How Much of Strip Search Decision-Making is Truly Evidence-Based?

Karrie Chou, Zekun Zhao
INF2178: Experimental Design for Data Science
Winter 2023

1 Introduction

1.1 Background and Literature Review

Ongoing discourse about North American policing practices points to the differences in treatment that arrested individuals who deviate from the appearance of the privileged North American: a white (Lund & Carr, 2010), male-presenting (Sang & Calvard, 2019), and heterosexual (Murray, 2014) individual receive from the police they interact with. We have previously reviewed cases where racial disparities have impacted policing practices and their execution (Hetey & Eberhardt, 2018) and gender differences have resulted in dissimilar criminal court decisions (Starr, 2012). The findings from these studies are all still part of today's reality when it comes to policing practices. Despite the shift towards police forces emphasizing the strength of their evidence-based decision making capabilities following increased investment into technological developments and integrations, these changes have been limited in magnitude (Koper et al., 2014).

The Toronto Police Service (TPS) is no stranger to such a rationale. Most recently, in February 2020, then-Chief of Police Mark Saunders came under fire when it was revealed that TPS members were using a "controversial facial recognition tool" to supplement their database of individuals (CBC News, 2020). While Saunders did not outwardly support the use of such a tool, and it is unknown who initially approved it, it is clear that there is a portion of TPS staff who will continue to explore technological advancements even at the risk of crossing ethical boundaries. When it comes to cases of arrests and strip searches, TPS does have a publicly available protocol for how they make the strip search decision, and how they collect, store, use, and disseminate data related to these events (Phan et al., 2022), but how policing technology, logical quantitative analysis, and human judgment each factor into that decision remains unclear.

TPS' strip search policy dictates that before a strip search can be conducted, a protective search and a frisk search have to also have been conducted, and following these "lower-level" searches of an arrested individual, "reasonable justification" has to be communicated in order to perform it (Phan et al., 2022). In October 2020, this policy was amended to include clauses that address "office accountability, training, and data management" (Phan et al., 2022). These changes

introduce additional administrative work to the strip search booking process, which ideally should reduce the chances of strip searches being performed liberally, as they have been previously criticized as a way for police to reinforce discriminatory social norms and hierarchies (Jones & Sheehy, 2021).

In our previous study, we looked at whether the decision to conduct a strip search on an arrested individual was based on evidence of their involvement in a crime where they could have incriminating items on their person. In other words, we wanted to see if we could determine that TPS administers strip searches using evidence-based decision making, or if there was an element of profiling involved. Ultimately, we concluded that there was not enough statistical evidence to say that the practice is definitely evidence-based, and that some element of profiling of the arrested individual factors into the strip searching decision.

1.2 Research Objective and Questions

For this study, we want to further dissect the motivations behind the strip search decision. Since we concluded that some element of profiling was involved in TPS making that decision, our research is now focused on understanding how arrest event-related information affects an arrested individual's subjection to strip searching and the nature of their interaction with police. This study will directly address TPS' assertion that strip searching was found to be an evidence-based practice, free of systemic bias in terms of the arrested individuals who were chosen to be strip searched despite the over-representation of arrests involving Black and White individuals in the dataset they collected and used for analysis (Phan et al., 2022).

Our guiding research questions are as follows:

> **RQ1:** Do any information sources that are valid for use in evidence-based policing practices factor into the strip search decision at all?
> **RQ2:** Since we established in our previous study that SearchReason did not have a significant effect on whether or not an individual was strip searched, and that to some extent, profiling of the arrested individual is involved in the decision, what arrest event-based information is actually taken into account in the strip search decision, if any?

2 Exploratory Data Analysis (EDA)

2.1 About the Dataset

Like our previous study, this study uses data from the "Arrests and Strip Searches (RBDC-ARR-TBL-001)" dataset that can be found on TPS's Public Safety Data Portal (Toronto Police Service, 2022). The observations in the dataset are all arrest events between January 2020

and December 2021 that TPS actioned. Before any exclusions are applied, there are n = 65,276 observations in the dataset.

Table 2.1.1. All variables in the raw dataset RBDC-ARR-TBL-001.

Variable Name	Variable Description
Arrest_Year	int – The year the arrest took place in (either 2020 or 2021).
Arrest_Month	str – The quarter the arrest took place in (either Jan-Mar, Apr-June, July-Sept, or Oct-Dec).
EventID	int – An identifier to specify details of the arrest event.
ArrestID	int – An identifier to specify details of the arrest event.
PersonID	int – An identifier to specify details of the arrest event.
Perceived_Race	str – The profiled race of the arrested individual.
Sex	str – The profiled sex of the arrested individual.
Age_group__at_arrest_	str – The age of the arrested individual, listed as a category (either under 17, 18-24, 25-34, 35-44, 45-54, 55-64, over 65)
Youth_at_arrest__under_18_years_	str – A dummy which indicates whether the arrested individual is classified as a youth (under 18 years).
ArrestLocDiv	str – An identifier to specify the location of the arrest event.
StripSearch	int – A dummy which indicates whether the arrested individual was strip searched.
Booked	int – A dummy which indicates whether the arrested individual was booked at a police facility within 24 hours of their arrest.
Occurrence_Category	str – The reason for arresting the individual.
Actions_at_arrest_Concealed_items	int – A dummy which indicates whether the arrested individual was uncooperative with the arresting officers by performing a certain action.
Actions_at_arrest_Combative	int – A dummy which indicates whether the arrested

_violent_or_spitter /biter	individual was uncooperative with the arresting officers by performing a certain action.
Actions_at_arrest_Resisted_defensive_or_escape_risk	int – A dummy which indicates whether the arrested individual was uncooperative with the arresting officers by performing a certain action.
Actions_at_arrest_Mental_instability_or_possibly_suicidal	int – A dummy which indicates whether the arrested individual was uncooperative with the arresting officers by performing a certain action.
Actions_at_arrest_Assaulted_officer	int – A dummy which indicates whether the arrested individual was uncooperative with the arresting officers by performing a certain action.
Actions_at_arrest_Cooperative	int – A dummy which indicates whether the arrested individual was completely cooperative with the arresting officers.
SearchReason_CauseInjury	int – A dummy which indicates the reason an arrested individual was booked for a strip search.
SearchReason_ArrestEscape	int – A dummy which indicates the reason an arrested individual was booked for a strip search.
SearchReason_PossessWeapons	int – A dummy which indicates the reason an arrested individual was booked for a strip search.
SearchReason_PossessEvidence	int – A dummy which indicates the reason an arrested individual was booked for a strip search.
ItemsFound	int – A dummy which indicates whether the strip search of an arrested individual resulted in items related to the crime event being found on their person.

We applied exclusion criteria by dropping observations which had missing values in ArrestID, Age_group__at_arrest_, Perceived_Race, and Occurrence_Category, as these are the only variables that had missing values in the raw dataset. After these exclusion criteria were applied, our dataset had n = 64,615 observations.

2.1.1 Feature Engineering

We used all of the same variables from our previous study, and created the following new variables in preparation for our subsequent EDA and analysis that is relevant to this study:

- **Age**, where each value was randomized to be a whole number within its corresponding age range in order to convert this predictor into a continuous one to support subsequent analysis.
- **SearchReasonCount**, which takes each of the four variables representing the reason why an arrested individual was booked for a strip search, and counts the number of reasons that were documented by a police officer.
- **UncooperativeActionsCount**, which takes each of the five variables representing uncooperative actions which affected the interaction between the arresting officer and arrested individual, and counts the number of uncooperative actions that the arrested individual took during the arrest.
- **StripSearch_count**, which counts how many times individuals within given demographic groups was strip searched during the data collection timeframe.

We also one-hot encoded all multilevel categorical predictors and created dummy predictors from them in order to improve their future interpretability.

Going forward in this report, we also categorize certain predictors into buckets based on which aspects of the arrest event they give us relevant data on. This is done in order to emphasize that this study is focused on understanding how the arrest event factors into the strip search decision, as a contrast to our previous study which focused on demographic predictors and their impacts. See the following table for more information:

Table 2.1.1.1. Defining terms for buckets of predictors that are used in later sections of the report.

Demographic predictors	Event-based predictors
Predictors that are generated based on characteristics of the arrested individual.	Predictors that are generated based on factors related to the arrest event, and not the arrested individual.

2.2 Visualizations

First, we assessed the distribution of categorical predictor values, including Perceived_Race, Age_group_clean, and Sex within our dataset.

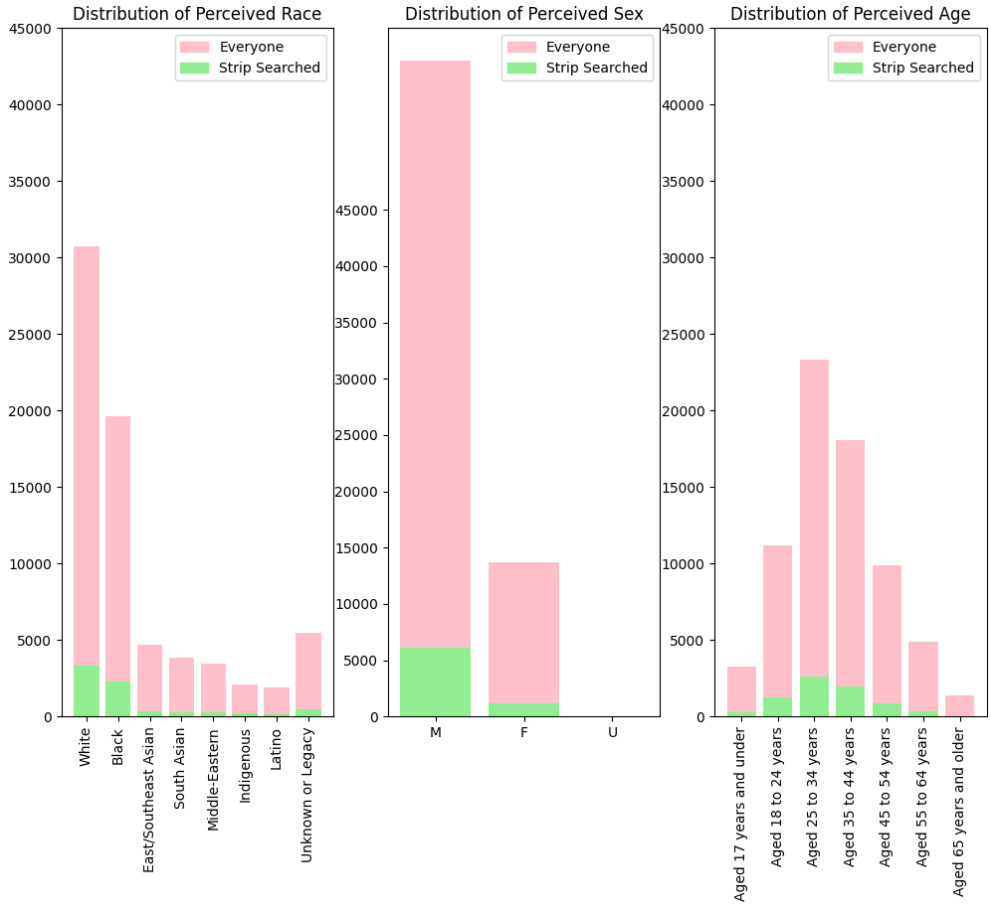


Figure 2.2.1. Distributions of perceived race, sex, and age within our dataset, sorted by whether or not an arrested individual was strip searched.

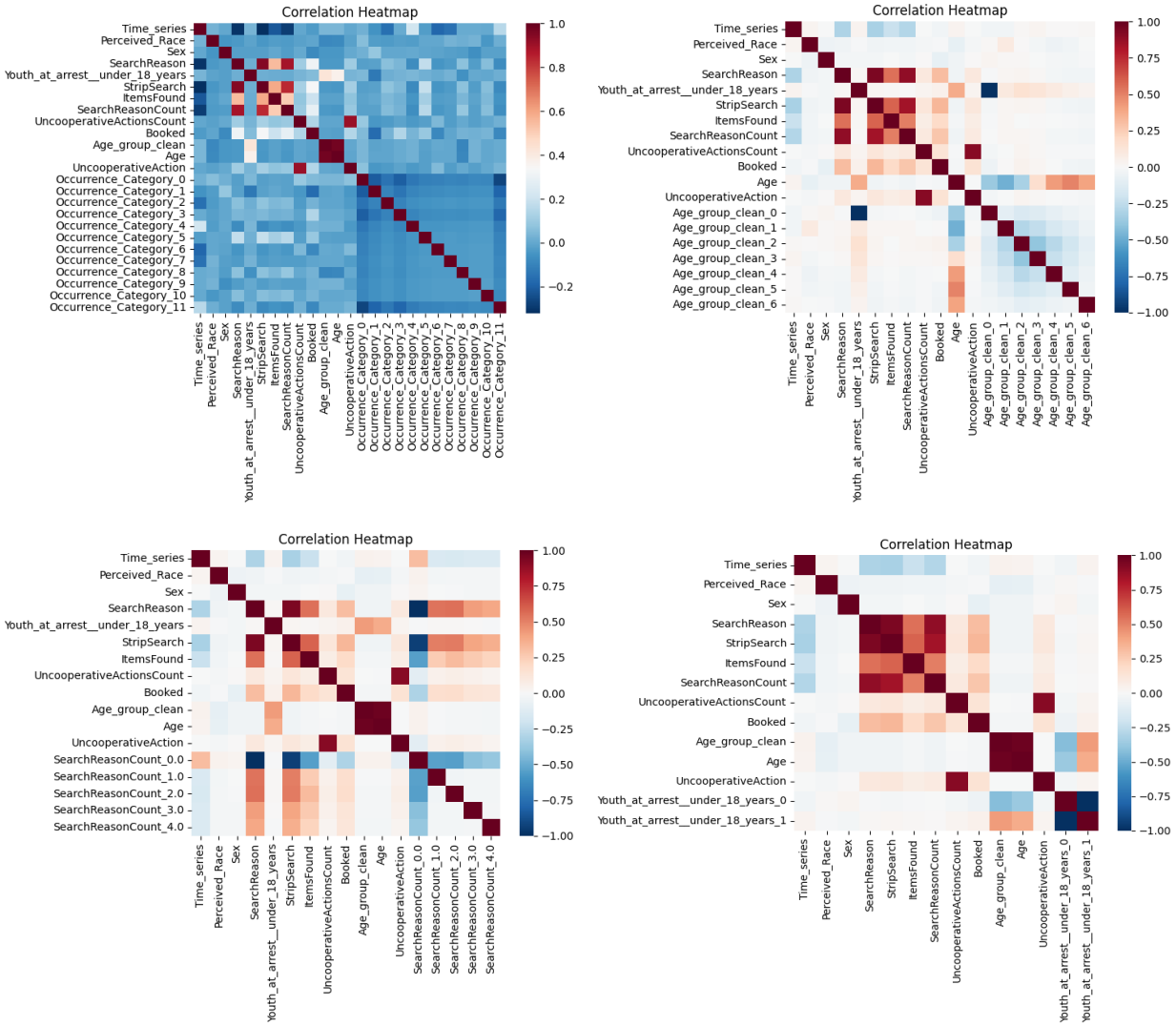
From these visualizations, we see that at a high level, the distributions for the entire dataset and strip searched individuals had similar shapes for each demographic characteristic. However, we can go deeper into this analysis, and look at the proportions of individuals who were strip searched in each of these demographic categories.

Table 2.2.1. Proportions of strip searched individuals in each demographic category.

Perceived Race		Perceived Sex		Perceived Age	
Category	**Proportion**	**Category**	**Proportion**	**Category**	**Proportion**
Black	0.132	Male	0.118	Age <17 years	0.125
White	0.122	Female	0.097	Aged 18 to 24 years	0.123
Indigenous	0.112	Unknown	0.000	Aged 25 to 34 years	0.128
South Asian	0.089			Aged 35 to 44 years	0.096
Middle Eastern	0.0077			Aged 45 to 54 years	0.074
East or Southeast Asian	0.074			Aged 55 to 64 years	0.087
Latino	0.071			Aged >65 years	0.027
Unknown or Legacy	0.099				

Based on the proportions above, we see that (1) Black, White, and Indigenous individuals, (2) individuals between the ages of 25 to 34, and (3) male individuals held the highest strip search proportions. These findings can help us build a case asserting that certain populations are more likely to be exposed to strip search decisions, even before factoring in their potentially criminal activity. Consequently, this assertion undermines the TPS' insistence that strip search administration is conducted "with little evidence of systemic biases" (Phan et al., 2022); while Phan et al.'s report highlights that "there was a positive relationship between rates of strip searches and items found in those strip searches by type of primary offence" (2022), it does not say anything about how police officers decide to book an individual for a strip search, and whether perceived demographic characteristics *do not play any role* in that decision being made. Our study will aim to clarify this missing piece of information.

During the EDA process, we wanted to also see if our one-hot encoded predictors had any collinearities with other predictors that we were considering for use in our final research design. The following correlation heat maps were generated, one for each multilevel categorical predictor that we split into one-hot encoded dummy predictors.



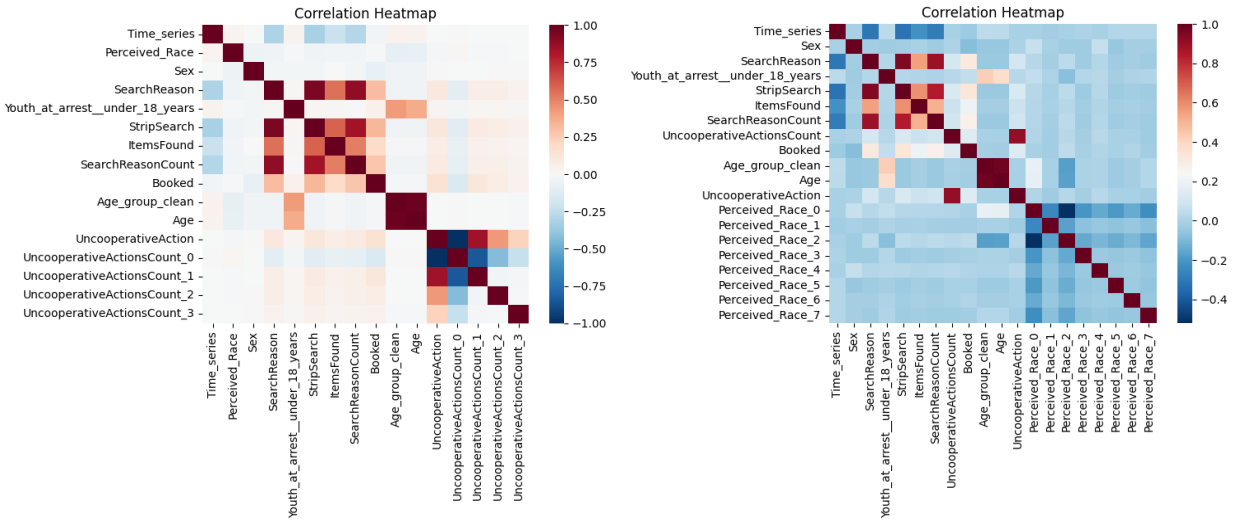


Figure 2.2.2. Correlation heat maps of one-hot encoded multilevel categorical predictors.

Overall, there is little correlation in either direction between all of our categorical predictors. These heat maps serve to assure us that we can assume that many of these predictors are independent of each other, which is beneficial for our subsequent analysis. In terms of relationships that we can see, the fact that predictors whose data are generated strictly from the arrest event (e.g. StripSearch, ItemsFound, and SearchReason) all have high positive correlations with each other suggests that to a certain extent, the individuals who TPS are choosing to strip search did have concealed items on their person, justifying the decision.

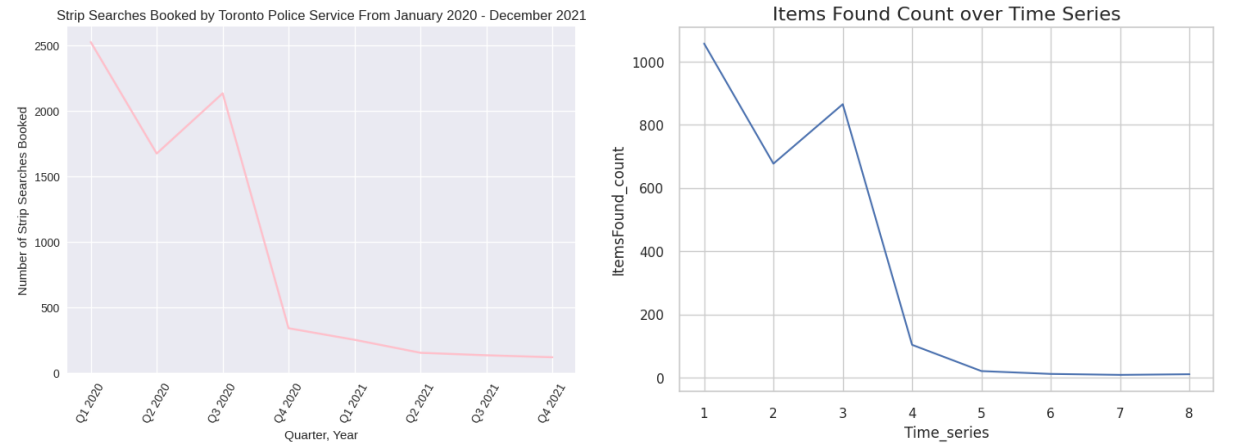


Figure 2.2.3. Left: a plot of strip searches conducted over time. Right: a plot of items found during strip searches over time. Note the similarity in the shapes of the two plots.

We also created the above two plots to understand how both strip search administration and strip search outcomes changed over time. The number of "successful" strip searches, that is, strip searches that uncovered items, decreased over time. In both visualizations, the plots follow similar trends in 2021 (periods 5-8 on the Time_series variable), which shows us that following

TPS' strip search policy amendments, not only are strip searches being performed less often, less items are also being uncovered during them. This may mean that lower level searches such as protective or frisk searches are becoming more successful at uncovering items, meaning that strip searches need to be conducted less often and that they don't reveal any additional evidence in the event that a case is being built against the arrested individual. However, this doesn't mean that strip searches are only being administered in cases where they are necessary; to determine whether this is true, we have to conduct more research into the matter.

2.3 Hypothesis Development and Welch's T-Tests

We use null hypothesis significance testing, backed by Welch's two-tailed t-tests, to assess whether certain characteristics of an individual who was strip searched had an impact on items being found, indicating that the police were justified in administering the strip search in a particular instance. The following sets of hypothesis tests were conducted, each with Items_found being the descriptive statistic.

Table 2.3.1. Summary of hypothesis test results.

	Test 1	**Test 2**	**Test 3**
Hypothesis	Individuals who were cooperative at their arrest and uncooperative individuals are equally likely to have items found during their strip search.	Strip searches that were conducted with a documented reason and strip searches without a documented reason are equally likely to result in items being found during a strip search.	Individuals who are cooperative during their arrest are more likely to be strip searched than individuals who are uncooperative.
H_0 and H_a	H_0: $\mu_{\text{Items found, Cooperative}} = \mu_{\text{Items found, Uncooperative}}$ H_a: $\mu_{\text{Items found, Cooperative}} \neq \mu_{\text{Items found, Uncooperative}}$	H_0: $\mu_{\text{Items found, Reason}} = \mu_{\text{Items found, No reason}}$ H_a: $\mu_{\text{Items found, Reason}} \neq \mu_{\text{Items found, No reason}}$	H_0: $\mu_{\text{Strip search, Cooperative}} = \mu_{\text{Strip search, Uncooperative}}$ H_a: $\mu_{\text{Strip search, Cooperative}} > \mu_{\text{Strip search, Uncooperative}}$
Test statistic	-1.426	1.916	-8.677
***p*-value**	0.154	0.056	0.000

Neither Test 1 nor Test 2 yield statistically significant results. Intuitively, this makes sense, as we have not established that an arrested individual of any given population would be more likely to have items on their person that have to be found through the means of a strip search. This further raises the question of how truthful TPS is in asserting that their strip search decisions are made

using evidence-based practices and rationales, especially because we also know from our correlation heat maps that these predictors, generated from events that take place during the arrest, don't have a strong relation with the success of a strip search.

Test 3 yields a statistically significant result. This means that individuals who were uncooperative during their arrest had a lower chance of being strip searched at all. This finding is interesting, as it indicates that TPS police are considering factors outside of valid strip search justifications to make their final decision. In this case, where uncooperativeness is a statistically significant factor in the difference between number of strip searches conducted for different subgroups in the dataset, we can explore TPS' consideration of its own staff's safety in arrest events and how that factors into a decision on the administrative process that an arrested individual will go through.

3 Research Design and Methodology

Our methodology for addressing our research questions (RQs) involves the use of power analyses, ANCOVA models, and a logistic regression classifier.

For both RQs, 3 different power analyses are conducted before any subsequent analysis took place in order to determine the sample size we need in order to establish that a statistically significant difference of either a small (Cohen's d = 0.2), medium (d = 0.5), or large (d = 0.8) effect size can be observed.

For RQ1, we constructed an ANCOVA model with StripSearch_count as the outcome. The model is as follows:

- StripSearch_count ~ Time_series + UncooperativeActionsCount + SearchReasonCount

SearchReasonCount and Time_series were chosen as predictors based on their high correlations with StripSearch_count as seen in Figure 2.2.2. Further, from our previous study and midterm report, we found that UncooperativeAction and, by extension, UncooperativeActionsCount, had a statistically significant impact on the strip search decision, meaning that TPS police conducted more strip searches on individuals who were labeled as uncooperative during their arrest. The sample that was used to create the ANCOVA model was constructed using grouped data that accounted for the total number of strip searches in different subsets of our original dataset. Data was grouped using Time_series, UncooperativeActionsCount, SearchReasonCount, and Age_group_clean.

For RQ2, we trained a logistic regression model whose construction is based on our second ANCOVA model to output whether or not an individual would have items on them that can be uncovered through a strip search. The logistic regression model is as follows:

- ItemsFound_Or_Not ~ Time_series + UncooperativeActionsCount + SearchReasonCount

The outcome variable ItemsFound_Or_Not takes on the same values for each observation as the ItemsFound variable from the original dataset. This model was trained on a training dataset which was constructed using 60% of the observations in the dataset with n = 64,615 that we used for analysis. The training was done with n = 1,000 bootstrapped samples. The remaining 40% of observations were set aside as a testing dataset so we could test the performance of our model and make a conclusion on how well it predicts strip search decisions.

3.1 Power Analysis Results

Our power analysis is conducted with the goal of achieving a power of 0.8 on any of our subsequent data analysis with significance level $\alpha = 0.05$. First, we calculated the sample sizes given different effect sizes, measured by Cohen's d, that we would need to meet the above conditions.

Table 3.1.1. Sample sizes needed, given power = 0.8, $\alpha = 0.05$, and Cohen's d = {0.2, 0.5, 0.8}. Sample sizes are rounded up to the nearest whole number.

For d = 0.2 (small effect size)	For d = 0.5 (medium effect size)	For d = 0.8 (large effect size)
n = 394	n = 64	n = 26

The number of samples needed to detect a larger effect size while power is consistently held at 0.8 decreases. This is intuitive; if the effect size is small (such as d = 0.2), but we expect the power of our statistical test to be high at 0.8, we would have to use more data to determine if a statistically significant relationship exists between our chosen predictors and outcomes.

Knowing the sample sizes needed to establish a statistical test with power = 0.8, we can construct power curves that show us the effects of different sample sizes on power level if we want to test for small, medium, and large effect sizes.

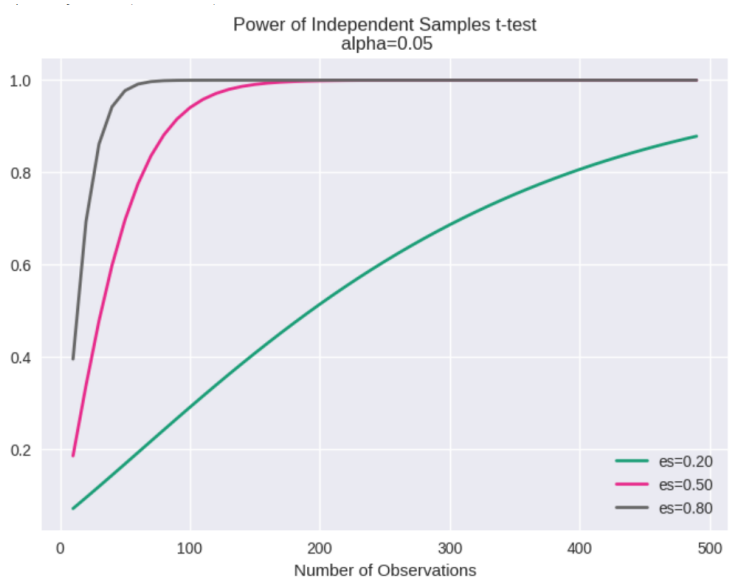


Figure 3.1.1. Plot of power curves based on different measures of Cohen's d.

We use the insights from our power curve to construct an ANCOVA model which satisfies the sample size requirements given by Table 3.1.1., thereby restricting the power of our statistical test to 0.8 or greater and increasing our confidence in its internal validity.

4 Results

4.1 ANCOVA Results

As mentioned in Section 3, we constructed the following ANCOVA model:

- StripSearch_count ~ Time_series + UncooperativeActionsCount + SearchReasonCount

Our final sample size used in the ANCOVA model is n = 639, which satisfies our power analysis results for all considered measures of Cohen's d and power = 0.8 or higher. The summary of the ANCOVA model is shown below in Table 4.2.1.

Table 4.1.1. Summary of ANCOVA model results.

	Coefficient	SE	95% CI	p
Intercept	36.391	2.744	31.002, 41.780	0.000
Time_series	-3.675	0.430	-4.519,	0.000

			-2.831	
UncooperativeActionsCount	-9.530	0.867	-11.232, -7.828	0.000
SearchReasonCount	0.300	0.702	-1.078, 1.679	0.669

At a significance level of $\alpha = 0.05$, Time_series and UncooperativeActionsCount are statistically significant predictors, as their p-values are below our established α. For Time_series, this means that on average, as we move further beyond Q1 of 2020, the number of strip searches made decreases across all observation unit subgroups of age, uncooperativeness, and strip search eligibility; this is an indication that TPS is following its redefined strip search policy and reducing the number of strip searches being conducted because the administrative process of booking one now requires more time and documentation of justifications. For UncooperativeActionsCount, this means that on average, as an individual of a given age group who is arrested during a given time period and who has a certain number of reasons documented justifying their booking for a strip search becomes more uncooperative during their arrest, they are less likely to even have a strip search conducted. This finding shows us that TPS police may account for their need to preserve their own safety and factor it into the strip search decision. Tolerance for uncooperativeness is not an arrest event-based predictor, and therefore we cannot conclude that TPS' strip search decision making is purely evidence-based. This is further corroborated by SearchReasonCount not being a statistically significant predictor. Knowing this fact, we continue supporting our assertion that documented reasons in favour of conducting a strip search don't have a significant impact on the final decision to strip search an arrested individual, which was formed during the hypothesis testing phase of our current study and supported by Test 3.

It's important to note that the adjusted R^2 value of our ANCOVA is 0.228, meaning that even though we identified statistically significant predictors, when they are combined, only 22.8% of variance in StripSearch_count was explained by the model. This low adjusted R^2 may stem from various factors, including an over-representation of individuals who were not strip searched, or that event-based predictors simply cannot explain the entirety of the strip-search decision making process because they don't represent information outside of the arrest taking place and the circumstances surrounding it. More research will be needed to determine how to create a model which accurately captures all factors that influence the strip search decision.

4.2 Logistic Regression Results

As mentioned in Section 3, our logistic regression model is as follows:

- ItemsFound_Or_Not ~ Time_series + UncooperativeActionsCount + SearchReasonCount

Below are the quantitative results of the model's performance on the testing dataset.

Table 4.2.1. Logistic Regression Results Table

Predictor Variable	Coefficient	Std. Error	z-value	P > |z|	95% Confidence Interval
Intercept	0.0546	0.067	0.810	0.418	-0.078, 0.187
Time_series	-0.2564	0.019	-13.853	0.000	-0.293, -0.220
Uncooperative ActionsCount	0.0634	0.039	1.624	0.104	-0.013, 0.140
SearchReason Count	0.0089	0.020	0.439	0.661	-0.031, 0.049

Table 4.2.2. Confusion matrix for the logistic regression model's performance on the testing dataset.

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive) 2	FN (False Negative) 1088
Actual Negative	FP (False Positive) 10	TN (True Negative) 1833

Table 4.2.3. Performance statistics of the logistic regression model with the testing data set. Items to pay attention to include precision, recall, and accuracy.

	Precision	Recall	F1-Score	Support
0	0.63	0.99	0.77	1843
1	0.17	0.00	0.00	1090
Accuracy			0.63	2933
Marco Avg	0.40	0.50	0.39	2933
Weighted Avg	0.46	0.63	0.48	2933

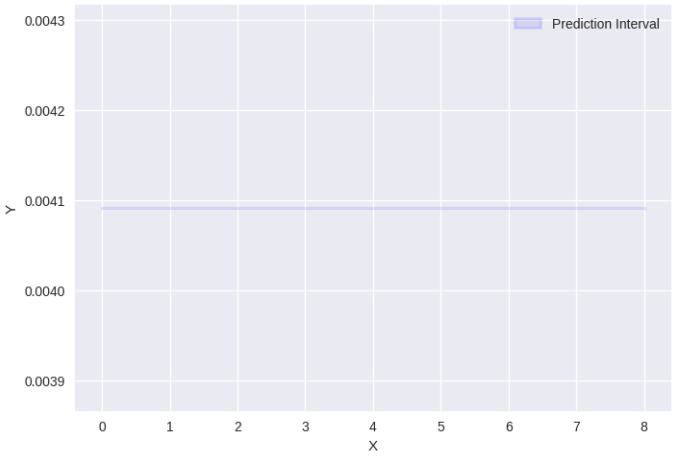


Figure 4.2.1. Interval plot of the model's predictions on the testing dataset. The distribution of the model's predictions can be summarized with $\mu = 0.004$ and σ close to 0.

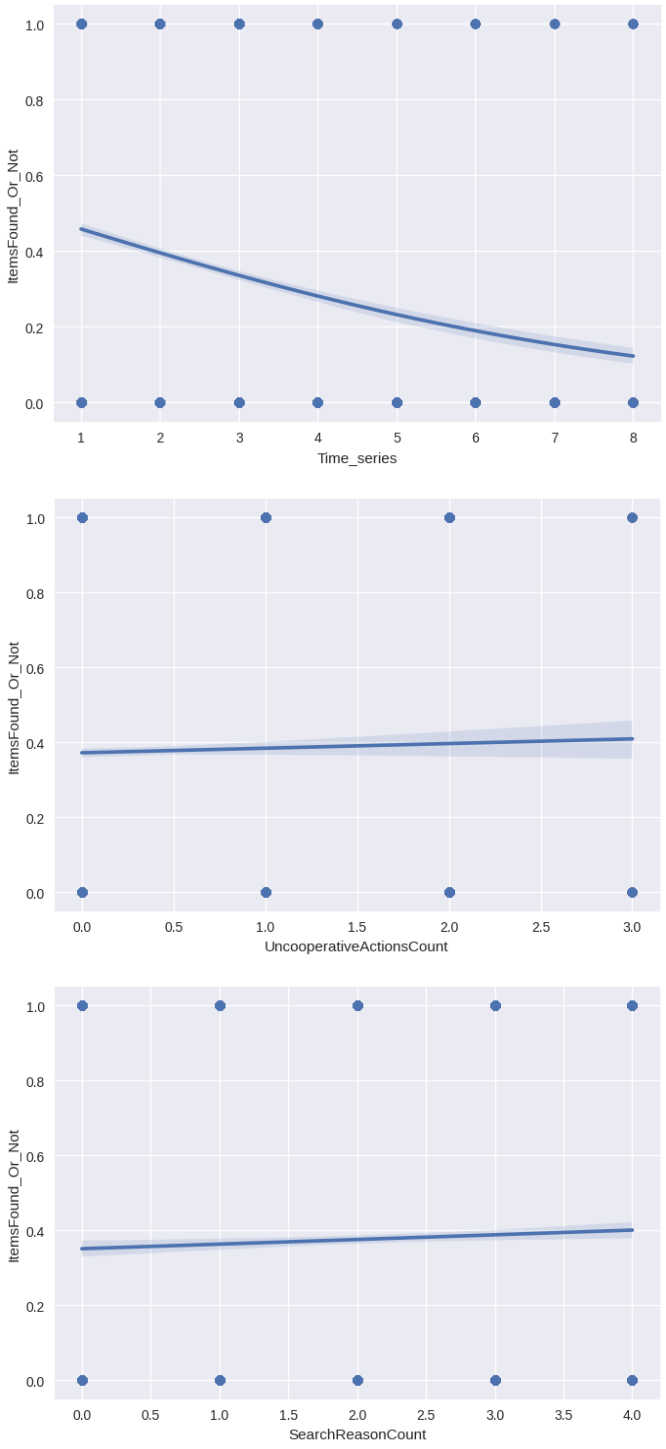


Figure 4.2.2. Plots of the logistic regression model's prediction holding all predictors constant and changing one (solid blue), the 95% confidence interval (shaded blue), and predictions from running the logistic regression model on the testing dataset (blue points). From top to bottom, the predictor whose value is being changed is Time_series, UncooperativeActionsCount, and SearchReasonCount.

Table 4.2.1. summarizes the results of our logistic regression model. The table shows the coefficients for each predictor variable, along with their standard errors, z-values, p-values, and 95% confidence intervals. The coefficient for Time_series is -0.2564, indicating that the odds of finding items during a strip search decreased from Q1 2020 to Q4 2021. The coefficient for UncooperativeActionsCount is 0.0634, which suggests that as the level of uncooperativeness of the individual being searched increased, the odds of finding items also increased, although this effect was not statistically significant ($p > 0.05$). Finally, the coefficient for SearchReasonCount is 0.0089, which indicates that the number of documented reasons for conducting the strip search did not have a significant effect on the odds of finding items ($p > 0.05$).

Based on these findings, we can conclude that the Time_series and the level of uncooperativeness of the individual being searched are important predictors of whether or not items are found during a strip search. The finding that the number of documented search reasons for conducting the strip search does not have a significant effect on the odds of finding items suggests that the decision to conduct a strip search is not purely evidence-based. However, it should be noted that the effect of uncooperativeness on the odds of finding items was not statistically significant, which may be due to the limited sample size or other confounding factors. Further research is needed to confirm and extend these findings.

Table 4.2.2. shows the confusion matrix from the model's predictions on the testing dataset. The model predicted 2 true positives (TP) and 10 false positives (FP), so it correctly predicted the discovery of the item in 2 cases, but incorrectly predicted the discovery of the item in 10 cases where it was actually not present. The model also predicted 1088 false negatives (FN) and 1833 true negatives (TN), indicating that there were 1088 cases where the item was discovered but the model failed to predict its presence, and 1833 cases where the item was not discovered and the model correctly predicted its absence. The precision of the model for predicting the presence of the item was 0.17, indicating that when the model predicted the discovery of the item, only 17% of the predictions were actually true. The recall of the model for predicting the presence of the item was 0.00, indicating that the model failed to correctly identify the presence of the item in any of the cases where it was actually discovered. The F1-score of the model was 0.00, indicating that the model performed poorly in predicting the discovery of the item.

In our model, if the value of ItemsFound_Or_Not is 1, items were found during the strip search, while a value of 0 indicates that items were not found. The precision and recall of the model can be used to evaluate its performance in predicting the absence or presence of the item.

Looking at Table 4.2.3., the low recall score of 0.00 indicates that the model failed to correctly identify the presence of the item in any of the cases where it was actually present, which is a serious issue in the context of the strip search process. The low precision score of 0.17 suggests that the model is not very accurate in predicting the discovery of the item, as it incorrectly

predicted its absence in a relatively large number of cases where it was actually discovered. Therefore, further investigation of optimal model choice may be necessary to improve its accuracy and reduce the number of false negatives and false positives.

5 Discussion

Based on our EDA and ANCOVA analyses, we attempted to uncover the relationship between the number of strip searches and three predictor variables: Time_series, UncooperativeActionsCount, and SearchReasonCount. Our previous research indicated that TPS was more likely to conduct strip searches on individuals who were less cooperative. However, as the level of uncooperativeness increased, TPS conducted fewer strip searches on them. One possible explanation is that when arrested individuals display more uncooperative actions, TPS officers may be hesitant to provoke them and make the situation worse.

To further investigate the decision-making process behind TPS's use of strip searches, we utilized the ItemsFound variable to examine the outcomes of these searches. Based on t-tests, we found that both cooperative and uncooperative arrested individuals had a similar likelihood of having items found during a strip search. This suggests that the decision to conduct a strip search is not purely based on factual evidence.

To explore the relationship between the search reason and the finding of items, we conducted t-tests on the ItemsFound variable for groups with and without a search reason documented. Results showed no significant difference between the two groups in terms of item findings at an alpha level of 0.05. However, when the alpha level was set to 0.1, we rejected the null hypothesis and found that search reasons do have a significant influence on the likelihood of finding items. This finding suggests that TPS should base their decisions to conduct strip searches on factual evidence rather than subjective impressions or feelings.

In our previous research, we found that police officers hold different attitudes towards conducting strip searches on youths and adults. To investigate this relationship further, we conducted two t-tests. The t-tests showed that police officers did conduct more strip searches on youths. However, when we compared the likelihood of finding items on adults and youths who underwent strip searches, we found that the probability of finding items was the same in both groups. This suggests that the police officers' differential attitudes towards conducting strip searches on youths and adults are questionable and may exacerbate negative impressions of the police among youths.

Furthermore, despite the observed correlations between strip search and Time_series, UncooperativeActionsCount, and SearchReasonCount in the heatmap, our ANCOVA analysis revealed that SearchReasonCount did not have a significant impact on strip search decisions.

This fact also resulted in a relatively low adjusted R^2 value of 0.228 for our model, even as the sample size increased.

As the outcome of strip search, we decided to use ItemsFound_Or_Not as the dependent variable to construct our logistic regression model. Based on the observation of the heatmap and previous studies, we selected Time_series, UncooperativeActionsCount, and SearchReasonCount as the independent variables.We also realized that our dataset was imbalanced, and therefore we attempted to shuffle and stratify our dataset. However, this did not improve our prediction accuracy.Therefore, we need to consider whether logistic regression is suitable for this dataset and explore the possibility of enriching the dataset by adding more relevant variables.

5.1 Limitations

For this study, we considered ways to mitigate the limitations that we discussed in our midterm report. From our midterm report, we identified that the principal limitations of our analysis came from (1) the dataset being entirely categorical and (2) the conflicting information on each person represented in the dataset due to multiple entries being associated with unique PersonID values. The continuous variables we created and used in this study, such as Age and StripSearchCount, were designed to address the challenges that these limitations caused in our previous study.

However, these approaches don't completely negate the difficulties caused by how the dataset was created and each predictor value labeled. Since each observation relies on TPS staff's perceptions of an arrested individual, there is a degree of reporting bias that cannot be completely controlled for in our analysis. Other issues, such as severe class imbalance in that there was a much lower number of strip searched individuals in the overall dataset, and an over-representation of certain demographics, mean that the external validity of our findings may only extend to TPS' policing practices. Class imbalance most notably affected the performance of our logistic regression model, which had an f1 score of 0.00 in predicting whether items would be found on an arrested individual, stemming from extremely low precision and recall. A dataset that is compiled using a mixture of information that is self-reported by arrested individuals and information that is supplied by police institutions, and that represents a larger geographic region, may help to reduce class imbalances and help us deliver findings that have stronger external validity.

Limitations in this study also stem from the applicability of our research methodology to the research questions we are studying. For instance, our logistic regression model yielded a relatively low accuracy, at 0.63. This low accuracy stems from the fact that the logistic model is an extension of the general linear model (GLM), which works best at predicting outcomes on linearly separable data, i.e. data that can easily be separated by class with a linear multi-dimensional hyperplane (M. Ataei, personal communication, January 30, 2023). With the grouped data that we used for our ANCOVA and our logistic regression containing 3 dimensions,

and our ungrouped clean data containing 16 dimensions excluding our one-hot encoded predictors, it's difficult to visualize a linear hyperplane that could separate observations based on the ItemsFound binary outcome variable. Future studies that build on our two research questions and the work we have done to understand the nuances of TPS' strip search decision making policy could incorporate non-GLM-derived classifiers to better understand the influencing factors for a strip search decision.

6 Conclusion

6.1 RQ1: Do any information sources that are valid for use in evidence-based policing practices factor into the strip search decision at all?

From our ANCOVA, we see that in TPS' case, the timing of an individual's arrest and the individual's uncooperativeness during their arrest did have a statistically significant impact on strip searches they were booked for. Taking the information we have about the changes TPS made to its strip search policy and how they are designed to impact the number of strip searches administered by the department into account, the trend of decreasing numbers of strip searches being conducted over the period that this dataset was collected makes sense. However, the impact of the arrested individual's uncooperativeness is a surprising finding, and highlights that there may be factors that are not strictly related to arrest event-based information that mediate the relationship between the police-arrestee interaction and the strip search decision. Further, documented justifications for a strip search did not have a statistically significant effect on the number of strip searches conducted on individuals within certain population subgroups, which is a potential flag for the inefficacy of TPS' policy amendments.

6.2 RQ2: Since we established in our previous study that SearchReason did not have a significant effect on whether or not an individual was strip searched, and that to some extent, profiling of the arrested individual is involved in the decision, what arrest event-based information is actually taken into account in the strip search decision, if any?

From our logistic regression model, we see that arrest event-based predictors did not do a good job at accurately predicting whether a strip search being conducted would uncover hidden items, justifying the strip search decision being made. Integrating our findings in this study with our conclusions from our previous study, where we concluded that police perceptions of arrested individuals did factor into the strip search decision, we can conclude now that while TPS has

made an effort to change its administration processes to incorporate the use of documented and verifiable evidence in the justification for booking a strip search, the organization has not successfully transitioned to an entirely evidence-based methodology. It is unclear still which aspects of arrest event-based data and information weigh into the strip search decision. Further research on how police's personal attitudes and self-preservation capacities mediate the relationship between event-based information and the strip search decision needs to be conducted to answer this question.

7 References

CBC News. (2020, February 13). *Toronto police admit using secretive facial recognition technology Clearview AI*. CBC. https://www.cbc.ca/news/canada/toronto/toronto-police-clearview-ai-1.5462785

Hetey, R. C., & Eberhardt, J. L. (2018). The Numbers Don't Speak for Themselves: Racial Disparities and the Persistence of Inequality in the Criminal Justice System. *Current Directions in Psychological Science*, *27*(3), 183–187. https://doi.org/10.1177/0963721418763931

Jones, D., & Sheehy, E. (2021). R v Desjourdy: A Narrative Of White Innocence And Racialized Danger. *The Canadian Bar Review*, *99*(3), Article 3. https://cbr.cba.org/index.php/cbr/article/view/4714

Koper, C. S., Lum, C., & Willis, J. J. (2014). Optimizing the Use of Technology in Policing: Results and Implications from a Multi-Site Study of the Social, Organizational, and Behavioural Aspects of Implementing Police Technologies. *Policing: A Journal of Policy and Practice*, *8*(2), 212–221. https://doi.org/10.1093/police/pau015

Lund, D. E., & Carr, P. R. (2010). Exposing Privilege and Racism in The Great White North: Tackling Whiteness and Identity Issues in Canadian Education. *Multicultural Perspectives*, *12*(4), 229–234. https://doi.org/10.1080/15210960.2010.527594

Murray, D. A. B. (2014). Real Queer: "Authentic" LGBT Refugee Claimants and Homonationalism in the Canadian Refugee System. *Anthropologica*, *56*(1), 21–32.

Phan, M. B., Dinca-Panaitescu, M., & Rebelo, N. (2022). *Understanding Strip Searches in 2020 Methodological Report*. Toronto Police Service.

Sang, K. J. C., & Calvard, T. (2019). 'I'm a migrant, but I'm the right sort of migrant': Hegemonic masculinity, whiteness, and intersectional privilege and (dis)advantage in

migratory academic careers. *Gender, Work & Organization*, *26*(10), 1506–1525. https://doi.org/10.1111/gwao.12382

Starr, S. B. (2012). *Estimating Gender Disparities in Federal Criminal Cases* (SSRN Scholarly Paper No. 2144002). https://doi.org/10.2139/ssrn.2144002

Toronto Police Service. (2022). *Arrests and Strip Searches (RBDC-ARR-TBL-001)*. https://data.torontopolice.on.ca/datasets/TorontoPS::arrests-and-strip-searches-rbdc-arr-tbl-001/about