Exploring Factors Affecting the Age of Arrest

A Final project submitted in conformity with the requirements.

Group 48: Yixin Li, Jianjuan Fan

For the Course of INF 2178

Faculty of Information

University of Toronto

Abstract

Society believes several factors contribute to criminal behavior. Among them, criminal behavior is often associated with age from a universal social perspective. A great deal of attention is also paid by society to the problem of juvenile delinquency. Using the arrest and strip search dataset from the Toronto Police Service, Exploratory Data Analysis (EDA) was conducted to examine the age, gender, location, cooperative behavior, and combative behavior of suspects in Toronto at the time of arrest. Data obtained from the EDA is used to determine whether factors influencing age and other criminal characteristics are interdependent. This final study focuses on the correction of arrest age group and other crime characteristics. Three preliminary conclusions were drawn based on power analysis, ANCOVA Test, and logistic regression. In power analysis, the gender feature shows a relatively small effect size, making it difficult to detect, while the race feature exhibits a better fit effect size, leading to more credible results in hypothesis testing. However, in ANCOVA, the gender feature impacts the age group at arrest, indicating that even after controlling for the effect of the race feature, significant gender differences still exist. Preliminary results from logistic regression show that there is no clear correlation between age group of arresting and arresting behavior. Different arresting ages showed no correlation with either cooperative or combative behavior at the time of arrest. Simultaneously, it is currently inconclusive from the perspective of logistic regression whether there is or is not a correlation between age group and sex.

Keywords: Age, gender, criminal behavior, teenager, power analysis, ANCOVA, logic regression

1. Literature Review

A number of efforts have been made by the Canadian government to understand the factors and patterns behind youth gangs and criminal behavior. A study published in 2017 by Dr. Dunbar examined the causes of the formation and development of youth gangs. In his research, he concluded that mainstream academia lacks an appropriate measure of the prevalence of youth gang involvement and activity in Canada (Dunbar, 2017). However, Canadian society has gained some understanding of certain influential groups that contribute to the formation of youth gangs, such as Aboriginal youth, new immigrants, and young female groups. According to Dr. Dunbar's research, most youth gang members in Canada are young males aged 12-17 and 18-24, with young adults serving as leaders (Dunbar, 2017). There are also age-related stable features in these organizational patterns, with female participation being lower than male participation, particularly among older females (Allen & Superle, 2016).

In light of the importance of youth crime as a social issue, it is vital to investigate the relationship between age and crime. The Arrests and Strip Searches (RBDC-ARR-TBL-001) dataset in Toronto offers a unique opportunity to examine the relationship between arrestee age and other characteristics. The purpose of this study is to examine the correlation between different age groups of arrestees and specific criminal characteristics, such as their age and location of arrest, their criminal behavior, and their likelihood of engaging in certain behaviors during arrest. Using the ANOVA test in the mid-term report of this study, we examined whether the presence of one attribute influences the birth and enhancement of another attribute when the age group attribute is associated with different characteristics. Based on the Midterm dataset, this study will also examine whether age groups are strongly correlated with other factors.

2. Introduction

RBDC-ARR-TBL-001 contains multiple significant attributes, including age groups, race, gender, location of arrest, type of crime, cooperative attitude during arrest, and combative attitude during the arrest. This study aimed to examine the relationship between youth groups and other recorded criminal characteristics. Specifically, it is intended to determine whether age groups (especially youth groups) are strongly correlated with specific features among arrested criminals. The midterm project observed that the dataset consisted primarily of adult male offenders based on preliminary EDA and ANOVA tests. Moreover, the age groups with the highest arrest probability among different racial groups fell within the 25-45 age bracket. Additionally, this dataset contained a significant data gap regarding youth crime issues, which could not be explored during the preliminary exploration phase of the midterm project.

In this study, we build on the results of the ANOVA test conducted during our midterm project. It further employs EDA, power analysis, ANCOVA, and logistic regression to analyze the correlation between different age groups and other factors, especially gender factors. We aim to investigate whether different age groups strongly correlate with gender and other crime factors using these three novel analytical methods. Especially in the midterm project, we discovered that most arrested individuals were male, with only a few females. It reveals some gender features in the dataset. We will further investigate the specific role of gender in the dataset and whether it interacts with the age group, which is the focus of our exploration. The ultimate aim is to examine whether early intervention in specific features can reduce the probability of youth group crimes when various features strongly correlate with age groups. Additionally, this research can be extended sociologically further to explore social morals, fairness, and respect. We aim to analyze the data's differences to identify if specific minorities are subject to police abuse and seek reasonable solutions.

Based on the background and existing research, we examined the factors that influence the age at arrest for crime, and our initial hypothesis is:

- H0: There is no difference in the mean of Age_group__at_arrest between the two levels of one certain attribute variable.
- H1: There is a difference in the mean of Age_group__at_arrest between the two levels of one certain attribute variable. We will further explore the dataset data through EDA, as well as examine the assumptions.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a method for understanding the interrelationship between variables as well as the relationship between variables and predicted values by understanding the dataset. The research will benefit from better feature engineering and model development as a result. It is a crucial step in data analysis. The tools used in this EDA include data science libraries (NumPy, Pandas, SciPy), and visualization libraries (Matplotlib, Plotly, Seaborn).

3.1 Data Overview

We first imported NumPy, Pandas, and SciPy so that the data in the dataset can be pre-processed. There are 65276 instances and 25 attributes in the dataset. In the dataset description, it is given that there are 65276 instances and 25 attributes in the dataset, and the dataset contains a mixture of categorical and numerical variables. Next, we have Identified the Null Values. Datasets are related to arrest and strip search, in which nine features contain Null values. What we need to pay attention is: ItermFound; SearchReason_PossessEvidence; SearchReason_PossessWeapons; SearchReason_AssistEscape and SearchReason_CauseInjury contain 57475 null Values, since these are binary variables, it is difficult to estimate missing values based on the available data. Meanwhile, the proportion of missing data is large, dropping all observations with missing data can result in a significant loss of data, which will introduce bias into the analysis and reduce the statistical power. Therefore, we decided to drop these features by columns. The Occurrence_Category contains 165 null values, Age_group__at_arrest_ contains 24 null values and ArrestID contains 469 null values, these are also difficult to estimate missing values based on the current dataset, but since the proportion of missing data is small (less than 5% of the dataset), dropping missing values is unlikely to have a significant impact on the analysis results. So, we dropped the missing value by rows. The final data size is 64615, and the dataset contains a mixture of categorical and numerical variables.

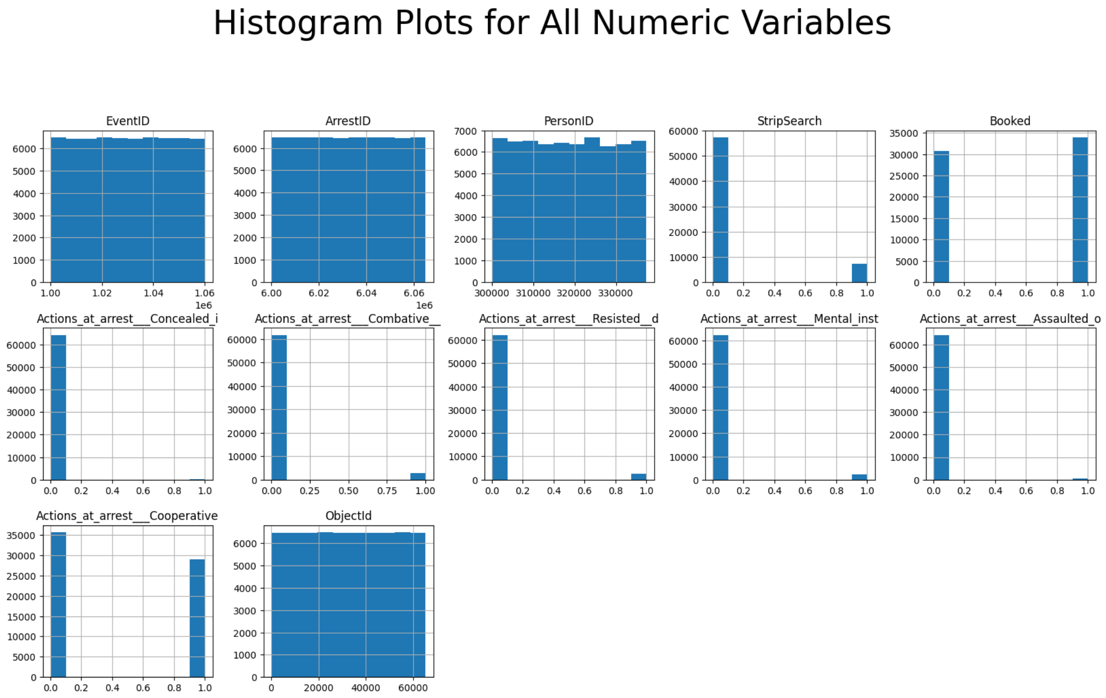
3.2 Explore Numerical Variables.

The dataset contains 13 numerical variables, which are:'Arrest_Year', 'EventID', 'ArrestID', 'PersonID', 'StripSearch', 'Booked', 'Actions_at_arrest___Concealed_i', 'Actions_at_arrest___Combative__', 'Actions_at_arrest___Resisted__d', 'Actions_at_arrest___Mental_inst', 'Actions_at_arrest___Assaulted_o', 'Actions_at_arrest___Cooperative', 'ObjectId'. However, since most of these numerical variables still play a categorical role and have no specific numerical meaning, we then focus on the distribution and the outlier.

3.2.1 Check Outliers. To check the outliers, we printed out descriptive statistics (such as mean, standard deviation, and quartiles) for the numerical columns in the DataFrame df rounded to two decimal places. On closer inspection, there are no large outliers in this dataset for numerical variables.

3.2.2 Histogram for All Numerical Variables. According to the histogram in Figure 1, it was clear that the data for each attribute did not appear to be normally distributed.

Figure 1



3.3 Explore Categorical Variables.

We checked the labels of each categorical variable and found that Age_group__at_arrest_ contained two groups of duplicates with different expressions but referring to the same thing, so we replaced “Aged 17 years and younger” with “Aged 17 years and under”, and “Aged 65 and older” with “ Aged 65 years and older”. Similarly, the Youth_at_arrest__under_18_years contains duplicate data. We replaced “Youth (aged 17 and younger)” with “Youth (aged 17 years and under)” . The number of unique value in each categorical variable as shown in Table 1

Table 1

The Number of the Unique Value in Categorical Variables

Categorical_Variables	Length(count)
Arrest_Month	4
Perceived_Race	8
Sex	3
Age_group_at_tarrest_	7
Youth_at_arrest_under_18	2
Arrest_Location	18
Occupancy_Category	31

3.4 Univariate Analysis - Explore Target Variable.

Since the target variable we are studying is Age_group_at_arrest_, we would like to further explore this target variable. From Figure 2 we can see that the number of unique values in Age_group__at_arrest_variable is 7.

Figure 2

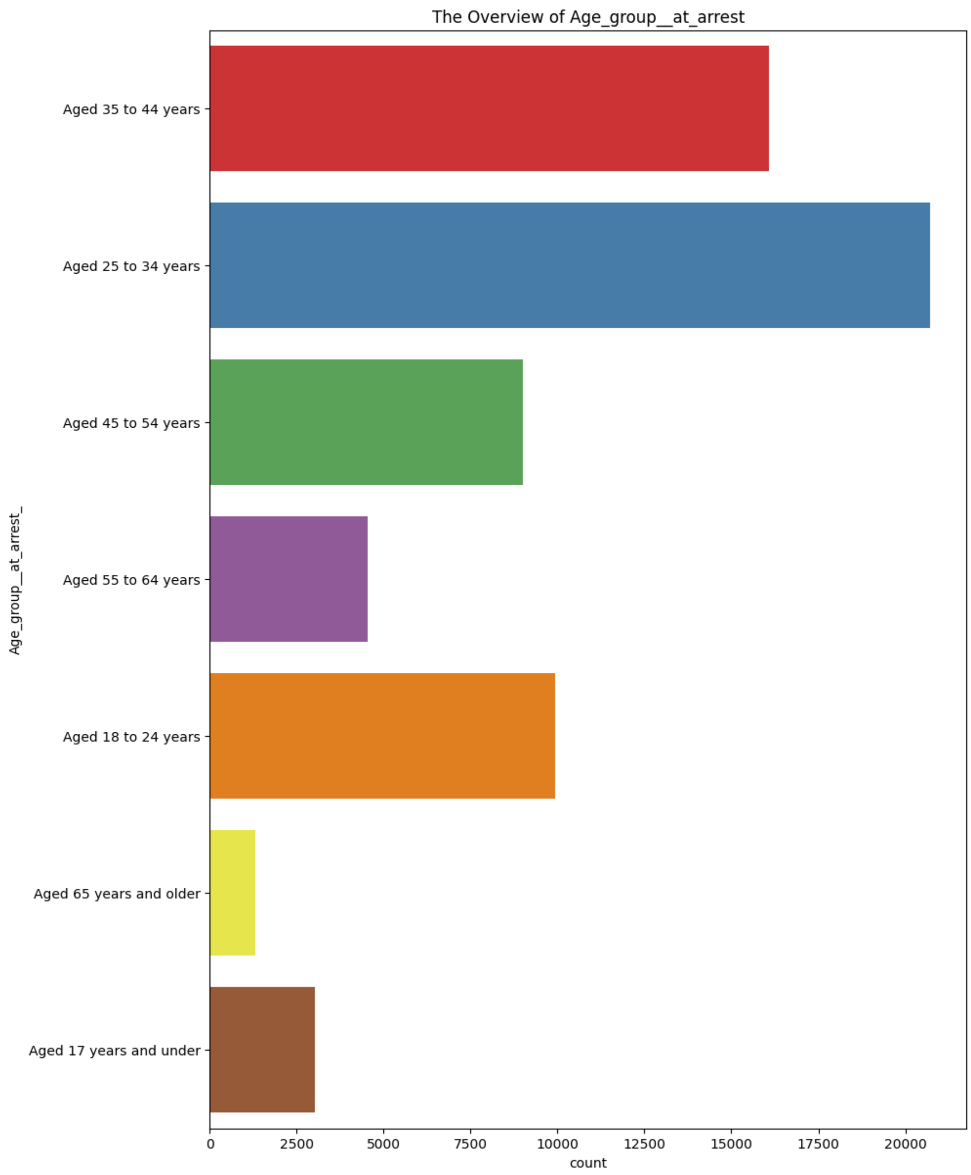


Table 2

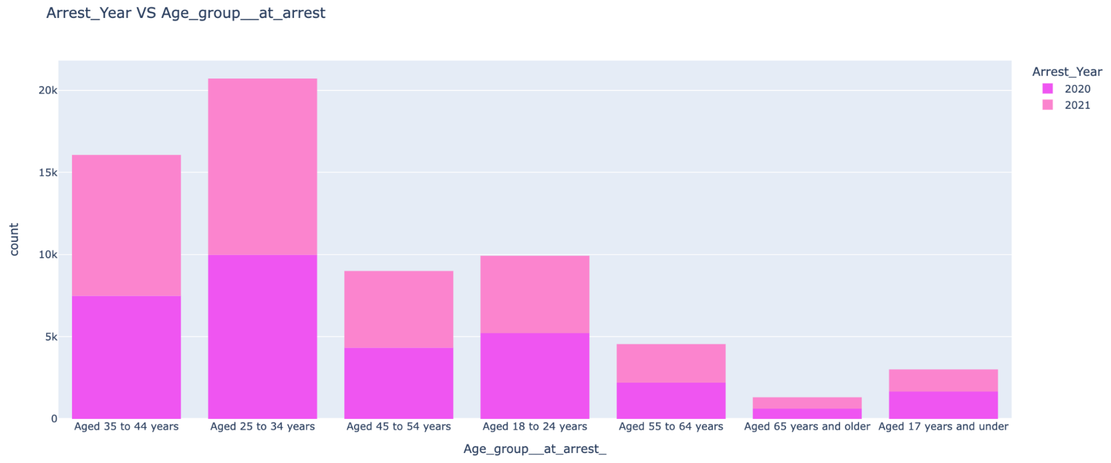
Age Group	Count	Percentage (%)
Aged 17 years and under	3012	0.046615
Aged 18 to 24 years	9934	0.153741
Aged 25 to 34 years	20725	0.320746
Aged 35 to 44 years	16072	0.248735

Aged 45 to 54 years	9003	0.139333
Aged 55 to 64 years	4553	0.070464
Aged 65 years and older	1316	0.020367
Sum	64615	1

Among those arrested and strip searched, young adults between the ages of 25 and 34 had the highest percentage, with 32.1%. Adults between the ages of 35 and 44 had the second highest percentage, with 24.9%. Young adults aged 18-24 ranked third. Middle-aged individuals aged 45-54 rank fourth with 13.9%. The graph indicates that young adults are the group most likely to be arrested and strip searched. There is a relatively lower likelihood of arrests (strip searches) for elders, teenagers, and kids. These graphs also show that the percentage of arrests of minors in Toronto is very low. Adults constituted 95.3% of the arrests.

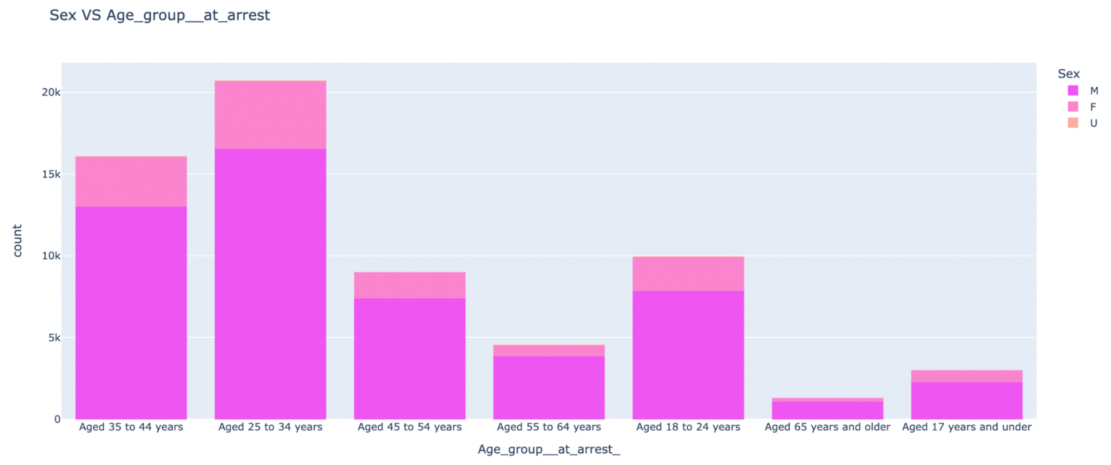
3.4.1 Arrest_Year in Age group Overview. The Toronto Police Department's dataset on arrests and strip searches focuses on the years 2020 and 2021. As can be seen in the bar chart, even though the age groups to which the arrestees belonged differed, they were equally likely to be arrested in 2020 and 2021. There is no year in the data set where the odds of being arrested are higher. It is also reasonable to infer that the probability of crime in 2020 and 2021 should be close to similar levels.

Figure 3



3.4.2 Sex in Age group Overview. According to chart visualization of the data for this attribute of gender, males constitute most suspects arrested and strip searched. As a result, 80.7% of the population is male. In comparison, the percentage of females is 19.3%. Besides, the number of female arrestees is low regardless of age group. The percentage of female arrestees in each age group is much lower than the percentage of males. At the same time, age-specific surveys among the female population found that they had the highest number and probability of arrests in the 25-34 age group. In this regard, males likewise presented the highest number and probability of arrest in the 25-34 age group. Thus, it can be said that 25-34 years old is the age group with the highest probability of occurrence of arrests for both males and females.

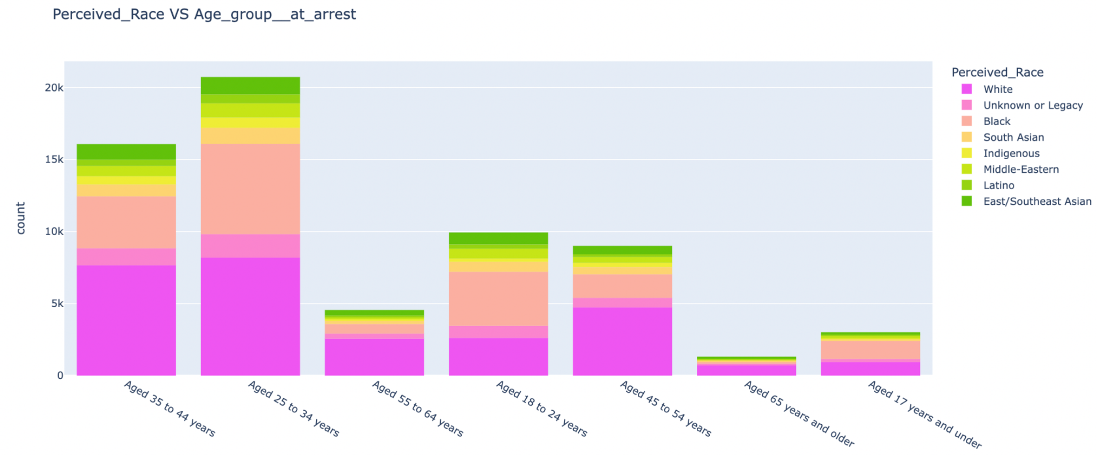
Figure 4



3.4.3 Race in Age Group Overview. The age groups 25-34 and 35-44 continue to be the two groups with the highest number of arrests for each race. However, racial disparities were found in the older arrest group. In the 55-64 age group, there is a significant decrease in the number of arrests for blacks overall, while the number of arrests for whites remains at a steady high. At the same time, racial differences emerge in the number of arrests of youth in the 18-24 age group. The number of arrests for blacks was higher in total, while the number of arrests for whites was lower.

Other Inspection: The visualization for the attribute of race indicates that the highest percentage of suspects arrested or strip-searched were white, at 42.5%. Blacks were followed by whites with 26.8%. Toronto was the primary location for data collection. Toronto is a multiracial area where whites account for most arrests (strip searches). However, whites constitute most of Toronto's population. Accordingly, it is difficult to claim that whites are more likely to be arrested purely based on proportionality. It is also necessary to conduct a comparative analysis of other factors. Particularly when compared to blacks, who are a minority group in Toronto based on their population percentage but have the second highest rate of being arrested (being strip searched), only less likely than whites (42.5% - 26.8%) = 15.7%.

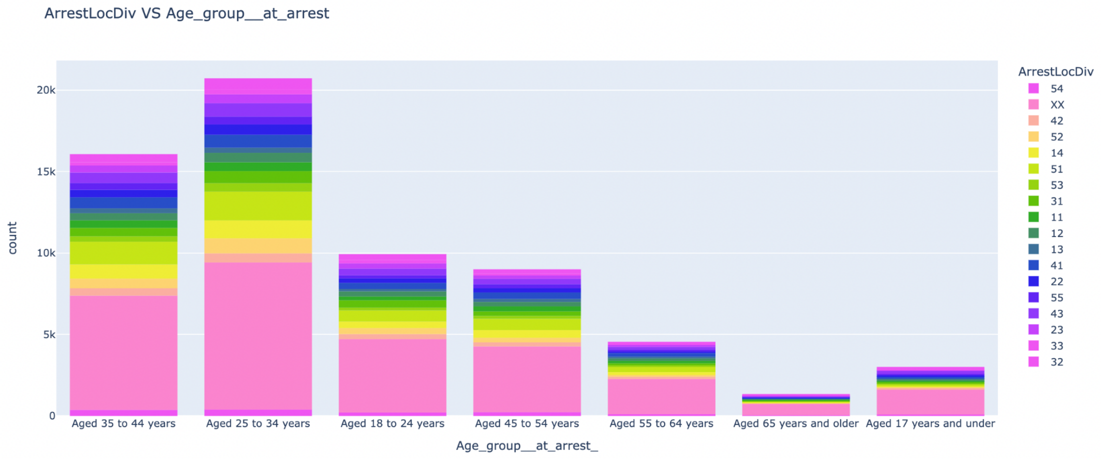
Figure 5



3.4.4 Location in Age Group Overview. In the comparison against area and age group, it can be found that the age group of 18-54 years old is most likely to be arrested in District 54.

Other Inspection: According to an analysis of the areas in which arrests were made, 45.4% of all arrests occurred outside of Toronto (or in areas that could not be identified). With a 7.7% arrest rate (strip search rate), 51 had the highest arrest rate among the identified Toronto areas. The other Toronto areas fluctuated within a small range of 2% to 4%.

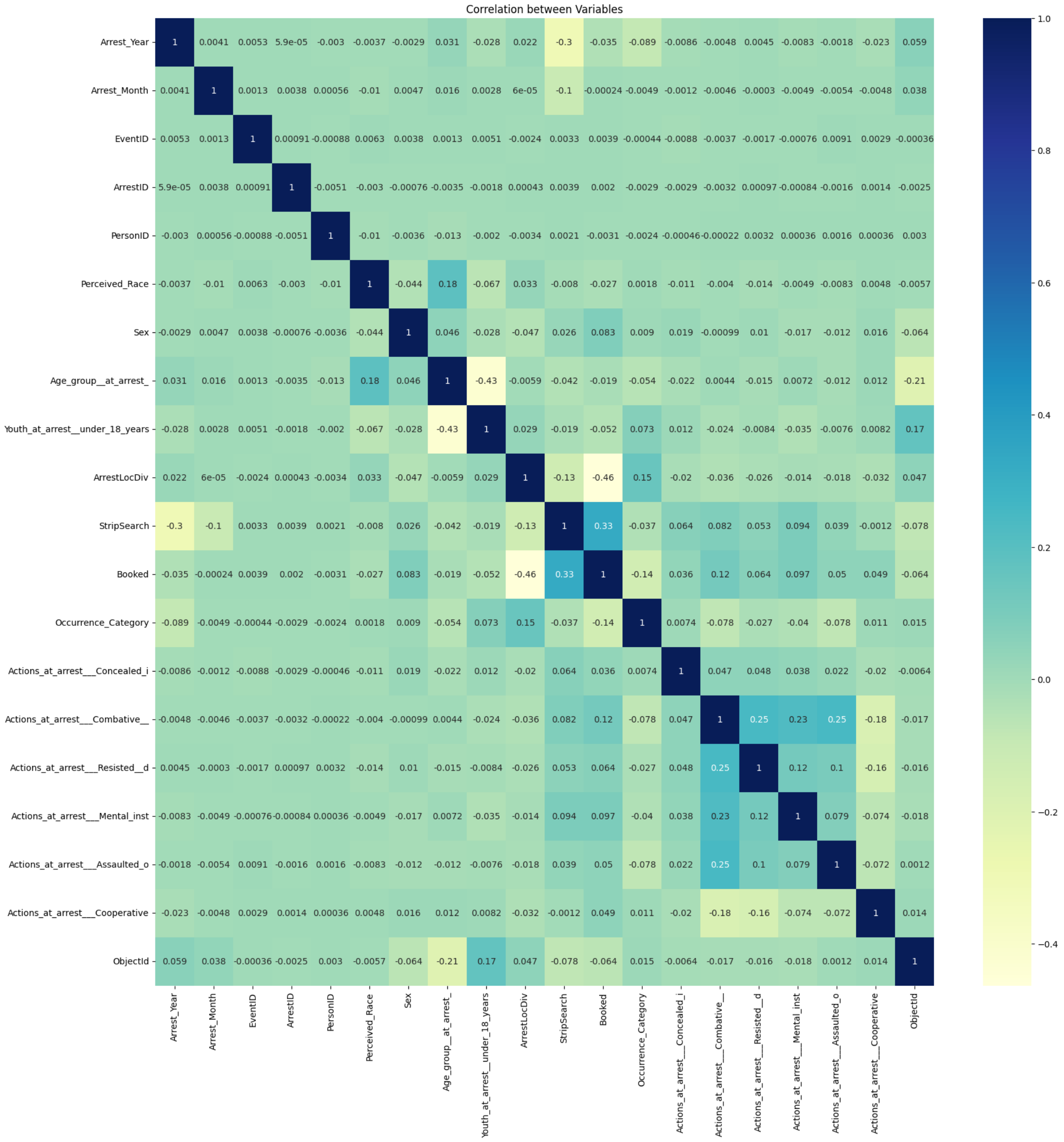
Figure 6



3.5 Multivariate Analysis

To discover patterns and relationships between variables in the dataset, we also perform multivariable analysis, which mainly focuses on the Heatmap. We first label encoded all categorical variables using label_encoder method, then generalize the heatmap by sns.heatmap method.

Figure 7



From the Heatmap figure we can see there is no strongly correlation between two variables, however, relatively speaking, booked and ArrestlocDiv, youth_at_arrest_under_18 and Age_group_at_arrest, Booked and StripSearch, arrest year and stripsearch, Actions_at_arrest___Combative__ and Actions_at_arrest___Assaulted_o, Actions_at_arrest___Combative__ and Actions_at_arrest___Resisted__d have weakly correlation.

3.6 Insights

Based on an exploratory analysis of the data, it was determined that the group that was arrested (strip searched) was primarily male and less female. There was a predominance of whites, followed by blacks. Except for areas outside of Toronto, Ward 51 was the most likely area to be arrested with well over 7% of 2%-4%. The majority of those arrested were between the ages of 18 and 54.

The different age groups showed some correlation in different attributes (region, race, and gender). In general, the 25-34 and 35-44 age groups show high arrest rates by gender, region, and ethnicity. However, the adolescent group seems to show different status in these different attributes. For example, it can be found in the dataset that Blacks have a relatively high percentage of arrests in the 18-24 age group, and a low probability of arrest in the 55+ age group. One thought is that a two-way anova test could be used to explore whether race, gender, and region have a mutual effect with age.

4. Methods

Based on the EDA, we have decided to investigate factors that affect the age of arrest, with a focus on two variables: Sex and Perceived Race. Specifically, we will group the data by the types of unique variables (unique values in Table 1) and study whether the mean of Age_group_at_arrest_ is equal among the groups. Since both Sex and Perceived Race have more than two unique values, we will start with an ANOVA test and proceed with further testing based on the results, such as power analysis to verify the probability of correctly rejecting the null hypothesis, ANCOVA and Tukey's test to further validate the results and enhance their credibility. Before conducting the ANOVA test, we need to first check the ANOVA assumptions. The assumptions for ANOVA test are:

1. Randomness and independence. The samples were randomly selected from the population and randomly assigned to each treatment group. Therefore, each observation is independent of the other observations.
2. Normality. The values in each sampling group are assumed to be from a normally distributed population. Through the histogram plot in EDA and the Shapiro test, we can obtain that the data do not satisfy the Gaussian distribution, so the results may be influenced, and the estimates of the within-group errors may deviate from the true values, leading to erroneous conclusions. We will try to improve the accuracy of the conclusions with multiple analyses.
3. Homogeneity of variance. All groups have equal variance. We conducted a Bartlett test for one of the main independent variables "Sex", but since the data were categorical variables after label encoding, which would affect the variance results, further hypothesis testing can attempt nonparametric methods to ensure the accuracy of the results.

5. Hypothesis Testing Results

5.1 One-Way ANOVA

After visualizing correlations between different variables in a dataset by Heatmap, we then used the one-way ANOVA for each feature to see if there were statistically significant differences between the variable means to further analyze the relationship between each feature and the age of arrest. We first selected Age_group__at_arrest_ as a dependent variable and set all features as independent variables respectively. We divide the data into multiple groups according to different categories/levels and then measure their mean of Age_group__at_arrest_. Although the data does not seem to be normally distributed, However, based on the large sample size, we believe that the ANOVA test will still give some statistical significance to the differences between the various groups.

Our analysis found that all attributes except Actions_at_arrest__Combative_ and Actions_at_arrest___Mental_inst have a statistically significant, since the p-values are less than 0.05, we have sufficient evidence to reject the null hypothesis that the means are all equal. We take the Sex feature to further explain. The hypothesis is:

- H0: There is no difference in the mean Age_group__at_arrest_ between the two levels of the Sex variable.
- H1: There is a difference in the mean Age_group__at_arrest_ between the two levels of the Sex variable.

Table 3

ANOVA Table for Sex Variable

Source	SS	DF	MS	F	p-unc	np2
Sex	243.2965	2	121.6482	68.57422	1.78E-30	0.002118
Within	114619.4	64612	1.773965	NaN	NaN	NaN

As a result (Table 3) of examining the relationship between the gender variable and the two characteristics of minors who were arrested (strip searched), the following ANOVA process can be derived.

1. The sum of squares of the gender variable is 243.2965.
2. The degrees of freedom (df) of the gender variable is 2, which is the number of levels of the gender variable minus 2.
3. The F-statistic (F) for the gender variable is 68.57422. As a ratio, this indicates the amount of age group variation that can be explained by gender. This is the amount of variation that cannot be explained by the gender variable.
4. The p-value (PR(>F)) for the gender variable is 1,78e-30, which is very small. The ratio represents the amount of variation in age groups that can be explained by the sex variable in comparison to the amount of variation that cannot be explained by it.

In this case, A p-value less than 0.05 is typically considered statistically significant, we can conclude that there is strong evidence for a difference in mean of age group between the two levels of the Sex variable.

Table 4

ANOVA Table for Perceived_Race Variable

Source	SS	DF	MS	F	p-unc	np2
Perceived_Race	4894.259	7	699.1798	410.7716	0.00E+00	0.04261
Within	109968.4	64607	1.702113	NaN	NaN	NaN

Similarly, for Perceived_Race , A p-value of 0 is less than 0.05, so we considered to reject the H0, there is evidence for a difference in mean age group between the two levels of the Perceived_Race.

5.2 Power Analysis

Power analysis is an important part of the experimental analysis phase, which gives us an indication of the reliability of the conclusions of the experiment given the available data. In power analysis, we need to pay close attention to the following four statistics: sample size, effect size, significance level and statistical power. Since we mentioned before that the ANOVA assumption is not fully satisfied, which leads to a possible bias in our hypothesis testing, it is more necessary to further investigate the probability of correctly rejecting a null hypothesis.

First, we conducted a power analysis on the Sex variable. Using the describe function, we found that the sample size being tested was 64615. The number of individuals in each group were group1 (Male): 52106, group0 (Female): 12500, group2 (Undefined): 9. In order to make the sample size more accurate, we used weighted averages with the number of individuals as weights. The resulting average sample size for each group was 44437. From the ANOVA table, we found that the eta-squared was 0.002118. By calculating eta_squared / (1 - eta_squared), we obtained the Cohen's f coefficient as the effect size. We set alpha as 0.05. Substituting the known variables into the FTestAnovaPower function in Python, we found that the statistical power was approximately 0.065, which is not high.

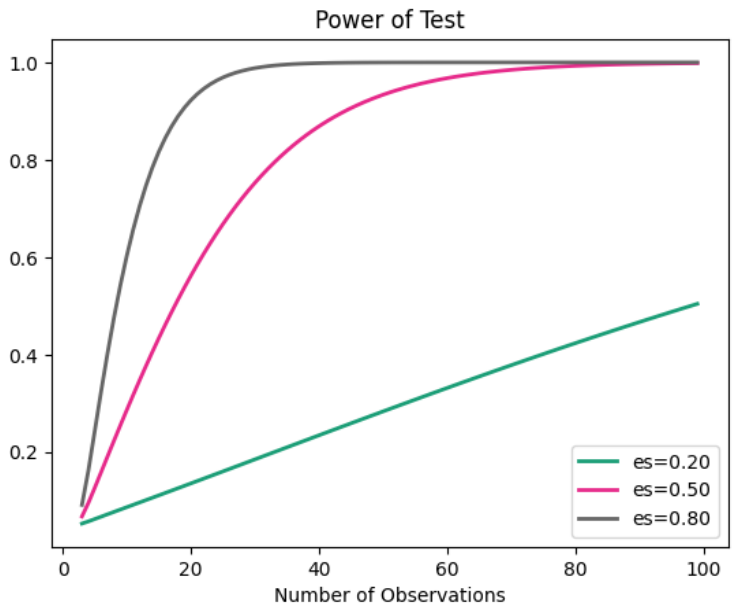
We then proceeded to further investigate. Assuming we wanted a higher statistical power, we set a target power of 0.8 and keeping all other variables constant, we used the FTestAnovaPower function in Python to determine the sample size required. As a result we found that we would need at least 2,138,675 samples for each group, which is a huge number. However, we noticed that the eta-squared value from the ANOVA table was very small. If we had chosen a normal effect size of 0.5, we would only need 41 samples for each group to achieve a power of 0.8. Small effect size may be associated with uneven variance and not being normally distributed.

When the sample variance is uneven, it may lead to false positive or false negative results. This is because in the group with larger variance, the observed variability may be greater than it is, leading to a false positive result, while in the group with smaller variance, the observed variability may be smaller than it is, leading to a false negative result. Similarly, if the data do not obey a normal distribution, this may lead to an underestimation or overestimation of the effect size.

For the Perceived_Race variable, the results are more optimistic. Through the ANOVA table, we found that its eta-squared is approximately 0.04261. We used the same method as Sex variable to calculate the weighted average sample size, which is 17433. We also calculated its Cohen's f coefficient and used FTestAnovaPower, from which we found that the power of the hypothesis test for this variable reached 0.997. We also conducted further tests to compare it with the Sex variable. Assuming that we need a power of 0.8, only 7252 samples are needed to achieve this for the Perceived_Race variable. So, our 17433 samples are more than enough.

Through the observation of the result for two variables, we found that the effect size and power are positively correlated with a sufficiently large sample size. We plotted the power of the test graph to validate this conclusion, setting the effect size to 0.2, 0.5, and 0.8 to represent weak, moderate, and strong effect sizes, respectively. The results are shown in Figure 8.

Figure 8



From Figure 8 we can observe that, with the same sample size, as the Effect Size increases, the power of the test also increases. On the other hand, with the same power, a larger Effect Size requires fewer samples. When the sample size is very small, the impact of a larger Effect Size on power is more sensitive. Ideally a larger sample size and a larger effect size can result in better statistical power in hypothesis testing.

Based on this finding, we further explored the eta-squared variable in power analysis. As this variable is a statistic that measures the effect size of variance differences between the dependent variable in different groups, the small sample size of the undefined group, which only had 9 samples, could greatly affect the accuracy of eta-squared. Therefore, we decided to remove this group and only perform hypothesis testing on the Male and Female groups. Since there are only two groups, we transformed the ANOVA test into a t-test.

We used the ttest_ind method to perform a t-test on the age group of the male and female groups, and the resulting t-value was approximately 11.71. The p-value was still less than 0.05, indicating that we could still reject the null hypothesis. We then conducted a power analysis based on the sample, using Cohen's d as the effect size. After calculation, the cohen's d was approximately 0.117, which was larger than the eta-squared in the ANOVA result. We then used the TTestIndPower method to calculate the power, which was found to be 1. Compared to the ANOVA result, the power has significantly increased. We also calculated the minimum sample size required to achieve 0.8 power, which is only 1139 samples per group. Therefore, our sample size is sufficient, and this result also confirms that extreme sample size can lead to bias in ANOVA results. Since the power of the ANOVA test was very low, more research is needed to confirm its conclusions, and we hope to further explore the validity of its conclusions through subsequent tests.

5.3 Two-Way ANOVA

Since the power of the hypothesis test for Sex variable is low, we then selected Sex and Perceived_Race variables and applied Two-Way ANOVA to determine combinations of factors that may be statistically significant. As the result shown in Table 5, we can see that the p-values for Sex and Perceived_Race turn out to be less than 0.05 which implies that the means of both the factors possess a statistically significant effect on Age_group__at_arrest_. The p-value for the interaction effect is also less than 0.05 which depicts that there is strong evidence to say there is significant interaction effect between Sex and Perceived_Race. However, it is noteworthy that the p-value for the sex variable is now 0.807688, which is no longer greater than 0.05, which means that we have evidence to accept the null hypothesis that the mean age group values are equal across different sex groups. Our explanation for this is related to the results of the power analysis. When considering the Sex variable alone, the effect of sex on age group may be significant. However, when adding the race variable to the two-way ANOVA model, the model is comparing the effects of both sex and race variable on age group. The Perceived_Race variable has a significant effect on age with super high power, and there is some correlation between the sex and race variables (p<0.05). Therefore, the effect of sex on age groups may be masked by the effect of race in the two-way ANOVA model. As a result, it shows that the effect of sex is no longer significant in the two-way ANOVA model. Therefore, we then selected Perceived_Race as a covariate to further do an ANCOVA test on Sex variable, adjusting for the effect of the covariate on the group means. This allows for more accurate comparisons between groups and gives more reliable conclusions regarding the association between variables.

Table 5

Two-way ANOVA Table for Sex and Perceived_Race Variables

	sum_sq	df	F	PR(>F)
C(Sex)	0.724252	2	0.21358	0.807688
C(Perceived_Race)	6914.3941	7	582.5816	0
C(Sex):C(Perceived_Race)	63.714855	14	2.684191	0.002762
Residual	109522.9	64596	NaN	NaN

5.4 ANCOVA

Unlike the ANOVA test, ANCOVA is used to analyze the relationship between the dependent and independent variables and considers the effect of covariates. It is usually used to control for the effects of certain covariates to reduce the variance of the errors, thus increasing the power of testing the relationship between the independent and dependent variables.

Based on the previous power analysis and ANOVA test, we selected the Sex variable as the independent variable, Perceived_Race as the covariate, and Age_group__at_arrest as the dependent variable for the ANCOVA test, and the result of the test is shown in Table 6.

Table 6

ANCOVA Table for Sex and Perceived_Race Variables

	Source	SS	DF	F	p-unc	np2
0	Sex	338.7917	2	98.95393	1.23E-43	0.003054
1	Perceived_Race	4014.023	1	2344.823	0.00E+00	0.03502
2	Residual	110605.4	64611	NaN	NaN	NaN

The results show that after considering the influence of perceived race, the p-value of the Sex variable is less than 0.05, indicating that the mean values of the age group at arrest among different sex groups are significantly different at the 95% confidence level. This is consistent with the results of the ANOVA test, and the ANCOVA analysis has increased the credibility of the ANOVA test under low statistical power.

The problem with ANOVA is that it only compares the means between groups and determines whether any of these means are statistically significantly different. In other words, it simply tells us that not all the group means are equal but doesn't tell us which groups are different from each other. We then performed the post hoc test (Tukey's test) to find out exactly which groups are different from each other.

5.5 Tukey's Test

We used the pairwise_tukeyhsd method to generalize the Tukey's test result for the Sex variable that we used in ANOVA. The result is shown as Table7.

Table 7

Tukey's Test Table for the Sex Variable

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	0.1553	0	0.1243	0.1864	TRUE
0	2	0.1076	0.9682	-0.9334	1.1485	FALSE
1	2	-0.0478	0.9936	-1.0884	0.9928	FALSE

According to the results of Tukey's post-hoc test, the mean difference between group0(Female) and group1(Male) is 0.1553, and the p-value is less than 0.05. Therefore, we can reject the null hypothesis of equal means between them and consider their difference to be significant. The mean differences between group0(Female) and group2(Undefined), and between group1(Male) and group2(Undefined) are 0.1076 and -0.0478, respectively, and both p-values are greater than 0.05. Therefore, we cannot reject the null hypothesis of equal means between them, and consider their differences to be insignificant. Overall, based on the results of this Tukey's post-hoc test, we can conclude that only the mean difference between group0(Female) and group1(Male) is significant.

Similarly, for the Tukey's test on perceived race, due to the large number of groups, we only display the groups with false rejects in the table below.

Table 8

Tukey's Test Table for the Perceived Race Variable

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-0.0273	0.9949	-0.1358	0.0812	FALSE
1	5	-0.079	0.1248	-0.168	0.0099	FALSE
2	5	-0.0517	0.858	-0.1637	0.0603	FALSE
2	6	-0.0684	0.5177	-0.1748	0.038	FALSE
3	4	-0.0899	0.2827	-0.2073	0.0276	FALSE
3	6	0.1097	0.0502	0	0.2195	FALSE
5	6	-0.0167	0.9991	-0.1031	0.0698	FALSE

Among the 28 pairs of groups, only 7 are not significant. The conclusion of the Tukey's test is consistent with the previous power analysis, indicating that higher power means higher statistical power of the study, that is, the ability to more accurately reject false null hypotheses.

6.Logistic Regression

6.1 Linear Regression

Before conducting logistic regression on the data, we employed linear regression to screen the age and other crime features preliminarily. The reasons for choosing linear regression first are as follows: Firstly, logistic regression is built upon linear regression and is essentially a generalized linear regression type. Secondly, linear regression can help us become familiar with and select data. We usually need exploratory data analysis to understand the distribution and features of the data. Linear regression can help us better understand data and relationships between variables. Thirdly, logistic regression requires a specific linear relationship between the independent and dependent variables. We can check whether the data meet this requirement first using linear regression. Finally, logistic regression usually requires the selection of some independent variables as input to the model. By performing linear regression first, we can select and filter the independent variables to choose the ones that significantly impact the dependent variable. Therefore, it is crucial to conduct linear regression to screen age and other features again before conducting logistic regression analysis.

During the EDA process, we found many missing or uncertain values in the dataset, making it difficult for us to conduct feature correlation research. After further exploration with linear regression, we determined that the variables "Actions_at_arrest___Combative__" and "Actions_at_arrest___Cooperative" had relatively abundant data and could be investigated concerning different age groups' criminal behavior. Therefore, after further exploring linear regression, we defined these two features and compared them linearly with varying age groups.

6.1.1 Best-Fit Line And Regression coefficient. In the definitions of "Actions_at_arrest___Combative__", and "Actions_at_arrest___Cooperative" for the age group, it can be Found in:

- Best-fit line: intercept a= 2.5755075851322244
- Regression coefficient b= [0.00864596]

We have encoded the age group categories in EDA, and Table 8 shows the numbering of each group.

Table 9

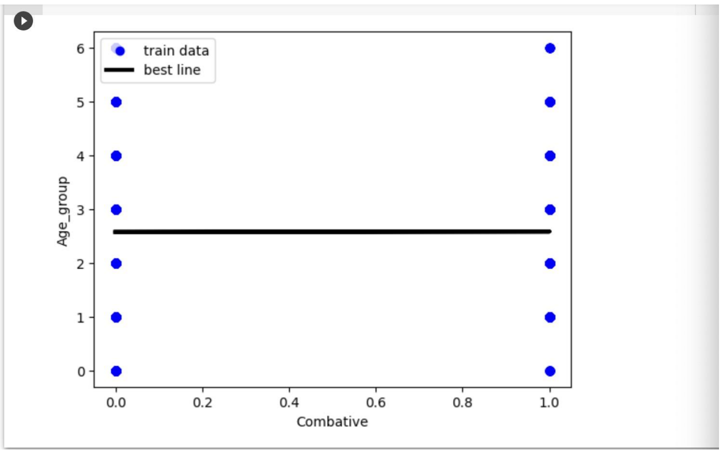
Numbering of the different age bands

Age Group	Encode

Aged 17 years and under	0
Aged 18 to 24 years	1
Aged 25 to 34 years	2
Aged 35 to 44 years	3
Aged 45 to 54 years	4
Aged 55 to 64 years	5
Aged 65 years and older	6

(1). Find the linear correlation between age into and combative as show below:

Figure 9

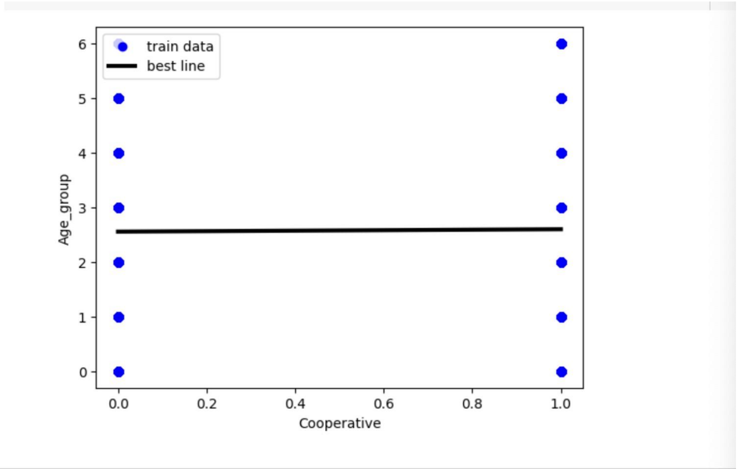


The initial predictive treatment of linear regression found that all values fell in age group 2, which is between 25-34. There is no other age group with "Actions_at_arrest___Combative__" that is of relevance to the display.

Discussion: From the figure, it can be seen that there is no strong correlation between age group and "Actions_at_arrest___Combative__" in terms of linear regression interpretation.

(2). Find the linear correlation between age into and cooperative as show below:

Figure 10

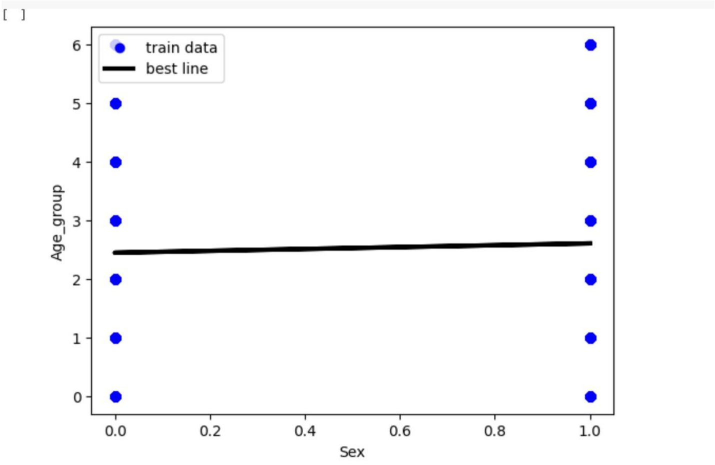


The initial predictive treatment of linear regression found that all values fell in age group 2, which is between 25-34. There is no other age group with "Actions_at_arrest___Cooperative__" that is of relevance to the display.

Discussion: From the figure, it can be seen that there is no strong correlation between age group and "Actions_at_arrest___Cooperative__" in terms of linear regression interpretation.

(3). Find the linear correlation between age into and sex as show below:

Figure 11



The initial predictive treatment of linear regression found that all values fell in age group 2, which is between 25-34. There is no other age group with "SEX" that is of relevance to the display.But compare with "Actions_at_arrest___Combative__" and "Actions_at_arrest___Cooperative__" , the lines in group2 is slightly tilted towards the direction of group3.

Discussion: From the figure, it can be seen that there is no strong correlation between age group and "SEX" in terms of linear regression interpretation but may have a slightly relatively strong correlation if compared with combative and cooperative actions.

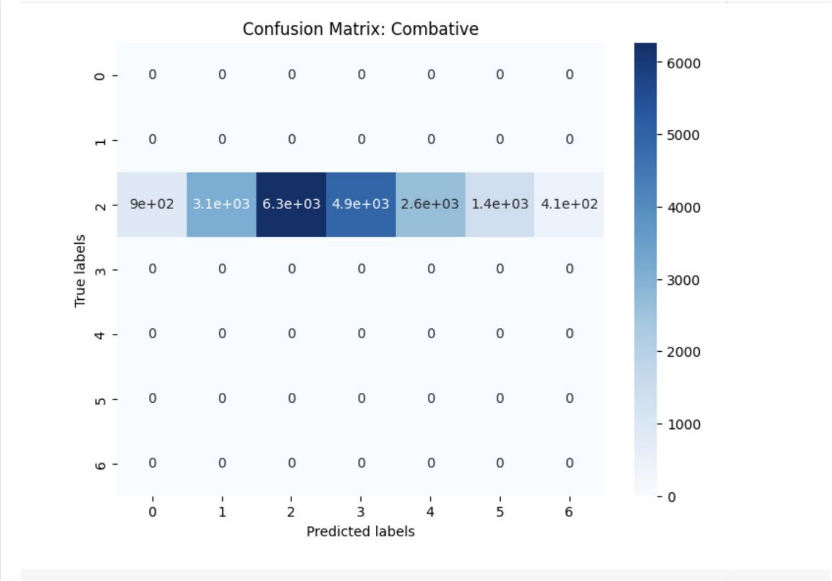
6.1.2 Overall Findings. From the exploration with linear regression, it can be seen that there is no clear correlation between either cooperative or combative behavior at the time of arrest and age. In other words, although the data for these two variables, "Actions_at_arrest___Combative__" and "Actions_at_arrest___Cooperative," is relatively abundant, there is no clear indication that they are related to age. Reasonable hypotheses can be proposed: The first possibility is that the sample size of age groups is insufficient. Cooperative and combative behavior at the time of arrest may be related to different age groups, but the sample size of age groups displayed in the dataset is insufficient and, therefore, cannot be explored in depth through linear regression. A clear indication of this is that during the EDA stage, we found that the proportion of the population under 18 was less than 5%, which also means that it is difficult to obtain specific data and conclusions on the behavior of juveniles when arrested from this dataset. The second possibility is that the sample size of cooperative and combative behavior at the time of arrest is insufficient. Although these two variables have less uncertain and missing values than others, the sample size is inadequate to support linear regression studies. Therefore, the next step is to conduct logistic regression studies based on the feedback from linear regression to verify whether there is indeed no particularly significant connection between behavior at the time of arrest and age groups.

6.2 Logistics Regression Prediction

We will use logistic regression to analyze the correlation between age group and gender, "Actions_at_arrest___Combative__" and "Actions_at_arrest___Cooperative." Logistic regression can help us explore the non-linear relationship between age and criminal behavior. Compared to linear regression, logistic regression is more suitable for dealing with cases where the dependent variable is binary and can better capture the complex relationship between age and criminal behavior. In this analysis, we will better clarify the positive, negative, or no correlation between age groups and criminal behavior.

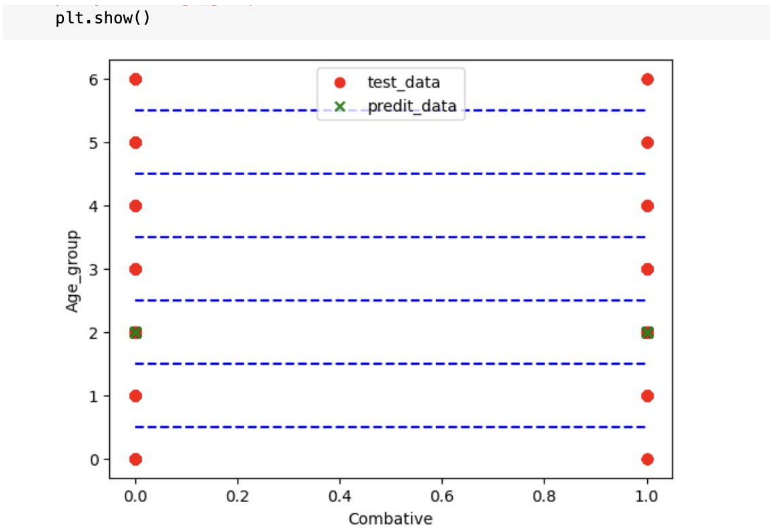
6.2.1 Using logistic regression prediction to determine whether there is an association between different age groups and combative action at the time of arrest. The final result using Heatmap shows as below:

Figure 12



This heatmap shows that the predicted values are primarily concentrated in group 2, corresponding to the age range of 25-34. Usually, if there is a correlation between age groups and combative action, the predicted values should be found in each age group, with varying degrees of density and strength. However, in the case of logistic regression, the predicted values only appear in age group 2. It suggests that there is currently no evidence of a correlation between age and combative action in terms of logistic regression prediction. Further depiction of logistic regression prediction is provided through scatter plots and line charts as follows.

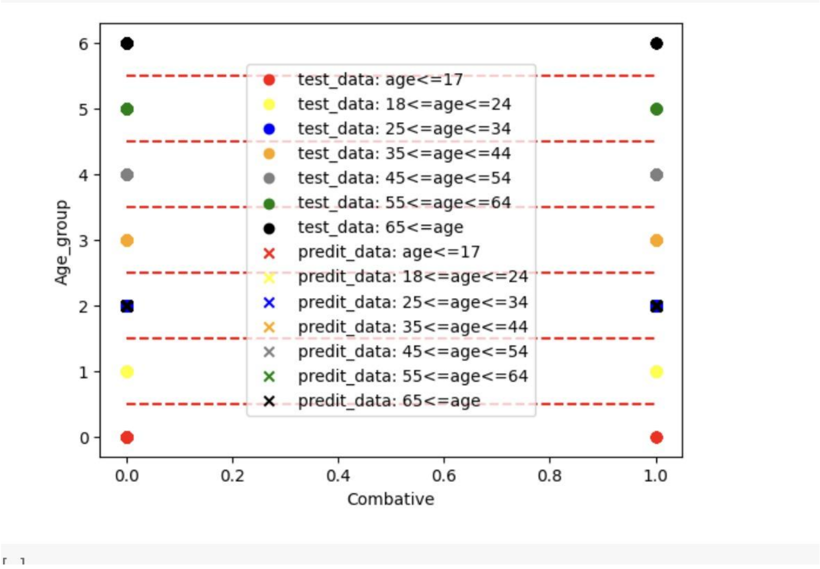
Figure 13



```
plt.show()
```

The symbol "o" represents the original data, while the symbol "x" represents the predicted values.

Figure 14

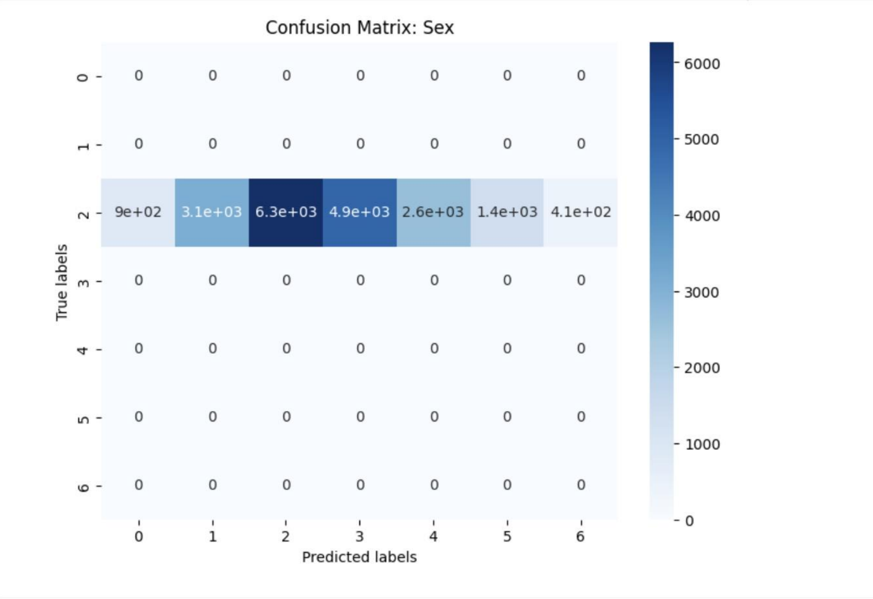


It demonstrates similar results to those presented in the heatmap above. The predicted values "X" are primarily concentrated within group 2, rendering it difficult for the dataset to elucidate the correlation between combative action and other age groups.

Discussion: From the displayed graph, it can be seen that there is no clear relationship between combative action at arrest and age group. The resistive behavior of arrestees appears to be random and sporadic across different age groups. However, it is impossible to conclude that there is no clear relationship between combative action at arrest and age group, as this dataset contains a significant amount of missing and uncertain values; thus, the information we can obtain is limited. It can only be said that under the existing dataset, it is difficult to explore the relationship between age group and combative action using logistic regression.

6.2.2 Using logistic regression to determine whether there is an association between different age groups and gender at the time of arrest. The final result using Heatmap shows as below:

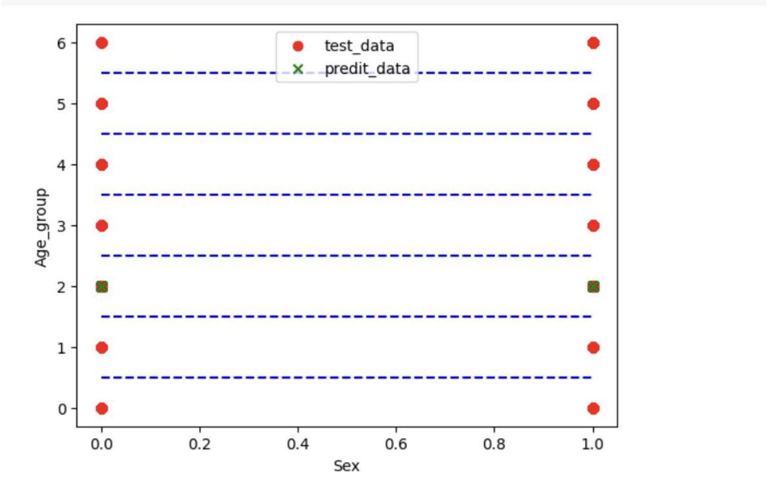
Figure 15



The presented heatmap displays the distribution of logistic regression predicted values, primarily clustered in group 2, indicating the age range of 25-34. If there exists a correlation between age groups and gender, the predicted values should be present across all age groups that include gender, albeit with differing intensities. However, the logistic regression model yields predicted values solely within the second age group. Nevertheless, it also exhibits a slight correlation with age group 3, which is statistically slightly strong compared to combative and cooperative actions.

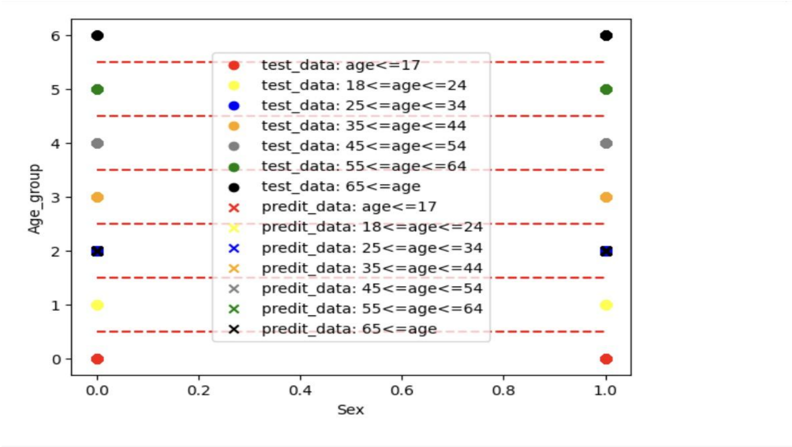
This observation suggests there is currently insufficient statistical evidence to infer a correlation between age and cooperative action as predicted by logistic regression. Further data support is needed to clarify this relationship. Further depiction of logistic regression prediction is provided as follows:

Figure 16



The symbol "o" represents the original data, while the symbol "x" represents the predicted values.

Figure 17



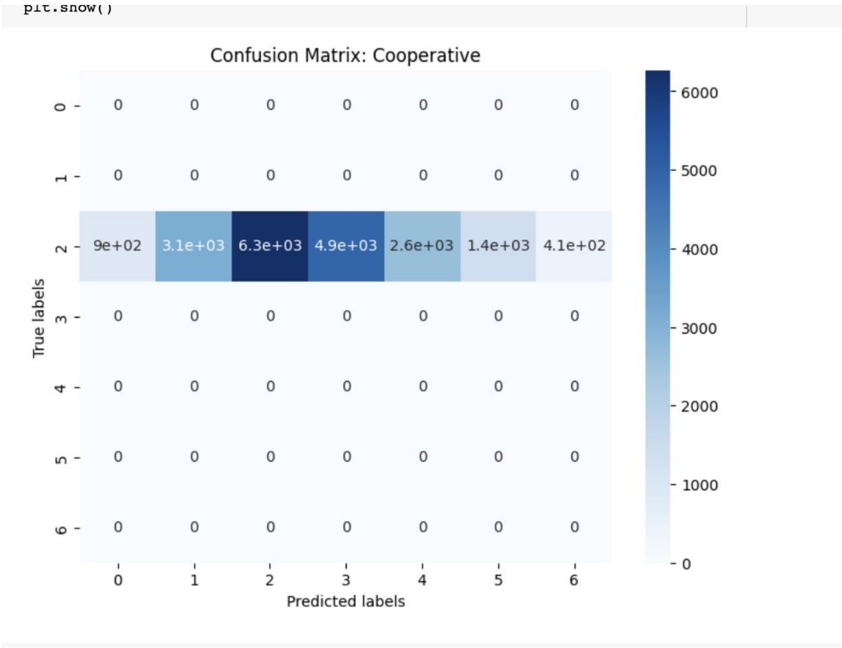
It demonstrates similar results to those presented in the heatmap above. The predicted values "X" are primarily concentrated within group 2, rendering it difficult for the dataset to elucidate the correlation between cooperative action and other age groups.

Discussion: The displayed chart shows that gender and age group are difficult to define as strongly correlated in the logistic regression analysis. One reason that needs to be considered is that in this dataset, the likelihood of women being arrested and searched is much lower than that of men. In each age group, the number of arrested women is minimal. It is precise that it is challenging to draw practical conclusions when analyzing the correlation between age and gender. Unless more female data appears in this dataset, logistic regression is unlikely to study the correlation between gender and age effectively.

Another point that can be understood is that in the initial EDA process, women were identified as a group not easily involved in crime and arrests. We identified a clear data feature in the midterm project: "Adult males mainly dominate this dataset." If we must study the relationship between gender and crime, we have already determined in the midterm project that women are often associated with low arrest rates, while men positively correlate with arrest rates. It can also explain why the correlation between gender and age group is difficult to reflect in the analysis of logistic regression - women as a gender themselves have little to do with crime and criminal age.

6.2.3 Using logistic regression to determine whether there is an association between different age groups and cooperative action at the time of arrest. The final result using Heatmap shows as below:

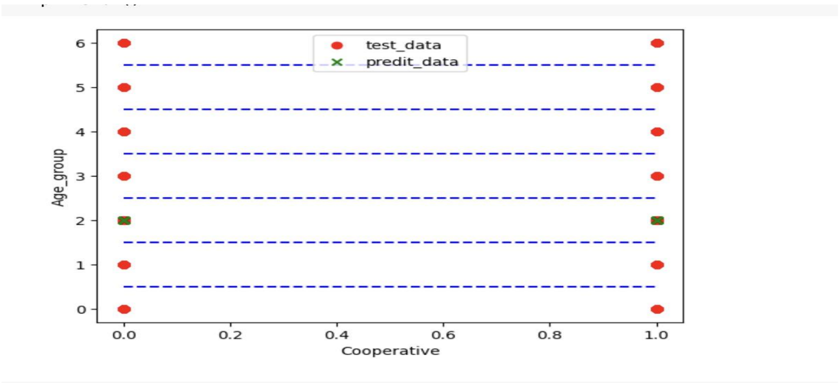
Figure 18



```
plt.show()
```

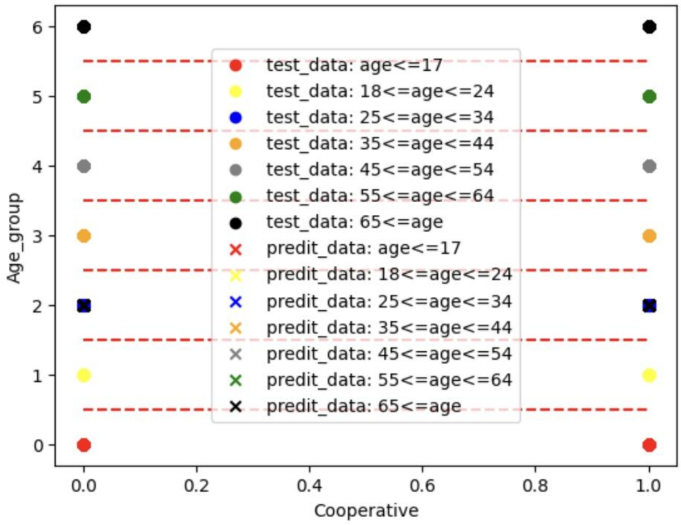
The presented heatmap displays the distribution of logistic regression predicted values, primarily clustered in group 2, indicating the age range of 25-34. If a correlation exists between age groups and cooperative action, the predicted values should be present across all age groups, albeit with differing intensities. However, the logistic regression model yields predicted values solely within the age group 2. This observation suggests that, at present, there is insufficient statistical support to infer a correlation between age and cooperative action as predicted by logistic regression. Further depiction of logistic regression prediction is provided as follows:

Figure 19



The symbol "o" represents the original data, while the symbol "x" represents the predicted values.

Figure 20



It demonstrates similar results to those presented in the heatmap above. The predicted values "X" are primarily concentrated within group 2, rendering it difficult for the dataset to elucidate the correlation between cooperative action and other age groups.

Discussion: The displayed chart reveals no clear correlation between cooperative action at the time of arrest and age group, as evidenced by the logistic regression analysis. Although there appears to be a higher likelihood of cooperative behavior among minors arrested, the overall trend of cooperative action among arrestees in different age groups is random and sporadic. Additionally, due to the considerable number of missing and uncertain values in this dataset, we cannot definitively conclude that there is no significant relationship between the cooperative action of arrestees and age group. The issue of missing data continues to hinder our logistic regression analysis, as the information available to us is severely limited. This limitation makes it difficult to explore the potential association between age group and cooperative action using logistic regression.

7. Discussion

7.1 Strength

The correlation between different age groups and crime factors are comprehensively analyzed using logistic regression, power analysis, and ANCOVA. As a result, it has the ability to study in a multidimensional manner the correlation between age groups and various crime features (particularly gender).

Following the preliminary data screening using linear regression, a further logistic regression analysis of the screened data assists in identifying the correlation and degree of influence between different age groups and behavior during the arrest. Using logistic regression, we can determine whether age is correlated with factors such as gender and behavior during arrests. Based on current data and sample size, power analysis can determine whether factors such as gender, race, and behavior at the time of arrest have sufficient statistical power. As a result, reliable tests can be conducted, and conclusions can be drawn. The critical difference between logistic regression and this approach is that it is not a binary classification algorithm and does not directly measure correlations between data using "yes" or "no." By controlling other variables (such as race), ANCOVA can be used to determine whether there are differences in crime factors between different age groups and whether other factors influence the relationship between age and specific characteristics. Combining these three methods provides a more comprehensive understanding of the relationship between different age groups and other aspects of the dataset. In this way, other potential factors influencing the outcome can be eliminated as well. These reports have enhanced data analysis for midterm projects.

7.2 Limitation

Certain limitations must be considered during analysis using the three methods. We believe these limitations are primarily due to the dataset's high degree of missing/null/uncertain data. When using various statistical methods, it is necessary to consider the constraints and assumptions of each way, but often these limitations cannot be determined precisely. For example, Since almost all the data are categorical, we have encoded the categorical data into numerical data for conducting hypothesis testing, which makes it difficult to satisfy the assumptions of hypothesis testing (normality and homogeneity of variance) completely. Therefore, it is hard to avoid experimental errors. Furthermore, we have found that using multiple analysis methods can lead to inconsistent results. For instance, in logistic regression, we were unable to draw any conclusion about the correlation between age group and gender, whereas a solid correlation between gender and age was observed in ANCOVA. Similarly, power analysis indicated a correlation between age group and ethnicity, which was not seen in logistic regression. Therefore, further exploration and explanation of the data are necessary. We cannot rule out the possibility that these different analysis methods may introduce errors and uncertainty. The best way to eliminate the errors and uncertainty is that the datasets need a lower degree of missing data to obtain more accurate results during research, which often requires larger sample sizes to get reliable results. Meanwhile, our opinion is that the most appropriate method for this dataset is to build multiple discrete data prediction models, such as decision trees, regression trees, bagging, random forests, etc., and compare metrics such as MAE, MSE, RMSE, accuracy, etc. to select the most suitable model for analysis and prediction.

8. Conclusion

This project is an extension of the data analysis conducted in the midterm, where we further explored the dataset using exploratory data analysis (EDA), power analysis, ANCOVA, and logistic regression. In the midterm project, we established that most of the arrests in the dataset were of "adult males" and that the probability of females and juveniles being arrested was relatively low. However, in the final project, there were some differences in the conclusions drawn from the three methods. Logistic regression revealed no significant relationship between the age group and gender of the arrestees, their level of cooperation during the arrest, or their resistance behavior during the arrest. However, logistic regression also indicated that the low probability of females in the arrestee population hindered the potential for a conclusion on the gender attribute. The power analysis concluded that the impact of the gender attribute was relatively small and may not be reliably detected. We confirmed through power analysis of the t-test that the reason for ANOVA's low power was highly correlated with the small number of the undefined samples, which caused bias. Due to the low statistical power of the ANOVA results, we further validated the hypothesis test conclusion by conducting ANCOVA and Tukey's test, which enhanced the credibility of the results. The ANCOVA analysis found that the racial attribute had a significant impact on the dataset, but there were still some gender influences present even after controlling for the racial attribute. It means that even after considering the effect of the racial attribute, gender still has a specific impact on the age group of arrestees. From the perspective of social and moral norms, this dataset may encourage researchers to pay more attention to the differences in the probability of crime among adult males, females, and juveniles. At the same time, more data or more appropriate sensitive tests are needed to improve the effectiveness of the test.

References

Allen, M. K., & Superle, T. (2016). Youth crime in Canada, 2014. *Juristat: Canadian Centre for Justice Statistics*, 1.

Dunbar, L. K. (2017). *Youth gangs in Canada: A review of current topics and issues*. Public Safety Canada= Sécurité publique Canada.

Seltman, H. J. (2018, July 11). *Experimental Design and Analysis*. CMU.EDU. https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

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