

# Big Data Investigation: Chicago Crime Dataset with Hadoop MapReduce

Emilio Singh	Francois van Greunen	Henry Wandera	Flip Nothling	Tafara Freddie Hove
University of Pretoria	University of Pretoria	University of Pretoria	University of Pretoria	University of Pretoria
u14006512	u18349863	u17253129	u71230191	u18278150
u14006512@tuks.co.za	fgreunen@gmail.com	u17253129@tuks.co.za	u71230191@tuks.co.za	u18278150@tuks.co.za

**Abstract**—Visualisation of Big Data is an important part of the big data analytics process, and can often reveal interesting patterns from a dataset. This report examines a Big Dataset, Chicago Crime statistics (2001-present), by using various visualization techniques.

## I. INTRODUCTION

The Chicago Crime Incidents dataset is a Big Dataset [1]. In order to visualise such a dataset, novel techniques should be employed to deal with the characteristics of big data. Novel big data visualization techniques allow exploration of the dataset in a visual manner, in order to find hidden patterns and to facilitate greater understanding of the dataset.

This report is a follow-up of a previous report, where MapReduce was employed in a variety of ways to process a big dataset. Some of that processed data can now be visualised, and the visualization thereof is part of this report.

The structure of the report is as follows. Section II will present the visualisations, grouped based on the algorithm used to create the data that is being visualised. Additional discussions are presented there as well. Section III will conclude by giving a brief and high-level summary of what has been learned.

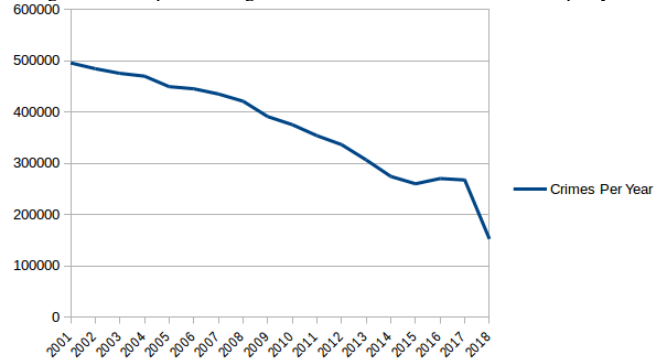
## II. VISUALISATIONS

The visualisations presented below are grouped into a number of categories based on which algorithm was used to create the data for visualisation, i.e. the results of the previous processing using MapReduce.

### A. MapReduce Aggregations and Summations

Presented in this section are the simpler visualisations of the aggregated dataset that can be produced with MapReduce. These visualisations are important because they can be used to establish a baseline for later visualisations which are designed to extract more involved insights. Additionally, some information from the dataset is difficult to extract with conventional means, in a timeous way due to its size, so MapReduce aids in acquiring these lower level insights faster.

Figure 1. Graph showing the number of crimes committed per year



What can be observed from Figure 1 is that crime is, in fact, decreasing over time. The general trend of the graph indicates that there is a steady downwards trend for crime incidents from the years of 2001–2014. In 2015, there is a slight increase in crime that dips sharply in 2018. Given that 2018 is not fully finished, data from the year is incomplete and this would account for the uncharacteristic decline. However, this information does correlate with the reporting that crime is declining within the city [2].

Figure 2. Graph showing the frequency of domestic crimes per year

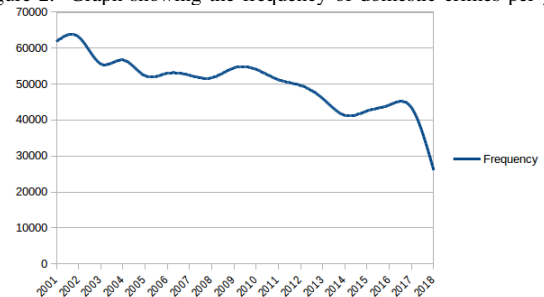


Figure 2 depicts the frequency of crimes of a domestic nature over time. Crimes of this type are never extremely common, as many victims are typically disinclined to report that crimes occurred against them. However, despite this, the graph does show general decline in the rate of crimes of a domestic nature. There are more peaks, periods of increase, than in Figure 1 which suggests that domestic crime is not as steadily decreasing as the rate of overall crime in the city.

Figure 3. Graph showing the arrests made against arrests failed

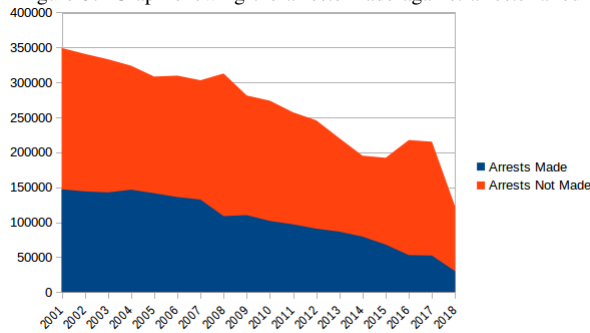


Figure 3 shows the number of crimes that led to an arrest, as well as those that did not. There are many more crimes where an arrest were made than where one was made. 2008 and 2016 stand out particularly in this graph as they are peaks for the number of arrests not made. In the case of 2008, the peak in unresolved crimes saw a corresponding dip in arrests made whereas in the case of 2016, the peak did not correspond to a significant trough.

Figure 4. Arrest ratio over the entire period (2001-2018)

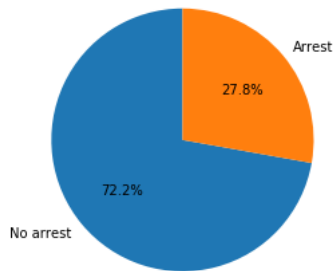


Figure 4 shows the percentage of arrested crimes over the entire period, and does not take into account any inter-year differences in arrest rate. Slightly more than a quarter of all crimes led to an arrest.

Figure 5. Arrest ratio, per year, over the entire period (2001-2018)

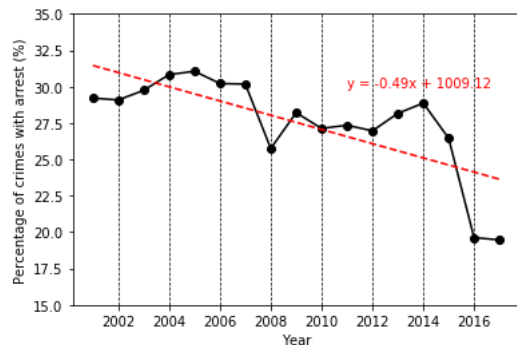
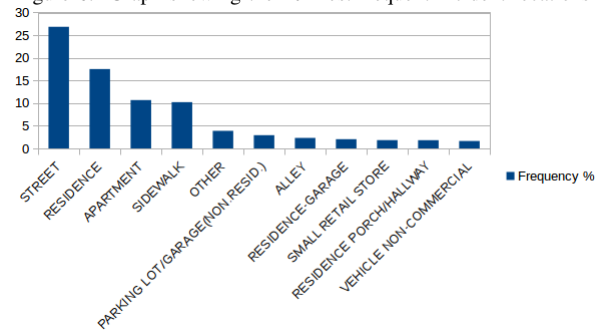


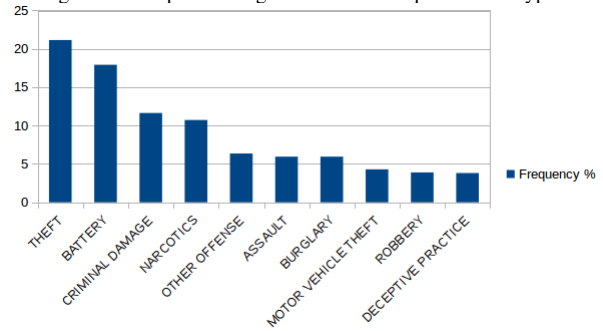
Figure 5 shows that the arrest rate is decreasing over time.

Figure 6. Graph showing the 10 most frequent incident locations



There are over a hundred types of location markers for incidents in the dataset. However, as Figure 6 shows, only a few of the location types account for most of the crime incidents. The top 10 incident location types, by frequency of occurrence in the dataset, account for over 81% of the incidents. In particular, street crime accounts for more than a quarter of all crime in the city and the other types involve similar urban components, like sidewalks and alleys.

Figure 7. Graph showing the 10 most frequent crime types



The dataset contains several hundred types of crime incident types. In Figure 7, it can be seen that the top 10, the 10 types with the highest frequency, account for over 91% of the crime in the city over the time period. The most frequent crime, at just over 20% is theft, followed closely by battery, criminal damages and narcotics. Interestingly enough, despite apartments and residential areas being more common as locations of crime, the crime of burglary, is not more common. Also deceptive practice making it onto the list when most of the other crimes involve theft or physical violence is interesting because it shows that there is a criminal element in the city that profits on the gullibility of the citizens if almost 5% of all crime in the city is related to deceptive practice.

## B. Geolocation Data

Presented here are visualisations of the geolocation data contained within the dataset. This merits a separate section due to the complexity of including geolocation data with other more conventional attributes. The visualisations that follow represent various cross-sections of the dataset, across different dimensions, that aim to show how crime is distributed across the city.

Figure 8. Crimes in ward 43 in 2017 with an arrest

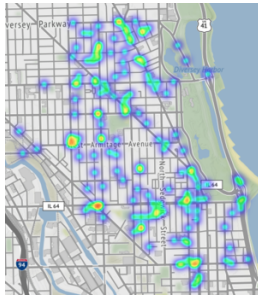
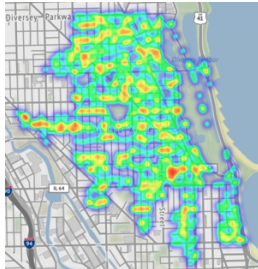


Figure 9. Crimes in ward 43 in 2017 with no arrest



Figures 8 and 9 show where arrests were made and not made in ward 43 in 2017. There are several small hotspots where arrests were not made, and this information could be used to increase the arrest rate.

Figure 10. Motor vehicle theft in Chicago at gas stations in 2017

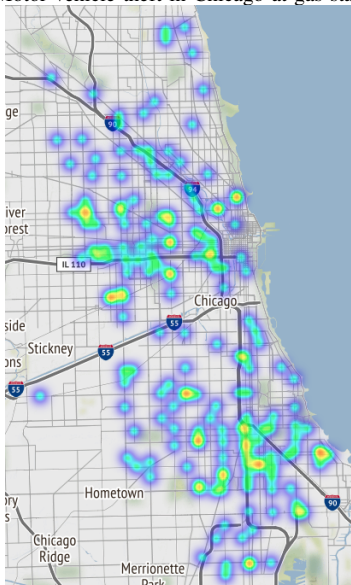


Figure 10 shows the locations and hotspots for motor vehicle theft at gas stations in 2017. This information could be used to improve policing at these locations, in order to lead to crime prevention or more arrests.

Figure 11. Burglary in Ward 24 in 2002

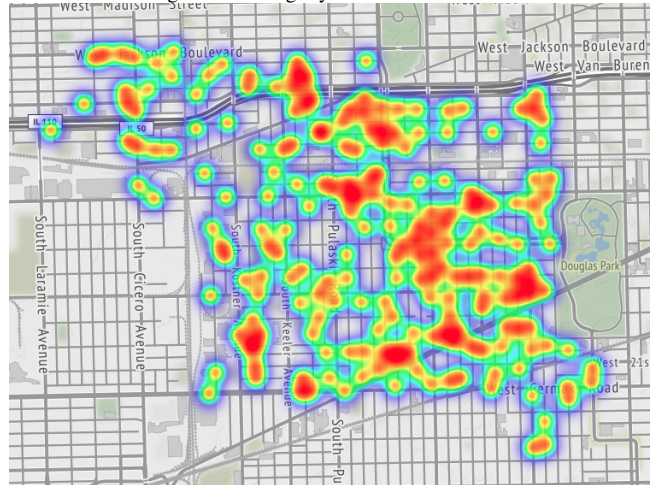


Figure 12. Burglary in Ward 24 in 2010

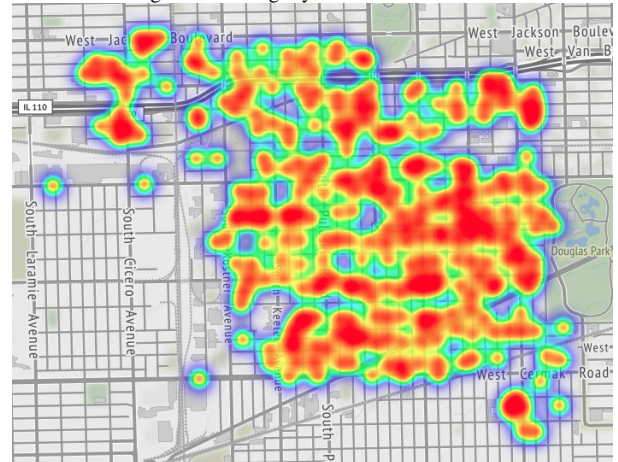
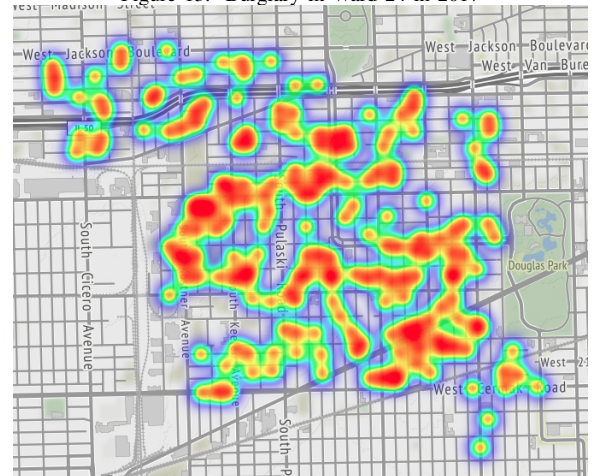


Figure 13. Burglary in Ward 24 in 2017



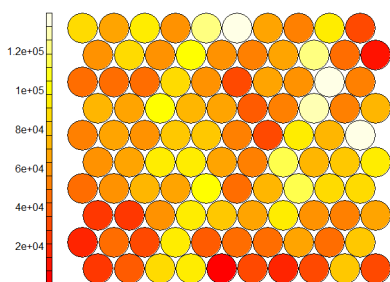
Figures 11, 12, and 13 show burglary events in ward 24 in 2002, 2010, and 2017, respectively. There is a clear difference in where these type of crimes occur in these different years. The heatmap for 2010 shows a drastic increase in the number of burglaries, especially

in the eastern regions of ward 24. The heatmap for 2017 shows that burglaries have decreased to 2002-levels, although the spread of burglary has changed when compared to 2002 - burglary is now occurring on the west side, where it was normally prevalent on the east side of ward 24.

### C. SOM

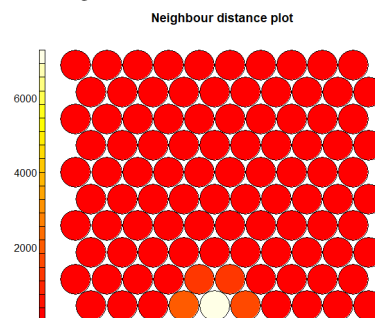
The Self Organising Map (SOM) discussed in an earlier report is useful in terms of being able to provide topological visualisations without strict geographic limitations. So the visualisations presented here provide more general insights that are more condensed, in part because of the compressive capacity of a SOM. In terms of the SOM, two visualisations are presented: a Count matrix and a U-Matrix. The former is useful because it demonstrates how many of the nodes the patterns in the datasets would map to. The latter demonstrates how far each of the nodes are to each other. The usefulness of a counts matrix is to demonstrate clusters in the map as neurons that receive many patterns, will help to define boundaries between the overall topology of the dataset. The U-Matrix also demonstrates the physical distance that separates nodes. Nodes that are close in value, helps to define the way in which the boundaries are formed; that is, are the nodes separated by large distances in value, or small. Also note that data has been standardised to ease the training process.

Figure 14. Count Matrix  
Counts plot



Graph 14 displays a number of interesting trends. There are 3 massive “hotspots” in terms of 3 nodes that have massive numbers of crimes mapped to them. These hotspots are located around similar regions of high intensity. There are a few lower incident spots, particularly in the left bottom corner which represents a cluster of nodes with very relatively fewer incidents than the surrounding nodes. Broadly speaking, two bands of higher incident density can be seen, more or less partitioning the map along vertical lines of high incidence with less dense incident frequency moving outward from these.

Figure 15. U-Matrix



The U-Matrix in Figure 15 demonstrates a very different situation than what was observed in Figure 14. In particular, the distances between units is generally very low, with one particular exception of a singular node that is extremely different to the other nodes surrounding it. Interestingly enough, the node which is extremely dissimilar to its neighbours, has relatively few incidents mapped to it. This suggests that this node is somehow a geographical outlier in terms of where crimes are committed.

### D. Apriori Algorithm

The Apriori algorithm implemented in the prior report produced a number of particular insights in the form of association rules. These rules can represent relatively hard to acquire insights, in part because of the size of the dataset. Visualisation of these rules is relatively difficult to do without disrupting their meaning and so for this section, the most pertinent rules are presented in a tabular format which shows the association rule. The duration specification simply implies that evidence supporting the rule can be found through those years in the dataset and this adds an additional dimension to the rules as they can be applied to specific time slices.

Table I  
ASSOCIATION RULES PRODUCED THROUGH APRIORI MINING OF  
DATASET

Rule	Duration
theft, motor vehicle, street =>no arrest	2001-2018
narcotics, street =>arrest	2001-2015
criminal damages =>street	2004-2018
Ward 012 =>no arrest	2007-2012
Apartment, burglary =>no arrest	2007-2018
Ward 011, arrests =>narcotics	2015-2018
Sidewalk, battery =>no arrest	2010-2018
Ward 008, Ward 014, Ward 025, street =>no arrest	2010-2018
Ward 018, Ward 019, theft =>no arrest	2010-2012

The rules presented here are mixed bag of insights. Later, more in-depth analysis of them will be presented. The common theme to all of them is of course arrest and types of crime. These rules make some suggestion of what types of crime need to be worked on, as well as suggest problem areas in which policing needs to improve. There are few long running rules, with the exception of the first two which span for 18 and 15 years respectively. A period of 2 to 8 years is what the other rules will typically span and a greater number of the refer to more recent dates. This of course is better as more recent insights are generally more applicable than older insights on data that has been in the more distant past.



### E. Naive Bayes

A Naive Bayes classifier employs probabilities to classify a crime as leading to an arrest, or not. Some of the probabilities for an arrest per feature are displayed and discussed below. The confusion matrix was already presented in a previous paper, and will therefore not be included here.

Figure 16. Arrest ratios for domestic and non-domestic crime types

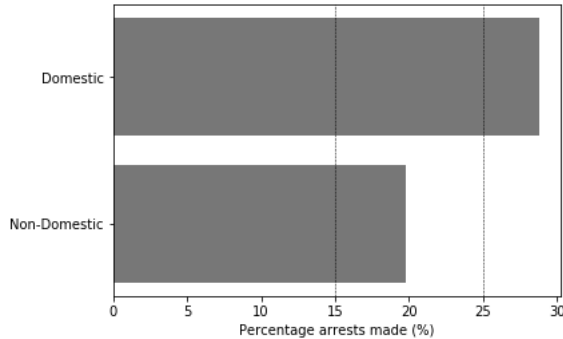


Figure 16 shows that domestic crimes are about 10% more likely to lead to an arrest being made.

Figure 17. Arrest ratios for the top 10 most frequent crime types

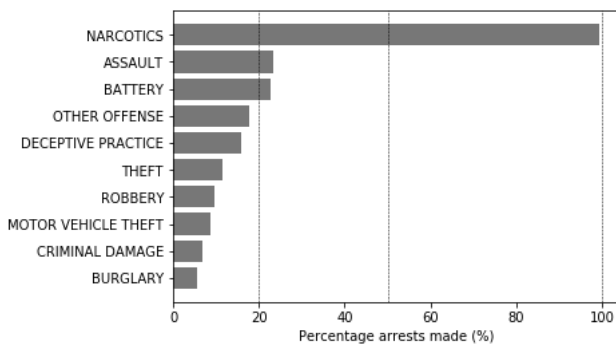


Figure 17 shows the arrest rate of various crime types (the top 10 most occurring types). Narcotics has an arrest rate of almost 100%. This could be explained by the fact that the discovery of narcotics on a person leads to an immediate arrest, unless that person escapes. However, subject matter expertise is required in order to understand this phenomenon better. Assault and battery have an arrest rate of more than 20%, which means that about 1 in 5 crimes of these crime types will lead to an arrest.

Figure 18. Arrest ratios for the 20 most frequent wards (geographic area)

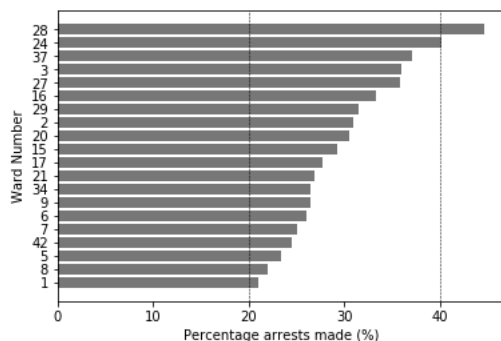


Figure 18 shows the arrest rate of the 20 most active wards. Ward 28 has the highest arrest rate at more than 40%. The lowest arrest rate of the top 20 most active wards is ward number 1, which has an arrest rate of just more than 20%.

Figure 19. Arrest ratios for the 20 most frequent crime locations

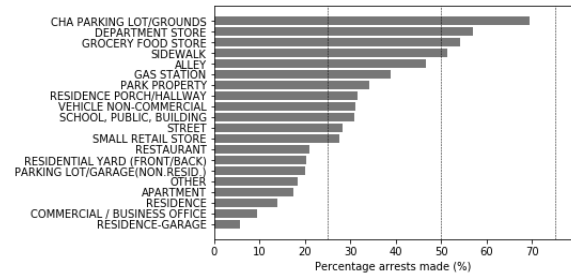
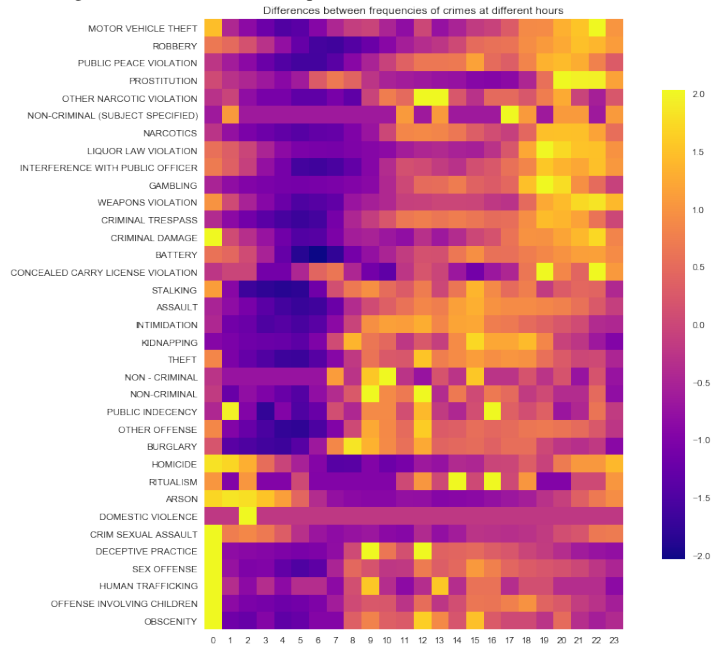


Figure 19 shows the arrest rate of the 20 most frequent crime locations. Parking lots experience an arrest rate of more than 60%, whilst apartments have an arrest rate of less than 20%.

### F. Heat Maps

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. Heat maps were used to perform the analysis of crime rates by location and type at different hours of the day.

Figure 20. Variations in frequencies of crimes at different hours



From Figure 20, the intensity of the colors vary from blue to yellow (negative values to positive values), representing a lower to higher frequency for the crime to occur at a given time respectively. The right upper part of the heat map shows crimes that experience high peaks at night. Other crimes like public indecency, ritualism, deceptive practice experience high peaks during day time between 9 and 15 Hours. There are low crime peaks in early mornings except for a few cases like arson, homicide and those at the left lower part of the map (midnight cases).

Figure 21. Crime occurrence in 24Hrs at the top 10 crime locations

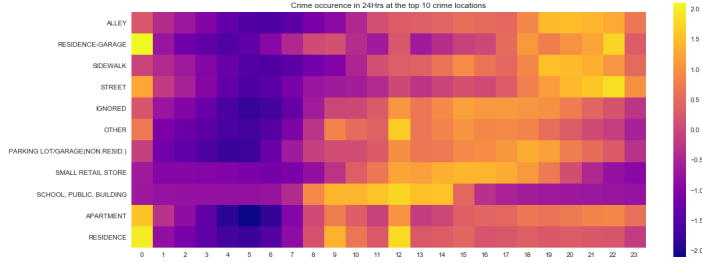
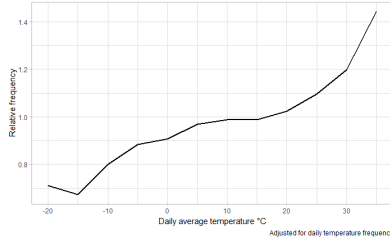


Figure 21 shows that locations such as school, public, building and small retail stores experience high peaks of crimes during day time. Morning hours between 1-7am are safe in these top 10 locations. Streets, alleys, sidewalk, and residence-garage experience higher peaks in evening and night hours. The variable named “IGNORED” is a collection of all locations that are not part of the top 10.

### G. Crime and Temperature

Combination of big data sets can lead to improved learning. The relationship between crime rates and temperature of United States cities have been investigated before. The New York Times recently reported on the effect of temperature on the crime rates in the United States [3]. For this visualisation, the Chicago Crime dataset was combined with the NOAA weather dataset daily temperatures. The temperatures measured by the weather station at the Chicago OHare International Airport, near the centre of Chicago, were used. The datasets were combined on the dates of the temperature readings and the dates on which crimes occurred. The effect of variations in the daily average temperature was eliminated from the combined dataset by dividing the crime frequency by the daily temperature frequency. The results show a strong correlation of temperature and the crime rate. Figure 22 shows the relative frequency of crime as it varies by temperature. On days with average temperatures of -20C, crimes occurred at a relative frequency that was 0.7 times the relative frequency of days with a temperature between 10C and 15C. On days with average temperatures of 30C, crimes occurred at a relative frequency 1.4 times the relative frequency of days with a temperature between 10C and 15C. However, this is just correlation between temperature and the crime rate and does not signify causality.

Figure 22. Correlation of Chicago Crime Rate and Daily Average Temperature



## III. CONCLUSION

Visualising big datasets requires a different approach to traditional dataset visualisation. For big datasets, the focus is more on data exploration and finding hidden information in the data. The use of heatmaps, geographic maps, and visualisations on processed data delivers insights into the dataset.

By using visualisation techniques on this dataset, various interesting patterns and anomalies were discovered. Simple visualisation of aggregations served as a starting point for analysis, and gave quick

feedback about the dataset. Geographic maps with heatmaps overlaid on top showed crime as it relates to location in Chicago, and those visualisations catered for longitudinal analysis over multiple time periods. The visualisation of the SOM and the Apriori table revealed interesting relationships that illuminate the dataset beyond what traditional visualisation could easily achieve. Visualising the probabilities of the Naive Bayes model provided insight into its workings, and shed some light about the difference in arrest rate among wards, for example. Heatmaps allowed visual exploration of the dataset with consideration for the time of day. Lastly, combining the dataset with weather data allowed a simple yet insightful visualisation of a correlation between temperature and crime frequency.

## REFERENCES

- [1] “Crimes — 2001 to present — city of chicago — data portal.” [Online]. Available: <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>
- [2] “Violent crime is down in Chicago,” May 2018. [Online]. Available: <https://bit.ly/2LPAqi6>
- [3] J. Asher, “A rise in murder? let’s talk about the weather - the new york times,” New York, sep 2018. [Online]. Available: <https://www.nytimes.com/2018/09/21/upshot/a-rise-in-murder-lets-talk-about-the-weather.html>