Exploratory Data Analysis & Visualization

Homicides Report 1980 - 2014 Dataset Interpretation

Yiyu Tao
Indiana University Bloomington
I422 Data Visualization

***Abstract*— Along with the advancement of technology in contemporary society, data analytics and visualization techniques are gradually adopted in criminology field. This project will present a simple version of technical process of analyzing homicide report data, focus on exploring possible meaningful combinations between two or multiple factors and perform an exhaustive exploratory data analysis on finding their hidden relationships. Then convert the numerical information into graphical representation for the purpose of achieving more efficient comparison and test the possibility of elevating the work to the next stage by applying basic machine learning classification methods.**

I. Introduction

A. Motivation

There are two major motivations that have contributed to the birth and topic choosing of this project. The first reason is the wide implementation of Big Data analytics in many aspects of people's lives including but not limited to business, sports, social science and science industry. It uses an impressive speed to enter the market and take control of huge role. People can even say nowadays, they're relying and depending on data and making almost all the decisions based on the results of data analysis. 53% of companies are using big data analytics today, up from 17% in 2015 with Telecom and Financial Services industries fueling the fastest adoption(Columbus 2017). Fifty-three percent means it already guarantees the dominant position, and the number

will only keep growing faster and faster. Not just business, the very similar situation is happening in sports. Ever since sabermetrics, statistical analysis of baseball game data, becomes a powerful tool than regular coaching and shows significant effect. Other types of sports like basketball, swimming, tennis are all adopting data analytics techniques to their game. 2016–2022 sports analytics forecasts estimate a massive 40.1% CAGR potentially reaching a value of USD 3.97 bn in 2022(Athithya 2019). Such increment rate really shows it's the necessary component in the future sports game.

Another motivating factor for this project is a remarkable series of achievements data analytics and visualization tools have accomplished in the field of criminology. In recent years they have been active on crime prevention, crime prediction, predictive policing and other works which are associated with social issues. The massive amount of data stored in the police database becomes the star of the show. IBM has been working with the police department of Manchester, New Hampshire, to combat crime ahead of time using IBM's SPSS Modeler software(IBM 2017). This is just one of the examples of countless crime analysis tools, each of them has their helpful features used to handle different cases. The most important thing is with the help of these softwares, the crime rate has significant reduction. Manchester, for instance, reported reductions of 12% in robberies, 21% in burglaries, and 32% in theft from vehicles, following the adoption of recommendations for preventive action from statistical analysis(IBM 2017). The number proves the promising future crime analysis has and encourages more data experts participate in further development.

B. Related Work

Among all the existing work and projects which are associated with crime data analysis and utilizing sophisticated visualization methods, three of them are picked due to its significance and reference values towards this project.

The first one is a records management system application which is used to conduct crime data analysis, developed by Sun Ridge System INC, a public safety technology company. It

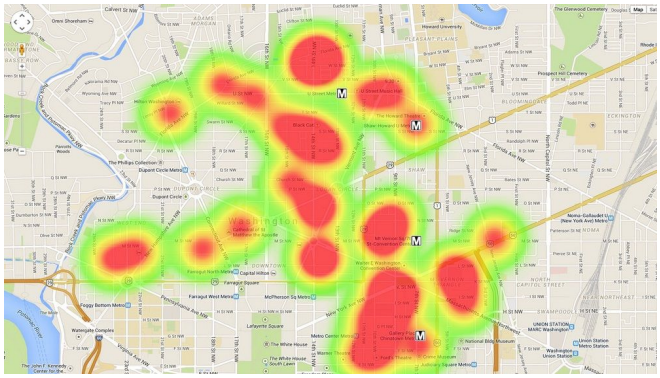


Fig 1. RIMS Crime Analysis

merges the powerful data mining features of our integrated applications with Google Maps and ESRI to visualize your research(SunRidgeSystem). The reason of choosing this as the representative existing work is the impressiveness and richness of its functions. This application is pretty mature on organizing and mapping out the crime data. It also provides customized service like changing the size of areas depends on needs, using different symbols to pin the information on map display and make the data clustered. The other advantage it has is easily sharing and transferring the data or research results to other systems that can avoid inconvenience in practical work. This is no doubtly an ideal tool for crime data analysis.

The second related work shares more similarity with this project on technical aspects. It is a data analysis project which uses chicago crime data during 2001 to 2017. It uses different data analysis and visualization techniques to express the data, including summary statistics with sql, heatmaps, line charts, bar graphs and pie charts. One of the great things about it is the creator sets a list of questions before starting doing actual work and each graph is designed to answer one specific question. This will improve the efficiency and accuracy of his work. But at the same time its shortcoming is very obvious, the failure of choosing the most suitable plot has seriously

influenced the viewer's perception and understanding towards the data. For example,

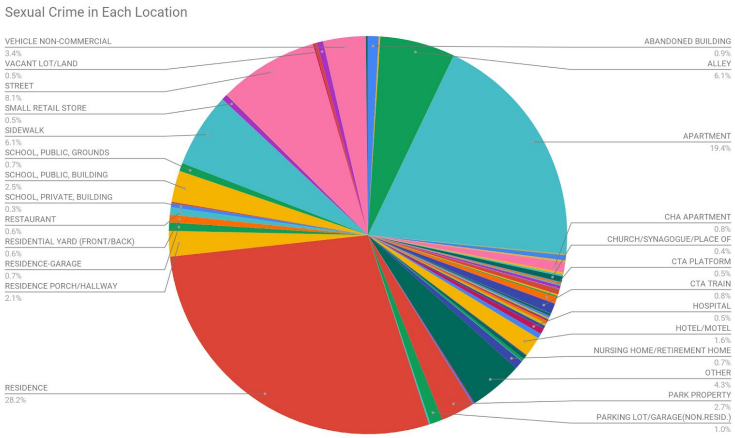


Fig 2. Sexual Crime in Each Location

the above plot completely shows the weakness of pie chart, it's losing the focus and becoming messy with so many sections, the annotations at the side don't really make that much difference either. Its fault definitely reminds of the importance of picking visualization methods wisely.

Lastly, the third work that has been chosen is an example of creating map based on dataset by using Folium. Folium is a powerful Python library that helps you create several types of Leaflet maps. The fact that the Folium results are interactive makes this library very useful for dashboard building(Roos 2015). The tricky thing about using this library is its strict requirement of quality of the dataset. Coordinate data is the foundation of crafting the interactive map which will be able to present the exact street address information. Unfortunately, the homicide report dataset doesn't contain any spatial data. Having such serious limitation is kind of one of the reasons this Folium being mentioned in this section. If there is a chance to add geometric data to the dataset, this technique will be an affirmative method to utilize. Another outstanding feature Folium has is its capability to filter data by adjusting the parameters. This is a quite helpful function because it will reduce a lot of work while extracting attributes information.

C. Contribution

The dataset will be used in this project is called "Homicides Report 1980-2014" and found on Kaggle. It contains approximately 63.8w records of homicides in the United States spans from 1980 to 2014. There are 24 columns in the file which are Record ID; Agency Code; Agency Name; Agency Type; City; State; Year; Month; Incident; Crime Type; Crime Solved;

Victim Sex; Victim Age; Victim Race; Victim Ethnicity; Perpetrator Sex; Perpetrator Age; Perpetrator Race; Perpetrator Ethnicity; Relationship; Weapon; Victim Count; Perpetrator Count and Record Source. Both of the geographical and demographical data of murders are included in this dataset, also it breaks down the data into the details of victim, perpetrator, relationship type and weapon type which are all useful and able to be applied cross analysis to discover certain properties like patterns, distribution and frequencies. This project will also compare different class of data to each other to get the sense of proving or disproving some social facts. Another objective of this project is investigating those special cases in terms of murder, extacting a group of data which shares same speical attributes and taking an in-depth look of it. Maybe it will reveal some new insights and reflections which are easy to neglect in the whole. For instance, serial killer, juvenile criminals and marriage issues. These are all the worthwhile things to dig into. If certain connections within the data can be found, this will push the project a step closer to generating valuable knowledge. In addition to do data cleansing and create visualizations, in the last part of the project a basic machine learning approach will be implemented to simulate the advanced operations of other predictive policing tools. The purpose of doing this is to test whether this dataset has the potential to incorporate new data and achieve higher level goals.

II. Process

A. Visualization Design Ideas

The mind map, which is shown below, contains the entire workflow of the visualization design of this project. It was completed as a guide before the actual plotting procedure started, all of the possible meaningful combinations were listed on it and classified by their category.

The first child node is one dimensional plot that illustrates some basic information, including both of the geography and time series data, the total amount of incidents will be broken down into each year, month, state and city. Also the plot that shows the count of every type of relationship except acquaintance, stranger and unknown which are sort of meaningless, will be generated for insights seeking purpose.

Since this dataset contains plenty of categorical data, the second child node will be the meaningful combinations of two of them. So this section will be filled with two dimensional

plots, there are twelve different combinations which means a total of twelve plots will be made to find the hidden meaning within them. The actual content includes the distribution of crime solving status, weapon choice of different relationship, race and age (both perpetrator and victim) by contrast, relationship type subdivision based on race, age (both perpetrator and victim) and crime solving status, lastly, the comparison between personal information like gender, age and race of perpetrator and victim.

Multidimensional plot may also contribute to generating new insights because of the two particular columns in this datasets, relationship and weapon. Each of the variable covers a variety of more than ten unique values. There is a possibility of getting new discovery by plotting these two with time series.

The next one is the map, the reason of having a map implementation is the advantage on completeness of the geographic information for this homicide report dataset. Also map gives the most direct sense of frequency comparison, therefore, the visualization will be valuable and aesthetic at the same time.

The last section of the visualization part of this project will focus on the special cases. In this project juvenile criminals and serial killers will be mentioned and analyzed. For criminals who are under 18, a two dimensional plot will be made based on their weapon usage and relationship with victims data. The reason for creating such a visualization is because the two most important questions for minor murderers should be who do they kill and how do they achieve that? On the other hand, serial killers, which have more investigation space to bring up hypothesis than previous case due to its complexity. Therefore, more visualization should be plotted for serial killers, including one dimensional plot of time, geography and personal information, also two dimensional plots like comparison between perpetrator and victim age, relation between weapon and relationship are optional graphs that are worth considering.

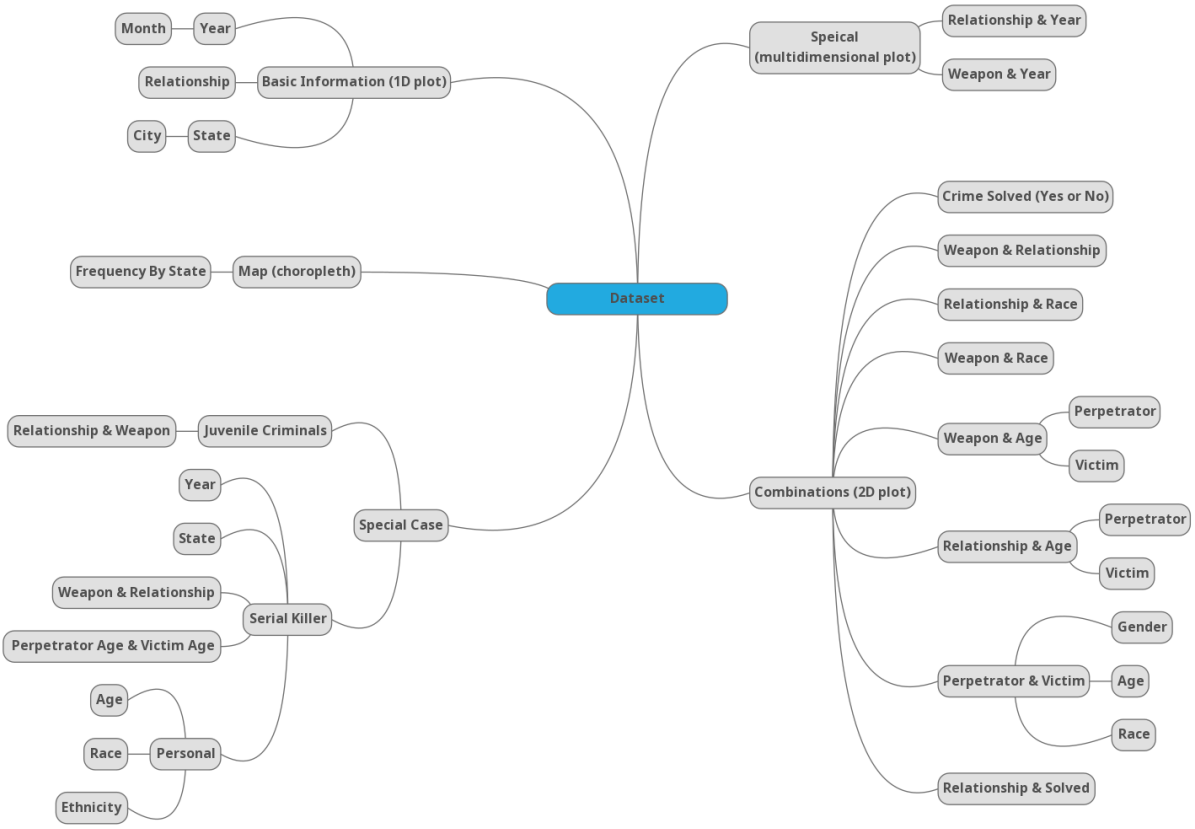


Fig 3. Visualization Design Mind Map

B. Methods Selection

All the data analysis and visualization will be accomplished by using Python and its libraries. The specific graph selection will be introduced by the order of its dimension.

Firstly, for one dimensional plot, obviously line chart is the perfect choice to plot time series data because they are excellent on demonstrating the trends of value changes. It has its slight disadvantages for easily losing details but this doesn't really matter in this case. And for state, city, relationship these categorical data, the ideal way to show the frequency distribution of discrete data that belongs to different categories is bar chart. It's easy to manipulate and great at presenting difference between kinds. If the data is sorted before plotting, bar chart will also offer direct feeling of order to the viewers.

Two dimensional plots have more options on choosing methods than one dimensional plots. Because for this section two of all types of data will be picked and perform comparison, everytime the situation will change based on the exact type of the data it used. For crime solving

status, pie chart is clearly the one because yes or no is a boolean condition with only two classes, also its ability to show the proportion of each class is very useful in this project. Next scenario is when both data are categorical data, this is the point that heatmap will have its advantages. The way of displaying non-numeric data heatmap uses is counting its frequency and using color-coding system to distinguish them. Its number annotation function can help viewer with getting deeper understanding of the data. But heatmap has several shortcomings that can even ruin the entire visualization. It's not applicable to data which has too many unique values because the graph will become crowded and messy. If there is a counted number way larger than others, heatmap will be meaningless because except that darkest matrix the rest won't have much difference. To conclude that, the creator must be careful with the data he uses to create the graph. The other scenario which is in contrast to the last one is the combination is formed by one categorical data and one numerical data. Seaborn's bar chart is a great pick here because it will automatically calculate the mean of numeric data, and this will provide the viewer the chance to exhibit new insights. It has the same issue on being influenced by poor data quality. For example, if there are only ten deaths caused by fire, the final result of average number will be inaccurate with the existence of even one unusual value. This will cause distortion to the visualization. Last scenario is when both of data is numerical data, at this point jointplot in seaborn will be in charge of making visualization. It shows the distribution of data in both axes well and is able to adjustment the plot kind depends on the needs. Its flexibility makes it an outstanding method of achieving the goal of giving graphic representation and doing comparison at the same time.

Next part is the multidimensional plot, radar plot is the one used in this project to plot the breakdown of weapon and relationship by year. The python library helps with making the visualization is Plotly. It produces interactive graph which allows the viewer to unselect any year he wants. The main issue in the process of using it is the unstable positions of each node, it does bother when the plot changes all the time along with running the codes.

The last plot is map, in this project Plotly choropleth map is used to create a statewide map that has color depth representing the frequencies and hover appearing while moving the mouse above it. Its has the drawback on ununified areas as a choropleth map. But in this project

the purpose of using it is just comparing the frequency of crime occurrence in each state, the issue doesn't really affect the result.

III. Results and Insights

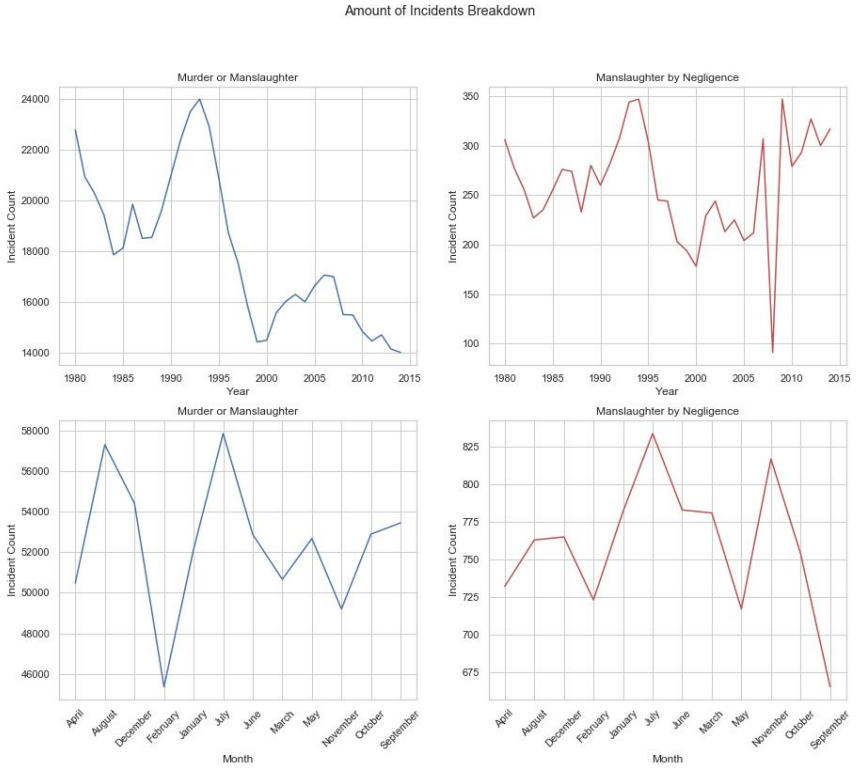


Fig 4. Incidents Amount Change by Time

This is the line chart which displays the change of incidents amount along with time, the interesting point of this is the time range from 1985 to 1995 and after 2000. After searching online, these ten years were somehow a chaotic period in American history, under the influence of the presidency of Lyndon B. Johanson, Vietnam war and drug issue outbreak the crime rate had increased a lot. Another notice is after 2000 the average of overall count is much lower than before, it shows at 21th century society has entered the age of peace.

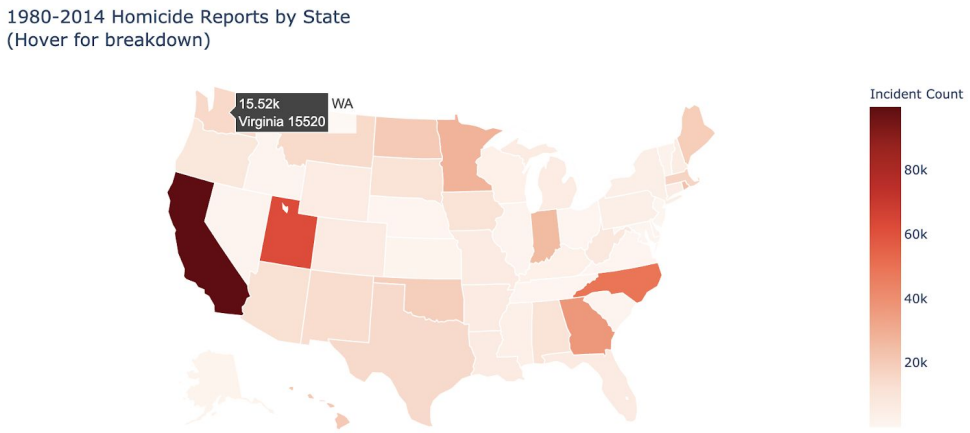


Fig 5. U.S. Homicides Choropleth Map by State

The result is pretty obvious, states like Califionia, Texas, Florida and New York has darker color than others, this basically obeys the common sense. Also while moving the mouse the viewer can check the specific number.

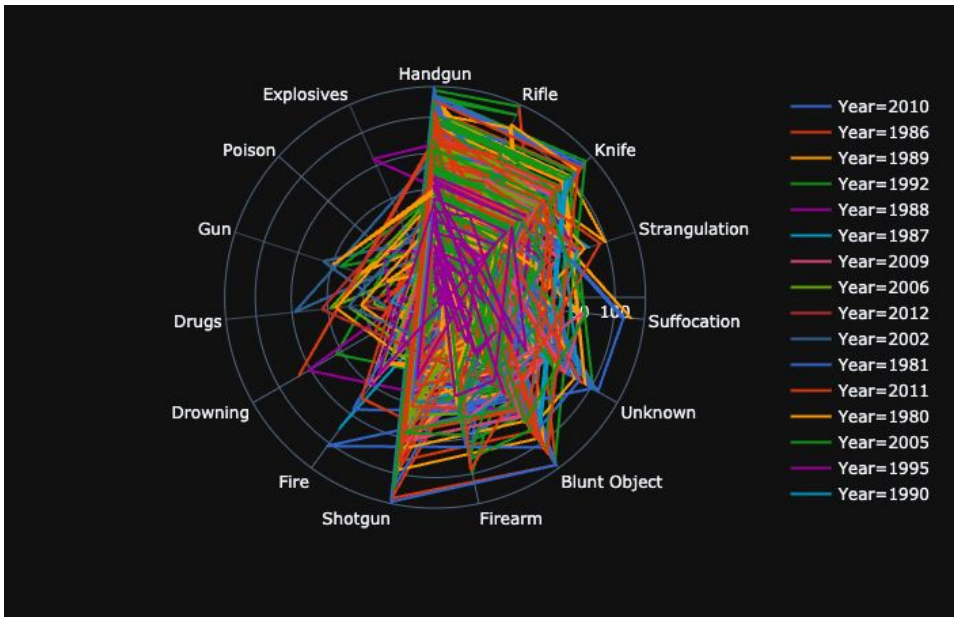


Fig 6. Radar Chart of Weapon Choice by Year

The viewer can see distinct differences between left and right side, around handgun, rifle, knife it has the most crowded area that can prove criminals does take advantage of gun free law in this country. Since the graph is interactive, when only 1985-1995 ten years of data is kept, they viewer will see much more points falling on handgun line which is also fit the research. Handgun-related homicides more than doubled between 1985 and 1990(Ford 2016)

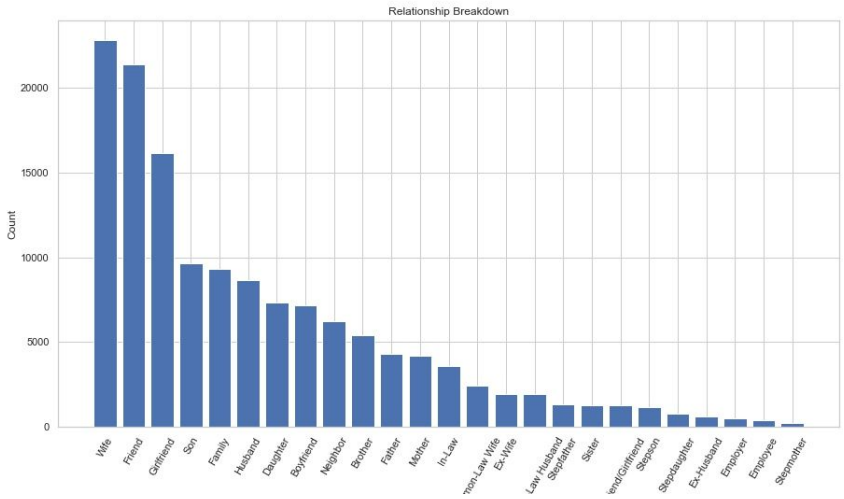
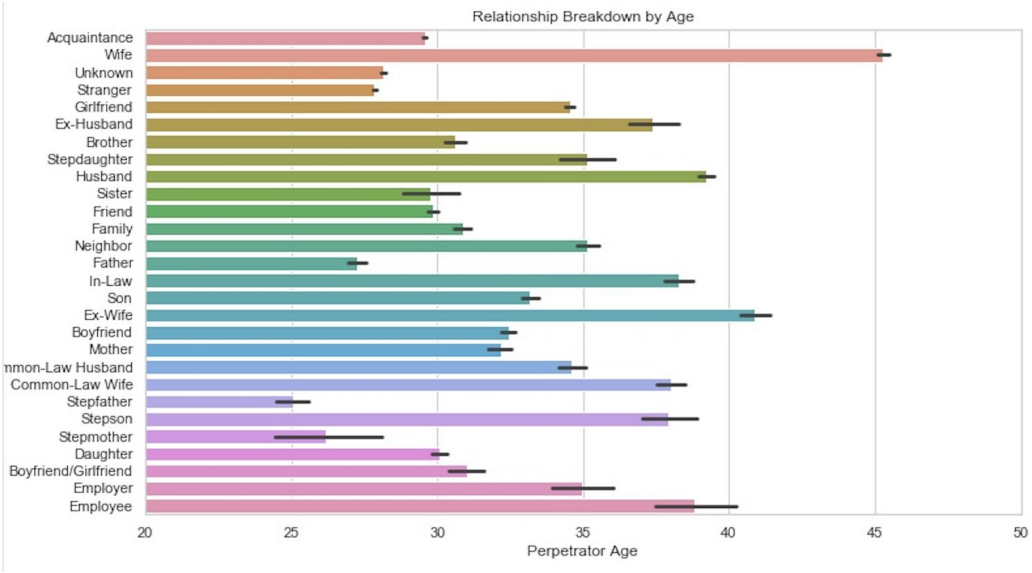


Fig 7. Relationship Breakdown

Other than acquaintance, stranger and unknown, the top three of relationship appearance frequency are wife, friend and girlfriend. Two of these are related to romantic relationship, confirms the fact that sometimes conflict between lovers can lead to a tragedy. Other than that, lots of family members are placed at the front position which means family issue is also very serious factor that triggers homicidal impulse.



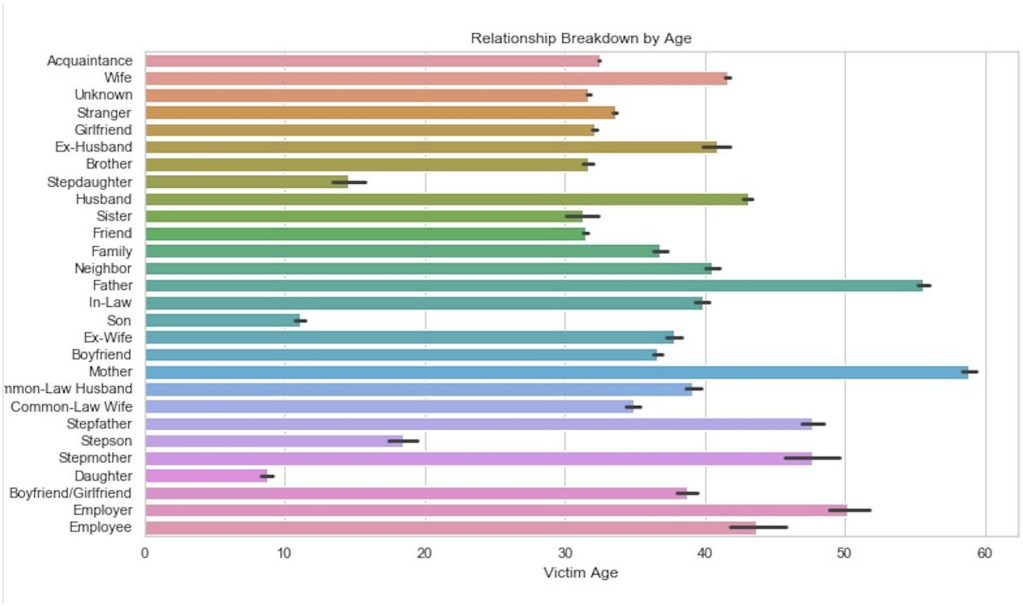


Fig 8, 9. Continued On Relationship

These two graphs have some very interesting results, in the first one, the average age of four labels - Husband, Wife, Ex-Husband, Ex-Wife is all late 30s or 40s. This shows most of the murders were happening after decades of marriage. When couples were sick of endless fights and arguments, some of them chose to kill the person they used to love. If they still made the same decision after they got divorced. That must mean the grievances has already been accumulated over the years and separation can't erase it.

If viewers look at the second graph, they will find out the average age for killing between parents and children is equally shocking, the results say parents victims are killed in their 50s and children victims are killed in their 10s. This means lots of well-educated and job owning adults would kill their parents at their mildlife, a time that they are about to start enjoying their retirement. And parents would kill their children who may just go to middle school and have fun in the childhood. This is just so sad.

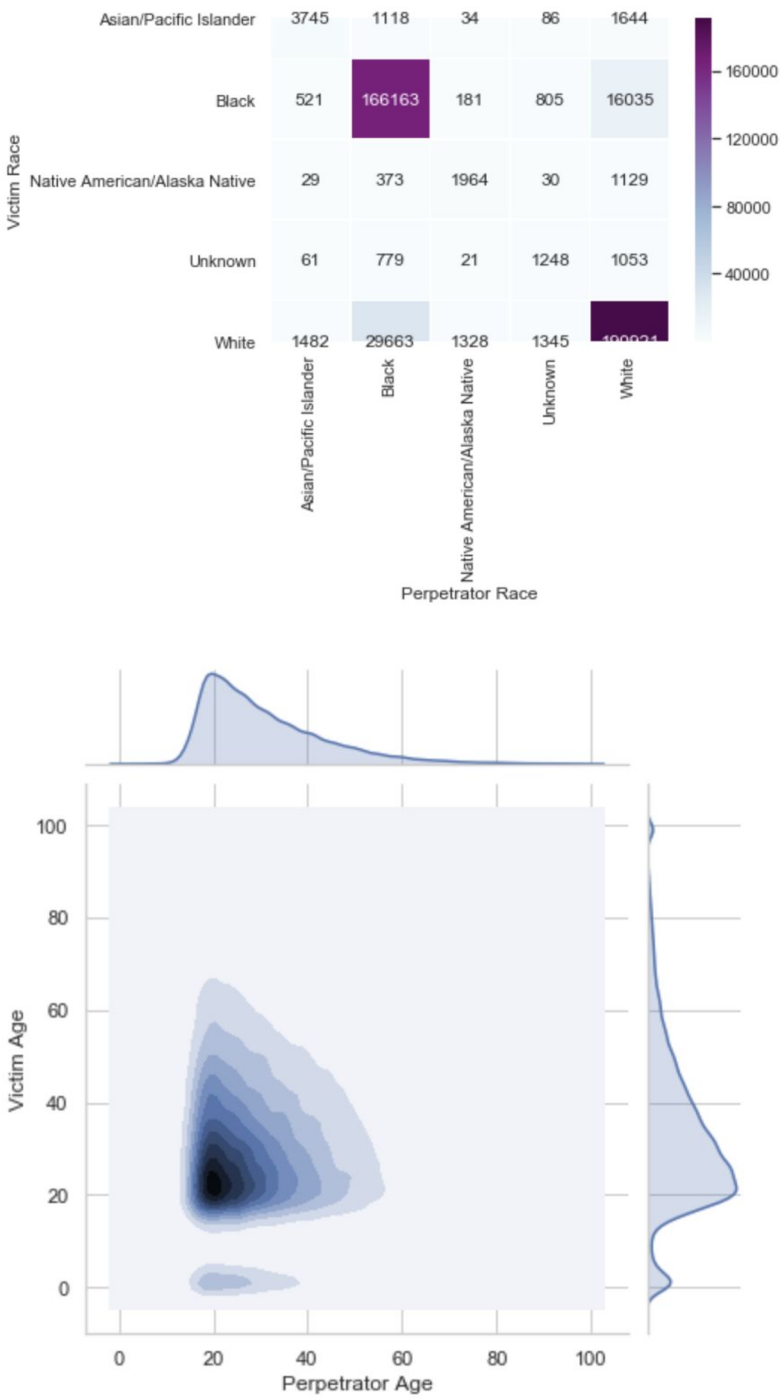


Fig 10, 11. Age and Race

From this kde jointplot of perpetrator and victim age, the viewer can tell the majority of the criminals are twenty to thirty years old, this shows the younger generation is easily infuriated by other people or lured by money or other stuff, and they don't really think through the

consequences before actually doing it. Also the fact that victim age range is wider than perpetrator age range proves the cruelty of people is just beyond the imagination.

About the contrast of perpetrator and victim race, even though most of the cases they are still have the same race, the black-white conflict is still obvious compared to other races, the possibility of racism issues being leading factor of murders still exists.

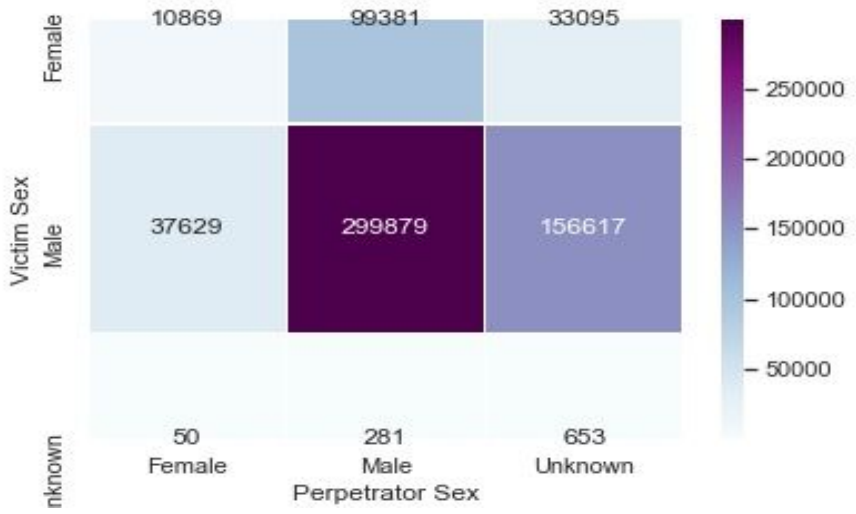
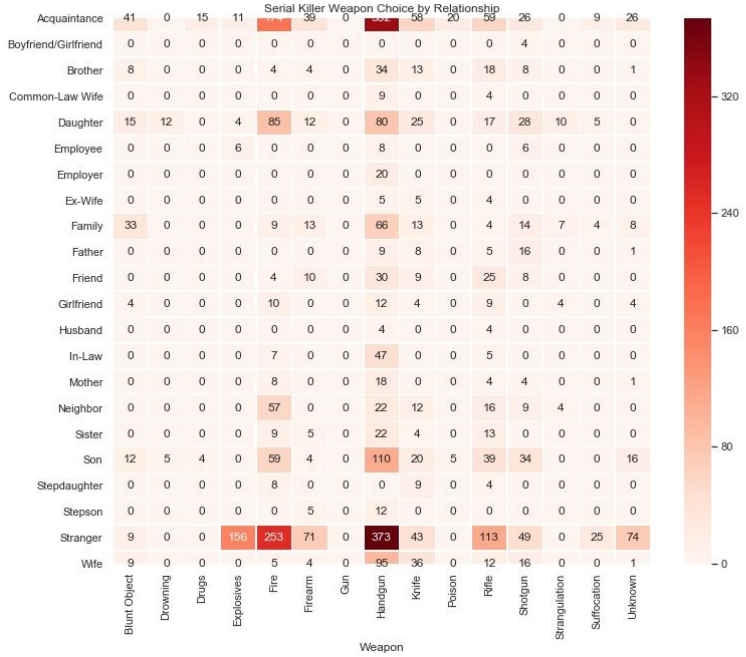


Fig 12. Gender Contrast

The incidents which have male perpetrators and female victims are a lot more than the other way around, this reflects severe social issues like male's dominance on violence and how they use physical advantage to hurt women. Also, the fact that ninety percent of total murderers except unknown is male, shows how likely men choose to use violent ways to solve problems instead of wisdom and reason.

Fig 13. Under 18 Weapon & Relationship

The two darkest matrices reflect issues of accessibility of guns for minors, purchasing weapons is actually prohibited by law for people who are under 21 in most states. Hence, their parents should pay attention to the existence of black market and their chance to use guns which are kept at home for safety purpose. Among all relationships, friend and girlfriend have highest frequency except stranger and acquaintance, shows sometimes closer relationship can be dangerous in their puberty. And the existence of wife and husband data will make people question the accuracy of dataset for sure.



Serial Killer Number of Incidents

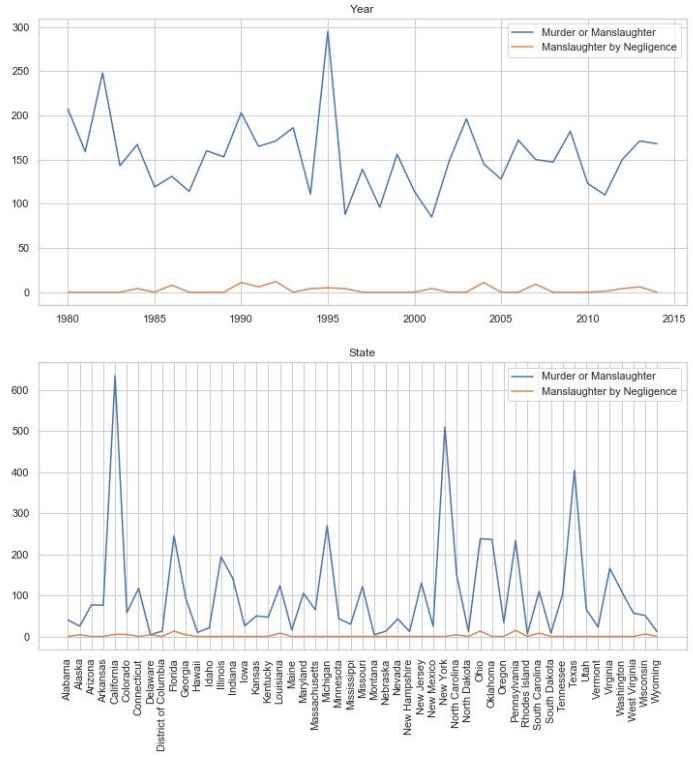


Fig 14, 15. Serial Killer

The peak points in the second plot of time series and location data are actually follow the rules of the whole dataset. The real interesting thing is actually happening in the first plot, check the color depth of two weapons called Fire and Explosives, it means some of the so called serial killer cases are possible to be misunderstandings unless all the perpetrators do that on purpose. Under these circumstances, some additional data like details of crime scene location or jobs information can really be helpful to make further judgement. And if the viewers look at the relationship, Son, Daughter, In-law, Family and Brother, the fact that these labels have relatively high frequency just shows family issues sometimes get more than one person involved.

IV. Discussion, Future Work

Even some new insights have been revealed during the analysis process, this entire project has serious limitations on bringing the work outcome to the next stage. It still remains at a superficial level as a data mining project due to various reasons, including the lack of numerical information, the questionable accuracy of the dataset, etc.

On the visualization aspect, this project also clearly has its pros and cons. For some of the graphs the details still need to be shaped, like the heatmap contains some junk information which is useless and taking space. The radar chart doesn't reach one hundred percent stratification either, because to make it readable sampling has been applied to the dataset and this may cause bias. Also some statements had been made based on the visualization requires further clarification and verification to become fully conclusive.

For future work, the dataset has to incorporate more data therefore advanced classification can be performed on it. The additional data can be the coordinates of each incident, visual images of the criminals, job status, income level and other informative data. With more data people can apply more machine learning or even deep learning algorithms to it to do more experiments. The ultimate goal of extension of this project can be building up accurate killer persona profile and predict the characteristics of those criminals in cold cases to solve them.

V. Conclusion

With the objectives of extracting attributes information, finding patterns, investigating special cases, mapping out data points and implementing machine learning classification, this project utilizes different types of data included in the homicides report dataset to create a variety of visualizations and provides some valuable insights based on the them. Even though there is still huge room for further improvement, this is a fine start and a great foundation for building up more work on the next level.

VI. References

Athithya, Vijay. "Scope of Analytics in Sports World." *Medium*, Towards Data Science, 1 May 2019, https://towardsdatascience.com/scope-of-analytics-in-sports-world-37ed09c39860.

Columbus, Louis. "53% Of Companies Are Adopting Big Data Analytics." *Forbes*, Forbes Magazine, 25 Dec. 2017, https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/#6d77785a39a1.

"Facing the Threat: Big Data and Crime Prevention." *Internet of Things Blog*, 19 Sept. 2017, https://www.ibm.com/blogs/internet-of-things/big-data-crime-prevention/.

Ford, Matt. "What Caused the Great Crime Decline in the U.S.?" *The Atlantic*, Atlantic Media Company, 15 Apr. 2016, https://www.theatlantic.com/politics/archive/2016/04/what-caused-the-crime-decline/477408/.

"RIMS Crime Analysis and Statistical Analysis Heat Mapping Pin Mapping." *Sun Ridge Systems, Inc.*, https://sunridgesystems.com/home/applications/records-management/crime-analysis/.

Roos, "Creating Interactive Crime Maps with Folium." *Data Science Blog by Domino*, 13 Oct 2015, https://blog.dominodatalab.com/creating-interactive-crime-maps-with-folium/.

Tafazoli, Sadaf. "My Notes on Chicago Crime Data Analysis." *Medium*, Medium, 19 Dec. 2018, https://medium.com/@stafa002/my-notes-on-chicago-crime-data-analysis-ed66915dbb20