Searching for Bias: An Analysis of Demographic Factors and Crime Severity on Strip Search Rates

Group 51
Yiyang Huang 1004902901
Mengchuan Wei 1009692101

A final paper submitted in conformity with the requirements
for the course of INF2178H

Faculty of Information
University of Toronto

2023.04.14

Abstract

Our primary goal is to investigate the potential biases and disparities in the use of strip searches on different demographic groups and potential difference in severity of the crimes that lead to an arrest committed by different demographic groups. We also aim to evaluate whether crime severity and number of actions at arrest are considered fairly in determining the necessity of strip searches. The study is based on 65276 arrest records containing information related to all arrests and strip searches, created by Toronto Police Service. The study was carried out by utilizing a combination of analytical methods, including exploratory data analysis, power analysis, ANCOVA tests, and logistic regression, to ensure a comprehensive investigation. Our study indicates significant gender differences in crime severity scores and racial disparities in strip searches during arrests. Indigenous and Black individuals were found to be more likely to be strip-searched. Furthermore, crime severity was not a key factor in strip search decisions, while actions taken during arrest playing a more significant role. Targeted measures should be implemented to minimize the risk of racial and gender discrimination during arrests and formalize the procedures.

Keywords: Strip Searches, Crime Severity, Under Age, Youth, Gender Disparity, Racial Disparities

Table of Contents

1. Introduction

The role of police officers in maintaining public order and safeguarding people's safety is critical, and this profession comes with significant power. However, it is crucial to monitor this power and ensure it is kept in institutional cages to avoid potential abuse. At the same time, it is essential to maintain a balance and not infringe upon the rights and interests of police officers, as this can demoralize the majority of police officers working on the front lines and exacerbate the profession's shortage. Today, there is a growing debate in society about police accountability and unbiased enforcement. According to *Public perceptions of the police in Canada's provinces*, published in the Canadian Centre for Justice and Community Safety Statistics, in 2019 only 42% of Canadians believe that their local police are able to treat people fairly (Dyna Ibrahim, 2020). Therefore, establishing a good interaction between police officers and citizens becomes essential for individuals and communities. The *RCMP's Bias-Free Policing Model Report* suggests that the justice system should extend the prohibition of racial profiling to all forms of bias-based profiling to support fairness in law enforcement practices. (CRCC, 2022).

In this study, we will evaluate potential biases in the use of strip searches, as well as demographic disparities in the severity of the crimes committed. Also, we aim to identify the factors associated with the use of strip searches and access whether crime severity and number of actions at arrest are considered fairly in determining the necessity of strip searches. The overall goal of the study is to provide a comprehensive analysis of the arrests and the use of strip searches in law enforcement and to identify any potential areas for improvement in policies and procedures.

This study is based on the *Arrests and Strip Searches* dataset Toronto Police Service, which includes 65,276 rows of records for 27 variables. To ensure the feasibility of the study, a literature review and power analysis are conducted prior to the experiment. This sample size is large enough to allow for statistical testing and more complex procedures. The research questions and hypotheses are then clearly stated. ANCOVA and logistic regression are implemented to identify the disparities and factors associated with the arrests and strip searches. In the analysis section, we will discuss the results obtained from ANCOVA and logistic regression and will test the hypothesis. In the limitations section, we will discuss any potential shortcomings or weaknesses of our study, including issues related to sample size and potential confounding variables.

2. Literature Review

A strip search is a procedure that is typically conducted by police officers on an arrested person, with clothing removed, to ensure that there are no weapons, drugs, or other contrabands are hidden(Lemke, 2022). This procedure can only be conducted when there is a valid suspicion that the arrested person is in possession of items that could pose a threat to staff and inmates in the correctional facility(Leach & Sabbatine, 1996). In recent years, there has been an ongoing debate about the fairness of using this extreme police power during investigations. According to race-based data released by the Toronto police in June 2022, strip searches are conducted more frequently in arrests involving certain identity groups, particularly Black and Indigenous people(Lemke, 2022). The study shows that although Black people account for 10% of the Toronto population, the likelihood of being strip-searched is over 30%(Lemke, 2022). Similarly, more than one-third of the arrested Indigenous people were subjected to strip searches(Lemke, 2022). This suggests that potential biases may be presented during the decision-making process of the strip search procedure.

According to the report produced by Toronto Police Service on the use of force and strip searches in 2020, there exists gender and race disparities in the use of strip searches by police officers(Phan, Dinca-Panaitescu, & Rebelo, 2021). In 2020, men are 3.3% more likely to be strip searched than women, with a strip search rate of 23% and 19.7% respectively(Phan et al., 2021). Study also shows that male in all perceived race group had a greater chance of being strip searched expect for white individuals. Disproportionality in Strip Search is particularly presented in Black men. Among all male arrests, 27.7% of those are black while they account for 33.4% of the male strip searches(Phan et al., 2021).

Research also shows that age can make a significant difference in criminal behaviour. According to U.S. Department of Justice, youth who ages 17 and under were involved in 15% of all crimes committed in the United States, including 26% of the robbery crimes and 16% of the violent crimes(Puzzanchera, 2010). Among all types of crime, the percentages of arson and vandalism committed by juveniles are significantly greater than other types of offences(47% and 38% respectively). In terms of all arrests related to homicide, 10% of them involved juveniles(Puzzanchera, 2010). Gender disparities also exist in juvenile crimes. Arrests of female juveniles accounted for 29% of the juvenile arrests and this percentage was lower for all types of crimes(Puzzanchera, 2010).

3. Research Hypothesis

3.1 Research Question

To address concerns about the fairness of police enforcement raised by Canadians, we are interested in investigating the potential biases and disparities in the use of strip searches on different ethnic groups. More specifically, we aim to examine the role of race in the frequency of strip searches during the arrest process, controlling for the number of arrests. Additionally, we are interested in identifying the groups of individuals that may be at a higher risk of committing severe crimes. To achieve this, we are going to investigate the potential difference in severity of the crimes that lead to an arrest committed by males and females, controlling for the number of arrests and the number of strip searches.

In addition to studying the impact of demographic factors on strip searches, we aim to examine the relationship between the number of actions taken during arrest, the severity of the crime, and the likelihood of being strip-searched. The goal is to investigate whether these factors are used fairly in determining the necessity of strip searches.

3.2 Research Hypothesis

1. There is a significant difference in the frequency of strip searches conducted on minority groups (Black or Indigenous) that are being arrested compared to other ethnic groups, controlling for the number of arrests.
2. The crimes committed by male tend to have a higher severity score than those committed by female, controlling for the number of arrests and number of strip searches
3. An increase in the number of actions taken during an arrest will result in an increased likelihood of being strip-searched.
4. An increase in severity of the crime will lead to an increased likelihood of being strip-searched.

4. Methods

4.1 Dataset

The primary source of the research is a dataset from Toronto Police Service. This dataset contains information related to all arrests and strip searches including Person ID, Perceived Race, Sex, Age Group, Youth at Arrest (Yes/No), Arrest Location Division, Strip Search (Yes/No), Booked (Yes/No), Occurrency Category, Actions at Arrest, and Search Reason.

Each row of information in his dataset is initially collected by each police division in and out of the City of Toronto, and then compiled and published by a private member of the Toronto Police Service. The Arrests_and_Strip_Searches_(RBDC-ARR-TBL-001).csv contains the records of 65276 arrests from the first quarter of 2020 to the fourth quarter of 2021, and in this research out main variables are:

`PersonID`: The integer data (Nominal) of the identity of the arrested person.

Perceived_Race: The string data (Nominal) of the perceived race of the arrested person.
Sex: The string data (Nominal) of the gender of the arrested person.
Youth_at_arrest__under_18_years: The string data (Nominal) of whether the person arrested was under 18 years old.
StripSearch: The integer data (Nominal) of whether the arrested person was strip-searched.
Booked: The integer data (Nominal) of whether the arrested person was strip-searched.
Occurrence_Category: The string data (Nominal) of the occurrence category of the arrested person.
Actions_at_arrest___Concealed_i: The integer data (Nominal) of whether the arrested person had actions of concealed items at arrest.
Actions_at_arrest___Combative__: The integer data (Nominal) of whether the arrested person had actions of combative, violent or spitter/biter at arrest.
Actions_at_arrest___Resisted__d: The integer data (Nominal) of whether the arrested person had actions of resisted, defensive or escape risk at arrest.
Actions_at_arrest___Mental_inst: The integer data (Nominal) of whether the arrested person had actions of mental instability or possibly suicidal at arrest.
Actions_at_arrest___Assaulted_o: The integer data (Nominal) of whether the arrested person had actions of assaulted officer at arrest.
Actions_at_arrest___Cooperative: The integer data (Nominal) of whether the arrested person had actions of cooperative at arrest.

4.2 Data Preparation

4.2.1 Data Cleaning

The dataset contained a number of NaN values, so we first needed to clean the dataset.

First, we excluded NaN values for *Sex*, *Perceive_Race*, and *Occurrency_Category*. Since there are 57,475 rows of Search Reasons (57,475 rows by 4 columns) that are all NaN values, accounting for 88.05% of the dataset, and we will not use these variables in this study, they will not be processed this time.

Second, we exclude the arrests involved unisex individuals (N = 9) from the dataset because we are only interested in studying male and female during this study.

Third, since the variable *Youth_at_arrest__under_18_years* has two values with the same meaning but different names, “Youth (aged 17 and younger)” and “Youth (aged 17 years and under)”, we merge them to unify the names.

Fourth, similar to the variable *Youth_at_arrest__under_18_years*, the variable *Occurrence_Category* has some same meaningful values but different names, including “Break & Enter” and “Break and Enter”, “FTA/FTC/Compliance Check/Parollee” and “FTA/FTC, Compliance Check & Parollee”, “Other Statute & Other Incident Type” and “Other Statute/Other

Incident Type", "Vehicle Related (inc. Impaired)" and "Vehicle Related", so we also merge them to unify the names.

After data cleaning, 65074 observations are obtained.

4.2.2 Data Wrangling

The dataset does not contain the continuous variables required in this study, so we needed to create the required new variables.

First, we created a numeric variable called *Score* based on *Occurrence_Category* to measure the severity of the incident that led to the arrest. We assigned a value to each type of incident that appears in *Occurrence_Category* based on the extent to which it would affect the safety of the population. Score is calculated by summing up the pre-assigned values based on the incident that led to each arrest.

Second, we created a quantity variable called *num_actions_at_arrest* based on 6 similar binary variables such as *Actions_at_arrest___Concealed_i*, *Actions_at_arrest___Combative__*, etc. to measure the total number of actions at the time of arrest. The total number is calculated by adding the original 6 binary variables together.

Third, we created *arrest_counts* and *strip_search_counts* by grouping *PersonID*, *Sex*, and *Perceived_Race* to measure the total number of times each person was arrested and the total number of times they were strip-searched, respectively. Then, we calculated the average score of crime severity for each group, represented by *avg_score*.

4.3 Measurement

Variable	Value Category
Independent Variables	
Perceived Race of Arrested Person (*Perceived_Race*, Nominal)	White, Unknown or Legacy, Black, South Asian, Indigenous, Middle-Eastern, Latino, East/Southeast Asian
Gender of Arrested Person (*Sex*, Nominal)	M, F
Number of Times Individual was Arrested (*Num_of_Arrests*, Ratio)	Ranging from 1 to 53 in this dataset
Number of Times Each Arrestee was Strip-searched (*Num_of_Strip_Searches*, Ratio)	Ranging from 0 to 13 in this dataset
Whether Arrested Person was Youth (*Youth_at_arrest__under_18_years*, Nominal)	Not a youth, Youth (aged 17 years and under)
Severity of Cases Involved (*Score*, Ratio)	Ranging non-negative in this dataset

Number of Actions Each Arrestee Had at Arrest (*num_actions_at_arrest*, Ratio)	0, 1, 2, 3
Dependent Variables	
Severity of Cases Involved (*Avg_Score*, Ratio)	Ranging non-negative in this dataset
Number of Times Each Arrestee was Strip-searched (*Num_of_Strip_Searches*, Ratio)	Ranging from 0 to 13 in this dataset
Whether the Arrestee was Strip-searched (StripSearch, Nominal)	0, 1
Source: The Arrests and Strip Searches Dataset from Toronto Police Service	

Table 1. Variables and Value Categories

The first independent variable is the perceived race of the arrested person (*Perceived_Race*, Nominal). This one describes the perceived race of the arrestee. There are multiple ways to describe this variable in real life. For our dataset, the variable has 8 different values, such as Black, White, East/Southeast Asian.

The second independent variable is the gender of the arrested person (*Sex*, Nominal). This one describes the sex of the arrestee. For our dataset, the values of the variable are male (coded as M) or female (coded as F).

The third independent variable is the number of times the individual was arrested (*Num_of_Arrests*, Ratio). For our dataset, the range of this variable is from 1 to 53.

The fourth independent variable, which is also the second dependent variable, is the number of times each arrestee was strip-searched (*Num_of_Strip_Searches*, Ratio). It can take any non-negative number. For our dataset, the range of this variable is from 0 to 13.

The fifth independent variable is whether the arrested person was under 18 years old or not at the time of arrest (*Youth_at_arrest__under_18_years*, Nominal). For our dataset, the values of the variable are no (coded as Not a youth) or yes (coded as Youth (aged 17 years and under)).

The sixth independent variable is the severity score of crimes (*Score*, Ratio). It is measured as a severity score and can take any non-negative number. For our dataset, the range of this variable is from 5 to 40. A higher value indicates a more serious case is involved by the arrestee.

The seventh independent variable is the number of actions each arrestee had at arrest (*num_actions_at_arrest*, Ratio). For our dataset, the range of this variable is 0, 1, 2, and 3.

The first dependent variable is the average severity score of crimes (*Avg_Score*, Ratio). For our dataset, it can take any non-negative number.

The third dependent variable is whether the arrestee was strip-searched (*StripSearch*, Nominal). For our dataset, the values of the variable are Yes (coded as 1) or No (coded as 0).

4.4 Methodologies

To begin with, we used Exploratory Data Analysis to show some of the basic statistics of our variables, as well as the sample in general.

Next, we conducted Power Analysis to ensure that the sample size reached the required value for our study to achieve the desired level of statistical power and to ensure that statistically significant effects were detected at a given level of confidence. We defined 2 functions to calculate the pooled standard deviation and Cohen's d respectively. We also set the significance level, chose the effect size, and determined the desired power, then used these to determine the sample size. Eventually, we evaluated the results and plotted power curves.

Then, since there were three research questions in this study, we did separate t-tests for the variables in each research question to verify whether there was a statistically significant relationship. All test results are statistically significant.

Next, we did 2 ANCOVA tests for the first 2 research questions separately. Before we did this, we checked the linear relationships. We found that there was a weak positive linear relationship between number of strip searches and number of arrests, there was a negative linear relationship between crime severity score and the number of strip searches, and there was a weak negative linear relationship between crime severity score and the number of arrests. Therefore, these 3 variables can be confounding variables during this study and we needed to use ANCOVA test to control for the confounding effect. The first ANCOVA test was to determine whether there was a significant difference in the frequency of strip searches (*Num_of_Strip_Searches*) conducted on minority groups (Black or Indigenous) that are being arrested compared to other ethnic groups, controlling for the number of arrests (*Num_of_Arrests*). The second ANCOVA test was to determine whether the crimes committed by male tend to have a higher severity score (*Avg_Score*) than those committed by female, controlling for the number of arrests (*Num_of_Arrests*) and number of strip searches (*Num_of_Strip_Searches*).

In addition, since the outcome is a binary variable, we conducted a Logistic Regression to determine the likelihood of being strip-searched and the factors that influence it. We also used SMOTE to balance the data so that the number of samples was equal when the output was the same as 0s and 1s. By doing so, each group was represented equally in the data and the model was able to learn equally from both groups. The dependent variable was whether the arrestee was strip-searched (*StripSearch*), and the independent variables were the perceived race of the arrestee (*Perceived_Race*), the gender of the arrestee (*Sex*), whether the arrestee was under 18 years old at arrest (*Youth_at_arrest__under_18_years*), the severity score of crimes (*Score*), and the number of actions each arrestee had at arrest (*num_actions_at_arrest*).

5. Exploratory Data Analysis

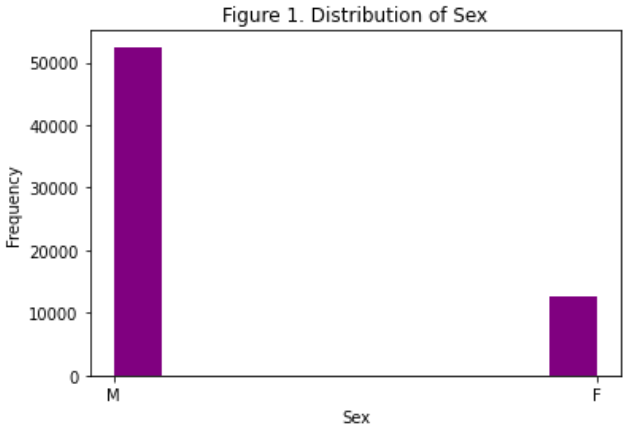
5.1 Descriptive Results

Variables	Median	Mean	SD	Range
Score	10	11.77	7.62	5-40
Num_actions_at_arrest	1	0.57	0.59	0-3
	Frequency	**%**		
StripSearch				
0	57275	88.02		
1	7799	11.98		
Youth_at_arrest__under_18_years				
Not a youth	62043	95.34		
Youth (aged 17 years and under)	3031	4.66		
Sex				
M	52499	80.68		
F	12575	19.32		
Perceived_Race				
White	27630	42.46		
Black	17487	26.87		
Unknown or Legacy	5041	7.75		
East/Southeast Asian	4402	6.76		
South Asian	3603	5.54		
Middle-Eastern	3227	4.96		
Indigenous	1926	2.96		
Latino	1758	2.70		

Table 2. Descriptive Statistics for Variables in ANCOVA and Logistic Regression (N=65074)

Descriptive statistics provide a fundamental summary of the sample and variables, serving as a preliminary basis for subsequent analysis. Table 2 presents the descriptive statistics of all the analytical variables. From this table, we can tell that around 12% of arrested individuals undergo a strip search. Additionally, adults account for 95% of all arrests. In terms of racial demographics, White individuals represent 42%, while Black individuals constitute 27% of the total arrests.

Next we conducted Exploratory Data Analysis (EDA). EDA allows us to become more familiar with the data, identify and correct obvious flaws, and understand the relationships between different variables.



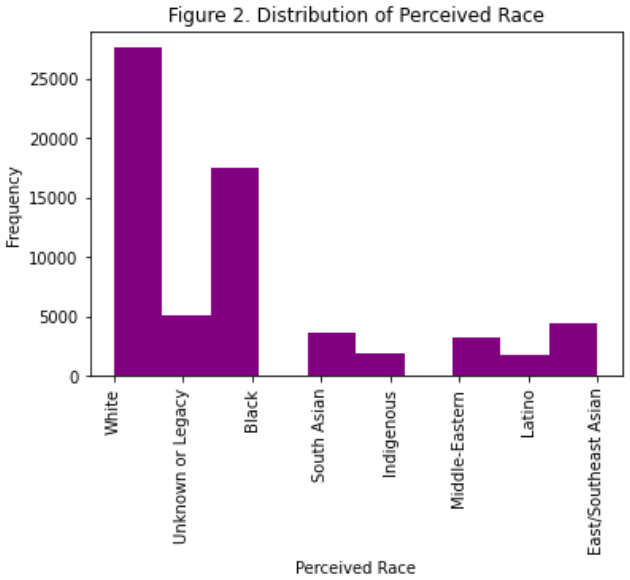


Figure 1 shows the distribution of sex, with the x-axis representing the sex and the y-axis representing the frequency. From this plot, we can tell male arrestees account for a greater proportion in all arrests compared to female arrestees. We can also tell that the data is not unbalanced. Figure 2 shows the distribution of perceived race, with the x-axis representing the perceived race and the y-axis representing the frequency. We can tell that the data is also not unbalanced.

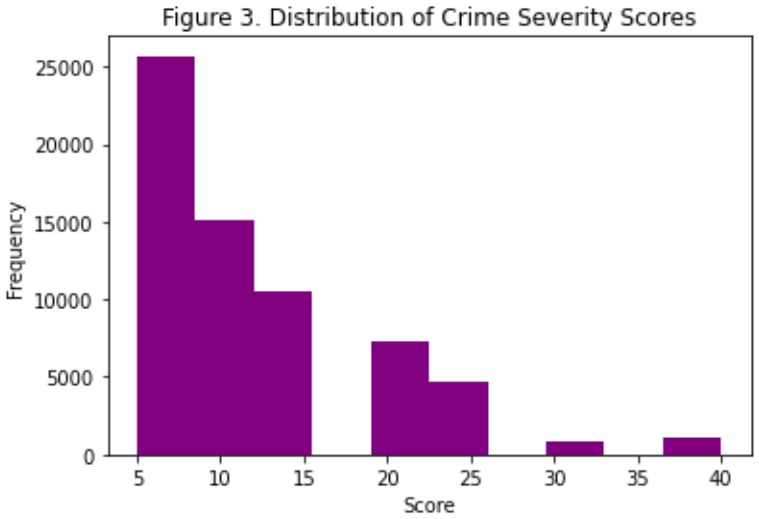


Figure 3 shows the distribution of crime severity scores, with the x-axis representing the scores and the y-axis representing the frequency. From this plot, we can tell the distribution of scores is heavily right-skewed, with more than half of the data points lying between 5 and 15. This suggests that most of the crimes committed are misdemeanors(~15) or below, such as theft and assault. The distribution of scores matches real-life scenarios, which suggests that the scoring system is valid.

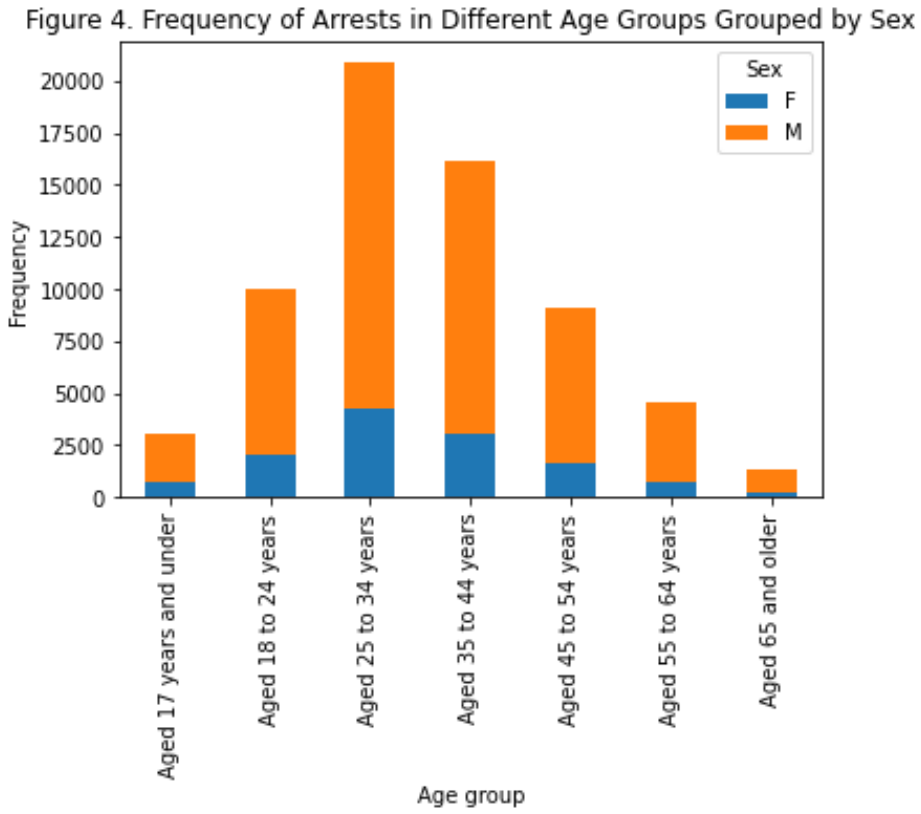


Figure 4 shows the number of arrests in each age group, grouped by sex. The bars are coloured by sex, with x-axis representing different age groups and y axis representing the frequencies. As shown in Figure 4, the number of arrests is significantly higher for males than females among all age groups. The distribution of age follows a normal distribution and the largest age group is 25-34 years old.

	Female	Male	Total
Not a youth	11843	50200	62043
Youth (aged 17 years and under)	732	2299	3031

Table 3. Contingency Table of Youth Status and Sex

Table 3 is a contingency table of youth status and sex. As shown in Table 2, 5% of all arrests involved youth who aged 17 years and under. Female youths account for a greater proportion in all youth arrests compared to female adults in all adult arrests. We can also tell that the data is highly unbalanced.

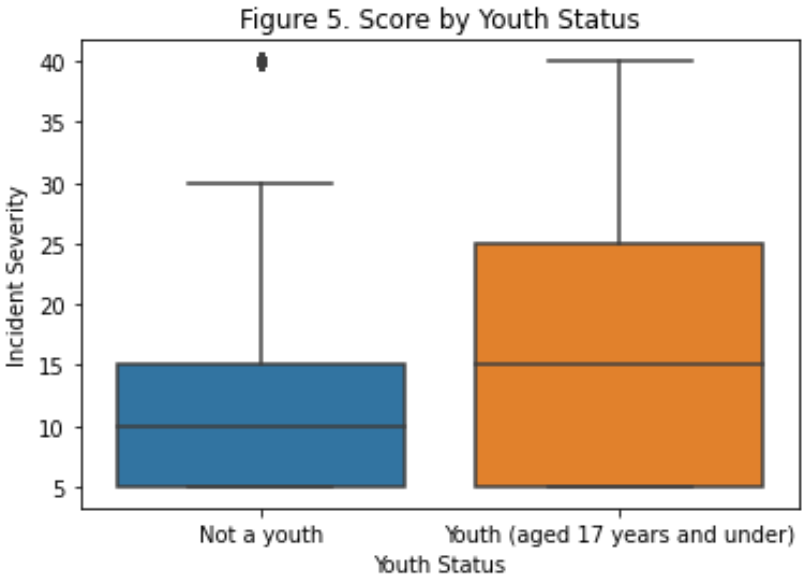


Figure 5 shows the distribution of the crime severity score grouped by youth status. As shown in this plot, the median and the third quantile of the severity score of the crimes committed by youth aged 17 or under are higher than those committed by adults. This potentially suggests that youth aged 17 or under tend to commit to more severe crimes than adults. To further investigate the difference in crime severity, we conducted a t-test and an anova test to validate this finding in later sections.

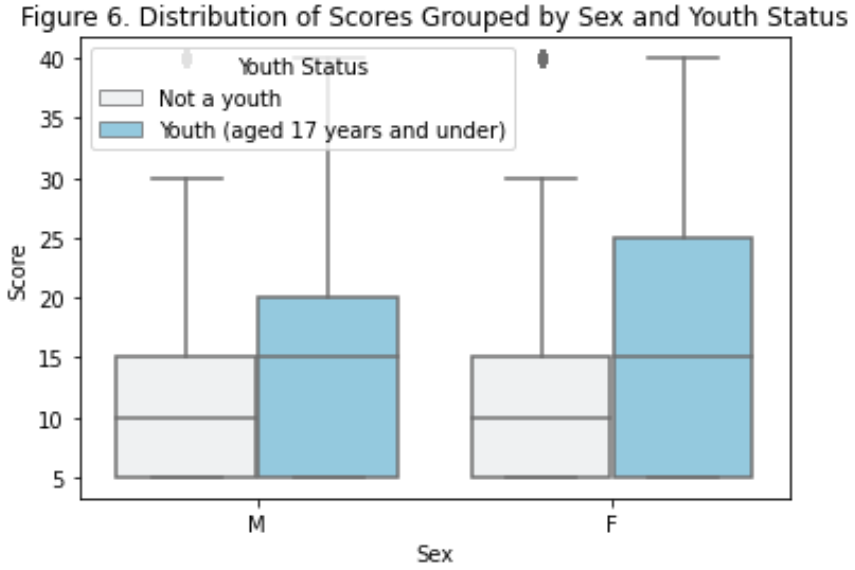


Figure 6 shows the distribution of the crime severity score grouped by sex and youth status. As shown in this plot, the distributions of the severity scores of crimes committed by male adults and female adults are similar, with a median score of 10 and some outliers around 40. The third quantile of the crime severity score for female youth is 25, which is higher than the third quantile for the other three groups. The significance of the difference in mean severity score will be further investigated by t-tests and ANOVA tests in the later section.

	No Strip Search	Strip Search	Strip Search Proportion
Male	45983	6516	0.124117
Female	11292	1283	0.102028

Table 4. Contingency Table of StripSearch and Sex

Table 4 shows the number of males and females who were strip-searched during the arrest process. As shown in this table, 12.4% of the males were strip-searched and 10.2% of the females were strip-searched. Therefore, we conducted further statistical analysis to investigate the gender disparity.

	No Strip Search	Strip Search	Total Number of Arrests	Strip_Searched_Prop
Indigenous	1620	306	1926	0.158879
Black	15053	2434	17487	0.139189
White	24064	3566	27630	0.129063
Unknown or Legacy	4506	535	5041	0.106130
East/Southeast Asian	4061	341	4402	0.077465
Latino	1626	132	1758	0.075085
South Asian	3346	257	3603	0.071329
Middle-Eastern	2999	228	3227	0.070654

Table 5. Contingency Table of StripSearch and Rerceived Race

Table 5 shows the number of strip searches, the total number of arrests, and the proportion of individuals who were strip-searched during the arrest process by their perceived race. As shown in the table, Indigenous, Black, and White Individuals had the highest likelihood of being strip-searched, with proportions at 15.9%, 13.9%, and 12.9%, respectively. South Asian and Middle-Eastern individuals have a relatively low risk of being strip-searched. Thus, we decided to conduct further statistical analysis to investigate the racial disparities.

Figure 7. Total Number of Arrests by Number of Actions at Arrest

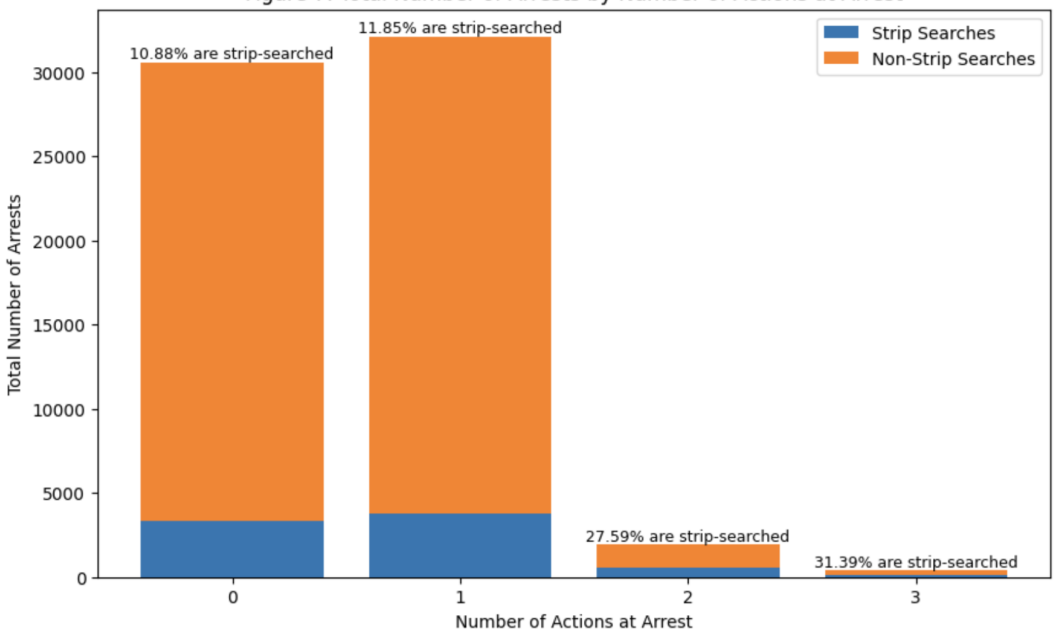


Figure 7 shows the plot of the total number of arrests grouped by the number of actions at arrest and the strip search indicator. From the figure, we can see that the majority of arrested individuals have zero or one action at arrest. Among those people, the likelihood of being strip-searched is around 11%. However, for individuals who have more than one action at arrest, the likelihood of being strip-searched increases to 29%. This suggests that the number of actions at arrest can be an important factor when determining the necessity of strip searches.

Figure 8. Crime Severity by Strip Search Indicator

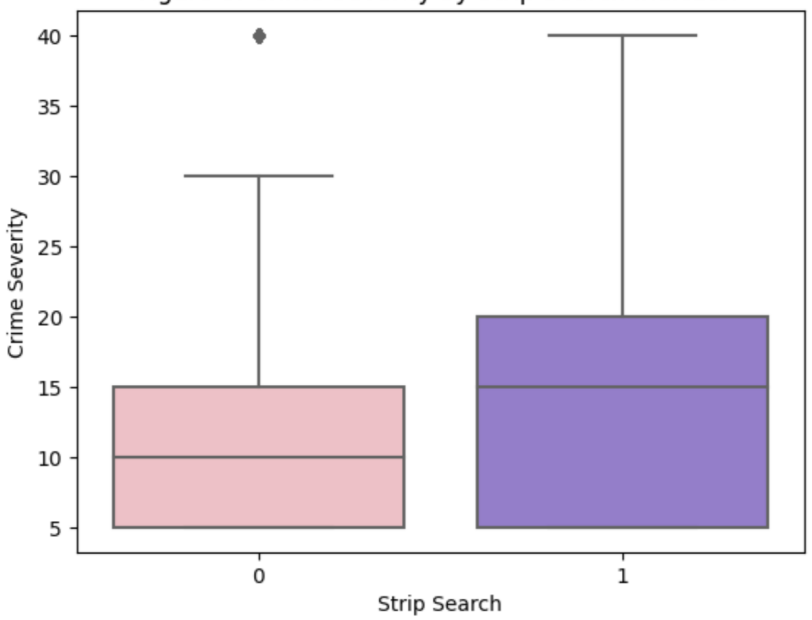


Figure 8 shows the box plot of the crime severity score by strip search indicator. The x-axis represents the strip search indicator, and the y-axis represents the crime severity score. From this plot, we can see that the median and third quartile of the crime severity score are 15 and 20, respectively, for individuals who are strip-searched, which is higher than for those who are not strip-searched. This suggests that the severity of the crime committed is likely to be another factor in determining the necessity of strip searches.

Fig 9. Scatterplots of Number of Arrests, Number of Strip Searches, and Crime Severity Score

Fig 9(a). Plot of Number of Strip Searches v.s. Number of Arrests

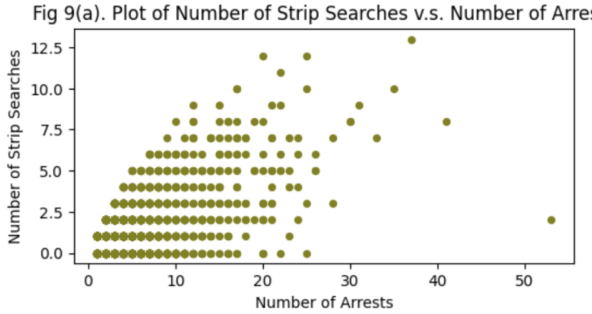


Fig 9(b). Plot of Crime Severity v.s. Number of Strip Searches

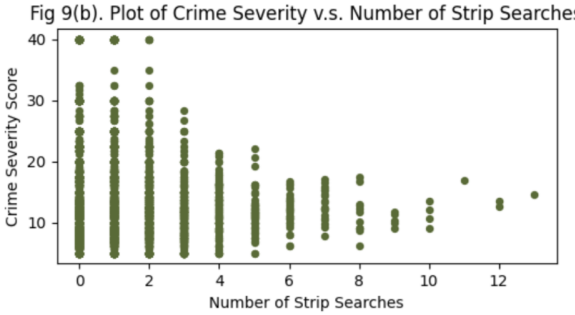


Fig 9(c). Plot of Crime Severity v.s. Number of Arrests

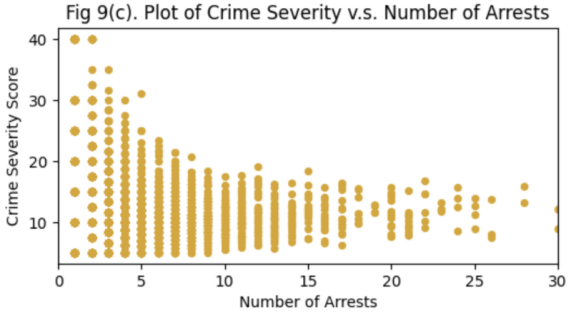


Figure 9 visualizes the relationship between number of arrests, number of strip searches, and crime severity score. Figure 9(a) suggests a weak positive linear relationship between number of strip searches and number of arrests. As the number of arrests that an individual increases, the number of strip searches tends to increase as well. Figure 9(b) suggests that there is a negative linear relationship between crime severity score and the number of strip searches experienced. Individuals who experienced more strip searches tend to commit less severe crimes. Figure 9(c) suggests that there is a weak negative linear relationship between crime severity score and the number of arrests experienced. Individuals who experienced more arrests tend to commit less severe crimes. These plots demonstrates that the number of arrests, number of strip searches, and crime severity score can be confounding variables during the study. Therefore, it is necessary to use ANCOVA instead of ANOVA to control for the confounding effect.

5.2 Welch's T-test

We are able to observe a difference in mean of number of strip searches and incident severity score in different demographic groups. To test the significance of differences, we ran several Welch's t-tests with the categorical features. Quantitative data are split into two groups based on the hypothesis that needs to be tested in each t-test. Before running the Welch's t-tests, we checked for the following three assumptions - Normality, Independency, and Homogeneity of variance. According to the Normal QQ-plot, the normality assumption is slightly violated, while the rest of the assumptions are fulfilled.

Race(Indigenous) and Strip Searches

We conducted a Welch's t-test to check the significance of the difference in mean frequency of strip search between Indigenous people and the other arrested individuals. Here are the hypotheses that are being tested:

> Null Hypothesis(H_0): There is no significant difference in the mean frequency of strip searches between Indigenous people and the other arrested individuals.($\mu_{SSInd} = \mu_{SSnotInd}$)
> Alternative Hypothesis(H_a): There is significant difference in the mean frequency of strip searches between Indigenous people and the other arrested individuals.($\mu_{SSInd} \neq \mu_{SSnotInd}$)

The test statistics is 5.40, which indicates that the difference in mean frequency of the strip searches between Indigenous people and the other arrested individuals is 5.40 standard errors away from zero. The p-value is 8.87×10^{-8}, which is less than the significance level of 0.05. The 95% confidence interval is [.118, .253], which does not include zero. Both of the p-value and 95% CI suggest that we are able to reject the null hypothesis and state that the difference is statistically significant. We may conclude that the difference in the mean frequency of strip searches between Indigenous people and the other arrested individuals is not equal to zero.

Race(Black) and Strip Searches

We conducted a Welch's t-test to check the significance of the difference in mean frequency of strip search between Black people and the other arrested individuals. Here are the hypotheses that are being tested:

> Null Hypothesis(H_0): There is no significant difference in the mean frequency of strip searches between Black people and the other arrested individuals. ($\mu_{SSBlack} = \mu_{SSnotBlack}$)
> Alternative Hypothesis(H_a): There is significant difference in the mean frequency of strip searches between Black people and the other arrested individuals. ($\mu_{SSBlack} = \mu_{SSnotBlack}$)

The test statistics is 7.78, which indicates that the difference in mean frequency of the strip searches between Black people and the other arrested individuals is 7.78 standard errors away from zero. The p-value is 7.89×10^{-15}, which is less than the significance level of 0.05. The 95% confidence interval is [.044, .074], which does not include zero. Both of the p-value and 95% CI suggest that we are able to reject the null hypothesis and state that the difference is statistically significant. We may conclude that the difference in the mean frequency of strip searches between Black people and the other arrested individuals is not equal to zero.

Race(White) and Strip Searches

We conducted a Welch's t-test to check the significance of the difference in mean frequency of strip search between White people and the other arrested individuals. Here are the hypotheses that are being tested:

> Null Hypothesis(H_0): There is no significant difference in the mean frequency of strip searches between White people and the other arrested individuals. ($\mu_{SSWhite} = \mu_{SSnotWhite}$)
> Alternative Hypothesis(H_a): There is significant difference in the mean frequency of strip searches between White people and the other arrested individuals.($\mu_{SSWhite} \neq \mu_{SSnotWhite}$)

The test statistics is 10.05, which indicates that the difference in mean frequency of the strip searches between White people and the other arrested individuals is 10.05 standard errors away from zero. The p-value is 1.05×10^{-23}, which is less than the significance level of 0.05. The 95% confidence interval is [0.056, 0.083], which does not include zero. Both of the p-value and 95% CI suggest that we are able to reject the null hypothesis and state that the difference is statistically significant. We may conclude that the difference in the mean frequency of strip searches between White people and the other arrested individuals is not equal to zero.

Sex and Crime Severity

We conducted a Welch's t-test to check the significance of the difference in mean severity score of crime committed by males and females. Here are the hypotheses that are being tested:

> Null Hypothesis(H_0): There is no significant difference in the mean severity scores of crimes committed by males and females. ($\mu_{ScoreMale} = \mu_{ScoreFemale}$)
> Alternative Hypothesis(H_a): There is significant difference in the mean severity scores of crimes committed by males and females. ($\mu_{ScoreMale} \neq \mu_{ScoreFemale}$)

The test statistics is -4.20, which indicates that the difference in mean frequency of the strip searches between White people and the other arrested individuals is -4.20 standard errors away from zero. The p-value is 2.67×10^{-5}, which is less than the significance level of 0.05. The 95% confidence interval is [-.460, -.167], which does not include zero. Both of the p-value and 95% CI suggest that we are able to reject the null hypothesis and state that the difference is statistically significant. We may conclude that the difference in the mean severity scores of crimes committed by male and female.

Strip Search Indicator and Number of Actions at Arrest

We conducted a Welch's t-test to check the significance of the difference in mean number of actions at arrest by strip search indicator. Here are the hypotheses that are being tested:

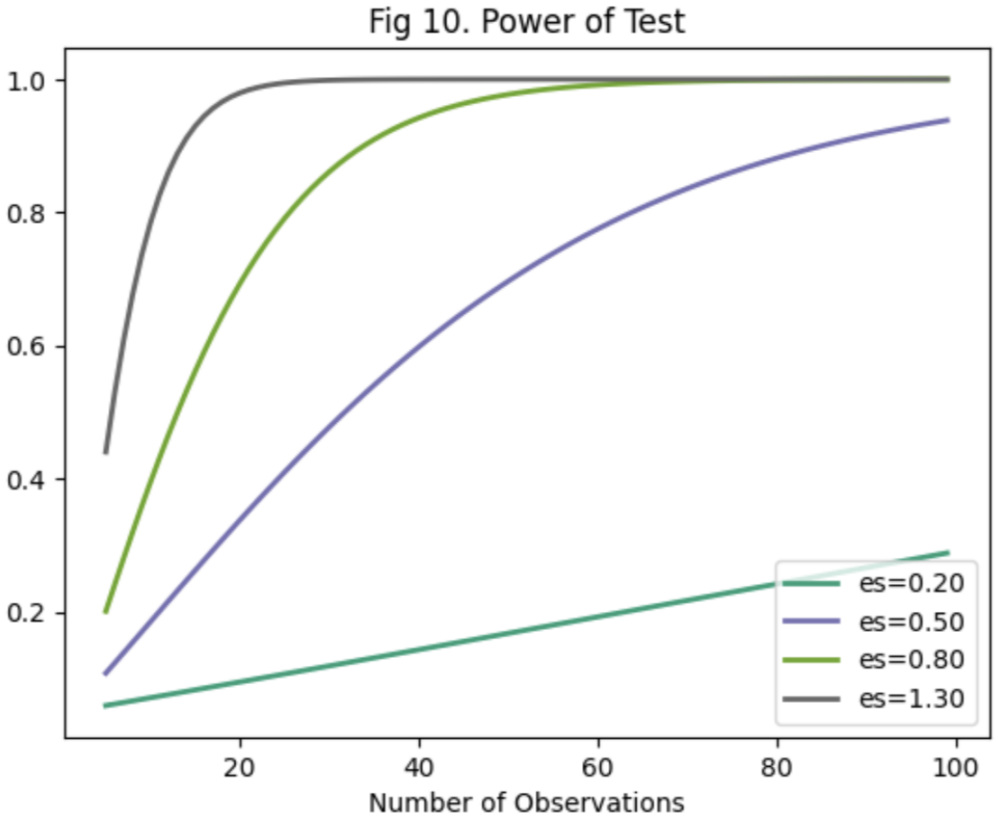
> Null Hypothesis(H_0): There is no significant difference in the mean number of actions at arrest by strip search indicator. ($\mu_{ActionsStripSearched} = \mu_{ActionsNotStripSearched}$)
> Alternative Hypothesis(H_a): There is significant difference in the mean number of actions at arrest by strip search indicator. ($\mu_{ActionsStripSearched} \neq \mu_{ActionsNotStripSearched}$)

The test statistics is 14.6, which indicates that the difference in mean number of actions at arrest between the individuals who undergo strip searches and those who do not is 14.6 standard errors away from zero. The p-value is 1.30×10^{-47}, which is less than the significance level of 0.05. The 95% confidence interval is [.101, .132], which does not include zero. Both of the p-value and 95% CI suggest that we are able to reject the null hypothesis and state that the difference is statistically significant. We may conclude that that there is a significant difference in the mean number of actions taken during the arrest process between individuals who undergo strip searches and those who do not.

6. Results

6.1 Power Analysis

Power analysis is a crucial step in experimental design as it enables us to determine the sample size needed to detect a statistically significant result. Before conducting the ANCOVA, we performed a power analysis based on our research questions and the dataset. This involves estimating the effect size, setting an alpha level, and determining the desired power level. Based on these parameters, we calculated the minimum sample size required to detect a significant effect.



Here is the power graph that shows the number of samples needed to reach a specific power level for different effect sizes. The x-axis represents the number of observations required, while the y-axis represents the power level. Each line on the graph corresponds to a particular effect size. From this plot, we can tell that when the effect size is small, more samples are needed to reach the same power level. For example, when the effect size is 0.8, 60 samples are needed to reach a power level of 1. However, when the effect size is 1.3, only 20 samples are needed to reach the same power level.

Sex and Average Crime Severity Score

	Required Sample Size	Actual Sample Size
Male	1285	30572
Female	5102	7700

Table 6. Comparison of Required Sample Size and Actual Sample Size($\alpha = 0.05$, power = 0.8, effect size = -0.08)

In this study, we would like to investigate the difference in the severity of crimes committed by males and females, while controlling for the number of arrests and strip searches. Before conducting the ANCOVA test, we needed to ensure that we had sufficient samples in each group to detect a significant effect.

To start, we computed the effect size (Cohen's d) for the severity score. The effect size was -0.08, indicating a small negative effect. This suggests that the difference in the average severity score between males and females is relatively small.

Next, we set the significance level to 0.05 and the desired power level to 0.8, and calculated the sample sizes needed for each group. The results showed that 1285 samples were needed for males and 5102 samples were needed for females. In actuality, we had a sample size of 30572 for males and 7700 for females. This indicates that we had enough samples to detect a statistically significant result.

Perceived Race and Strip Search

In addition to investigating the severity of crimes committed by males and females, our study aims to examine the number of strip searches conducted on different ethnic groups when they are arrested, while controlling for the number of arrests.

The dataset includes eight different ethnic groups. To calculate the required sample sizes for each ethnic group, we set the significance level to 0.05 and the power to 0.8. We assumed that the difference in the number of strip searches conducted on different ethnic groups would be small, so we set the effect size to 0.2. Based on these parameters, we calculated that a sample size of 366 would be needed for each ethnic group.

Perceived Race	Actual Sample Size
Black	17032
East/Southeast Asian	4111
Indigenous	1888
Latino	1636
Middle-Eastern	3059
South Asian	3392
White	26866
Unknown or Legacy	4887

Table 7. Actual Sample Size of Each Ethnic Group

Table 7 displays the number of observations in each ethnic group. It indicates that all groups have a number of observations greater than the required sample size of 366. This suggests that we have sufficient sample sizes to detect a significant effect in our investigation of the number of strip searches conducted on different ethnic groups, while controlling for the number of arrests

6.2 ANCOVA Tests

According to the t-test performed earlier, we can conclude that the difference in average crime severity scores is statistically significant between males and females. Additionally, the frequency of strip searches conducted on minority groups (Black or Indigenous) is statistically different from those conducted on other ethnic groups. Therefore, we conducted ANCOVA tests to further investigate the differences while controlling for potential confounding variables.

Assumption Check

Before conducting the ANCOVA test, we performed assumption check to check for normality, homogeneity of variances, and linearity assumptions.

Shapiro-Wilk test is used to assess the normality of the data. The null hypothesis for the test is that the data is normally distributed. The p-value we get for each group is less than the significance level, this suggests that we should reject the null hypothesis and states that the data do not follow a normal distribution.

The Levene's test is used to assess the homogeneity of variances across groups. The null hypothesis for the test is that the variances are equal across groups. In our case, the p-values for the Levene's test on number of strip searches by perceived race and severitu score by sex are 6.66e-100 and 1.19e-08, respectively, which is less than 0.05. This indicates that the homogeneity of variances is slightly violated.

The scatterplots (Appendix) are generated to visually represent the relationships between the dependent variables (Num_of_Strip_Searches and Avg_Score) and the covariate (Num_of_Arrests) for each level of the between-subjects factors (Perceived_Race and Sex). From these two plots, we can observe a weak linear relationship between the different variables, suggesting that the linearity assumption is met.

Sex and Crime Severity Score

(Controlling for the Number or Arrests and Strip Searches)

Here are the hypotheses that are being tested:

Null Hypothesis (H_0):

There is no significant difference in the mean severity of crime by male and female while controlling for number of arrests and number of strip searches. ($\mu_{SeverityCrime\_Male} = \mu_{SeverityCrime\_Female}$)

Alternative Hypothesis (H_a):

There is a significant difference in the mean severity of crime by male and female while controlling for number of arrests and number of strip searches. ($\mu_{SeverityCrime\_Male} \neq \mu_{SeverityCrime\_Female}$)

	Sum of Squares	DF	F-Value	Uncorrected p-value
Sex	2.27e+03	1.0	44.52	2.55e-11
Num_of_Arrests	2.09e+04	7.0	409.40	1.40e-90
Num_of_Strip_Searches	1.14e+04	1.0	224.54	1.26e-50
Residual	2.05e+06	40299	NaN	NaN

Table 8. One-way ANCOVA Results for Sex and Crime Severity

Table 8 shows the results obtained by running the ANCOVA. The uncorrected p-value for sex is 2.55×10^{-11}, which is less than 0.05. This suggests that we can reject the null hypothesis and conclude that the difference in mean severity of crime between males and females, while controlling for the number of arrests and strip searches, is statistically significant.

This finding implies that sex is an important factor to consider when analyzing criminal behavior and the severity of crimes committed. It indicates that there may be underlying differences between males and females that affect their propensity to engage in criminal activities and commit more severe crimes. By considering sex factor, law enforcement agencies and policymakers can better understand and address the root causes of criminal behavior and develop targeted interventions for specific needs of each gender. It also highlights the need for further research into the factors that contribute to gender differences in criminal behavior.

Perceived Race and Strip Searches

(Controlling for the Number of Arrests and Crime Severity Score)

Here are the hypotheses that are being tested:

Null Hypothesis (H_0):

There is no significant difference in the mean number of strip searches by different perceived races, while controlling for the number of arrests. ($\mu_{StripSearchNumber\_i} = \mu_{StripSearchNumber\_j}\ for\ any\ i \neq j$)

Alternative Hypothesis (H_a):

There is a significant difference in the mean number of strip searches by different perceived races, while controlling for the number of arrests. ($\mu_{StripSearchNumber\_i} \neq \mu_{StripSearchNumber\_j}\ for\ any\ i \neq j$)

	Sum of Squares	DF	F-Value	Uncorrected p-value
Preceived_Race	20.68	7	12.08	1.68e-15
Num_of_Arrests	5891.57	1	24090.24	0.00
Avg_Score	54.04	1	220.98	7.50e-50
Residual	9854.16	40293	NaN	NaN

Table 9. One-way ANCOVA Results for Preceived Race and Strip Searches

Table 8 shows the results obtained by running the ANCOVA. The uncorrected p-value for preceived race is 1.68×10^{-15}, which is less than 0.05. This indicates that there is a statistically significant difference in the mean number of strip searches conducted for individuals of different perceived races, while controlling for the number of arrests and the crime severity score.

In the ANCOVA test, we controlled for the potential confounding factor—the number of arrests and crime severity score—to obtain a more robust result. In practical terms, the finding suggests that perceived race can be an important determinant of the likelihood of being strip-searched. This might indicate the existence of potential biases during arrests, leading to higher rates of strip searches for certain racial groups. Such findings raise concerns about potential racial discrimination in law enforcement practices and highlight the need for further investigation and potential policy changes to address these disparities. Policymakers should also consider implementing specific policies to reduce the risk of racial discrimination during the arrest procedure.

6.3 Logistic Regression

We aim to further investigate the impact of demographic factors on strip searches. Additionally, we are interested in examining the relationship between the number of actions taken during arrest, the severity of the crime, and the likelihood of being strip-searched. The outcome variable, StripSearch, is a binary variable that takes the value of 1 if the arrest resulted in a strip search and 0 otherwise.

Before fitting the logistic regression model, we apply the Synthetic Minority Oversampling Technique (SMOTE) to address the imbalance issue. Prior to resampling, there are 7,799 samples with StripSearch equal to one and 57,275 samples with StripSearch equal to zero. In the resampled data, each group has 57,275 samples.

We construct the logistic regression model with severity score, number of actions at arrest, perceived race, sex, and youth status as independent variables, and the strip search indicator as the outcome variable. Here is a summary of the logistic regression modeling results:

	Coefficient	Standard Error	t	P >| t |	[2.5%	97.5%]
const	0.3664	0.006	66.139	0.000	0.356	0.377
Score	0.0072	0.000	39.195	0.000	0.007	0.008
Num_actions_at_arrest	0.0649	0.003	25.060	0.000	0.060	0.070
Sex_M	0.0683	0.004	16.204	0.000	0.060	0.077
Youth_aged 17 years and under	-0.1683	0.009	-19.199	0.000	-0.180	-0.147
Perceived_Race_East/Southeast Asian	-0.1697	0.007	-22.990	0.000	-0.184	-0.155
Perceived_Race_Indigenous	0.0373	0.009	4.014	0.000	0.019	0.055
Perceived_Race_Latino	-0.2187	0.011	-19.121	0.000	-0.241	-0.196
Perceived_Race_Middle-Eastern	-0.2000	0.009	-23.121	0.000	-0.217	-0.183
Perceived_Race_South Asian	-0.1992	0.008	-24.149	0.000	-0.215	-0.183
Perceived_Race_Unknown or Legacy	-0.0902	0.007	-13.497	0.000	-0.103	-0.077
Perceived_Race_White	9908.20	0.005	-6.344	0.000	-0.032	-0.017

Table 9. Summary of the Logistic Regression Modeling Results

From this table, we can see that the coefficients for male, youth, and different perceived race categories are all statistically significant with p-values close to 0. The confidence intervals do not include zero. Both p-values and confidence intervals suggest that these demographic features are all statistically significantly associated with the likelihood of being strip-searched. Individuals with different demographic characteristics may have different chances of being strip-searched. Positive coefficients indicate higher log odds of being strip-searched compared to the reference category, while negative coefficients indicate lower log odds. Similarly, the crime severity score and the number of actions at arrest are also significantly associated with the likelihood of being strip-searched. However, interpreting log odds in a logistic regression is not usually meaningful, so we should calculate the odds ratios for further investigation.

	Odds Ratio
const	1.44
Score	1.007
Num_actions_at_arrest	1.067
Sex_M	1.071
Youth_aged 17 years and under	0.849
Perceived_Race_East/Southeast Asian	0.844
Perceived_Race_Indigenous	1.038
Perceived_Race_Latino	0.804
Perceived_Race_Middle-Eastern	0.819
Perceived_Race_South Asian	0.819
Perceived_Race_Unknown or Legacy	0.914
Perceived_Race_White	0.976

Table 10. Odds Ratios of Features

Table 10 displays the odds ratios for each feature, which are more interpretable than the log odds. In terms of perceived race, the reference group is Black. According to the results, we can see that only Indigenous individuals have higher odds of being strip-searched. The likelihood of an Indigenous individual being strip-searched is 3.8% higher than for Black individuals. White individuals have 2.4% lower odds of being strip-searched compared to the reference group. Latino individuals have the lowest likelihood of being strip-searched (19.6% lower than the reference group). Additionally, males have 7.1% higher odds of being strip-searched compared to females, and individuals aged 17 years and under have 15.1% lower odds of being strip-searched compared to those above 17 years. The findings suggests that there exist demographic disparities in strip searches. The significant odds ratios for perceived race categories suggest that racial biases may be present in strip search decisions. Also, there might exist gender-based differences in law enforcement practices.

Table 10 also indicates that for each unit increase in the severity score, the odds of being strip-searched increase by 0.7%. This suggests that the crime severity score might not be a key factor to consider for strip searches. However, for each additional action taken during arrest, the odds of being strip-searched increase by 6.7%. This shows that actions taken during arrest positively affect the likelihood of being strip-searched. The findings demonstrate a potential lack of consideration of crime severity in the decision-making process for strip searches.

Fig 11. Confusion Matrix for Logistic Regression

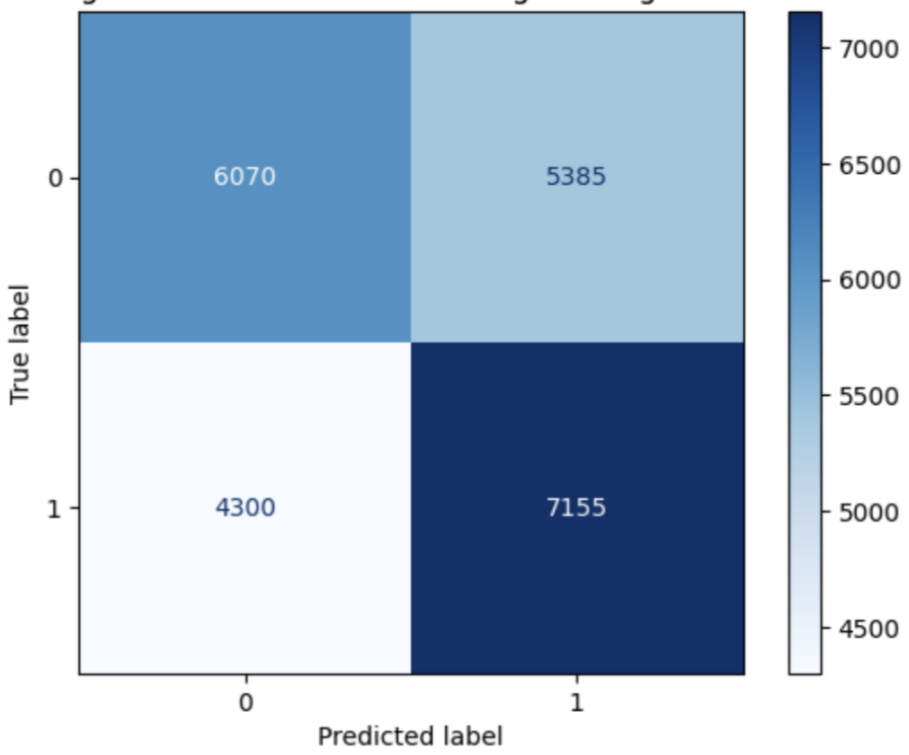


Figure 11 is the confusion matrix of the prediction generated by logistic regression, where x-axis represents the predicted label and y-axis represents the true label. Among 11455 negative cases, 6070 of them are predicted correctly. Among 11455 positive cases, 7155 of them are predicted correctly. The accuracy of the prediction is 0.58 and the precision is 0.57. We also examined the percentage of target values in the test data were within the prediction intervals. The result shows that 99.91% of the prediction intervals contain true target.

7. Discussion

7.1 Implications

In this study, we investigated the impact of demographic factors on the likelihood of being strip-searched during an arrest. Our findings reveal a statistically significant demographic disparities in strip search practices, with particular emphasis on perceived race, sex, and youth status. This suggests that there exists potential biases in law enforcement practices and raises concerns about possible racial and gender-based discrimination during the arrest process. Based on our findings, we believed that policymakers should implement targeted measures to minimize the risk of racial and gender discrimination during arrest procedures. This may include develop training programs for officers that focus on understanding equitable treatment of individuals from different demographic groups. Also, recruiting a diverse workforce within law enforcement agencies can promote understanding of different demographic groups, which may be helpful in addressing discrimination issues.

Moreover, our results indicate that crime severity may not be a key determinant in the decision to conduct a strip search. Instead, actions taken during arrest seem to have a more substantial impact on the likelihood of a strip search being performed. Law enforcement agencies should consider emphasizing crime severity and relevant situational factors in decision-making during the arrest procedures. Also, extra training can be provided to officers on addressing the potential consequences of conducting unnecessary strip searches.

The logistic regression model used in our study achieved a prediction accuracy of 0.58 and a precision of 0.57, indicating that the efficiency of the model in predicting strip search occurrences is not ideal. To achieve a better performance, performing a feature selection process to identify the most relevant and informative predictors might be helpful. This procedure will eliminate irrelevant and redundant features that may negatively impact the model's accuracy.

7.2 Limitations

Our study have certain limitations that should be acknowledged. Firstly, the arrests and strip searches dataset is a large dataset that contains more than 60,000 observations. During the study, we used 0.05 as a threshold for significance. However, when working with large datasets, a p-value less than 0.05 may not always be an appropriate standard to determine statistical significance. As sample size increases, there is a higher chance of obtaining a significant result, which means, in some cases, we may be seeing significance by accident(Bruhn, 2021). This suggests that rejecting every null-hypothesis based on whether or not p-value is less than 0.05 may not accurately reflect the true significance of the result we obtained(Bruhn, 2021). We should adjust the significance threshold based on sample size accordingly to get more valid results.

Secondly, the normality assumption is not met before conducting t-test and ANCOVA, and this could potentially affect the accuracy of the result. To address this issue, applying data transformations or employing robust statistical methods can help mitigate the impact of the violated normality assumption and improve the reliability of the analyses.

Lastly, crime severity scores are assigned based on the severity of the case associated with average jail time as a reference, which could be biased. For example, different types of sexual related crimes can have different degrees of severity. Assigning a single value to all such cases may fail to accurately reflect the true severity of the incident. It would be better to have a column recording the severity of each case at the time of arrest for future studies.

8. Conclusion

Our study aimed to address concerns about the fairness of police enforcement in Canada by investigating potential biases and disparities in the use of strip searches on different ethnic groups. The study is based on Arrests_and_Strip_Searches Dataset from Toronto Police Service. We examined the role of race in the frequency of strip searches during the arrest process, controlling for the number of arrests. Also, we explored the potential differences in crime severity between males and females, controlling for the number of arrests and strip searches. Then, we examined the relationship between the number of actions taken during arrest, crime severity, and the likelihood of being strip-searched.

We found that demographic factors, such as race, sex, and youth status, significantly influenced strip search practices, suggesting potential biases and disparities in law enforcement. Moreover, our results indicated that crime severity may not be the primary determinant for conducting strip searches, with actions taken during arrest having a greater impact on the likelihood of strip searches.

However, our study has limitations, including the use of a 0.05 significance threshold with a large dataset and violations of the normality assumption before conducting t-tests and ANCOVA. Future studies could address these limitations by adjusting the significance threshold based on sample size and employing robust statistical methods or data transformations to mitigate the impact of violated normality assumptions.

In conclusion, our findings suggest that there may be biases and disparities in the use of strip searches during arrests. Policymakers and law enforcement agencies should consider implementing targeted measures to minimize discrimination, such as developing training programs focused on equitable treatment and promoting workforce diversity in law enforcement agencies. Future research should continue to investigate the factors influencing strip search decisions and explore ways to address the observed disparities and potential biases in law enforcement practices.

References

Bruhn, A. (2021). Stop Using p<0.05. Retrieved 25th Feb, 2023 from https://towardsdatascience.com/stop-using-p-0-05-9743e5cddc21

Civilian Review and Complaints Commission. (2022). Review of the RCMP's Bias-Free Policing Model Report. https://www.crcc-ccetp.gc.ca/en/review-rcmps-bias-free-policing-model-report

Dyna, I. (2020). Public perceptions of the police in Canada's provinces, 2019. Retrieved 25th Feb, 2023 from https://www150.statcan.gc.ca/n1/pub/85-002-x/2020001/article/00014-eng.htm

Lemke, M. (2022). Policing Toronto: Strip Searching in a Divided City. Retrieved 25th Feb, 2023 from https://www.ojp. gov/ncjrs/virtual-library/abstracts/new-strip-search-paradigm#:~:text=Under%20the%20proposed%2 0paradigm%2C% 20a,so%20as%20to%20protect%20the

Leach, D. & Sabbatine , R. (1996). New Strip Search Paradigm. Retrieved 25th Feb, 2023 from https://www.ojp.gov/ncjrs/ virtual-library/abstracts/new-strip-searchparadigm#:~:text=Under% 20the%20proposed%20 paradigm %2C%20a,so%20as% 20to%20protect%20the

Mary K. A. & Tamy S. (2016). Youth crime in Canada, 2014. Retrieved 25th Feb, 2023 from https://www150.statcan.gc.ca/n1/pub/85-002-x/2016001/article/14309-eng.htm

Phan, M. B., Dinca-Panaitescu, M., & Rebelo, N. (2021). Understanding Strip Searches in 2020. Retrieved 25th Feb, 2023 from https://www.tps.ca/media/filer_public/e4/b1/e4b1 b125-2a2e -4d69-ad02-77ab3f3d5878/4e217e01-3cd6-4fe8-8898-39 cf8693e871.pdf

Puzzanchera, C. (2010). Juvenile Arrests (2007). Diane Publishing.

List of Appendices

Figure A1. Normal QQ Plot - Severity Score(Male)

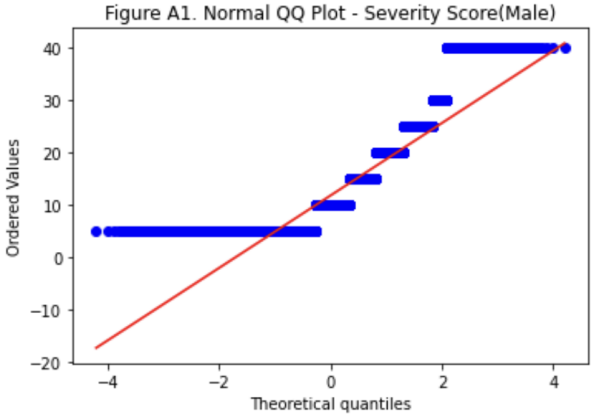
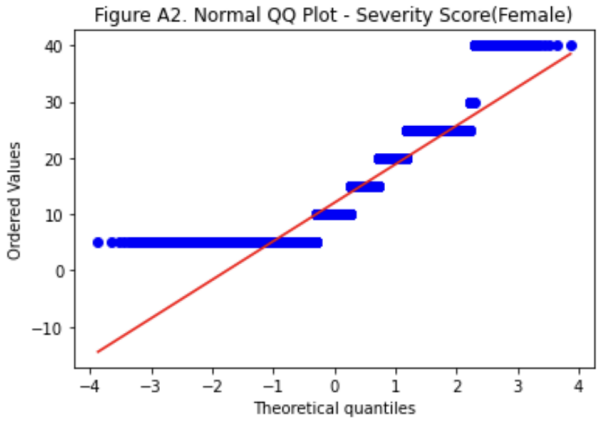
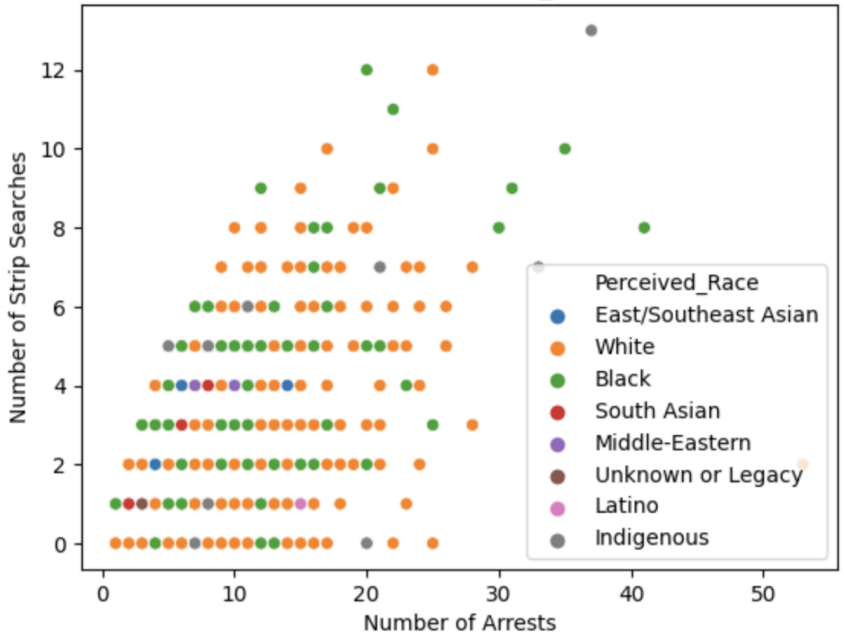


Figure A2. Normal QQ Plot - Severity Score(Female)



A3. Scatterplot of Number of Strip_Searches vs. Arrests



A4. Scatterplot of Crime Severity Score vs. Number of Arrests