INF2178: Experimental Design for Data Science

Kuan Yi Chou

1003850674

Introduction

In a racial sense of view, police have been portrayed as an abuser of power. In this study, we will try to investigate whether we observe significant evidence than backup the view of police targeting minorities. We also want to protect front line officers by indicating criminals that are booked more frequently previously so that there are indications that standard of procedure may be edited in response.

Using the data collected by Toronto Police Service during 2020-2021, we used the booked variable to create a booked_sum variable that indicates the total number of bookings a unique Person ID received during the period of time. Using Booked and Booked_Sum as dependent variables we used power analysis to determine the effect size and sample size. Afterwards, we used Ancova to analyze the effect each independent variable has on the total number of bookings (booked_sum). Finally, using Logistic regression we were able derive the change in likelihood of booking from each unit change of independent variables.

Literature Review

Toronto Police have a long history of targeting racial minorities. According to the BBC, Blacks are more likely to encounter and have force used upon them by the police BBC News. (BBC News, 2021). While the Police Chief of Toronto publicly admitted that there is systematic discrimination in the Toronto police station. Furthermore, on data that is not recorded on the public dataset, Black and Asians both are over 50% more likely to have a firearm drawn upon them during routine arrests. This is especially disturbing since there is no indication of firearms drawn by the arrestees. Toronto police are just more scared of the race of the suspect (Charles, C. 2021, June 10).

Police misconduct is often overlooked, unreported, and concealed with the assistance of fellow officers and labor unions. While top officials may call for change, it is essential that officers on the ground take action to implement reform. Currently, change remains an idea with little progress made.

It is important to note that the long history of targeting and mistreatment does make minorities more prone to disrespect authority or have a negative attitude towards the police, which may result in an increase of violent or non-cooperative actions towards the police (Dunham, R. G. 1976).

Police reforms are very crucial in the sense that the reformation of police, the increased diversity in the police force may be able to water down the systematic discrimination in place today. It will be very interesting to see how the increase in minority police can help change the police force (Reiner, R., & Spencer, J. 1996).

EDA

Dataset Description

Our dataset is collected by the Toronto Police department that documents the Arrests, Booking and Strip Search conducted by police officers upon individuals from different social-demographic backgrounds carried on between the year 2020 & 2021. Describing the dataset, there is “Arrest_Year” and “Arrest_Month” which specifies the period of which the individual is arrested. There are three Identification Numbers, firstly the EventID is where the Identification number of the arrest. In our assumption ArrestID is most likely the identification number of the police officer. Furthermore in our assumption, PersonID is the identification number of the individuals being arrested. The assumption is based on mostly matching social demographic background following the identical PersonID. Following the identification numbers, there exhibits social demographic backgrounds such as “Perceived_Race” stating the racial background of the arrestee (White, Black, Indigenous, South Asian, East & South East Asian, Middle Eastern, Latino or Unknown & Legacy). Following is two age indicators, one is “Age_group__at_arrest_” and another is “Youth_at_arrest__under_18_years” categorizing arrestees into different age level and whether they are a youth or not. “ArrestLocDiv” is a column stating the location of the division that made the arrest, if the location is in an exterior jurisdiction, it will be marked with a XX. The columns “StripSearch” and “Booked” records further steps taken after the arrest, it is important to note that while the dataset does not show, if Strip Search happened, a booking took place. Some areas of the data set have a 1 for Strip Search and 0 for Booked, which is incorrect. The column “Occurrence_Category” indicates why the arrest occurred, ranging from assault, warrant, fraud to murder. The following columns are related towards the actions the individuals took after being arrested, most are aggressive stances with one cooperative option. Then it gave the officer columns to record why they performed the search and if there were any items found.

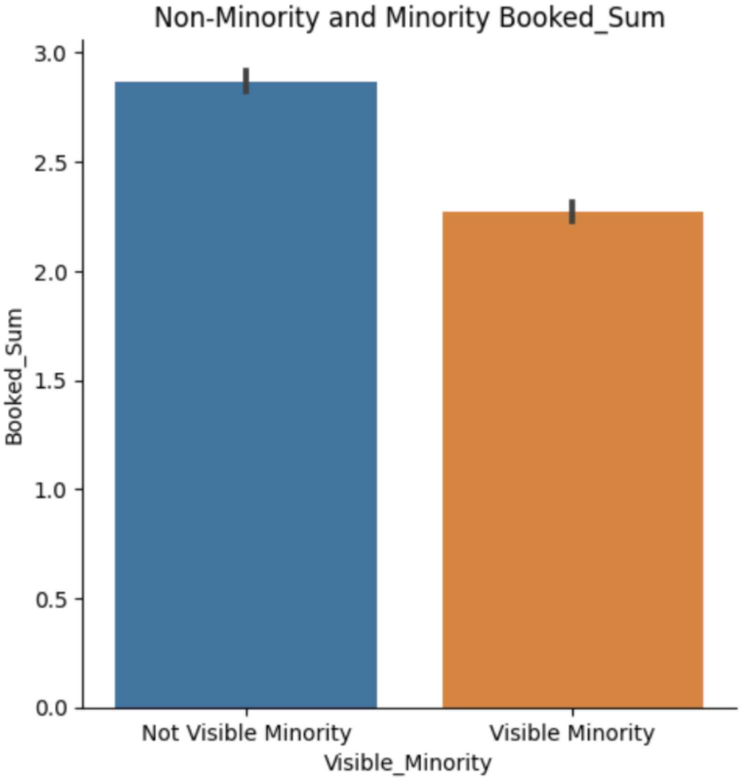
Data Cleaning

I performed several data cleaning steps to better analyze the dataset as it is quite messy. Firstly, I addressed the issue of 9 ‘U’ genders by removing the observations as they will skew our analysis and would not be significant to our final result. Secondly, I addressed the issue mentioned above where there are occasions when Strip Searched happens but Booked is not recorded by automatically changing Booked to “1” if Strip Search equals to 1. Moving forward, I replaced all occurrence categories to either

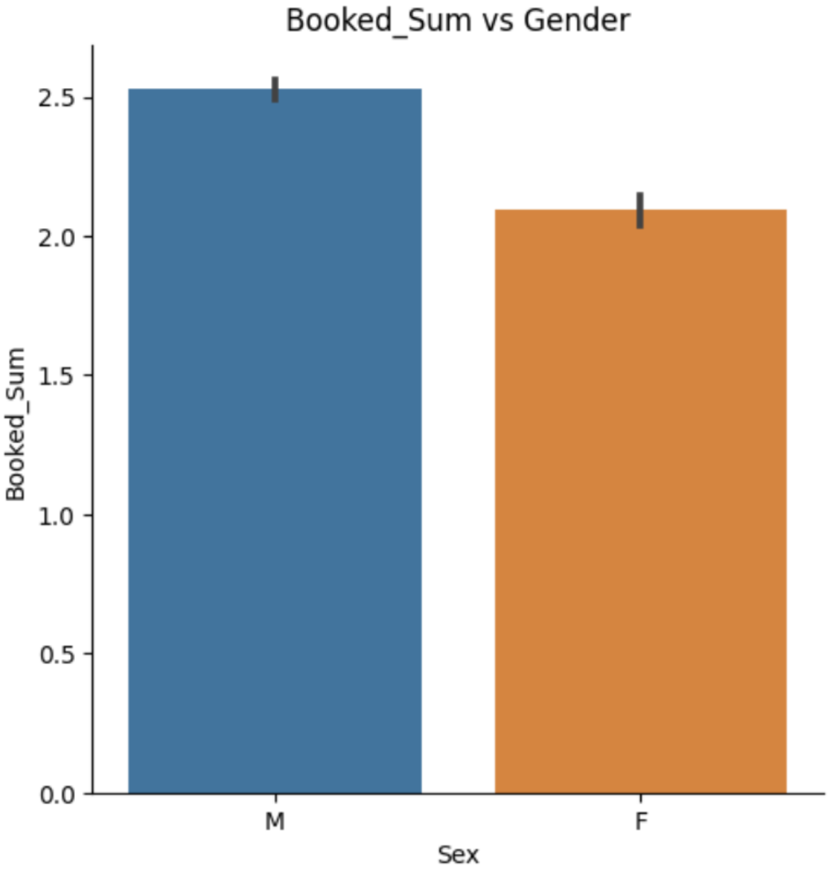
Violent Crimes or Non Violent Crimes. I also changed race into Not Visible Minorities which includes White and Indigenous and Visible Minorities which includes every other race as per the Canadian guidance on race. I created several new variables, the first new variable created is “Resisted Arrest” which records the sum of the number of records between the following variables 'Actions_at_arrest___Concealed_i', 'Actions_at_arrest___Combative__', 'Actions_at_arrest___Resisted__d', 'Actions_at_arrest___Mental_inst', 'Actions_at_arrest___Assaulted_o'. I also created another variable named ‘Booked_Sum’ which is the total amount of Booking a person with an unique person ID received.

Descriptive Statistics

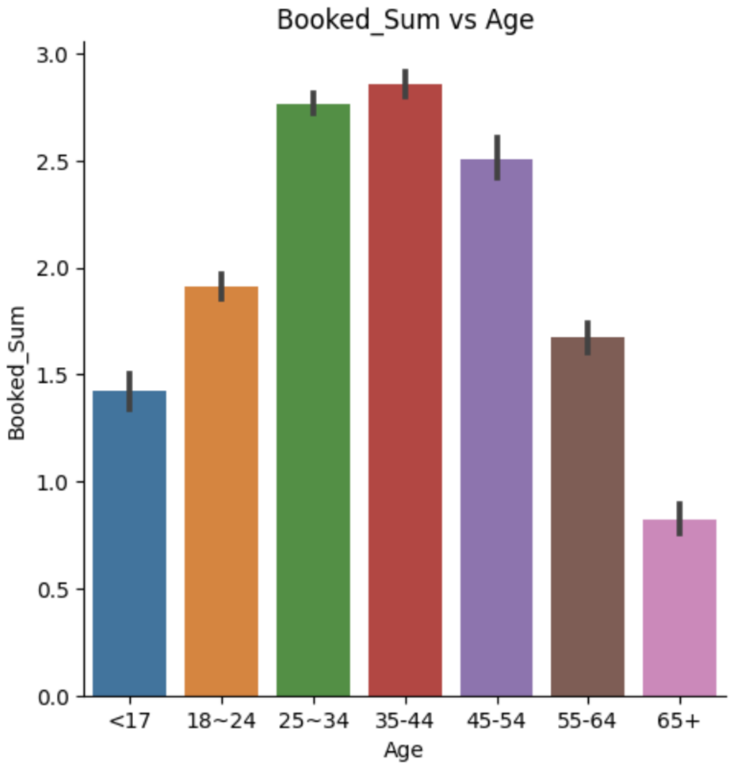
For the analysis, there are several key statistics that we want to highlight. First and foremost, the analysis would be centered around Police Bookings, which refers to the Booked and Booked_Sum variables. The independent variables that we would like to analyze include Visible Minority, Gender, Age, Occurrence_Category, Resisted_Arrest and Cooperative at Arrest.



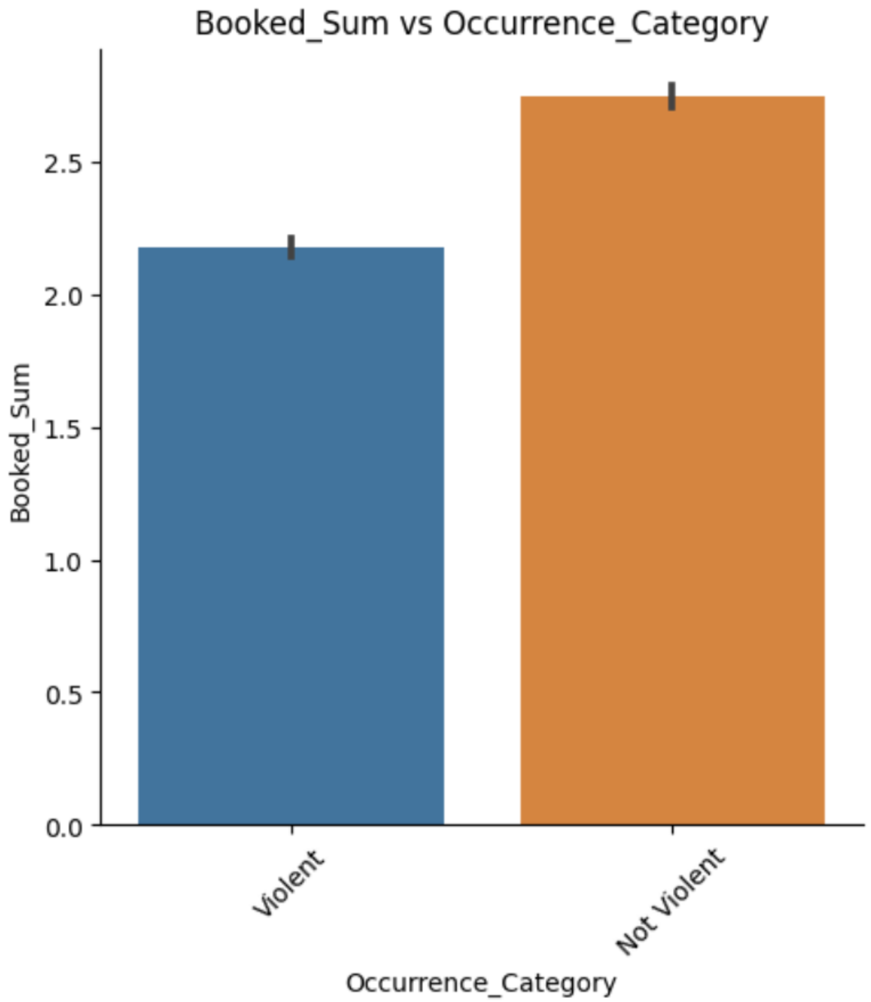
Let us begin with the variable “Visible Minority”, those who were identified as a visible minority are marked on the right as Visible Minority and those who were identified as White or Indigenous are marked on the left as Not Visible Minority. Upon preliminary investigation, it is to seem that Minorities receive less sum booking overall in comparison to those who are Not Visible Minority. However, as we see later on in the study, this may be a misleading indicator as we uncover more information through further analysis.



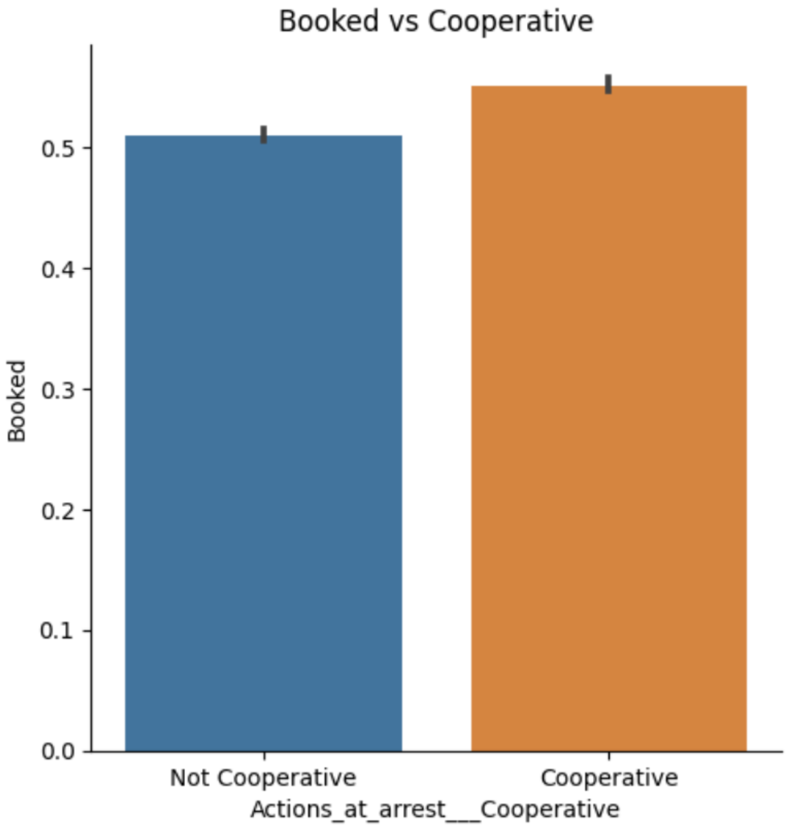
From the variable Gender, we observed that those who are identified as Male are likely booked more often than those who are identified as Female. From which, we would be inclined to include it in our further analysis because there seems to be a significant difference between the two.



Age is an interesting factor as it presents an increase in Booked Sum before 35 years old and a decrease in booked sum after 44 years old, creating a maximum point in the range between 35-44 years old.



Contrary to what I hypothesized beforehand, individuals who have not committed a violent crime are booked more than the individuals who did. This may be related to the survivorship bias, as individuals who have committed violent crimes are more likely to be locked away. Furthermore, Not Violent crime does include actions such as parole check which would be repeated visits upon the same individual.



From the result of the histogram, we see that individuals who are cooperative during arrests are more likely to be booked. This is a very interesting phenomenon however the difference is not as significant between the Cooperative and Not Cooperative.

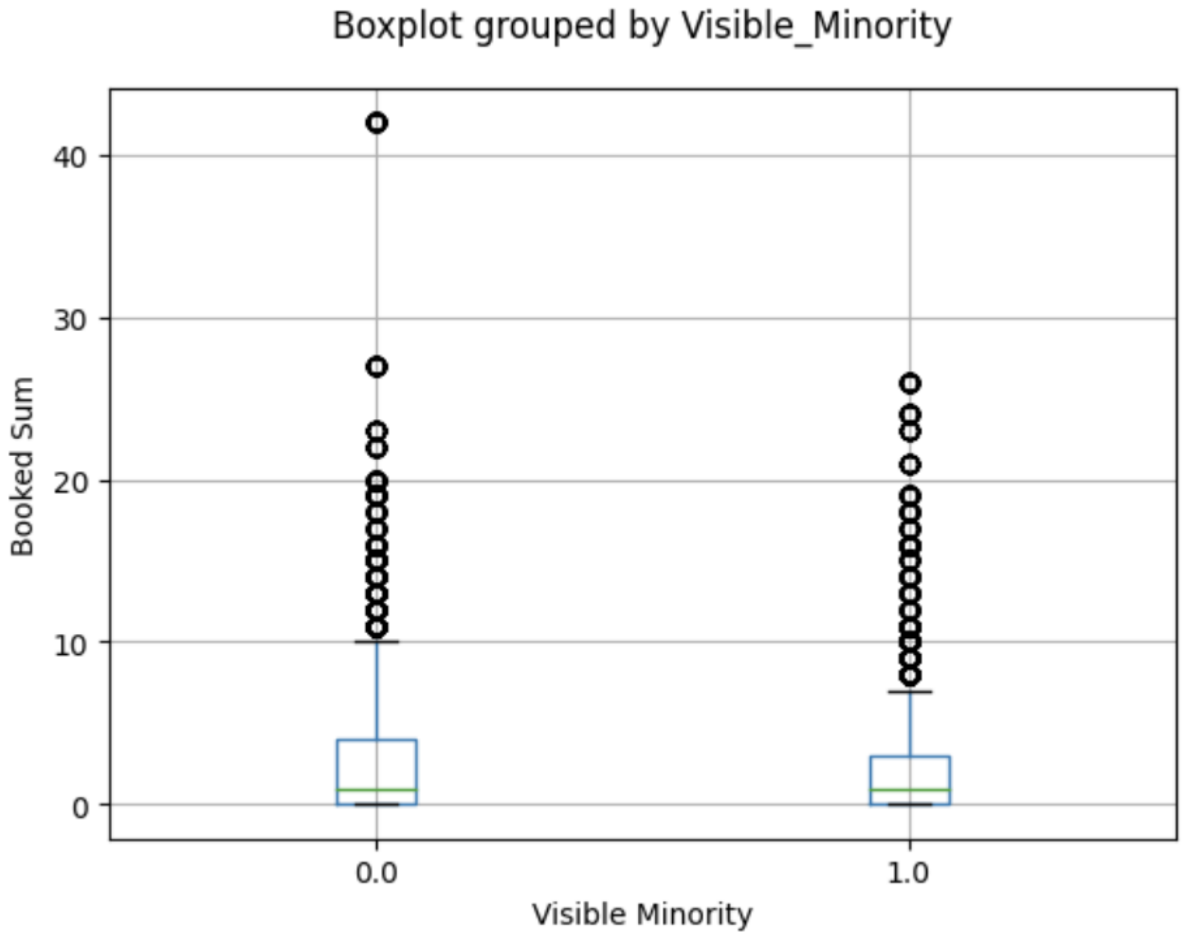
Two Sample T Tests

We would like to conduct several two sample T Tests to examine if there truly exhibit a difference between the two variables.

Two Sample T Test: Visible Minority

T Statistic	-18.60
P Value	5.8326e-77

According to the P Value, there is a significant difference between the Booked_Sum number for visible minority and non visible minority. Specifically, the amount of booked_sum for visible minority is lower than those who are not described as a visible minority.To better illustrate the condition, we use a box plot to demonstrate the difference.

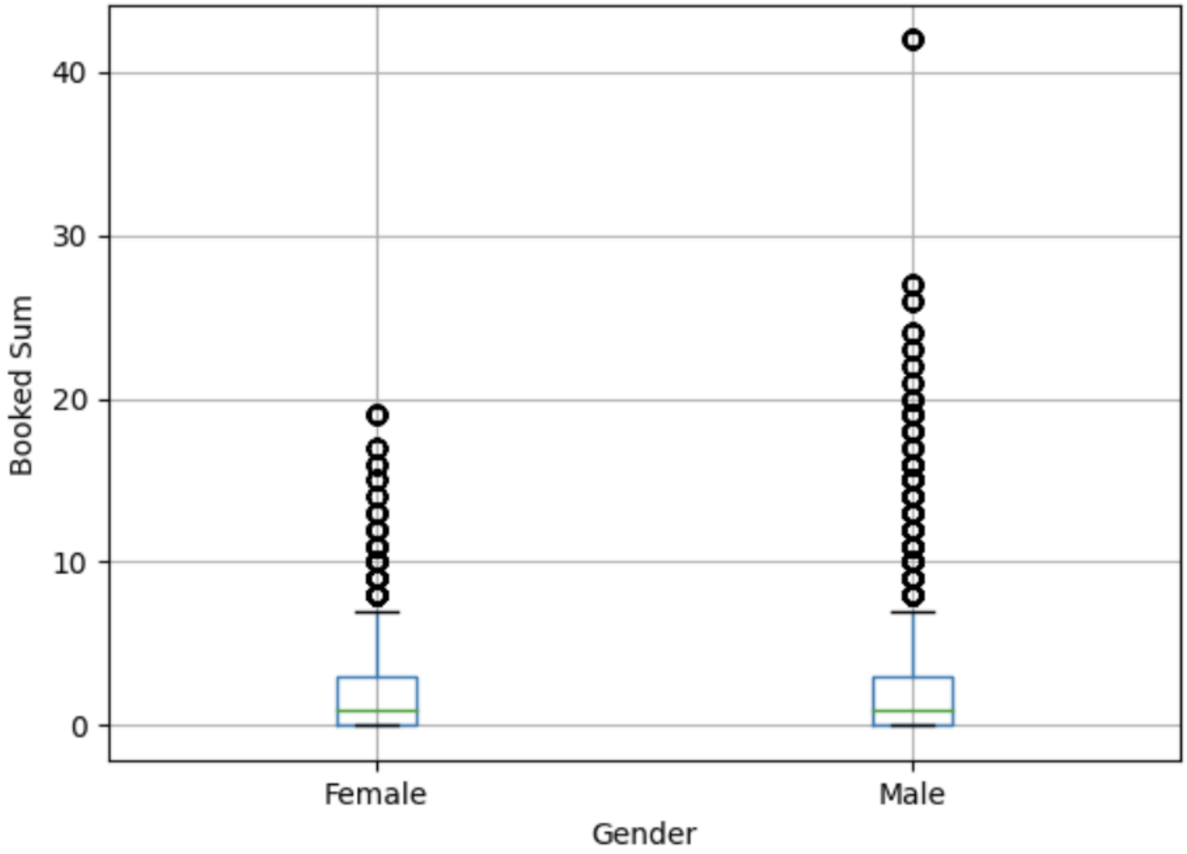


The box plot shows that those who have been identified by the police as a Visible Minority have a lower Booked_Sum rate than those who have not been identified as a Visible Minority. However, it is important to note that there are lots of outliers in both cases that skews the plot.

Moving on to Gender, which is another variable of interest. We want to use Two Sample T Test to determine the significance of the variable. As shown below, the P Value for the gender variable is very tiny, suggesting that there exists a significant difference for the variable Gender. The T Statistic suggests that there is an increase in the mean of booked_sum when gender changes from female to male. We then used Boxplot to visualize the difference between Male and Female.

Two Sample T Test: Gender

T Statistic	13.41
P Value	8.1426e-41



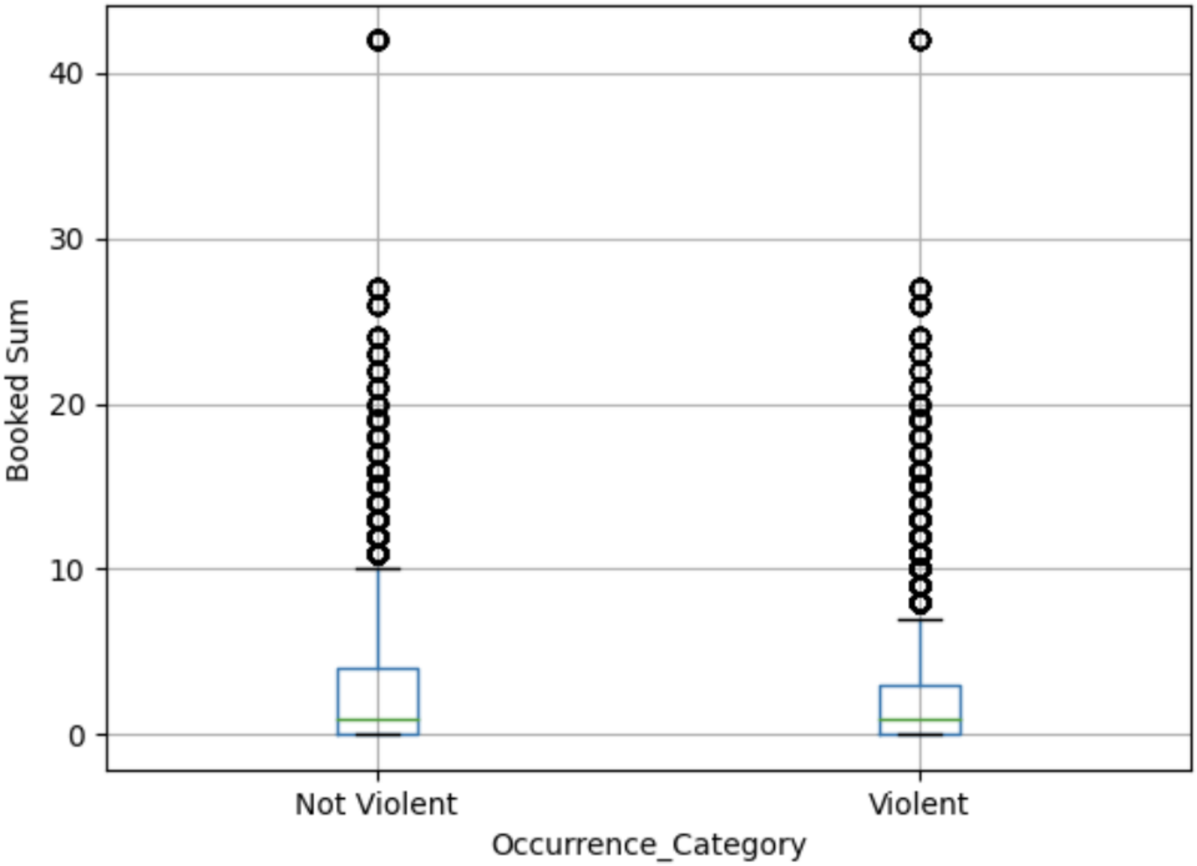
As we see from the boxplot, there is not a very significant difference between the mean of both genders. However, we examine a higher amount of outliers from the Male side. Due to the outliers, it may alter the plot making it seem like the difference between the means are small.

I was unable to produce the two sample T Test for the Occurrence Category, the violent and non violent crime indicator is uncertain.

Two Sample T Test: Occurrence Category

T Statistic	nan
P Value	nan

Using the boxplot function, we produced the boxplot of booked_sum of violent and nonviolent crimes. However, the boxplot does not generate sufficient information as well.



Method

To analyze the dataset, I will be using three different models. Initially, we will check the assumptions for the variables. Then we will go to the first model, which is the Power Analysis, of which we will examine the effect size of each variable. Afterwards we will be answering research question 1 through Ancova. Finally, we will use Logistics regression to answer research question 2.

Assumption Checking

In order to proceed with the analysis, we need to examine whether the assumptions conditions are met or not. There are five main assumptions that we need to check, the first is Independence of Observation, the second Normality, the third is Homogeneity of Variance, the fourth is Linearity, the fifth is Parallel of Regression Slope.

The first assumption check can be completed through examination of the dataset. Each observation is isolated from each other, thus the dataset satisfies the first assumption.

The second assumption check is a normality check. In order to check normality of the independent variables, we apply the Anderson Darling test to check if the data is normal. In response, all of the variables that have been mentioned from our EDA are not normal thus it does not satisfy our assumption.

The third assumption test is the homogeneity of variance. This can be conducted through the levene test. From our result, none of the variables satisfy the levene test therefore the assumption is not met as well.

The fourth and fifth assumptions are Linearity and Parallel of Regression Slope. Most of the data we have included has a binary result, thus we do not have to investigate the linearity and the parallelism of the regression slope.

Thus, in conclusion, the data does not satisfy the assumptions required for Ancova analysis. However due to the big amount of data, we can overlook the violation of assumption.

Power Analysis

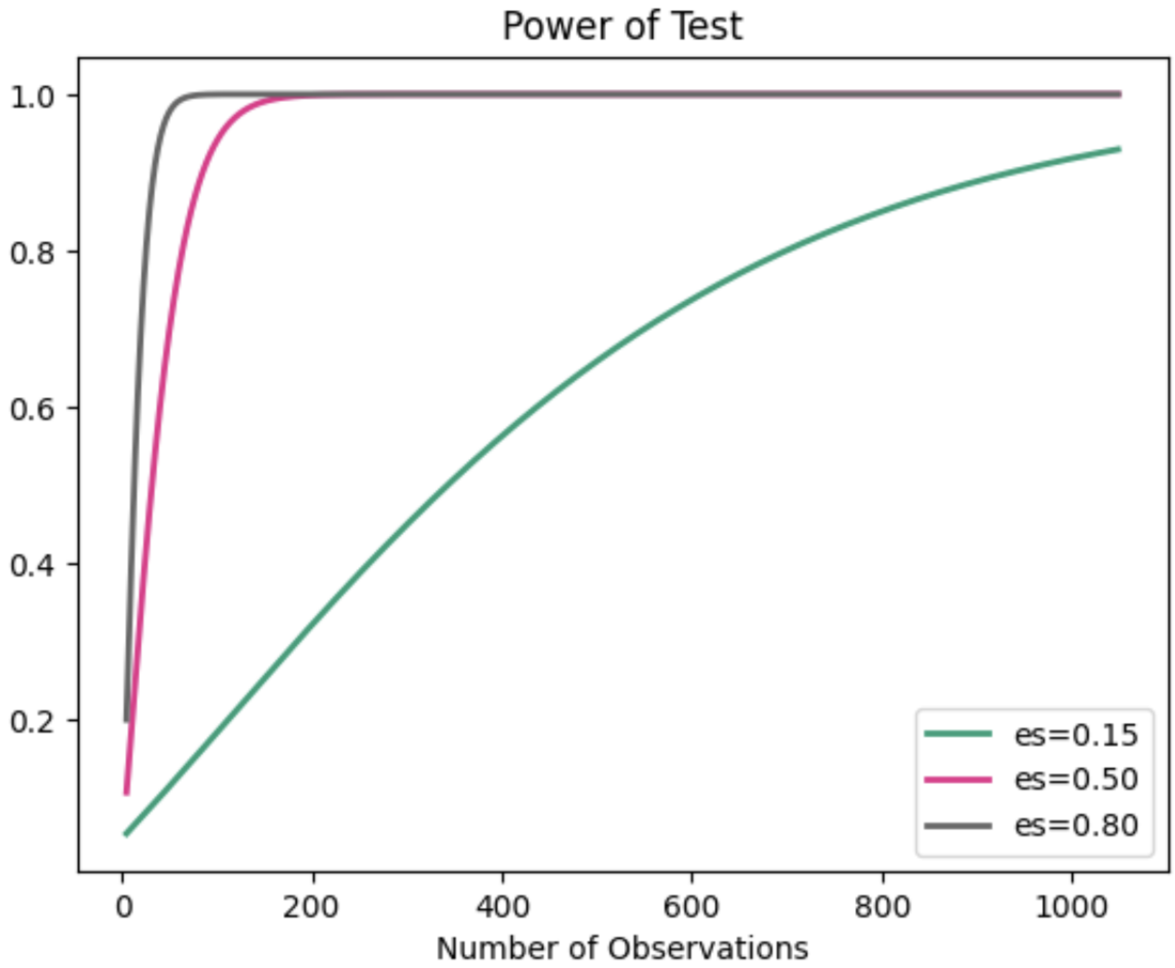
Prior to completing the Ancova analysis and Logistic Regression, we want to analyze the effect size of the variables that we are interested in and the sample size necessary for analysis. In this section, we will analyze five different variables, Visible Minority, Resisted Arrest, Age, Gender and Cooperative Response.

For the variable Visible Minority, we first investigated the effect size of the variable. Under the condition of alpha = 0.05 and power = 0.8, we are able to calculate that the variable has an effect size of 0.1598. Which suggests that the variable has a small effect and influence on the outcome. We then want to examine whether the sample size is great enough for our analysis.

Sample Size Analysis for Power = 0.8

Sample Size Needed Group 1	706.365
Actual Size Group 1	29652
Sample Size Needed Group 2	545.663
Actual Size Group 2	22906

From our analysis, we discovered that the sample size available is a lot greater than the sample size required. Thus we can safely assume that the sample size available is able to generate sufficient power.



As the plot above shows, for an Effect Size of 0.15 and Power of 0.8, it needs around 700 observations. Our dataset include more than 700 observations, again concurring with our analysis.

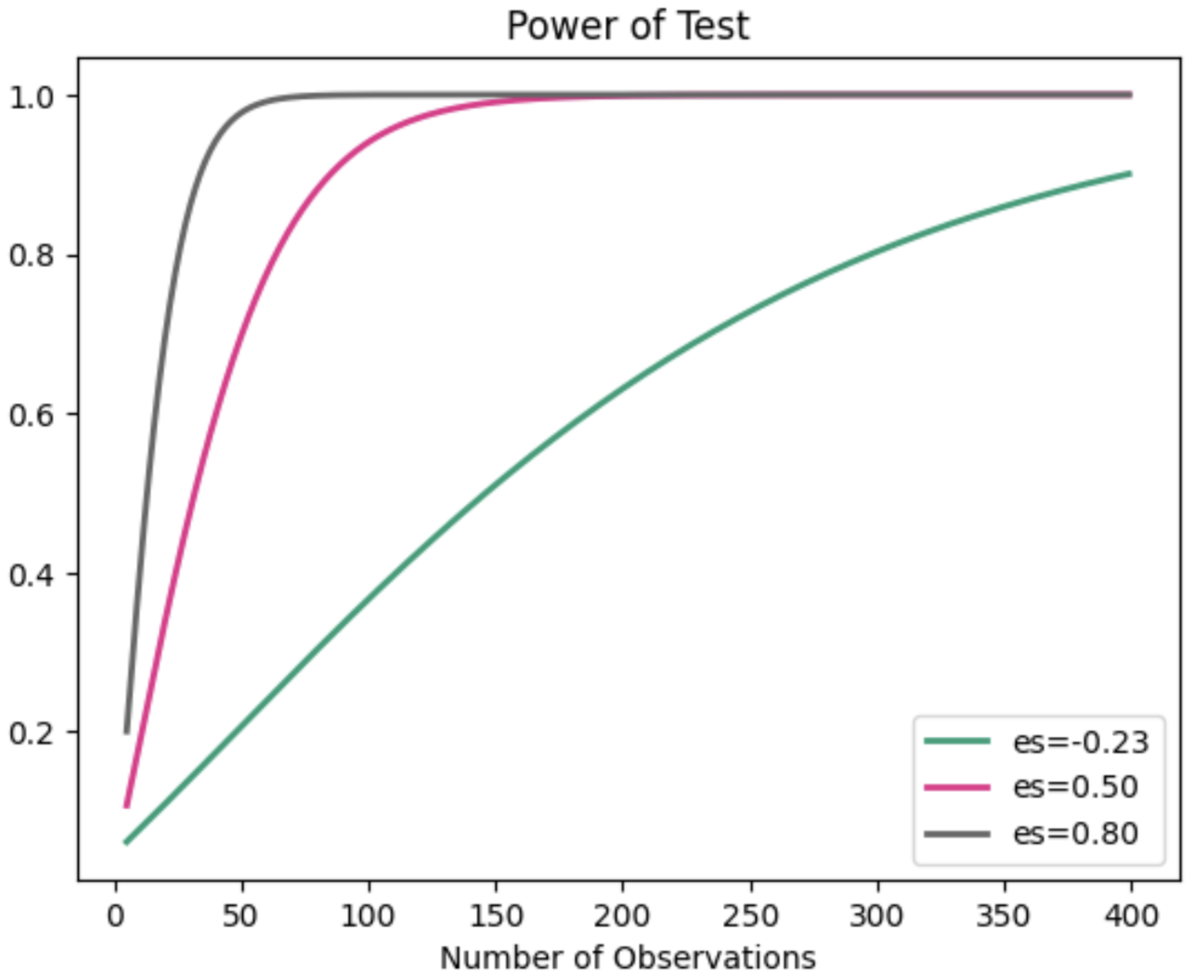
For the variable Resisted Arrest, we first investigated the effect size of the variable. Under the condition of alpha = 0.05 and power = 0.8, we are able to calculate that the variable has an effect size of -0.238. Which suggests that the variable has a negative and small effect and influence on the outcome. We then want to examine whether the sample size is great enough for our analysis.

Sample Size Analysis for Power = 0.8

Sample Size Needed Group 1	328.27
Actual Size Group 1	37563
Sample Size Needed Group 2	242.11

Actual Size Group 2	27704

From our analysis, we discovered that the sample size available is a lot greater than the sample size required. Thus we can safely assume that the sample size available is able to generate sufficient power.



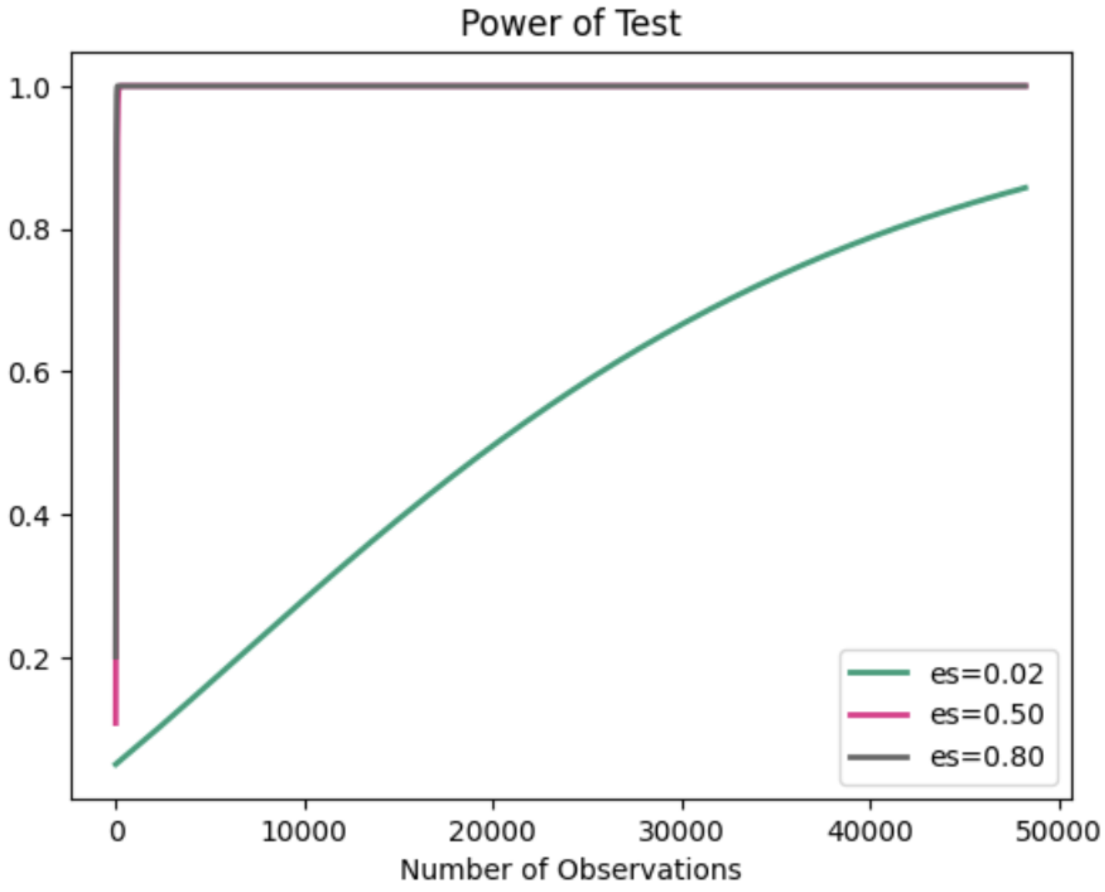
As the plot above shows, for an Effect Size of -0.23 and Power of 0.8, it needs around 325 observations. Our dataset include more than 325 observations, again concurring with our analysis.

For the variable Age, we first investigated the effect size of the variable. Under the condition of alpha = 0.05 and power = 0.8, we are able to calculate that the variable has an effect size of 0.02. Which suggests that the variable has a tiny effect and influence on the outcome. Due to the small size of the effect, we may not need to include the variable in our future analysis. We then want to examine whether the sample size is great enough for our analysis.

Sample Size Analysis for Power = 0.8

Sample Size Needed Group 1	48252.86
Actual Size Group 1	37563
Sample Size Needed Group 2	35588.14
Actual Size Group 2	27704

From our analysis, we discovered that the sample size available does not satisfy the need of the for the power to be 0.8. Thus we will not be able to include the factor for future analysis.



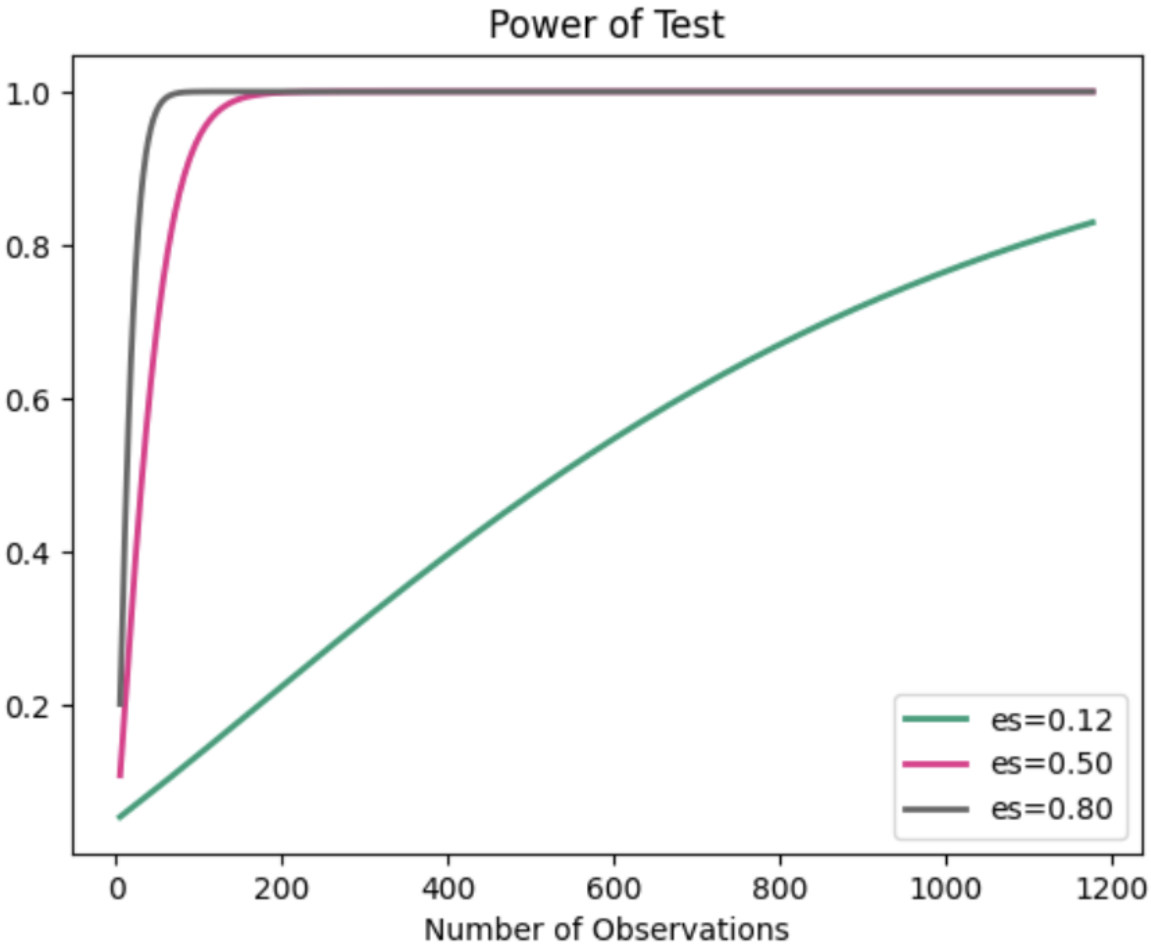
As the plot above shows, for an Effect Size of 0.002 and Power of 0.8, it needs around 45000 observations. Our dataset include less than 45000 observations, again concurring with our analysis that we cannot include the Age factor in our model.

For the variable Cooperative Response, we first investigated the effect size of the variable. Under the condition of alpha = 0.05 and power = 0.8, we are able to calculate that the variable has an effect size of 0.122. Which suggests that the variable has a small effect and influence on the outcome. We then want to examine whether the sample size is great enough for our analysis.

Sample Size Analysis for Power = 0.8

Sample Size Needed Group 1	1178.925
Actual Size Group 1	36161
Sample Size Needed Group 2	948.917
Actual Size Group 2	29106

From our analysis, we discovered that the sample size available is a lot greater than the sample size required. Thus we can safely assume that the sample size available is able to generate sufficient power.



As the plot above shows, for an Effect Size of 0.12 and Power of 0.8, it needs around 1250 observations. Our dataset include more than 1250 observations, again concurring with our analysis.

Of the five variables we tested with our power analysis, we are able to pass the test upon four different variables and isolated age as the sole indicator to not include in further analysis. We will use this information to our advantage by not including Age in our Ancova and Logistic Regression Research Question.

Ancova Research Question

The research question set up for Ancova is to determine the factors that influence the amount of Booking an individual receives.
Formally, the research question is as follows “*What is the relationship between the total number of times an individual is booked and their occurrence category, visible minority status, whether they resisted arrest, and their gender*?"

The research question includes both social demographic information along with action conducted by each individual. The justification for the research question is further strengthened by its ability to flag repeated offenders, the crime they commit and how dangerous they are to frontline officers.

Our model is as follows
Booked_Sum ~ Occurrence_Category + Visible_Minority + Resisted_Arrest + Gender.

From our Ancova analysis, we would like to highlight several results that we find particularly interesting.

The first result that we are interested in is the R-squared value, which is an indicator of how well the independent variables explain the dependent variable. The R-squared value for this model is 0.031 which suggests that the model only explains 3.1% of the dependent variable booked_sum.

The second result we want to interpret is the intercept of the model. We see that the mean amount of Booked_Sum is at 2.60. Which suggests that each unique PersonID on average is booked 2.6 times during the two years of observational period.

The third result that we want to interpret is the variables coefficient and p-value. In order to better fit the research question, we will separate the four variables into social-demographic background and actions at arrest.

Lets begin with the social-demographic background related variables,individuals who are identified as visible minorities have recorded less amount of booked_sum than those who are not identified as visible minorities. The difference between them is 0.6124 bookings. The p-value for this coefficient is less than 0.05, which suggests that this is a statistically significant relationship. The analysis on the variables gender suggests that male individuals have an average of 0.5420 more bookings than female individuals. The p-value is also less than 0.05, which suggests that this is a statistically significant relationship.

Moving on towards actions at arrests. We observe that violent criminals are booked less than non-violent criminals. The difference between the two is 0.5665 average less bookings during the period. This may be due to the amount of sentencing an individual receives after committing a violent crime. The p-value for this coefficient is less than 0.05, which suggests that this is a statistically significant relationship. Lastly, we want to investigate the variable Resisted_Arrest, individuals that have resisted arrest have a higher amount of booked_sum during the period of time, the difference is very significant compared to the previous three. Individuals who have resisted arrest on average are booked 1.08 times more than those who do not resist arrest. The p-value for this coefficient is less than 0.05, which suggests that this is a statistically significant relationship.

Logistic Regression Research Question

In relation towards research question 1, we want to investigate variables that affect how likely it is for an individual that is arrested to be booked.

Formally, the research question is as follows. "*How is the likelihood of being booked affected by occurrence category, visible minority, cooperative actions at arrest, gender and Resisting Arrest.*"

Firstly, we want to justify the research question by indicating that we want to analyze if there is any prejudice in social demographic background such as being a visible minority or by gender. Furthermore, we hypothesize that violent crimes and reactions at crime scenes may be a stronger indicator of bookings.

The Rsquared value of the logistics regression model is really low as it sits at 0.02346, which suggests that the model only explains 2.35% of the variance from the booked variable. The model does not explain the dependent variable well.

The reference group for our logistic regression refers to a non-violent crime, non visible minority, female, did not cooperate but did not resist arrest as well. Now let's dive into the analysis deeper.

While responding to a violent or non violent crime, there is a small difference between bookings and non-bookings. While violent criminals are 1.013 times more likely to be booked, the p-value sits at 0.534, which suggests that the difference is not statistically significant.

Sex does in fact have a strong influence on whether or not the arrest turns into a booking. The odds ratio is 1.487 which suggests that male is 1.487 times more likely to be booked than female suspects, the p-value is below 0.05 which indicates that there is statistically significant evidence that there is a difference between male and female.

Being cooperative at arrest does not help an individual avoid being booked. The odds ratio is 1.304 which suggests that those who are cooperative are 1.304 times more likely to be booked than those who are not cooperative, the p-value is below 0.05 which indicates that there is statistically significant evidence that there is a difference between cooperative and non-cooperative.

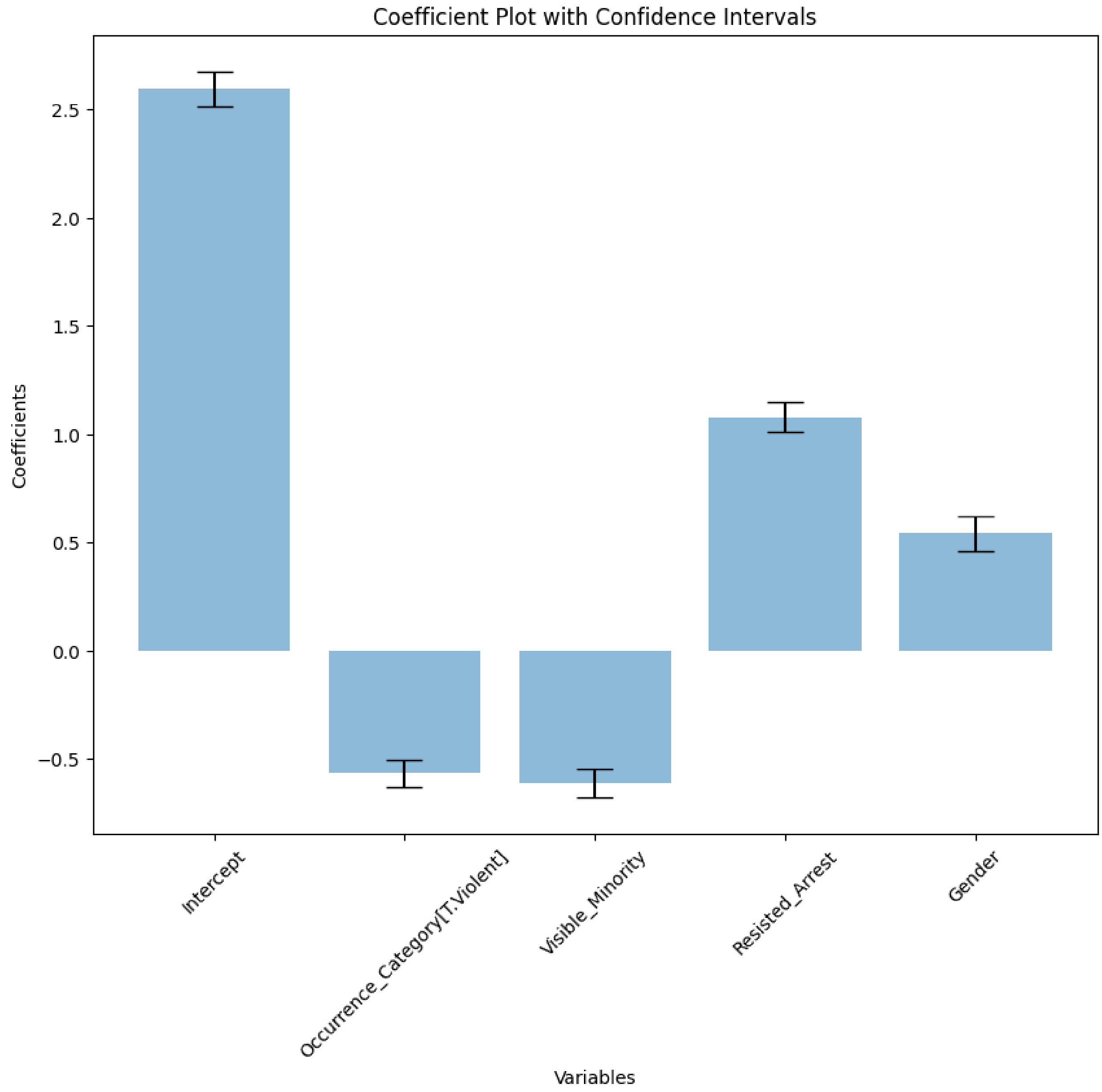
The variable that impacts the likelihood of arrest the most is resisting arrest, resisting arrest have a very high odds ratio of 2.325, which suggests that for every additional resisting arrest action taken, the likelihood of the individual being booked increased by 2.325 times. The p-value is less than 0.05 which suggests that the result is statistically significant.

Results & Findings

After running the Ancova analysis, we were able to derive some insightful results. Even though our model does not explain the dependent variable well, we are able to make some preliminary findings and formulate the bigger picture for the future. Firstly, those who resisted arrest are often more dangerous than those who do not resist arrest. There is a significant indicator that male are more likely to have more bookings during the two year period than females. Individuals that are identified as visible minorities (Black, Asian, and Latino) have less number of bookings in the two years period than non-visible minorities (Indigenous and White). Individuals who committed a violent crime have less number of bookings than those who did not commit a violent crime.

Ancova Analysis Result

	Coefficient	1st Quantile	3rd Quantile	P-Value
Intercept	2.5958	2.514	2.677	0.000
Occurrence_Category[T.Violent]	-0.5665	-0.63	-0.503	0.000
Visible_Minority	-0.6124	-0.676	-0.549	0.000
Resisted_Arrest	1.0800	1.009	1.151	0.000
Gender	0.5420	0.463	0.621	0.000

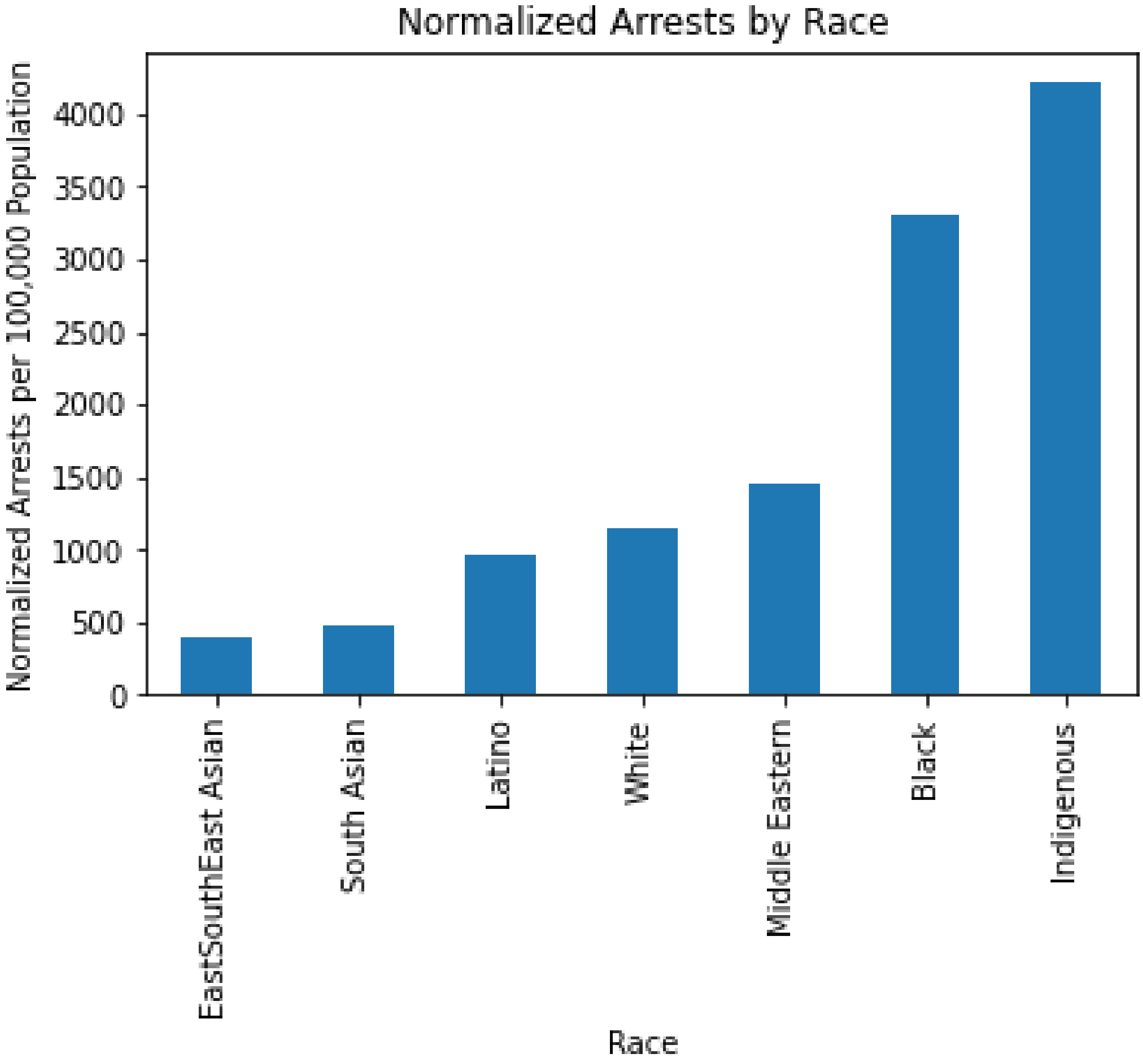


In addition to discussing the sum value, we want to investigate whether each of the variables affect the likelihood of receiving a booking. In this scenario, most variables stayed the same response.

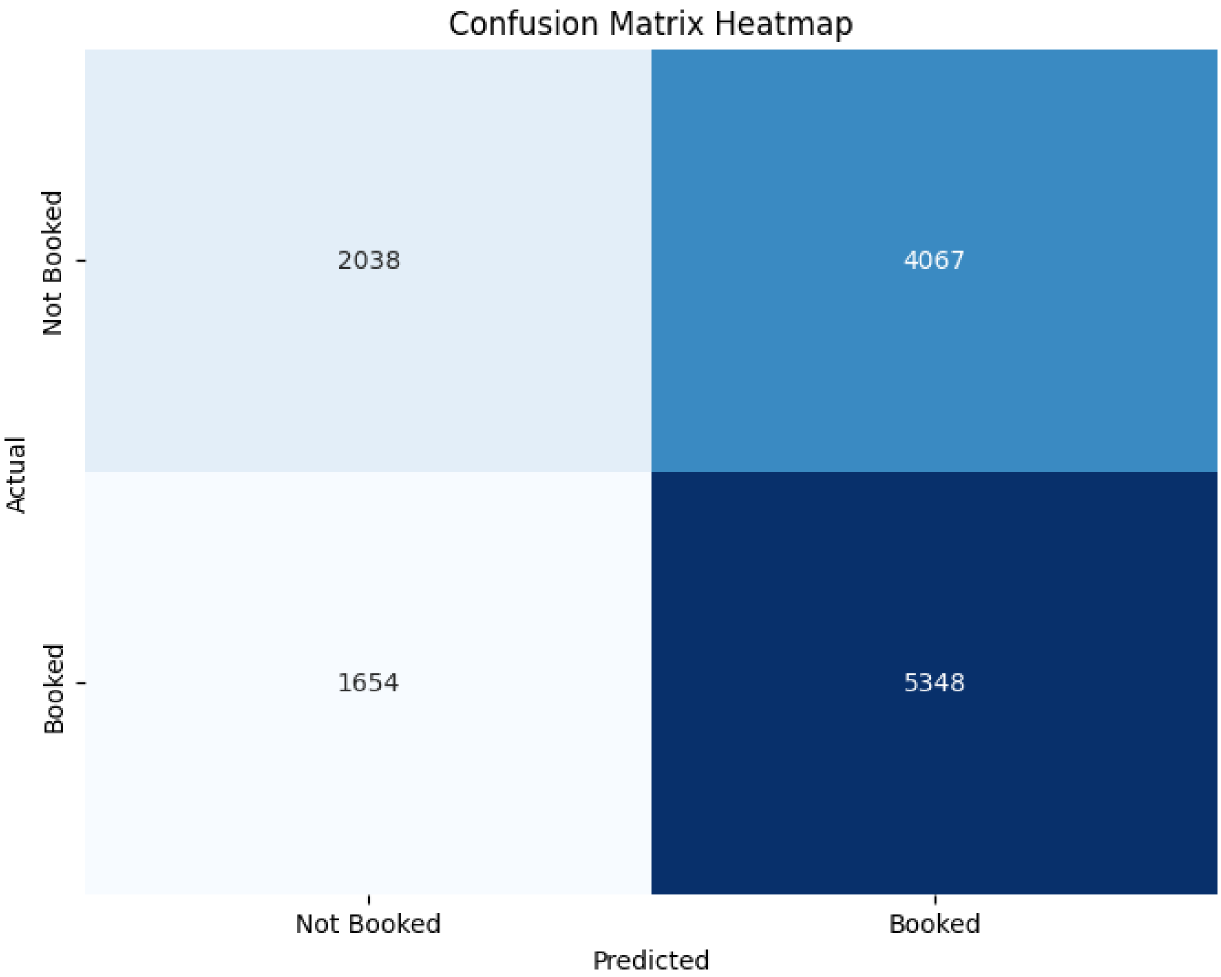
Logistic Regression Results

	Lower CI	Upper CI	OR	P Value
Intercept	0.628	0.702	0.664	0.000
Occurrence_Category[T.Violent]	0.973	1.055	1.013	0.534
Sex[T.M]	1.414	1.564	1.487	0.000
Actions_at_arrest___Cooperative	1.251	1.359	1.304	0.000
Visible_Minority	1.026	1.113	1.068	0.002
Resisted_Arrest	2.191	2.467	2.325	0.000

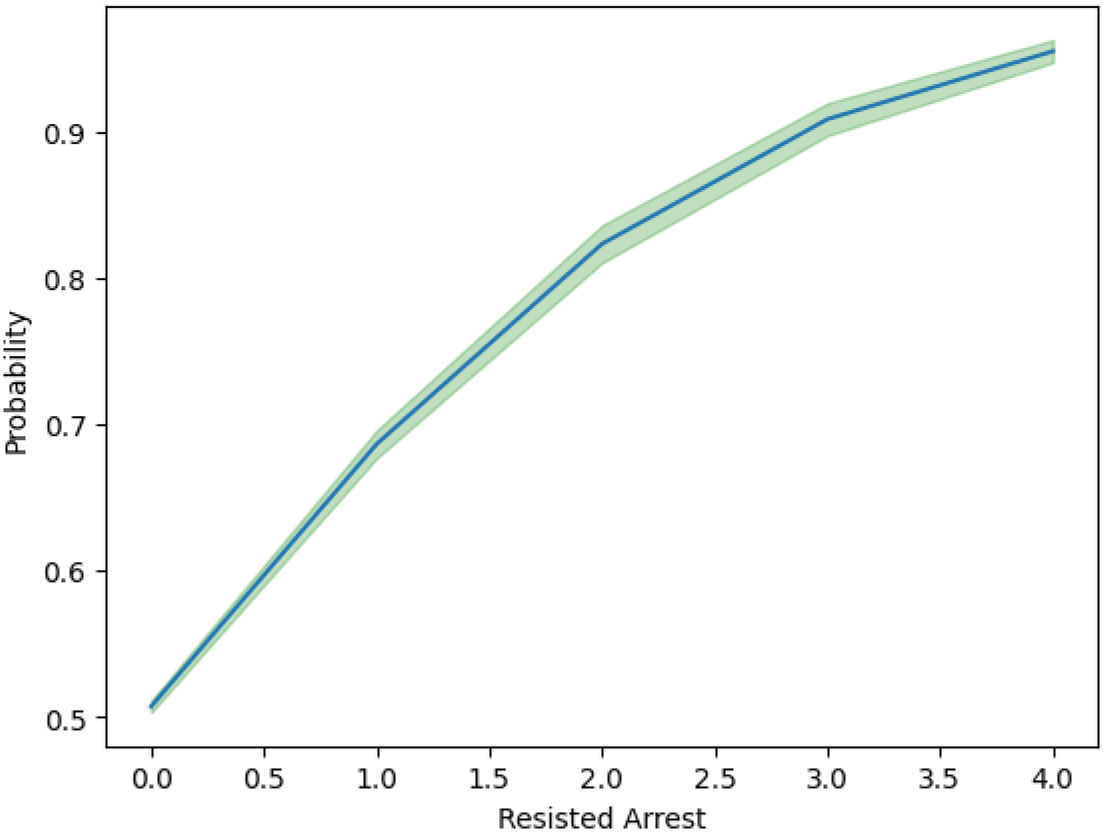
However, the variable Visible Minority gave a different response. While looking at individual arrests and bookings, we see officers are more likely to book a visible minority than those who are not a visible minority. If we are establishing a just-cause analysis, when officers have a just-cause to stop and arrest an individual, they are more likely to book them if they are a visible minority. Which may be a result of racial profiling or other issues as we have observed previously that visible minorities commit less crime as a population. From the previous midterm study, police target Black, Middle Eastern and Indigenous individuals the most after normalizing the data.



We also reached a conclusion that the dataset is very complex and arresting, booking and stripsearching is a very complex issue.



The overall prediction accuracy for the Logistic Regression is not very high as it sits at 55%. However, the model is anticipating individuals to be booked and not booked which suggests that it is still rather active in participation.



Using solely Resisted Arrest to test the likelihood of being arrested, one can determine if an individual shows all signs of resisting arrest they will most likely be booked. With each additional instance of resistance showing greatly increasing the likelihood of being arrested.

Limitations

There are two major limitations from the dataset that we face as major obstacles. The first is that the dataset is messy and not normal. This is acceptable due to the big amount of sample size which outweighs the con of having not normal data. Furthermore, most of the analysis is done with binary yes or no indicators, thus negating more of the pressure off of not normal data.

Another limitation that we have to address is that we do not know what happened after the bookings. We do not have data that shows if the arrestee is locked up or not. This is the first idea that came to mind after we observe the fact that violent criminals are booked less than non violent criminals. However, it is important to note that the likelihood to be arrested does not have a statistically significant difference between violent and non violent criminals.

The data collection method and entry can be perfected furthermore as there are inconsistent inputs of variables such as gender and age. There are no suitable descriptions available for some columns and variables which makes it difficult to be processed with sound logic.

Conclusion

In this study, we conducted Power Analysis to determine the effect size of several variables including Visible Minority, Resisted Arrest, Age, Gender and Cooperative Response. After investigating the variables against the dependent variable Booked_Sum, we decided to omit Age as the variable has a very limited effect size. We then move on to Ancova analysis to determine the effect that each variable has on the total amount of individual bookings in the two years period. Afterwards, we conducted an analysis on how the variables affect the likelihood of an individual getting booked. Furthermore, we created a confusion matrix to visualize the accuracy of the model along with a prediction interval for the variable Resisted Arrest.

Overall, the model does not portray the dependent variable perfectly as the r-square is low, however, it does give some interesting insight to build off for future studies.

Discussion

There are some routes for future studies that can be built off of this study. The first is to create a better model including more variables to measure a higher amount of variance in our dataset. The second is to create an in depth analysis based on more data collected focusing on the difference in booked sum and booking likelihood of violent and non violent crimes as it defies logical reasoning that those who are suspected to have committed violent crimes are able to be not booked. Thirdly, we want to investigate why colored and visible minorities are booked less frequently than non visible minorities while having a larger chance of being booked.

Citation

BBC News. (2021, June 7). Toronto police apologise for historic mistreatment of LGBTQ2S+ community. https://www.bbc.com/news/world-us-canada-61818396

Charles, C. (2021, June 10). The Toronto police apology for its treatment of racialized people is meaningless without action. The Conversation. https://theconversation.com/the-toronto-police-apology-for-its-treatment-of-racialized-people-is-meaningless-without-action-185262

Dunham, R. G. (1976). Minorities and police confrontation in America. National Criminal Justice Reference Service. https://www.ojp.gov/ncjrs/virtual-library/abstracts/minorities-and-police-confrontation-america

Reiner, R., & Spencer, J. (1996). The myth of the "clean pair of hands": The politics of policing reform in Britain. The British Journal of Criminology, 36(1), 109-127. https://academic.oup.com/bjc/article-abstract/36/1/109/606112?login=false

Statistics Canada. (2021). Census Profile, 2021 Census. Retrieved from https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/page.cfm?Lang=E&GENDERlist=1&STATISTIClist=1&HEADERlist=0&DGUIDlist=2021A00053520005&SearchText=toronto