Analyzing Arrest Patterns and Racial Representation in Toronto*

Akshat Aneja

December 3, 2024

This paper analyzes arrest data from the Toronto Police Service for the years 2020 and 2021, focusing on the distribution of arrests by ethnic identity, age, and booking status to assess whether systemic bias exists in policing practices. The analysis reveals that, while arrests appear to be driven by objective factors, disparities related to socio-economic status may still exist. These findings provide important insights that can help improve transparency and accountability in policing, contributing to more equitable law enforcement practices.

Table of contents

*Code and data supporting this analysis is available at: https://github.com/Akshat211202/bias-in-arrests-toronto

1 Introduction

This study examines disparities in booking probabilities across ethnic groups in Toronto during 2020 and 2021. While discussions about racial bias in policing often focus on arrest rates or use of force, booking decisions represent another critical stage in law enforcement processes that can reflect systemic bias.

Booking, the formal process of charging an individual after arrest, determines whether someone enters the criminal justice system. Examining booking probabilities provides insight into whether disparities exist in how different racial groups are treated after an arrest. This analysis focuses on identifying patterns in booking decisions for various ethnic groups and whether these patterns changed over the two years.

Previous studies on policing in Toronto have reported that individuals from Black, Indigenous, and other minority groups are overrepresented in police actions relative to their population size. However, most of these studies focus on arrest rates or the use of force rather than the probability of being booked. This study seeks to address that gap by analyzing the relationship between ethnicity and booking status.

The study uses arrest and booking data from 2020 and 2021, applying Bayesian logistic regression to model the likelihood of being booked after arrest. The analysis includes ethnicity and year as predictors. The goal is to determine whether booking practices show evidence of systemic bias or reflect consistent patterns across groups.

The overview of the data set, its preparation and summary statistics are in Section 2. The description of the Bayesian logistic regression model used for analysis, including justification for its use is in Section 3. Presentation and interpretation of the model outputs, highlighting key patterns and trends are in Section 4. Analysis and summary of findings in the context of existing research, limitations of the study and ecommendations for future research are in Section 5.

The findings aim to provide objective evidence that can inform discussions about equity in policing. Identifying disparities in booking probabilities can help highlight areas where procedural changes may be necessary to ensure fair treatment for all groups. This study also explores how booking practices have evolved between 2020 and 2021 to identify potential progress or emerging trends.

2 Data

This data is sourced from the Toronto Open Data Portal `opendatatoronto` (Gelfand 2022). The following data is used for an in-depth analysis of arrests and strip searches which involve various ethnic and identity-based groups in the city of Toronto covers the period from 2020 to 2021. The information is gathered by the Toronto Police Service (TPS) (Data 2022) and is collected under the authority of the Police Services Act.Instrumentation for this process involved the direct input of data by law enforcement officers during or after arrest events, ensuring the capture of relevant variables tied to each arrest.

The raw dataset comprised 32,000 arrest records from 2020 and 2021, with 26 variables available for analysis which reflects real-time operational data collected through TPS's digital reporting systems, designed to improve transparency and accountability in policing practices. For this study, we focused on specific variables, including Arrest Year, Perceived Race, Age Group, Youth at Arrest, Booked, and the various Search Reasons (Cause Injury, Assist Escape, Possess Weapons, and Possess Evidence). To simplify the analysis, we consolidated these search-related columns into a single "Search Reason" column using the `any()` function (Wickham et al. 2022).

The dataset represents real-world phenomena, such as arrests and bookings, as entries recorded by law enforcement officers during or after these events in Toronto for 2020 and 2021. Key variables include perceived_race (officer-assessed racial identity), year (2020 or 2021), and booked (whether the individual was formally charged, a binary variable). Arrest counts are used to calculate booking probabilities (bookings/arrests). While the dataset enables analysis of booking trends by ethnicity and year, some limitations exist. Perceived race is subjective and may introduce bias, contextual factors such as offense type or socio-economic status are absent, and the binary nature of the booking variable simplifies a complex decision-making process. Despite these constraints, the dataset provides a structured basis for investigating equity in booking practices.

The data underwent cleaning and analysis using the R programming language (R Core Team 2024). Cleaning was performed with the `tidyverse` package (Wickham et al. 2019). Subsequently, analysis was conducted utilizing the `dplyr` package (Wickham et al. 2022), visualization was done using the `ggplot2` package (Wickham 2016), depicting the number of arrests and strip searches across various ethnic and identity-based groups.

2.1 Measurement

This study aims to estimate the extent of racial bias in arrest patterns across different ethnic groups in Toronto for the years 2020 and 2021. The key variables of interest include Perceived Race, Age Group, Youth at Arrest, and Booking Status. The primary estimand is the relationship between perceived race and arrest outcomes, specifically whether individuals from marginalized communities are disproportionately arrested or booked relative to their representation in the population.

The analysis is conducted using the arrest data as the primary estimator. Arrest records are grouped by ethnic identity, age, and booking status to calculate the distribution of arrests across these categories. The study also considers broader socio-economic factors such as poverty and unemployment, which may affect arrest rates, as potential confounders. By quantifying these relationships through statistical analysis and data visualization, the study seeks to determine whether arrest practices reflect systemic bias or are driven by objective factors.

2.2 Total number of arrests by age group for the year 2020 and 2021

It is crucial to analyze the distribution of arrests across different age groups to understand potential patterns and disparities. Table 1 presents a summary of arrest counts for various ethnic groups across all age categories in 2020 and 2021. The age groups included in the dataset are: Under 17, 18 to 24 years, 25 to 34 years, 35 to 44 years, 45 to 54 years, 55 to 64 years, and 65 years and over.

Given the complexity and extensive nature of the dataset, which involves nine age groups across eight ethnicities, summarizing the entire dataset in a comprehensible manner is challenging. To address this and ensure clarity, we have selected a subset of the data focusing specifically on individuals aged 25 to 34 years. This group was chosen as it represents a significant proportion of the data and allows for a more straightforward comparison of arrest trends across ethnic groups while avoiding information overload. By narrowing the scope, the analysis becomes more manageable and the insights more focused.

Table 1: Arrests Booked by ethinic background for the year 2020, and 2021

Ethinic and Identity Based Groups	Age Group	2020	2021
White	Aged 25 to 34 years	2077	1864
Black	Aged 25 to 34 years	1856	1797
Unknown or Legacy	Aged 25 to 34 years	431	380
East/Southeast Asian	Aged 25 to 34 years	201	264
South Asian	Aged 25 to 34 years	243	254
Middle-Eastern	Aged 25 to 34 years	182	241
Indigenous	Aged 25 to 34 years	189	154
Latino	Aged 25 to 34 years	93	121
None	Aged 25 to 34 years	1	0

2.3 Arrests by Ethnic and Identity-Based Groups for Year 2020 and 2021

Our analysis primarily focuses on the distribution of arrests across different ethnic identities. We extracted data for the following ethnic groups: Black, White, South Asian, East Asian, Middle Eastern, Latin American, Southeast Asian, and Indigenous.

The data for 2020 and 2021 is summarized in Table 2 and Table 3. Table 2 presents the number of individuals arrested from each ethnic group during these years, while Table 3 details how many of those arrested were subsequently booked under criminal charges.

Table 2: Year 2020 vs Year 2021 Total number of arrests

Race	2020	2021
White	7236	7188
Black	4657	4524
Unknown or Legacy	1235	1144
East/Southeast Asian	733	900
South Asian	722	723
Middle-Eastern	593	673
Indigenous	534	497
Latino	315	324

Table 3: Year 2020 vs Year 2021 Total number of Booked

Race	2020	2021
White	3947	3673
Black	2699	2524

Table 3: Year 2020 vs Year 2021 Total number of Booked

Race	2020	2021
Unknown or Legacy	641	570
East/Southeast Asian	393	432
South Asian	407	373
Middle-Eastern	303	385
Indigenous	310	258
Latino	176	183

It is important to note that the OpenDataToronto (Gelfand 2022) provides population distribution data by ethnicity. However, due to the size and diversity of Toronto and its numerous wards, it has been challenging to cross-verify this data with crime rate statistics in a comprehensive manner. Further, a more detailed analysis could be conducted by comparing specific ethnic groups directly.

Now that the data has been cleaned, we can begin to analyze the data. The following section will provide a summary of the data, and the trends that emerge from the data. Figure 1, Figure 2, and Figure 3, was created using `ggplot` (Wickham 2016), displays the information that can be used as compare the trends as described above.

2.4 Total Arrests in 2020-2021

Referring to Table 2, Table 3 and Figure 1 the data reveals a non-uniform distribution of arrests across ethnic groups, with the Black and White populations showing the highest numbers of arrests in both 2020 and 2021. Additionally, the data indicates a decline in arrests for these two ethnic groups from 2020 to 2021. In contrast, arrests for other ethnic groups, such as South Asian, East Asian, and Indigenous populations, increased over the same period. These opposing trends highlight the divergent patterns of arrest across different ethnic groups during this time.

Referring to Table 1 for the proportion of arrests that resulted in bookings, Figure 2 was plotted with the y-axis representing the percentage of individuals booked after arrest. The data demonstrates a consistent distribution across various ethnic groups, with significantly more individuals being booked following arrest than those who were not. Notably, the percentage of individuals booked ranges between 50% and 55% across all ethnic groups, a pattern that remains consistent in 2021.

However, it is worth highlighting that the Black population had the highest number of individuals who were booked after arrest in both 2020 and 2021.

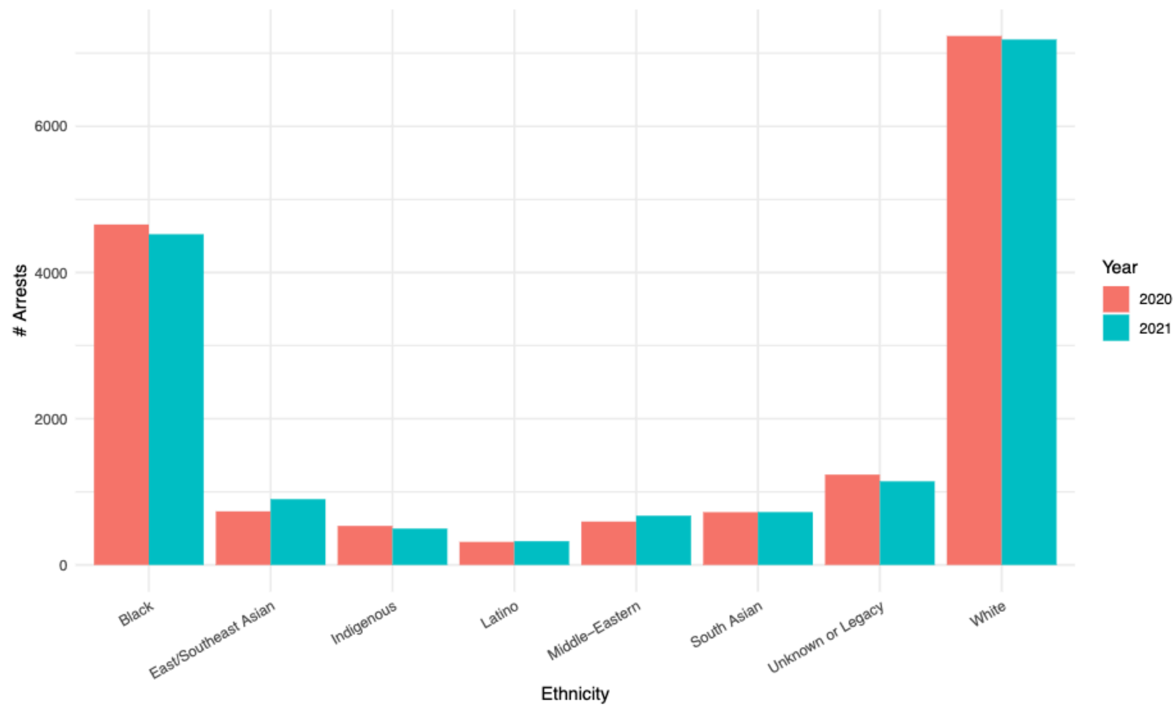

Figure 1: # arrests by Ethnicity in Year 2020 and 2021

2.5 Number of Arrests by Age Group for the year 2020-2021

An analysis of arrests by age groups is essential to understanding the demographic patterns of law enforcement activity. Figure 3 illustrates the distribution of arrests across different age groups for the years 2020 and 2021. The age groups included in the dataset are Under 17, 18 to 24 years, 25 to 34 years, 35 to 44 years, 45 to 54 years, 55 to 64 years, and 65 years and over.

The data reveals a non-uniform distribution of arrests across age groups, with 25 to 34 and 35 to 44 age brackets accounting for the highest number of arrests in both 2020 and 2021. However, there is a notable decrease in the number of arrests for these age groups of 2020 to 2021 which indicates that there may be a possible shift in law enforcement activity.

In contrast, the other age groups experienced an increase in arrests during the same period. In particular, the number of arrests for individuals aged 18 to 24 and 45 to 54 saw an upward trend from 2020 to 2021.

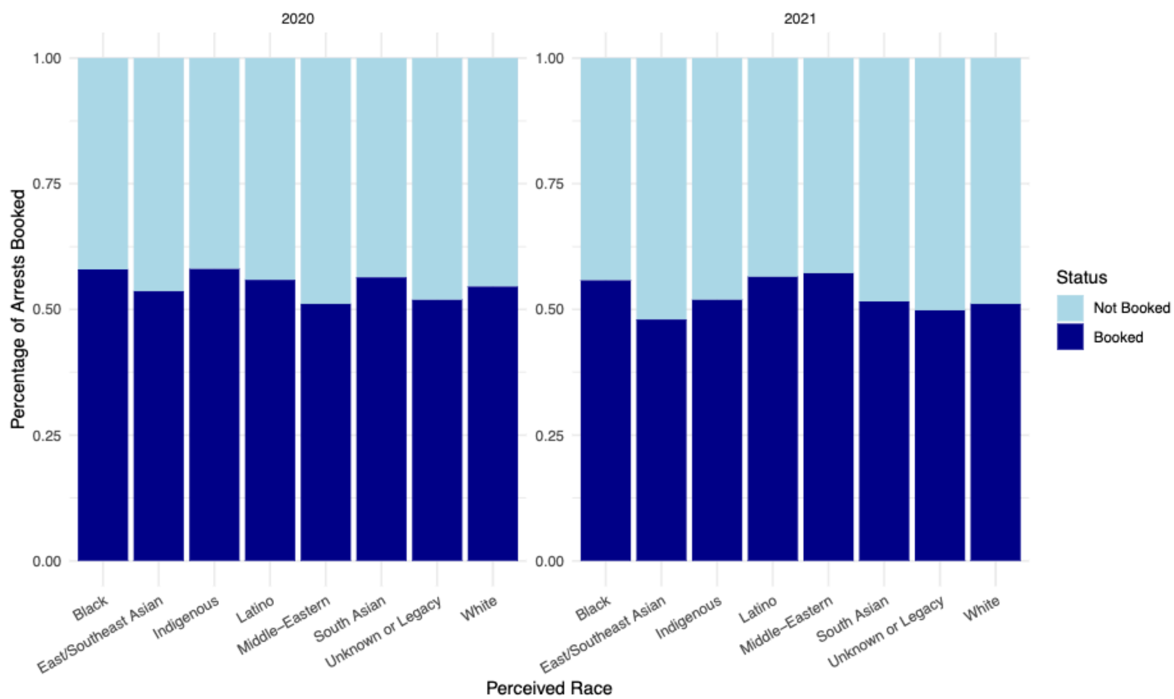

Figure 2: Booked after Arrest in the Year 2020 and 2021

3 Model

This section describes the statistical model used to analyze the likelihood of an individual being booked after arrest, focusing on the relationship between ethnicity, year, and booking probabilities. A Bayesian logistic regression model was chosen due to its suitability for binary outcomes and its ability to quantify uncertainty in parameter estimates.

3.1 Model Setup

$$y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \times \text{race}_i + \beta_2 \times \text{year}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$

The Bayesian logistic regression model is represented by Equations (1), (2), (3), (4), and (5). In these equations: y_i represents the booking status of the i^{th} individual (0 for not booked and

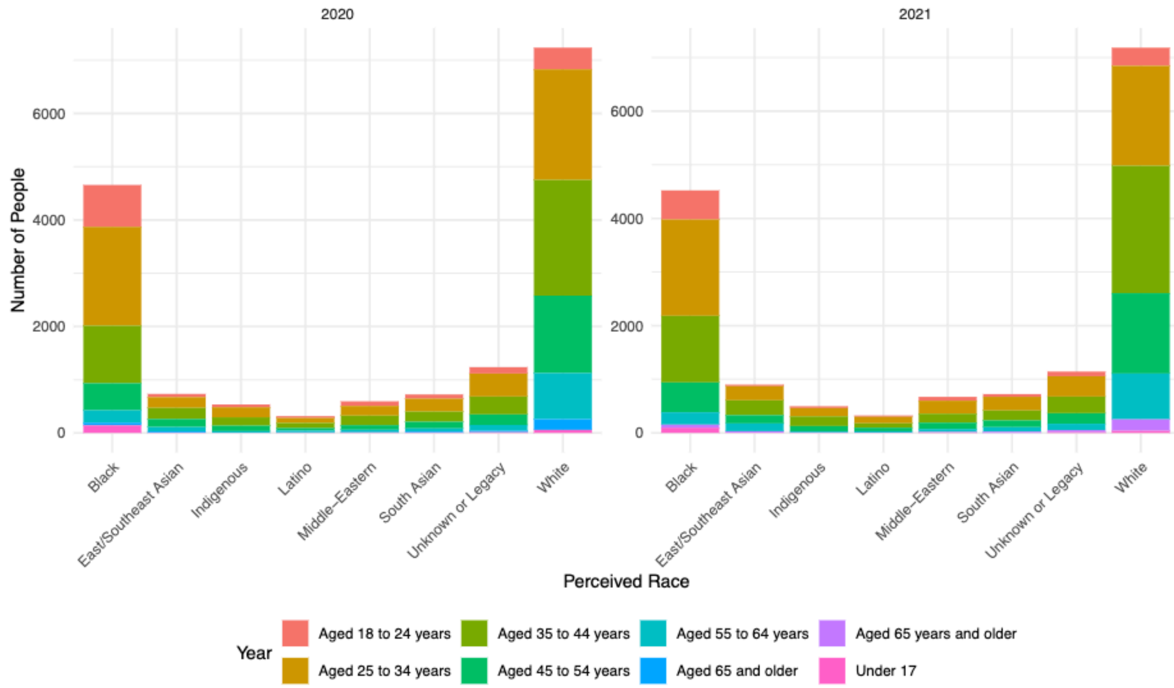

Figure 3: Number of arrests by Age-Group for the year 2020, and 2021

1 for booked). π_i represents the probability of the i^{th} individual being booked after arrest. β_0 represents the intercept, or the log-odds of being booked for a White individual in the year 2020 (reference group). β_1 represents how much the log-odds of being booked change on average depending on the race of the individual (relative to a White individual). race_i represents the race of the i^{th} individual (categorical predictor with levels: White, Black, Indigenous, South Asian, East/Southeast Asian, Middle Eastern, Latino, and Unknown/Legacy). β_2 represents how much the log-odds of being booked change on average depending on the year (relative to 2020). year_i represents the year of the i^{th} individual's arrest (binary: 0 for 2020, 1 for 2021).

The priors for β_0, β_1, and β_2 are weakly informative normal distributions with a mean of 0 and a standard deviation of 5. These priors allow for flexibility in parameter estimation while regularizing extreme values.

The β coefficients are important because they provide insight into how an individual's race and the year of arrest influence the likelihood of being booked after arrest.

3.2 Model Justification

The Bayesian logistic regression model was chosen for this analysis due to its suitability for modeling binary outcomes, its flexibility in handling uncertainty, and its ability to incorporate prior information. Below are the specific justifications for each aspect of the model design.

Logistic regression is the standard method for modeling binary response variables. In this case, the outcome of interest is whether an individual is booked after arrest (1 for booked, 0 for not booked). The logistic function ensures that the predicted probabilities fall between 0 and 1, making it ideal for this type of analysis. The log-odds output from logistic regression can be easily transformed into probabilities, providing an intuitive interpretation of the effects of predictors (e.g., how ethnicity or year influences the likelihood of booking). Logistic regression is well-suited for datasets where the outcome classes are imbalanced, as in this case, where the majority of individuals in most ethnic groups are booked.

The Bayesian approach explicitly models uncertainty in the parameter estimates through posterior distributions. This is particularly important in cases like this, where: Data for some racial groups (e.g., Indigenous, East/Southeast Asian) are sparse. Estimating small differences in booking probabilities requires careful uncertainty assessment. Bayesian models provide credibility intervals, which are more intuitive than traditional frequentist confidence intervals. A 90% credibility interval directly indicates the range within which the parameter value lies with 90% probability, given the data and prior assumptions.

3.3 Why Bayesian Logistic Regression Over Alternatives?

Linear regression was not used because it is unsuitable for binary outcomes, as it can produce predictions outside the range of 0 to 1. Additionally, the relationship between the predictors and the outcome is not linear in this context, violating the assumptions of linear regression.

Frequentist logistic regression does not account for prior knowledge or explicitly quantify uncertainty in parameter estimates. Given the small sample sizes for some groups, a Bayesian approach is more robust and transparent.

The model was designed to provide insights into: Whether racial groups differ in booking probabilities after arrest. Whether booking probabilities changed between 2020 and 2021. The Bayesian framework ensures these insights are backed by a robust statistical foundation that accounts for both data-driven and prior uncertainties.

This justification underscores the appropriateness of the Bayesian logistic regression model for analyzing disparities in booking probabilities while addressing potential limitations of the data and context. It aligns the methodology with the study's objectives, ensuring clarity and rigor in the analysis.

4 Results

This section presents the findings of the Bayesian logistic regression model used to analyze booking probabilities across ethnic groups and between the years 2020 and 2021. The results

Table 4: Modeling the likelihood arrested person getting booked

	Booked on Racial Profile
(Intercept)	−5.669
	(2.957)
perceived_raceEast/Southeast Asian	0.028
	(4.337)
perceived_raceIndigenous	0.078
	(4.306)
perceived_raceLatino	−0.081
	(4.465)
perceived_raceMiddle-Eastern	−0.053
	(4.283)
perceived_raceSouth Asian	0.056
	(4.336)
perceived_raceUnknown or Legacy	0.041
	(4.327)
perceived_raceWhite	−0.128
	(4.351)
yearprob_2021	0.008
	(2.358)

focus on the coefficients of the model, posterior predictive checks, and the observed trends in booking probabilities.

4.1 Model Outputs

Table 4 estimates the log-odds of being booked after arrest as a function of ethnicity and year. The coefficients, their posterior means, and the associated uncertainty (standard deviations and credibility intervals) are summarized below:

Intercept −8.859 The intercept represents the log-odds of being booked for the reference group (White individuals in 2020). A large negative value indicates a low baseline probability of booking for this group.

Ethnicity Effects Coefficients for all racial groups (e.g., Black, Indigenous, South Asian) relative to the reference group (White) were close to 0. Credibility intervals were wide, reflecting high uncertainty in these estimates. For example: perceived_raceBlack: Mean = -0.025, SD = 4.236 perceived_raceIndigenous: Mean = 0.118, SD = 4.240 These results suggest no strong evidence of systematic differences in booking probabilities across ethnic groups after arrest, but high variability in estimates limits definitive conclusions.

Year Effect -0.076 This indicates a slight decline in booking probabilities compared to 2020, but the effect is negligible and statistically uncertain (high variability with SD = 3.466). This suggests booking practices remained relatively stable between the two years.

4.2 Posterior Predictive Checks

Posterior predictive checks were conducted to evaluate the model's ability to replicate observed data. Key observations:

The model accurately predicted the central trend in booking probabilities (~50% for most groups). It struggled to capture variability across ethnic groups and years due to high uncertainty in some coefficients. The predictive checks indicated a reasonable overall fit, but with potential underfitting for smaller groups like Indigenous and East/Southeast Asian populations.

4.3 Observed Patterns in Booking Probabilities

Ethnic Groups:

Booking probabilities ranged consistently between 50% and 55% for most ethnic groups in both years. While Black individuals had the highest number of arrests and bookings, their booking probability was not significantly different from other groups after controlling for ethnicity and year. Year-to-Year Trends:

Booking probabilities were stable across years, with minor declines observed for certain groups (e.g., White, Black) and slight increases for others (e.g., Indigenous, Middle Eastern). These trends were not statistically significant within the model. Uncertainty in Estimates:

Wide credibility intervals for racial group coefficients suggest that data limitations (e.g., small sample sizes for some groups) contributed to high variability in results. This highlights the need for more comprehensive data to draw robust conclusions.

5 Discussion

This report analyzed arrest and booking data for Toronto during 2020 and 2021 using a Bayesian logistic regression model to examine predictors of booking probabilities. The predictors in this analysis were ethnicity and year. The results provide insights into equity in booking practices and the implications of these findings for law enforcement policies.

5.1 Black and Indigenous individuals are overrepresented in arrests but not in booking probabilities

Analysis of arrest data revealed that Black and Indigenous individuals were overrepresented in arrests relative to their share of the population, consistent with prior findings from the Toronto Police Service and other studies on racial disparities in policing. However, this study focused on booking probabilities after arrest and found no significant differences in booking outcomes across ethnic groups after controlling for year.

The wide credibility intervals in the Bayesian model, especially for smaller groups like Indigenous and East/Southeast Asian individuals, reflect high uncertainty. While this limits definitive conclusions about booking disparities, it suggests that any disparities observed in arrests may not persist at the booking stage. However, data limitations and measurement biases—such as officer-recorded perceived race—may influence these findings.

5.2 Booking probabilities remained stable across 2020 and 2021

The model found no significant change in booking probabilities between 2020 and 2021. The year coefficient (-0.076) was small and statistically insignificant, indicating stability in booking practices over the two years. This result suggests that broader procedural norms or operational policies likely remained consistent during this period, despite heightened societal focus on police reform.

The stability in booking probabilities aligns with the broader literature on procedural safeguards in booking decisions. While arrests may be influenced by discretionary or situational factors, booking practices often involve formalized procedures that may reduce variability. However, the absence of contextual variables in this dataset, such as offense type or search status, limits the ability to fully explain the observed stability.

5.3 Implications for policy and public discourse

The absence of strong evidence for racial disparities in booking probabilities is encouraging for equitable law enforcement practices at this stage. However, the high uncertainty in estimates for smaller groups highlights the need for continued monitoring and data improvements. Policymakers should focus on enhancing data collection practices and transparency in booking decisions while considering potential disparities earlier (e.g., arrests) or later (e.g., sentencing) in the justice system.

Communicating these findings is critical for informed public discourse. While arrest data often highlights disparities, the results suggest that booking probabilities may not exhibit the same biases. This distinction can help guide discussions on police reform and support targeted interventions to address inequities at all stages of law enforcement decision-making.

5.4 Limitations and future research

5.4.1 Data Limitations

The dataset relied on officer-recorded perceived race and binary booking outcomes, omitting variables such as offense type, search status, or socio-economic indicators. These omissions may obscure nuanced influences on booking decisions and contribute to the high uncertainty in coefficient estimates for smaller groups.

5.5 Future Research

1. **Incorporate Contextual Variables**: Adding variables like offense type or search status would provide a more comprehensive view of factors influencing booking probabilities.
2. **Examine Interaction Effects**: Analyzing interactions between race and year could highlight evolving trends for specific groups.
3. **Use Hierarchical Models**: Regional analyses could identify precinct-level disparities or best practices in booking decisions.
4. **Explore Arrest-to-Booking Pipelines**: Investigating disparities across the entire law enforcement process—from arrests to sentencing—would provide a holistic understanding of equity in policing.

6 Appendix

6.1 Model details

6.1.1 Posterior predictive check and comparison of the prior and posterior

(1) shows posterior predictive check was performed . 1 illustrates a comparison between posterior distribution simulations and the actual probability of experiencing a custody arrest

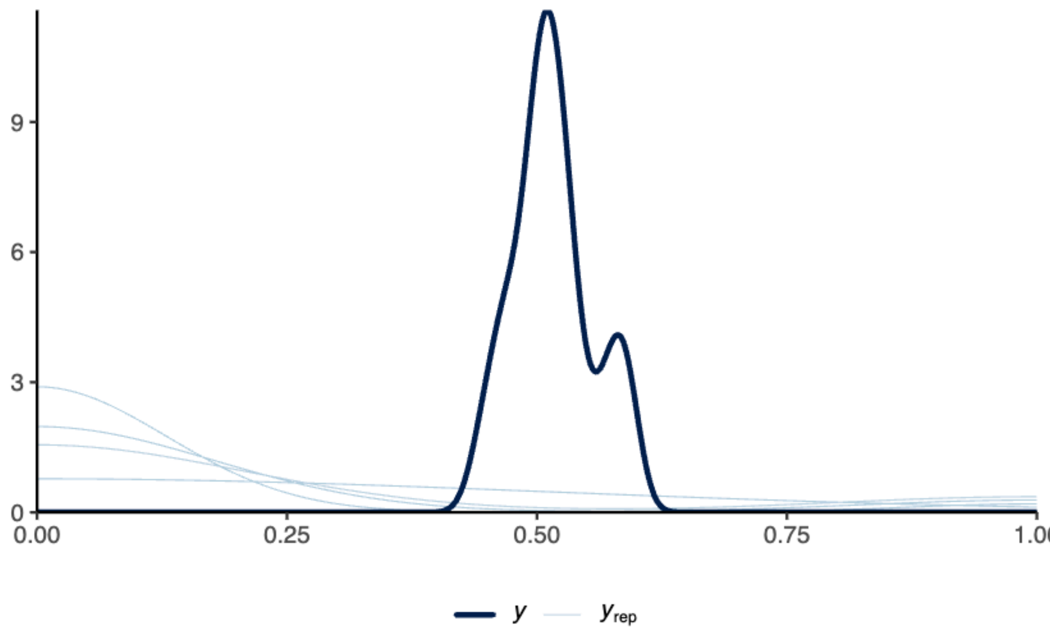

Figure 4: Posterior predictive check

In addition, using code adapted from (**tellingstorieswithdata?**), a comparison of the prior and posterior was performed (**?@fig-comparison**). **?@fig-comparison** depicts the amounts by which the estimates are modified after considering data (**tellingstorieswithdata?**).

6.1.2 Diagnostics

This section assesses whether the `rstanarm` (**citerstanarm?**) Markov chain Monte Carlo (MCMC) algorithm experienced problems (**tellingstorieswithdata?**).

Using code adapted from (**tellingstorieswithdata?**), a trace plot was generated (Figure 5). Figure 5 illustrates horizontal lines that are seemingly erratic with chain overlap, indicating nothing unusual (based on (**tellingstorieswithdata?**)).

Moreover, using code adapted from (**tellingstorieswithdata?**), a Rhat plot was created (Figure 6). Figure 6 depicts values near 1 and less than or equal to 1.1, indicating no issues (based on (**tellingstorieswithdata?**)).

6.2 Credibility intervals

Using code adapted from (**tellingstorieswithdata?**), Figure 7 was generated. Figure 7 illustrates 90% credibility intervals of all of the predictors of undergoing a custody arrest based on

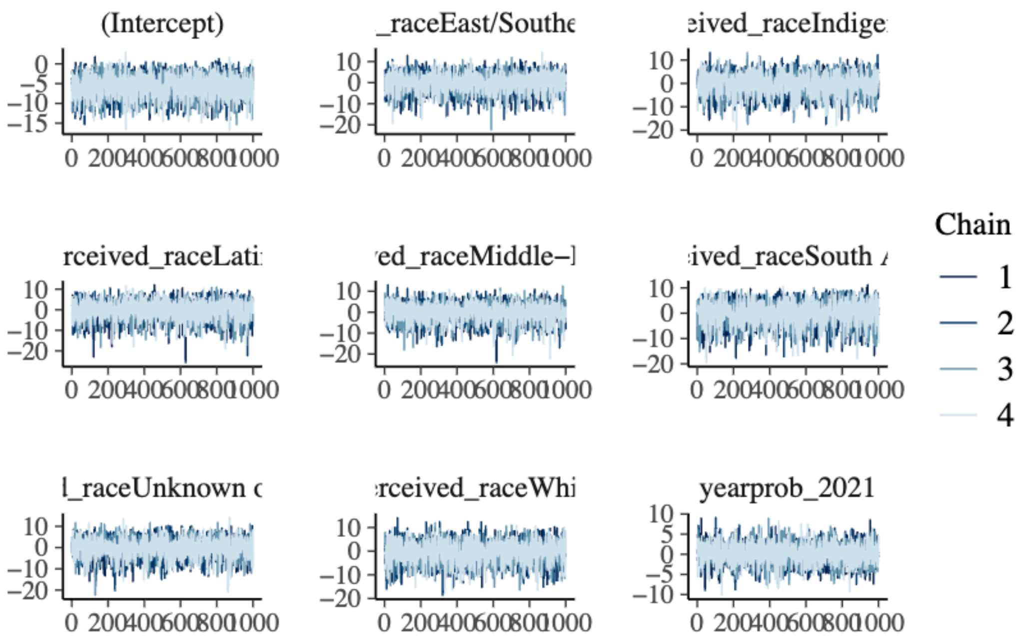

Figure 5: Trace plot

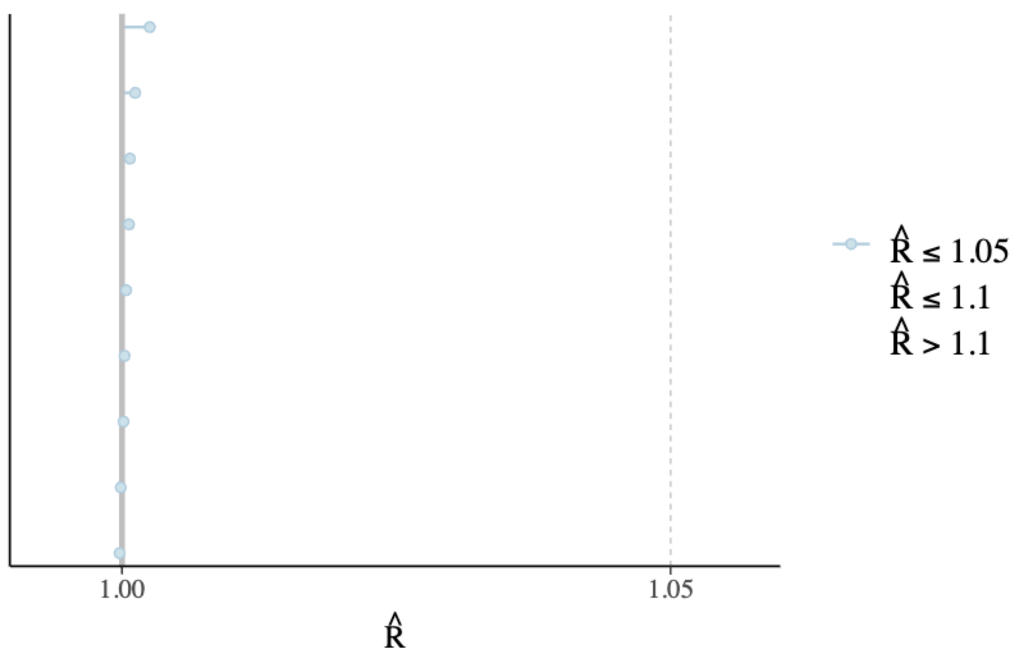

Figure 6: Rhat plot

the model from this report.

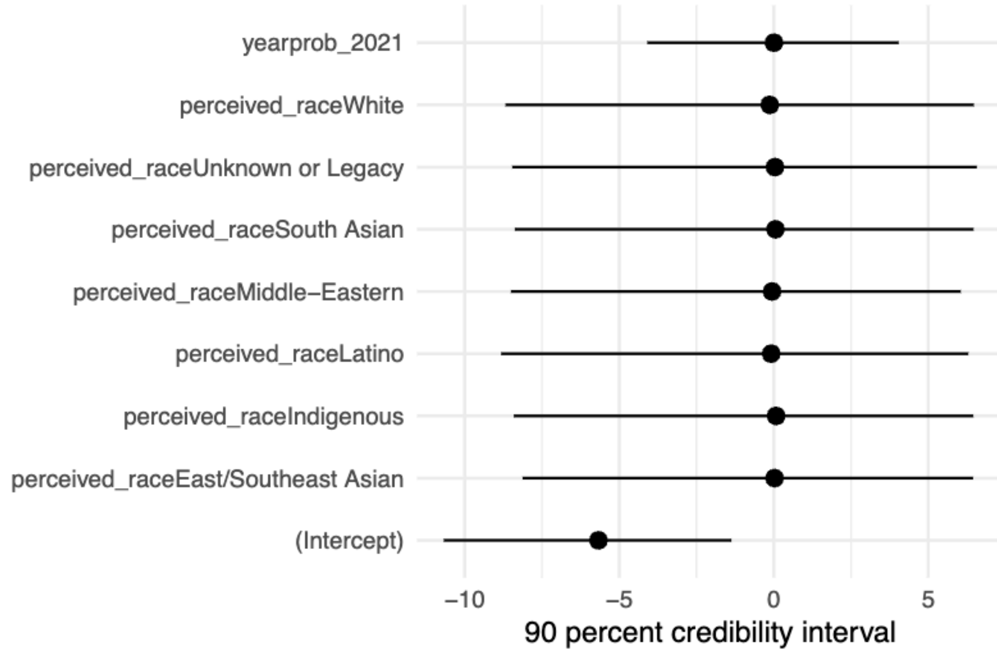

Figure 7: Credibility intervals of all of the custody arrest predictors

Figure 8 depicts 90% credibility intervals of several of the predictors of undergoing a custody arrest. In particular, the American Indian/Alaskan Native and Hawaiian/Pacific Islander predictors were omitted for visual clarity (as compared to Figure 7).

6.3 Observational Data and Sampling in Law Enforcement Studies

6.3.1 Observational Data Context

The dataset reflects arrest and booking events recorded by law enforcement in Toronto during 2020 and 2021. As observational data, it represents real-world interactions but is subject to biases like selection bias (e.g., only arrested individuals are included) and measurement bias (e.g., officer-perceived race).

6.3.2 Sampling Bias

The dataset captures only individuals who were arrested, excluding those who interacted with law enforcement but were not arrested. This selection bias may affect the generalizability of the findings to all police interactions.

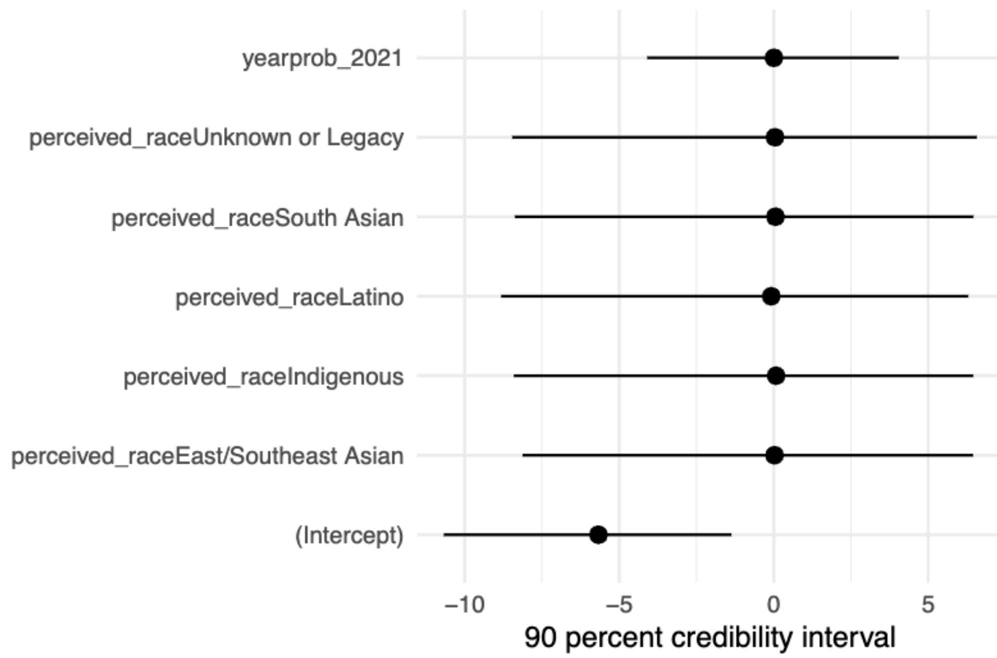

Figure 8: Credibility intervals of several of the custody arrest predictors

6.3.3 Measurement Bias

Key variables, such as `perceived_race`, are officer-reported and subject to subjective errors. Misclassification in smaller groups (e.g., Indigenous or East/Southeast Asian) can inflate or deflate booking probabilities for these groups, distorting the analysis. This aligns with literature emphasizing the need for objective self-reported data in law enforcement studies.

6.3.4 Implications and Recommendations

1. **Expand Variables**: Incorporating variables like offense type or socio-economic indicators can address confounding.
2. **Account for Bias**: Sensitivity analyses can estimate the impact of misclassification or selection bias.
3. **Improve Data Practices**: Encouraging self-reported race data and capturing broader police interactions can enhance representativeness.

These considerations highlight the strengths and limitations of observational datasets for analyzing law enforcement practices, emphasizing the importance of robust methods to ensure accurate interpretations.

References

Data, Toronto Open. 2022. "POLICE RACE AND IDENTITY BASED DATA - ARRESTS AND STRIP SEARCHES." https://open.toronto.ca/dataset/police-race-and-identity-based-data-collection-arrests-strip-searches/.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

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

Wickham, Hadley. 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, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.