Data Analysis Report

The evolution of criminality in the US since 1980

Louis Chauvet - Alexandre Cros
5ISS - INSA Toulouse
January 8th 2020

Prelude

All of our code and the dataset we used can be followed following this link:
https://github.com/acros1/data-analysis
This repository includes the dataset as well as the code used to obtain the graphs in this report. It also includes the generation of other graphs we did not end up using in this report, but were useful to getting to our conclusions.

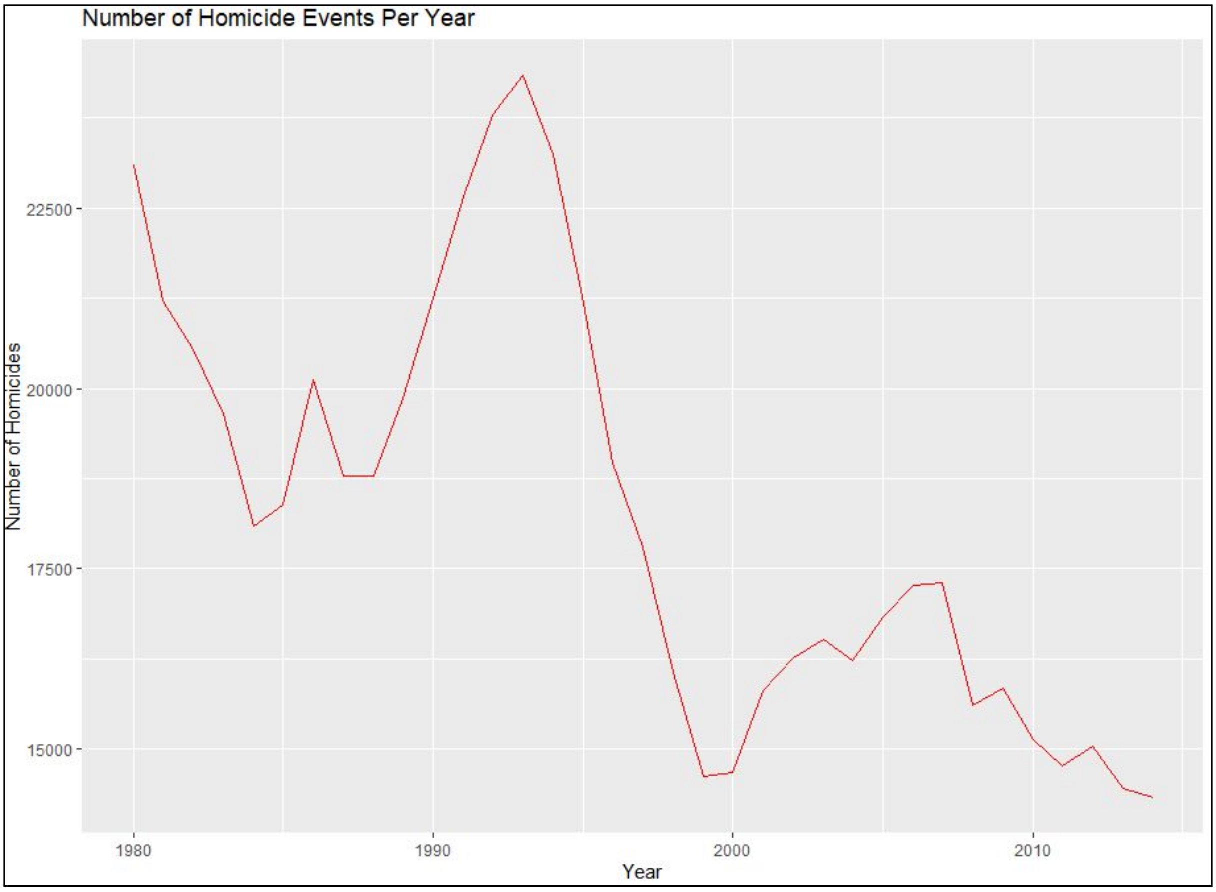
Introduction

For this project, we browsed Kaggle to find a dataset we'd like to analyse, and ended up choosing a dataset detailing every homicide in the USA from 1980 to 2014. We chose it because of curiosity, but also because of how detailed it seemed to be: 24 unique caracteristics (symbolised by the 24 columns), and 638 854 entries (symbolised by that many rows). Such a large amount of entries leads to a more detailed analysis, and more reliable conclusions. A description of the dataset as well as a link to it on Kaggle can be found in the readme file of the repository linked in the Prelude.

The angle we decided to study the dataset from was following this question: Does this data tell us anything about how homicidal crimes have evolved over the course of 34 years?
We'll be detailing our findings through a series of commented graphs, trying to take away as much information from them as we can.

Analysis

We decided to plot many different pieces of data, and expose our most interesting findings here. Since we want to look at how these statistics evolved over the year, we had to discard a lot of data which, while it was interesting from an informative point of view, didn't show any meaningful evolution over the years. Some points were also harder to analyze since a relatively high proportion of crimes are, unfortunately, never solved. We were still able to extract some very interesting information, so let's get to it.

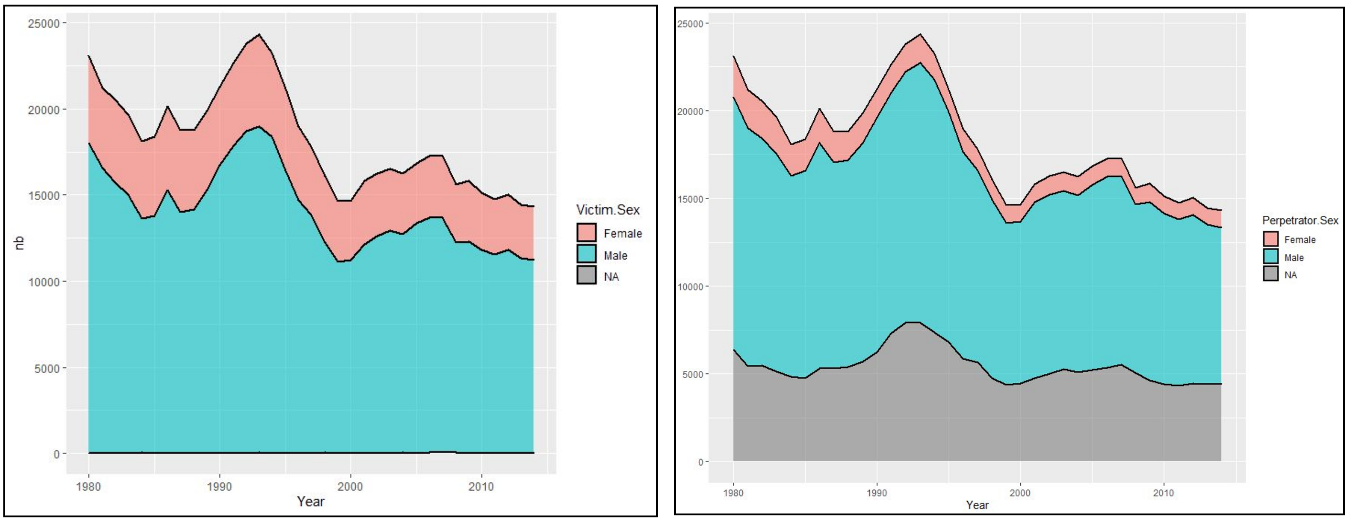


Plot 1: Number of homicide cases per year

The first and most basic graph we could extract was a simple evolution of the number of homicides per year. Those values were also useful for a variety of other calculations, since percentages usually give us a better idea of evolutions than hard values.

Even though, we can definitely notice this spike and drop in the early 90s, so apparent it has its own wikipedia article. There is no clear explanation for this drop, and is usually associated with economic prosperity in the country.

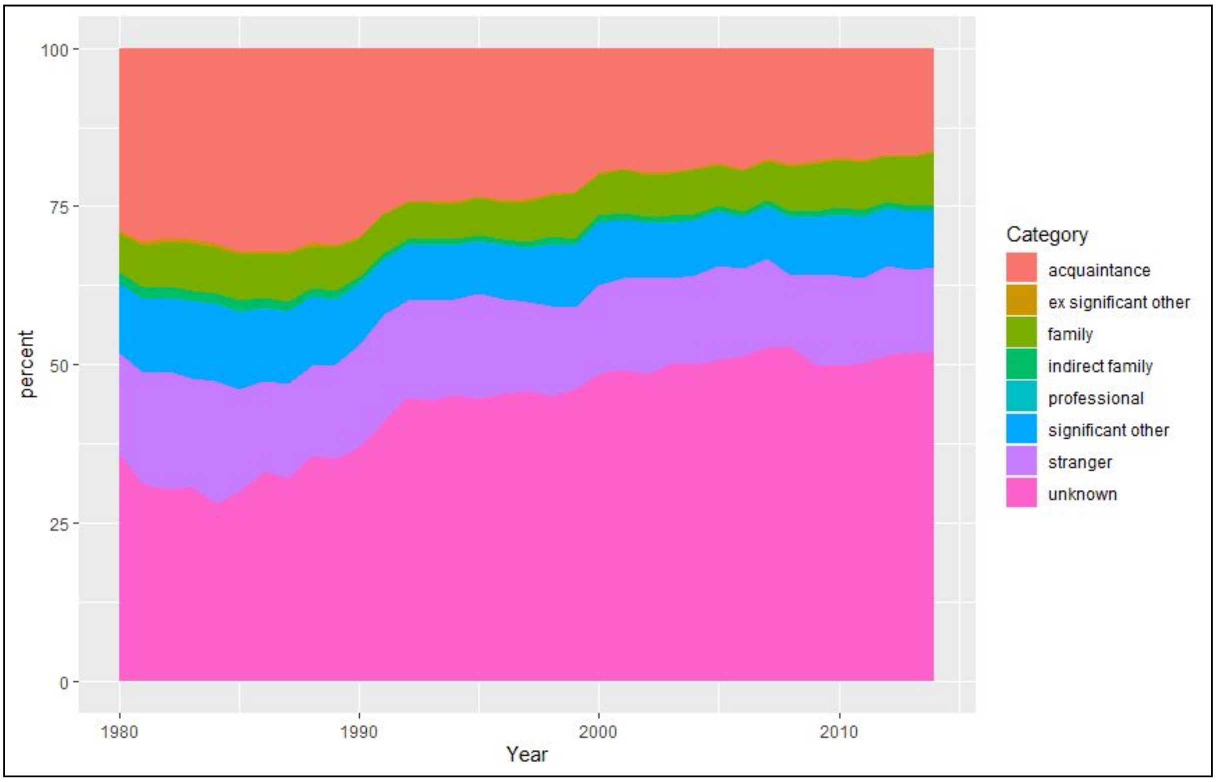
Another interesting basic measurement is the proportion of male and females among perpetrators and victims.



Plot 2 and 3: Number of male and female Victims(left) and Perpetrators(right)

While we can’t see a clear evolution of these proportions over the years, it’s easily noticed that males are the overwhelming majority of both perpetrators and victims. The unknown value in perpetrator sex is also almost exactly the same as the number of unsolved cases, as sex is one of the most basic pieces of data you can collect on someone: there are almost no cases of a solved case with unknown perpetrator sex, and it could be interesting to see how this statistic evolves in an age of more fluid concepts of gender.

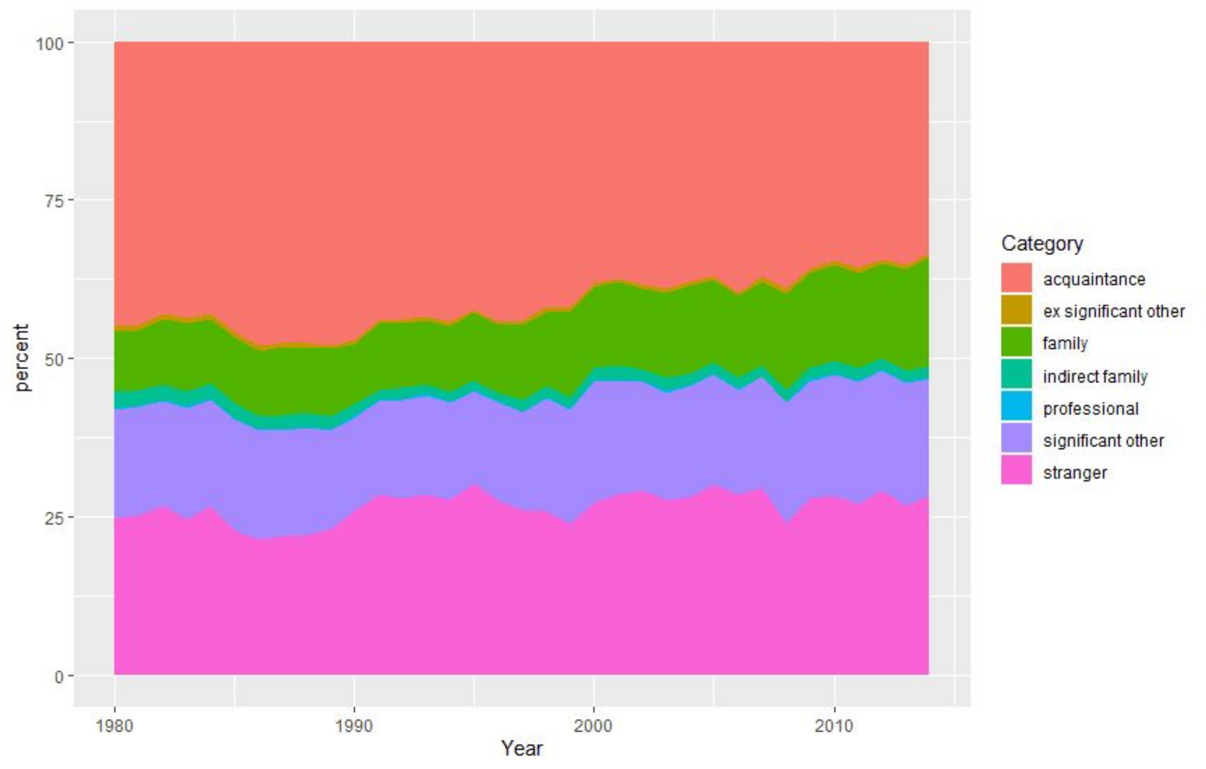
The bulk of the project was centered around the relationship between the perpetrator and the victim: we examined the relationship entry in the dataset, and studied the proportion of each relationship along the years. The dataset included over 20 different types of relationship, which gave us a very pretty rainbow chart but no exploitable information. Because of this, we decided to group all these specific categories into larger groups, the specifics of which you can find in the project’s code. We were left with 8 main categories:



Plot 4:Relationship between perpetrator and victim in percents

This first graph does not give too much clear information because of the amount of space taken by the “unknown” value. However, this ended up giving us an interesting insight on how crimes get solved: the further back we go, the more we know about the relationship between perpetrator and victim. This increase in unknowns could partly be explained by the fact that some cases are still getting solved to this day, so this unknown status might be changed at a later day if the case is ever solved. This might also be a change in criminality, where the witnesses and family of the victim cannot confirm that the perpetrator is known or unknown to the victim. Finaly, this might be a change in how the american police files cases, prefering to leave “unknown” than to put “stranger” or “acquaintance” without being 100% certain.

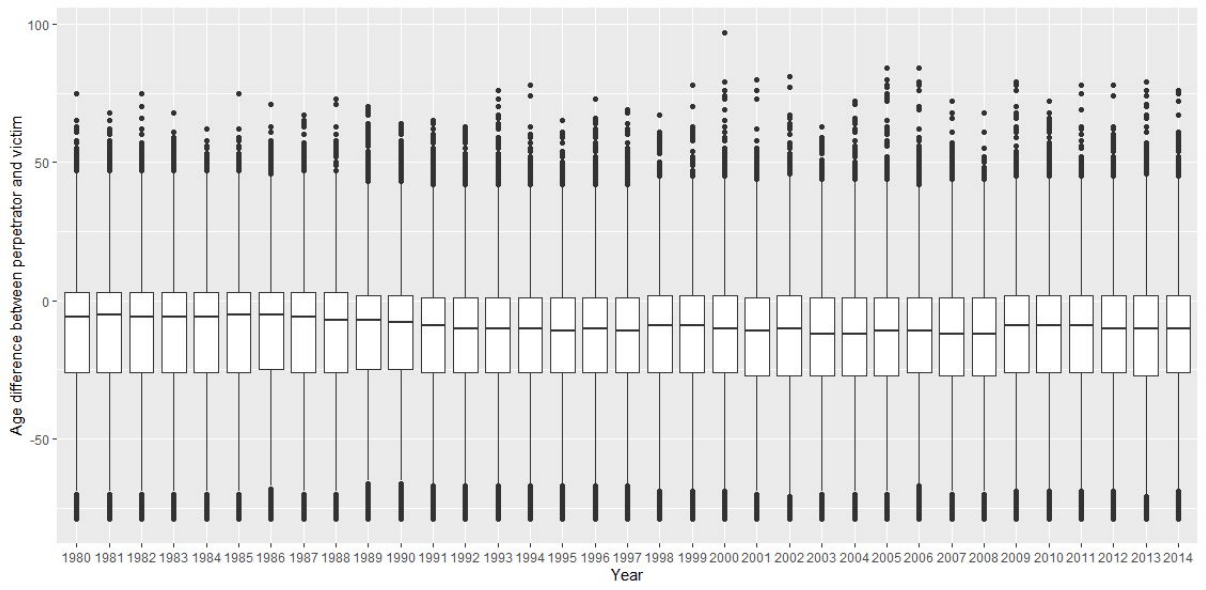
To have a better idea of the evolution of relationships we wanted to observe though, we decided to create this graph again after removing all of those unknown values.



Plot 5: Known relationship between perpetrator and victim in percents

This graph gives us much more information about the evolution of the relationship between perpetrator and victim. Most categories have stayed pretty much stable, except for two: we can see a decline in acquaintance homicides, and a sharp increase in the proportion of homicides between family members. It's a little hard to make any hard conclusions about this figure without some sociological studies to accompany it, but it might be linked to the rapid evolution of the family unit during the last few decades. In any case, this change is significant enough to be noticed.

The last statistic we decided to include in this report was the age difference between perpetrator and victim:



Plot 6: Box plot of age difference between perpetrator and victim per year

We decided to check out this statistic after a stupid thought: what age group should I be weary of? well as much as we can't tell you for sure (you might want to look out for people 16 years older than you though), we got two interesting pieces of information from this graph: firstly, there are a couple of issues with the dataset, since a few perpetrators were a thousand year older than their victims, which seems pretty far-fetched. It showed the importance of looking at data critically rather than aways taking it at face value. Secondly, it showed the importance of varying graph types: we originaly plotted the age difference per year as a simple line graph of the average age difference, giving us pretty much no information other than "be careful of people 16 years older than you". Through this box plot though, we can see that even if the average age difference never moves by more than a couple of year, the distribution of the middle 50% of values has: the median value has gone down pretty continuously towards the middle, meaning less homicides are being commited by peole of a similar age to their victim. We could not have found this clear trend without trying a variety of plotting methods.

Conclusion

This project was both very instructive and enjoyable. It was a great opportunity to introduce the tools widely used in data analysis today, and encourage us to solve our own problem.

With the freedom of chosing our own dataset to analyse came the burden to justify its advantages and flaws. This dataset was great to exploit because of the sheer amount of information it provided: many different categories and almost 700 000 entries. The subject matter was also extremelly interesting, since it was so serious and thorough. However, because of the angle we chose to analyse this data from, that being the evolution over time of different parameters, we were limited in the types of graphs we could use. There were also some graphs we decided not to share for personnal reasons, such as statistics based on race as we both did not agree with this classification and did not feel comfortable discussing them. Finally, while having so much data was great for the precision of our results, some calculations took over half an hour to compute, which was definitelly a weak point of the project.

Overall, the freedom we were given for this project made it an incredible learning opportunity, and used our curiosity as a powerful tool to learn a new language and field of study.