https://colab.research.google.com/drive/1ooM-21u6tJ4jxarVxIkiYnPSnNlfg2j0

UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

INF2178 Final Project

by the group40 of

Hung Wang
Rohit Pandit

Faculty of Information

April 15, 2023

Table of Contents

Introduction

1.1 Introduction

In 2019, a Sunday school teacher from Oregon was falsely accused of being a drug smuggler and got strip-searched at Vancouver International Airport. After a body cavity search and an X-ray, no drugs were found, and the woman stated she was traumatized (Thomas&Gidda, 2023). This controversial incident hit the news headline and caused a heated debate on the legality and legitimacy of strip searches by the Police on suspects, especially on females. Additionally, one study by Liberty Investigates has revealed that black girls are three times more likely to be subjected to invasive strip searches by the Police than white girls in the UK (Johnson, 2019). The report also discovered that the Police use strip searches more frequently on women than on men. Consequently, such searches have traumatizing effects on the victims and violate their human rights. This also implies that sex could impact the individual likelihood of being strip searched.

Besides the topic of strip searches, in Canada, studies have shown that sex can impact an individual's likelihood of being arrested. According to the Government of Canada, one report had shown that females only accounted for 1 in 4 persons accused in police-reported crime incidents in Canada in 2017(Laura, 2019). Chesney-Lind and Pasko's (2013) research has shown that women are typically apprehended for criminal offenses such as theft, drug offenses, fraud, and impaired driving. In cases where women participate in violent behaviour, it is usually of less severe intensity, taking place in private settings and directed towards individuals known to them, as per Schwartz's (2013) findings. Conversely, male aggression tends to result in more harm, happening more frequently in public spaces, and directed toward strangers, as per Schwartz's (2013) observations. As a result, this suggests that the likelihood of being arrested can be influenced by sex.

This study aims to explore how a personal attribute of sex impacts the likelihood of being arrested in Canada. Additionally, the study seeks to examine whether the personal attribute of sex is associated with the individual likelihood of strip search. For this study, we will use a dataset on arrests and strip searches published by Toronto Police Service. By investigating the influence of sex on arrest rates and strip search rates, the study hopes to highlight the potential for discriminatory practices within law enforcement. The dataset analysis and the subsequent evaluation of the evidence aim to provide valuable insights for policymakers, criminal justice professionals, and scholars in the field to better understand and address issues of bias and discrimination within the criminal justice system.

1.2 Literature Review

The study will reference two studies on the criminal justice system and outcomes in the United States and Canada for the comparison in the following sections of this paper. The first study *Sex Differences in the Likelihood of Arrest* by Stolzenberg & D'Alessio, examined data from the National Incident-Based Reporting System (NIBRS) from FBI to determine whether sex differences exist in the likelihood of arrest in nineteen states and the District of Columbia during 2000. Seven offenses were analyzed including kidnapping, forcible rape, forcible fondling, robbery, aggravated assault, simple assault, and intimidation. The results found that males are significantly more likely to be arrested for kidnapping, forcible fondling and intimidation than females. Similarly, tying back to the first study, this study also showed that Black females are more likely to be arrested for aggravated and simple assault compared to White females. However, in some cases, the offender's sex on the probability of arrest was more pronounced when an individual was black. By analyzing the data, the analyses suggest that the sex of a criminal offender has an influence on the police decision-making process. Overall, the results of the study indicate that the reason behind the lower arrest rate for females is partially due to the fact that law enforcement officials show more leniency towards women. The authors suggest that these findings have important implications for law enforcement policies and practices, including the need for more gender-specific approaches to policing.

The second study, *"It's Sexual Assault. It's Barbaric": Strip Searching in Women's Prisons as State-Inflicted Sexual Assault* by Jessica Hutchison explored the female prisoners' experience of being striped search in Canadian prisons and provided evidence that strip-searching women prisoners is a form of state-inflicted sexual assault. In the context of the #MeToo movement, the controversy of sexual assault has gained prominence in public and academic discourse. However, imprisoned women have been largely overlooked in these conversations, particularly concerning their experiences of strip searches. This study aimed to increase such awareness by interviewing five cisgender women who had been strip-searched while being incarcerated in Canada. The study's findings indicate that strip-searching is a form of sexual assault. These incarcerated women could not decline the strip search due to the power imbalances and the fear of severe repercussions they faced. Furthermore, their past experiences of sexual victimization made the strip search especially detrimental. The study concludes that the structural violence within prisons reflects the structural violence in society. The state's development and implementation of strip search policies mean that strip searching constitutes state-inflicted sexual assault. The author theorizes that strip-searching is not recognized as sexual assault because incarcerated women are treated as subhuman and not entitled to humane treatment. The implications of the study suggest the need to reduce the harms of strip-searching, with the ultimate goal of eliminating this practice in women's prisons.

1.3 Research Questions and Objective

In our project, we used a dataset that gives information about arrests and strip searches in Toronto. The dataset comprises information about 65,275 entries for 37347 unique people including 25 different attributes. Each column attribute name represents a variable and each row is an observation. The cell represents the value.

For research question 1 using ANCOVA, we seek to shed light on the potential influence of the personal attribute of sex on the number of arrests. In order to achieve this goal, we aim to closely scrutinize whether a discernible relationship exists between sex and the number of arrests. To minimize potential confounds, we will utilize a collective covariate of negative behaviour at arrest, which combines concealment of items, combative behaviour, violence, spitting or biting, resistance, defensive or escape tendencies, mental instability, suicidal tendencies, and assault of law enforcement personnel. By controlling for these additional variables, we can more accurately isolate the effect of sex on arrest frequency and arrive at a more nuanced understanding of the underlying factors at play.

For research question 2 using logistic regression, we want to investigate the influence an individual's sex and the number of arrests had with the individual being strip searched. We shortlisted features like the arrest year, arrest month, age of the individual, if the individual was a youth, and the race of the individual and grouped them to count the number of arrests. These features draw more background information related to the individual's sex that would play an important role in helping us understand the relationship with the individual being strip searched better.

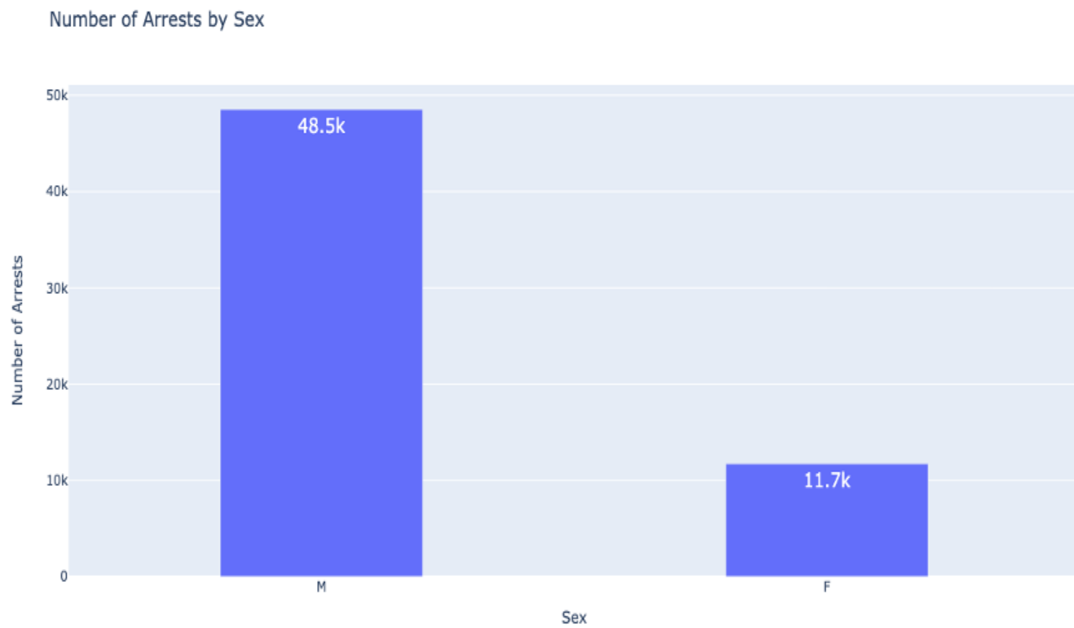
By exploring the two research questions outlined, our primary objective is to furnish an enhanced comprehension of the dataset, predicated on the notion that the personal attribute of sex may exert influence on both the frequency of arrests and the likelihood of being strip searched. The aim of our investigation is to scrutinize the effect of sex in relation to these outcomes, which in turn will allow us to pinpoint potential patterns or biases that may be demonstrated to the criminal justice system. To facilitate our inquiry, we conducted a review of the relevant literature, as well as engaged in a comprehensive analysis of descriptive statistics, t-tests, ANCOVA, and logistic regression techniques with respect to the dataset in question. Drawing on our preliminary findings, we have obtained significant insights into the fundamental characteristics of the data, and will leverage these insights to guide our ongoing exploration.

2 Exploratory Data Analysis

2.1 Descriptive Statistics

As a preliminary step towards obtaining a comprehensive understanding of the dataset related to the number of arrests, we generated a count chart that depicts the number of male and female arrestees (Figure 1). As expected, the result revealed a notable difference in the number of arrests between males and females, with the former having a significantly greater proportion of the total arrests recorded. This finding underscores the importance of examining potential sex-based differences in the context of criminal justice system situations.

Figure 1.

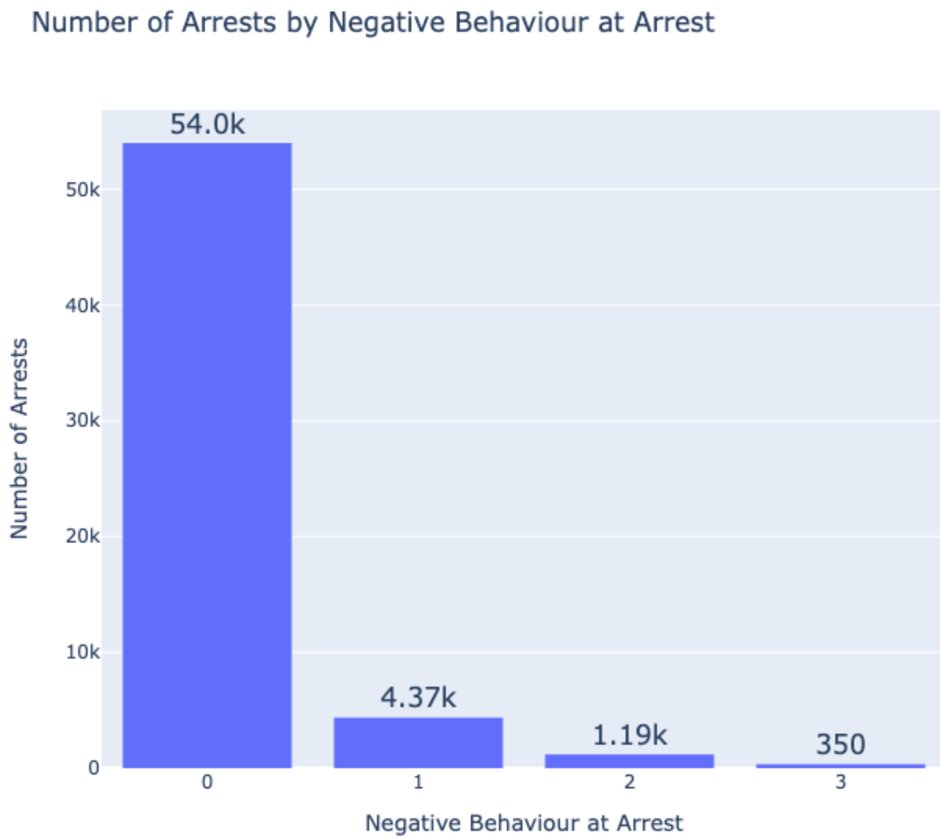


In addition to the count chart analysis, we further explored potential differences in the number of arrests between males and females by generating boxplots (Figure 2). Our examination of the boxplots revealed a higher median value for males, indicating that on average, males have a greater number of arrests than females. Additionally, the range of values for males was wider than that of females, meaning a greater variation in the number of arrests recorded among males. Conversely, the boxplot for females displayed a notably left-skewed distribution, indicating a higher proportion of female arrestees with a lower number of arrests compared to males. Again, these findings provided additional insights into the potential disparities in the number of arrests between males and females and can serve as a valuable reference for further analysis.

Figure 2

Regarding the covariate of negative behaviour at arrest, we produced a count chart to display the frequency of different levels of negative behaviour observed (Figure 3). The bar chart revealed that the majority of arrestees scored 0 on the negative behaviour, indicating that a considerable proportion of arrestees did not exhibit any negative behaviour during their arrest. Furthermore, the second most common score was 1, which was less frequent than 0. This suggests that some arrestees displayed minor negative behaviour during their arrest. In contrast, scores of 2 and 3 had lower frequencies than 0 and 1, indicating that fewer arrestees exhibited more intense negative behaviour. Overall, the distribution of scores for negative behaviour suggests that most arrestees did not display significant negative behaviour during their arrest. However, we still want to include negative behaviour as the covariate for our ANCOVA analysis as we think that the display of negative behaviour may exert an impact on the arrest determinations made by law enforcement. These findings could be valuable for understanding the prevalence and severity of negative behaviour during arrests.

Figure 3



Additionally, we produced a stacked bar chart, a regular bar chart and an interaction plot showing the distribution of negative behaviour at arrest broken down by male and female (Figure 4&5&6). Based on our analysis of the negative behaviour at arrest dataset, it is clear that males exhibit more negative behaviour than females across all scores from 0 to 3. The score of 0 had the highest proportion for both males and females, but the proportion for males was still higher. The data suggests that there is a notable difference in the negative behaviour exhibited by males and females during arrests. However, we are also aware that more male arrestees are recorded in this dataset.

Figure 4

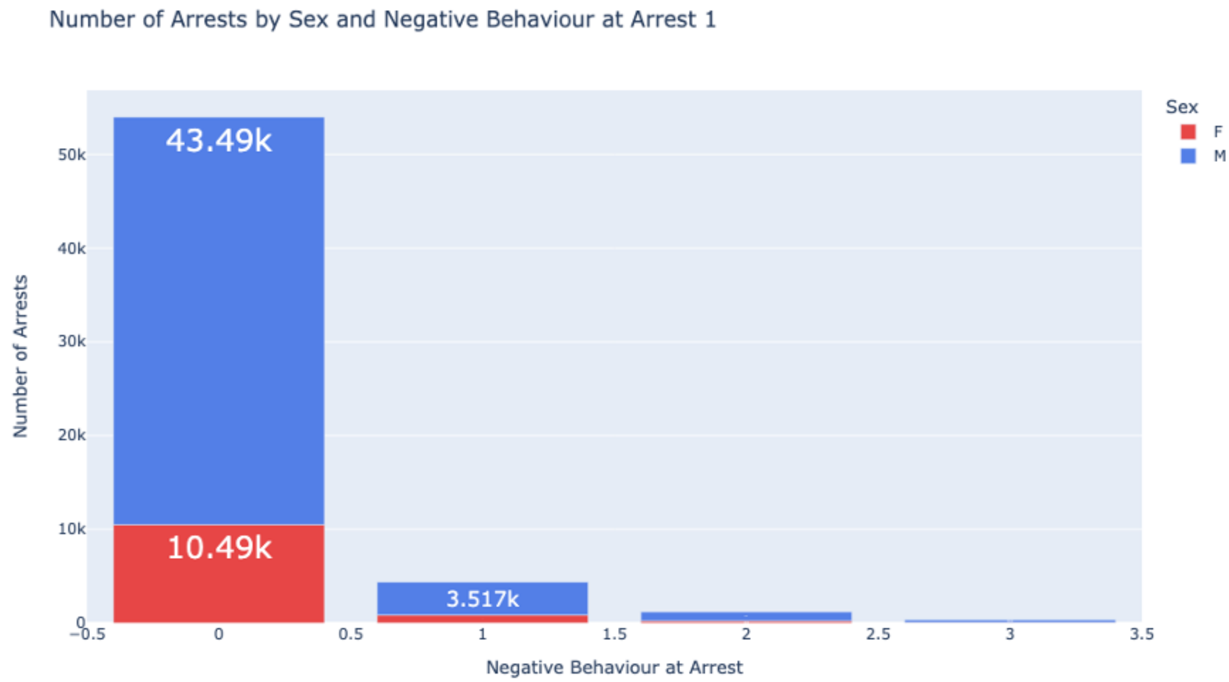


Figure 5

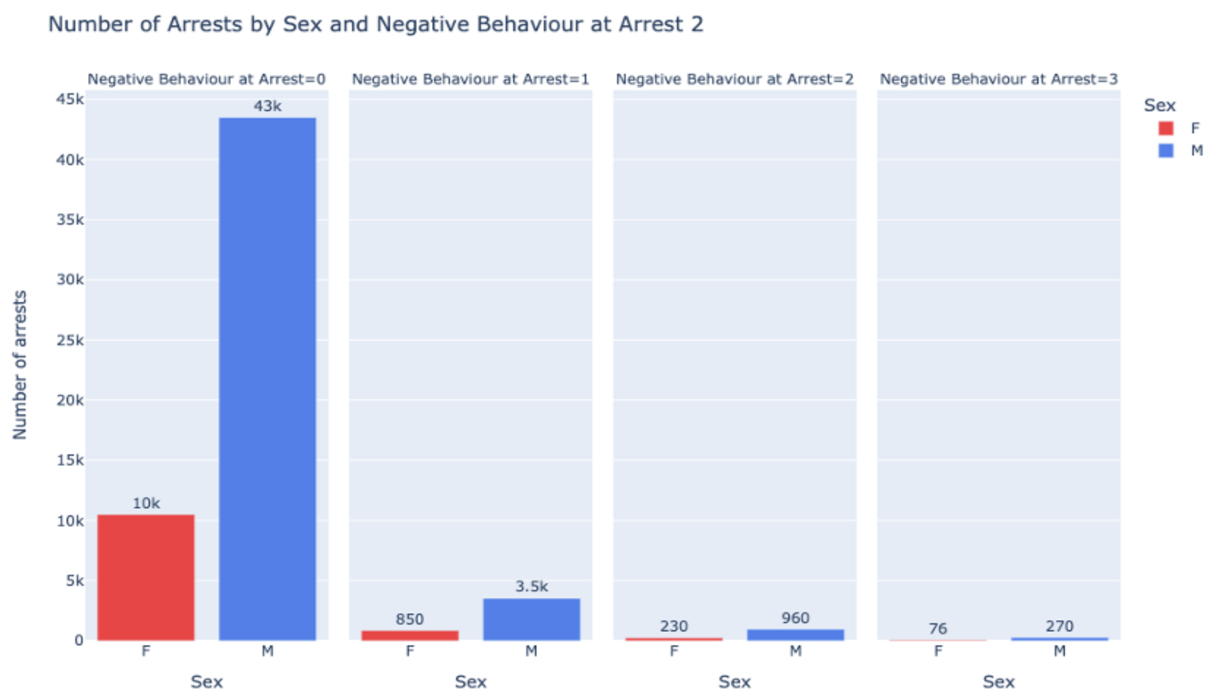
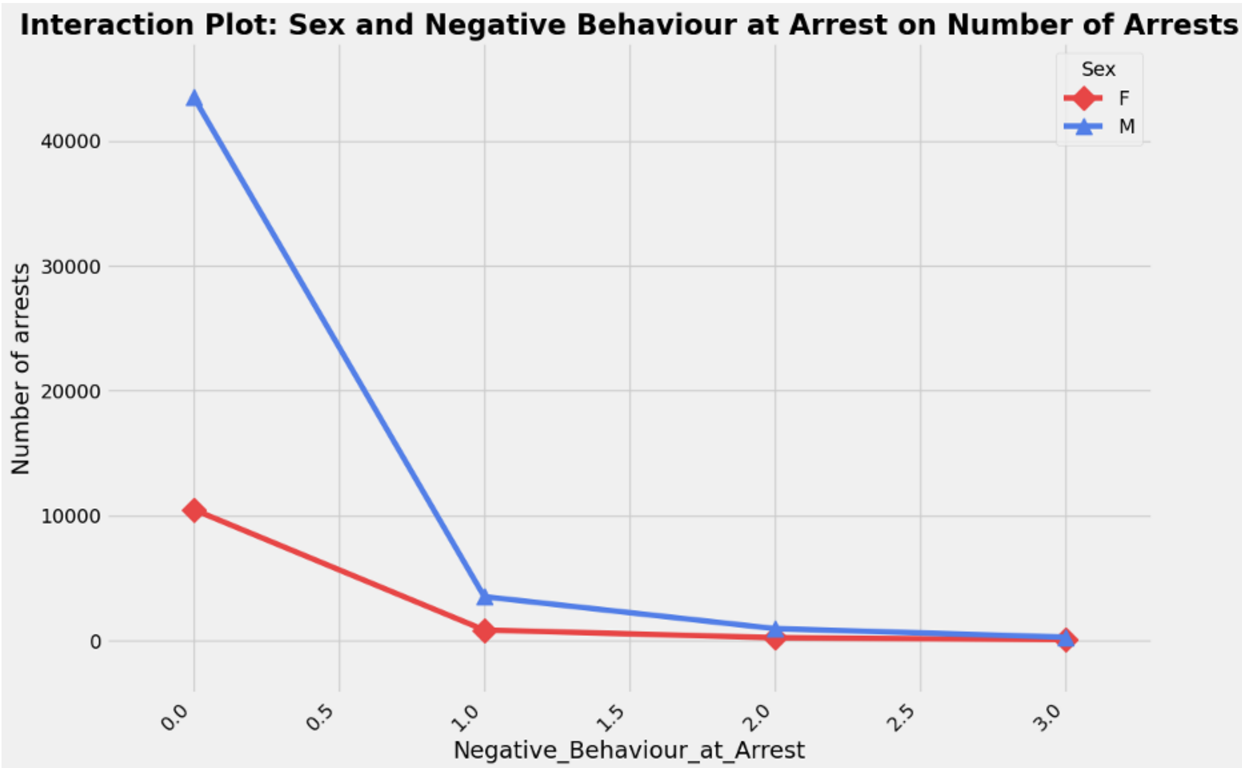
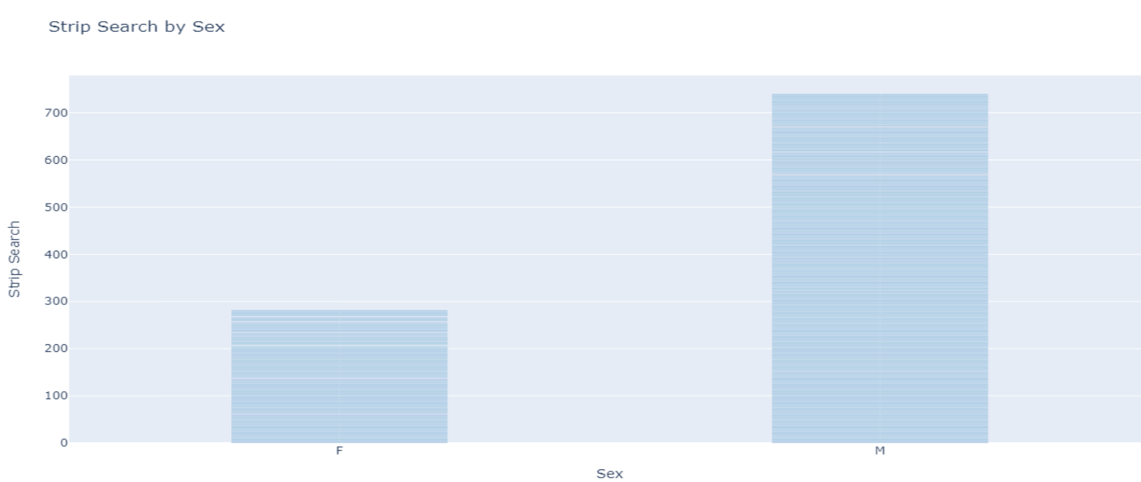


Figure 6



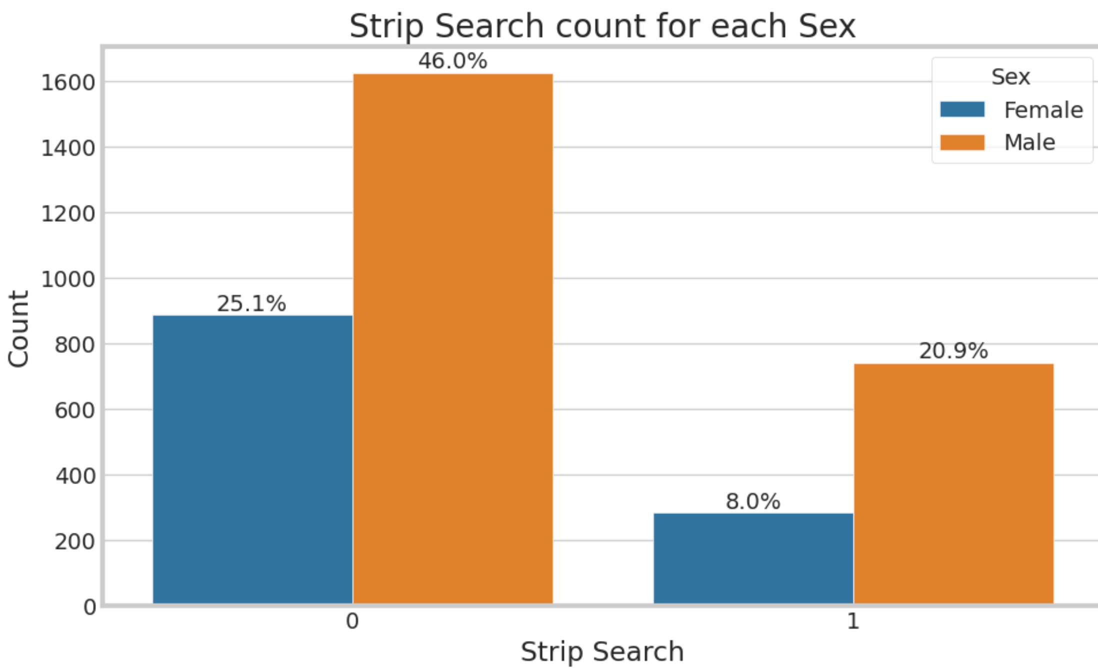
We generated a bar chart to plot the distribution of strip search as per the sex (Figure 7). As expected, the result revealed a notable difference in the strip search count between the males and females.

Figure 7



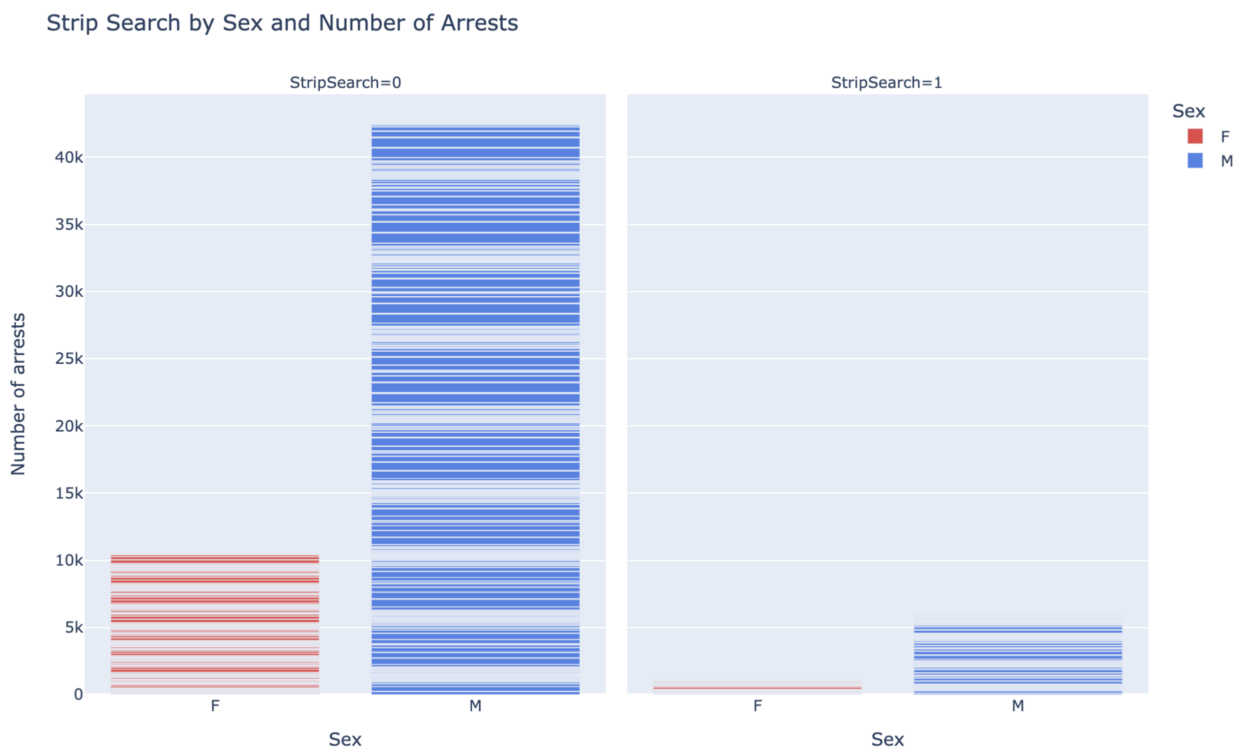
We further generated a count chart to count the distribution of strip search as per the sex for either scenario of the strip search (Figure 8). The results revealed a notable difference in the strip search count between males and females for both the cases of strip searches, with the former having a significantly greater proportion of the strip search recorded in both scenarios.

Figure 8



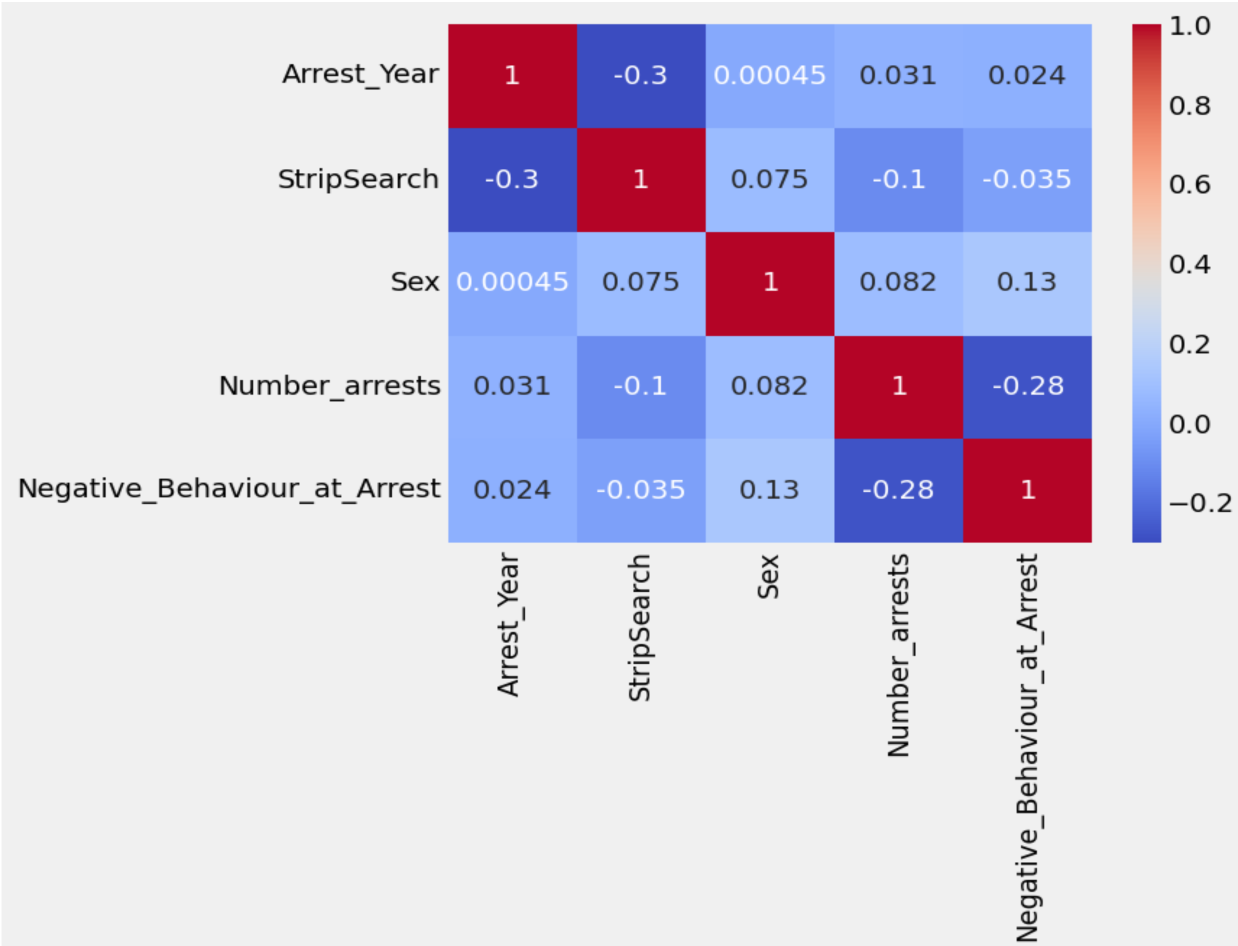
Additionally, we produced a stacked bar chart plot showing the distribution of Strip search broken down by male and female and the number of arrests (Figure 9). Based on our analysis of the Strip search dataset, it is clear that males are more likely to be arrested as compared to females. Moreover, the strip search count of males was comparatively higher than the count for the females. The data suggests that there is a notable difference in the strip search by males and females during arrests. However, we are also aware that more male arrestees are recorded in this dataset.

Figure 9



As referred to in the literature review, we were interested in identifying the relationship strip search had with sex and the number of arrests. Before plotting the correlation matrix all the categorical variables were encoded to numerical. The correlation matrix (Figure 10) supported our understanding of demographics related to the "StripSearch". A correlation between Strip Search and Sex (0.075) and Strip Search and Number of Arrests (-0.1) variables which would be of interest in the Logistic Regression was observed. The selection of our independent variables was further validated as the values were not close to 0 signifying a lower level of correlation. Hence, we finalized Sex and Number of Arrests as our independent variables.

Figure 10



2.2 T-test

In order to further investigate the data, we conducted Welch's t-tests between sex and the number of arrest to determine whether there were significant differences in the means. Welch's t-test was selected over one-sample or two-sample t-tests due to the unequal sample sizes of the two groups since the sample sizes of the two groups are different. In addition, by adjusting for unequal variances, Welch's t-test provides a more precise measure of differences between the two groups. These tests were conducted separately on the 2 different subsets that we had created addressing the two research questions respectively.

Before running the T-test, we ensured the following assumptions were fulfilled.

1. A nominal explanatory independent variable with two levels.

2. A dependent variable is measured on a continuous scale.
3. Normality assumption: The Shapiro-Wilk test shows that the dataset provided is not normally distributed, meaning the normality assumption is violated. However, when the sample size is larger than 50, the Central Limit Theorem suggests that the sample mean will be normally distributed regardless of the distribution of the population. In this case, we argue that the t-test assumption of normality is met and proceed with the t-test.
4. Independence of observations: The observations used in the following t-tests are independent of each other, meaning that the values in one sample do not affect the values in the other sample.

Sex and Number of Arrests

We calculated the mean number of arrests for males and females respectively. Upon observation, we noticed the male average was greater than the female average. Following that, we conducted a Welch's T-test to explore whether there is a significant difference in the mean number of arrests between the male and female arrestees. The following is our hypothesis:

- Null hypothesis H0: The population mean of the two independent groups, male and female are equal in terms of the number of arrests.
- Alternative hypothesis H1: The population mean of the two independent groups, male and female, are not equal in terms of the number of arrests.

Our statistical analysis, with a significance level (alpha) of 0.05 and a 95% confidence interval, yielded a p-value of 1.56e-09, which is less than the predetermined significance level. This result allows us to reject the null hypothesis and conclude that there is a significant difference in the average number of arrests between males and females in the population. Therefore, if we were to randomly select individuals from the dataset, it is likely that we would observe a difference in the average number of arrests between males and females. As a result, we will include the variable 'Sex' in our analysis ANCOVA to determine the impact of sex on the number of arrests.

3 Methodology

3.1 Data Cleaning

We checked for NULL, NA or NaN values in all the columns of the dataset. Since the number of missing value rows is too many, instead of deleting them, we replaced the missing values in numerical columns including Arrest ID, Age, Occurrence, various Search reasons and Items Found with zeros. All the blank rows of categorical variables were replaced by the modal values of those respective columns. After ensuring that the dataset did not have any missing or blank values, we performed the describe and info function to view the summary statistic and data type of each column.

3.2 Dataset Description

In our project, we used a dataset that gives information about arrests and strip searches in Toronto. The dataset comprises information about 65,275 entries for 37347 unique people including 25 different attributes. Each column attribute name represents a variable and each row is an observation. The cell represents the value.

For research question 1 using ANCOVA, we created one subset of the main dataset. The subset, Number of Arrests, comprised 8 columns: Arrest_Year , Arrest_Month, Perceived_Race, Sex, Youth_at_arrest__under_18_years, Age_group__at_arrest_, Number_arrests, and Negative_Behaviour_at_Arrest. We counted the number of times each of these combinations of variables was booked for any offences and renamed to Number_arrests. For Negative Behavior at the time of Arrests, we summed up the values of all action attributes, including Actions at arrest Concealed i, Actions at arrest combative, Actions at arrest Resisted d, Actions at arrest Mental inst and Actions at arrest Assaulted o under one column named Negative Behavior at Arrest. This helped us analyze the patterns of the number of arrests, understand the demographics of the arrested population, and control their negative behaviour at arrest for our first research question.

For research question 2 using Logistic Regression, we created a subset of the main dataset. The subset named- number of arrests comprised 6 columns: Arrest Year, Arrest Month, StripSearch, Perceived Race, Sex, Youth at arrest under 18 years, Age group at arrest, Number of Arrests and

Negative Behavior at Arrest. We summed up the values of all action attributes including Actions at arrest Concealed i, Actions at arrest combative, Actions at arrest Resisted d, Actions at arrest Mental inst and Actions at arrest Assaulted o under one column named Negative Behavior at Arrest. We counted the number of times each of these combinations of variables were booked for any offenses and renamed them under a new column Number_Arrests. We wanted to investigate the reasons that contribute to "StripSearch".

3.3 Research Question 1 using ANCOVA

For ANCOVA analysis, we checked the following 5 test requirements:

1. Data is independently and randomly sampled.
2. The level of measurement is interval/ratio.
3. Normality assumption: Populations are normally distributed. However, this is not very strict, especially since our groups have a large sample size.
4. Independence of observations: The observations in each group should be independent of one another. This means that the value of one observation should not affect the value of another
5. Homogeneity of regression slopes: The relationship between the covariate and the dependent variable should be linear, and the slopes of the regression lines should be equal across groups of the independent variable.

3.3.1 Power analysis

Power is a metric that measures the probability of correctly identifying a positive result. It is calculated as the complement of the probability of failing to identify a true effect, also known as the Type 2 error rate. In practice, a power value of 0.8 is commonly employed. The potential statistical power of an experiment can be determined by taking into account the significance level, sample size and estimated effect size. Therefore, one can also calculate the desired sample size required to achieve a desired statistical power, for example, 0.8, for their statistical experiment, which is the objective of this section.

To understand the ideal sample size for our variable Sex in terms of number of arrests to achieve desired statistical power, we first defined a function called pooled_standard_deviation to calculate the pooled standard deviation of two samples, in this case, males and females. The function would calculate the sample sizes of the two samples, and then calculate the sample variances using np.var function from the NumPy library with the degree of freedom equal to 1. Next, the function calculated the pooled standard deviation. Additionally, a second function called Cohens_d was created to calculate Cohen's d, which is a measure of effect size for the difference between two means, in our case, males and females. The function calculates the mean values of the two samples using np.mean, and then calls the pooled_standard_deviation function to calculate the pooled standard deviation.

Next, we wanted to calculate the effect size for the number of arrests. First, we imported the TTestIndPower class from the statsmodels.stats.power module, which calculated the statistical power of a two-sample t-test. The previously-defined Cohens_d function was used here to calculate the effect size between males and females. The following lines set the significance level, the standard of 0.05, and the desired statistical power, which was 0.8, and calculated the ratio of the sample sizes of males to females. The result will give us the effect size for the number of arrests.

In order to calculate the ideal sample size to achieve the desired statistical power for males and females, respectively, we imported the TTestIndPower class from the statsmodels library. Then, we used the solve_power method of the analysis object to calculate the sample size required to achieve the desired statistical power for a given effect size, significance level, and ratio of sample sizes. The result gave us the required sample size for both males and females to achieve statistical power. We also created a power curve graph for two samples to illustrate the relationship between sample size and effect size, holding power at 0.8 and alpha at 0.05.

3.3.2 ANCOVA

In our ANCOVA analysis, the dependent variable (Number_arrests) was measured on a Continuous scale, and the independent variable (Sex) and a covariate (Negative_Behaviour_at_Arrest) were considered. To perform ANCOVA analysis, we first created a new data frame called “number_arrests__sex_1way” that includes only the variables “Sex,” “Negative_Behaviour_at_Arrest’, and “Number_arrests” from the original data frame. It is also important to ensure the data is correctly stored in the new data frame. Next, we used the ANCOVA function from the pingouin library on the Number_arrests variable, with Negative_Behaviour_at_Arrest as a covariate and Sex as the independent variable. The output of this function provided the conditional probability of predicting Number_arrests in terms of the sex of the individual, holding Negative_Behaviour_at_Arrest constant.

3.4 Research Question 2 using Logistic Regression

For Logistic Regression analysis, we checked the following 6 test requirements:

1. Assumption of a binary dependent variable.
2. Number of observations in each class of the binary dependent variable is roughly equal.
3. Absence of multicollinearity in the independent variables.
4. Assumption that all the observations of our dataset are independent and were randomly sampled.
5. Assumption of no strongly influential outliers in the dataset.
6. Linearity of the independent variables and logs-odds

3.4.1 Logistic Regression:

In Our Logistic Regression analysis, the dependent variable (StripSearch) was measured and had binary values, and the independent categorical variable (Sex) and a continuous variable (Number_arrests) were considered. The value 1 of the dependent variable represented that the individual was strip searched and the value 0 represented that the individual was not strip searched. The categorical variable sex had 2 levels M and F respectively where M represents individuals identified as males and F represents individuals identified as females. The number of arrests was derived by summing the categories of Arrest year, Arrest Month, Race, Sex, Youth or Not and Age group at the time of arrest. Then, to perform Logistic Regression analysis, we created a new data frame called "number_arrests1" that includes only the variables "StripSearch", "Sex", and "Number_arrests" from the original data frame. We used label encoder to encode the categorical variable Sex. The output of this function provided the probability of an individual being strip searched or not in terms of the sex of the individual and the number of arrests. The dimensions of the newly formed subset were 3533 rows and 3 columns. We performed a train-test split on the dataset and split the dataset in a ratio of 80:20. We fit the logistic regression on the larger subset (training data) to train the model. The smaller subset (test data) was further used to evaluate the model performance. After predicting the test data, we computed prediction intervals using following steps:

1. We used the predict_proba method of the logistic regression model to predict the probabilities of the test set.
2. We then computed the variance of the predicted probabilities using the formula:

   variance = predicted_probability * (1 - predicted_probability).

   We further computed the standard error.
3. We then used the norm package of scipy to calculate the z-score. We set the confidence level of 0.95 and computed the z-score for this confidence level. We then calculated the lower and upper bounds of the prediction intervals using the below formulas:

a. lower_bound = predicted_probability - z_score * standard error

b. upper_bound = predicted_probability + z_score * standard error

We could not plot the prediction intervals for the test data as one of our independent variable was categorical and the other was continuous.

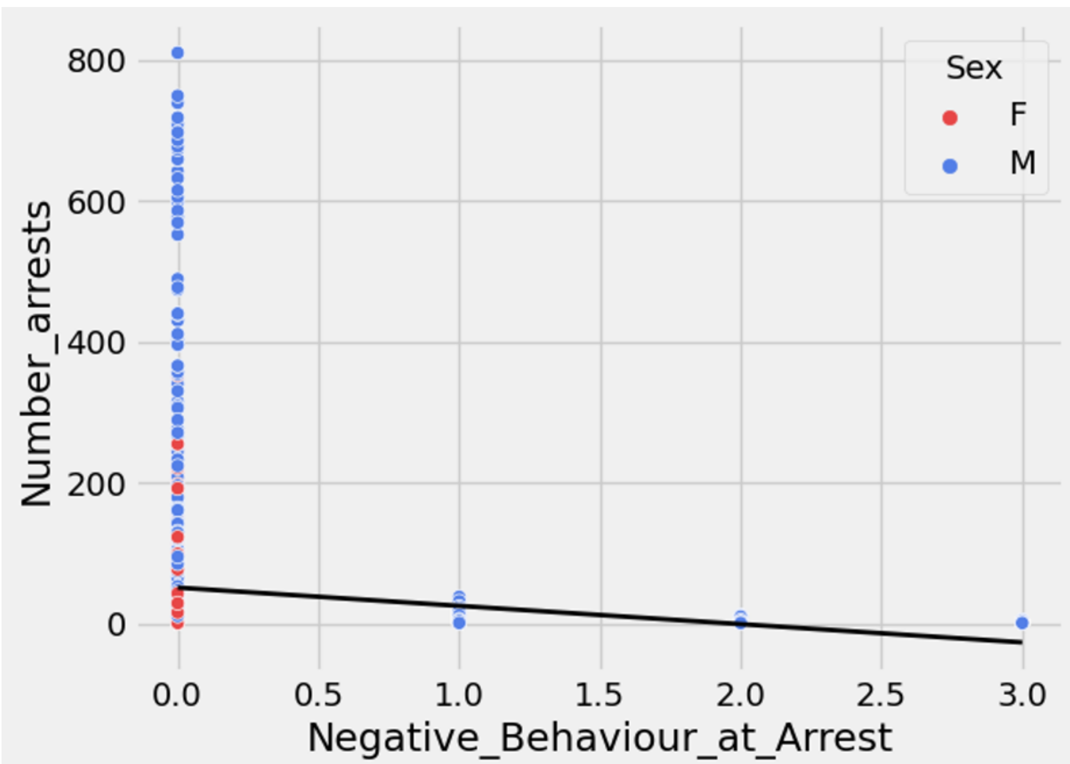
4 Results and Findings

We derived some interesting insights about the underlying data considering several aspects and factors of the data. The first research question was aimed to analyze how the number of arrests differed on the basis of a person's sex after controlling their negative behavior at arrest. The second research question was aimed to examine the relationship between strip search with the sex and number of individuals who got arrested. Overall, the results drew light on some interesting insights about the factors that affected the number of arrests and strip search.

4.1 Research Question 1 using ANCOVA

For ANCOVA assumption checks, we found that the data is independently and randomly sampled. The level of measurement is also interval/ratio. The Shapiro-Wilk test shows that the dataset provided is not normally distributed, meaning the normality assumption is violated. However, when the sample size is larger than 50, the Central Limit Theorem suggests that the sample mean will be normally distributed regardless of the distribution of the population. In this case, we argue that the t-test assumption of normality is met and proceed with the t-test. Furthermore, using Levene's test, we found that the population variances are not equal for the number of arrests. As demonstrated by the scatterplot (Figure 11), the relationship between the covariate and the dependent variable is non-linear, and the regression line slopes differ across the independent variable groups. For the sake of this project, we will continue our analysis, but we will also address these limitations in the discussion section below.

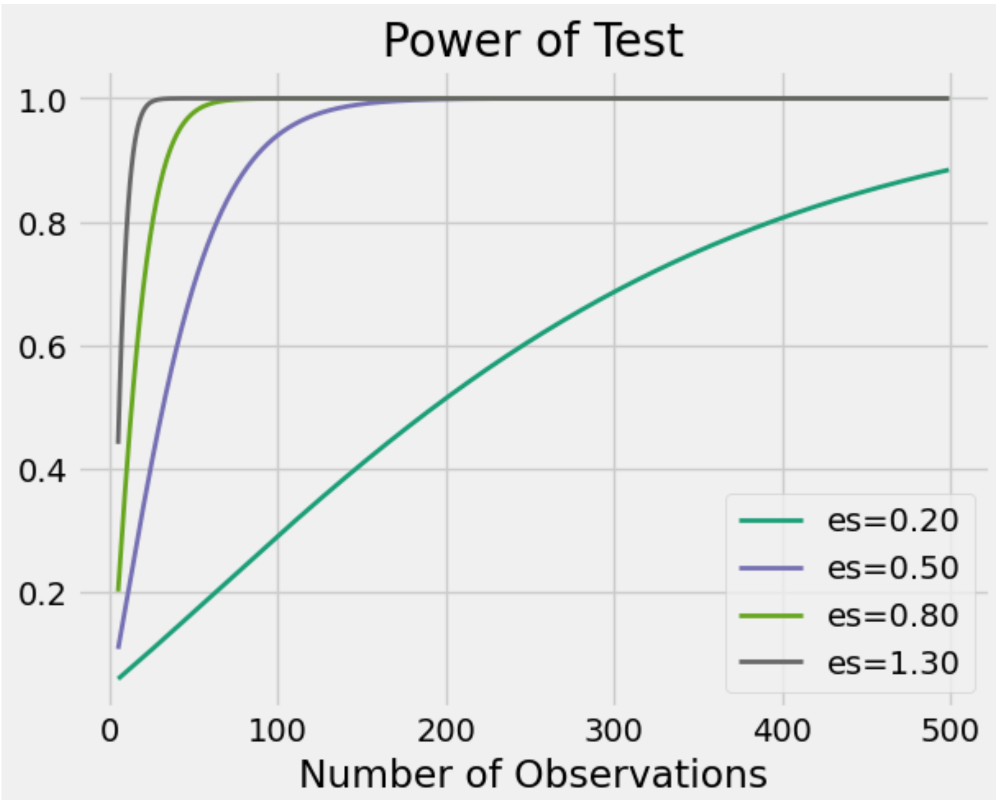
Figure 11



4.1.1 Power Analysis

The power curve graph (Figure 12) illustrates how statistical power varies across different sample sizes for each effect size and each sex group. For instance, at a small effect size of 0.20, the power curve graph shows that a sample size of approximately 400 participants is needed to achieve a statistical power of 0.80 or higher. At a medium effect size of 0.5, a sample size of less than 100 observations is needed to achieve a statistical power of 0.80 or higher. At a large effect size of 1.3, a sample size of less than 50 participants is needed to achieve a statistical power of 0.80 or higher. In other words, as the effect size increases, the number of samples needed to find significance reduces.

Figure 12



In our experiment, we calculated that our effect size for the difference in the mean number of arrests between males and females is 0.18. According to Cohen's guidelines, an effect size of 0.2 is small, 0.5 is medium, and 0.8 is large. As a result, the effect size of 0.18 in our case indicates that there is a small difference in the number of arrests between males and females. With the effect size of 0.18 at an alpha of 0.05 and power of 0.8, we derived that the required sample size for males and females is 335.11 and 643.63 while the actual sample size we have for males and females are 1840 and 958. This indicates that conducting the ANCOVA statistical experiment to test our hypothesis using our current dataset can yield a strong statistical power of more than 0.8. Therefore, we continue our experiment with confidence.

Table 1

Sex	Required Sample Size	Actual Sample Size
Male	335.11	1840
Female	643.63	958

4.1.2 ANCOVA

According to the ANCOVA results of our analysis, Sex (Independent Variable) of a person did have a significant effect on our dependent variable number of arrests after controlling the individual's negative behaviour at arrest as a covariate. In more detail, "uncorrected p-value" for Sex is less than 0.05. Therefore, we can reject the null hypothesis that each of the sex, including males and females, respectively, results in the same number of arrests, even after controlling for their negative behaviour at arrests. In terms of practical interpretation, we hypothesized that the individual's sex could predict one's likelihood of being arrested. Drawing the insight from our analysis, we see that there is a statistically significant relationship between an individual's sex and the likelihood of being arrested when controlling for their negative behaviour at arrest.

Table 2

	Source	p-unc
0	Sex	5.357803e-15
1	Negative_Behaviour_at_Arrest	2.306641e-74
2	Residual	NaN

4.2 Research Question 2 using Logistic Regression

For the Logistic Regression assumption check, we selected the Dependent Variable StripSearch containing binary values 0 and 1 which satisfies the condition of the binary dependent variable. To eliminate the imbalance in our dataset we implemented the oversampling method. The results of the correlation plot suggest a low multicollinearity value between our independent variables Sex and Number of arrests (Figure 10). We could not find any evidence that suggested the lack of outliers. We conducted the Box-Tidwell Test to check the assumption of linearity of independent variables and log-odds Leung, K. (2022, September 13). The test checks for linearity between the predictors and the logit. In our case, the Number_arrests was the continuous independent variable.

Table 3

Continuous IDV	**P>|Z|**
Number_arrests: Log_Number_arrests	0.000

The statistical significance of the interaction term is checked based on the p-values. The p-value of our continuous variable Number_arrests: Log_Number_arrests is statistically significant as the $p \leq 0.05$. We could thus conclude that there is presence of **non-linearity** between the variable Number_arrests and the logit.

4.2.1 Logistic Regression

We could understand the relationship between the independent variables Sex and Number of Arrests with the dependent variable Strip search. From the coefficient value of Sex 0.2796, it can be concluded that there exists a significant relationship between sex and Strip search. Relatively higher proportion of males were strip searched as compared to females. Also, the coefficient for number of arrests was -0.0119 indicating it has a negative relationship with strip search.

The formula used for the Logistic Regression model was **StripSearch ~ Sex + Number_arrests.**

Table 4

	Coef	Significance (P>|Z|)
Intercept	-0.0634	0.268
Sex	0.2796	0.000
Number of Arrests	-0.0119	0.000
Pseudo R-square	0.02472	

Interpretation:

Intercept: The log odds of a person being strip searched is -0.0634 for individuals who are identified as Males (Reference level for IDV Sex) and have a number of arrests score of 0. The effect is however statistically insignificant as the p-value is greater than 0.05.

Coefficient for Sex:

The coefficient for the number of arrests is 0.2796. Compared with individuals identified as Females, the log odds of getting the individuals who are strip searched and identified as Males are 0.2796 higher, controlling for other Sex. Again, this feature is statistically significant.

Coefficient for Number of Arrests:

The coefficient for number of arrests is -0.0119, which means when holding sex constant, increasing the number of arrests by one unit decreases the predicted log of the odds of individuals being strip searched by 0.0119. Again, this feature is statistically significant.

Pseudo R- Square:

This value gives us an indication of how much variation in the dependent variable Strip Search can be explained by our Independent variables Sex and Number of Arrests according to our model. The Pseudo R-squared value is 0.02472, which indicates that only a small proportion of variation in the dependent variable strip search is explained by independent variables sex and number of arrests.

Table 5 Odds Ratio:

	Odds Ratio	Lower CI	Upper CI
Intercept	0.938561	0.839004	1.049932
Sex	1.322623	1.156526	1.512573
Number Arrests	0.988123	0.985175	0.991080

We performed a logistic regression to examine the effects of sex, number of arrests on the likelihood that an individual is strip searched.

As the sex is statistically significant the following is its interpretation: The odds ratio indicates that the odds of strip searching for individuals identified as females are 1.34 times the odds of strip searching an individual identified as male holding number arrests as constant.

As the number of arrests is statistically significant the following is its interpretation: Increasing Number of Arrests was associated with decreased likelihood of an individual being strip searched. The odds ratio is approximately 0.98. For each additional Number of Arrests increase for the individual, the odds that the individual is being strip searched decreases by about 0.98 times.

Table 5

Confusion Matrix:

<table>
<tr><th colspan="2" rowspan="2">Confusion Matrix</th><th colspan="2">Actual Value</th></tr>
<tr><th>Positive (1)</th><th>Negative (0)</th></tr>
<tr><td rowspan="2">Predicted Value</td><td>Positive (1)</td><td>True Positive
249</td><td>False Positive
257</td></tr>
<tr><td>Negative (0)</td><td>False Negative
66</td><td>True Negative
135</td></tr>
</table>

After training the model on the training dataset, we got the following results of prediction for our test data set consisting of 707 samples:

1. True Positive (TP):

   249 individuals that were strip searched were correctly identified.

2. False Positive (FP):

257 individuals that were not strip searched were incorrectly identified as being strip searched.

3. False Negatives (FN):

   66 individuals that were strip searched were incorrectly identified as not being strip searched.

4. True Negatives (TN):

   135 individuals that were not being strip searched were correctly identified as individuals who were not strip searched by the model.

Precision:

TP/ (TP + FP) = 249/ 249 + 257 = 0.492

Proportion of all positive cases that were correctly predicted by model is 49.2%.

Recall :

TP/ (TP + FN) = 249/ 249 + 66= 0.79

Proportion of actual successful cases correctly predicted by model is 79%.

Test Accuracy:

The logistic regression model that we designed to classify if a person was strip searched or not could accurately classify **54.31%** individuals of the test dataset.

Prediction Intervals:

A prediction interval estimates the range of values within which we expect the actual values to be plotted with a certain confidence level. It is used to estimate uncertainty around every single predicted outcome.

5 Discussion and Limitations

ANCOVA Limitations

Our interpretation of how the number of arrests differed on the basis of a person's sex after controlling their negative behavior at arrest maybe limited due to following factors:

1. We observed that samples of the dataset had a relatively higher number of individuals who had low scores of negative behavior at the time of arrest as compared to other levels of the negative behaviors at the time of arrests (Figure 3).
2. The results of the Levene's test highlighted that the population of variances were unequal for the number_arrests. So we could not satisfy the assumption of the independence of the observation.
3. We could not satisfy the assumption of homogeneity of regression slopes. The relationship of the covariate with the dependent variable was non-linear. Also, the regression line slope differed across the independent variables negative_behavior_at_time_of_arrest and sex.

Logistic Regression Limitations

Our interpretation of finding the likelihood of a person being strip searched based on sex and the number of arrests maybe be limited due to following factors:

1. Model may have lower predictive power due to very low correlation between the dependent variable Strip search and independent variables Sex and Number of arrests. Moreover, this factor can affect the reliability of the coefficients.
2. As the data was imbalanced, we implemented oversampling which carries some limitations like overfitting. This further reduces the generalizability and the model can produce biased results.
3. The results of the Box-Tidwell test detected non-linearity between Number_arrests and the logit.
4. Outliers were observed in the data for strip search distribution based on number of arrests. This violates the assumption and might affect the model performance.

To overcome these limitations, larger sample size specific to the association under study can be utilized along with more complex statistical techniques like non-linear transformations, sampling techniques need to be implemented to better understand the relationship between strip search and Number of arrests and Sex.

Possible improvements:

The quality of the dataset can be enhanced by improving the balance between different interrelated features. Gathering more data around this topic may have certain limitations but options including exploring different mediums of data sources like surveys and questionnaires, data gathered from conducting interviews of the focused groups and using different sampling techniques can enhance the overall data quality. The sampling technique can be selected while considering the nature of the research question and the characteristics as it might have an impact on the prediction results. Furthermore, we believe that the experimentation and exploration of data can also help in understanding the relationship between the Strip Search and Number of arrests, Sex variables better. The possibility of exploring the Regularization techniques like L1 and L2 will also help in eliminating the problems arising due to overfitting. We can also evaluate the performances of different classification models to overcome the constraints arising due to the logistic regression model. In order to overcome the limitation of homogeneity of regression for ANCOVA, we can use statistical methods that models the non-linear relationships between the dependent variable and the covariates. Methods like the polynomial regression Agarwal, A. (2018, October 8). As per the research of Olmayan, P., Analizinde, K., Metotlar, K., Cangür, Ş., Sungur, A., & Ankarali, H. (2018) we can also consider the non-parametric methods to deal with this limitation.

6 Conclusion

In our first research question, we aimed to analyze how the number of arrests of an individual varied with respect to the individual's sex after controlling the individual's negative behavior at the time of arrest using the ANCOVA analysis. This finding supports our hypothesis that an individual's sex could predict the likelihood of being arrested. However, we faced several limitations while conducting the analysis. We can further take these drawbacks into account and explore additional factors contributing to the number of arrests. The results of the ANCOVA analysis suggested that the independent variable sex had a statistically significant effect on the dependent variable number of arrests. Our second research question aimed to investigate the relationship that the sex and number of arrests of an individual had with the likelihood of the individual being strip searched using the logistic regression analysis. The results of the logistic regression provided evidence that the likelihood of strip search would slightly increase for individuals identified as males as compared to females whereas individuals whose number of arrests was higher had slightly lesser likelihood of being strip searched. The relationships between the dependent variables and the independent variables were found to be statistically significant.

Our model could correctly identify the individuals that were strip searched with an accuracy of 54.31%. The predictions should however be interpreted with caution as our model faced several limitations due to the imbalance of the samples collected in the datasets and the detected outliers.

7 References

Agarwal, A. (2018, October 8). *Polynomial regression*. Medium. Retrieved April 15, 2023, from https://towardsdatascience.com/polynomial-regression-bbe8b9d97491

Chesney-Lind, M., & Pasko, L. (2013). *The female offender: Girls, women, and crime (*3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Hutchison, J. (2019). “it’s sexual assault. it’s barbaric”: Strip searching in women’s prisons as state-inflicted sexual assault. *Affilia*, *35*(2), 160–176. https://doi.org/10.1177/0886109919878274

Laura Savage. 2019. Female offenders in Canada, 2017. (January 2019). Retrieved February 27, 2023 from https://www150.statcan.gc.ca/n1/pub/85-002-x/2019001/article/00001-eng.htm

Johnson, E. (2019, June 10). *'it was traumatizing': Sunday school teacher strip-searched at Vancouver Airport | CBC News*. CBCnews. Retrieved April 15, 2023, from https://www.cbc.ca/news/canada/british-columbia/sunday-school-teach-strip-searched-at-vancouver-airport-1.5161802

Leung, K. (2022, September 13). Assumptions of Logistic Regression, Clearly Explained. Medium. https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290#:~:text=One%20of%20the%20critical%20assumptions

Lisa Stolzenberg and Stewart J. D'Alessio. 2004. Sex differences in the likelihood of arrest. Journal of Criminal Justice 32, 5 (2004), 443–454. DOI:http://dx.doi.org/10.1016/j.jcrimjus.2004.06.006

Olmayan, P., Analizinde, K., Metotlar, K., Cangür, Ş., Sungur, A., & Ankarali, H. (2018). Araştırma Makalesi / Research Article Düzce Tıp Fakültesi Dergisi. Duzce Medical Journal, 20(1), 1–6. https://dergipark.org.tr/tr/download/article-file/542296#:~:text=ABSTRACT%20Aim%3A%20Nonparametric%20covariance%20analysis

Schwartz, J. (2013). A "new" female offender or increasing social control of women's behaviour? Cross-national evidence. *Feminist Studies, 39*(3), 790-821.

Thomas , T., & Gidda, M. (2023, April 6). *Black girls three times more likely to undergo invasive strip-search by met police*. The Guardian. Retrieved April 15, 2023, from https://www.theguardian.com/uk-news/2023/apr/06/black-girls-three-times-more-likely-to-undergo-invasive-strip-search-by-met-police