**Investigating Crime Information in the United States from 1980 to 2014**

Motivation and Background:

The recent spike in crimes caused by social unrest motivated us to consider crime incidences –specifically homicides--and see what inferences I can make about trends of the data. With a deteriorating criminal justice system in the United States of America, I believe that by using data analysis and big data we can come up with a more effective system where marginalized and affected groups can be protected beforehand.

Contrary to what Laurel Eckhouse - a researcher with the Human Rights Data Analysis Group’s Policing Project, and a doctoral candidate in political science at the University of California at Berkeley, who believes that big data be reinforcing racial bias in criminal justice system, I think these data-driven tools remove human bias from the system, making it more fair as well as more effective.

As I write this paper, Big data is expanding to the criminal justice systems. According to The Washington Post, in [Los Angeles](http://www.foxla.com/news/local-news/46625629-story), “police use computerized “predictive policing” to anticipate crimes and allocate officers. In [Fort Lauderdale](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing), Fla., machine-learning algorithms are used to set bond amounts. In states across the country, [data-driven estimates](https://fivethirtyeight.com/features/prison-reform-risk-assessment/) of the risk of recidivism are being used to set jail sentences.”

I acknowledge that this method has a lot of drawbacks like data-driven decision-making risks exacerbating, rather than eliminating, racial bias in criminal justice. Black defendant is more likely to have a prior conviction than the white one, despite identical conduct because of racial biases in arrest and conviction rates. According to Laurel Eckhouse, a risk assessment relying on racially compromised criminal-history data will unfairly rate the black defendant as riskier than the white defendant. Risk-assessment tools typically evaluate their success in predicting a defendant’s dangerousness on rearrests — not on defendants’ overall behavior after release. If our two defendants return to the same neighborhood and continue their identical lives, the black defendant is more likely to be arrested. Thus, the tool will falsely appear to predict dangerousness effectively, because the entire process is circular: Racial disparities in arrests bias both the predictions and the justification for those predictions.

But, these tools use archaic algorithms, which can and have to be updated. With improved algorithms, we can accurately predict marginalized groups in certain areas, and we can change policies, improve the justice system by updating judges with the new information and have a more inclusive and safe society.

In this paper, I tried to find trends based on demographics—such as if a certain age, sex, or race is more likely to be affected than others, then I can use the information to predict future cases. Additionally, by looking at incidence rates of homicides, I can get a better idea of which states have lower crime rates than others and to see if there are generally different policies in certain states that have lower homicide rates, then their policies may be implemented elsewhere to reduce crime. I can also do the same for states that have very high rates of homicide—what policies may or may not be in place in these states that could cause these high rates? I was also curious to see if there was a trend in weapons used overtime—to see if more guns and automated weapons are used now than they were in 1980 or vice versa.

Summary and Research Questions:

This research project considers trends and insights I can make from a dataset that consists of information regarding homicides in the US from 1980-2014. I would like to take a close look at demographics that might be more at risk than others, as well as general trends such as weapon type over time. In detail, these are my research questions and the answers I arrived at after using my program to compute values:

1. What is the most common victim age, sex, and race affected in the entire US?

2. What is the state which has the highest number of incidents? What is the state with the lowest number?

3. Are specific relationships between perpetrator and victim more susceptible? Such as wife- husband, or boyfriend-girlfriend?

4. Have controversial world events caused “spikes” in homicide rates? If so, which events could be attributed to three biggest spikes in incidents per month?

5. How do most common victim race, age, and sex vary between two states? Such as Florida and

Washington? California and Georgia? How does victim race vary over all the states?

6. Does the use of handguns as a murder weapon increase from 1980 to 2014?

Dataset:

The Murder Accountability Project is the most complete database of homicides in the United States currently available. This dataset includes murders from the FBI's Supplementary Homicide Report from

1976 to the present and Freedom of Information Act data on more than 22,000 homicides that were not reported to the Justice Department. This dataset includes the age, race, sex, ethnicity of victims and perpetrators, in addition to the relationship between the victim and perpetrator and weapon used.

• [https://goo.gl/Y0nTdb (Ac](https://goo.gl/Y0nTdb)tual CSV file, need to download before viewing)

My dataset has abundant information of each crime, with these factors labeled for each incident, with a total of 638,455 cases:

|  |  |  |
| --- | --- | --- |
| 1. | Record number | 13. Victim Age |
| 2. | Agency Code | 14. Victim Race |
| 3. | Agency Name | 15. Victim Ethnicity |
| 4. | Agency Type | 16. Perpetrator Sex |
| 5. | City | 17. Perpetrator Age |
| 6. | State | 18. Perpetrator Race |
| 7. | Year | 19. Perpetrator Ethnicity |
| 8. | Month | 20. Relationship (between victim and perp.) |
| 9. | Incident | 21. Weapon |
| 10. | Crime Type | 22. Victim Count |
| 11. | Crime Solved | 23. Perpetrator Count |
| 12. | Victim Sex | 24. Record Source |

Methodology (python code):

To find most common victim age, race, and sex, and relationship:

1. First, look through each row of the file. Each row of the file will include the column name with its respective value for that case number. With my columns of interest being “Victim Age”, “Victim Sex”, “Victim Race”, and “Relationship”, look through each row and find the value associated with this column name. Accumulate values of each row for these specific column names. For example, the “Victim age” category would accumulate every single valid age.

2. Using the data points accumulated for each column, find the most common value of the data points associated with each column of interest.

To find most common victim age, race, sex, and relationship for an individual state:

1. Using the same method as above, find the modes for selected categories, but first categorize by state. Look through data file and accumulate modes for each category for each state.

For finding spikes of incidents:

1. First look through each row of the file—each row has the number of incidents for that case number. For every year from 1980 to 2014 accumulate the total incidents per month for that year.

2. Using the incident values from month to month, compare each value with next value. If the next

value is equal to or greater than 150% of the previous value, then that indicates that there is a spike in the data.

3. Record the months in which the spike took place between. For example, if the total incidents for March totaled up to 150% times the total incidents for February in the year 1991, record that there is a spike from February to March of 1991.

To find three highest spikes:

1. When you find that there is a “spike” between two months, keep track of how much the count

increases by. The three highest differences are the three largest spikes. To find if the use of handguns increases over time:

1. Look through each row of the file, find rows in which weapon type used was “handgun”.

Accumulate total number of these cases for every year.

2. Plot graph of number of cases using handgun as weapon type vs. year. Look at trend in graph to see if number of cases increases or decreases.

To find which states have the highest number of incidents:

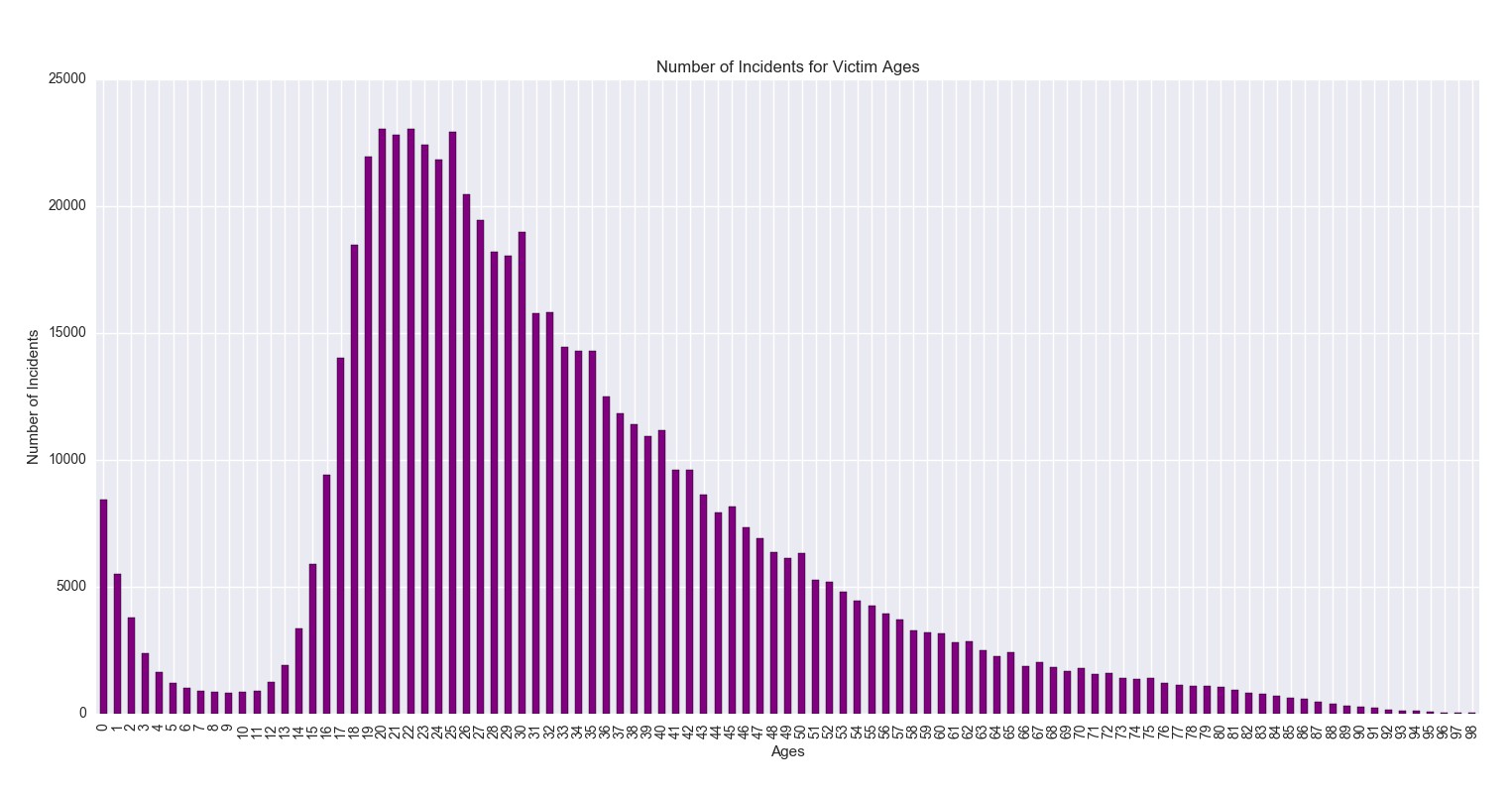
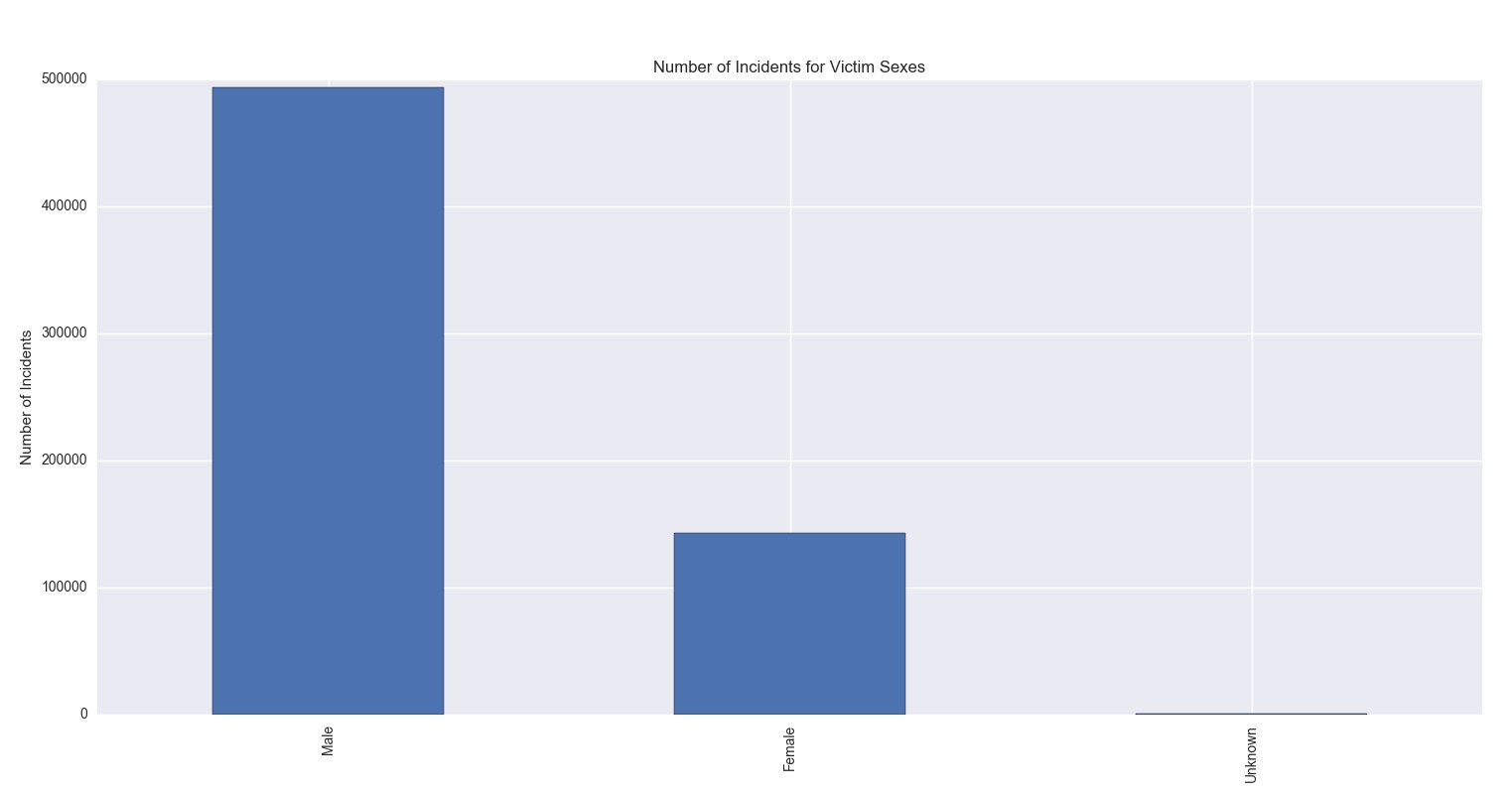
1. Look through each row of the csv file. For every state accumulate incident totals. Graph incident totals vs. state.

2. Look at the graph to see which states have highest incidents and which states have lowest incidents.

Results:

1. What is the most common victim age, sex, and race affected in the entire US?

From my data, I found that the most common victim age was 22 at 3.66%, most common victim sex is Male at 77.51%, and most common victim race was White at 50.24%. In theory that would mean that the most at-risk demographic would be white males who are around 22 years old. However, this does not take into consideration the overall demographics of the United States. If this demographic is prevalent through the entire US, that could skew this information as they would naturally be the greatest at risk when simply looking at chance. This information does not account for different regions of the US, as there could be different demographics more at risk in different parts of the US. Below are two graphs that show the number of incidents for each sex, and number of incidents for ages 0-98.



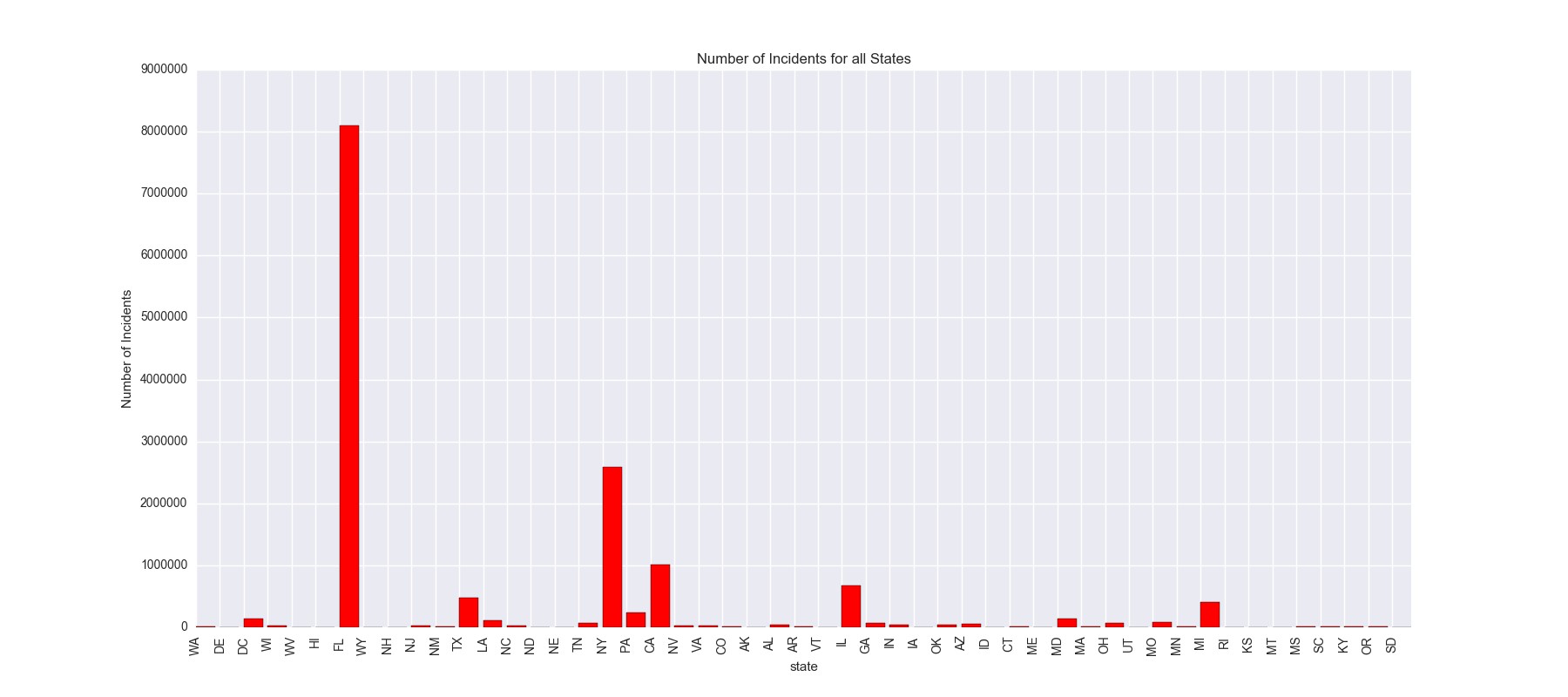
2. What is the state which has the highest number of incidents? What is the state with the lowest number?

The state with the highest total incidents from 1980 – 2014 was Florida, and the state with the lowest number was North Dakota. North Dakota was least only by a couple hundred counts from the states with the second and third lowest incidents—which was not of much importance to us. What I found interesting about this data was that the margin at which Florida was the state with the highest incidents. Compared to New York, the state with the second highest total incidents, Florida was an increase of over

5,000,000 counts. This stood out to us, and I researched more into homicides in Florida through the

1980’s to 2014. I discovered that Florida was a central part of drug war that took place between Colombia and the US in the 1980’s. During this time, there were record numbers of drug-related homicides in South Florida, which could be a reason for this drastically high incidence rate in this state.

Below is a graph of the incident count for every state in the US. This graph shows better how there is a great difference between the states with the first and second highest total incidents.



3. Are specific relationships between perpetrator and victim more susceptible? Such as wife- husband, or boyfriend-girlfriend?

The most common relationship that I found was “acquaintance,” meaning that the victim and perpetrator may have known each other but were not relatives, in a relationship, or close friends. This is important because it shows that the likelihood of someone you do not know of attacking you is low—it is more likely someone you have talked to or know already. That might be frightening, but the overall chance of this happening is in general relatively low.

4. Have controversial world events caused “spikes” in homicide rates? If so, which event could be attributed to the highest increase in homicides?

The three greatest spikes occurred from: November to December 1983, February to March 2007, and

November to December 1980.

Originally, when I were looking for spikes in crime, I believed that I would find periods of time which were spread apart and that could be attributed to various causes and world events. Two of the spikes that I found occurred within three years of each other, so I tried to look rather at trends of events that occurred in those years rather than looking at individual causes for these spikes. For the year

2007, I looked at specific events.

When doing my research as to what could have caused these spikes, I found that starting from the

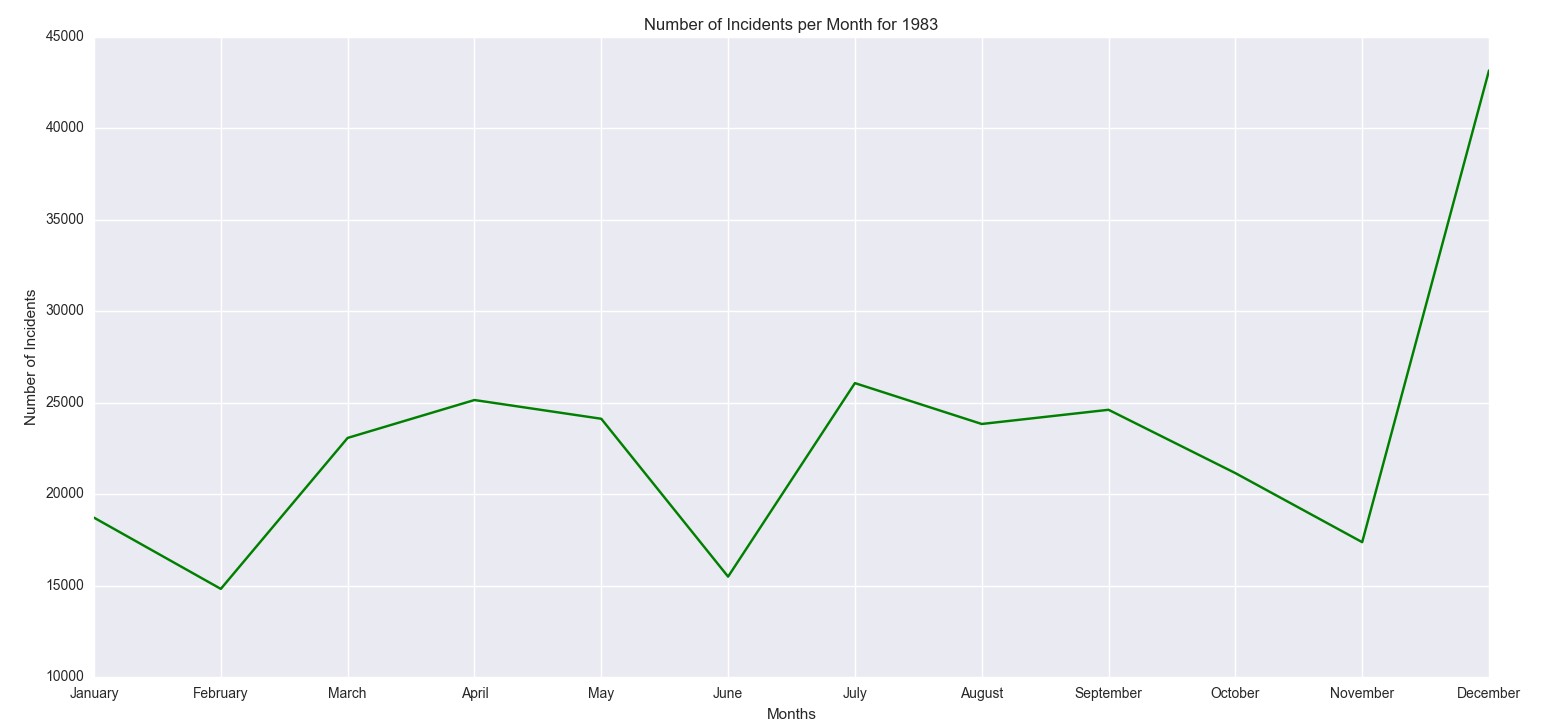
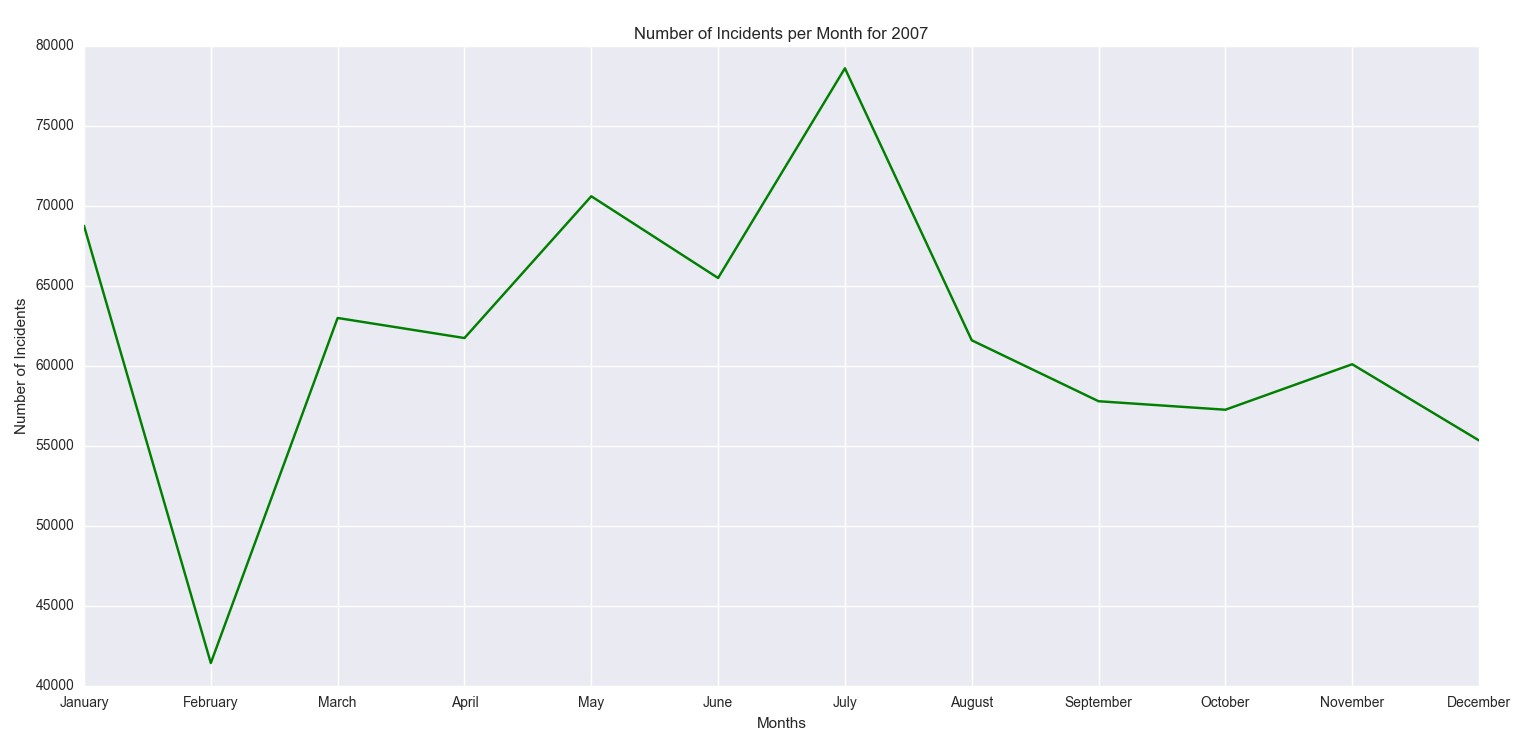
1980’s, gang-violence was heavily prominent, especially in urban cities in the US. This caused a severe incline in violent crime in the US and sparked action against gang-violence shortly after during the late

1990’s. There was however, not any specific event that I found that could be attributed to the two spikes in the early 80’s.

In 2007, there was a marked increase in violent crime. Criminologists wrote about multiple theories as to why homicide rates were going back up, and they suggested that the spike could be a result of an increase in the juvenile population, growing numbers of released prison inmates and the rise of serious gang problems in smaller jurisdictions.

Inserted below are the graphs for the years 1983 and 2007, where you can clearly see the spikes. In the program, the graphs for all years with spikes are saved, so you can look at those graphs as well.

Page 6



5. How do most common victim race, age, and sex vary between two states? Such as Florida and

Washington? California and Maine? How does victim race vary over all the states?

a. Between Florida and Washington:

i. Washington: Age: 22 at 3.47%, Race: White at 50.24%, Sex: Male at 77.51%

ii. Florida: Age: 25 at 3.30%, Race: White at 55.04%, Sex: Male at 75.31%

b. Between California and Georgia:

i. California: Age: 20 at 4.31%, Race: White at 66.08%, Sex: Male at 81.21%

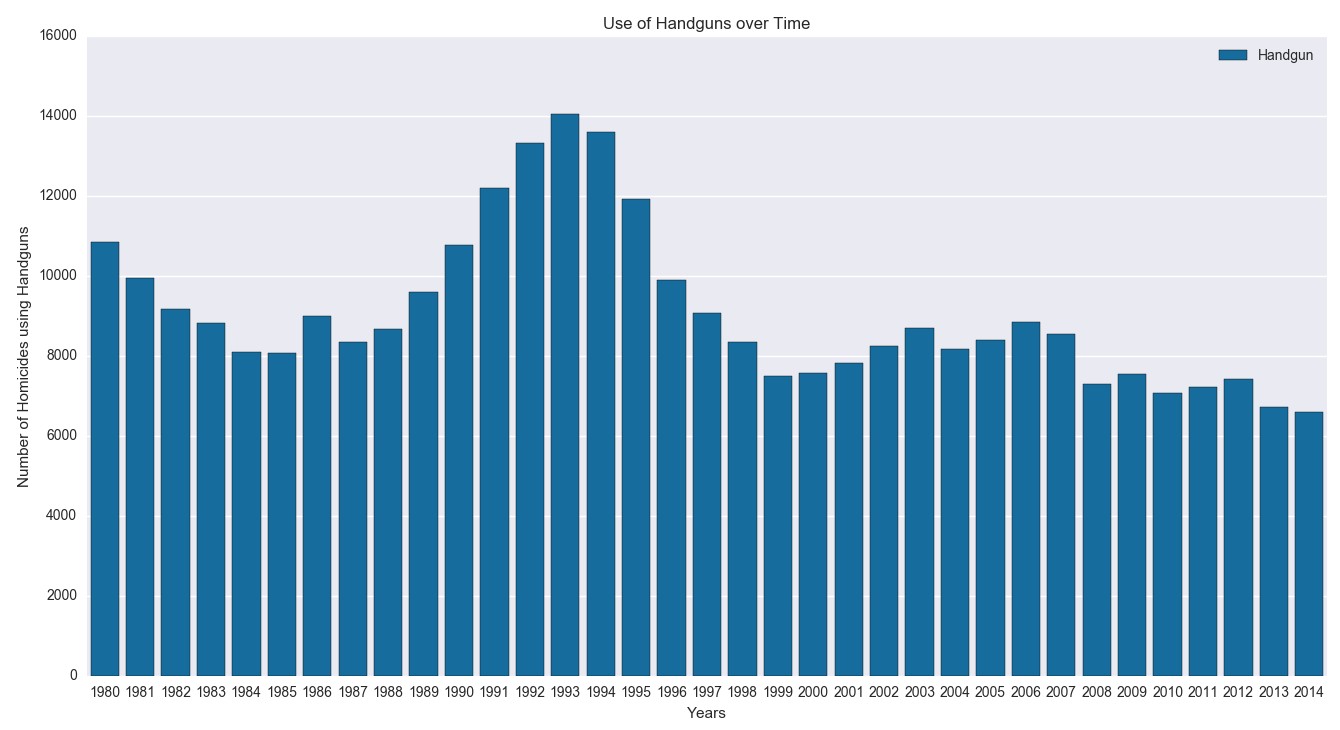
ii. Georgia: Age: 23 at 3.62%, Race: Black at 67.90%, Sex: Male at 75.56%

When looking at the differences between the states chosen, it becomes clear what demographics are at risk in various areas of the United States. For example, when comparing what races are affected in California and Georgia, two states on opposite sides of the country, black people are more affected in Georgia compared to mostly white people affected in California. Of course, this information has many outlying circumstances—such as population demographics per state and violence rates among different races in specific states, but this gives us a general idea of what people are being affected in what states. This is important, as this information can tell us who to focus on when trying to predict crime across the US in different states, and keep people safer country-wide.

6. Does the use of handguns increase over time?

From the graph, I can see that the use of handguns decreases after a sharp increase between the years

1988-1993. The rapid decline could be attributed to stricter gun laws, as well as decline in drug-related crimes, which, from my research, seemed to be a large proportion crimes which involved guns. Additionally, with the prevalence of gang-activity in urban areas during these years, the use of guns as a weapon in homicides was very high. Long-term solutions were implemented these years—involving reducing number of youth involved in gang activity, which could attribute to this decline in gun use after the mid-90’s.



Testing:

To test my program, I created a smaller dataset called “crime\_data\_small.csv” that consisted about only

15 rows out of the entire larger dataset with various state names, ages, races, etc. This allowed us to initially manually calculate the modes and frequencies that I were looking for, and make sure that the results that I were getting out of the program were correct. Most of my testing was done this way, in addition to print statements that I inserted in the code throughout the process to see what values were being computed and transferred to different data structures and functions. Left in the code are some assert statements that ensure that user input is valid.

I made sure that my graphs were accurate by graphing smaller datasets, and again, doing the graphs ourselves by hand to make sure the output graph looks the same as my hand drawn graphs. In the end, my results are accurate, and I can assure that all values computed and graphed are correct.

References:

1. Pattavina, April. Information Technology and the Criminal Justice System. Thousand Oaks, Calif.: Sage Publications, 2005. Web.
2. McKean, Jerome B., and Byers, Bryan. Data Analysis for Criminal Justice and Criminology : Practice and Applications. Boston: Allyn and Bacon, 2000. Web.
3. Eckhouse, Laurel. "Big data may be reinforcing racial bias in the criminal justice system." The Washington Post. 10 Feb 2016. Web. 27 May 2017.