Crime in Chicago and Vancouver Analysis

Mahsa Daneshmand, Jacob Lower, Mohamad Almahmood, Max Gunton, Zhendong Su

INTRODUCTION

1.1 Project Description

In this project, we compare the crime rates of Vancouver and Chicago from 2003 through 2017. We compare the types of crime, GPS location of each crime, time and date, as well as the overall amount of crime in each city. We used three tools to visualize the data: Tableau, Bokeh, and D3. The Tableau visualization focused on the amount of murder, theft, and assault on each holiday of each year. The Bokeh visualization provided four basic charts for all crimes in the data sets. The D3 visualization mapped each crime to their corresponding geographical location. We found that Tableau provides exciting and creative diagrams that are functionally limited while Bokeh provides basic and traditional charts and diagrams that are informative and versatile. Thus, Tableau focuses on aesthetics but not functionality, while Bokeh focuses on functionality but not aesthetics. However, D3 combines the positive properties of both Tableau and Bokeh providing creative, informative, versatile, and aesthetically pleasing visualizations.

1.2 Research Question

Through our work, we tried to answer the following research question: “How to compare two large multidimensional datasets that have uncertainty/missing data, over time with minimal lie factor?”. We found that D3 is the most appropriate tool to answer our research question as it is versatile and focuses on both aesthetics and functionality.

1.3 Project Motivation

We were interested in comparing two very similar cities in two different countries and see if the difference in government and culture would impact the amount and type of crime in each city. To this end, we had some predictions as to which city would have more homicides, drug crimes, and thefts. We wanted to do this from the perspective of an individual. Typically, when a traveler wants to know how dangerous a location is, they are concerned over the specific locations they will be visiting. So, we have focused on what a hypothetical traveler between Chicago and Vancouver would want to know. We feel as though this type of visualization is underrepresented, as it is easy to find models of total crime and global crime, but not many visualizations of crime between specific locations.

2. MOTIVATION, DATA AND DATA QUESTION

2.1 Data Description

Each dataset covers the same topic but uses different headers. For the Vancouver data, the information dimensions are type of crime, year, month, day, hour, minute, location by block, neighborhood, and latitude and longitude. In total, this dataset contains 590,738 individual cases. The Chicago dataset is similarly formatted. It includes as data dimensions; ID, case number, date, location by block, IUCR, crime type, description of crime, location description, successful arrest (True/False), domestic, beat, district, ward, community, FBI code, latitude and longitude, year, updated on. Common dimensions between he datasets are primarily descriptive; crime type, time, location. The Chicago crime report is much more robust but contains more errors. This was uncovered during pre-processing of the data. The Vancouver dataset is more concise but less informative.

2.2 Data Questions

After analyzing the data, we decided on these primary data questions:

- We want to see if there is a correlation between time of day and amount and type of crimes in each city.
- What trends do we find in the types of crimes around holidays?
- Are our predictions correct that we will see more violent crime in Chicago and drug crime in Vancouver?
- Is there a common trend in location and violent crime between these two unrelated cities?

3. RELATED WORK

Multidimensional data visualizations have been studied broadly in the literature. Pillat and Elmqvist et al. proposed an interactive approach for multidimensional data exploration by scatterplots. Their technique takes advantage of using a matrix of scatterplots that provides an overview of the possible configurations, thumbnails of the scatterplots. It also supports interactive navigation in the multidimensional space. Their technique is implemented (SCATTERDICE) in Java by using the InfoVis Toolkitx [1]. This technique is relevant to our data as it allows us to interactively view our data and navigate its multidimensional space. CirculeView [2] is another technique used to compare and represent multidimensional continuous data with changing characteristics over time using a combination of visualization techniques such as treemaps, and pie charts (Keim et al, 2004). This technique is relevant to our data as the data is large, multidimensional, and temporal with changing characteristics over time which matches the goal of this technique. A 2011 study by Tran Van Long and Lars Linsen [3] attempts to understand the distribution of records based on density and to find clusters using areas of high density that exhibit correlations between dimensions or variables. The multidimensional clusters are then visualized using star coordinate visualization which allows for interactive analysis of the distribution of clusters and understanding the relations between clusters and dimensions. This technique would help in visualizing our data as the data represent different areas in each city and space-time clustering of crime events and neighborhood characteristics would help us answer our research and data questions.

4. DESIGN JUSTIFICATION

4.1 Choice of Visual Encodings

For Tableau we chose length, radial angle, and shape. Furthermore, for Bokeh we chose 2D Position, color hue, length, height, frequency, and shading. In addition, for D3 we included 2D Position, color hue, number/amount which are used on points in the main map visualization. Moreover, D3 used Size, X-position, Y-position, and color hue as visual encodings.

4.2 Layout

We chose to represent our Tableau visualization using a circle divided into 365 sections each representing a day of the year and the days corresponding to the holidays are labeled with the names of the holidays. We chose a circle because it is efficient in representing those days in a compact form. A bar stems from each day section and the bar's color indicates the type of crime while its length indicates the number of records for that specific crime in that specific day. This visualization can answer one of our data questions by providing the overall view of crime type and amount for each holiday through a given year. This allows us to see the trends of a specific crime type within a year.

To best emphasize the comparison of crime types between the two cities, Bokeh forgoes more aesthetic choices and instead affords a simple, familiar layout. Four graphs representing the selected crime's over the entire dataset, each year, each month, and each day. Each graph has a small legend in the top left, showing red for Chicago and blue for Vancouver. A line of each cities color shows the progression of each crime type. An array of crime types as selectable buttons on top of the graphs afford the user easy selection. This layout is simple, line graphs are familiar to most users and comparison is immediately obvious even between graphs. The selectors are simple, labels that darken when selected. Only one type is able to be selected at a time to ensure that comparison remains simple.

In the D3 visualization comparing the data using location was the focus. For this reason, the size of both maps is the same, and the layout is intentionally symmetric to avoid focus on one map over the other. Labels are put over the maps to indicate the city presented, but aside from this, the text is intentionally kept to a minimum. The maps tiles used were chosen because they provide important information but are not overly stimulating or distracting from the data being presented.

4.3 Interaction

In the tableau design, there are a few interactivity capabilities. The first is the selection of a year from the *select year* box. The second is the selection of a crime type from Type legend. The third one enables us to hover over each day to show more detailed information in an overlaid window. We chose to represent each crime type with a corresponding color which are Red, Cyan and Green. We chose those colors because they are easier to differentiate between them as they are distinctly different.

This visualization has simple interaction. On each individual graph, the user can pan and zoom over to get greater detail on the occurrences and exact date and time of each crime. Selection is done by clicking on the title of the crime type, which changes the data in each graph

This visualization allows for pan and zoom, drop down menu for selecting crimes, a checkbox for keeping crimes on a map, hover shows additional details of crime, and clicking on a crime shows all the crime in that immediate area (with 100m).

4.4 Style and Aesthetics

We chose a circle to represent a year inspired from the shape of a clock. We decided to represent our visualization in the form of a circle as it is more creative and aesthetically appealing than using a straight line.

Aesthetics in the Bokeh representation are limited to simple lines and dull shades. This was chosen to emphasize the comparison as it is, two scaling values. Aesthetics are limited, and so no information is occluded.

In the D3 representation, we chose to emulate a map. It's a familiar scene that anyone should be able to look at and understand. Color to show various types of crime were chosen based on culture: red for murder because it is an aggressive act and color for example and green for robbery because of money.

5. IMPLEMENTATION

To show the data on a circle in tableau, we defined two parameters. The Radial Inner parameter calculates the diameter of the inner circle and the Radial Outer parameter calculates the diameter of the outer circle. We also defined several measures. The Radial Field is defined based on the "Number of Records" column of the data. Therefore, the length of lines on the circle are representing the number of records of each crime type. Radial X and Radial Y represent the X and Y coordinate of the place we draw the line. Moreover, Holiday label dimension is also defined. This dimension labels the holidays' dates with the names of the holidays. To add interactivity capability of selecting the year, we created a dashboard action that filters the data upon selection a year.

The Bokeh visualization was implemented using the Python Bokeh libraries, some minor HTML scripting, and a large amount of data pre-processing in python. Bokeh provided intuitive functions and classes that were quick to implement and showed results quickly. Bokeh made it possible to arrange multiple visualizations of data into structures that could be displayed on a webpage with ease, resulting in the selection list and four graphs used in the final model. We primarily approached the designed with the intent to make something easy to understand, and so toggle one-of buttons were used for type selection. The graphs themselves were split into four time periods; total, yearly, monthly, and daily crime analysis.

Our third visualization, with maps was implemented using html, css, and javascript (with D3 library). This visualization was focused on location in addition to comparison. In order to lessen the amount the user is restricted, the maps support panning and zooming, while preventing the user from zooming too far out. This allows the user to explore the map how they like, but also keeps them focused on the area of interest. Because the raw data is quite large choosing a subset was important for the overall functionality of the visualization. If all points are plotted the visualization begins to lag and becomes too much of a burden on the browser and CPU. The selection and filtering functionality were implemented using a drop-down menu and checkbox, and although these were not the preferred methods, they were chosen in order to move forward with development.

6. DATA INSIGHTS

Each of our visualizations helped to answer the data questions through their unique designs and emphasis on their strengths. The Tableau visualizations were useful in answering the holiday crime problem for its specific crime types. Holidays are labeled clearly, and the amount and time of crime on that holiday displayed for each year. The visualization itself highlights dates and the labels of holidays if a crime of the type has occurred, making it clear as to which dates had more of which crime. Interestingly, we found that the beginning of the year, in general, had less crime overall. Perhaps the cold weather and general festivities centered around the family reduced the occurrences of homicide. However, the more commercial holidays had increases in thefts, especially around the infamous Black Friday and Boxing Days of the States and Canada respectively.

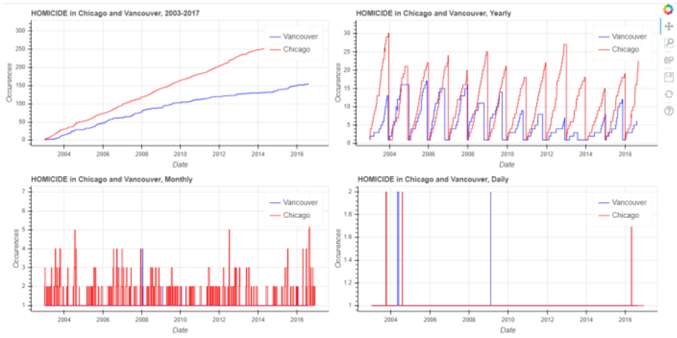


Fig 6.1. Bokeh homicide graphs.

The Bokeh visualization was useful in answering the question of the expected crimes between the cities; namely, homicide being greater in Chicago and drug crimes more so in Vancouver. Surprisingly, homicide rates are not that different in Vancouver than in Chicago. Although Chicago has more homicide, there was not much more than in Vancouver.

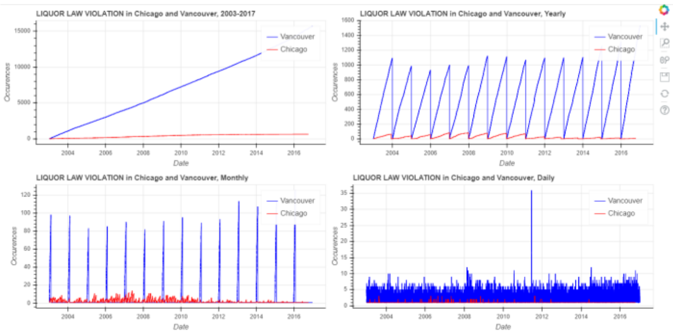


Fig 6.2. Bokeh liquor law violation graphs.

It was also surprising to see how much more alcohol-related crime there was in Vancouver than in Chicago. This leads us to believe that something was wrong, especially looking at the enormous spikes at the beginning of each month for Vancouver. This was a prime example of how the data was malformed, a lot of the points were clustered without days for the Vancouver dataset. This became even more clear when looking at other narcotics violations.

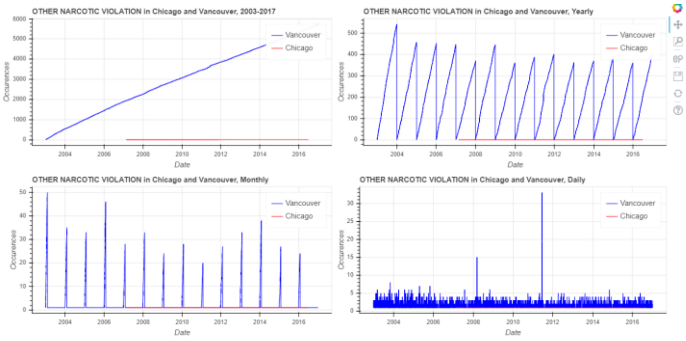


Fig. 6.3. Bokeh narcotic violation graphs.

The Chicago dataset does not start recording this type of crime until 2007, which makes this comparison useless. An interesting note, the one spike for Vancouver corresponds to June 15th, 2011, the date of the Stanley Cup Riot.

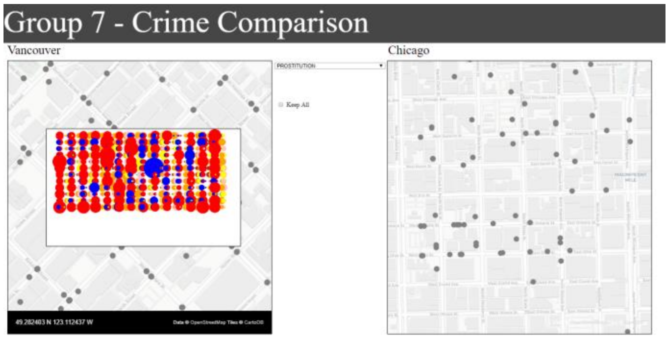


Fig 6.4. D3 crime comparison.

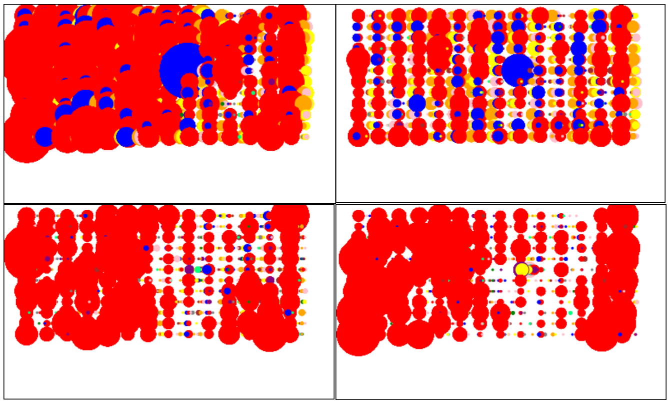


Fig 6.5. D3 visualization showing various Vancouver downtown location Data. Year is on the x-axis from 2003 on left to 2016 on right and month on the y-axis with January at the top and December at the bottom. each color represents a different type of crime and the larger the dot indicates a higher number of that specific crime in that month and year.

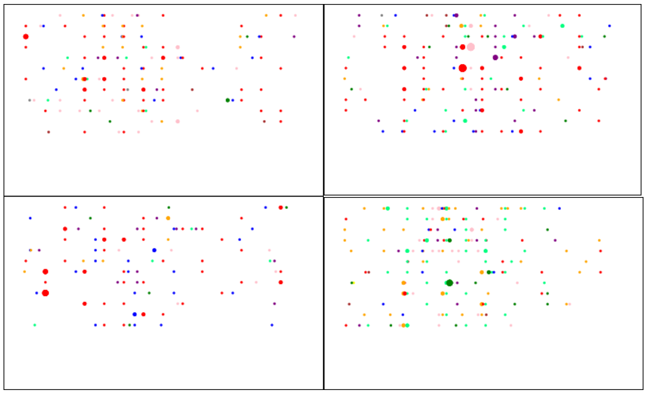


Fig 6.6. D3 visualization showing various Chicago downtown location Data.

One very interesting thing that the D3 visualization allowed us to discover was the dispersal pattern of crime in both Chicago and Vancouver. When looking at the data for specific locations there was a recurring pattern where Vancouver contained areas of very high crime as indicated by the first set of images above, and for whatever reason this was something that just wasn't seen in the Chicago data as indicated by the second set. In general, Vancouver was much more polarized in the areas of high crime, with a higher crime density. This is an interesting finding and given more time would be an interesting avenue for further research to determine the reason.

7. RESEARCH FINDINGS

Preprocessing of data was crucial to our understanding of the data. First, we had to make minor adjustments, such as to dimensions (date and time figures). Tableau required its own version of the data, had to be cut down and added new dimension for a count. Bokeh unveiled how the raw data was malformed, this was useful and proved to us that to understand data, a simple visualization is needed. We then better processed our data for use in D3. Each visualization, created in sequence; Tableau then Bokeh then D3, revealed difference challenges and helped the next address our research question. To reiterate our research question, we want to find useful methods to comparing large, multidimensional datasets, with malformed and missing data. Starting off in Tableau, we wanted to take a snapshot of the data and compare each set within itself. We believed that before we could compare these datasets, we needed to understand each dataset on its own. Each Tableau visualization selected three crime types and created a clock-like visualization over each year. We were able to watch as the crime rose and fell over the years in a noticeable pattern. Some data points seemed off, there were consistent themes were cases of homicide would occur only on the first of each month, as an example of the strange structure of the datasets. We then needed to look at the data more in depth, without aesthetics that could cloud the truth. Bokeh was used to perform this analysis, primarily to answer the research question in its most basic form: to compare the data in its raw state between the sets. Doing just do, as simple line graphs, it became clear that there were missing data points, types, and even entire segments set incorrectly (such as strings instead of integers and so on). After this realization, we cleaned up the data with a hefty amount of python data processing, pruning malformed sets revealed by Bokeh. Finally, we found which dimensions were most accurate; time, position, and a smaller set of types. We made the D3 visualization around these factors, emphasizing what was most correct, and displayed them side by side. In this way, by eliminating malformed data, and showing not only the most accurate but he most accurate components of each data case, we were successful in displayed these two, large, malformed datasets with malformed values with minimal inaccuracies.

8. DISCUSSION

The Tableau visualization can answer one of our data questions about crime types and holidays. Tableau is a handy and appealing software and is the best way of getting an initial idea about the data. However, getting to know all its capabilities takes a lot of time. Regarding our research question, tableau failed to answer to our research question as we did not put the data of Vancouver and Chicago together and each data was imported to tableau separately. Hence, the final visualization for each city is independent from the other one and placed on separate dashboard. Therefore, we cannot compare the data of two cities on the same view with each other.

The Bokeh visualization was instrumental in providing us with information on our datasets. We discovered, through our simple, straightforward design, which data points were invalid and how to categorize and extract them. The simple line graphs and quick ability to change between crime types made it easy to figure out which cases were not accurate. We were able to eliminate these invalid cases to reduce the lie factor of our final visualization. This being a key point of our research question, we were happy to find that Bokeh proved to be a great tool for processing and understanding large datasets with malformed values. We imagine that one could implement what we have made here to help in analyzing other large, malformed datasets with ease. If the names of each dimension are provided, any CSV can be visualized across a temporal dimension for faults and missing values. Some major limitations of our design are due to its simple nature. The implementation only looks at differences over time, changing this to another value would be difficult. The lines can overlap, giving the illusion that they are in sync when it is very possible that some missing values are obscured by the overlap. When there are major differences in the short-term analysis of data points, the graphs can become hectic and hard to parse, making it nearly impossible to decipher. Even so, this visualization was an important step in answering our researching question, shedding light on the importance of using simple visualizations to better understand our data and weed out inaccuracies.

The D3 visualization was able to provide answers to some of our data questions as well as touch on our research question. Comparing the crime data on the maps allowed us to clearly see which areas were subject to the highest amounts of a crime, and how they compared to each other. The use of maps also means that answering some more complex questions like, "Is there more crime in the industrial areas?" can be left to the interpretation of the user. More specifically we did not have to categorize the areas in which crimes happened to answer this question. This was also beneficial in minimizing the lie factor as we were able to use the raw location data (after scaling and shifting). One of the more interesting results/findings from this visualization is performing comparisons on a per location basis. This proved to be a good way to compare two disjoint sets, while also providing insight into the density of crime at a specific location.

Future work should focus on making the Tableau visualization more readable and should allow for side by side comparison of both cities. The visualization includes all days of the year; thus, the days are too tightly packed together making it difficult to read. Therefore, the holiday labels on lines overlap and this makes reading the labels and moving through days difficult. Finding a way to prevent having overlaps of labels could be another future work. Moreover, differentiating the holidays from non-holiday days also can make the visualization more readable, for example, making the lines corresponding to holidays wider or using a different color to represent them, is a possible solution to distinguishing between holidays and non-holiday days. To get more from the visualization, a percentage can be added over the lines corresponding to holidays. This percentage would represent the crime amount difference between a holiday and its previous holiday or the average crime amount over the whole year.

For the Bokeh visualization, we would want to improve upon its layout and some running time improvements. For runtime, we want to find a way to pre-load each dataset faster or access it in such a way that lag is minimized. For improvements upon the design itself, a way to have the lines in each visualization get out of each other's way is needed. As it is now, noise can obscure the lines and make it impossible to compare. Also, a way to detect which graph will have a smaller y-axis value over all, so the shorter one can be placed in from would be helpful.

One of the major hurdles in our project was time, and this meant that our final D3 visualization was left with loose ends and unimplemented features. In the future we would like to add more filters (such as date and time) and allow the user to select these as a range rather than a single value. We imagined using calendar to implement the date where the user would simply click and drag over the days of interest. And for the time we envisioned a similar selection process with a clock. The pop-up locational data that is presented when an individual point is selected currently doesn't have any labels on the axis; this is something that we would like to add. In addition to this when there is high crime density at a location the larger points can overwrite the smaller so adding transparency is something worth looking into to prevent this. One final change we would like to implement is point clustering (represent multiple points in a given radius as a single point with the number of subpoints encoded in its diameter). This would allow the user to have more data represented on the map without the added burden on the CPU.

9. HIGH-RESOLUTION IMAGES

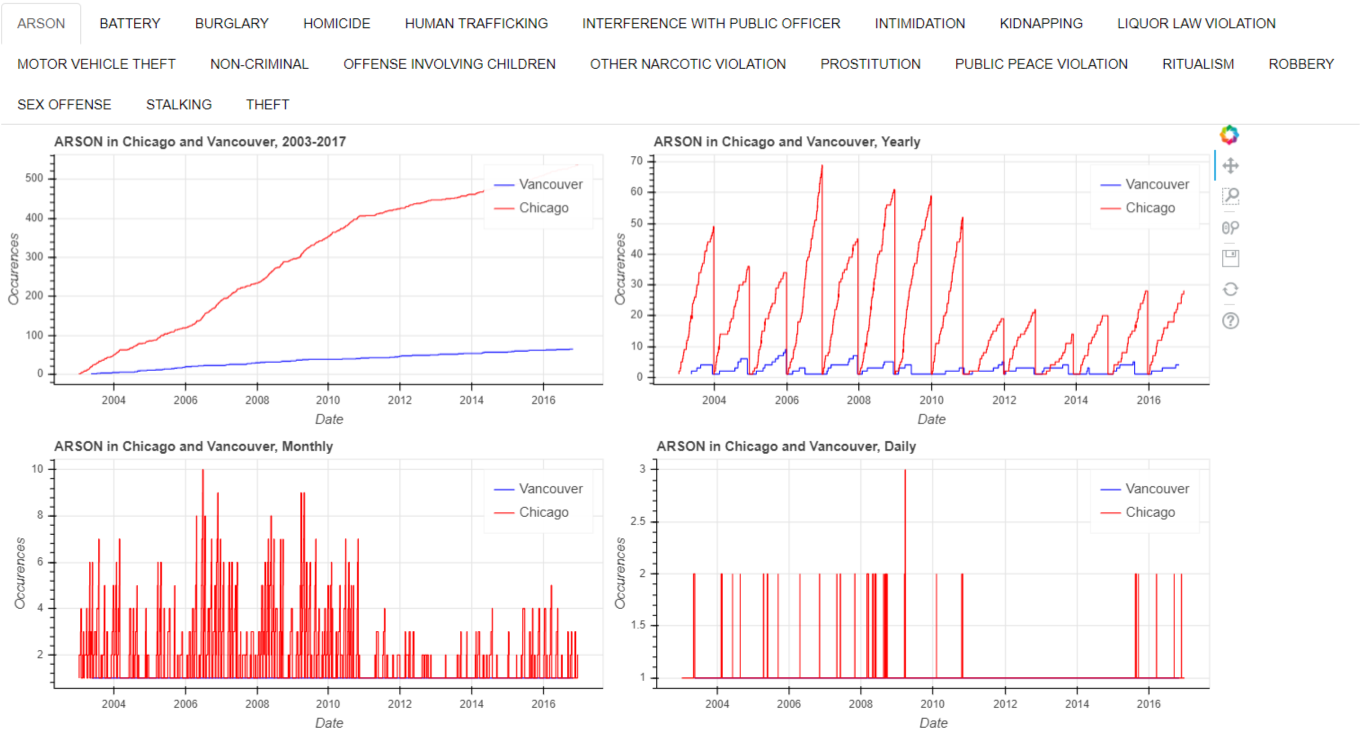
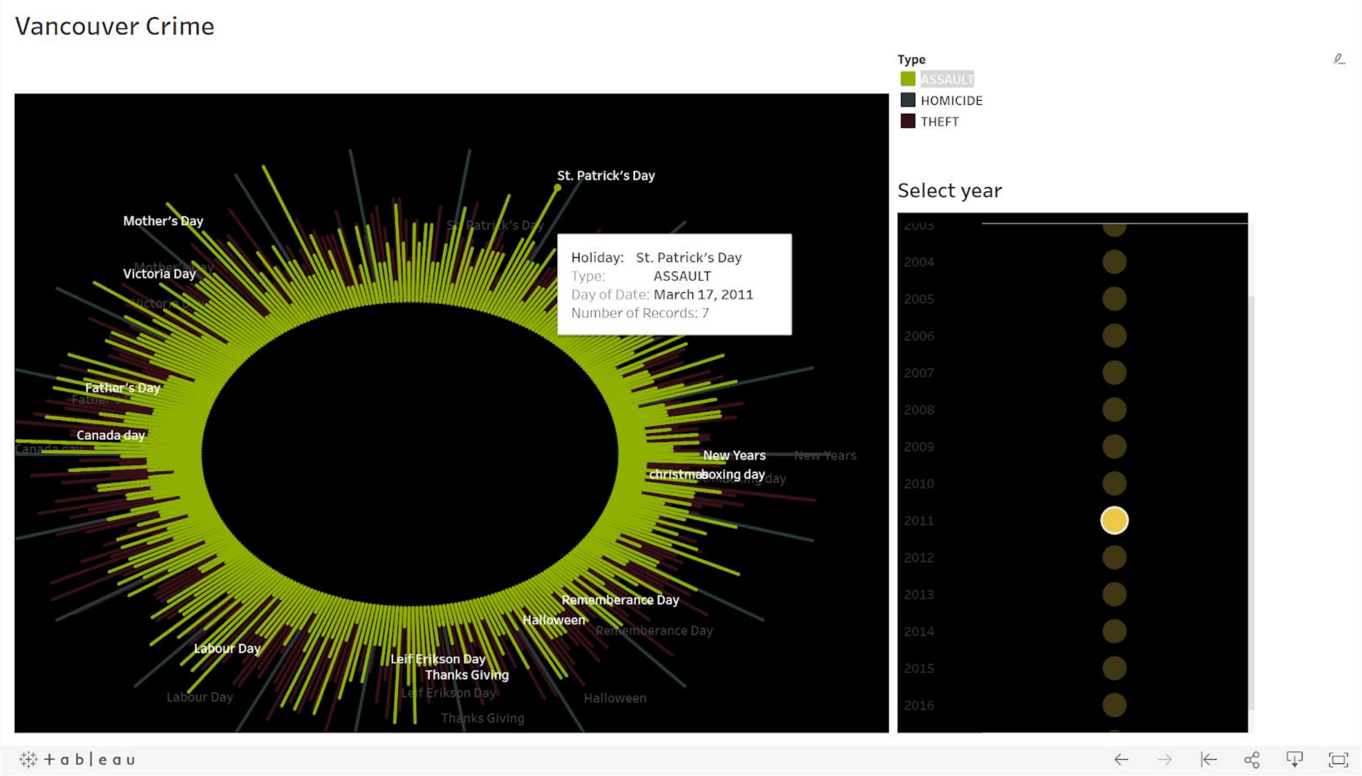


Fig 10.1. This figure Emphasizes raw data comparison and uses simple line graphs for quick comparison of the type of crime over the entire dataset (top left), each year (top right), each month (bottom left), and each day (bottom right). Red represents Chicago, blue represents Vancouver. Y-axis shows occurrences of selected crime type, and X-axis shows the date. Each crime type (top of figure) can be selected to change the current data type to the selected type. The toolbar next to the top right graph has selectable tools (top to bottom); pan, area zoom, scroll zoom, save image, reset, bokeh help.



Vancouver Crime

Type

- 🟩 ASSAULT
- ⬛ HOMICIDE
- 🟥 THEFT

Holiday: St. Patrick's Day
Type: ASSAULT
Day of Date: March 17, 2011
Number of Records: 7

Select year

Fig 10.2. Tableau visualization highlighting assault crimes in 2011. Assault is represented in Green, Homicide is represented in Cyan, and theft is represented in Red. The number of assaults corresponding to each day is represented using columns. The more assaults happened in that day the longer the column. Each holiday is labeled and hovering over each day shows a window containing further details (e.g. Holiday, Type, Date, Number of Records). In addition, the type of crime can be selected using the Type legend on the top right and the year can be selected using the 'Select Year' box on the bottom right.

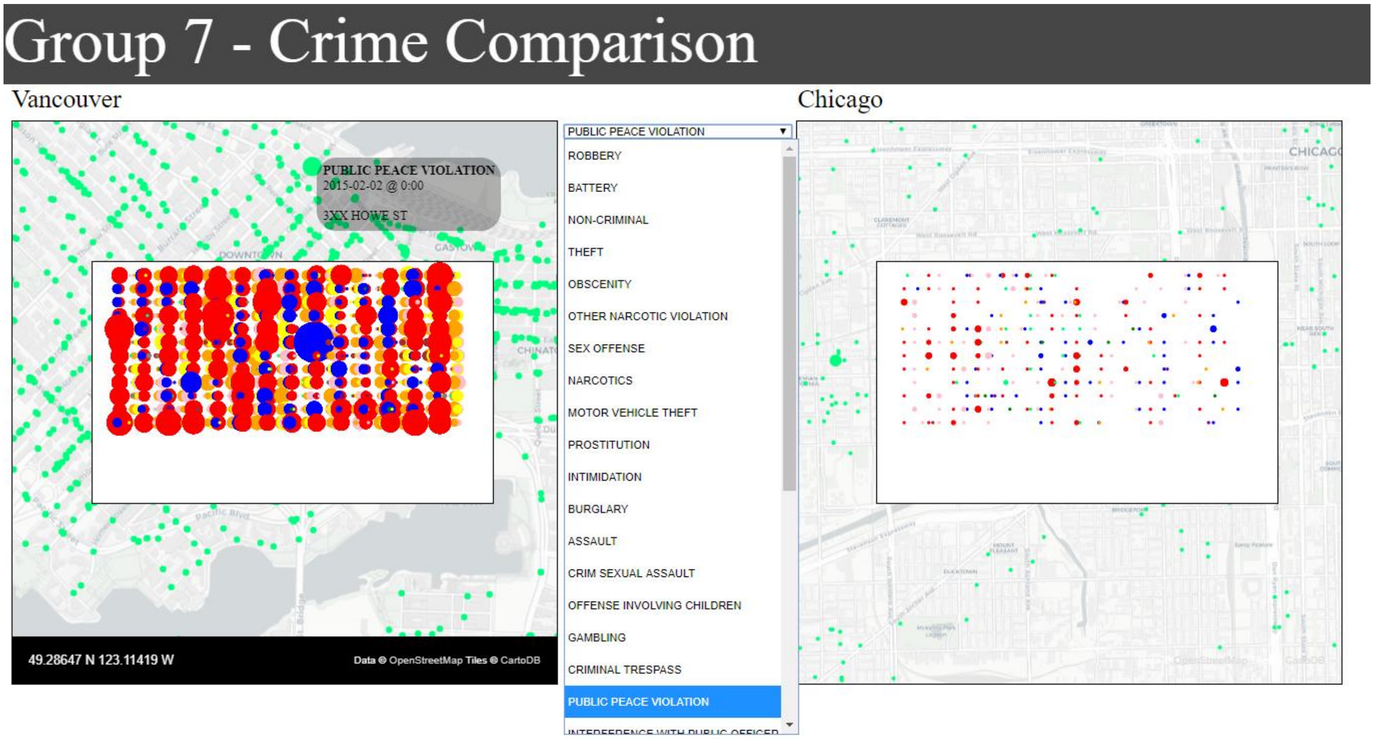


Fig 10.3. This shows Vancouver on the left and Chicago on the right. The user can select the crime they are interested in. The maps are then populated with color coded dots representing the crimes committed. If hovered over a pop-up is shown describing the details of the crime, and if clicked on, the crime history for that location (within 100m) is presented. This data is shown with years on the x-axis and months on the y-axis. The color of the dot represents the type and its size represents the amount of that type committed in that year and month.

REFERENCES

[1] J.-D. Fekete. The infovis toolkit. In IEEE Symposium on Information Visualization, pp. 167–174. IEEE, 2004.
[2] D. A. Keim, J. Schneidewind, and M. Sips. Circleview: a new approach for visualizing time-related multidimensional data sets. In Proceedings of the working conference on Advanced visual interfaces, pp. 179–182. ACM, 2004.
[3] T. Van Long and L. Linsen. Visualizing high density clusters in multidimensional data using optimized star coordinates. Computational Statistics, 26(4):655, 2011.