Arrests and Strip Searches in Toronto:

A Case Study on race and violent arrests

Final Report

Jean Paul Ngezigihe

Jian Zhang

INF2178H S LEC0101: Experimental Design

Professor Shion Guha

April 16th, 2023

Table of contents

Introduction

"Stop and frisk" is a police practice of temporarily detaining, questioning, and at times searching civilians and suspects on the street for weapons. It became a subject of racial profiling criticism as ninety percent of those stopped in 2017 were African-American or Latino, mostly aged 14–24 (New York Civil Liberties Union, 2019). Seventy percent of those stopped were later found to be innocent (New York Civil Liberties Union, 2019). A similar method also appeared in Canada, specifically in Ontario, as "Carding". Local police enforcement and city officials have largely used carding as a way to "protect the streets" and provide a false sense of security to win votes while negating its psychological impacts on victims (Johnson & al., 2021). It was officially discontinued in the summer of 2014. However, some argue the effect of this colonial method is still pervasive today in Toronto Police Service.

Moreoever, there is evidence to suggest that race may have an impact on how individuals are perceived as combative during an arrest. Research has shown that Black individuals are more likely to be seen as aggressive during police encounters than White individuals, even when controlling for factors such as age and gender (Gau and Brunson, 2010; Piquero et al., 2017). Furthermore, studies have found that Black individuals are more likely to experience use of force during arrests compared to White individuals (Cesario et al., 2018; Fryer et al., 2019). These findings raise important questions about the role that race plays in police interactions and highlight the need for further research on this topic. This is why we believe it is important to understand whether perceptions of race influence police officers' decisions to use force and whether implicit biases towards certain racial groups contribute to violent actions at arrests. In this report, we will investigate the effect of demographic attributes on the Toronto Police arrests and strip searches, using the publicly available dataset "Arrests and Strip Searches".

Our first research question examines: "*Is there a significant difference in strip search based on perceived race*?"

The second research question examines: "*Is there a significant difference in strip search based on perceived race and age?"*

Our third research question examines: "*Does race have an impact on violent actions" during an arrest*?"

Literature review

According to the Toronto Police Service, a strip search refers to a search conducted by a police officer on a person, which includes the removal of some or all clothing and a visual inspection of the body. Scholars from the International Journal for Crime, Justice and Social Democracy describe strip searches as a police misuse of power with potential harm on individuals and the society. While scholars agree on strip searches being abusive, others have questioned why governments are funding police departments more than ever (Spurrier, M. (2023). Police departments are now increasingly turning to algorithms to respond to such criticisms (Joh, E. E. (2017). By using artificial intelligence to decide which neighborhoods to increase efforts, they argue their practices to be fair and non-discriminatory. However, such practices still find themselves rooted in colonial practices (Barabas & al., 2020). Users of algorithmic tools for decision making need to develop more robust frameworks for understanding their work and inquire their own perspective and position of power (Barabas & al., 2020).

Additionally, The bias towards Black people being seen as violent is a well-documented phenomenon in the field of social psychology. A study conducted by Correll et al. (2002) found that participants, regardless of their own race, were more likely to shoot a Black target than a White target in a simulated shoot/don't shoot task. Another study by Eberhardt et al. (2004) found that participants were more likely to perceive ambiguous objects as weapons when they were held by Black individuals compared to White individuals. These findings suggest that there is a stereotype in society that associates Black individuals with violence, leading to biased perceptions and behaviors towards them. Additionally, a study by Goff et al. (2008) found that Black boys as young as 10 years old are more likely to be seen as older, guilty, and deserving of punishment compared to White boys. These biases can have serious consequences, as they can contribute to police brutality, racial profiling, and unjust treatment of Black individuals in the criminal justice system.

This report aims to use a studying up lenses while investigating our research questions,

Exploratory Data Analysis

Descriptive statistics

Using df.head(), we see the dataset consists of 25 columns with information on individual characteristics such as sex, age group as well as information about the arrest; action at arrest, whether items were found, etc. More information on the dataset in the following section. Furthermore, we identified there were 65276 entries in this dataset and obtained information about the column data types and counts. See Table 1 below for first 10 variables:

Table 1: Data type and count for first 10 variables

#	Column	Non-Null	Count	Datatype
0	Arrest_Year	65276	Non-Null	int64
1	Arrest_Month	65276	Non-Null	object
2	EventID	65276	Non-Null	int64
3	ArrestID	64807	Non-Null	object
4	PersonID	65276	Non-Null	int64
5	Perceived_Race	65276	Non-Null	object
6	Sex	65276	Non-Null	object
7	Age_group_at_arrest	65276	Non-Null	object
8	Youth_at_arrest	65276	Non-Null	object
9	ArrestLocDiv	65276	Non-Null	int64
10	StripSearch	65276	Non-Null	int64

This table informs our method choice as we know the datatypes of our variables we choose to investigate. Next, we performed a descriptive summary of our numeric variables

Table 2: Descriptive statistics of some of the numeric variables in the dataset

index	StripSearch	Booked	Actions_at_arrest___Concealed	Actions_at_arrest___Combative	ItemsFound
count	65276	65276	65276	65276	7801
mean	0.11	0.51	0.004	0.04	0.37
std	0.32	0.49	0.06	0.20	0.48
min	0.0	0.0	0.0	0.0	0.0
25%	0.0	0.0	0.0	0.0	0.0
50%	0.0	1.0	0.0	0.0	0.0
75%	0.0	1.0	0.0	0.0	1.0
max	1.0	1.0	1.0	1.0	1.0

From this table, we see many of our numeric variables range from 0 to 1. Therefore, descriptive statistics such as minimum or maximum do not provide much information.

Preprocessing

From this initial exploratory analysis, we realized there were many variables we did not need. Therefore, we preprocessed the data to remove unnecessary variables. Additionally, our dependent variable of interest is strip search. However, it is non-continuous. Therefore, we performed some preprocessing to make it a continuous variable and perform our General Linear Model (GLM). We then create different tables for analysis. See below Table 3 for Strip search by race and groupand booking.

Table 3: Strip search by race and group

Arrest_Year	Perceived_Race	Age_group__at_arrest_	StripSearch
2020	Black	Aged 17 years and younger	147
2020	Black	Aged 18 to 24 years	596

2020	Black	Aged 25 to 34 years	807
2020	Black	Aged 35 to 44 years	427
2020	Black	Aged 45 to 54 years	175

Table 4: Booked by race and age group

Arrest_Year	Perceived_Race	Age_group__at_arrest_	Booked
2020	Black	Aged 17 years and younger	332
2020	Black	Aged 18 to 24 years	1181
2020	Black	Aged 25 to 34 years	1798
2020	Black	Aged 35 to 44 years	954
2020	Black	Aged 45 to 54 years	426

Then, we combined both tables to form our data frame in Table 5 “Strip search and Booked by race and age group” which we used for more exploratory data analysis.

Table 5: Dataframe combining Strip search and Booked by race and age group

Arrest_Year	Perceived_Race	Age_group__at_arrest_	StripSearch	Booked
2020	Black	Aged 17 years and younger	147	332
2020	Black	Aged 18 to 24 years	596	1181
2020	Black	Aged 25 to 34 years	807	1798

2020	Black	Aged 35 to 44 years	427	954
2020	Black	Aged 45 to 54 years	175	426

Then, we determine if there are any missing data points or duplicate rows in our dataset. There were none. Then, we verified for missing values. Table 6 shows there were no missing values.

Table 6: DataFrame missing values

Variable	Missing values
Arrest_Year	0
Perceived_Race	0
Age_group__at_arrest	0
StripSearch	0
Booked	0

Table 7 shows the total number of incidents in which the arrestee was considered combative or strip-searched, broken down by year of arrest, perceived race, and age group. This information can help identify patterns and potential confounding variables, and we will use this information to analyze the data for the ANCOVA.

Table 7: Dataframe combining Action_at_arrest_Combative and StripSearch by race and age group

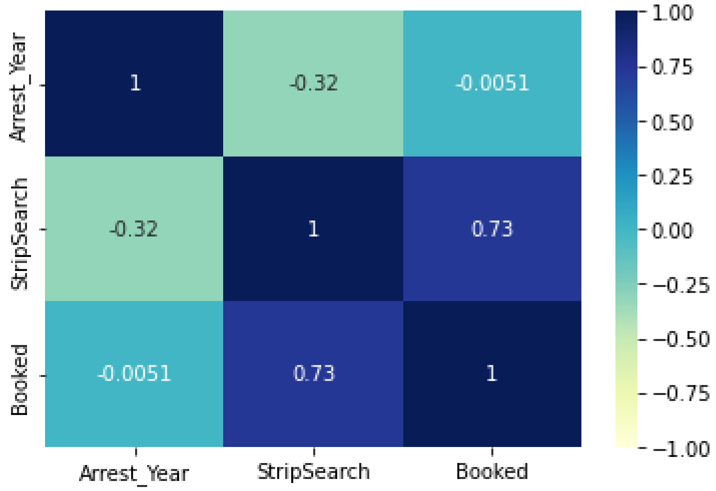
Arrest_Year	Perceived_Race	Age_group__at_arrest_	Actions_at_arrest___Combative__	StripSearch
2020	Black	Aged 17 years and younger	15	147
2020	Black	Aged 18 to 24 years	99	596

2020	Black	Aged 25 to 34 years	183	807
2020	Black	Aged 35 to 44 years	100	427
2020	Black	Aged 45 to 54 years	31	175

Data Visualization

We identified which variables had a higher correlation with each other.

Figure 1: HeatMap

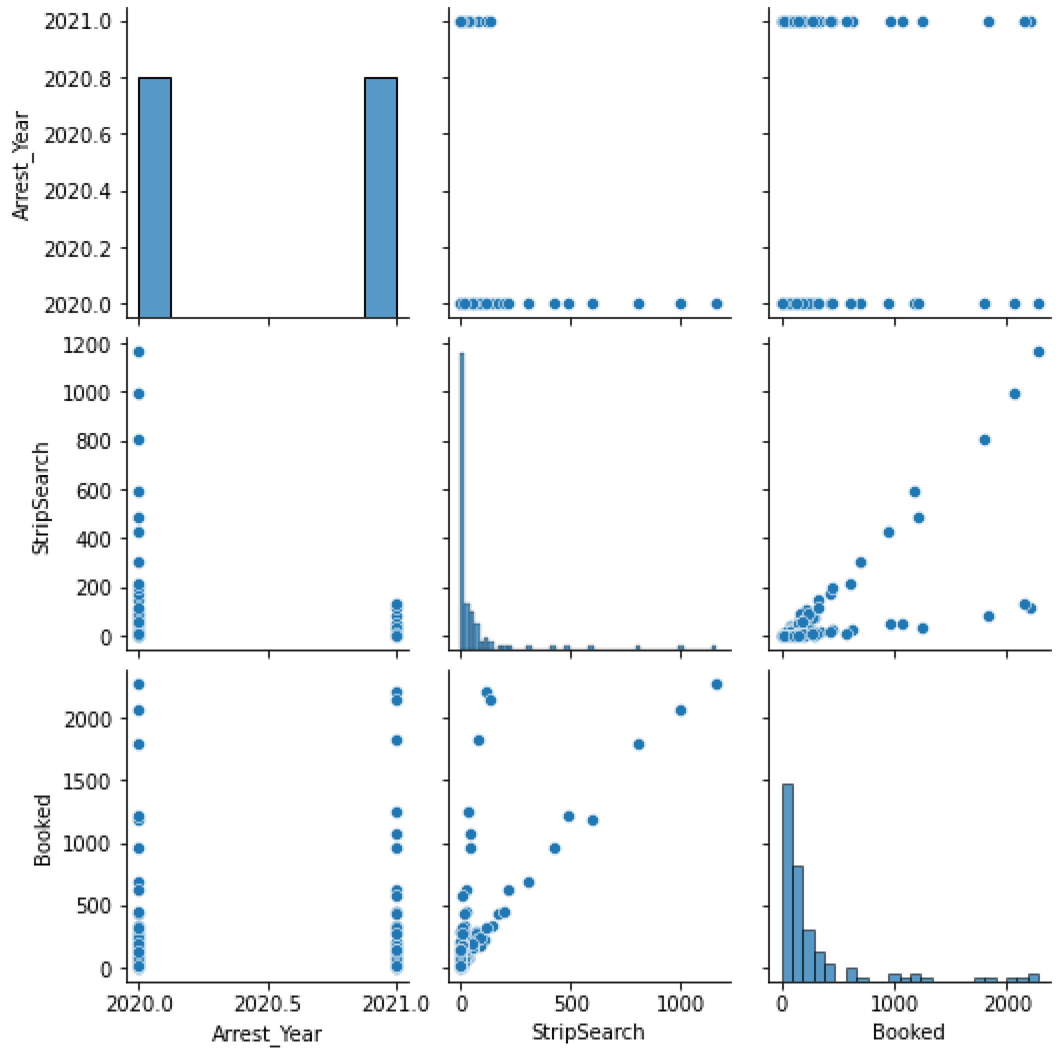


Starting with the first variable from top left we see Arrest year has a correlation value of -0.32 and -0.0051 with StripSearch or Booked. This indicates no correlation with these variables. This informs us that police officers' arrests and bookings are not impacted by the year.

StripSearch has a correlation value of -0.32 and 0.73 with Arrest_year and Booked. This indicates no correlation with Arrest_Year and a positive correlation with Booked. This informs us that police strip searches are correlated to being subsequently booked. However, we may not make inference here since this is not causation and more analysis is required.

Booked has a correlation value of -0.0051 and 0.73 with Arrest_year and StripSearch. This indicates no correlation with Arrest_Year and a positive correlation with StripSearch. This informs us that police bookings are correlated to having been striped searched before. Again, this information cannot tell us whether strip search causes bookings, but informs it could be a potential analysis step.

Figure 2: Pairplots between numeric variables



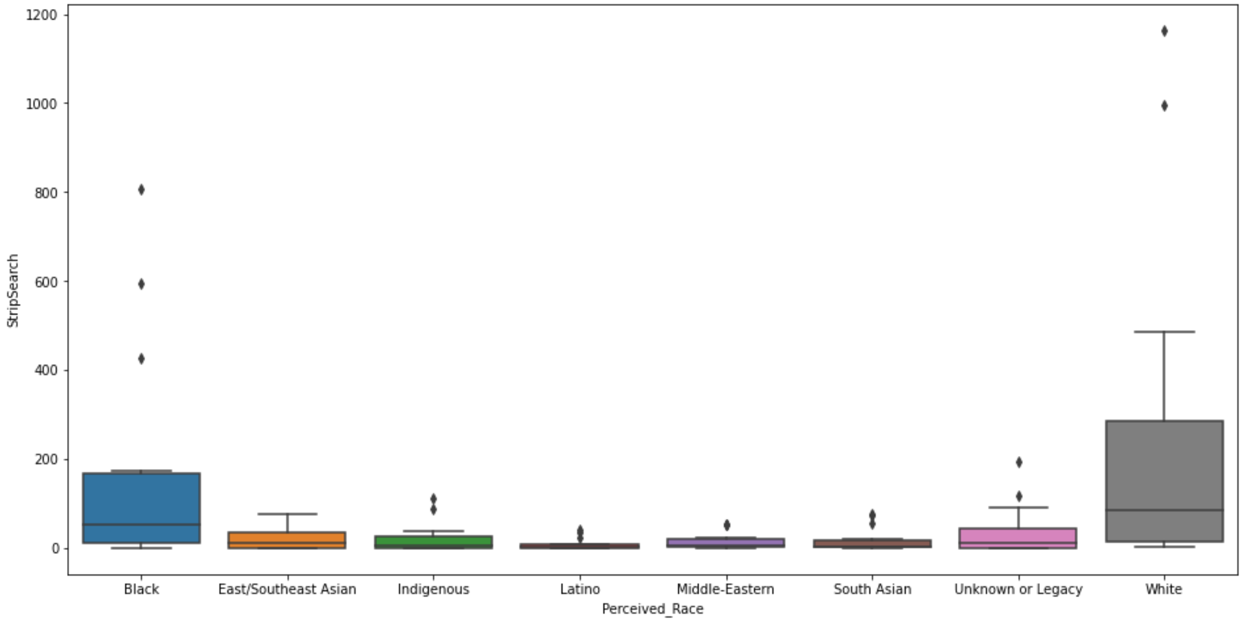
The pair plots allow us to graphically see the relationship between our numeric variables. There is no trend in the plots associated with Arrest_Year. In the Strip Search and Booked plots, we see the data points moving upward to the right. This displays there is a positive relationship between StripSearch and Booked. This confirms our previous analysis from the heat map and informs us we might need to investigate them further.

Box Plots

According to the Arrest and Strip Search (RBDC-ARR-TBL-001) data set, there may be some difference in the likelihood of being booked for ethnic groups that are not white compared to white race. That is, racial groups that are non-white may have a greater likelihood of being booked compared to White racial groups.

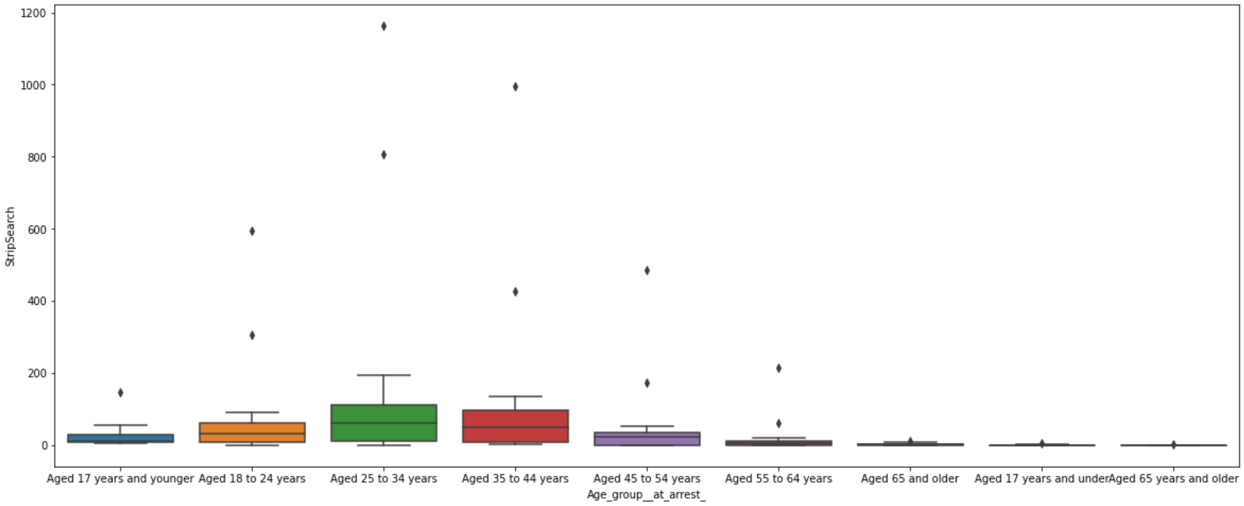
Based on the Arrest and Strip Search (RBDC-ARR-TBL-001) dataset, we found some variation in the likelihood of strip searches based on race and age. In detail, ethnic groups that are not white may have a higher probability of being strip searched compared to white race. Also, younger groups are more likely to be strip searched than older groups.

Figure 3: Distribution of Race in Strip Search Boxplot



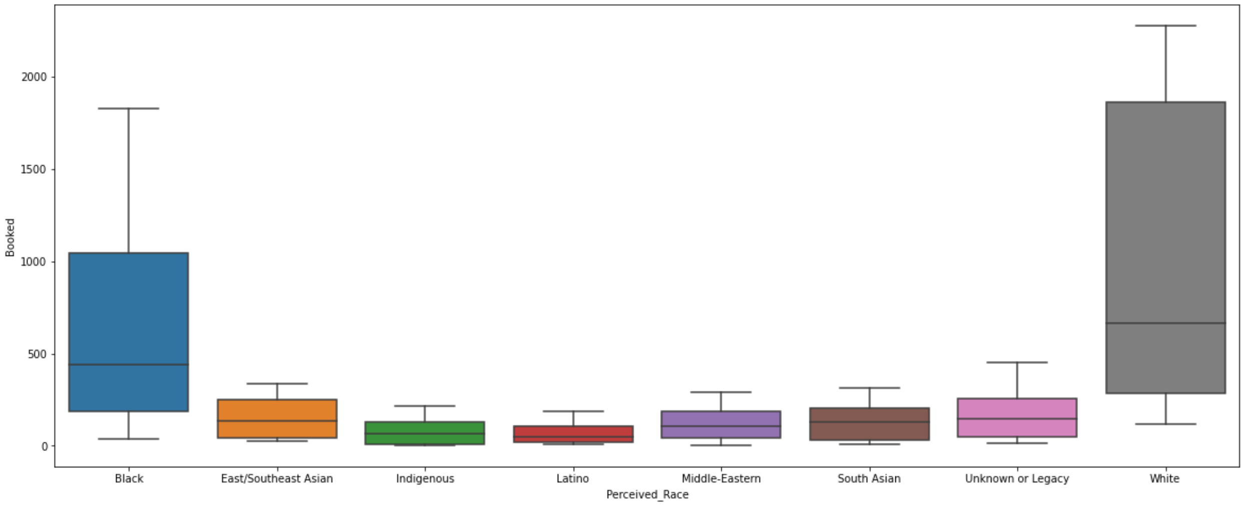
From the boxplot, we can observe the distribution of strip search results by perceived race, and we find that there are differences in the probability of being strip-searched by different perceived races. For example, the boxes for non-whites are higher and may have some outliers than those for whites, indicating that non-whites are more likely and more often strip-searched at the time of arrest than whites. This finding also implies that there may be some racial bias in the law enforcement practices of the officers.

Figure 4: Boxplot of Strip Searches by Age Group at Arrest



The boxplot expresses the distribution of strip searches of individuals by age group at the time of the arrest. We can see that the probability distribution is wider for the younger group compared to the older group, suggesting that the younger group has a relatively greater variation in the probability of being strip-searched, while the older group is likely to receive a relatively consistent search procedure. In addition, we can see boxes for the 18-24 and 25-34 age groups (with the 26-35 age group having the highest IQR) compared to the other age groups), which also suggests that these two age groups are more likely to be strip searched compared to the other age groups. Overall, there was some variability across age groups regarding the probability of being strip-searched, suggesting that age could be a factor in the decision to strip-search.

Figure 5: Boxplot of Booked by Perceived Race at Arrest



This Boxplot shows the variable "Booked" distribution across the different categories of "Perceived_Races". We can discover from the graph that the Black group has the highest median value in the arrested, so perceived race and their IQR are also larger, indicating a more significant variation in Book for this group. This boxplot also suggests that there may be some racial differences in the likelihood of being registered by the officers.

Figure 6: Boxplot of Combative Actions at Arrest by Perceived Race

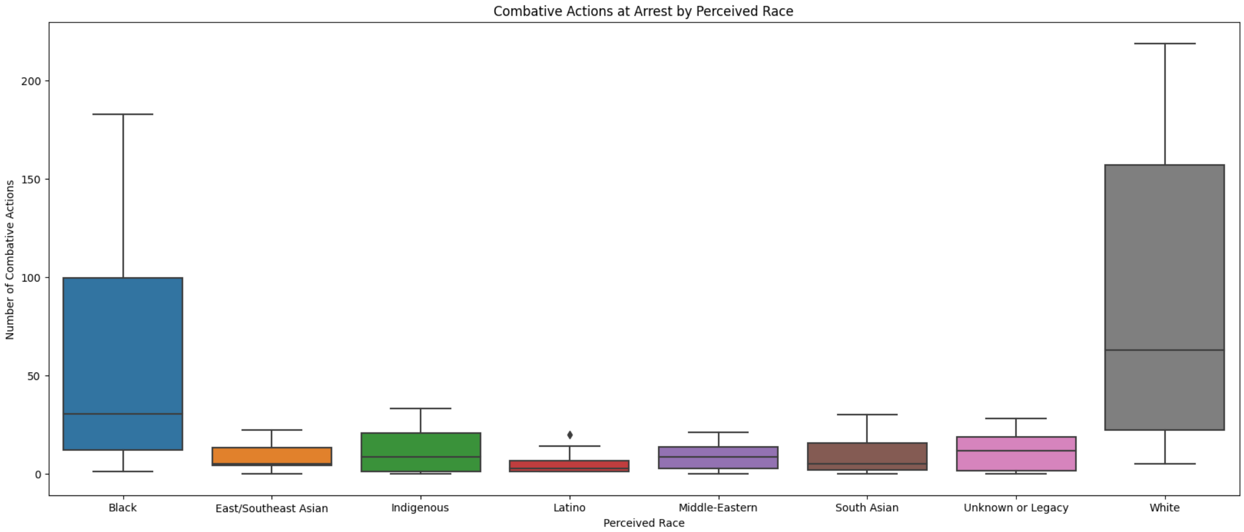


Figure 6 shows the distribution of combat actions taken at arrest across the different perceived racial categories, allowing us to observe more visually the distribution of combat actions at arrest for each perceived racial category. The x-axis shows the eight racial categories and the y-axis shows the number of combat actions. We can learn from the graph that the other races have a higher median number of combat actions at the time of arrest compared to individuals perceived as white. However, within each racial category, the number of combat actions taken varied significantly. This sign indicates that there may be racial differences in the number of combat actions taken at the time of the arrest. Also, it suggests that perceived race may be one of the essential factors in combat actions taken at the time of the arrest.

Figure 7: Combative action at Arrest by Perceived Race and Strip Search

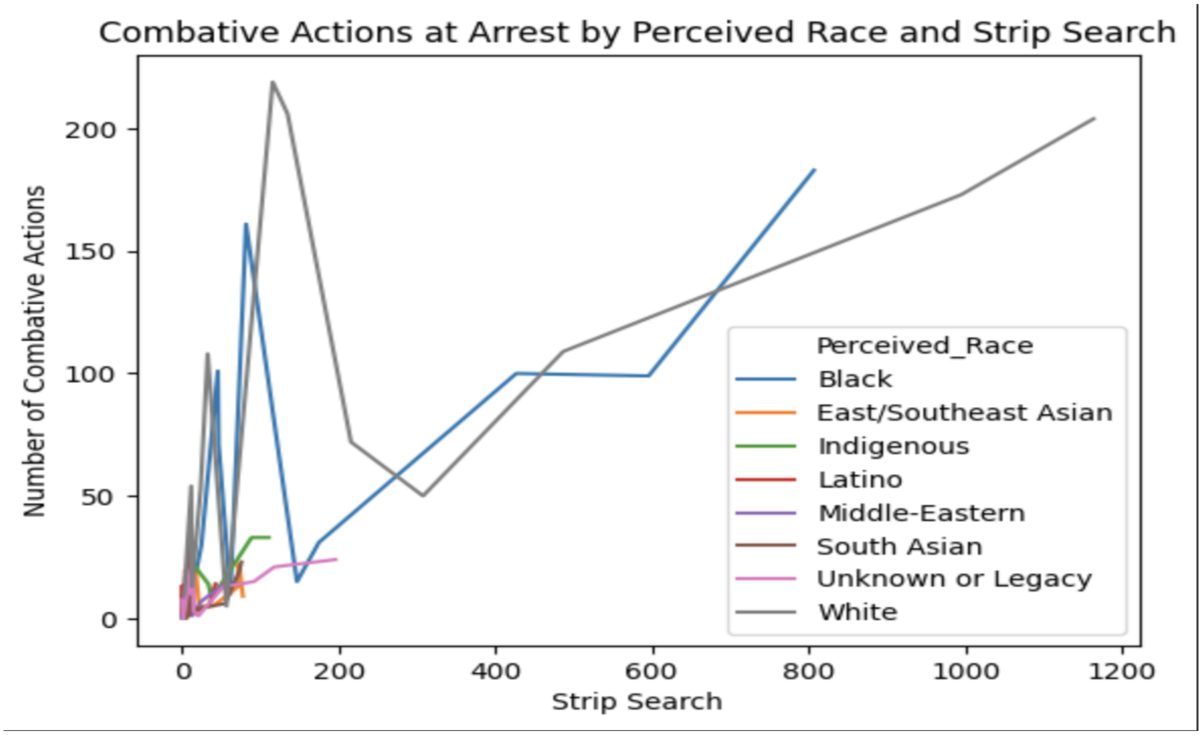


Figure 7 representation of the chart shows how the number of confrontational actions at the time of arrest varies depending on whether a strip search was conducted, and how this relationship varies across perceived racial groups. X-axis "*StripSearch*" indicates whether a strip search was conducted at the time of arrest. y-axis "*action_at_arrest_combative*" indicates the number of combative actions taken at the time of arrest. Each line indicates the "Perceived_Races" who engaged in confrontational behavior at the time of arrest. The graph indicates that the number of combative actions taken at the time of arrest was higher when strip searches were conducted and that this effect was consistent across all perceived racial groups. However, there are still limitations to answering our third research question because the data still have potential confounding variables that may affect the relationship between strip searches, adversarial actions, and perceived race. Therefore, we will build ANCOVA and logistic regression models to investigate the interplay between the variables that influence "*action_at_arrest_combative*"

Power Analysis

Now, prior to computing any statistical test to answer our research questions, we calculate the effect size of the explanatory variable using Cohen's D metric. We chose Cohen’s D because it provides a standardized measure of the magnitude of the effect, which we will use later to determine how many samples we need. We are interested in strip search and actions at arest_combative as outcome variables . Therefore, prior to the t-tests to analyze whether strip searched (outcome variable) differed between perceived white race and perceived black race, we ran two Cohen’s D effect size tests. We obtain the results below

Table 8: Effect size results for Action_at_arrest_Combative and Perceived race

Outcome variable	Effect Size
Actions_at_arrest___Combative__	0.006
Perceived_Race	0.030

After obtaining the effect size, the required sample size was computed using the obtained effect size and establishing the statistical power at 80%. Indeed, it is an essential step as it allows us to determine the appropriate sample size needed to detect a statistically significant effect. The results in Table 9 indicate that a sample size of 22,182 was required for Black Race, while a sample size of 14,023 was required White Race. The actual sample size provided in the dataset are 117 526 and 27,723 respectively, which is only 4656 missing for Black race and (13.700) in extra for White race. Therefore, our sample size for the outcome variable perceived race is almost enough to have enough statistical power. Parallely, the results in Table 9 indicate that a sample size of 525,340 was required for Black Race, while a sample size of 332,11 was required White Race. This is significant because the sample size provided in the dataset are 117 526 and 27,723 respectively, which is considerably lower. This low sample size impacts the reliability of the results and may limit experiment internal validity and future claims on this outcome variable Actions_at_arrest___Combative__.

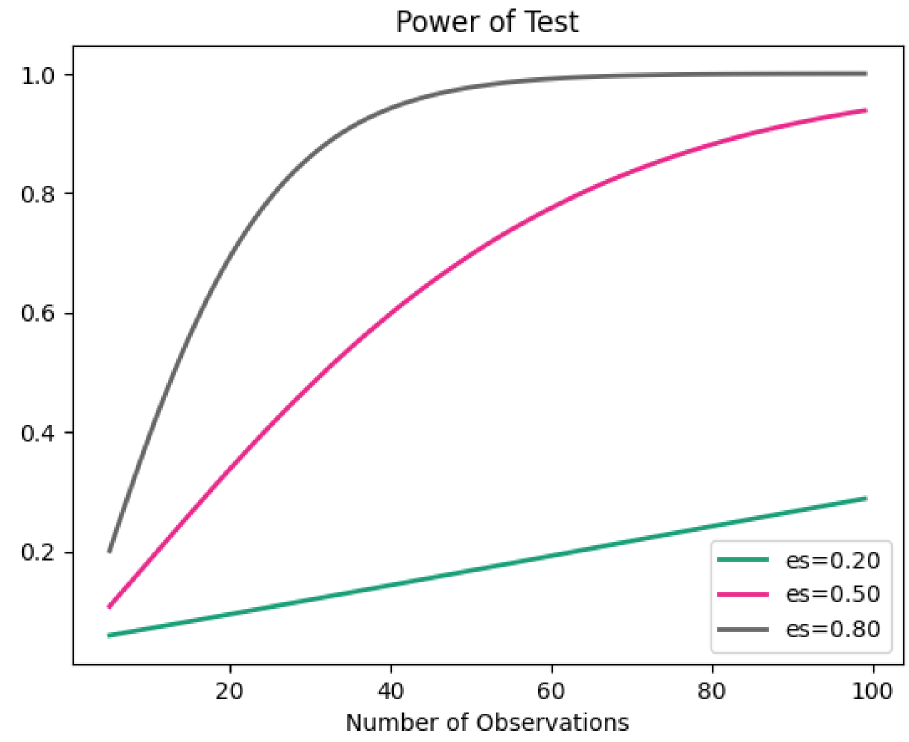
Table 9: Sample size needed for Action_at_arrest_Combative and Perceived race

Actions at arrest Combative	**Sample size needed**	**Actual Sample Size**	**Difference(n)**
Black race	525,340	17,526	50,7814
White race	332,111	27,723	304,388
Perceived race	**Sample size needed**	**Actual Sample Size**	**Difference(n)**
Black race	22,182	17,526	4656
White race	14,023	27,723	(13,700)

According to Howard J. Seltman, there are three ways to improve statistical power;, reducing variance, increasing *n*, and increasing the spacing between population means Seltman, H. J. (2012). The dataset is already collected therefore, we do not have the ability to obtain more samples or manipulate the experiment as we are post-hoc. Alternatively, we may adjust the level of significance effect size to match our sample size needed, but this approach may increase the risk of Type I or Type II errors. The graph in figure 6 displays the power of the test with different effect sizes. By choosing an effect size of 0.80 earlier, we ensured that our experiment will have the best statistical power at 1.0. This will minimize the likelihood of incorrectly rejecting the null hypothesis if some alternative hypothesis is really true.

Ultimately, because of the lack of experiment design modification available since the data was already collected, and to retain a high statistical power, we accept the limitation in number of samples for the action_at_arrest_combative outcome variable and will consider it in further statistical tests and methods.

Figure 8: Power Graph



T-tests

From the EDA, we see a high proportion of strip search on perceived race Black. We wonder if there is a significant difference in the means compared to the overall groups. Additionally, we see a high count of strip searches for young adults. We are curious to see if there is significant difference in the means compared to the overall groups. In this section, we will perform T-Tests to test these hypotheses.

Hypothesis for t-tests

Is there a significant difference in the means of Black compared to the overall population?

H0: There is no significant difference in the means of Black compared to the overall population
H1: There is a significant difference in the means of Black compared to the overall population

Is there a significant difference in the means of Indigenous compared to the overall population?

H0: There is no significant difference in the means of Indigenous compared to the overall population
H1: There is a significant difference in the means of Indigenous compared to the overall population

* The hypothesis for the other t-test would follow the same convention *

Assumptions check

As we will be using t-test, we need to check if it meets its assumptions.

1. **The assumption for a t-test follows a continuous or ordinal scale**. Yes, our data measurement strip search is continuous.
2. **The second assumption made is that of a simple random sample.** Yes, our sample collected a representative, randomly selected portion of the total population.
3. **The data is a normal distribution, bell-shaped distribution curve.**

Shapiro test result

Subsequently, to assess the distribution of our data, we performed a Shapiro test. The p-value is 0.0 which is less than the alpha of 0.05. We therefore reject the null hypothesis. We have sufficient evidence to say that the sample does not come from a normal distribution. Although this assumption is not met, we will continue our test and note it as a limitation.

4. **The data is a reasonably large sample size**. Yes, our data is reasonably large with above 500 data points.
5. **Data sample is independent.** Individuals arrested are independent of each other. Yes, our data is independent.

Overall, more than one assumption is violated. However, we will still run the statistical test and note it as a limitation.

Results

Using a two-tailed t test, we find most groups are significantly different. However, we use a one-tail t-test, alternative ‘greater’ to determine if the mean of the group is greater than the population mean, which is our interest of research. The results in Table 10 indicate that the mean strip search for the Black group is higher than the mean of the population (M=9.04). With alpha established at 0.05, this is a statistically significant difference as the p-value (0.003) is less than 0.05, 95% CI [12.93, 31.33]. Therefore, we can reject the null hypothesis that there is no difference in strip search for the Black group and the population. However, it is important to note we do have a limitation on this test since it fails the normally distributed assumption

Table 10: One sample one-tail 'greater' T-Test results

Group	P Value	Significant/Not Significant
Black	0.003	Significant
Latino	1.0	Not Significant
Indigenous	1.0	Not Significant
Middle-Eastern	1.0	Not Significant

Method

Dataset description

The Arrest and Strip Search dataset (RBDC-ARR-TBL-001) collects demographic information on individuals arrested and strip-searched by relevant law enforcement agencies. Each row represents an arrest with demographic information and information about whether they were strip searched or booked at a police station within 24 hours. The dataset contains a mix of integer and string. The datatypes are mostly categorical such as age and arrest month with a few continuous columns. The dataset can be found here.

The goal of our method analysis is to demonstrate statistically whether arrestees who resulted in strip searches were influenced by their age. For example, are younger individuals perceived as more likely to engage in criminal activity, increasing the likelihood of being strip searched.

Another research question examines whether people of different racial backgrounds are more likely to be arrested and booked by the Toronto Police Service. Our methods provide us with a quantitative argument to answer our research question and potentially make other correlation claims on potential bias and discrimination in the Toronto Police department.

It is worth mentioning that the "perceived race" values in the arrest and strip search dataset (RBDC-ARR-TBL-001) are essentially based on the personal opinions of the officers involved in the arrest and strip search at the time. For example, the accuracy of perceived race is largely limited by the officers' biases and assumptions about race at the time and/or time of day. In other words, because of some bias of the officers at the time, it is likely that the actual race of the arrestee i not be inaccurate to the race registered may include for error

ANOVA

One Way

So far, the EDA demonstrated some perceived groups such as Black are searched more often than other groups (Figure 3). Additionally, younger age groups seemed to be striped search more often (figure 4). T-Test is a good start to find correlation between groups. However, we need a more robust model to formulate conclusions. With ANOVA (Analysis of Variance), we can determine whether the variation between the means of the groups is larger than the variation within the groups, and whether this difference is statistically significant. We will perform a one-way ANOVA to answer our first research question. The independent variable will be perceived race. The dependent variable will be strip search.

H0: $\mu 1 = \mu 2 = \mu 3 = \mu ..$ (where μ = mean) There is no significant difference in strip search for perceived race
H1: There is a significant difference in strip search for perceived race

Two Way

We are also interested in the combined effect on age groups. We will run a two-way ANOVA to answer our second research question. It will inform if there is a significant effect of each factor on the dependent variable, as well as whether there is an interaction effect between the two factors. The two factors will perceive race and age groups at arrests as independent variables. Strip search will be our dependent variable.

H0: There is no significant difference in strip search by perceived race.
H1: There is no significant difference in strip search by age.
H2: There is no significant interaction between perceived race and age, or their main effects, on strip search.

Assumptions checks

1. **Independence**: The observations within each group of our dataset are independent of each other. This assumption is passed.
2. **Normality**: The data within each group is not normally distributed. See Figure 6. This assumption fails.
3. **Homogeneity of Variance**: The variance of the data within each group is equal.
   Levene's test

To assess the homogeneity of variance assumption, we perform the Levene's test. The p-value is 6.22e-43. This is below the alpha of 0.05. Therefore we reject the null hypothesis that the variance among groups is equal. This assumption fails.

4. **Random Sampling**: The groups are formed by random sampling from the population. Indeed, the arrests came from arrests that were not controlled. This assumption is passed.

The assumptions are violated. However, we will still run the statistical test and note it as a limitation.

Post-hoc tests e.g, Tukey's HSD

We will run a post-hoc Tukey HSD (Honestly Significant Difference) test for each ANOVA. It will allow us to identify which specific groups in our dataset are significantly different from each other after conducting our ANOVA test that showed a significant difference between the means of some or all of the groups.

ANCOVA

For the purpose of investigating the answer to our third research question, namely whether race had an effect on being perceived as more "*combative*" during the arrest.

We conducted an ANOVA study on *action_at_arrest_combative, Perceived_Race*, and s*trip searches*, with the dependent variable *action_at_arrest_combative* indicating whether combat actions were taken at the time of arrest, and the independent variable *Perceived_Race* while

StripSearch was used as a covariate in this study. We will use ANCOVA to analyze whether there are differences in the "Actions_at_arrest___Combative__" variable between racial groups based on "Perceived_Race" while controlling for the effect of "StripSearch" as a continuous covariate. However, the outcome variable needs to be continuous when conducting an ANOVA. Then. Consequently, with the purpose of creating a continuous variable for the dependent variable, we divided the total number of times arrestees were perceived as combative by year of arrest, perceived race, and age group. In summary, our ANCOVA was designed to explore whether perceived race had a significant effect on being perceived as more "combative" during the arrest while controlling for potential confounding effects of strip searches.

Logistic Regression

Until now, we have discussed statistical tests to answer our first and second research question. For the third research question; *Does race have an impact in being deemed as more“combative” during an arrest?*, a logistic regression model will be performed to test the relationship between multiple quantitative and/or categorical explanatory variables, including perceived race and our binary categorical outcome; combative at arrest. Logistic regression models the relationship between the predictor variables and the outcome using a logistic function (Stephenson, B., Cook, D., Dixon, P., Duckworth, W.& al., (2008)). This function is "S" shaped and ranges from 0 to 1, which is a good fit for binary outcomes, as it is the case for our outcome variable . Secondly, it is appropriate for this experiment because of its flexibility in handling both categorical and continuous predictor variables, making it well-suited for our dataset.

To perform the method, we have chosen the following features: A*rrest_Month, Perceived_Race, Sex,* and *age_group_at_arrest*. However, after consideration, we dropped Arrest_Month from the features as they were not of focus in our research and would be noise. We then performed a one-hot-encoding on the features because their data types were non-numerical. Additionally, to overcome the unbalancing of classes, we have used a stratified sampling when splitting our dataset using the Sklearn stratify function. Then, the dataset is separated into 80% training and 20% validation. This enables enough data points for our model while keeping unseen data for validation. The findings will be presented in the section below.

Results

This section presents the key findings of methods that examine the questions of race and violent arrests in the city of Toronto.

ANOVA

The results in Table 11 show an F-statistic value of 16.48 for One-Way ANOVA, which demonstrates a statistically significant difference in the mean number of strip searches. Furthermore, the P-value of 5.389566540799918ee-17 is less than 0.05, indicating strong evidence for the rejection of the null hypothesis. Overall, the results based on the one-way ANOVA showed significant differences in the number of strip searches based on perceived race.

Table 11: One-Way ANOVA results

ANOVA	F StatisticF	p–value	Reject /Accept
One-Way	16.48	6.91e-18	Reject

To assess which perceived races differed significantly in the mean number of strip searches, we tested using Tukey's HSD test. The results in Appendix 1 shows the Black group has a p-value of 0.001 against all the groups except White. This is lower than the significant threshold of 0.05. Therefore, the Tukey HSD test confirms that the Black group were significantly more likely to be strip searched on average than all other minority groups. Furthermore, we can find other groups that have a p-value of lower than 0.05 such as Indigenous compared to White.

Table 12: Two-Way ANOVA results

	Sum squared	DF	F	PR(>F)
C(Perceived_Race)	102570.07	7.0	19.37	5.27e-24
C(Age_group__at_arrest_)	48742.18	8.0	8.05	1.72e-10
C(Perceived_Race):C(Age_group__at_arrest_)	89384.06	56.0	2.11	7.96e-06
Residual	598168.62	791.0	NaN	NaN

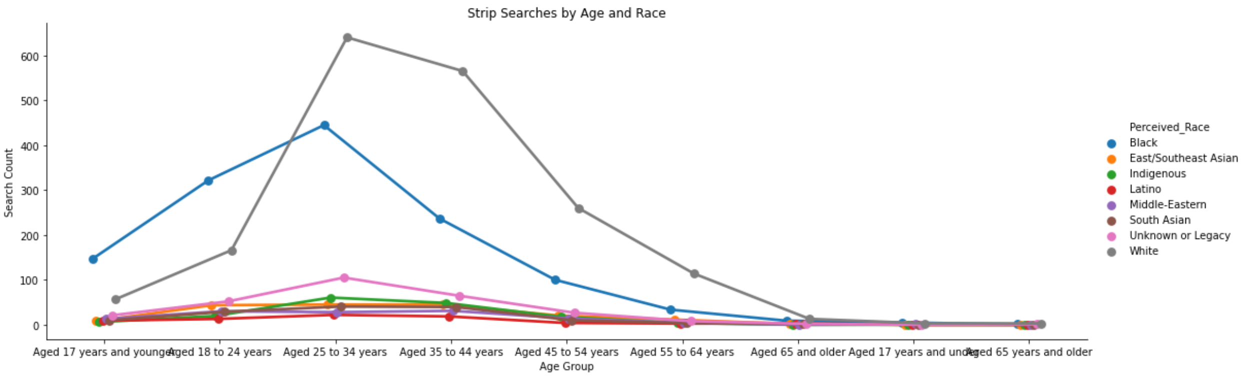
We conducted a two-way ANOVA to explore the effect of perceived race and age group on the likelihood of being strip-searched. The two-way ANOVA included two independent variables, "Perceived_Race" and "age group at the time of arrest," with the dependent variable being "strip search".

The p-values for Perceived Race and Age group at arrest is 5.27e-243 and 1.72e-10 respectively. This is less than 0.05 which implies that the means of both the factors possess a statistically significant effect on Strip searches. The p-value for the interaction effect is 7.96e-06. This is lower than 0.05 which depicts that there is a significant interaction effect between Perceived Race and Age group. We reject all three null hypotheses.

Two-Way ANOVA Post Hoc Test - Tukey HSD

The findings from the Post Hoc Test, with a sample appearing in Appendix II, display the Black - Aged 17 under, 18-24, 25-34, up to 65 years old, has a p-value of lower than 0.005 against the White group 17. This is lower than the significant threshold of 0.05. Therefore, we reject the null hypothesis. We find similar results for the following minority groups; Indigenous, Latino and Middle-Eastern.The Tukey HSD test confirms that the race and age group were significantly more likely to be strip searched on average.

Figure 9 : Interaction plot of the relationship between the number of strip searches conducted by age and race



The interaction plot above illustrates the relationship between the number of strip searches conducted by age and race. This interaction plot includes the number of strip searches plotted on the y-axis and the age group plotted on the x-axis and uses different colors to indicate the different perceived races of the searched individuals.

The interaction plot shows that for most age groups, the number of strip searches increases with age for arrestees younger than 35 years old, but for arrestees older than 35 years old, it just starts to decrease slowly with age.

In addition, the interaction plot shows that the perceived Black race is relatively more likely to be strip searched than other perceived races. Overall, we can use the information expressed in the plot to understand that perceived race and age group have an interactive effect on the rate of strip searches. This is because the probability of being strip searched varies for different perceived racial groups depending on their age.

In general, the p-values for the three effects of C(Perceived_Race), C(Age group at arrest), and C(Perceived_Race):C(Age_group__at_arrest_) were all less than the significance level (0.05), indicating that we can reject the null hypothesis. Therefore, we can conclude that there is a significant difference in the mean value of "StripSearch" in different levels of each factor and their combination.

ANCOVA

We conducted an analysis of covariance to examine the relationship between "Actions_at_arrest___Combative__" and "Perceived_Race" while controlling for the effect of "StripSearch". The results from Table 13 show that at 95% confidence intervals, the independent variable "Perceived_Race" (F= 5.16, P<0.05) and the covariate "StripSearch" (F= 560.76, P<0.05) had a statistically significant effect on the dependent variable "Actions_at_arrest___Combative __" both had a statistically significant effect. This suggests that both "Perceived_Race" and "StripSearch" were strong and significant predictors regarding "Actions_at_arrest___Combative__" after controlling for the effects of each other. In addition, the effect of "Strip Search" would be even greater than the effect on "Actions_at_arrest___Combative __"

Table 13: ANCOVA results

Source	SS	DF	F	p-unc	np2
Perceived_Race	29701.82	7	5.16	4.546413e-05	0.25
StripSearch	49906.32	1	60.79	5.345797e-12	0.37
Residual	84556.67	103	NaN	NaN	NaN

Logistic regression

We performed a logistic regression to examine the effects of race, sex and age on the likelihood that individuals arrested are deemed combative. The results from Table 14 indicate none of the features are statistically significant.

Table 14: Logistic Regression results

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.12	0.07	-41.98	0.00	-3.27	-2.98
race_encoded	-0.009	0.007	-1.34	0.17	-0.02	0.004
sex_encoded	-0.01	0.05	-0.19	0.84	-0.11	0.09
age_encoded	0.02	0.01	1.67	0.09	-0.004	0.05

Subsequently, we calculated the odds ratio in table 15. The odds ratio is approximately 0.99, 0.98 and 1.02 for race, sex and age. If the predictors were significant, for each additional unit increase, the odds that the individual would be deemed combative increases by about 0.99, 0.98 and 1.02 for times respectively.

Table 15: Logistic Regression Odds ratio

Group	Odds ratio

Race	0.99
Sex	.98
Age	1.02

After building the model, we assessed it with metrics and a confusion matrix. The findings in Table 16 and 17 indicate an accuracy of 95% , precision of 95%, recall of 100% and f1-score of 98%.

Table 16: Logistic Regression Model Metrics

	Accuracy	f1-Score	Precision	Recall
Result (%)	95	98	95	100

Table 17: Confusion Matrix

	Predicted not Combative	Predicted Combative	Total
Actual Not deemed Combative	12466	0	12466
Actual deemed Combative	590	0	590
Total	13056	0	

Discussion

ANOVA

Overall, the results of our statistical analysis reveals there are significant differences in strip searches based on demographic attributes of Race. The two way ANOVA also shows there is a significant difference in the means of strip search based on Race and Age as well as a combined effect on strip search. From the Tukey HSD test on our Two-Way ANOVA, we see the groups with significant differences in means are minority groups. Indeed, the results demonstrate that while "Carding" practices are discontinued, the Toronto Police still dispositionally target vulnerable communities. Further investigation should be done to identify the reasons behind this discriminatory approach. It is known that police tend to increase more efforts in certain racialized communities, which therefore increase the number of arrests (Giwa, S, 2018). We recommend the Toronto Police to reassess their arrests and strip searches practices based on the evidence brought by this paper that there exists some relationships between strip search and protected demographic attributes. Debiasing techniques such as intra-processing methods could be helpful in debiasing (Savani & al., 2020). The "Artificial Intelligence Fairness 360" toolkit could be implemented as well as it provide a good starting bias framework for organizations (Bellamy & al., 2019)

ANCOVA

The results of ANCOVA show that "Perceived_Race" has an effect on "*action_at_arrest_combative*", which is similar to the results of our previous study. This result may suggest that the police may implement different tactics when making arrests depending on the race of the arrestee, and so contribute to the reason why different races are belligerent when being arrested. However, controlling for "StripSearch" as a covariate, the effect of "strip search" is in turn somewhat larger than the effect of "Perceived_Race" (0.371153>0.259952) which indicate "perceived race" goes not cause the most significant factor of "action at arrest". This would suggest that strip searches are a significant predictor of combative action during arrest, possibly because the blow to the arrestee's heart from the strip search may be the likelihood of causing them to be combative at the time of the arrest. However, this ANCOVA only controls for "StripSearch" as a covariate to detect the effect between the variables "Perceived_Race" and "*action_at_arrest_combative*." Moreover, there are other variables in the data that may also affect "*action_at_arrest_combative*". Also, the scope of this ANCOVA on the data is limited to the Toronto Police Department, so there may not be a large enough volume of data to more accurately demonstrate the true cause of arrest at the time of arrest.

However, it is important to note that we can observe in the ANCOVA that the effect of "Perceived_Race" on "*action_at_arrest_combative*" will change when controlling for

“StripSearch” as a covariate. This effect indicates that we did not directly conclude the causality of “*action_at_arrest_combative*” in ANCOVA, but we can also conclude from the results that “Perceived_Race” and "strip search" have a significant effect on “*action_at_arrest_combative*”. In summary, we can use ANCOVA to investigate further the complex and mutually influential relationships among the variables that cause “*action_at_arrest_combative*”. These observed inter-influential relationships between variables are then used to reduce violence and bias in future arrests as much as possible.

Logistic Regression

The initial exploratory data analysis had shown a trend of black individuals deemed as combative, violent or spitter/biter. For example, in figure 7, we see the perceived black race as being the second race with the most violent arrests. This first step led us to go deeper in our analysis since we could not yet make causal claims on solely t-tests. This is why we needed a model to make further inference. However, the logistic regression results showed that race was not a factor of significance. This is against our initial hypothesis that race was in fact a factor in police officers deeming individuals this character. Indeed, we believed it could also be an excuse for officers to then proceed to strip searches and bookings because of this behavior. While individual factors may influence the officers perception, other circumstances may lead to such situations. For example, one potential explanation for why race may not be the cause of combative behavior at the time of arrest is that other factors, such as socioeconomic status or mental health, may play a more significant role. Research has shown that individuals from disadvantaged backgrounds, such as those living in poverty or experiencing homelessness, are more likely to exhibit combative behavior at the time of arrest (Motley Jr, R. O., & Joe, S. (2018)). Similarly, individuals with mental health conditions or substance use disorders may be more likely to resist arrest or act combatively.. Another possible explanation for why race may not be the cause of combative behavior at the time of arrest is that police officers may be more likely to use force or be aggressive with certain types of individuals, regardless of their race. Indeed, the implicit bias, which is the unconscious tendency to associate certain characteristics or behaviors with particular groups of people may happen for other causes than just race. For example, police officers may be more likely to perceive an individual as combative if they are wearing certain types of clothing, such as baggy pants or a hoodie, rather than if they are wearing business attire. ditions or substance use disorders may be more likely to resist arrest or act combatively.

In addition, the time of day could be a factor. Probably arrests at nights or weekends where there is a higher likelihood of individuals being intoxicated may lead to offensive arrests. In conclusion, while the dataset demonstrates some initial prejudice based on race, it is not conclusive and there exists potential explanations for why race may not be the cause of combative behavior at the time of arrest. We therefore draw causal effects about the relationship between race and combative behavior at the time of arrest.

Additionally, even if the prector was significant, the initial statistical power analysis revealed we needed considerably more samples to obtain a strong statistical power (power 80%). Therefore, in order to mitigate this limitation, we may compile datasets from previous years, up until we have available. This will allow us to more samples and potentially discover new data patterns.

Furthermore, the odds ratio did not provide any information since the factors were not significant. However, we could see the greater odds for age. This makes sense as probably younger individuals may have more energy to be combative relative to older arrestees.

Finally, the high accuracy and f1 score of our model, at 95 and 98% respectively indicate a robust and reliable model. While accuracy is a straightforward metric that measures the proportion of correct predictions made by our Logistic Regression, f1 score is a more nuanced metric that takes into account both precision and recall. We realize that our scenarios where the data is imbalanced, with more data with the negative label (not combative, violent). Therefore, a f1 score is more appropriate as it takes into account both the number of true positives and the number of false negatives.

Confusion matrix

The confusion matrix results in Table 15 for the logistic regression with the outcome variable Actions_at_arrest___Combative_ suggest our model performed well. The confusion matrix shows that there were 12466 true negatives, indicating that the model correctly identified a large number of individuals as not being combative during their arrest. Additionally, there were no false negatives, suggesting that the model did not incorrectly classify any individuals as not being combative when they actually were.

On the other hand, there were 590 true positives, indicating that the model correctly identified a smaller number of individuals as being combative during their arrest. However, there were no false positives, suggesting that the model did not incorrectly classify any individuals as being combative when they actually were not.

Both precision and recall are important metrics to consider. A high precision score indicates our model is accurately identifying individuals who are actually combative during arrest, and not labeling non-combative individuals as combative. A high recall score would indicate our model is accurately identifying as many individuals who are actually combative during arrest as possible, and not missing any of them. Given the research question of whether race has an impact on an individual being deemed as more "combative" during an arrest, it may be more important to prioritize recall over precision. This is because missing cases where an individual is actually combative during arrest could lead to biased results and inaccurate conclusions

Conclusion

In this study, we used three different statistical methods to analyze these data: ANCOVA, logistic regression, and power analysis to derive the factors that perceived race would cause arrestees to become belligerent during the arrest.

The ANCOVA results showed that the effect of “Perceived_Race” on “*action_at_arrest_combative*” changed when controlling for "StripSearch" as a covariate, signifying that strip search may be a significant predictor of combative action at arrest. In addition, although the initial exploratory data analysis showed a tendency for Blacks to be perceived as combative, logistic regression results showed that race was not a significant factor in police perceptions of arrestees as combative at the time of the arrest.

The statistical analyses we designed for this purpose provide valuable insight into the complex relationships among variables that influence strip search and combative behaviour at arrest. However, this only suggests a correlation between Perceived_Race and belligerence at arrest and cannot determine whether it is causal.

Our study's results may also highlight the potential impact of individual characteristics on arrest procedures, which could inform future research and policy development. Further research could explore additional variables that play a more important role in combativeness at the time of arrest, such as gender or socioeconomic status, to provide a more comprehensive understanding of the factors that influence arrest procedures.

As further possible action, we could ask the Toronto Police Department to collect more granular dataset with for example the police officer ID could help us detect if the prejudices of strip searches are done by the same police workers. Additionally, hidden bias training could be a good use of public funds, as a way to create self-awareness of officers and potentially reduce the biases in targeting ethnic minority groups.

References

Barabas, C., Doyle, C., Rubinovitz, J. B., & Dinakar, K. (2020, January). Studying up: reorienting the study of algorithmic fairness around issues of power. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 167-176).

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4-1.

Giwa, S. (2018). Community policing in racialized communities: A potential role for police social work. Journal of Human Behavior in the Social Environment, 28(6), 710-730.

Grewcock, M., & Sentas, V. (2021). Strip searches, police power and the infliction of harm: An analysis of the New South Wales strip search regime. International journal for crime, justice and social democracy, 10(3), 191-206.

Joh, E. E. (2017). Feeding the machine: Policing, crime data, & algorithms. Wm. & Mary Bill Rts. J., 26, 287.

Johnson, V. E., Nadal, K. L., Sissoko, D. G., & King, R. (2021). "It's not in your head": Gaslighting,'Splaining, victim blaming, and other harmful reactions to microaggressions. Perspectives on psychological science, 16(5), 1024-1036.

Motley Jr, R. O., & Joe, S. (2018). Police use of force by ethnicity, sex, and socioeconomic class. Journal of the Society for Social Work and Research, 9(1), 49-67.

Savani, Y., White, C., & Govindarajulu, N. S. (2020). Intra-processing methods for debiasing neural networks. Advances in Neural Information Processing Systems, 33, 2798-2810.

Seltman, H. J. (2012). Experimental design and analysis.

Spurrier, M. (2023). The police are perpetrating harm, so why is the government giving them more power?. bmj, 380.

Stephenson, B., Cook, D., Dixon, P., Duckworth, W., Kaiser, M., Koehler, K., & Meeker, W. (2008). Binary response and logistic regression analysis. available at:< a href=" http://www. stat. wisc. edu/mchung/teaching/MIA/reading/GLM. logistic. Rpackage. pdf"> http://www. stat. wisc. edu/mchung/teaching/MIA/reading/GLM. logistic. Rpackage. pdf</a>(last access: 30 August 2014).

Stop-and-Frisk Data". New York Civil Liberties Union. January 2, 2012. Retrieved November 30, 2019

Appendix I: Sample of One-Way ANOVA Post Hoc Test - Tukey HSD

Multiple Comparisons of Means - Tukey HSD, FWER = 0.05						
group1	group2	meandiff	p-adj	lower	upper	reject
Black	East/Southeast Asian	-19.0826	0	-31.0518	-7.1135	TRUE
Black	Indigenous	-19.0048	0.0001	-31.3904	-6.6193	TRUE
Black	Latino	-20.8203	0	-33.1082	-8.5325	TRUE
Black	Middle-Eastern	-19.9964	0	-32.1034	-7.8895	TRUE
Black	South Asian	-19.7027	0	-31.8386	-7.5669	TRUE
Black	Unknown	-17.3839	0.0003	-29.3268	-5.441	TRUE
Black	White	8.6141	0.3492	-3.2524	20.4806	FALSE
East/Southeast Asian	Indigenous	0.0778	1	-12.2555	12.4111	FALSE
East/Southeast Asian	Latino	-1.7377	0.9999	-13.973	10.4975	FALSE
East/Southeast Asian	Middle-Eastern	-0.9138	1	-12.9673	11.1397	FALSE
East/Southeast Asian	South Asian	-0.6201	1	-12.7027	11.4624	FALSE
East/Southeast Asian	Unknown	1.6987	0.9999	-10.19	13.5875	FALSE
East/Southeast Asian	White	27.6967	0	15.8848	39.5087	TRUE

Appendix II: Sample of Two-Way ANOVA Post Hoc Test - Tukey HSD

Multiple Comparisons of Means - Tukey HSD, FWER = 0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Black Aged 17 years and under	Black Aged 17 Years and younger	3	1	-174.736	180.7356	FALSE
Black Aged 17 years and under	Black Aged 18 To 24 Years	17	1	-136.924	170.9235	FALSE
Black Aged 17 years and under	Black Aged 25 To 34 Years	12	1	-141.924	165.9235	FALSE
Black Aged 17 years and under	Black Aged 35 To 44 Years	14	1	-139.924	167.9235	FALSE
Black Aged 17 years and under	Black Aged 45 To 54 Years	10.5	1	-143.424	164.4235	FALSE
Black Aged 17 years and under	Black Aged 55 To 64 Years	5	1	-148.924	158.9235	FALSE
Black Aged 17 years and under	Black Aged65andolder	1	1	-176.736	178.7356	FALSE