Visualizing Co-Occurring Factors in Homicide Using Tableau

Nghia Nguyen

Introduction / Background

Around 40% of homicides go unsolved, and it is estimated that there are currently 30 active serial killers in the United States.[1,2] Therefore, serial killers pose an active threat to public safety, and if they do commit murder, many of them will get away with it. However, serial killers do not operate completely randomly. Many follow an M.O., or a habit of working. Additionally, a serial killer may have a preferred weapon of choice or a preferred targeted demographic. Consequently, their preferences may manifest as recurring patterns in data. Therefore, we can visualize these co-occurring observations in homicides in a co-occurrence heatmap. These observational factors can be plotted along the axes of a heatmap (even multidimensionally). Each square in the heatmap would therefore represent a combination of factors and the color of each square would indicate how often it occurs in the data. With this visualization, we would be able to see a high frequency in which factors occur together, such as victim demographics and weapon of choice, and thus, we could observe a potential serial killer.

Audience / Purpose of Work

The best audience for this work would be people who are actively searching for unknown serial killers or people who are looking to connect unknown homicides with recorded serial killers. Therefore, this work would be most useful for law enforcement organizations, such as the FBI and local enforcement agencies, and investigators. The purpose of these visualizations is to help people better perceive recurring factors of a serial killer homicide, e.g. weapon, sex, race, so that these enforcement agencies, as they occur together. There are currently active serial killers, many of whom could try to avoid creating a repeating pattern, such as location. Consequently, local law enforcement agencies may not view treat a single, unknown murder as a widespread public threat. However, this could bring to light recurring patterns in homicides, which could then be used as sufficient reason for law enforcement to investigate. Additionally, since a large number of homicides are unsolved, this could help law enforcement connect unknown murders with existing cases, or it could help bring closure to a victim's loved ones.

Dataset

The dataset used in this work was downloaded from https://www.kaggle.com/murderaccountability/homicide-reports . It contains 638455 rows x 24 columns. As there are a large number of columns, I will describe them. The first column is the record number. A few of the columns are devoted to identifying the reporting agency, and personnel who reported the incident. Several columns are devoted to reporting the perpetrator and victim demographics, such as age, sex, and race. A few columns are dedicated to when and where the crime happened. Other columns include weapon, victim count, and relationship between victim and perpetrator, and record source.

There are some problems that are associated with this dataset. For example, some of the reported values are unknown. Sometimes a victim's sex may be unknown, or the weapon used may be unknown. Additionally, since these came from numerous data sources, there may be discrepancies between

reporting tactics. One of the things I noticed were that gun, handgun, and rifle, were different weapon categories in this dataset when they could just be interpreted as "gun". A good thing about this dataset, however, is that it is that it appears to be already cleaned. Another problem I also noticed with this dataset while working on it is that there is a non-random count (974) of people aged 999. I cannot find any information about it, even looking on the dataset's creator's page, so I am assuming that age 999 is just another way to represent "unknown" as a number. Also the victim and perpetrator names aren't displayed, perhaps for privacy purposes, so if the user is looking for a specific person, they would have to contact the people who compiled this database, which includes the Murder Accountability Project and the FBI.

Visualization Methods

For the first section of the visualization, I chose to display the filters I intended to use. These filters include a map of the United States, a slider for displaying a range of years, and check boxes for the relationship between the victim and perpetrator, age, race, sex, and weapon.

Figure 1.

I used the map as a filter so we could filter the location. Rather than having the map's click option as select by state, I wanted the user to be able to specifically select which ever area they wanted to select, such as by clicking on a certain dot or by highlighting certain areas using the select tools provided in the map. Additionally, I added the year slider so the users can filter it by a range of dates. The relationship checkbox is specifically added so I could filter "strangers" and "unknown" relationships from the dataset, and this dimension should be read as victim to perpetrator, so "boyfriend" means the victim is the boyfriend to the perpetrator. Additionally, I added the victim sex, age, race, and weapon used as a way to

filter out the dimensions of the heatmap. Also having a filter for age addressing the problem of the age 999, mentioned in the Dataset section, which can lead to skew in the histogram.

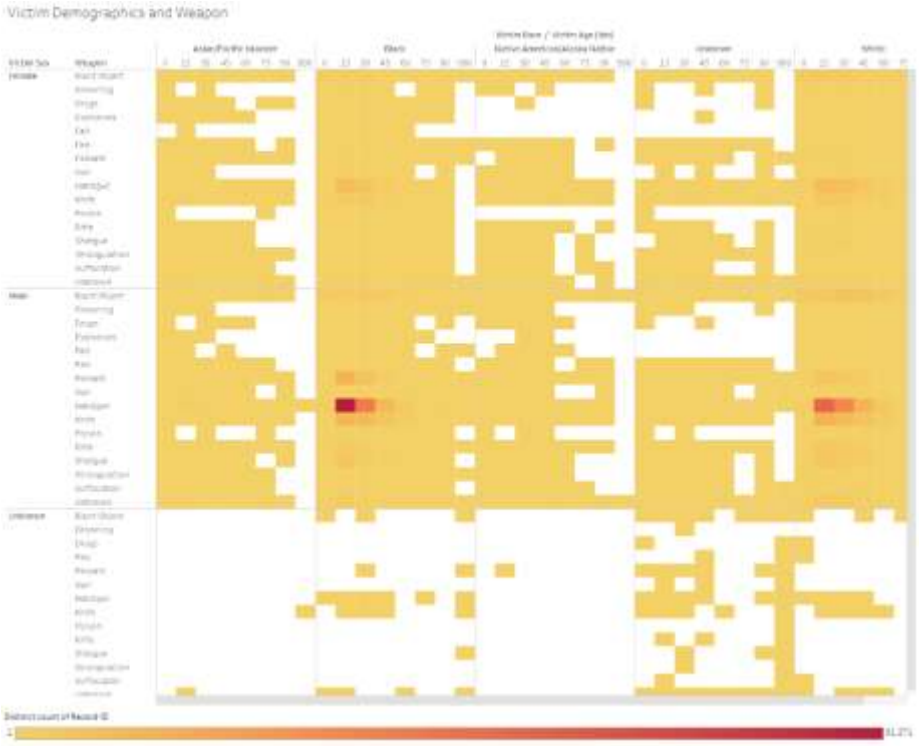


Figure 2.

The main visualization is the heatmap depicted in figure 2. The heatmap is made with 4 dimensions that allows the user to see what the most common occurrence amongst those 4 dimensions is. I chose 4 dimensions since it is the most that this heatmap can allow for without being unreadable. I chose race, sex, age, and weapon as my dimensions since these made sense for grouping together victims, and also, these dimensions have small enough categories that do not make it too hard to view the heatmap. As you can see, the most common across the whole data set is black male deaths, aged 15-29, due to a handgun. The heatmap also makes it easy to see this occurrence. Additionally, I used this heatmap as a filter so the user is able to use the heatmap filter the location distribution in figure 1 and the frequency distribution in figure 4. Furthermore, this heatmap can be replaced by a specialized chord diagram. See, figure 3, from: http://www.jamesmiller.com/Daisy2011/daisychart.html . This was my initial inspiration. However, the visualization in the link requires tableau knowledge that I do not know, knowledge that would be difficult to search up, and would probably run very slowly if I tried to code it myself.

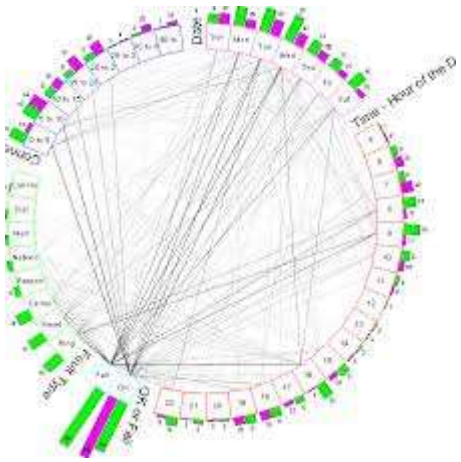


Figure 3.

In figure 3, all possible lines are drawn between each sub-category, and each line between the sub-categories gets thicker as the number of co-occurrences between the two increases. The dimensions could be put along the circumference, with histograms and bar charts above each dimension as depicted in figure 3.

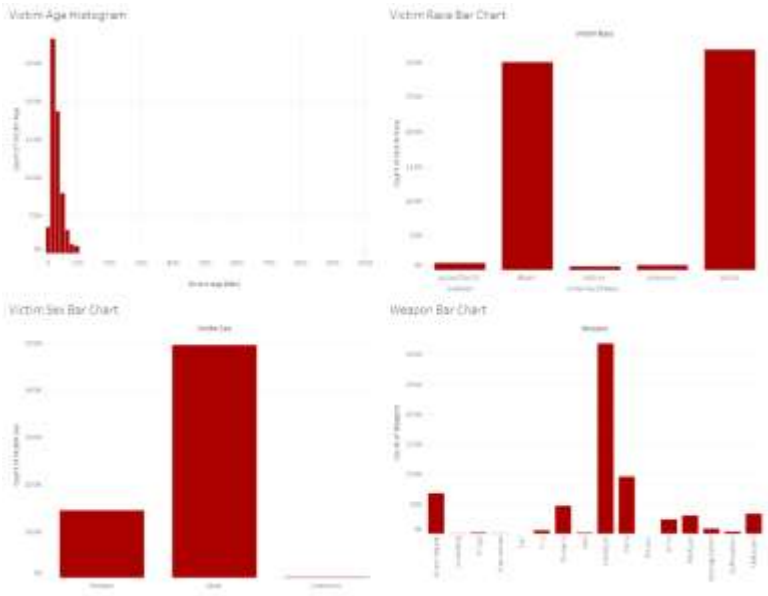


Figure 4.

The last part of this visualization is the histogram and bar charts in figure 4. Similar to figure 3, I took the bar charts from figure 3 and separated them out to its own section, which turned out to be more useful since it is easier to read. This allows the user to be able to see how frequent categories of these dimensions are occurring; thus, it helps them understand whether the co-occurrences in the heatmap are out of the ordinary or due to some other underlying factor. For example, we could see that black males, aged 15-29, murdered by handguns occur more compared to white males of similar demographics, yet there is a lower population of black victims compared to white victims. This alludes to the fact that young black men are more likely to die from handguns, since there is not a higher population of black males to account for the increased frequency.

I used Tableau since it was a GUI, making these visualizations relatively simple to construct as opposed to coding. Furthermore, I thought that the online Tableau community was helpful for me to create this. I found help from researching Tableau "shopping cart" visualization. Also, the software was able to process 600,000+ data points in a 4D matrix quickly, which was useful for the heatmap. Additionally, Tableau allowed for me to easily publish the workbook online.

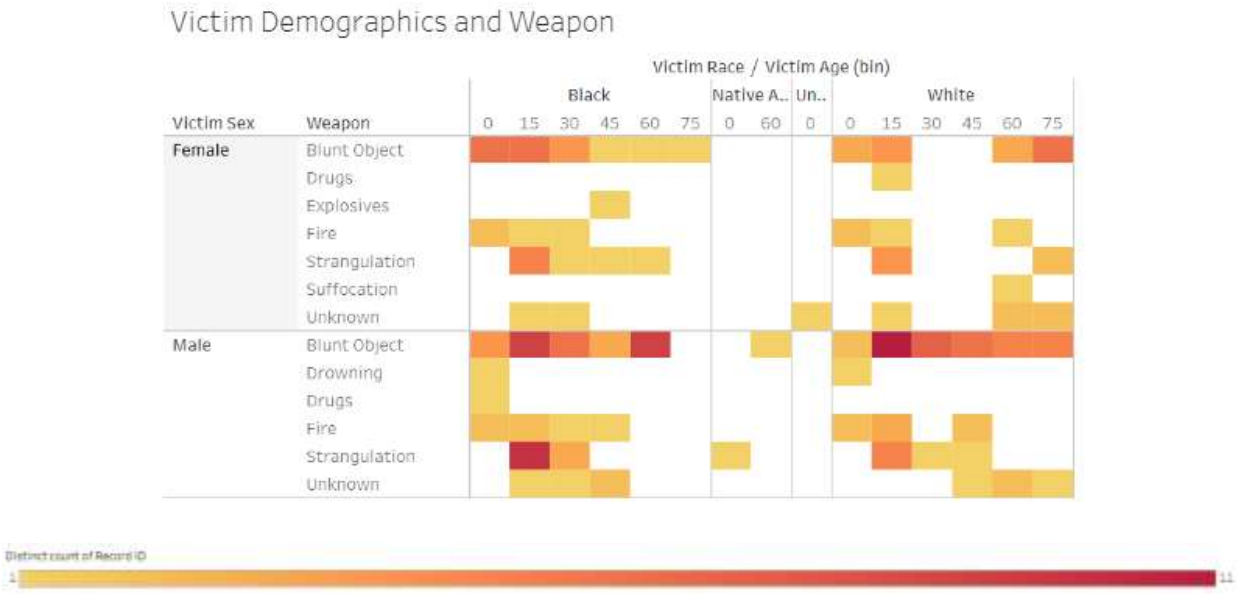
Results

The visualization will be available at this link:
https://public.tableau.com/profile/nghia.nguyen#!/vizhome/Final-HomicideData-Workbook/Dashboard1

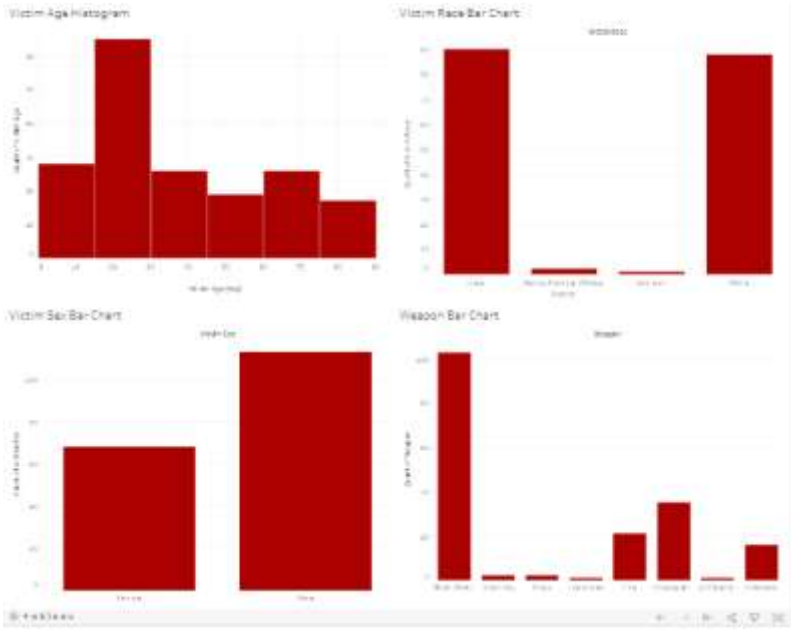
I noticed that this visualization does not work well if one does not have an idea that they are searching for. Therefore, I will demonstrate by example. Rather than trying to find an unknown serial killer, I will demonstrate this visualization with a known serial killer. Say we are in the Milwaukee police department in the early 1990s and we are noticing a certain recurring pattern in non-gun related deaths. We will be able to set certain parameters to our advantage. So, as the user, we could select Milwaukee as our location and set the date range from 1980-1991:

Next, we believe that these victims do not have a strong connection to their murderers. So we will check strangers and unknown. Additionally, we will check "acquaintances". This is since this murderer is already found and probably would have been filed under strangers and unknown at the time. And since gun and knife deaths occur fairly regularly and randomly (i.e. without connection to a calculated serial killer), we will filter them out too:

The heat map comes out like this:

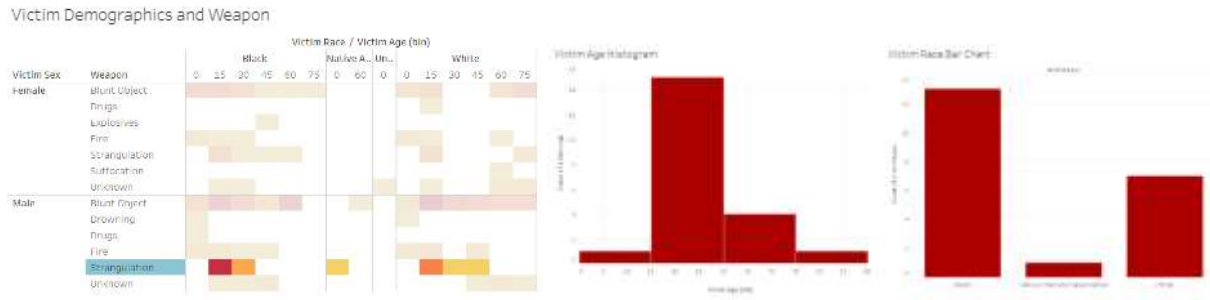


Consequently, we are able to see some certain patterns occurring, between different ages, sex, races, and the weapon used. For example: young black females, blunt object; 15-29 year old black males, blunt object; 60-74 year black males, blunt object, 15-29 year old white males, blunt object, and 15-29 year old black males, strangulation. You could also hover over the boxes to learn how many times these occurred, under "Distinct count of Record ID". The Milwaukee police would be able to see these as different problems to be address. The one that we will focus on specifically is 15-29 year old black males, strangulation.



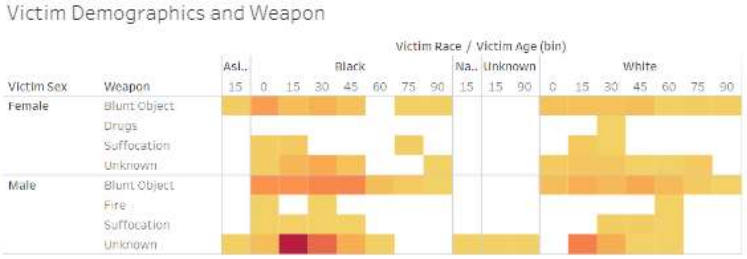
Here are the bar charts (above) for the parameters set before. As we can see, it is not irregular that the murders that occurred amongst 15-29 year old black males who died by strangulation could be accounted for by the heighten frequency of deaths amongst 15-29 year olds and males. However, we

could see that the amount of black and white men is near equal, so why would these strangulations occur more amongst black men? Additionally, the amount of strangulations appears to occur in about 1/3 of deaths compared to those of blunt objects, yet the heatmap shows the shade of the box as on par with the boxes in the heatmap associated with death by blunt object. We could be looking at a potential serial killer.

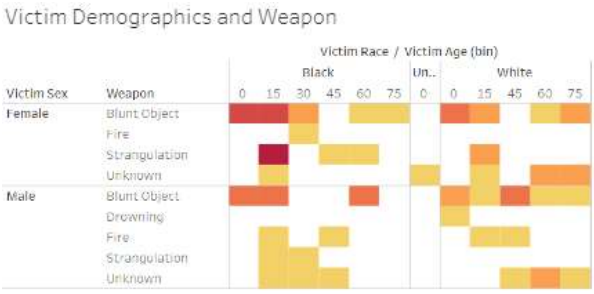


Additionally, we could use the heatmap to filter male strangulation, as shown above, and it shows that these occur more disproportionally in black 15-29 year olds as compared to the graphs shown before without filtering male strangulation.

The serial killer that we are dealing with is Jeffrey Dahmer, which commited crimes around 1978-1991 in Milwaukee. Should the police station had all of this data and this program available to them, perhaps they may have been able to investigate and catch Dahmer sooner. From this data, we could expect that this serial killer *only* kills young black men. However, the added bonus of this program is that the user may be able to investigate further and connect these strangulations to those that also occurred in white men of the same age group, since Dahmer had both white and black victims. On a further note, here is the heatmap after Dahmer was arrested (after 1991):



As we can see, there are no strangulations. However, I also find it weird how there are no strangulations amongst both men and women during this time. I reset the years to 1980-1992 and checked only unknown murderers and found that there were also a large number of young, black female strangulations, and I wonder if they are connected to Dahmer in any way, even though his targeted demographic is known as men. See below:



Along with the link, https://public.tableau.com/profile/nghia.nguyen#!/vizhome/Final-HomicideData-Workbook/Dashboard1, I will be submitting a zip file of the data and workbook.

Discussion

In this work, I have created a visualization in order to help law enforcement and investigators to identify serial killers as well as try to connect known homicides with other unknown homicides. The visualization allows the user to filter location, year, relationship of victim to perpetrator, race, sex, age, and weapon used. The user would be able to visually discern the frequency of how these factors occur together. Additionally, the user would be able to further analyze the data with provided bar charts.

There are some future features that I want for this visualization. For example, one of the shortcomings of this method is that it does not allow for many dimensions to be viewed in the heatmap. Therefore being able to create the daisy chart chord diagram, mentioned in Visualization Methods under figure 3., would allow us to see maybe 5 or 6 dimensions. Another feature that I wanted for this visualization is a way to display the record numbers of the victims selected using a filter. This would allow the user to identify certain record numbers of victims to research further. However, I tried to do this, but Tableau would not allow it due to the large amount of data. Another limitation is the ability to pick different and more or less dimensions and measures. For instance, it would be helpful to add different dimensions and measures from the dataset and add a menu that allows the user to change which dimension to view on the heatmap. Additionally, I might want to reduce the number of dimensions I displayed in the heat map, such as combining male, female, and unknown into one category. This visualization also had the limitation that it was difficult in highlighting potential serial killers from a broad scope. As shown in my Results section, I had to pick a specific city and specific time frame and rule out certain relationships and weapons. I would want a better visualization or algorithm that even when looking at the entire map of the United States, would highlight certain homicides that are occur regularly enough to be attributed to a serial killer.

Works Cited

1. https://nypost.com/2018/09/25/a-shocking-number-of-us-murders-went-unsolved-last-year/
2. https://www.bustle.com/articles/112070-how-many-active-serial-killers-are-there-right-now

Dataset. https://www.kaggle.com/murderaccountability/homicide-reports

Figure 3. http://www.jamesmiller.com/Daisy2011/daisychart.html