Write-up

1. Introduction

With the economic downturn and inflation brought on by the pandemic, Toronto's streets are beginning to see large numbers of homeless people and the number of indiscriminate attacks is gradually increasing. In the first week of February alone, there were three cases of indiscriminate assault in the Toronto area. One of the victims died, while another was a student at the University of Toronto. The research team hopes to find out the impact of gender and age group on suspects (those searched), as well as the distribution area and gender of mentally unstable suspects (usually the perpetrators of indiscriminate attacks are mentally unstable people) through the strip search dataset released by the Toronto Police Service. The method we would use is exploratory data analysis, an approach to analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. The specific hypotheses related to the research questions will be tested by ANOVA, which is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups difference by comparing each group's means and includes spreading the variance into diverse sources.

2. Exploratory Data Analysis (EDA)

2.1 Descriptive Analysis

A total of 65,276 cases were included in the dataset from 2020 to 2021, of which 2,179 were mentally unstable or suicidal suspects. The number of cases in which the suspect was a female mentally unstable person ranged from a low of 1 to a high of 105, while the number of males

mentally unstable suspects ranged from a low of 6 to a high of 254. The average number of female cases was 13.9 and the average number of male cases was 46.4.

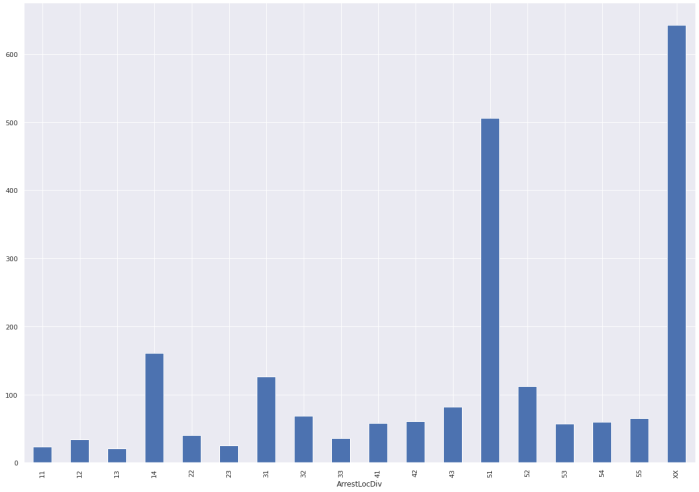

Figure 2-1

Block 51 had the highest number of mentally unstable persons of all cases recorded in the search, more than double that of Block 14, which ranked second. In addition, the bar chart also told us that the police have a large number of cases involving mentally unstable suspects where the location was not recorded or where it occurred is vague.

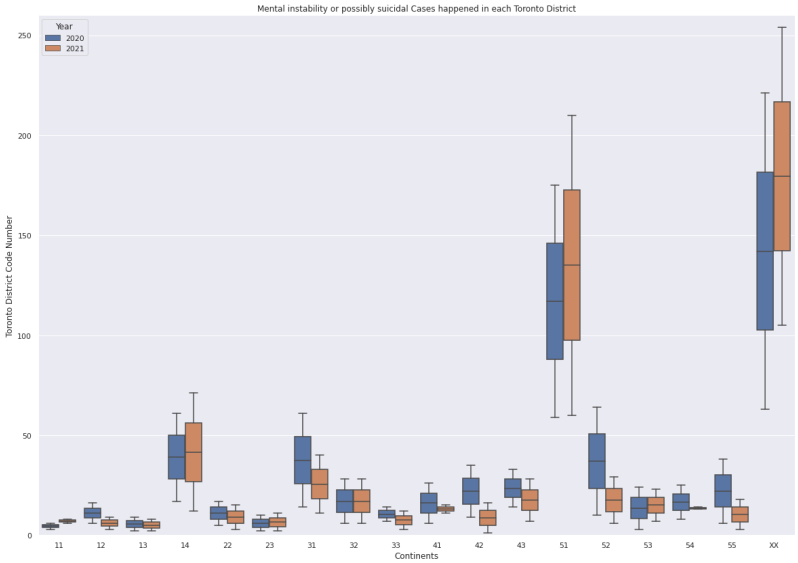

Figure 2-2

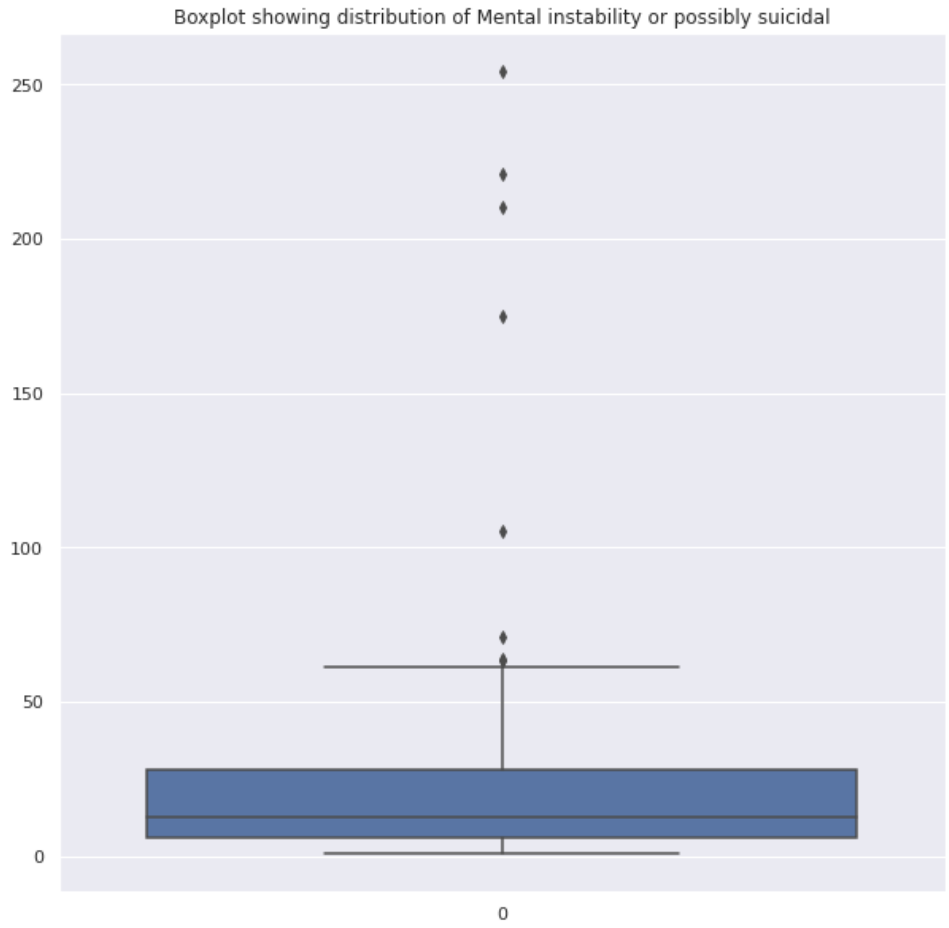

Figure 2-3

Box plots show the distribution of quantitative data to facilitate comparisons between variables or between levels of categorical variables. Boxes show the quartiles of the data set, while whisker extensions show the remainder of the distribution, except for points identified as "outliers" using the interquartile function method. The box plots show cases involving mentally unstable or suicidal people in 17 Toronto blocks (excluding unknown location XX) in 2020 versus 2021. The advantage of the box plot is that the observer can visualize the range between the median and quartiles of the corresponding data, as well as its outliers. The box plot shows that the median number of cases occurring in each neighborhood is relatively stable and stays below 50, except for Block 51. If we consider the cases involving mentally unstable people in

2020 and 2021 as a whole, it has a small difference between the upper and lower quartiles, and the larger outliers should be brought about by Block 51 and unknown locations.

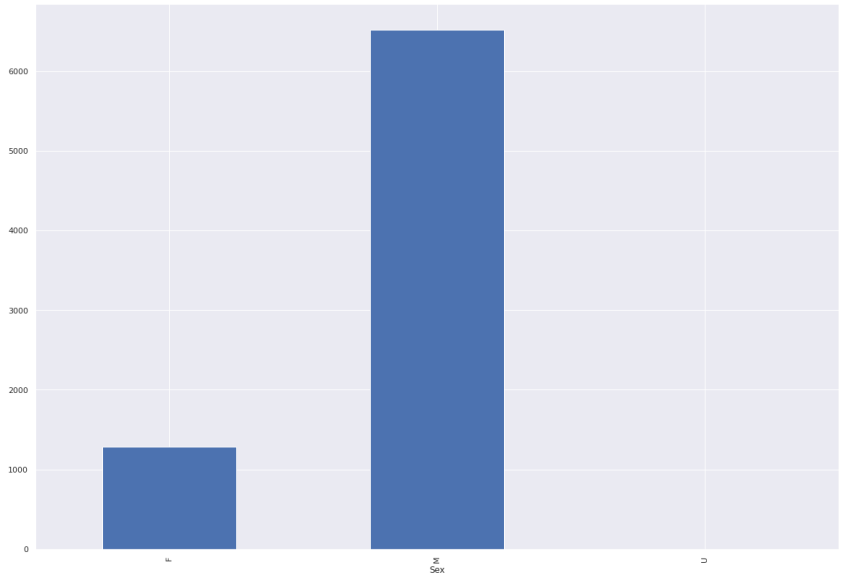

Figure 2-4 Strip Search on different sex

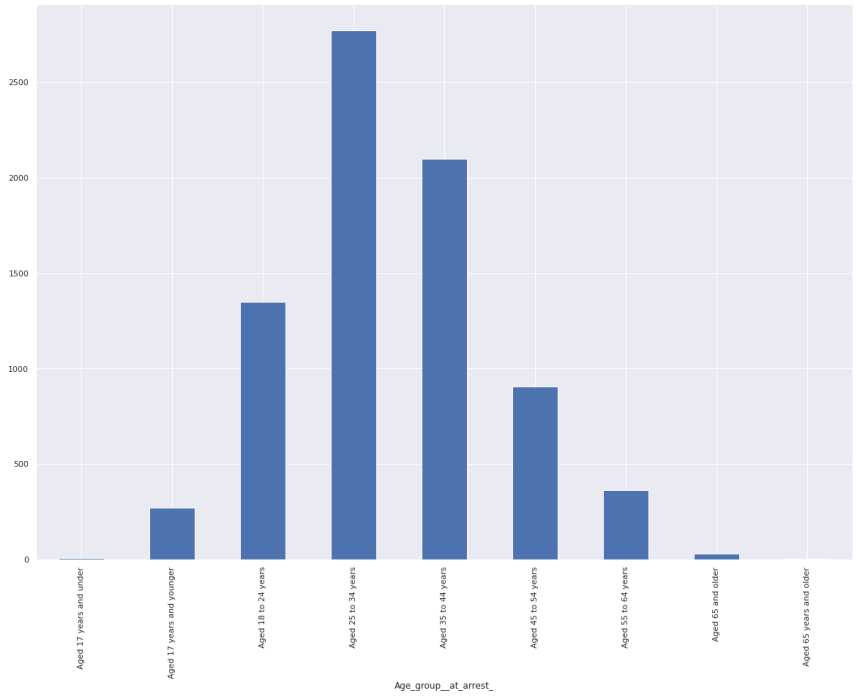

Figure 2-5 Strip Search on different age groups

Besides, in the 65276 cases, there were 7801 cases with a strip search, including 1283 for females and 6518 for males. For different age groups, 25-34 and 35-44 have the most cases of a strip search, while teenagers and elders have the least cases.

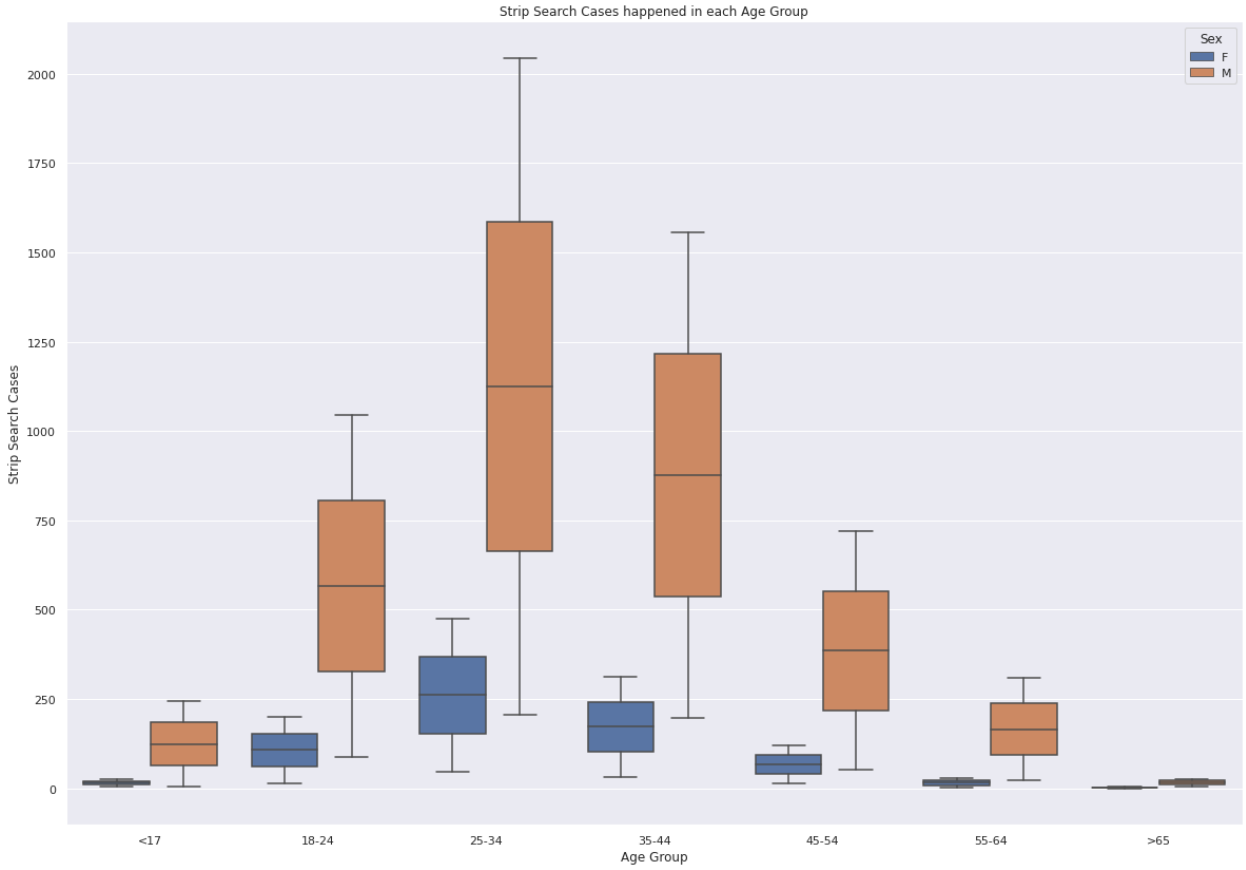

Figure 2-6

2.2 T-test

The results showed that there were more males (M=46.4, SD=63.4) than females (M=13.9, SD=22.1) in all police-recorded cases involving mental instability and suicidal ideation. With an alpha value determined to be 0.05, this is a statistically significant difference as the p-value (0.006) is less than 0.05, 95% CI [9.97, 55.08]. Therefore, we can reject the null hypothesis that

there is no difference between men and women in all police-recorded cases involving mental instability and suicidal tendencies.

Then, we do unpaired two-sample T-tests to see whether Strip Search varies in different groups of a variable. Our hypotheses are H0: There is no significant difference in strip search cases between males and females. H1: There is a significant difference in strip search cases between males and females. As a result, we can see that the p-value is equal to 0.044, less than 0.05, so we can reject the null hypothesis, and conclude that there is a significant difference in strip search between males and females.

3. Method

3.1 Dataset description

The primary source of the research is a dataset from Arrests and Strip Searches by Toronto Police Service; this dataset contains information on all arrests and strip searches. The dataset also includes information on whether a person was taken to a police station within 24 hours of their arrest. However, some records may indicate a strip search but not a booking due to issues with the booking template. In these cases, it should be assumed that a booking did take place. This database was initially collected and created by Toronto Police Service. The 'Arrests_and_Strip_Searches_(RBDC-ARR-TBL-001).csv' contains the records of 65,276 cases from 2020 to 2021.

In this research, our main variables are:

strip search: a type of search conducted by a police officer, wherein they remove some or all of a person's clothing and visually inspect their body.

Age group: The age of the person who was arrested or strip-searched is their age at the time of the arrest, as reported to the arresting officer.

Year: 2020 and 2021

Sex: Male (M) and Female (F)

In order to make this dataset available, we compute the amount of each group (divided by year, sex, and age group) and create a new data frame to do the ANOVA test.

3.2 ANOVA test

Then, we presented the two-way ANOVA to test for the amount of strip search among different sex and age groups. In the first ANOVA test, the Null hypothesis is “There is no significant main effect of Sex or Age group at arrest on StripSearch.”, while the Alternative hypothesis is “There is a significant main effect of Sex or Age group at arrest on StripSearch" and "There is a significant interaction effect between Sex and Age group at arrest on StripSearch.” And the hypothesis of the other two ANOVA tests is similar to the first one, only changing the independent variable to Age group and Year / Sex and Year.

	sum_sq	df	F	PR(>F)
C(Sex)	9.79E+05	1	3.893754	0.068542
C(Age_group__at_arrest_)	1.56E+06	6	1.033677	0.444279
C(Sex):C(Age_group__at_arrest_)	6.09E+05	6	0.403802	0.864393
Residual	3.52E+06	14	NaN	NaN

Table 3-1 Two-way ANOVA: Effect of Sex and Age group at arrest on StripSearch

From Table 3-1, all three p-values are higher than 0.05, so we cannot reject the null hypothesis. Therefore, there is no significant main effect of the Sex or Age group at arrest and no significant interaction effect between the two factors on StripSearch.

	sum_sq	df	F	PR(>F)
C(Age_group__at_arrest_)	1.56E+06	6	1.383141	0.28775
C(Year)	1.48E+06	1	7.857849	0.014092
C(Age_group__at_arrest_):C(Year)	1.00E+06	6	0.887887	0.529137
Residual	2.63E+06	14	NaN	NaN

Table 3-2 Two-way ANOVA: effect of Year and Age group at arrest on StripSearch

From Table 3-2, two p-values are higher than 0.05, so we cannot reject the null hypothesis. Therefore, there is a significant main effect of Year on StripSearch, but no significant main effect of Age group at arrest on StripSearch.

	sum_sq	df	F	PR(>F)
C(Sex)	9.79E+05	1	6.622096	0.016678
C(Year)	1.48E+06	1	9.98732	0.004228
C(Sex):C(Year)	6.64E+05	1	4.490743	0.044619
Residual	3.55E+06	24	NaN	NaN

Table 3-3 Two-way ANOVA: effect of Sex and Year on StripSearch

From Table 3-3, all three p-values are lower than 0.05, so we can reject the null hypothesis. Therefore, there is a significant main effect of Sex or Year on StripSearch, and also a significant interaction effect between Sex and Year on StripSearch.

3.3 Post-hoc tests

Then, we do Tukey's HSD as the posthoc tests for the strip search cases grouped by Sex and Year.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
female_2020	female_2021	-151.2857	0.8746	-718.202	415.6306	False
female_2020	male_2020	681.8571	0.0143	114.9408	1248.7735	True
female_2020	male_2021	-85.2857	0.9	-652.202	481.6306	False
female_2021	male_2020	833.1429	0.0024	266.2265	1400.0592	True
female_2021	male_2021	66.0	0.9	-500.9163	632.9163	False
male_2020	male_2021	-767.1429	0.0053	-1334.0592	-200.2265	True

Table 3-4 Tukey HSD: Means of StripSearch Grouped by Sex and Year

Thus, we would conclude that there is a statistically significant difference between the means of groups female_2020 and male_2020, groups female_2021 and male_2020, and groups male_2020 and male_2021, but not a statistically significant difference between the means of groups female_2020 and female_2021, groups female_2020 and male_2021, and groups female_2021 and male_2021.

4. Results

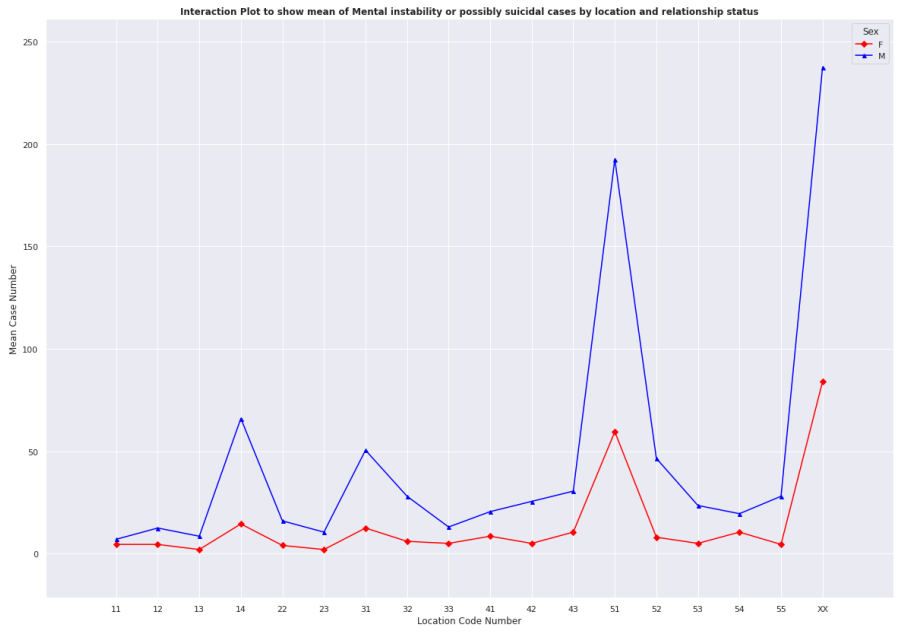

Figure 4-1

Although the interaction chart does not provide any information about statistically significant differences, the chart shows that. (a) males are more present in mentally unstable and suicidal cases compared to females; (b) the higher the number of cases in the neighborhood, the greater the difference in the percentage of males and females involved

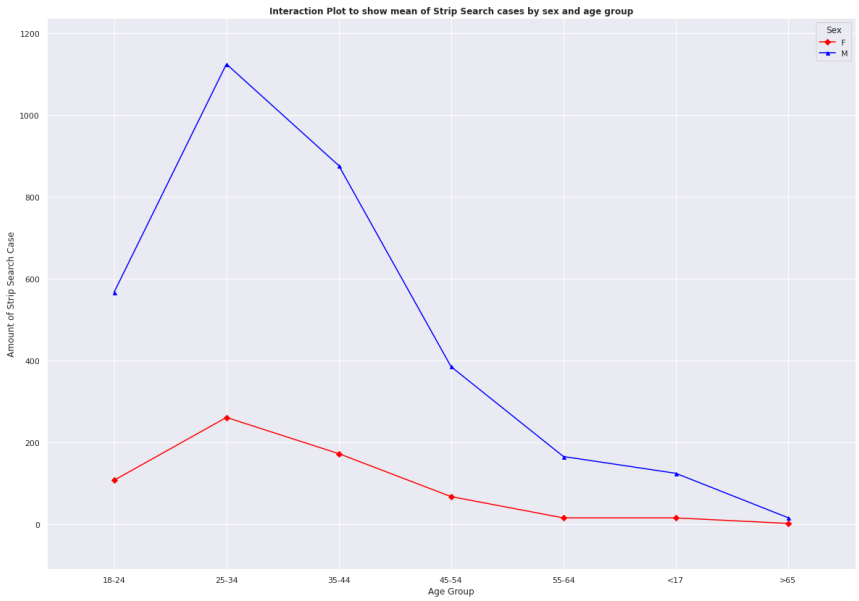

Figure 4-2

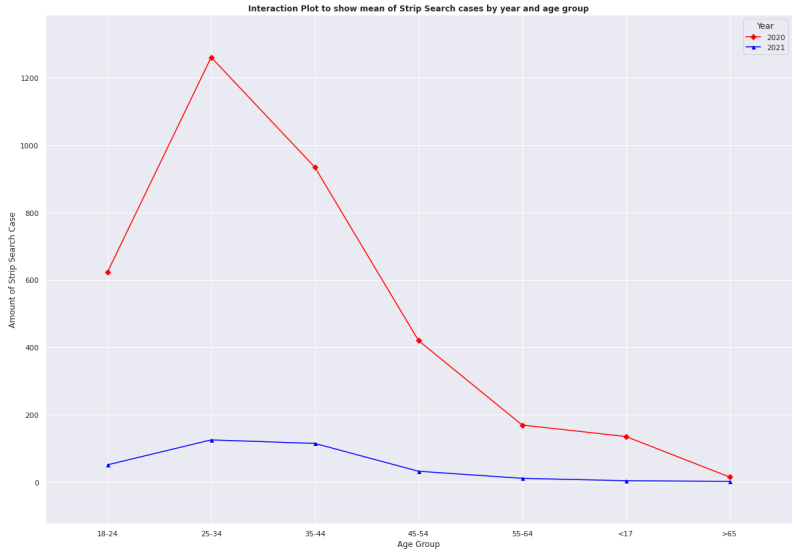

Figure 4-3

Based on the research, we found that although there is a significant difference in strip search between males and females, there is no significant main effect of the Sex or Age group at arrest and no significant interaction effect between the two factors on strip search. When we consider the Year and Age group as independent variables, only Year (instead of Age group and the interaction between Year and Age group) shows the main effect on StripSearch.

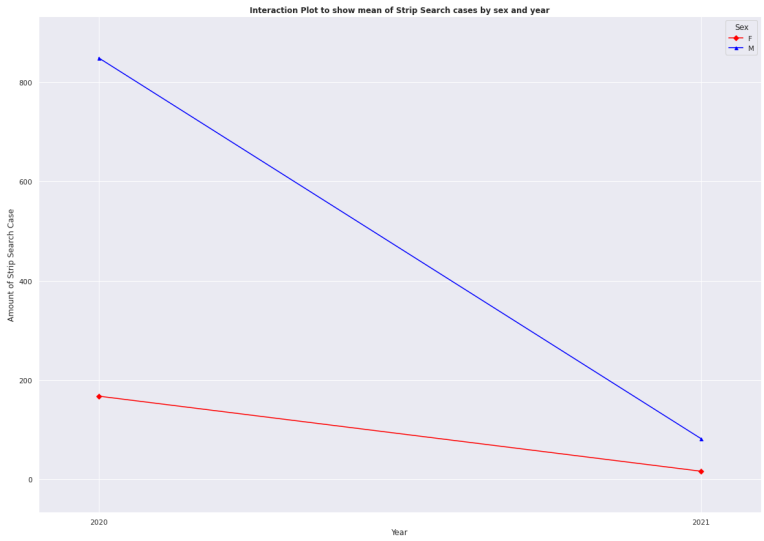

Figure 4-4

However, there is a significant main effect of Sex or Year on StripSearch, and also a significant interaction effect between Sex and Year on StripSearch, especially between the means of groups female_2020 and male_2020, groups female_2021 and male_2020, and groups male_2020 and male_2021.

5. Discussion and Conclusion

The findings of this project have to be seen in light of some limitations. Most importantly, the inequality in our group-level data. In this analysis, we chose to perform two-way ANOVA to test the differences between group means. However, the sample sizes of each group are not equal. Although it is not one of the assumptions made in an ANOVA, it may still lead to problems, such

as it may reduce the statistical power or the robustness to unequal variance. Despite the data inequality, several improvements could be made for future studies. For example, a three-way ANOVA test could be used later to solve the problem of unbalanced data: regarding "Year", "Sex", and "Age Group" as the three categorical variables, which may show the difference among each group clearer and more detailed.

In conclusion, after the icon, t-test, and ANOVA tests, we were able to determine that: firstly, gender and year had a significant effect on the strip search data while different age groups did not. Second, among all strip searches, cases involving mentally unstable and suicidal individuals were significantly influenced by gender and district. Toronto's 14 and 51 Blocks were the most frequent locations for related cases. In addition, the number of strip searches decreases each year, but the number of mentally unstable suspects increases each year. This is the reason for the increase in indiscriminate attacks felt by Toronto citizens in the news. Based on the data set and preliminary analysis, the research team recommends that Toronto citizens take care to avoid Block 14 and Block 51 and pay more attention to men who are acting abnormally when walking outside.