My title*

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

Emily Su

December 1, 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 2 of the largest cities in the United States, New York 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

For the data analysis portion of our paper, we used the statistical programming language R (R Core Team 2024),

We use the statistical programming language R (R Core Team 2024)…. Our data (The Washington Post 2018)…. Following Alexander (2023), we consider…

Overview text

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

Limitation of dataset. - There's only data available from 2010 onwards for both New York and Los Angeles - Not all victims were able to be identified and they were removed from the dataset during data cleaning

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**), from (**palmerpenguins?**).

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.4 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

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.7 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 "New York" 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.7.

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.

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

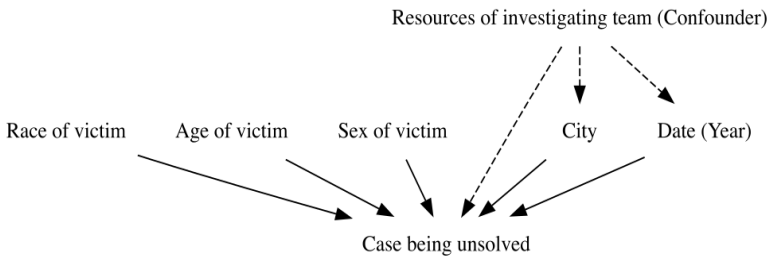


Figure 1: Causal relationship between homicide case information and homicide case being unsolved

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 New York and Los Angeles (2010 to 2017)

4.1.1 Date (Month and Year)

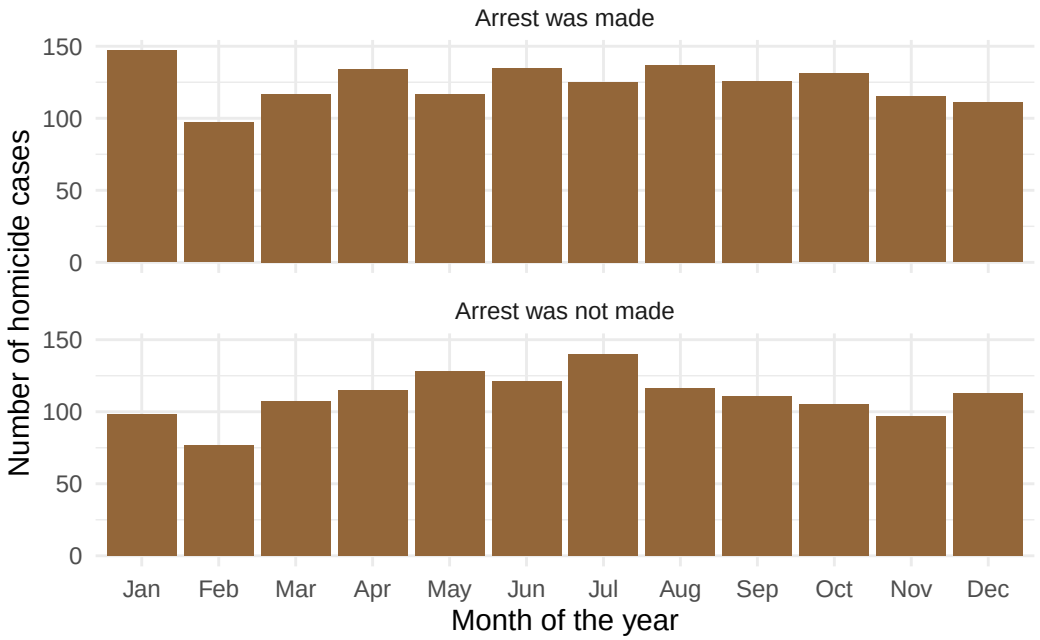


Figure 2: Number of solved and unsolved homicides across the 12 months of a year in Los Angeles and New York (2010 to 2017)

- TODO: Add in summary statistics table

4.1.2 City

Table 1: Proportion and number of solved and unsolved homicides in Los Angeles and New York (2010 to 2017)

City	Status of the homicide case	Number of cases	Proportion of cases
Los Angeles	Arrest was made	1109	0.51
Los Angeles	Arrest was not made	1087	0.49
New York	Arrest was made	383	0.61

Table 1: Proportion and number of solved and unsolved homicides in Los Angeles and New York (2010 to 2017)

City	Status of the homicide case	Number of cases	Proportion of cases
New York	Arrest was not made	241	0.39

4.1.3 Disposition

Table 2: Disposition of homicide cases in New York and Los Angeles (2010 to 2017)

City	Disposition of the homicide case	Number of cases
Los Angeles	Closed by arrest	1109
Los Angeles	Open/No arrest	1087
New York	Closed by arrest	383
New York	Closed without arrest	17
New York	Open/No arrest	224

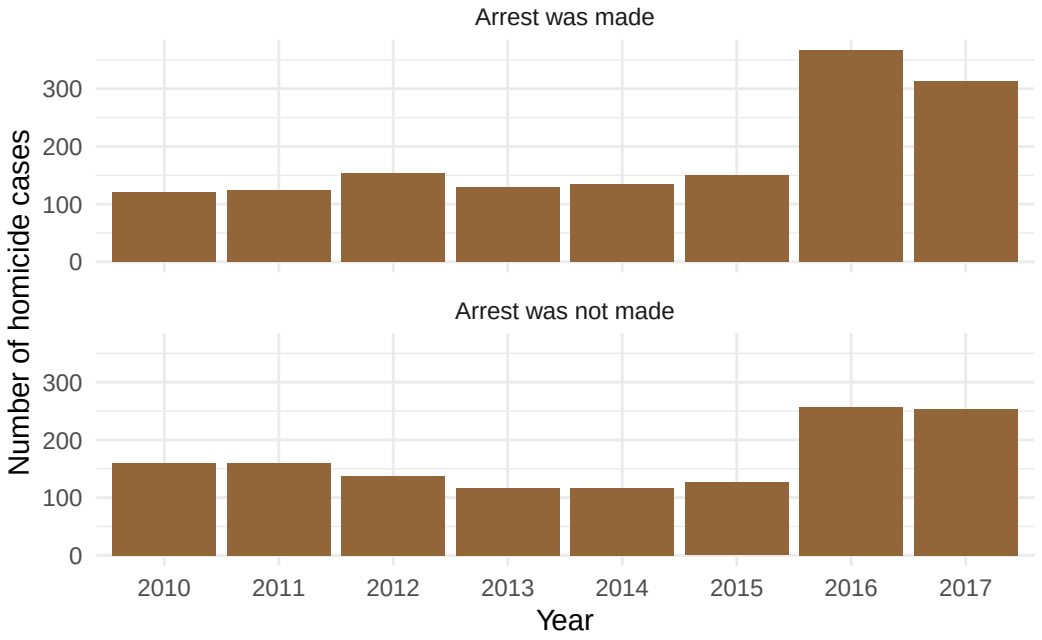


Figure 3: Number of solved and unsolved homicides from 2010 to 2017 in Los Angeles and New York

4.1.4 Victim's Age

4.1.5 Victim's Sex

Table 3: Proportion and number of homicide cases per sex in New York and Los Angeles (2010 to 2017)

Victim's sex	Status of the homicide case	Number of cases	Proportion of cases
Female	Arrest was made	269	0.67
Female	Arrest was not made	135	0.33
Male	Arrest was made	1223	0.51
Male	Arrest was not made	1193	0.49

4.1.6 Victim's Race

Table 4: Number of homicide cases per sex in New York and Los Angeles (2010 to 2017)

Victim's race	Status of the homicide case	Number of cases
White	Arrest was made	164
White	Arrest was not made	84
Asian	Arrest was made	41
Asian	Arrest was not made	17
Black	Arrest was made	604
Black	Arrest was not made	612
Hispanic	Arrest was made	645
Hispanic	Arrest was not made	597
Other	Arrest was made	38
Other	Arrest was not made	18

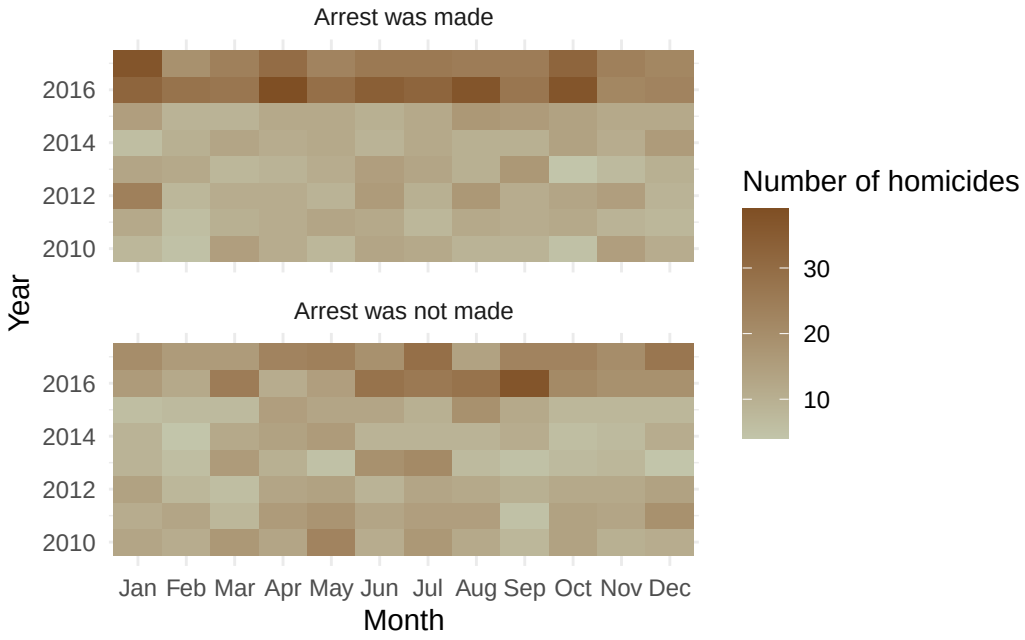


Figure 4: Number of solved and unsolved homicides from January to December from 2010 to 2017 in Los Angeles and New York

4.2 Model Results

TODO: - Mention the software used to implement the model, and provide evidence of model validation and checking—such as out-of-sample testing, RMSE calculations, test/training splits, or sensitivity analyses—addressing model convergence and diagnostics (although much of the detail make be in the appendix). RSME calculation

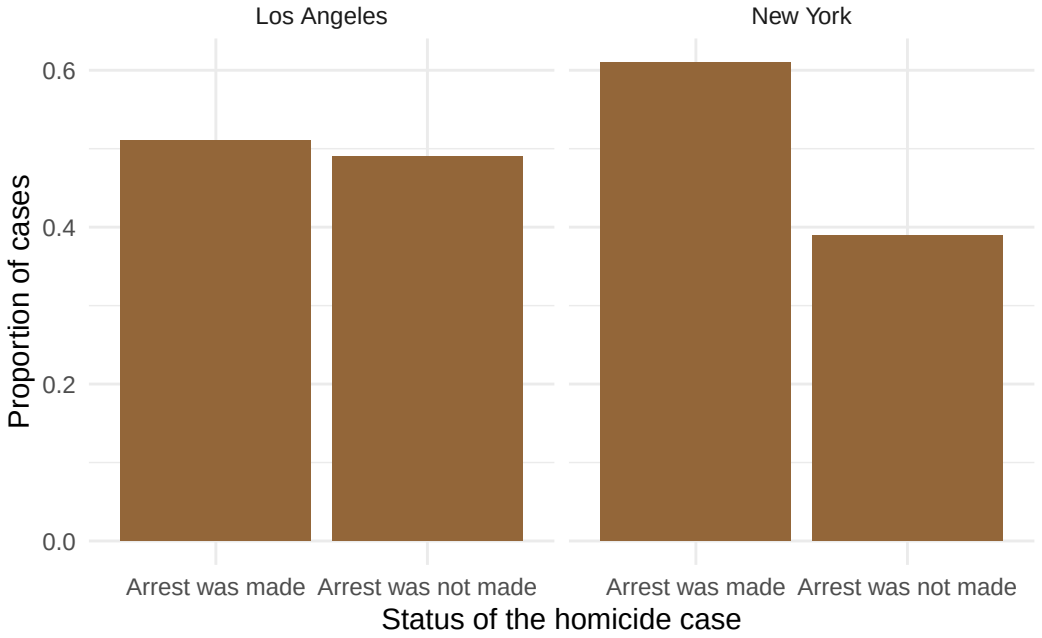


Figure 5: Proportion of solved and unsolved homicides in Los Angeles and New York (2010 to 2017)

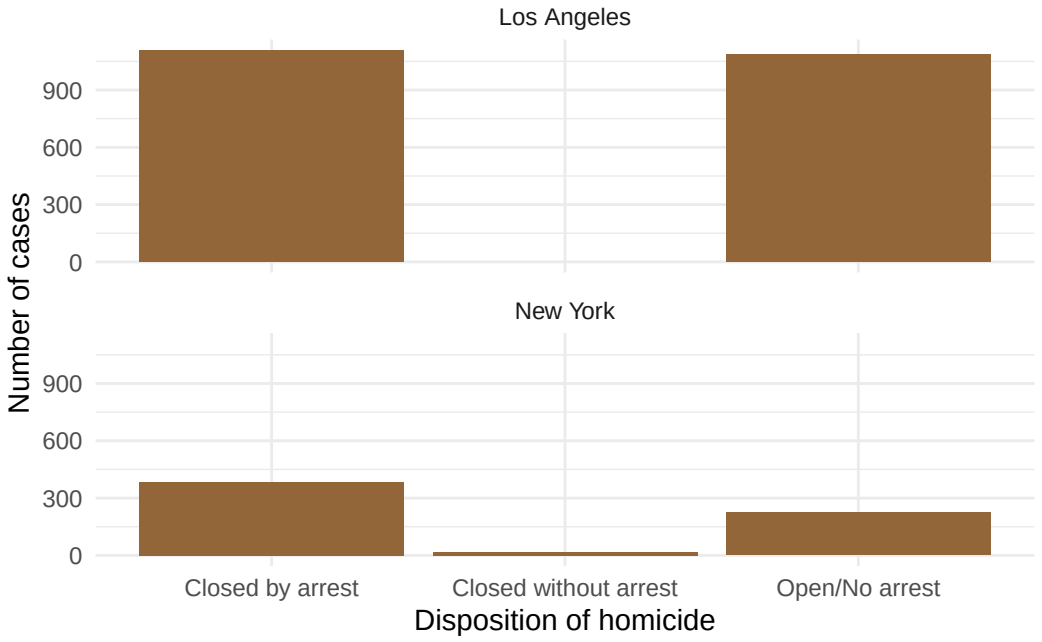


Figure 6: Disposition of homicide cases in New York and Los Angeles (2010 to 2017)

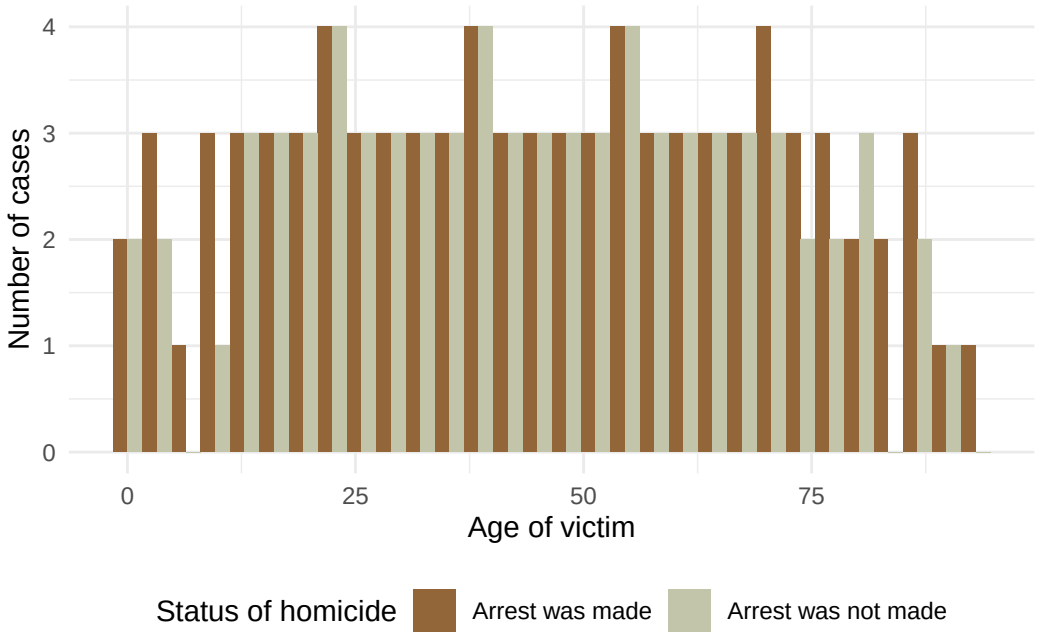


Figure 7: Distribution of victim's age in solved and unsolved homicides in New York and Los Angeles (2010 to 2017)

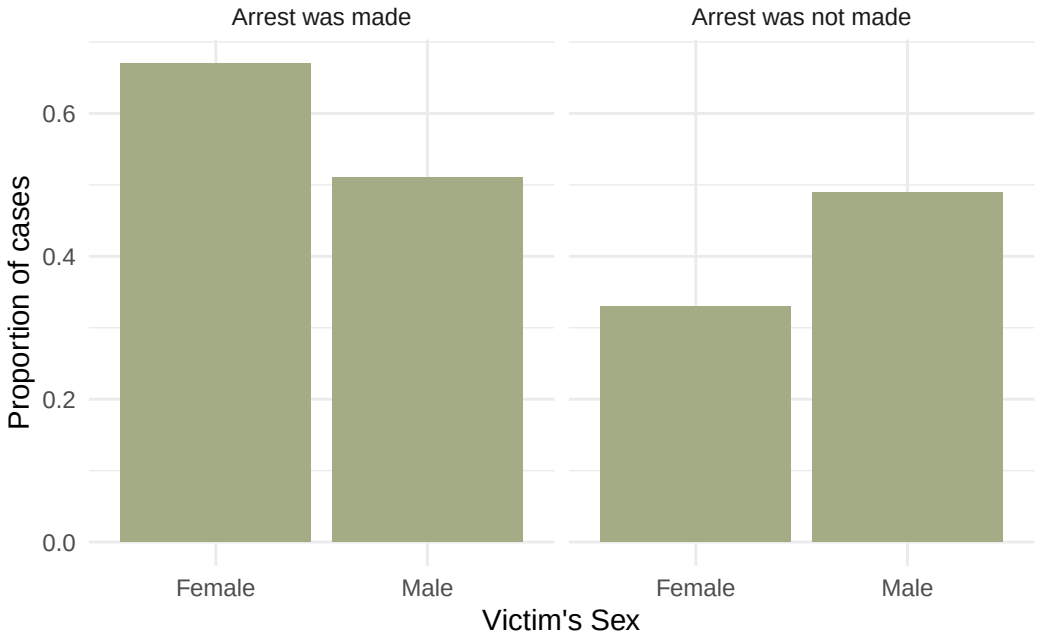


Figure 8: Proportion of homicide cases per sex in New York and Los Angeles (2010 to 2017)

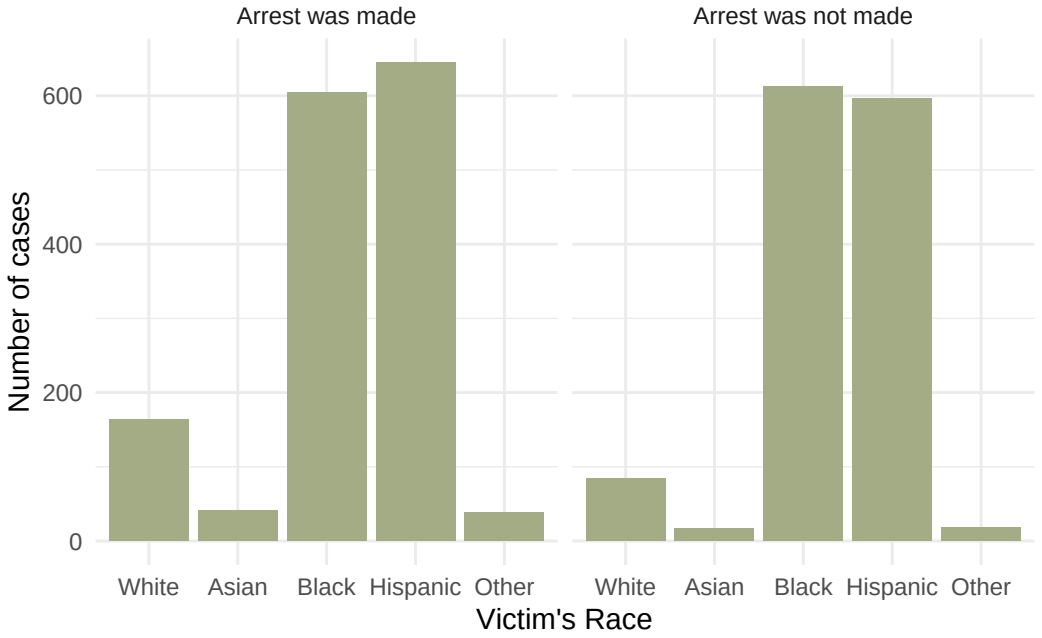


Figure 9: Number of homicide cases per race in New York and Los Angeles (2010 to 2017)

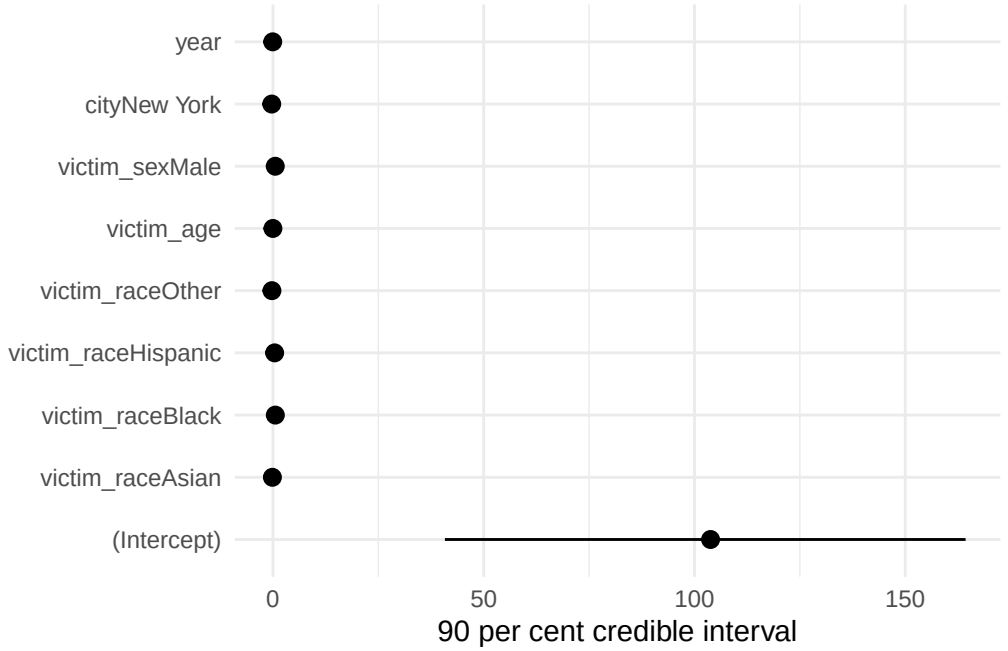


Figure 10: The credible intervals (line) for coefficient estimates (dot) of predictor variables for homicides that go unsolved from 2010 to 2017.

Table 5: 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)	103.817
	(36.659)
victim_raceAsian	−0.125
	(0.325)
victim_raceBlack	0.570
	(0.154)
victim_raceHispanic	0.406
	(0.154)
victim_raceOther	−0.239
	(0.317)
victim_age	−0.004
	(0.003)
victim_sexMale	0.548
	(0.116)
cityNew York	−0.282
	(0.111)
year	−0.052
	(0.018)
Num.Obs.	2820
R2	0.035
Log.Lik.	−1902.792
ELPD	−1912.0
ELPD s.e.	10.1
LOOIC	3824.0
LOOIC s.e.	20.1
WAIC	3824.0
RMSE	0.49

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 6: Number of solved and unsolved homicides across the 12 months of a year in Los Angeles and New York (2010 to 2017)

Status of the homicide case	Month	Number of cases in the month
Arrest was made	Jan	147
Arrest was made	Feb	97
Arrest was made	Mar	117
Arrest was made	Apr	134
Arrest was made	May	117
Arrest was made	Jun	135
Arrest was made	Jul	125
Arrest was made	Aug	137
Arrest was made	Sep	126
Arrest was made	Oct	131
Arrest was made	Nov	115
Arrest was made	Dec	111
Arrest was not made	Jan	98
Arrest was not made	Feb	77
Arrest was not made	Mar	107
Arrest was not made	Apr	115

Table 6: Number of solved and unsolved homicides across the 12 months of a year in Los Angeles and New York (2010 to 2017)

Status of the homicide case	Month	Number of cases in the month
Arrest was not made	May	128
Arrest was not made	Jun	121
Arrest was not made	Jul	140
Arrest was not made	Aug	116
Arrest was not made	Sep	111
Arrest was not made	Oct	105
Arrest was not made	Nov	97
Arrest was not made	Dec	113

Table 7: Number of solved and unsolved homicides from 2010 to 2017 in Los Angeles and New York

Status of the homicide case	Year	Number of cases in the year
Arrest was made	2010	121
Arrest was made	2011	124
Arrest was made	2012	154
Arrest was made	2013	129
Arrest was made	2014	134
Arrest was made	2015	150
Arrest was made	2016	367
Arrest was made	2017	313
Arrest was not made	2010	160
Arrest was not made	2011	160
Arrest was not made	2012	137
Arrest was not made	2013	117
Arrest was not made	2014	117
Arrest was not made	2015	126
Arrest was not made	2016	257
Arrest was not made	2017	254

A.5 Idealized Survey and Methodology

- Link to literature

A.5.1 Idealized Survey Objectives

A.5.2 Sampling Approach

A.5.3 Respondent Recruitment

A.5.4 Data Validation

A.5.5 Idealized Survey Design

A.5.6 Link to Idealized Survey

- Using Google Forms

A.5.7 Limitations

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 Overview and Evaluation of The Washington Post's Dataset

- TODO: Make sure to link evaluation to literature

A.6.1 Overview

A.6.2 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;

A.6.3 Strengths and limitations

- what is good and bad about the sampling.

A.7 Model details

A.7.1 Variance Inflation Factor

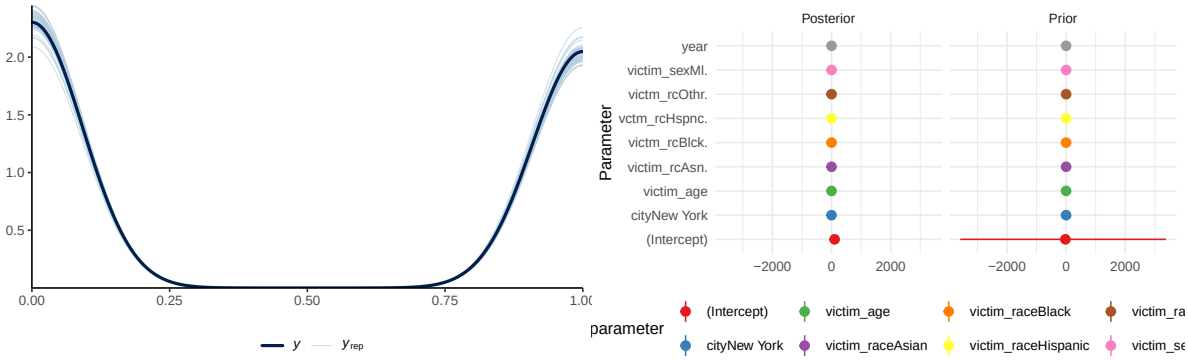
Table 8: Valence inflation factor (VIF) of each predictor for unsolved homicide model from 2010 to 2017

	GVIF	Df	GVIF^(1/(2*Df))
victim_race	1.146735	4	1.017262
victim_age	1.080917	1	1.039672
victim_sex	1.027160	1	1.013489
city	1.415212	1	1.189627
year	1.365517	1	1.168553

A.7.2 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

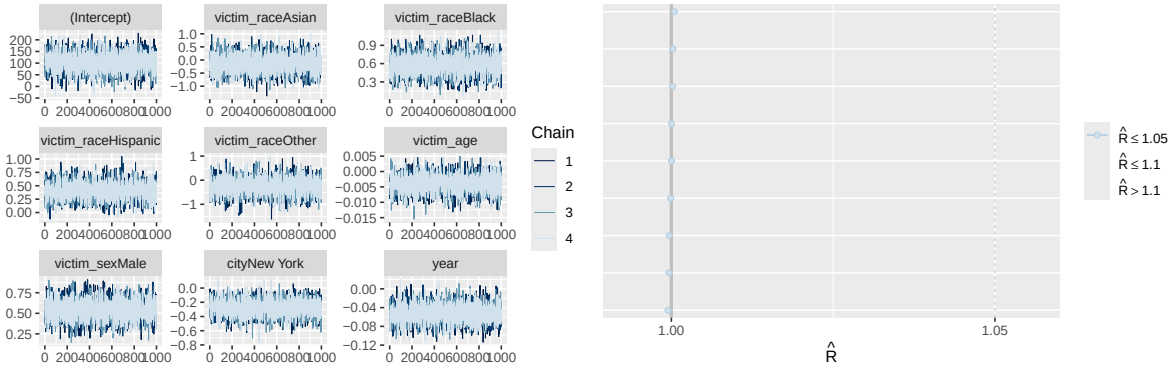
(b) Comparing the posterior with the prior

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

A.7.3 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/.

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

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

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.

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

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

The Washington Post. 2018. *Unsolved Homicide Database*. https://github.com/washingtonpost/data-homicides.