# Case Study 0

Austin Kao

1/31/2021

Introduction

The New York City police department has a Stop-Question-Frisk (SQF) policy where individuals can be stopped by police to be questioned and, in some cases, frisked based on the reasonable suspicion that the individual will commit a crime or is committing a crime. However, opponents of the policy protest that black and Hispanic residents are stopped by police at a significantly higher rate than white and Asian residents, making the practice unconstitutional. Now that a class-action lawsuit has led the NYPD to require justification for any SQF event, we seek to examine any racial disparities that remain in the police rationale to stop residents for violent crimes. We define violent crime to be any crime that involves force or the threat of force, which includes crimes like murder, robbery, and assault.

The data used for this case study comes directly from the NYPD Stop, Question, and Frisk database from the years of 2017 to 2019. Officers who conduct an SQF event are supposed to complete a form recording details about the event. These forms are then entered into the electronic database, which the NYPD releases annually. We also use NYC police precinct data adapted from data compiled by John Keefe, an editor for the New York Times. To determine what justification police officers are providing for executing an SQF event, we used the "SUSPECTED_CRIME_DESCRIPTION" field in the Stop, Question, and Frisk database, which lists the exact crime being investigated. We then classified the listed crimes into violent and nonviolent categories for a new variable called "SUSPECTED_VIOLENCE", which will act as our response variable. "SUSPECTED_VIOLENCE" has a value of 1 if the suspected crime is violent in nature, and 0 if the suspected crime is nonviolent.

Because we are examining why police officers decide to conduct an SQF event, we want to consider any potential details officers will have access to or will have considered before commencing a stop, which are details like height, weight, age, sex, body build type, and how the stop was initiated. We will also consider race as a factor and compare its influence against these other factors to establish how great an impact race has in these decisions and also to examine what racial disparities exist. After an EDA about how each of these factors affects police suspicion of violent crime, we determined that height, age, sex, and how the stop was initiated to be the most significant factors, so we will use these factors in our model. The location of the event can also have an effect, especially since the racial makeup of the city changes by precinct. To account for this fact, we decided to add the proportion of Black and Hispanic residents within the precinct to the model. The variables we have chosen are represented by the fields "SUSPECT_HEIGHT", "SUSPECT_REPORTED_AGE", "SUSPECT_SEX", "STOP_WAS_INITIATED", "PRECINCT_PROP_BLACK", "PRECINCT_PROP_HISP", and "RACE" in the data.

On preliminary examination of the data, it would seem that some racial disparity does exist. When comparing the ratio of stops for suspected violent and nonviolent crimes for each race, black and Hispanic residents are stopped for suspected violent crimes at higher rates than white residents.

Ratio Between Suspected Violent and Nonviolent Crimes by Race
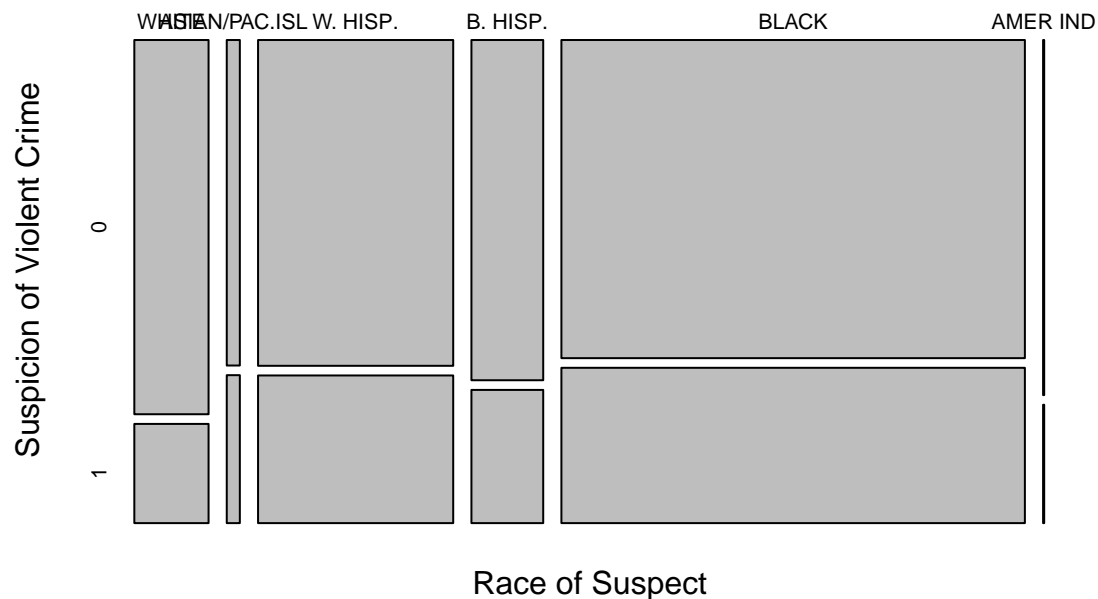
Methodology

Because we are primarily concerned with the race of the suspects at SQF events, we filtered out 427 observations where the race was listed as null or male. Also, to examine SQF events across the years between 2017 and 2019, we combined the data sets from 2017 to 2019 into one data set. Any fields that are present in one data set but absent in another were excluded from the overall data set. The fields are named "SUPERVISING_ACTION_CORRESPONDING_ACTIVITY_LOG_ENTRY_REVIEWED" and "STOP_LOCATION_PREMISES_NAME"are both present in the 2017 and 2018 data sets but absent in the 2019 data, while the fields named "ASK_FOR_CONSENT_FLG" and "CONSENT_GIVEN_FLG"are both present in the 2019 data set but absent in both the 2017 and 2018 data. None of the fields being excluded affect police decisions to execute an SQF event except for perhaps the premises, but other stop location fields are more comprehensive and can act as sufficient substitutes.

To create the "SUSPECTED_VIOLENCE" field, which has a value of 1 if the suspected crime is violent in nature, and 0 if nonviolent, we used the data in "SUSPECTED_CRIME_DESCRIPTION". The crimes we decided were violent are assault, menacing, forcible touching, murder, rape, terrorism, reckless endangerment, and robbery, while all other crimes were classified as nonviolent.

We also removed an additional 3130 observations due to missing or impossible data. If a suspect was missing height or age data, that data point was removed. There were also data points removed that were missing the "STOP_WAS_INITIATED" field. There were also impossible data points, such as a suspect who was recorded as being 60.11 feet tall. This data point was also removed since it is unclear if the officer was using inches or added an extra zero.

To fix some issues with the "SUSPECT_RACE_DESCRIPTION" field, we created a new field called "RACE" where we simplified the names being used and combined several categories that were deemed identical. We combined the "AMER IND", "AMERICAN INDIAN/ALASKAN NATIVE", and "AMERICAN INDIAN/ALASKAN N" values into "AMER IND", and combined "ASIAN/PACIFIC ISLANDER" and "ASIAN/PAC.ISL" values into "ASIAN/PAC.ISL".

In regards to the sex of the suspect, there were some null values present in the data. We decided to label these values as "UNSPECIFIED" since the suspect could have any gender, including non-binary genders that the data cannot capture.

To allow for meaningful interpretation of the intercept, we centered the variables "SUSPECT_HEIGHT" and "SUSPECT_REPORTED_AGE" about their means.

We proceeded to take a simple random sample of 3000 observations out of the remaining 32539 observations to serve as the basis for our model. This also satisfies our model assumption that the data is random.

To address concerns about the independence of the predictor variables, particularly with how race can be affected by where the suspect is stopped, we calculated the proportion of Black residents and the proportion of Hispanic residents in each precinct, based on the 2020 precinct population data. We then added that information in the form of the variables "PRECINCT_PROP_BLACK" and "PRECINCT_PROP_HISP". By adding these variables to our model, we hope to help account for the relationship between location and race in our model.
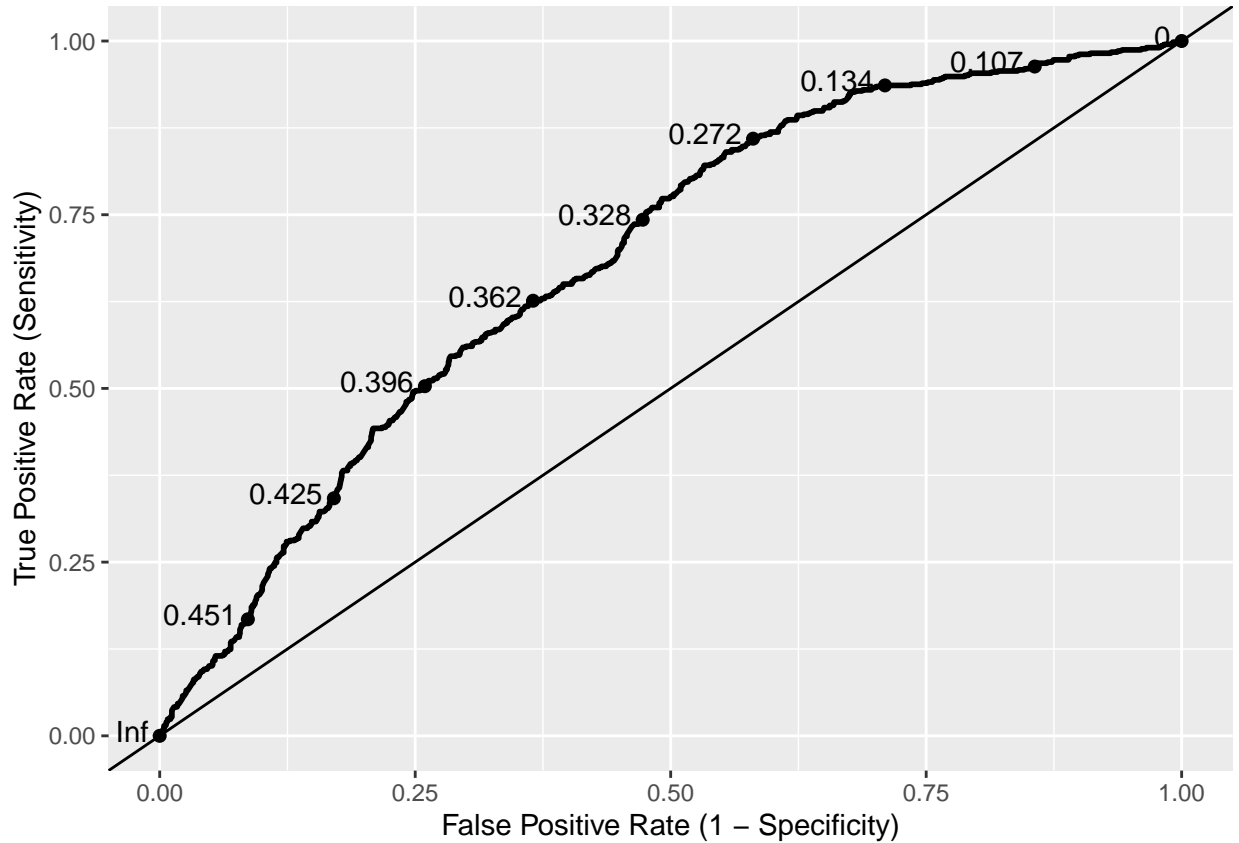
Results

Based off the predictor variables we chose, we generated a model for the log odds that a suspect will be stopped by the police for a violent crime.

```
## # A tibble: 14 x 5
##    term                                  estimate std.error statistic  p.value
##    <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                              -2.38     0.258    -9.22  2.87e-20
##  2 RACEASIAN/PAC.ISL                         0.769    0.536     1.44  1.51e- 1
##  3 RACEW. HISP.                              0.318    0.220     1.45  1.48e- 1
##  4 RACEB. HISP.                              0.0719   0.260     0.277 7.82e- 1
##  5 RACEBLACK                                 0.367    0.205     1.79  7.40e- 2
##  6 RACEAMER IND                            -11.8    325.       -0.0365 9.71e- 1
##  7 SUSPECT_HEIGHT                           -0.113    0.133    -0.848  3.96e- 1
##  8 SUSPECT_REPORTED_AGE                     -0.0209   0.00439  -4.75   2.04e- 6
##  9 SUSPECT_SEXFEMALE                         0.0423   0.174     0.243  8.08e- 1
## 10 SUSPECT_SEXUNSPECIFIED                    0.760    1.06      0.719  4.72e- 1
## 11 STOP_WAS_INITIATEDBased on Radio Run     1.56     0.146    10.7    1.39e-26
## 12 STOP_WAS_INITIATEDBased on C/W on Scene  1.77     0.175    10.1    6.38e-24
## 13 PRECINCT_PROP_BLACK                      -0.106    0.175    -0.603  5.47e- 1
## 14 PRECINCT_PROP_HISP                        0.0434   0.254     0.171  8.64e- 1
```

The model has the form $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13}$, where β_0 is the intercept, x_1 is an indicator of whether or not the suspect is Asian/ Pacific Islander, x_2 is an indicator of whether or not the suspect is a white Hispanic, x_3 is an indicator of whether or not the suspect is a Black Hispanic, x_4 is an indicator of whether or not the suspect is Black, x_5 is an indicator of whether or not the suspect is an American Indian/ Alaskan Native, x_6 is the suspect's height from the mean, x_7 is the suspect's reported age from the age, x_8 is an indicator of whether or not the suspect is female, x_9 is an indicator of whether or not the suspect has an unspecified gender, x_{10} is an indicator of whether or not the police commenced the stop based off of someone else's call over the radio, x_{11} is an indicator of whether or not the police commenced the stop based off of a cooperating witness at the scene, x_{12} is the proportion of Black residents in the precinct, and x_{13} is the proportion of Hispanic residents in the precinct, and β_1 through β_{13} are the corresponding coefficients.
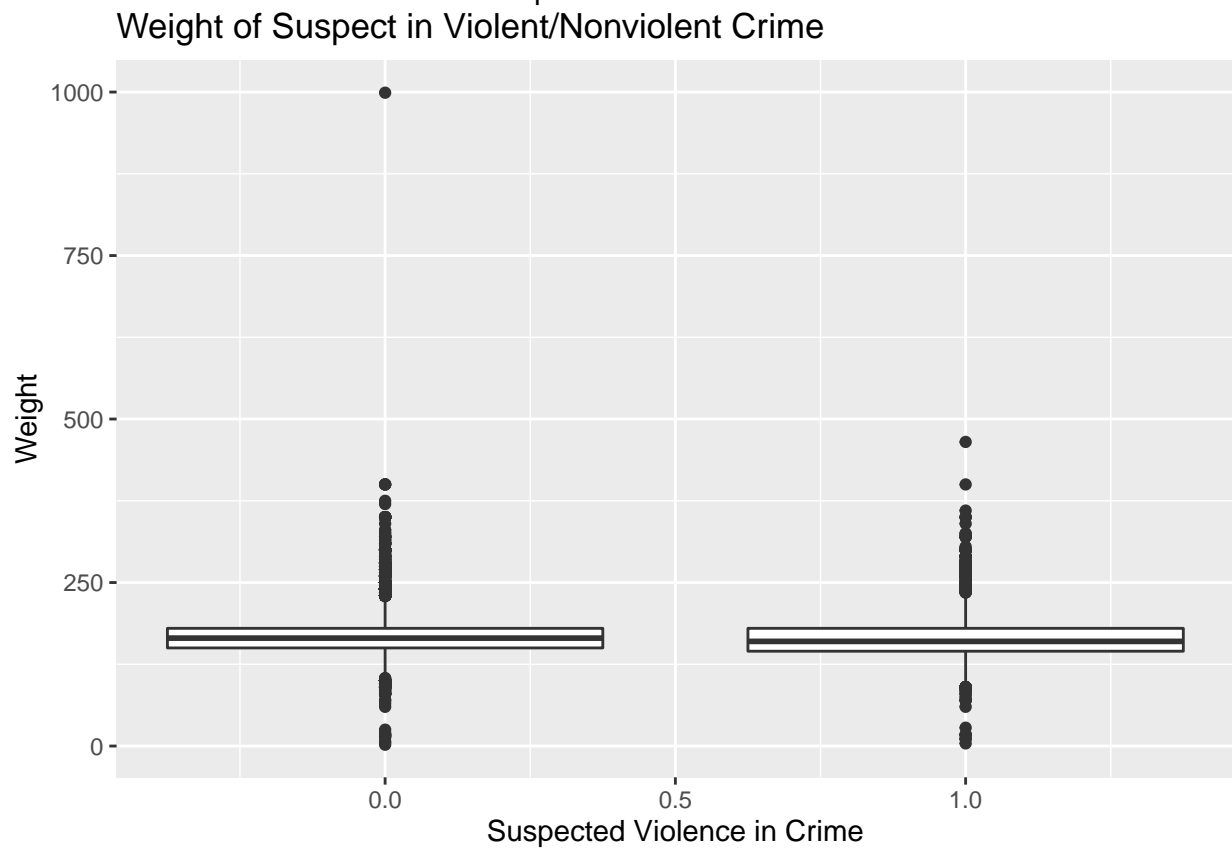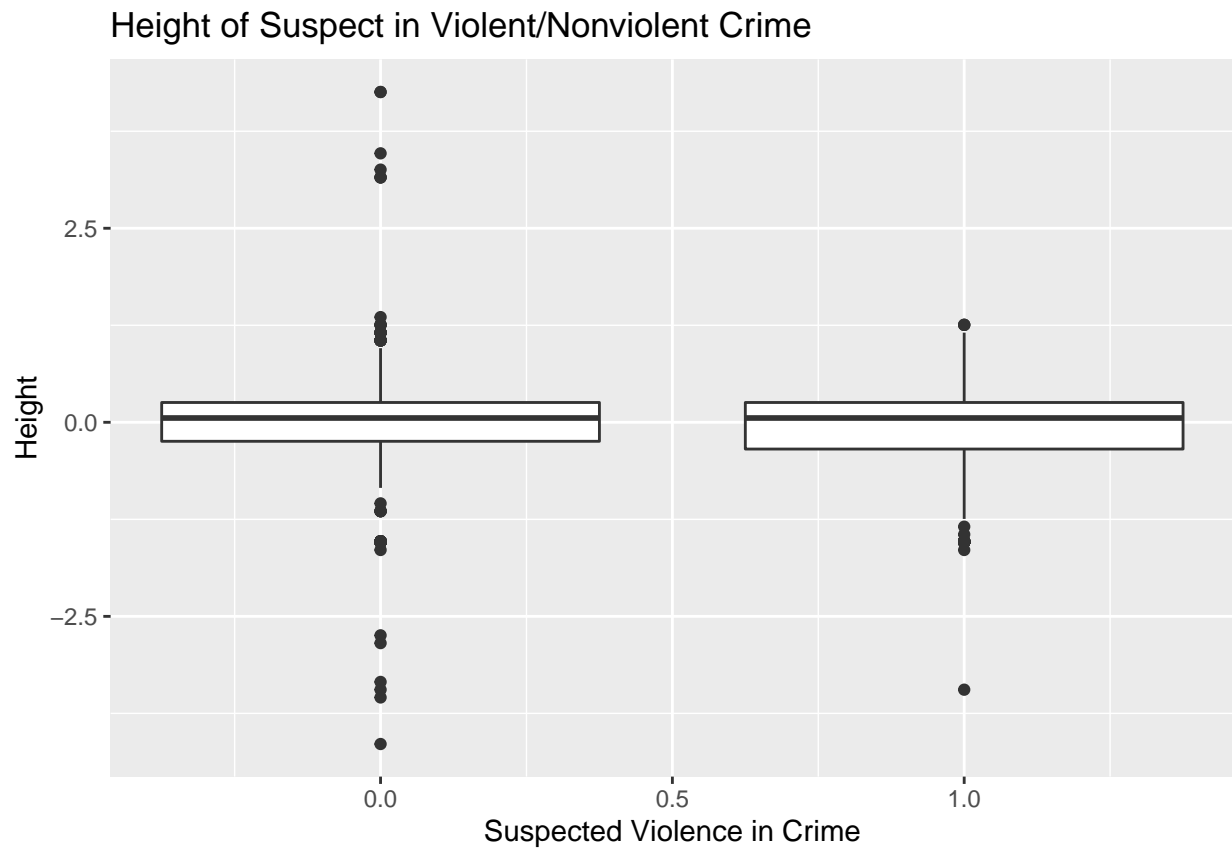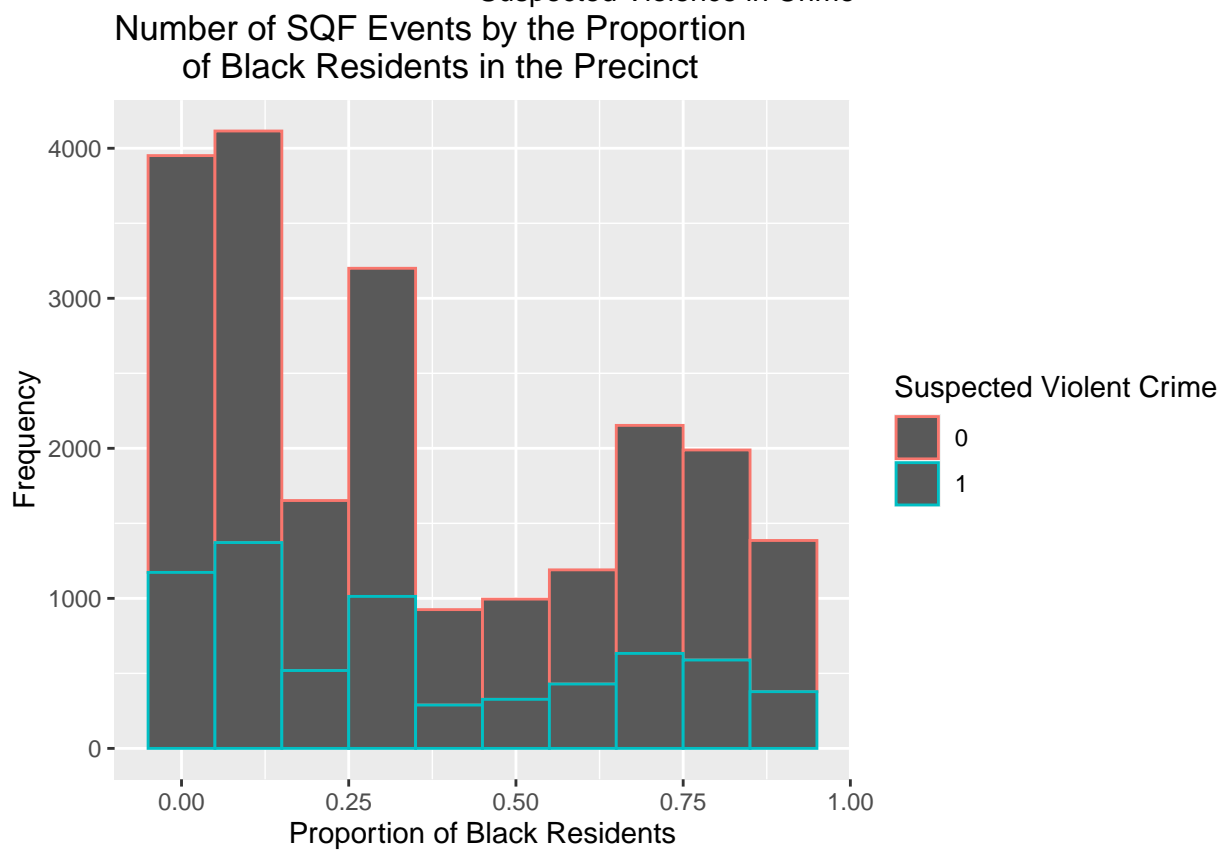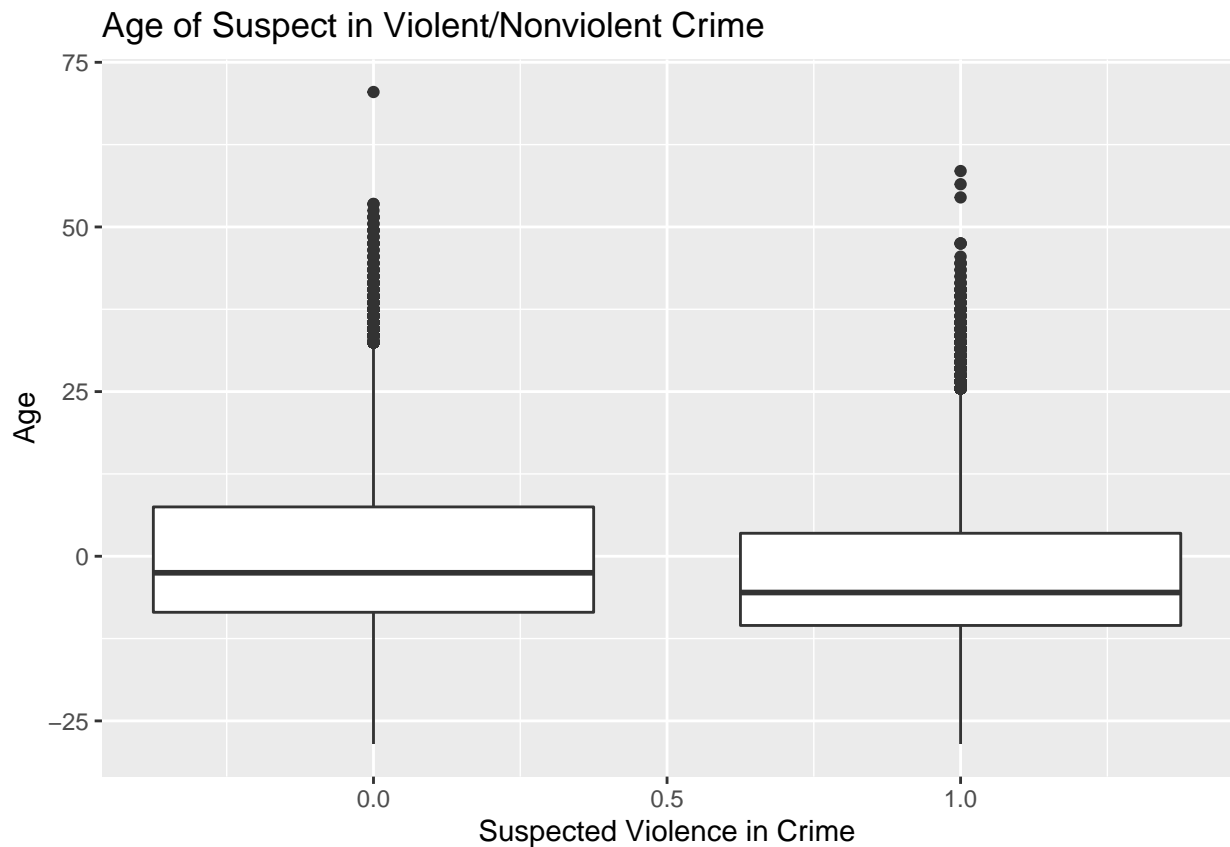
Discussion

From these results, we can see how race can have a significant impact on police decisions. In general, it is the most significant factor behind why an officer will suspect a violent crime, outside of how the officer decides to initiate the stop. For a white male with mean height and age who was stopped based on the officer's own initiative in a precinct with no Black or Hispanic residents, the odds that the officer will suspect him of a violent crime is $e^{-2.38056024}$. The odds of police suspicion of a violent crime are multiplied by a factor of $e^{0.36663303}$ for a Black resident than for a white resident. However, it is worth noting that the odds of police suspicion of a violent crime are multiplied by a factor of $e^{1.56279470}$ when the officer is called in than when the officer self initiates the stop, and the odds of police suspicion of a violent crime are multiplied by a factor of $e^{1.76761269}$ when the officer has a cooperative witness on the scene than when the officer self initiates the stop.

Appendix

Note: To reproduce all the results in this case study, simply knit the Analysis.rmd document.

Height of Suspect in Violent/Nonviolent Crime

Weight of Suspect in Violent/Nonviolent Crime

Age of Suspect in Violent/Nonviolent Crime

Number of SQF Events by the Proportion
of Black Residents in the Precinct

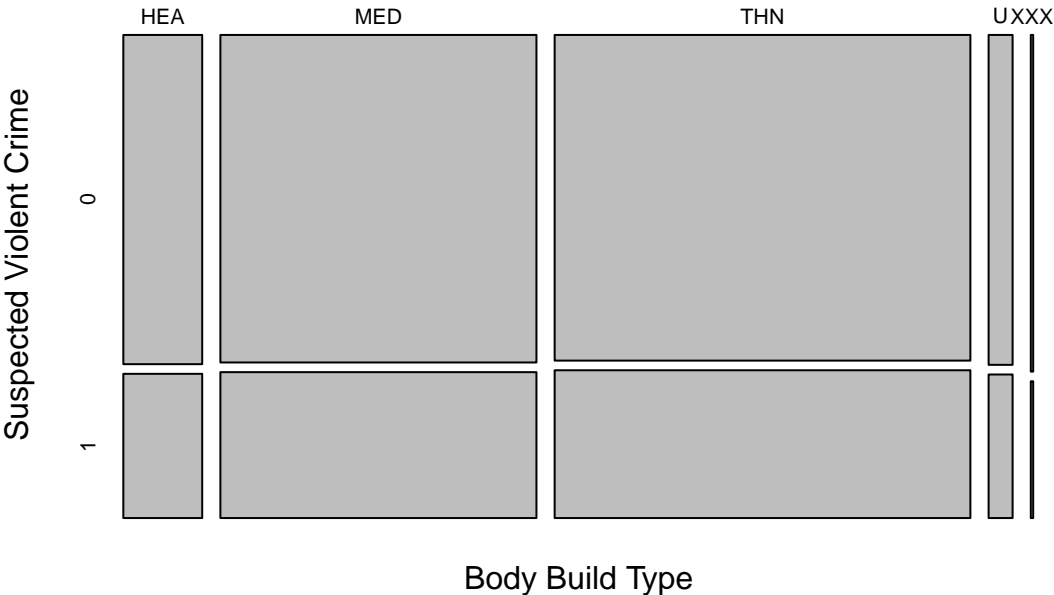## Number of SQF Events by the Proportion of Hispanic Residents in the Precinct
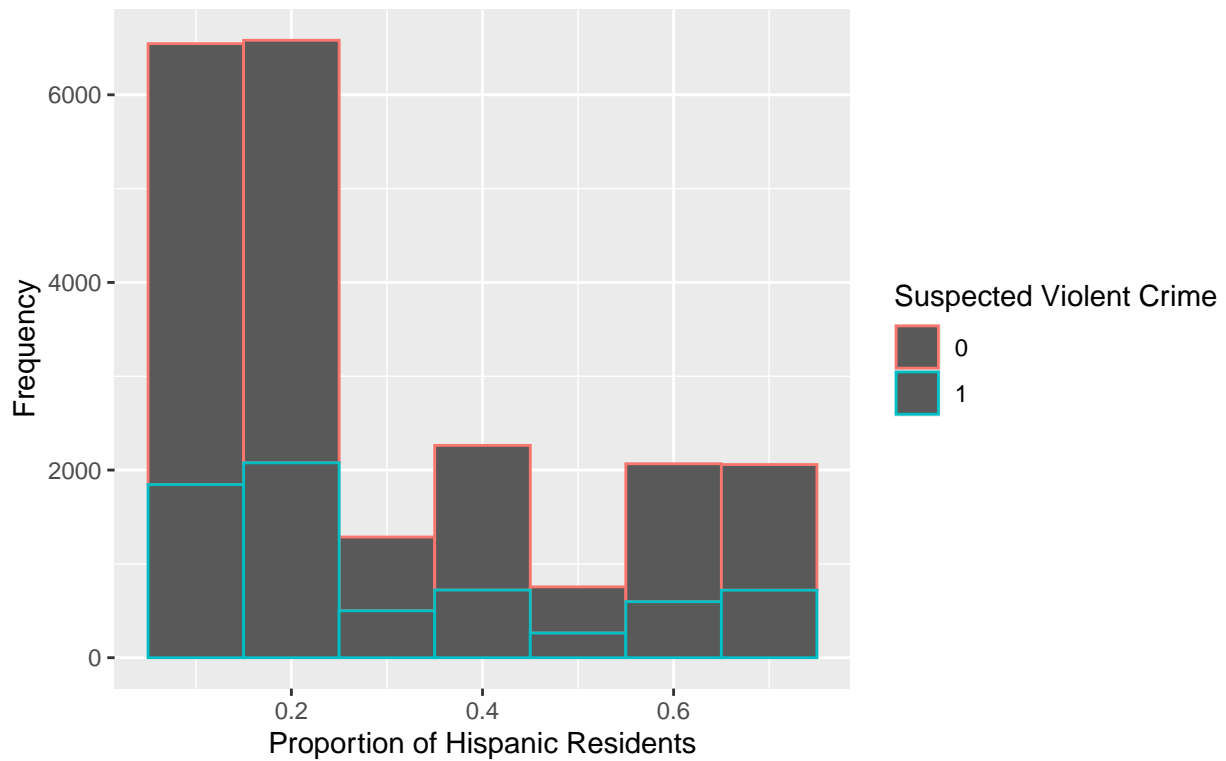
Ratio of Violent and Nonviolent Crime by Sex
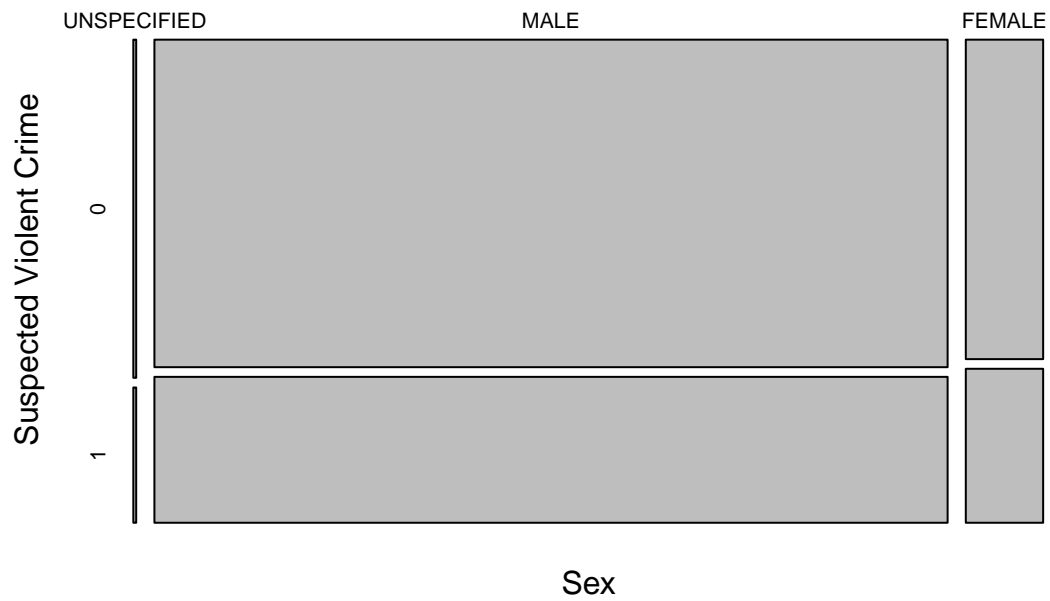
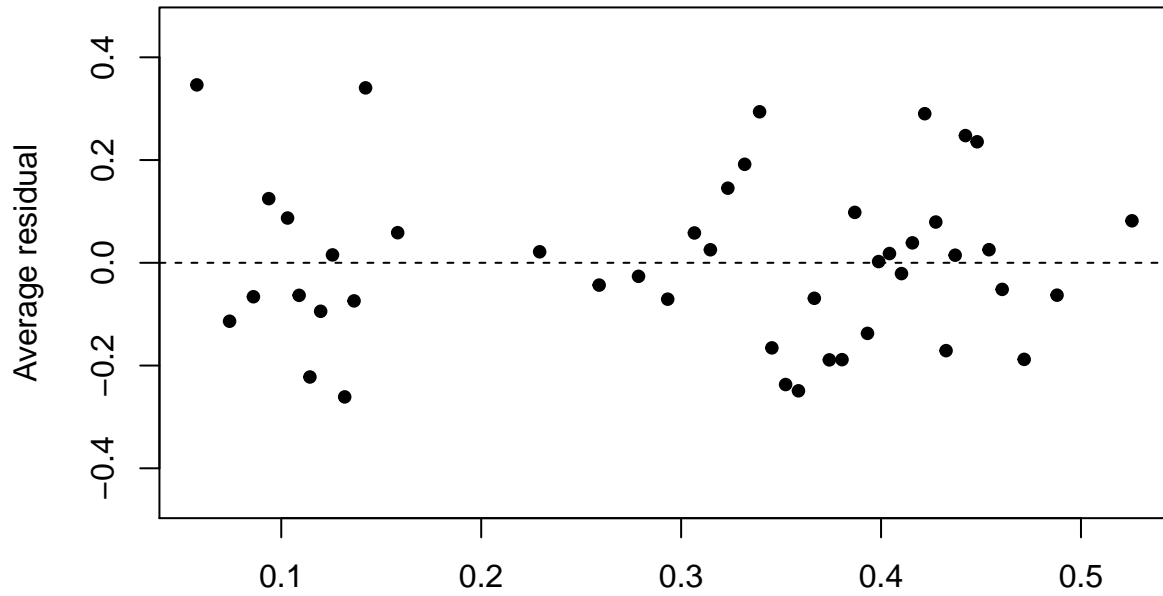## Ratio of Violent and
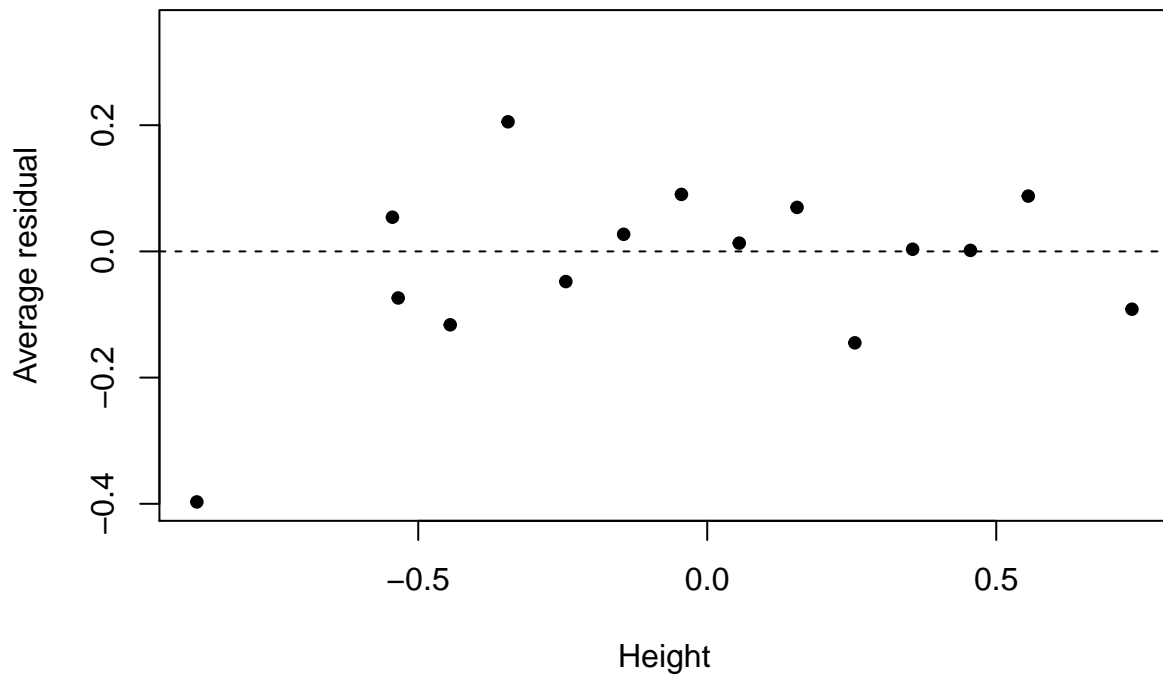## Nonviolent Crime by How Stop was Initiated

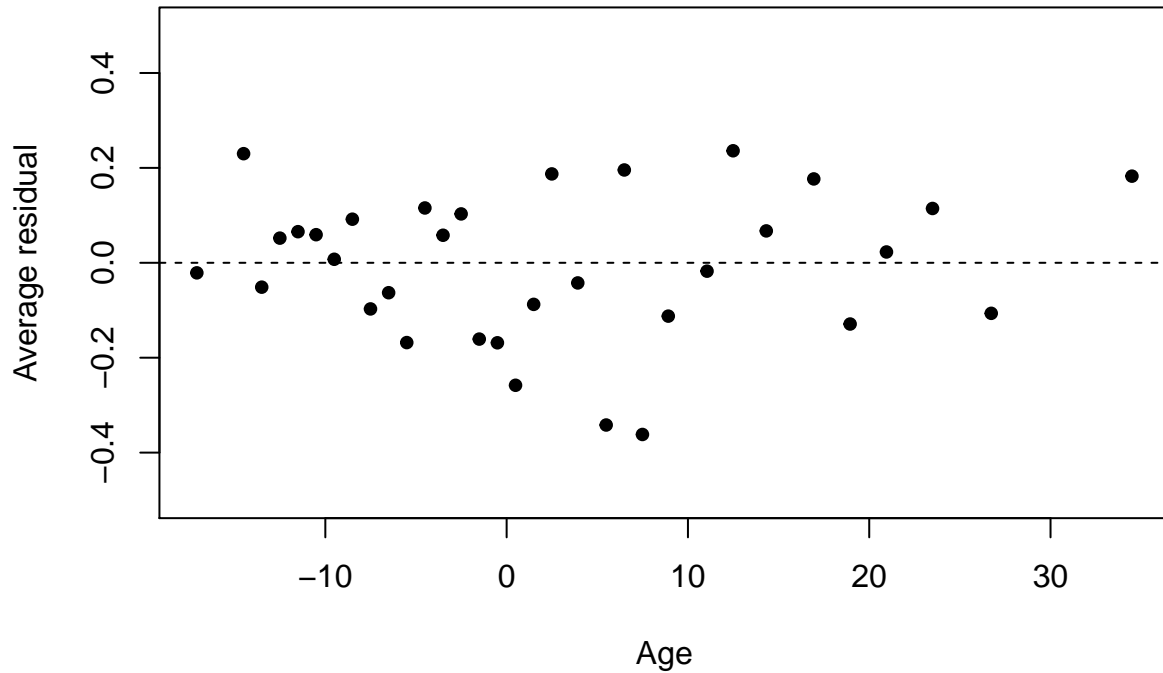## Ratio of Violent and Nonviolent Crime by Body Build Type
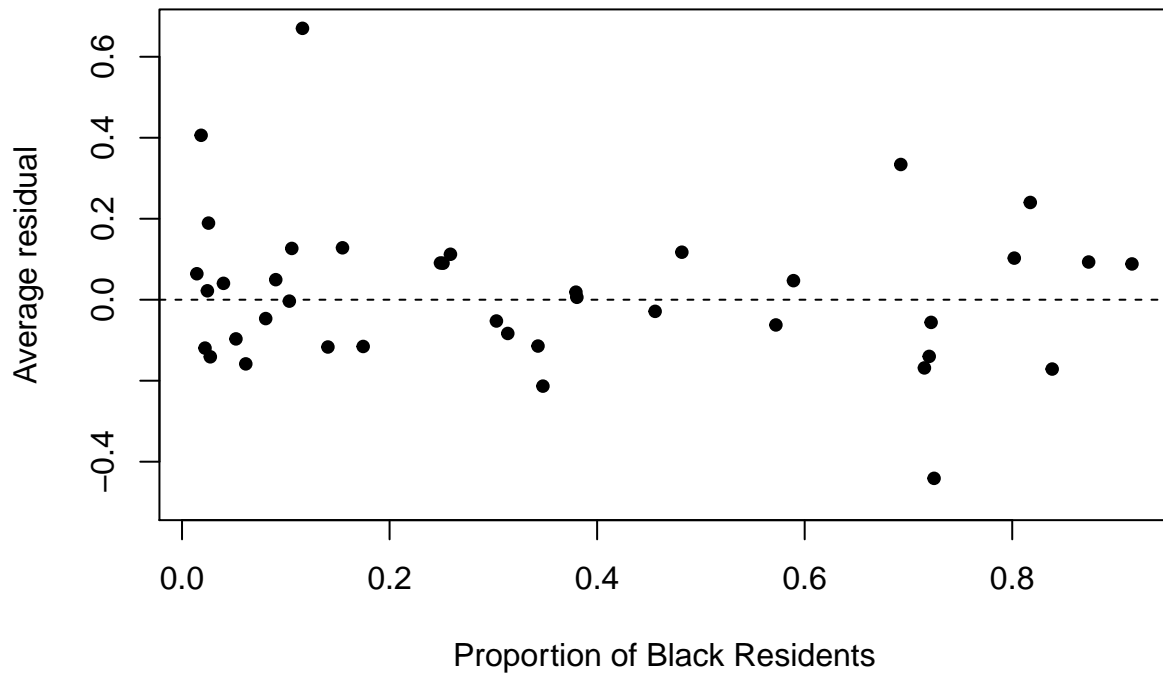
Binned Residuals vs. Probabilities
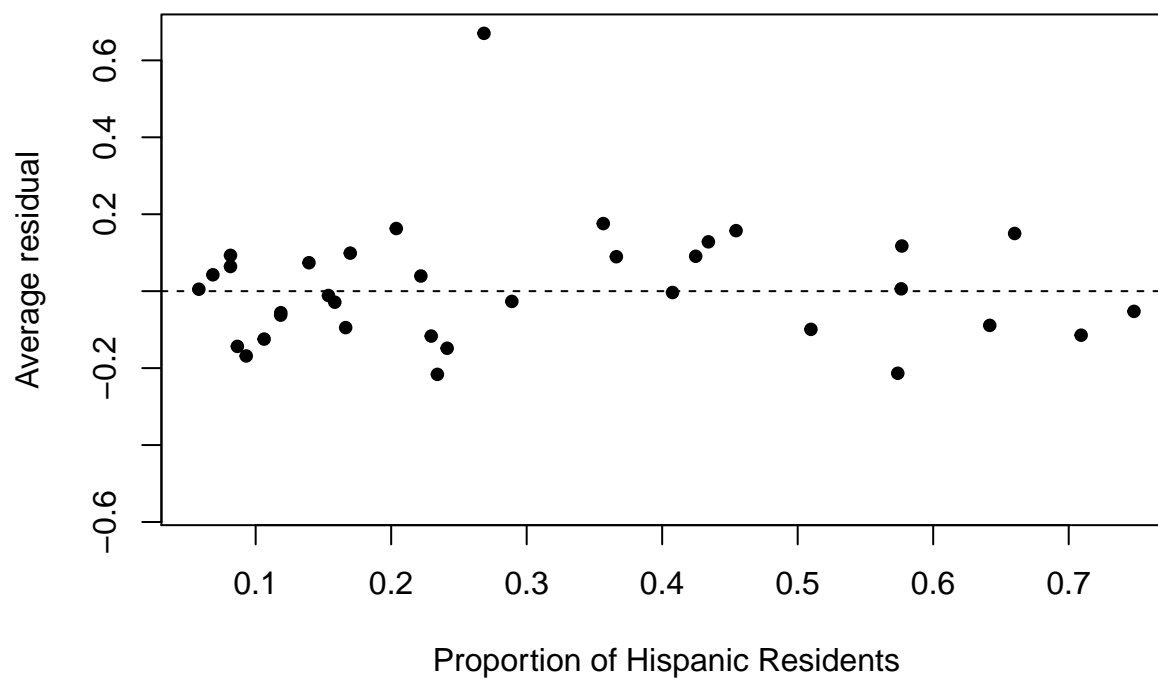

Binned Residuals vs. Height

Binned Residuals vs. Age

Binned Residuals vs. Proportion of Black Residents

Binned Residuals vs. Proportion of Hispanic Residents

Bibliography

"Stop-and-Frisk Data". *New York Civil Liberties Union*, https://www.nyclu.org/en/stop-and-frisk-data.