Exploring Factors Affecting the Age of Arrest

A midterm 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

There is a consensus that criminal behavior is often correlated with other factors in society. Age tends to have a correlation with the presence of criminal behavior and delinquency. Society often focuses its attention on juvenile delinquency as well. A study was conducted to examine the age, gender, and location of arrest etc. in Toronto through Exploratory Data Analysis (EDA) using the arrest and strip search dataset from the Toronto Police Service. Using the data provided by the EDA, it is investigated whether there are factors that influence each other regarding age and other criminal characteristics. A major focus of this paper is the mutual impression of arrest age and other characteristics of crime. Two preliminary conclusions were drawn based on Heatmap and both One-Way and Two-Way ANOVA Test: First, the two attributes of arrest month and location of arrestees in Toronto are mutually influencing each other. Second, there is evidence of a significant relationship between the arrest age and the multiple features, such as arrest months and arrest locations.

Keywords: Age, sex, race, location, criminal behavior, teenager

1. Literature Review

The Canadian government has been trying to understand the causes and trends of youth gangs and criminal behavior. Dr. Dunbar explored the causes of youth gang formation and development in his 2017 academic literature on youth gangs in Canada. In his article, he argues that there is currently no accurate measure in mainstream academia of the prevalence of youth gang involvement and activity in Canada (Dunbar, 2017). However, with the advancement of social practices and the collection and analysis of a large amount of data, Canadian society has gained some understanding of several vital groups that influence the formation of youth gangs. These key groups include Aboriginal youth, new immigrant youth, and young female groups. Researchers have found that Canadian youth gangs tend to have young adults as leaders, with a majority of underage males aged 12-17 and 18-24 members (Dunbar, 2017). There are age-oriented stable features in these organizational patterns. At the same time, female participation is much lower among youth gangs in Canada than among males. Within the same group of females, older females are far less involved than younger and underage females. Age and gender show very distinct characteristics in Canadian youth gangs.

In addition, national statistics reflect the nature of youth crime and violence in Canada and youth gang membership. As a whole, youths aged 12 to 17 and young adults aged 18 to 24 accounted for more than one-third of the defendants in gang-related crimes committed by Canadian police in 2014 (Dunbar, 2017). Therefore, it can be reasonably speculated whether age and gang criminal behavior are interconnected. Dr. Dunbar believes that a deeper understanding of specific risk factors could help develop future solutions to address the issue of youth gang involvement and gang-related activities in Canada.

Allen and Superle's study found that in most cases, police report of youth crime involved relatively minor offenses. The most common criminal offenses committed by youth were theft, mischief, and common assault (one of the less violent and severe

offenses). (Allen & Superle, 2016) Meanwhile, in 2014 statistics found that minors were charged with crimes at a much lower rate than 18-24-year-olds in Canada. Meanwhile, adults between the ages of 25 and 54 have higher odds of committing a crime. In turn, their crime rate tends to decrease as they age into old age (Allen & Superle, 2016). Again, this shows that age may strongly influence the probability of crime.

Another more notable finding is a correlation between juvenile delinquency and the location of the crime. One in ten incidents in which juveniles were charged occurred during school hours or at a supervised activity. Juveniles were more likely to be charged with violent crimes (19%) and drug crimes (27%) at school than property crimes. Possession of marijuana and common assault was the most common offenses that juveniles were accused of occurring at school (Allen & Superle, 2016). Schools and locations near schools have become locations with high juvenile crime rates.

Youth crime is one of the critical social topics. From both Dr. Dunbar's and Allen & Superle's studies, it was found that age showed some correlation with some specific offenses and crime phenomena. The Arrests and Strip Searches (RBDC-ARR-TBL-001) dataset contains specific information on many arrestees in Toronto. Therefore, three questions can be explored from this dataset: first, whether the age and crime phenomenon of arrestees in Toronto, the number one city in Canada, is consistent with Canada as a whole; second, starting from the Toronto Police Department data, one explores whether specific characteristics are highly correlated (e.g., age and location of arrest) among criminals arrested by the police; third, Further investigation is required to determine whether the presence of one attribute influences the birth of another when different characteristics are associated with it (e.g., whether the underage group is associated with a particular criminal activity). Since the dataset contains data on both adults and minors, the age of the arrestees will be the primary focus of this investigation. The purpose of this study is to investigate whether the age of arrestees and other attributes are interacting.

2. Introduction

The Arrests and Strip Searches (RBDC-ARR-TBL-001) dataset includes a variety of data categories. The attributes of age group, race, gender, arrest location, and type of crime are significant symbolic in society and are the focus of this study. The age of arrestees, the gender share of arrestees, the race share of arrestees, the probability of offenses occurring in different geographical areas, and the frequency of different types of offenses are often controversial and hot topics in social issues. People explore whether specific characteristics are highly correlated among criminals arrested by police. From there, one further explores whether one attribute's presence also influences another's birth when different characteristics are strongly related. A reduction in crime rates may result if early intervention of specific characteristics can reduce the creation of another criminal characteristic with which it is highly correlated. Further, society may also explore social morality, fairness, and respect topics. For example, whether a particular race is associated with high arrest rates. Similarly, whether people in a particular area are associated with a particular type of crime, and whether a particular age group is highly correlated with certain types of crime. People explore the differences presented by the data to determine if some specific minority groups are being mistreated in the police department. They also make reasonable claims and solutions as a result.

Based on age groups, this study will examine whether other factors in the dataset, such as gender, race, and location, strongly correlate with age at crime. In this study, exploratory analysis and preprocessing of the data will be performed with Python's visualization package. Correlations between age at crime and other attributes can be analyzed using heatmaps and both one-way and two-way ANOVA tests.

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 analyzing. The tools used in this EDA include data science libraries (NumPy, Pandas, SciPy), and visualization libraries (Matplotlib, Plotly, Seaborn).

3.1 Data Overview

(i). The first step in this study is to examine the dataset. This step involves importing NumPy, Pandas, and SciPy so that the data in the dataset can be pre-processed.

Figure 1:

```
import numpy as np
import pandas as pd
# LabelEncoder(String to int)
from sklearn.preprocessing import LabelEncoder
# Visulization
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import scipy.stats as stats
# get ANOVA table as R like output
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```
[100] df=pd.read_csv('/content/Arrests_and_Strip_Searches_(RBDC-ARR-TBL-001).csv')
      print(df)
```

```
        Arrest_Year Arrest_Month  EventID   ArrestID  PersonID  \
0              2020    July-Sept  1005907  6017884.0    326622
1              2020    July-Sept  1014562  6056669.0    326622
2              2020      Oct-Dec  1029922  6057065.0    326622
3              2021      Jan-Mar  1052190  6029059.0    327535
4              2021      Jan-Mar  1015512  6040372.0    327535
...             ...          ...      ...        ...       ...
```

(ii). From Figure 2 we can see that 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. Next, we look at the dataset to gain more insight about it.

Figure 2:

```
[103] #shape of dataset
      df.shape

      (65276, 25)
```

(iii). To determine the exact form of the data, use the .head and tail methods of Pandas.

Figure 3:

```
#First Five Dataset
df.head()
```

	Arrest_Year	Arrest_Month	EventID	ArrestID	PersonID	Perceived_Race	Sex	Age_group__at_arrest_	Youth_at_arrest__under_18_years
0	2020	July-Sept	1005907	6017884.0	326622	White	M	Aged 35 to 44 years	Not a youth
1	2020	July-Sept	1014562	6056669.0	326622	White	M	Aged 35 to 44 years	Not a youth
2	2020	Oct-Dec	1029922	6057065.0	326622	Unknown or Legacy	M	Aged 35 to 44 years	Not a youth
3	2021	Jan-Mar	1052190	6029059.0	327535	Black	M	Aged 25 to 34 years	Not a youth
4	2021	Jan-Mar	1015512	6040372.0	327535	South Asian	M	Aged 25 to 34 years	Not a youth

5 rows × 25 columns

Figure 4:

```
#Last Five Rows of Dataset
df.tail()
```

	Arrest_Year	Arrest_Month	EventID	ArrestID	PersonID	Perceived_Race	Sex	Age_group__at_arrest_	Youth_at_arrest__under_18_yea
65271	2021	Oct-Dec	1055609	6044336.0	316123	Indigenous	F	Aged 25 to 34 years	Not a you
65272	2021	Oct-Dec	1032758	6031692.0	307736	South Asian	M	Aged 35 to 44 years	Not a you
65273	2021	Oct-Dec	1021067	6064396.0	324057	White	F	Aged 45 to 54 years	Not a you
65274	2021	Oct-Dec	1008998	6008662.0	331870	Unknown or Legacy	M	Aged 17 years and under	Youth (aged 17 years and und
65275	2021	Oct-Dec	1033395	6032145.0	310583	Latino	M	Aged 18 to 24 years	Not a you

5 rows × 25 columns

(iv). Identify the Null Values.

Datasets are related to arrest and strip search, in which nine features contains 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.

Figure 5:

```
#identify the null values
df.isnull().sum()
```

```
Arrest_Year                          0
Arrest_Month                         0
EventID                              0
ArrestID                           469
PersonID                             0
Perceived_Race                       4
Sex                                  0
Age_group__at_arrest_               24
Youth_at_arrest__under_18_years      0
ArrestLocDiv                         0
StripSearch                          0
Booked                               0
Occurrence_Category                165
Actions_at_arrest___Concealed_i      0
Actions_at_arrest___Combative__      0
Actions_at_arrest___Resisted__d      0
Actions_at_arrest___Mental_inst      0
Actions_at_arrest___Assaulted_o      0
Actions_at_arrest___Cooperative      0
SearchReason_CauseInjury         57475
SearchReason_AssistEscape        57475
SearchReason_PossessWeapons      57475
SearchReason_PossessEvidence     57475
ItemsFound                       57475
ObjectId                             0
dtype: int64
```

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.

(v). The current dataset information is as Figure 6. It can see that the dataset contains mixture of categorical and numerical variables.

Figure 6:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 64615 entries, 0 to 65275
Data columns (total 20 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Arrest_Year                       64615 non-null  int64
 1   Arrest_Month                      64615 non-null  object
 2   EventID                           64615 non-null  int64
 3   ArrestID                          64615 non-null  float64
 4   PersonID                          64615 non-null  int64
 5   Perceived_Race                    64615 non-null  object
 6   Sex                               64615 non-null  object
 7   Age_group__at_arrest_             64615 non-null  object
 8   Youth_at_arrest__under_18_years   64615 non-null  object
 9   ArrestLocDiv                      64615 non-null  object
 10  StripSearch                       64615 non-null  int64
 11  Booked                            64615 non-null  int64
 12  Occurrence_Category               64615 non-null  object
 13  Actions_at_arrest___Concealed_i   64615 non-null  int64
 14  Actions_at_arrest___Combative__   64615 non-null  int64
 15  Actions_at_arrest___Resisted__d   64615 non-null  int64
 16  Actions_at_arrest___Mental_inst   64615 non-null  int64
 17  Actions_at_arrest___Assaulted_o   64615 non-null  int64
 18  Actions_at_arrest___Cooperative   64615 non-null  int64
 19  ObjectId                          64615 non-null  int64
dtypes: float64(1), int64(12), object(7)
memory usage: 10.4+ MB
```

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.

(i) Check Outliers.

On closer inspection, there is no large outliers in this dataset for numerical variables.

Figure 7:

	Arrest_Year	EventID	ArrestID	PersonID	StripSearch	Booked
count	64615.0	64615.0	64615.0	64615.0	64615.0	64615.0
mean	2021.0	1029991.0	6032400.0	318602.0	0.0	1.0
std	0.0	17319.0	18706.0	10814.0	0.0	0.0
min	2020.0	1000000.0	6000000.0	300000.0	0.0	0.0
25%	2020.0	1014988.0	6016200.0	309220.0	0.0	0.0
50%	2021.0	1029987.0	6032402.0	318594.0	0.0	1.0
75%	2021.0	1044998.0	6048596.0	327922.0	0.0	1.0
max	2021.0	1060002.0	6064804.0	337346.0	1.0	1.0

	Actions_at_arrest___Concealed_i	Actions_at_arrest___Combative__
count	64615.0	64615.0
mean	0.0	0.0
std	0.0	0.0
min	0.0	0.0
25%	0.0	0.0
50%	0.0	0.0
75%	0.0	0.0
max	1.0	1.0

	Actions_at_arrest___Resisted__d	Actions_at_arrest___Mental_inst
count	64615.0	64615.0
mean	0.0	0.0
std	0.0	0.0
min	0.0	0.0
25%	0.0	0.0
50%	0.0	0.0
75%	0.0	0.0
max	1.0	1.0

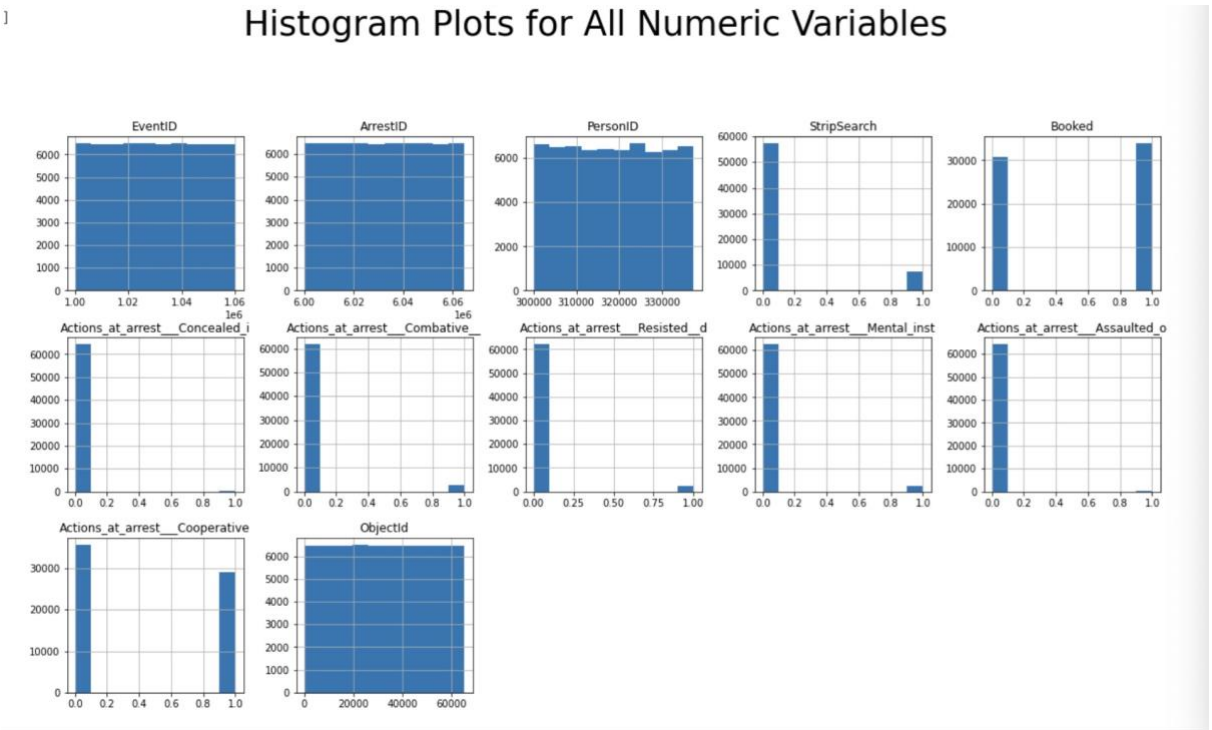
	Actions_at_arrest___Assaulted_o	Actions_at_arrest___Cooperative
count	64615.0	64615.0
mean	0.0	0.0
std	0.0	0.0
min	0.0	0.0
25%	0.0	0.0
50%	0.0	0.0
75%	0.0	1.0
max	1.0	1.0

(ii). Histogram for All Numerical Variables

According to the histogram in Figure 7, it was clear that the data for each attribute did not appear to be normally distributed.

Figure 8:

]



3.3 Explore Categorical Variables.

We checked the labels in the variable. Through the Table 1 we can see that these categories are small in length, and we have 64615 data, so we don't need to drop them.

Table 1:

Categorical_Variables	Length
Arrest_Moth	4
Perceived_Race	8
Sex	3
Age_group_at_tarrest_	9
Youth_at_arrest_under_18	3
Arrest_Location	18
Occupancy_Category	31

3.4 Univariate Analysis - Explore Target Variable.

Since the target variable we are studying is Age_group_at_tarrest_, we would like to further explore this target variable. From the below Figure 9 and Figure 10 we can see that the number of unique values in Age_group__at_arrest_variable is 9.

Figure 9:

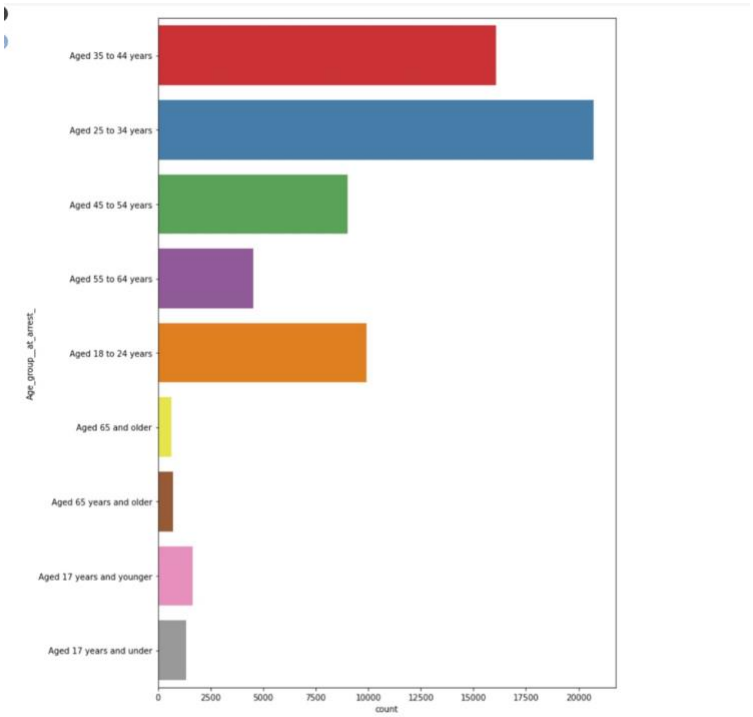
```
[50] df['Age_group__at_arrest_'].value_counts()

     Aged 25 to 34 years           20725
     Aged 35 to 44 years           16072
     Aged 18 to 24 years            9934
     Aged 45 to 54 years            9003
     Aged 55 to 64 years            4553
     Aged 17 years and younger      1663
     Aged 17 years and under        1349
     Aged 65 years and older         693
     Aged 65 and older               623
     Name: Age_group__at_arrest_, dtype: int64
```

```
[51] df['Age_group__at_arrest_'].value_counts()/len(df)

     Aged 25 to 34 years           0.320746
     Aged 35 to 44 years           0.248735
     Aged 18 to 24 years           0.153741
     Aged 45 to 54 years           0.139333
     Aged 55 to 64 years           0.070464
     Aged 17 years and younger     0.025737
     Aged 17 years and under       0.020878
     Aged 65 years and older       0.010725
     Aged 65 and older             0.009642
     Name: Age_group__at_arrest_, dtype: float64
```

Figure 10:

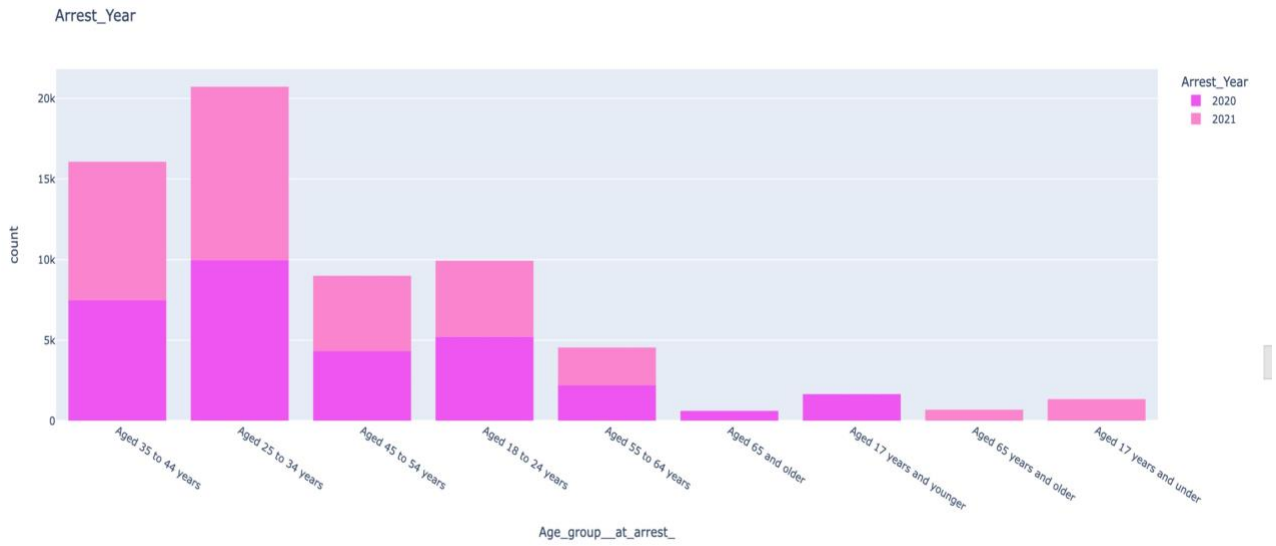


Among those arrested and strip searched, young adults between the ages of 25 and 34 had the highest percentage, with 32.1%. Adults between the ages of 35 and 44 had the second highest percentage, with 24.9%. Young adults aged 18-24 ranked third. Middle-aged individuals aged 45-54 rank fourth with 13.9%. 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 shows 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:

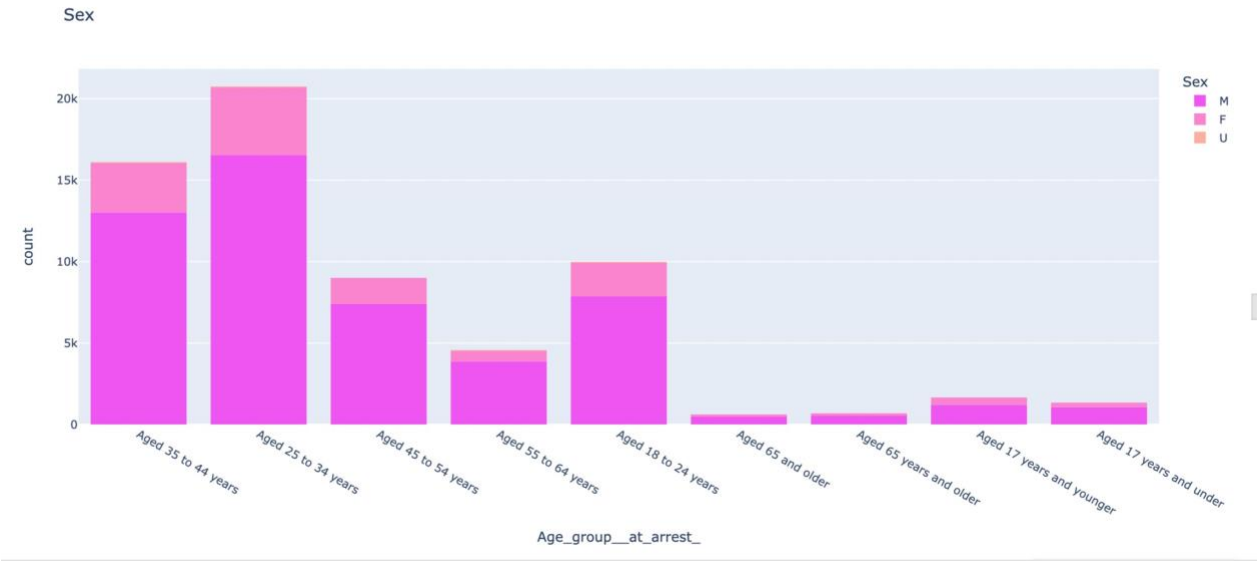
Figure 11:



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.

3.4.2 Sex in Age group Overview

Figure 12:



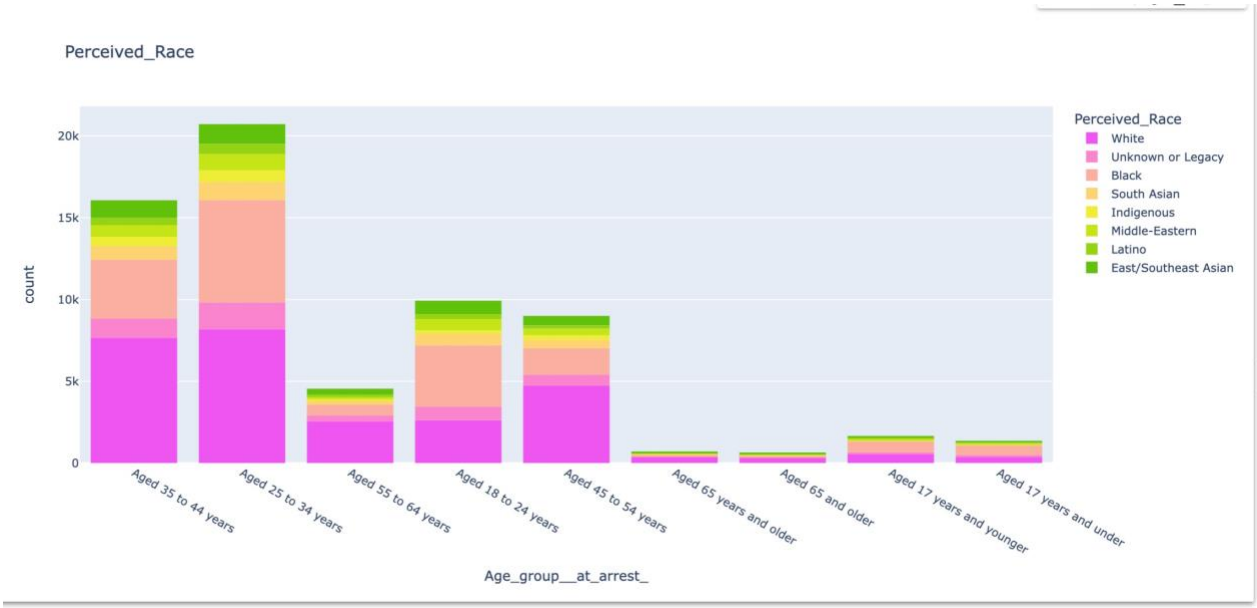
According to chart visualization of the data for this attribute of gender, males constitute most suspects arrested and strip searched. As a result, 80.7% of the population is male. In comparison, the percentage of females is 19.3%.

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.

3.4.3 Race in Age Group Overview

Figure 13:



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.

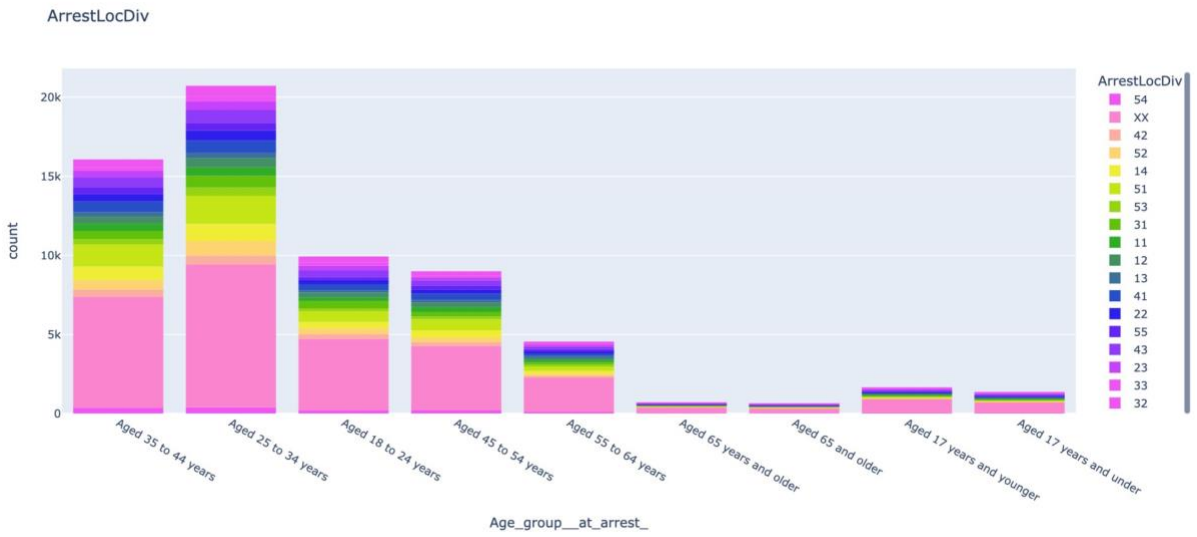
At the same time, racial differences emerge in the number of arrests of youth in the 18-24 age group. The number of arrests for blacks was higher in total, while the number of arrests for whites was lower.

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 the 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%.

3.4.4 Location in Age Group Overview

Figure 14:

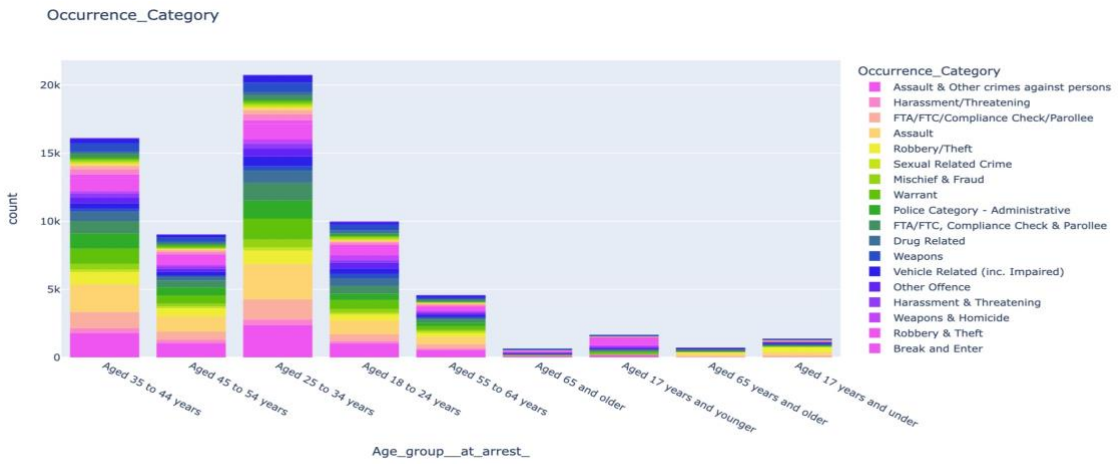


In the comparison against area and age group, it can be found that the age group of 18-54 years old is most likely to be arrested in District 54.

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%.

3.4.5 Occurrence Category in Age group Overview:

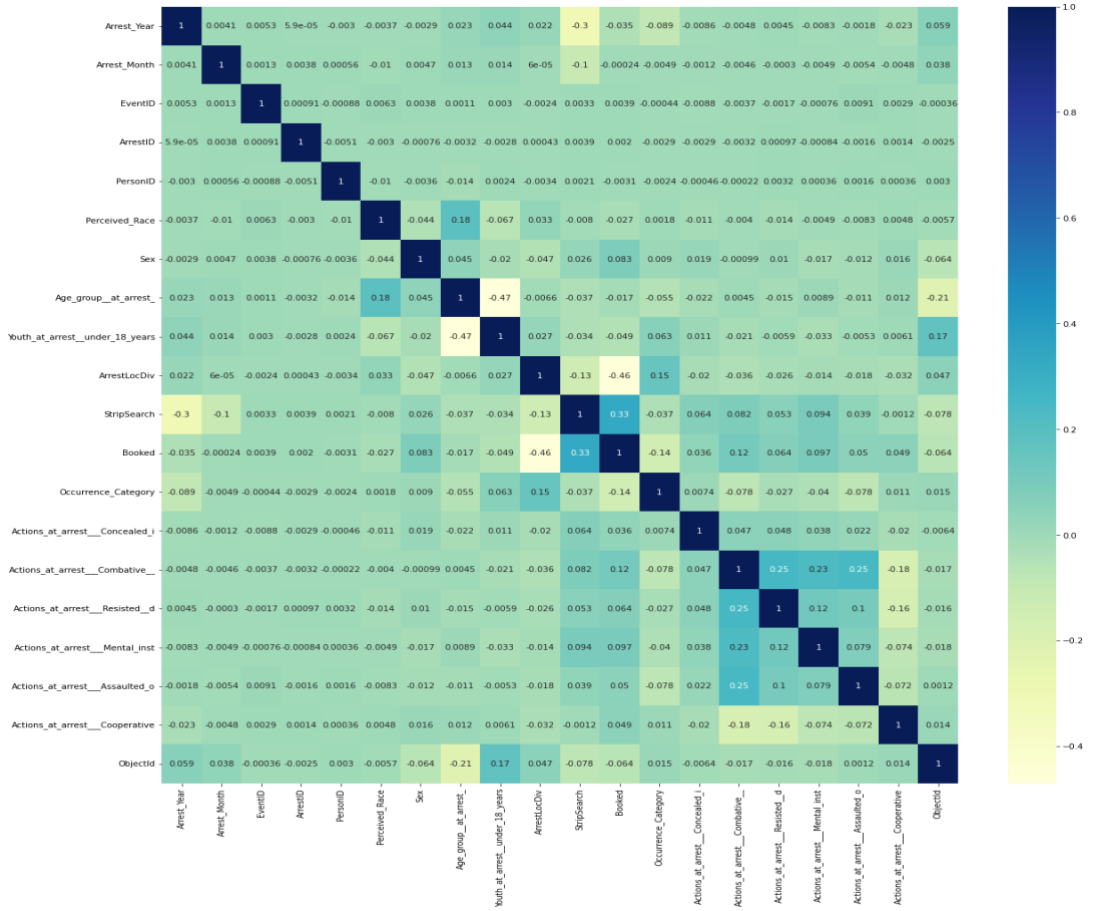
Figure 15



3.5 Multivariate Analysis

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

Figure ***16***



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. Hypothesis

(i) H0: There is **no** difference in the mean of another attribute between the two levels of one certain attribute variable.

H1: There is **a** difference in the mean of another attribute between the two levels of one certain attribute variable.

(ii) Specific basic Attribute：**Age group**

Deeply the null hypothesis can be explained:

Ho: 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.

5. ANOVA & Tukey's Test

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 dependent variable and set all features as independent variable 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__ had 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.

Figure 17

```
                        sum_sq       df          F        PR(>F)
C(Arrest_Year)       66.301592      1.0  33.331895  7.805229e-09
Residual         128523.887157  64613.0        NaN           NaN
                        sum_sq       df          F        PR(>F)
C(Arrest_Month)      74.215452      3.0  12.437195  3.973529e-08
Residual         128515.973297  64611.0        NaN           NaN
                          sum_sq       df           F  PR(>F)
C(Perceived_Race)    5340.498591      7.0  399.924033     0.0
Residual           123249.690158  64607.0         NaN     NaN
                  sum_sq       df         F        PR(>F)
C(Sex)       259.422198      2.0  65.30697  4.636308e-29
Residual  128330.766551  64612.0       NaN           NaN
                        sum_sq       df         F        PR(>F)
C(ArrestLocDiv)      409.422297     17.0  12.13702  1.755158e-34
Residual          128180.766452  64597.0       NaN           NaN
                        sum_sq       df          F        PR(>F)
C(StripSearch)       178.586017      1.0  89.859313  2.640021e-21
Residual          128411.602731  64613.0        NaN           NaN
                    sum_sq       df          F    PR(>F)
C(Booked)        35.158274      1.0  17.670888  0.000026
Residual    128555.030475  64613.0        NaN       NaN
                               sum_sq       df          F          PR(>F)
C(Occurrence_Category)     2318.202169     30.0  39.522825  4.452455e-228
Residual                 126271.986580  64584.0        NaN            NaN
                                          sum_sq       df           F  \
C(Actions_at_arrest___Concealed_i)       61.337231      1.0  30.834964
Residual                               128528.851518  64613.0        NaN

                                           PR(>F)
C(Actions_at_arrest___Concealed_i)  2.820319e-08
Residual                                   NaN
                                          sum_sq       df         F    PR(>F)
C(Actions_at_arrest___Combative__)       2.622122      1.0  1.31757  0.251033
Residual                               128587.566627  64613.0       NaN       NaN
                                          sum_sq       df          F  \
C(Actions_at_arrest___Resisted__d)       28.053427      1.0  14.099144

                                    PR(>F)
C(Actions_at_arrest___Resisted__d)  0.000174
Residual                                 NaN
                                          sum_sq       df         F    PR(>F)
C(Actions_at_arrest___Mental_inst)       10.163838      1.0  5.10745  0.023827
Residual                               128580.024911  64613.0       NaN       NaN
                                          sum_sq       df         F    PR(>F)
C(Actions_at_arrest___Assaulted_o)       16.216177      1.0  8.149206  0.004309
Residual                               128573.972572  64613.0       NaN       NaN
                                          sum_sq       df         F    PR(>F)
C(Actions_at_arrest___Cooperative)       17.680074      1.0  8.884968  0.002876
Residual                               128572.508674  64613.0       NaN       NaN
                                        sum_sq       df           F  PR(>F)
C(Youth_at_arrest__under_18_years)  29387.324213      2.0  9570.156068     0.0
Residual                            99202.864536  64612.0          NaN     NaN
```

We take the Sex feature as an example. The **hypothesis can be:**

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.

As a result 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.

- The sum of squares of the gender variable is 259.422198.
- The degrees of freedom (df) of the gender variable is 2, which is the number of levels of the gender variable minus 2.
- The F-statistic (F) for the gender variable is 65.30697. 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.
- The p-value (PR(>F)) for the gender variable is 4.636308e-29, which is 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 age group between the two levels of the Sex variable.

5.2 Two-Way ANOVA

We also selected Arrest_Month and Location variables and applied **Two-Way ANOVA** to determine combinations of factors that may be statistically significant.

Figure 18

	sum_sq	df	F	PR(>F)
C(Arrest_Month)	72.590104	3.0	12.208206	5.551819e-08
C(ArrestLocDiv)	407.796949	17.0	12.102936	2.294821e-34
C(Arrest_Month):C(ArrestLocDiv)	183.758539	51.0	1.817913	3.228460e-04
Residual	127924.417809	64543.0	NaN	NaN

From the table, we can see that the p-values for Arrest_Month and ArrestLocDiv 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 sufficient evidence to say there is significant interaction effect between Arrest_Year and ArrestLocDiv.

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.

5.3 Post-Hoc Analysis (Tukey's Test)

We used the pairwise_tukeyhsd method to generalize the Tukey's test result for the two features (Arrest_Month, ArrestLocDiv) that we used in Two-Way ANOVA. We selected the Arrest_Month results for illustration, the result for the Arrest Month is showing below:

Figure 19

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================
group1 group2 meandiff p-adj   lower   upper  reject
---------------------------------------------------
     0      1  -0.0576 0.0014  -0.098 -0.0172   True
     0      2   0.0221 0.5023 -0.0186  0.0629  False
     0      3   0.0245 0.4267 -0.0169   0.066  False
     1      2   0.0797  0.001  0.0404   0.119   True
     1      3   0.0821  0.001  0.0421  0.1221   True
     2      3   0.0024    0.9  -0.038  0.0428  False
---------------------------------------------------
```

From the table we can see that there is statistically significant difference between the means of *groups* 0 *and* 1*, groups* 1*and* 2*,* and groups 1 and 3, but not a statistically significant difference between the means of other groups.

6. Discussion

6.1 Strength

This study utilized three methods for exploring data: EDA and ANOVA Test. Combining the three methods allowed us to identify some characteristics of arrestees in Toronto. The number of arrests of males is significantly higher than the number of arrests of females in Toronto. Particularly, the age group of arrestees did not appear to be a separate variable. According to the study, it interacts with gender, behavior at the time of arrest, and the reason for arrest.

6.2 Limitation

There are limitations to this study, including the possibility of exploring the interplay between age and gender, behavior at the arrest site, and reason for arrest. However, we do not provide any further explanations to two questions: whether age correlates positively or negatively with other specific factors, and whether age correlates positively or negatively with other specific characteristics cannot be determined. As well, the number of adolescents present in this dataset is relatively low, accounting for less than 5% of the total sample size. As a result, the findings from this study cannot be considered "precise" and can only be considered indicative of a general trend.

7. Conclusion

To achieve the objective of exploring *the low age of crime* and figuring out the features that have a significant impact on arrest age, we used two types of methods in general: one was exploratory data analysis (EDA) method, another one used ANOVA test. In EDA, we conducted some descriptive statistical analyses (such as distribution check, null value check, normality check, and correlation analysis) via the usage of graphical techniques, including pie chart, tables, heatmaps, then followed by the hypothesis testings (ANOVA test and OLS). The findings of this study indicate that the age attribute is not a stand-alone variable. A correlation was found between its presence and gender, behavior at the time of arrest, and the reason for the arrest. Considering that this ANOVA analysis consisted only of a simple correlation analysis, it is difficult to draw

further conclusions. However, researchers can build on this analysis to further explore how minor offenders are specifically associated with their gender, behavior at arrest, the reason for being searched, and how strong the association is. Additionally, when investigating the behavioral factors associated with arrest scenes for arrestees under 18, the factors were associated mainly with less harmful action such as escape, concealing objects, or well-intentioned behaviors, such as cooperation. It may also enable researchers to explore whether there are better ways to protect minors and correct their errors when crimes are committed against.

References

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

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

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

Howard J, S. (2018). *Experimental Design and Analysis*.