Differences in Homicide Case Characteristics Indicates Why Justice is Not Served*

An analysis of solved and unsolved homicides from 2010 to 2017 in one of the United States's 2 largest cities, Chicago and Los Angeles

Emily Su

December 2, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/moonsdust/unsolved-murders.

1 Introduction

Overview paragraph

This led to us investigate the following question in our paper: what are the differences in homicide case information like the year and city the homicide took place and the victims' perceived characteristics (age, sex, and race) between solved and unsolved homicides in one of the 2 of the largest cities in the United States, Chicago and Los Angeles, from 2010 to 2017?

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

2 Data

2.1 Overview

The dataset we used for this paper comes from The Washington's Post's GitHub repository, "How The Post mapped unsolved murders", which is also known as the "Unsolved Homicide Database" (The Washington Post 2018b). We used the statistical programming language R (R Core Team 2024), tidyverse (Wickham et al. 2019), janitor (Firke 2023), lubridate (Grolemund and Wickham 2011), dplyr (Wickham et al. 2023), ggplot2 (Wickham 2016), arrow (Richardson et al. 2024), testthat (Wickham 2011), and knitr (Xie 2024) to retrieve, clean, test, and analyze the dataset. To construct, test, and analyze our model, we used the following packages: rstanarm (Goodrich et al. 2024), modelsummary (Arel-Bundock 2022), and car (Fox and Weisberg 2019). The causal model diagram created to understand the relationship between predictor variables, the outcome variable, and a confounder used DiagrammeR (Iannone and Roy 2024), rsvg (Ooms 2024), DiagrammeRsvg (Iannone 2016) and png (Urbanek 2022). To style our scripts, we used lintr (Hester et al. 2024) and styler (Müller and Walthert 2024).

The Washington Post's Unsolved Homicide Database is a dataset compiled by Washington Post reporters that contains over 52000 homicides in the United States (US) from 50 of the largest US cities from 2007 to 2017 (The Washington Post 2018b). This dataset includes information about the victim such as their name, sex, race, and age as well as geographic and temporal information of the homicide. The Washington Post was interested in using the information compiled to map unsolved homicides across the United States in major cities from 2007 to 2017 (The Washington Post 2018a). Another dataset we had considered using was another one compiled by The Washington Post on school shootings across the US and approaching our problem with a different perspective on the characteristics of the perpetrator. However, due to there only being 416 observations and numerous observations missing information, we forgo using the dataset.

The raw dataset we retrieved from The Washington Post using a script that downloads the CSV file from their GitHub repository contains 52179 observations with one observation being a homicide case. However, since we narrowed our scope to focus on the two of the most populated US cities, Los Angeles and Chicago, the number of observations in our cleaned dataset ended up being 6307. Our data look as follows:

Table 1: Preview of dataset on solved and unsolved homicides (2010 to 2017) with the original dataset compiled by The Washington Post

victim_race	victim_age	victim_sex	city	disposition	year	month	arrest_was_not_made
Black	61	Female	Chicago	Closed by arrest	2010	1	0
Hispanic	27	Male	Chicago	Open/No arrest	2010	1	1
Black	49	Male	Chicago	Open/No arrest	2010	1	1
Black	21	Male	Chicago	Closed by arrest	2010	1	0
Hispanic	17	Male	Chicago	Closed by arrest	2010	1	0
Hispanic	20	Male	Chicago	Open/No arrest	2010	1	1

Table 2: Number of observations, minimum, maximum, median, mean, 1st and 3rd quartile of variables in dataset on solved and unsolved homicides (2010 to 2017) excluding victim_sex, city, and disposition

	victim_race	victim_age	year	month	arrest_was_not_made
	White : 376	Min. : 1.00	Min. :2010	Min. : 1.000	Min. :0.000
	Asian : 43	1st Qu.:21.00	1st Qu.:2012	1st Qu.: 4.000	1st Qu.:0.000
	Black :4063	Median :27.00	Median :2014	Median : 7.000	Median :1.000
	Hispanic:1763	Mean :30.31	Mean :2014	Mean : 6.688	Mean :0.674
	Other : 62	3rd Qu.:37.00	3rd Qu.:2016	3rd Qu.: 9.000	3rd Qu.:1.000
	NA	Max. :94.00	Max. :2017	Max. :12.000	Max. :1.000

Table 1 and Table 2 indicates our variables of interest, which are the following: victim_race, victim_age, victim_sex, city, disposition, year, month, and arrest_was_not_made. victim_race represents the race of the homicide victim, which can be "White" (376 observations), "Hispanic" (1763 observations), "Black" (4063 observations), "Asian" (43 observations), and "Other" (62 observations). victim_age signifies the age of the homicide victim at the time of their death and is defined as a whole number. Table 2 reveals that the mean victim_sex is the sex of the homicide victim where they are either identified as a "female" or "male". The city variable defines the city the victim is reported to have been found in. The disposition variable is the specific status of a homicide case where a case can fall in either of the following three status: "Closed by arrest", "Open/No arrest", and "Closed without arrest". The year variable represents a year from 2010 to 2017 that indicates the year the homicide took place. Following this, the month variable represents the month a homicide took place. arrest_was_not_made is a variable that was constructed based on the disposition variable with "Closed by arrest" being converted to a 0 and "Open/No arrest" and "Closed without arrest" being converted to a 1. arrest_was_not_made indicates the status of a homicide case as either being unsolved, which is denoted by a 1, and solved, which is denoted by a 0.

However, our dataset has limitations. There was only data available from 2010 onwards for Los Angeles provided by The Washington Post and so that limited the number of homicides we could look at for both Chicago and Los Angeles. Also, not all victims were able to be identified in some homicide cases and unknown attributes of the victim such as their age, sex, and gender were indicated with the text "Unknown" in the dataset. However, this causes issues with data type compatibility such as the value "Unknown" being a character type being under the victim_age column, which has a data type of integer. This would lead to issues with our model providing accurate estimates. We decided to remove cases during our data cleaning where victim's demographic information is missing at least one of the three columns, victim_age, victim_sex, victim_race.

2.2 Measurement

The Federal Bureau of Investigation (FBI) has a program called the Uniform Crime Reporting (UCR) Program to generate statistics for the public (Federal Bureau of Investigation 2024). The FBI originally had a system under the UCR called Summary Reporting System (SRS), which obtained details about different crimes taking place from law enforcement agencies nationwide such as victim information (Federal Bureau of Investigation 2024). However, the SRS was replaced with a new system called the National Incident-Based Reporting System (NIBR) in 2021 that obtained more details about various crimes (Federal Bureau of Investigation 2024). For The Los Angeles Police Department, after a homicide case occurs, the investigating team handwrites information into physical crime reports with details like the type of crime, the premise the crime occured at, and the age and ethnicity of the victim (City of Los Angeles 2024). The crime reports are then transcribed into a digital format and then sent to the SRS monthly (City of Los Angeles 2024). After a homicide occurs in Chicago, The Chicago Police Department uses a system called the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system to report on details such as the victim, if an arrest was made or not, and the location the crime took (City of Chicago 2024). This information is then reported to the FBI under the UCR program.

Reporters from the Washington Post then obtained the data from the FBI specifically on homicides from 50 large US cities from 2007 to 2010, which can be access through the UCR publications page on the FBI website (The Washington Post 2018b). They also obtain data about homicide counts and closure rates through papers and compare these values with the ones from the FBI dataset for accuracy (The Washington Post 2018b). The Washington Post would also use public records like medical examiner reports, death certificates, and court records, to fill in any missing information since some departments only report partial information to the FBI (The Washington Post 2018b). The Washington Post defined cases to be closed without arrest when they are reported by the police as "exceptionally cleared" where there is evidence of who the perpetrator is but an arrest is not possible because of reason like they has died (The Washington Post 2018b). They also define cases to be closed by arrest if the police reported it to be and other cases are defined to be open/no arrest (The Washington Post 2018b).

2.3 Outcome variable

The outcome variable we are interested in looking at with our model and our analysis is the arrest_was_not_made variable. We use this variable to compare homicide case characteristics of solved and unsolved homicides.

2.4 Predictor variables

The predictor variables for our model are the following: victim_race, victim_age, victim_sex, city, and year. The variables, disposition and month are not predictor variables in our model but they are used to investigate trends between homicide case information and the status of the homicide being solved or unsolved.

3 Model

The model we implemented was a Bayesian logistic regression model. This model was constructed after we saw patterns between homicide case information and a homicide case going unsolved in our analysis. We are interested in seeing if certain characteristics of a homicide case such as the victim's perceived characteristics (sex, gender, age) and the year and city the victim is found impacts the likelihood of their case going unsolved.

3.1 Model set-up

With our model, we will make the assumption that there is a relationship between homicide case information like the victim's race, victim's age, victim's sex, the city, and the year with a homicide case being unsolved. We also assume that the predictor variables are independent from one another, which we check in Section A.6 using variance inflation factor (VIF) and it indicates the predictors are not highly correlated with each other. We define our model as follows:

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) \\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \times \text{victim\_race}_i + \beta_2 \times \text{victim\_age}_i + \beta_3 \times \text{victim\_sex}_i + \beta_4 \times \text{city}_i + \beta_5 \times \text{year}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5) \\
\beta_4 &\sim \text{Normal}(0, 2.5) \\
\beta_5 &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

We define y_i to be the status of the homicide case where 1 means the homicide is unsolved (case is still open / been closed without arrest) and 0 means the homicide is solved (case has been closed with arrest). π_i represents the probability of the homicide being solved. $\text{logit}(\pi_i)$ indicates the log-odds of the homicide being unsolved. Now looking at the coefficients b_i and predictor variables, β_0 is the intercept of our model and is the log-odds when all predictor variables are equal to 0. victim_race_i signifies the race of the victim, which could be either "Asian", "Black", "Hispanic", "White", and "Other". In our model, we use "White" as the baseline for victim_race_i to see if being part of a minority impacts if the case goes unsolved or not. β_1 is the coefficient that represents the log-odds when victim_race_i changes. victim_age_i is the victim's age, which is a whole number. β_2 is the log-odds when victim_age_i increases by 1 year. victim_sex_i represents the sex of the victim, where in the dataset it is either "female" or "male" and β_3 is the log-odds when victim_sex_i changes. city_i is the city the victim was reported to be killed in, which from our dataset could be either "Chicago" or "Los Angeles". β_4 is the coefficient that stands for the log-odds when city_i changes. We define year_i to be the year the homicide was reported to have occured from 2010 to 2017. β_5 indicates the log-odds when year_i increases by 1 year.

Our model runs in R (R Core Team 2024) using the rstanarm package (Goodrich et al. 2024). For our model's priors, we use the default priors provided by rstanarm (Goodrich et al. 2024). Diagnostics for the model such as in posterior predictive check, posterior versus prior comparison, trace and Rhat plots can be found in Section A.6.

3.2 Model justification

We used a logistic regression model in a Bayesian framework due to the fact that our outcome is binary and predicts if a homicide is unsolved or not. Another model we considered is a logistic regression model with an instrumental variable. Introducing a instrumental variable into our model could have potentially provided a more accurate model and given us more consistent coefficient estimates as noted by Cameron and Trivedi (2005). However with the available information we had about each case, there was no candidate instrumental variable that impacted at least one variable in our data and not influence the outcome of the case being solved or unsolved. Thus, the model would fail the "Exclusion Restriction" assumption mentioned by Alexander (2023). We also went through different pairs and groupings of variables and how there was not a strong relationship between variables that is relevant and consequently fail Alexander (2023)'s "Relevance" assumption. We also have known treatment variable/predictor variables that can be used to measure the outcome variable and therefore, the instrumental variable is less likely to be necessary (Alexander 2023).

In our model, we assumed there is a relationship between the outcome variable, homicide is unsolved, with homicide case information like demographic (sex, race, and age of victim), geographic (city), and temporal (year) information, which are the predictor variables. For the demographic data, we decided to keep the grouping provided by The Washington Post such as for sex it was "female" and "male" and race it was "White", "Black", "Hispanic", "Asian", and "Other. However, we did remove victims who has any demographic information that falls under the"Unknown" grouping from our dataset and did not run our regression model on these observations. The reason for this is due to factors such as the data type of the predictor and keeping consistency across all observation and removing any unknown values. For example, our predictor variable, victim's age is a integer data type but it contained the string "Unknown" in the raw dataset. So the values "Unknown" is removed in our final dataset and not used to train our model. Since our outcome variable is binary for our model, we constructed the arrest_was_not_made column for it based on the disposition variable to reflect it and considered dispositions with values like "Open/No arrest" and "Closed without arrest" to be 1 and "Closed by arrest" to be 0.

Performing root mean square deviation (RSME) calculations on our model in Table 7 yields a RSME value of about 0.45. RSME represents the difference between the model's predicted values and observed values with 0 meaning that the predicted and observed values are the same (Frost 2023). Since the model's RSME value is close to 0, we can say our model is able to predict values with less error compared to other models that have a higher RSME value.

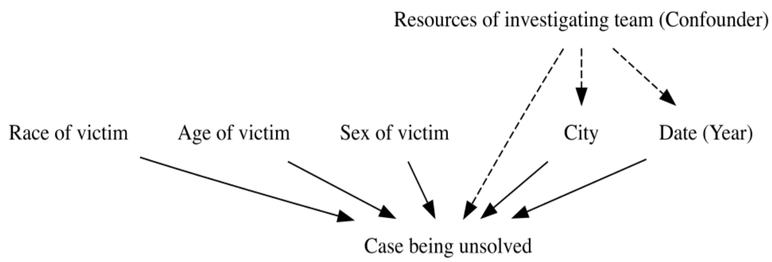


Figure 1: Causal relationship between homicide case information and homicide case being unsolved

However, our model has limitations and there are situations where this model would not work. Figure 1 shows that there is a confounder, "resources of investigating team" between the predictor variables, city and year and the outcome variable, case being unsolved. We define "resources of the investigating team" to include any of the following: the amount of people on the team investigating the case, time available allotted to investigate the case, amount of open cases for the team, cost, skill and education levels of members, etc. We currently do not have any information available about the resources of the investigating team for the case and further investigation is needed. We proposed an idealized survey we would conduct to collect the necessary data to further understand the relationship between homicide case characteristics and unsolved homicides in Section A.5. If we have information available about the resources of the investigating team for the case, the current model would not work and would need to be revised. This is due to the interaction between the city, year, and outcome variables with the confounder.

4 Results

4.1 Differences in Homicide Case Information Between Solved and Unsolved Cases in Chicago and Los Angeles (2010 to 2017)

4.1.1 Date (Month and Year)

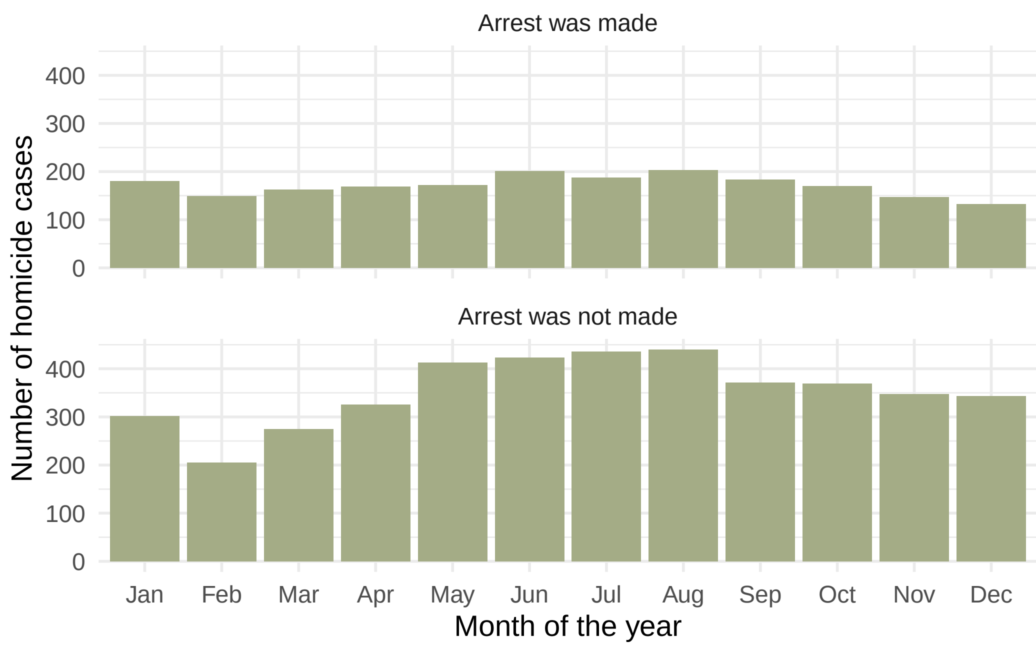


Figure 2: Number of solved and unsolved homicides across the 12 months of a year in Los Angeles and Chicago (2010 to 2017)

- TODO: Add in summary statistics table

4.1.2 City

Table 3: Proportion and number of solved and unsolved homicides in Los Angeles and Chicago (2010 to 2017)

City	Status of the homicide case	Number of cases	Proportion of cases
Chicago	Arrest was made	947	0.23
Chicago	Arrest was not made	3164	0.77
Los Angeles	Arrest was made	1109	0.51

Table 3: Proportion and number of solved and unsolved homicides in Los Angeles and Chicago (2010 to 2017)

City	Status of the homicide case	Number of cases	Proportion of cases
Los Angeles	Arrest was not made	1087	0.49

4.1.3 Disposition

Table 4: Disposition of homicide cases in Chicago and Los Angeles (2010 to 2017)

City	Disposition of the homicide case	Number of cases
Chicago	Closed by arrest	947
Chicago	Closed without arrest	216
Chicago	Open/No arrest	2948
Los Angeles	Closed by arrest	1109
Los Angeles	Open/No arrest	1087

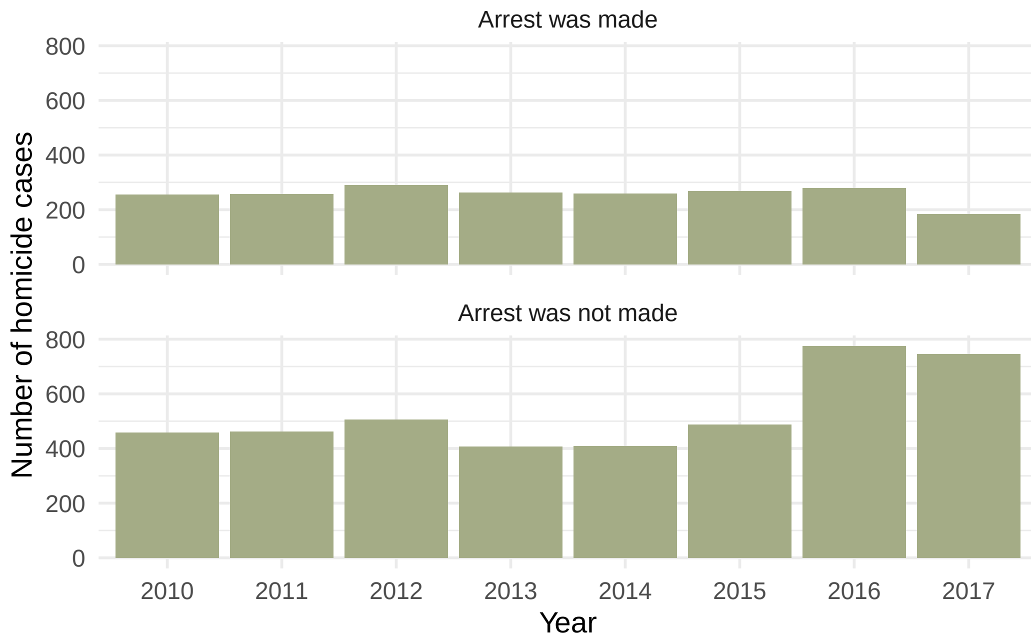


Figure 3: Number of solved and unsolved homicides from 2010 to 2017 in Los Angeles and Chicago

4.1.4 Victim's Age

4.1.5 Victim's Sex

Table 5: Proportion and number of homicide cases per sex in Chicago and Los Angeles (2010 to 2017)

Victim's sex	Status of the homicide case	Number of cases	Proportion of cases
Female	Arrest was made	335	0.49
Female	Arrest was not made	348	0.51
Male	Arrest was made	1721	0.31
Male	Arrest was not made	3903	0.69

4.1.6 Victim's Race

Table 6: Number of homicide cases per sex in Chicago and Los Angeles (2010 to 2017)

Victim's race	Status of the homicide case	Number of cases
White	Arrest was made	191
White	Arrest was not made	185
Asian	Arrest was made	22
Asian	Arrest was not made	21
Black	Arrest was made	1103
Black	Arrest was not made	2960
Hispanic	Arrest was made	699
Hispanic	Arrest was not made	1064
Other	Arrest was made	41
Other	Arrest was not made	21

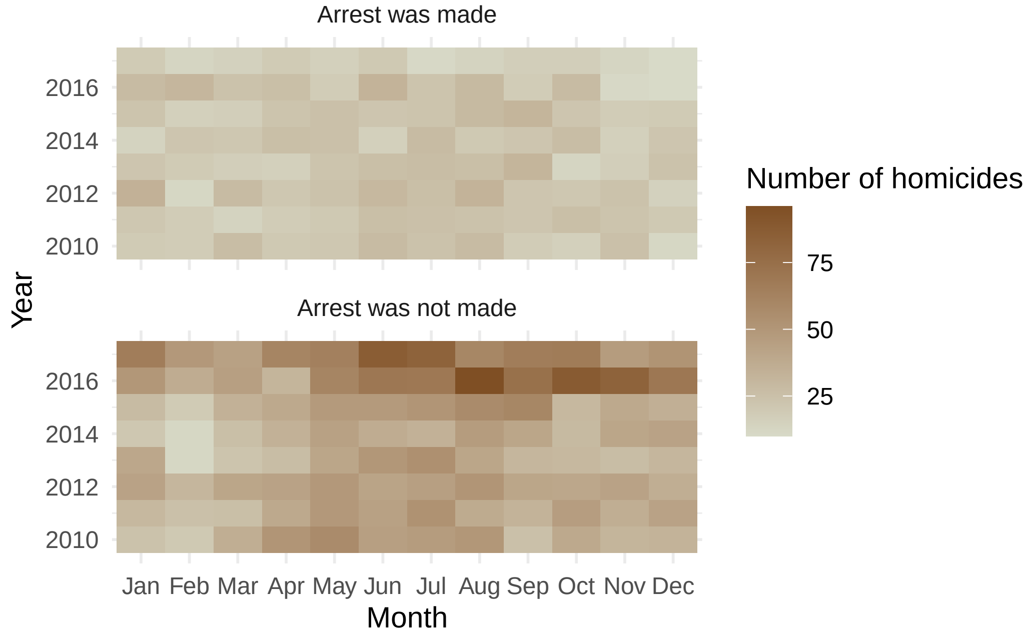


Figure 4: Number of solved and unsolved homicides from January to December from 2010 to 2017 in Los Angeles and Chicago

4.2 Model Results

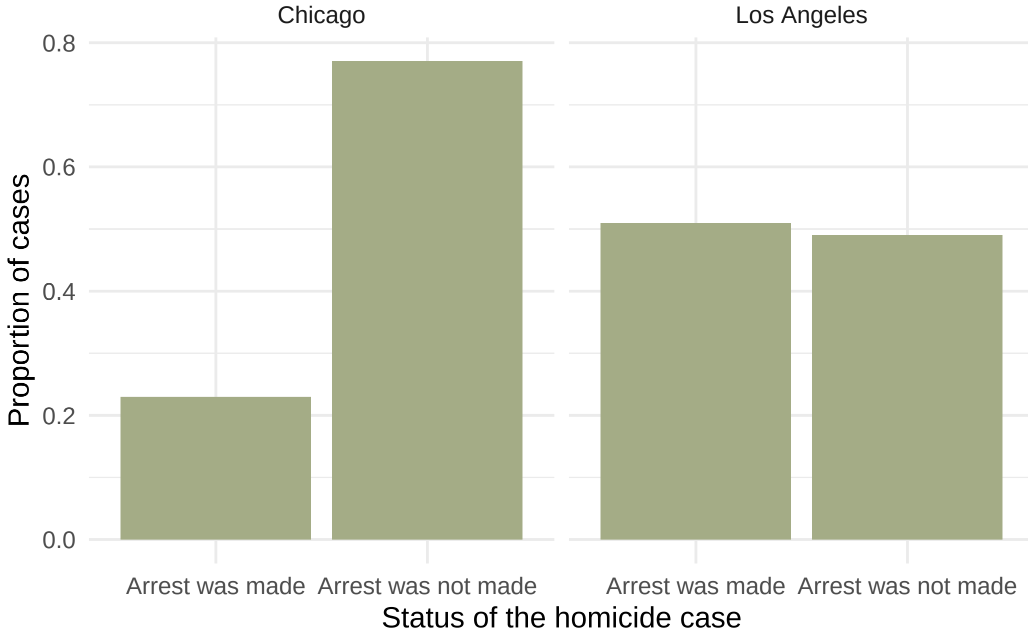


Figure 5: Proportion of solved and unsolved homicides in Los Angeles and Chicago (2010 to 2017)

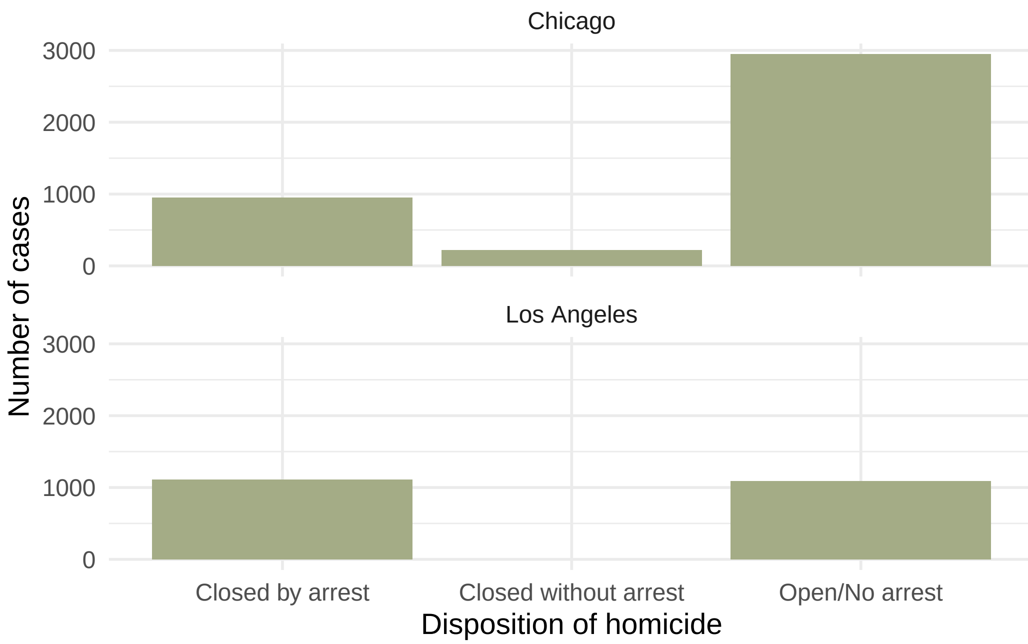


Figure 6: Disposition of homicide cases in Chicago and Los Angeles (2010 to 2017)

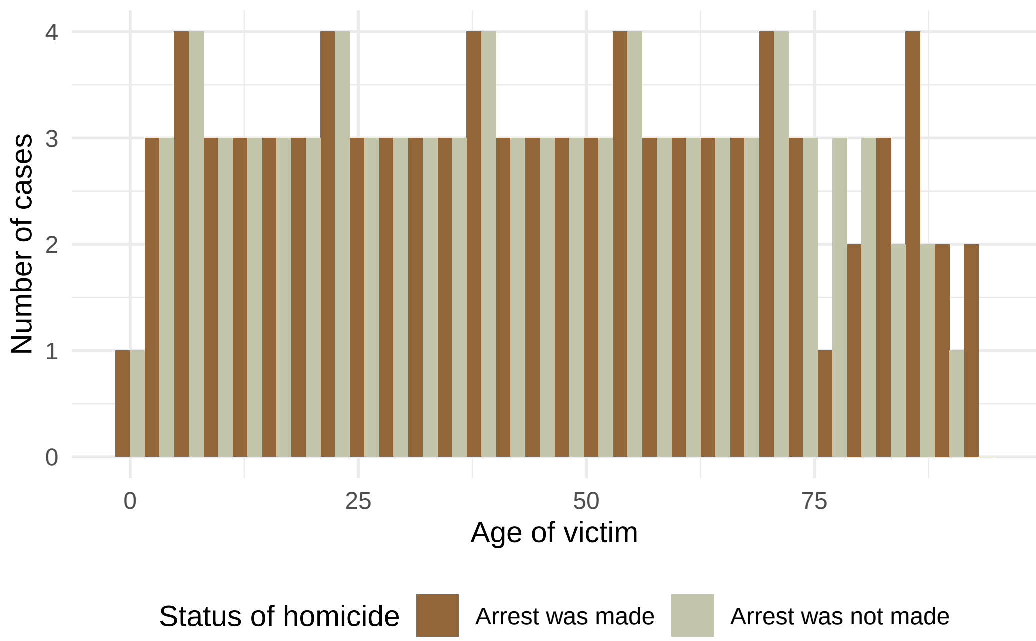


Figure 7: Distribution of victim's age in solved and unsolved homicides in Chicago and Los Angeles (2010 to 2017)

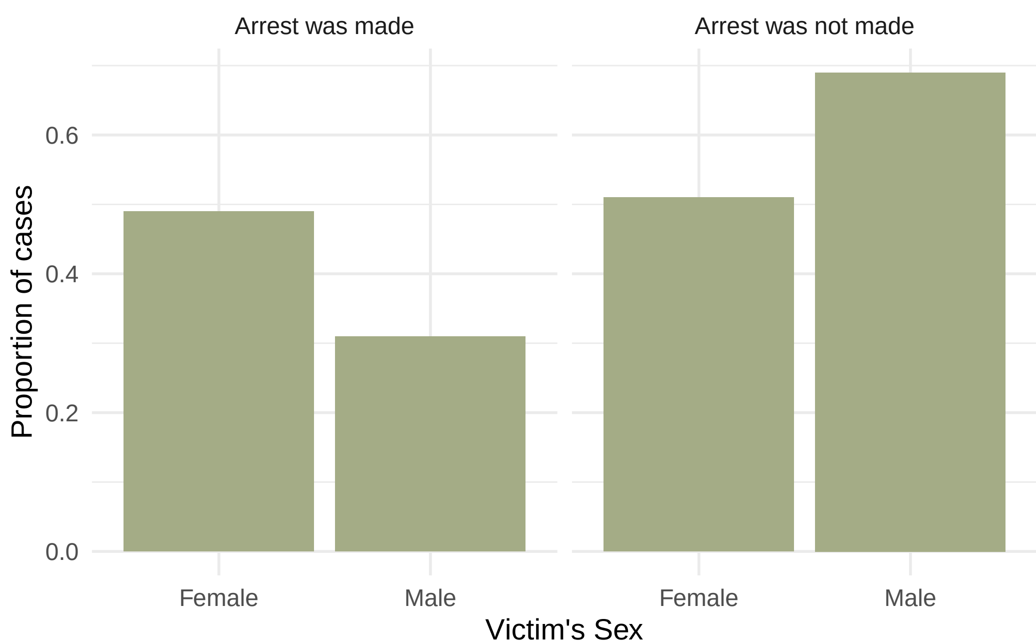


Figure 8: Proportion of homicide cases per sex in Chicago and Los Angeles (2010 to 2017)

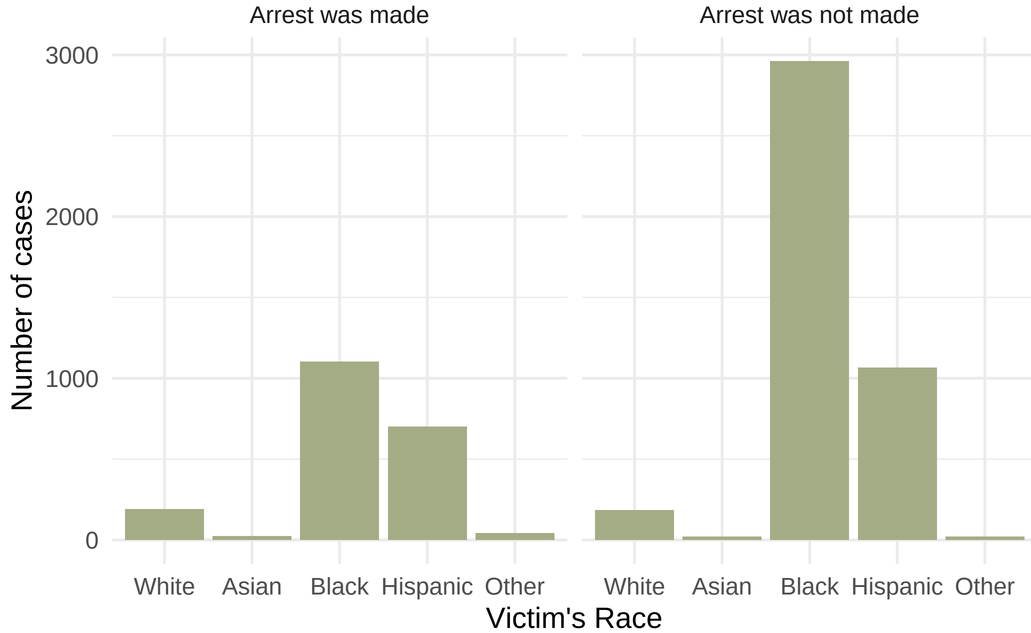


Figure 9: Number of homicide cases per race in Chicago and Los Angeles (2010 to 2017)

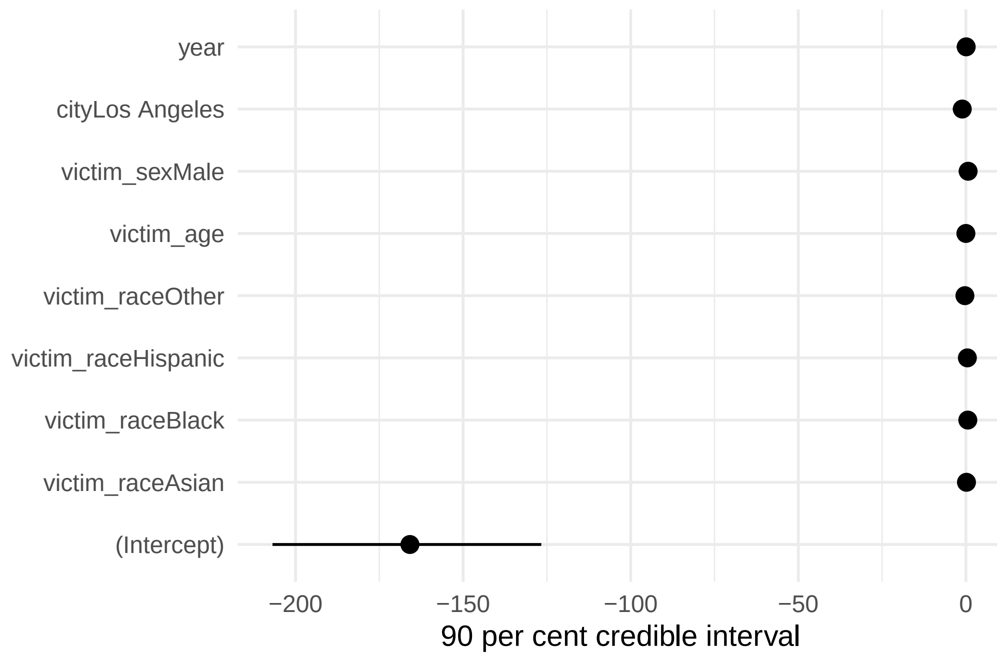


Figure 10: The credible intervals (line) for coefficient estimates (dot) of predictor variables for homicides that go unsolved from 2010 to 2017.

Table 7: Relationship between a homicide being unsolved from 2010 to 2017 with the city and year a victim is found in/on and the race, age, and sex of a victim. Mean absolute deviation (MAD) values are in parenthesis.

	Unsolved homicides (2010 to 2017)
(Intercept)	−165.865
	(24.869)
victim_raceAsian	0.168
	(0.347)
victim_raceBlack	0.567
	(0.124)
victim_raceHispanic	0.434
	(0.126)
victim_raceOther	−0.296
	(0.283)
victim_age	−0.006
	(0.002)
victim_sexMale	0.662
	(0.084)
cityLos Angeles	−1.089
	(0.066)
year	0.083
	(0.012)
Num.Obs.	6307
R2	0.105
Log.Lik.	−3657.680
ELPD	−3666.8
ELPD s.e.	34.2
LOOIC	7333.7
LOOIC s.e.	68.4
WAIC	7333.6
RMSE	0.45

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Areas of improvement and next steps

Weaknesses and next steps should also be included.

A Appendix

A.1 Note on Reproducing

In order to reproduce the results in the paper, first run the 00-install_packages.R in the scripts folder located in this paper's GitHub repository. Then run the other scripts based on the number at the beginning of the script name.

A.2 Acknowledgments

We would like to thank Alexander (2023) for providing assistance with the R code used to produce the tables and graphs in this paper.

A.3 Code styling

Code written in the scripts was checked and styled with lintr (Hester et al. 2024) and styler (Müller and Walthert 2024).

A.4 Additional Tables

Table 8: Number of solved and unsolved homicides across the 12 months of a year in Los Angeles and Chicago (2010 to 2017)

Status of the homicide case	Month	Number of cases in the month
Arrest was made	Jan	180
Arrest was made	Feb	149
Arrest was made	Mar	163
Arrest was made	Apr	169
Arrest was made	May	172
Arrest was made	Jun	201
Arrest was made	Jul	187
Arrest was made	Aug	203
Arrest was made	Sep	183
Arrest was made	Oct	170
Arrest was made	Nov	147
Arrest was made	Dec	132
Arrest was not made	Jan	302
Arrest was not made	Feb	205
Arrest was not made	Mar	275
Arrest was not made	Apr	326

Table 8: Number of solved and unsolved homicides across the 12 months of a year in Los Angeles and Chicago (2010 to 2017)

Status of the homicide case	Month	Number of cases in the month
Arrest was not made	May	413
Arrest was not made	Jun	423
Arrest was not made	Jul	436
Arrest was not made	Aug	440
Arrest was not made	Sep	371
Arrest was not made	Oct	369
Arrest was not made	Nov	348
Arrest was not made	Dec	343

Table 9: Number of solved and unsolved homicides from 2010 to 2017 in Los Angeles and Chicago

Status of the homicide case	Year	Number of cases in the year
Arrest was made	2010	256
Arrest was made	2011	257
Arrest was made	2012	290
Arrest was made	2013	262
Arrest was made	2014	259
Arrest was made	2015	269
Arrest was made	2016	279
Arrest was made	2017	184
Arrest was not made	2010	459
Arrest was not made	2011	462
Arrest was not made	2012	505
Arrest was not made	2013	407
Arrest was not made	2014	409
Arrest was not made	2015	488
Arrest was not made	2016	775
Arrest was not made	2017	746

A.5 Idealized Survey and Methodology

A.5.1 Idealized Survey Objectives

The objective of our survey is to collect information from Chicago Police Department and Los Angeles Police Department about their education background, reason for closing a homicide case without arrest, difficult homicide case they have done, how many cases they have to cover at a time, etc.

regarding resources available

A.5.2 Sampling Approach

- what is the population, frame, and sample;
- how is the sample recruited;
- what sampling approach is taken, and what are some of the trade-offs of this;
- how is non-response handled;

The sampling approach we plan to take is stratified sampling, which is a type of probabilistic sampling (Alexander 2023)

A.5.3 Respondent Recruitment

A.5.4 Data Validation

A.5.5 Idealized Survey Design

- Using an online survey in cases.

A.5.6 Link to Idealized Survey

- Using Google Forms

A.5.7 Limitations

- what is good and bad about the sampling.

A.5.8 Idealized Survey Questions

- Should have an introductory section and include details of a contact person
- Question type should be varied and appropriate.
- Have a final section that thank the respondents

A.6 Model details

A.6.1 Variance Inflation Factor

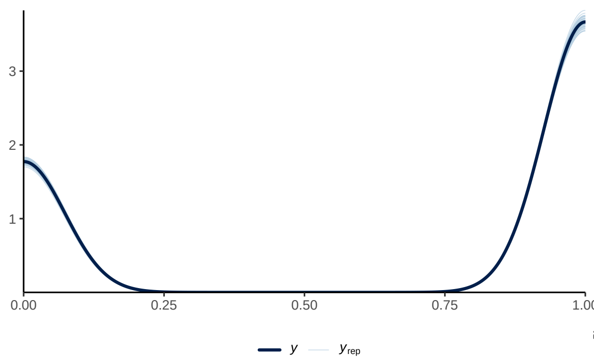
Table 10: Valence inflation factor (VIF) of each predictor for unsolved homicide model from 2010 to 2017

	GVIF	Df	GVIF^(1/(2*Df))
victim_race	1.268056	4	1.030131
victim_age	1.120664	1	1.058614
victim_sex	1.016122	1	1.008029
city	1.208373	1	1.099260
year	1.009956	1	1.004966

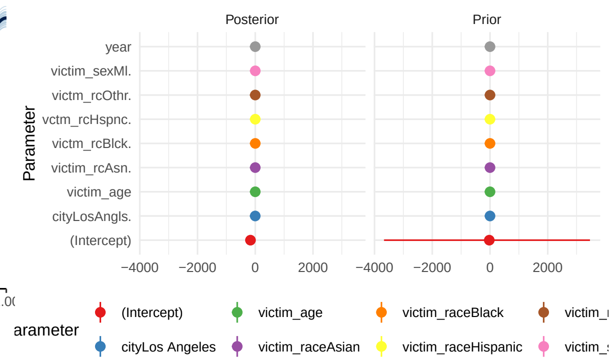
A.6.2 Posterior predictive check

In Figure 11a we implement a posterior predictive check. This shows...

In Figure 11b we compare the posterior with the prior. This shows...



(a) Posterior prediction check



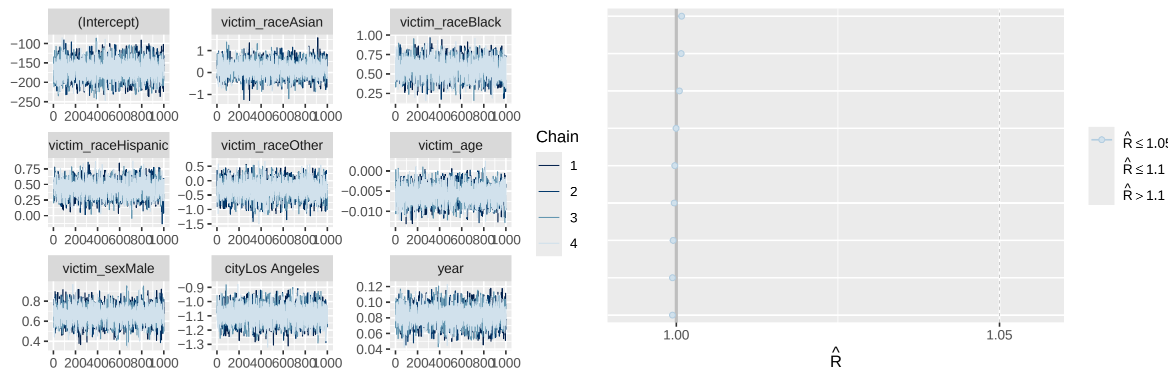
(b) Comparing the posterior with the prior

Figure 11: Examining how the model fits, and is affected by, the data

A.6.3 Diagnostics

Figure 12a is a trace plot. It shows... This suggests...

Figure 12b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 12: Checking the convergence of the MCMC algorithm

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. https://cameron.econ.ucdavis.edu/e240a/ch04iv.pdf.

City of Chicago. 2024. *Crimes - 2001 to Present*. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data.

City of Los Angeles. 2024. *Crime Data from 2010 to 2019*. https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z/about_data.

Federal Bureau of Investigation. 2024. *Crime/Law Enforcement Stats (Uniform Crime Reporting Program)*. https://www.fbi.gov/how-we-can-help-you/more-fbi-services-and-information/ucr.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

Frost, Jim. 2023. *Root Mean Square Error (RMSE)*. https://statisticsbyjim.com/regression/root-mean-square-error-rmse/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hester, Jim, Florent Angly, Russ Hyde, Michael Chirico, Kun Ren, Alexander Rosenstock, and Indrajeet Patil. 2024. *Lintr: A 'Linter' for r Code*. https://CRAN.R-project.org/packag e=lintr.

Iannone, Richard. 2016. *DiagrammeRsvg: Export DiagrammeR Graphviz Graphs as SVG*. https://CRAN.R-project.org/package=DiagrammeRsvg.

Iannone, Richard, and Olivier Roy. 2024. *DiagrammeR: Graph/Network Visualization*. https://CRAN.R-project.org/package=DiagrammeR.

Müller, Kirill, and Lorenz Walthert. 2024. *Styler: Non-Invasive Pretty Printing of r Code*. https://CRAN.R-project.org/package=styler.

Ooms, Jeroen. 2024. *Rsvg: Render SVG Images into PDF, PNG, (Encapsulated) PostScript, or Bitmap Arrays*. https://CRAN.R-project.org/package=rsvg.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

The Washington Post. 2018a. *Homicide Database: Mapping Unsolved Murders in Major US Cities*. https://www.washingtonpost.com/graphics/2018/investigations/unsolved-homicide-database/.

———. 2018b. *How the Post Mapped Unsolved Murders: Unsolved Homicide Database*. https://github.com/washingtonpost/data-homicides.

Urbanek, Simon. 2022. *Png: Read and Write PNG Images*. https://CRAN.R-project.org/package=png.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2024. *_Knitr: A General-Purpose Package for Dynamic Report Generation in r_*. https://yihui.org/knitr/.