National Incident-Based Reporting System (NIBRS) Data: A Practitioner's Guide

Jacob Kaplan

2021-05-16

Contents

1 Preface **5**

1.1 Goal of the book . . . 7

1.2 Structure of the book . . . 8

1.3 Citing this book . . . 8

1.4 Pronunciation . . . 8

1.5 Sources of NIBRS data . . . 9

1.5.1 Where to find the data used in this book . . . 10

1.6 Recommended reading . . . 10

2 Overview of the Data **12**

2.1 Choosing a unit of analysis . . . 12

3 Administrative- Window Exceptional Clearance Segment **14**

3.1 Important variables . . . 14

3.1.1 The incident date or report date . . . 14

3.1.2 Hour of incident . . . 14

3.1.3 Exceptional clearance . . . 16

3.1.4 Number of other segments . . . 16

4 Offense Segment **20**

4.1 Important variables . . . 20

4.1.1 Crime category . . . 20

4.1.2 Offense subtype . . . 20

4.1.3 Drug or alcohol use . . . 20

4.1.4 Crime location . . . 20

4.1.5 Weapons 20
4.1.6 Hate crime indicator (bias motivation) 20

5 Offender Segment **21**
5.1 Important variables 21
5.1.1 Age 22
5.1.2 Sex 22
5.1.3 Race 23

6 Victim Segment **25**
6.1 Important variables 25
6.1.1 Crime category 25
6.1.2 Victim type 25
6.1.3 Injury 25
6.1.4 Relationship to offender 25
6.1.5 Residence status 25
6.1.6 Age 26
6.1.7 Sex 26
6.1.8 Race 27
6.1.9 Ethnicity 27
6.1.10 Homicide type 28

7 Arrestee and Window Arrestee Segment **29**

8 Property and Window Property Segment **30**

9 Group B Arrest Reports Segment **31**

List of Figures

3.1 The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019. . . . 15

3.2 The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019. . . . 15

3.3 The percent of crimes that are reported each hour for all agencies reporting to NIBRS in 2019. . . . 16

3.4 The percent of crimes that are reported each hour for all agencies reporting to NIBRS in 2019. . . . 17

3.5 The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019. . . . 17

3.6 The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019. . . . 18

3.7 The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019. . . . 18

3.8 The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019. . . . 19

Chapter 1

Preface

Nearly a century ago the FBI started collecting data on crime that occurred in the United States as a way to better understand and respond to crime. This data, the Uniform Crime Reporting (UCR) Program Data, is a monthly count of the number of crime incidents (in cases where more than one crime happens per incident, only the most serious crime is included) in each police agency that reports data.[1] Other than for homicides, only the number of crimes that occurred is included. So we know, for example, the number of robberies in a city but nothing about who the victims or offenders were, when that month (day or time of day) the robberies occurred, or the type of location where they happened. To address these limitations the FBI started a new dataset in 1991, the National Incident-Based Reporting System data, which is known by its abbreviation NIBRS, and is the topic of this book.

NIBRS data provides detailed information on every crime reported to the police, including victim and offender demographics, whether the offender was arrested (and type of arrest or type of "exceptional clearance"), the crime date and hour, victim-offender relationship, and the crime location (as a location type, not the exact address). It also covers a far wider range of crimes than UCR data did. With the exception of UCR data on assaults against police officers, all NIBRS data can be converted back to UCR data, making it fully backwards compatible - and, therefore, comparable to UCR data. In many ways NIBRS data is a massive improvement to UCR data. We now have a far deeper understanding of crime and this has led to an explosion of research that allows a far more detailed analysis of crime and crime-policies than the blunt UCR data.

However, this is a major limitation to this data: most agencies don't use it. According to the FBI only about 8,500 police agencies, covering about 45% of the US population, reported NIBRS data in 2019 (the latest year currently available), fewer than half of the about 18,000

[1]This data has been expanded since it began in 1929 to include information on arrests, hate crimes, and stolen property. For more on this, please see my book Uniform Crime Reporting (UCR) Program Data: A Practitioner's Guide at ucrbook.com.

police agenices in the United States. This is an even larger problem that it seems are the agencies that do report - especially in earlier years of the data - are disproportionately small and rural. So we're missing out of data from major cities. A number of states don't have any agencies reporting, making this data relatively biased at least in terms of geography and city size. **Even so, the FBI has said that they are moving entirely to NIBRS data starting in 2021, and will no longer even collect UCR data.** While NIBRS can be converted to UCR data, meaning we can have consistent statistics over time, for agencies that don't report to NIBRS, we have no information on their crimes. In effect, unless the majority of agencies suddenly switch to NIBRS - which, given the high level of detail relative to UCR data is a costly and timely switch - we will be flying blind for most crime in the country.

So there are really three major problems with NIBRS data, both related to the lack of reporting. First, we are potentially looking at a massive loss of data when UCR data ends in 2020 - it takes over a year for data to be released so even though I'm writing this is Spring 2021, 2019 UCR and NIBRS data are the latest years available. Especially given the huge crime changes during 2020 - and whose violent crime increases and continuing into 2021 - losing standardized crime data for most cities is a very bad thing. The second problem is that even if suddenly all agencies do start reporting in 2021, we'd only have a single year of data available. Even for agencies that already report, we generally don't have too many years of data for them. This really limits the kind of research since we can do since it's hard to know if a finding is based on a trend or is just a weird outlier without having many years of data available. This means that for the next several years at least we'll be mostly using NIBRS data as UCR-like datasets, aggregated to the month- or year-level so we can compare it with UCR data from the past. Luckily, this problem will be alleviated the longer we wait as more years of data will become available.

The final issue is that this data is massive. A single year of 2019 data, with <50% of agencies, and few large agencies, reporting has about 6.5 million crime incidents recorded. Since each crime incident can have multiple victims, offenders, and crimes, there are more rows for these datasets. Once all agencies report - though it's doubtful that'll ever occur - we're looking at tens of millions of rows per year. And even now if we wanted to look at a decade of data we're going to be dealing with over 50 million rows of data. So this data requires both good hardware - a strong laptop or server is necessary - and good programming skills, which most academics sorely lack.

While people generally refer to NIBRS just as "NIBRS data" it is actually a collection multiple different datasets all corresponding to a single crime incident. For example, if you care about victim info you'll look in the victim file called the "Victim Segment" (each of the datasets are called "Segments" since they are part of the whole picture of the crime incident) and likely will merge it with other data, such as when are where the crime occurred which

is in the "Offense Segment". In most cases you'll merge together multiple datasets from the NIBRS collection to be able to answer the question that you have. This means that you'll need to understand how to deal with multiple datasets, and subset and merge them as needed.

Relative to the FBI's UCR data there are far fewer "weird things" in NIBRS data. Still, we'll cover instances of the "weirdness" in the data, such as the why crime always goes up on the 1st of the month, or why there are more crimes at noon than at nearly all other hours of the day. We'll also be discussing how much of the detailed information that should be available in the data is missing, and when that affects which questions we can answer.

A word of caution. To date fewer than half of agencies report NIBRS data. As they rush to comply with the FBI's order that they only will accept NIBRS data, there will likely be more mistakes made and erroneous data included in NIBRS data later than is covered in this book, which ends with 2019 data as 2019 is the most recent year available. So while I always urge caution when using any data - caution that should be accompanied by a thorough examination of your data before using it - NIBRS data from 2020 and beyond merits extra attention.

1.1 Goal of the book

By the end of the book you should have a firm grasp on NIBRS data and how to use it (or as is often the case, choose not 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. As the FBI has moved to only use NIBRS data starting in 2021, I expect the uses of this data - and thus the number of emails I get - to grow very quickly. This is an increasingly popular dataset used by criminologists (and by other fields studying crime) and yet there are still occasions where papers are using the data incorrectly.[2] 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. "NIBRS

[2]Though given that the data is fairly complicated and requires good programming knowledge, the bar is higher to use it. So there are far fewer bad uses of this data than there is for UCR data.

data began in 1991". 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 nine chapters: this chapter, an intro chapter briefly summarizing each segment files and going over overall issues with NIBRS data, and seven chapters each covering one of the seven UCR datasets. Each chapter will follow the same format: we'll start with a brief summary of the data such as its possible uses and pitfalls. And then, 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). *National Incident-Based Reporting System (NIBRS) Data: A Practitioner's Guide*. https://nibrsbook.com/.

BibTeX format:

```
@Manual{ucrbook,
  title = {National Incident-Based Reporting System (NIBRS) Data: A Practitioner's Guide
  author = {{Jacob Kaplan}},
  year = {2021},
  url = {https://nibrsbook.com/},
}
```

1.4 Pronunciation

This data is usually just called NIBRS, and generally there's no distinction between segment files since they work in unison as they are pieces of the overall criminal incident. "NIBRS" is generally pronounced as "NIE-BERS". It rhymes with "HIGH-BERS". I've also heard it pronounced - usually by non-academics - using a soft i like in "timber" so it sounds like "nih-bers". I prefer the "NIE-BERS" saying but it really doesn't make a difference.

1.5 Sources of NIBRS 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 - NIBRS 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. They've also put together what they call "Extract Files" which are where they merged some of the NIBRS segments together, saving you the effort of doing so yourself. These extract files essentially take every potential unit of analysis - incident, victim, offender, and arrestee (some crimes have no victims, only arrestees) - and merge it with the segment which has info about the incident such as the time of day or the outcome, and information about the reporting agency. This source only has data through 2016 which means that the most recent years (NIBRS 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 (NIBRS data is at the day-level so this is very aggregated data) for certain crimes (and not even all the crimes actually available in the data). This site has only a small subset of the available data and is already aggregated so you're dealing with a subset of data in a unit of analysis that you may not want. For example, this site lets you see the annual age of offenders for certain crimes in age brackets such as aged 20-29. As the data provides the exact age (in years) of each offender, this is much less useful than the full data. The crimes on this site are also limited to only the eight "Index Crimes" (Murder, rape, robbery, aggravated assault, arson, burglary, theft, and motor vehicle theft) so are only a tiny share of the crimes actually reported in NIBRS data. For more on what Index Crimes are, please see here. This data source is potentially 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 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 into R and saved the files in R and Stata files for easy use. The main advantage is that all my data has standard variable names and column names and can be read into modern programming languages R and Stata (this is also true of recent NACJD years, but early years come as fixed-width ASCII files). 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 so people using languages such as SAS or SPSS cannot use this data.[3]

1.5.1 Where to find the data used in this book

The data I am using in this book is the cleaned and concatenated data that I put together from the raw data that the FBI releases. The raw data that the FBI releases is available here. The data that I have released is available on the data hosting site openICPSR 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 openICPSR to download the data; it's not available through this book's GitHub repo.

1.6 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 "2019 National Incident-Based Reporting System 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 takes down the PDF.[4] 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 NIBRS data.

[3] I am not sure if SAS or SPSS can read in R or Stata data files.

[4] 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.

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

The average American had in 2019 about a 1 in 20,500 chance of being murder, 1 in 1,223 chance of being robbed, and a 1 in 64 chance of having something they own stolen. Getting these numbers is extremely simple. We take the number of crimes reported to the police and divide it by the number of people living in the United States that year. For example, there were about 16,000 murders in 2019 and 328 million people in the county - 16,000 / 328 million = ~1/20,500. You'll more commonly see this - in news articles, in political speeches, in research articles, on TV, etc. - reported as the rate per 100,000 people but that's just a matter of conversion, the numbers are the same. This is, however, totally wrong. It assumes - and let's now just pretend that there's no underreporting of crimes to the police - that every single person has the exact same risk of victimization. We know this is wrong intuitively. There are the "bad parts of town" or people who "run with the wrong crowd". Research in criminology backs this up by finding that crime is generally concentrated among a small group of people and within a small geographic area (usually a small number of streets or neighborhoods in a city). From surveys that ask if people have been victims of a crime we also know that victimization rates differ by age, race, gender, income, and city type. Indeed, think of a personal characteristic (e.g. risk tolerance, athleticism, frequency outdoors) and there will probably be large differences in the likelihood of being a victim within these groups.

So why do people so frequently talk about crime as rates per total population, why assume that everyone has equal risk of being a victim?

2.1 Choosing a unit of analysis

Given the detail of this data

Consider, for example, an incident where four men rape a single woman. If you're interested in measuring robbery you could do so in several different ways, each of which addresses a

different part of crime measurement and will lead to different answers to your questions: the number of crime incidents, the number of victims, the number of offenders, and the number of crimes. . First, we can follow the old UCR measure of incident-level and say that this is one rape since only one crime incident occurred (even though there were multiple offenders). Second, we could look at the victim-level, which again is one rape as there was only one victim. Or at the offender-level, which now has four rapes since each offender would be responsible the robbery. Finally we could look at the offense-level.

Chapter 3

Administrative- Window Exceptional Clearance Segment

The Administrative Segment provides information about the incident itself, such as how many victims or offenders there were. In practice this means that it tells us how many other segments - offense, victim, offender, and arrestee segments - there are for this particular incident. It also has several important variables at the incident-level such as what hour of the day the incident occurred and whether the incident date variable is actually just the date the incident was reported. Finally, it tells us whether the case was cleared exceptionally and, if so, what type of exceptional clearance it was. This can tell us, for example, how many crimes was cleared because the offender died or the victim refused to cooperate. As the UCR data doesn't differentiate between normal clearances (i.e. arrest the offender) and exceptional clearances, this provides a far deeper understanding of case outcomes.

3.1 Important variables

3.1.1 The incident date or report date

3.1.2 Hour of incident

An extremely important aspect of crime data is when exactly the crime occurs. If, for example, crime always spikes when the local high school ends their day that would likely indicate that high school students are involved with crime (both as victims-offenders). In my own research on daylight saving time-crime I only care about the sunset hours, which is when daylight saving time would affect outdoor lighting. When crime happens also would

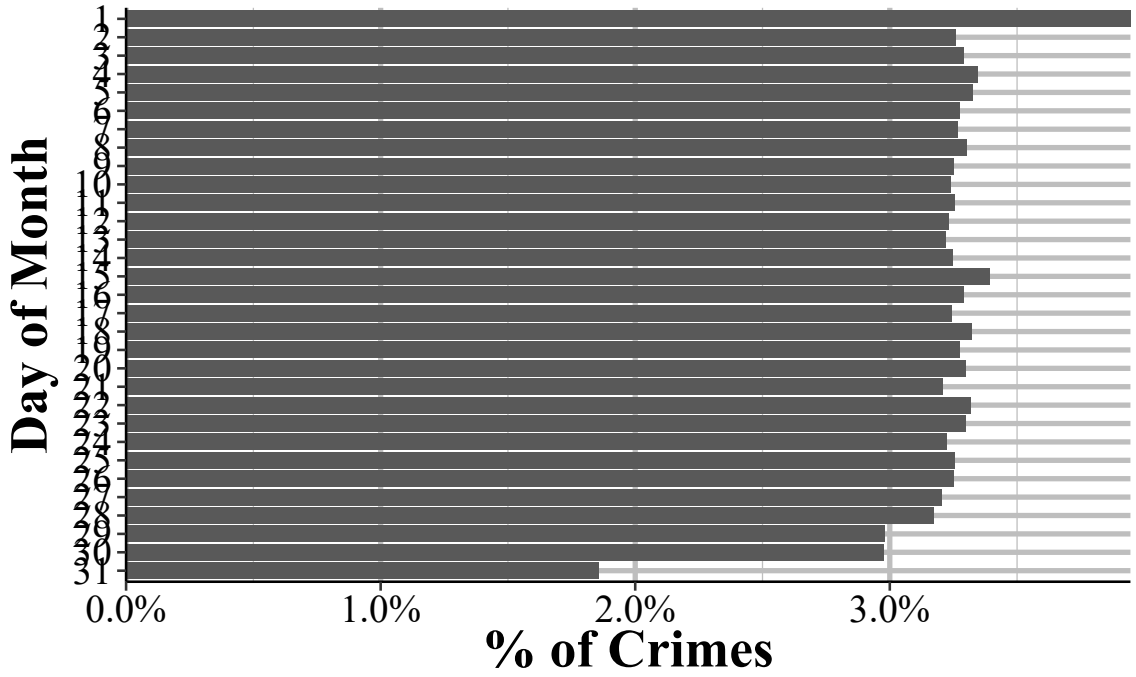


Figure 3.1: The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019.

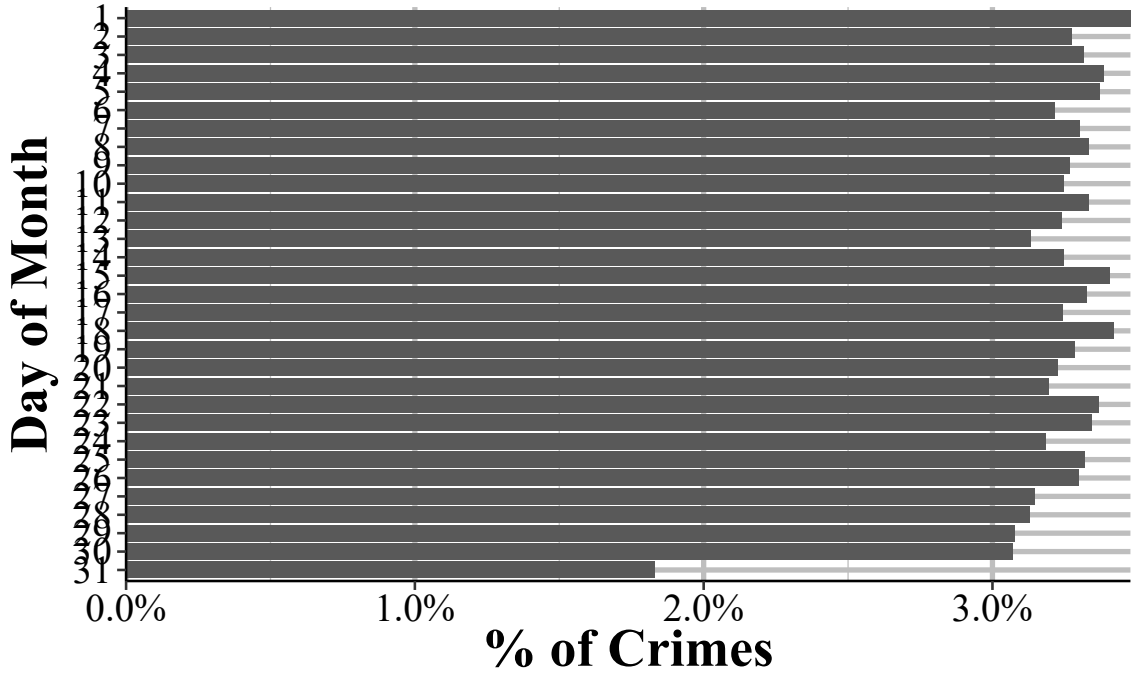


Figure 3.2: The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019.

affect police behavior as they'd likely increase patrol during times of elevated crime. Luckily NIBRS data does have the time of each incident, though it's only at the hour level.

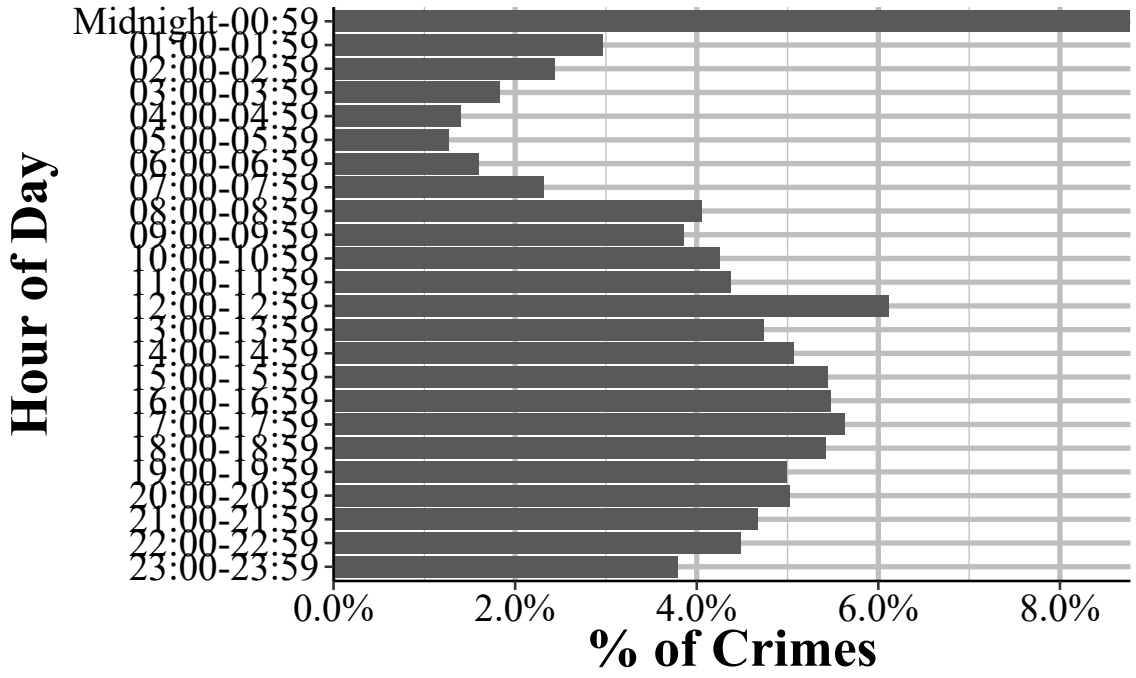


Figure 3.3: The percent of crimes that are reported each hour for all agencies reporting to NIBRS in 2019.

3.1.3 Exceptional clearance

3.1.4 Number of other segments

3.1.4.1 Offense segments

3.1.4.2 Victim segments

3.1.4.3 Offender segments

3.1.4.4 Arrestee segments

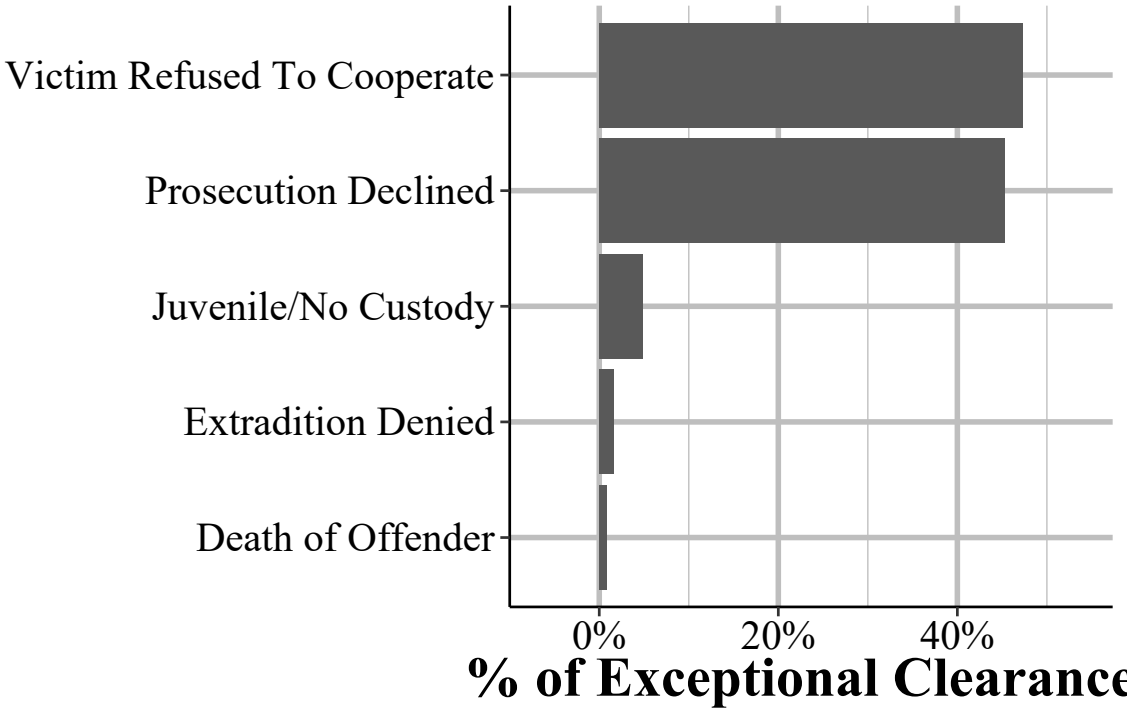


Figure 3.4: The percent of crimes that are reported each hour for all agencies reporting to NIBRS in 2019.

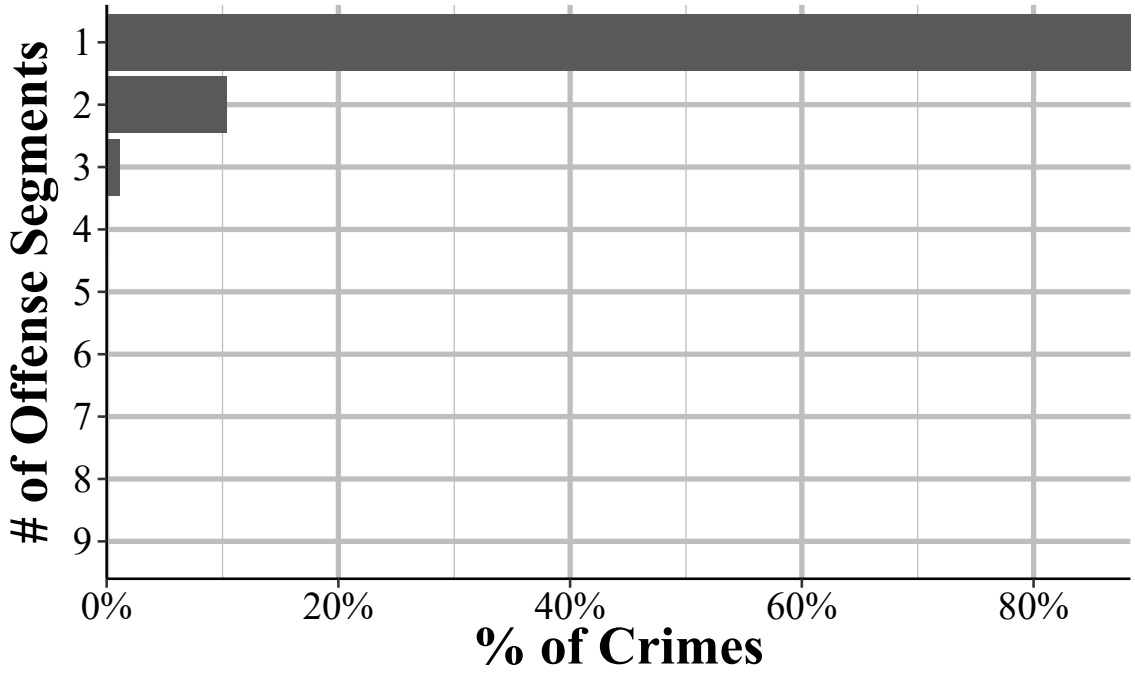


Figure 3.5: The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019.

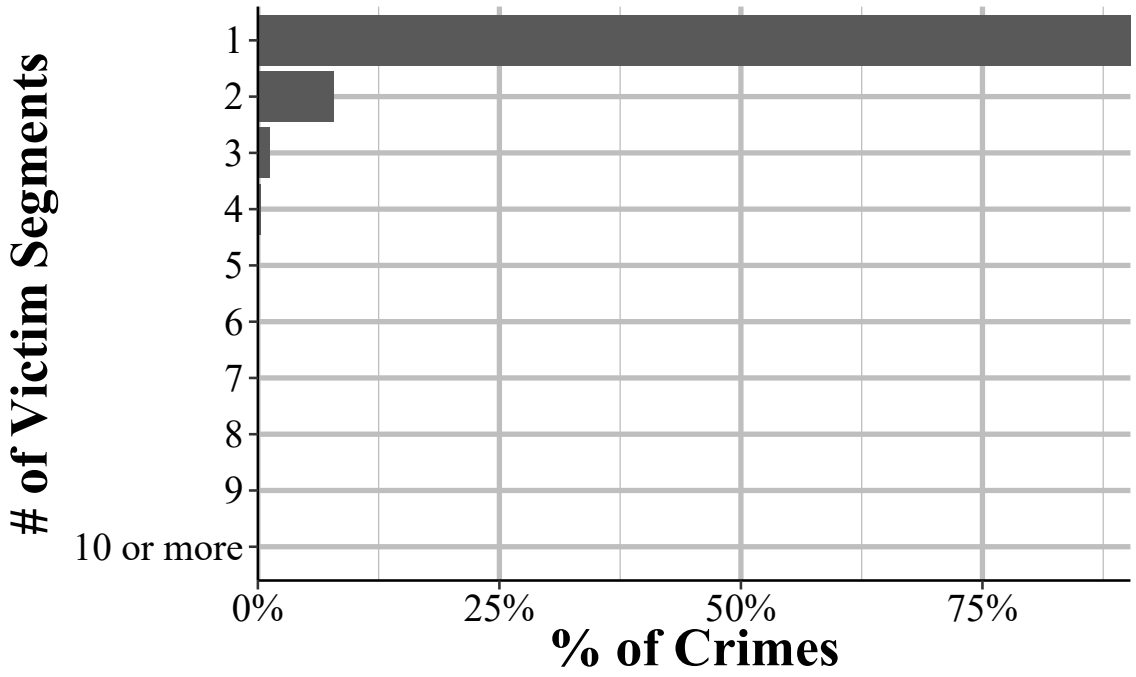


Figure 3.6: The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019.

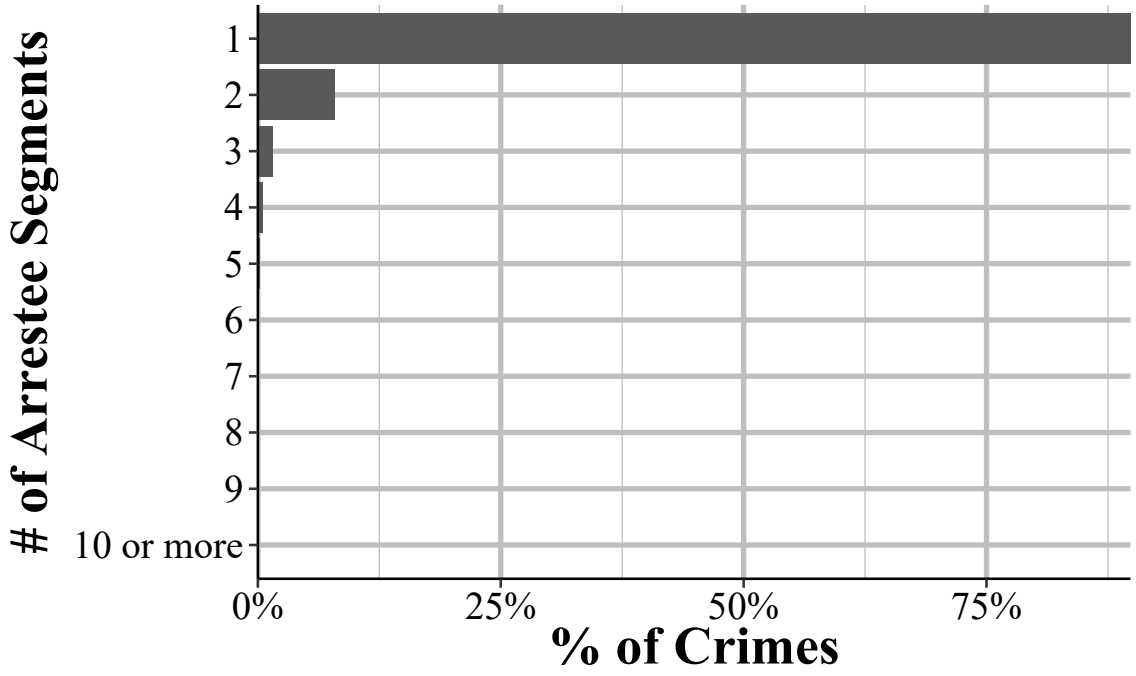


Figure 3.7: The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019.

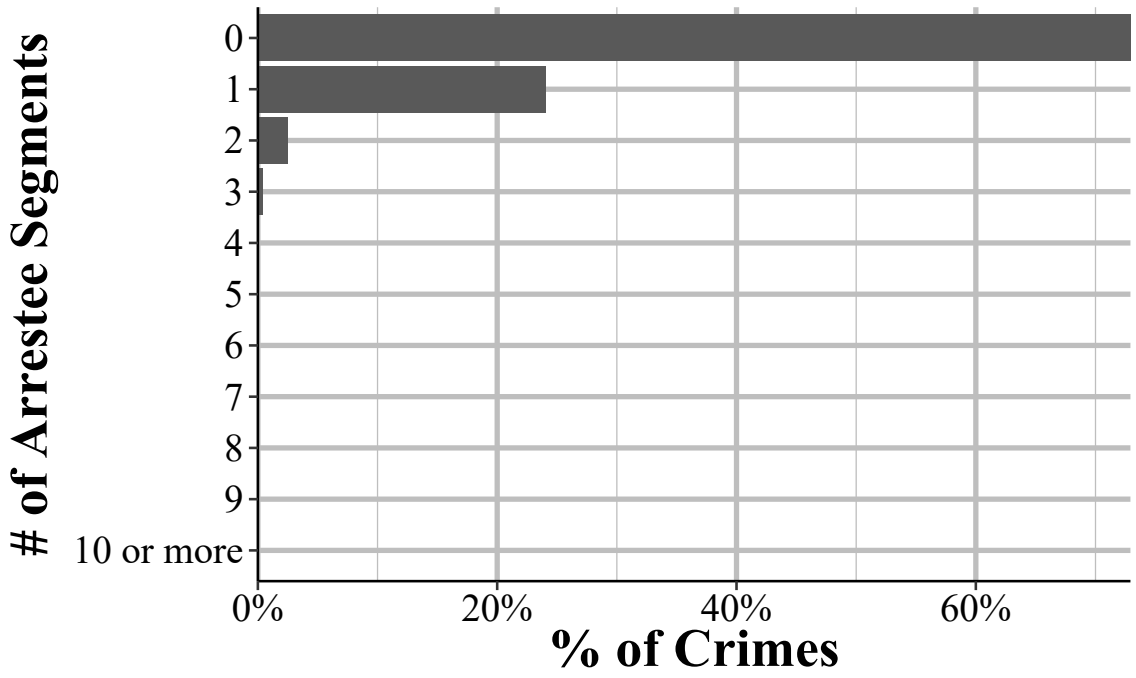


Figure 3.8: The percent of crimes that are reported each day of the month for all agencies reporting to NIBRS in 2019.

Chapter 4

Offense Segment

```
load("data/nibrs_offense_segment_segment_2019.rda")
offense <- nibrs_offense_segment_segment_2019
rm(nibrs_offense_segment_segment_2019)
gc()
#>             used   (Mb) gc trigger   (Mb)  max used   (Mb)
#> Ncells   5081061  271.4    9818648  524.4   5098871  272.4
#> Vcells 192420688 1468.1  268965598 2052.1 192464051 1468.4
```

4.1 Important variables

4.1.1 Crime category

4.1.2 Offense subtype

4.1.3 Drug or alcohol use

4.1.4 Crime location

4.1.5 Weapons

4.1.6 Hate crime indicator (bias motivation)

Chapter 5

Offender Segment

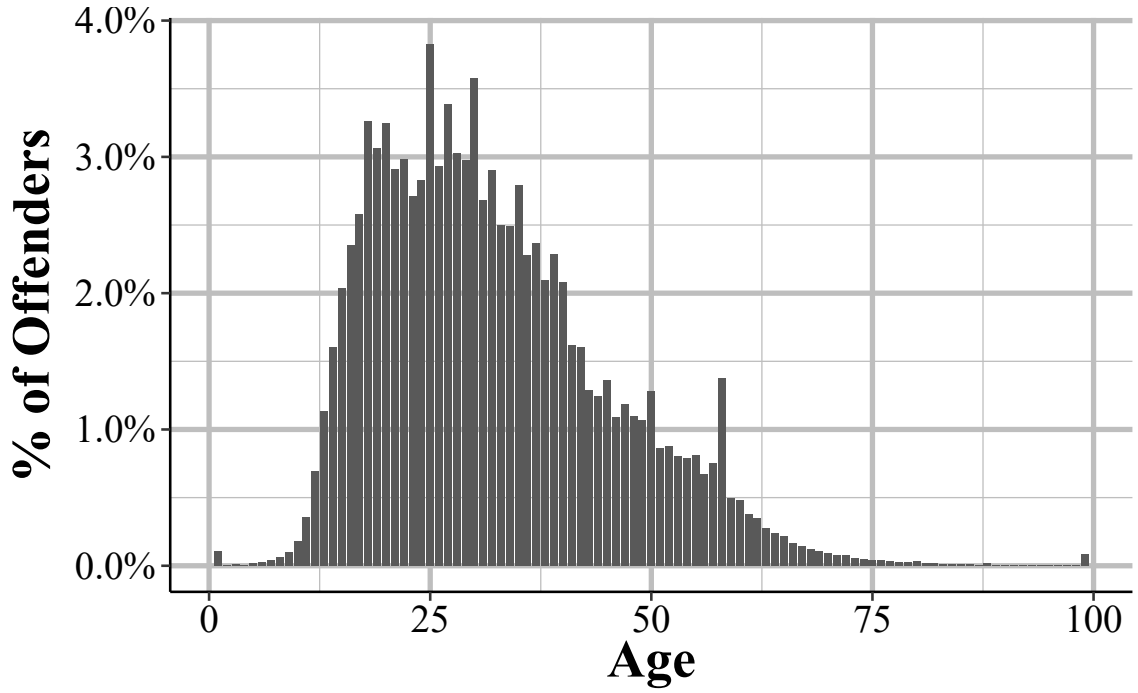
```
load("data/nibrs_offender_segment_segment_2019.rda")
offender <- nibrs_offender_segment_segment_2019
rm(nibrs_offender_segment_segment_2019)
gc()
#>            used  (Mb) gc trigger  (Mb) max used  (Mb)
#> Ncells  5081047 271.4    9819981 524.5  5098857 272.4
#> Vcells 88748472 677.1  108845140 830.5 88791835 677.5
```

As might be expected, the Offender Segment provides information about who the offender is for each incident, though this is limited to only demographic variables. So we know the age, sex, and race of each offender but nothing else. This means that important variables such as criminal history, ethnicity, socioeconomic status, and motive are missing. In the Victim Segment we learn about the relationship between the victim and offender, and in the Offense Segment we learn which weapon (if any) the offender used. So there is some other data on the offender in other segments but its quite limited. This data has one row per offender so incidents with multiple offenders have multiple rows. In cases where there is no information about the offender there will be a single row where all of the offender variables will be "unknown". In these cases having a single row for the offender is merely a placeholder and doesn't necessarily mean that there was only one offender for that incident.

5.1 Important variables

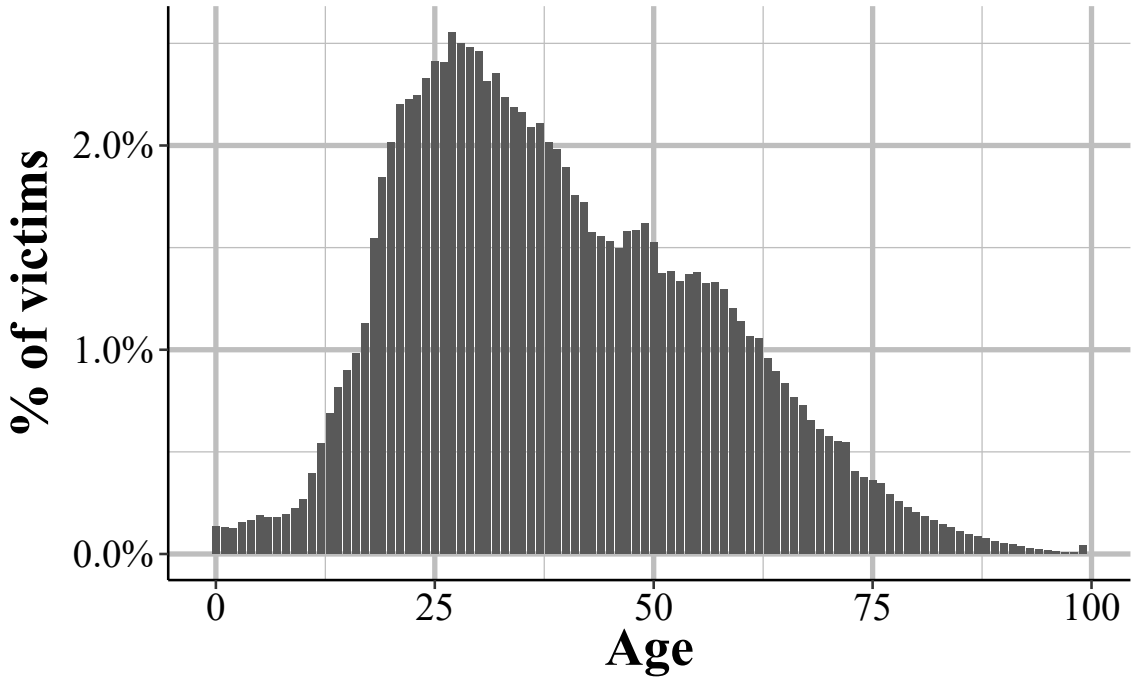
5.1.1 Age

```
offender$age_of_offender[offender$age_of_offender %in% "over 98 years old"] <- 99
offender$age_of_offender[offender$age_of_offender %in% "unknown"] <- NA
offender$age_of_offender <- as.numeric(offender$age_of_offender)
make_stat_count_plots(offender, "age_of_offender", count = FALSE, ylab = "% of Offenders
#> Warning: Removed 2943237 rows containing non-finite values (stat_count).
```



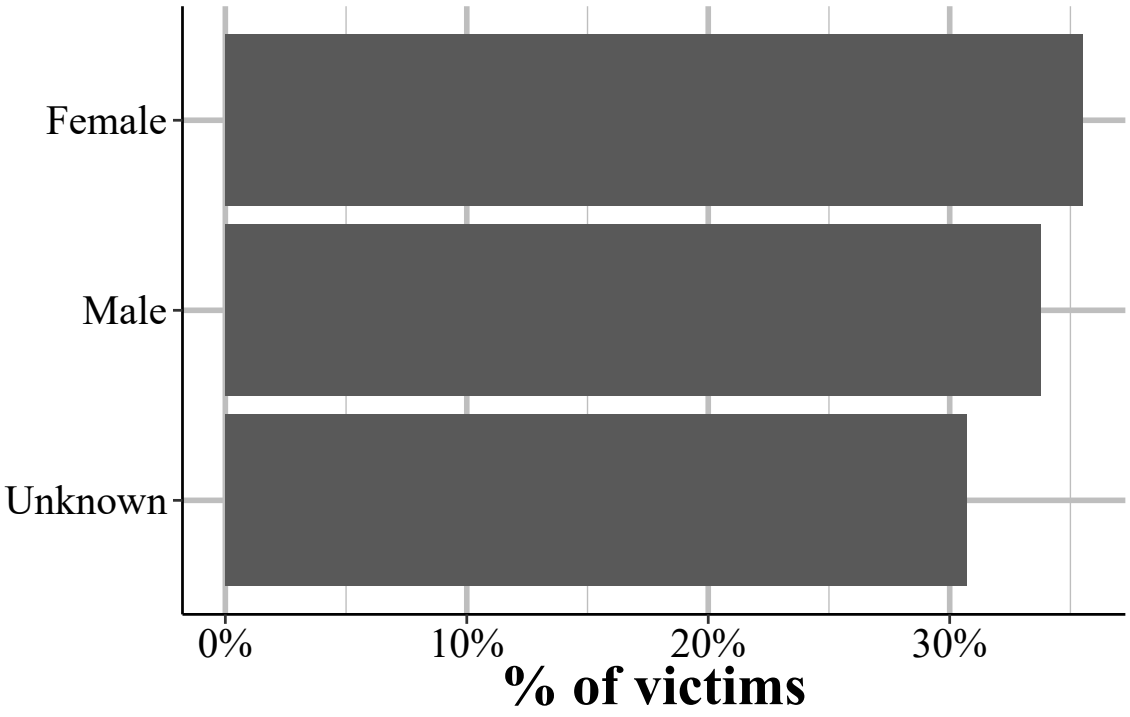
5.1.2 Sex

```
offender$sex_of_offender[offender$sex_of_offender == "m"] <- "male"
offender$sex_of_offender[is.na(offender$sex_of_offender)] <- "unknown"
offender %>%
  mutate(sex_of_offender = capitalize_words(sex_of_offender)) %>%
crimeutils::make_barplots("sex_of_offender", count = FALSE, ylab = "% of Offenders") +
  ggplot2::scale_y_continuous(labels = scales::percent)
#> Scale for 'y' is already present. Adding another scale for 'y', which will
#> replace the existing scale.
```



5.1.3 Race

```
offender$race_of_offender[is.na(offender$race_of_offender)] <- "unknown"
offender %>%
  mutate(race_of_offender = capitalize_words(race_of_offender)) %>%
crimeutils::make_barplots("race_of_offender", count = FALSE, ylab = "% of Offenders") +
  ggplot2::scale_y_continuous(labels = scales::percent)
#> Scale for 'y' is already present. Adding another scale for 'y', which will
#> replace the existing scale.
```


Female
Male
Unknown
0%
10%
20%
30%
% of victims

Chapter 6

Victim Segment

6.1 Important variables

6.1.1 Crime category

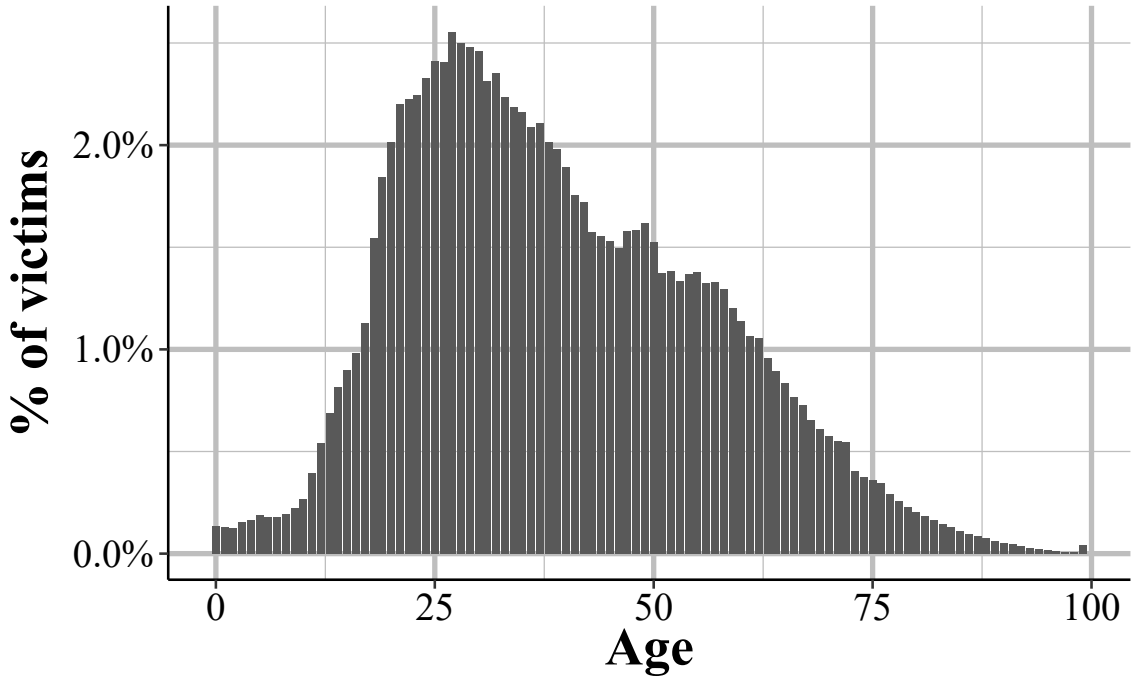
6.1.2 Victim type

6.1.3 Injury

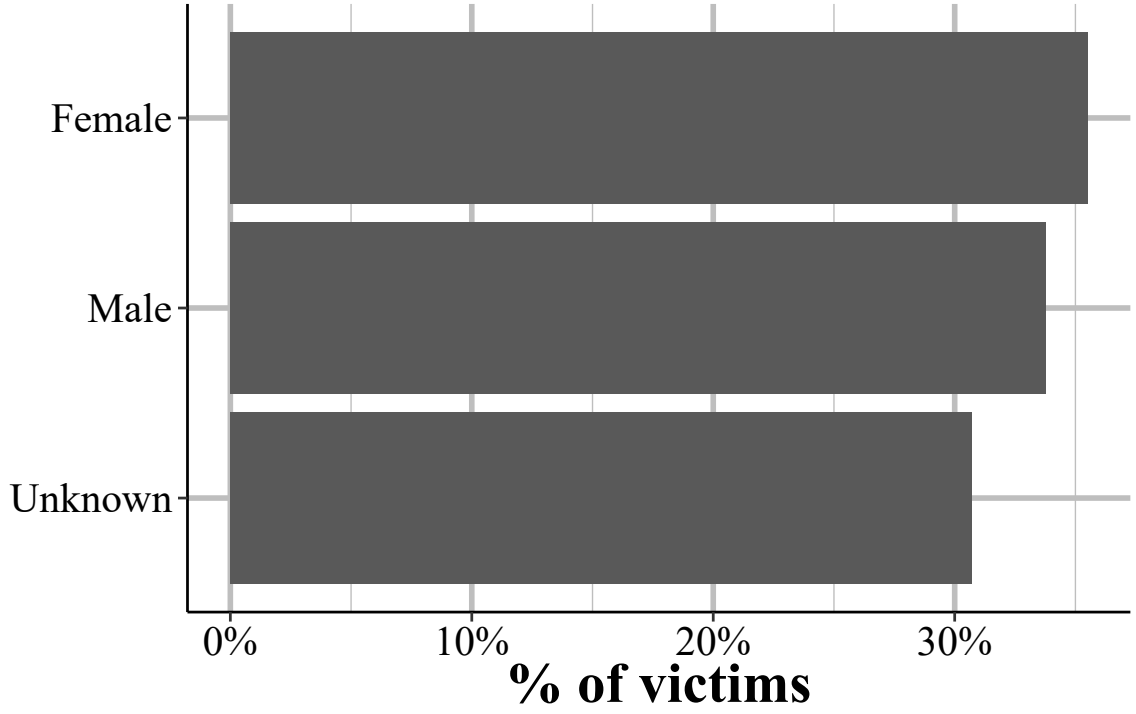
6.1.4 Relationship to offender

6.1.5 Residence status

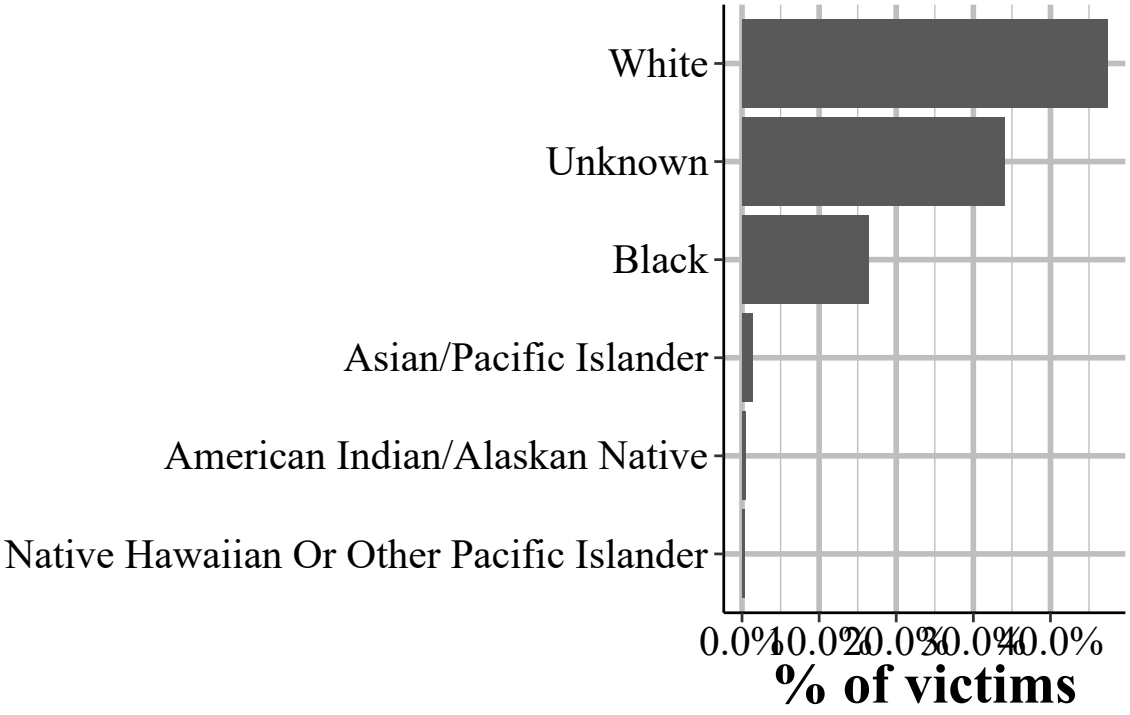
6.1.6 Age



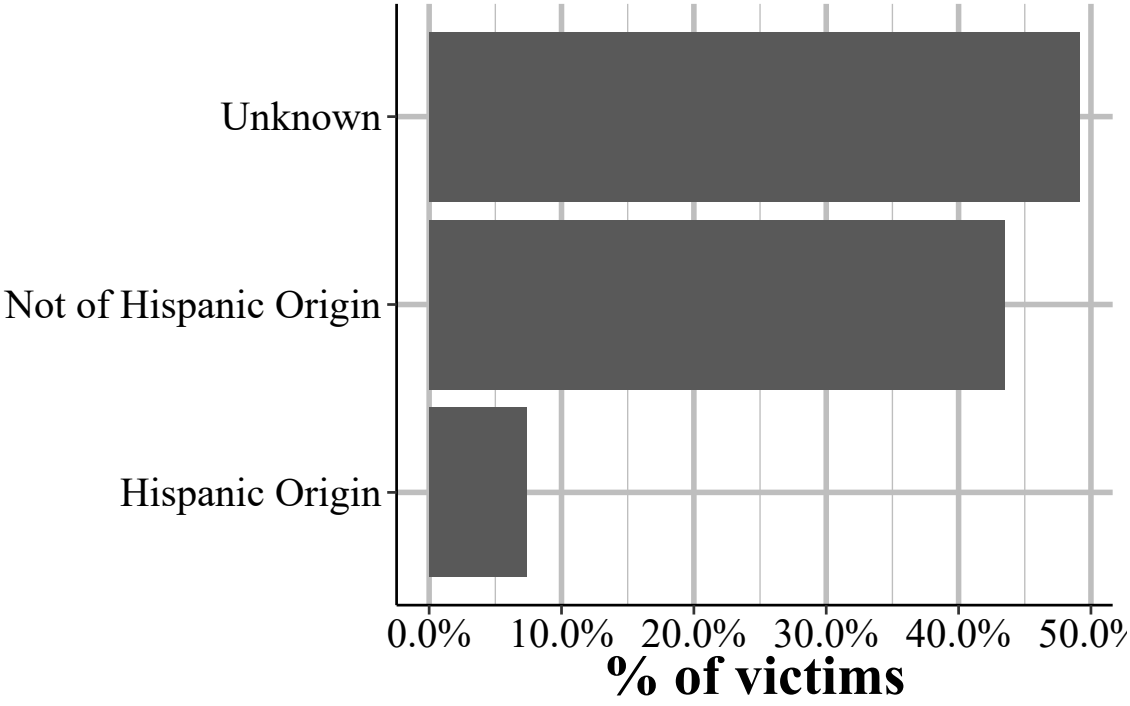
6.1.7 Sex



6.1.8 Race



6.1.9 Ethnicity



6.1.10 Homicide type

Chapter 7

Arrestee and Window Arrestee Segment

Chapter 8

Property and Window Property Segment

Chapter 9

Group B Arrest Reports Segment