University of Toronto - Faculty of Information

INF 2178 Final Project

Statistical Analysis on Data Provided by the Toronto Police Service

Group 13: Anja Zhang & Lok Lam Wong

April 16, 2023

Introduction

The purpose of this study is to analyze the "Arrest and Strip Search" dataset from the Toronto Police Service's Public Service Data Portal to better understand strip search practices and their impact on minority groups in Canada. The study will begin with a literature review that examines the definition of unlawful strip searches, the legal framework governing such searches, and the psychological and emotional impact of such searches on minority groups.

Unlawful strip searches have been a topic of concern in Canada for a long time, especially for minority groups. This literature review aims to explore what constitutes an unlawful strip search, the legal regulations surrounding it, and the emotional and psychological consequences for minority groups in Canada. Unlawful strip searches are defined as searches that involve removing some or all of an individual's clothing without proper justification or violating their rights. The Toronto Police define a strip search as removing or rearranging clothing to visually inspect private areas, such as genitals, buttocks, breasts, or undergarments (Lemke, 2022). This type of search is conducted without the person's consent and under police supervision.

Although strip searches are allowed when an individual is being admitted to a correctional facility, they must be conducted in a minimally intrusive manner. However, marginalized groups, including Indigenous peoples, Black Canadians, and people with disabilities, have reported experiencing strip searches that are degrading, humiliating, and unnecessary. Indigenous women and girls have been subjected to strip searches in a discriminatory and harmful manner, leading to calls for the reform of strip search policies in Canada (Ghobrial, 2021). Minority groups, such as Black Canadians, Indigenous people, and people with visible disabilities, are disproportionately subjected to strip searches when there is no reasonable suspicion. According to a 2019 report by Ontario's Independent Police Review Director, approximately 22,000 strip searches are conducted by police each year in Ontario, with the majority taking place in Toronto. This has led to concerns that strip searches are being used as a tool for harassment and discrimination against minority groups (Gillis, 2020).

The Toronto Police Force has implemented measures to address these issues, including reviewing procedures, training, and accountability. The force has also overhauled when and how officers perform strip searches, which has led to a significant drop in numbers. This includes mandatory steps officers must take before conducting a strip search, such as conducting a less invasive "frisk" search. The force has also implemented higher levels of reporting and data collection requirements to better understand the impact of strip searches on minority groups and to minimize their negative effects.

In conclusion, unlawful strip searches are a significant issue in Canada, particularly for minority groups. While they may be necessary for certain situations, it is crucial that they are conducted with respect for individuals' rights and minimize physical and emotional harm. More reporting and analysis are needed to develop strategies to minimize the negative impact of strip searches

on affected individuals, especially marginalized communities. This study will then construct three research questions to tackle the data from different perspectives and conduct exploratory data analysis using graphical and visualization methods. To test hypotheses, the study will employ statistical analysis methodologies, such as ANCOVA and logistic regression. Finally, the results will be synthesized to draw a conclusion and identify areas for improvement in future work. The study aims to contribute to the reform of strip search policies in Canada by shedding light on the current state of affairs and highlighting areas that need improvement.

Data Description

The Toronto Police Service contributed the dataset for this study, "Arrests and Strip Searches," which was used in the analysis. It contains 65,276 records with 24 qualities and details regarding strip searches and arrests made by the police between 2010 and 2019. There are 12 categorical and 12 numerical variables in the dataset, with the numbers 1 and 0 denoting the numerical variables. The variables include the person's age, gender, and race as well as the time, place, and nature of the offense that led to their arrest. They also include whether or not a strip search was carried out. The collection also contains details about the kind of strip search conducted, where it took place, and how it turned out.

The dataset can be used to examine trends in the Toronto Police Service's arrests and strip searches, including any potential inequalities in the utilization of strip searches in relation to particular demographics. Researchers can also investigate any biases in the data by looking at the causes of arrests and the results of strip searches. Due to internal documentation concerns, the dataset may contain certain errors, such as cases where a person had a strip search but was not recorded as having been booked.

Since each record in the dataset has a "Arrest_ID," the team assumes that all of them represent arrests. Overall, the dataset is a helpful tool for academics, decision-makers, and community members trying to understand how Toronto police use strip searches and how it affect disadvantaged neighborhoods.

Power Analysis

To better understand the ideal sample size and its related effect size of female and male strip searches, we performed a power analysis. We assume that power is 0.8 and alpha is 0.05, which are some usual standards.

Table 1: Power analysis result (Strip Search, Sex)

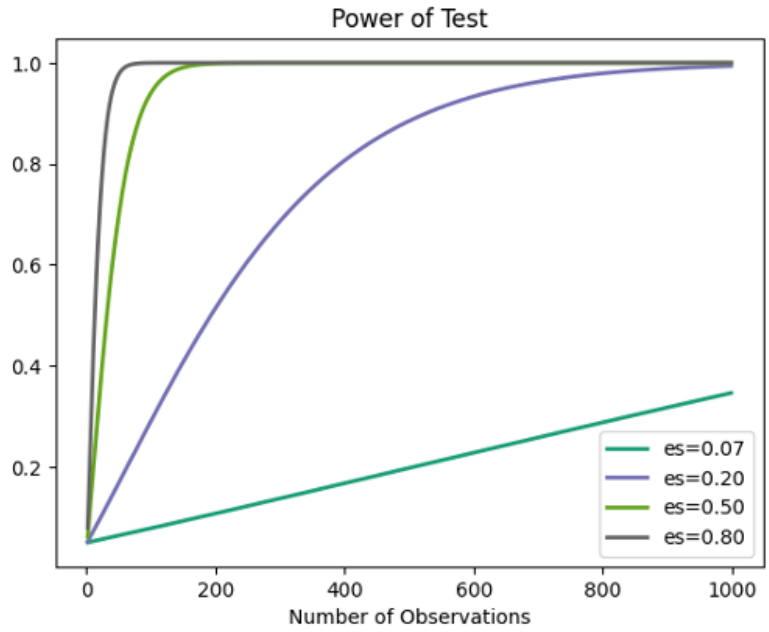
	Female	Male
Effect Size (Cohen's D)	0.06808694832729541	

Actual Sample Size	12609	52634
Required Sample Size	2099.052	8762.115

Before performing the statistical analysis on the statistical significance of sex on strip searches, we first computed the ratio of male and female strip search and their effect size. The resulted effect size of the sex groups is 0.07. The necessary sample size was then obtained in accordance with this effect size, the predetermined alpha and the power level. Based on the findings, a sample size of 2,099.052 females and 8,762.115 males is needed for each group. Since the dataset contains an adequate sample size of 12,609 females and 52,634 males, statistical tests can be run to detect this level of alpha, power, and effect size.

Meanwhile, the actual sample size in the dataset allows us to verify the statistical power. When the effect size and alpha are set to 0.8 and 0.05 respectively, the powers for both female and male strip searches are 1.00, indicating that there is a confirmed likelihood of detecting a significant difference in strip searches between the sex groups if the effect is present. It may be due to the large sample size in the dataset.

Figure 1: Power curve of male and female strip searches



With the industry standard of low (i.e., 0.2), medium (i.e., 0.5), and high (i.e., 0.8) effect sizes, we constructed a power curve to illustrate the effect size for male and female strip searches. The effect size in this dataset is extremely small, 0.07. Therefore, larger sample sizes are required in order to acquire the necessary statistical power. The current sample size in the dataset is sufficient to achieve the statistical power.

Additionally, another power analysis was performed to determine whether this study has adequate statistical power to perform the logistic analysis at hand. The result indicates that the power analysis was successful and that the power estimate is 1.000 with the significance level set at 0.05, desired effect size at 0.5 and sample size at 200. A power of 1.000 indicates that the study has sufficient statistical power to detect the effect size specified in the power analysis with a probability of 1, which means that there is a 100% chance of rejecting the null hypothesis when the effect size is present in the population. In other words, a power of 1.000 suggests that the study is well-designed and has enough statistical power to detect the effect size of interest, which means that the study is likely to produce reliable and meaningful results. It is important to note, however, that a power analysis alone does not guarantee the validity or generalizability of a study's results, and other factors such as sample representativeness, study design, and data quality should also be considered.

Table 2: Statistical power of the dataset used to build logistic regression

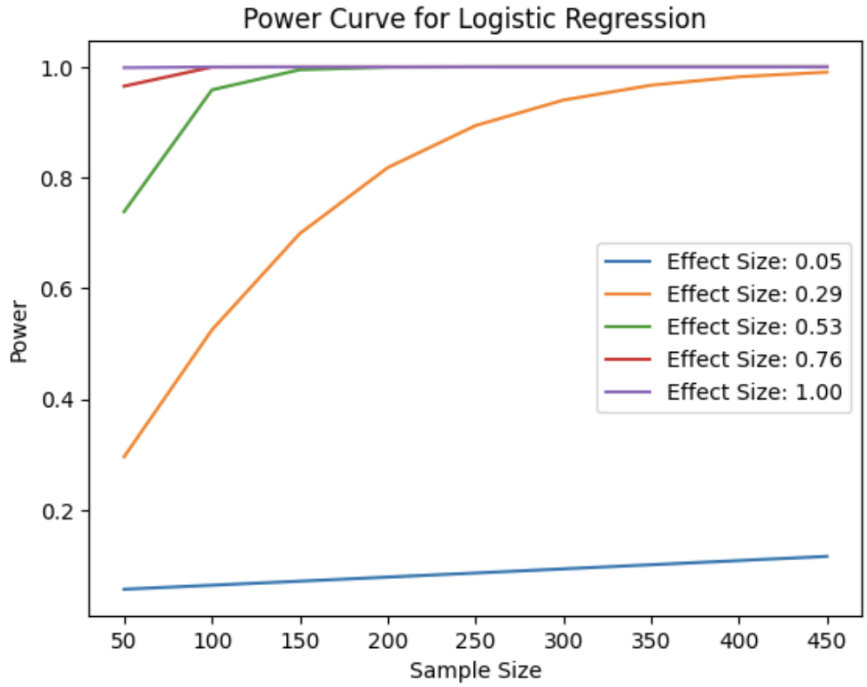
Power	1.00

The estimated sample size needed to achieve a desired statistical power for a two-sample t-test at 0.05 significance level, with a desired effect size of 0.5 and a sample size of 200 is 64.

Table 3: Required sample size to achieve desired statistical power

Sample size	63.765610587854056

Figure 2: Power curve for resampled data of logistic regression



Lastly, a power curve was created to visualize the effect size for each level of power. At the significance level (alpha), the effect sizes (a range of values from 0.05 to 1 with 5 values) and the sample sizes (an array of values from 50 to 450 with a step of 50) are defined.

Research Question 1

Research Objective & Question

As the literature review research supports, the police forces may have practiced discriminatory measures against certain profiles of people. The goal of the first research question is to understand if attributes such as age, race, and sex are important factors when analyzing if an arrested individual should be strip searched. If the result of the study proves to be positive to the question, there could be an indication that the Toronto Police Service is using profiling based on the individual's age, race, and sex to determine whether the same individual should be strip searched or not.

Research Question 1: Are the age, race, and sex of the arrested individual crucial factors when predicting the likelihood of this individual being strip searched?

The data that will be utilized to conduct the power analysis and logistic regression will be the attributes of "Age," "Race," "Sex," and "Strip Search.". The group predicts that these attributes will have significant importance when determining whether an individual will be strip searched. To perform the analysis, there are various assumptions made about this dataset which includes:

1. Linearity: There should be a linear relationship between the logit of the dependent variable and the independent variables. This means that the effect of the independent variables on the log-odds of the dependent variable should be constant across the range of the independent variables.
2. Independence of observations: The observations are independent of each other. This means that the data points should not be correlated with each other.
3. Sample size: The sample size is large enough to achieve stable and precise estimates of the regression coefficients.
4. No outliers: Outliers are not been taken into consideration when constructing the logistic regression
5. Normality of the residuals: The residuals are normally distributed. Deviation from normality may indicate the presence of outliers or influential observations, which can affect the model fit.
6. The dependent variable should be binary: Logistic regression assumes that the dependent variable is binary, it can take on only two values, 0 or 1.

Data Preprocessing

The relevant data provided by the "Arrests and Strip Searches" by the Toronto Police Service are relatively extensive. Since 'Age_group__at_arrest_' and 'Youth_at_arrest__under_18_years' have duplicated categories, it is required to preprocess the data. For 'Age_group__at_arrest_', 'Aged 17 years and under', 'Aged 17 years and younger' and 'Aged 65 and older', 'Aged 65 years and older' have the same meaning, therefore merging them into the same group respectively. Besides, there were 24 blank records, which is a small number compared to a total of 65,276 records, so it was removed. A new column ,'Age', is created with the cleaned data of 'Age_group__at_arrest_'. For 'Youth_at_arrest__under_18_years', 'Youth (aged 17 and younger)' and 'Youth (aged 17 years and under)' share the same meaning, therefore they were integrated into one group. A new column, 'Youth', is created with the cleaned data of 'Youth_at_arrest__under_18_years'. Then these categorical variables have been converted into numerical variables ready to be used for building the logistic regression

EDA

There are three main factors in which the study wants to investigate which includes, the age, race, and sex of an individual who have been conducted a strip search. Using exploratory data analysis, this research will perform a preliminary study on understanding the data pattern of each variable in relationship to the rate of being strip searched.

Table 4: Frequency of age groups and Strip Search rates

	17 and younger	18 - 24	25 - 34	35 - 44	45 - 54	55 - 64	65 and older	Total

0	2762	8692	18178	14144	8161	4228	1286	57451
1	280	1349	2771	2098	905	362	36	7801
Total	3,042	10,041	20,949	16,242	9,066	4,590	1,322	65,252
% Strip Searched	9%	13%	13%	13%	10%	8%	3%	12%

Table 4 shows the frequency of individuals that have been strip searched based on age groups. The total # of individuals who were strip searched in this dataset is 7801, which is 12% of the arrests (under the assumption that all records in the dataset represent an arrest). Diving into each age group, it appears that 18-24, 25-34, and 35-44 age groups have the highest rate of strip searches at 13%. This result can hint that the strip search practice between different age groups is conducted differently by the Toronto Police Force.

Figure 3: Bar chart of strip search rates by age group

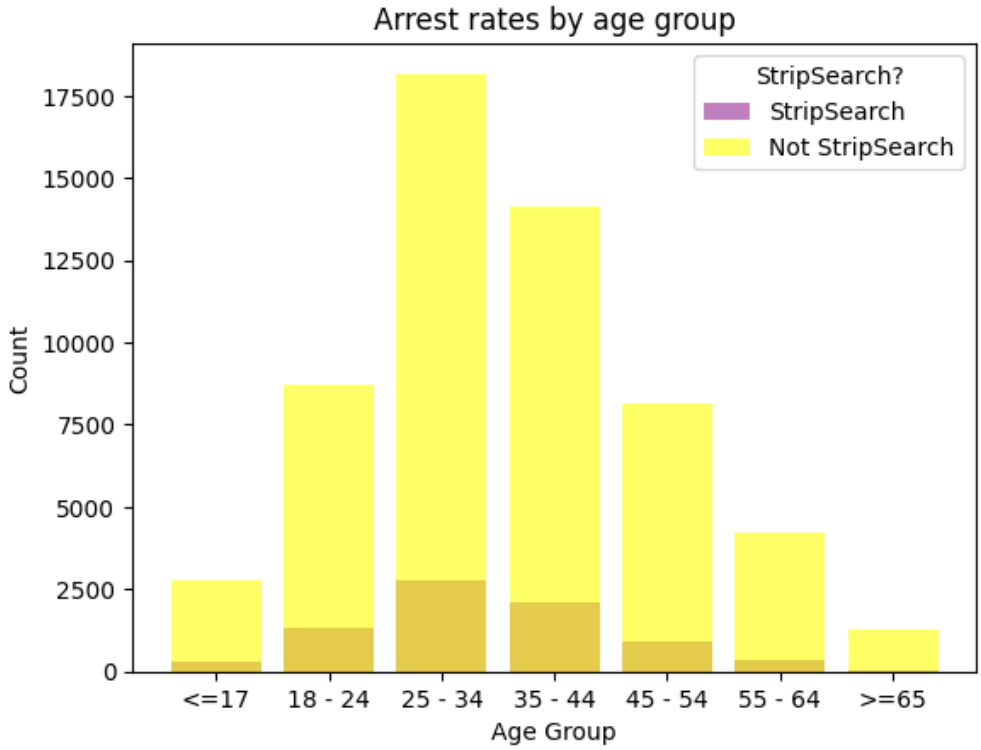


Table 4 is visualized as a bar chart to better understand the distribution of arrested and non-arrested people. Since people aged between 25 and 34 and 35 and 44 were the most populous in the dataset, they made up the majority of both strip searched and not strip searched cases. It also showed that young adults and middle-aged people (i.e., 18 – 24, 25 – 34, 34 – 44, 45 – 54) were more likely to be arrested. The largest disparity in arrests for a crime was among those aged 25 to 34, followed by those aged between 35 and 44. Meanwhile, teenagers (i.e.,

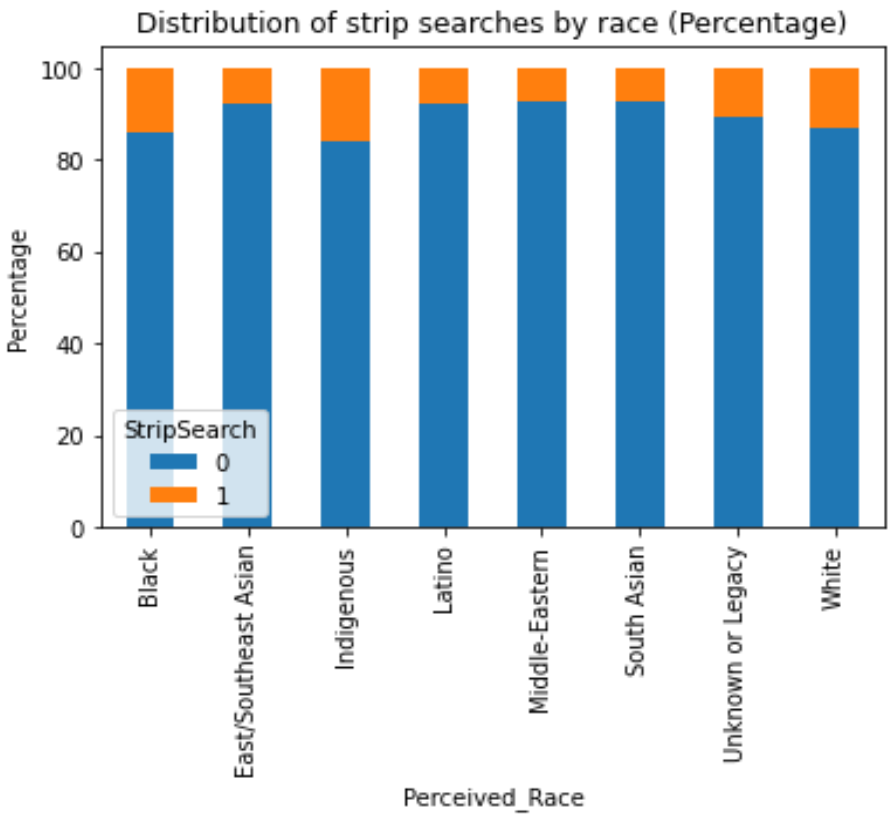
<=17) and the elderly (i.e., >=55) showed opposite results. Hence, these tables and figures align with our hypothesis that the Toronto Police Force may consider age when making strip search and arrest decisions, especially for youth and the elderly.

Table 5: Frequency of race groups and strip searches

	0	1	Total	% of StripSearched
Black	15,084	2,434	17,518	14%
East/Southeast Asian	4,071	341	4,412	8%
Indigenous	1,626	306	1,932	16%
Latino	1,636	132	1,768	7%
Middle-Eastern	3,009	228	3,237	7%
South Asian	3,356	257	3,613	7%
Unknown or Legacy	4,519	536	5,055	11%
White	24,147	3,566	27,713	13%
Total	57,448	7,800	65,248	12%

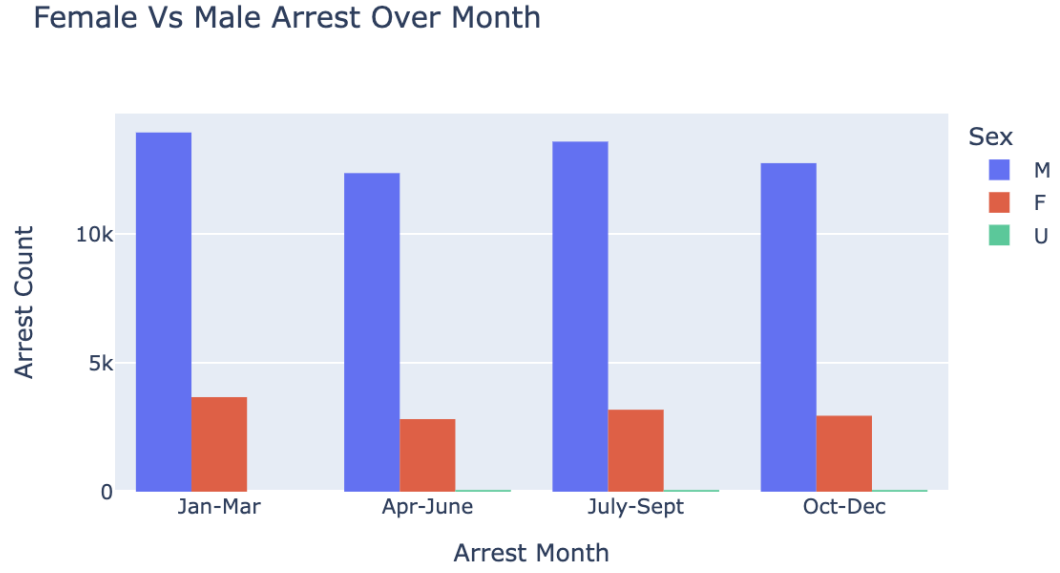
About 90% of the people in the dataset had not experienced a strip search, and only 10% had. It can be assumed from this finding that the act of strip search is not a normal practice. The Toronto Police Service may have used the act of strip search with consideration. Meanwhile, there were significant differences in ethnic distribution. Indigenous people experience 16% strip searches out of total arrest, Black experienced 14% and white experience 13%. White people contributed more than 40% of the dataset, and the remaining 7 racial groups made up the remaining 60%. The Indigenous, Black and Whites had the most strip searches. This can suggest there could be discriminatory measures when making the decision on conducting strip searches on these specific ethnicities.

Figure 4: Distribution of strip searches by race (Percentage)



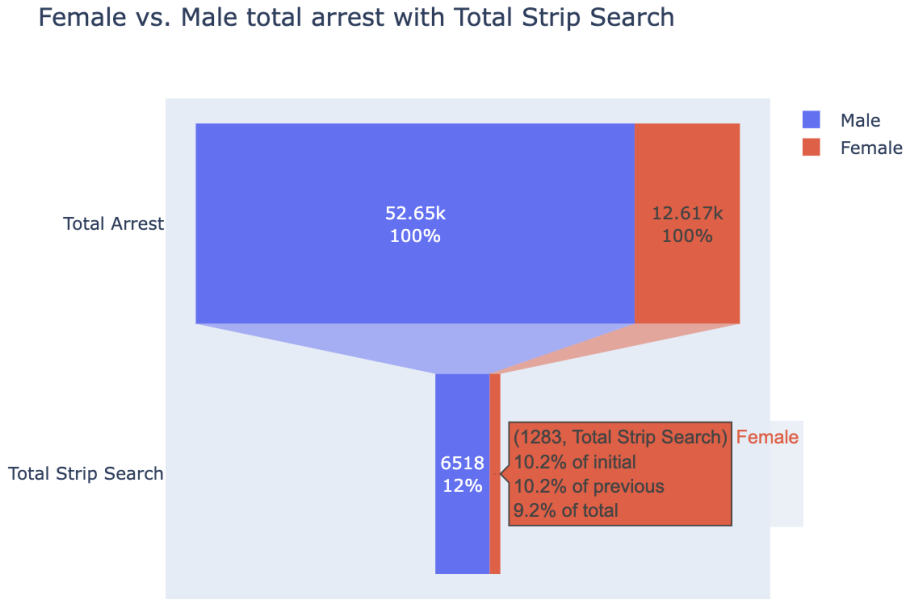
The study performed percentage statistics to normalize each ethnic group to see the percentage distribution of their strip searches. Figure showed that all groups experienced similar strip search percentages, with Indigenous, Black, and White experiencing the highest rates. This could be an indication that there was minor discrimination against specific racial groups when conducting strip searches.

Figure 5: Female Vs Male Arrests across each quarter for 2020 and 2021



While working under the assumption that all records with an unique “Arrest_ID” will be classified as arrested. It appears that there is no difference between female and male individuals and is not only apparent for strip searches. It is evident in figure 2 that the rate of arrest of male individuals is also significantly higher. This then triggers the question whether the rate of arrest to the rate of strip search have any significant difference between male and female.

Figure 6: Female vs. Male Total Arrest to Total Strip Search fro 2020 and 2021



The funnel in figure 6 shows the rate of percentage of total female vs. male individuals which were arrested and the rate of percentage from that arrested total that performed a strip search. Numerical percentage and total are listed in the below table.

Table 6: Female vs. Male numerical total details on arrested and strip searches and the rate of strip search for 2020 and 2021.

Sex	Strip Search	Total Strip Search	Total Arrested	Rate of Strip Search
Female	0	0	11,334	
Female	1	1283	1,283	10.17%
Male	0	0	46,132	
Male	1	6518	6,514	12.38%

The above results from figure 3 and table 2 shows that although in number the total of male individuals arrested and strip searched are significantly higher than female individuals. The percentage of rate of strip search from the arrest total is in comparison have a much less significant difference. The rate of strip search from total arrest for female individuals is 10.17%, where a total of 1283 individuals were strip searched from the total of 12617 individuals that were arrested. The rate of strip search from total arrest for male individuals is 12.38%, where a total of 6518 individuals were strip searched from the total of 52646 individuals that were arrested.

Methods

Logistic Regression

The approach which this study took to investigate the research question is using logistic regression to measure if the attributes including 'Age', 'Race' and 'Sex' are able to predict factors when it comes to measuring if an individual will be strip searched.

> **Null hypothesis (H_0): Attributes include 'Age', 'Perceived_Race', and 'Sex' can NOT perdict the 'StripSearch' with great accuracy**
> **Alternative hypothesis (H_1): Attributes include 'Age', 'Perceived_Race', and 'Sex' can perdict the 'StripSearch' with great accuracy**

The process of creating a logistic regression model started with splitting the dataset into train and test dataset with the data that has already been cleaned. To prevent overfitting of the model, the decision was to split 30% of the dataset for testing and 70% for training. The data was split at random. The independent variables identified from the dataset were 'Age', 'Perceived_Race', and 'Sex'. The target predictor was 'Strip Search'. The training of the logistic regression model was done using the 'sklearn.linear_model' package and its LogsticRegression function. The trained model was named 'lr_model'. Once the model has been completed, it needs to be tested for accuracy and other metrics.

The first set of metrics is the accuracy score.

Table 7: Accuracy score of the original logistic regression model

Accuracy Score	0.879648549243972

The accuracy score of 0.8796 indicates the proportion of correctly classified instances in the test data by the logistic regression model. The model predicted the outcome correctly for approximately 88% of the test instances. This is a good result, however accuracy alone may not be sufficient to evaluate the performance of a model, especially if the classes are imbalanced or the cost of misclassification varies significantly between the classes.

To understand the results better, a confusion matrix was also created. And through the confusion matrix, it was noticed that there is an imbalance in the sampling. This could be the result of no or too less sampling to negative results in the trained testset. As you can see, there are 17220 true negatives (TN), 0 false positives (FP), 2356 false negatives (FN), and 0 true positives (TP), which means the model can not identify true positive cases. Therefore in this case, the F1-score would also equal 0.

Table 8: Confusion matrix of original logistic regression model

Actual/Predicted	Negative (0)	Positive (1)
Negative (0)	17220	0
Positive (1)	2356	0

Knowing this inaccuracy in the modeling of this dataset. The decision was made to oversample the minority (1 for strip search) to gain a more balanced dataset. After performing this process, the results are drastically different and have a better representation for the accuracy of the logistic regression mode. Actual results of the corrected model will be discussed more in detail in the “Results & Finding” section.

After which, to ensure that there is no overfitting in the model. A cross-validation test was taken to ensure that there is no overfitting for the logistic regression model create. he cross-validation was performed using a K-fold cross-validation with K=5, meaning the data was split into 5 equal parts and the model was trained and evaluated 5 times, with each part being used once as the validation set and the remaining parts as the training set.The cross-validation scores are a list of the accuracy scores obtained for each of the 5 iterations. In this case, the scores are [0.88046893, 0.88039231, 0.88045977, 0.88045977, 0.88045977]. The results suggest that the model is performing consistently and there is no sign of overfitting.

Table 9: Cross-validation score for oversampled logistic regression model

Cross-validation scores
0.88046893
0.88039231
0.88045977
0.88045977
0.88045977

Next, a prediction interval is calculated for the logistic regression model to estimate the range of values in which a future observation is likely to fall, given the 95% level of confidence. Unlike a confidence interval, which estimates the precision of the mean value of a random variable, a prediction interval estimates the range of values that a new observation will likely fall in, given the values of the input features. This is particularly useful in logistic regression, where the output variable is binary (e.g., 0 or 1) and the model predicts the probability of the positive class (e.g., 1). Results for the actual prediction interval will be discussed in the Results & Findings section. The results of the prediction interval are then plotted into a visualization to show trends.

Furthermore, odds ratios are being calculated to measure the strength of association between the predictor variables (also known as independent variables or features) and the response variable (also known as the dependent variable or outcome). More specifically, the odds ratio measures the change in odds of the response variable associated with a one-unit change in the predictor variable, while holding all other predictors constant. odds ratio greater than 1 indicates a positive association between the predictor variable and the response variable, while an odds ratio less than 1 indicates a negative association. In this case, both odds ratios are greater than 1, which suggests that both predictor variables are positively associated with the positive outcome.

Lastly, the confidence interval is calculated to tell us the range of values that the true value of the model parameter is likely to fall within, with a 95% confidence level. The confidence interval gives us a measure of the uncertainty around the estimate of the model parameter. A narrower interval indicates less uncertainty and more precise estimate, while a wider interval indicates more uncertainty and less precise estimate.

Results & Findings

Table 10: Confusion Matrix of the updated logistic regression model

Actual/Predicted	Negative (0)	Positive (1)
Negative (0)	8307	8913
Positive (1)	967	1369

The confusion matrix of the logistic regression model shows that there are 1389 cases of true positive, 8307 cases of true negative, 8913 cases of false positive, and 967 cases of false negative. Which means the model correctly predicted 1389 cases that were actually positive, correctly predicted 8307 cases that were actually negative cases, incorrectly predicted 8913 cases that there were positive cases when there actually weren't, and incorrectly predicted 967 cases that there were negative cases when there actually were positive cases. In conclusion the logistic regression model did not perform as well as the hypothesis had predicted based on the age, race, and sex attributes.

Table 11: Error rate, Precision, Recall & F1-Score for the updated logistic regression mode

Accuracy	0.4953003677973028
Error rate	0.5046996322026972
Precision	0.13482818870122307
Recall	0.5895585738539898
F1-score	0.21946595038710698

The above metrics to the logistic regression model are all showing disappointing results towards the performance. The accuracy shows the proportion of correctly classified instances out of the total number of instances. An accuracy score of 0.495 indicates that the model is only par with random at predicting the rate of strip search. The error rate on the other hand, is the proportion of incorrectly classified instances out of the total number of instances. An error rate of 0.505 means that the model is on par than random at predicting the target variable. The precision is the rate of true positives (correctly classified instances of the positive class) out of all instances classified as positive. A precision score of 0.135 indicates that of all instances predicted as positive, only 13.5% were actually positive. The recall, conversely, is the rate of true positives out of all actual positive instances. A recall score of 0.590 indicates that of all actual positive instances, the model correctly identified 59% of them. Last, the F1-score is the harmonic mean of precision and recall, and provides a balanced measure of the two metrics. An F1-score of 0.219 indicates that the model is not performing well on both precision and recall. In summary, all metrics have proven the model has been underperforming.

Table 12: Classification Report of the updated logistic regression model

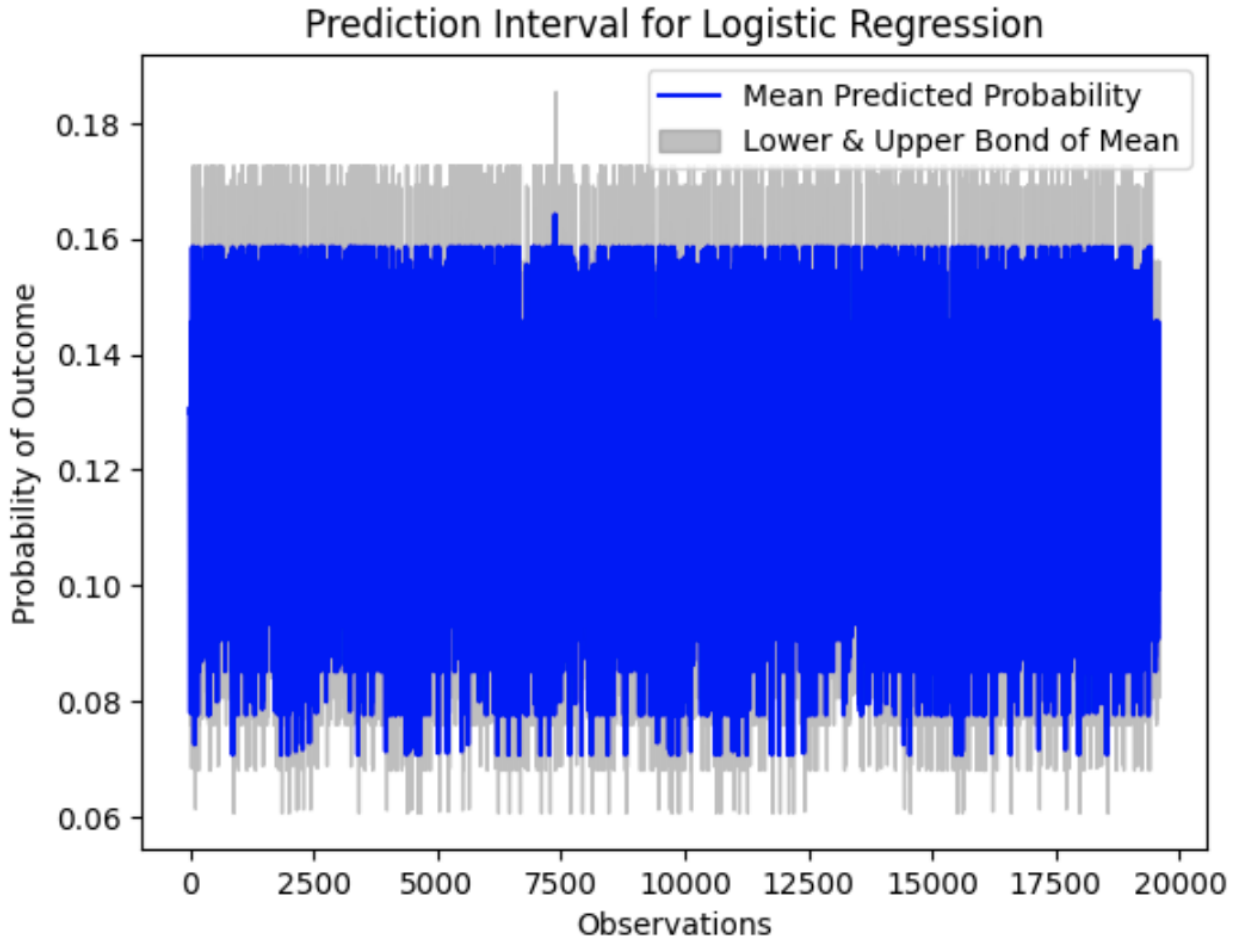
	Precision	Recall	F1-Score	Observations
No StripSearch	0.90	0.48	0.63	17220
StripSearch	0.13	0.59	0.22	2356

The classification report provides an evaluation of the logistic regression model more in detail. The precision for "No StripSearch" is relatively high at 0.90, which means that when the model predicts if an individual is not going to be strip searched, it is correct 90% of the time. However, the recall for "No StripSearch" is relatively low at 0.48, which means that the model is not able to identify all cases where a person will not be strip searched.

Furthermore, the precision for "StripSearch" is extremely low at 0.13, which means that when the model predicts if an individual will be strip searched, it is correct only 13% of the time.

However, the recall for "StripSearch" is higher in comparison at 0.59, which means that the model is able to identify a majority of the cases where a person will be strip searched. The F1-score measure of the balance between the two. The F1-score for "No StripSearch" is 0.63, which indicates a moderate balance between precision and recall, while the F1-score for "StripSearch" is only 0.22, which indicates that there is a significant imbalance between precision and recall for this category. Overall, this could mean that with the existing data, age, sex, and race are not good indicators for predicting if an individual will be strip searched.

Figure 7: Prediction curve plot for updated logistic regression



Using the data from the prediction interval, the above figure of the prediction curve provides the mean predicted probability with 95% confidence level (alpha = 0.05). The mean predicted probability ranges from 0.07 to 0.16 and the lower and upper boundary of the prediction interval ranges from 0.06 to 0.17. The prediction interval further proves that the logistic regression is not performing well and the null hypothesis could be proven to be true.

Table 13: Logit regression results for updated logistic regression model

	coef	std err	z	P>|z|	0.025	0.975
const	0.0789	0.023	3.437	0.001	0.034	0.124
Age	-0.1227	0.006	-21.36	0.0	-0.134	-0.111
Perceived_ Race	0.0078	0.002	3.341	0.001	0.003	0.012
Sex	0.2385	0.019	12.872	0.0	0.202	0.275

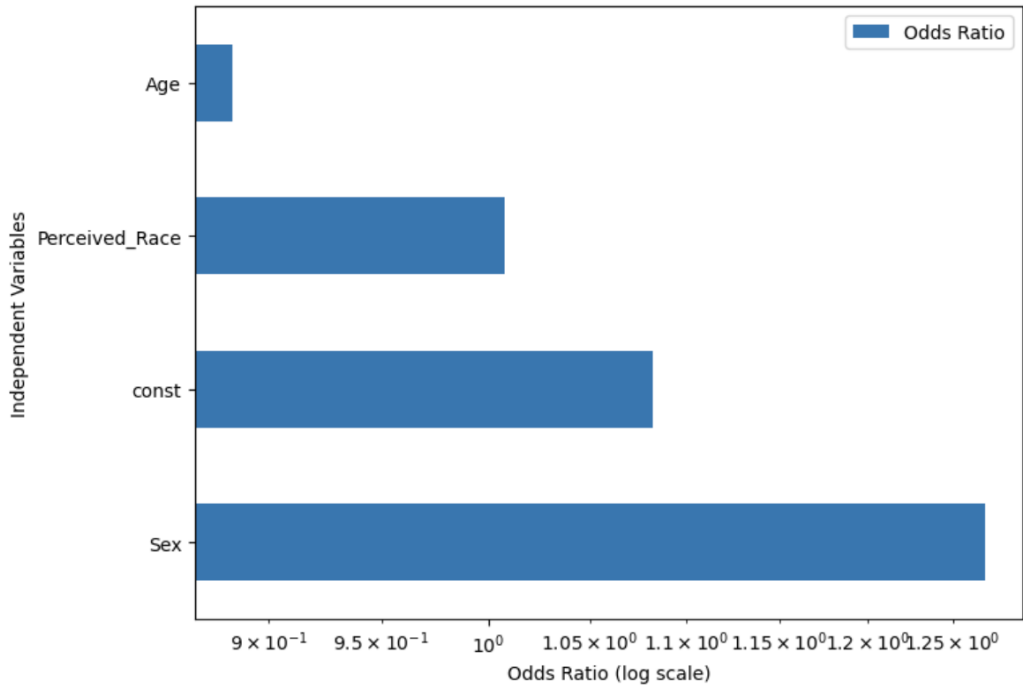
The above table can help us understand the performance of each attribute better. The intercept (const) is 0.0789 which is statistically significant ($P<0.05$) and has a 95% confidence interval of [0.034, 0.124]. Age has a negative coefficient (-0.1227) which is statistically significant ($P<0.05$) and has a 95% confidence interval of [-0.134, -0.111]. Perceived_Race has a positive coefficient (0.0078) which is statistically significant ($P<0.05$) and has a 95% confidence interval of [0.003, 0.012]. Sex has a positive coefficient (0.2385) which is statistically significant ($P<0.05$) and has a 95% confidence interval of [0.202, 0.275].

Table 14: Odds ratios for updated logistic regression model

Age	Perceived Race	Sex
0.88449636	1.00787037	1.26931151

The odd ratio means that the ‘Age’ variable increases by one unit, the odds of the dependent variable, ‘Strip Search’ occurring decrease by a factor of 0.88449636. If the ‘Perceiced_Race’ variable increases by one unit, the odds of the dependent variable occurring increase by a factor of 1.00787037. Similarly, if the ‘Sex’ variable increases by one unit, the odds of the dependent variable occurring increase by a factor of 1.26931151. This could mean the older an individual is the less likely the individual will be strip searched. Male individuals are more likely to be strip searched and the perceived race of the individual does not greatly increase the rate of strip search.

Figure 8: Odds ratio per attribute to the outcome



When plotting the odds ratio in a visualized format, it is apparent that the factor with the heaviest influence on the probability of being strip searched is the sex. On average male individuals who are arrested have a greater chance of being strip searched.

Research Question 2

Research Objective & Question

To further investigate the issue of gender discrimination, we want to assess the difference in strip search counts between females and males while accounting for the impact of age. As we learned in the previous study, age was a crucial variable that could influence the number of strip searches. Much fewer strip searches were conducted among the youngest and oldest populations. Therefore, if we do not account for the age variable, the results of the relationships between females and males could be unreliable. For instance, the dataset might contain a greater proportion of middle-aged males and younger females, which led to higher male strip searches. In that situation, we might conclude that gender discrimination occurred when strip searches were performed, which might be inaccurate. We aim to identify any statistically significant differences in strip search counts between males and females that are not the result of age effects by adjusting for age. We assume that females might undergo fewer strip searches than males while adjusting for the influence of age because it is a common stereotype that men are more violent and vicious than women. Therefore, our second research question is:

> **Research Question 2: Is there a difference in the strip search counts between female and male, controlling for age?**

Data Preprocessing

Similar to the previous report, we combined duplicated categories with the same meaning, such as ‘Aged 17 years and under’ and ‘Aged 17 years and younger’. Besides, we removed the blank records in age (i.e., 24 rows) and created a new column, ‘Age’, with the cleaned data of ‘Age_group__at_arrest_’. Also, we created a dummy column for Age (i.e., Age_Dummy) that serves as the control variable for the ANCOVA analysis.

Additionally, the statistical analysis in this paper seeks to examine the differences between females and males and whether one of these groups was being discriminated against by the Toronto Police Service. Therefore, ‘U’ in the ‘Sex’ was removed and would not be considered in the following research.

EDA

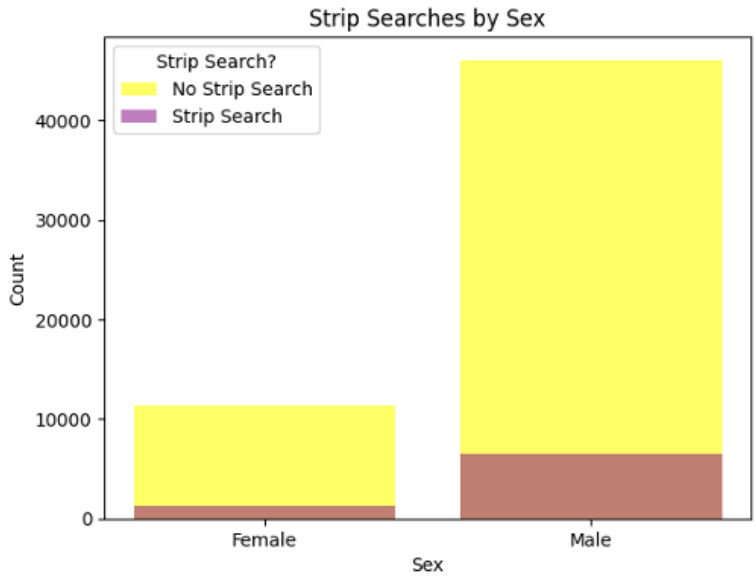
To better comprehend the three variables, we carried out a thorough descriptive analysis.

Table 15: Frequency of sex and strip searches

Sex/ Strip Search	0	1	Total
F	11,326	1,283	12,609
M	46,116	6,518	52,634
Total	57,442	7,801	65,243

In the dataset, more than 80% of the records were male and about 12% had experienced strip searches (1).

Figure 9: Bar chart of strip searches by sex



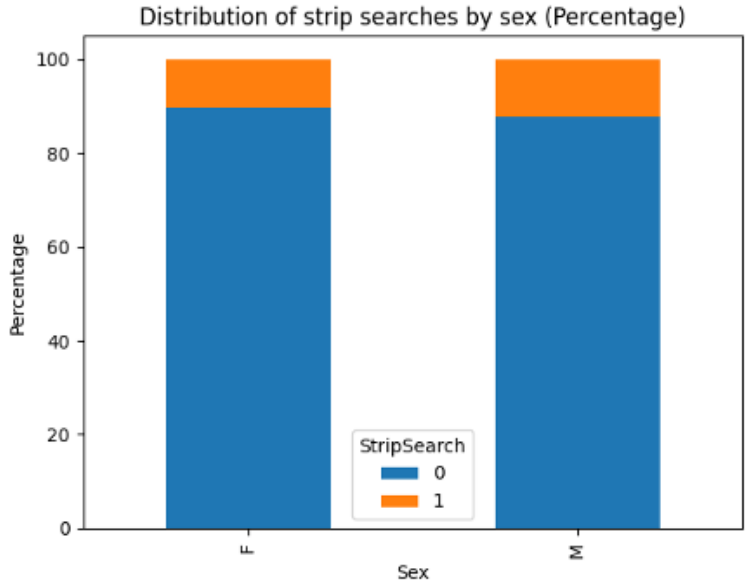
To better understand the distribution of strip searches among males and females, we presented table 15 as a bar chart. Male strip searches were more common than female ones since there were more men than women in the population. Figure 9 supports our hypothesis that men were subjected to more strip searches than women.

Table 16: Percentage of strip searches grouped by sex

Sex/ Strip Search	0	1
F	89.824728	10.175272
M	87.616370	12.383630

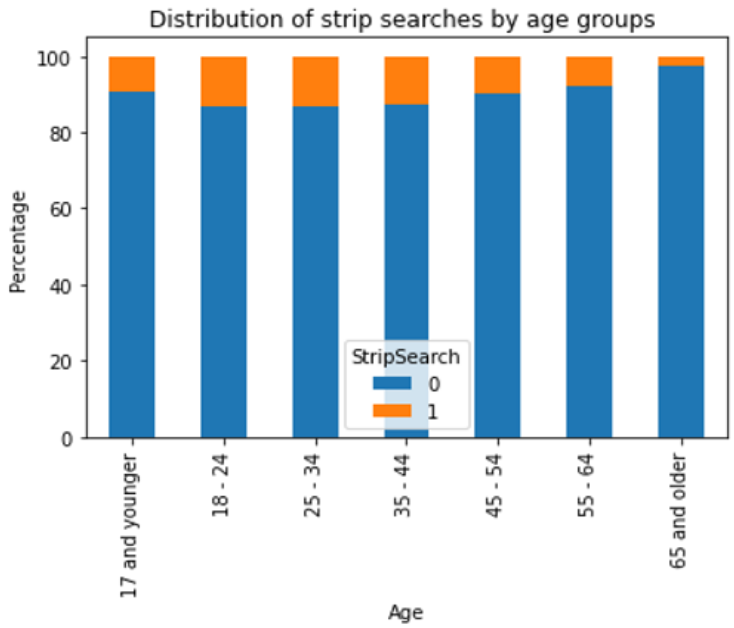
Since the number of females and males differ significantly, we presented percentage statistics for 'Male' and 'Female' to distinguish the differences of strip searches between them. Table 16 revealed that the percentage of strip searches was similar between females and males, with males having a slightly higher percentage of strip searches.

Figure 10: Distribution of strip searches by sex



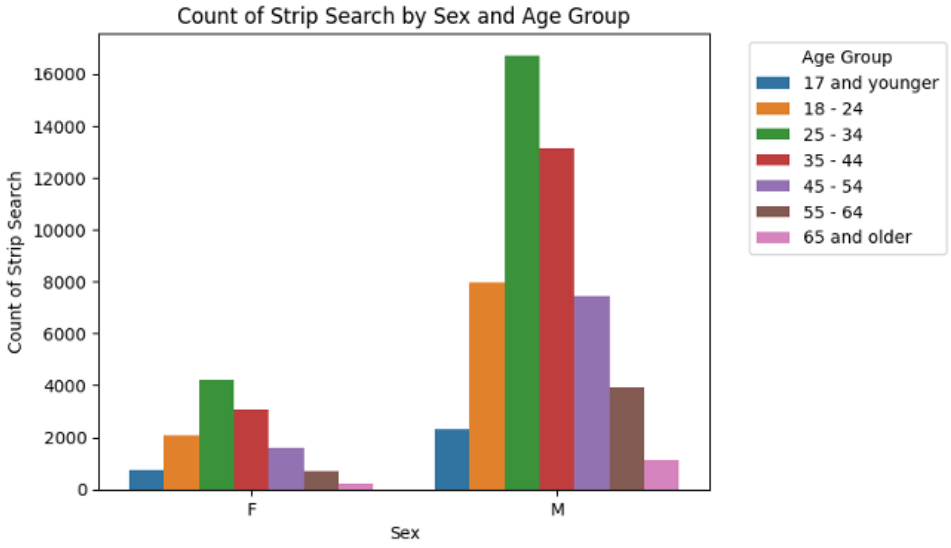
We visualized table 16 as a bar chart to better observe the variations between male and female strip searches. Figure 10 showed that the male strip searches had a slightly higher percentage than that of the females.

Figure 11: Distribution of strip searches by age groups (Percentage)



As with the sex groups, the frequency of each age group was uneven, with people aged 18 to 44 contributing more than 70% of the dataset. Hence, we have to translate them into percentages to better understand the distribution of strip searches per age group. The percentage statistics show that people aged 18 to 44 had the highest strip searches while youth (i.e., 17 and younger) and the elderly (i.e., 65 and older) had the lowest. Along with this descriptive analysis and the ANOVA test in the previous report, age had a statistically significant influence on strip searches, notably for youth and the elderly. Therefore, we would like to control for the age variable and concentrate on the examination of the relationship between sex and strip searches.

Figure 12: Strip searches by sex and age group



We visualized sex, age, and strip searches as a bar chart to investigate their relationship. We can observe that males had higher strip searches for all age groups. It might be due to the fact that males had an overall greater count. Therefore, figure 12 does not indicate how age affects the relationship between sex and strip searches. To do this, we must run the statistical tests.

To acquire more understanding into these variables, we performed three t-tests to measure their differences in strip searches. 'StripSearch' is the dependent variable, while 'Age' and 'Sex' are the independent variables. The two independent variables are in independent groups.

In the following t-tests, we assume that the data was continuous, randomly sampled, had similar or same variances, and had a roughly normal distribution in each independent group. Moreover, the data did not have outliers. and the sample size was sufficient.

Table 17: t-test result (Strip searches for males and females)

t-test	
t-statistic	p-value
-7.237309885776752	4.73893523159573e-13

Null hypothesis (H_0): There is no significant difference in the number of strip searches between males and females.
Alternative hypothesis (H_1): There is a significant difference in the number of strip searches between males and females.

The p-value is extremely small, 4.73e-13, implying a statistically significant difference in the average number of strip searches between females and males. Hence, the null hypothesis is rejected. We have significant evidence to conclude that the number of strip searches was different between females and males. Also, the t-statistics is -7.237, which means that females had fewer mean strip searches than males. While there could be many other factors, such as the type of crime, contributed to this result, it might imply that police treat females and males differently when conducting strip searches, with more strip searches for males compared to females.

Table 18: t-test result (Strip searches for youth and non-youth)

t-test	
t-statistic	p-value
-5.343426792919216	9.715337378803637e-08

Null hypothesis (H_0): There is no significant difference in the number of strip searches between youth (i.e., 17 and younger) and non-youth.
Alternative hypothesis (H_1): There is a significant difference in the number of strip searches between youth (i.e., 17 and younger) and non- youth.

The p-value is extremely small, 9.72e-08, implying a statistically significant difference in the average number of strip searches between youth (i.e., 17 and younger) and non-youth. Hence, the null hypothesis is rejected. We have significant evidence to conclude that the number of strip searches was different between youth and non-youth. Also, the t-statistics is -5.343, which means that youth (i.e., 17 and younger) had fewer mean strip searches than non-youth. While there could be many other factors, such as the type of crime, contributed to this result, it might imply that police treat youth (i.e., 17 and younger) better than non-youth when strip searches were considered.

Table 19: t-test result (Strip searches for the elderly and non-elderly)

t-test	
t-statistic	p-value
-20.221367204581707	6.970766903447176e-81

Null hypothesis (H_0): There is no significant difference in the number of strip searches between the elderly and non-elderly.
Alternative hypothesis (H_1): There is a significant difference in the number of strip searches between the elderly and non-elderly.

The p-value is extremely small, 6.97e-81, implying a statistically significant difference in the average number of strip searches between the elderly and non-elderly. Hence, the null hypothesis is rejected. We have significant evidence to conclude that the number of strip searches was different between the elderly and non-elderly. Also, the t-statistics is -20.221, which means that the elderly had significantly fewer mean strip searches than the non-elderly. While there could be many other factors, such as the type of crime, contributed to this result, it might imply that police treat the elderly better than the non-elderly when strip searches were considered.

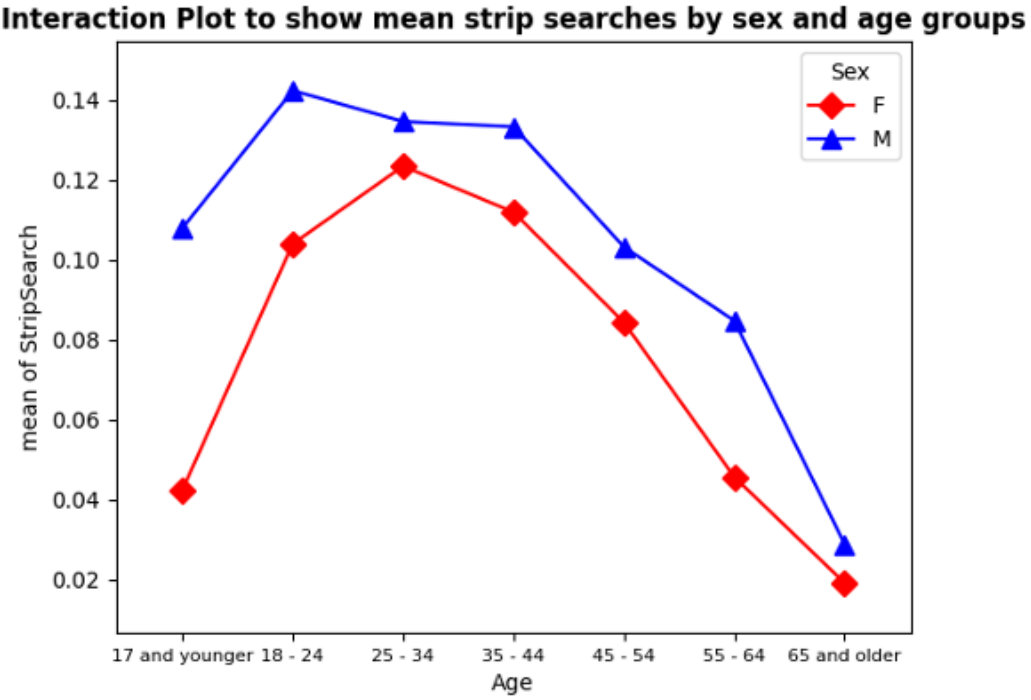
Therefore, according to the three t-tests, we found that the elderly and youth experienced significantly fewer strip searches, which align with our previous study. Also, females were subjected to much fewer strip searches than male, which is consistent with our hypothesis that the Toronto Police Service might practise gender discrimination when considering strip searches. Hence, these have piqued our curiosity in looking into the relationship between strip searches and sex while controlling for age groups.

Research Method

We first developed an interaction plot to visualize the relationships of strip search, age, and sex groups. An ANCOVA test was then conducted to investigate the relationship between strip searches and sex groups after controlling the effect of the age groups. 'StripSearch' is the dependent variable, 'Sex' is the independent variable, and 'Age' is the controlling variable. It helps to determine whether mean strip searches between females and males are statistically significant while controlling for age groups.

Results & Findings

Figure 13: Interaction plot of mean strip searches by sex and age groups



We created an interaction plot to visualize the relationship between strip search, sex, and age groups. The higher the mean of strip search, the higher probability of being strip searched, vice versa. Figure 13 shows that, across all age groups, females were less likely than males to be subjected to a strip search. While the elderly had the smallest gap, teenagers had the biggest difference in mean strip search counts between males and females. However, it does not reveal whether males had a statistically significant greater strip search count than females if we control the age groups. Therefore, we need to conduct the ANCOVA test.

We conducted an ANCOVA test to assess the strip searches between females and males while controlling for age groups. 'StripSearch' is the dependent variable, 'Sex' is the independent variable, and 'Age' is the controlling variable.

In the following ANCOVA test, we assume that the data was randomly sampled, had similar or same variances, and had a roughly normal distribution in each independent group. Additionally, the dependent (i.e., strip search) and control variables (i.e., age) have a linear relationship. Also, there are no interactions between the independent (i.e., sex) and control variables (i.e., age). Moreover, the control variables (i.e., age) are not strongly correlated to one another.

Null hypothesis (H_0):
There is no significant effect of sex on strip searches, controlling for the age groups (i.e., Mean strip search counts for females = Mean strip search counts for males).
Alternative hypothesis (H_1):
There is a significant effect of sex on strip searches, controlling for the age groups (i.e., Mean strip search counts for females ≠ Mean strip search counts for males).

Table 20: ANCOVA test result (Strip Search, Sex, Age)

	SS	DF	F	PR(>F)	np2
Sex	5.740380	1	54.676508	1.437907e-13	0.000837
Age	13.865291	1	132.065410	1.549181e-30	0.002020
Residual	6849.420948	65240	NaN	NaN	NaN

The ANCOVA test compared the differences in mean strip searches between females and males to see if they were statistically significant after controlling for age groups. The F-statistics for the sex were as high as 54.68, meaning large differences in strip searches between females and males with a lower probability of observing it by chance. Besides, the sex variable has a sum of squares of 5.74, degree of freedom of 1. The corresponding p-value is extremely low at 1.44e-13, which indicates a significant effect of sex on strip searches when controlling for the age. Hence, the null hypothesis is rejected. We have significant evidence to conclude that the strip search counts differed significantly between females and males, controlling for the age groups. Additionally, the F-statistics for the age were as high as 132.07, meaning large differences in strip searches between different age groups with a lower probability of observing it by chance. Besides, the age variable has a sum of squares of 13.87, degree of freedom of 1. The corresponding p-value is extremely low at 1.55e-30, which indicates a significant effect of age on strip searches. It also explains why age was the control variable in our study, which looked solely at the impact of sex on strip searches. Moreover, the residual sum of squares is 6849, which is relatively large. It indicates that there are many unexplained noises in the model. Therefore, the test result suggests that sex is a significant predictor of strip searches while controlling for age, which is the same as our initial hypothesis. However, more variables are needed to explain the noises, such as the type of crime.

From the descriptive analysis, more males than females had strip searches across all age categories. Yet, the percentages of male and female strip searches were roughly the same. Therefore, it is difficult to conclude whether the Toronto Police Office handled a sex group differently. From the t-tests, females had statistically significantly fewer strip searches than male, which is consistent with our hypothesis that the Toronto Police Service might practise gender discrimination when conducting strip searches. Meanwhile, age also played a role in the decision to do a strip search. Hence, the ANCOVA test was conducted to examine the sole effect of sex on strip search after considering the impacts of age groups. In the ANCOVA test, the null hypothesis is rejected. We have strong evidence to conclude that the strip search counts differed significantly between females and males, controlling for the age groups. Therefore, there was a strong relationship between sex groups and strip searches, which aligns with our hypothesis. One of the reasons for this phenomena might be due to the gender discrimination of the Toronto Police Office when conducting strip searches. Therefore, the descriptive analysis, t-tests and ANCOVA test agreed that females were considerably less likely than males to undergo strip searches.However, it should be highlighted that these analyses can only indicate that sex groups were a predictor of the strip search, not that they caused the decision to do the search.

Discussion

The analysis of a logistic regression model to predict whether a person will be subjected to a strip search based on their age, sex, and race. The confusion matrix demonstrates that the model incorrectly predicted 8913 cases that were positive when they actually weren't, and incorrectly predicted 967 cases that were negative when they actually were positive cases. The model correctly predicted 1389 cases that were actually positive, 8307 cases that were actually negative cases, and 967 cases that were actually negative when they were actually positive. The model underperformed, as evidenced by its low error rate, precision, recall, and F1-score. According to the classification report, a person's age, sex, or race are not reliable determinants of whether they may be subjected to a strip search.The results of the logit regression and the prediction curve provided additional evidence for the model's poor performance. The logistic regression model did not perform as well as the hypothesis had indicated, in order to sum up. Which may also imply that it is not necessary to consider an individual's age, race, or sexual orientation when determining whether a strip search will take place. Therefore, there is insufficient evidence to conclude from this research that the Toronto Police Service uses profiling to decide whether to conduct strip searches.

In all age groups, girls were less likely than males to be subjected to a strip search, according to the descriptive analysis of the data and a plot. If age groups were taken into account, the plot did not indicate whether males underwent considerably more strip searches than females. So, controlling for age groupings, an ANCOVA test was used to investigate the association between sex and strip searches. Even after accounting for age groupings, the research still showed a

substantial sex effect on strip searches. The study reveals that women were significantly less likely than men to be subjected to strip searches, and the article draws the conclusion that there may be gender discrimination in the strip search procedures used by the Toronto Police Office. It's crucial to remember that the research can only show that sex groups were a predictor of the strip search, not that they were the reason the search was decided to be conducted.

Limitation and Future Work

As this dataset does not contain any continuous variables, we chose the boolean attributes, 'StripSearch', as the dependent variables for the research questions. We considered creating new columns for the dataset, such as percentage counts of female strip searches, but it still relies on the 'StripSearch' column, which is more of a descriptive analysis. Since the dependent variables are not continuous, we were unable to develop different types of graphs and analyzes, such as boxplots and standard deviations, which limits the diversity of EDA. Besides, the boolean variables may reduce the accuracy of all models and tests because they can only be viewed as probabilities of occurrence rather than true continuous variables.

Additionally, there are more variables than gender, age, and race that affect the number of strip searches, especially the type of crime and actions at arrest. Hence, we can continue to study the relationship between more variables by building more statistical models, such as the chi-squared model of the relationship and dependence between the type of crime and race, or incorporate more variables into the logistic regression model. These tests will provide a deeper understanding of their relationship.

Besides, several of the statistical tests' assumptions may not hold true given the dataset. The data may not have the same or similar variance as assumed by the T-test, logistic regression, and ANCOVA tests. Additionally, the ANCOVA test makes the false assumption that the dependent (i.e., strip search) and control variables (i.e., age) have a linear relationship. These could make statistical tests less accurate.

Also, this report provides a preliminary analysis and investigation of the relationship among gender, age, race, and the number of strip searches. It does show that there is a strong relationship between them and predicts both dependent variables. Yet, these analyses could not draw conclusions about casual relationships. We can only know which groups have significantly higher strip searches, but we cannot conclude that Toronto Police Service discriminates against them.

Therefore, after confirming their relationships, we can determine causality by conducting experiments, such as randomized controlled trials, which helps to answer our research questions of finding out whether specific groups are being discriminated against by the Toronto Police Service when deciding to strip search.

Conclusion

This report examines whether the Toronto Police Service practiced discriminatory measures against certain groups, such as the male individuals, youth, the elderly, and the Black, when conducting strip searches. Although none of the analyses were able to confirm a causal relationship, the number of strip searches were statistically significantly higher among male, while controlling for the age groups. The combinations of age, sex, and race, however, may not be accurate predictors of strip searches because they yield underwhelming results, according to the logistic regression model. These results partly align with our initial hypothesis that these groups were statistically significantly different from the others. Also, many noises affected the accuracy of our models. Hence, it is suggested that further research using more variables and different models is needed to determine causality and to see if the Toronto Police Service discriminates against any groups. We hope the Toronto Police Service can use these insights to improve its training and prevent discrimination when making decisions.

Bibliography

[1]Adrian Ghobrial. 2021. CityNews. *toronto.citynews.ca*. Retrieved February 21, 2023 from https://toronto.citynews.ca/2021/03/02/strip-searches-by-toronto-police-drop-dramatically-in-february/

[2]Wendy Gillis. 2020. “Clearly, we were doing it wrong”: Toronto police are doing far fewer strip searches under strict new rules, Interim chief says. *thestar.com*. Retrieved February 20, 2023 from https://www.thestar.com/news/gta/2020/11/23/clearly-we-were-doing-it-wrong-toronto-police-doing-far-fewer-strip-searches-under-new-rules-interim-chief-says.html

[3]Monika Lemke. 2022. Policing Toronto: Strip Searching in a Divided City - The Bullet. *Socialist Project*. Retrieved February 21, 2023 from https://socialistproject.ca/2022/07/policing-toronto-strip-searching-in-a-divided-city/