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

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

Emily Su

December 14, 2024

Despite Chicago Mayor Brandon Johnson announcing in 2024 that the Chicago Police Department had the highest homicide clearance rate of 54% in years, homicide clearance rates across the United States (US) have been decreasing since 1980, with more homicides going unsolved. This paper looks at patterns of homicide case information with the case resolution from 2010 to 2017 in two of the largest cities in the United States, Chicago and Los Angeles. We found that unsolved homicides are more likely to occur in the middle and later half of the year and in Chicago compared to Los Angeles with the majority of victims being Black or Hispanic males. These findings can inform the public and US police departments of populations more vulnerable to having unsolved cases however further investigation is needed on the investigators involved in each unsolved homicide case.

Table of contents

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

1 Introduction

On November 20, 2024, Chicago Mayor Brandon Johnson announced that the Chicago Police Department had the highest homicide clearance rate of 54% in recent years (Bradley and Schroedter 2024). Despite celebrating this accomplishment, Bradley and Schroedter (2024) found that if only the cases with arrest were considered and not closed cases where the police believed they had identified the perpetrator without an arrest being made, their clearance rate was instead 23% this year. This concern about homicide clearance rate was expressed by Clayton (2023), who mentioned that homicide clearance rates from 1980 to 2020 have decreased from 71% to 50% in the United States (US). As of 2023, there are currently no up-to-date databases keeping track of homicide resolutions, and current crime databases do not indicate a victim's race and what led to a homicide, making it difficult to figure out why homicides are going unsolved (Clayton 2023). The ongoing concern about homicide clearance raises the following question: what are the differences in homicide case information like the year and city where the homicide took place and the victim's perceived characteristics (age, sex, and race) between solved and unsolved homicides in the 2 of the largest cities in the United States, Chicago and Los Angeles, from 2010 to 2017?

Our estimand is the homicide case information like the victim's demographic information (age, sex, and race) and the year and city the homicide occurred in/on if the homicide case is unsolved or solved.

In this paper, to see what factors could impact the likelihood of a homicide case going unsolved, we analyzed data provided by The Washington Post (The Washington Post 2018b) that was compiled by their reporters on homicide cases and their status from 2010 to 2017. Previously, The Washington Post used the data they assembled to construct a map and pinpoint places where unsolved homicides occur (The Washington Post 2018a). In another study, Magee et al. (2020) looked at factors that contributed to the low homicide clearance rates in Indianapolis, Indiana across different neighbourhoods. They found that there was an association between neighbourhoods with higher arrest rates and having a higher chance of homicide clearance (Magee et al. 2020). However, specific analysis looking at the differences in case characteristics between two US cities, Chicago and Los Angeles, has not been done previously. Our findings indicate that a majority of unsolved homicide victims from 2010 to 2017 are Black or Hispanic and are likely to be male but their age does not impact if their case goes unsolved or not. We also found that unsolved homicides are more likely to occur in Chicago compared to Los Angeles from 2010 to 2017. From our analysis, most unsolved homicide cases occurred during the middle and later half of the year with there being a jump in unsolved cases from 2016 onwards. These findings could inform the public and police departments about populations

more vulnerable to homicide cases going unsolved and cities to be more alert about such as Chicago with homicide clearance rates.

For the remainder of the paper, the data section (Section 2) goes over the data we used and its features and limitations, how the data was obtained, our variables of interest for our model and preliminary analysis, and a bit of our data cleaning process. The model section (Section 3) discusses the type of model we used, the relationships between variables in our model, the justification for our model including the root mean square error (RMSE) for our model. The results section (Section 4) contains tables and graphs created based on our data and the results from our model. The discussion section (Section 5) goes over what we did in our results, explaining the meaning of our findings, the implications of our findings to the real world, and areas of improvement and suggestions for future works. The appendix (Section A) includes a link to a dashboard that was created with interactive versions of the graphs in this paper, additional tables, notes on reproducing and code styling, model diagnostics and validation, our idealized survey and methodology to gather data on investigators responsible for homicide cases, and our evaluation of The Washington Post's methodology.

2 Data

2.1 Overview

The dataset we used for this paper comes from The Washington 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 the predictor variables, the outcome variable, and a confounder used DiagrammeR (Iannone and Roy 2024), rsvg (Ooms 2024), DiagrammeRsvg (Iannone 2016) and png (Urbanek 2022).

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 (The Washington Post 2024). However, due to there being only 416 observations with numerous observations missing information, we forgo using the dataset.

We retrieved the raw dataset from The Washington Post using a script that downloads the CSV file from their GitHub repository. The raw dataset contains 52,179 observations with each observation being a homicide case. However, since we narrowed our scope to focus on one of the two most populated US cities, Los Angeles and Chicago, the number of observations in our cleaned dataset ended up being 6,307. 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 indicate 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" (1,763 observations), "Black" (4,063 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 shows that the mean victim_sex is the sex of the homicide victim where they are 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 statuses: "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 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 handwrite information into physical crime reports with details like the type of crime, the premise the crime occurred 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 use 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 place at (City of Chicago 2024). This information is then reported to the FBI monthly under the UCR program (City of Chicago 2024).

Reporters from the Washington Post then obtained the data from the FBI specifically on homicides from the 50 largest US cities from 2007 to 2017, which can be accessed through the UCR publications page on the FBI website (The Washington Post 2018b). They selected the 50 largest cities based the size of the city in 2012 (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 also used 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 (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 from 2010 to 2017.

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 impact 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.8 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 a posterior predictive check, trace and Rhat plots can be found in Section A.8.

3.2 Model justification

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 "female" and "male" for the victim's sex and "White", "Black", "Hispanic", "Asian", and "Other" for the victim's race. However, we removed victims who had any demographic information with the "Unknown" value 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 our predictors and keeping consistency across all observations by 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 containing "Unknown" are removed from 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 indicate if an arrest was made or not or in order words, the homicide is solved or unsolved. We considered dispositions with values like "Open/No arrest" and "Closed without arrest" to be 1 and "Closed by arrest" to be 0 for the arrest_was_not_made column.

We used a logistic regression model in the Bayesian framework since our outcome is binary and it 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 found there was not a single strong relationship between variables that is relevant and consequently failing 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).

Performing root mean square error (RSME) calculations on our model in Table 8 yields a RSME value of about 0.45. RSME represents the difference between the model's predicted

values and observed values with 0 indicating 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.

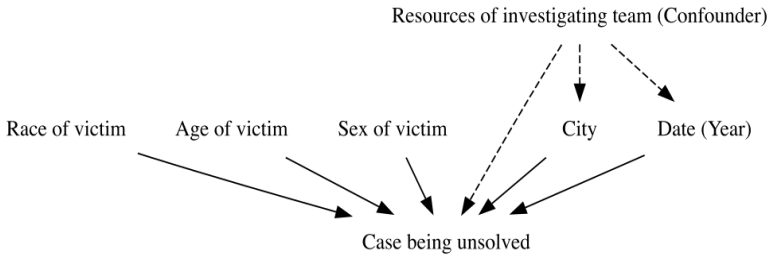


Figure 1: Causal relationship between homicide case information and homicide case being unsolved

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 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.6. 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 confounder with the city, year, and outcome variable.

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)

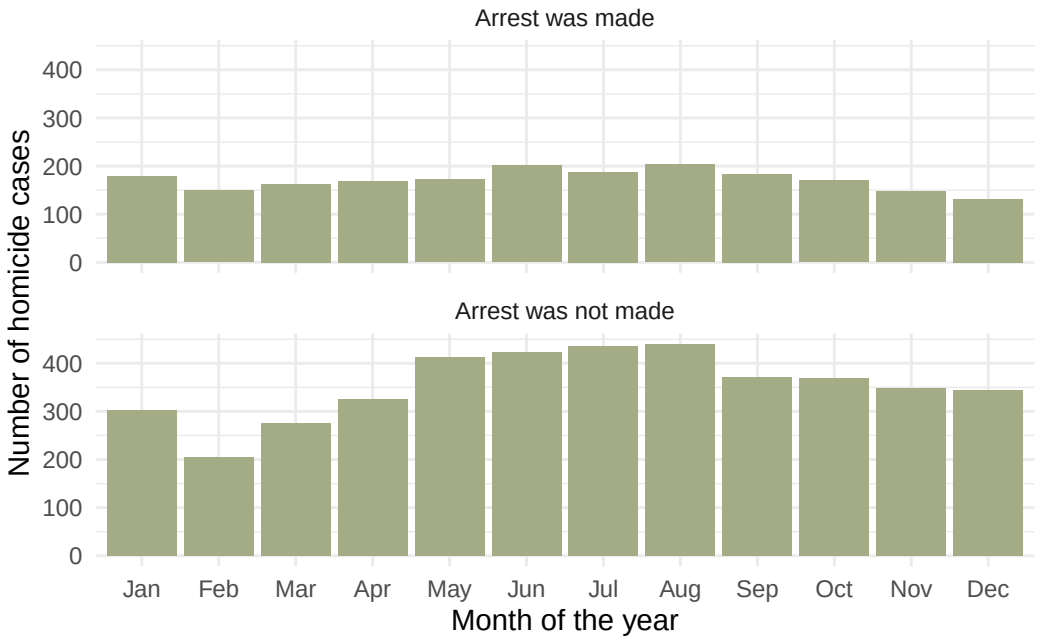


Figure 2: Number of solved and unsolved homicides across the 12 months of a year in Los Angeles and Chicago (2010 to 2017)

Figure 2 and Table 9 indicates that most of the unsolved homicide cases in Los Angeles and Chicago from 2010 to 2017 were reported in the summer months with there being 436 unsolved cases in July and 440 for August. Figure 2 shows that most unsolved homicide cases occur during the middle and later half of the year. On the other hand, Table 9 and Figure 2 shows that the number of cases that were reported with arrest also happen during August with 203 cases and June with 201 cases. However, there also are fewer arrest made compared to the number of cases where an arrest was not made.

Figure 3 and Table 10 shows that there was sudden increase in the number of unsolved homicides reported in 2016 and 2017 compared to previous years with 775 cases for 2016 and 746 cases for 2017. However, for the cases with an arrest made, most of the cases were reported back in 2012 with 290 cases and 2016 and 279 cases.

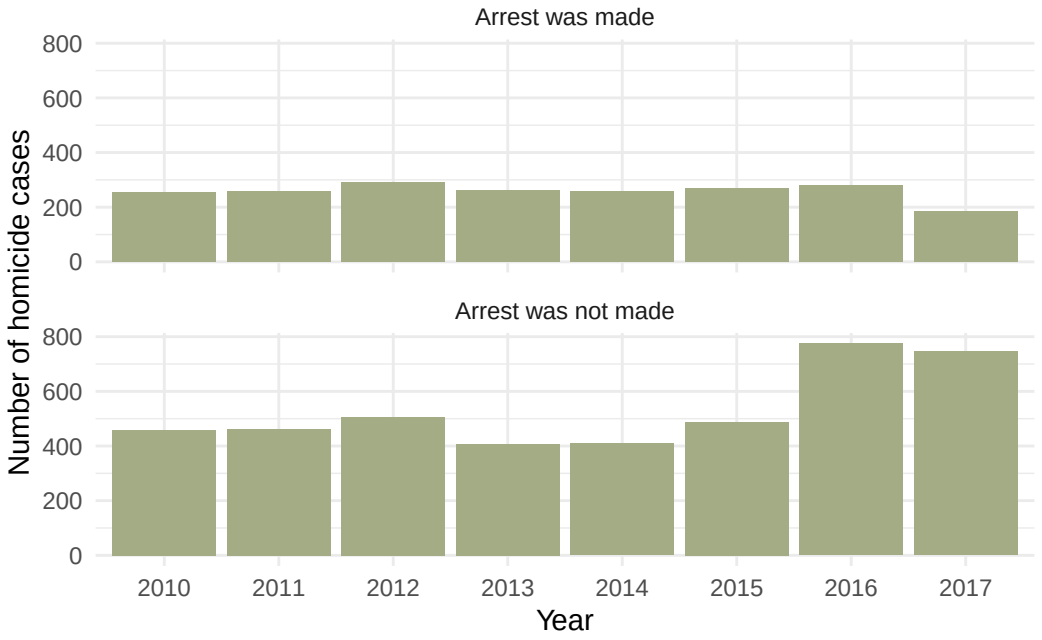


Figure 3: Number of solved and unsolved homicides from 2010 to 2017 in Los Angeles and Chicago

Based on where the dark brown colour appears in Figure 4, majority of homicides that go unsolved occur in the middle and later half of 2016 while for 2017 most of them happen middle of 2017. For the homicides that go solved, the lightness appears to be uniform across with the more dark parts in the middle of Figure 4.

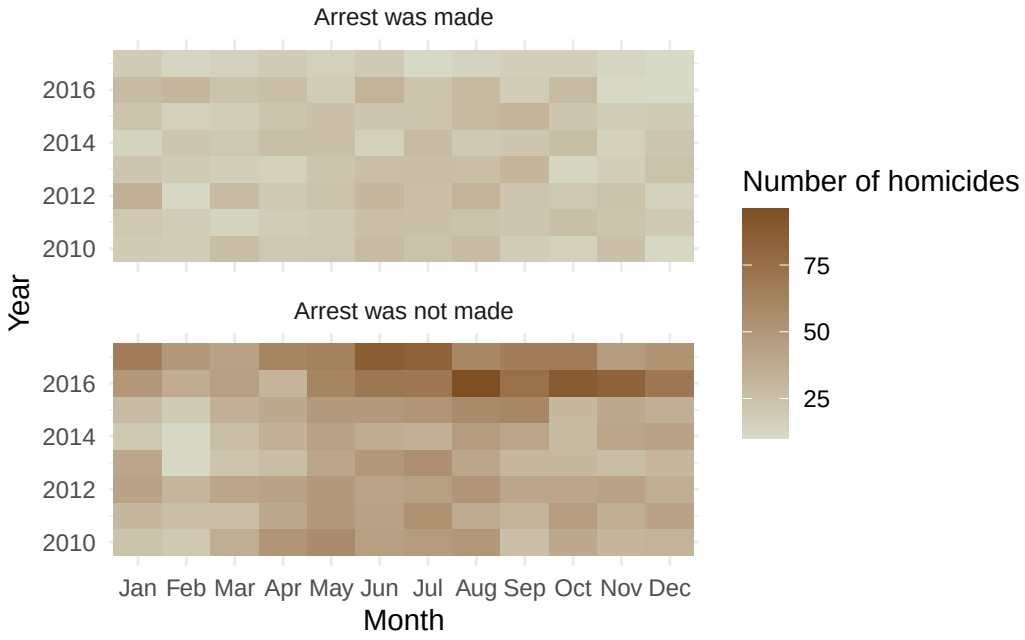


Figure 4: Number of solved and unsolved homicides from January to December from 2010 to 2017 in Los Angeles and Chicago

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
Los Angeles	Arrest was not made	1087	0.49

Figure 5 and Table 3 shows that 77% of Chicago's homicide cases are unsolved (with 3,164 cases) while for Los Angeles, 49% (1,087 cases) of their homicide cases are unsolved from 2010 to 2017. Only 23% (947 cases) of homicide cases are solved with an arrest made for Chicago while for Los Angeles, 51% (1,109 cases) of homicides are solved with an arrest made. Chicago has a higher percentage and higher number of unsolved homicide cases compared to Los Angeles.

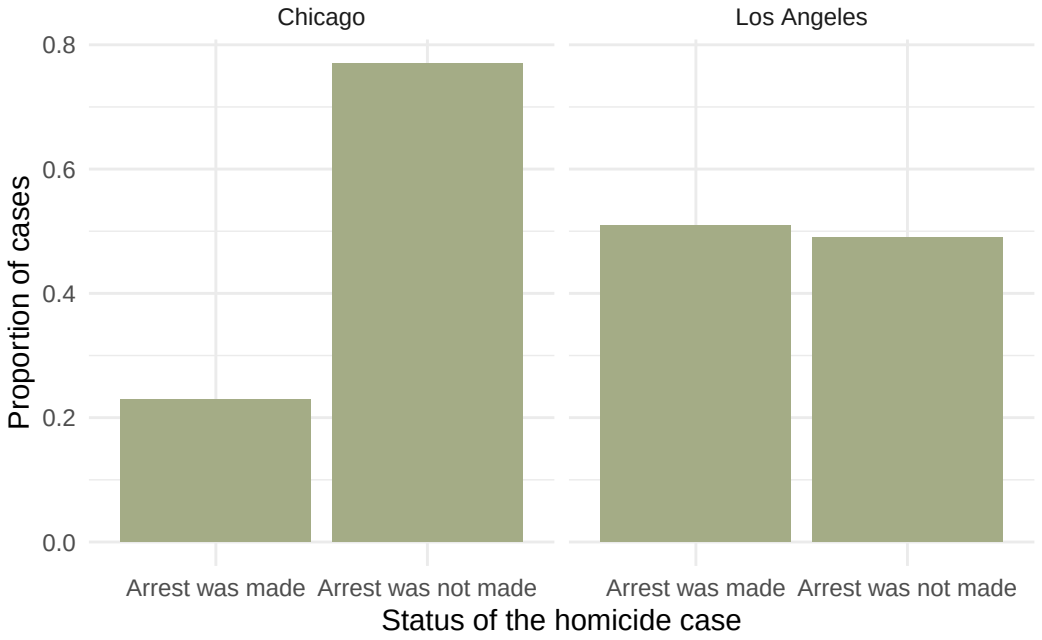


Figure 5: Proportion of solved and unsolved homicides in Los Angeles and Chicago (2010 to 2017)

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 6 and Table 4 shows that the number of homicide cases that are closed by arrest at 1,109 cases is close to the number of homicide cases that are open at 1,087 cases from 2010 to 2017. On the hand, there is a larger difference for Chicago between the number of homicides that are closed by arrest at 947 cases with the ones that are closed without arrest and open/no arrest at 216 and 2,948 cases, respectively.

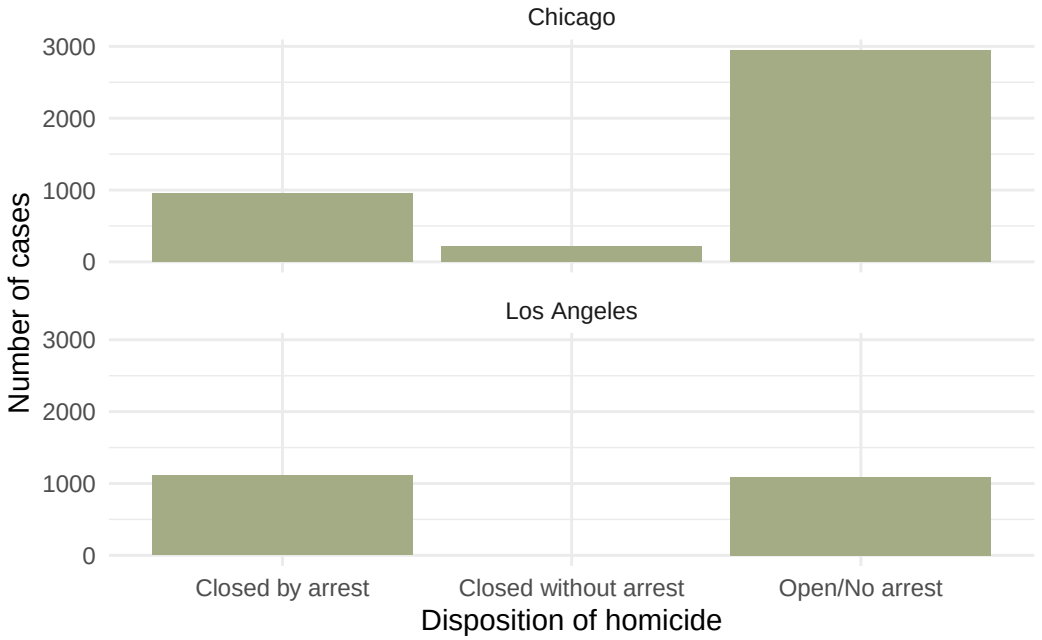


Figure 6: Disposition of homicide cases in Chicago and Los Angeles (2010 to 2017)

4.1.4 Victim's Age

Table 5: 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's Age
Min. : 1.00
1st Qu.:22.50
Median :44.00
Mean :44.55
3rd Qu.:66.00
Max. :94.00

Figure 7 shows that the distribution of age of homicide victims is relatively uniform. It appears that the victims of unsolved and solved homicides are about the same age if one were to not consider ages above 75. However, if we were to only consider age 75 onwards, it appears that the distribution is a bit right-skewed. We can see that from Table 5, the youngest homicide victim is 1 years old and the oldest is 94 years old. The median age of homicide victims is 44 years old with the mean age being 45 years old.

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

4.1.5 Victim's Sex

Table 6: 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

Figure 8 and Table 6 shows that most victims of homicide are male with 1,721 solved cases and 3,903 unsolved cases in comparsion to female victims with 335 solved cases and 348 unsolved cases. The proportion of cases with arrest and not is almost equal for homicides with female victims at 49% and 51%, respectively. On the other hand, the proportion of cases with arrest (31%) and not (69%) has a 38% difference for homicides with female victims.

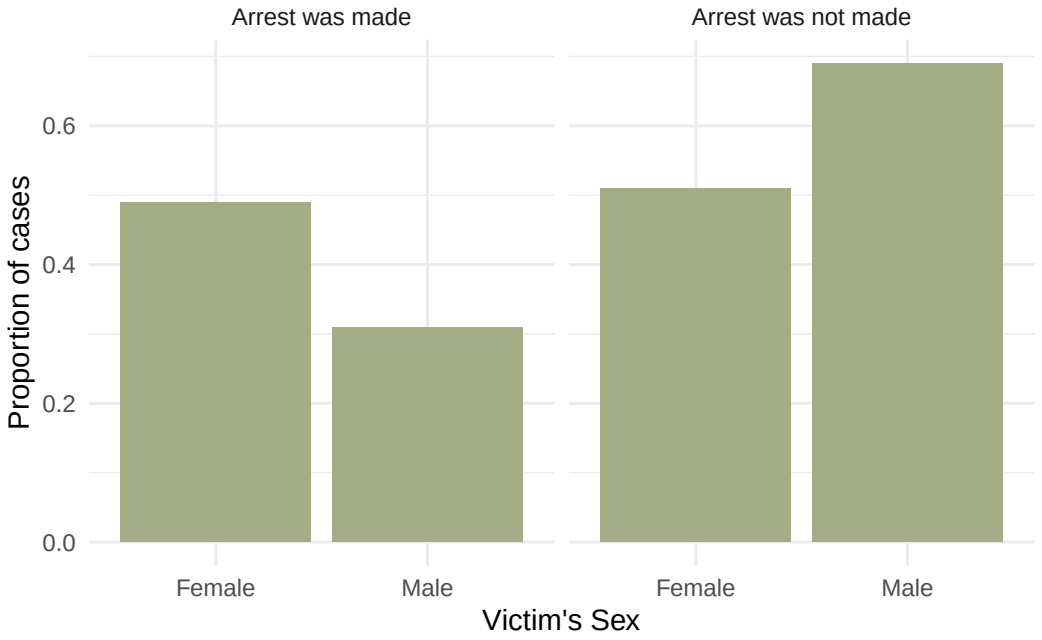


Figure 8: Proportion of homicide cases per sex in Chicago and Los Angeles (2010 to 2017)

4.1.6 Victim's Race

Table 7: 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

From Figure 9 and Table 7, they show that there were more homicide case solved when the victims identified as "White" (191 cases), "Asian" (22 cases), and "Other" (41 cases). However, a disproportionate number of homicide cases have victims who are Black or Hispanic. For both

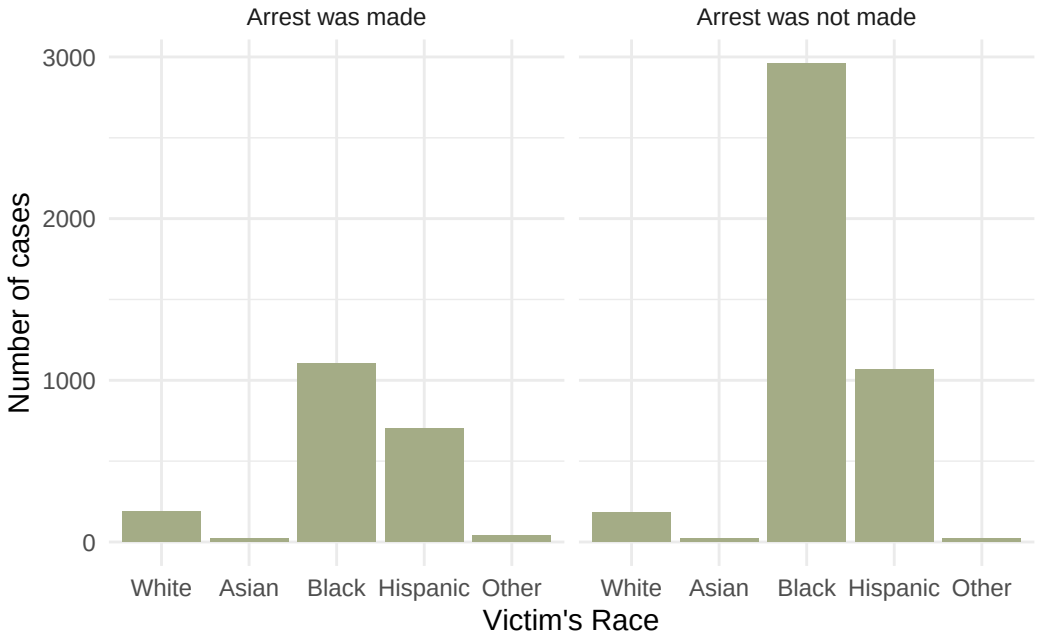


Figure 9: Number of homicide cases per race in Chicago and Los Angeles (2010 to 2017)

Black and Hispanic victims, they have more homicide cases that go unsolved with 2,960 and 1,064 cases compared to homicide cases that are solved with 1,103 and 699 cases, respectively.

4.2 Model Results

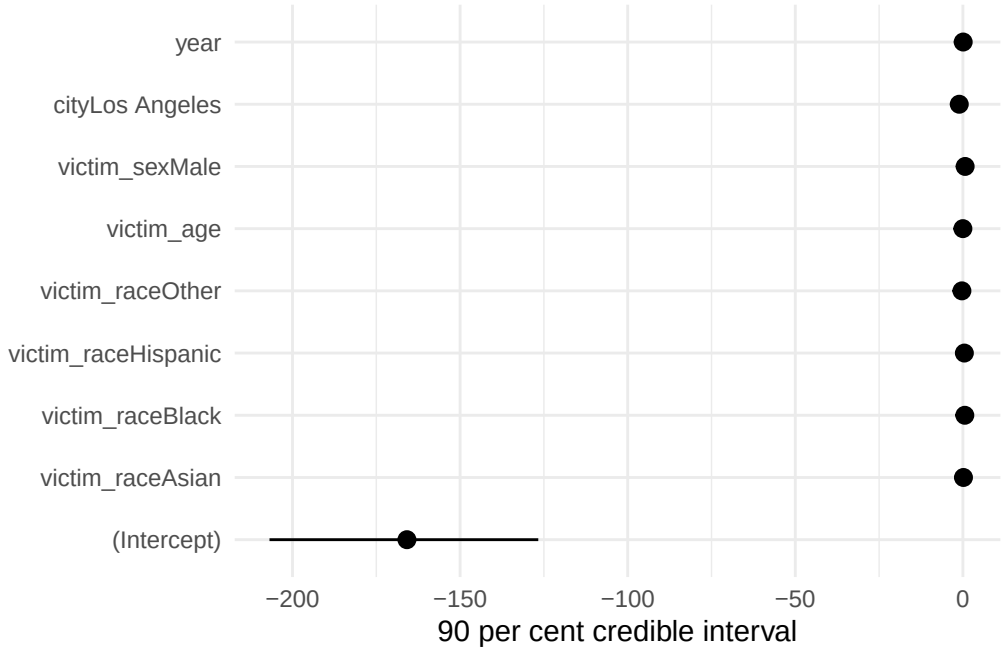


Figure 10: The credible intervals (line) for coefficient estimates (dot) of predictor variables for homicides that go unsolved from 2010 to 2017.

Table 8 and Figure 10 present the results from our logistic regression model in the Bayesian framework and the 90% credible intervals of each predictor, respectively. From Table 8, the intercept β_0 of -165.865 indicates that when the homicide victim is White, since "White" is the baseline for the victim_race predictor, and victim_age is 0 and year being 2010 with its coefficient being 0 and the victim's sex is female, the homicide is more likely to be solved. Table 8 indicates that the coefficient estimate of victim_race β_1 is 0.168 when the victim is Asian. This means that the log odds of a homicide case being unsolved increase by 0.168 when the victim is Asian while other predictor variables stay fixed. Following this, when the victim's race is Black or Hispanic, the likelihood of a homicide being unsolved increases by 0.567 and 0.434, respectively. This indicates that the likelihood of a homicide going unsolved increases if the victim's race is Asian, Black, or Hispanic relative to a victim's race being White. On the other hand, when the victim's race falls under "Other", the log odds of a homicide case being unsolved decreases by 0.296. This indicates that the likelihood of a homicide being unsolved decreases if the victim's race falls under "Other" relative to a victim's race being White. However, since the 90% credible interval for the coefficient estimate of victim_race appears to be close to 0 as seen in Figure 10, this means that a victim's race has a weak likelihood of impacting the outcome of their case being unsolved.

Table 8 shows that since the coefficient estimate of victim_age β_2 is -0.006, this means the log

Table 8: 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

odds of a homicide case being unsolved decreases by 0.006 when the age of the victim increases by 1 year as other predictor variables stay constant. This means victims of unsolved homicides are likely on the younger side. Figure 10 also indicates that β_2's 90% credible interval is close to 0 or includes 0 in its interval, indicating there is a chance the victim's age does not have an influence on the outcome of a homicide case being unsolved. With the coefficient estimate of victim_sex β_3 being 0.662 in Table 8, it indicates that the log odds of a homicide case being unsolved increases by 0.662 when the victim is male while other predictors are fixed. This suggests that the likelihood of a homicide case being unsolved increases when the victim is a male. Looking at Figure 10, the credible interval for β_3 is also slightly above 0. Table 8 also shows that the log odds of a homicide being unsolved decreases by 1.089 as indicated by the coefficient estimate β_4 when the city is set to Los Angeles with other predictors being fixed. This indicates that more unsolved homicide cases are likely to occur in Chicago instead. Figure 10 also shows that β_4's credible interval is below 0. Looking at the year predictor and its coefficient estimate, the log odds of a homicide case being unsolved increases by 0.083 as noted by the coefficient estimate β_5 when the year increases by 1 year while the other predictors stay constant. This indicates that cases are more likely to be unsolved in the later years between 2010 to 2017. Figure 10 also shows that β_5's credible interval is slightly above 0 or includes 0 indicating that the year has a weak likelihood of indicating if a homicide is unsolved.

5 Discussion

In Section 4, we analyzed homicide case information like the victim's age, sex, and race and the year and city where the homicide case occurred to see if there were any differences between homicide case information between unsolved and solved homicides from the year 2010 to 2017. We also built a Bayesian logistic regression model in Section 4.2, to see the likelihood of the homicide case information impacting the homicide case outcome of going unsolved.

5.1 Chicago is likely to have more unsolved homicides than Los Angeles with a surge in unsolved cases from 2016 onwards

From Figure 5 and Table 3, it indicates that while Los Angeles has slightly more homicide cases that are solved than unsolved, 77% of Chicago's homicide cases are unsolved from 2010 to 2017. This is also seen in Table 8, where our model indicates that unsolved homicides in Chicago are likely to go unsolved compared to Los Angeles. This suggests potential issues with how the Chicago Police Department is managing homicide case closure. From Figure 2 and Figure 3, we saw that most unsolved homicide cases occur in the middle and later half of the year and that there was a jump in unsolved homicides in 2016 and 2017 compared to past years. Ansari and Flores (2017) reported on January 2, 2017, that 2016 was the year with the most number of homicides in the past 19 years due to gun violence in the city. This suggests that most of the unsolved cases could have occurred due to police departments like the Chicago Police Department being unable to keep up with the surge of homicide cases.

5.2 Age of the homicide victim likely does not impact the outcome of their case being unsolved however their gender may

Figure 7 indicates that the age of the homicide victim likely does not impact the outcome of their case be unsolved due to the uniformity of the distribution of ages. This is further seen in Figure 10, where there is a chance the victim's age does not influence the outcome of a homicide case in our model. This implies that the age of the homicide victim does not impact if their case is unsolved or not. However, Figure 10 and Figure 8 indicate that female victims had a higher proportion of their homicide cases being solved while on the other hand, male victims had a higher proportion of their homicide cases going unsolved. This suggesst that males are more likely to have their cases go unsolved.

5.3 Disproportionate number of homicide victims in unsolved cases in Chicago and Los Angeles from 2010 to 2017 are Black and Hispanic

From Figure 9 and Table 7 it shows that Black and Hispanic victims make up more than 1,000 unsolved cases compared to other victims of other races who have under 1,000 unsolved cases.

However, Black victims make up the majority of unsolved homicides at 2,960 cases with only 1,103 solved homicides. Based on Table 8 and Figure 10, our model shows that there is a slight likelihood that being Black and Hispanic relative to being White impacts a homicide case going unsolved. Chavis (2021) had found that the Chicago Police Department was under scrutiny for another issue with their gang databases being inaccurate and targeting mostly Black and Latinx residents. Chavis (2021) also mentions that Chicago is one of the United States' most segregated cities both economically and racially. This suggests that Black and Hispanic victims could be facing systematic biases in homicide case closures and their cases being solved.

5.4 Areas of improvement

The dataset used in the paper was compiled in 2018 and we currently do not know which homicide cases have been solved as of 2024, which limits the years we were able to analyze. There are other factors like the resources of the police department that could have also impacted the outcome of homicides being unsolved, which we did not have information on and that could have reduced the accuracy of our model. Since some information in the dataset was aggregated by humans from not only FBI data but also papers, human error could have occurred when inputting information into the dataset. For future works on unsolved homicides, we recommend compiling a new dataset for the years 2018 to 2024 similar to what The Washington Post did and considering other factors like the resources of the police departments (The Washington Post 2018b).

A Appendix

A.1 Dashboard for Interactive Visualizations

A dashboard containing interactive versions of the graphs found in this paper was developed using shiny (Chang et al. 2024), shinydashboard (Chang and Borges Ribeiro 2021), and plotly (Sievert 2020). The link to the shiny app can be found here: https://49z7k8-emily-su.shinyapps.io/unsolved-homicides-app/.

A.2 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's name.

A.3 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.4 Note on Code styling

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

A.5 Additional Tables

Table 9: 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

Table 9: 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	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
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 10: 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.6 Idealized Survey and Methodology

A.6.1 Idealized Survey Objectives

The objective of our survey is to obtain information about the investigators from US police departments and their experience with dealing with homicide cases. This information would be used to take into account any potential factors about police officers that could impact if a homicide case ends up unsolved. Currently, we only have state collected homicide data that is collected by the FBI with basic demographic, temporal, and geographic information. Johnson (2021) looked at prison data and noted that the standards in place with state collected data do not help us fully understand the cause and impact of the US prison system. The data collected often only looked at the individuals in prison and not considered structural factors like the public's opinion to crime policies (Johnson 2021). This can also be applied to homicide data where we can not say we fully understand the cause of homicide cases going unsolved by only looking at the victim's characteristics and where they were found. Thus, this led us to conducting this survey to further understand the crime case system better and what interventions may be needed, which is also noted by Johnson (2021) on community sourced prison data. Community source data involves obtaining data from the people living in or with the system itself, as defined by Johnson (2021), and in our case, our survey aims to collect data from people who maintain the homicide case system. In the following idealized methodology, we will cover our sampling approach and respondent recruitment, data validation, the design of our survey, limitations of our survey, a link to how the survey could look, and the survey questions themselves.

A.6.2 Sampling Approach and Respondent Recruitment

Our target population are all investigators from US police departments as defined by (Alexander 2023). However, since we can not survey every investigator in the US, we will narrow down our group to be investigators from 5 US police department across the United States who are responsible for homicide cases and are located in large cities, which is our sampling frame. More specifically, our sample will be all the investigators from 5 US police department in large cities who are responsible for homicide cases and who we can gather data on (Alexander 2023).

The sampling approach we plan to take is a mix of non-probabilistic and probabilistic sampling. Conducting non-probabilistic sampling would ensure that we are obtaining the necessary information from police departments in more populated cities (Alexander 2023). However some limitations with non-probabilistic sampling is that our findings from the survey might not generalize to other police departments nationwide (Alexander 2023). We also plan to conduct probabilistic sampling within the non-probabilistic sample we have so that any person in our sampling frame has a chance of being picked for to complete our survey (Alexander 2023).

We will first select 5 police departments based on their city's population. To obtain information about the most populated cities in 2023, we will use the United States Census Bureau's website to research the 5 most populated city in 2023 or the most recent data they have published (United States Census Bureau 2024). Before proceeding to contact members under the detective bureau, we will reach out to public affairs office of each department and prepare the necessary documents and a proposal about our study to them. If no approval is received, we will reach out to another police department based on the population of the city they are in.

One we receive approval from the office, we will manually go through each of the department's website or reach out to the public affaris office for information and collect the emails, ranking, and names of investigators on the homicide case team. After information about each investigator has been collected from 5 departments, we will perform stratified sampling. Stratified sampling is a type of probabilistic sampling where we divide our sample into a strata based on their ranking and eventually form one stratum (Alexander 2023). We will then take a random sample from each strata and contact the individuals that were sampled. We aim to have at least 5 respondents from each strata complete our survey. We chose stratified sampling since ensures every ranking in a police department working on homicide cases is represented in our survey (Neyman 1934). However, a drawback of stratified sampling is that we may receive a non-response from individuals we sampled (Alexander 2023). As defined by Alexander (2023), non-response is a measurement error that happens when one refuses to respond to a survey or leaves questions/a question missing. To address non-response errors, respondents will receive monetary compensation for completing the survey to the end and we will also implement respondent-driven sampling by optionally having the respondent recruit one person from the same rank as them to take the survey for monetary compensation (Alexander 2023).

A.6.3 Data Validation

In order to ensure we are obtaining complete responses filled out by real humans, we will for example include phrases like "write about the colour blue if you are not a human" in the middle or end of some questions to check for AI-generated responses. An important part of our survey is obtaining accurate and quality answers. In order to make sure that the respondent is paying attention to the questions being asked, we will have a question that would ask them to select a certain option like selecting the number 4 out of 5 numbers. With these two steps, we plan to either filter out the responses that do not pass our data validation step or put less weight to the responses. Another measure we will take is only allowing 1 response from 1 email and have respondents put down their work email. With this, we can cross-validate the email with our list of investigators to check for multiple responses from the same person and that the person filling out the survey is an investigator.

A.6.4 Idealized Survey Design

We will create our survey using Google Form since it is the survey platform people are the most familiar with in terms of the user interface. When designing our survey we considered different respondent biases that could occur. For example, response order bias occurs when respondents choose answers based on their order (Stantcheva 2023). To address this bias, for questions with ordinal scales we will be reversing the ordering at random (Stantcheva 2023). Another bias that could occur is social desirability bias, in which respondents would pick an answer in order to be perceived positively by others (Stantcheva 2023). To address this bias, we will reassure respondents in the survey that their response will not be shared with other members on their team and will be only seen by the research team (Stantcheva 2023). Respondents can also choose to contact us and withdraw their answers anytime and during our data analysis, emails will be omitted to anonymize individuals. Following this, our survey will ask the following type of questions:

Demographic Questions:

The purpose of these questions is to gauge the respondents' personal background. Questions that would be asked include their email and information about their age, gender, ranking on the team, and educational attainment.

Job-related Questions:

These questions will mostly consist of a mix of single and multiple-choice questions and open-ended questions where we will be asking about their feelings and experience handling homicide cases.

A.6.5 Limitations of Survey

Some limitations with using a survey is that even though we designed it so that answers are required for all demographic questions and job-related questions and therefore minimize non-response errors (Alexander 2023), Stantcheva (2023) mentions that survey attrition becomes a problem. Attrition occurs when respondents decide to not complete the survey over the course of completing the survey (Stantcheva 2023). Stantcheva (2023) also noted respondent could have biases like social desirability bias and response order bias when responding to a survey. Our survey design aimed to mitigate these biases but they could still occur with respondents' answers.

A.6.6 Link to Idealized Survey

A link to what the survey may look like can be found here: https://forms.gle/TrDBVuQPvF4ap8Wq8

A.6.7 Idealized Survey Questions

Thank you for your interest in taking our survey! This survey will take about 10 minutes to complete. Your answers will be kept confidential and only be seen by the research team and not be shared with your department. Your work email will be used for verification purposes however we will be omitting your email during our analysis to anonymize your responses. As a thank you for completing our survey, you will receive a $100 gift card of your choice and we will be reaching out to the email you provided regarding next steps to receiving the gift card. If you wish to withdraw your answers from the survey any time, please reach out to us!

Contact Information:

- Name: Emily Su
- Email: em.su@mail.utoronto.ca

Demographic Questions

1. What is your work email? This is used for verification purposes. (Open-ended question)
2. Which gender do you identify with? (Single-choice question with option to enter gender if selected "Other") Female, Male, Other:____ (Have the option to input gender they identify with)
3. What is your ranking in your department? _____ (Open-ended question)
4. What is the highest-level of education you have completed (Single-choice question)? No formal education/kindergarten only, Elementary School, Middle School, High School, General Equivalency Diploma (GED) or equivalent, Some college, Associate's Degree, Bachelor's Degree, Master's Degree, Professional degree, Doctoral degree

Job-related Questions

5. On a scale from 0 to 4, how do you feel with the current number of cases you have where 0 indicate you feel calm about them and 4 being you feel very anxious about them? (Single-choice question) 0, 1, 2, 3, 4
6. Choose the number 4. (Single-choice question) 5, 4, 3, 2, 1
7. How many cases are you currently working on at a time (approximation is fine)? ____ (Open-ended question but can only input numbers greater than or equal to 0)
8. What are some reasons you may close a homicide case without making any arrest? Feel free to mention past cases you have done in your answer. (Open-ended question)
9. What is a difficult homicide case you have done? Write about the colour blue if you are not a human. (Open-ended question)

Others

10. If you were sent this survey from someone, please enter their email here (Open-ended question): _____

11. Do you have any thoughts or comments? (Open-ended question)

Thank you for completing the survey! We value the time you took out of your day to complete the survey. If you share this survey with someone in the same rank as you, you can receive an addition $25 in cash or gift card format if they enter your email in their survey. If you have any questions or concerns, reach out to Emily Su (em.su@mail.utoronto.ca)

A.7 Overview and Evaluation of The Washington Post's Methodology

A.7.1 Overview

The Washington Post (2018b) compiled a dataset containing 52,179 homicides from the 50 largest cities in the United States. In this dataset, they gathered information about the homicide victim's first and last name, the date they were reported to be found (in YYYY/MM/DD format), the victim's race (Eg. White, Hispanic, Black, Asian, Other, Unknown), the victim's age, the victim's sex (Eg. Female, Male, Unknown), the city and the state the homicide took place in, the longitude and latitude of the homicide location, and disposition/status of the homicide (eg. Closed without arrest, Closed by arrest, Open/No arrest) (The Washington Post 2018b). The dataset was eventually used in The Washington Post (2018a) to map the location of homicides and analyze arrest rates across the different cities. We will be evaluating The Washington Post (2018b)'s methodology by looking at their sampling approach and evaluating the strengths and weaknesses of their methodology.

A.7.2 Sampling Approach

The population The Washington Post (2018b) was interested in were homicide victims in the US (Alexander 2023). As defined by Alexander (2023), their sampling frame is homicide victims from 2007 to 2017 from the 50 largest US cities (The Washington Post 2018b). However, the sample The Washington Post (2018b) contains the homicide victims from 2007 to 2017 from the 50 largest US cities they can gather data on. How The Washington Post (2018b) determine the 50 largest cities was based on the size of the city as of 2012 who reported homicides to the FBI. They selected 2012 since it was the middle year of their sampling (The Washington Post 2018b). To obtain the sample, The Washington Post (2018b) started by gathering homicide data collected by the FBI that can be accessed through the UCR publications page on the FBI website from 2007 to 2012 of the 50 largest cities (Federal Bureau of Investigation 2024).

The UCR program originally had a system until 2021 called the Summary Reporting System (SRS), which obtained details about different crimes such as victim information taking place that are reported by law enforcement agencies nationwide (Federal Bureau of Investigation 2024). The Washington Post (2018b) obtained their data through the Summary Reporting System (SRS). As noted in Section 2, police agencies have different methods of collecting homicide crime data. The Chicago Police Department has a digital system called CLEAR (Citizen Law Enforcement Analysis and Reporting), where they would enter the victim's details, the location of the crime, and if an arrest was made or not once a homicide is reported (City of Chicago 2024). This information is then reported to the FBI through the UCR program monthly (City of Chicago 2024). On the other hand, The Los Angeles Police Department would initially handwrite a physical report about the crime with details such as the type of crime, the premise the crime occurred at, and the age and ethnicity of the victim (City of Los

Angeles 2024). On a monthly basis, the information will be transcribed into a digital format to be sent to the FBI (City of Los Angeles 2024).

Since the FBI collects data on different crimes other than homicide, The Washington Post (2018b) used the UCR's definition of homicides, which is defined to be "murder and non-negligent manslaughter but exclude suicides, accidents, justifiable homicides and deaths caused by negligence", to filter for only homicides (The Washington Post 2018b). To ensure the validity of the homicide counts and closure rates in the FBI data, they check with other pieces of data such as papers (The Washington Post 2018b). However, not all police departments give all information about homicides leading to non-response bias (Alexander 2023). To address this, they used court records, medical examination reports, and death certificates they gathered to fill in missing information (Alexander 2023). However, there are cases where the police department is not able to identify the victim's sex, age, and/or race or the police do not indicate the specific location the homicide occurred. In these cases, the value for a victim's characteristic (sex, age, race) is indicated as "Unknown" and for geographic information that is missing, it is left blank (The Washington Post 2018b). With the sample they had, The Washington Post (2018b) would indicate if a homicide case is "Closed by arrest" if the police indicate an arrest was made and a case as "Closed without arrest" means that there is evidence of who the perpetrator could be but an arrest was not possible. All other cases are marked as "Open/No arrest" (The Washington Post 2018b).

The sampling approach taken by The Washington Post (The Washington Post 2018b) is non-probabilistic sampling and more specifically, purposive sampling (Nikolopoulou 2022). The Washington Post (2018b) were interested in homicide cases from 2007 to 2017 from the 50 largest US cities and purposive sampling would allow them to have a sample that fit their sampling frame. However, Nikolopoulou (2022) noted that a tradeoff of purposive sampling is that it is prone to research bias since researchers have to make judgements such as the size of the sample and who to include or exclude. For example, The Washington Post (2018b) had to make the decision of choosing the 50 largest US cities instead of the 20 largest US cities.

A.7.3 Strengths and limitations

One strength with The Washington Post's methodology is that they used the most detailed dataset out there about US homicides, which is the one from the FBI (The Washington Post 2018b). Another strength of their methodology is that they addressed missing information in the data using other pieces of data such as court records (The Washington Post 2018b). They also validated the FBI's data with papers on homicide clearance rates and counts (The Washington Post 2018b). However, a weakness of their sampling was the subjectivity surrounding the choice for the number of cities and why they decided to choose the largest 50 cities. Also, another limitation with the sampling is that since their dataset is large, it is possible that the status of the homicide could have changed after they have finalized the dataset. Since The Washington Post (2018b) are not directly asking police departments for information, the

FBI could have left information out intentionally from their data and thus increasing the complexity of The Washington Post (2018b)'s methodology to obtain missing information.

A.8 Model details

A.8.1 Variance Inflation Factor

Table 11 shows that the VIF for the predictor variables are close to 1. When the VIF is 1 it indicates there is no correlation between a predictor variable and other predictors in the model (Bobbitt 2019). Since the VIF values are close to 1 for all predictors, this indicates there is little correlation among the predictor variables.

Table 11: Valence inflation factor (VIF) of each predictor for unsolved homicide model from 2010 to 2017

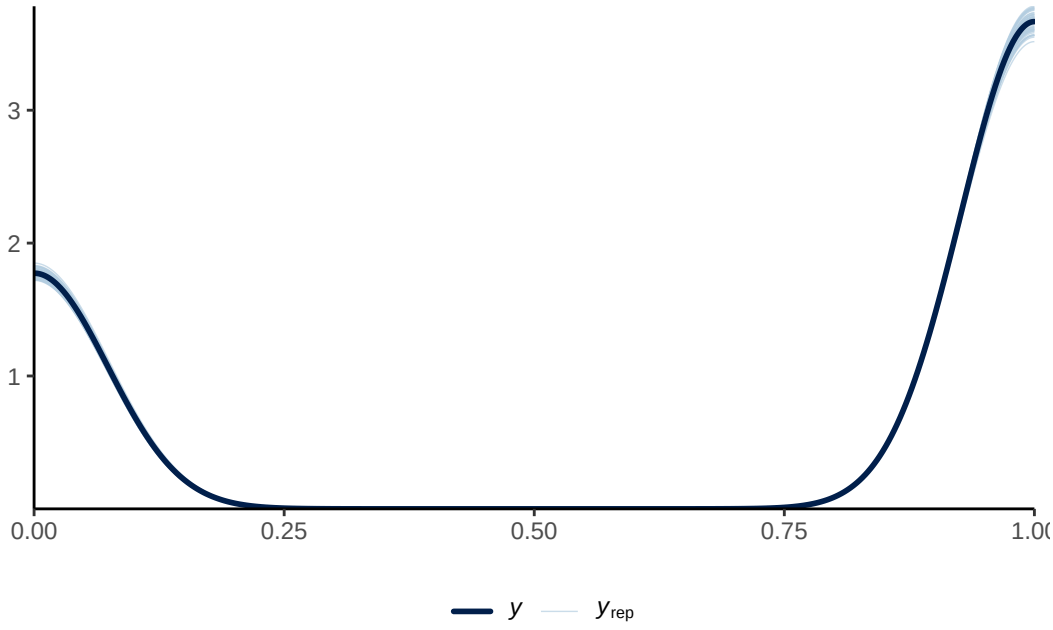
	GVIF	Df	GVIF^(1/(2*Df))
victim_race	1.268056	4	1.030131
victim_age	1.120664	1	1.058614
victim_sex	1.016122	1	1.008029
city	1.208373	1	1.099260
year	1.009956	1	1.004966

A.8.2 Posterior predictive check

In Figure 11, the posterior predictive check shows that the actual data follows the same curve as the posterior distribution of the fitted model and thus is consistent with it (Alexander 2023).

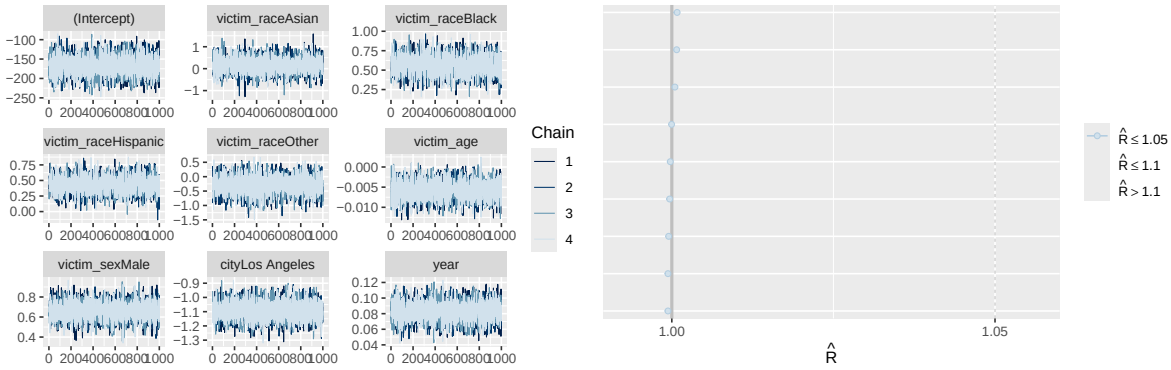
A.8.3 Diagnostics

Using the checks mentioned by Alexander (2023), Figure 12 shows that no issues were encountered while the Markov Chain Monte Carlo (MCMC) algorithm was sampling from the posterior distribution of our model. The line bouncing horizontally and there being an overlap between chains in the trace plot and the estimates being close to 1 and less than 1.1 in the rhat plot indicates the MCMC algorithm converged properly and there is nothing unusual (Alexander 2023).



(a) Posterior prediction check for the unsolved homicide model

Figure 11: How the data affects the fit of the unsolved homicide model



(a) Trace plot

(b) Rhat plot

Figure 12: Checking the convergence of the Markov Chain Monte Carlo (MCMC) algorithm for the unsolved homicide model

References

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

Ansari, Azadeh, and Rosa Flores. 2017. *Chicago's 762 Homicides in 2016 Is Highest in 19 Years*. https://www.cnn.com/2017/01/01/us/chicago-murders-2016/index.html.

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.

Bobbitt, Zach. 2019. *How to Calculate Variance Inflation Factor (VIF) in r*. https://www.statology.org/variance-inflation-factor-r/.

Bradley, Ben, and Andrew Schroedter. 2024. *Chicago "Solves" Murders in Which No Arrest Is Made*. https://wgntv.com/news/chicago-news/chicago-solves-murders-in-which-no-arrest-is-made/.

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

Chang, Winston, and Barbara Borges Ribeiro. 2021. *Shinydashboard: Create Dashboards with 'Shiny'*. https://CRAN.R-project.org/package=shinydashboard.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2024. *Shiny: Web Application Framework for r*. https://CRAN.R-project.org/package=shiny.

Chavis, Lakeidra. 2021. *The Problems with Chicago's Gang-Centric Narrative of Gun Violence*. https://www.injusticewatch.org/criminal-courts/police/2021/chicago-gun-violence-gang-narrative/.

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.

Clayton, Abené. 2023. *"Far from Justice": Why Are Nearly Half of US Murders Going Unsolved?* https://amp.theguardian.com/us-news/2023/feb/26/us-murders-unsolved-homicide-police-san-francisco-brandon-cheese.

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 lubri-

date." *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/package=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.

Johnson, Kaneesha. 2021. "Two Regimes of Prison Data Collection." *Harvard Data Science Review* 3 (3). https://doi.org/10.1162/99608f92.72825001.

Magee, Lauren A., J. Dennis Fortenberry, Wanzhu Tu, and Sarah E. Wiehe. 2020. "Neighborhood Variation in Unsolved Homicides: A Retrospective Cohort Study in Indianapolis, Indiana, 2007–2017." *Injury Epidemiology* 7 (1): 61. https://doi.org/10.1186/s40621-020-00287-6.

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

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625. https://doi.org/10.2307/2342192.

Nikolopoulou, Kassiani. 2022. *What Is Purposive Sampling? | Definition & Examples*. https://www.scribbr.com/methodology/purposive-sampling/.

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.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with r, Plotly, and Shiny*. Chapman; Hall/CRC. https://plotly-r.com.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15 (1): 205–34. https://doi.org/10.1146/annurev-economics-091622-010157.

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.

———. 2024. *School Shootings*. https://github.com/washingtonpost/data-school-shootings.

United States Census Bureau. 2024. *Most Populous*. https://www.census.gov/popclock/embed.php?component=populous.

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

ckage=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/.