earlier using the crime data to see if we can spot any correlations between the two.

**Analysis:** If we compare Fig. 4 with Fig. 2, we can immediately notice certain correlations between the two maps. In Fig. 4, residentials areas are denoted in green, commercial zones in blue, industrial zones in yellow, and red is used for everything outside of those categories. We can clearly see that the blue commercial zones fall in line almost perfectly with hotspots on the heatmap. Notable areas include the mall, which is shown as a blue rectangle on the zoning map towards the southwest, and of course, the downtown area. If we compare Fig. 5 with Fig. 3, we can begin to understand why the thefts are spread out the way they are. Again, incidents seem to be concentrated in the commercial zones, denoted in blue. In other words, many of these thefts tend to occur in or around the shops and businesses that line the streets.

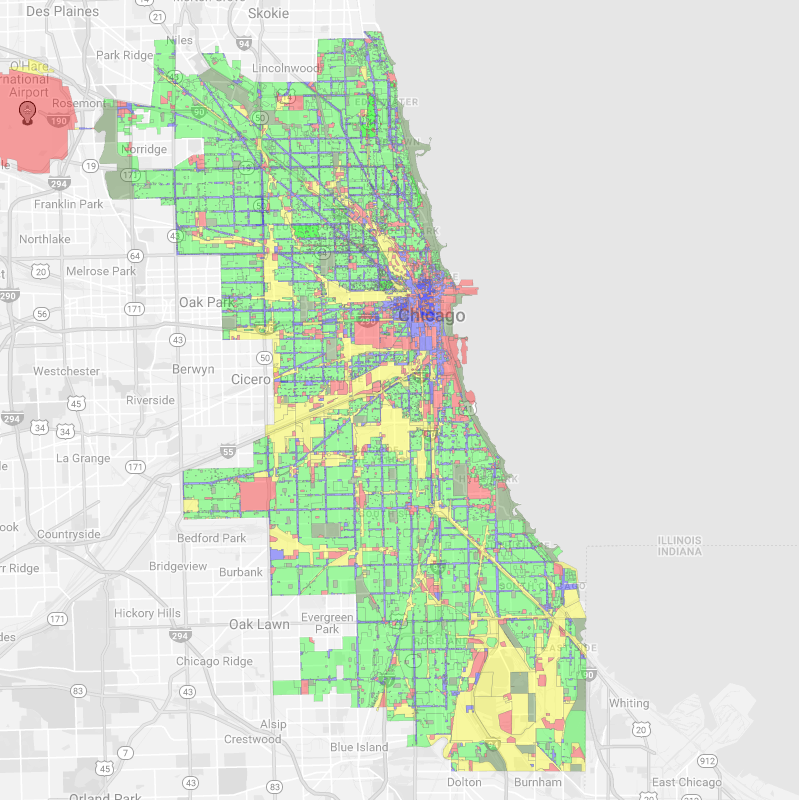


Figure 4. Zoning map for the city of Chicago, 2017.

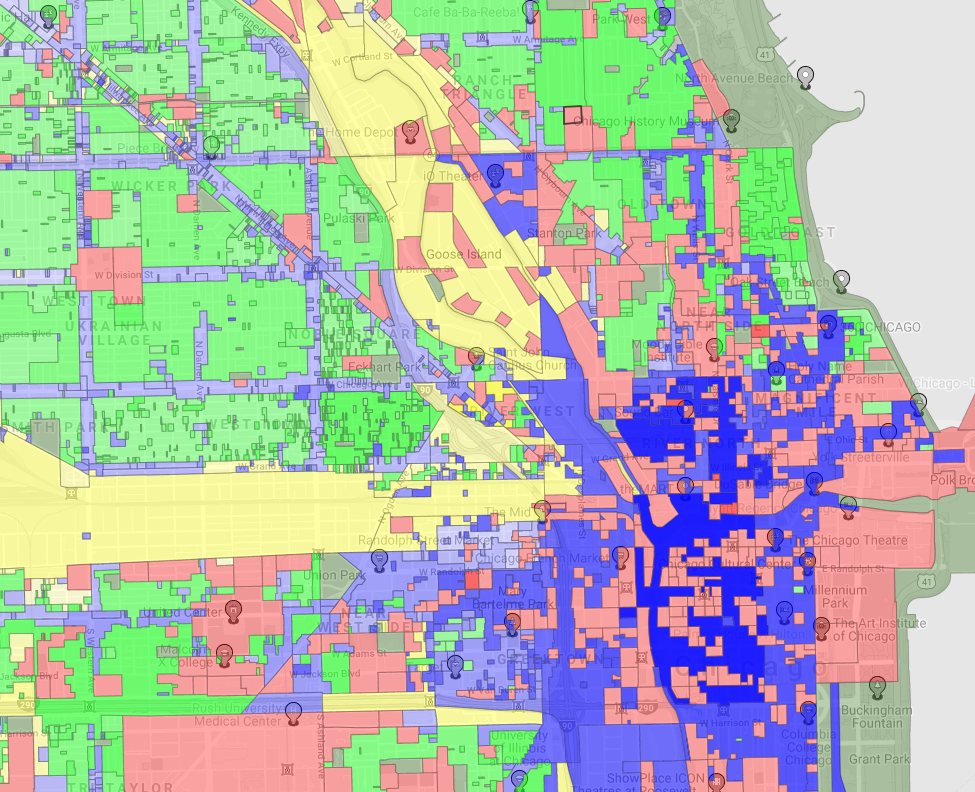


Figure 5. Zoning map of downtown Chicago, 2017.

**5. Conclusion**

Overall, we believe our analysis was sufficient enough to answer the questions we initially asked. By querying the data, we were able to learn that theft has been the most common form of crime in Chicago. And of all theft crimes committed, thefts valued at $500 and under have been the most common. A vast majority of thefts occur between the hours of 9AM and 9PM, while also spiking around noon and midnight. Thanks to our spatial analyses, we learned that most thefts tend to occur in areas where large numbers of people come together, such as shopping and dining areas. Our heatmap showed us that the downtown area had the highest concentration of theft incidents by far. Upon closer inspection, we can see that many of them are happening along busy streets. By using a zoning map, we discovered that these areas of high theft concentration were primarily commercial zones.

### References

[1] Mangipudi, Vivek. “Analysis of Crimes in Chicago 2001 – 2017.” *Amazon Web Services*, 28 July 2017, <https://rstudio-pubs-static.s3.amazonaws.com/294927_b602318d06b74e4cb2e6be336522e94e.html>.

[2] Storey-Fisher, Ry, et al. “An Analysis of Crime in Chicago.” *RPubs*, RStudio, 23 May 2017, <https://rpubs.com/ry_lisa_elana/chicago>.

[3]”2nd City Zoning.” *2nd City Zoning*, DataMade, September 2017, <https://secondcityzoning.org/>.

1. CLEAR: Citizen Law Enforcement Analysis and Reporting [↑](#footnote-ref-1)
2. Data URL[: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2](:%20https:/data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2) [↑](#footnote-ref-2)
3. Google Fusion Tables: <https://support.google.com/fusiontables/answer/2571232> [↑](#footnote-ref-3)