TPS Arrest Counts and Arrest Outcomes by Age Group, Sex and Perceived Race

INF2178 Winter 2023

Final Submission

Paul King, Zijian Zhang

Colab link:
https://colab.research.google.com/drive/1iupBYhLnrOpqEKH2tRTl2kh_5o1oJpRS?usp=sharing

Introduction

Toronto is the most populous city in Canada, and the Toronto Police Service is the third-largest police force in Canada, making up 7.4% of Toronto's operating budget in 2022 (Jones, 2022). TPS has been criticized for discriminatory behaviour (Omstead, 2023), as well as violent incidents and deaths that disproportionately impact racialized people. In 2013, 18-year old Sammy Yatim was shot 9 times by Constable James Forcillo, then tased (Kari, 2015). In 2020, 29-year-old Regis Korchinski-Paquet, who was a Black and Indigenous woman, fell to her death in the presence of 6 TPS officers, who were never criminally charged (Hayes, 2020). Korchinski-Paquet's death is concerning in that we see that her race and gender intersect in how they impacted her interaction with police, a pattern that is known as Misogynoir (Thompson, 2018) and was identified in the 1995 Report of the Commission on Systemic Racism in the Ontario Criminal Justice System. Melchers claimed that differences in how groups are treated by the police is not proof of discriminatory practices in and of itself (2003). However, more recently, Mensah et al. concluded that "the criminal justice system [in Canada] allocates substantial amounts of time and energy to exercising discretion in service of selective targeting that produces unjustifiable racial disparities with respect to arrests, charges, the distribution of negative credentials, and so on, to Blacks, Indigenous people, and other racialized populations." (2021) An independent report also concluded that White, Black and Indigenous populations were over-represented in strip searches when compared to the proportion of arrests (Foster and Jacobs, 2022). TPS continues to receive large budget increases yearly in spite of criticism, lack of investment in social and community infrastructure, and calls to defund the police and introduce alternative non-police responses (Omstead, 2023 and Jones, 2022). Given TPS's disproportionate budget and widespread accusations of discrimination, there is a need to examine arrest data to look for patterns of discriminatory behaviour.

Based on our dataset and the results of our exploratory data analysis, we sought to answer these four research questions:

> **RQ 1.** Is there a difference in mean arrest counts between perceived race categories?
> **RQ 2.** Is there a difference in stripsearch probabilities between age groups?
> **RQ 3.** Is there a difference in booking probabilities between perceived race categories and sex?
> **RQ 4**. Can we predict whether an arrest event will lead to a stripsearch based on age group, sex, and perceived race?

We used T-tests, Interaction Plots, One- and Two-Way ANOVA, ANCOVA, and Tukey's HSD tests to show that there are indeed patterns of discriminatory behaviour in the TPS in terms of arrest outcomes, with differences being observed in all three outcome variables, disproportionately impacting Black, Indigenous, and White people. However, year, month, and actions on arrest were far more significant in determining the outcome of a particular arrest event.

Dataset Description

To investigate our research questions, we used the Arrests and Strip Searches dataset, available freely from the Toronto Police Service Public Safety Data Portal (2022). Each observation represents an arrest of a person by a police officer, and whether or not the person was stripsearched or "booked" (brought to the police station) in the arrest, along with their perceived race, age group, and sex. Because each arrest lists a unique ID for the person being arrested, we were able to calculate the number of arrest events for a particular individual, as well as the probability of a particular individual being stripsearched or booked. There are limitations and ambiguities to the dataset. The difference between arrests and bookings is not defined, besides that an arrest is not necessarily a booking. Sex is left as an ambiguous term, since we do not know if it refers to legal gender, biological characteristics, or sex as perceived by the arresting officer. It's unclear if "U" sex refers to nonbinary individuals; given the lack of definition, and that there are only 9 observations with "U" sex, they have been removed from our analysis. Similarly, perceived race is dependent on the arresting officer, and can vary from arrest to arrest for the same individual. Because of this, when determining the race of a particular individual, we used the most common perceived race as reported by arresting officers.

Exploratory Data Analysis

Plots

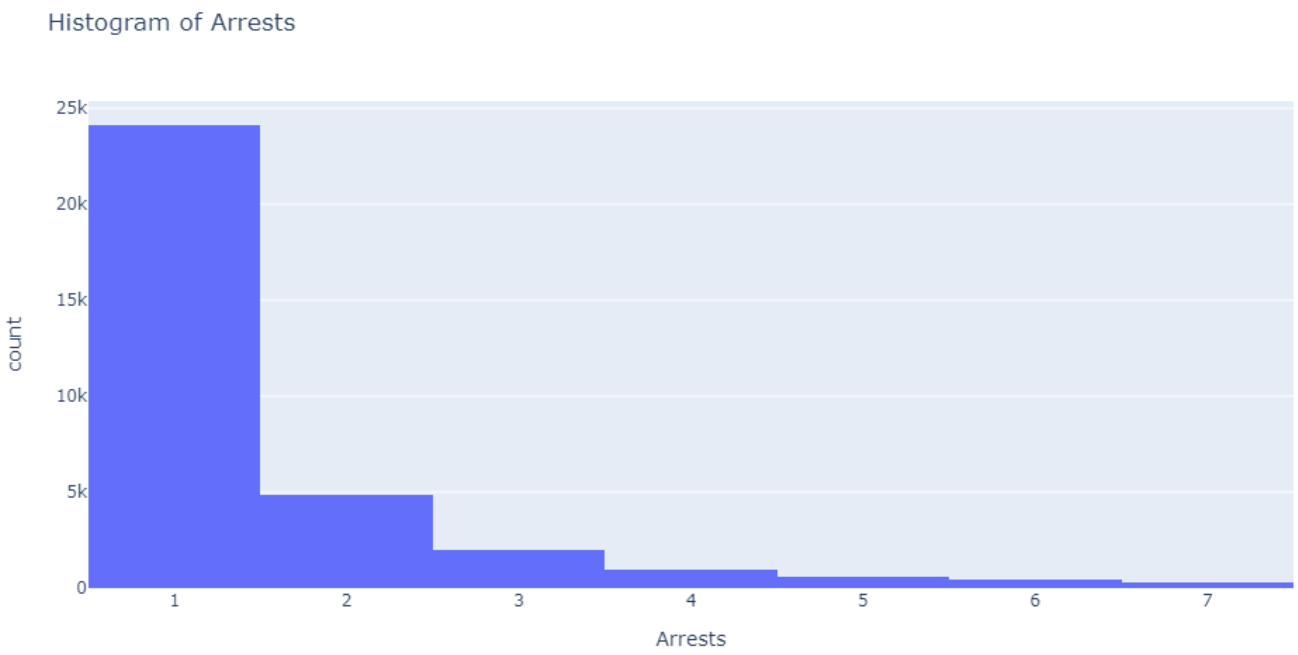

Fig1. Histogram of Arrests

In a histogram of arrest counts, we found that the data skewed heavily to the left and that our data is not normally distributed.

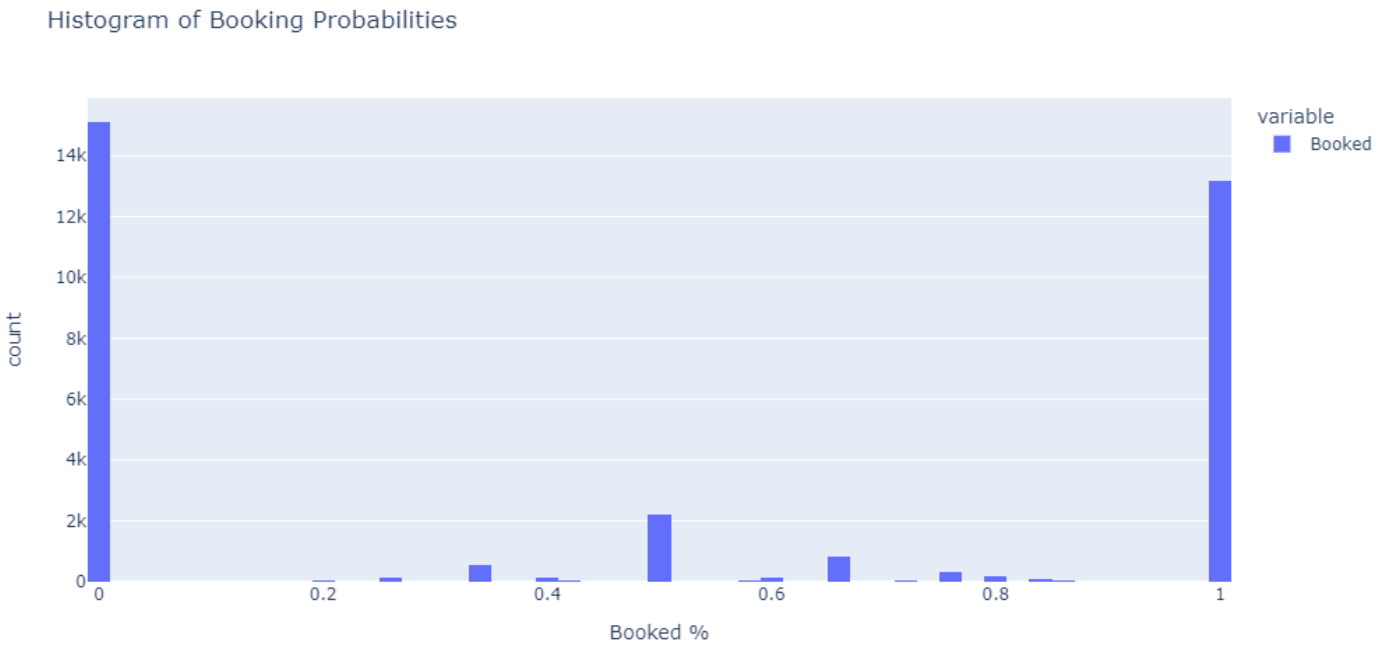

Fig2. Histogram of Booking Probabilities

Booking probabilities were similarly not normal, with most people arrested being consistently booked or not booked.

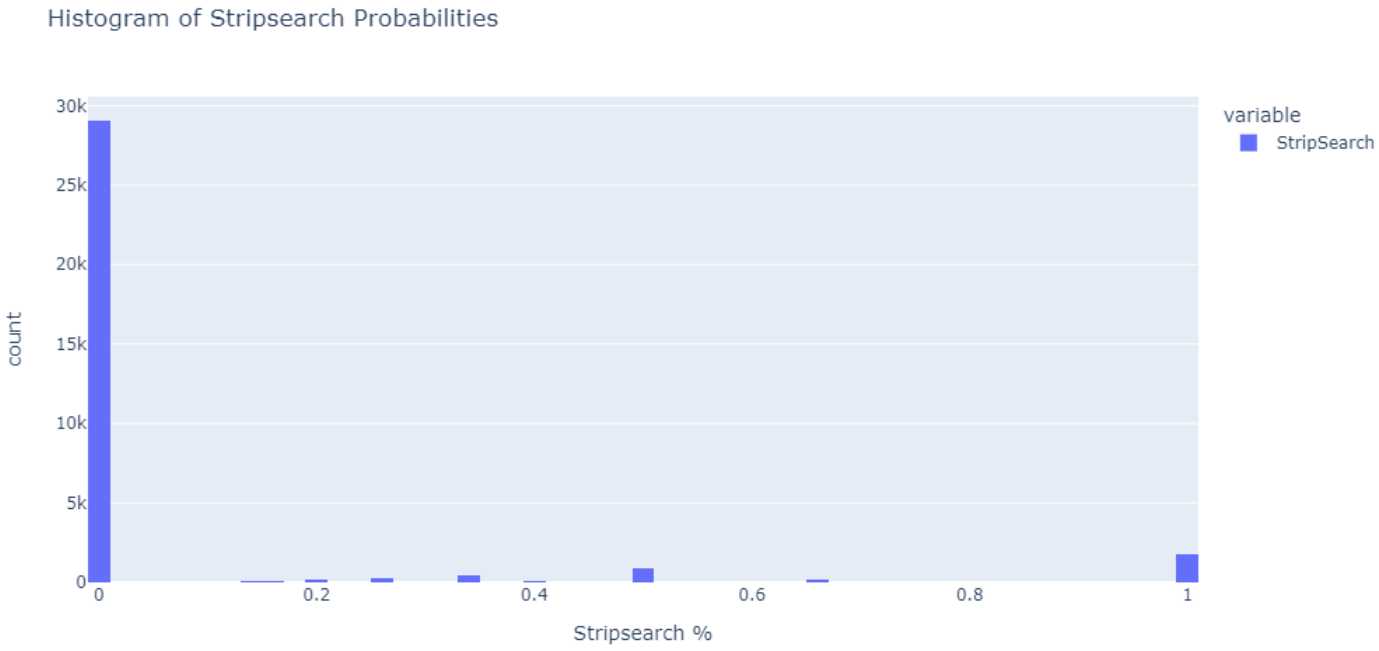

Fig3. Histogram of Stripsearch Probabilities

In a histogram of stripsearch probabilities, we see that the majority of arresting events do not result in a stripsearch.

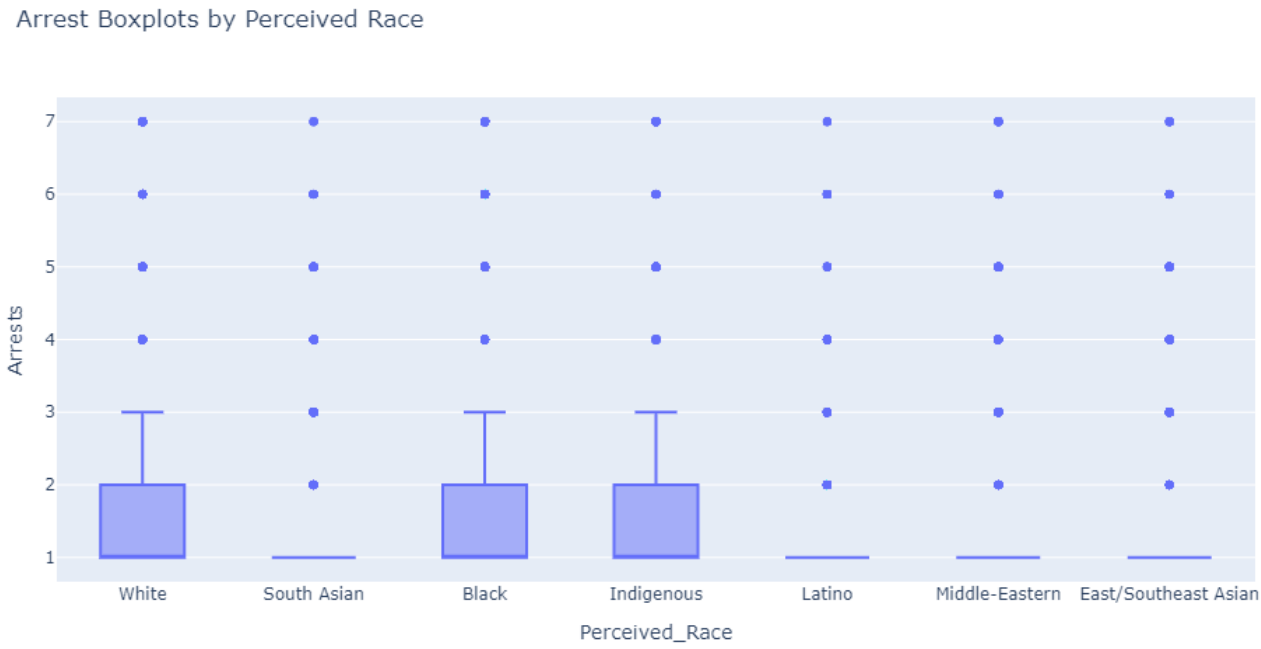

Fig4. Boxplot of Arrests by Perceived Race

We see that mean arrest counts are consistently higher for White, Black and Indigenous people.

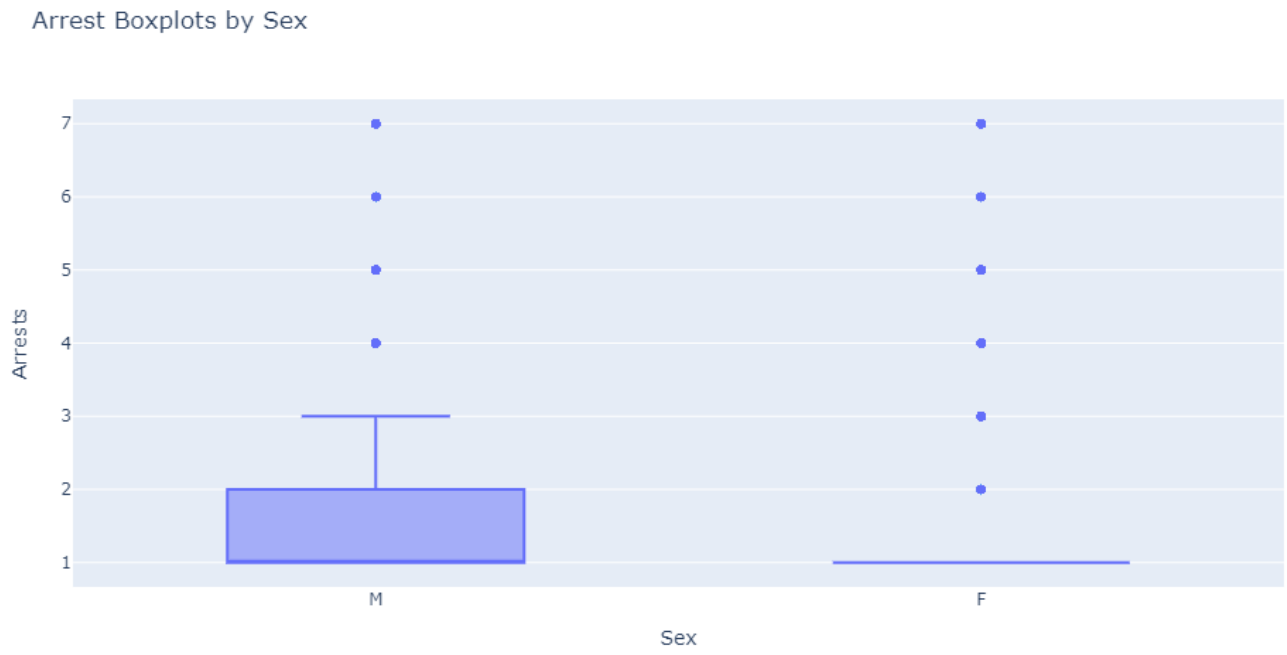

Fig5. Boxplot of Arrests by Sex

Arrest counts are also higher for men than women.

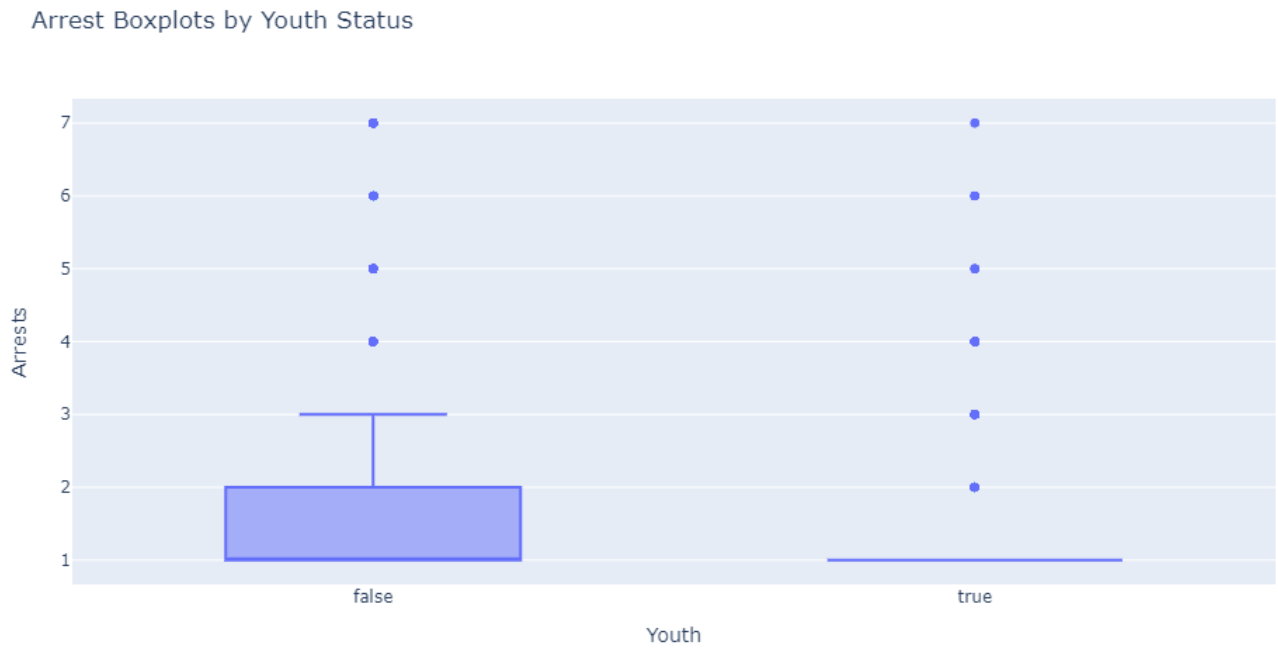

Fig6. Boxplot of Arrests by Youth Status

Non-youths are more likely to have higher arrest counts than non-youths.

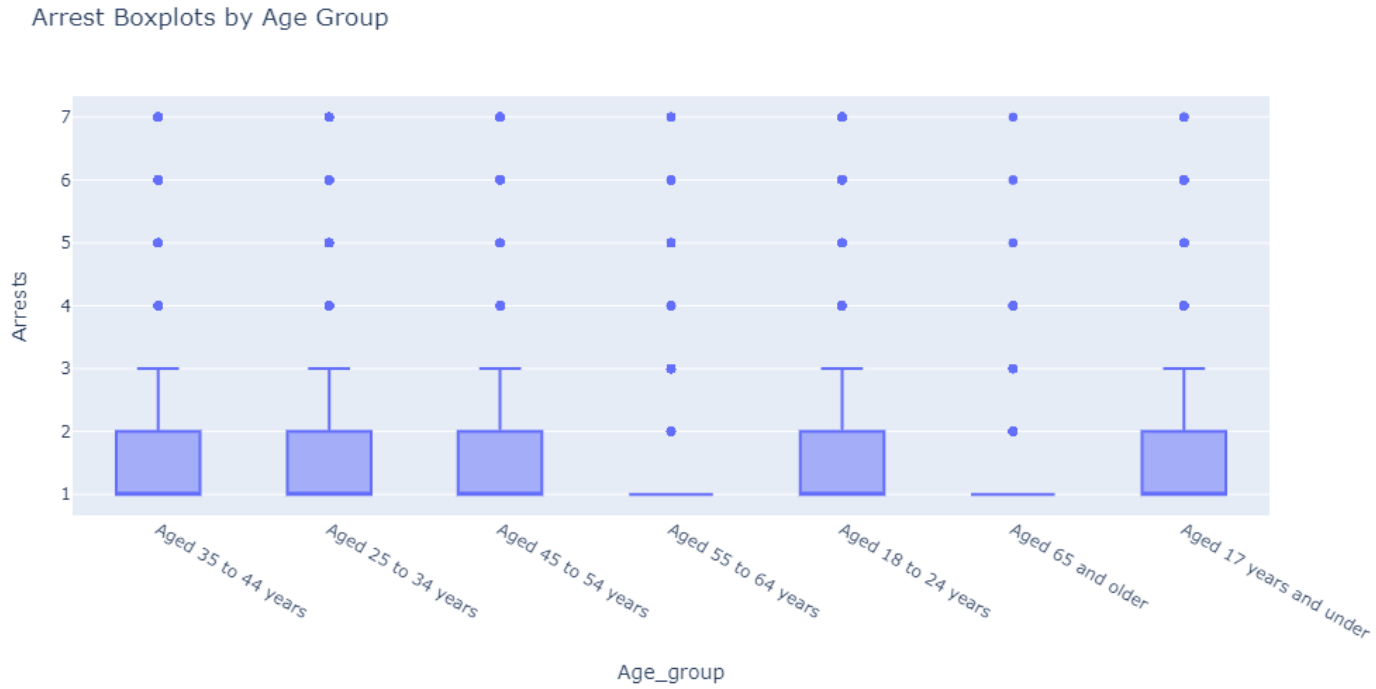

Fig7. Boxplot of Arrests by Age Group

By contrast, this boxplot shows that arrest counts are actually lower for ages 55 and up.

Power analysis

We used power analysis to calculate the sample size needed for each sample group to achieve a statistical power of 0.8. A power level of 0.8 is a commonly used threshold for statistical power in many fields of research, as it balances the risk of Type I and Type II errors while still providing a reasonable chance of detecting a true effect. For sex as an explanatory variable, we calculated its effect size for each outcome variable, using Cohen's D. We then computed the the required sample size using the obtained effect size and establishing the statistical power at 80%. Our data's actual sample size is 26293 for males and 6885 for females, which is larger than any of the required sample sizes we calculated. We also graphed our effect sizes against small, medium and large effect sizes for different numbers of observations.

For booking probability, we found that sex had an effect size of 0.28. Our required sample size is 124 males and 472 females.

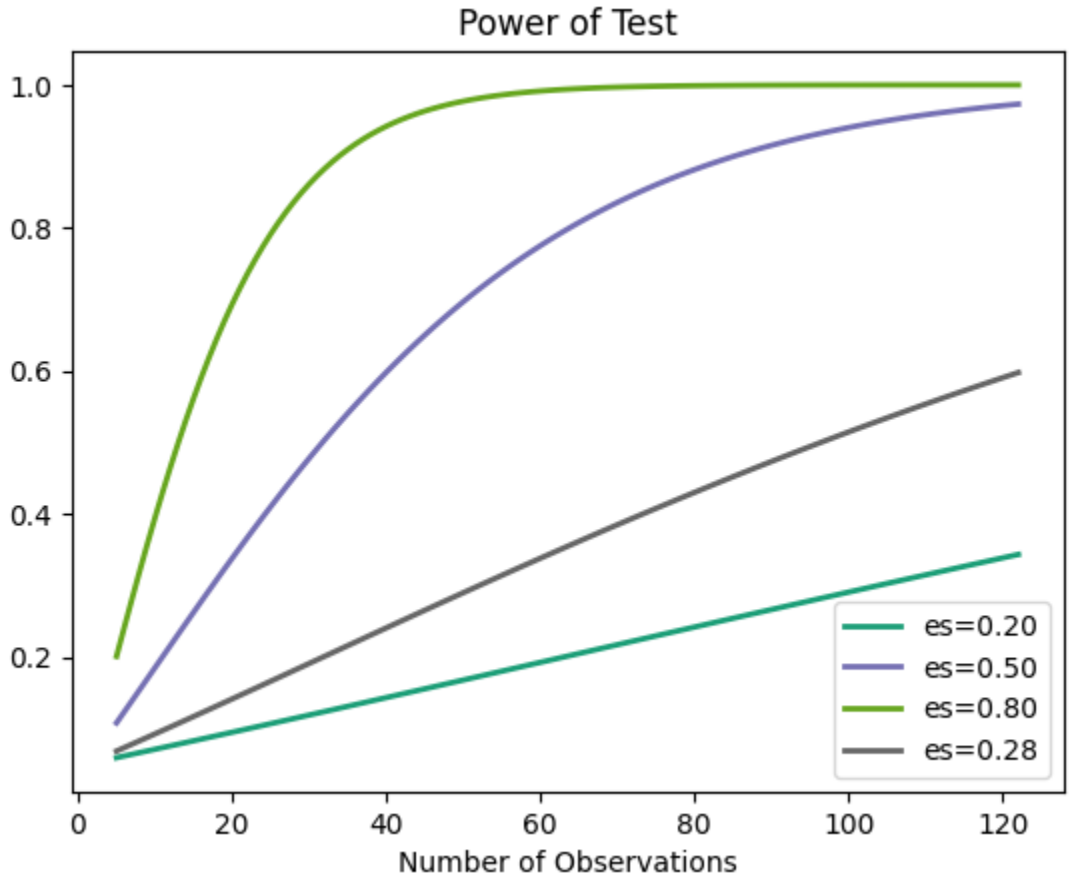

Fig8. Power of Tests for booking probability vs sex

We conclude that our test's effect size for booking probability is between medium and small.

For stripsearch probability, we found that sex had an effect size of 0.08. Our required sample size is 1607 males and 6137 females.

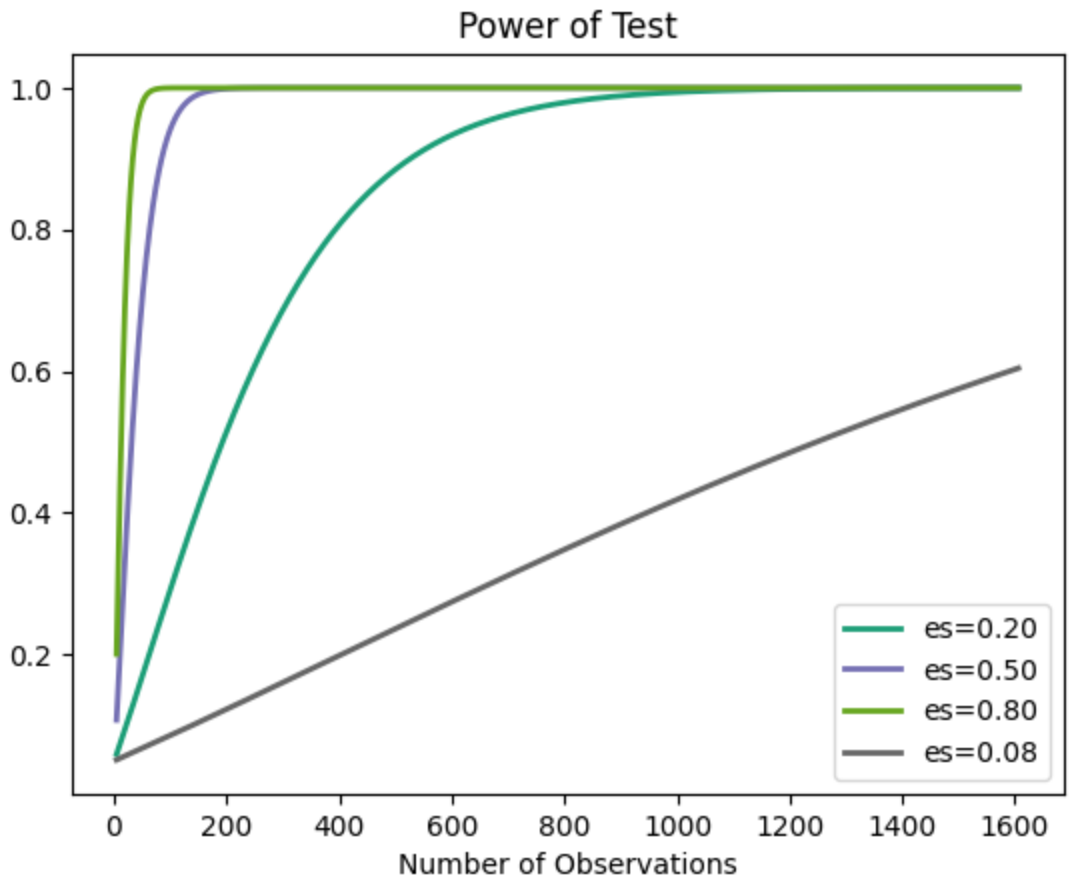

Fig9. Power of test for stripsearch probability vs sex

We conclude that our test's effect size for stripsearch probability is small.

For arrest count, we found that sex had an effect size of 0.10. Our required sample size is 928 males and 3540 females.

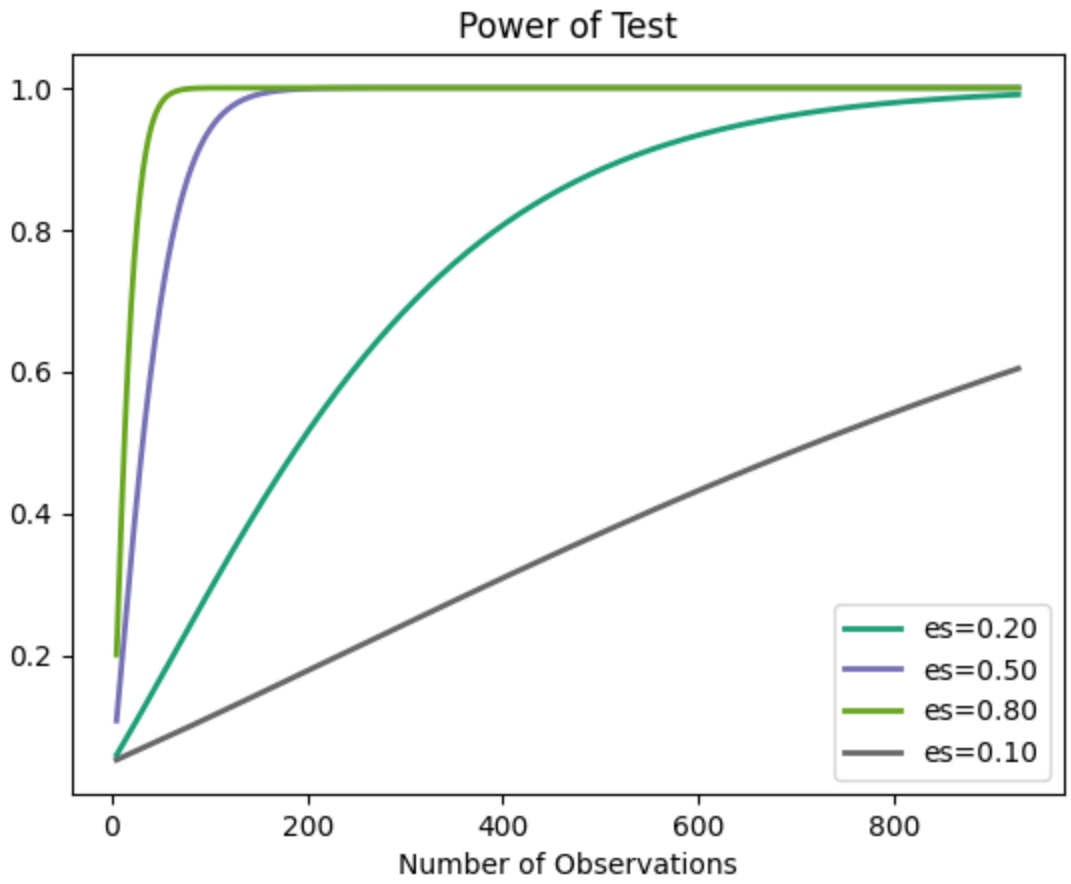

Fig10. Power of Tests for arrest time vs sex

We conclude that our test's effect size for arrest counts is small.

T-tests

For stripsearch probability, we conducted T-tests between groups with the following hypotheses:

> Null hypothesis: There is no difference between group means of stripsearch probability.
> Alternative hypothesis: There is a difference between group means of stripsearch probability.

Group 1	Group 2	T-statistic	P-value
Youth	**Not a youth**	**-26.504**	**1.6E-124**
South Asian	**Black**	**-11.47**	**4.32E-30**
White	Black	-6.111	1.01E-09
South Asian	Indigenous	-5.160	3.12E-07
White	Indigenous	-2.203	0.0279
Indigenous	Latino	5.140	3.32E-07
Indigenous	Middle-Eastern	5.252	1.93E-07
Indigenous	East/Southeast Asian	5.253	1.95E-07
Male	Female	6.161	7.45E-10
White	Latino	6.258	5.16E-10
White	Middle-Eastern	7.673	2.23E-14
White	South Asian	7.728	1.39E-14
White	East/Southeast Asian	8.545	1.65E-17
Black	**Latino**	**9.295**	**4.45E-20**
Black	**Middle-Eastern**	**11.290**	**4E-29**
Black	**East/Southeast Asian**	**12.456**	**2.98E-35**

Table1. Statistically significant T-test comparisons for Mean Stripsearch Probability

For stripsearch probability, the largest T-statistics were for youth status, followed by Black people compared to South Asian, East/Southeast Asian, Middle-Eastern and Latino people.

For booking probability, we conducted T-tests between groups with the following hypotheses:

Null hypothesis: There is no difference between group means of booking probability.
Alternative hypothesis: There is a difference between group means of booking probability.

Group 1	Group 2	T-statistic	P-value
Youth	**Not a youth**	**-9.344**	**8.36E-20**
White	**Black**	**-8.825**	**1.18E-18**
South Asian	Black	-5.845	5.52E-09
South Asian	Latino	-3.312	9.41E-04
White	Latino	-3.247	1.20E-03
South Asian	Indigenous	-2.403	1.64E-02
White	Indigenous	-2.173	3.02E-02
South Asian	East/Southeast Asian	2.204	2.76E-02
Indigenous	Middle-Eastern	2.553	1.08E-02
Latino	Middle-Eastern	3.455	5.61E-04
Indigenous	East/Southeast Asian	3.913	9.80E-05
White	East/Southeast Asian	4.008	6.23E-05
Latino	East/Southeast Asian	5.151	2.86E-07
Black	Middle-Eastern	5.835	5.98E-09
Black	**East/Southeast Asian**	**9.392**	**8.63E-21**
Male	**Female**	**21.235**	**4.28E-98**

Table2. Statistically significant T-test comparisons for Mean Booking Probability

For booking probability, the largest T-statistic was obtained for Males compared to Females, followed by Black people compared to East/Southeast Asian and White people, and then youth status.

For arrest count, we conducted T-tests between groups with the following hypotheses:

> Null hypothesis: There is no difference between group means of numbers of arrests.
> Alternative hypothesis: There is a difference between group means of numbers of arrests.

Group 1	Group 2	T-statistic	P-value
South Asian	**Black**	**-11.86**	**5.21E-32**
South Asian	**Indigenous**	**-10.85**	**1.99E-25**
Black	Indigenous	-7.375	5.03E-13
White	Indigenous	-6.755	3.20E-11
Youth	Not a youth	-4.215	2.76E-05
South Asian	Middle-Eastern	-2.815	4.90E-03
South Asian	Latino	-2.425	1.54E-02
Latino	East/Southeast Asian	2.604	9.30E-03
White	Black	2.739	6.16E-03
Middle-Eastern	East/Southeast Asian	3.075	2.12E-03
Black	Latino	4.910	1.02E-06
White	Latino	6.290	4.27E-10
Black	Middle-Eastern	6.678	2.83E-11
Male	Female	8.027	1.10E-15
White	Middle-Eastern	8.559	1.78E-17
Indigenous	Latino	9.001	1.36E-18
Indigenous	Middle-Eastern	9.494	2.87E-20
Indigenous	East/Southeast Asian	10.972	6.69E-26
Black	**East/Southeast Asian**	**12.914**	**1.03E-37**
White	**South Asian**	**14.335**	**1.33E-45**
White	**East/Southeast Asian**	**15.638**	**4.16E-54**

Table3. Statistically significant T-test comparisons for Mean Arrest Count

For arrest counts, we found the largest T-statistic for White people compared to South Asian and East/Southeast Asian people, followed by Black people compared with East/Southeast Asian people, followed by South Asian people compared with Black and Indigenous people.

For perceived race, we found widespread differences in means in all three outcomes. Similarly, for sex, all outcomes differed between males and females, and for youth status, all outcomes differed between youths and non-youths.

Correlation Graph

We created a graph of all features and their correlations

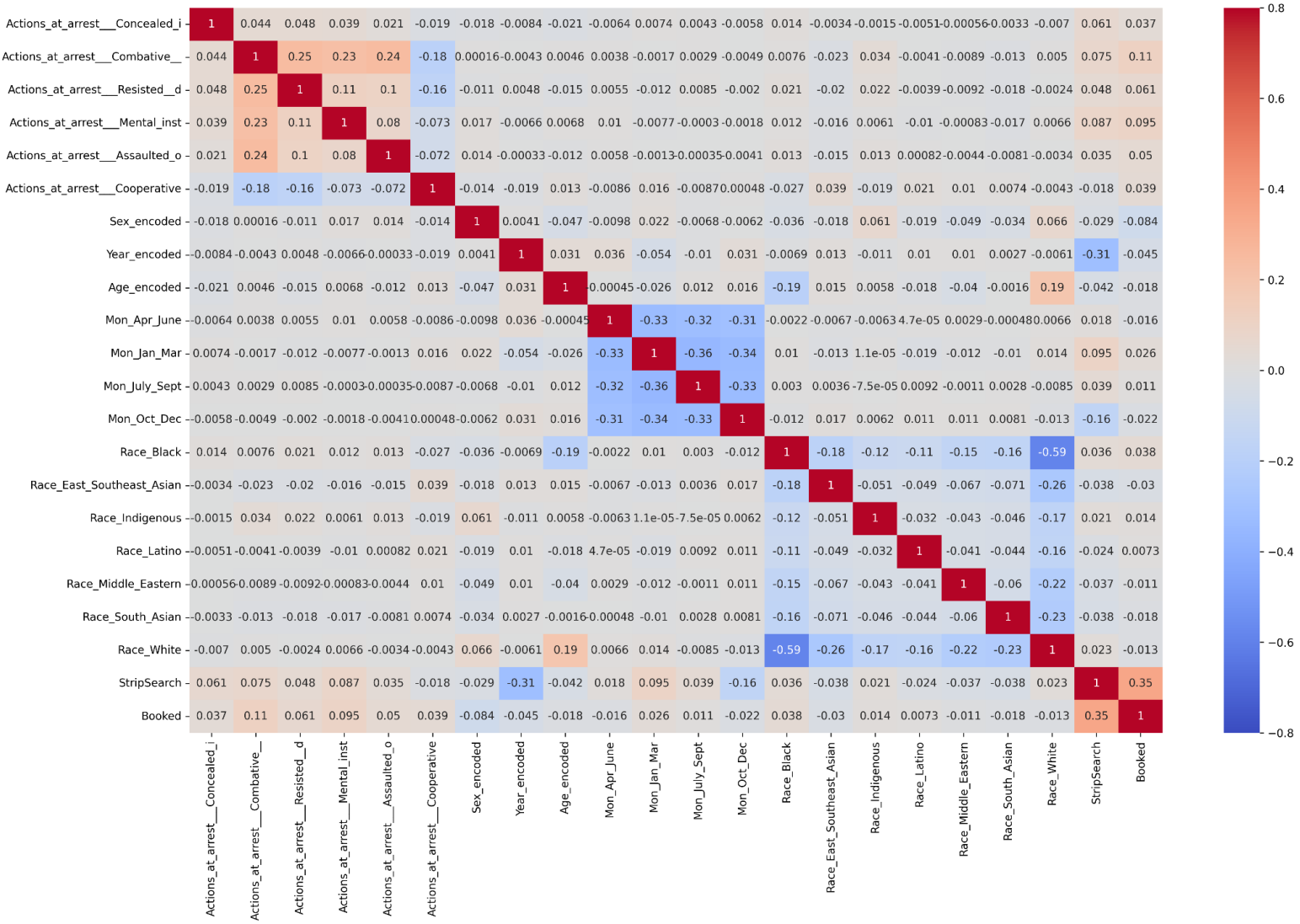

Fig11. Correlation graph of all features

From the graph, we can see that the year is the most correlated feature with strip search, and that combative actions are the most correlated feature with booking, suggesting that timing and actions at arrest are strong confounding factors when predicting the result of an arrest event.

Methods

We produced Interaction Plots for all three research questions, comparing perceived race and sex or perceived race and youth status for all three outcomes. We used one-way ANOVA for research questions 1 and 2, and two-way ANOVA for research question 3 to find differences in mean outcomes for categories. For research questions 1 and 2, after rejecting the null hypothesis with the results of our ANOVA, we used Tukey's HSD test to look for statistically significant differences between particular categories. We used ANCOVA in research questions 2 and 3 to verify our results with number of arrests as a covariate. For research question 4, we used logistic regression to model the probability of a stripsearch during an arrest event and identify which features of an arrest event were most important, and compared accuracy and area under curve for logistic regression with all features and samples, logistic regression with undersampling, and logistic regression with undersampling and less important features removed.

Results

Research Question 1

Is there a difference in mean arrests between racial categories?

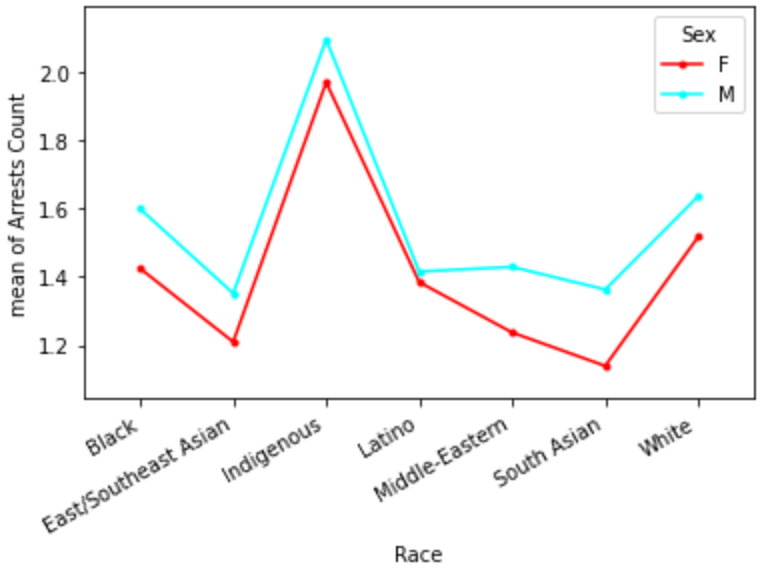

Fig12. Interaction Plot for mean Arrest Count by Perceived Race and Sex

This interaction plot shows similar outcomes for men and women for all perceived races.

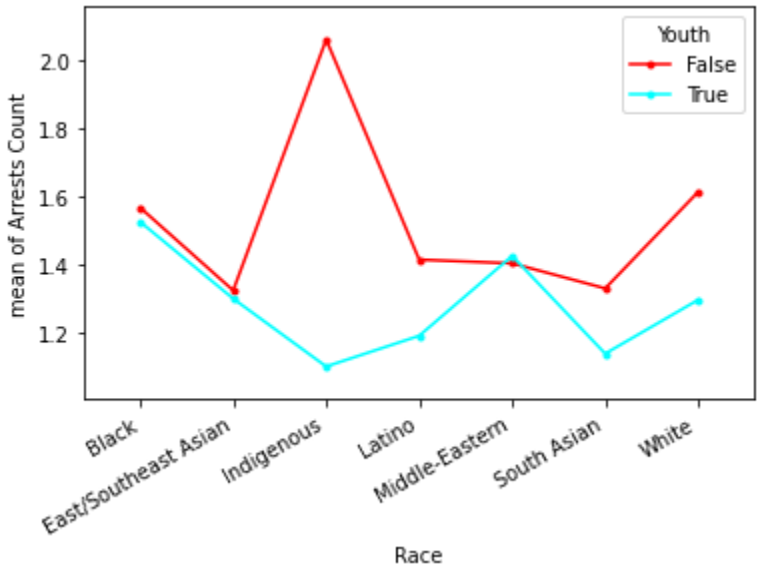

Fig13. Interaction Plot for mean Arrest Count by Perceived Race and Youth Status
This interaction plot shows that arrest numbers are significantly higher for Indigenous and White non-youths.

We performed a one-way ANOVA between perceived race groups with the following hypotheses:

Null hypothesis: There is no difference in means of arrest counts among perceived race groups
Alternative hypothesis: There is at least one perceived race group that differs significantly from the overall mean of arrest counts.

	Sum of Squares	df	F	P-value
Perceived Race	540.527	6.0	72.573	2.683e-90
Residual	41176.501	33171.0		

Table4. ANOVA results for difference in mean arrest counts by perceived race

With a significance value of 2.683e-90 (< 0.05), the ANOVA test suggests that we can reject the null hypothesis and that means there is a difference in arrest counts between perceived race groups.

Group 1	Group 2	Mean Difference	P-value	Lower Bound	Upper Bound
East/Southeast Asian	**Indigenous**	**0.724**	**0.001**	**0.578**	**0.869**
Black	Indigenous	0.481	0.001	0.344	0.618
East/Southeast Asian	White	0.285	0.001	0.219	0.350
South Asian	White	0.280	0.001	0.208	0.352
Middle-Eastern	White	0.204	0.001	0.126	0.282
Latino	White	0.198	0.001	0.0949	0.301
Black	Latino	-0.156	0.001	-0.261	-0.0516
Black	Middle-Eastern	-0.162	0.001	-0.242	-0.0822
Black	South Asian	-0.239	0.001	-0.313	-0.164
Black	East/Southeast Asian	-0.243	0.001	-0.311	-0.175
Indigenous	White	-0.439	0.001	-0.575	-0.304
Indigenous	**Latino**	**-0.637**	**0.001**	**-0.803**	**-0.471**
Indigenous	**Middle-Eastern**	**-0.643**	**0.001**	**-0.794**	**-0.492**
Indigenous	**South Asian**	**-0.720**	**0.001**	**-0.868**	**-0.571**

Table5. Tukey's HSD statistically significant results for difference in mean arrest counts by perceived race

We can see that the means of fourteen pairs of groups are significantly different from each other. Our Tukey's HSD test showed the biggest significant differences with a mean difference of 0.72 arrests between East/Southeast Asian people and Indigenous people, followed by Indigenous people compared to South Asian, Middle-Eastern and Latino people.

Research Question 2

Is there a difference in stripsearch probabilities between age groups?

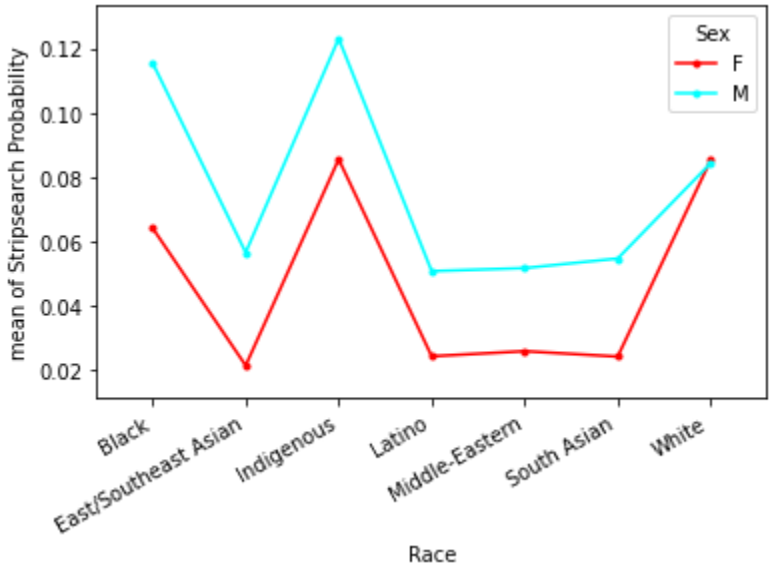

Fig14. Interaction Plot for mean Stripsearch Probability by Perceived Race and Sex

This interaction plot shows greater probabilities of stripsearch for males of all perceived races besides White males.

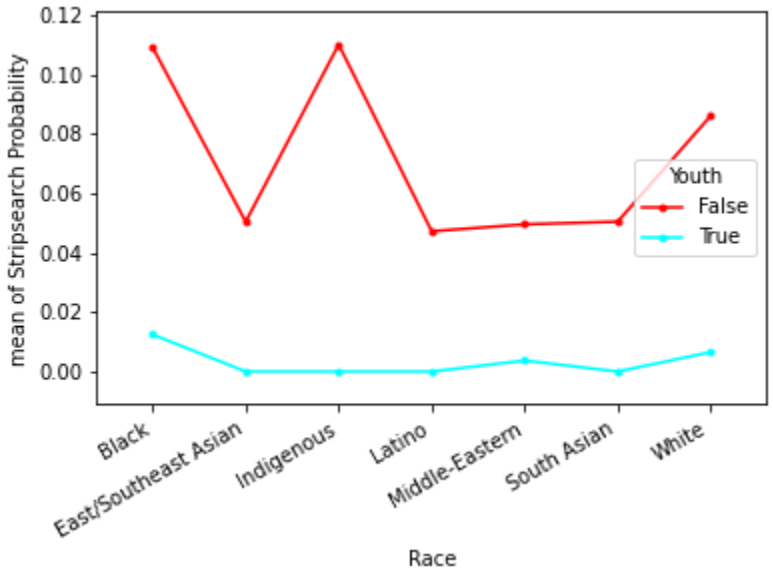

Fig15. Interaction Plot for mean Stripsearch Probability by Perceived Race and Youth Status

This interaction plot shows greater probabilities of stripsearch for Black, Indigenous, and White non-youths, mirroring the independent report by Foster and Jacobs (2020).

We performed a one-way ANOVA between age groups with the following hypotheses:

> Null hypothesis: There is no difference in stripsearch probabilities between age groups
> Alternative hypothesis: There is at least one age group that differs significantly from the overall mean of stripsearch probabilities.

	Sum of Squares	df	F	P-value
Age Group	9.227	6.0	25.947	5.783e-31
Residual	1965.999	33171.0		

Table6. ANOVA results for difference in mean stripsearch probability by age group

With a significance value of 5.783e-31 (< 0.05), the ANOVA test suggests that we can reject the null hypothesis and that means there is a difference in stripsearch probabilities among the perceived race groups.

	Sum of Squares	df	F	Uncorrected P-value	Partial eta-squared
Age Group	8.227	6	23.553	6.09e-28	0.00424
Arrests	34.854	1	598.657	4.77e-131	0.0177
Residual	1931.143	33170			

Table7. ANCOVA results for difference in mean stripsearch probability by age group while controlling for number of arrests

With a significance value of 6.09e-28 (< 0.05), our ANCOVA supports this result, rejecting the null hypothesis that there is no difference in stripsearch probabilities among perceived race groups, even after accounting for number of arrests.

Group 1	Group 2	Mean Difference	P-value	Lower Bound	Upper Bound
Aged 18 to 24 years	**Aged 65 and older**	**-0.0774**	**0.001**	**-0.103**	**-0.0523**
Aged 25 to 34 years	**Aged 65 and older**	**-0.0609**	**0.001**	**-0.0852**	**-0.0366**
Aged 35 to 44 years	**Aged 65 and older**	**-0.0554**	**0.001**	**-0.0801**	**-0.0307**
Aged 18 to 24 years	Aged 55 to 64 years	-0.0476	0.001	-0.0645	-0.0308
Aged 45 to 54 years	Aged 65 and older	-0.0432	0.001	-0.0687	-0.0177
Aged 17 years and under	Aged 65 and older	-0.0346	0.0063	-0.0632	-0.0061
Aged 18 to 24 years	Aged 45 to 54 years	-0.0343	0.001	-0.0484	-0.0201
Aged 25 to 34 years	Aged 55 to 64 years	-0.0311	0.001	-0.0468	-0.0154
Aged 55 to 64 years	Aged 65 and older	-0.0298	0.0203	-0.0569	-0.0027
Aged 35 to 44 years	Aged 55 to 64 years	-0.0256	0.001	-0.0419	-0.0093
Aged 18 to 24 years	Aged 35 to 44 years	-0.022	0.001	-0.0347	-0.0094
Aged 25 to 34 years	Aged 45 to 54 years	-0.0177	0.001	-0.0305	-0.005
Aged 18 to 24 years	Aged 25 to 34 years	-0.0165	0.001	-0.0284	-0.0046

Aged 17 years and under	Aged 35 to 44 years	0.0208	0.0165	0.0022	0.0393
Aged 17 years and under	Aged 25 to 34 years	0.0263	0.001	0.0082	0.0443
Aged 17 years and under	Aged 18 to 24 years	0.0428	0.001	0.0238	0.0619

Table8. Tukey's HSD statistically significant results for difference in mean stripsearch probability by age group

We can see that the means of sixteen pairs of groups are significantly different from each other. We found significant differences in arrest probability between people aged 65 and older and people aged 18 to 24 years, 25 to 34 years and 35 to 44 years of age, with older people being less likely to be stripsearched.

Research Question 3

Is there a difference in booking probabilities between perceived race categories and sex?

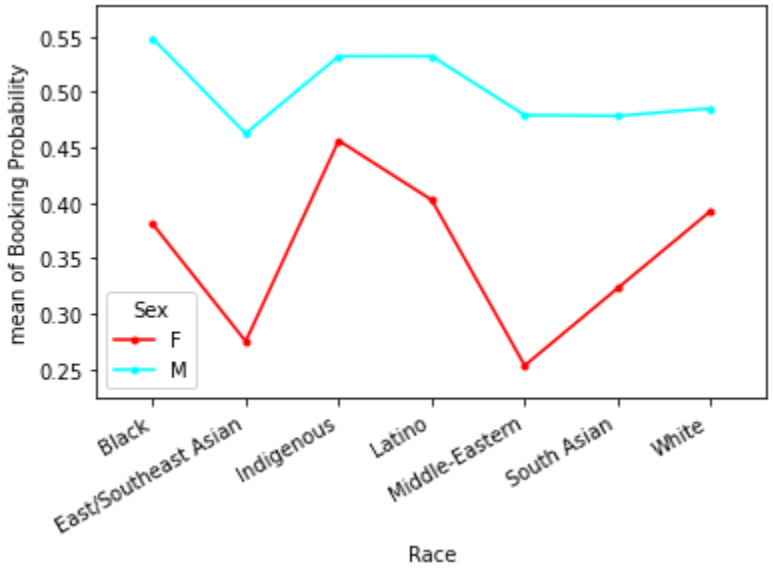

Fig16. Interaction Plot for mean Booking Probability by Perceived Race and Sex

This interaction plot shows lower probability of booking for East/Southeast Asian and Middle-Eastern women.

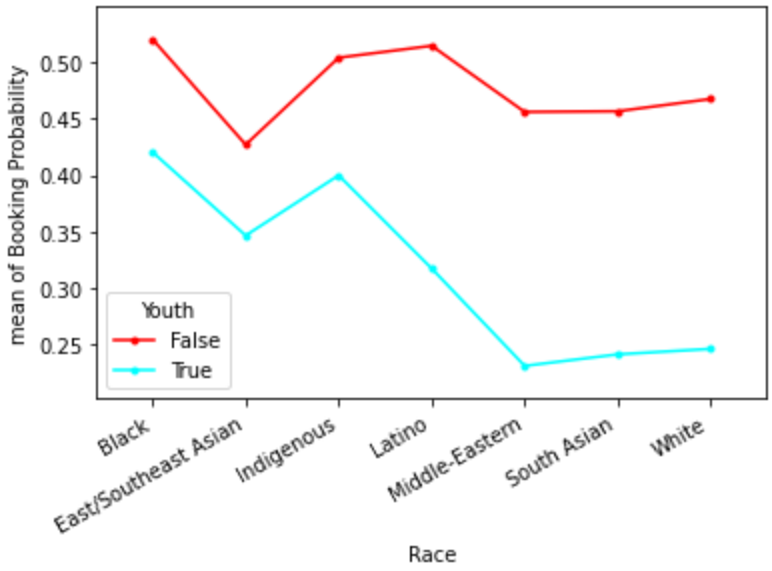

Fig17. Interaction Plot for mean Booking Probability by Perceived Race and Youth Status

This interaction plot shows lower probability of booking for Middle-Eastern, South Asian, and White youths.

We performed a two-way ANOVA between perceived race and sex groups with the following hypotheses:

Null hypotheses:

- There is no booking probability difference in group means at any level of the sex variable.
- There is no booking probability difference in group means at any level of the perceived race variable.
- The effect of sex variable does not depend on the effect of the perceived race variable (ie. there is no interaction effect).
- There is at least one age group that differs significantly from the overall mean of the stripsearch probabilities.

Alternative hypotheses:

- There is a difference in booking probability between the two sex groups.
- There is a difference in booking probability in at least one perceived race group.
- There is an interaction effect between perceived race and sex.

	Sum of Squares	df	F	P-value
Sex	119.0425	2.0	241.470	3.288e-105
Perceived Race	14.740	7.0	8.543	6.900e-07
Sex and Perceived Race	21.871	14.0	6.338	8.373e-10
Residual	16084.531	65253.0		

Table9. ANOVA results for difference in mean booking probability by perceived race and sex

With a significance value of 3.288e-105(< 0.05), the two way-ANOVA test suggests that we can reject the first null hypothesis and that means there is a booking probability difference between sex groups.
With a significance value of 6.900e-07(< 0.05), the two way-ANOVA test suggests that we can reject the second null hypothesis and that means there is a booking probability difference between perceived race groups.
With a significance value of 8.373e-10(< 0.05), the two way-ANOVA test suggests that we can reject the third null hypothesis and that means there is an interaction effect between sex groups and perceived race groups.

	Sum of Squares	df	F	Uncorrected P-value	Partial eta-squared
Perceived Race	26.625	6	20.940	1.19e-24	0.00377
Arrests	100.247	1	473.053	3.72e-104	0.0141
Residual	7029.242	33170			

Table10. ANCOVA results for difference in mean booking probability by perceived race group while controlling for number of arrests

With a significance value of 1.19e-24 (< 0.05), our ANCOVA supports our results, rejecting the null hypothesis that there is no difference in booking probabilities among perceived race groups, even after accounting for number of arrests.

	Sum of Squares	df	F	Uncorrected P-value	Partial eta-squared
Age Group	85.632	1	407.570	4.32e-90	0.0121
Arrests	96.294	1	458.314	5.41e-101	0.0136
Residual	6970.234	33175			

Table11. ANCOVA results for difference in mean booking probability by age group while controlling for number of arrests

With a significance value of 4.32e-90 (< 0.05), our ANCOVA supports our results, rejecting the null hypothesis that there is no difference in booking probabilities among age groups, even after accounting for number of arrests.

Research Question 4

Can we predict whether an arrest will lead to a stripsearch based on age group, sex, and perceived race?

	Accuracy	Area under curve
Logistic regression	88%	0.53
Logistic regression with undersampling	78%	0.78
Logistic regression with undersampling and unimportant features removed	78%	0.78
Logistic regression with undersampling and age, sex and perceived race as features	57%	0.58

Table12. Accuracy and Area Under Curve results for logistic regressions

Logistic regression with all samples resulted in an 88% accuracy but only 0.53 AUC, owing to most events being predicted as not resulting in a stripsearch.

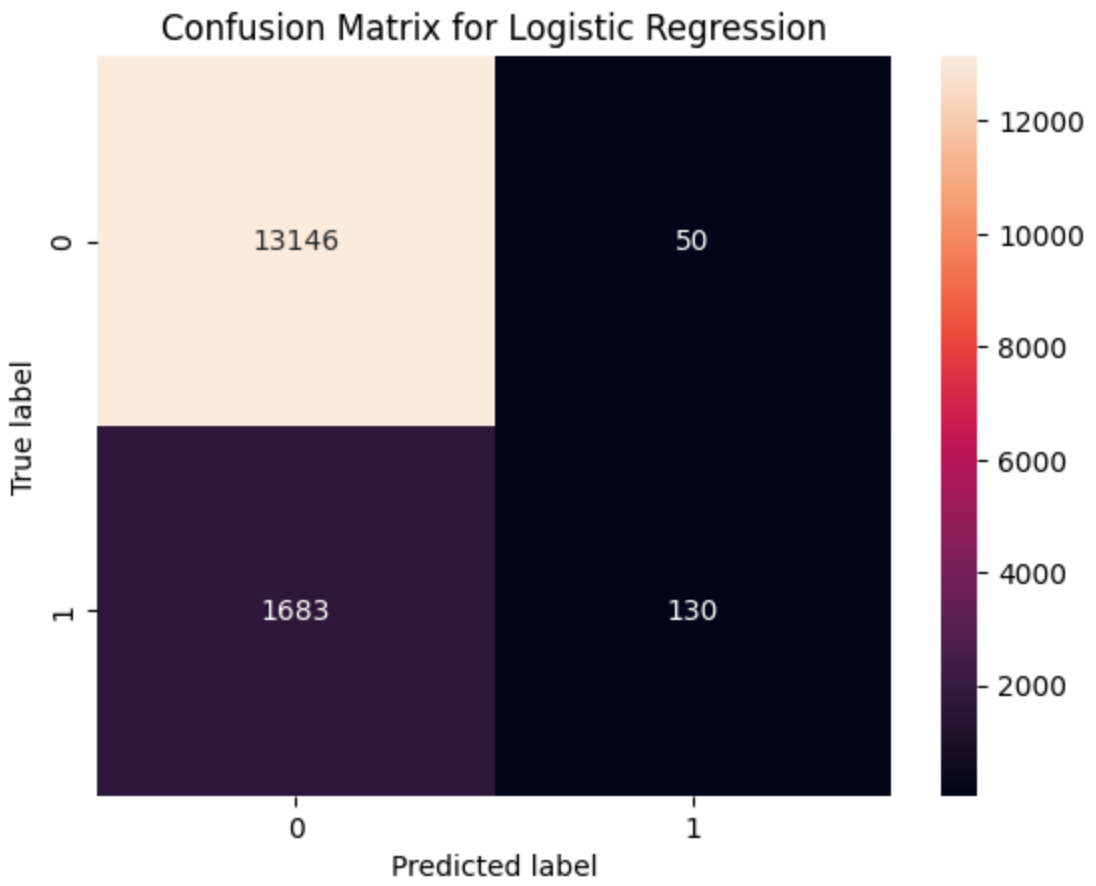

Fig18. Logistic regression confusion matrix with all features and no undersampling

	Odds Ratio	Lower Confidence Interval	Upper Confidence Interval
Concealed Items	5.51	3.86	7.84
Combative	1.71	1.49	1.97
Resisted arrest	1.45	1.25	1.68
Mental instability	2.73	2.37	3.15
Assaulted officer	1.46	1.04	2.04
Cooperative	0.98	0.92	1.05
Sex	0.72	0.66	0.78
Year	0.07	0.07	0.08
Age	0.91	0.89	0.93
April to June	5.24	4.62	5.96
January to March	6.54	5.78	7.39
July to September	5.54	4.89	6.27
Black	1.86	1.60	2.15
Indigenous	2.26	1.83	2.78
Latino	1.08	0.84	1.39
Middle Eastern	0.83	0.67	1.03

South Asian	0.94	0.77	1.16
White	1.84	1.59	2.13

Table13. Logistic regression odds ratios and confidence intervals with all features and no undersampling

Our dataset is imbalanced, with only 12% of arrest events resulting in stripsearches. We implemented undersampling, obtaining a 78% accuracy but an improved 0.78 AUC.

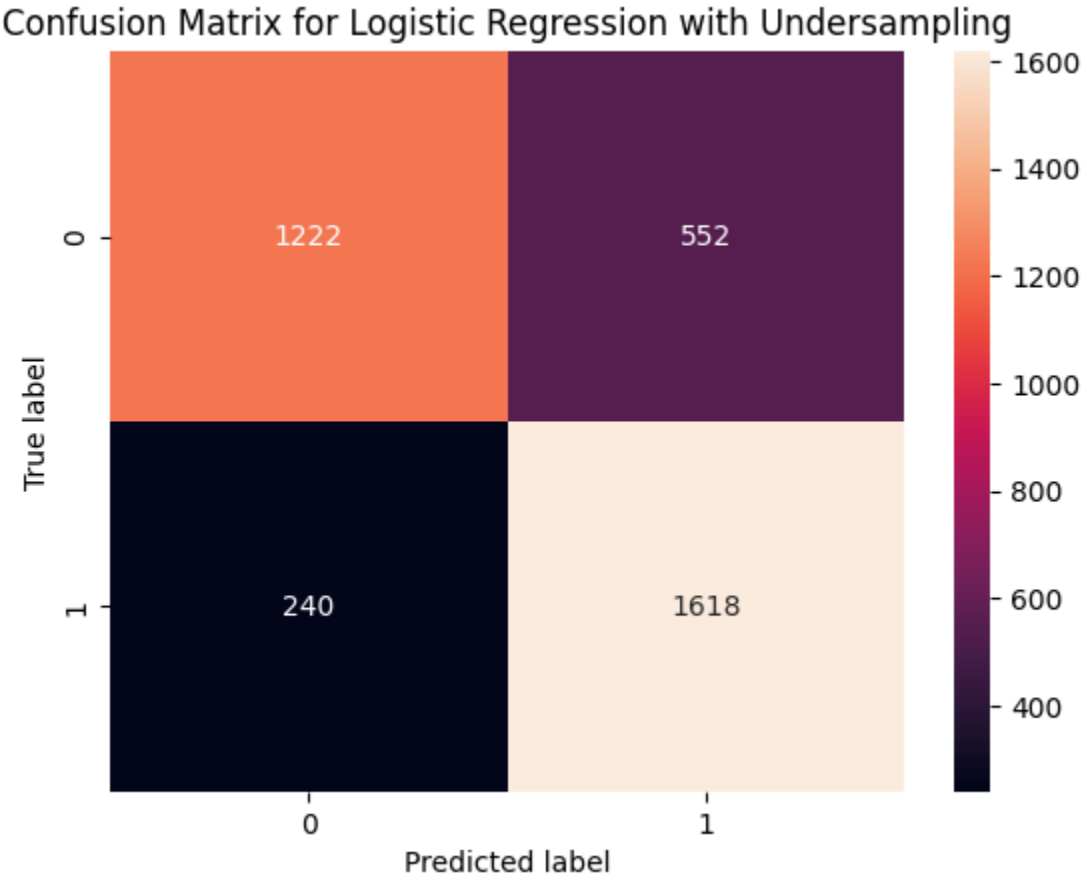

Fig19. Logistic regression confusion matrix with all features using undersampling

	Odds Ratio	Lower Confidence Interval	Upper Confidence Interval
Concealed Items	14.74	5.96	36.43
Combative	1.86	1.48	2.34
Resisted arrest	1.47	1.17	1.86
Mental instability	2.40	1.90	3.03
Assaulted officer	1.78	1.02	3.11
Cooperative	1.05	0.95	1.15
Sex	0.73	0.64	0.82
Year	0.07	0.06	0.08
Age	0.89	0.86	0.92
April to June	4.94	4.20	5.80
January to March	5.94	5.09	6.93
July to September	4.57	3.91	5.34
Black	1.41	1.15	1.74
Indigenous	1.80	1.33	2.46
Latino	0.72	0.51	1.01
Middle Eastern	0.66	0.50	0.89

South Asian	0.68	0.52	0.91
White	1.54	1.26	1.88

Table14. Logistic regression odds ratios and confidence intervals with all features and using undersampling

Using the results of the last logistic regression, we extracted the coefficients of features to determine which features the model considered most important.

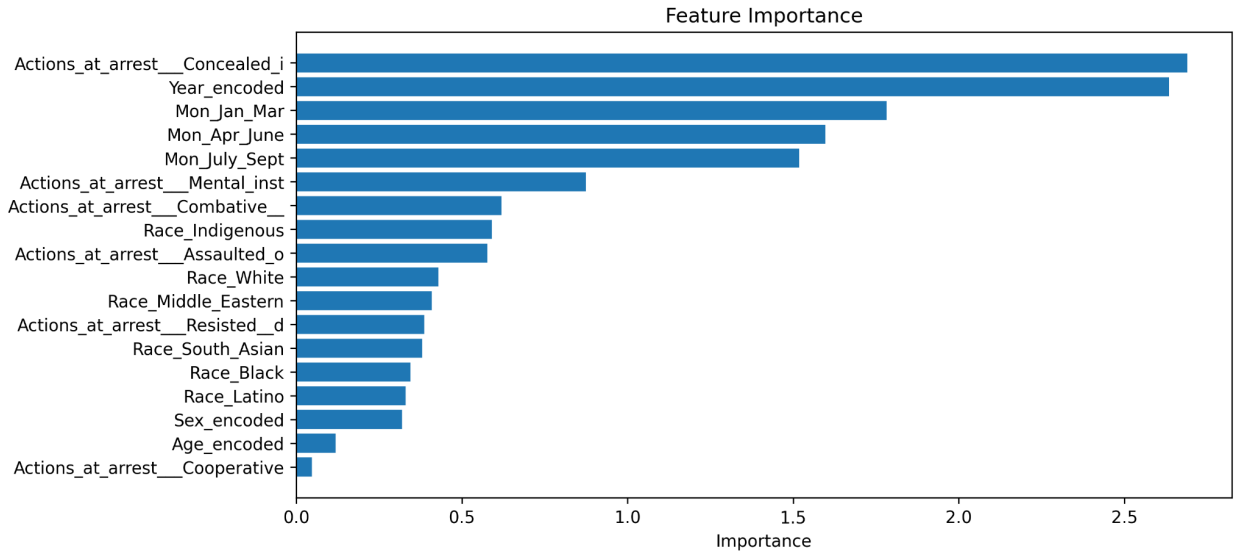

Fig20. Feature importance sorted from most to least important

The logistic regression model weighted actions at arrest, year and month as most important. To confirm this, we ran the logistic regression again, removing the features of age, sex, and perceived race. This resulted in a nearly identical score of 78% accuracy and 0.78 AUC, indicating that date and year were significantly more relevant to predicting a stripsearch.

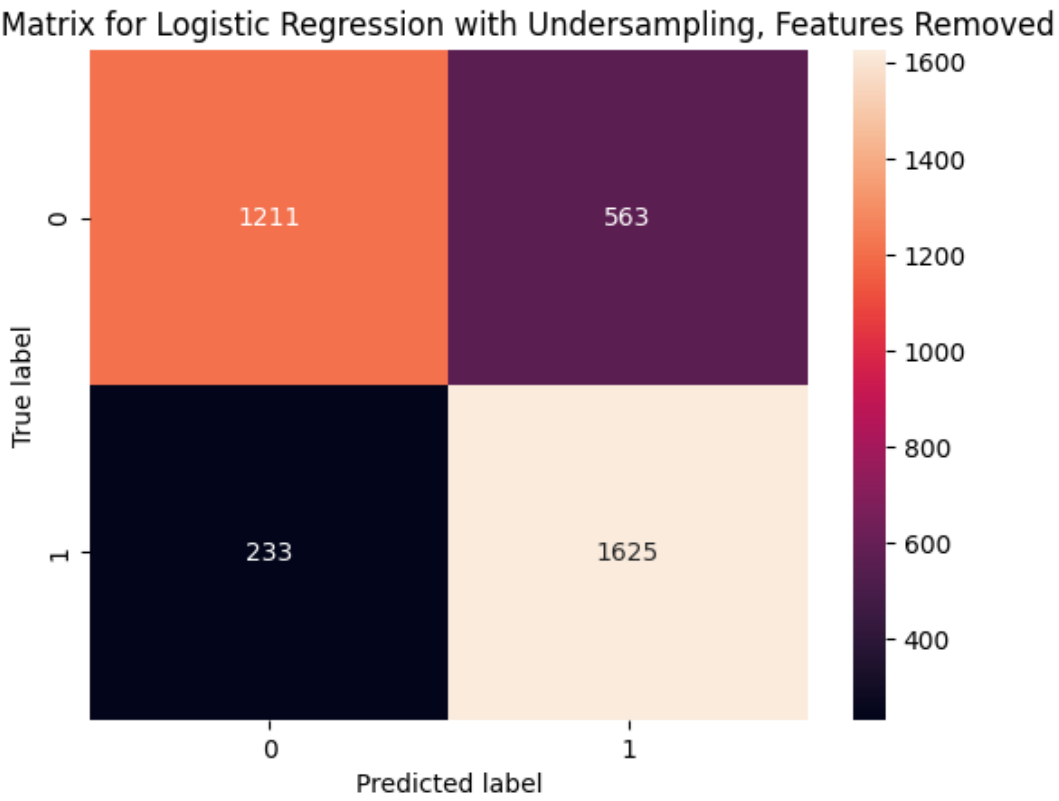

Fig21. Logistic regression confusion matrix with unimportant features removed and using undersampling

	Odds Ratio	Lower Confidence Interval	Upper Confidence Interval
Concealed Items	16.58	6.65	41.34
Combative	1.87	1.49	2.34
Resisted arrest	1.52	1.21	1.91
Mental instability	2.32	1.84	2.93
Assaulted officer	1.66	0.95	2.90
Year	0.07	0.06	0.08
April to June	4.97	4.24	5.83
January to March	6.03	5.17	7.03
July to September	4.56	3.91	5.32

Table15.Logistic regression odds ratios and confidence intervals with important features and using undersampling

Repeating the logistic regression and removing month, year and actions, but leaving in age, sex, and perceived race, we found an accuracy of 57% and AUC of 0.57.

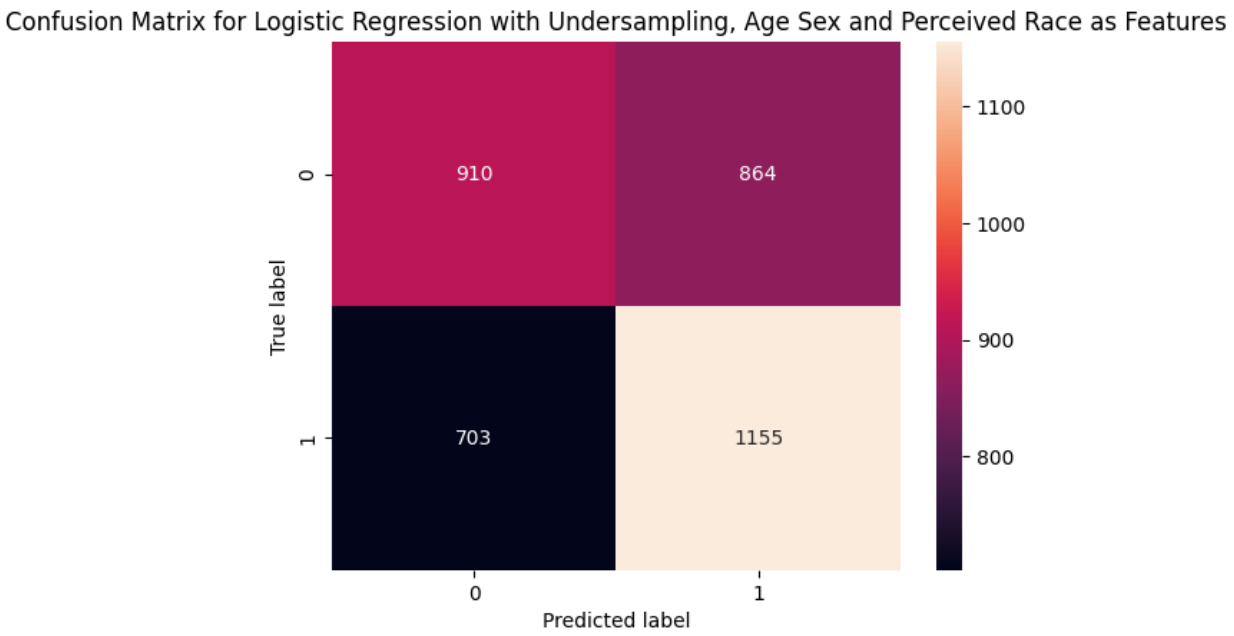

Fig22. Logistic regression confusion matrix with age, sex and perceived race as factors and using undersampling

	Odds Ratio	Lower Confidence Interval	Upper Confidence Interval
Sex	0.74	0.67	0.82
Age	0.88	0.85	0.91
Black	1.68	1.42	1.99
Indigenous	2.18	1.69	2.80
Latino	0.88	0.66	1.18
Middle Eastern	0.79	0.62	1.00
South Asian	0.83	0.65	1.05
White	1.72	1.46	2.02

Table16.Logistic regression odds ratios and confidence intervals with age, sex and perceived race as factors and using undersampling

Discussion

The results of our T-tests, ANOVAs, ANCOVAs, and Tukey HSD tests all point to widespread differences in arrest numbers and outcomes depending on age group, perceived race, and sex. We observed disproportionate numbers of Indigenous people, especially non-youths, being arrested multiple times. We saw greater probabilities of stripsearch for Black, Indigenous, and White non-youths. Our results also support the misogynoir observed by Thompson (2018) or the 1995 Report of the Commission on Systemic Racism, showing an interaction effect between perceived race and sex that affects probability of being booked, with Black, Indigenous, Latino and White females more likely to be booked, even though males are more likely to be booked overall. Overall, our results support the conclusion of Foster and Jacobs (2022) in that arrest counts and outcomes for Black, Indigenous, and White people are different from the mean. However, the results of our logistic regression were inconclusive. Because our logistic regression was based on individual arrest events rather than all arrest events linked to a specific individual, the confounding factors of year, month, and actions of arrest were found to be significantly more important than age, sex and perceived race

Conclusion

We disagree with Melchers' claim that data is not proof of discriminatory practice (2003); our results show systematic differences in arrest counts and outcomes that back up accusations of discrimination in the literature and in the media. There is a need for further examination of the behaviour of arresting officers that is causing these differences, and need for intervention in the TPS to reduce these differences.
There are limitations to our analysis. T-Tests and ANOVA both make assumptions that outcomes are normally distributed, whereas our outcomes all skewed heavily to the left. Our logistic regression was based on individual arrest events, which resulted in large confounding factors. It is also a concern that data about TPS is supplied by TPS, given that information can be misreported by arresting officers, and that the data may be presented in an altered state due to data cleaning or other processes. Unfortunately, there is a lack of data from a neutral party. Of particular interest is how the variables of sex and perceived race depended on arresting officer. For future directions, we believe it is important to "Study Up" on arresting officers, in the spirit of Barabas et al.'s work studying judges (2020), to determine how officers make their decisions in arresting events. More data on arresting officers would allow us to create binary classifiers to see if the outcome of an arresting event is more dependent on the individual being arrested, or the arresting officer.

References

Arrests and Strip Searches (RBDC-ARR-TBL-001). (2022, November 10). Toronto Police Service Public Safety Data Portal. https://data.torontopolice.on.ca/datasets/TorontoPS::arrests-and-strip-searches-rbdc-arr-tbl-001/about

Barabas, C., Doyle, C., Rubinovitz, J., & Dinakar, K. (2020). Studying up: Reorienting the study of algorithmic fairness around issues of power. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 167–176. https://doi.org/10.1145/3351095.3372859

Foster, L. & Jacobs, L (2020, May). Independent Expert Assessment Report: Toronto Police Service Race-Based Data Collection Strategy Phase I. *Toronto Police Services Board* and *Toronto Police Service.* https://www.tps.ca/media/filer_public/1f/e7/1fe76fd7-1bde-4c4d-bd90-091a5ccba5da/2bea48b6-40a7-4f43-ba92-ce59df850c73.pdf

Hayes, M. (2020, August 26). Ontario police watchdog clears Toronto officers in death of Regis Korchinski-Paquet. *The Globe and Mail*. https://www.theglobeandmail.com/canada/article-ontario-police-watchdog-clears-toronto-officers-in-death-of-regis/

Kari, S. (2015, December 16). Toronto officer says he tased Sammy Yatim to get medical help faster. *The Globe and Mail*. https://www.theglobeandmail.com/news/toronto/toronto-officer-says-he-tased-sammy-yatim-to-get-medical-help-faster/article27794484/

Melchers, R. (2003). Do Toronto Police Engage in Racial Profiling? *Canadian Journal of Criminology and Criminal Justice*, *45*(3), 347–366. https://doi.org/10.3138/cjccj.45.3.347

Mensah, J., Firang, D., J. Williams, C., & Afrifah, M. (2021). Racial Discrimination in the Canadian Criminal Justice System: How Anti-Black Racism by the Toronto Police Harms Us All. *Canadian Social Work Review*, *38*(2), 63–86. https://doi.org/10.7202/1086120ar

News ·, R. P. J. · C. (2022, October 21). *The police budget is one of Toronto's largest expenses. Here's what you need to know about it | CBC News*. CBC. https://www.cbc.ca/news/canada/toronto/toronto-police-spending-1.6623747

Press ·, J. O. · T. C. (2023, January 9). *Toronto police board approves proposed $48M funding increase despite criticism | CBC News*. CBC. https://www.cbc.ca/news/canada/toronto/toronto-police-board-budget-review-1.6707656

Report of the Commission on Systemic Racism in the Ontario Criminal Justice System. (n.d.). Retrieved February 22, 2023, from https://books-scholarsportal-info.myaccess.library.utoronto.ca/en/read?id=/ebooks/ebooks5/gov5/2020-01-07/1/reportracismont00comm

Thompson, C. (2018, Spring). Misogynoir in Canada. *Herizons*, *32*(1), 21.

Appendix: Measurement

Variables and Value categories	
Variable	Category
Dependent variable	
StripSearch (probability of a person being strip searched)	0 - 1
Booked (probability of a person being booked)	0 - 1
Arrests	1 - 7 times
Independent variable	
Perceived_Race	White South Asian Black Indigenous Latino Middle-Eastern East/Southeast Asian
Sex	M F
Age_group	Aged 17 years and under Aged 18 to 24 years Aged 25 to 34 years Aged 35 to 44 years Aged 45 to 54 years Aged 55 to 64 years Aged 65 and older
Youth	True False

Table17. Breakdown of variables

StripSearch
One of the dependent variables. The StripSearch variable from the original dataset is a boolean variable. It has 0 and 1. 0 indicates that the subject was not subjected to a strip search. 1 indicates that the subject was subject to a strip search. We were able to determine the total number of arrest occurrences for a certain person since each arrest records a special ID for the person being arrested. To show the likelihood of someone getting striped searched, we altered StripSearch. This makes it easier for us to conduct our tests.

Booked
One of the dependent variables. The Booked variable from the original dataset is a boolean variable. It has 0 and 1. 0 indicates that the subject was not booked with an officer. 1 indicates that the subject was booked with an officer. We were able to determine the total number of arrest occurrences for a certain person since each arrest records a special ID for the person being arrested. To show the likelihood of someone getting booked with an officer, we altered Booked. This makes it easier for us to conduct our tests.

Arrests
One of the dependent variables. This column records the number of times a specific person is arrested.

Perceived_Race
One of the independent variables. This column contains a total of seven races as perceived by the arresting officer.

Sex
One of the independent variables. We removed “U” sex from the original data due to a small sample size. This column now contains M (male) and F (female) refers to the sex of the person.

Age_group
One of the independent variables. The original data contains nine age groups. Some groups were combined due to overlap in data.

Youth
One of the independent variables. Youth is a boolean variable. True means the person is aged 18 and older. False means the person is aged under 18.