Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide

Jacob Kaplan

2021-07-27

Contents

List of Tables

List of Figures

Chapter 1

Preface

If you've read an article about crime or arrests in the United States in the last half century, in most cases it was referring to the FBI's Uniform Crime Reporting Program Data, otherwise known as UCR data. UCR data is, with the exception of the more detailed data that only covers murders, a *monthly number of crimes or arrests reported to a single police agency* which is then gathered by the FBI into one file that includes all reporting agencies. It is actually a collection of different datasets, all of which have information about crimes and arrests that occur in a particular jurisdiction. Think of your home town. This data will tell you how many crimes were reported for a small number of crime categories or how many people (broken down by age, sex, and race) were arrested for a (larger) set of crime categories in that city (if the city has multiple police agencies then each agency will report crimes/arrests under their jurisdiction though the largest agency - usually the local police department - will cover the vast majority of crimes/arrests in that city) in a given month.

This is a very broad measure of crime, and its uses in research - or uses for understanding crime at all - is fairly limited. Yet is has become over much of the last century - and will likely remain among researchers for at least the next decade - the most important crime data in the United States.[1]

UCR data is important for three reasons:

1. The definitions are standard, and all agencies (tend to) follow them so you can compare across agencies and over time.[2]

[1]The FBI has said they will no longer accept UCR data after 2020, instead only accepting the more detailed National Incident-Based Reporting System (NIBRS) data. However, only about half of agencies reported NIBRS data in 2019 and this number decreases steadily for earlier years. This means that UCR data has the longevity that NIBRS doesn't have, as most agencies have reported for decades, and will still be useful even though the data becomes increasingly outdated.

[2]We'll see many examples of when agencies do not follow the definitions, which really limits this data.

2. The data is available since 1960 (for most of the datasets) so there is a long period of available data.[3]
3. The data is available for most of the 18,000 police agencies in the United States so you can compare across agencies.

More than many other datasets, there will be times when using UCR data that you'll think "that's weird". This book will cover this weirdness and when we think the weirdness is just an odd - but acceptable - quirk of the data, and when it is a sign of a big problem in the data or in that particular variable and that we should avoid using it. For most of this book we'll be discussing the caveats of the above reasons - or, more directly, why these assumptions are wrong - but these are the reasons why the data is so influential.

1.1 Goal of the book

By the end of each chapter you should have a firm grasp on the dataset that is covered and how to use it properly. However, this book can't possibly cover every potential use case for the data so make sure to carefully examine the data yourself for your own particular use.

I get a lot of emails from people asking questions about this data so my own goal is to create a single place that answers as many questions as I can about the data. Again, this is among the most commonly used crime datasets and there are still many current papers published with incorrect information about the data (including such simple aspects like what geographic unit data is in and what time unit it is in). So hopefully this book will decrease the number of misconceptions about this data, increasing overall research quality.

Since manuals are boring, I'll try to include graphs and images to try to alleviate the boredom. That said, I don't think it's possible to make it too fun so sorry in advanced. This book is a mix of facts about the data, such as how many years are available, and my opinions about it, such as whether it is reliable. In cases of facts I'll just say a statement - e.g. "the offenses data is available since 1960". In cases of opinion I'll temper the statement by saying something like "in my opinion…" or "I think".

1.2 Structure of the book

This book will be divided into ten chapters: this chapter, an intro chapter briefly summarizing each dataset and going over overall issues with UCR data, and seven chapters each covering one of the seven UCR datasets. The final chapter will cover county-level UCR data,

[3]While the original UCR data first reported in 1929, there is only machine-readable data since 1960.

a commonly used but highly flawed aggregation of UCR data that I recommend against using. Each chapter will follow the same format: we'll start with a brief summary of the data such as when it first because available and how it can be used. Next we'll look at how many agencies report their data to this dataset, often looking at how to measure this reporting rate a couple of different ways. Finally, we'll cover the important variables included in the data and how to use them properly (including not using them at all) - this will be the bulk of each chapter.

1.3 Citing this book

If this data was useful in your research, please cite it. To cite this book, please use the below citation:

Kaplan J (2021). *Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide*. https://ucrbook.com/.

BibTeX format:

```
@Manual{ucrbook,
  title = {Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide},
  author = {{Jacob Kaplan}},
  year = {2021},
  url = {https://ucrbook.com/},
}
```

1.4 Sources of UCR data

There are a few different sources of UCR data available today. First, and probably most commonly used, is the data put together by the National Archive of Criminal Justice Data (NACJD)). This a team out of the University of Michigan who manages a huge number of criminal justice datasets and makes them available to the public. If you have any questions about crime data - UCR or other crime data - I highly recommend you reach out to them for answers. They have a collection of data and excellent documentation available for UCR data available on their site here. One limitation to their data, however, is that each year of data is available as an individual file meaning that you'll need to concatenate each year together into a single file. Some years also have different column names (generally minor changes like spelling robbery "rob" one year and "robb" the next) which requires more work to standardize before you could concatenate. They also only have data through 2016 which

means that the most recent years (UCR data is available through 2019) of data are (as of this writing) unavailable.

Next, and most usable for the general public - but limited for researchers - is the FBI's official website Crime Data Explorer. On this site you can chose an agency and see annual crime data (remember, UCR data is monthly so this isn't as detailed as it can be) for certain crimes (and not even all the crimes actually available in the data). This is okay for the general public but only provides a fraction of the data available in the actual data so is really not good for researchers.

Finally, I have my own collection of UCR data available publicly on openICPSR, a site which allows people to submit their data for public access. For each of these datasets I've taken the raw data from the FBI (for early years of homicide data this is actually from NACJD since the FBI's raw data is wrong and can't be parsed. For later years of homicide data this is from the FBI's raw data.) and read it into R. Since the data is only available from the FBI as fixed-width ASCII files, I created a setup file (we'll explain exactly how reading in this kind of data works in the next chapter) and read the data and then very lightly cleaned the data (i.e. only removing extreme outliers like an agency having millions of arsons in a month). For each of these datasets I detail what I've done to the data and briefly summarize the data (i.e. a very short version of this book) on the data's page on openICPSR. The main advantage is that all my data has standard variable names and column names and, for data that is small enough, provide the data as a single file that has all years. For large datasets like the arrest data I break it down into parts of the data and not all years in a single file. The downside is that I don't provide documentation other than what's on the openICPSR page and only provide data in R and Stata format. I also have a similar site to the FBI's Crime Data Explorer but with more variables available - that site is available here.

It's worth mentioning a final source of UCR information. This is the annual Crimes in the United States report released by the FBI each year around the start of October.[4] As an example, here is the website for the 2019 report. In this report is summarized data which in most cases estimates missing data and provides information about national and subnational (though rarely city-level) crime data. As with the FBI's site, it is only a fraction of the true data available so is not a very useful source of crime data for quality research. Still, this is a very common source of information used by researchers.

1.4.1 Where to find the data used in this book

The data I am using in this book is the cleaned (we'll discuss in more detail exactly what I did to clean each dataset in the dataset's chapter, but the short answer is that I did very

[4] They also release a report about the first 6-months of the most recent year of data before the October release but this is generally an estimate from a sample of agencies so is far less useful.

little) and concatenated data that I put together from the raw data that the FBI releases. That data is available on my website here. I am hosting this book through GitHub which has a maximum file size allowed that is far smaller than these data, so you'll need to go to my site to download the data; it's not available through this book's GitHub repo.

1.5 Recommended reading

While this book is designed to help researchers use this data, the FBI has an excellent manual on this data designed to help police agencies submit their data. That manual, called the "Summary Reporting System (SRS) User Manual" provides excellent definitions and examples of many variables included in the data. In this book when I quote the FBI, such as defining a crime, I quote from this manual. The manual is available to download as a PDF on the FBI's site and I've also posted it on my GitHub page in case the FBI ever take down the PDF.[5] The link on my GitHub page is here. I highly recommend that you read this manual before using the data. That manual, alongside this book which tries to explain when and how the agencies don't follow the manual, will provide a solid foundation for your understanding of UCR data.

1.6 How to contribute to this book

If you have any questions, suggestions (such as a topic to cover), or find any issues, please make a post on the Issues page for this book on GitHub. On this page you can create a new issue (which is basically just a post on this forum) with a title and a longer description of your issue. You'll need a GitHub account to make a post. Posting here lets me track issues and respond to your message or alert you when the issue is closed (i.e. I've finished or denied the request). Issues are also public so you can see if someone has already posted something similar.

For more minor issues like typos or grammar mistakes, you can edit the book directly through its GitHub page. That'll make an update for me to accept, which will change the book to include your edit. To do that, click the edit button at the top of the site - the button is highlighted in the below figure. You will need to make a GitHub account to make edits. When you click on that button you'll be taken to a page that looks like a Word Doc where you can make edits. Make any edits you want and then scroll to the bottom of the page. There you can write a short (please, no more than a sentence or two) description of what you've done and then submit the changes for me to review.

[5] This is far more likely to happen as a result of standard government changing a site and forgetting to update the link rather than intentionally making the manual unavailable.

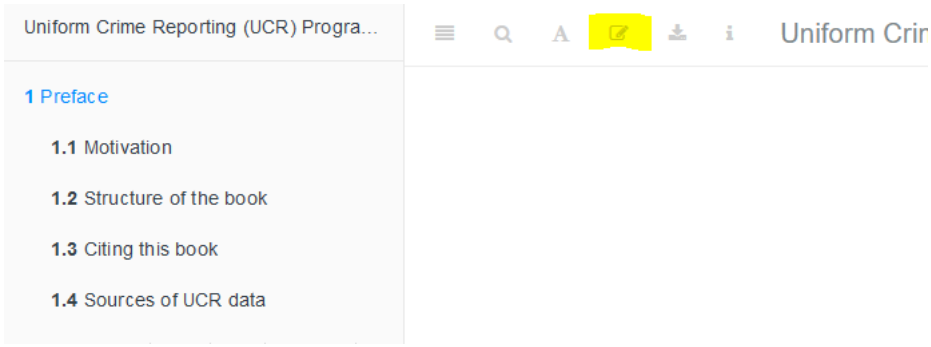


Figure 1.1: The edit button for how to make edits of this book.
(#fig:unnamed-chunk-1)

Please only use the above two methods to contribute or make suggestions about the book. Don't email me. While it's a bit more work for you to do it this way, since you'll need to make a GitHub account if you don't already have one, it helps me. I wrote this book, in part, to help my career so having evidence that people read it and are contributing to it is important to me. It's a way to publicly measure the book's impact.

About the Author

Jacob Kaplan holds a PhD and a master's degree in criminology from the University of Pennsylvania and a bachelor's degree in criminal justice from California State University, Sacramento. He is the Chief Data Scientist at the Research on Policing Reform and Accountability RoPRA at Princeton University. His current research portfolio includes evaluating police policy and reforms, place-based crime prevention, measuring spatial crime concentration, and simulating how firing 'bad apples' affects police complaints and uses of force. In the past he's written on the effect of marijuana decriminalization on domestic violence, how increasing the number of police officers affects prison trends, how outdoor lighting affects crime and perception of safety, and public perceptions of forensic science techniques. He is the author of several R packages that make it easier to work with data, including fastDummies and asciiSetupReader. His website allows for easy visualization of crime-related data and he has released over a dozen crime data sets (primarily FBI UCR data) on openICPSR that he has compiled, cleaned, and made available to the public.

For a list of papers he has written (including working papers), please see here.

For a list of data sets he has cleaned, concatenated, and made public, please see here.

For a list of R packages he has created, please see here.

Chapter 2

Overview of the Data

One of the first, and most important, questions I think people have about crime is a simple one: is crime going up? Answering it seems simple - you just count up all the crimes that happen in an area and see if that number of bigger than it was in previous times.

However, putting this into practice invites a number of questions, all of which affect how we measure crime. First, we need to define what a crime is. Not philosophically what actions are crimes - or what should be crimes - but literally which of the many thousands of different criminal acts (crimes as defined by state or federal law) should be considered in this measure. Should murder count? Most people would say yes. How about jaywalking or speeding? Many would say probably not. Should marital rape be considered a crime? Now, certainly most people (all, I would hope) would say yes. But in much of the United States it wasn't a crime until the 1970s (**??**).

Next, we have to know what geographic and time unit to measure crimes at since these decisions determine how precise we can measure crime and when it changed. That is, if you are mugged on Jan 1st at exactly 12:15pm right outside your house, how do we record it? Should we be as precise as including the exact time and location (such as your home address)? Out of privacy concerns to the victim, should we only include a larger time unit (such as hour of the day or just the day without any time of day) or a larger geographic unit (such as a Census Tract or the city)?

The final question is that when a crime occurs, how do we know? That is, when we want to count how many crimes occurred do we ask people how often they've been victimized, do we ask people how often they commit a crime, do we look at crimes reported to police, crimes charged in a criminal court? Each of these measures will likely give different answers as to how many crimes occurred.[1]

[1]The Bureau of Justice Statistics does measure crime by asking a random sample of people whether they were the victim of a crime. For more on this, please see their National Crime Victimization Survey reports

The FBI answered all of these questions in 1929 when they began the Uniform Crime Reporting (UCR) Program Data, or UCR data for short. **Crime consists of eight crime categories - murder, rape, robbery, aggravated assault, burglary, motor vehicle theft, theft, and simple assault - that are reported to the police and is collected each month from each agency in the country.** So essentially we know how many of a small number of crime categories happen each month in each city (though some cities have multiple police agencies so even this is more complicated than it seems). These decisions, born primarily out of the resource limitations of 1929 (i.e. no computers), have had a major impact on criminology research. The first seven crime categories - known as "Index Crimes" or "Part 1 crimes" (or "Part I" sometimes) - are the ones used to measure crime in many criminology papers, even when the researchers have access to data that covers a broader selection of crimes than these.[2] These are also the crimes that the news uses each year to report on how crime in the United States compares to the previous year. The crime data actually also includes the final crime, simple assault, though it is not included as an index crime and is, therefore, generally ignored by researchers - a relatively large flaw in most studies that we'll discuss in more detail in Section @ref(indexCrimes).

2.1 Crimes included in the UCR datasets

UCR data covers only a subset - and for the crime data, a very small subset - of all crimes that can occur. It also only includes crimes reported to the police. So there are two levels of abstraction from a crime occurring to it being included in the data: a crime must occur that is one of the crimes included in one of the UCR datasets (we detail all of these crimes below) *and* the victim must report the incident to the police. While the crimes included are only a small selection of crimes - which were originally chosen since at the time the UCR was designed these were considered important offenses and among the best reported - this is an important first step to understanding the data.

UCR data should be understood as a loose collection of data that seeks to understand crimes and arrests in the United States. There are seven datasets in the UCR collection that each cover a different topic or a subset of a previously covered topic: crimes, arrests, property crimes specifically, homicides specifically, police officers killed or assaulted, arson specifically, and hate crimes.[3] In this section we'll go over the crimes included in the two main UCR datasets: Offenses Known and Clearances by Arrest and (which I like to call the "crime"

[2] Arson is also an index crime but was added after these initial seven were chosen and is not included in the crimes dataset (though is available separately) so is generally not included in studies that use index crimes.

[3] There is also a human trafficking dataset though this is a newer dataset and rarely used so I will not cover it in this book.

dataset) and the Arrests by Age, Sex, and Race (the "arrests" dataset). These are the most commonly used UCR datasets and the stolen property and homicide datasets are simply more detailed subsets of these datasets. The hate crime dataset can cover a broader selection of offenses than in the crimes or arrests data, so we'll discuss those in the hate crimes chapter.

2.1.1 Crimes in the Offenses Known and Clearances by Arrest dataset

As mentioned above - and as most criminology papers will tell you - the crimes included in the UCR's Offenses Known and Clearances by Arrest data are the seven index crimes (eight when including arson, though arson is only reported in its own dataset so is usually excluded) - homicide, rape, robbery, aggravated assault, burglary, theft, and motor vehicle theft - as well as simple assault. This is true but incomplete. The data also includes subcategories for all crimes other than simple assault and theft - though theft has its own UCR dataset which goes into detail about the thefts. Both robbery and aggravated assault, for example, have subcategories showing which weapon the offender used (if any) during the crime. This allows for a more detailed understanding of the crime than looking only at the broad category. I'm not sure why most research includes only the broader categories and doesn't tend to look at subcategories, but that seems to be the case in most studies that I have read. Some police agencies only report the broader categories and don't report subcategories, but most report subcategories so this is an under-exploited source of data.

1. Homicide

- Murder and non-negligent manslaughter
- Manslaughter by negligence

2. Rape

- Completed rape
- Attempted rape

3. Robbery

- With a firearm
- With a knife of cutting instrument
- With a dangerous weapon not otherwise specified
- Unarmed - using hands, fists, feet, etc.

4. Aggravated Assault (assault with a weapon or with the intent to cause serious bodily injury)

- With a firearm
- With a knife of cutting instrument
- With a dangerous weapon not otherwise specified
- Unarmed - using hands, fists, feet, etc.

5. Burglary

- With forcible entry
- Without forcible entry
- Attempted burglary with forcible entry

6. Theft (other than of a motor vehicle)

7. Motor vehicle theft

- Theft of a car
- Theft of a truck or bus
- Theft of an other type of vehicle

8. Simple Assault

2.1.2 Crimes in the Arrests by Age, Sex, and Race dataset

The crimes included in the Arrests by Age, Sex, and Race - the "arrest" data tells you how many people were arrested for a particular crime category - are different than those in the crime data. The arrest data covers a wider variety of crimes, including drug and alcohol crimes, gambling, and fraud. However, it is also less detailed than the crime dataset when it comes to violent crime. While it covers the same broad categories of violent crimes as the crimes data - murder, rape, robbery, aggravated assault, and simple assault - it doesn't include the more detailed breakdown that is available in the crime data. For example, in the crime data robbery is included as well as the subcategories of robbery with a gun, robbery with a knife, robbery with another dangerous weapon, and robbery without a weapon. In comparison, the arrest data only includes robbery without any subcategories.

1. Homicide

- Murder and non-negligent manslaughter
- Manslaughter by negligence

2. Rape
3. Robbery
4. Aggravated assault
5. Burglary
6. Theft (other than of a motor vehicle)
7. Motor vehicle theft
8. Simple assault
9. Arson
10. Forgery and counterfeiting
11. Fraud
12. Embezzlement

13. Stolen property - buying, receiving, and possessing
14. Vandalism
15. Weapons offenses - carrying, possessing, etc.
16. Prostitution and commercialized vice
17. Sex offenses - other than rape or prostitution
18. Drug abuse violations - total

- Drug sale or manufacturing
  - Opium and cocaine, and their derivatives (including morphine and heroin)
  - Marijuana
  - Synthetic narcotics
  - Other dangerous non-narcotic drugs
- Drug possession
  - Opium and cocaine, and their derivatives (including morphine and heroin)
  - Marijuana
  - Synthetic narcotics
  - Other dangerous non-narcotic drugs

19. Gambling - total

- Bookmaking - horse and sports
- Number and lottery
- All other gambling

20. Offenses against family and children - nonviolent acts against family members. Includes neglect or abuse, nonpayment of child support or alimony.
21. Driving under the influence (DUI)
22. Liquor law violations - Includes illegal production, possession (e.g. underage) or sale of alcohol, open container, or public use laws. Does not include DUIs and drunkenness.
23. Drunkenness - i.e. public intoxication
24. Disorderly conduct
25. Vagrancy - includes begging, loitering (for adults only), homelessness, and being a "suspicious person."
26. All other offenses (other than traffic) - a catch-all category for any arrest that is not otherwise specified in this list. Does not include traffic offenses. Very wide variety of crimes are included - use caution when using!
27. Suspicion - "Arrested for no specific offense and released without formal charges being placed."

28. Curfew and loitering law violations - for minors only.
29. Runaways - for minors only.

2.1.3 What is an index (or part 1) crime?

One of the first (and seemingly only) thing that people tend to learn about UCR crime data is that it covers something called an "index crime."[4] Index crimes, sometimes written as Part 1 or Part I crimes, are the seven crimes originally chosen by the FBI to be included in their measure of crimes as these offenses were both considered serious and generally well-reported so would be a useful measure of crime. Index crimes are often broken down into property index crimes - burglary, theft, and motor vehicle theft (and arson now, though that's often not included and is reported less often by agencies) - and violent index crimes (murder, rape, robbery, and aggravated assault). The "index" is simply that all of the crimes are summed up into a total count of crimes (violent, property, or total) for that police agency.

The biggest problem with index crimes is that it is simply the sum of 8 (or 7 since arson data usually isn't included) crimes. Index crimes have a huge range in their seriousness - it includes, for example, both murder and theft - so summing up the crimes gives each crime equal weight. This is clearly wrong as 100 murders is more serious than 100 thefts. This is especially a problem as less serious crimes (theft mostly) are far more common than more serious crimes. In 2017, for example, there were 1.25 million violent index crimes in the United States. That same year had 5.5 million thefts. So using index crimes as your measure of crimes fails to account for the seriousness of crimes. Looking at total index crimes is, in effect, mostly just looking at theft. Looking at violent index crimes alone mostly measures aggravated assault. This is especially a problem because it hides trends in violent crimes. As an example, San Francisco, shown in Figure @ref(fig:sfThefts), has had a huge increase - about 50% - in index crimes in the last several years. When looking closer, that increase is driven almost entirely by the near doubling of theft since 2011. During the same years, index violent crimes have stayed fairly steady. So the city isn't getting more dangerous - at least not in terms of violent index crimes increasing - but it appears like it is due to just looking at total index crimes.

Many researchers divide index crimes into violent and nonviolent categories, which helps but even this isn't entirely sufficient. Take Chicago as an example. It is a city infamous for its large number of murders. But as a fraction of index crimes, Chicago has a rounding error worth of murders. Their 653 murders in 2017 is only 0.5% of total index crimes. For violent index crimes, murder made up 2.2% of crimes that year. As seen in Figure @ref(fig:chicagoMurder), in no year where data is available did murders account for more

[4] Index crimes are sometimes capitalized as "Index Crimes" though I've seen it written both ways. In this book I keep it lowercase as "index crimes."

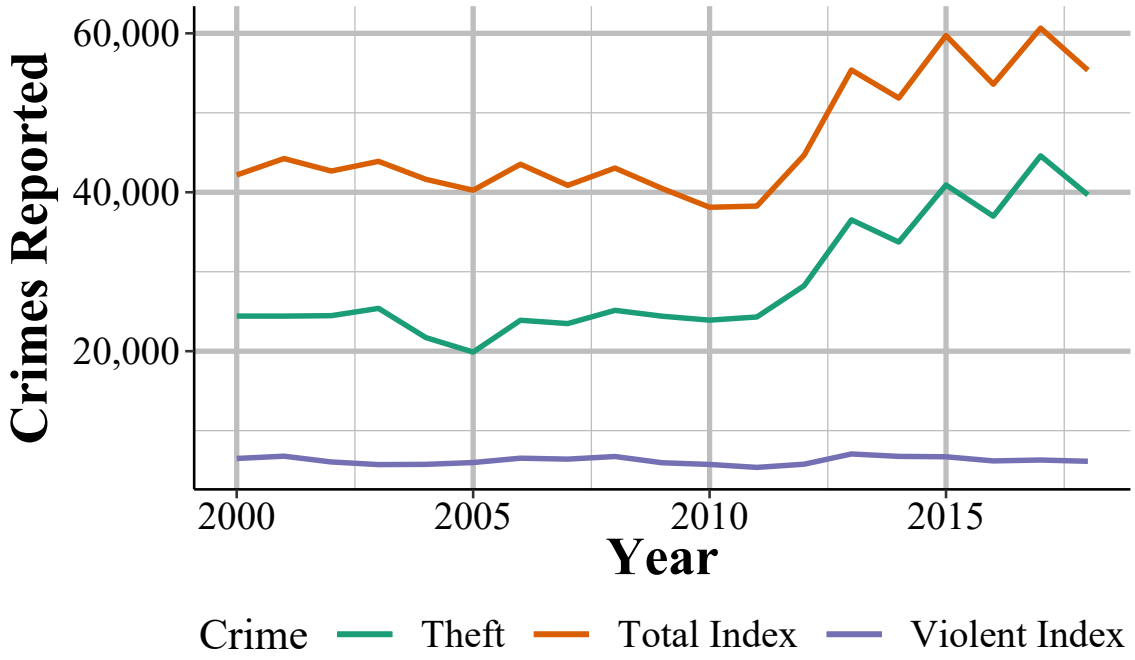


Figure 2.1: Thefts and total index crimes in San Francisco, 2000-2018.
(#fig:sfThefts)

than 3.5% of violent index crimes; and, while murders are increasing as a percent of violent index crimes they still account for no more than 2.5% in most recent years. What this means is that changes in murder are very difficult to detect. If Chicago had no murders this year, but a less serious crime (such as theft) increased slightly, we couldn't tell from looking at the number of index crimes, even from violent index crimes. As discussed in the below section, using this sample of crimes as the primary measure of crimes - and particularly of violent crimes - is also misleading as it excludes important - and highly common relative to index crimes - offenses, such as simple assault.

2.1.3.1 What is a violent crime?

An important consideration in using this data is defining what a "violent crime" is. One definition, and the one that I think makes the most sense, is that a violent crime is one that uses force or the threat of force. For example, if I punch you in the face, that is a violent crime. If I stab you, that is a violent crime. While clearly different in terms of severity, both incidents used force so I believe would be classified as a violent crime. The FBI, and most researchers, reporters, and advocates would disagree. Organizations ranging from the FBI itself to Pew Research Center and advocacy groups like the Vera Institute of Justice and the ACLU all define the first examine as a non-violent crime and the second as a violent crime. They do this for three main reasons.

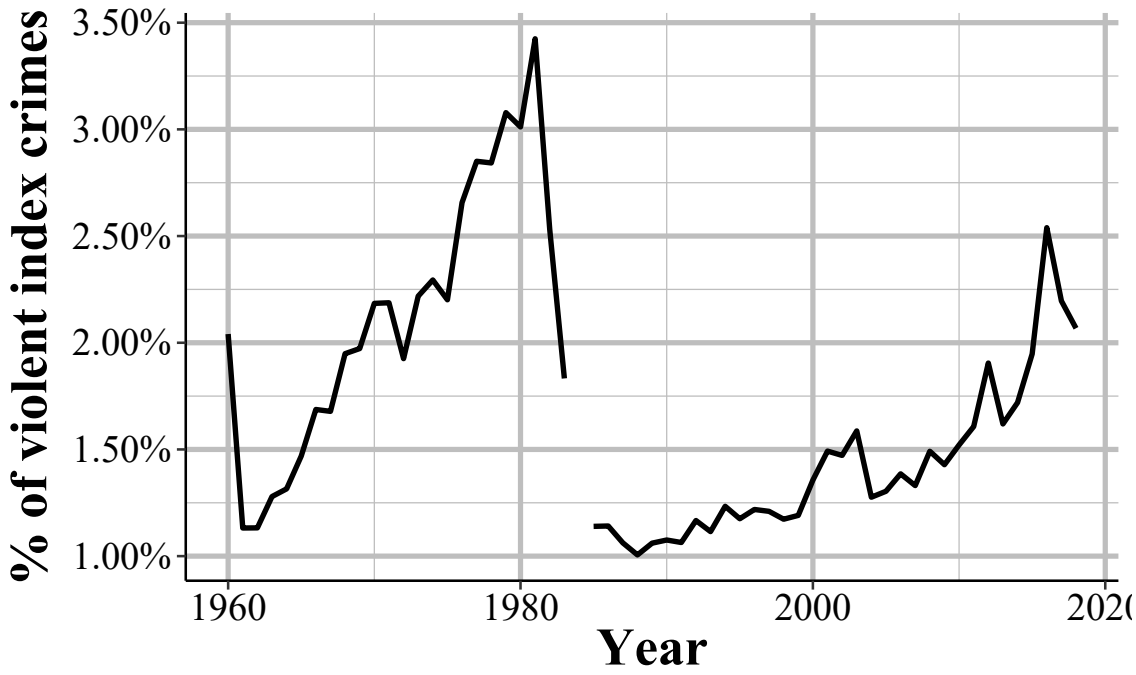


Figure 2.2: Murders in Chicago as a percent of violent index crimes, 1960-2018. (#fig:chicagoMurder)

The first reason is that simple assault is not an index crime, so they don't include it when measuring violent crime. It is almost a tautology in criminology that you use index crimes as the measure of crime since that's just what people do. As far as I'm aware, this is really the main reason why researchers justify using index crimes: because people in the past used it so that's just what is done now.[5] This strikes me as a particularly awful way of doing anything, more so since simple assault data has been available almost as long as index crime data.[6]

The second reason - and one that I think makes sense for reporters and advocates who are less familiar with the data, but should be unacceptable to researchers - is that people don't know that simple assault is included, or at least don't have easy access to it. Neither the FBI's annual report page not their official crime data tool website include simple assault since they only include index crimes. For people who rely only on these sources - and given that using the data itself requires at least some programming skills, I think this accounts for most users and certainly nearly all non-researchers - it is not possible to access simple assault crime data (arrest data does include simple assault on these sites) from these official sources.

[5]I've also seen the justification that aggravated assault is more serious since it uses a weapon, though as the UCR subcategory of aggravated assault without a weapon clearly shows, aggravated assault does not require use of a weapon.

[6]Simple assault is first available in 1964. Index crime data is available since 1960.

The final reason is that it benefits some people's goals to classify violent crime as only including index crimes. This is because simple assault is extremely common compared to violent index crimes - in many cities simple assault is more common than all violent index crimes put together - so excluding simple assault makes it seem like fewer arrests are violent than they are when including simple assault. For example, a number of articles have noted that marijuana arrests are more common than violent crime arrests (**?????**) or that violent crime arrests are only 5% of all arrests (**??**).[7] While true when considering only violent index crimes, including simple assault as a violent crime makes these statements false. Some organizations call the violent index crimes "serious violent crimes" which is an improvement but even this is a misnomer since simple assault can lead to more serious harm than aggravated assault. An assault becomes aggravated if using a weapon or there is the *potential* for serious harm, even if no harm actually occurs.[8]

As an example of this last point, Figure @ref(fig:simpleIndex) shows the number of violent index crimes and simple assaults each year from 1960 to 2018 in Houston, Texas (simple assault is not reported in UCR until 1964, which is why 1960-1963 show zero simple assaults). In every year where simple assault is reported, there are more simple assaults than aggravated assaults. Beginning in the late 1980s, there are also more simple assaults than total violent index crimes. Excluding simple assault from the being a violent crime greatly underestimates violence in the country.

2.2 Issues common across UCR datasets

In this section we'll discuss issues common to most or all of the UCR datasets. For some of these, we'll come back to the issues in more detail in the chapter for the datasets most affected by the problem.

2.2.1 Negative numbers

One of the most common questions people have about this data is why there are negative numbers, and if they are a mistake. Negative numbers are not a mistake. The UCR data is monthly so police agencies report the number of monthly crimes that are known to them - either reported to them or discovered by the police. In some cases the police will determine that a crime reported to them didn't actually occur - which they called an "unfounded crime."

[7] The FBI's report for arrests does include simple assault so the second reason people may not include it does not apply here.

[8] UCR data provides no information about the harm caused to victims. The new FBI dataset NIBRS actually does have a variable that includes harm to the victim so if you're interested in measuring harm (an understudied topic in criminology), that is the dataset to use.

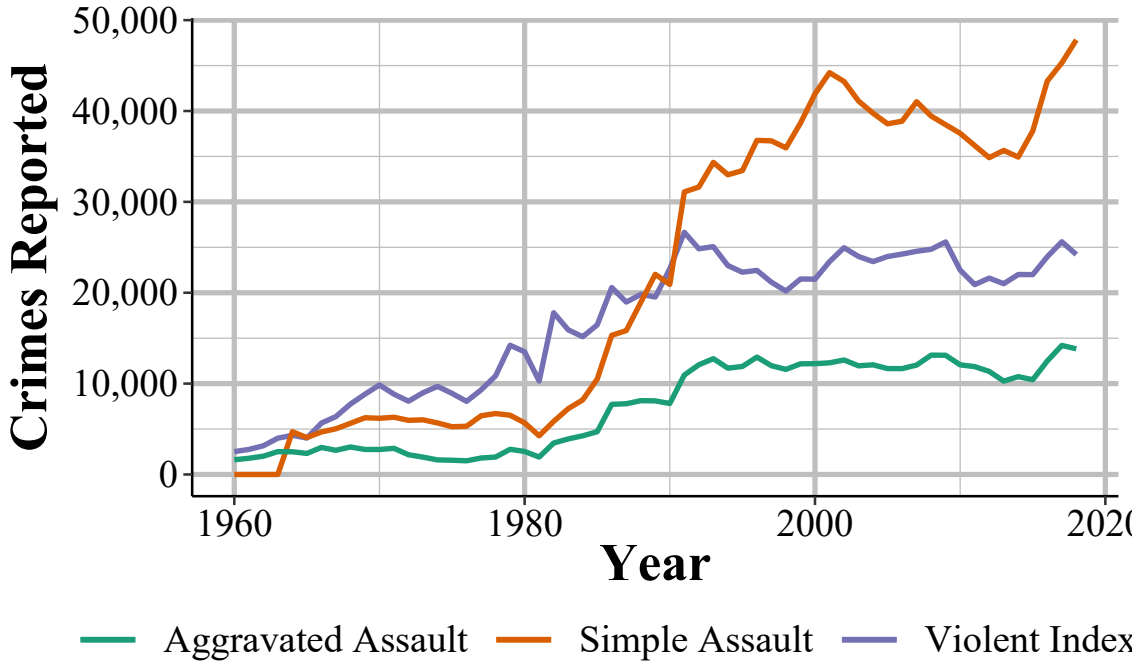


Figure 2.3: Reported crimes in Houston, Texas, from 1960 to 2018. Violent index crimes are aggravated assault, rape, robbery, and murder.
(#fig:simpleIndex)

An example that the FBI gives for this is a person reports their wallet stolen (a theft) and then later finds it, so a crime was initially reported but no crime actually occurred.

How this works when the police input the data is that an unfounded crime is reported both as an unfounded crime and as a negative actual crime - the negative is used to zero out the erroneous report of the actual crime. So the report would look like 1 actual theft (the crime being reported), -1 theft (the crime being discovered as not have happened), and 1 unfounded theft. If both incidents occurred in the same month then this would simply be a single unfounded theft occurring, with no actual theft - literally a value of $1 + -1 = 0$ thefts.

Negative values occur when the unfounding happens in a later month than the crime report. In the theft case, let's say the theft occurred in January and the discovery of the wallet happens in August. Assuming no other crimes occurred, January would have 1 theft, and August would have -1 thefts and 1 unfounded theft. There is no way of determining in which month (or even which year) an unfounded crime was initially reported in. When averaging over the long term, there shouldn't be any negative numbers as the actual and unfounded reports will cancel themselves out. However, when looking at monthly crimes - particularly for rare crimes - you'll still see negative numbers for this reason. Since crimes can be unfounded for reports in previous years, you can actually see entire year's crime counts be negative, though this is much rarer than monthly values.[9]

[9]From 1960-2018, there were 39 agency-years with a negative count of murders.

2.2.2 Agency population value

Each of the UCR datasets include a population variable that has the estimated population under the jurisdiction of that agency.[10] This variable is often used to create crime rates that control for population. In cases where jurisdiction overlaps, such as when a city has university police agencies or county sheriffs in counties where the cities in that county have their own police, UCR data assigns the population covered to the most local agency and zero population to the overlapping agency. So an agency's population is the number of people in that jurisdiction that isn't already covered by a different agency.

For example, the city of Los Angeles in California has nearly four million residents according to the US Census. There are multiple police agencies in the city, including the Los Angeles Police Department, the Los Angeles County Sheriff's Office, the California Highway Patrol that operates in the area, airport and port police, and university police departments. If each agency reported the number of people in their jurisdiction - which all overlap with each other - we would end up with a population far higher than LA's four million people. To prevent double-counting population when agency's jurisdictions overlap, the non-primary agency will report a population of 0, even though they still report crime data like normal. As an example, in 2018 the police department for California State University - Los Angeles reported 92 thefts and a population of 0. Those 92 thefts are not counted in the Los Angeles Police Department data, even though the department counts the population. To get complete crime counts in Los Angeles, you'd need to add up all police agencies within in the city; since the population value is 0 for non-LAPD agencies, both the population and the crime sum will be correct.

The UCR uses this method even when only parts of a jurisdiction overlaps. Los Angeles County Sheriff has a population of about one million people, far less than the actual county population (the number of residents, according to the Census) of about 10 million people. This is because the other nine million people are accounted for by other agencies, mainly the local police agencies in the cities that make up Los Angeles County.

The population value is the population who reside in that jurisdiction and does not count people who are in the area but don't live there, such as tourists or people who commute there for work. This means that using the population value to determine a rate can be misleading as some places have much higher numbers of non-residents in the area (e.g. Las Vegas, Washington D.C.) than others.

[10] Jurisdiction here refers to the boundaries of the local government, not any legal authority for where the officer can make arrests. For example, the Los Angeles Police Department's jurisdiction in this case refers to crimes that happen inside the city or are otherwise investigated by the LAPD - and are not primarily investigated by another agency.

2.2.3 Reporting is voluntary ... so some agencies don't (or report partially)

When an agency reports their data to the FBI, they do so voluntarily - there is no national requirement to report.[11] This means that there is inconsistency in which agencies report, how many months of the year they report for, and which variables they include in their data submissions.

In general, more agencies report their data every year and once an agency begins reporting data they tend to keep reporting. The UCR datasets are a collection of separate, though related, datasets and an agency can report to as many of these datasets as they want - an agency that reports to one dataset does not mean that they report to other datasets. Figure @ref(fig:agenciesReporting) shows the number of agencies that submitted at least one month of data to the Offenses Known and Clearances by Arrest data in the given year. For the first decade of available data under 8,000 agencies reported data and this grew to over 13,500 by the late 1970s before plateauing for about a decade. The number of agencies that reported their data actually declined in the 1990s, driven primarily by many Florida agencies temporarily dropping out, before growing steadily to nearly 17,000 agencies in 2010; from here it kept increasing but slower than before.

There are approximately 18,000 police agencies in the United States so recent data has reports from nearly all agencies, while older data has far fewer agencies reporting. When trying to estimate to larger geographies, such as state or national-level, later years will be more accurate as you're missing less data. For earlier data, however, you're dealing with a smaller share of agencies meaning that you have a large amount of missing data and a less representative sample.

Figure @ref(fig:bigAgenciesReporting) repeats the above figure but now including only agencies with 100,000 people or more in their jurisdiction. While these agencies have a far more linear trend than all agencies, the basic lesson is the same: recent data has most agencies reporting; old data excludes many agencies.

This voluntariness extends beyond whether they report or not, but into which variables they report. While in practice most agencies report every crime when they report any, they do have the choice to report only a subset of offenses. This is especially true for subsets of larger categories - such as gun assaults, a subset of aggravated assaults, or marijuana possession arrests which is a subset of drug possession arrests. As an example, Figure @ref(fig:nycGunAssaults) shows the annual number of aggravated assaults with a gun in New York City. In 2003 the New York Police Department stopped reporting this category of offense, resuming only in 2013. They continued to report the broader aggravated assault

[11] Some states do mandate that their agencies report, but this is not always followed.

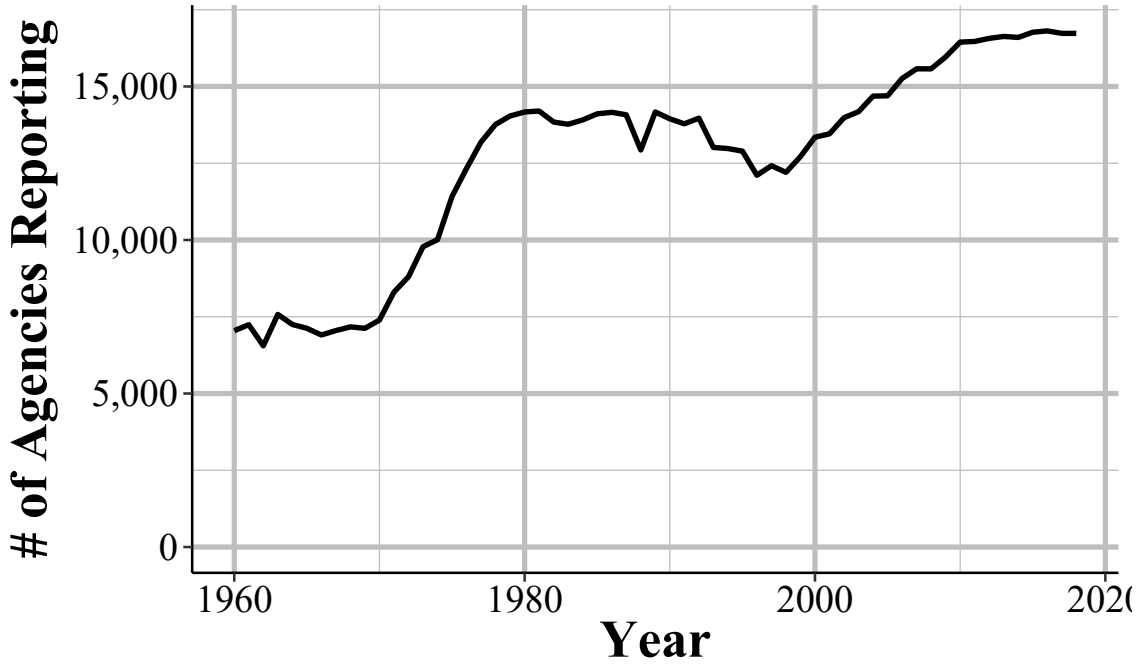


Figure 2.4: The annual number of agencies reporting to the Offenses Known and Clearances by Arrest dataset. Reporting is based on the agency reporting at least one month of data in that year.
(#fig:agenciesReporting)

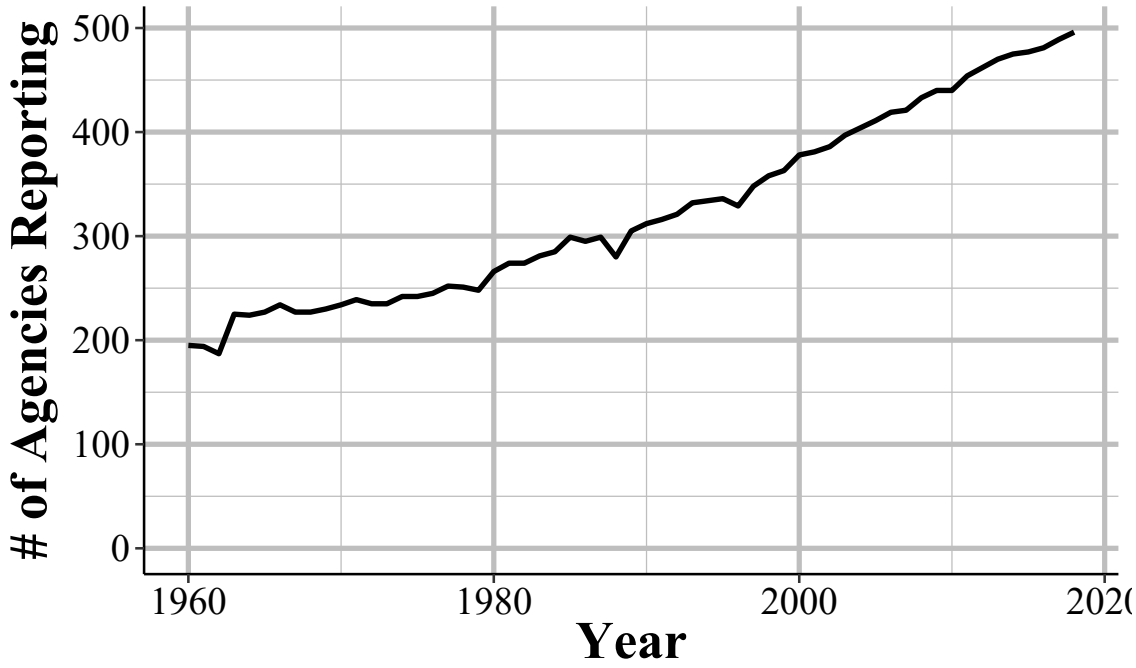


Figure 2.5: The annual number of agencies with a population of 100,000 or higher reporting to the Offenses Known and Clearances by Arrest dataset. Reporting is based on the agency reporting at least one month of data in that year.
(#fig:bigAgenciesReporting)

category, but not any of the subsections of aggravated assaults which say which weapon was used during the assault.

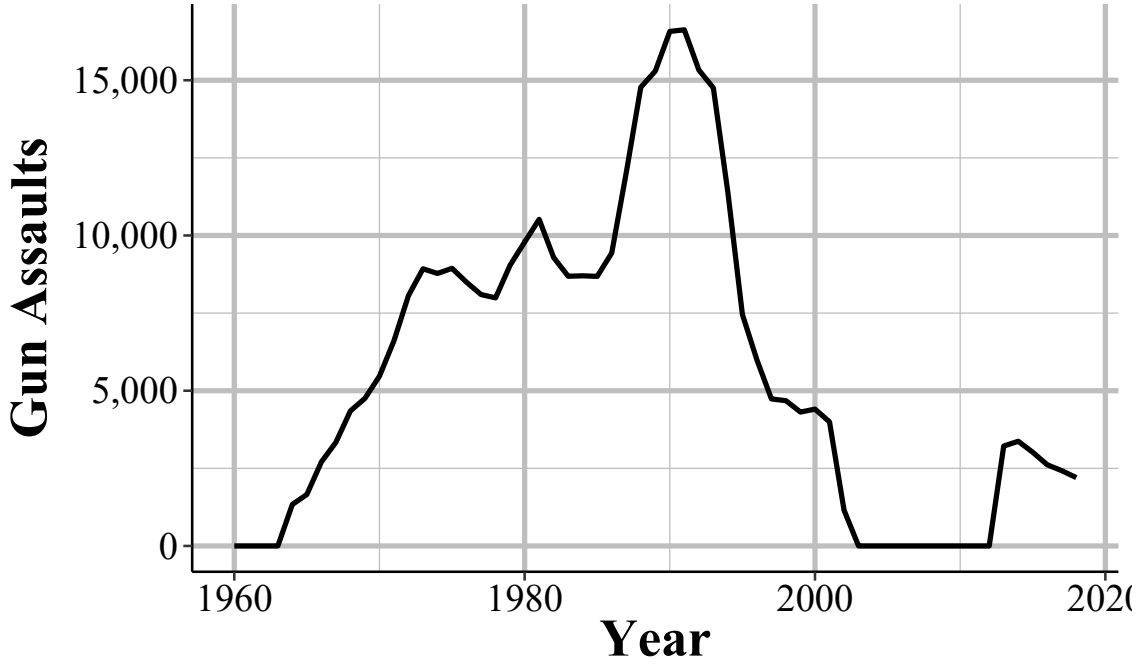


Figure 2.6: Monthly reports of gun assaults in New York City, 1960-2018. (#fig:nycGunAssaults)

Given that agencies can join or drop out of the UCR program at will, and report only partial data, it is highly important to carefully examine your data to make sure that there are no issues caused by this.

Even when an agency reports UCR data, and even when they report every crime category, they can report fewer than 12 months of data. In some cases they simply report all of their data in December, or report quarterly or semi-annually so some months have zero crimes reported while others count multiple months in that month's data. One example of this is New York City, shown in Figure @ref(fig:nycMurderMonthly), in the early-2000s to the mid-2010s where they began reporting data quarterly instead of monthly.

When you sum up each month into an annual count, as shown in Figure @ref(fig:nycMurderYearly), the problem disappears since the zero months are accounted for in the months that have the quarterly data. If you're using monthly data and only examine the data at the annual level, you'll fall into the trap of having incorrect data that is hidden due to the level of aggregation examined. While cases like NYC are obvious when viewed monthly, for people that are including thousands of agencies in their data, it is unfeasible to look at each agency for each crime included. This can introduce errors as the best way to examine the data is manually viewing graphs and the automated method, looking for outliers through some kind of comparison to expected values, can be incorrect.

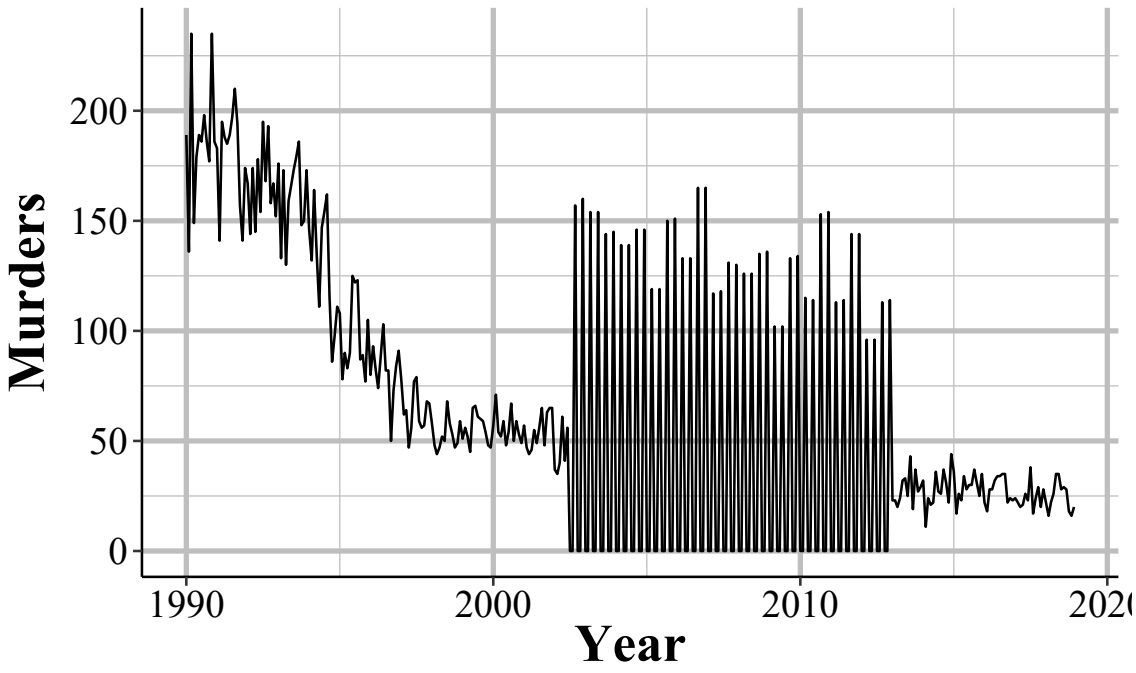


Figure 2.7: Monthly murders in New York City, 1990-2018. During the 2000s, the police department began reporting quarterly instead of monthly and then resumed monthly reporting.
(#fig:nycMurderMonthly)

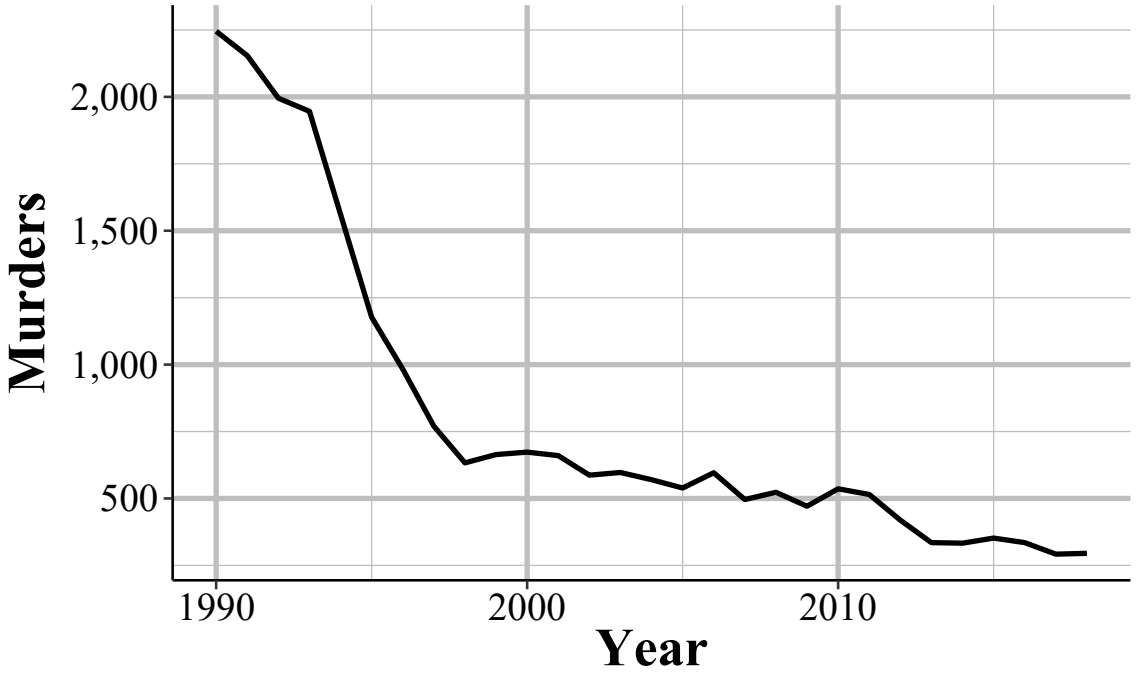


Figure 2.8: Annual murders in New York City, 1990-2018.
(#fig:nycMurderYearly)

In other cases when agencies report fewer than 12 months of the year, they simply report partial data and as a result undercount crimes. Figure @ref(fig:miamiDadeMurderAnnual) shows annual murders in Miami-Dade, Florida and has three years of this issue occurring. The first two years with this issue are the two where zero murders are reported - this is because the agency didn't report any months of data. The final year is in 2018, the last year of data in this graph, where it looks like murder suddenly dropped significantly. That's just because Miami-Dade only reported through June, so they're missing half of 2018.

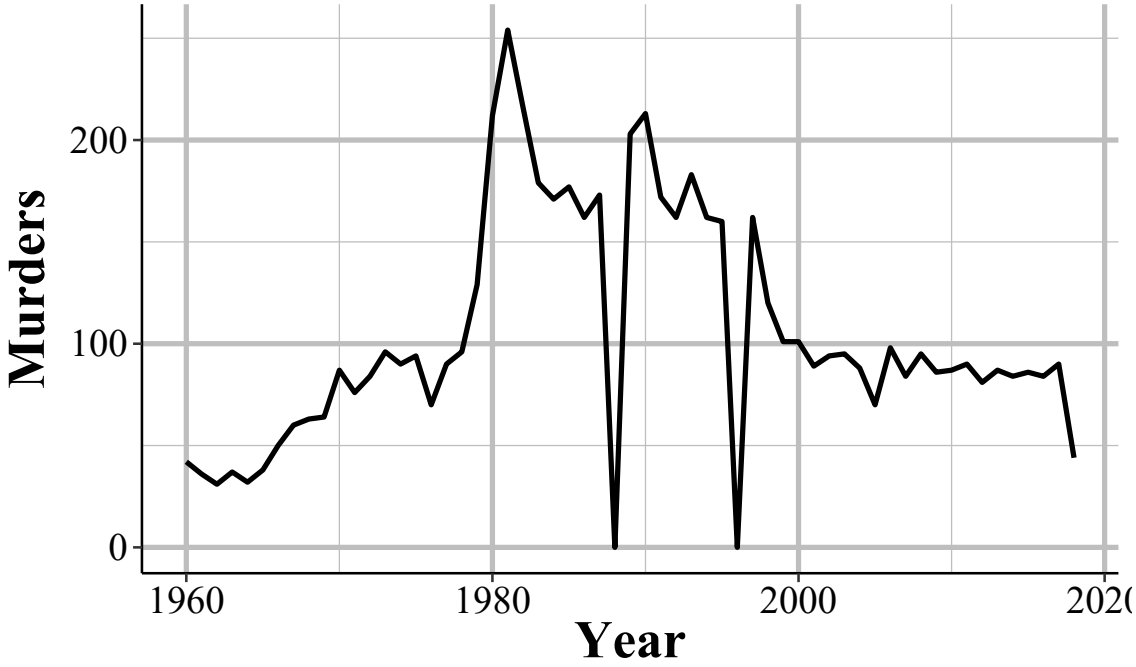


Figure 2.9: Annual murders in Miami-Dade, Florida, 1960-2018. (#fig:miamiDadeMurderAnnual)

2.2.4 Zero crimes vs no reports

When an agency does not report, we see it in the data as reporting zero crimes, not reporting NA or any indicator that they did not report. In cases where the agency says they didn't report that month we can be fairly sure (not entirely since that variable isn't always accurate) that the zero crimes reported are simply that the agency didn't report. In cases where the agency says they report that month but report zero crimes, we can't be sure if that's a true no crimes reported to the agency or the agency not reporting to the UCR. As agencies can report some crimes but not others in a given month and still be considered reporting that month, just saying they reported doesn't mean that the zero is a true zero.

In some cases it is easy to see when a zero crimes reported is actually the agency not reporting. As Figure @ref(fig:nycGunAssaults) shows with New York City gun assaults,

there is a massive and sustained dropoff to zero crimes and then a sudden return years later. Obviously, going from hundreds of crimes to zero crimes is not a matter of crimes not occurring anymore, it is a matter of the agency not reporting - and New York City did report other crimes these years so in the data it says that they reported every month. So in agencies which tend to report many crimes - and many here can be a few as 10 a year since going from 10 to 0 is a big drop - a sudden report of zero crimes is probably just non-reporting.

Differentiating zero crimes and no reports becomes tricky in agencies that tend to report few crimes, which most small towns do. As an example, Figure @ref(fig:danvilleRape) shows the annual reports of rape in Danville, California, a city of approximately 45,000 people. The city reports on average 2.8 rapes per year and in five years reported zero rapes. In cases like this it's not clear whether we should consider those zero years as true zeros (that no one was raped or reported their rape to the police) or whether the agency simply didn't report rape data that year.

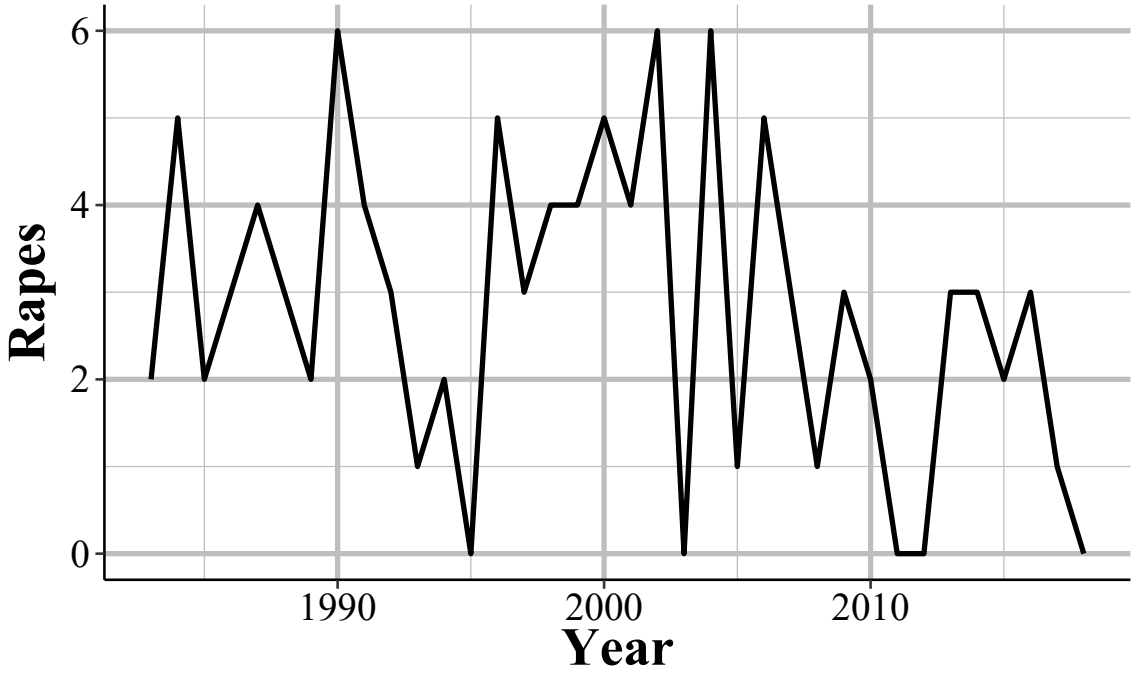


Figure 2.10: Annual rapes reported in Danville, CA, 1960-2018. (#fig:danvilleRape)

2.3 A summary of each UCR dataset

The UCR collection of data can be roughly summarized into two groups: crime data and arrest data. While there are several datasets included in the UCR data collection, they are all extensions of one of the above groups. For arrest data, you have information about who

(by race and by age-gender, but not by race-gender or race-age other than within race you know if the arrestee is an adult or a juvenile) was arrested for each crime. For crime data, you have monthly counts of a small number of crimes (many fewer than crimes covered in the arrest data) and then more specialized data on a subset of these crimes - information on homicides, hate crimes, assaults or killings of police officers, and stolen property.

Each of these datasets will have its own chapter in this book where we discuss the data thoroughly. Here is a very brief summary of each dataset which will help you know which one to use for your research. I still recommend reading that data's chapter since it covers important caveats and uses (or misuses) of the data that won't be covered below.

2.3.1 Offenses Known and Clearances by Arrest

The Offenses Known and Clearances by Arrest dataset - often called Return A, "Offenses Known" or, less commonly, OKCA - is the oldest and most commonly used dataset and measures crimes reported to the police. For this reason it is used as the main measure of crime in the United States, and I tend to call it the "crimes dataset." This data includes the monthly number of crimes reported to the police or otherwise known to the police (e.g. discovered while on patrol) for a small selection of crimes, as well as the number of crimes cleared by arrest or by "exceptional means" (a relatively flawed and manipulable measure of whether the case is "solved"). It also covers the number reported but found by police to have not occurred. Since this data has monthly agency-level crime information it is often used to measure crime trends between police agencies and over time. The data uses something called a Hierarchy Rule which means that in incidents with multiple crimes, only the most serious is recorded - though in practice this affects only a small percent of cases, and primarily affects property crimes.

2.3.2 Arrests by Age, Sex, and Race

The Arrests by Age, Sex, and Race dataset - often called ASR or the "arrests data" - includes the monthly number of arrests for a variety of crimes and, unlike the crime data, breaks down this data by age and gender. This data includes a broader number of crime categories than the crime dataset (the Offenses Known and Clearances by Arrest data) though is less detailed on violent crimes since it does not breakdown aggravated assault or robberies by weapon type as the Offenses Known data does. For each crime it says the number of arrests for each gender-age group with younger ages (15-24) showing the arrestee's age to the year (e.g. age 16) and other ages grouping years together (e.g. age 25-29, 30-34, "under 10"). It also breaks down arrests by race-age by including the number of arrestees of each race (American Indian, Asian, Black, and White are the only included races) and if the arrestee is a juvenile (<18

years old) or an adult. The data does technically include a breakdown by ethnicity-age (e.g. juvenile-Hispanic, juvenile-non-Hispanic) but almost no agencies report this data (for many years zero agencies report ethnicity) so in practice the data does not include ethnicity. As the data includes counts of arrestees, people who are arrested multiple times are included in the data multiple times - it is not a measure of unique arrestees.

2.3.3 Law Enforcement Officers Killed and Assaulted (LEOKA)

The Law Enforcement Officers Killed and Assaulted data, often called just by its acronym LEOKA, has two main purposes.[12] First, it provides counts of employees employed by each agency - broken down by if they are civilian employees or sworn officers, and also broken down by gender. And second, it measures how many officers were assaulted or killed (including officers who die accidentally such as in a car crash) in a given month. The assault data is also broken down into shift type (e.g. alone, with a partner, on foot, in a car, etc.), the offender's weapon, and type of call they are responding to (e.g. robbery, disturbance, traffic stop). The killed data simply says how many officers are killed feloniously (i.e. murdered) or died accidentally (e.g. car crash) in a given month. The employee information is at the year-level so you know, for example, how many male police officers were employed in a given year at an agency, but don't know any more than that such as if the number employed changed over the year. This dataset is commonly used as a measure of police employees and is a generally reliable measure of how many police are employed by a police agency. The second part of this data, measuring assaults and deaths, is more flawed with missing data issues and data error issues (e.g. more officers killed than employed in an agency).

2.3.4 Supplementary Homicide Reports (SHR)

The Supplementary Homicide Reports dataset - often abbreviated to SHR - is the most detailed of the UCR datasets and provides information about the circumstances and participants (victim and offender demographics and relationship status) for homicides.[13] For each homicide incident it tells you the age, gender, race, and ethnicity of each victim and offender as well as the relationship between the first victim and each of the offenders (but not the other victims in cases where there are multiple victims). It also tells you the weapon used by each offender and the circumstance of the killing, such as a "lovers triangle" or a gang-related murder. As with other UCR data, it also tells you the agency it occurred in and the month and year when the crime happened.

[12] This data is also sometimes called the "Police Employees" dataset.

[13] If you're familiar with the National Incident-Based Reporting System (NIBRS) data that is replacing UCR, this dataset is the closest UCR data to it, though it is still less detailed than NIBRS data.

2.3.5 Hate Crime Data

This dataset covers crimes that are reported to the police and judged by the police 'to be motivated by hate. More specifically, they are, first, crimes which were, second, motivated - at least in part - by bias towards a certain person or group of people because of characteristics about them such as race, sexual orientation, or religion. The first part is key, they must be crimes - and really must be the selection of crimes that the FBI collects for this dataset. Biased actions that don't meet the standard of a crime, or are of a crime not included in this data, are not considered hate crimes. For example, if someone yells at a Black person and uses racial slurs against them, it is clearly a racist action. For it to be included in this data, however, it would have to extend to a threat since "intimidation" is a crime included in this data but lesser actions such as simply insulting someone is not included. For the second part, the bias motivation, it must be against a group that the FBI includes in this data. For example, when this data collection began crimes against transgender people were not counted so if a transgender person was assaulted or killed because they were transgender, this is not a hate crime recorded in the data (though it would have counted in the "Anti-Lesbian, Gay, Bisexual, Or Transgender, Mixed Group (LGBT)" bias motivation which was always reported).[14]

So this data is really a narrower measure of hate crimes than it might seem. In practice it is (some) crimes motivated by (some) kinds of hate that are reported to the police. It is also the most under-reported UCR dataset with most agencies not reporting any hate crimes to the FBI. This leads to huge gaps in the data with some states having extremely few agencies reporting crime - see, for example Figure @ref(fig:hateCrimes) for state-level hate crimes in 2018 - agencies reporting some bias motivations but not others, agencies reporting some years but not others. While these problems exist for all of the UCR datasets, it is most severe in this data. This problem is exacerbated by hate crimes being rare even in agencies that report them - with such rare events, even minor changes in which agencies report or which types of offenses they include can have large effects.

2.3.6 Property Stolen and Recovered (Supplement to Return A)

The Property Stolen and Recovered data - sometimes called the Supplement to Return A (Return A being another name for the Offenses Known and Clearances by Arrest dataset, the "crime" dataset) - provides monthly information about property-related offenses (theft, motor vehicle theft, robbery, and burglary), including the location of the offense (in broad categories like "gas station" or "residence"), what was stolen (e.g. clothing, livestock, firearms), and how much the stolen items were worth.[15] The "recovered" part of this dataset covers the

[14]The first year where transgender as a group was a considered a bias motivation was in 2013.

[15]It also includes the value of items stolen during rapes and murders, if anything was stolen.

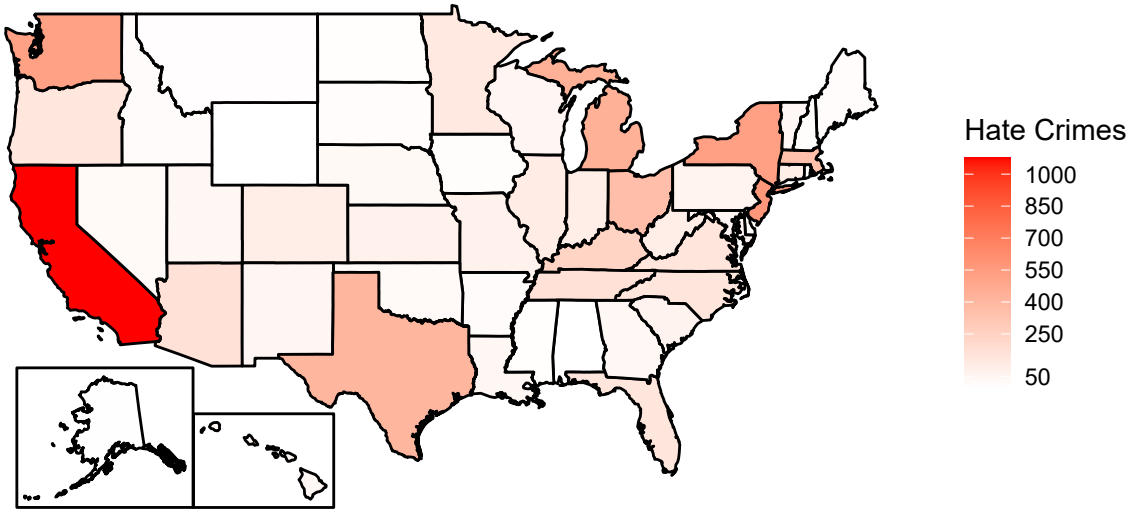


Figure 2.11: Total reported hate crimes by state, 2018.
(#fig:hateCrimes)

type and value of property recovered so you can use this, along with the type and value of property stolen, to determine what percent and type of items the police managed to recover. Like other UCR datasets this is at the agency-month level so you can, for example, learn how often burglaries occur at the victim's home during the day, and if that rate changes over the year or differs across agencies. The data, however, provides no information about the offender or the victim (other than if the victim was an individual or a commercial business [based on the location of the incident - "bank", "gas station", etc.]). The value of the property stolen is primarily based on the victim's estimate of how much the item is worth (items that are decreased in value once used - such as cars - are supposed to be valued at the current market rate, but the data provides no indication of when it uses the current market rate or the victim's estimate) so it should be used as a very rough estimate of value.

2.3.7 Arson

The arson dataset provides monthly counts at the police agency-level for arsons that occur, and includes a breakdown of arsons by the type of arson (e.g. arson of a person's home, arson of a vehicle, arson of a community/public building) and the estimated value of the damage caused by the arson. This data includes all arsons reported to the police or otherwise known to the police (e.g. discovered while on patrol) and also has a count of arsons that lead to an arrest (of at least one person who committed the arson) and reports that turned out to not be arsons (such as if an investigation found that the fire was started accidentally).

For each type of arson it includes the number of arsons where the structure was uninhabited or otherwise not in use, so you can estimate the percent of arsons of buildings which had

the potential to harm people. This measure is for structures where people normally did not inhabit the structure - such as a vacant building where no one lives. A home where no one is home *at the time of the arson* does not count as an uninhabited building.

In cases where the arson led to a death, that death would be recorded as a murder on the Offenses Known and Clearances by Arrest dataset - but not indicated anywhere on this dataset. If an individual who responds to the arson dies because of it, such as a police officer or a firefighter, this is not considered a homicide (though the officer death is still included in the Law Enforcement Officers Killed and Assaulted data) unless the arsonist intended to cause their deaths. Even though the UCR uses the Hierarchy Rule, where only the most serious offense that occurs is recorded, all arsons are reported - arson is exempt from the Hierarchy Rule.

2.4 How to identify a particular agency (ORI codes)

In the UCR and other FBI data sets, agencies are identified using **OR**iginating Agency **I**dentifiers or an ORI. An ORI is a unique ID code used to identify an agency.[16] If we used the agency's name we'd end up with some duplicates since there can be multiple agencies in the country (and in a state, those this is very rare) with the same name. For example, if you looked for the Philadelphia Police Department using the agency name, you'd find both the "Philadelphia Police Department" in Pennsylvania and the one in Mississippi. Each ORI is a 7-digit value starting with the state abbreviation (for some reason the FBI incorrectly puts the abbreviation for Nebraska as NB instead of NE) followed by 5 numbers.[17] When dealing with specific agencies, make sure to use the ORI rather than the agency name to avoid any mistakes.

For an easy way to find the ORI number of an agency, use this page on my site. Type an agency name or an ORI code into the search section and it will return everything that is a match.

2.5 The data as you get it from the FBI

We'll finish this overview of the UCR data by briefly talking about format of the data that is released by the FBI, before the processing done by myself or NACJD that converts the data to a type that software like R or Stata or Excel can understand. The FBI releases their data

[16] I will refer to this an "ORI", "ORI code", and "ORI number", all of which mean the same thing.

[17] In the NIBRS data (another FBI data set) the ORI uses a 9-digit code - expanding the 5 numbers to 7 numbers.

as fixed-width ASCII files which are basically just an Excel file but with all of the columns squished together. As an example, Figure @ref(fig:ascii) shows what the data looks like as you receive it from the FBI for the Offenses Known and Clearances by Arrest dataset for 1960, the first year with data available. In the figure, it seems like there are multiple rows but that's just because the software that I opened the file in isn't wide enough - in reality what is shown is a single row that is extremely wide because there are over 1,500 columns in this data. If you scroll down enough you'll see the next row, but that isn't shown in the current image. What is shown is a single row with a ton of columns all pushed up next to each other. Since all of the columns are squished together (the gaps are just blank spaces because the value there is a space, but that doesn't mean there is a in the data. Spaces are possible values in the data and are meaningful), you need some way to figure out which parts of the data belong in which column.

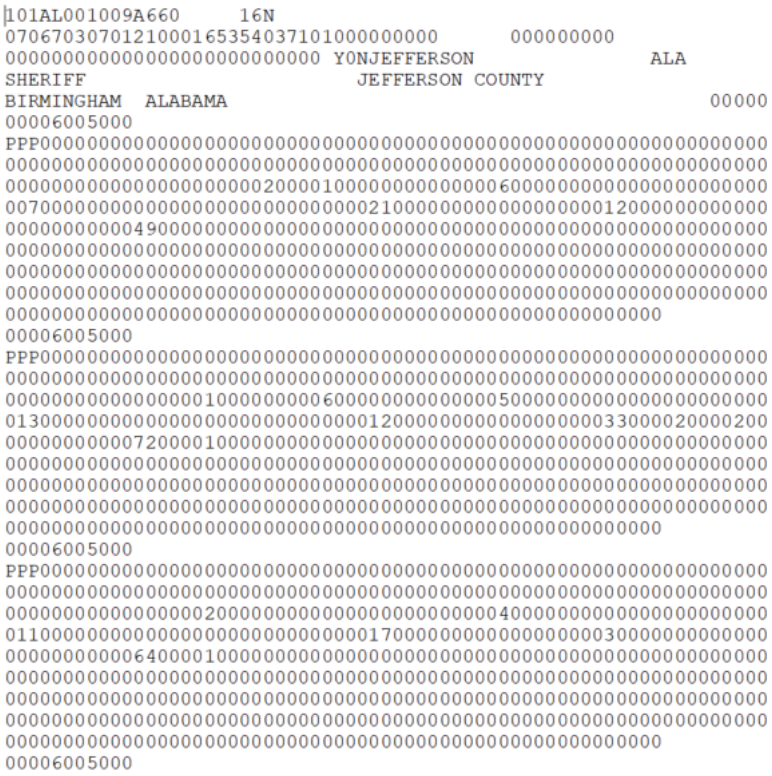


Figure 2.12: Fixed-width ASCII file for the 1960 Offenses Known and Clearances by Arrest dataset.
(#fig:ascii)

The "fixed-width" part of the file type is how this works (the ASCII part basically means it's a text file). Each row is the same width - literally the same number of characters, including blank spaces. So you must tell the software you are using to process this file - by literally writing code in something called a "setup file" but is basically just instructions for whatever software you use (R, SPSS, Stata, SAS can all do this) - which characters are certain columns.

For example, in this data the first character says which type of UCR data it is (1 means the Offenses Known and Clearances by Arrest data) and the next two characters (in the setup file written as 2-3 since it is characters 2 through 3 [inclusive]) are the state number (01 is the state code for Alabama). So we can read this row as the first column indicating it is an Offenses Known data, the second column indicating that it is for the state of Alabama, and so on for each of the remaining columns. To read in this data you'll need a setup file that covers every column in the data (some software, like R, can handle just reading in the specific columns you want and don't need to include every column in the setup file).

The second important thing to know about reading in a fixed-width ASCII file is something called a "value label."[18] For example, in the above image we saw the characters 2-3 is the state and in the row we have the value "01" which means that the state is "Alabama." Since this type of data is trying to be as small as efficient as possible, it often replaces longer values with shorter one and provides a translation for the software to use to convert it to the proper value when reading it. "Alabama" is more characters than "01" so it saves space to say "01" and just replace that with "Alabama" later on. So "01" would be the "value" and "Alabama" would be the "label" that it changes to once read.

Fixed-width ASCII files may seem awful to you reading it today, and it is awful to use. But it appears to be an efficient way to store data back many decades ago when data releases began but now is extremely inefficient - in terms of speed, file size, ease of use - compared to modern software so I'm not sure why they *still* release data in this format. But they do, and even the more *modern* (if starting in 1991, before I was born, is modern!) NIBRS data comes in this format. For you, however, the important part to understand is not how exactly to read this type of data, but to understand that people who made this data publicly available (such as myself and the team at NACJD) must make this conversion process.[19] **This conversion process, from fixed-width ASCII to a useful format is the most dangerous step taken in using this data - and one that is nearly entirely unseen by researchers.**

Every line of code you write (or, for SPSS users, click you make) invites the possibility of making a mistake.[20] The FBI does not provide a setup file with the fixed-width ASCII data so to read in this data you need to make it yourself. Since some UCR data are massive, this involves assigning the column width for thousands of columns and the value labels for hundreds of different value labels.[21] A typo anywhere could have potentially far-reaching

[18]For most fixed-width ASCII files there are also missing values where it'll have placeholder value such as -8 and the setup file will instruct the software to convert that to NA. UCR data, however, does not have this and does not indicate when values are missing in this manner.

[19]For those interested in reading in this type of data, please see my R package asciiSetupReader.

[20]Even highly experienced programmers who are doing something like can make mistakes. For example, if you type out "2+2" 100 times - something extremely simple that anyone can do - how often will you mistype a character and get a wrong result? I'd guess that at least once you'd make a mistake.

[21]With the exception of the arrest data and some value label changes in hate crimes and homicide data,

consequences, so this is a crucial weak point in the data cleaning process - and one in which I have not seen anything written about before. While I have been diligent in checking the setup files and my code to seek out any issues - and I know that NACJD has a robust checking process for their own work - that doesn't mean our work is perfect.[22] Even with perfection in processing the raw data to useful files, decisions we make (e.g. what level to aggregate to, what is an outlier) can affect both what type of questions you can ask when using this data, and how well you can answer them.

the setup files remain consistent so a single file will work for all years for a given dataset. You do not need to make a setup file for each year.

[22]For evidence of this, please see any of the openICPSR pages for my detail as they detail changes I've made in the data such as decisions on what level to aggregate to and mistakes that I made and later found and fixed.

Chapter 3

Offenses Known and Clearances by Arrest

The Offenses Known and Clearances by Arrest dataset - often called Return A, "Offenses Known" or, less commonly, OKCA - is the oldest and most commonly used dataset and measures crimes reported to the police. For this reason it is used as the main measure of crime in the United States, and I tend to call it the "crimes dataset." This data includes the monthly number of crimes reported to the police or otherwise known to the police (e.g. discovered while on patrol) for a small selection of crimes, as well as the number of crimes cleared by arrest or by "exceptional means" (a relatively flawed and manipulable measure of whether the case is "solved"). It also covers the number reported but found by police to have not occurred. Since this data has monthly agency-level crime information it is often used to measure crime trends between police agencies and over time. The data uses something called a Hierarchy Rule which means that in incidents with multiple crimes, only the most serious is recorded - though in practice this affects only a small percent of cases, and primarily affects property crimes.

3.1 Which crimes are included?

This data set contains information on the number of "Index Crimes" (sometimes called Part I crimes) reported to each agency. These index crimes are a collection of eight crimes that, for historical reasons based largely by perceived importance and reliable of their reporting in the 1920's when the UCR program was first developed, are used as the primary measure of crime today. Other data sets in the UCR, such as the Arrests by Age, Sex, and Race data and the Hate Crime data have more crimes reported.

The crimes are, in order by the Hierarchy Rule (which we'll discuss next):

1. Homicide
    - Murder and non-negligent manslaughter
    - Manslaughter by negligence
2. Rape
    - Rape
    - Attempted rape
3. Robbery
    - With a firearm
    - With a knife of cutting instrument
    - With a dangerous weapon not otherwise specified
    - Unarmed - using hands, fists, feet, etc.
4. Aggravated Assault (assault with a weapon or with the intent to cause serious bodily injury)
    - With a firearm
    - With a knife of cutting instrument
    - With a dangerous weapon not otherwise specified
    - Unarmed - using hands, fists, feet, etc.
5. Burglary
    - With forcible entry
    - Without forcible entry
    - Attempted burglary with forcible entry

6. Theft (other than of a motor vehicle)

7. Motor Vehicle Theft

    - Cars
    - Trucks and buses
    - Other vehicles

8. Arson

9. Simple Assault

Arson is considered an index crime but is not reported in this data - you need to use the separate Arson data set of the UCR to get access to arson counts. See Chapter @ref(arsonChapter) for an overview of the Arson data. The ninth crime on that list, simple assault, is not considered an index crime but is nevertheless included in this data.

Each of the crimes in the list above, and their subcategories, are included in the UCR data. In most news and academic articles, however, you'll see them reported as the total number of index crimes, summing up categories 1-7 and reporting that as "crime." These index crimes are often divided into violent index crimes - murder, rape, robbery, and aggravated assault - and property index crimes - burglary, theft, motor vehicle theft. For more on index crimes, and the drawbacks of using them, please see Section @ref(indexCrimes).

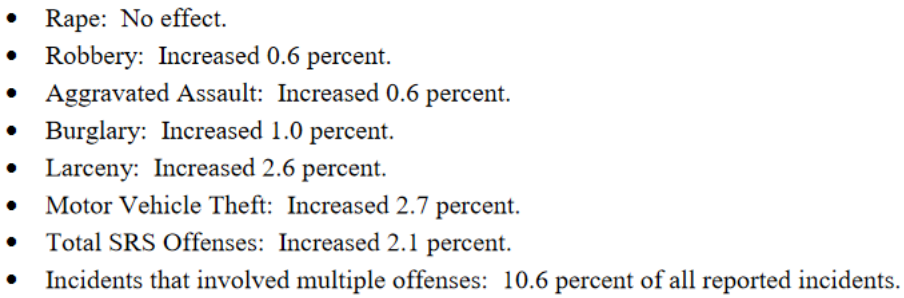
3.1.1 Hierarchy Rule

This dataset uses what is called the Hierarchy Rule where only the most serious crime in an incident is reported (except for motor vehicle theft and arson, which are always included). For example if there is an incident where the victim is robbed and then murdered, only the murder is counted as it is considered more serious than the robbery. That the data uses the Hierarchy Rule is an oft-cited (by academics, reporters, random people on Twitter) criticism of the data that is, in my opinion, overblown.

In practice, the Hierarchy Rule has only modest effects on the data, undercounting few crimes. Though the Hierarchy Rule does mean this data is an undercount, data from other sources indicate that it isn't much of an under count. The FBI's other data set, the National Incident-Based Reporting System (NIBRS) contains every crime that occurs in an incident (i.e. it doesn't use the Hierarchy Rule). Using this we can measure how many crimes the

Hierarchy Rule excludes (Most major cities do not report to NIBRS so what we find in NIBRS may not apply to them). In over 90% of incidents, only one crime is committed. Additionally, when people talk about "crime" they usually mean murder which, while incomplete to discuss crime, means the UCR data here is accurate on that measure.

The FBI also released a report available here in 2015 that directly examined this issue by taking NIBRS data from 2014 and examined how NIBRS data (which includes all crimes) compares to using the Hierarchy Rule and keeping only the most serious crime. Figure @ref(fig:fbiHierarchy) is a screenshot from their report showing the percent increases in crimes when including all crimes in an incident relative to following the Hierarchy Rule. They find that 10.6% of incidents have multiple crimes occurring, which is similar to other years of data that I have examined myself. For violent crime, murder and rape have no change; for the remaining violent crimes - robbery and aggravated assault - crimes increased by 0.6%.[1] Burglary increased by 1% and the largest increases came from theft and motor vehicle theft, increasing by 2.6% and 2.7%, respectively. Curiously motor vehicle theft increased even though the FBI's documentation for this data says that it is exempt from the Hierarchy Rule and should always be reported. This suggests either non-compliance or that the manual is incorrect.



- Rape: No effect.
- Robbery: Increased 0.6 percent.
- Aggravated Assault: Increased 0.6 percent.
- Burglary: Increased 1.0 percent.
- Larceny: Increased 2.6 percent.
- Motor Vehicle Theft: Increased 2.7 percent.
- Total SRS Offenses: Increased 2.1 percent.
- Incidents that involved multiple offenses: 10.6 percent of all reported incidents.

Figure 3.1: The FBI's findings of how crime reporting changes when using the Hierarchy Rule using NIBRS 2014 data.
(#fig:fbiHierarchy)

So using the Hierarchy Rule does undercount crime, but this is a small undercounting and is primarily led by property crime. Violent crime is only slightly undercounted. Please keep in mind that this is for crimes that the police record and is unaffected by outside decisions like what the district attorney charges or what the defendant is ultimately convicted of.

[1] Murder is not shown in this figure since murder is always reported so cannot change.

3.2 Agencies reporting

Figure @ref(fig:offensesAgenciesReporting) shows the annual number of police agencies that reported at least one month that year. With data starting in 1960, there were a little under 7,500 agencies reporting a year until 1970 when the number of agencies began increasing. This continued until the late 1970s when about 14,000 agencies reported, and this remained steady for over a decade before declining in the 1990s until about 12,500 in the latter half of the decade. Then the number of agencies reporting increased steadily until about 16,500 agencies reported in 2010. The number of agencies has slowly increased since then, adding a few hundred agencies from 2010 to 2018. While this data is available through 2019, this graph only shows data through 2018. As this graph shows agencies that report at least one month, it overcounts reporting agencies as some report fewer than every month in the year. Agencies that do report, however, tend to report all 12 months of the year so this problem is not necessarily serious, though it depends on the agencies you're looking at.

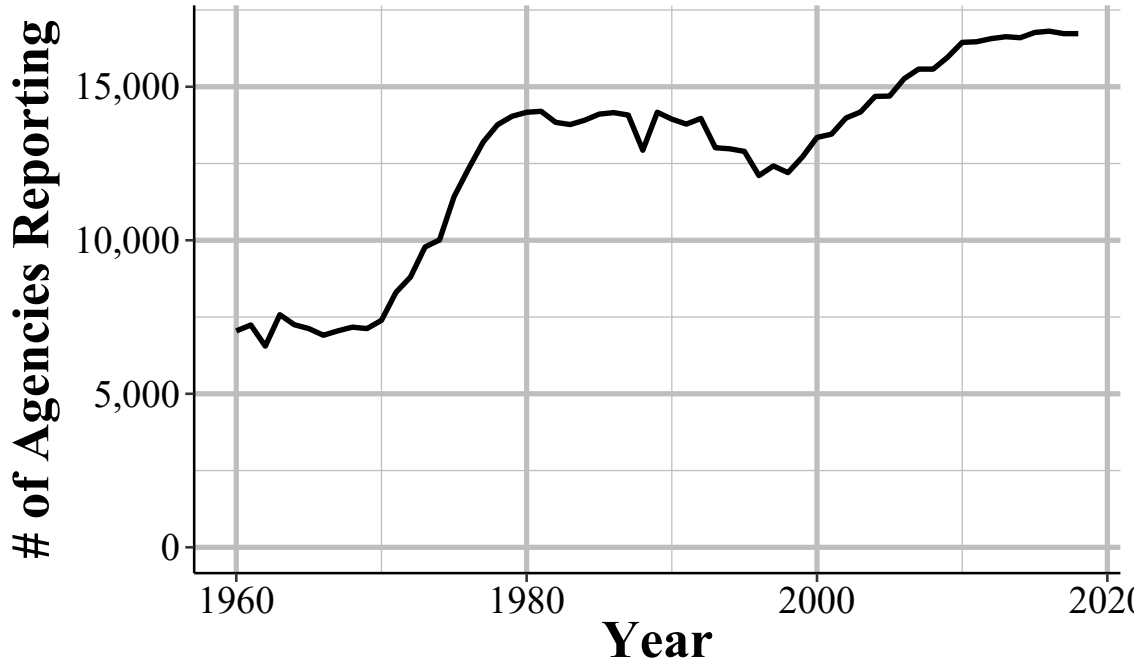


Figure 3.2: The annual number of agencies reporting to the Offenses Known and Clearances by Arrest dataset. Reporting is based on the agency reporting at least one month of data in that year.
(#fig:offensesAgenciesReporting)

Figure @ref(fig:offensesbigAgenciesReporting) repeats the above figure but now including only agencies with 100,000 people or more in their jurisdiction. While these agencies have a far more linear trend than all agencies, the basic lesson is the same: recent data has most agencies reporting; old data excludes many agencies.

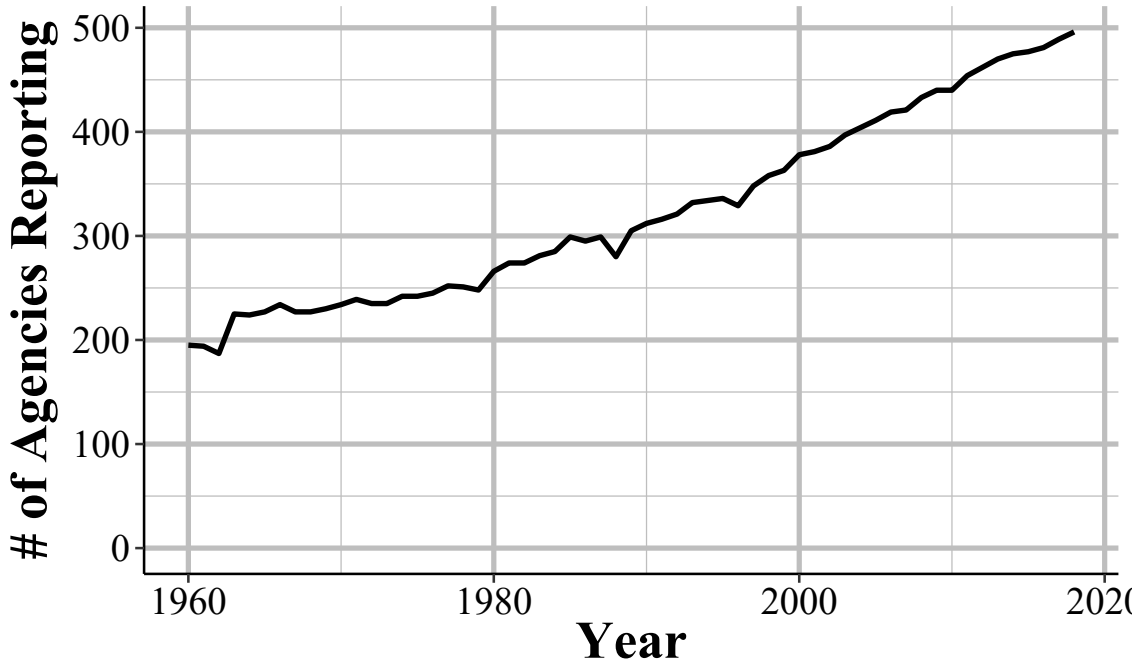


Figure 3.3: The annual number of agencies with a population of 100,000 or higher reporting to the Offenses Known and Clearances by Arrest dataset. Reporting is based on the agency reporting at least one month of data in that year.
(#fig:offensesbigAgenciesReporting)

3.3 Important variables

For each crime we have four different categories indicating the number of crimes actually committed, the number cleared, and the number determined to not have occurred. Like other UCR data, there are also variables that provide information about the agency - ORI codes, population under jurisdiction - the month and year that the data covers, and how many months reported data.

3.3.1 Actual crimes

This is the number of offenses that *actually* occurred - where *actually* means that a police investigation found that the crime report was accurate. Crimes that are reported that the police find did not occur (e.g. report of an arson but turns that the fire began accidentally) are called "unfounded" crimes. So this variable is the one people use to measure *crime*. For example if 10 people are murdered in a city the number of "actual murders" would be 10. A crime is a crime incident, regardless of how many offenders there were. If there are multiple victims in a case, such as a double murder, then it would count as multiple crimes.

3.3.2 Total cleared crimes

A crime is cleared when an offender is arrested or when the case is considered cleared by exceptional means. To be more specific, this data counts crime as a crime incident, regardless of the number of offenders. For example, if 3 people robbed a person, that is one crime of robbery, not 3 separate crimes. This crime is cleared when one of the three robbers is arrested - no matter the outcome for the other two robbers. Arresting all 3 still counts as a single robbery cleared.

Even though this dataset is formally named "Offenses Known and Clearances by Arrest" it does include clearances where no one is arrested, which are called "exceptional clearances" or "clearances by exceptional means." For exceptional clearances, police must have identified the offender, have sufficient evidence to arrest that offender, know where they are (so they can actually be apprehended) and only then be unable to make the arrest. Exceptional clearances include cases where the offender dies before an arrest (by any means, including suicide, accidental, illness, killed by someone including a police officer) or when the police are unable to arrest the person since they are already in custody by another jurisdiction (including out of the country or in custody of another agency) and extradition is denied. Two other potential causes of exceptional clearance are when prosecution of the case cannot go forward because the district attorney refuses to prosecute the case, for reasons other than lack of evidence, or when a victim refuses to assist the prosecution in the case.

Unfortunately, this data does not differentiate between clearances by arrest or by exceptional means so there's no way to determine how many clearances mean the offender is actually arrested - and even more problematic, how many clearances have all of the offenders arrested.[2] There is also evidence that police agencies report classify large numbers of clearances as clearances through exception means (again, we have no way to tell which is which using this data) even though exceptional clearances should be rare given that times where the offender is known but cannot be arrested are likely far less common than when they are arrested. This is particularly a problem in the case of rapes where in some cases there are far more exceptional clearances than clearances by arrest (**?**). For an excellent investigation into this issue, please read the **?** article available on ProPublica's website here

Clearances are reported in the month that they occur, regardless of when the crime they are clearing occurred. In practice, however, most crimes are cleared in the month that they occur. According to the 2019 NIBRS, it takes on average 7 days between the incident and the arrest (median = 0 days) date when averaging across all crimes - for individual crimes these values will be different. This means that most of the clearances will be for the same month as the initial crime - though less so as the month comes to a close. Of course, police agencies can solve older cases - and even target cold cases to be solved - so this is still a

[2] NIBRS data does differentiate these types of clearances

degree of uncertainty for which month these clearances are for.

Still, there are occasionally months - and even years - where there are more reported crimes cleared than crimes that occur. Figure @ref(fig:montgomeryClearances) shows the number of actual and cleared murders from the Montgomery County Police Department in Maryland. Several years have more murders cleared than were committed - a sign that the month of clearance does not correspond to the month of occurrence, rather than the police solving crime before it happened.[3]

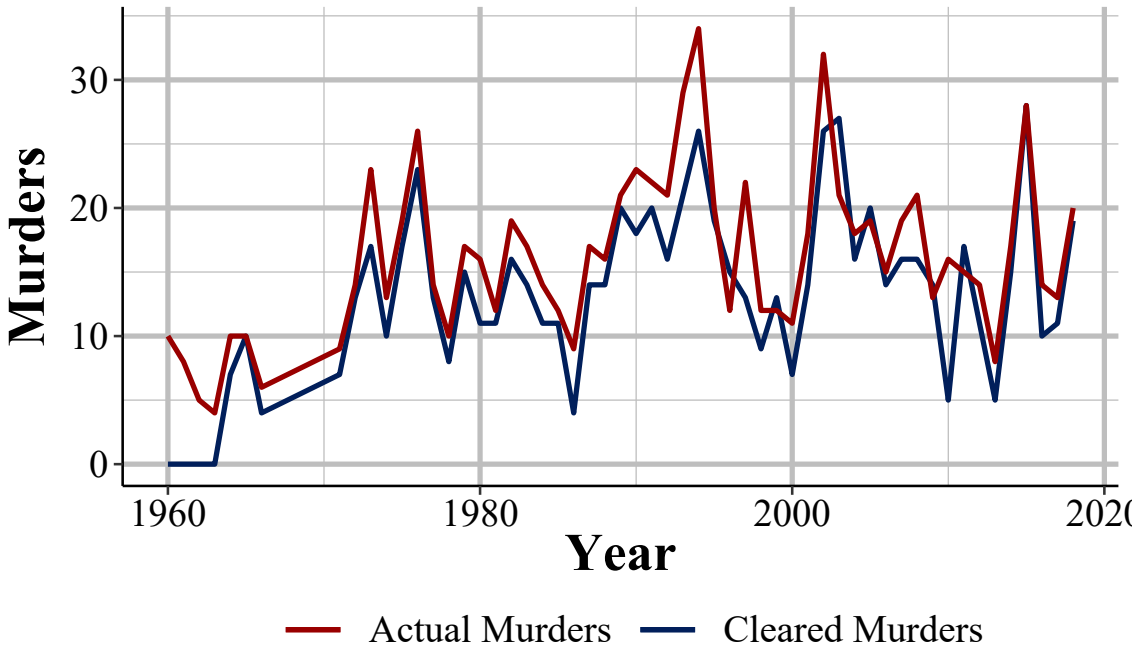


Figure 3.4: The annual number of actual and cleared murders from the Montgomery County Police Department, MD, 1960-2018.
(#fig:montgomeryClearances)

3.3.3 Crimes cleared where all offenders are under 18 years old

This variable is a subset of the Total Cleared variable and only includes clearances for offenses in which **every** offender is younger than age 18. Since this requires that the police know, or at least believe, the age of every offender, it is probably highly inaccurate. This category includes cases where the juvenile is given a citation to show up in court for their trial and is not formally arrested and taken into custody.

[3]The Minority Report is not yet a documentary.

3.3.4 Unfounded crimes

An unfounded crime is one in which a police investigation has determined that the reported crime did not actually happen. For example I observed during a ride-along a report of a burglary where the homeowners said that they came home, and the front door was open and they thought it might have been their son who forgot to close it but were worried that it could be a burglar, so they called the police just in case. This would be recorded as a burglary and if it turned out to be the son, the police would then record this as an unfounded burglary.

Other unfounded crimes would include when someone reports a crime but later says that the report wasn't true. For example, a person could report a burglary to the police to collect insurance money on the items they claim was stolen. If the police discover this they would unfound the case - and the lying to the police and fraud would not be counted as neither of those are crimes included in this dataset. Unfounding crimes are especially common for rapes. If a person reports that they were raped and then later say that this report isn't true - and doing so is relatively common, especially for child sexual abuse victims, even when the rape did actually happen - then the police will unfound the case.

Figure @ref(fig:phillyRapeUnfound) provides one example of this by showing the number of actual - that is, rapes that the police say actually occurred - and unfounded rapes annually from 1960-2018 in Philadelphia, PA. Interestingly, the spike in actual rapes in 2013 due to the new rape definition that year (discussed below) does not correspond to a spike in unfounded rapes - which suggests that this variable is not being reported properly. It is unlikely that the number of rapes would spike so much without any corresponding increase - and an actual decrease - in unfounded rapes. Since unfounded rapes are so common, especially compared to other crimes - and evidence that police also misreport clearances for rape - this variable is likely to be manipulated by the police to make it appear that there are fewer rapes than there actually is. As such, this variable - and relatedly the actual crimes variable" is not reliable, especially for rapes. How unreliable, of course depends on the specific offense and the agency reporting.

Another way to look at this data is the percent of rapes that were unfounded. Since unfounded and actual rapes are distinct categories, we'll need to divide unfounded rapes by the sum of actual and unfounded rapes and then multiple the result by 100. Please note that unfounding occurs in the month the police discover the case to be unfounded, it is not tied to the month of the original report. So this graph isn't really showing the percent that year that were unfounded - but for our purposes of getting a general sense it is acceptable. Usually about 10% of reported rapes are unfounded in Philadelphia. The 44% in 1983, the first year with data, is probably high due to it accounting for unfounded cases in previous years that weren't reported in the data until 1983.

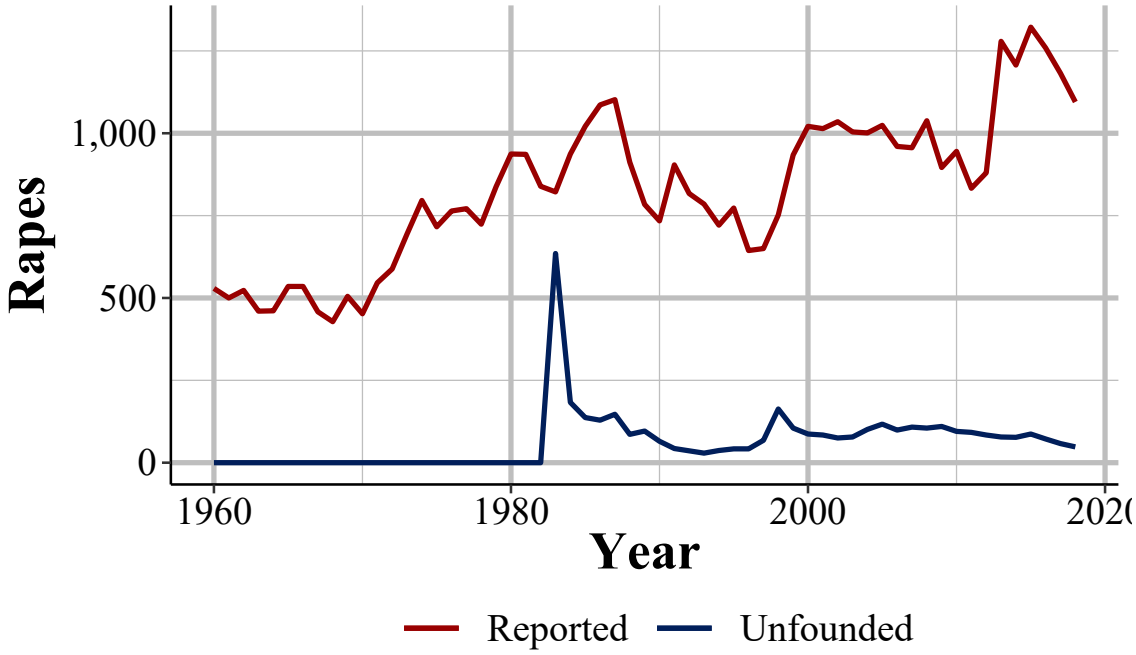


Figure 3.5: The annual number of actual and unfounded rapes in Philadelphia, PA, 1960-2018.
(#fig:phillyRapeUnfound)

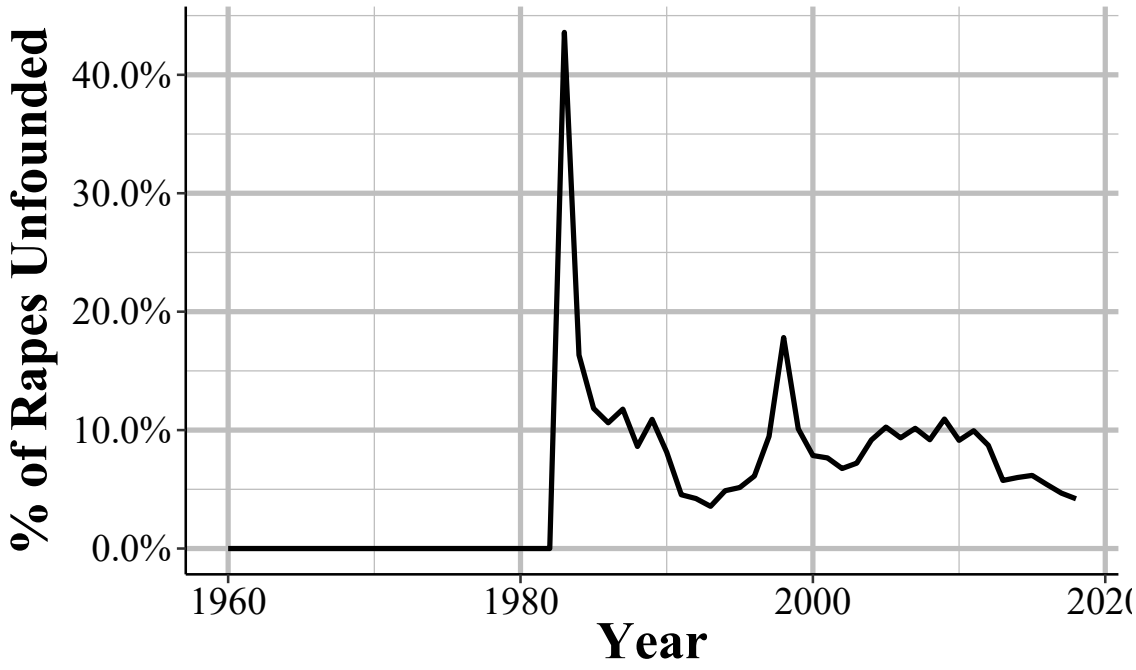


Figure 3.6: The percent of reported rapes that the police recorded as unfounded in Philadelphia, PA, 1960-2018.
(#fig:phillyRapeUnfoundPercent)

As a comparison, Figure @ref(fig:phillyRobberyUnfoundPercent) shows the percent of robberies that are recorded as unfounded in Philadelphia over the same time period. While the percent has been increasing, for most of the data it is under 1% and never exceeds 3%. To me this suggests that the Philly Police, and they aren't unique, are overcounting rapes as unfounded to improve their rape statistic data.

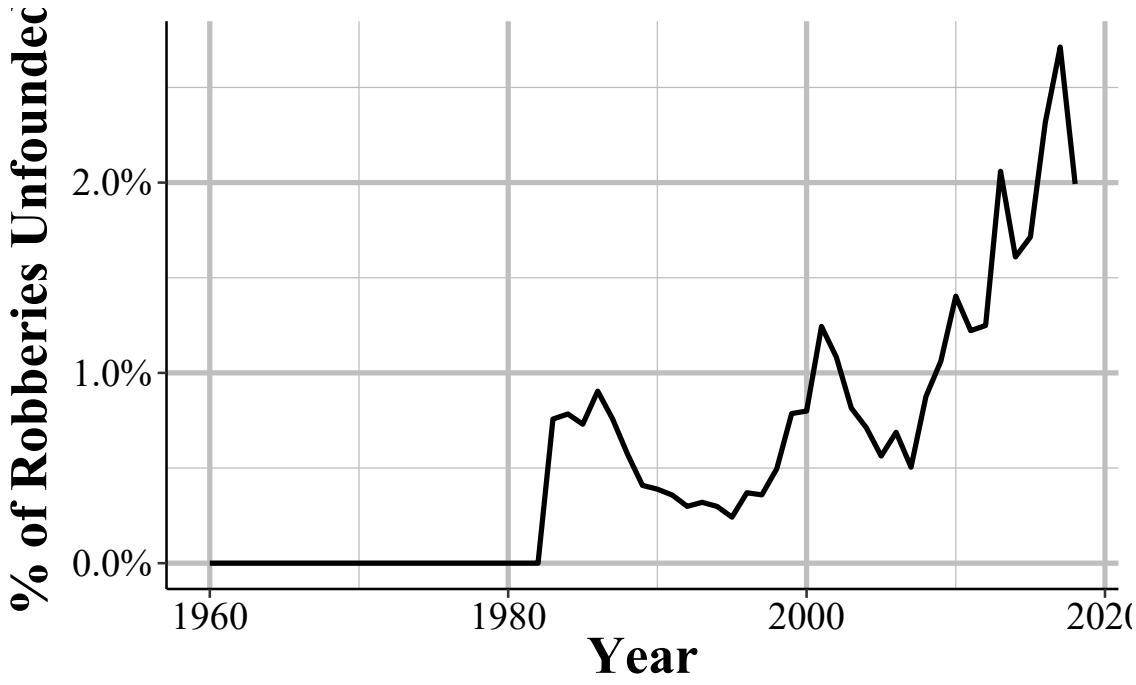


Figure 3.7: The percent of reported robberies that the police recorded as unfounded in Philadelphia, PA, 1960-2018.
(#fig:phillyRobberyUnfoundPercent)

3.4 Important issues

3.4.1 Rape definition change

The FBI changed the definition of rape for UCR data starting in 2013 to a broader definition than the older definition, which is commonly called the "legacy definition" or "legacy" or "historical" rape. The legacy definition is "the carnal knowledge of a female **forcibly** and against her will" (emphasis added). This means that only rape is only included in UCR data when it is a female (or any age, there is no differentiation for child victims) forcibly vaginally penetrated by a penis. This is a narrow definition and excludes a number of sexual acts that people may consider rape such as forced oral or anal sex, and cases with a male victim.

The new (and current) definition "penetration, no matter how slight, of the vagina or anus

with any body part or object, or oral penetration by a sex organ of another person, without the consent of the victim." Starting in 2013, rape has a new, broader definition in the UCR to include oral and anal penetration (by a body part or object) and to allow men to be victims. The new definition is: "Penetration, no matter how slight, of the vagina or anus with any body part or object, or oral penetration by a sex organ of another person, without the consent of the victim." The previous definition included only forcible intercourse against a woman. This definition is far broader and is effectively any non-consensual sexual act. It also includes male victims though the data does not differentiate between male or female (or any other gender) victims.

Both the current and legacy definitions exclude statutory rape and incest other than forcible incest. They both also include lack of consent as cases where the victim cannot give consent, such as if they are too young or are mentally or physically incapacitated - the FBI specifically give the example of being temporarily incapacitated through drugs or alcohol.

As this revised definition is broader than the original one post-2013, rape data is not comparable to pre-2013 data. 2013, however, is simply the year that the FBI changed the definition which means that agencies should have changed their reporting to the new definition. As might not be too surprising, not all agencies followed this requirement. We'll look at four examples to show when there is clear evidence that the agency did change their definition in 2013, when it's clear they did so a year later, when it's unclear exactly when they made the change, and when the agency seems to not follow the change at all.

We'll start with the Philadelphia Police Department in Philadelphia, PA, shown in Figure @ref(fig:rapePhilly) which shows the annual number of rapes from 2000-2018. It's declining slowly but steadily over the 2000-2012 time period until spiking sharply in 2013. Since the rape definition change in 2013 is far broader than previous year's definition, this makes sense. A broader definition should lead to a sudden increase in reported rapes if the agency is reporting correctly.

In comparison, New York City has the sudden spike a year later, which indicates that they didn't start using the new definition until 2014. Figure @ref(fig:rapeNYC) shows that rape is fairly steady, though increasing, in the years leading up to 2013 and has almost no change from 2012 to 2013, but a huge increase in 2014 and then steadily increases from there, spiking again in 2018. This seems like a fairly clear indicator that NYC simply didn't follow the new definition until 2014.

Less clear is what's happening in San Francisco, California, shown in Figure @ref(fig:rapeLA). Here we do see an increase in 2013 which while it appears small on the graph is actually a 49% increase from 2012. Then there is a much larger spike in 2014 - a 120% increase - which may suggest that part of the agency started following the new definition in 2013 and the remainder followed in 2014. However, large increases or decreases are relatively common

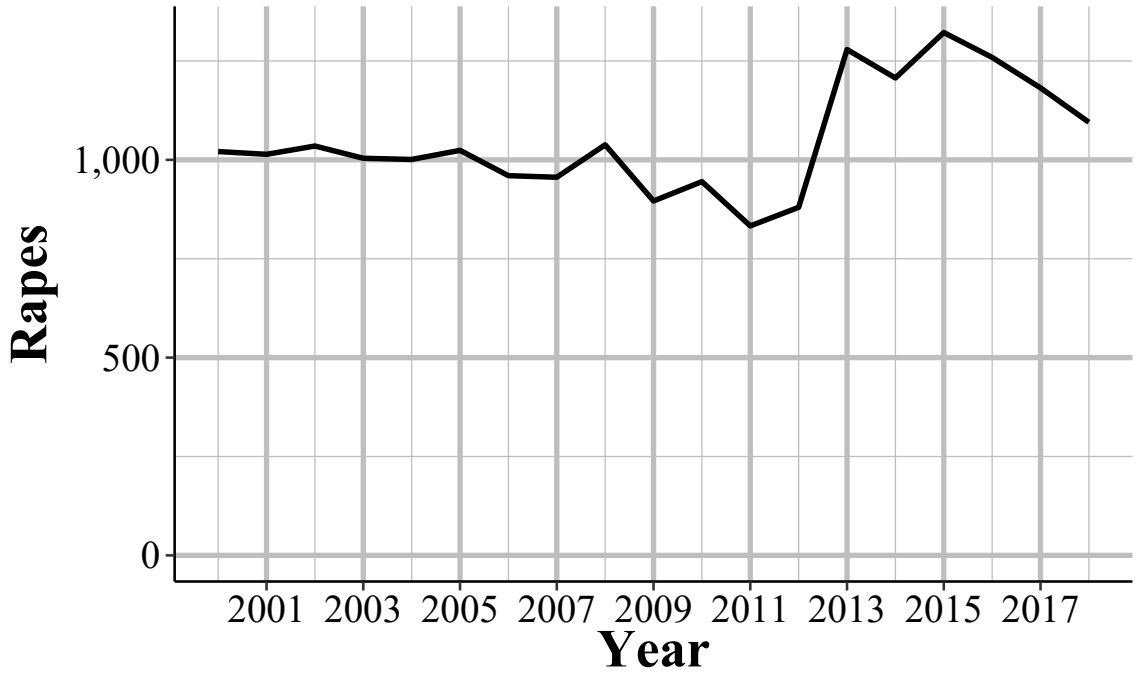


Figure 3.8: The annual number of rapes reported in Philadelphia, Pennsylvania, 2000-2018. (#fig:rapePhilly)

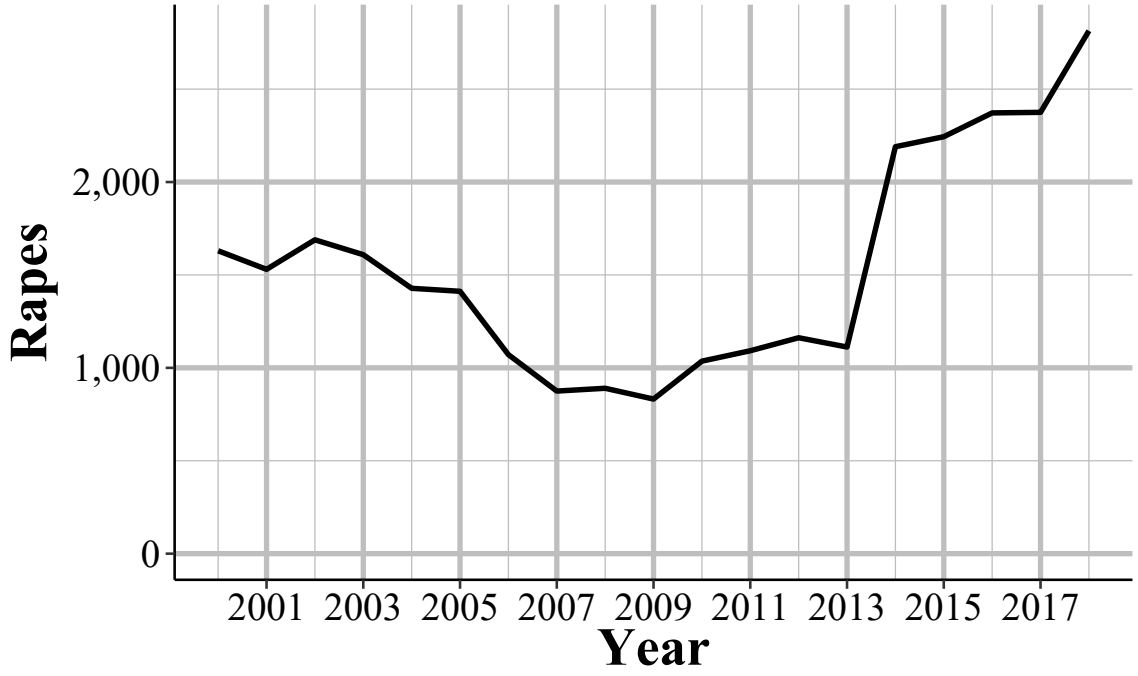


Figure 3.9: The annual number of rapes reported in New York City, 2000-2018. (#fig:rapeNYC)

in San Francisco so it could also be that the agency only switched to the new definition in 2014 and the spike in 2013 is just a coincidence.

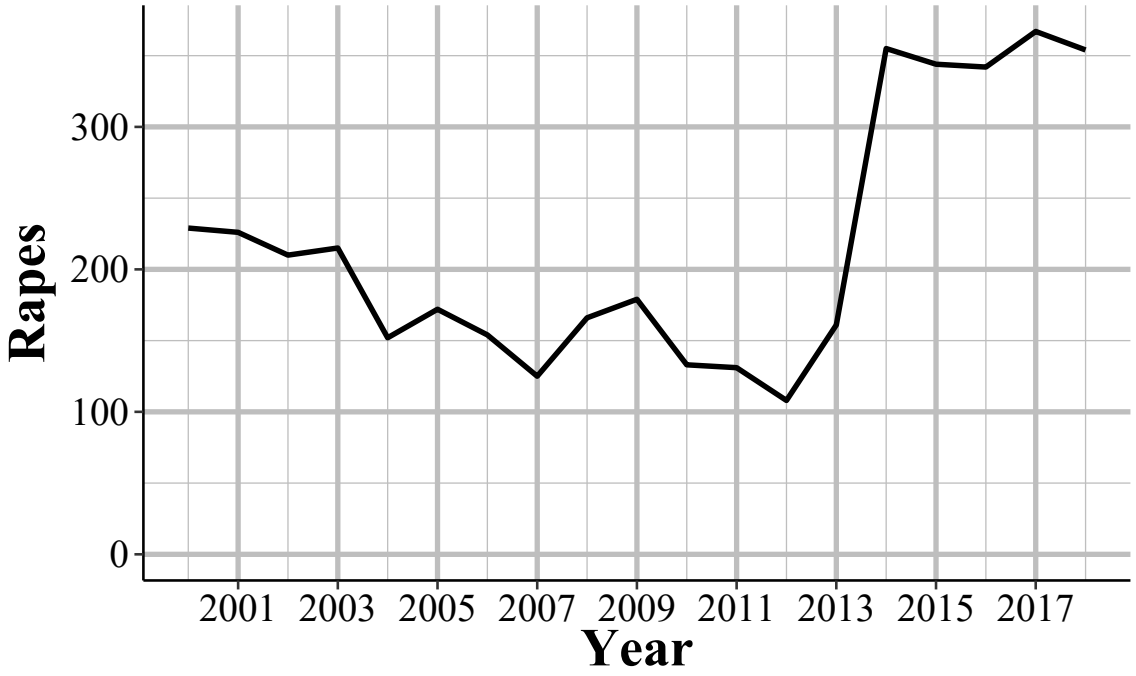


Figure 3.10: The annual number of rapes reported in San Francisco, California, 2000-2018. (#fig:rapeLA)

Finally, we'll look at Jackson Police Department in Mississippi where the definition change seems to have had no effect. As seen in Figure @ref(fig:rapeJackson), reported rapes start to undulate in 2010 with 2013 data perfectly in line with the before and after trends - no sign that there is a change in reporting. This suggests that Jackson simply did not follow the definition change and continues to report using the old definition.

My takeaway from this is that rape should not be used at all for years after 2012. While the definition change makes pre-2013 and 2013+ years non-comparable, the differences in agency responses to this change - i.e. if they follow the rules or not - is such a mess that the data is too flawed to use.

3.4.2 The decline of manslaughter

This data contains two different crime subcategories for homicide: murder and non-negligent manslaughter, and manslaughter by negligence. The first is our measure of murder, and it includes everything we traditionally think of when it comes to murder - shootings, stabbings, strangulation, basically any intentional killing of another person.[4] Suicides, killing

[4]Attempted murder is usually classified as an aggravated assault.

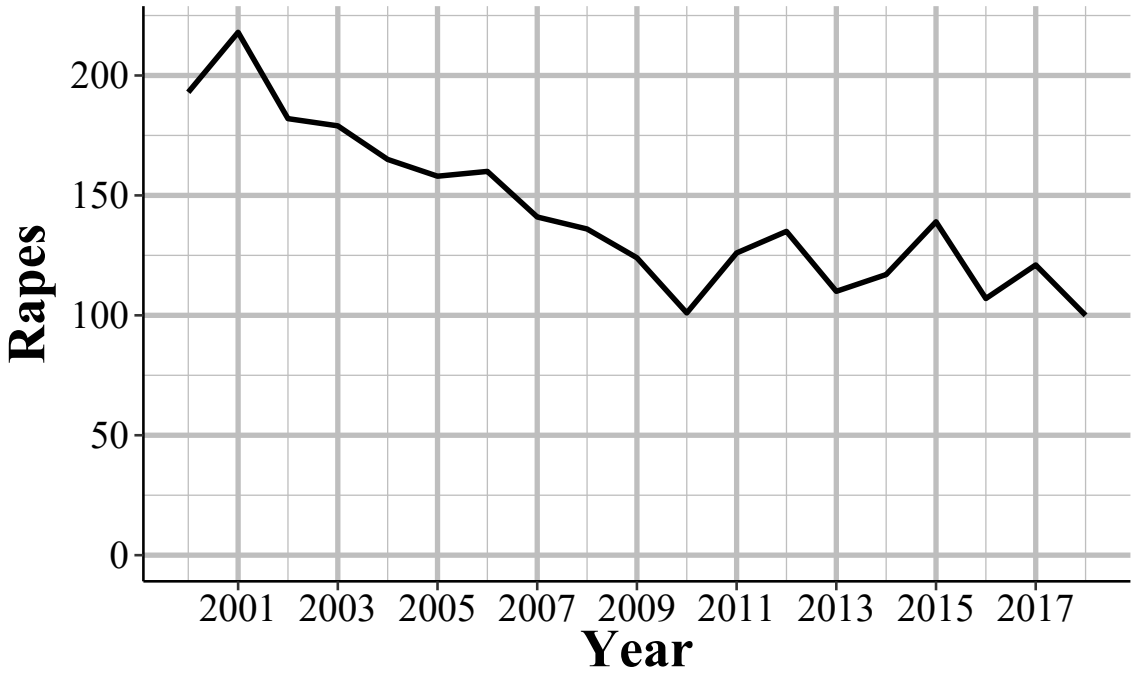


Figure 3.11: The annual number of rapes reported in Jackson, Mississippi, 2000-2018. (#fig:rapeJackson)

a fetus, and accidental killings (e.g. car crashes) are not considered murders.[5] The second, manslaughter by negligence - usually called just "manslaughter" - is when someone kills another person through "gross negligence" but does not kill them intentionally. This can include accidental killings when the death was caused by gross negligence. The FBI provide examples of this as kids playing with guns and shooting each other (and not knowing the gun was loaded) or a hunter accidentally shooting someone while hunting. After the late 1970s this excludes deaths from traffic accidents caused by negligence, such as hit and runs or DUIs. However, prior to this, these deaths were included, which is the cause of the very high number of manslaughters in the 1960s and 1970s.

Figure @ref(fig:manslaughterVsMurder) shows the annual number of murders, manslaughter, and the sum of the two nationwide from 1960-2018. This just sums up the total reported counts from every agency each year so part of the increase is simply due to more agencies reporting as the year gets closer to the present day - so please pay attention to the diverging paths of each crime, not the trend for the individual crime over time. Murder is always more common than manslaughter but these values are not that far apart in the early decade of data and manslaughter doesn't become rare until the end of the 1970s.

Figure @ref(fig:manslaughterPercent) shows another way to look at this data: manslaughter as a percent of reported murder. In the early years of our data manslaughter was fairly

[5]Even the intentional killing of a fetus is classified as an aggravated assault against the mother, not a murder of the fetus.

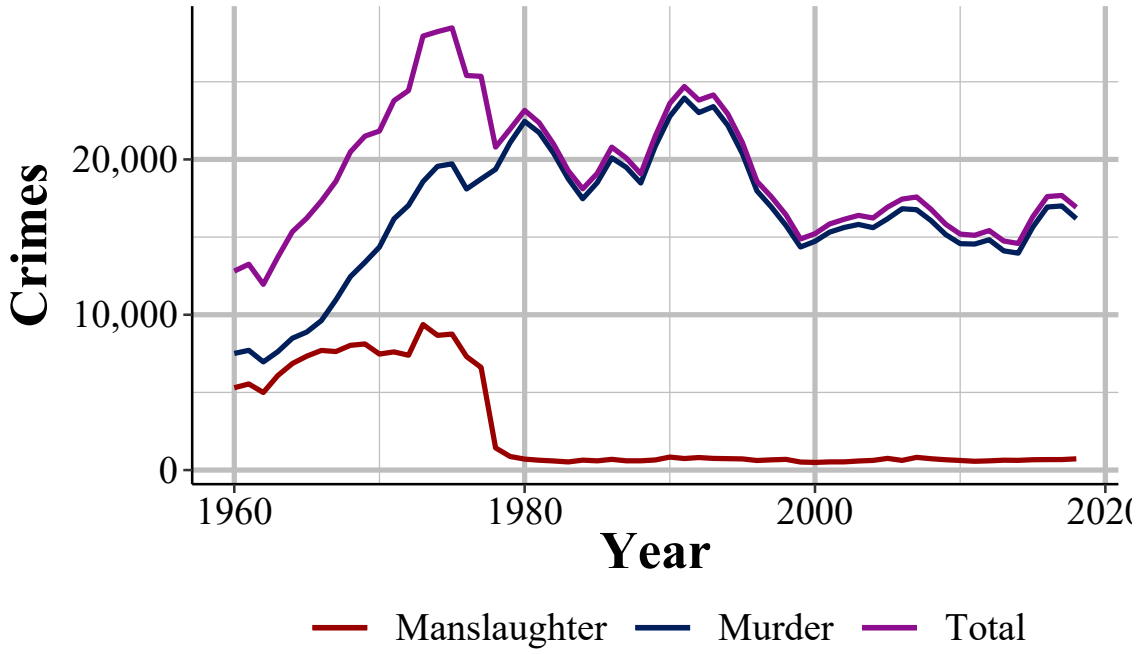


Figure 3.12: The annual number of murder and non-negligent manslaughter, manslaughter by negligence, and the sum of the two, nationwide from 1960-2018.
(#fig:manslaughterVsMurder)

common, with about 70-80% as many manslaughters reported as murders. This declined sharply in the mid-1960s until there were around 45% as many manslaughters as murders in the mid-1970s. Again this declined until it was about 4% in 1980, and it has remained around there ever since. As police behavior could reduce traffic fatalities - and arrests for DUIs and traffic tickets are designed to improve public safety - it is unfortunate the we no longer have data on traffic deaths.

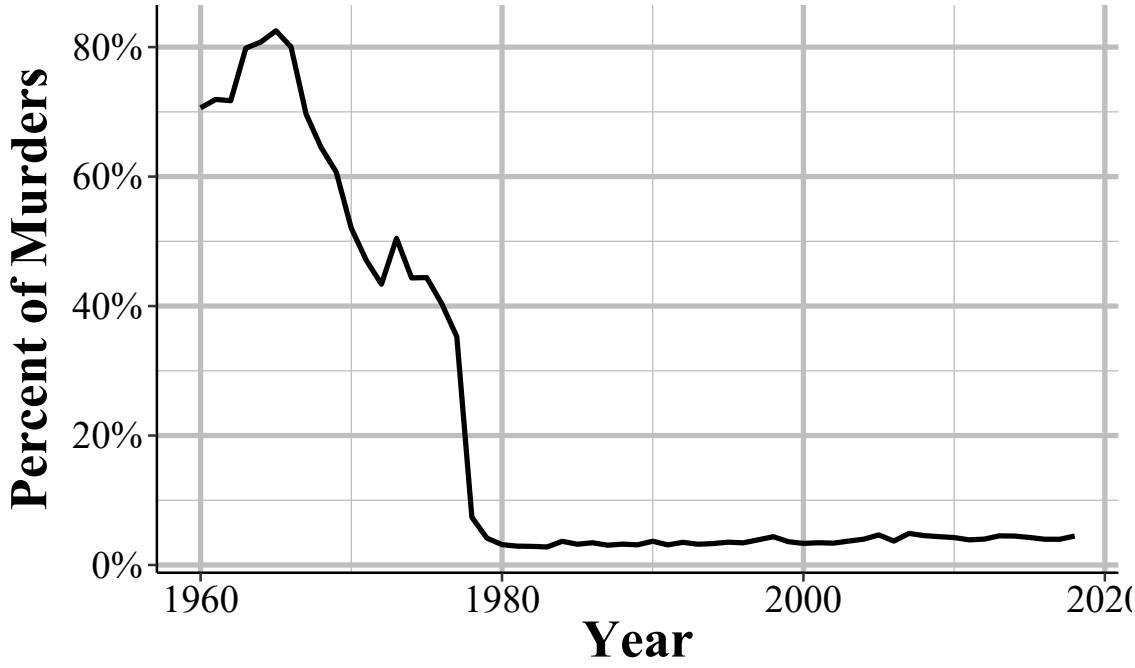


Figure 3.13: Reported manslaughter by negligence as a percent of reported murder and non-negligent manslaughter, nationwide 1960-2018.
(#fig:manslaughterPercent)

Chapter 4

Property Stolen and Recovered (Supplement to Return A)

The Property Stolen and Recovered data - sometimes called the Supplement to Return A (Return A being another name for the Offenses Known and Clearances by Arrest dataset, the "crime" dataset) - provides monthly information about property-related offenses (theft, motor vehicle theft, robbery, and burglary), including the location of the offense (in broad categories like "gas station" or "residence"), what was stolen (e.g. clothing, livestock, firearms), and how much the stolen items were worth.[1] The "recovered" part of this dataset covers the type and value of property recovered so you can use this, along with the type and value of property stolen, to determine what percent and type of items the police managed to recover. Like other UCR datasets this is at the agency-month level so you can, for example, learn how often burglaries occur at the victim's home during the day, and if that rate changes over the year or differs across agencies.

The data, however, provides no information about the offender or the victim (other than if the victim was an individual or a commercial business [based on the location of the incident - "bank", "gas station", etc.]). The value of the property stolen is primarily based on the victim's estimate of how much the item is worth (items that are decreased in value once used - such as cars - are supposed to be valued at the current market rate, but the data provides no indication of when it uses the current market rate or the victim's estimate) so it should be used as a very rough estimate of value.

Before getting into the details of this data, let's look at one example of how this data can measure property crime in a few different ways. This data is highly useful to use as a rough measure of the cost of crime. This cost is limited to just the value of the property stolen - so excludes things like injuries, mental health effects of victimization, etc. - but is still better

[1] It also includes the value of items stolen during rapes and murders, if anything was stolen.

than nothing. Since this data includes, for some types of property stolen, both the number of offenses per month (broken down by type of theft and items stolen) and the value of the stolen property, we can also see if the average value of these thefts change over time.

We'll look at home burglaries that occur during the day in Philadelphia. First, we can look at the number of these kinds of burglaries each month or year. Figure @ref(fig:phillyHomeBurglaryCount) shows the annual number of daytime home burglaries and in recent years it has declined sharply into having the fewest on record in 2019. So citywide, Philadelphia has never been safer when it comes to daytime home burglaries. But when considering the cost of crime, we also want to know the actually monetary cost of that incident. This encompasses a lot of different values including physical and emotional harm to the victim (including harm to structures such as broken doors and windows) and changes in people's behavior (e.g. quit job to avoid going to a certain part of town). With this data we only have the value of the physical items stolen, so while it is still a cost of crime it's a rather shallow one.

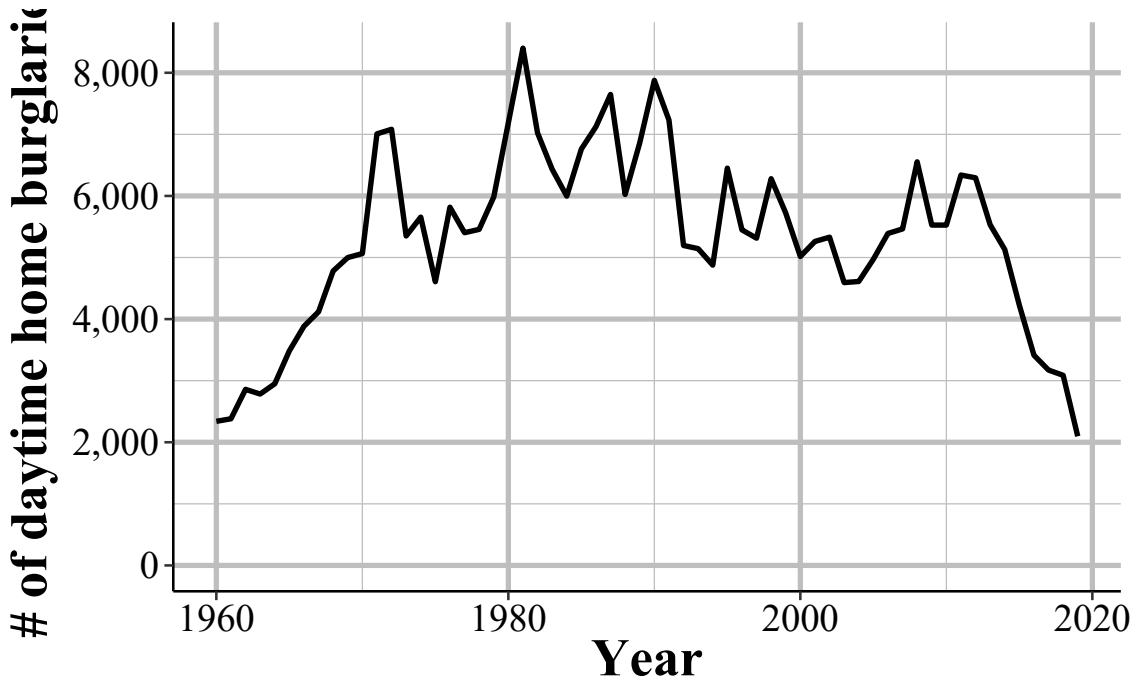


Figure 4.1: The annual number of daytime home burglaries reported in Philadelphia, PA, 1960-2019.
(#fig:phillyHomeBurglaryCount)

Figure @ref(fig:PhillyBurglaryCost) measures this cost of crime by showing the annual sum of the value of the property stolen for daytime home burglaries in Philadelphia. The trend here is different than the previous graph which showed undulations in the number of burglaries but not major trend changes until the 2010s; here is a steady increase over the long term, though with varying speed of increase, until it peaked in the late 2000s/early 2010s before

declining substantially in recent years. While the number of burglaries peaked in the early 1980s, the total value of burglaries didn't peak until the early 2010s, so the cost of this crime (even this very narrow measure of cost) can't be ascertained from knowing the number of burglaries alone. From this measure we can say that daytime home burglaries were worse in the early 2010s and are substantially better currently.

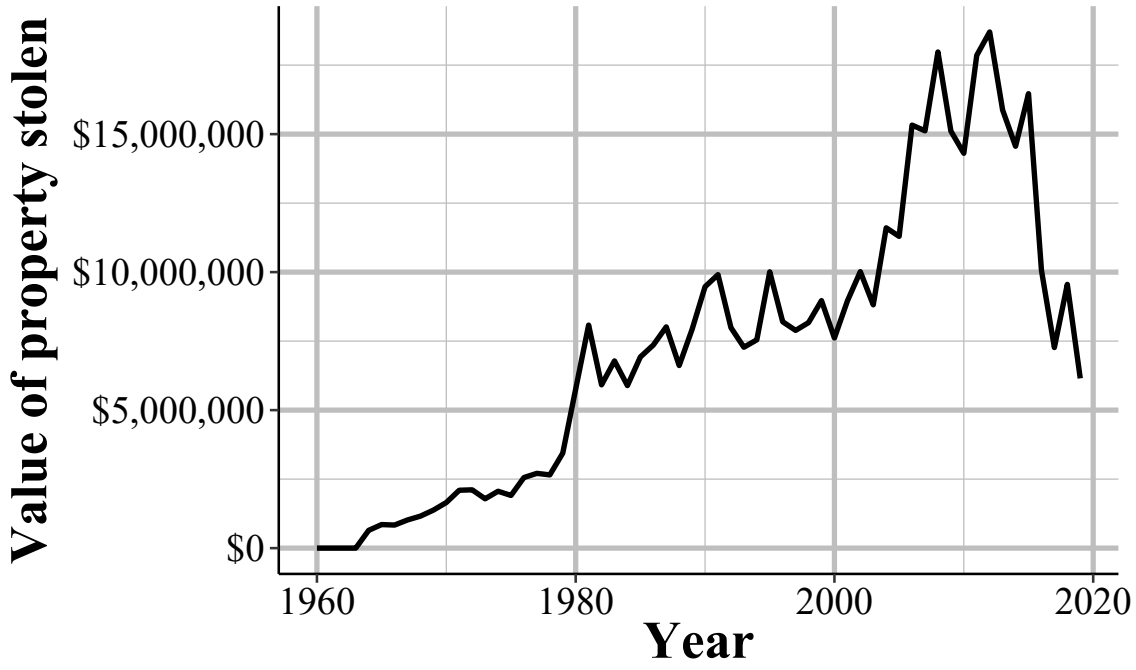


Figure 4.2: The total annual cost of daytime home burglaries in Philadelphia. (#fig:PhillyBurglaryCost)

The final way we can measure daytime home burglaries is to put the previous variables together to look at the cost per burglary. This will give us the average amount of property stolen from each burglary victim. Figure @ref(fig:phillyHomeCostPerBurglary) shows the cost average cost per burglary for each year of data available. Now we have a different story than the other graphs. Even though the number of daytime home burglaries declined substantially over the last decade and the total cost is around the level seen in the 1980s, the cost per burglary remains at record highs in recent years, though down from the peak in the mid-2010s. This suggests that while burglaries are down, the burden on each burglary victim has continued to grow.

Part of this - and part of the long-term increase seen in Figure @ref(fig:PhillyBurglaryCost) - is simply due to inflation. $1 in 1963, the first year we have data on the value of burglaries, is worth $8.28 in 2019, according to the Bureau of Labor Statistics. The values in this data are *not* adjusted for inflation so you need to do that adjustment yourself before any analyses, otherwise your results will be quite a bit off. When we adjust all values to 2019 dollars, as shown in Figure @ref(fig:phillyHomeCostPerBurglaryInflation), the trend becomes changes

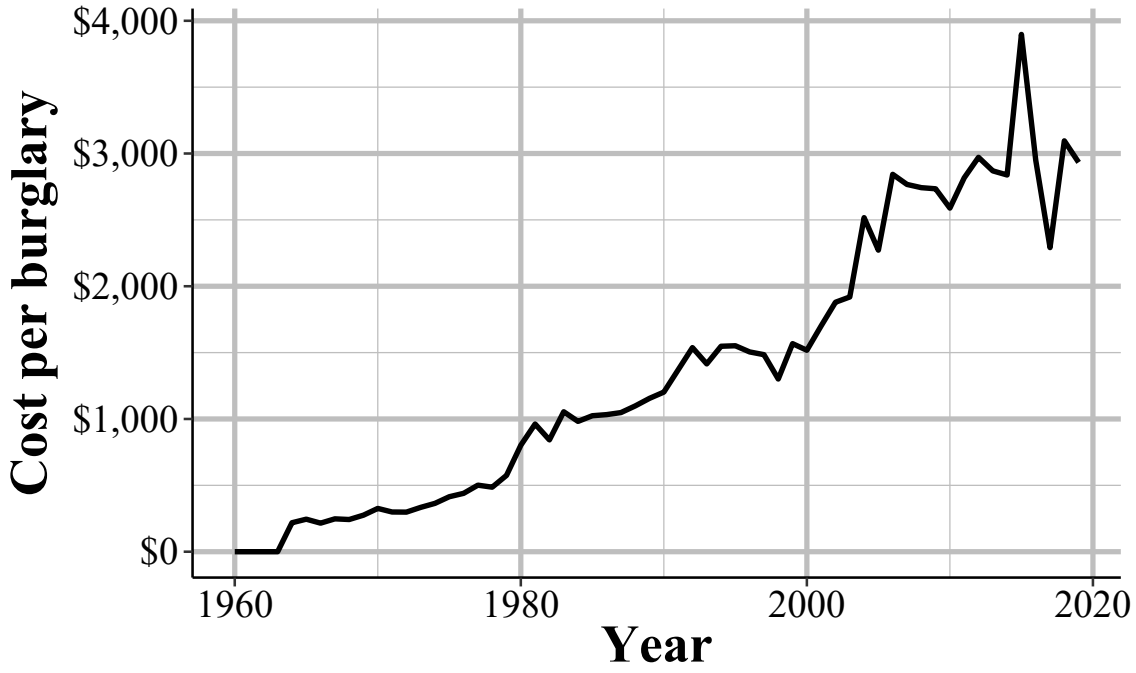


Figure 4.3: The annual number of burglaries and cost per burglary for daytime home burglaries in Philadelphia, 1960-2019.
(#fig:phillyHomeCostPerBurglary)

a bit. There's still a steady increase in cost per burglary over time but it is far more gradual than in Figure @ref(fig:phillyHomeCostPerBurglary). And now the difference from the cost in early years and late years are far smaller.

4.1 Agencies reporting

We'll start by looking at which agencies report. The data is available from 1960 through 2019 though the columns about the value of the property only begin in 1964. Figure @ref(fig:propertyAgencies) show the number of agencies each year that reported at least one month during that year. In the first several years of data barely any agencies reported data and then it spiked around 1966 to over 6,000 agencies per year then grew quickly until over 12,000 agencies reported data in the late 1970s. From here it actually gradually declined until fewer than 12,000 agencies in the late 1990s before reversing course again and growing to about 15,000 agencies by 2019 - down several hundred agencies from the peak a few years earlier.

Since this data is called the "Supplement to Return A" we would expect that the agencies that report here are the same as the ones that report to the Offenses Known and Clearances by Arrest data, which is also called the Return A dataset. Figure @ref(fig:agenciesInBoth)

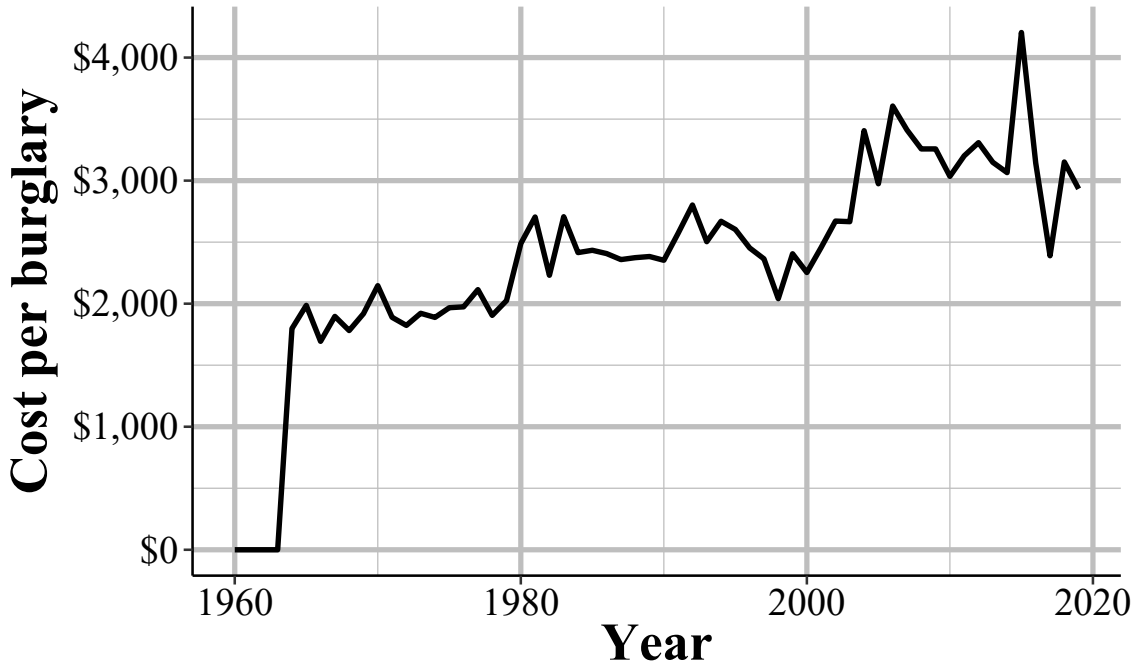


Figure 4.4: The inflation-adjusted annual number of burglaries and cost per burglary for daytime home burglaries in Philadelphia, 1960-2019.
(#fig:phillyHomeCostPerBurglaryInflation)

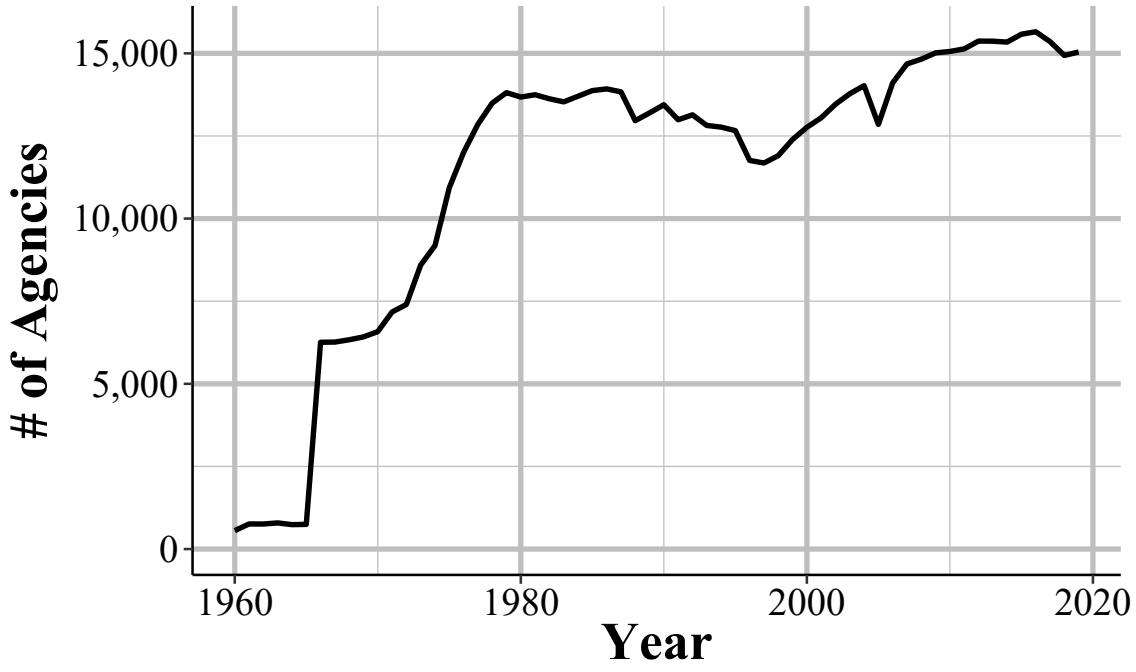


Figure 4.5: The annual number of police agencies that report at least month of data that year.
(#fig:propertyAgencies)

shows the percent of agencies in this dataset that are also in the Return A data. Except for the first several years of data in the 1960s, we can see that most years have nearly all agencies reporting to both, though this has declined in recent years. Since the late 1970s over 90% of agencies that report to the Offenses Known data also report to this dataset.

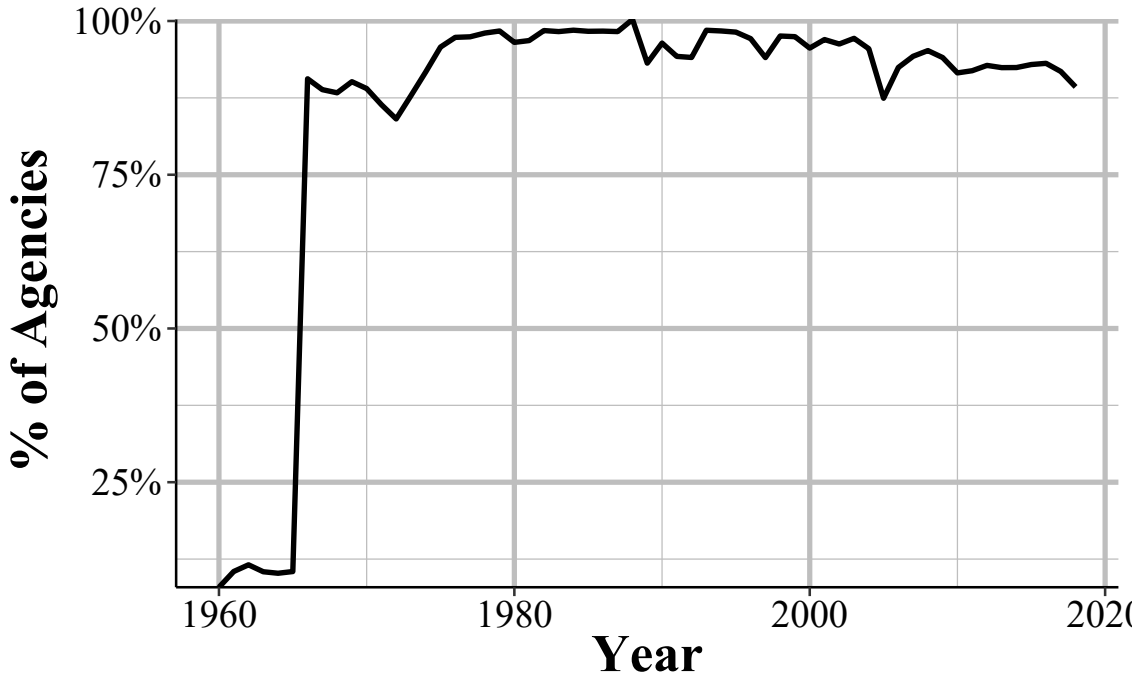


Figure 4.6: The percent of agencies in the Supplement to Return A data that are also in the Offenses Known and Clearances by Arrest (Return A) data in that year, 1960-2019. (#fig:agenciesInBoth)

4.2 Important variables

This data can really be broken into two parts: the monthly number of property (including robbery) crimes that occur that are more detailed than in the Offenses Known data, and the value of the property stolen (and recovered) from these crimes.[2] For each category I present the complete breakdown of the available offenses/items stolen and describe some of the important issues to know about them. Like other UCR data, there are also variables that provide information about the agency - ORI codes, population under jurisdiction - the month and year that the data covers, and how many months reported data.

[2]I'm really not sure why this data isn't just merged with the Offenses Known data, especially the first part where it's just counts of crimes.

4.2.1 A more detailed breakdown of property (and robbery) crimes

The first part of this data is just a monthly (or yearly in the annual data that I've released - the FBI, however, only releases data monthly so I just aggregated the months together) number of crimes of each type reported to the police (and that a police investigation discovered actually happened. For more on this process, please see Chapter @ref(actual)). There are six crimes and their subsets included here: burglary, theft, robbery, and motor vehicle theft. The remaining two are rape and murder, but they don't break down these crimes into any subcategories and are only available starting in 1974.

Burglary is reported based on whether the building burgled was the victim's residence or not, and also the time of the burglary. Time is either during the day (6am-5:59pm) or night (6:pm-5:59am) or if the time is unknown. Data is available since 1960 for both the day and night burglaries, but only since 1964 for the unknown time burglaries. For robbery, the subcategories are based on where the robberies occurred such as if it happened in a bank, a gas station, or a street.

Theft is divided into two groups. The first group is based on the value of items stolen: less than $50, $50-$199, and $200 and up. The second group of thefts is broken into the type of theft such as a shoplifting or stealing from someone's car. All theft variables begin in 1960 other than thefts from a coin machine and from a building which start in 1964 and the miscellaneous "All other thefts" variable that has data starting in 1961. Finally, motor vehicle theft is broken into where the stolen vehicle was stolen and found: stolen in the reporting agency's jurisdiction and found by the same agency, stolen in the reporting agency's jurisdiction and found by a different agency, and stolen in a different agency's jurisdiction and found by the reporting agency.

The complete list of each subsection and a brief definition for the non-obvious ones are presented below.

- Burglary
  - Home/residence during the day (6:00am - 5:59pm)
  - Home/residence during the night (6:00pm - 5:59am)
  - Home residence at unknown time
  - Non-residence (i.e. all buildings other than victim's home) during the day (6:00am - 5:59pm)
  - Non-residence (i.e. all buildings other than victim's home) during the night (6:00pm - 5:59am)
  - Non-residence (i.e. all buildings other than victim's home) at unknown time

- Theft/larceny (excluding of a motor vehicle)
  - <$50
  - $50-$199
  - $200 and up
  - Pick pocket
  - Purse snatching
  - Shoplifting
  - Stealing from a car (but not stealing the car itself)
  - Stealing parts of a car, such as the car battery or the tires
  - Stealing a bicycle
  - Stealing from a building where the offender is allowed to be in (and is not counted already as shoplifting)
  - Stealing from a "coin operated machine" which is mainly vending machines
  - All other thefts
- Robbery
  - Highway - This is an old term to say a place is outside and in generally accessible and visible areas. This includes robberies on public streets and alleys.
  - Commercial building - This is robberies in a business other than ones stated below. Includes restaurants, stores, hotels, bars.
  - Gas station
  - Chain/convenience store - a neighborhood store that generally is open late and sells food
  - Home/residence
  - Bank
  - Miscellaneous/other - This is all other robberies not already covered.
- Motor vehicle theft
  - Stolen in current agency's jurisdiction and found by that agency
  - Stolen in current agency's jurisdiction and found by another agency
  - Stolen in another agency's jurisdiction and recovered by current agency
- Murder
- Rape

4.2.2 The value of property stolen in property (and robbery) crimes

The next set of variables is the value of the property stolen in each crime, as well as the value of property stolen broken down by the type of property (e.g. clothing, electronics, etc.). To be clear, this is *only* the value of the property stolen during the crime. The cost of any injuries (mental or physical) or any lasting cost to the victim, their family, and society for these crimes are not included. This, of course, vastly underestimates the cost of these crimes so please understand that while this is a measure of the cost of crime, it is only a narrow slice of the true cost.

The cost is the cost for the victim to replace the stolen item. So the current market price for that item (though factoring in the current state the item is in, e.g. if it's already damaged) and, for businesses, the cost to buy that item and *not* the cost they sell it for. While the police can ask the victim how much the property was worth, they aren't required to use the exact amount given as victims may exaggerate the value of items. This is not an exact science, so I recommend only interpreting these values as estimates - and potentially rough estimates.

The value of the property stolen is broken into two overlapping categories: by crime type, and by type of property that was stolen. These are the exact same categories as covered in Section @ref(propertycount) but now is the dollar amount of the items stolen from those types of crimes. The second group is what type of items, based on several discrete categories, was stolen. Please note that multiple items can be stolen in each category - i.e. there is no Hierarchy Rule used here (for more in the Hierarchy Rule that is used in other UCR data, please see Section @ref(hierarchy)). And it counts the property stolen for each crime type *as well as* for each item type. So if you sum up all of the crime variables and all of the item type variables together you'll overcount the value of property stolen. Each of the item types and their definitions - some categories include or exclude things that you may not expect them to - are detailed below.

Though this data starts in 1960, not all variables are available then. 1963 and 1964 saw many new variables added - the values in these variables are reported as 0 in previous years - and in 1974 and 1975 even more variables are added. In 1963 the value of burglaries where the time of the burglary was known, thefts broken down into categories based on the value of property taken, thefts of car parts, theft from cars, shoplifting, and "other" thefts was added to the data. In the following year this data began including the value of property stolen from burglaries where the time of the burglary was unknown was added as well as thefts of bicycles, from "coin operated machines" (i.e. vending machines), purse snatching, and pick pocketing. The value of property stolen during rapes and murders was first reported in 1974. Finally, 1975 was the last year with new variables, with this year including consumable

goods, stolen guns, household goods, livestock, office equipment and electronics, and sound and picture equipment.

- Burglary
  - Home/residence during the day (6:00am - 5:59pm)
  - Home/residence during the night (6:00pm - 5:59am)
  - Home residence at unknown time
  - Non-residence (i.e. all buildings other than victim's home) during the day (6:00am - 5:59pm)
  - Non-residence (i.e. all buildings other than victim's home) during the night (6:00pm - 5:59am)
  - Non-residence (i.e. all buildings other than victim's home) at unknown time
- Murder
- Rape
- Robbery
  - Highway - This is an old term to say a place is outside and in generally accessible and visible areas. This includes robberies on public streets and alleys.
  - Commercial building - This is robberies in a business other than ones stated below. Includes restaurants, stores, hotels, bars.
  - Gas station
  - Chain/convenience store - a neighborhood store that generally is open late and sells food
  - Home/residence
  - Bank
  - Miscellaneous/other - This is all other robberies not already covered.
- Theft/larceny (excluding of a motor vehicle)
  - Pick pocket
  - Purse snatching
  - Shoplifting
  - Stealing from a car (but not stealing the car itself)
  - Stealing parts of a car, such as the car battery or the tires
  - Stealing a bicycle
  - Stealing from a building where the offender is allowed to be in (and is not counted already as shoplifting)
  - Stealing from a "coin operated machine" which is mainly vending machines
  - All other thefts

- Currency
  - This includes all money and signed documents that can be exchanged for money (e.g. checks). Blank checks and credit and debit cards are not included (they are in the Miscellaneous/other category)
- Jewelry and "precious metals"
  - Only metals that are considered high value are included here. Metals that are generally worth little are counted in the Miscellaneous/other category.
- Clothing and fur
  - This also includes items that you take with you when leaving the house (except for your phone): wallet, shoes, purse, backpacks.
- Motor vehicle stolen in current agency's jurisdiction
  - This includes only vehicles than can be driven on wheels so excludes trains and anything on water or that can fly.
- Office equipment and electronics
  - This includes "typewriters" and "magnetic tapes" but is essentially any kind of equipment needed to run a business. So printers, computers, cash registers, computer equipment like a monitor or a mouse, and computer software. These items do not have to be stolen from a commercial building to be included in this category.
- Sound and picture equipment
  - This is a kind of odd category that is a product of its time. Anything that produces noise or pictures (including the fancy motion pictures) is included. This includes TVs, cameras, projectors, radios, MP3 players (but not phones that can play music) and (since again, this is a very old dataset) VHS cassettes.
- Guns
  - This includes all types of firearms other than toys or BB/pellet/paintball guns.
- Home furniture
  - This includes all of the "big things" in a house: begs, chairs, AC units, washer/dryer units, etc. However, items that are in the "Office equipment and electronics" category do not apply.
- Consumable goods

  - This is anything that can be consumed such as food, drinks, and drugs, or anything you use in the bathroom.

- Livestock
  - This is all animals other than ones that you would consider a pet

- Miscellaneous/other
  - Anything that is not part of the above categories would fall in here. Cell phones and credit cards are included.

4.2.3 Value of recovered property by type of item stolen

For the below subset of items stolen, this data includes the value of the items that were recovered. This is a subset of the section above - the value of property stolen. I don't know why but this data has more values for the property stolen than for property recovered. For example, we know how much was stolen during bank robberies but not how much was recovered from bank robberies. The only info we have for the value of recovered property is for the breakdown in the items themselves - not breakdowns of crimes such as robbery or burglary. So we can know the value of guns recovered, but not if those guns were taken from a burglary, a robbery, a theft, etc.

While this dataset began in 1960, the recovered property variables begin later, and in different years. For clothing and fur, currency, jewels and precious metals, motor vehicles, miscellaneous/other, and the variable that sums up all of the recovered property variables, the first year with data was 1961. The remaining variables - consumable goods, guns, household goods/home furniture, livestock, office equipment and electronics, and sound and picture equipment - began in 1975.

Another issue is that it uses the value of the property as it is recovered, not as it is stolen. For example, if someone steals a laptop that's worth $1,000 and then the police recover it damaged and it's now worth only $200, we'd see $1,000 for the stolen column for "office equipment and electronics" and only $200 for the recovered column for that category. So an agency that recovers 100% of the items that were stolen can appear to only recover a fraction of them since the value of recovered items could be substantially lower than the value of stolen items - and there's no way to know how many items are actually recovered, we must rely on the value of stolen and recovered.[3]

The full list of items recovered and their definitions are below:

[3]Even if we look at the Offenses Known and Clearances by Arrest data, that only says if there was an arrest or exceptional clearance in the case, not if the property stolen was recovered

- Currency
  - This includes all money and signed documents that can be exchanged for money (e.g. checks). Blank checks and credit and debit cards are not included (they are in the miscellaneous/other category)
- Jewelry and "precious metals"
  - Only metals that are considered high value are included here. Metals that are generally worth little are counted in the miscellaneous/other category.
- Clothing and fur
  - This also includes items that you take with you when leaving the house (except for your phone): wallet, shoes, purse, backpacks.
- Motor vehicle stolen in current agency's jurisdiction
  - This includes only vehicles than can be driven on wheels so excludes trains and anything on water or that can fly.
- Office equipment and electronics
  - This includes "typewriters" and "magnetic tapes" but is essentially any kind of equipment needed to run a business. So printers, computers, cash registers, computer equipment like a monitor or a mouse, and computer software. These items do not have to be stolen from a commercial building to be included in this category.
- Sound and picture equipment
  - This is a kind of odd category that is a product of its time. Anything that produces noise or pictures (including the fancy motion pictures) is included. This includes TVs, cameras, projectors, radios, MP3 players (but not phones that can play music) and (since again, this is a very old dataset) VHS cassettes.
- Guns
  - This includes all types of firearms other than toys or BB/pellet/paintball guns.
- Household goods/Home furniture
  - This includes all of the "big things" in a house: begs, chairs, AC units, washer/dryer units, etc. However, items that are in the "Office equipment and electronics" category do not apply.
- Consumable goods

  - This is anything that can be consumed such as food, drinks, and drugs, or anything you use in the bathroom.

- Livestock
  - This is all animals other than ones that you would consider a pet

- Miscellaneous/other
  - Anything that is not part of the above categories would fall in here. Cell phones and credit cards are included.

4.3 Data errors

This dataset comes with a considerable number of data errors - basically enormous valuations for stolen property. Some of the values are so big that it is clearly an error and not just something very expensive stolen. However, this gets tricky since there can in fact simply be very expensive items stolen and make it look like an error. Some of the stolen property include variables for both the number of items of that type stolen and the total value of the items. From this we can make an average value per item stolen which can help our understanding of what was stolen. However, some items only have the value of the property stolen and the value of property recovered so we actually don't know how many of those items were stolen. These cases make it even harder to believe that a large value is true and not just a data error since we don't know if the number of these thefts increased, causing the increase in the value reported. We'll look at a couple examples of this and see how difficult it can be to trust this data.

First, we'll look at the value of livestock thefts in New York City. Livestock is one of the variables where we know the value stolen and recovered but not how many times it happened. Being a major urban city, we might expect that there are not many livestock animals in the city so the values should be low. Figure @ref(fig:nycLivestock) shows the annual value of livestock thefts in NYC. There are two major issues here. First, in all but two years they report $0 in livestock thefts. This is likely wrong since even New York City has some livestock (even just the police horses and the horse carriages tourists like) that probably got stolen. The second issue is the massive spike of reported livestock theft value in 1993 with over $15 million stolen (the only other year with reported thefts is 1975 with $87,651 stolen). Clearly NYC didn't move from $0 in thefts for decades to $15 million in a year and then $0 again so this is a blatant data error.

It gets harder to determine when a value is a mistake when it is simply a big spike - or drop - in reported value in a dataset that otherwise looks correct. Take, for example, the

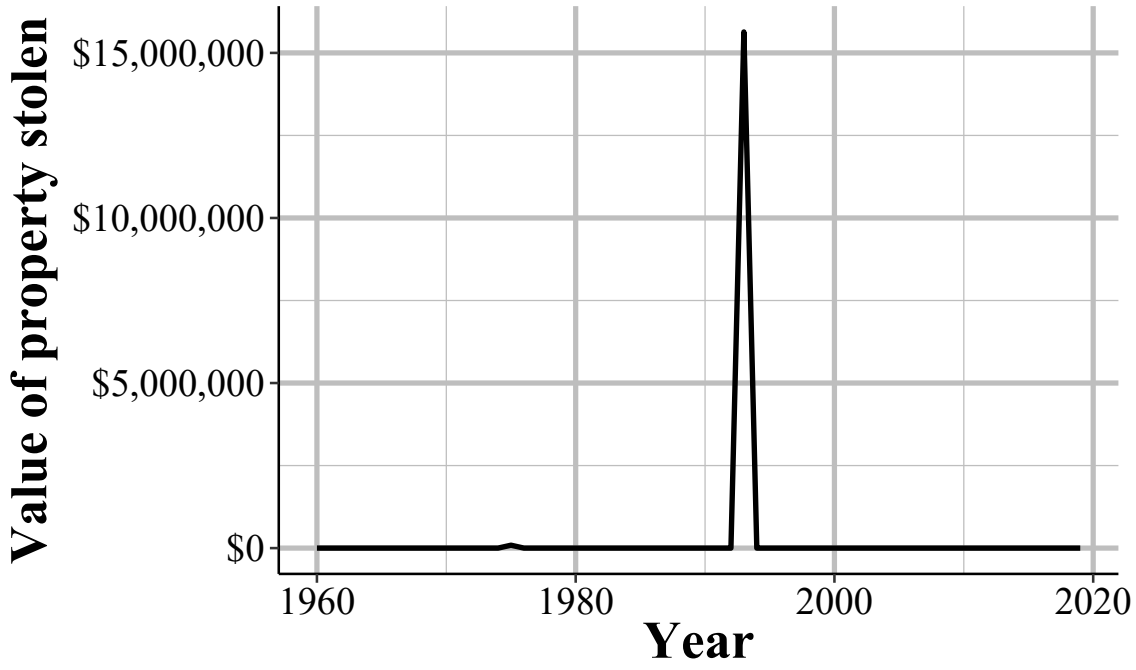


Figure 4.7: The annual value of stolen livestock in New York City, 1960-2019. (#fig:nycLivestock)

annual value of stolen clothing and fur in Philadelphia from 1960-2019, shown in Figure @ref(fig:phillyFurValue). The annual value of these stolen items more than doubled in 1989 compared to the previous year and then declined rapidly in the following year. Is this real? Is it a data error? It's really hard to tell. Here we don't know how many clothing/fur thefts there were, only the value of the total thefts that month (which is aggregated annually here). It continues a multi-year trend of increasing value of thefts but it larger than previous increases in value. And while the spike is very large in percent terms, it's not so extraordinarily huge that we immediately think it's an error - though some outlier detection methods may think so if it's beyond the expected value for that year.

It's also important to have some understanding of what the data *should* look like when trying to figure out what data point may be incorrect. In this figure we see a huge spike in 1989. If we know, for example, that a ring of fur thieves were active this year, then that makes it far more likely that the data is real. This may be a rather odd example, but it is helpful to try to understand the context of the data to better understand when the "weird" data is an error and when it's just "weird but correct".

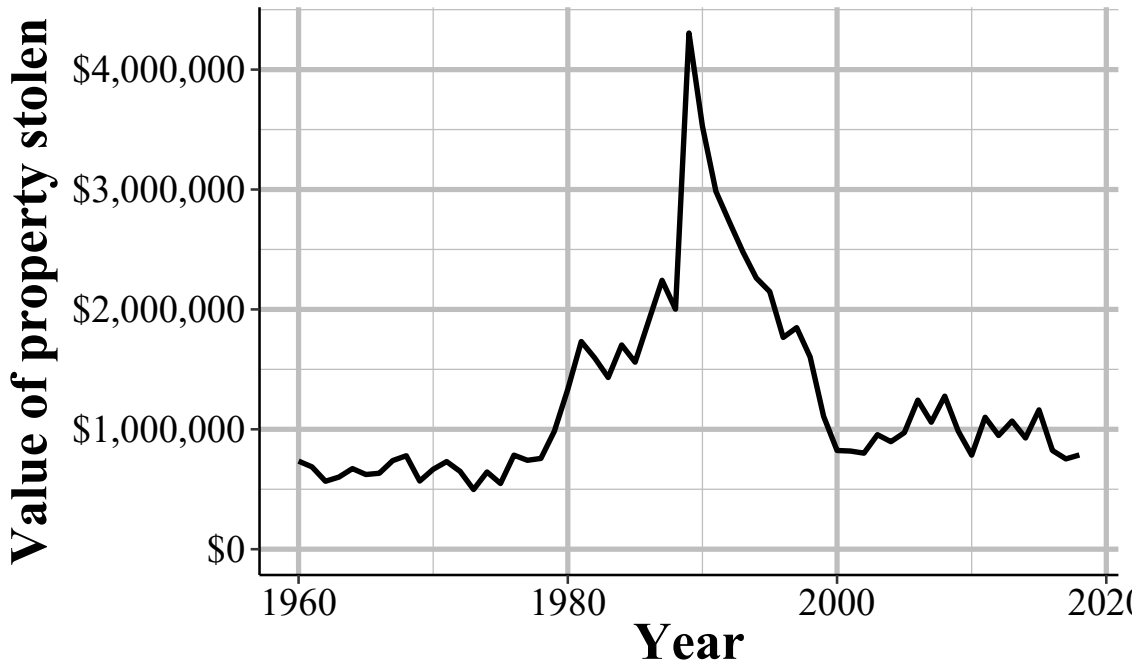


Figure 4.8: The annual value of stolen clothing and fur in Philadelphia, PA, 1960-2019 (#fig:phillyFurValue)

Chapter 5

Arrests by Age, Sex, and Race

The Arrests by Age, Sex, and Race dataset - often called ASR, or the "arrests data", or sometimes the "Arrests by Age, Sex, Race, and Ethnicity (ASRE) though this is really misleading since most years don't even report ethnicity data - includes the monthly number of arrests for a variety of crimes and, unlike the crime data, breaks down this data by age and gender. This data includes a broader number of crime categories than the crime dataset (the Offenses Known and Clearances by Arrest data) though is less detailed on violent crimes since it does not breakdown aggravated assault or robberies by weapon type as the Offenses Known data does. For each crime it says the number of arrests for each gender-age group with younger ages (15-24) showing the arrestee's age to the year (e.g. age 16) and other ages grouping years together (e.g. age 25-29, 30-34,"under 10"). It also breaks down arrests by race-age by including the number of arrestees of each race (American Indian, Asian, Black, and White are the only included races) and if the arrestee is a juvenile (<18 years old) or an adult. The data does technically include a breakdown by ethnicity-age (e.g. juvenile-Hispanic, juvenile-non-Hispanic) but almost no agencies report this data (for most years zero agencies report ethnicity) so in practice the data does not include ethnicity. As the data includes counts of arrestees, people who are arrested multiple times are included in the data multiple times - it is not a measure of unique arrestees.

5.1 Agencies reporting

This data is available from 1974 through 2019 and Figure @ref(fig:arrestsAgenciesReporting) shows how many agencies reported at least one month of the year for each of these years. I'm not sure why there's a dip in 1980. Since it immediately reverses itself in the next year I think it's just a data issue, not a real decrease in the number of agencies that report. The first year of data has about 9,000 agencies reporting and that increases strongly to a little

over 13,000 in 1979. Following the odd blip in 1980, the number of agencies remain steady for nearly a decade before declining to a local low of about 11,000 in 1998 before again increasing steadily until the end of our data where nearly 15,000 agencies report. This 15,000, however, still remains under the estimated 18,000 police agencies in the United States and below the reporting rates of UCR data such as the Offenses Known and Clearances by Arrest data. This data is also missing some important cities such as New York City which hasn't reported even a single month since 2002 and Chicago which tends to only report a single month if at all.

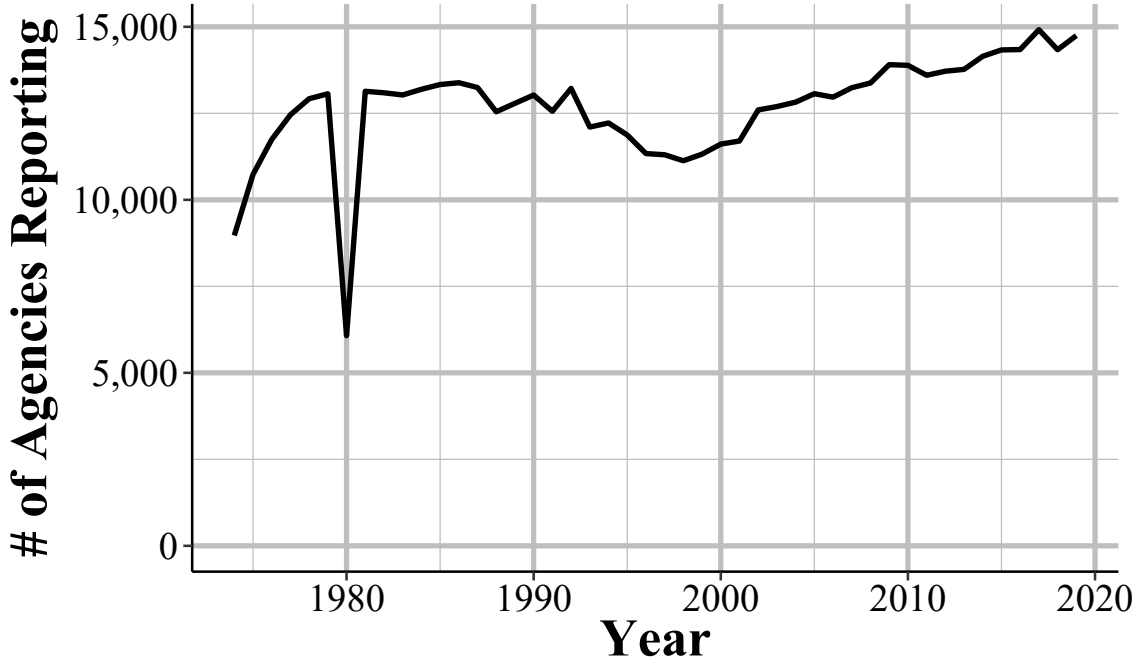


Figure 5.1: The annual number of agencies reporting at least one month of data in that year. (#fig:arrestsAgenciesReporting)

If we look at agencies that report all 12 months of the year, as seen in Figure @ref(fig:arrestsAgenciesReporting12Months), now far fewer agencies report. In almost every year about 45-55% of agencies that report at all report all 12 months. This means that we're missing data for about half of agencies. When an agency reports at all, they tend to report more months than fewer. For example, there are about twice as many agency-years with 11 months reported than with only 1 month reported. So having only about half of agencies report all 12 months of the year doesn't mean that we're missing a ton of data from the remaining half of agencies - but does merit close attention to which agencies you use in your research and how much missingness there is for those agencies.

Another issue is that agencies can report only some crimes. So, for example, they may report how many people were arrested for theft but not for murder, and it's unclear when that means that truly zero people were arrested for murder and when the agency just didn't

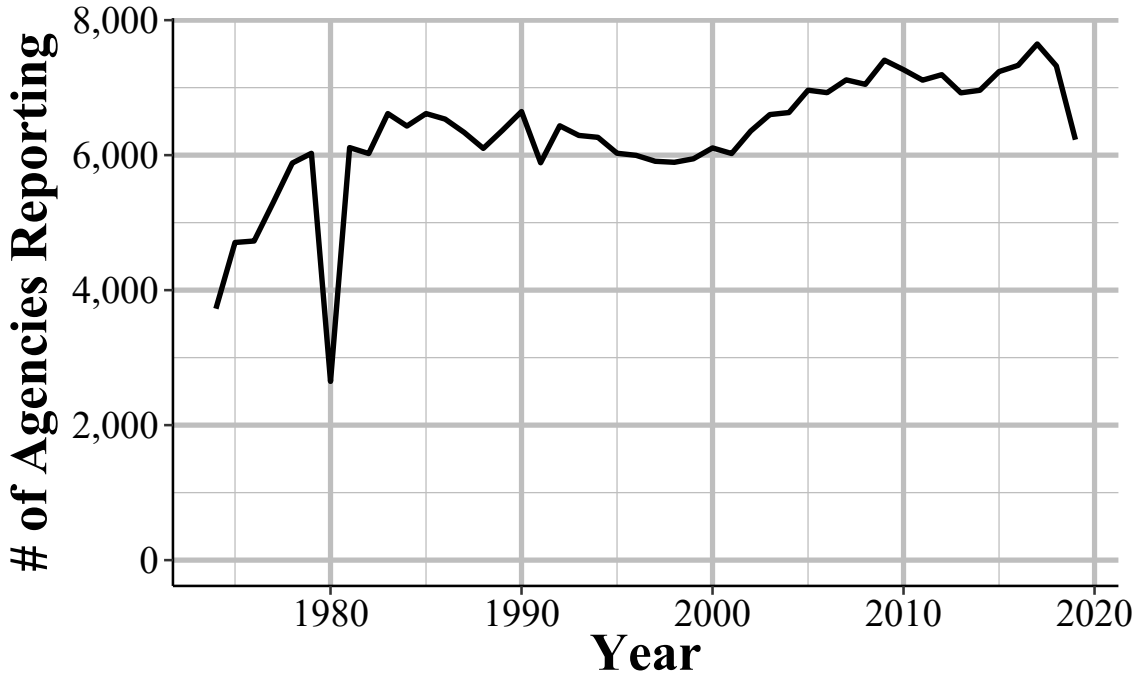


Figure 5.2: The annual number of agencies reporting 12 months of data in that year. (#fig:arrestsAgenciesReporting12Months)

report. This isn't in the original data but in my version of this data (available here), I added variables for the number of months reported (based on the agency reporting at least arrest for that crime in that year) for each crime category. Again, please note that when there are fewer than 12 months reported in that year for that variable, that could simply mean that there were no arrests for that crime that year (The FBI does tell agencies not to even submit a report in these cases) - but we don't know that for sure.

5.2 What is an arrest? (what unit is this data in?)

A key part of understanding this data is knowing what unit of analysis it's in. This data is the monthly number of *total arrests for a given crime, with only the most serious crime per incident included*. Consider for example, a person who robs a bank, shooting and killing a guard and pointing their gun at other people in the bank. They are arrested and then released from jail (just imagine that this is true) and are rearrested the next day for shoplifting. And let's further assume that both arrests were in the same month in the same agency. How many arrests are here? They committed multiple crimes in the first incident (murder, robbery, aggravated assault) but in this dataset they would only be classified as an arrest for the most serious crime, murder. And then separately they'd also be an arrested for shoplifting. So assuming that no other arrests occurred in that police agency that month, there would

be two arrests reported: one for murder and one for shoplifting.

There's no way to tell how many unique people were arrested, or of those arrested multiple times which crimes they were arrested for. So if you have 100 arrests there may be 1 person arrested 100 times or 100 people arrested once - though, of course, the true number is likely somewhere in between. This means that studies that try to use this data as a measure of unique people or even the percent of arrestees by group (age, gender, or race) relative to some base rate of the population such as the number of people living in that city are going to be wrong - though how wrong is unclear.[1]

Common uses of this data - more common in more news articles or advocacy group reports than in peer-reviewed research articles - compare the percent of arrestees of a certain group to the percent of a city's population of that group. Any differences between the arrestee percent and the resident percent is, according to these reports, evidence of a disparity (this is most common for looking at differences by race). Since these analyses are generally looking at annual data, it assumes that people of each group (usually they look at race but gender and age data is also available) are rearrested at the same rate. That is, White people, for example, are rearrested in the same year for the same crime at the exact same rate as Black people. If not, then you're be comparing different things since one group would have more overall people arrested while the other would have fewer people but who are arrested more frequently. Whether this distinction between arrests and unique people arrested affects your interpretation of the data depends on the study you are doing, but it's important to consider in your research. One way to address this is to use other data on the rate of rearrest by group, though you'd have to be very careful to not extrapolate the results of the other study too far beyond what they could tell you of the specific time and place they studied.

5.2.1 The Hierarchy Rule

In incidents where the arrestee commits multiple crimes, this data uses something called the Hierarchy Rule which says that only the most serious crime is counted as the crime that the person was arrested for. For a comprehensive overview of the Hierarchy Rule, please see Section @ref(hierarchy). Basically, the FBI chose seven crimes in 1929 that they call Index Crimes - or sometimes called Part I crimes - and these were considered the most important crimes to be recorded. For more on Index Crimes, please see Section @ref(indexCrimesOffensesKnown). If a person is arrested for multiple crimes and an Index Crime is one of those crimes, then the Index Crime at the top of the Hierarchy is the one recorded in this data. Section @ref(arrestsCrimesIncluded) shows all crimes included and

[1] While all studies are going to be estimates of the real effect, that's no reason to be flippant in using data (UCR and other data). Without having a high level of confidence that your estimates are close to the actual value, you shouldn't do that study.

the crimes 1-7 as well as 9 (arson) are the Index Crimes. The top of the Hierarchy is the crime with the lowest number. So murder is always reported in incidents where there's a murder; rape is always reported when there's an incident with rape but no murder; etc.

In incidents where the arrestee committed both an Index Crime and a Part II crime, then only the top Index Crime is recorded. This can lead to rather silly results since some Part II crimes are certainly more serious than some Index Crimes. Consider, for example, a person arrested for simple assault, carrying a firearm, pimping, and of theft. The first three crimes are, in my opinion, clearly more serious than theft. But since theft is an Index Crime, this person would be considered to have been arrested for theft.

The remaining crimes are called Part II crimes and are not arranged in any particular way. So a lower value numbered crime is not higher on the Hierarchy than a higher value number - Part II crimes don't follow the Hierarchy. If all of the crimes in an incident are Part II crimes then the agency must decide for themselves which crime is the most serious. This can lead to agencies deciding their own hierarchy differently than others which makes this data much less comparable across agencies than if there was a standard rule.

5.3 Crimes included

1. Homicide

- Murder and non-negligent manslaughter
- Manslaughter by negligence

2. Rape
3. Robbery
4. Aggravated assault
5. Burglary
6. Theft (other than of a motor vehicle)
7. Motor vehicle theft
8. Simple assault
9. Arson
10. Forgery and counterfeiting
11. Fraud
12. Embezzlement
13. Stolen property - buying, receiving, and possessing
14. Vandalism
15. Weapons offenses - carrying, possessing, etc.

16. Prostitution and commercialized vice
17. Sex offenses - other than rape or prostitution
18. Drug abuse violations - total

- Drug sale or manufacturing
  - Opium and cocaine, and their derivatives (including morphine and heroin)
  - Marijuana
  - Synthetic narcotics
  - Other dangerous non-narcotic drugs
- Drug possession
  - Opium and cocaine, and their derivatives (including morphine and heroin)
  - Marijuana
  - Synthetic narcotics
  - Other dangerous non-narcotic drugs

19. Gambling - total

- Bookmaking - horse and sports
- Number and lottery
- All other gambling

20. Offenses against family and children - nonviolent acts against family members. Includes neglect or abuse, nonpayment of child support or alimony.
21. Driving under the influence (DUI)
22. Liquor law violations - Includes illegal production, possession (e.g. underage) or sale of alcohol, open container, or public use laws. Does not include DUIs and drunkenness.
23. Drunkenness - i.e. public intoxication
24. Disorderly conduct
25. Vagrancy - includes begging, loitering (for adults only), homelessness, and being a "suspicious person."
26. All other offenses (other than traffic) - a catch-all category for any arrest that is not otherwise specified in this list. Does not include traffic offenses. Very wide variety of crimes are included - use caution when using!
27. Suspicion - "Arrested for no specific offense and released without formal charges being placed."
28. Curfew and loitering law violations - for minors only.
29. Runaways - for minors only.

5.4 Important variables

This data has the standard set of variables describing the agency that is reporting. This includes the agency ORI - which is the unique ID for that agency - the agency name, their state, the population under their jurisdiction, and the month and year of the data.

For each crime this data provides the number of arrests in that month (or year for the annual data) broken down by age (within this, by gender), by race (within this, by if they are a juvenile or an adult), and by ethnicity though this is an enormously flawed variable. Finally, we also know the number of juvenile arrests that ended in a few possible outcomes, though we don't know the crime that led to these arrests. We'll get into each of these variables below.

5.4.1 Age

For each crime the data provides the number of people of each gender by age, with several years in the peak offending age given as the specific age and younger and older ages broken into groups. Only female and male genders are available, and there is no variable for "unknown" gender. So to get a total arrests for that crime for that age, just add the female and male variables together. Below are the ages or age categories included in the data, and these are the same for female and male arrestees.

- Female
  - Under 10
  - 10-12
  - 13-14
  - 15
  - 16
  - 17
  - 18
  - 19
  - 20
  - 21
  - 22
  - 23
  - 24
  - 25-29
  - 30-34
  - 35-39

  - 40-44
  - 45-49
  - 50-54
  - 55-59
  - 60-64
  - 65 and older

- Male
  - Under 10
  - 10-12
  - 13-14
  - 15
  - 16
  - 17
  - 18
  - 19
  - 20
  - 21
  - 22
  - 23
  - 24
  - 25-29
  - 30-34
  - 35-39
  - 40-44
  - 45-49
  - 50-54
  - 55-59
  - 60-64
  - 65 and older

One way to use this data is to look at the age-crime curve of offending. The age-crime curve is a criminological finding that crime trends to grow in the early teenage years to peaking around age 18 before declining sharply. So essentially people commit crime as teenagers and then tend to fizzle out as they get older. Figure @ref(fig:phillyRapeAge) shows this trend for male arrestees of rape in Philadelphia from 1974-2019, which is every year of data we have available. A major problem with this figure is that some of the ages are for single years and some are for age categories. In the graph there were 748 arrests for rape for people aged 24. The next age is the category of aged 25-29 and there were 3,442 arrests for this age group.

One way to address this is to assume that each age in the category has the same number of arrests, so dividing 3,442 by 5 gives us about 688 arrests per age. Assuming equal arrests by age, however, is not consistent with either the literature on the age-crime curve or the findings in this figure for previous ages, as the number of arrests by age is, overall, going down since age 18. So instead of assuming equality, would we assume that older ages have fewer arrests than younger ages (maybe taking the percent change from the previous years where we do have individual ages available)? This is a tricky question to answer and it makes these kinds of analyses really hard to do - and very imprecise since all of your assumptions will be wrong, though hopefully not *too* wrong.

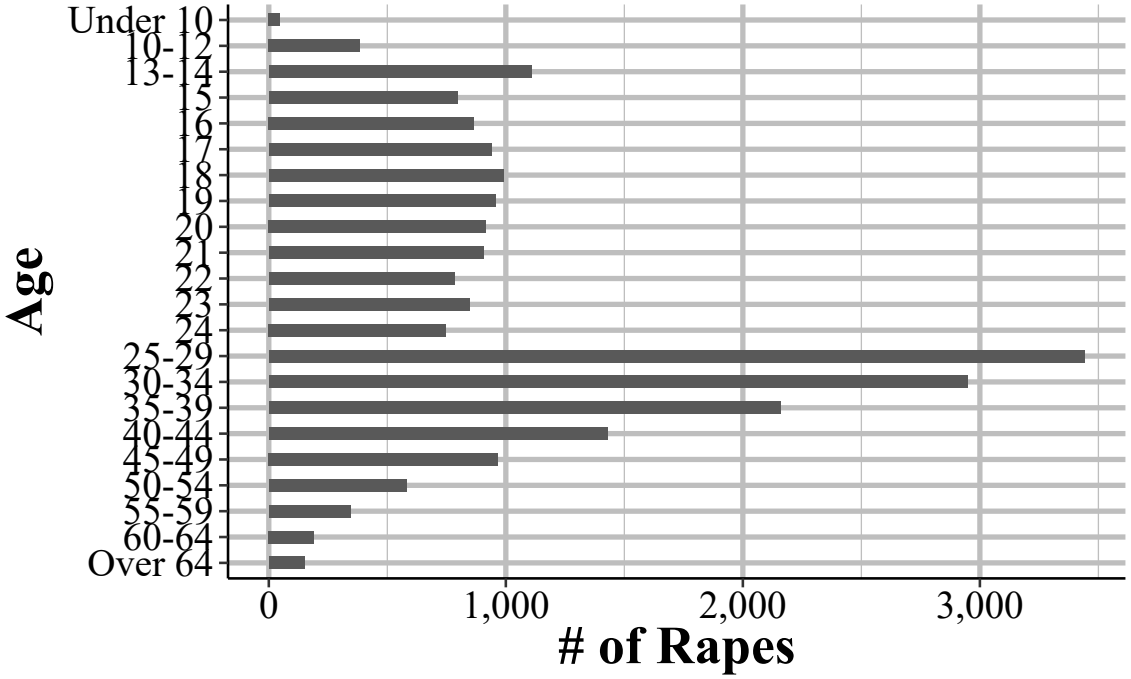


Figure 5.3: The total number of rapes by male arrestees reported by arrestee age in Philadelphia, 1974-2019.
(#fig:phillyRapeAge)

5.4.2 Race

The data also breaks down the number of arrests for each crime by race, with the only included races being American Indian, Asian, Black, and White. This is further broken down into if the arrestee was an adult (18 years or older) or a juvenile (under 18). Whether the arrestee is Hispanic is in a separate (and nearly universally non-reported variable). Since the ethnicity variable is separate, and since the data is not at the arrestee-level unit, there's no way to interact the race and ethnicity variables. So, for example, there is no way to

determine how many White-Hispanic or White-Non-Hispanic arrestees. Just total White arrestees and total Hispanic arrestees.

As with race variables in other UCR datasets - and, really, any dataset - you should be cautious about using this variables since it is the officer's perception of the arrestee's race - though of course some arrests do have other data about the arrestee's race such as what they tell the officer.

Even though there is information about the specific age of arrestee (or the age range, depending on the arrestee's age) and their gender, there is no gender information combined with race and no age beyond the adult/juvenile binary. If you add up all arrests that are broken down by gender-age and compare it to the sum of all of the arrests broken down by adult/juvenile-race here, in some cases these numbers don't add up. That's because while most agencies do report the age variables, not all agencies report the race variables. So summing up the race variables will actually undercount the total number of arrests.

- Adult
  - American Indian
  - Asian
  - Black
  - White
- Juvenile
  - American Indian
  - Asian
  - Black
  - White

Figure @ref(fig:phillyMarijuanaRacePercent) shows one example of an analysis of this data by showing the percent of arrests of adults for marijuana possession by the arrestee's race in Philadelphia, PA for all years of data we have available, 1976-2019. At the bottom are American Indian and Asian arrestees who make up nearly none of the arrests for this crime. Black arrestees, shown in purple, make up the bulk of arrests with only a few years making up under 60% of arrests and growing to around 80% of arrests since the mid-2000s. As White arrestees, shown in black, are the only other race category included, they make up a near perfect mirror image of Black arrestees, composing of around 40% of arrests until decreasing starting in the 1990s to end up with about 20% of arrests in recent years.

Interestingly, while the disparity between Black-White arrests has grown dramatically in recent decades, the total number of arrests have a very different trend as shown in Figure

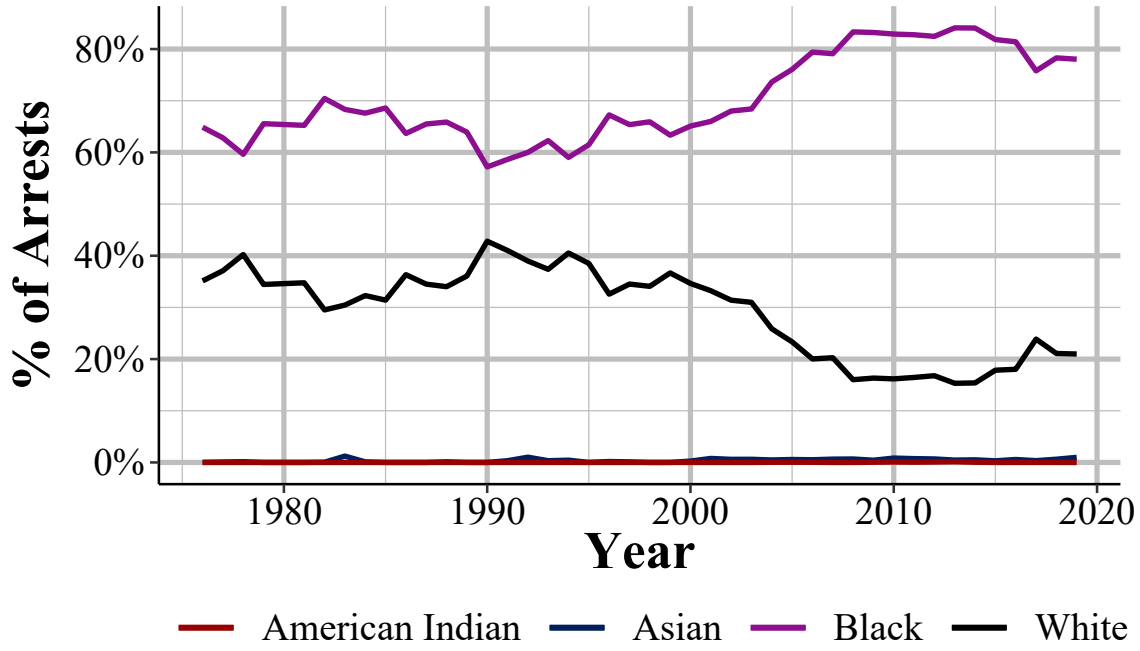


Figure 5.4: The annual percent of adult marijuana possession arrests in Philadelphia by arrestee race, 1976-2019.
(#fig:phillyMarijuanaRacePercent)

@ref(fig:phillyMarijuanaRaceCount). Total marijuana possession arrests declined in the mid-1980s then increased in the mid-1990s from only a few hundred arrests in the early 1990s to nearly 6,000 arrests in the late-2000s before dropping precipitously to about 700 each year in the late-2010s. Yet throughout this latter period as a percent of arrests, Black people consistently grew for years before plateauing around 2007 with a small decline in the last few years. Philadelphia decriminalized marijuana possession in 2014 under Mayor Nutter which is right when the steepest decline in arrests happened, though in the last couple of years also saw a decline in arrests. This suggests that who is arrested, in terms of race, is relatively unrelated to the total number of arrests, at least for marijuana.

5.4.3 Ethnicity

While technically included, the ethnicity variable is largely useless since for most years no agencies reported it and for the years where agencies do report ethnicity, not all agencies do so. The ethnicities included are Hispanic and non-Hispanic are broken down by if the arrestee is an adult (18+ years old) or a juvenile (<18 years old).

- Adult
  - Hispanic

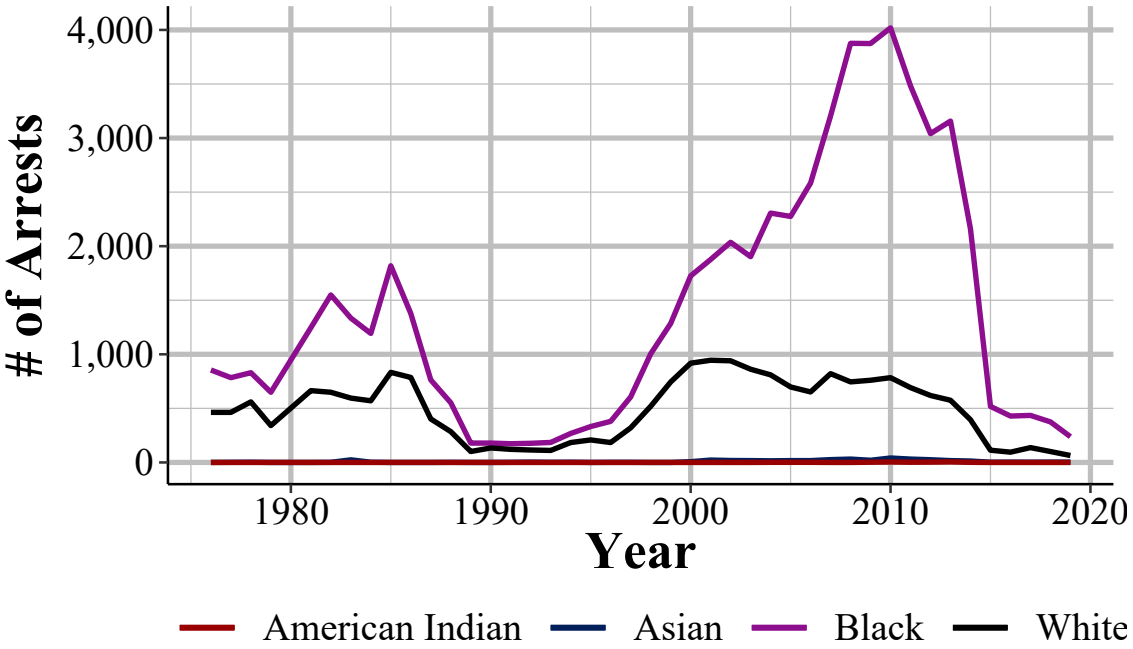


Figure 5.5: The annual number of adult marijuana possession arrests in Philadelphia by arrestee race, 1976-2019.
(#fig:phillyMarijuanaRaceCount)

  - Non-Hispanic

- Juvenile
  - Hispanic
  - Non-Hispanic

Figure @ref(fig:theftHispanic) shows the annual number of Hispanic arrestees for theft for all agencies that reported any data that year.[2] For several years no agencies reported until the number of Hispanic arrestees start climbing in 1980 and peaks in 1986 at about 136,000 arrestees. Then there are zero Hispanic arrestees for a few years, four Hispanic arrestees in 1990, and then again zero Hispanic arrestees, this time for decades. Only in 2017 do the number of Hispanic theft arrestees begin to creep up. From 2017 to 2019 (the last year available at the time of this writing) there are Hispanic arrestees reported every year, though now only about 30,000 per year.

Perhaps a better way to look at this data is to see what percent of agencies report ethnicity data. Figure @ref(fig:theftHispanicPercentAgencies) show the percent of agencies each year that report at least one Hispanic or non-Hispanic (which are the only choices, but showing only Hispanic arrests would exclude agencies where no Hispanic people truly were arrested)

[2]Theft is used as it's one of the most common crimes.

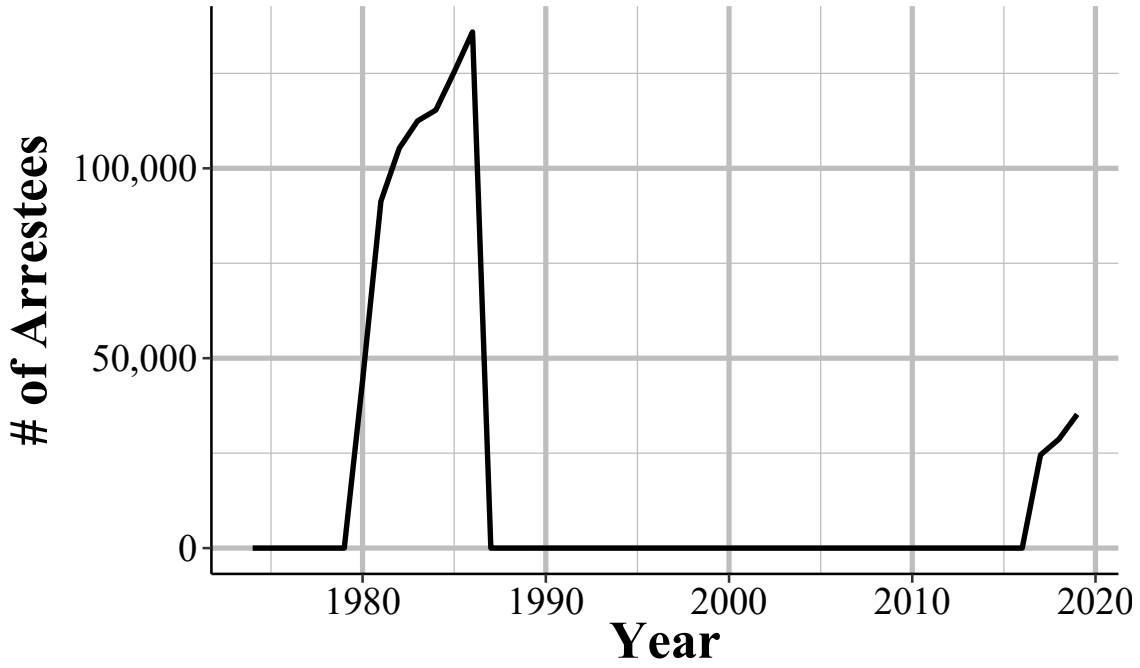


Figure 5.6: The national annual number of Hispanic arrestees for theft. This includes all agencies that year that reporting any number of months. Hispanic arrestees include both juvenile and adult arrestees
(#fig:theftHispanic)

arrest for theft, of all agencies that reported theft data. About 90-95% of agencies reported ethnicity data in the early 80s and then only a couple agencies report in 1990 and 1991. Other than those agencies, none report between 1987 and 2017. Starting in 2017, 36% of agencies report and this number has grown by about five percentage points a year to 46% in 2019. Since fewer than half of agencies currently report, I strongly recommend against using these variables, even for the recent years of data.

5.4.4 Juvenile referrals

The final variable of interest are five mutually exclusive outcomes for juveniles who are arrested by the police for a crime that if they were adults would have been counted as a formal arrest. Unlike the rest of this dataset where juvenile is defined as being under the age of 18, these variables allow states to use their own definition of juvenile. So potentially the limit for who is a juvenile could be below the age of 18, and nothing in the data indicates when this is so - you'd have to check each state to see their definition and if it changed over time. There is no breakdown by crime so this gives you the outcomes for juveniles arrested for all crimes in that agency. Please note that the number of juveniles in other variables and the number here do not always line up, which is a mix of underreporting of this variable, arrests

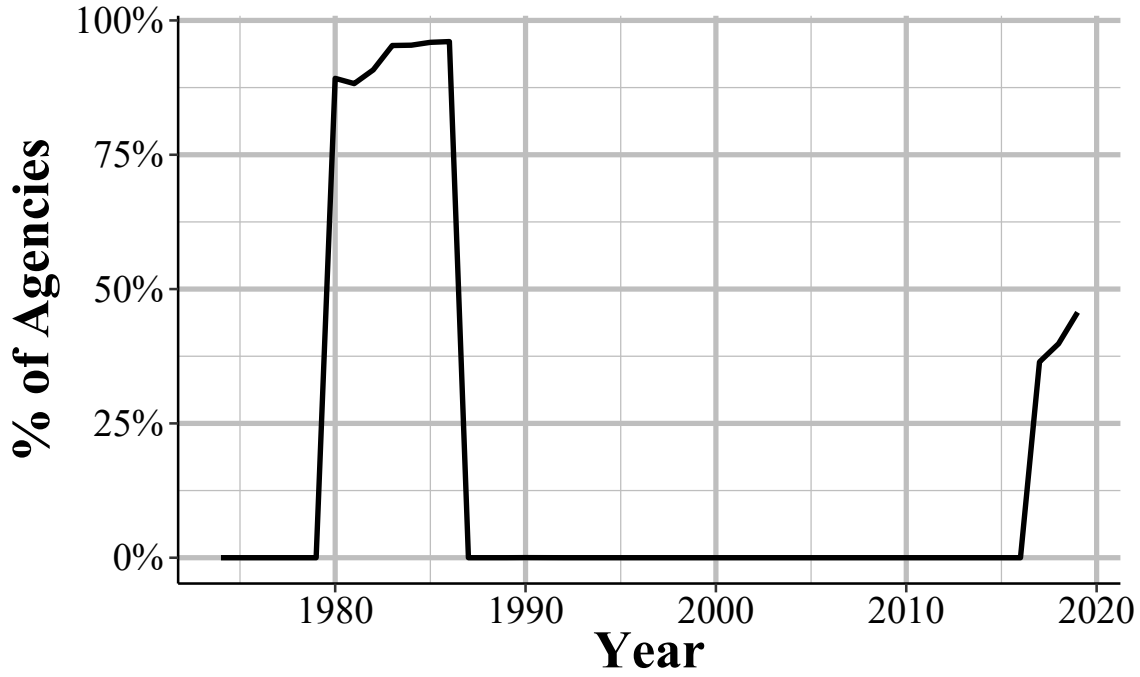


Figure 5.7: The annual percent of agencies that report theft arrests that reported at least one Hispanic person or one non-Hispanic person arrested for theft. Arrestees include both juvenile and adult arrestees.
(#fig:theftHispanicPercentAgencies)

for other jurisdictions are not counted as an arrest in the above variables, and different age definitions for who is a juvenile. A juvenile may potentially get multiple referrals, such as being released and then later referred to court. But in this data only the *initial* referral is included. Below are the five potential outcomes and definitions of each:

- Handled within department and released
  - Juvenile is arrested but then released without any formal charges. Generally released to adult relatives with a warning but no formal charge.
- Referred to juvenile court or probation department
- Referred to welfare agency
- Referred to other police agency
  - This includes when the agency makes an arrest on behalf of a different agency, such as when the juvenile committed a crime in that different agency's jurisdiction. People arrested in this category are also not included in the other variables for juvenile arrests (e.g. arrests by age) as that only includes people who committed a crime in the agency's own jurisdiction.
- Referred to criminal or adult court

– These are juveniles who are referred to be tried in criminal court as adults. This is for states that allow juveniles to be tried as adults. This is the police's recommendation that they be tried as adults, regardless of the decision of the district attorney or court for whether that juvenile is ultimately tried as an adult.

We can look at an example of this in Figure @ref(fig:phillyJuvenileReferrals) which shows the annual number of referral types in Philadelphia from 1974-2019. For all the first couple of years almost all of the referrals have either been that the agency handles the arrest internally and releases the juvenile without any formal charges, or that the juvenile is formally arrested and referred to juvenile court. This is a pretty consistent trend with those two categories being the predominant outcome for juveniles arrested in other agencies and over time. In most years handled internally is the most common outcome though juvenile court is occasionally more common, including in recent years.

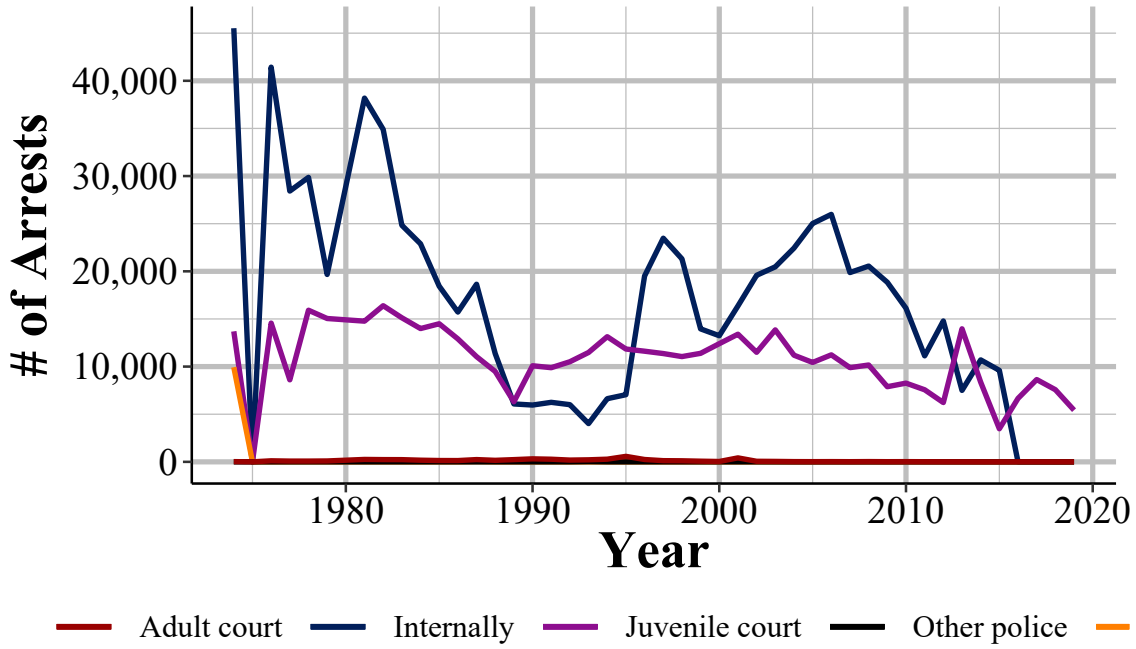


Figure 5.8: The annual number of juvenile referrals in Philadelphia by referral type, 1974-2019.
(#fig:phillyJuvenileReferrals)

Chapter 6

Supplementary Homicide Reports (SHR)

The Supplementary Homicide Reports dataset - often abbreviated to SHR - is the most detailed of the UCR datasets and provides information about the circumstances and participants (victim and offender demographics and relationship status) for homicides.[1] For each homicide incident it tells you the age, gender, race, and ethnicity of each victim and offender as well as the relationship between the first victim and each of the offenders (but not the other victims in cases where there are multiple victims). It also tells you the weapon used by each offender and the circumstance of the killing, such as a "lovers triangle" or a gang-related murder. As with other UCR data, it also tells you the agency it occurred in and the month and year when the crime happened.

6.1 Agencies reporting

This data only has a report when the agency has a murder that year and since murder is rare it is difficult to measure underreporting. One way we can look at reporting is to compare murders in the SHR data with that of other datasets. We'll look at two of them: the Offenses Known and Clearances by Arrest which is covered in detail in Chapter @ref(offensesKnown), and the Center for Disease Control and Prevention (CDC) data on national deaths from homicide.[2] Both this dataset and the Offenses Known and Clearances by Arrest data are UCR datasets so you may think that the numbers of murders from each dataset should be the same. That is a perfectly reasonable assumption, but since this is UCR data we're

[1] If you're familiar with the National Incident-Based Reporting System (NIBRS) data that is replacing UCR, this dataset is the closest UCR data to it, though it is still less detailed than NIBRS data.

[2] CDC WONDER data is available here: https://wonder.cdc.gov/

talking about, you'd be wrong. Police agencies are free to report to either, both, or neither dataset so while the number of murders are close for each dataset, they are never equal. CDC WONDER data aggregates mortality data (among other data) from state death certificates which reduces the issue of voluntary reporting that we have in UCR data.

Figure @ref(fig:shrVsOffenses) shows the annual number of murder victims from each of these datasets starting in 1976 for the UCR data and in 1999 for the CDC data.[3] For the UCR data, in every year the numbers are fairly similar and the trends are the same over time, but the number of murders is never equal. The numbers have actually gotten worse over time with the difference between the datasets increasing and the Offenses Known data having consistently more murders reported than the SHR data since the late 1990s. Compared to the CDC data, however, both UCR datasets - and in particular the SHR data - undercount the number of murders. While trends are the same, UCR data reports thousands fewer murders per year than the CDC data, indicating how much of an issue underreporting is in this data.

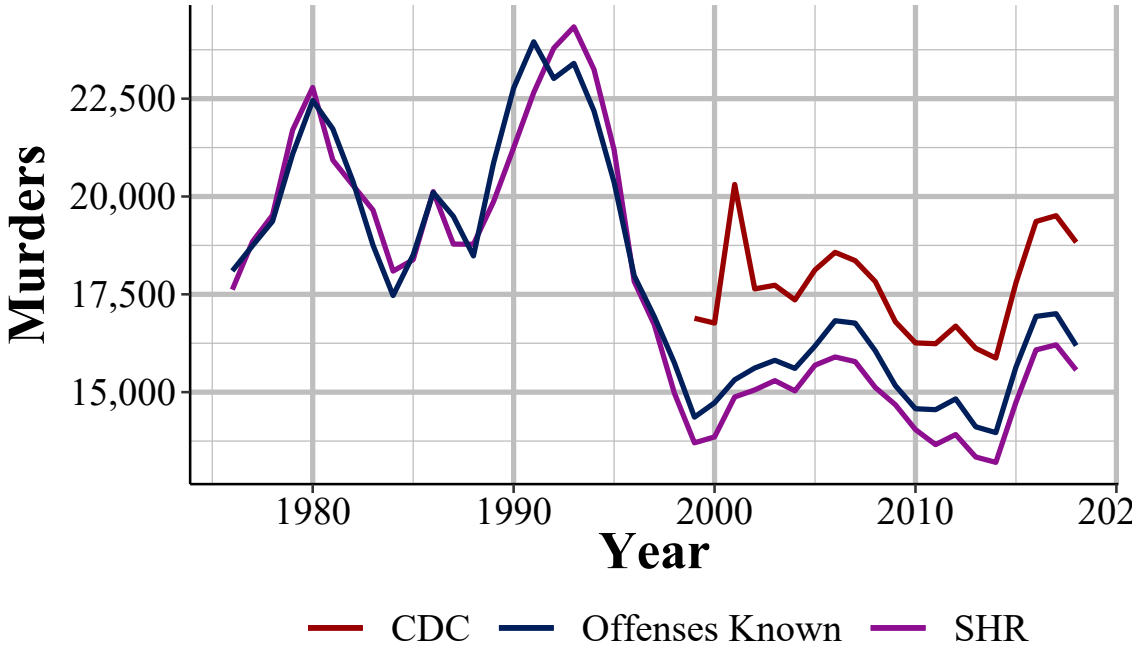


Figure 6.1: The annual number of murders from the Supplementary Homicide Report and the Offenses Known and Clearances by Arrest dataset. Numbers differ because agencies voluntarily report and may not report to both datasets.
(#fig:shrVsOffenses)

[3] 1975 is actually the first year that the Supplementary Homicide Reports data is available but that dataset only has info for a single victim and offender - all later years has info for up to 11 victims and offenders - so 1976 is often used as the first year of data

6.2 Important variables

The data has demographic information for up to 11 victims and 11 offenders, as well as the information on the weapon used by each offender, the relationship between the first victim and each offender, and the circumstance of the murder. The data also has the traditional UCR set of variables about the agency: their ORI code, population, state, region and the month and year of this data. One key variable that is missing is the outcome of the homicide: there is no information on whether any of the offenders were arrested.

While there is information on up to 11 victims and offenders, in most cases, there is only a single victim and a single offender in each incident. To see how the breakdown for the number of victims in each incident looks, Figure @ref(fig:numberSHRVictims) shows the percent of incidents with each possible number of victims.[4] In nearly all incidents - 96.1% - there was only a single victim. This drops to 3.2% of incidents for two victims, 0.5% for three victims, and only about 0.2% of incidents have four or more victims.

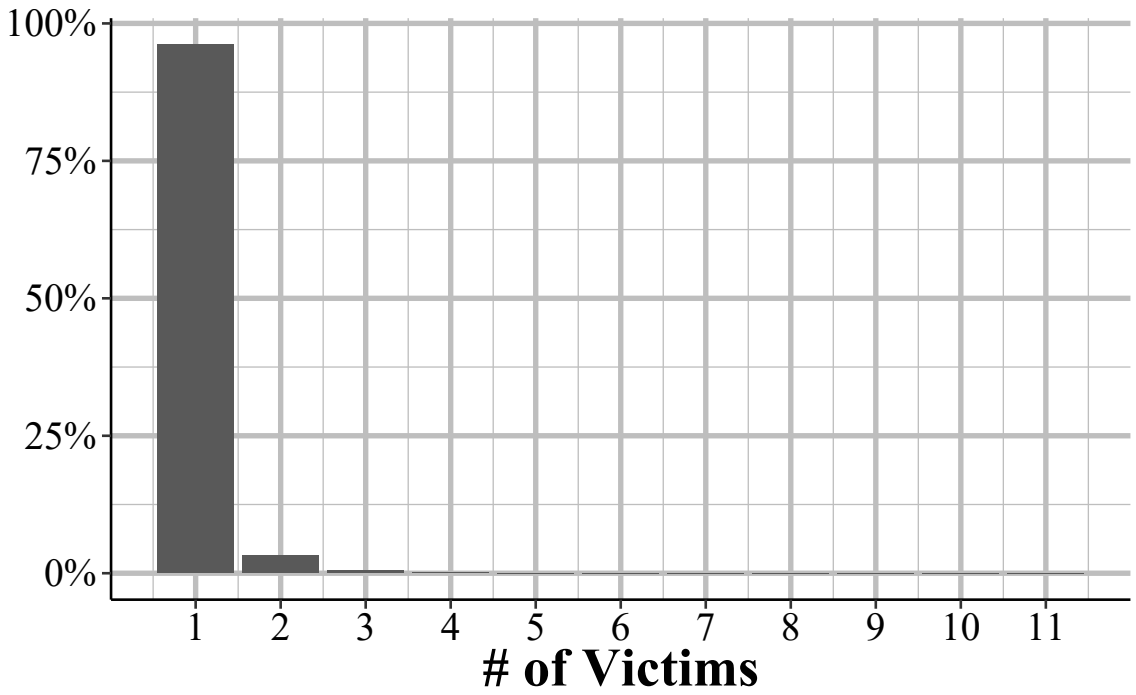


Figure 6.2: The percent of incidents from 1976-2018 that have 1-11 victims. (#fig:numberSHRVictims)

Figure @ref(fig:numberSHROffenders) shows the breakdown of the number of offenders per homicide incident.[5] It's a little less concentrated than with victims but the vast majority of

[4]There is one incident where there are a reported 12 victims. For simplicity of the graph, this incident is not included.

[5]There is one incident where there are a reported 22 offenders and one with 12 offenders. For simplicity of the graph, these incidents are not included.

homicides are committed by one offender - or at least the police only report one offender. About 88% of homicides have only one offender, 8.2% have two, 2.4% have three, and fewer than 1% have four. Fewer than 0.5% of homicides have more than four offenders. However, this is all a bit misleading. In cases where there is no information about the offender, including how many offenders there is, the data simply says that there is a single offender. So the number of homicides with a single offender is an overcount while the number with more offenders is an undercount.

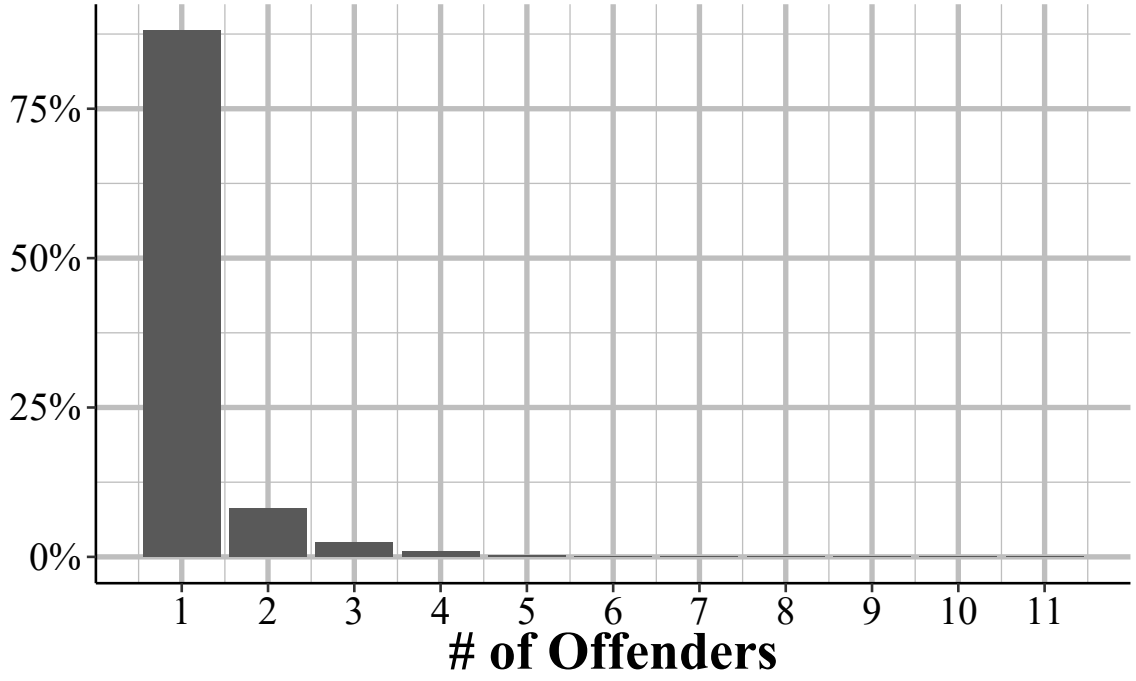


Figure 6.3: The percent of incidents from 1976-2018 that have 1-11 offenders. (#fig:numberSHROffenders)

The variable "situation" says what type of victim-offender number combination the incident is - e.g. "multiple victims/single offender", "single victim/multiple offenders", etc. - and does indicate if the number of offenders is unknown (though curiously there are over 4,000 instances where the number of offenders is unknown but they still say there are two offenders) so you can use this variable to determine if the police don't know how many offenders there is. You're still limited, of course, in that the number of offenders is always what the police think there are, and they may be wrong. So use this variable - and anything that comes from it like the percent of offenders of a certain race - with caution.

We'll now look at a number of important variables individually. Since the data can potentially have 11 victims and 11 offenders - but in practice has only one each in the vast majority of cases - we'll only look at the first victim/offender for each of these variables. Therefore, the results will not be entirely accurate, but will still give you a good overview of the data. The

figures below will use data for all homicides from 1976 to 2018 so will cover all but the most recent year of data.

6.2.1 Demographics

There are two broad categories of variables that we'll cover: demographics of the victim and offenders, and characteristics of the case. We start with demographics.

6.2.1.1 Age

This data includes the age (in years) for each victim and each offender. For those under one years old, it also breaks this down into those from birth to six days old "including abandoned infant" and those seven days old to 364 days old. So there's a bit more info on baby murders. It also maxes out the age at 99 so for victims or offenders older than that we don't get their exact age, just text that says "99 years or older" (which I turn to the number 99 in the figures below).

Figure @ref(fig:shrOffenderAge) shows the percent of murders from 1976-2018 where the first offender in the case is of each age from 0-99. Offenders with unknown ages are excluded from this graph and make up about 27% of cases. The average (mean) age is 30.9 years old (shown in orange) which is due to a long right tail; the median age is 28 years old. If you look closely at the left side of the graph you can see that there are some very young offenders, with at least one offender for each year of age from 0 to 10 included in the data. It's not clear from this alone that these ages are a data entry error. While a two-year-old certainly couldn't murder someone, the data does include deaths caused by "children playing with gun" (homicide circumstances will be discussed in Section @ref(circumstance)) so these ages could potentially be correct.

If you're familiar with the age-crime curve in criminology - which basically says crime peaks in late teen years then falls dramatically - this shows that exact curve, though is older and doesn't decline as the offender ages as quickly as we see with less serious crimes.

Figure @ref(fig:shrVictimAge) repeats Figure @ref(fig:shrOffenderAge) but with victim age rather than offender age. The mean victim age (shown in orange) is 32.9 and the median age is 30. Though the age victim age is a bit younger than the average offender age, trends are relatively similar for teenagers and older where deaths spikes in the late teen years and then declines steadily. The major difference is the U-shape for younger victims - for victims under age 15, homicides peak at age 0 (i.e. younger than their first birthday) with ~1.5% of all homicides being this this age. They then decline until plateauing at around age 6 before increasing again in the early teen years.

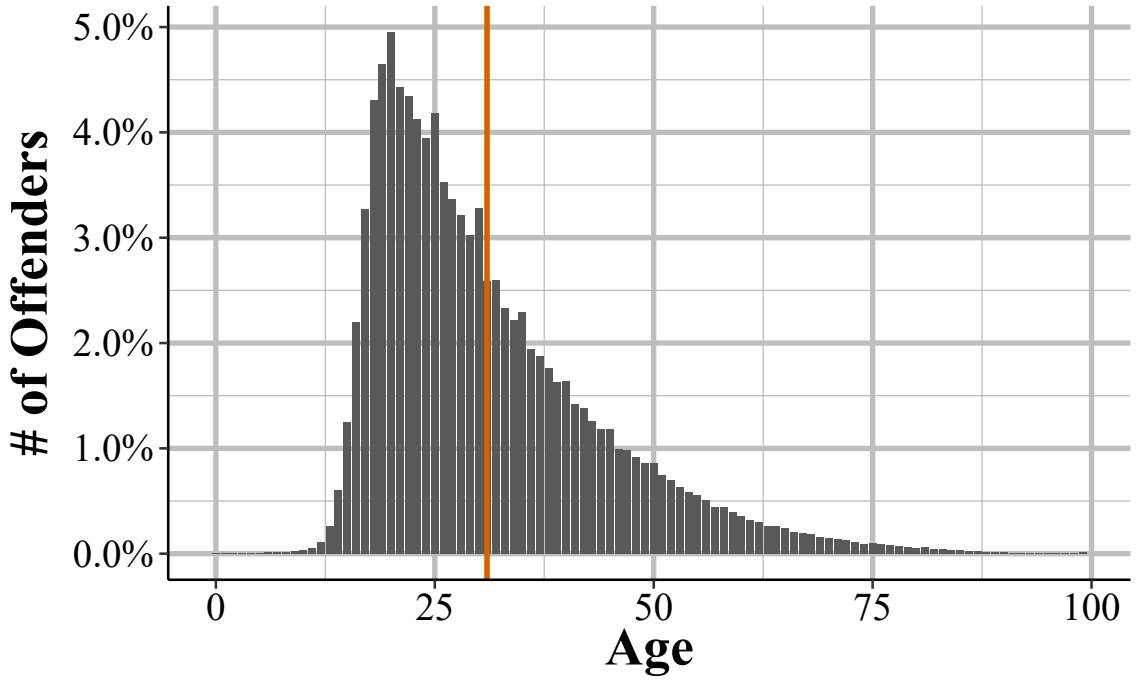


Figure 6.4: The age of homicide offenders, based on the first offender in any homicide incident. Offenders under age 1 (classified as 'birth to 7 days old, including abandoned infant' and '7 days to 364 days old') and considered 0 years old. Offenders reported as '99 years or older' are considered 99 years old.
(#fig:shrOffenderAge)

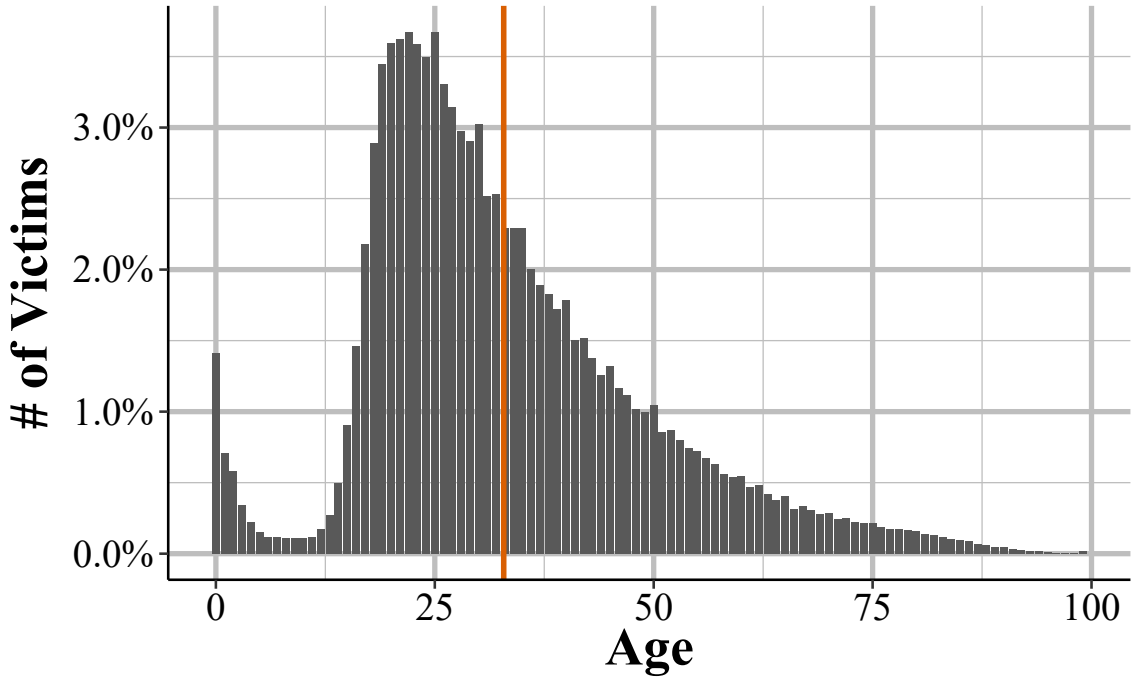


Figure 6.5: The age of homicide victims, based on the first victims in any homicide incident. Victims under age 1 (classified as 'birth to 7 days old, including abandoned infant' and '7 days to 364 days old') and considered 0 years old. Victims reported as '99 years or older' are considered 99 years old.
(#fig:shrVictimAge)

6.2.1.2 Sex

We'll next look at victim and offender sex, a simple variable since only male and female are included. About 62% of offenders, as seen in Figure @ref(fig:shrOffenderSex), are male and about 8% are female, indicating a large disparity in the sex of homicide offenders. The remaining 30% of offenders do not have sex data available because the police do not know the sex of this individual. For offenders who aren't arrested, this variable may be inaccurate since it is perceived sex of the offender.

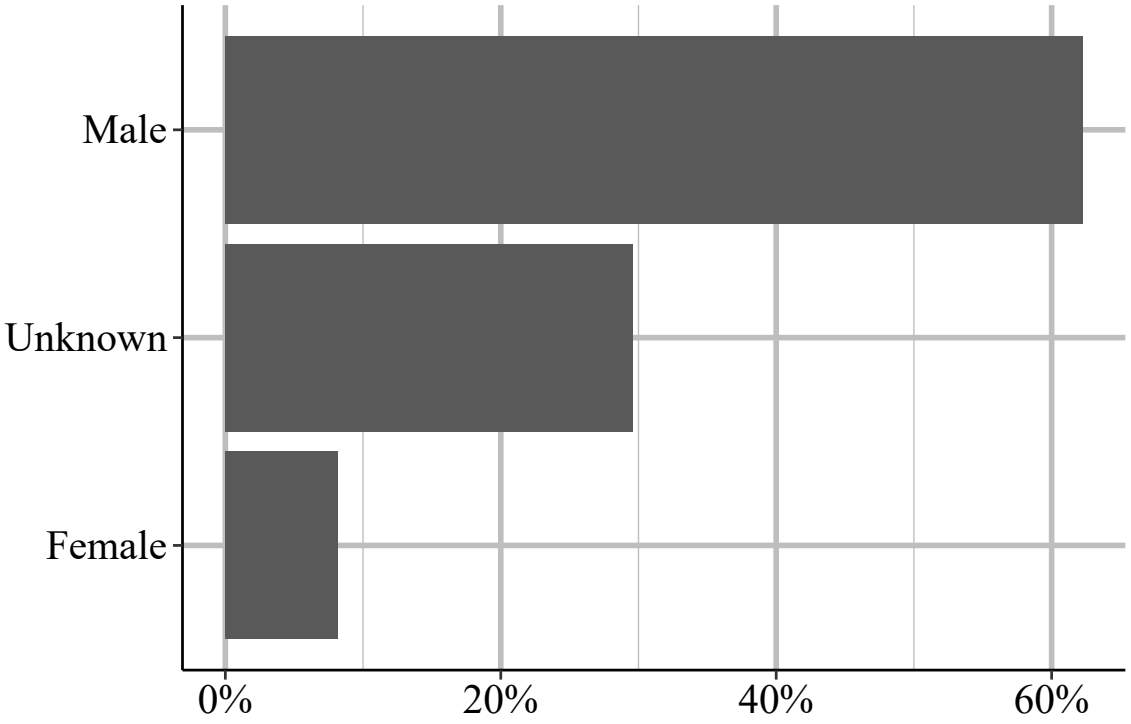


Figure 6.6: The sex of offenders, 1976-2018.
(#fig:shrOffenderSex)

There is far less uncertainty for victim sex, with under 0.15% of victims having an unknown sex. Here again there is a large disparity between male and female with about 78% of victims being male and 22% being female. While there are more male victims than male offenders, this is probably just due to there being so many unknown offenders.

6.2.1.3 Race

This data also includes the race of the victims and offenders. This includes the following races: American Indian or Alaskan Native, Asian, Black, Native Hawaiian or Other Pacific Islander, and White. These are the only races included in the data; Hispanic is considered an ethnicity and is available as a separate, though flawed, variable. There is no category for

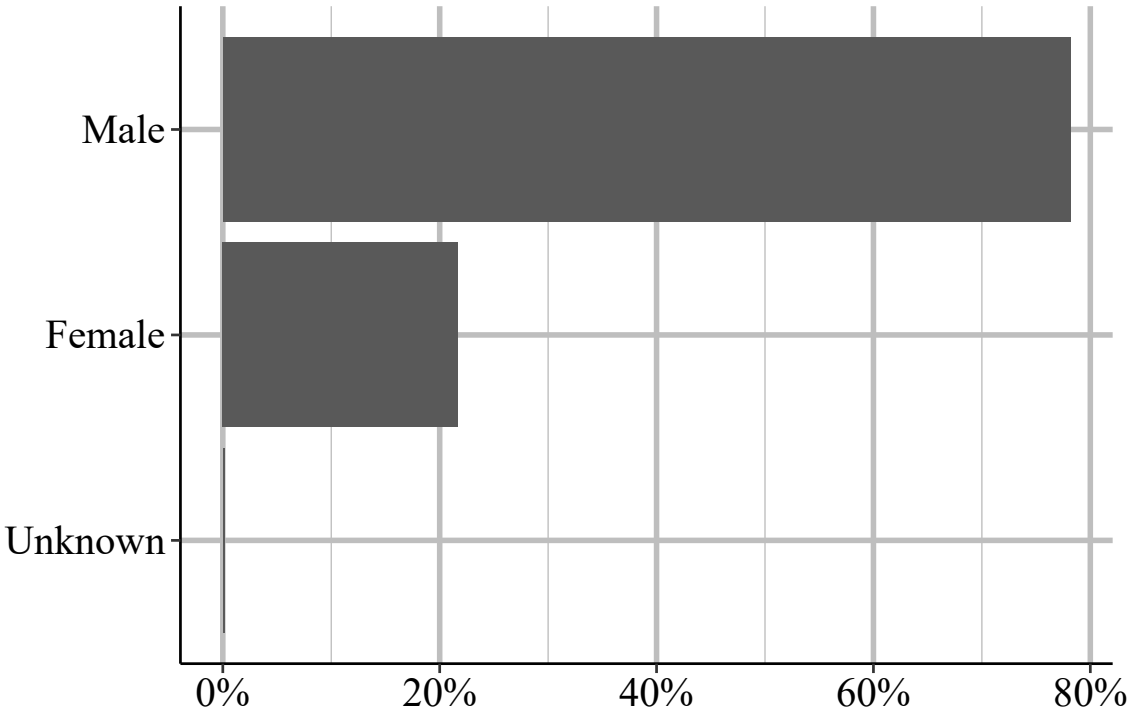


Figure 6.7: The sex of victims, 1976-2018.
(#fig:shrVictimSex)

bi- or multi-racial. As with other demographics info for offenders, in cases where no arrest is made (and we don't know in this data if one is made), there's no way to confirm the person's race (and race itself is hard to put into discrete boxes like done here) so these results may not be entirely accurate.

Figure @ref(fig:shrOffenderRace) shows the percent of homicides in the data by race. Black and White victims are included are similar percentages, at 34.2% and 33.8% of victims, respectively. The next most common group is Unknown at about 30.5% of offenders. Given that so many offenders have an unknown race, the reliability of the other race measures is limited. The remaining races are Asian at 0.9% of offenders, American Indian or Alaskan Native at 0.6%, and Native Hawaiian or Other Pacific Islander at 0.004%.

For victim race, seen in Figure @ref(fig:shrVictimRace), only about 1% of victim races are unknown. This means we can be a lot more confident in the race of the victims than in the race of the offender. Or, at least the challenge of categorizing people by race is the major problem, not missing data As with offenders, White and Black victims are the two most common races, with 49% and 47.7% of victims, respectively. There are almost double the percent of Asian victims than Asian offenders at 1.51% of victims (and 0.9% of offenders). American Indian or Alaskan Natives make up 0.8% of victims while Native Hawaiian or Pacific Islanders make up 0.005% of victims.

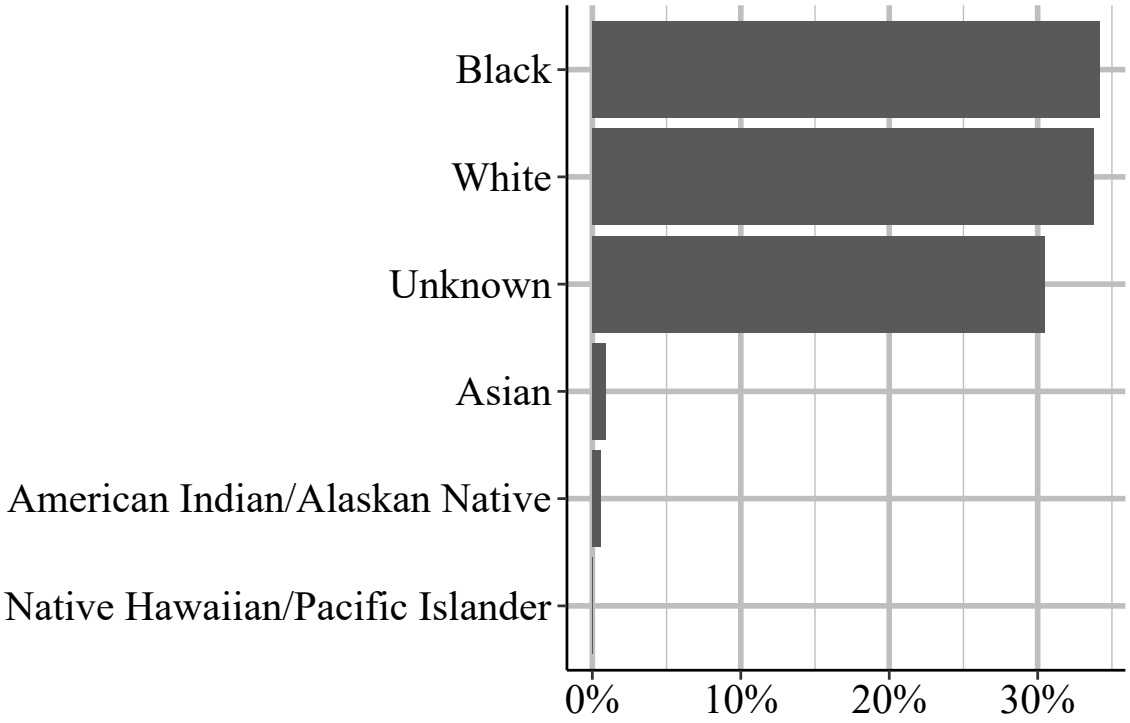


Figure 6.8: The race of offenders, 1976-2018.
(#fig:shrOffenderRace)

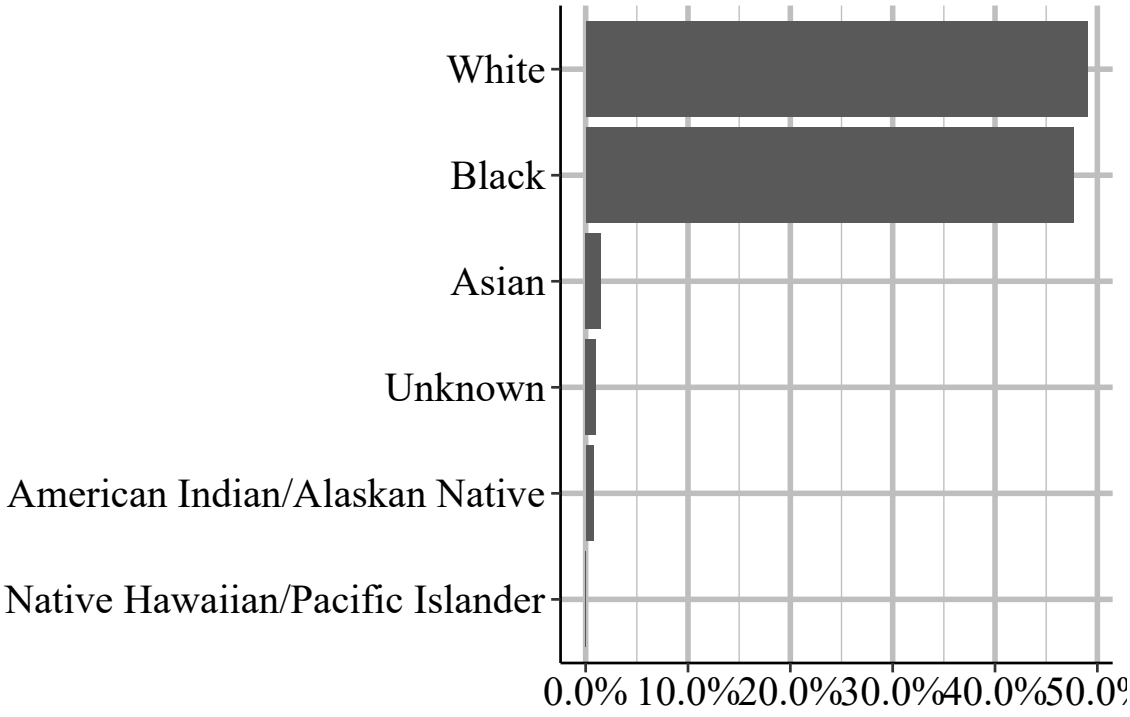


Figure 6.9: The race of victims, 1976-2018
(#fig:shrVictimRace)

6.2.1.4 Ethnicity

The final demographic variable is ethnicity which is whether the victim or offender is Hispanic or not Hispanic. The UCR data has a weird relationship with this variable (which is also in the Arrests by Age, Sex, and Race dataset, discussed in Chapter @ref(arrests)) where ethnicity is technically a variable in the data but very rarely collected. As such, this is an unreliable variable that if you really want to use needs careful attention to make sure it is being reported consistently by the agencies that you are looking at.

The vast majority - 71.6% - of offenders have an unknown ethnicity while 21.5% are not Hispanic and 6.9% are Hispanic.

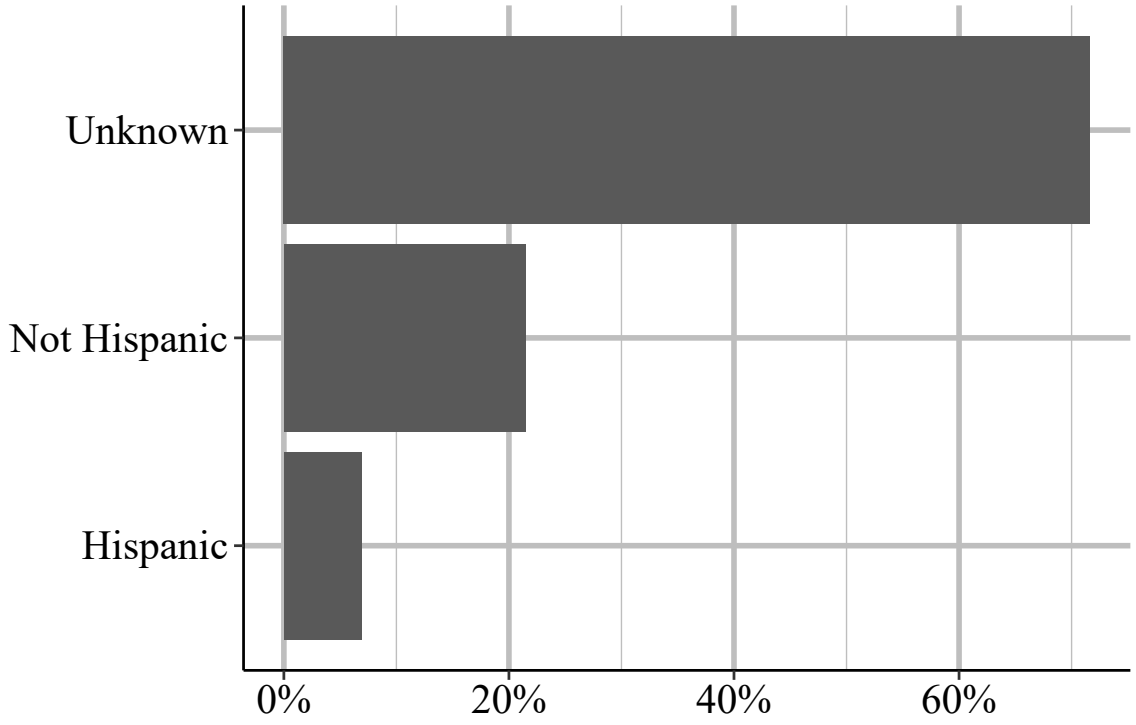


Figure 6.10: The ethnicity of offenders, 1976-2018.
(#fig:shrOffenderEthnicity)

Unlike the other demographic variables, there is still a huge amount of underreporting when it comes to victim ethnicity, though still less than for offender ethnicity. 59% of victims have an unknown ethnicity. Nearly 30% of victims are reported as not Hispanic while 10.8% are reported as Hispanic.

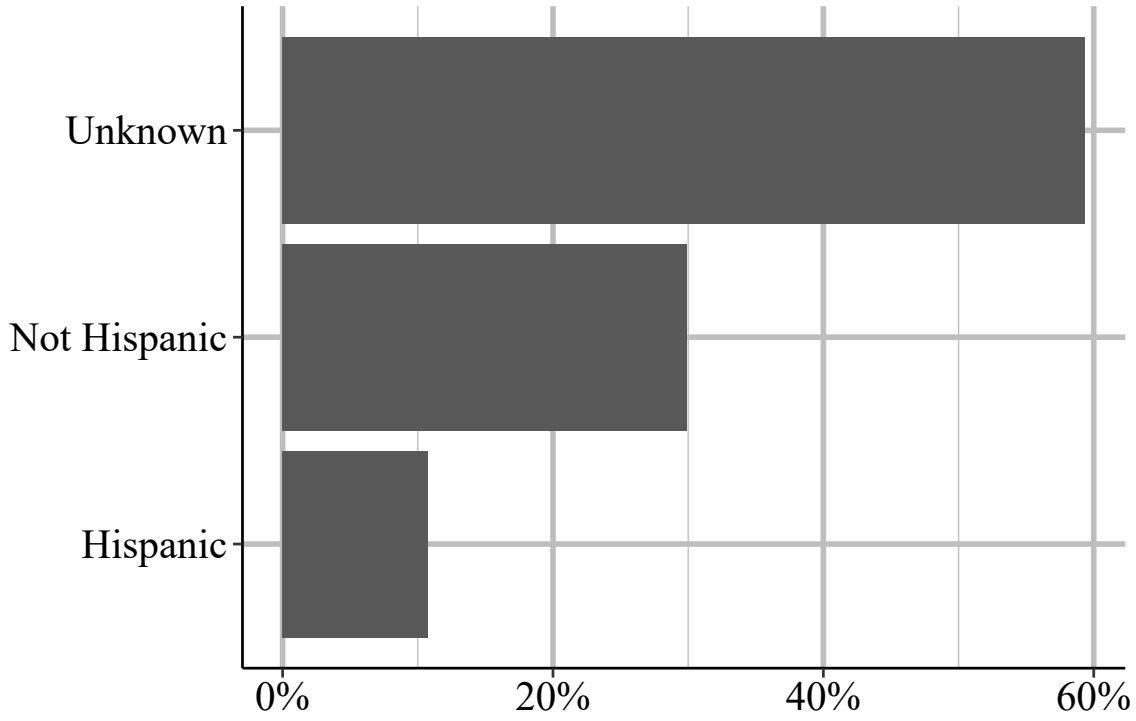


Figure 6.11: The ethnicity of victims, 1976-2018.
(#fig:shrVictimEthnicity)

6.2.2 Case characteristics

6.2.2.1 Weapon used

The first variable we'll look at is the weapon used by each offender. Table @ref(tab:shrWeapon) shows the weapon used by the first offender in every incident from 1976 to 2018. Each offender can only be reported as having a single weapon, so this table essentially shows the number (and percent) of murders caused by this weapon. This isn't entirely true since in reality an offender could use multiple weapons and there can be multiple offenders. In these cases the police include what they believe is the "primary" weapon used by this offender.

The most commonly used weapon is a handgun, which is used in nearly half of murders. This is followed by a knife or other sharp weapon used to cut at 15% of murders, and then by "firearm, type not stated" which is just a firearm where we don't know the exact type (it can include handguns) at 7.3% of murders. The fourth most common weapon is "personal weapons" at nearly 6% of murders. "Personal weapons" is a weird term to mean that there was no weapon - the "weapon" was the offender who beat the victim to death. Shotguns are involved in 5% of murders and all other weapons are involved in fewer than 5% of cases. In total there are 19 different weapons included though most are very uncommon.

Table 6.1: (#tab:shrWeapon)The weapon used in a homicide incident. In cases where there are multiple offenders, shows only the primary weapon for the first offender.

Weapon	# of Incidents	% of Incidents
Handgun	359,188	49.75%
Knife Or Cutting Instrument	109,670	15.19%
Firearm, Type Not Stated	52,638	7.29%
Personal Weapons - Includes Beating	42,763	5.92%
Shotgun	36,072	5.00%
Other Or Unknown Weapon	35,073	4.86%
Blunt Object	33,291	4.61%
Rifle	26,314	3.64%
Strangulation - Includes Hanging	9,619	1.33%
Fire	5,089	0.70%
Asphyxiation - Includes Death By Gas	4,420	0.61%
Other Gun	2,668	0.37%
Narcotics/Drugs - Includes Sleeping Pills	2,058	0.29%
Drowning	1,367	0.19%
Other Or Type Unknown	586	0.08%
Poison - Does Not Include Gas	467	0.06%
Explosives	374	0.05%
Pushed Or Thrown Out of Window	251	0.03%
Narcotics Or Drugs	48	0.01%
Total	721,956	100%

6.2.2.2 Relationship between first victim and offenders

An interesting and highly useful variable is the relationship between the first victim and each offender. To be clear, this is only for the first victim; we don't have the relationship between other victims and offenders. However, as seen earlier, this isn't *too much* of an issue since nearly all incidents only have a single victim. There are 29 possible relationship types (including "unknown" relationship) which are broken into three categories: legal family members, people known to the victim but who aren't family, and people not known to the victim. Table @ref(tab:shrRelationship) shows these relationships and the number and percent of murders with these relationships. If you're looking at this on a computer you can sort or search through this table.

The most common relationship, with a little over a third of murders, is that the police don't know the relationship. So there is a good deal of uncertainty in the relationship between victims and offenders. Next is that the victim is the offender's acquaintance at 20.5% or is a stranger at 15.5%. The next is "other - known to victim" which is similar to being an acquaintance at almost 5% of murders. The 5th most common relationship, at 3.65% is that the victim is the wife of the offender, so is murdered by her husband, and is the first familial relationship of this list. This is followed by the victim being the friend of the murderer at 3.62%. The remaining relationships all make up fewer than 3% of all murders.

Table 6.2: (#tab:shrRelationship)The relationship between the first victim and the first offender in a homicide incident.

Relationship	Category	# of Incidents	% of Incidents
Unknown		263,237	36.46%
Acquaintance	Not family (but known)	147,741	20.46%
Stranger	Not known	111,955	15.51%
Other - Known To Victim	Not family (but known)	33,426	4.63%
Wife	Family	26,353	3.65%
Friend	Not family (but known)	26,143	3.62%
Girlfriend	Not family (but known)	19,194	2.66%
Husband	Family	11,700	1.62%
Other Family	Family	10,400	1.44%
Son	Family	10,342	1.43%
Boyfriend	Not family (but known)	9,041	1.25%
Neighbor	Not family (but known)	7,569	1.05%
Daughter	Family	7,436	1.03%
Brother	Family	6,403	0.89%
Father	Family	5,059	0.70%
Mother	Family	4,578	0.63%
In-Law	Family	4,392	0.61%
Common-Law Wife	Family	3,209	0.44%
Common-Law Husband	Family	2,690	0.37%
Ex-Wife	Not family (but known)	2,205	0.31%
Stepfather	Family	1,693	0.23%
Homosexual Relationship	Not family (but known)	1,663	0.23%
Stepson	Family	1,381	0.19%
Sister	Family	1,344	0.19%
Ex-Husband	Not family (but known)	876	0.12%

Relationship	Category	# of Incidents	% of Incidents
Stepdaughter	Family	747	0.10%
Employer	Not family (but known)	530	0.07%
Employee	Not family (but known)	422	0.06%
Stepmother	Family	227	0.03%
Total		721,956	100%

6.2.2.3 Homicide circumstance

We also have information on the type of the murder, which this data calls the "circumstance". This comes as relatively broad categories that leave a lot to be desired in our understanding of what led to the murder. Table @ref(tab:shrCircumstance) shows the number and percent of each circumstance for the first victim of each murder from 1976 to 2018. This data has 33 possible circumstances which it groups into four main categories: murders that coincide with committing another crime ("felony type" murders), murders that don't coincide with another crime ("non-felony type" murders), justifiable homicides, and negligent manslaughter.

The felony type murders are simply ones where another crime occurred during the murder. While this is called "felony type" it does include other crimes such as theft and gambling (which isn't always a felony) so is a bit of a misnomer. The "non-felony type" are murders that happen without another crime. This includes gang killings (where, supposedly, only the murder occurred), children killed by babysitters, fights among intoxicated (both of alcohol and drugs) people, and "lover's triangle" killings. Justifiable homicides are when a person (civilian or police officer) kill a person who is committing a crime.[6] Negligent manslaughter includes accidental shootings such as when children find and shoot a gun, but excludes deaths from traffic accidents.

The most common circumstances, accounting for 27.4%, 26.4%, and 12.5%, respectively, are "other arguments", "unknown", and "other non-felony type - not specified". Since the data includes "argument over money or property" as one category, the "other arguments" mean that it's an argument for a reason other than over money or property. The "other non-felony type" one does not mean that the murder did not occur alongside another crime, but also doesn't fall into the non-felony categories included. Robbery is the only remaining circumstance with more than 5% of murders, at 8%.

[6]This dataset is one source of information on how many people police kill each year. But it is a large undercount compared to other sources such as the Washington Post collection, so is not a very useful source of information on this topic.

Table 6.3: (#tab:shrCircumstance)The circumstance of the homicide for the first offender in a homicide incident.

Circumstance	Category	# of Incidents	% of Incidents
Other Arguments	Non-Felony Type	197,905	27.41%
Unknown		190,837	26.43%
Other Non-Felony Type - Not Specified	Non-Felony Type	90,203	12.49%
Robbery	Felony Type	57,312	7.94%
Narcotic Drug Laws	Felony Type	26,489	3.67%
Juvenile Gang Killings	Non-Felony Type	23,292	3.23%
Felon Killed By Police	Justifiable Homicide	16,394	2.27%
Brawl Due To Influence of Alcohol	Non-Felony Type	15,174	2.10%
Argument Over Money Or Property	Non-Felony Type	14,667	2.03%
Other Felony Type - Not Specified	Felony Type	13,902	1.93%
All Suspected Felony Type	Felony Type	12,743	1.77%
Felon Killed By Private Citizen	Justifiable Homicide	12,376	1.71%
Lovers Triangle	Non-Felony Type	10,372	1.44%
Burglary	Felony Type	6,052	0.84%
Brawl Due To Influence of Narcotics	Non-Felony Type	4,704	0.65%
Gangland Killings	Non-Felony Type	4,693	0.65%
All Other Manslaughter By Negligence Except Traffic Deaths	Negligent Manslaughter	4,339	0.60%
Rape	Felony Type	4,086	0.57%
Other Negligent Handling of Gun Which Resulted In Death of Another	Negligent Manslaughter	3,266	0.45%
Arson	Felony Type	2,982	0.41%
Other Sex Offenses	Felony Type	1,408	0.20%

Circumstance	Category	# of Incidents	% of Incidents
Child Killed By Babysitter	Non-Felony Type	1,297	0.18%
Children Playing With Gun	Negligent Manslaughter	1,272	0.18%
Motor Vehicle Theft	Felony Type	1,259	0.17%
Institutional Killings	Non-Felony Type	1,055	0.15%
Gambling	Felony Type	1,031	0.14%
Larceny	Felony Type	753	0.10%
Prostitution And Commercialized Vice	Felony Type	601	0.08%
Other - Not Specified	Felony Type	554	0.08%
Sniper Attack	Non-Felony Type	474	0.07%
Victim Shot In Hunting Accident	Negligent Manslaughter	329	0.05%
Gun Cleaning Death - Other Than Self-Inflicted	Negligent Manslaughter	125	0.02%
Abortion	Felony Type	10	0.00%
Total		721,956	100%

6.2.2.4 Homicide subcircumstance

The "subcircumstance" just tells you more information about justifiable homicides. This includes the circumstance leading up to the "felon" - which is how the person killed is described, though technically they don't need to have committed a felony - was killed. It includes if this person attacked an officer (the one who killed them), a different officer, a civilian, or was committing or fleeing a crime.

Table 6.4: (#tab:shrSubCircumstance)The circumstance for the first offender in a homicide incident in cases where the offender is killed. This includes incidents where the only person who dies in the offender.

Subcircumstance	# of Incidents	% of Incidents
Felon Killed In Commission of A Crime	10,320	35.87%
Felon Attacked Police Officer	8,553	29.73%

Subcircumstance	# of Incidents	% of Incidents
Felon Attacked A Civilian	4,498	15.63%
Not Enough Information To Determine	2,423	8.42%
Felon Resisted Arrest	1,226	4.26%
Felon Attacked Fellow Police Officer	951	3.31%
Felon Attempted Flight From A Crime	799	2.78%
Total	28,770	100%

Chapter 7

Law Enforcement Officers Killed and Assaulted (LEOKA)

The Law Enforcement Officers Killed and Assaulted data, often called just by its acronym LEOKA, has two main purposes.[1] First, it provides counts of employees employed by each agency - broken down by if they are civilian employees or sworn officers, and also broken down by gender. And second, it measures how many officers were assaulted or killed (including officers who die accidentally such as in a car crash) in a given month. The assault data is also broken down into shift type (e.g. alone, with a partner, on foot, in a car, etc.), the offender's weapon, and type of call they are responding to (e.g. robbery, disturbance, traffic stop). The killed data simply says how many officers are killed feloniously (i.e. murdered) or died accidentally (e.g. car crash) in a given month. The employee information is at the year-level so you know, for example, how many male police officers were employed in a given year at an agency, but don't know any more than that such as if the number employed changed over the year. This dataset is commonly used as a measure of police employees and is a generally reliable measure of how many police are employed by a police agency. The second part of this data, measuring assaults and deaths, is more flawed with missing data issues and data error issues (e.g. more officers killed than employed in an agency).

7.1 Agencies reporting

Figure @ref(fig:leokaAgencies) shows the annual number of police agencies that reported at least one month that year. The first year of data available, 1960, has about 8,400 agencies reporting though this quickly drops to a trough of around 4,800 agencies that last for several years. After some undulations in the 1970s, reporting agencies steadily increases to nearly

[1] This data is also sometimes called the "Police Employees" dataset.

14,000 agencies in the 1980s and remains steady until declining to around 12,000 by the late 1990s. Then reporting again steadily increases for the rest of our data to about 16,000 agencies by the end.

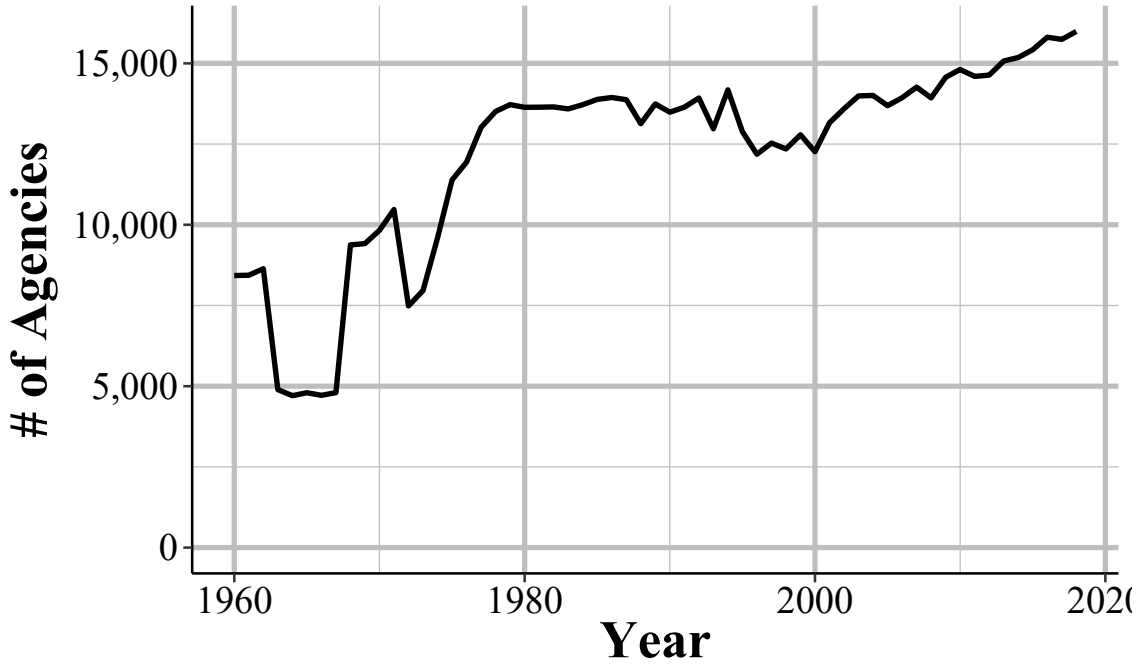


Figure 7.1: The annual number of police agencies that report at least month of data that year, 1960-2018
(#fig:leokaAgencies)

7.2 Important variables

The important variables can be divided into two sections: information on people employed by the department, and information about assaults against officers. The employee information is a snapshot in time during the year while the assault information tells you the number of assaults, broken down several different ways, for each month of the year. Like other UCR data, there are also variables that provide information about the agency - ORI codes, population under jurisdiction - the month and year that the data covers, and how many months reported data.

7.2.1 Number of employees

This data includes the number of people employed by the department with breakdowns by if they are civilian employees or sworn officers (i.e. carries a gun and badge and can make

arrests) as well as by gender. The only genders available are female and male. This is the number of employees as of Halloween that year so it is a single point in time. Though this helps us as it is consistent every year, we don't know exactly when certain officer classes start, which we'd likely see through a jump in employment that year, or if employment or hiring patterns change over the year.

- Female employees
  - Officers
  - Civilians
- Male employees
  - Officers
  - Civilians

We'll look first at the number of employees that are civilian and that are sworn officers through examining Philadelphia in Figure @ref(fig:leokaCivilianOfficers). The number of civilian employees has remained at a little under 1,000 employees from about 1970 through the end of our data, though declining very slightly since the middle 2000s. This is curious since the city's population and crime trends have changed dramatically over this time and the ability of civilian employees to contribute has also changed, such as that they now have computers.[2] In contrast, the number of police has changed far more than civilians, growing rapidly in the 1960s and 1970s to peaking at a little over 8,000 officers in the mid-1970s before declining substantially to the 6,000s. in the late-1980s. As with many agencies nationwide, the number of officers increased in the 1990s and then has decreased steadily in ensuing years. By recent years there are about as many officers as in the late-1980s, even though the city's population has grown substantially since then.

We can also look at the number of officers (or civilian employees) by gender. Figure @ref(fig:leokaOfficersGender) shows the percent of Philadelphia police officers by gender. For the first decade of data all female officers (or civilians) were recorded as male, so that variable should be interpreted as "total officers" until 1971 when it is split into gender. Starting at basically 0% of officers in 1971, female officers grew until they made up about a quarter of officers in 2000 and then has declined slowly since then.

7.2.2 Officers killed

There is almost no information about officers killed. The data only breaks this down into if they died "feloniously" which just means that someone killed them on purpose (e.g. shooting

[2]The last time I heard, which was several years ago, patrol officers in Philadelphia still had to write up certain reports using typewriters. So tech apparently is still about 1960 level.

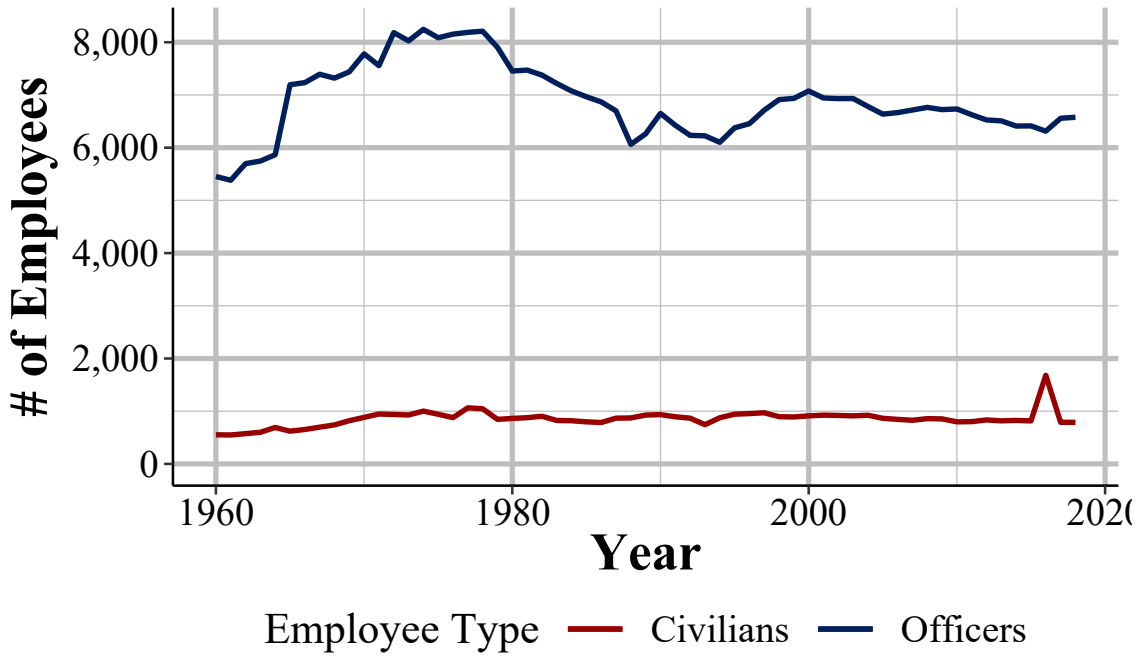


Figure 7.2: The number of civilian employees and sworn officers in Philadelphia, 1960-2018 (#fig:leokaCivilianOfficers)

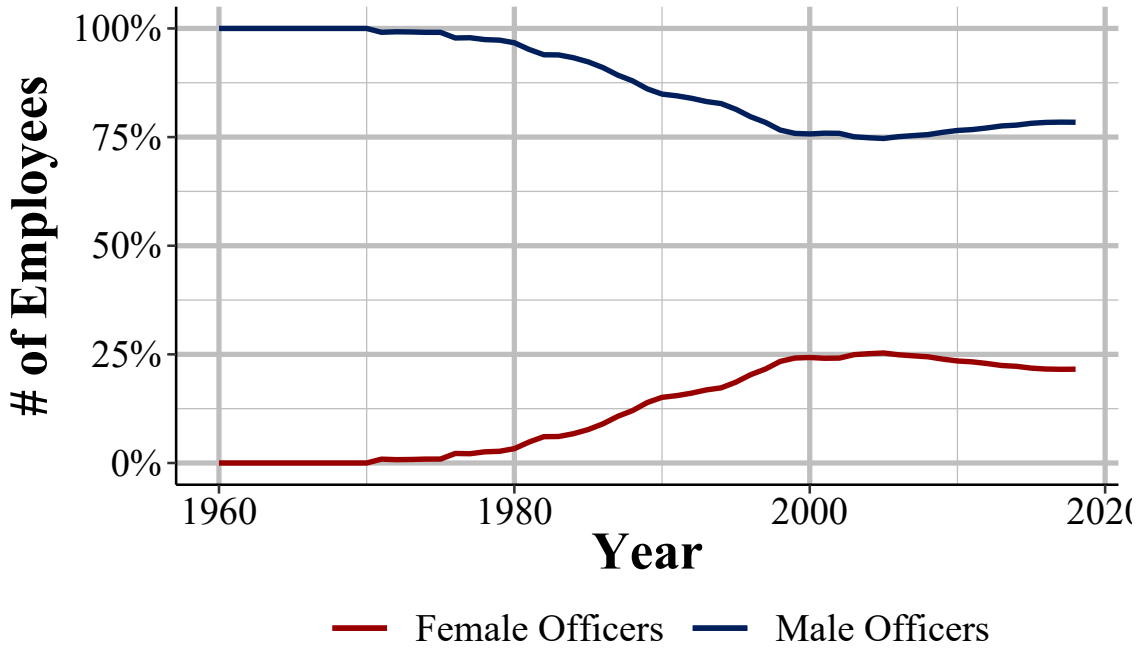


Figure 7.3: The percent of female and male sworn officers in Philadelphia, 1960-2018 (#fig:leokaOfficersGender)

them, intentionally hitting them with a car) or if they died "accidentally" such as if they die during a car crash while on duty. The FBI actually collects more information on officer deaths than they release in this data. This includes the circumstances of each death such as the type of death (e.g. car crash, shooting, ambush, etc.), what weapon the offender had if feloniously killed, and even a detailed written summary of what occurred for each officer killed. They post this information in their annual LEOKA report which is part of their Crime in the United States report. The 2019 report, the latest year available, can be found on their site here.

We can look at what data is available through Figure @ref(fig:leokaOfficersKilled) which shows the number of Los Angeles Police Department officers killed over time. There are no accidental killings until 1975 though this is misleading because that accidental killings variable is not reported until 1971, which is a year in which many other variables in this data began reporting. So we actually have no idea how many officers were killed accidentally from 1960-1970 since this variable is always reported as 0. In general it seems like there is about one officer killed per year in recent decades while the period from 1980 to 2000 was the time of highest danger with as many as five officers killed in a single year. We can also see some trend changes with felonious killings more common than accidental killings in the 1990s and then accidental killings becoming far more common starting in 2000.

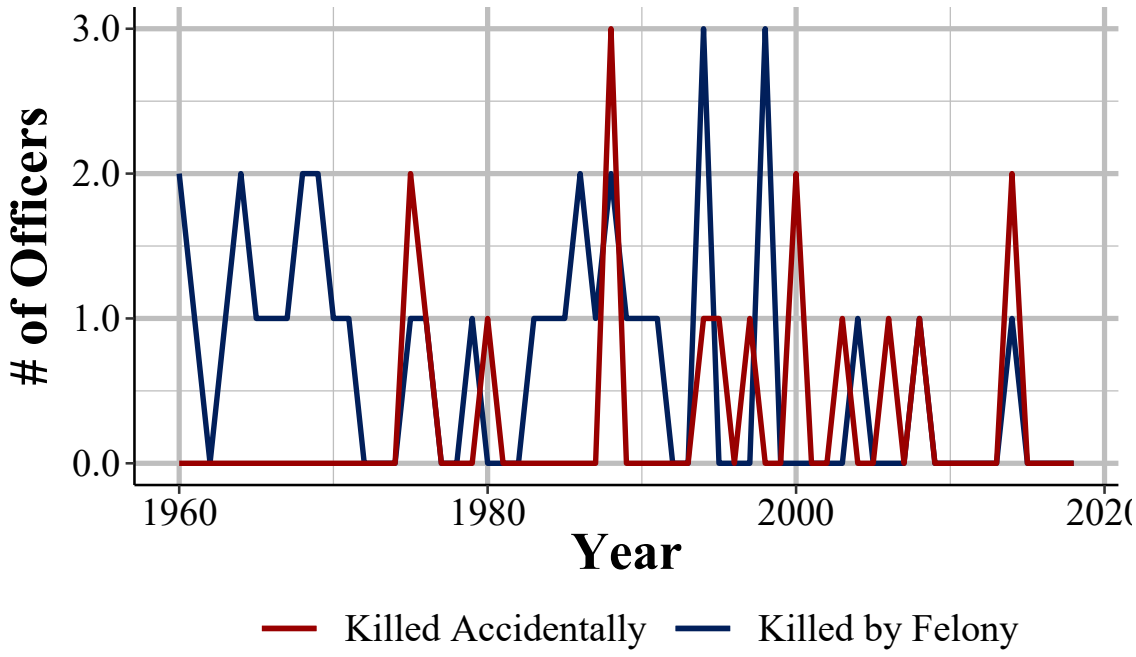


Figure 7.4: The number of officers killed by felony and killed accidentally in Los Angeles, 1960-2018
(#fig:leokaOfficersKilled)

We can also look at the national number of officers killed as in Figure @ref(fig:leokaOfficersKilledNational).

Please note that this is simply summing up the number of officers killed by all agencies that report that year so changes over time are certainly partially due to different agencies reporting each year. Therefore, we'll focus on interpreting the different between felony and accidental killings rather than counts over time - though even this may be off if agencies that reported more felony or more accidental killings differ in their reporting over time. Again we see that there are no officers killed accidentally, due to that variable not being reported, until 1971. The difference between officers killed by felony and killed accidentally is widest are the start of our data and narrows considerable until there are only several more felonious killings than accidental killings by the late 1990s. Though this trend reverses in the early 2010s with accidental killings decreasing and felonious killings increasing again.

What can we make of this? It's hard to say. Interpreting this properly requires adding some other key variables such as the number of officers employed, the number of circumstances they respond to (e.g. are they patrolling more, apprehending violent offenders more, etc.), the number of guns on the street, the quality or availability or body armor, among others.

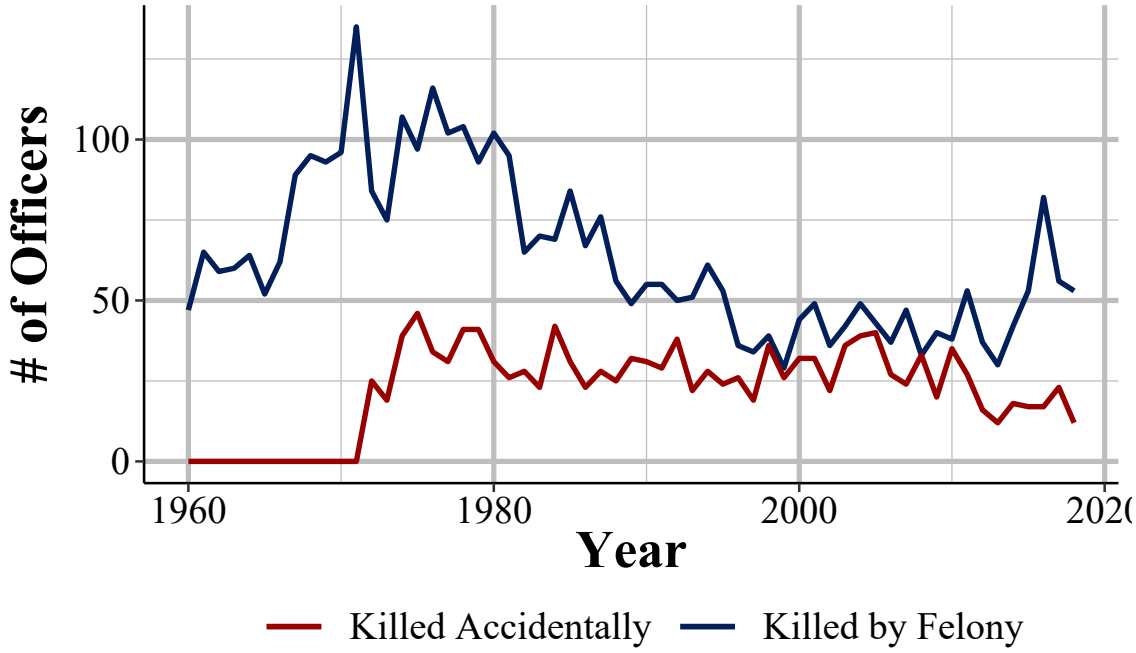


Figure 7.5: The national number of officers killed by felony and killed accidentally, 1960-2018 (#fig:leokaOfficersKilledNational)

7.2.3 Assaults by injury and weapon

This data breaks down the monthly number of assaults on officers in a few different ways. Here, we'll look at the number of assaults where the officer is injured or not injured and within these categories by which weapon the offender had. This is the number of officers

assaulted so if an incident has three officers assaulted, that will count as three different assaults. If the offender used multiple weapons then only the most serious weapon would be counted. For example, if an offender used a knife and a gun during the assault, the assault would be counted as a gun assault. Unfortunately we only know if an officer was injured or not and not the severity of the injury. So we can't tell if the officer is merely bruised or was shot or stabbed.

- Assaults with injury
  - Offender has firearm
  - Offender had knife
  - Offender had other weapon
  - Offender was unarmed
- Assaults without injury
  - Offender has firearm
  - Offender had knife
  - Offender had other weapon
  - Offender was unarmed

We can start by looking at the breakdown of assaults by injury and weapon type for officers in the Los Angeles Police Department. Figure @ref(fig:leokaAssaultTypeInjury) shows the number of assaults from all years reported for these categories. Over the complete time period there were almost 40,000 officers assaulted with about example three-quarters of these assaults - 30,000 assaults - leading to no injuries. This data shows the number of officers assaulted, not unique officers, so an officer can potentially be included in the data multiple times if they are assaulted multiple times. A little under a quarter of assaults lead to officer injury with most of these from unarmed offenders. Interestingly, there are far more gun and knife assaults where the officer is not injured than where the officer is injured.

We can also look at assaults over time. Figure @ref(fig:leokaAssaultsInjuryYear) shows the number of assaults, assaults with injury, and assault without injury for the Los Angeles Police Department from 1960 to 2018. We can immediately see some data issues are there are years with no assaults recorded. And in the late-2000s there is a sudden drop from about 250 assaults with injuries per year in the previous few decades to nearly zero officer injuries reported a year. This strongly suggests some change in reporting rather than a true decrease in assaults with injuries. For the decades where the data is less obviously wrong, there is a consistent trend of most assaults leading to no injuries, though the distance between the number of injury and non-injury assaults fluctuates over time.

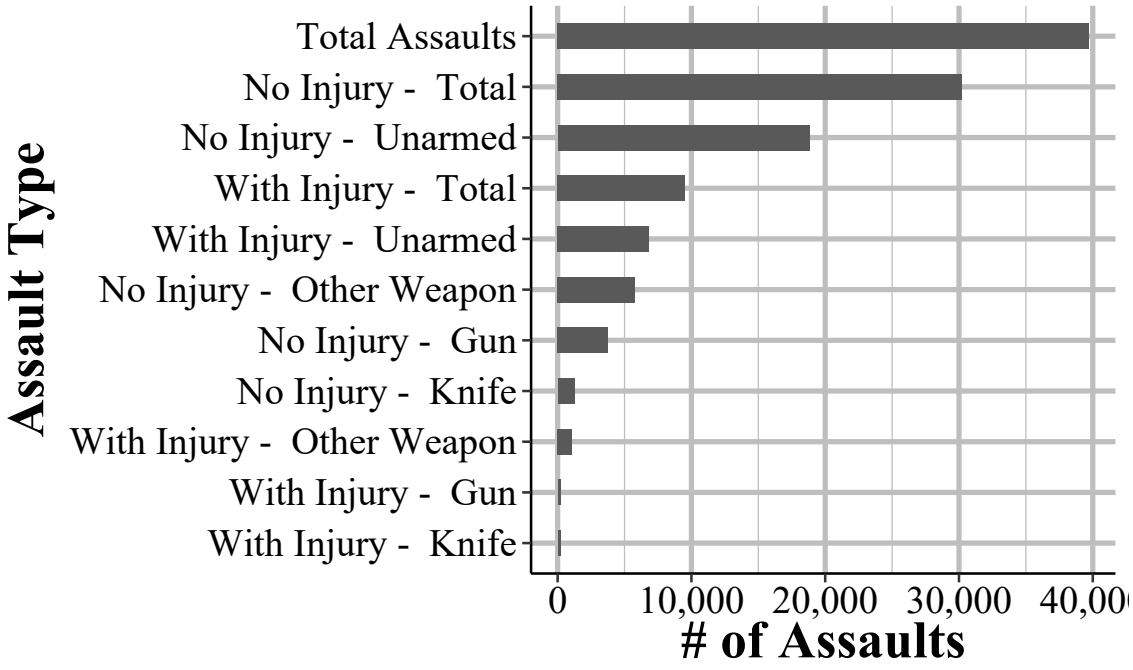


Figure 7.6: The total number of assaults on officers by injury sustained and offender weapon in Los Angeles, 1960-2018.
(#fig:leokaAssaultTypeInjury)

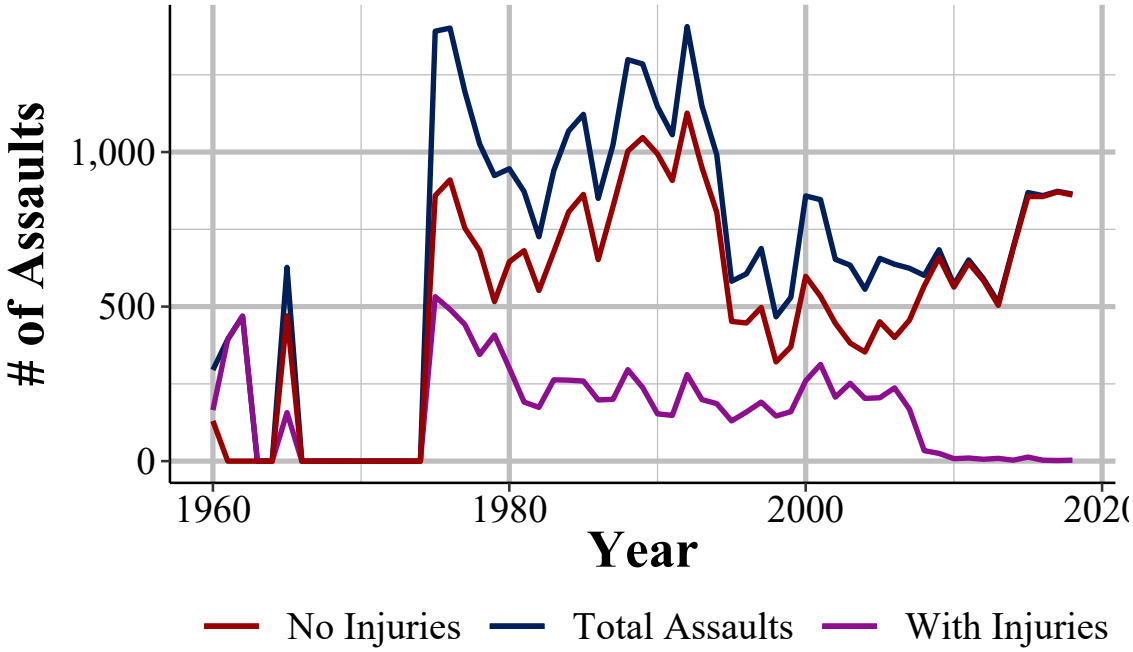


Figure 7.7: The annual number of assaults on officers by injury sustained in Los Angeles, 1960-2018.
(#fig:leokaAssaultsInjuryYear)

7.2.4 Assaults by call type

The next group of ways that assaults are broken down is by the type of call the officer is assigned when they are assaulted. For example, if an officer is responding to a burglary report, any assault they experience on that call will be classified as "burglary" related. In addition, we know how many assaults were cleared by arrest or cleared through exceptional means (for more on this, please see @ref(clearedCrimes)) though it doesn't differential between the two. Since assaults are based on the number of officers assaulted, not the number of incidents where officers are assaulted, arresting a single person can clear multiple assaults. The possible call types are below:

- Disturbance call (e.g. domestic violence, person carrying a gun in public)
- Burglary
- Robbery
- Officers arresting someone for another crime
- Civil disorder
- Officer has custody of prisoners
- Suspicious persons
- Officers are ambushed
- Mentally deranged person
- Traffic pursuit and traffic stops
- All other call types
- Total - sum of all call types

Figure @ref(fig:leokaAssaultCallType) shows the number of assaults on Los Angeles Police Department officers by the type of call for 1960-2018. There were about 38,000 assaults against Los Angeles Police Department officers with a little over 31,000 of these assaults cleared. An important thing to note is that the number of assaults here is less than the nearly 40,000 assaults for the same agency over the same time period we saw in Figure @ref(fig:leokaAssaultCallType). This is because some variables are not reported for all years and agencies are free to report which variables they want to report in any given year. This makes it massively tricky to use this data since even simple statistics for the same agency for supposedly the same variable (here it's technically different variables but should still be the total number of officers assaulted) can be different. The most common type of call where officers are assaulted are disturbance calls which include domestic violence and reports of dangerous individuals such as people carrying guns in public. The least common call type is ambush calls, though in these calls the police are called to a scene by the offender who intends to assault or kill the officers, so is likely far more dangerous than other call types, even though it is rare.

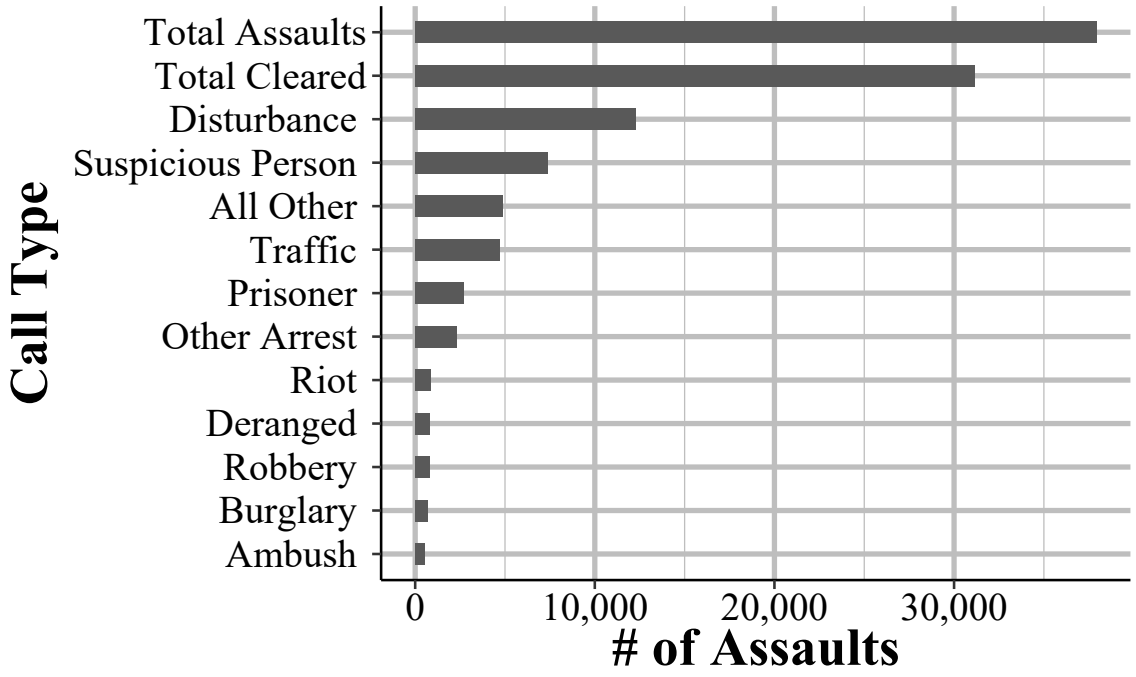


Figure 7.8: Assaults on Los Angeles Police Department officers by type of call where they were assaulted at, 1960-2018.
(#fig:leokaAssaultCallType)

Within these call types is also a breakdown by offender weapon use, with the same weapons as above, and the type of officer assignment which is essentially if they are alone or not and if they're on foot or not. Finally, it says how many assaults are cleared by arrest or cleared through exceptional means though it doesn't differential between the two. The shift assignment is essentially how they go through their normal day, if this is in a vehicle, alone, as a detective, or under a different assignment (including being off-duty). For example, being in a vehicle with two officers means that their normal assignment is driving in a vehicle, not that they were actually assaulted in said vehicle. This also doesn't necessarily mean that these are the only officers at the scene. It is simply the shift assignment of the officer who is assaulted. For example, if an officer who normally works alone in a vehicle shows up to a scene where other officers are present, and who are under different shift assignments, and gets assaulted - and no one else gets assaulted - that is an assault for officers "in a vehicle alone".

- Offender weapons
  - Offender has firearm
  - Offender had knife
  - Offender had other weapon
  - Offender was unarmed

- Type of officer shift assignment
  - In a vehicle with two officers
  - In a vehicle alone
  - In a vehicle alone but assisted by other officers
  - Detective or special unit alone
  - Detective or special unit assisted by other officers
  - Other assignment alone
  - Other assignment assisted by other officers
- Number of assaults on police cleared

We'll look specifically at disturbance calls since they are the most common call type, at least for the Los Angeles Police Department. Figure @ref(fig:leokaDisturbanceWeapon) shows the total number of disturbance assaults by offender weapon in Los Angeles. Most assaults have an unarmed offender with a sharp decline to the number of offenders with a weapon other than a gun or knife. Assaults by a gun and by a knife are the least common.

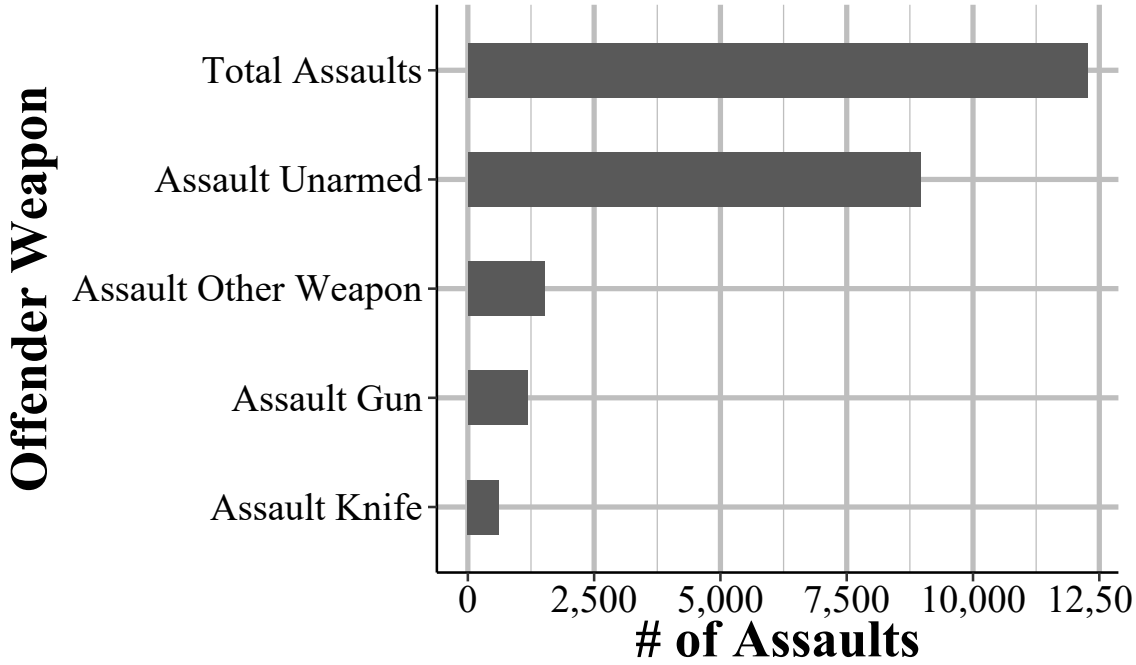


Figure 7.9: The number of assaults on Los Angeles Police Department officers in disturbance calls by the injury sustained by the officer, 1960-2018.
(#fig:leokaDisturbanceWeapon)

Again using disturbance calls for the Los Angeles Police Department, we can look at assaults by the officer assignment, as seen in Figure @ref(fig:leokaShiftAssignment). In the vast majority of assaults it is of officers who are in a vehicle along with a partner. This drops

very sharply to several hundred assaults on detectives who are assisting other officers and then increasingly declines to the other shift assignments to the least common assault being against detectives who are acting alone.

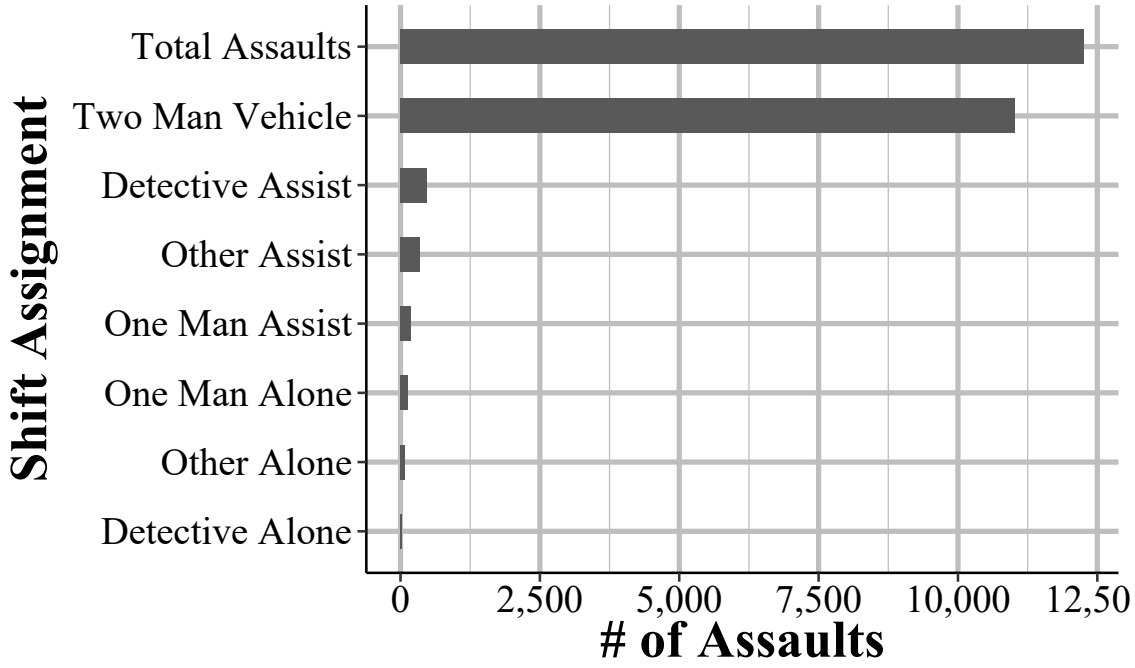


Figure 7.10: The number of assaults on Los Angeles Police Department officers in disturbance calls by the injury sustained by the shift assignment of the officer, 1960-2018. (#fig:leokaShiftAssignment)

7.2.5 Assaults by time

The final breakdown in assaults is by the time they occur, divided into 12 two-hour chunks starting at 12:01am. Like some other variables this data is only available starting in 1971. There is no more information than total assaults in this time so we don't know if the assaults led to injuries, the type of call or shift assignment the officer was on, or the offender's weapons.

- 12:01am - 2:00am
- 2:01am - 4:00am
- 4:01am - 6:00am
- 6:01am - 8:00am
- 8:01am - 10:00am
- 10:01am - 12:00pm
- 12:01pm - 2:00pm

- 2:01pm - 4:00pm
- 4:01pm - 6:00pm
- 6:01pm - 8:00pm
- 8:01pm - 10:00pm
- 10:01pm - 12:00am

We'll look at these time chunks in Figure @ref(fig:phoenixAssaultTimes) which shows the total number of assaults by time of day from 1971 to 2018 in Phoenix, Arizona. The most common times for officers to be assaulted looks like a mirror image of when crime is highest: late night and early morning. The 12:01am to 2am chunk is the most common time followed by 10pm to midnight, with assaults increasing at the day grows later and at its lowest point from 6-8am. This strongly suggests that officers are assaulted at crime scenes, such as responding to crimes or making arrests.

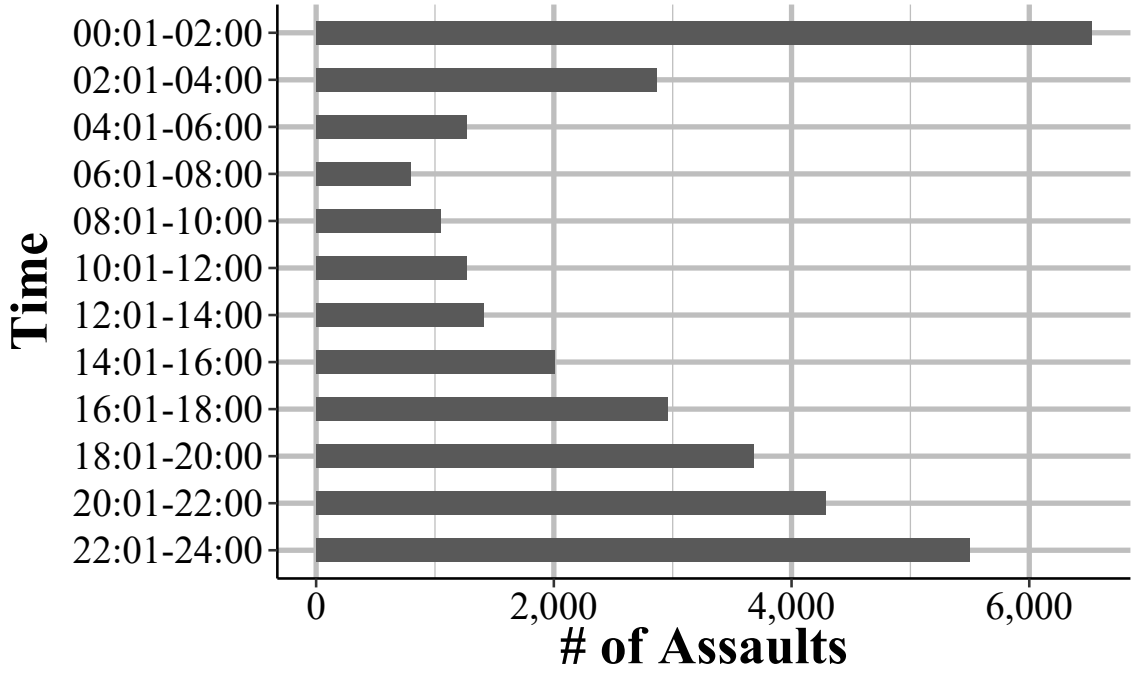


Figure 7.11: The number of assaults against Phoenix Police Department officers by hourly grouping for all years with data available, 1971-2018.
(#fig:phoenixAssaultTimes)

Chapter 8

Arson

The arson dataset provides monthly counts at the police agency-level for arsons that occur, and includes a breakdown of arsons by the type of arson (e.g. arson of a person's home, arson of a vehicle, arson of a community/public building) and the estimated value of the damage caused by the arson. This data includes all arsons reported to the police or otherwise known to the police (e.g. discovered while on patrol) and also has a count of arsons that lead to an arrest (of at least one person who committed the arson) and reports that turned out to not be arsons (such as if an investigation found that the fire was started accidentally).

For each type of arson it includes the number of arsons where the structure was uninhabited or otherwise not in use, so you can estimate the percent of arsons of buildings which had the potential to harm people. This measure is for structures where people normally did not inhabit the structure - such as a vacant building where no one lives. A home where no one is home *at the time of the arson* does not count as an uninhabited building.

In cases where the arson led to a death, that death would be recorded as a murder on the Offenses Known and Clearances by Arrest dataset - but not indicated anywhere on this dataset. If an individual who responds to the arson dies because of it, such as a police officer or a firefighter, this is not considered a homicide (though the officer death is still included in the Law Enforcement Officers Killed and Assaulted data) unless the arsonist intended to cause their deaths. Even though the UCR uses the Hierarchy Rule, where only the most serious offense that occurs is recorded, all arsons are reported - arson is exempt from the Hierarchy Rule.

8.1 Agencies reporting

Figure @ref(fig:arsonAgencies) shows the annual number of police agencies that reported at least one month that year. With data starting in 1979, there were a little over 12,000 agencies

reporting a year until the early 2000s where it recovered from a sharp drop in agencies to steadily increase to about 16,000 a year. While arson data is available through 2019, this graph only shows data through 2018.

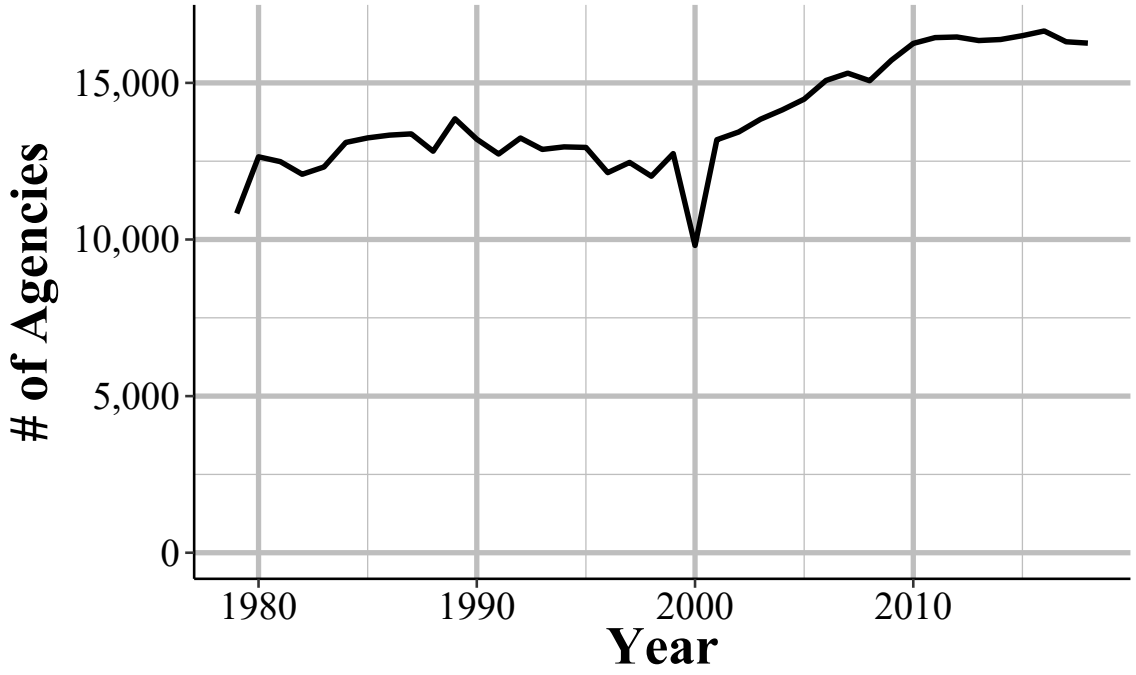


Figure 8.1: The annual number of police agencies that report at least month of data that year.
(#fig:arsonAgencies)

8.2 Important variables

Similar to the Offenses Known and Clearances by Arrest data, which is covered in Chapter @ref(offensesKnown), this data shows the monthly number of crimes - in this case only arsons - reported or discovered by the police, the number found by police to be unfounded, and the number cleared by the police. In addition, it includes the number of arsons in structures that were uninhabited, and the estimated cost of the damage caused by the arsons. For each variable, the arsons also broken into several categories of arsons, which we'll talk about in Section @ref(arsonType). Like other UCR data, there are also variables that provide information about the agency - ORI codes, population under jurisdiction - the month and year that the data covers, and how many months reported data.

8.2.1 Actual arsons

This variable includes the number of arsons either reported to the police or otherwise known to the police in that month and that the police determine to actually be arsons - that is, reports that are not unfounded. This is based only on the police's determination that an arson occurred, not the decision of other groups like a court or the conviction of someone for the crime.

8.2.2 Unfounded arsons

This variable shows the number of arsons reports that the police determined to not actually be arsons. For example, if a house burns down and police think it was arson but later determine that it was an accident, it would be an unfounded arson.

8.2.3 Reported arsons

This variable is the sum of actual arsons and unfounded arsons - it is the total number of arsons that were reported or known to the police, even if a later investigation found that it wasn't actually an arson. Since this is the sum of two already present variables - and is less informative than the two variables are as separate variables - I'm not exactly sure why it's included.

8.2.4 Cleared arsons

This shows the number of arsons where at least one offender is arrested or when an arrest cannot be made for some reason, but the police know who the offender is - the latter option is called "exceptional clearances" or "clearances by exceptional means." There is no way to tell if a clearance is due to an arrest or due to exceptional means.[1].

For exceptional clearances, police must have identified the offender, have sufficient evidence to arrest that offender, know where they are (so they can actually be apprehended) and only then be unable to make the arrest. Exceptional clearances include cases where the offender dies before an arrest (by any means, including suicide, accidental, illness, killed by someone including a police officer) or when the police are unable to arrest the person since they are already in custody by another jurisdiction (including out of the country or in custody of another agency) and extradition is denied. Two other potential causes of exceptional clearance is when prosecution of the case cannot go forward because the district attorney

[1]NIBRS data does tell more information about the type of arrest, but UCR data does not

refuses to prosecute the case, for reasons other than lack of evidence, or when a victim refuses to assist the prosecution in the case.

Please note that this data is at the incident-level which means that having multiple people arrested for an incident still only is a clearance for that single incident. Clearances are also reported in the month they occur, not in the month that the initial crime happened. This can lead to cases where there are more clearances than crimes, incorrectly leading to the conclusion that police solve >100% of crimes.

Figure @ref(fig:arsonClearance) shows the number of actual arsons (reports that are founded) and clearances for single-family home arsons in League City, Texas, a city of about 100,000 outside of Houston. In most years there were fewer clearances than arsons, but in four years (1982, 1981, 1992, and 2007) there were more clearances than arsons. This is simply because clearances are reported in the month they occur, regardless of when the arson they are clearing occurred.

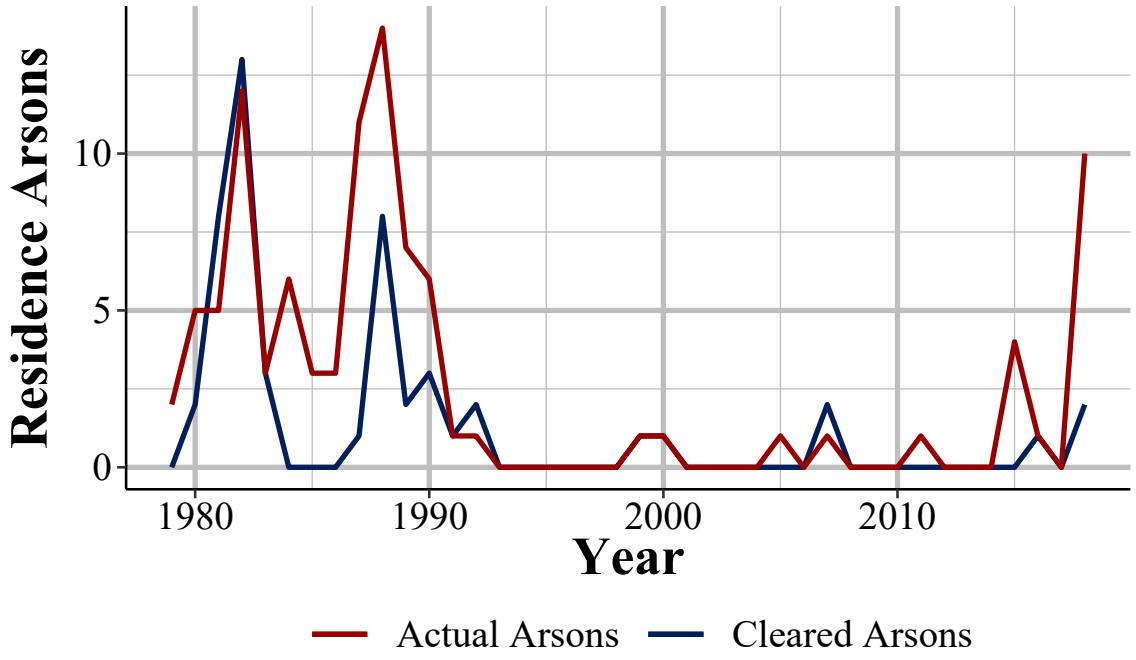


Figure 8.2: The annual number of single-family home arsons and clearances in League City, Texas.
(#fig:arsonClearance)

It is rare that there are more clearances than crimes in a given month though this is partially just due to few cases every being cleared. According to the 2019 NIBRS, it takes on average 7 days between the incident and the arrest (median = 0 days) for all crimes, and an average of 11.6 days from incident to arrest for arsons specifically (median = 0 days) for offenders who get arrested. This means that most clearances will be for crimes committed that month, though certainly not all. Therefore, use caution when trying to use this variable to measure

crime clearance rates.

8.2.5 Cleared for arsons where all offenders are under age 18

This variable is the same for normal clearances but only for arsons where every offender is under the age of 18. In these cases a clearance by arrest can include citing the juvenile offender with an order to go to court to stand trial, even if the juvenile is not actually taken into police custody. As this variable requires that the police know every offender (to be able to determine that they are all under 18), it is likely highly flawed and not a very useful variable to use.

8.2.6 Uninhabited building arsons

This data also includes the number of arsons that occur in uninhabited structures. These structures must be uninhabited in the long-term, not simply having no one inside them when the arson occurs. The FBI defines these are structures that are "uninhabited, abandoned, deserted, or not normally in use" at the time of the arson. For example, a vacation home whose owners aren't living in at the time would be "not normally in use" so would be an uninhabited building. A business that is closed when the fire started, but is open during the day, is not an uninhabited building.

8.2.7 Estimated damage cost

The final variable is the estimated cost of the arson. This is how much the police estimates the value (in dollars) of the damaged or destroyed property is. Since this is the value of property damage, injuries to people (including non-physical injuries such as trauma or mental health costs) are not included. Since this is estimated damage it may be inaccurate to some degree. This variable is the sum of monthly estimated cost so while you can get the average cost by dividing this by the number of actual offenses, this average may be significantly off due to having extremely small or large values in your data. This value may be $0 since arsons include attempted arsons which may cause little or no damage. Please note that this value is not inflation-adjusted so you will have to adjust the value yourself.

8.3 Types of arsons

For each of the arson categories above, this dataset has information for ten different *types* of arson. The type is based on where the arson occurred, not based on how the fire was

initiated, how far or fast it spread, or any other information about the arson - nothing is actually known about the arson outcome other than if the arson was cleared and the estimated damage. There are seven arsons types for buildings, two for vehicles, and one as an "other" category that includes arsons of outdoor areas like parks or forests (though this group does not have any subcategories so all you know is the arson is neither of a building or a vehicle). For both the buildings and the vehicle arson types there is also a "total buildings" and "total vehicles" category that is just the sum of each subcategory; there is also a "grant total" variable that sums all building, vehicle, and other arsons.

1. Building arsons
    - Single occupancy home
    - Other residential
    - Storage
    - Industrial/manufacturing building
    - Other commercial building
    - Community or public building
    - All other structures
2. Vehicles
    - Motor vehicles
    - Other vehicles
3. All other arsons

Some arsons can burn down multiple types or structures or cars - fire, after all, tends to spread. This data defines the arson based on where the fire originated. So an arson that starts in someone's house and also burns down their car in the garage would be a single-family home arson, not a vehicle arson. This is true even if the damage is more severe for a type other than where the fire started. So if someone threw a Molotov cocktail at a car parked outside a house and it lightly damaged the car but spread to homes nearby and destroyed those homes, it is still a vehicle arson (though the damage recorded would include the homes burned.)

8.4 Data errors

Like the other UCR data, there are some cases where there are obvious data entry errors leading to impossible numbers of reported arsons or the price of an arson. As an example, Figure @ref(fig:residenceArson) shows the annual number of single-family home arsons in

Byron City, Illinois, which has a population of slightly over 3,600 people. Every year with data available there are zero arsons reported until 2003 when 469 arsons are reported. Since it is exceedingly unlikely that suddenly an eighth of the city suddenly burned down, and the city never again had a residence arson, this is almost certainly a data entry error. As arsons are relatively rare, having errors - and especially ones like this - can drastically change the results of your analysis so it is important to check your data carefully for more errors. For those using the data that I have cleaned and concatenated, the complete list of obvious outliers that I removed is available on the data's page on openICPSR.

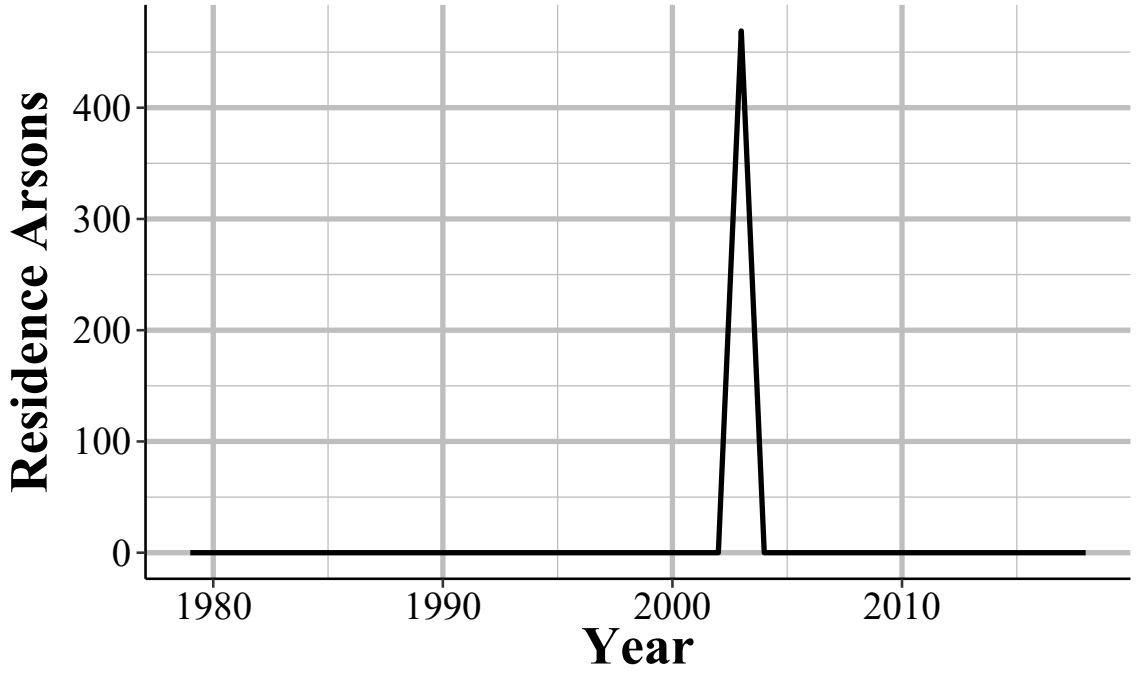


Figure 8.3: Annual single-family home arsons in Byron City, Illinois. The sudden spike to over 400 arsons in a single year is an example of data errors in this dataset. (#fig:residenceArson)

There are also cases where it is less clear when a value is a data error or is simply due to an outlier - but real - value. For example, Figure @ref(fig:arsonCost) shows the annual average cost of a single-family home fire in Los Angeles, California. In most years the damage is low - since an arson can damage only part of the house, these low values likely mean that on average only part of the house was damaged, not the entire house. In 2009, however, the average damage is about $540,000 per arson. Is this a data entry error that simply inputs a damage value that is too high? It certainly appears to be a data error since it's a sudden huge jump in damage value. However, it could also be that some extraordinarily expensive homes were destroyed that year. In 2009 Los Angeles only reported 63 single-family home arsons so having one, or a few, super mansions - which LA has plenty of - destroyed could mean that this huge value is real.

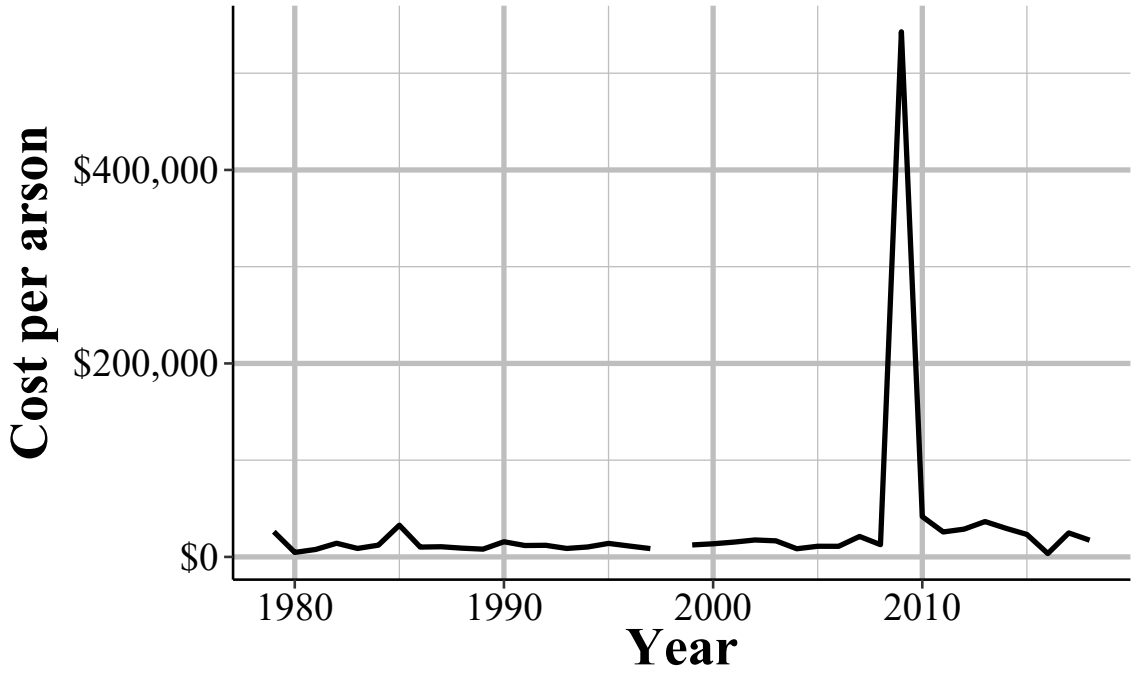


Figure 8.4: The annual average cost per single-family home arson in Los Angeles, California. (#fig:arsonCost)

Chapter 9

Hate Crime Data

This dataset covers crimes that are reported to the police and judged by the police to be motivated by hate. More specifically, they are, first, crimes which were, second, motivated - at least in part - by bias towards a certain person or group of people because of characteristics about them such as race, sexual orientation, or religion. The first part is key, they must be crimes - and really must be the selection of crimes that the FBI collects for this dataset. Biased actions that don't meet the standard of a crime, or are of a crime not included in this data, are not considered hate crimes. For example, if someone yells at a Black person and uses racial slurs against them, it is clearly a racist action. For it to be included in this data, however, it would have to extend to a threat since "intimidation" is a crime included in this data but lesser actions such as simply insulting someone is not included. For the second part, the bias motivation, it must be against a group that the FBI includes in this data. For example, when this data collection began crimes against transgender people were not counted so if a transgender person was assaulted or killed because they were transgender, this is not a hate crime recorded in the data (though it would have counted in the "Anti-Lesbian, Gay, Bisexual, Or Transgender, Mixed Group (LGBT)" bias motivation which was always reported).[1]

So this data is really a narrower measure of hate crimes than it might seem. In practice it is (some) crimes motivated by (some) kinds of hate that are reported to the police. It is also the most under-reported UCR dataset with most agencies not reporting any hate crimes to the FBI. This leads to huge gaps in the data with some states having zero agencies report crime, agencies reporting some bias motivations but not others, agencies reporting some years but not others. While these problems exist for all of the UCR datasets, it is most severe in this data. This problem is exacerbated by hate crimes being rare even in agencies that report them - with such rare events, even minor changes in which agencies report or which types of offenses they include can have large effects.

[1]The first year where transgender as a group was a considered a bias motivation was in 2013.

My main takeaway for this data is that it is inappropriate to use for most hate crime research. At most it can be used to look at within-city within-bias-motivation trends, while keeping in mind that even this narrow subset of data is limited by under-reporting by victims and potential changes in police practices of reporting such as how many months of data they report per year.

9.1 Agencies reporting

We'll start by looking at how many agencies report hate crime each year. This is a bit tricky since there can be multiple ways to examine how agencies report, and since agencies can truly have no hate crimes in a year so it's hard to differentiate the true zeroes from the non-reporters. Figure @ref(fig:hateAgencies) shows the number of agencies that report at least one hate crime incident in that year. From the start in 1991 there were about 750 agencies reporting and that grew steadily to about 2,000 agencies in year 2000. From there it increased a bit over the next decade before declining to below 1,750 in the early 2010s and rising again to around 2,000 agencies at the end of our data.

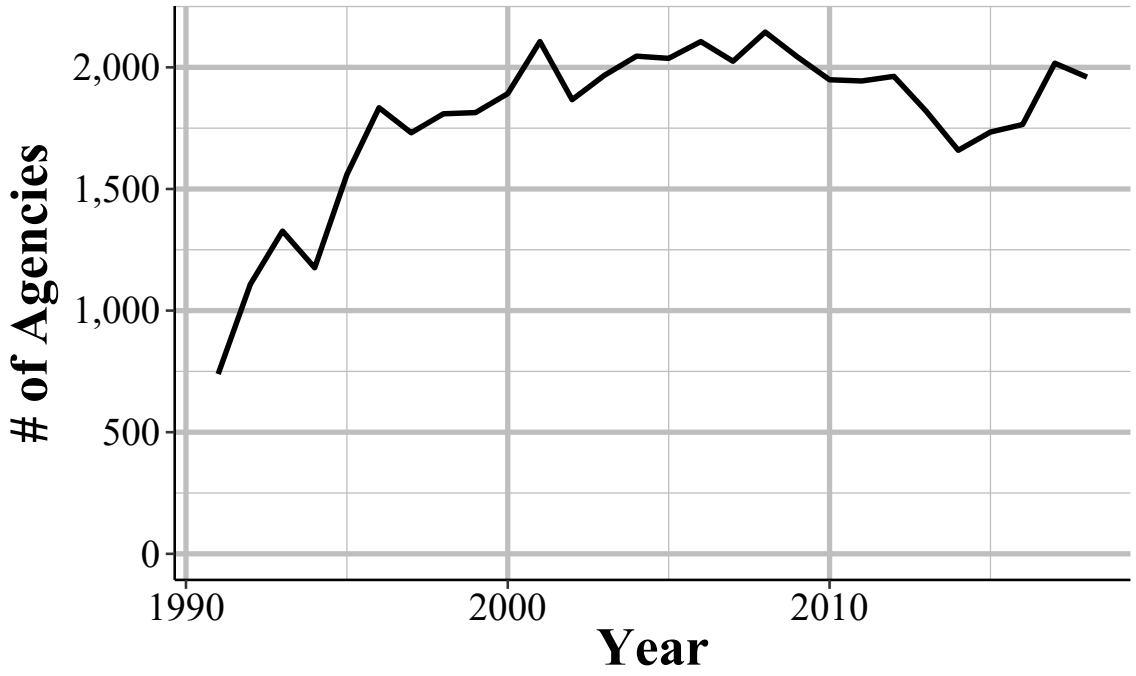


Figure 9.1: The annual number of police agencies that report at least one hate crime incident in that year.
(#fig:hateAgencies)

The 2,000 or so agencies that report each year are not the same every year. Figure @ref(fig:hateCrimesEver) shows the cumulative number of agencies that have reported at

least one hate crime between 1991 and 2018. There is a steady growth in the cumulative number of agencies, with about 350 new agencies each year. This means that each year some new agencies report hate crimes while some agencies that reported a hate crime in the previous year don't report any hate crimes in the current year.

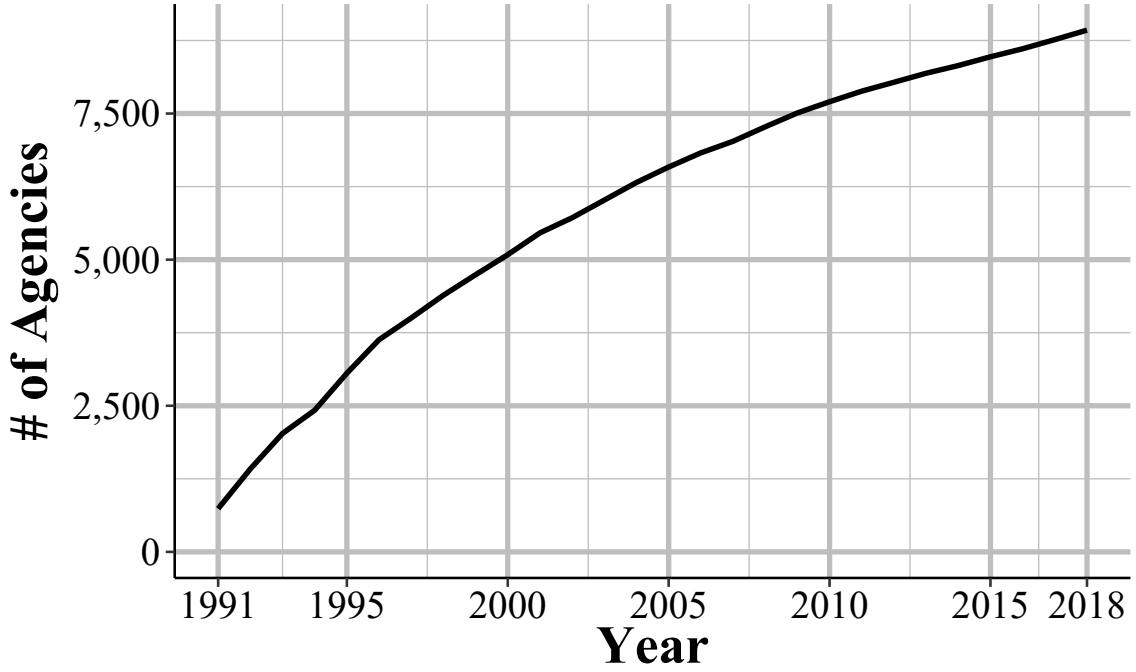


Figure 9.2: The cumulative number of agencies that have reported one or more hate crimes between 1991 to 2018.
(#fig:hateCrimesEver)

Figure @ref(fig:hateCrimesPreviousYear) puts this into hard numbers by showing the percent of agencies who reported a hate crime in a certain year who *also* reported a hate crime in the previous year. For most years between 50% and 60% of agencies which reported a hate crime in the year shown on the x-axis also reported a hate crime in the previous year, indicating somewhat high consistency in which agencies have hate crimes.

Another way to understand reporting is to look at the number of reported hate crimes by state and see which states report and which don't. Figure @ref(fig:hateCrimesMap) does this for 2018 data by showing the number of reported hate crime incidents by state. While every state other than Wyoming reporting at least one hate crime in 2018, there are large differences between states because even in states that have reporting agencies, not all agencies in that state report. For example, Alabama reported only two hate crimes in 2018, the least of any state other than Wyoming.

Since the number of state-wide hate crimes is partially influenced by population, we'll also look at it as the percent of agencies in the state that report at least one hate crime. Again this is limited by population as agencies in each state cover different populations - and most

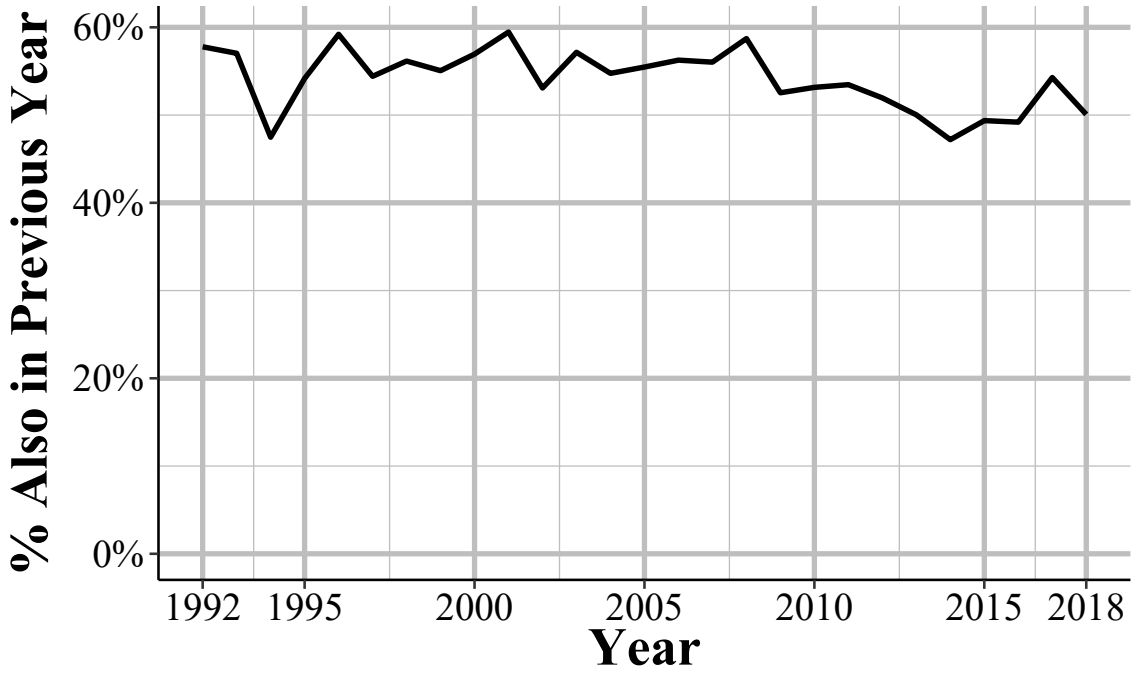


Figure 9.3: The percent of agencies that report a hate crime in a given year that also reported a hate crime in the previous year, 1992-2018.
(#fig:hateCrimesPreviousYear)

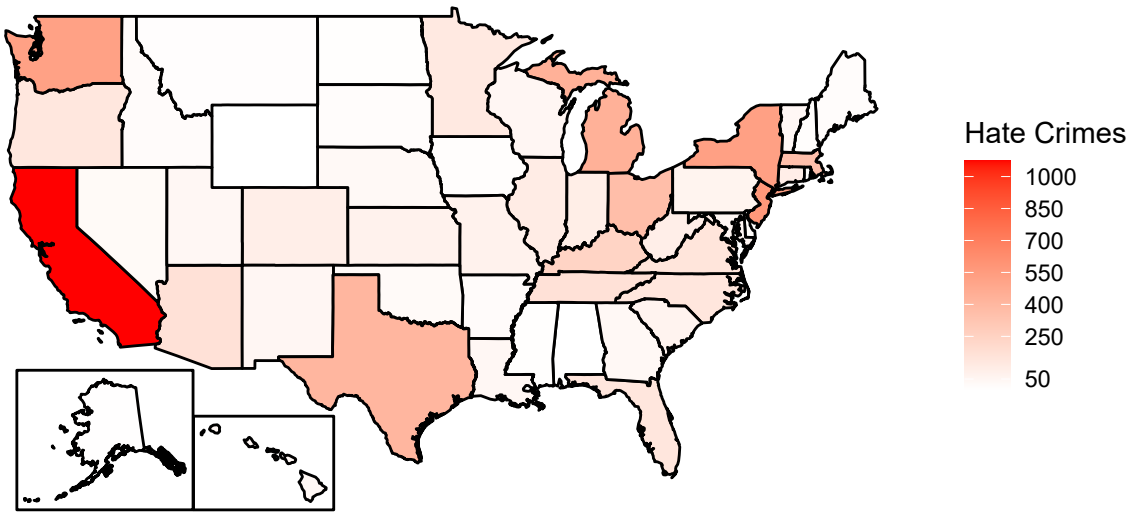


Figure 9.4: Total reported hate crimes by state, 2018.
(#fig:hateCrimesMap)

agencies are small so a few large agencies can account for a large share of crimes, including hate crimes. Still, this is a useful exercise as we can easily see where very few agencies report. Figure @ref(fig:statePercentReporting) shows the percent of agencies for each state that reported at least one hate crime in 2018. In Connecticut, the state with the highest percent of agencies reporting, 28% of agencies reported at least one hate crime. In 34 states, fewer than 10% of agencies reported a hate crime, and in 19 states fewer than 5% of agencies did so. One interesting finding from this graph is the more liberal states - Connecticut, New Jersey, Vermont, California, Washington - have the highest share of agencies reporting a hate crime, indicating that the culture of the state may influence either the propensity of hate crimes, whether victims report, or whether agencies report hate crimes.

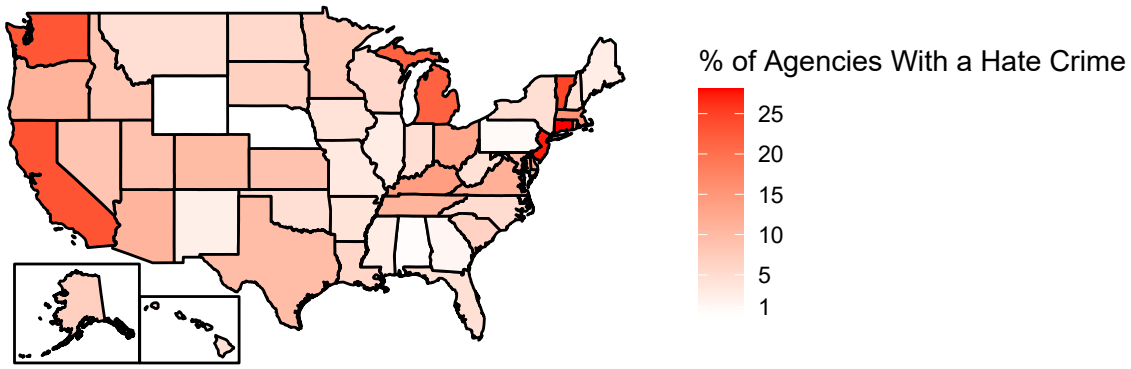


Figure 9.5: The percent of agencies in each state that reported at least one hate crime in 2018.
(#fig:statePercentReporting)

9.2 Important variables

This data has the standard set of variables describing the agency that is reporting. This includes the agency ORI - which is the unique ID for that agency - the agency name, their state, and the population under their jurisdiction.

They also have a variable with the exact date that the hate crime occurred on - though not the date it was reported on. Figure @ref(fig:hateCrimesByDay) shows the percent of hates crimes between 1991 and 2018 that occurred on each day of the week. Interestingly, the most common days for a hate crime to occur is on Friday and Saturday, which is also when non-hate crimes most frequently occur. This suggests that hate crimes do follow the same trends - at least partially - as other crimes.

Each hate crime incident can cover up to 10 different crimes occurring - e.g. a person who burglarizes a synagogue and spray paints a swastika on the wall would have both vandalism and burglary reported in this data. With each crime, this data has the bias motivation for

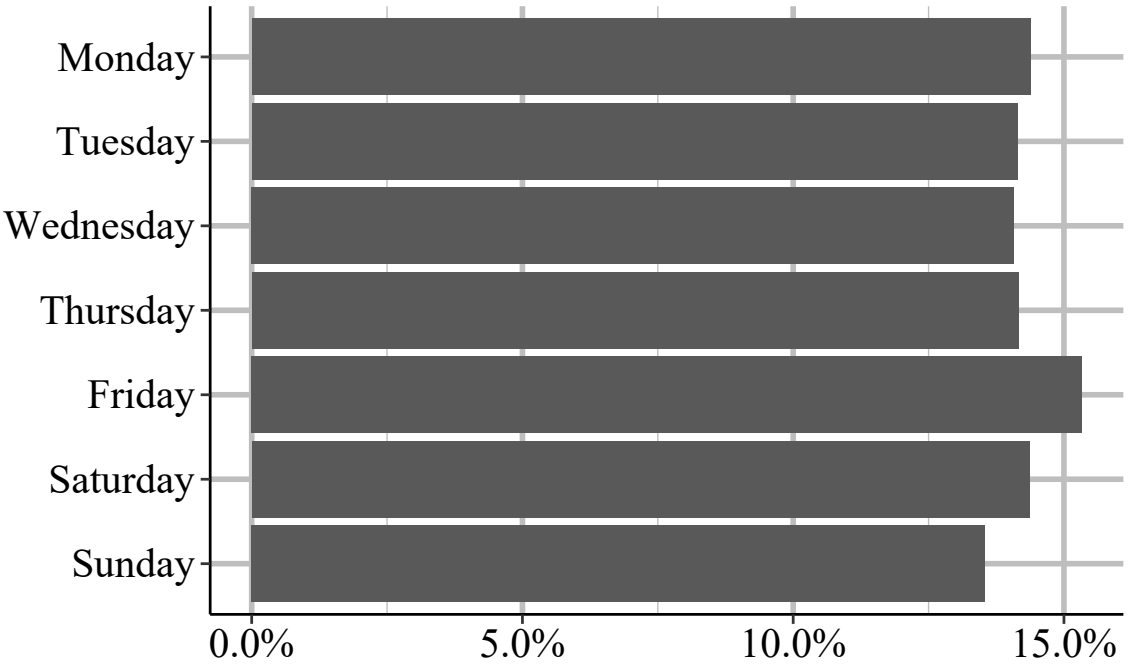


Figure 9.6: The day of the week that hate crimes occurred on, 1991-2018. (#fig:hateCrimesByDay)

that crime, the location of the crime (in broad categories, not the actual location in the city like a street address would have), and the number of victims for that offense. In practice, in most hate crimes with multiple offenses recorded, the bias motivation, location, and victim count is the same for each offense.

Figure @ref(fig:crimesPerHateCrime) shows the number of crimes per incident for each hate crime reported between 1991 and 2018. In 96.6% of cases, there is only one offense in that incident.[2] This drops sharply to 3.2% of incidents having two offenses, 0.21% having three offenses, 0.016% having four offenses, and 0.002% having five offenses. Even though this data does allow up to 10 offenses per hate crime incident, there has never been a recorded case with more than five offenses.

9.2.1 The bias motivation (who the hate is against)

The most important variable in this data is the "bias motivation" which is the FBI's term for the cause of the hate. A hate crime targeted against Black people would be an "anti-Black" bias motivation. For the police to classify an incident as a hate crime, and to assign a particular bias motivation, the police must have *some* evidence that the crime was motivated by hate. The victim saying that the crime is a hate crime alone is not sufficient - though if

[2] In 0.0005% of hate crimes there is no recorded offense. This is not shown in the graph.

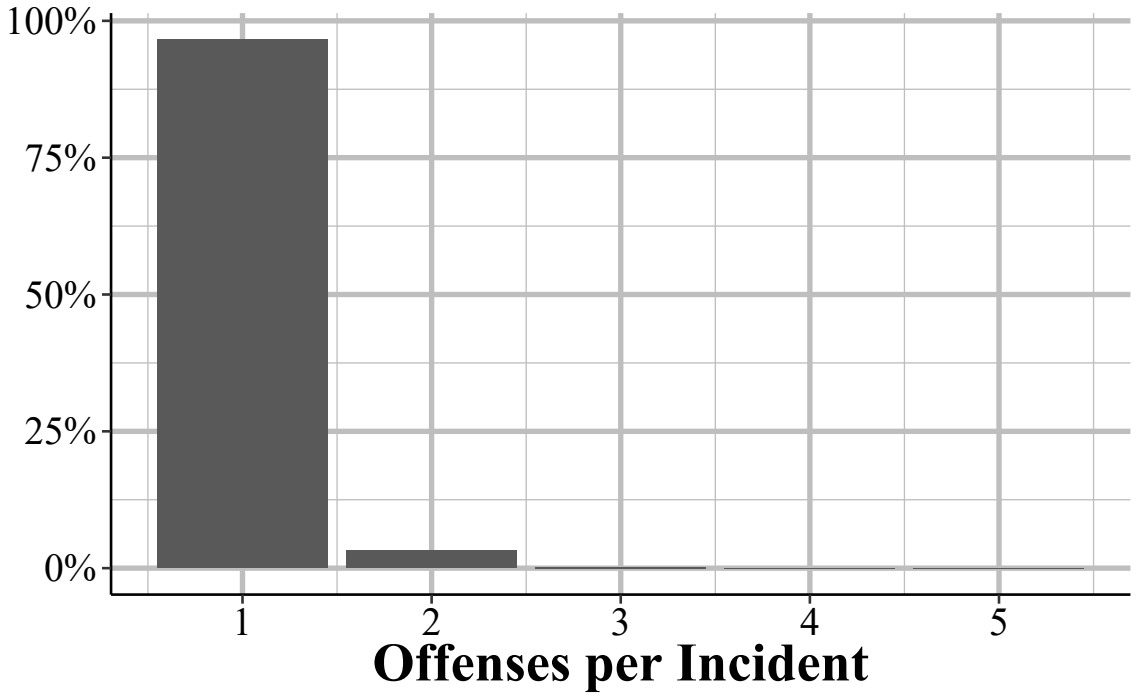


Figure 9.7: The number of offenses per hate crime incident. (#fig:crimesPerHateCrime)

large portions of the victim's community believe that the crime is a hate crime, this is a factor in the police's assessment. The evidence required is not major, it includes things as explicit as racial slurs said during an incident and less obvious factors like the victim is celebrating their community (e.g. attending a holiday event) or the crime occurring on an important holiday for that community (e.g. Martin Luther King Day, religious holidays). The FBI also encourages police to consider the totality of the evidence even if none alone strongly suggests that the crime was a hate crime in making their determination about whether the incident was a hate crime or not.

This also means that many (likely most) hate crimes will not be recorded as hate crimes since there is no evidence that the crime is motivated by hate. For example, if a man committed a crime against Asian people for crimes because they are Asian, that would in reality be a hate crime. However, if the offender does not say anything anti-Asian to the victim, which is the mostly likely thing to indicate that this is a hate crime, the crime would not likely be recorded as a hate crime. For example, at the time of this writing (Spring, 2021), there are numerous media reports discussing an increase in anti-Asian hate crimes as a result of racism relating to the pandemic.[3] This data would likely undercount both anti-Asian behavior and anti-Asian hate crimes. First, if someone walked to an Asian person and

[3]This dataset for 2020 won't be released until Fall 2021 so this data is not useful for measuring urgent policies about current events.

called them an anti-Asian slur, that is clearly a hateful act and would be classified as a hate crime under some organization's collections methods. However, as hateful as this incident is, *this alone* would not be classified as a hate crime in this dataset as a slur is not a crime. If accompanied by other criminal behavior, or if it continues to the point where it can be considered intimidation, it would then be classified as a hate crime. Second, crimes against Asian victims that are in fact hate crimes, but have no evidence that they are hate crimes would not be classified as hate crimes in this data.

Bias motivation is based on the offender's perceptions of the victim so even if they are incorrect in who their victim is, if they intended to target someone for their perceived group membership, that is still a hate crime. For example, if a person assaults a man because they think he is gay, that is a hate crime because the assault was motivated by hate towards gay people. Whether the victim is actually gay or not is not relevant - the offender perceived him to be gay so it is an anti-gay hate crime. To make this even more complicated, the offender must have committed the crime because they are motivated, at least to some degree, by their bias against the victim. Being biased against the victim but targeting them for some other reason means that the crime is not a hate crime.

The biases that the FBI includes in this data has grown over time, with new bias motivations being added in 1997, 2012, 2013, and 2015. Table @ref(tab:hateBiasMotivation) shows each bias motivation in this data, the year it was first reported, how many hate crimes there were for this bias motivation from 1991-2018 and what percent of hate crimes that bias motivation makes up. For ease of seeing how bias motivations changed over time, the table is sorted by year and by frequency of incident within that year. If you're viewing this on a computer or phone, you can sort the table yourself. The year is the first year with that bias motivation - as hate crimes for certain groups are very rare, the bias motivation could have technically been available in previous years. The last column in this table shows the percent of hate crimes from 1991-2018, though this is a very rough measure since some groups are so small that even accounting for a small percent of total hate crimes can mean that are massively disproportionately targeted.

Table 9.1: (#tab:hateBiasMotivation)The bias motivation for hate crime incidents. In incidents with multiple bias motivation, this shows only the first bias motivation recorded.

Bias Motivation	First Year Reported	# of Incidents	% of Incidents
Anti-Eastern Orthodox (Greek, Russian, Etc.)	2015	131	0.07%
Anti-Other Christian	2015	130	0.07%

Bias Motivation	First Year Reported	# of Incidents	% of Incidents
Anti-Sikh	2015	80	0.04%
Anti-Hindu	2015	40	0.02%
Anti-Mormon	2015	38	0.02%
Anti-Jehovahs Witness	2015	20	0.01%
Anti-Buddhist	2015	19	0.01%
Anti-Transgender	2013	513	0.26%
Anti-Male	2013	77	0.04%
Anti-Native Hawaiian Or Other Pacific Islander	2013	49	0.02%
Anti-Gender Non-Conforming	2012	152	0.08%
Anti-Female	2012	142	0.07%
Anti-Mental Disability	1997	944	0.47%
Anti-Physical Disability	1997	503	0.25%
Anti-Black	1991	68,588	34.35%
Anti-Jewish	1991	25,943	12.99%
Anti-White	1991	23,366	11.70%
Anti-Male Homosexual (Gay)	1991	20,188	10.11%
Anti-Hispanic	1991	12,725	6.37%
Anti-Ethnicity Other Than Hispanic	1991	10,016	5.02%
Anti-Lesbian, Gay, Bisexual, Or Transgender, Mixed Group (LGBT)	1991	5,990	3.00%
Anti-Asian	1991	5,891	2.95%
Anti-Multi-Racial Group	1991	4,811	2.41%
Anti-Female Homosexual (Lesbian)	1991	4,221	2.11%
Anti-Muslim	1991	3,561	1.78%
Anti-Other Religion	1991	3,269	1.64%
Anti-American Indian Or Native Alaskan	1991	2,195	1.10%
Anti-Catholic	1991	1,458	0.73%
Anti-Protestant	1991	1,196	0.60%
Anti-Multi-Religious Group	1991	1,118	0.56%
Anti-Arab	1991	1,091	0.55%
Anti-Heterosexual	1991	542	0.27%
Anti-Bisexual	1991	527	0.26%
Anti-Atheism/Agnosticism	1991	149	0.07%
Total	NA	199,683	100%

2015 is the most year of new bias motivations, as of data through 2018. This year added a

number of religions such as Anti-Buddhist, Anti-Sikh, and Anti-Jehovah's Witness. In 2013, anti-Transgender was added and this is the most common of the bias motivations added since data began in 1991 with 513 hate crimes between 2013-2018 - 0.26% of all hate crimes from 1991-2018. That year also added anti-male and Anti-Native Hawaiian or Other Pacific Islander, which is the most recent racial group added. In 2012, anti-gender non-conforming and anti-female were included, while in 1997 both anti-mental and anti-physical disability were added. In part due to having fewer years of data available, these newer bias motivations make up a small percent of total hate crimes, accounting for ~1.4%.

The original hate crimes - that is, those in the data in 1991 when this dataset was released - are far more common. The most common bias motivation is anti-Black at 34% of hate crimes, anti-Jewish at 13%, anti-White at 12%, anti-male homosexual (gay) at 10%, anti-Hispanic at 6%, and anti-ethnicity other than Hispanic (this group means a crime against an ethnic group that isn't Hispanic, though it is occasionally reported as anti-non-Hispanic which is incorrect.) at 5%. All other bias motivations are less than 5% of hate crimes and consist of a variety of ethnic, racial, religious, or sexual orientation. Some hate crimes can potentially fall in multiple categories. For example, there is a bias motivation of "anti-male homosexual (gay)" and of "anti-lesbian, gay, bisexual, or transgender, mixed group (LGBT)" so there is some overlap between them.

9.2.2 The crime that occurred

The "crime" part of hate crimes is which criminal offense occurred during the incident. A hateful act where the action is not one of the crimes that the FBI records would not be considered a hate crime. This is likely most common when considering something like a person calling someone by a hateful slur (e.g. "You're a [slur]", "go back to your own country") but where the action is not technically a crime. Another layer of difficulty in using this data is that not all crimes that the FBI includes were initially included when data become available in 1991. Every several years the FBI adds new crimes to be included in this data. Table @ref(tab:hateOffense) shows each crime in the data, the first year that this crime was reported, the total number of these crimes reported between 1991 and 2018, and the percent of all incidents this crime makes up.[4] This table is sorted with the most recent years first, and within year by how common the crime was.

Nearly all hate crimes are vandalism/destruction of property (31%), intimidation (30%), and simple assault (20%) or aggravated assault (11%) with no remaining crime making up more than 2% of total hate crimes.

[4]This tables uses only the first offense in an incident so counts are slightly lower than if all crimes in every incident is used.

Table 9.2: (#tab:hateOffense)The offense type for hate crime incidents. In incidents with multiple offense types, this shows only the first offense type recorded.

Offense	First Year Reported	# of Incidents	% of Incidents
Human Trafficking - Commercial Sex Acts	2017	1	0.00%
Fraud-Other	2016	17	0.01%
Bribery	2014	2	0.00%
Assisting Or Promoting Prostitution	2013	4	0.00%
Purchasing Prostitution	2013	2	0.00%
Wire Fraud	2006	18	0.01%
Impersonation	2001	108	0.05%
Prostitution	2001	12	0.01%
Statutory Rape	1999	12	0.01%
Theft From Coin-Operated Machine Or Device	1999	11	0.01%
Negligent Manslaughter	1999	3	0.00%
False Pretenses/Swindle/Confidence Game	1997	230	0.12%
Extortion/Blackmail	1997	41	0.02%
Incest	1997	7	0.00%
Stolen Property Offenses - Receiving, Selling, Etc.	1996	94	0.05%
Sexual Assault With An Object	1996	23	0.01%
Pocket-Picking	1996	19	0.01%
Welfare Fraud	1996	2	0.00%
Drug Equipment Violations	1995	204	0.10%
Credit Card/Atm Fraud	1995	125	0.06%
Embezzlement	1995	56	0.03%
Forcible Sodomy	1995	52	0.03%
Pornography/Obscene Material	1995	34	0.02%
Purse-Snatching	1995	23	0.01%
Theft From Building	1994	478	0.24%
Kidnapping/Abduction	1994	99	0.05%
All Other Larceny	1993	1,670	0.84%
Drug/Narcotic Violations	1993	862	0.43%
Theft From Motor Vehicle	1993	642	0.32%
Shoplifting	1993	521	0.26%
Weapon Law Violations	1993	289	0.14%
Theft of Motor Vehicle Parts/Accessories	1993	174	0.09%
Counterfeiting/Forgery	1993	165	0.08%

Offense	First Year Reported	# of Incidents	% of Incidents
Forcible Fondling - Indecent Liberties/Child Molest	1993	142	0.07%
Motor Vehicle Theft	1992	401	0.20%
Destruction of Property/Vandalism	1991	62,502	31.30%
Intimidation	1991	59,750	29.92%
Simple Assault	1991	39,035	19.55%
Aggravated Assault	1991	22,213	11.12%
Robbery	1991	3,748	1.88%
Burglary/Breaking And Entering	1991	3,363	1.68%
Arson	1991	1,224	0.61%
Theft-Other	1991	796	0.40%
Murder/Non-Negligent Manslaughter	1991	265	0.13%
Forcible Rape	1991	242	0.12%
Total	NA	199,681	100%

Agencies that report to the FBI's National Incident-Based Reporting System (NIBRS) can also report bias motivations for their crimes, and these reports are included in this dataset. One tricky thing is that the crimes included are different depending on if the agency reported through NIBRS or to the dataset directly, and are not NIBRS reporting agencies. The above table shows all crimes included, but agencies who report directly can only submit a subset of crimes. For these agencies the only possible crimes are the ones that are included Offenses Known and Clearances by Arrest dataset, that is detailed in Chapter @ref(offensesKnown), as well as vandalism/destruction of property and human trafficking.

9.2.3 The location of the crime

This data is interesting because it includes the location - in categories for types of places, not actual location in the city - of the incident. This is important since the type of location can be a factor in whether the incident is classified as a hate crime. For example, spray paint on a synagogue or a mosque is much more likely to be a hate crime than spray paint on a wall of an abandoned building. Table @ref(tab:hateLocations) shows the locations of hate crimes sorted by the year that location was first included in the data and then by frequency of hate crimes. Each hate crime incident can have multiple locations since each offense can have its own incident, but in most cases (96.6%) a hate crime only has a single location.

As with the crime and the bias motivation, the available locations have increased as time went on, though these newer locations are relatively uncommon. One important change in location is that starting in 2010 the location of school/college was split to have one location

be for elementary and high schools and another location be for colleges and universities. The majority of hate crimes occur in the victim's home (30%), on a road or alley (19%), in an unknown location (13.6%), and in a parking lot or parking garage. All other locations occur in fewer than 5% of hate crimes.

Table 9.3: (#tab:hateLocations)The location of hate crime incidents. In incidents with multiple locations, this shows only the first location recorded.

Location	First Year Reported	# of Incidents	% of Incidents
Military Installation	2015	3	0.00%
Community Center	2013	75	0.04%
Dock/Wharf/Freight/Modal Terminal	2012	14	0.01%
Shelter - Mission/Homeless	2011	68	0.03%
Rest Area	2011	44	0.02%
Arena/Stadium/Fairgrounds/Coliseum	2011	35	0.02%
Auto Dealership New/Used	2011	34	0.02%
Abandoned/Condemned Structure	2011	31	0.02%
Daycare Facility	2011	23	0.01%
Farm Facility	2011	16	0.01%
Amusement Park	2011	14	0.01%
Atm Separate From Bank	2011	9	0.00%
Tribal Lands	2011	6	0.00%
School - Elementary/Secondary	2010	1,748	0.88%
School - College/University	2010	1,248	0.63%
Park/Playground	2010	646	0.32%
Shopping Mall	2010	142	0.07%
Industrial Site	2010	65	0.03%
Camp/Campground	2010	23	0.01%
Gambling Facility/Casino/Race Track	2010	21	0.01%
Residence/Home	1991	59,606	29.86%
Highway/Road/Alley	1991	37,988	19.03%
Other/Unknown	1991	27,178	13.61%
School/College	1991	17,249	8.64%
Parking Lot/Garage	1991	11,440	5.73%
Church/Synagogue/Temple	1991	7,530	3.77%
Commercial/Office Building	1991	4,502	2.25%
Restaurant	1991	3,978	1.99%
Bar/Nightclub	1991	3,549	1.78%

Location	First Year Reported	# of Incidents	% of Incidents	
Government/Public Building		1991	2,894	1.45%
Convenience Store		1991	2,561	1.28%
Specialty Store - Tv, Fur, Etc.		1991	2,375	1.19%
Air/Bus/Train Terminal		1991	1,931	0.97%
Field/Woods		1991	1,917	0.96%
Service/Gas Station		1991	1,830	0.92%
Grocery/Supermarket		1991	1,628	0.82%
Department/Discount Store		1991	1,516	0.76%
Drug Store/Doctors Office/Hospital		1991	1,448	0.73%
Jail/Prison		1991	1,262	0.63%
Hotel/Motel		1991	1,214	0.61%
Construction Site		1991	507	0.25%
Bank/Savings And Loan		1991	418	0.21%
Liquor Store		1991	346	0.17%
Lake/Waterway		1991	326	0.16%
Rental Storage Facility		1991	188	0.09%
Total		NA	199,646	100%

9.2.4 Number and race of offenders

There are two variables that have information about the people who commit the hate crime: the number of offenders and the race of the offenders (as a single value with the race of the group if all are of the same race or it will say a "multi-racial" group). Unfortunately, important information like the age of the offenders, their criminal history, their relationship to the victim, their gender, or whether they are arrested are completely unavailable in this dataset.

When the police do not have any information about the number of offenders (which is common in cases of property crimes such as vandalism but rare in violent crimes), this data considers that to have zero offenders. The zero is just a placeholder that means that the police have no idea how many offenders there are, not that they think there were actually no offenders. Figure @ref(fig:hateCrimeOffenderNumber) shows the percent of hate crimes from 1991-2018 that have each number of offenders recorded. In the actual data it says the actual number of offenders, with the largest group in the current data going to 99 offenders - in this graph I group 10 or more offenders together for simplicity. I also relabel zero offenders as "Unknown" offenders since that's more accurate. The most common number of offenders per hate crime is one offender, at about 46% of hate crimes from 1991-2018 having only one

offender. This drops sharply to 9% of hate crimes having 2 offenders and continues to drop as the number of offenders increase. However, about a third (36.8%) of hate crimes have an unknown number of offenders.

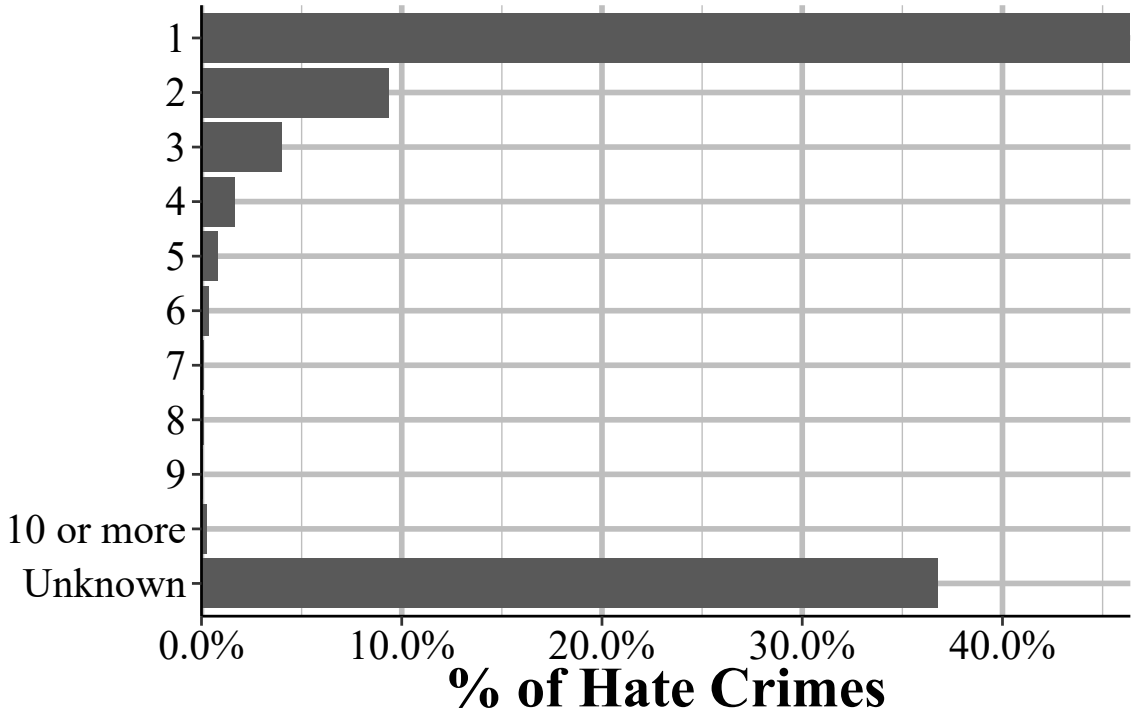


Figure 9.8: The race of offenders, as a group, for hate crime incidents, 1991-2018. (#fig:hateCrimeOffenderNumber)

The data also includes the race of the offenders as a group, though not the ethnicity (Hispanic or non-Hispanic) of the offenders. In cases where there are multiple races among offenders, that will be classified as a "multi-racial group". As shown in Figure @ref(fig:hateCrimeOffenderNumber) The most common racial group is "unknown" since the police do not know the race of the offenders. Then are White offenders at nearly 40% of hate crimes followed by Black offenders at nearly 13% of hate crimes. The remaining racial groups are rare with about 2% of hate crimes being committed by a multi-racial group of offenders and 0.72% of hate crimes committed by Asian or Pacific Islander offenders and 0.54% committed by American Indian or Native Alaskan offenders.

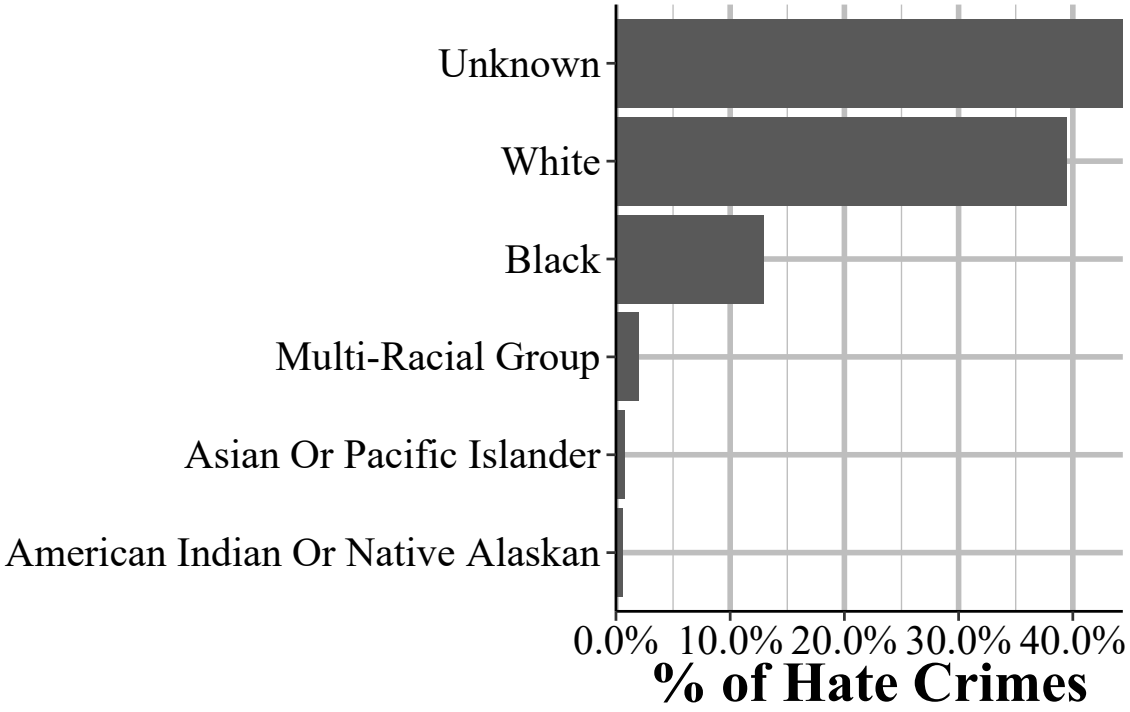


Figure 9.9: The race of offenders, as a group, for hate crime incidents, 1991-2018. (#fig:hateCrimeOffenderRace)

Chapter 10

County-Level UCR Data

UCR data is only available at the agency-level.[1] This has caused a lot of problems for researchers because many variables from other datasets (e.g. CDC data, economic data) is primarily available at the county-level. Their solution to this problem is to aggregate the data to the county-level by summing all the agencies in a particular county.[2]

More specifically, nearly all researchers who use this county-level UCR data use the National Archive for Criminal Justice Data (NACJD)'s datasets which have done the aggregation themselves (full disclosure, I used to have my own version of this data available on openICPSR and followed NACJD's method. My reasoning was that people were using it anyways and I wanted to make sure that they knew the problem of the data, so I included the issues with this data in the documentation when downloading it. However, I decided that the data was more flawed than I originally thought so I took down the data.).[3] These are not official FBI datasets but "UCR staff were consulted in developing the new adjustment procedures".[4] The "new" procedures is because NACJD changed their missing data imputation procedure starting with 1994 UCR data, and for this chapter I'll only focus on this "new" procedure.

There are two new problems with county-level UCR data: 1) agencies in multiple counties, 2) and agencies with missing data. I say main new problems as other problems with the data such as definition changes for crimes and underreporting of crimes are inherent in the agency-level data and thus transferred to the county-level data.

[1] Even for county-level agencies such as Sheriff's Offices, the data is only for crimes in that agency's jurisdiction. So the county sheriff reports crimes that they responded to but not crimes within the county that other agencies, such as a city police force, responded to.

[2] Because the county-level data imputes missing months, this dataset is *only* available at the annual-level, not at the monthly level.

[3] These files are known on NACJD's website as "Uniform Crime Reporting Program Data: County-Level Detailed Arrest and Offense Data."

[4] To be very clear, this chapter isn't a critique of NACJD, which is a great organization with a very talented team, but merely of a single dataset that they released using imputation methods from decades ago.

The first issue is in distributing crimes across counties when an agency is in multiple counties. If, for example, New York City had 100 murders in a given year, how do you create county-level data from this? UCR data only tells you how many crimes happened in a particular agency, not where in the jurisdiction it happened. So we have no idea how many of these 100 murders happened in Kings County, how many happened in Bronx County, and so on. UCR data does, however, tell you how many counties the agency is in and the population of each. They only do this for up to three counties so in cases like New York City you don't actually have every county the agency is part of.[5] NACJD's method is to distribute crimes *according to the population of the agency in each county.* In the New York City example, Kings County is home to about 31% of the people in NYC while Bronx County is home to about 17%. So Kings County would get 31 murders while Bronx County gets 17 murders. The problem with this is the crime is not evenly distributed by population. Indeed, crime is generally extremely concentrated in a small number of areas in a city. Even if 100% of the murders in NYC actually happened in Bronx County, only 17% would get assigned there. So for agencies in multiple counties - which is very common among the largest agencies in the country, and seems more common in some regions than others - the crimes assigned may be massively incorrect.

The second problem is that not all agencies report data, and even those that do may report only partially (e.g. report fewer than 12 months of the year). So by necessity the missing data has to be filled in somehow. While all methods FOR imputing missing data are wrong - as they will be different than the real data, though sometimes not by much - the ways done to impute the missing data in this dataset are particularly incorrect. We'll go into more detail about why they're incorrect in this chapter.

In this chapter, I'll then talk about the data as it is with a focus on why the data is flawed and shouldn't be used. To be clear, all data has measurement errors and much of this book is dedicated to talking about the errors in the other UCR datasets.

The problems in the county-level data, however, are egregious enough to merit special warning. This is by no means the first warning about this data. Most famously is Maltz and Targonski's (2002) paper in the Journal of Quantitative Criminology about the issues with this data. They concluded that "Until improved methods of imputing county-level crime data are developed, tested, and implemented, they should not be used, especially in policy studies" which is a conclusion I also hold. These warnings, as well as those from others have been largely (it appears) ignored.[6]

[5]For New York City specifically NACJD does distribute to all five counties, and does so by county population.

[6]Hopefully the warnings deterred at least some people.

10.1 Current usage

Even with the well-known flaws of this data, it remains a popular dataset. A search on Google Scholar for "county-level UCR" returns 3,780 results as of this writing in summer 2021. About half of these results are from 2015 or later. In addition to use by researchers, the county-level UCR data is used by organizations such as the FBI in their annual Crimes in the United States report (which is essentially the report that informs the media and the public about crime, even though it's actually only a subset of actual UCR data) and Social Explorer, a website that makes it extremely convenient to examine US Census data. Based on my reading of UCR papers there are also some differences in data usage by field.[7]

This data is most widely used (as a share of papers in the field) by economics researchers and fields other than criminology such as the relatively rare psychologist or political scientist studying crime.[8] Criminologists, however, tend to focus more on agency-level data rather than county-level data, though many criminologists still do use county-level UCR data.[9]

10.2 How much data is missing

A major problem with county-level UCR data is that some data is missing, and is then imputed (poorly). So to understand how much of an issue missing data is in the county-level UCR data, we'll now look at missingness in the Offenses Known and Clearances by Arrests dataset which is the "crime data" of the UCR. County-level data also aggregates arrests from the Arrests by Age, Sex, and Race dataset, which has lower reporting (and thus more missingness) than the Offenses Known and Clearances by Arrest data. But for simplicity we'll only look at the crime dataset. We'll do so in a number of different ways to try to really understand how much data is missing and how it changed over time.

For each of the below graphs and tables we use the Offenses Known and Clearances by Arrest data for 1960-2018 and exclude any agency that are "special jurisdictions". Special

[7] I used to have a Google Scholar alert for UCR papers but turned it off since so many papers either used my data but didn't cite me, used the data improperly, or both.

[8] Given economist's policy of not citing other fields it's likely that they haven't read Maltz and Targonski's paper. Normally I'd be concerned with criticizing entire fields, but no economist will read this as I'm not an economist myself so it's fine.

[9] I note these differences not to join the ego-driven feud between the fields. As a criminologist who currently works in a political science department, these feuds are quite stupid. I note it, however, to show that this data usage is partially driven by subcultures within fields. For example, criminology has long known about which may explain why it is more commonly used in non-criminology fields. It is therefore important for researchers to talk to (or collaborate with {though to be clear I'm not volunteering or requesting to collaborate with you}) people who have content-expertise about the data that you're using. And that open communication between the fields (including collaborations and submitting to journals from fields other than your own) will strengthen both fields.

jurisdiction agencies are, as it seems, special agencies that tend to have an extremely specific jurisdiction and goals. These include agencies such as port authorities, alcohol beverage control, university police, and airport police. These agencies tend to cover a tiny geographic area and have both very low crime and very low reporting rates.[10] So to prevent missingness being overcounted due to these weird agencies I'm excluding them from the below examples. I'm also excluding federal agencies (though UCR only has 7 reporting federal agencies) as these operate much the same as special jurisdiction agencies.

We'll first look at how many months are reported in the example year of 2018, though we'll see below that 2018 is pretty similar to other years. Table @ref(tab:countyMonthsReportedDefinitions) shows the number of months reported using two definitions. The first is how the FBI, and NACJD when imputing data, classifies number of months reported and this is actually just the last month reported. So if an agency reports only in July they are classified as reporting 7 months; if they report only in December they are classified as reporting 12 months. Whether they actually report previous months doesn't matter based on this definition.

The second definition is my own - and available in the data I've released on openICPSR - and is based on how many months the data says the agency reported. That is, for each month the UCR data actually says if they received a report or not. So this is a superior way of measuring though not full-proof as some months say they have a report but don't and some say they don't have a report but do have crime recorded. Also, in 2018 the FBI changed how they classify this variable so now every month is reported for every agency, making the measurement useless for 2018 and more recent data.

The table shows what percent of agencies that reported data had data for each possible number of months: 0 through 12 months. Column 2 shows the percent for the 1st method while column 3 shows the percent for my method. And the final column shows the percent change from moving from the 1st to 2nd measure.

Ultimately the measures are quite similar though systematically overcount reporting using the 1st method. Both show that about 23% of agencies reported zero months. The 1st method has nearly 73% of agencies reporting 12 months while the 2nd method has 69%, a difference of about 5% which is potentially a sizable difference depending on exactly which agencies are missing. The remaining nearly 4% of agencies all have far more people in the 2nd method than in the first, which is because in the 1st method those agencies are recorded as having 12 months since they reported in December but not actually all 12 months of the year. There are huge percent increases in moving from the 1st to 2nd method for 1-11 months reported though this is due to having very few agencies report this many months. Most months have only about 50 agencies in the 1st method and about 70 in the 2nd, so the actual difference is not that large.

[10] Even though these are unusual agencies, in real analyses using UCR data at the county-level you'd like want to include them. Or justify why you're not including them.

Table 10.1: (#tab:countyMonthsReportedDefinitions)The number of months reported to the 2017 Offenses Known and Clearances by Arrest data using two definitions of months reported. The 'Last Month' definition is the preferred measure of months reported by both the FBI and researchers, though this overcounts months.

Months Reported	Last Month Definition	Months Not Missing Definition	Percent Difference
0	4,481 (23.01%)	4,514 (23.18%)	+0.74
1	93 (0.48%)	112 (0.58%)	+20.43
2	36 (0.18%)	70 (0.36%)	+94.44
3	40 (0.21%)	156 (0.8%)	+290.00
4	44 (0.23%)	66 (0.34%)	+50.00
5	42 (0.22%)	78 (0.4%)	+85.71
6	48 (0.25%)	78 (0.4%)	+62.50
7	55 (0.28%)	81 (0.42%)	+47.27
8	56 (0.29%)	87 (0.45%)	+55.36
9	56 (0.29%)	102 (0.52%)	+82.14
10	158 (0.81%)	242 (1.24%)	+53.16
11	199 (1.02%)	419 (2.15%)	+110.55
12	14,162 (72.74%)	13,465 (69.16%)	-4.92

We can look at how these trends change over time in Figure @ref(fig:countyAnyMonthReported) that shows the annual number of agencies that reported at least one month of data in that year. About 5,500 agencies reported at least on month throughout the 1960s and then grew rapidly over the next decade until about 12,500 agencies reported in the end of the 1970s. This declined over the next two decades before again increasing in the mid-late 1990s where it steadily increased to about 14,000 agencies in 2010 and has stagnated since then, with a small dip in 2018. Out of the approximately 18,000 police agencies in the United States, this is relatively low reporting even as far as recent decades.

We saw in Table @ref(tab:countyMonthsReportedDefinitions) that when agencies do report they tend to report all 12 months of the year. Figure @ref(fig:countyDecemberReported) examines whether this is true by showing the number of agencies in each year that reported in December, which the way that county-level UCR data assumes that all 12 months are reported. The trends are nearly identical to Figure @ref(fig:countyAnyMonthReported) but with several hundred fewer agencies reporting in December each year than reporting at least one month. This shows that the trend of when agencies do report they tend to report all 12

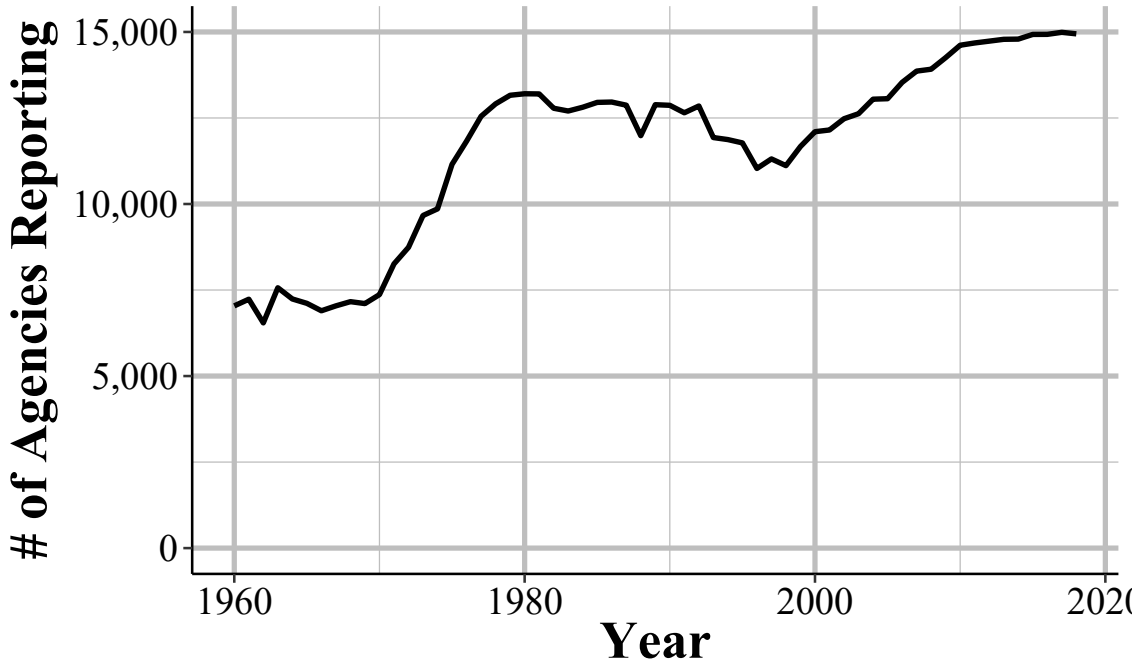


Figure 10.1: The annual number of agencies that reported at least one month of data in that year.
(#fig:countyAnyMonthReported)

months (or at least report in December) is consistent over time.

Another way to look at this is to examine, as Figure @ref(fig:countyDecemberPercent) does, the percent of agencies that report in December of agencies that report at least one month. On average about 92% of agencies that do report any month also report in December, and this number has steadily grown over time, though declined in 2018 to only 90%.

Since the number of agencies reporting changes every year - generally increasing over time - we can look at the percent of agencies that reported in December out of all agencies that reported data (and this includes reporting zero months of data) to UCR. Figure @ref(fig:countyDecemberPercentAnyAgency) shows this trend over time and about 75% of agencies that submit to UCR each year report in December. Reporting rates undulate over this time period - the low is 62% in 1997 and the high is 83% in 1979 - but tend to return to ~75% reporting before trending in the opposite direction again.

Not all agencies report data to UCR, even to say that they're not reporting any months of data. In 1960, for example, only 8,452 agencies reported data to UCR and 1,406 of these reported zero months of data. To get accurate data on county crime you'll want data from all agencies, not just ones that reported (or told the FBI they weren't reporting) data. Figure @ref(fig:countyDecemberPercentAllAgencies) shows the annual percent of agencies that reported in December of each year out of the 19,036 agencies that ever reported to UCR. This 19,036 is higher than the ~18,000 agencies often discussed (including in this

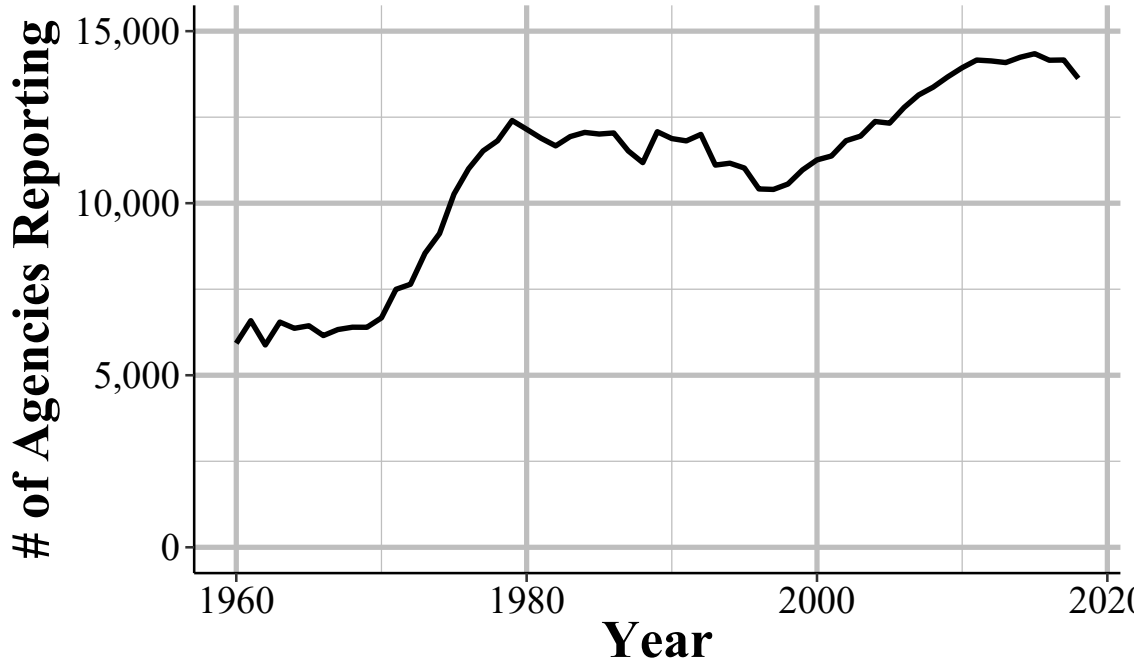


Figure 10.2: The annual number of agencies that reported data in December of that year (which by the FBI's definition would mean they reported 12 months of the year.
(#fig:countyDecemberReported)

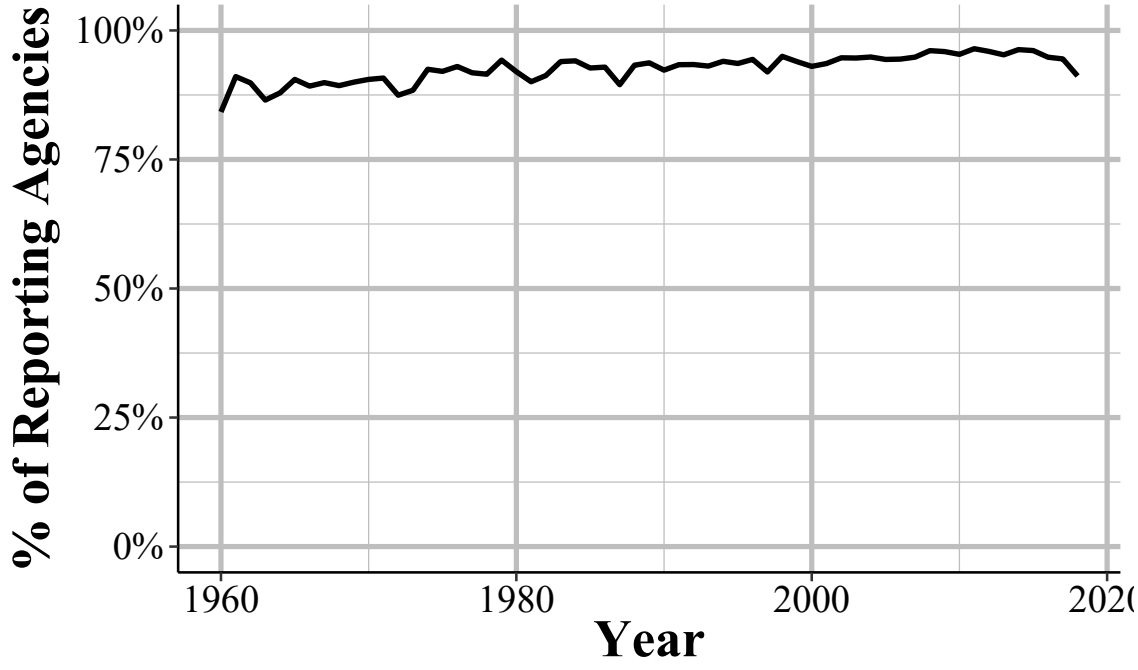


Figure 10.3: The annual percent of agencies that reported in December of that year of those that reported at least one month of data.
(#fig:countyDecemberPercent)

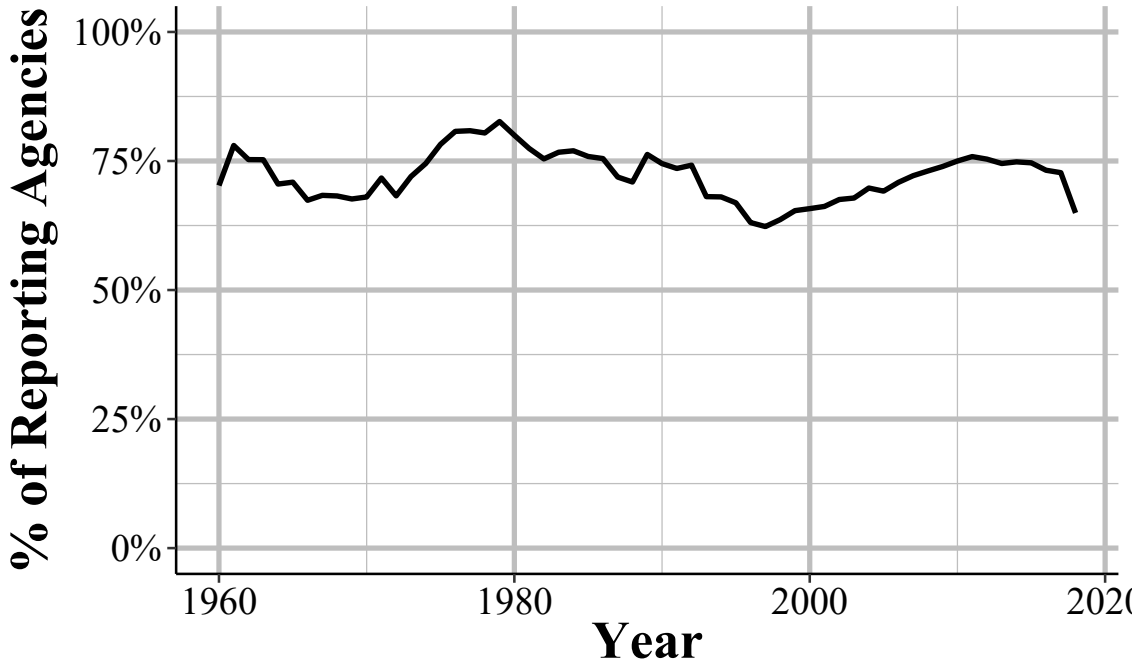


Figure 10.4: The annual percent of agencies that reported in December of that year including those that did not report any data that year.
(#fig:countyDecemberPercentAnyAgency)

book) by academics as the approximate number of agencies. I believe that the number here is higher because it includes agencies that may have closed or been swallowed by a larger or nearby agency.

The trends in Figure @ref(fig:countyDecemberPercentAllAgencies) are very similar to those in Figure @ref(fig:countyAnyMonthReported). In the first decade of data only about 25% of agencies reported December data and this rose rapidly to over 60% of agencies in the late 1970s. From here it rather steadily declined until it bottomed out at nearly 50% in 1993 and began rising again. Reporting rates peaked at the mid-60% starting in 2010 and remained stagnant until declining to 62% in 2018.

The zero reporting agencies are largely in smaller agencies - with the average sized agency in 2018 having a population of 1,682 and the largest having a population of 463,545 - and, in 2018, amounts to agencies covering about 13 million people, or about 4% of the United States population.

10.3 Current imputation practices

There are three paths that the county-level UCR data takes when dealing with the agency-level data before aggregating it to the county-level. The path each agency is on is dependent

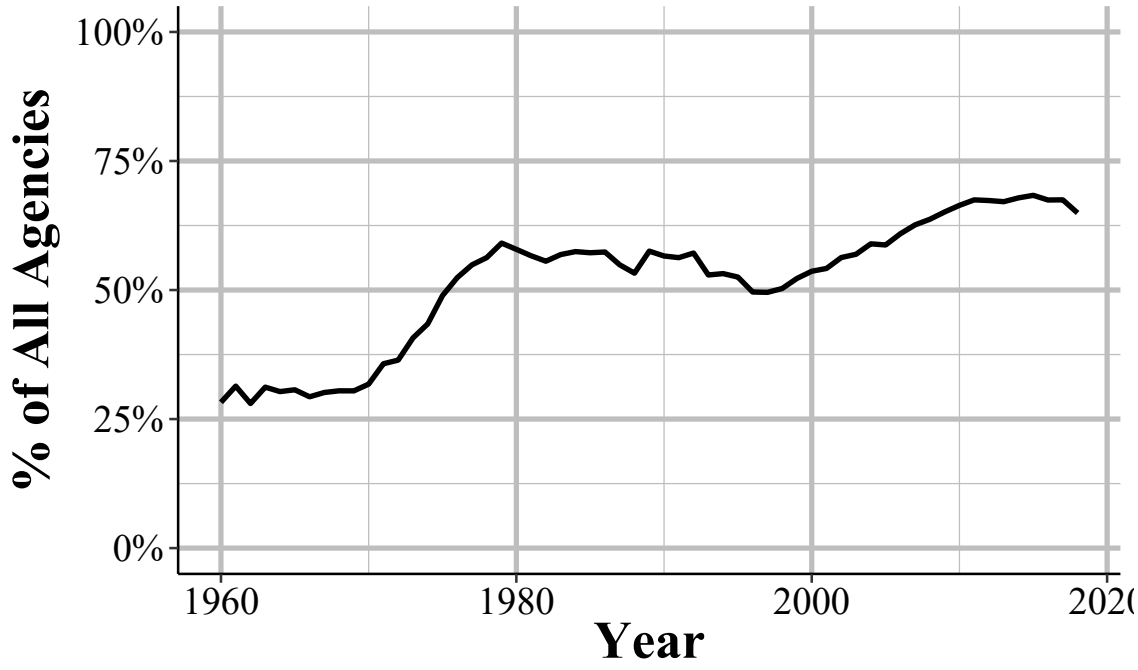


Figure 10.5: The annual percent of agencies that reported in December of that year out of all agencies that ever reported to the Offenses Known and Clearances by Arrest dataset (N=19,036).
(#fig:countyDecemberPercentAllAgencies)

on how many months of data they report. Figure @ref(fig:countyImputation) shows each of these three paths. We'll look in detail at these paths in the below sections, but for now we'll briefly summarize each path.

First, if an agency reports only two or fewer months, the entire agency's data (that is, any month that they do report) is deleted and their annual data is replaced with the average of agencies that are: 1) in the same state, 2) in the same population group (i.e. very roughly the same size), 3) and that reported all 12 months of the year (i.e. reported in December but potentially not any other month).

When an agency reports 3-11 months, those months of data are multiplied by 12/numbers-of-months-reported so it just upweights the available data to account for the missing months, assuming that missing months are like the present months.

Finally, for agencies that reported all 12 months there's nothing missing so it just uses the data as it is.

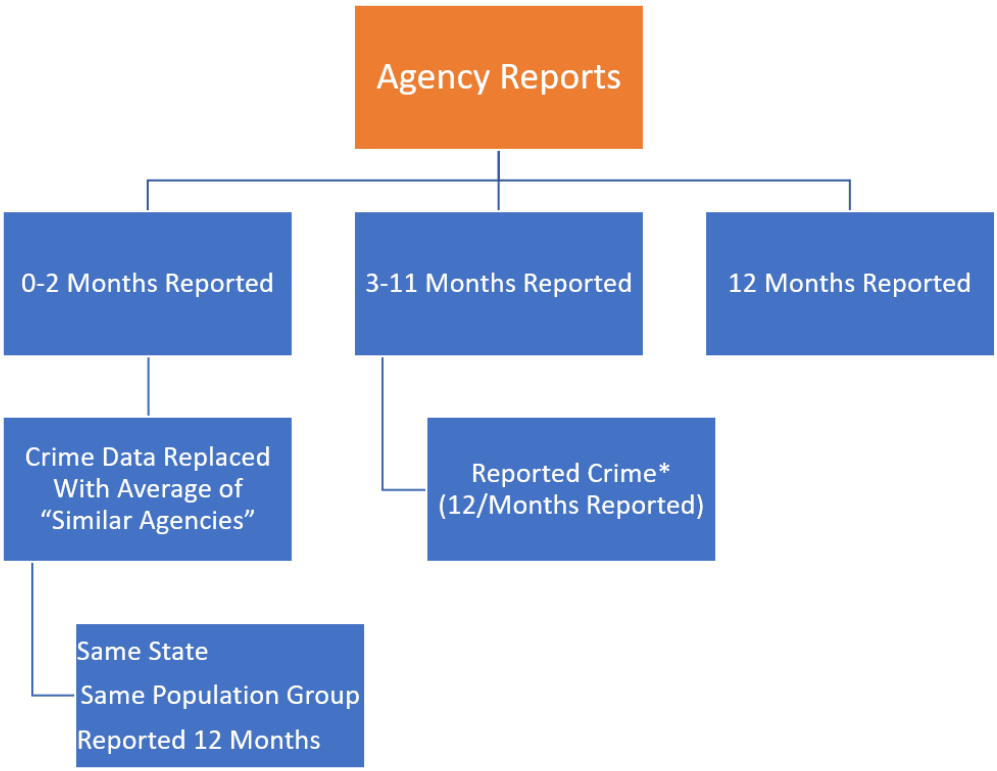


Figure 10.6: The imputation procedure for missing data based on the number of months missing.
(#fig:countyImputation)

10.3.1 1-9 months missing

When there are 1-9 months reported the missing months are imputed by multiplying the reported months of data by 12/numbers-of-months-reported, essentially just scaling up the reported months. For example, if 6 months are reported then it multiplies each crime values by 12/6=2, doubling each reported crime value. This method makes the assumption that missing values are similar to present ones, at on average. Given that there are seasonal differences in crime - which tends to increase in the summer and decrease in the winter - how accurate this replacement is depends on how consistent crime is over the year and which months are missing. Miss the summer months and you'll undercount crime. Miss the winter months and you'll overcount crime. Miss random months and you'll be wrong randomly (and maybe it'll balance out but maybe it won't).

We'll look at a number of examples and simulations about how accurate this method is. For each example we'll use agencies that reported in 2018 so we have a real comparison when using their method of replacing "missing" months.

Starting with Table @ref(tab:CountyPhillyMurders), we will see the change in the actual annual count of murders in Philadelphia when replacing data from each month. Each row shows what happens when you assume that month - and only that month - is missing and interpolate using the current 12/numbers-of-months-reported method. Column 1 shows the month that we are "replacing" while column 2 shows the actual number of murders in that month and the percent of annual murders in parentheses. Column 3 shows the actual annual murders which is 351 in 2018; column 4 shows the annual murder count when imputing the "missing" month" and column 5 shows the percent change between columns 3 and 4.

If each month had the same number of crimes we'd expect each month to account for 8.33% of the year's total. That's not what we're seeing in Philadelphia for murders as the percentages range from 5.13% in both January and April to 12.25% in December. This means that replacing these months will not give us an accurate count of crimes as crime is not distributed evenly across months. Indeed, as seen in column 5, on average, the annual sum of murders when imputing a single month is 1.85% off from the real value. When imputing the worst (as far as its effect on results) months you can report murder as either 4.27% lower than it is or 3.5% higher than it is.

Table 10.2: (#tab:CountyPhillyMurders)The imputed number of murders in Philadelphia in 2018 when missing a single month. This shows how different the imputed value is to the real value for each month missing.

Month	Murders That Month	Actual Annual Murders	Imputed Annual Murders	Percent Change
January	18 (5.13%)	351	363	+3.50
February	26 (7.41%)	351	354	+1.01
March	27 (7.69%)	351	353	+0.70
April	18 (5.13%)	351	363	+3.50
May	33 (9.40%)	351	346	-1.17
June	26 (7.41%)	351	354	+1.01
July	27 (7.69%)	351	353	+0.70
August	41 (11.68%)	351	338	-3.65
September	32 (9.12%)	351	348	-0.85
October	27 (7.69%)	351	353	+0.70
November	33 (9.40%)	351	346	-1.17
December	43 (12.25%)	351	336	-4.27

Part of the reason for the percent difference for murders when replacing a month found above is that there was high variation in the number of murders per month with some months having more than double the number as other months. We'll look at what happens when crimes are far more evenly distributed across months in Table @ref(tab:countyPhillyThefts). This table replicates Table @ref(tab:CountyPhillyMurders) but uses thefts in Philadelphia in 2018 instead of murders. Here the monthly share of thefts ranged only from 6.85% to 9.16% so month-to-month variation is not very large. Now the percent change never increases above an absolute value of 1.62 and changes by an average of 0.77%. In cases like this, the imputation method is less of a problem.

Table 10.3: (#tab:countyPhillyThefts)The imputed number of thefts in Philadelphia in 2018 when missing a single month. This shows how different the imputed value is to the real value for each month missing.

Month	Thefts That Month	Actual Annual Thefts	Imputed Annual Thefts	Percent Change
January	2,720 (7.36%)	36,968	37,361	+1.06

Month	Thefts That Month	Actual Annual Thefts	Imputed Annual Thefts	Percent Change
February	2,532 (6.85%)	36,968	37,566	+1.62
March	2,598 (7.03%)	36,968	37,494	+1.42
April	3,200 (8.66%)	36,968	36,837	-0.35
May	3,215 (8.70%)	36,968	36,821	-0.40
June	3,226 (8.73%)	36,968	36,809	-0.43
July	3,312 (8.96%)	36,968	36,715	-0.68
August	3,454 (9.34%)	36,968	36,560	-1.10
September	3,287 (8.89%)	36,968	36,742	-0.61
October	3,386 (9.16%)	36,968	36,634	-0.90
November	2,906 (7.86%)	36,968	37,158	+0.52
December	3,132 (8.47%)	36,968	36,912	-0.15

Given that the imputation method is largely dependent on consistency across months, what happens when crime is very rare? Table @ref(tab:countyDanvilleVehicle) shows what happens when replacing a single month for motor vehicle thefts in Danville, California, a small town which had 22 of these thefts in 2018. While possible to still have an even distribution of crimes over months, this is less likely when it comes to rare events. Here, having so few motor vehicle thefts means that small changes in monthly crimes can have an outsize effect. The average absolute value percent change now is 7.3% and this ranges from a -15.68% difference to a +9.1% difference from the real annual count. This means that having even a single month missing can vastly overcount or undercount the real values.

Table 10.4: (#tab:countyDanvilleVehicle)The imputed number of motor vehicle thefts in Danville, California, in 2018 when missing a single month. This shows how different the imputed value is to the real value for each month missing.

Month	Vehicle Thefts That Month	Actual Annual Vehicle Thefts	Imputed Annual Vehicle Thefts	Percent Change
January	3 (13.64%)	22	20	-5.77
February	0 (0.00%)	22	24	+9.09
March	1 (4.55%)	22	22	+4.14
April	2 (9.09%)	22	21	-0.82
May	3 (13.64%)	22	20	-5.77
June	1 (4.55%)	22	22	+4.14

Month	Vehicle Thefts That Month	Actual Annual Vehicle Thefts	Imputed Annual Vehicle Thefts	Percent Change
July	1 (4.55%)	22	22	+4.14
August	0 (0.00%)	22	24	+9.09
September	5 (22.73%)	22	18	-15.68
October	5 (22.73%)	22	18	-15.68
November	1 (4.55%)	22	22	+4.14
December	0 (0.00%)	22	24	+9.09

In the above three tables we looked at what happens if a single month is missing. Below we'll look at the results of simulating when between 1 and 9 months are missing for an agency. Table @ref(tab:countyPhillyMurderMonthsMissing) looks at murder in Philadelphia again but now randomizes removing between 1 and 9 months of the year and interpolating the annual murder count using the current method. For each number of months removed I run 10,000 simulations.[11] Given that I am literally randomly choosing which months to say are missing, I am assuming that missing data is missing completely at random. This is a very bold assumption and one that is the best-case scenario since it means that missing data is not related to crimes, police funding/staffing, or anything else relevant. So you should read the below tables as the most optimistic (and thus likely wrong) outcomes.

For each number of months reported the table shows the actual annual murder (which never changes) and the imputed mean, median, modal, minimum, and maximum annual murder count. As a function of the randomization, the imputed mean is always nearly identical to the real value. The most important columns, I believe, are the minimum and maximum imputed value since these show the worst-case scenario - that is, what happens when the month(s) least like the average month is replaced. Since as researchers we should try to minimize the harm caused from our work if it is wrong, I think it is safest to assume that if data is missing it is missing in the worst possible way. While this is a conservative approach, doing so otherwise leads to the greatest risk of using incorrect data, and incorrect results - and criminology is a field important enough to necessitate this caution.

As might be expected, as the number of months missing increases the quality of the imputation decreases. The minimum is further and further below the actual value while the maximum is further and further above the actual value.

[11] This is actually more than I need to run to get the same results but it's easier to run many simulations than to math out how many I actually need to find all possible combinations of missing months.

Table 10.5: (#tab:countyPhillyMurderMonthsMissing)A simulation showing how the imputed values of murders in Philadelphia in 2018 changes as the number of months to impute changes. For each number of months missing (and thus, imputed) 10,000 simulations are run for removing and imputing those months of data.

# of Months Missing	Actual Murder	Mean Imputed Murder	Median Imputed Murder	Modal Imputed Murder	Min Imputed Murder	Max Imputed Murder
1	351	351.00	353.45	353.45	336.00	363.27
2	351	351.10	350.40	350.40	320.40	378.00
3	351	350.84	352.00	353.33	312.00	385.33
4	351	350.96	349.50	370.50	301.50	394.50
5	351	351.05	351.43	353.14	289.71	404.57
6	351	350.79	350.00	376.00	284.00	418.00
7	351	350.71	350.40	348.00	276.00	436.80
8	351	351.71	354.00	312.00	264.00	450.00
9	351	351.40	348.00	344.00	248.00	468.00

This problem is even more pronounced when looking at agencies with fewer crimes and less evenly distributed crimes. Table @ref(tab:countyDanvilleBurglaryMonthsMissing) repeats the above table but now looks at motor vehicle thefts in Danville, California. By the time 5 months are missing, the minimum value is nearly half of the actual value while the maximum value is a little under 50% larger than the actual value. By 9 months missing, possible imputed values range from 0% of the actual value to over twice as large as the actual value.

Table 10.6: (#tab:countyDanvilleBurglaryMonthsMissing)A simulation showing how the imputed values of motor vehicle thefts in Danville, California, in 2018 changes as the number of months to impute changes. For each number of months missing (and thus, imputed) 10,000 simulations are run for removing and imputing those months of data.

# of Months Missing	Actual Vehicle Theft	Mean Imputed Vehicle Theft	Median Imputed Vehicle Theft	Modal Imputed Vehicle Theft	Min Imputed Vehicle Theft	Max Imputed Vehicle Theft
1	22	22.01	22.91	22.91	18.55	24.00
2	22	22.00	22.80	25.20	14.40	26.40
3	22	22.03	22.67	21.33	12.00	29.33
4	22	22.07	22.50	19.50	9.00	31.50
5	22	22.02	22.29	22.29	6.86	34.29
6	22	22.00	22.00	20.00	6.00	38.00
7	22	22.04	21.60	24.00	4.80	43.20
8	22	21.91	21.00	21.00	3.00	48.00
9	22	22.11	24.00	24.00	0.00	52.00

10.3.2 10-12 months missing

In cases where there are more than 9 months of data missing, the current imputation method replaces the entire year of data for that agency with the average of the crime for agencies who reported 12 months of data, are in the same state and in the same population group as the given agency. Considering that when an agency reports data it tends to report every month of the year - and about a quarter of agencies still do not report any months of data - this is a far bigger issue than when agencies are missing 1-9 months of data. The imputation process is also far worse here.

Whereas with 1-9 months missing the results were at least based on the own agency's data, and were actually not terribly wrong (depending on the specific agency and crime patterns) when only a small number of months were missing, the imputation for 10+ months missing is nonsensical. It assumes that these agencies are much like similarly sized agencies in the same state.

There are two major problems here. First, similarly sized agencies are based on the popu-

lation group which is quite literally just a category indicating how big the agency is when grouped into rather arbitrary categories. These categories can range quite far - with agencies having millions more people than other agencies in the same category in some cases - so in most cases "similarly sized" agencies are not that similarly sized. The second issue is simply the assumption that population is all that important to crime rates. Population is certainly important to crime counts; New York City is going to have many more crimes than small towns purely due to its huge population, even though NYC has a low crime rate. But there is still huge variation in crimes among cities of the same or similar size as crime tends to concentrate in certain areas. So replacing an agency's annual crime counts with that or other agencies (even the average of other agencies) will give you a very wrong count.

For this method of replacing missing data to be accurate agencies in the same population group in each state would need to have very similar crime counts. Otherwise it is assuming that missing agencies are just average (literally) in terms of crime. This again assumes that missing data is missing at random, which is unlikely to be true.

In each of the below examples we use data from 2018 Offenses Known and Clearances by Arrest and use only agencies whose final month reported was December. This makes it the actual agencies in each population group that would replace agencies that are missing 10 or more months of data in 2018. As agencies can - and do - report different numbers of months each year, these numbers would be a little different if using any year other than 2018.

For each population group we'll look at the mean, median, and maximum number of murders plus aggravated assaults with a gun.[12] This is essentially a measure of the most serious violent crimes as the difference between gun assaults and murders is, to some degree, a matter of luck (e.g. where the person is shot can make the difference between an assault and a murder).[13] This is actually not available in NACJD's county-level UCR data as they don't separate gun assaults from other aggravated assaults, though that data is available in the agency-level UCR data. If we see a wide range in the number of murders+gun-assaults in the below table, that'll indicate that this method of imputing missing data is highly flawed.

Table @ref(tab:countyPopulationGroupStatsNational) shows these values for all agencies in the United States who reported 12 months of data (based on the "December last month reported" definition) in 2018. The actual imputation process only looks at agencies in the same state, but this is still information at seeing broad trends - and we'll look at two specific states below. Column 1 shows each of the population groups in the data while the remaining columns show the mean, median, minimum, and maximum number of murders+gun-assaults in 2018, respectively.[14] For each population group there is a large range of values, as seen

[12] Aggravated assaults with a gun include but are not limited to shootings. The gun does not need to be fired to be considered an aggregated assault.

[13] Attempted murders are considered aggravated assaults in the UCR.

[14] The agency-level UCR data actually has more population groups than this list, but NACJD has grouped

from the minimum and maximum values. There are also large differences in the mean and median values for larger (25,000+ population) agencies, particularly when compared to the top and bottom of the range of values.[15] Using this imputation method will, in most cases (but soon we'll see an instance where there's an exception) provide substantially different values than the real (but unknown) values.

Table 10.7: (#tab:countyPopulationGroupStatsNational)The mean, median, minimum, and maximum agency size nationwide for all population groups in the 2018 Offenses Known and Clearances by Arrests data.

Population Group	Mean Murder + Gun Assault	Median Murder + Gun Assault	Minimum Murder + Gun Assault	Max Murder + Gun Assault
City Under 2,500	0	0	0	108
City 2,500-9,999	1	0	0	55
City 10,000-24,999	6	2	0	220
City 25,000-49,999	13	4	0	381
City 50,000-99,999	34	15	0	608
City 100,000-249,999	131	64	1	1,235
City 250,000+	1,093	638	4	6,209
MSA Counties And MSA State Police	19	3	-1	1,701
Non-MSA Counties And Non-MSA State Police	3	1	0	295

To better see the distribution, Figure @ref(fig:countyPopulationGroupsBoxplot) shows boxplots for the rate per 100,000 population of murders+gun-assaults in each population group in 2018. Since there can't be a rate without any population, this excludes all agencies with a population of zero. To make the graph simpler to see, it also excludes all agencies with a rate about 500, which is 57 agencies. Each population group has a relatively large range of values with the size of the boxplot growing as the size of the population group increases. This means that there are a wider variety of rates in larger population groups than in smaller population groups. Each population group also has a good number of agencies that are outliers with a much higher rate than is normal.

some together. Given that some states may have few (or no) agencies in a population group, combining more groups together does alleviate the problem of having no comparison cities but at the tradeoff of making the comparison less similar to the given agency.

[15]The negative number for minimum crimes in the Non-MSA Counties and Non-MSA State Police is due to a reporting quirk of UCR, covered in Chapter 2, and is not a mistake in the data.

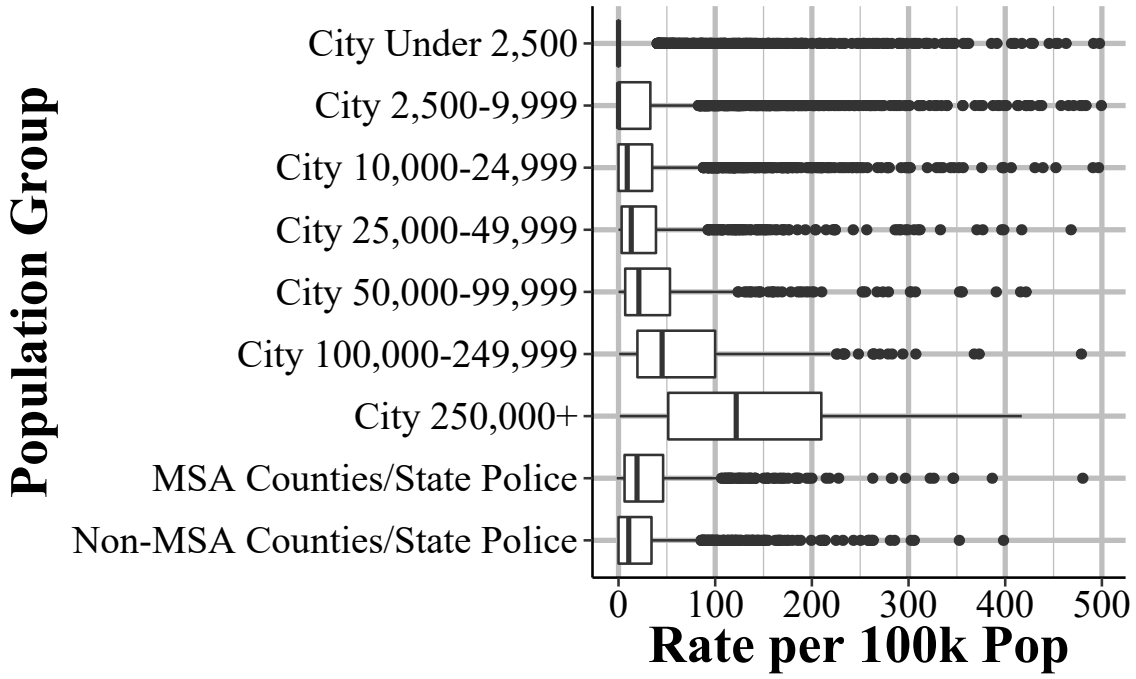


Figure 10.7: Boxplots showing the distribution of annual murders+gun-assaults for each population group for all agencies that reported December data in 2018. (#fig:countyPopulationGroupsBoxplot)

Since the actual imputation process looks only at agencies in the same state, we'll look at two example states - Texas and Maine - and see how trends differ from nationally. These states are chosen as Texas is a very large (both in population and in number of jurisdictions) state with some areas of high crime while Maine is a small, more rural state with very low crime. Table @ref(tab:countyPopulationGroupStatsTexas) shows results in Texas. Here, the findings are very similar to that of Table @ref(tab:countyPopulationGroupStatsNational). While the numbers are different, and the maximum value is substantially smaller than using all agencies in the country, the basic findings of a wide range of values - especially at larger population groups - is the same.

Table 10.8: (#tab:countyPopulationGroupStatsTexas)The mean, median, minimum, and maximum agency size for agencies in Texas for all population groups in the 2018 Offenses Known and Clearances by Arrests data.

Population Group	Mean Murder + Gun Assault	Median Murder + Gun Assault	Minimum Murder + Gun Assault	Max Murder + Gun Assault
City Under 2,500	0	0	0	6

Population Group	Mean Murder + Gun Assault	Median Murder + Gun Assault	Minimum Murder + Gun Assault	Max Murder + Gun Assault
City 2,500-9,999	2	1	0	17
City 10,000-24,999	7	5	0	38
City 25,000-49,999	17	13	0	61
City 50,000-99,999	37	27	1	141
City 100,000-249,999	105	77	12	467
City 250,000+	1,343	669	76	4,919
MSA Counties And MSA State Police	43	6	0	1,701
Non-MSA Counties And Non-MSA State Police	2	1	0	28

Now we'll look at data from Maine, as shown in Table @ref(tab:countyPopulationGroupStatsMaine). Here, results are much better: there is a narrow range in values meaning that the imputation would be very similar to the real values. This is driven mainly by Maine being a tiny state, with only one city larger than 50,000 people (Portland) and Maine being an extremely safe state so most places have zero murders+gun-assaults. In cases like this, where both crime and population size are consistent across the state (which is generally caused by everywhere having low crime), this imputation process can work well.

Table 10.9: (#tab:countyPopulationGroupStatsMaine)The mean, median, minimum, and maximum agency size for agencies in Maine for all population groups in the 2018 Offenses Known and Clearances by Arrests data.

Population Group	Mean Murder + Gun Assault	Median Murder + Gun Assault	Minimum Murder + Gun Assault	Max Murder + Gun Assault
City Under 2,500	0	0	0	0
City 2,500-9,999	0	0	0	2
City 10,000-24,999	1	1	0	7
City 25,000-49,999	6	4	0	14
City 50,000-99,999	12	12	12	12
City 100,000-249,999	-	-	-	-
City 250,000+	-	-	-	-

Population Group	Mean Murder + Gun Assault	Median Murder + Gun Assault	Minimum Murder + Gun Assault	Max Murder + Gun Assault
MSA Counties And MSA State Police	0	0	-1	2
Non-MSA Counties And Non-MSA State Police	2	0	0	19

10.4 Final thoughts

County-level UCR data should not be used for research because it handles missing data poorly and distributes crimes for agencies in multiple counties very poorly, so will give inaccurate results. This is the conclusion that Maltz and Targonski had in 2002 in their paper examining the data and is, in my opinion, the conclusion that still holds today as the problems have not improved (or even changed since 2002). There are two main problems: distributing crimes for agencies in multiple counties and missing data. The first, given UCR data constraints, seems unfixable though is not a problem for agencies in only one county. For the second, if an imputation method was used that properly (i.e. got as close to the right answer as possible) handled missing data that would solve this issue.

Given the decades of data and agencies that release their own crime data on their websites that can be used as a check on how close imputed data is to the real data (as some agencies, such as Philadelphia in 2019, don't report all months of data to the UCR but still release incident-level data publicly on their city's 'open data' websites) this is a solvable problem.[16] Missing data is not a problem only in criminology, and is actually addressed in some criminology papers, so the fact that so much research relies on this data that does such a poor job of handling missing data and literally assumes that crime is not at all concentrated is a sorry testament to this field.[17] We can do better.

[16] But not by me, so please don't ask.

[17] There are also a good number of criminology papers that try to interpolate missing data from survey papers, which in general I think is a mistake since they often assume that the data is missing at random.