INF 2178 Midterm Project

Exploration of the Arrest Pattern and Performance of the Toronto Police Service

By: Group 32

Si Cheng (1003834158),
Ruolan Zhang (1004711010)

Instructor: Prof. Shion Guha

Table of Contents

1. Introduction

1.1 Background Information

Throughout history, Canadians have always relied on their law enforcement for keeping them safe, maintaining law and order, protecting lives and property, preventing crime, and detecting and apprehending criminals. Based on the data collected from the 2019 General Social Survey (GSS), it was found that 90% of Canadians living in the provinces were reported to have either a high degree or some level of trust in the police (Ibrahim). However, things seemed to change after the breakout of the Covid-19 pandemic, as another survey has shown a drop in those positive perceptions of the police service after March 2020 (Ruddell). This could cause a potential problem that individuals are less likely to comply with the law, report crimes, share information with the police, and most importantly, to collaborate with officers. There has been significant debate over the police system, as numerous controversies have arisen in recent times. These controversies have brought into question the performance of the law enforcement, their use of force, and their treatment of vulnerable populations, and so on. There are voices saying Toronto police officers are underworked yet overpaid, and are too powerful. According to information gathered by the Ontario Municipal Benchmarking Initiative in 2014, Toronto police officers had the lightest workloads compared to their counterparts in the study, with an average of 18 cases per year per officer (compared to Ottawa's 24 and Hamilton's 28). The city also did not perform well in solving violent crimes, which indicates the low productivity and potential lack of ability of Toronto police officers to deal with violent suspects (Preville). Furthermore, issues of injustice such as racial discrimination have also been brought to the forefront of public attention. Reports in the news have highlighted that compared to the overall population in the city, minority groups were found to be disproportionately affected by both use-of-force incidents and strip searches (Honderich).

1.2 Motivation

All these controversies have led to discussions about the police performance, which inspired us to assess the current Toronto police service system by exploring three questions of interest regarding their productivity across different divisions in previous years, questionable unjust treatment when it comes to minorities, and levels of aggression demonstrated from the suspects they deal with. We hypothesize that analysis of metrics such as caseload, individual’s level of cooperation under arrest, and strip search frequencies due to different reasons based on race, would provide the public with more insights into the police service we are relying on. We will use a dataset on arrests and strip searches happening in the city of Toronto from 2020 to 2021.

1.3 Research Objective and Questions

The objective of our research is to assess the performance of the Toronto police service in managing caseloads and ensuring equitable treatment for all individuals, by analyzing caseload variations across different years and divisions, the frequency of strip searches across different racial groups and reasons. Additionally, we aim to offer constructive advice by exploring how the demographic information of suspects, such as gender, relates to their level of cooperation during the arrest.

Our analysis will be structured around the following three research questions, which were developed based on the insights gained from the background research, literature review, and the statistical analysis with an informal, exploratory examination of the data (including the descriptive statistics, …diagrams, and T-tests in the EDA section).

- Research Question 1: How does the caseload of individual police officers change between 2020 and 2021, and are there any variations in caseloads between different divisions? Is there a significant interaction effect between year and division with respect to caseload?

- Research Question 2: How does the frequency of searches vary across different racial groups and search reasons? Are certain racial groups more likely to be searched for specific reasons?

- Research Question 3: How does the level of cooperation during an arrest vary across different age and gender groups?

2. Literature Review

Three research questions put different focuses on the dataset, "Arrests and Strip Searches", to dig into some meaningful criminal tendencies as well as potential biases for the criminology field. A British study conducted between May 1999 to September 2000 in north London collected data from people who were arrested and held in police custody in a single police station, including demographic characteristics, the reason for arrest, whether strip searches were conducted etc. It emphasized that the reason for the arrest is most likely the major factor for determining whether a strip search was conducted throughout every age group. Specifically, people associated with concealing evidence such as drugs and robbery own the highest two percentages of arrests that were strip-searched. This study also pointed out that the lack of trust and confidence towards minority ethnic communities, for example, African-Caribbean, triggered the local police to apply strip searches more frequently than other races. Different from our dataset, the outcome of arrest is another emphasized factor within their consideration. The same study shows arrests resulting in further charges have a 14% of strip search rate although the covariance is relatively lower.

From the Criminal Code booklet(C-46) amended by the Government of Canada, section 129 and section 170 define someone who assaulted police officers or resisted being arrested as guilty. This official document works as background research and provides concrete support for exploring the selected research question. In addition, a study related to police practice and research, called "Resisting Arrest: Predictions of Suspect Non-compliance and Use of Force against Police Officers" used logistic regression to predict that whether a witness is presented is a significant factor for suspects to be non-cooperative when they get arrested, after examining 1,220 resisting arrest situations. Surprisingly, resisting arrest mostly occurs when witnesses are presented if only considering situation variables. Furthermore, a typical portrait of the person who is likely to resist arrest without the effect of drugs or alcohol is an above 30-year-old white male.

As background research on counting the performance of Toronto police officers, "Reported Crime Statistics in Toronto can be misleading" points out the concern that the number of incidents reported in

Toronto in recent years is actually underrated. Approximately two-thirds of citizens chose not to report less serious incidents to local police, and 83% of sexually assaulted cases, which can be considered serious criminal offenses, are under-reported. Therefore, it will be pretty difficult to evaluate police performance from currently available datasets. According to the geographical division of the Toronto area, this study points out that divisions 52, 51 and 55 in the downtown area ranked as the three divisions with the highest crime rates since there are clusters of night-time entertainment.

3. Exploratory Data Analysis

3.1 Data Cleaning

To conduct analyses for our three research questions, we created three sub-datasets that contain the corresponding dependent variables and predictors. In particular, aggregation methods such as group by were applied, since we are mainly interested in the case frequencies regarding different features. For the first sub-dataset, the left join was used to get all unique police IDs and replace NA with 0 in the count column for officers who did not manage any case in either 2020 or 2021. For the second sub-dataset, we reset the data type of all four search reason columns to integer for further calculation. Additionally, only records with the searchStrip equal to 1 were extracted. For the third sub-dataset, we exclude the records where gender was undefined since they constitute only a negligible proportion of the entire dataset. Moreover, it is notable that we merged “Age 17 years and under” and “Age 17 years and younger” together into one group, and merged “Aged 65 and older” and “Aged 65 years and older” into one group, for further interpretation.

3.2 Descriptive Statistics

Summary statistics are computed to show the sum-up features of each independent variable in the research question, and provide a basic understanding of values recorded in the dataset, including mean, count, standard deviation and 95% confidence interval(Table 3.1 to Table 3.6).

Table 3.1 *Summary Statistics for Arrest Year*

Arrest Year	Mean	Count	std	ci95_lo	ci95_hi
2020	1.220852	26194	0.703666	1.212330	1.229374
2021	1.237209	26913	0.726025	1.228535	1.245883

Table 3.2 *Summary Statistics for ArrestLocDiv*

ArrestLocDiv	Mean	Count	std	ci95_lo	ci95_hi
11	1.188704	1505	0.619072	1.157427	1.219982
12	1.152727	1650	0.494673	1.128858	1.176596
13	1.177350	936	0.566969	1.141028	1.213673

14	1.240844	2512	0.696263	1.213616	1.268072
22	1.137062	1627	0.495015	1.113008	1.161116
23	1.117685	1555	0.443358	1.095648	1.139722
31	1.170455	1936	0.540719	1.146368	1.194541
32	1.141607	1829	0.486314	1.119320	1.163895
33	1.068233	894	0.318939	1.047326	1.089140
41	1.157387	2281	0.493154	1.137149	1.177626
42	1.132515	1630	0.468385	1.109777	1.155254
43	1.199724	2173	0.647747	1.172489	1.226959
51	1.406775	3572	1.030473	1.372981	1.440569
52	1.145374	2043	0.491766	1.124050	1.166699
53	1.127869	1220	0.432529	1.103598	1.152140
54	1.237893	1160	0.729541	1.196810	1.280776
55	1.166917	1330	0.546515	1.137545	1.196289
XX	1.274877	23254	0.808584	1.264485	1.285270

Table 3.3 *Summary Statistics for Race*

Race	Mean	Count	std	ci95_lo	ci95_hi
Black	159.322581	31	187.947047	93.160260	225.484901
East/Southeast Asian	22.000000	29	24.124676	13.219514	30.780486
Indigenous	22.962963	27	25.069874	13.506552	32.419374
Latino	22.217391	23	11.700647	7.435474	16.999308
Middle-Eastern	15.379310	29	18.162783	8.768732	21.989889
South Asian	20.875000	24	22.158152	12.009893	29.740107
Unknown or Legacy	35.172414	29	41.617363	20.025239	50.319588
White	218.968750	32	276.739715	123.083331	314.854169

Table 3.4 *Summary Statistics for SearchReason*

SearchReason	Mean	Count	std	ci95_lo	ci95_hi

AssistEscape	60.818182	44	119.686854	25.452977	96.183387
CauseInjury	99.533333	60	211.275352	46.073296	152.993371
PossessEvidence	52.709677	62	104.703545	26.646845	78.772510
PossessWeapons	60.965517	58	124.346709	28.963573	92.967462

Table 3.5 *Summary Statistics for* Age_group_at_rest_

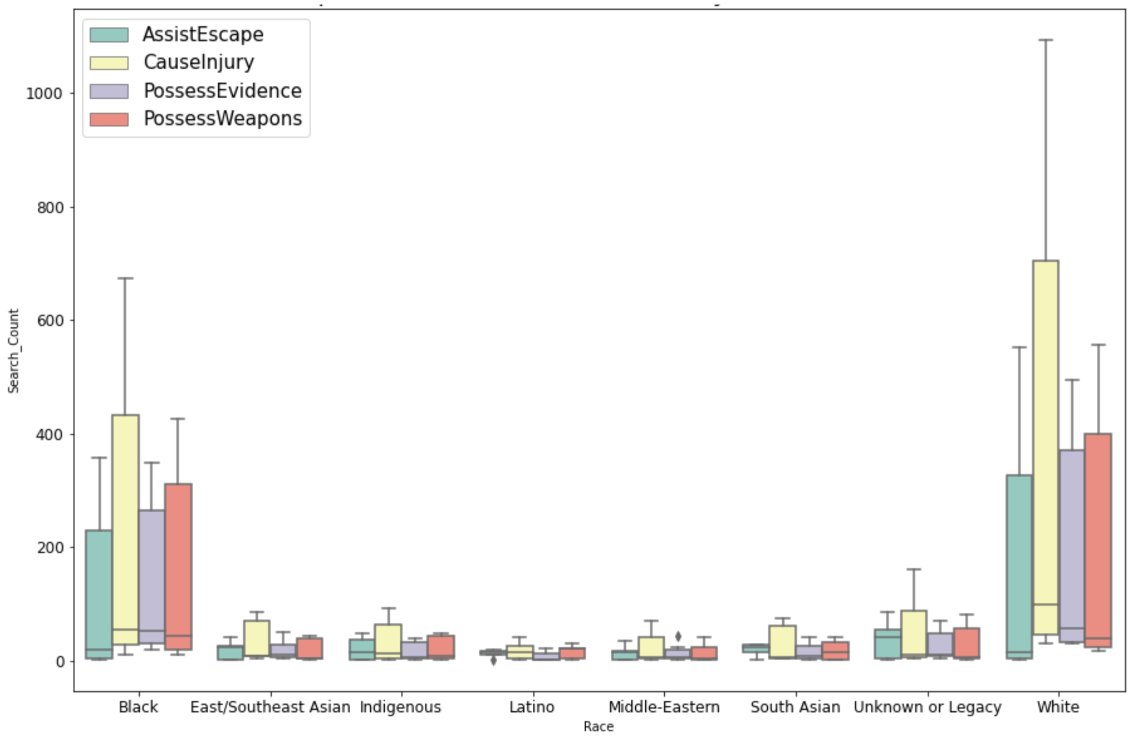
Age_group_at_rest_	Mean	Count	std	ci95_lo	ci95_hi
Aged 17 years and younger	50.812500	32	74.721607	24.922784	76.702216
Aged 18 to 24 years	351.250000	16	457.525737	127.062389	575.437611
Aged 25 to 34 years	738.312500	16	1046.426727	225.563404	1251.061596
Aged 35 to 44 years	564.062500	16	876.379177	134.636703	993.488297
Aged 45 to 54 years	305.000000	16	511.947914	54.145522	555.854478
Aged 55 to 64 years	156.250000	16	290.649961	13.831519	298.668481
Aged 65 years and older	24.178571	28	39.164471	9.671848	38.685295

Table 3.6 *Summary Statistics for* Sex

Sex	Mean	Count	std	ci95_lo	ci95_hi
F	105.602941	68	207.635813	56.251041	154.954841
M	402.222222	72	740.839441	231.097042	573.347402

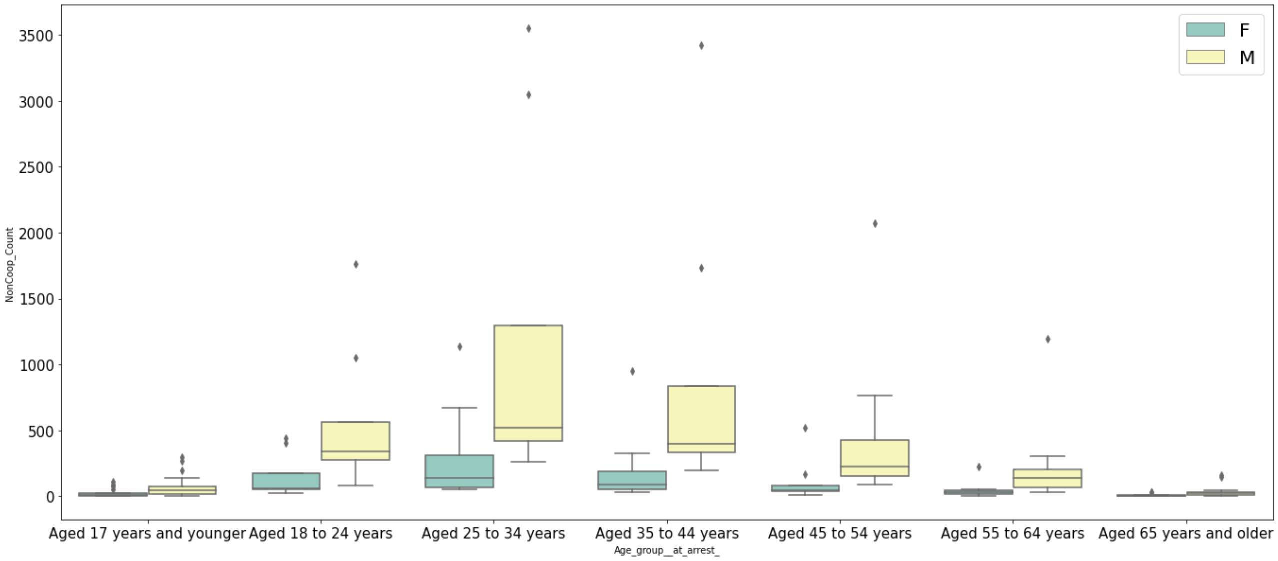
This barplot (Diagram 3.1) shows the count of strip searches for each racial group, and is further grouped by four search reasons. Most of the distributions are pretty right skewed and cause the imbalance of whisker length. From an at-a-glance aspect of this plot, people who were implemented strip searches being perceived as white and black take dominant shares of the sample. Cause injury is the search reason that most likely leads to being strip searched.

Diagram 3.1 *Barplot of Strip Search Count for Each Race by Search Reasons*



This barplot below(Diagram 3.2) displays the number of people who are non-cooperative when being arrested, grouped by age groups with nine levels. Another independent variable selected for this boxplot is Sex, and we are able to compare the difference in distributions of males and females through boxes and the position of whisker marks. Generally, we can see suspects between the age of 25 to 34 have the largest number of non-cooperative arrests. As shifting toward younger or older age levels, significant decreases in counts can be observed for both sexes. While males tend to actively resist arrest more frequently than females under every age group.

Diagram 3.2 *Barplot of Non-cooperative Count for Each Age Group by Gender*



3.3 T-test

All selected independent variables for three research questions are categorical with two or more than two levels, while dependent variables are continuous measurements. For those independent variables with exactly two levels, we ran an independent two-sample t-test and a dependent paired t-test to decide whether the outcome means for the two groups differed from each other using hypothesis testing. We will first check four assumptions for performing a t-test as supportive evidence for concrete results. Then, propose null and alternative hypotheses we are going to test in each research question. Moving on to calculate test statistics using Python packages, interpret results and decide whether we should reject or fail to reject the null hypothesis. Two t-tests are going to be performed to check whether differences in means are statistically significant: the first pair of independent-dependent variables is Sex and the count of non-cooperative arrests, and the second one is Year and the number of incidents each police handled.

3.3.1 Assumption Check

Two-sample t-test

For Sex and the count of non-cooperative arrest, we will perform a two-sample two-sided t-test. The assumption for independence is automatically checked since two groups are assembled by gender, and no same individual is appearing in both samples. Next, we use the Anderson-Darling Normality test to check whether both samples are normally distributed. This test sets up the null hypothesis that data is not different from normal. The test statistics of the male sample is approximately 11.67. We compare this value to each critical value that corresponds to each significance level to see if the test result is significant. For example, the critical value for alpha equals 0.1 is 1.039, and our test result is greater than this critical value. Therefore, this test statistic is significant at this level. After completing the comparison in every significance level, we can see the test statistics are significant for all levels, and we have sufficient evidence to reject the null hypothesis that sample data is normal. Thus, the normality assumption is violated in this case. The same procedure is performed on the female sample, and the result is pretty much similar to that of the male one which is not normally distributed. However, we will still carry out our test since a real-life dataset is impossible to be perfectly normal under any circumstance. The analysis result should not be rejected for the reason of non-normality, but have to mention this point as a limitation. Then, use the rule of thumb to check the homogeneity of variances. When the ratio of the larger variance to the smaller variance is less than 4, we will consider those two samples to have roughly equal variance. However, since the calculated ratio does not satisfy this rule, the homogeneity of variance is also violated. As a result, we will perform Welch's t-test which does not assume the equal variance of two populations instead of the Student's t-test. Last but not the least, the dataset is obtained using a random sampling method, so assumption four talking about random sampling is also passed.

Paired t-test

Because we are trying to compare the performance of every Toronto local police officer between 2020 and 2021, a paired t-test is a proper option to estimate whether there exist mean differences in the same group in 2020 and 2021. The number of Incidents that each police handle is measured twice, and later paired with each in a one-to-one manner. Through the Anderson-Darling Normality Test, two groups are also non-normally distributed. While in this case, the homogeneity of variance is checked. Additionally, sample data is randomly collected from the population.

3.3.2 Hypothesis

Hypothesis for one-sided two-sample t-test

- H0 (Null Hypothesis):
  The mean number of people who were not cooperative under arrests does not vary by gender
- Ha (Alternative hypothesis):
  The mean number of male who were not cooperative under arrest is greater than that of the female's.

Hypothesis for paired t-test

- H0 (Null hypothesis):
  The mean number of cases each police officer handled does not vary by year(true mean difference of paired sample equals zero)
- Ha (Alternative Hypothesis):
  The mean number of cases each police officer handled varied by year(true mean difference of paired sample does not equals zero)

3.3.3 Result

Gender and the number of non-cooperative people

Calculated summary statistics display that the mean number of males who show non-cooperative actions(concealed item, combative, resisted, mental instability, assaulted officer) when being arrested(Mean=402.22) is significantly greater than that of females(Mean=105.60). To examine whether the difference is statistically significant, we apply Welch's t-test with non-equal variance. The resulting p-value for this test is 0.0008 which is smaller than the chosen significance level of 0.05. Therefore, we have sufficient evidence to reject the null hypothesis that the true mean difference equals zero.

Year and the number of incidents each police handled

In the paired t-test, exploring whether the mean number of incidents each police officer handled(continuous variable) differs by year(2-level categorical variable), we measured each police officer twice in 2020 and 2021 respectively and paired both together in a one-to-one manner. The mean number of incidents each police handled in 2021(Mean=0.67) is slightly higher than that of in 2020(Mean=0.65). P-value equals 0.0010 shows that the probability of observing the test result under the null hypothesis is 1%, which is statistically significant. Therefore, we have sufficient evidence to reject the null hypothesis that the true mean difference of paired samples equals zero.

4. Method

4.1 Dataset Description

The dataset used for this project is the Arrests and Strip Searches uploaded on Toronto Police Service (TPS) Public Safety Data Portal (Arrests and Strip Searches (RBDC-ARR-TBL-001)). The dataset contains details of all arrest and strip search cases that happened in 2020 and 2021, where the demographic data of people involved, arrest divisions, actions at arrest, reasons for a strip search, and other relevant information were entered. The dataset has a size of 65277 and 25 features, indicating that there were 65277 cases in total in these two years and information regarding 24 different attributes was recorded. Some of the important personal and procedural characteristics comprise the suspect's race, sex, age group, arrest location, whether being strip-searched, actions they took while being arrested, and the specific reason why they were asked for strip searches. 12 out of 24 features are numerical, others are text. It is worth mentioning in Table 5.6 Tukey's HSD test result for Q2 that most of the numerical features are in binary format, and all measurements are on a nominal scale. It is worth noting that further analysis in this report assumes the variable 'PersonID' as the identification number of the police officer who is responsible for the arrest since no clarification about this variable is provided in the dataset summary.

4.2 ANOVA Tests

RQ1: Caseload vs. Year and Division

The plot presented in the EDA section and the one-sided t-test result of the first research question showed that there is a significant difference in the mean number of cases managed by each police officer in 2020 and 2021. We will use a two-way ANOVA test to further explore how the mean number of cases (dependent variable) varies from each division (multi-level explanatory variable), and how the interaction between the two years and each division throughout Toronto affects police officers' caseload.

The hypothesis being tested are the following:

> **H0**: There is no difference in the mean number of cases managed by each police officer between 2020 and 2021 for all divisions throughout Toronto.
> **Ha**: There is a difference in the mean number of cases managed by each police officer between 2020 and 2021 for all divisions throughout Toronto.
>
> **H0**: There is no difference in the mean number of cases managed by each police officer across various divisions in Toronto for the two years.
> **Ha**: There is a difference in the mean number of cases managed by each police officer across various divisions in Toronto for the two years.
>
> **H0**: There is no interaction between the year and division regarding the mean number of cases managed by each police officer.
> **Ha**: There is an interaction between the year and division regarding the mean number of cases managed by each police officer.

RQ2: Search Frequency vs. Race and Search Reasons

From the side-by-side boxplot generated based on the search against race and reason sub-dataset, we observe a difference existed between racial groups regarding the mean number of times being searched, among which the white and black groups seemed to be notably different from other groups. Additionally, another slight difference can be noted by comparing the mean number of searches conducted for various reasons. We will use a two-way ANOVA test to investigate how the mean number of searches (dependent variable) differs depending on race (multi-level explanatory variable) and search reasons (multi-level explanatory variable), and how the effect of change in search reasons depends on the level of race.

The hypothesis being tested are the following:

> **H0**: For all search reasons, there is no difference in the mean number of searches carried out for each racial group.
> **Ha**: For all search reasons, there is a difference in the mean number of searches carried out for each racial group.
>
> **H0**: Across all racial groups, there is no difference in the mean number of searches conducted for different search reasons.
> **Ha**: Across all racial groups, there is a difference in the mean number of searches conducted for different search reasons.
>
> **H0**: There is no interaction between the race and search reason regarding the mean number of searches.
> **Ha**: There is an interaction between the race and search reason regarding the mean number of searches.

RQ3: Level of Cooperation vs. Age and Sex

From a side-by-side boxplot generated based on the uncooperative individual counts against the age and sex sub-dataset, we observe a significant difference between the two gender groups in terms of the level of cooperation during arrest, where males tend to be much more difficult to deal with than women. This observation can also be verified by the one-tailed t-test performed above. The result of the t-test indicates that the mean number of uncooperative males is greater than the mean number of uncooperative females. Moreover, another notable difference can also be found between age groups, where people in early adulthood and middle age seem to be more uncooperative. We will use a two-way ANOVA test to investigate how the mean number of uncooperative individuals during arrest (dependent variable) differs from both age (multi-level explanatory variable) and sex (two-level explanatory variable), and how the effect of change in sex depends on the level of age.

The hypothesis being tested are the following:

> **H0**: The mean number of people who were not cooperative during arrest does not vary by gender for all age groups.
> **Ha**: The mean number of people who were not cooperative during arrest does vary by gender for all age groups.
>
> **H0**: There is no difference in the mean number of people who were not cooperative during arrest among different age groups for both genders.
> **Ha**: There is a difference in the mean number of people who were not cooperative during arrest among different age groups for both genders.

H0: There is no interaction between the age and gender regarding the mean number of people who were not cooperative during arrest.
Ha: There is an interaction between the age and gender regarding the mean number of people who were not cooperative during arrest.

Each of these three two-way ANOVA tests is also accompanied by a corresponding interaction plot, to visualize how two predictors of interest impact each other regarding the outcome.

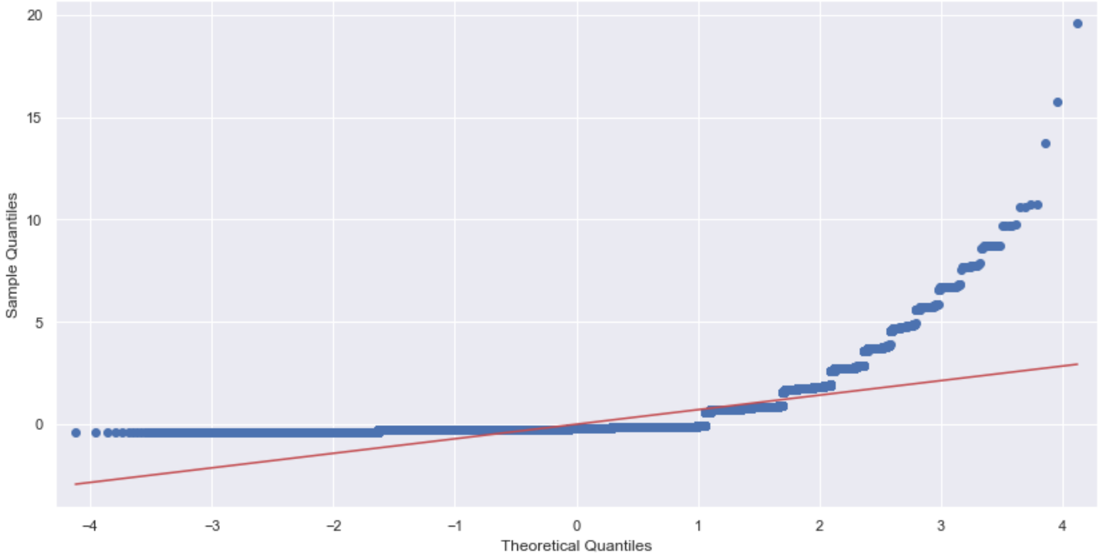
4.4 Assumption Check

In order for the test results to be considered valid, we need to perform the assumption checks to make sure that the following assumptions are met:

Independence
The raw dataset is transformed and aggregated into three distinct sub-datasets, each corresponding to one of the three research questions. All groups are mutually exclusive and there are no repeated measures. For each research question, two categorical variables are independent of each other, which means one does not have an influence on the other one.

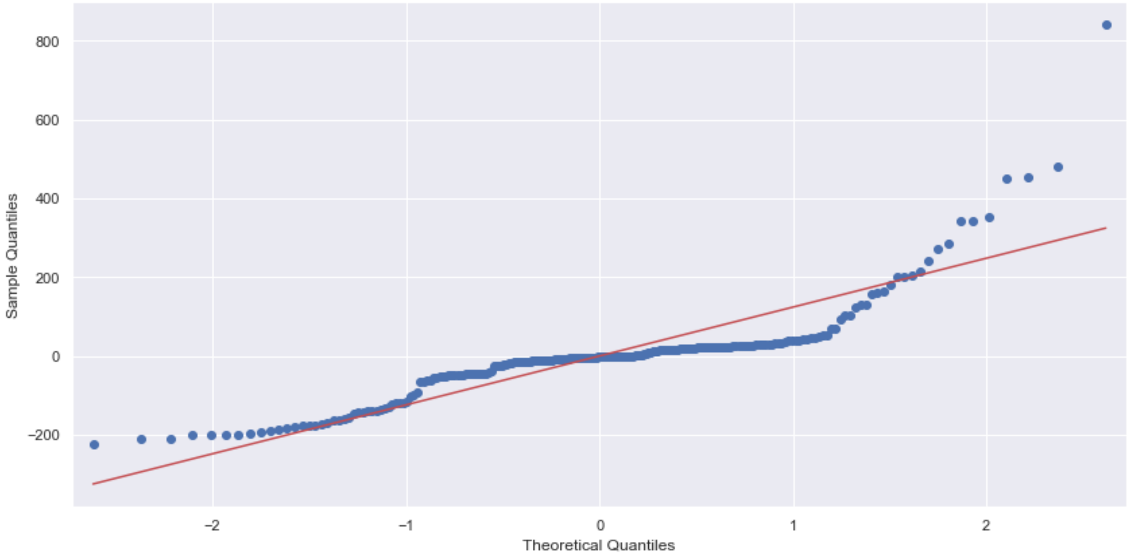
Normality
Normality assumption can be checked using a visual approach, for example, a residual plot. QQ-plot is drawn for each research question by plotting residuals of the additive model that is created when performing two-way ANOVA tests.

Diagram 4.1 Normal QQ-Plot for Research Question 1



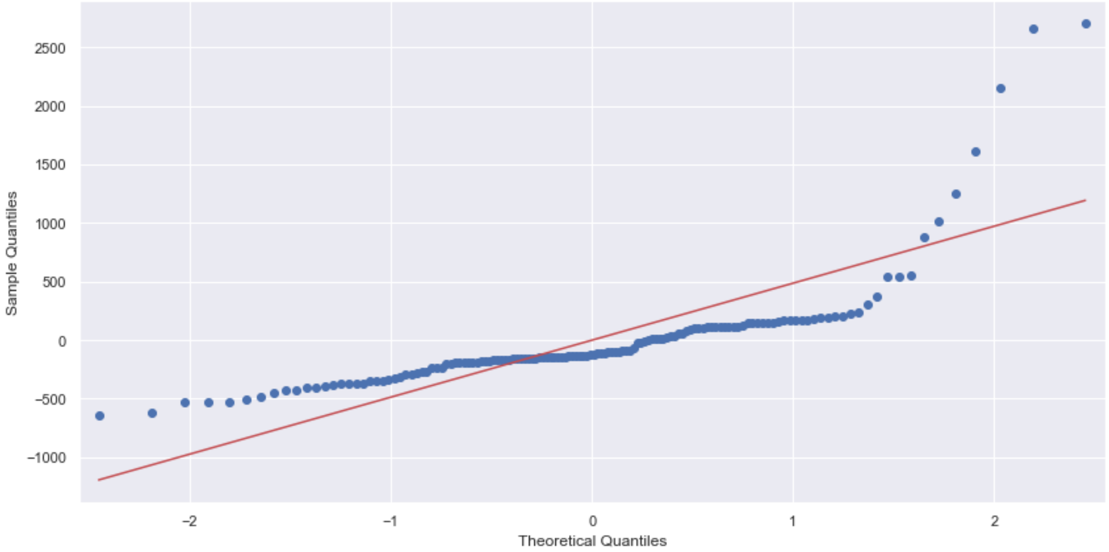
In Diagram 4.1, there are pretty many values on the right tail that significantly deviate from the red line, and the plot is quite right-skew. Therefore, the normality assumption for research question 1 is still violated.

Diagram 4.2 Normal QQ-Plot for Research Question 2



Compared to the last one, the QQ-plot for research question 2 (Diagram 4.2) looks more normal. However, there are still many outliers located at two tails of the plot that deviate from the linear red line. Therefore, the dataset is still not normally distributed.

Diagram 4.3 Normal QQ-Plot for Research Question 3



QQ-plot for research question 3 (Diagram 4.3) is not a normal plot since values do not fit the line well. There are many points deviating from the overall linear trend. Therefore, the normality assumption is violated.

Homogeneity of Variance

As all data is drawn from non-normally distributed datasets as we check in the last section, Levene's Test is selected to check the homogeneity of variance for each categorical variable inside every research question. We will only check categorical variables with more than two levels in this section since other two-level categorical variables are already checked before applying the t-test.

In the first research question, examining how year and location division affect a police officer's performance, 'division' is a categorical variable with multiple levels. Applying Levene's Test, we propose the null hypothesis that each level of location division has an equal variance. The resulting p-value equals 2.368e-120 which is significantly smaller than 0.05. As we have sufficient evidence to reject the null hypothesis, the homogeneity of variance assumption is violated for this arrest division. However, another categorical variable is tested to have equal variance.

Both categorical variables in the second research question have more than two levels, so we have to perform Levene's test twice. We propose the null hypothesis that each level of the racial group has an equal variance. The resulting p-value is 7.209e-12, indicating we have sufficient evidence to reject the null hypothesis. Therefore, the equal variance assumption for 'Race' is violated. In comparison, repeating a similar process on the 'Search Reason' variable, we got a p-value equal to 0.312 which is larger than the chosen significance level. So we can conclude that the 'Search Reason' passes this assumption.

Similarly, implementing Levene's Test for the 'Age' variable in the third research question, the resulting p-value is smaller than the chosen significance level, which equals 9.763e-04. Therefore, we can reject the null hypothesis that each level of age group has equal variance and make the conclusion that 'Age' violates the homogeneity of variance.

4.5 Post-hoc Tests

Performing Tukey's HSD test subsequent to each two-way ANOVA test enables us to examine the specific pairwise associations between various group combinations in greater detail, revealing which groups exhibit significant differences compared to other groups.

5. Result and Finding

RQ1: Caseload vs. Year and Division

Within the result of the two-way ANOVA test for our first research question displayed in Table 5.1, the interaction between the explanatory variables arrest year and arrest division is not significant (F = 1.402, p = 0.124), so we have no evidence to reject the additive model, and we conclude that the arrest year effects on the outcome are the same for all eighteen levels of divisions, and arrest division effects on the outcome are the same for both years. Therefore, we re-run the two-way ANOVA test with a different model, i.e., an additive model where the interaction term is removed.

Table 5.1 *Two-way ANOVA test result with interaction for RQ1*

	Sum of Squares	df	F	PR(>F)
Arrest_Year	3.187982	1.0	6.306177	1.203475E-02

ArrestLocDiv	313.828527	17.0	36.51689	3.129573E-120
Arrest_Year*ArrestLocDiv	12.046501	17.0	1.401723	1.242035E-01
Residual	26829.15468	53071	NaN	NaN

The result for the two-way ANOVA without interaction is shown in Table 5.2. Using a significance level of $\alpha = 0.05$, we can observe that both arrest year and arrest division affect the outcome. The mean number of cases managed by each police officer was significantly different for 2020 and 2021 ($F = 6.306$, $p = 0.012$), and it is separate for levels of division because the additive model is an adequate model. This result is consistent with the result of the T-test performed in the EDA section. Besides, the F-statistic and p-value of arrest division ($F = 36.517$, $p < 0.001$) indicates a statistically significant result that at least one level of divisions differs from the other levels regarding the mean number of cases managed by each police officer, and this is not affected by the year due to the additive model. Table 5.2 also tells us that the standard deviation in any group is approximately 0.711 (square root of Sum of Squares/df).

Table 5.2 *Two-way ANOVA test result without interaction for RQ1*

	Sum of Squares	df	F	PR(>F)
Arrest_Year	3.187982	1.0	6.305366	1.204025E-02
ArrestLocDiv	313.828527	17.0	36.512193	3.250680E-120
Residual	26841.201178	53088.0	NaN	NaN

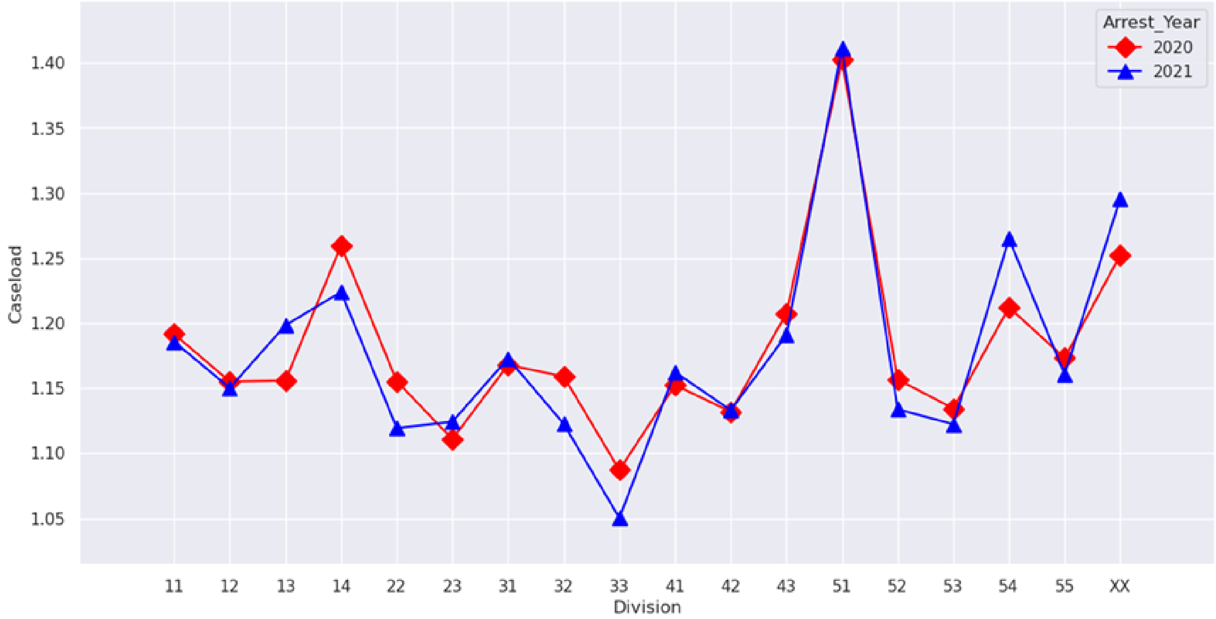
With Tukey's HSD post hoc test result displayed in Table 5.3, further important information about which levels of divisions are significantly different can be inferred by checking the adjusted p-value for each comparison, i.e., the mean differences between 153 possible combinations of groups. The differences lie between divisions 11 and 33 (p-adj = 0.0076), divisions 11 and 51 (p-adj < 0.001), divisions 11 and XX (p-adj = 0.0007), and 47 more comparisons listed in Table 5.3, whose adjusted p-value are all less than the 0.05 level of significance.

Table 5.3 *Tukey's HSD test result for RQ1*

Comparison	Group1	Group2	MeanDiff	p-adj	Reject
8	11	33	-0.1205	0.0076	TRUE
12	11	51	0.2181	0	TRUE
17	11	XX	0.0862	0.0007	TRUE
…	…	…	…	…	…
153	55	XX	0.108	0	TRUE

For the corresponding interaction plot (Diagram 5.1), the first impression is that there is no interaction between year and division. The caseload differs from most levels of divisions, among which officers working at division 33 manage the lowest average number of cases, and officers working at division 51 manage the highest average number of cases, and this phenomenon is not affected by years. Additionally, there is a slight difference in the average number of cases between 2020 and 2021.

Diagram 5.1 Interaction Plot to Show Individual Police Officer's Workload by Division and Year



RQ2: Search Frequency vs. Race and Search Reasons

Looking at the result of the two-way ANOVA test for our second research question displayed in Table 5.4, the interaction between the explanatory variables race and search reason is not significant (F = 0.428, p = 0.987), so we have no evidence to reject the additive model, and we conclude that the race effects on the outcome are the same for all four levels of search reasons, and search reason effects on the outcome are the same for all 8 levels of races. Therefore, we re-run the two-way ANOVA test with a different model, i.e., an additive model where the interaction term is removed.

Table 5.4 *Two-way ANOVA test result with interaction for RQ2*

	Sum of Squares	df	F	PR(>F)
Race	1.354282E+06	7.0	11.286626	5.937831E-12
SearchReason	9.316626E+04	3.0	1.811717	1.464011E-01
Race*SearchReason	1.542077E+05	21.0	0.428391	9.873658E-01
Residual	3.291154E+06	192.0	NaN	NaN

The result for the two-way ANOVA without interaction is shown in Table 5.5 below. Using a significance level of α = 0.05, we can observe that only race affects the outcome, while search reason does not affect the outcome. The mean number of searches varies significantly from different racial groups (F = 11.961, p < 0.001)., and this phenomenon exists for all search reasons, as the interaction term is not significant and hence not included in the additive model we use. This result is consistent with the boxplot generated in the EDA section. Nevertheless, the F-statistic and p-value of the search reason (F = 1.920, p = 0.127) indicate that we fail to reject the null hypothesis and search reason has no statistically significant effect on the mean number of searches. Table 5.5 also tells us that the standard deviation in any group is approximately 127.183 (square root of Sum of Squares/df).

Table 5.5 *Two-way ANOVA test result without interaction for RQ2*

	Sum of Squares	df	F	PR(>F)
Race	1.354282E+06	7.0	11.960681	7.643598E-13
SearchReason	9.316626E+04	3.0	1.919916	1.273450E-01
Residual	3.445362E+06	213.0	NaN	NaN

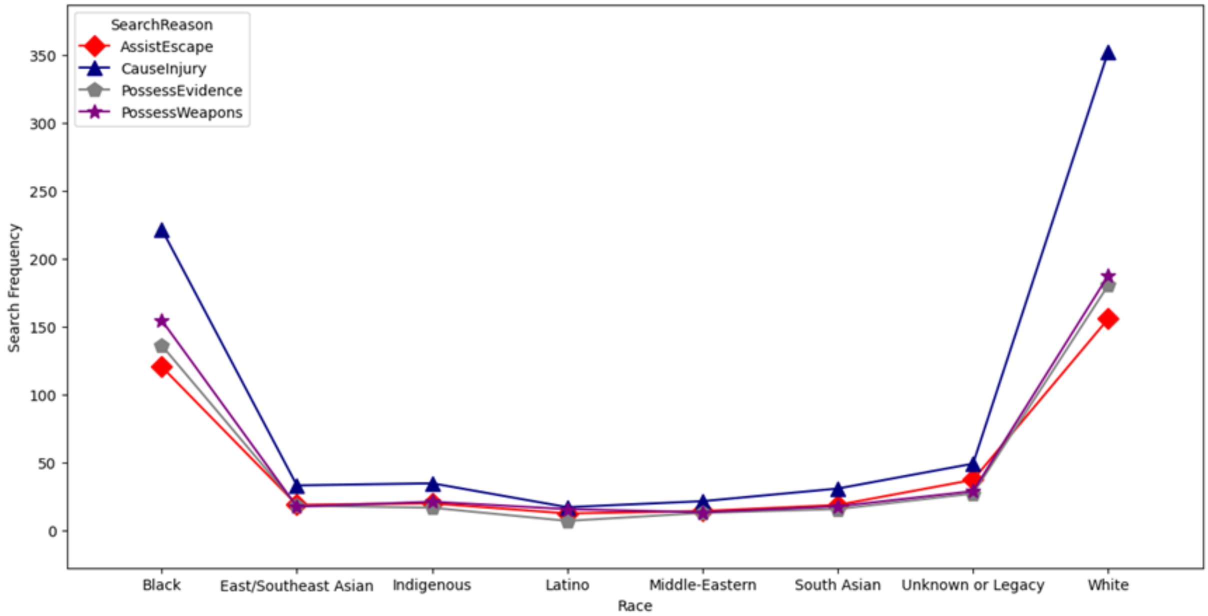
Tukey’s HSD post hoc test result displayed in Table 5.6 shows that differences are presented in 12 out of 28 possible groups' mean comparisons. It is worth mentioning that among those 12 comparisons, either the Black or White racial group was involved in each one. Moreover, the adjusted p-value for comparisons involving White groups was considerably smaller than the adjusted p-value for other comparisons. E.g., p = 0.0012 (Black and Latino) compared with p < 0.001 (White and Latino).

Table 5.6 *Tukey’s HSD test result for RQ2*

Comparison	Group1	Group2	MeanDiff	p-adj	Reject
1	Black	East/ Southeast Asian	-137.3226	0.0012	TRUE
2	Black	Indigenous	-136.3596	0.0018	TRUE
3	Black	Latino	-147.1052	0.0011	TRUE
4	Black	Middle-Eastern	-143.9433	0.0005	TRUE
5	Black	South Asian	-138.4476	0.0024	TRUE
6	Black	Unknown or Legacy	-124.1502	0.0054	TRUE
13	East/ Southeast Asian	White	196.9688	0.0	TRUE
18	Indigenous	White	196.0058	0.0	TRUE
22	Latino	White	206.7514	0.0	TRUE
25	Middle-Eastern	White	203.5894	0.0	TRUE
27	South Asian	White	198.0938	0.0	TRUE
28	Unknown or Legacy	White	183.7963	0.0	TRUE

Looking at the corresponding interaction plot(Diagram 5.2), there does not seem to exist an interaction between race and search reason. Notably, white people and black people are much more likely to be searched no matter what the specific reason is. Besides, although a minor difference can be observed when comparing the frequency of searches carried out for four different reasons, the interaction plot does not offer any indication of statistically significant differences. It is necessary to combine it with the ANOVA and Tukey's HSD tests to analyze and interpret the result.

Diagram 5.2 Interaction Plot to Show the Search Frequency by Race and Search Reason



RQ3: Level of Cooperation vs. Age and Sex

Looking at the result of the two-way ANOVA test for our third research question displayed in Table 5.7, the interaction between the explanatory variables age and sex is not significant (F = 2.085432, p = 0.059420), indicating that there is not enough evidence to reject the additive model, and we conclude that the sex effects on the outcome are the same for all nine levels of ages, and age effects on the outcome are the same for both genders. Therefore, we re-run the two-way ANOVA test with a new additive model without including the interaction term.

***Table 5.7** Two-way ANOVA test result with interaction for RQ3*

	Sum of Squares	df	F	PR(>F)
Age	8.729473E+06	6.0	6.083378	0.000013
Sex	3.371773E+06	1.0	14.098288	0.000264
Age*Sex	2.992535E+06	6.0	2.085432	0.059420
Residual	3.013440E+07	126.0	NaN	NaN

The result for the new two-way ANOVA without interaction is shown in Table 5.8. Using a significance level of α = 0.05, we can observe that both age and sex affect the outcome. The mean number of people who were not cooperative during arrest was significantly different for male and female groups (F =

13.435414, p < 0.001), and this is not affected by the individual's age as the model we use is an additive model. This result aligns with the result of the T-test performed previously. Besides, the mean number of people who were not cooperative during arrest was significantly different for age groups (F = 5.797349, p < 0.001), and the phenomenon is not influenced by sex. Table 5.8 also shows that the standard deviation in any group is approximately 500.961 (square root of Sum of Squares/df).

Table 5.8 *Two-way ANOVA test result without interaction for RQ3*

		df	F	PR (> F)
Age	8.729473E+06	6.0	5.797349	0.000022
Sex	3.371773E+06	1.0	13.435414	0.000357
Residual	3.312693E+07	132.0	NaN	NaN

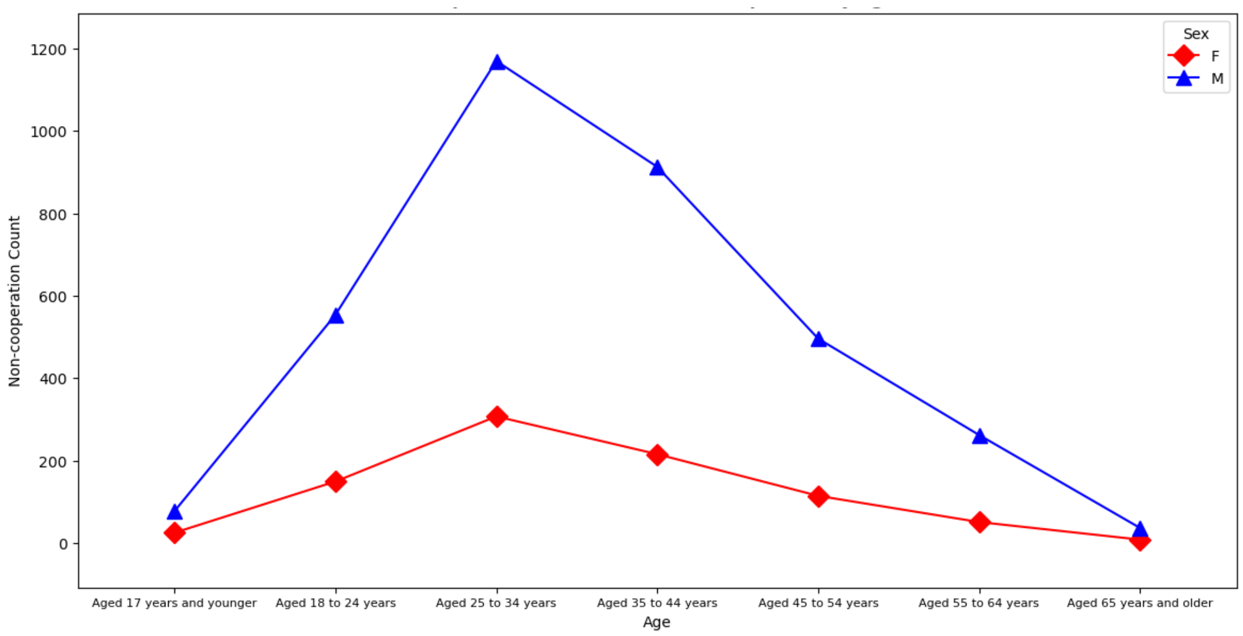
Tukey's HSD post hoc test result displayed in Table 5.9 shows that only 5 out of 21 comparisons are statistically significant. Differences lie between Aged 17 years and younger and Aged 25 to 34 years, Aged 17 years and younger and Aged 35 to 44 years, Aged 25 to 34 years and Aged 55 to 64 years, Aged 25 to 34 years and Aged 65 years and older, Aged 35 to 44 years and Aged 65 years and older. It is worth mentioning that the Aged 25 to 34 years are involved in 3 of these 5 differences, illustrating that this group is significantly different from other age groups regarding the average number of people who were not cooperative during arrest.

Table 5.9 *Tukey's HSD test result for RQ3*

	Group1	Group2	MeanDiff	p-adj	Reject
2	Aged 17 years and younger	Aged 25 to 34 years	687.5	0.0007	TRUE
3	Aged 17 years and younger	Aged 35 to 44 years	513.25	0.0279	TRUE
14	Aged 25 to 34 years	Aged 55 to 64 years	-582.0625	0.0329	TRUE
15	Aged 25 to 34 years	Aged 65 years and older	-714.1339	0.0005	TRUE
18	Aged 35 to 44 years	Aged 65 years and older	-539.8839	0.0214	TRUE

By analyzing the corresponding interaction plot(Diagram 5.3), we cannot really determine if there exists an interaction between age and sex. However, there is a considerable big difference in the average number of people who were not cooperative during arrest between the female and male groups with the same age from 25 to 34. The plot also shows that males are more uncooperative than females in general. Besides, the level of cooperation varies among different age groups, where middle-aged people seem to be most uncooperative.

Diagram 5.3 Interaction Plot to Show the Level of Cooperation by Age and Sex



6. Discussion

As public members living in Toronto, people inevitably get in touch with and rely on police sectors for their daily demands, including but not limited to crime prevention, law maintenance and community policing. Elevating police performances and preventing police brutality are able to efficiently earn public trust towards law enforcement as well as expose positive influences on the relationship between citizens and police officers. The most essential part is that citizens themselves have to ensure that they receive fair treatment from the police. This report provides an analysis of arrest cases in the Toronto area from 2020 to 2021, examining the interaction of suspects' demographics and relevant situational variables, and estimating police caseloads. As stated in the title of the dataset, the strip search is one of our focuses to study whether there exists any bias when conducting such offensive action without the consent of the person.

Synthesizing results we get from T-test and two-way ANOVA test, significant differences in caseloads can be observed between various divisions regardless of the variable ‘Year’. According to the division boundary provided by Toronto police, the division with the largest caseloads in both years, which is Division 51, is the Toronto downtown area with large amounts of infrastructure and entertainment. There is no surprise for this result because the downtown area has only 8.594 square kilometres but with relatively high population density. Tourism attraction also contributes to high caseloads. The downtown area is the destination for international tourists and there are a couple of famous landmarks existing inside this area such as the CN tower, the Royal Ontario Museum and the University of Toronto. With intuition, crimes are much easier to be committed in the area with a large population flow every day. This matches a part of the literature reviews associated with Toronto crime rates in recent years, and clusters of night bars can be another possible reason to explain the high caseload of police officers in Division 51.

Adding the influence of year into our consideration, caseload variation is inconspicuous within most divisions in 2020 compared to that of 2021 in the interaction table, which matches the calculated small mean difference of caseload between years displayed in the summary statistic. Synthesizing together with the ANOVA statistics, although the result is statistically significant with the suggestion of p-value(0.04) presented in the t-test, the caseload variation is actually not that significant. It is worth mentioning that the ANOVA and t-test results would become not statistically significant if the significance level was set to 0.01, and the new results would be more aligned with the descriptive statistics and interaction plot. On the other hand, small fluctuations did appear within a few divisions such as Division 14, 33, and 54, but the overall caseload of police officers does not experience any major changes. Therefore, we want to conclude that police performance is relatively stable in 2020 and 2021, but in order to obtain a more comprehensive evaluation of Toronto police, more data from recent years is still required.

To ensure citizens in Toronto are treated with respect and dignity in public security systems, analysis is performed on race-based strip search data to test whether Toronto police follow their proposed bias-free principles. Based on ANOVA test results, the mean number of strip searches varies a lot by racial group, no matter which search reason is applied. White and Black are two racial groups having the largest two mean numbers of strip searches in both 2020 and 2021. However, one limitation of this analysis method is only the frequency of the strip search is computed but not combined with the arrest frequency to calculate the strip search rate for each racial group. Since population bases for each racial group living in Toronto are significantly dissimilar, and the number of arrests frequency is unknown in our analysis, more strip search cases do not stand for a higher search rate for a specific race. 43.5% of the population living in Canada are White whereas only 9.6% of it are Black in 2021. We would consider adding statistics about arrest frequency for each racial group to compute strip search rate, so we are able to better compare the test statistics.

In addition, the goal of exploring the demographic characteristics of non-cooperative suspects is to generalize into a public portrait that may impose a potential threat to police officers. Intuitively, males would act more offensively than females when they are under arrest, and this perspective perfectly matches our test result that the mean number of non-cooperative males is significantly larger than that of the females. As we collaborate with another statistically significant variable, age group suspects aged between 25 to 34 display significant mean differences in the number of non-cooperative arrests compared to other age groups. Therefore, we generalize that males aged between 25 to 34 are inclined to display combative actions or assault the police. However, studying criminology should not be limited to basic demographic information, other variables such as education level, job class and income are also important indicators that can help to analyze crime motivations and provide a reference for police officers about the distribution of police power.

One concern for the above analysis is that two of three assumptions that we suppose to be held for both the ANOVA tests and T-test are actually violated. The good news is, our additive models created in ANOVA tests are pretty robust against the violation of the normality assumption. For a real-world dataset, data scientists cannot expect the distribution to be perfectly normal, and they have to accept this imperfection. At this time, the Type one error rate will be relatively close to the alpha level set in the test. A remedy for non-normal distribution is to transform the dataset using algorithms into a normal shape, but this method does not apply to every dataset. In comparison, the violation of the equal variance assumption should be treated with more concern. We used Welch’s T-test instead of the Student’s T-test since the former one is specifically designed for two groups with unequal variances. For two-way ANOVA tests, the violation of equal variance can be tolerated when sample sizes are close to each other, but we cannot ensure samples have an equal number of observations in this dataset. Therefore, our next step would be searching for some methods that are applicable to samples with unequal variances. In general, we can conclude that our ANOVA models and T-test results are robust to current light violations in assumptions but we can still try to find other concrete methods to improve our analysis.

7. Conclusion

The outcomes of our exploratory data analysis and ANOVA tests provide several significant insights regarding the caseload of particular police officers and the application of strip searches in Toronto. First, according to our data, there is not much difference in the number of cases handled by each police officer in different divisions and years. However, even small differences can have significant implications for the performance of individual police officers. Second, we discovered that while strip searches were carried out equally for each of the four reasons, black and white people were discovered to be subjected to them more frequently. This implies that there might be racial inequities in the police system, and more investigation is required to fully understand this problem. Finally, our data show that males in their early to late thirties tend to be less cooperative when they are being arrested. This suggests that police officers should take extra precautions when dealing with these individuals to ensure their safety and that of others involved. Overall, our research indicates that the Toronto Police Service appears to handle a small number of cases, but there are issues with the way different racial groups are treated. These results have significant implications for police operations and regulations and emphasize the need for more study to eliminate racial inequities in law enforcement and enhance Toronto's policing standards.

Reference

Arrests and strip searches (RBDC-arr-TBL-001). Toronto Police Service Public Safety Data Portal. (n.d.). Retrieved February 27, 2023, from https://data.torontopolice.on.ca/datasets/TorontoPS::arrests-and-strip-searches-rbdc-arr-tbl-001/about

Bedre, R. (2022, March 6). *ANOVA using python (with examples)*. Data science blog. Retrieved February 27, 2023, from https://www.reneshbedre.com/blog/anova.html

Caldwell, A. R., Lakens, D., Parlett-Pelleriti, C. M., Prochilo, G., & Aust, F. (2022, March 31). *Power analysis with superpower*. Chapter 12 Violations of Assumptions. Retrieved February 27, 2023, from https://aaroncaldwell.us/SuperpowerBook/violations-of-assumptions.html

Crawford, C., & Burns, R. (2002). Resisting arrest: Predictors of suspect non-compliance and use of force against police officers. *Police Practice and Research*, *3*(2), 105–117. https://doi.org/10.1080/15614260290033611

guest_blog. (2022, June 23). *Introduction to ANOVA for statistics and data science*. Analytics Vidhya. Retrieved February 27, 2023, from https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/

Honderich, H. (2022, June 15). *Toronto police use more force against ethnic minorities - report*. BBC News. Retrieved February 27, 2023, from https://www.bbc.com/news/world-us-canada-61818396

Ibrahim, D. (2020, November 25). Public perceptions of the police in Canada's provinces, 2019. Retrieved February 27, 2023, from https://www150.statcan.gc.ca/n1/pub/85-002-x/2020001/article/00014-eng.htm

Newburn, T., Shiner, M., & Hayman, S. (2004). RACE, CRIME AND INJUSTICE? Strip Search and the Treatment of Suspects in Custody. The British Journal of Criminology, 44(5), 677–694. http://www.jstor.org/stable/23639161

Police divisions. Toronto Police Service Public Safety Data Portal. (n.d.). Retrieved February 27, 2023, from https://data.torontopolice.on.ca/datasets/43ef8c93684f44a78eade66b3350ce9f_0/explore?location=43.683069%2C-79.219002%2C11.02

Preville, P. (2020, December 11). *Toronto's billion-dollar problem: Our cops*. Toronto Life. Retrieved February 27, 2023, from https://torontolife.com/from-the-archives/toronto-police-service-vs-everybody/

Race-based-data. Toronto Police Service Public Safety Data Portal. (n.d.). Retrieved February 27, 2023, from https://data.torontopolice.on.ca/pages/race-based-data

Ruddell, R. (2022). The changing context of Canadian policing: An examination of the public's perceptions after 2020. *Journal of Community Safety and Well-Being*, *7*(2), 47–52. https://doi.org/10.35502/jcswb.260

Westin, M. (2021). Reported Crime Statistics in Toronto can be Misleading.

Zach. (2020, December 20). *Welch's t-test: When to use it + examples*. Statology. Retrieved February 27, 2023, from https://www.statology.org/welchs-t-test/