Racial Disparities in Traffic Stops/Citations

Keohane sQUAD: Chris Liang, Andrew Qin, Bob Qian, and Katie Nash

2020-11-05

Introduction and Data

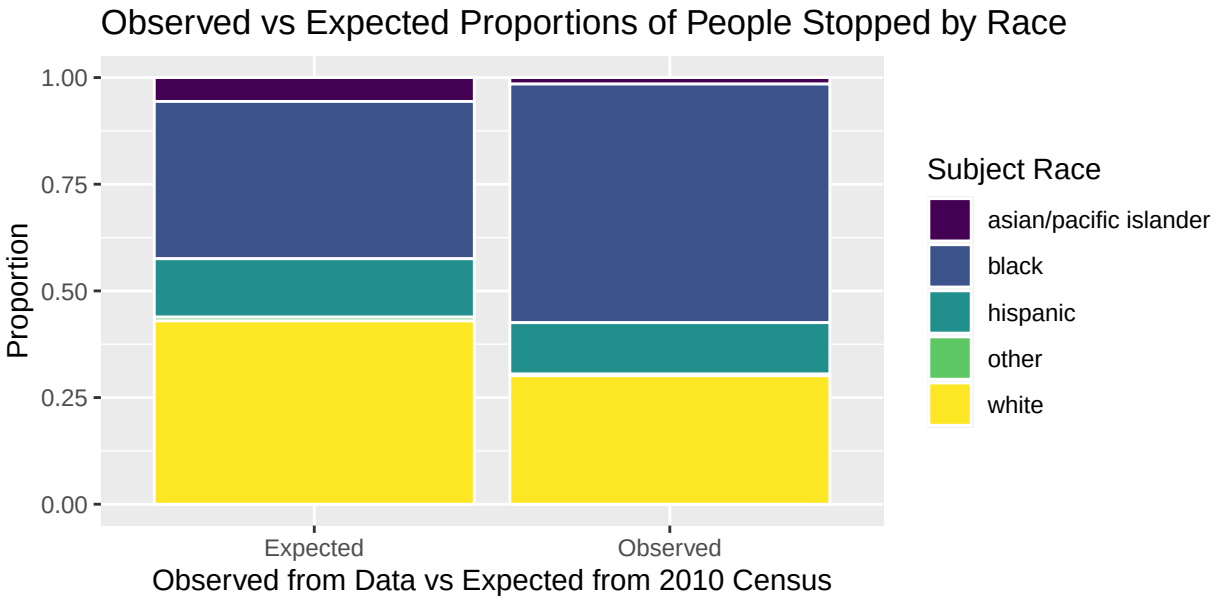
Our data is a census of individual police stops in Durham created by the Stanford Open Policing project. The Stanford Open Policing Project "[collects] and [standardizes] data on vehicle and pedestrian stops from law enforcement departments across the country"(https://openpolicing.stanford.edu/). We would like to see if that same kind of racial bias is evident in police stops in Durham. In doing so, we also wish to examine if other demographic characteristics (primarily sex) influence traffic stops. Our general research question is the following: what is the relationship between a subject's demographic attributes (primarily sex or race) and the likelihood of being stopped by police in traffic in Durham?

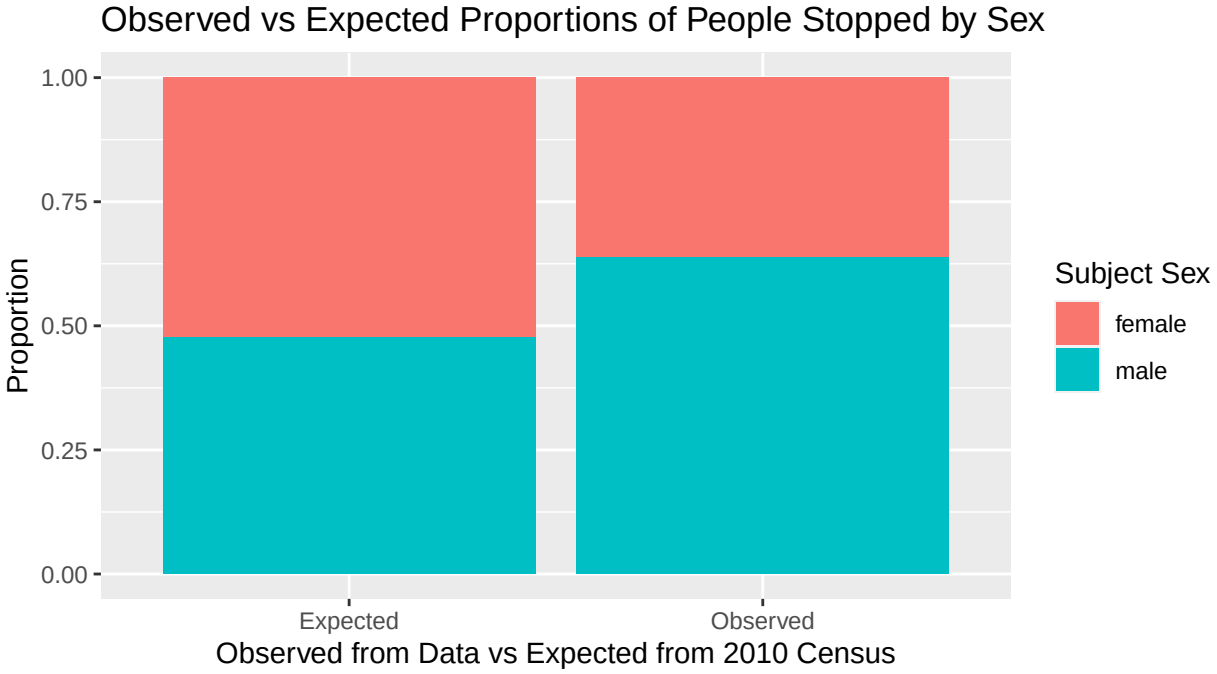
We hypothesize that race and the likelihood of being stopped by police in traffic in Durham are related, with black people representing disproportionately more of the people being stopped relative to their proportion within the population. We also hypothesize that sex has no significant relationship with being stopped in traffic. To find the true population proportions of people by race, sex, and age in Durham, we will utilize the 2010 Durham census data (https://www.census.gov/quickfacts/fact/table/durhamcountynorthcarolina/RHI625219#RHI625219).

Additionally, we will examine whether race, sex, or age are related to the outcome of the traffic stop (whether a citation will be issued). We hypothesize that race and the likelihood of receiving a citation are related, with black people more likely to receive a citation upon being stopped. We additionally hypothesize that younger people have a higher chance of receiving a citation upon being stopped and that sex has no significant relationship with being stopped in traffic.

The dataset has 29 variables and 323147 observations, and each observation in the data set is an individual police stop recorded in Durham during 2001 to 2015. A categorical variable in the data set is `subject_race`, which describes the race of the subject involved in the traffic stop. A discrete numerical variable in the data set is `subject_age`, which describes the age of the subject at the time of the traffic stop. A continuous numerical variable in the data set is `time`, which describes the hour, minute, and second that the stop was recorded. Other variables in the data set include `outcome`, which is what resulted from the stop (a warning or a citation, for example); `reason_for_stop`, which describes what the violation leading to the stop was; and `search_conducted`, whether a search of the subject was conducted during the stop.

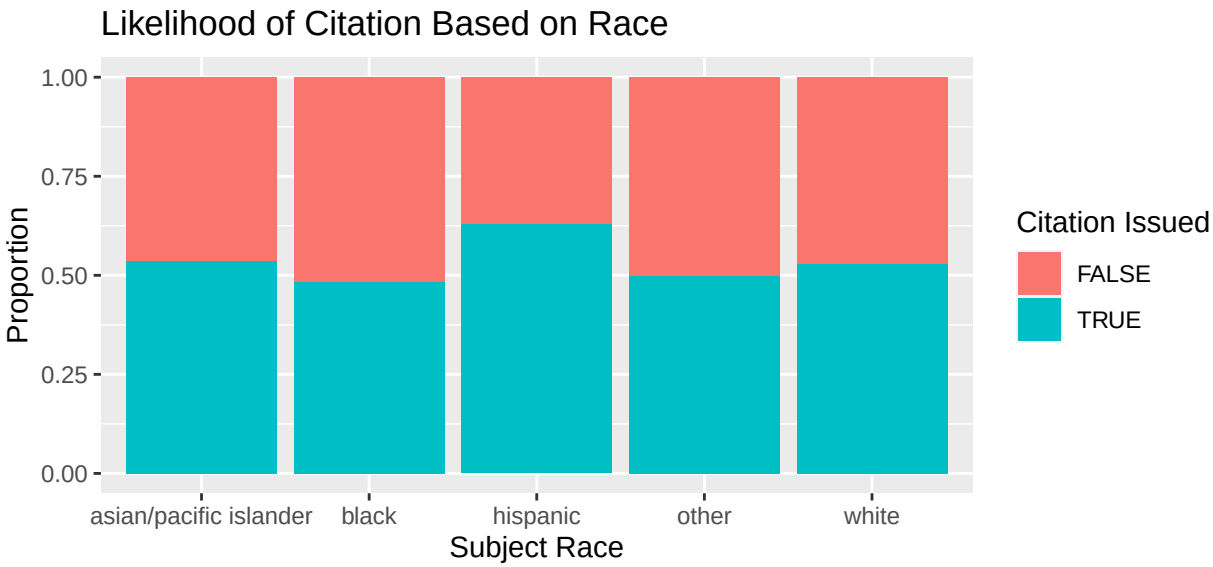
Methodology





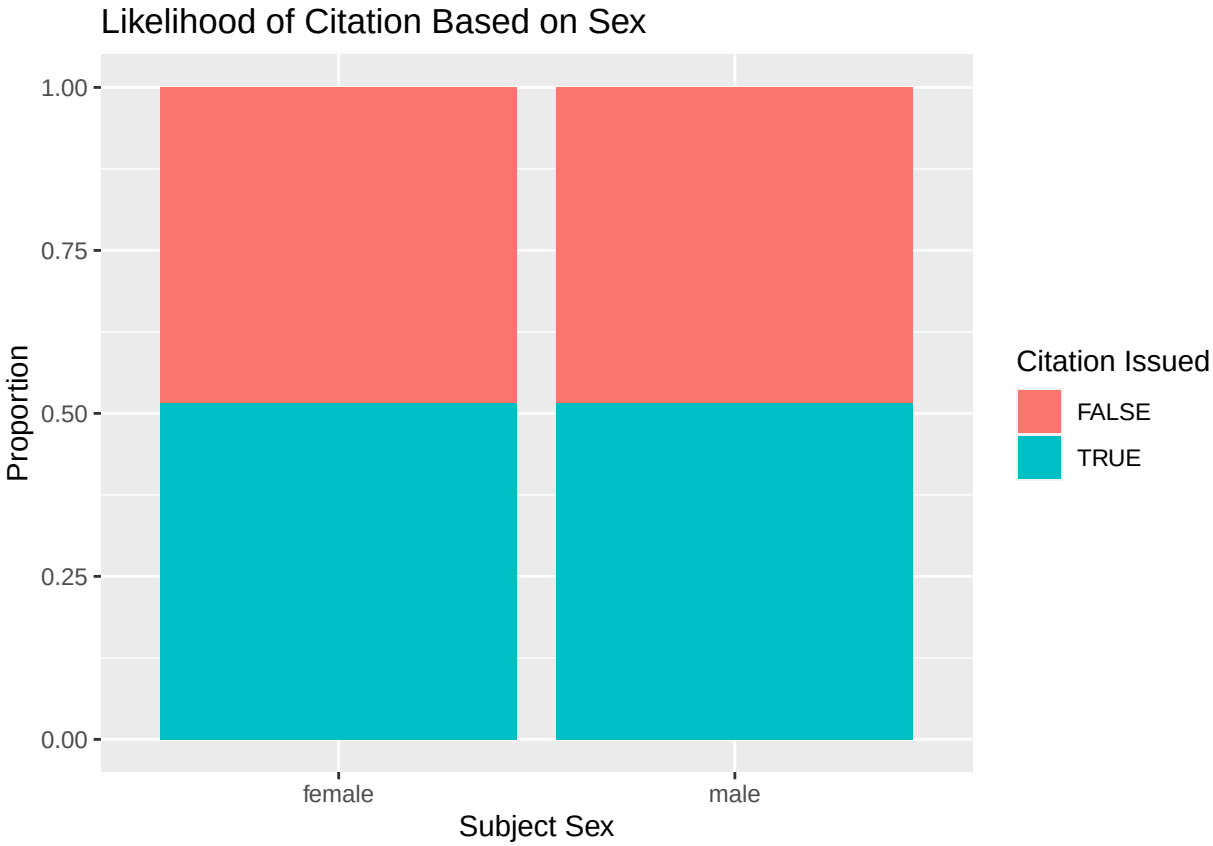
The variables we use to address the research question are `subject_race`, `subject_sex`, and `citation_issued`. For these variables we filter out any unknown and NA values.

To begin with, we visualize a segmented bar graph with the probability of citation based on race below.



According to the chart, it appears that Hispanics are the race with the highest proportion of citations issued.

We then visualize a segmented bar graph with the probability of citation based on sex below.



Based on this chart, the proportion of citations issued for females and males appears to be roughly equal.

ADD NULL DISTRIBUTION STUFF HERE (HOW DID WE SIMULATE)

To answer our research question, we utilize the chi-square test. We selected this test because we want to determine whether there is an association between two variables where we have more than two samples.

We perform two chi-square tests. For the first, we ask whether there is an association between someone's race status and whether a citation was issued. For the second, we ask whether there is an association between someone's sex and whether a citation was issued.

With each chi-square test, we compare observed versus the expected counts that we would expect if each H_0 were true. If these total differences are "large enough," then we reject the null hypothesis. We will perform each chi-square test at the $\alpha = 0.05$ significance level.

ADD LOGISTIC REGRESSION STUFF HERE - how we investigate age's relationship to citation issued.

Results

We first investigated the research question in reference to stop rates. Our exploratory data analysis indicated that black people appeared to be stopped at a disproportionately higher rates compared to their proportion within the Durham County population. Since the attempted census of 323147 observations is far too large to create a bootstrapped null distribution to check the statistical significance of the proportions, we created a stratified proportional sample of 3231 observations, roughly 1% of the original dataset. We decided to check if this difference between the observed and expected proportion (based on Durham County population) was statistically significant through simulation:

Let ρ equal the true proportion of stopped drivers who were black within Durham County.

$H_0 : \rho = 0.369$. The true proportion of stopped drivers who were black within Durham County is equal to the true proportion of black people within Durham County (0.369).

$H_A : \rho > 0.369$. The true proportion of stopped drivers who were black within Durham County is greater than the true proportion of black people within Durham County.

$\alpha = 0.05$

```
## [1] 0
```

Because our p-value of 0 is less than our α of 0.05, we reject the null hypothesis. There is sufficient evidence to indicate that the true proportion of people who are stopped within Durham County that are black is greater than the proportion of black people within the Durham County population (0.369). This indicates that black people are disproportionately stopped at a higher rate.

To investigate the second element of the research question (in reference to sex), we created another stratified proportional sample with 3232 observations, or 1% of the observations in the data frame. Our exploratory data analysis indicated that females were associated with a lower likelihood of being stopped, so we conducted a one-tailed test for difference in proportions to determine if the difference between the observed and expected proportions is statistically significant.

Let ρ equal the true proportion of stopped drivers who were female within Durham County.

$H_0 : \rho = 0.523$. The true proportion of stopped drivers who were female within Durham County is equal to the true proportion of female people within Durham County (0.523).

$H_A : \rho < 0.523$. The true proportion of stopped drivers who were female within Durham County is not equal to the true proportion of female people within Durham County.

$\alpha = 0.05$

```
## [1] 0
```

Because our p-value of 0 is less than our α of 0.05, we reject the null hypothesis. There is sufficient evidence to indicate that a female person is disproportionately less likely to be stopped by police in traffic in Durham County relative to their proportion within the larger population. Conversely, this also means that a male person is more likely to be stopped by police in traffic relative to their proportion within the larger population.

We then investigated the second element of our research question and conducted a series of chi-square tests of independence to determine if a person's race or sex is associated with a higher chance of receiving a citation upon being stopped.

H_0 : Race and the likelihood of receiving a citation upon being stopped are not associated.

H_A : Race and the likelihood of receiving a citation upon being stopped are associated.

$\alpha = 0.05$.

```
## # A tibble: 1 x 3
##   statistic chisq_df p_value
##       <dbl>    <int>   <dbl>
## 1     2785.        4       0
```

The chi-squared test for independence outputted a statistic of 2785.354. The distribution of the test statistic is a chi-squared distribution, which is unimodal and right-skewed with 4 degrees of freedom.

Since our p-value of 0 is less than our α of 0.05, we reject the null hypothesis. There is sufficient evidence to indicate that race and the likelihood of receiving a citation upon being stopped are associated.

We then tested if a driver's sex is associated with the likelihood of receiving a citation.

H_0 : Sex and the likelihood of receiving a citation upon being stopped are not associated.

H_A : Sex and the likelihood of receiving a citation upon being stopped are associated.

$\alpha = 0.05$.

```
## # A tibble: 1 x 3
##   statistic chisq_df p_value
##       <dbl>    <int>   <dbl>
## 1  0.000593        1   0.981
```

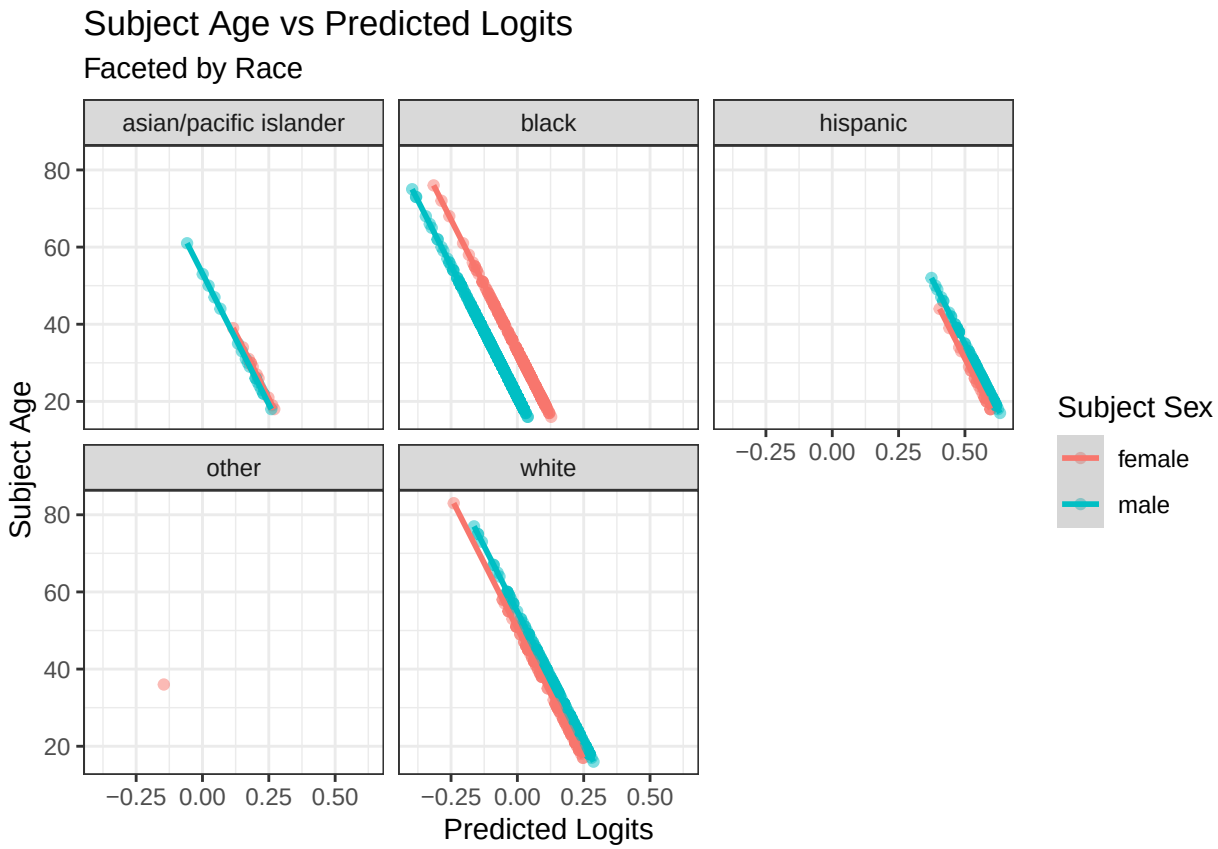
The chi-squared test for independence outputted a statistic of 0.001. The distribution of the test statistic is a chi-squared distribution, which is unimodal and right-skewed with 1 degree of freedom.

Since our p-value of 0.981 is greater than our α of 0.05, we fail to reject the null hypothesis. The data does not provide sufficient evidence to indicate that sex and the likelihood of receiving a citation upon being stopped are associated.

By itself, the chi-squared test only provides evidence for the association of two variables but does not inform us of the exact nature of the association. In order to investigate the nature of the association between race and the likelihood of receiving a citation, we created a logistic regression model. The logistic regression model has four variables to predict the log-odds of receiving a citation upon being stopped: The subject's race, the subject's age, the subject's sex, and an interaction variable between a subject's race and subject's sex (See Appendix 1 for the reasoning why this logistic regression model was chosen).

Conditions of Logistic Regression:

1. Independence - Each traffic stop is independent of other traffic stops; one traffic stop resulting in a citation does not affect the likelihood that other traffic stops result in citations.
2. Linearity - Below, we have depicted scatterplots of the relationship between age and the log-odds of receiving a citation, faceting by race and coloring by sex to isolate the linear predictor. The Linearity Assumption is met because there is a linear relationship between the age of a subject and the log-odds of receiving a citation when other predictors (race and sex in this context) are held constant.



Conditions met. Proceed with a logistic regression model.

```
## # A tibble: 11 x 5
##    term                                estimate std.error statistic   p.value
##    <chr>                                  <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                           0.373    0.0146      25.5   2.94e-143
##  2 subject_raceasian/pacific islander    0.0306   0.0483       0.634 5.26e-  1
##  3 subject_raceblack                    -0.127    0.0129      -9.90  4.19e- 23
##  4 subject_racehispanic                  0.357    0.0247      14.5   2.25e- 47
##  5 subject_raceother                    -0.253    0.105       -2.40  1.63e-  2
##  6 subject_age                          -0.00738  0.000279   -26.5   2.86e-154
##  7 subject_sexmale                       0.0323   0.0132       2.44  1.46e-  2
##  8 subject_raceasian/pacific islander:su~ -0.0435  0.0608     -0.715 4.74e-  1
##  9 subject_raceblack:subject_sexmale    -0.121    0.0164      -7.36  1.79e- 13
## 10 subject_racehispanic:subject_sexmale -0.00388  0.0285      -0.136 8.92e-  1
## 11 subject_raceother:subject_sexmale     0.151    0.125        1.21  2.28e-  1
```

The logistic regression model has outputted the following equations to predict the likelihood of receiving a citation:

Predicted Citation Log-Odds = 0.373 + 0.003 * subject_raceasian/pacific islander - 0.007 * age + 0.032 * male - 0.043 * subject_raceasian/pacific islander * male.

Predicted Citation Log-Odds = 0.373 - 0.127 * subject_raceblack - 0.007 * age + 0.032 * male - 0.121 * subject_raceblack * male.

Predicted Citation Log-Odds = 0.373 + 0.357 * subject_racehispanic - 0.007 * age + 0.032 * male - 0.004 * subject_racehispanic * male.

Predicted Citation Log-Odds = 0.373 - 0.253 * subject_raceother - 0.007 * age + 0.032 * male + 0.151 *

subject_raceother * male.

This model yields a few key conclusions:

1. Holding age and sex constant, we expect the odds that a Hispanic person will receive a citation upon being stopped by police in Durham County to be 1.4290359 times the odds that a white person will receive a citation upon being stopped by police. The coefficient is also statistically significant (p-value < 0.01), meaning there is less than a 1% chance such a coefficient or more extreme would be found in the data if race and the likelihood of receiving a citation were not associated.
2. Holding age and sex constant, we expect the odds that a black person will receive a citation upon being stopped by police in Durham County to be 0.8807337 times the odds that a white person will receive a citation upon being stopped by police. The coefficient is also statistically significant (p-value < 0.01).
3. Holding race and sex constant, for every additional year of age, we expect the odds of receiving a citation upon being stopped to multiply by 0.9926402. The coefficient is also statistically significant (p-value < 0.01), indicating that the data does provide sufficient evidence that the driver's age and the likelihood of receiving a citation are associated (slope 0).
4. Holding race and age constant, we expect the odds that a male person will receive a citation upon being stopped by police in Durham County to be 1.0328067 times the odds that a female person will receive a citation upon being stopped. This coefficient, however, is not statistically signifiacnt (p-value > 0.01).
5. In most cases, the interaction variable between sex and race does not result in a statistically significant coefficient, the critical exception being the case of black people. Holding age and race constant, upon being stopped, a black man's odds of receiving a citation are expected to be 0.9151497 times the odds that a black woman will receive a traffic citation upon being stopped (also factored in the nonsignificant "sex" variable by itself in odds calculation; without that coefficient, a black man's odds of receiving a citation are expected to be 0.8863285 times the odds that a black woman will receive a citation upon being stopped). Thus, unlike nearly every other race listed, black people are the only race where women are statistically significantly more likely to receive a citation upon being stopped.

The implications of this model will be further discussed in the "Discussion" section of the report.

Discussion

Throughout our research, we have learned that males are more likely to be stopped than females when compared to the expected proportions found from the 2010 Census. Although there is a stereotype that women are worse drivers than men, the higher proportion of males being pulled over than females helps support research that males tend to take more risks when driving than woman. (https://www.sciencedirect.com/science/article/pii/S1369847814001727)

Though black people are disproportionately more likely to be stopped for a traffic violation, they were not the most likely to receive a citation upon being stopped. We found that Hispanics were the most likely to receive a citation after being stopped. It is unclear as to why this occurs because we only have observational data, which is limitation. All this provides evidence for is that Blacks are more likely to be stopped. This does not mean that police officers are racially profiling because there can be lurking variables. However, our data is consistent with past literature.

For most races, there is very little change in data between citations being given by sex. However, significantly more Black females are given citations than there male counterparts. This could possibly show unspoken discrimination against Black females. However, with our limitations it is unknown if there are any lurking or unknown variables.

The Stanford policing project separated races differently than the census data As a result, our proportions are slightly skewed. In the future we would attempt to match the categories over the different datasets. If we were able to standardize the groupings we would have less ambiguous data and more accurate proportions and conclusions.

We are unable to establish if the trends we found were due to causation or if they were due to discriminatory policing practices. We are only able to take note of trends and not isolate causation.

Appendix 1

This appendix is devoted to showing why we chose the logistic regression model that we did. Utilizing the demographic characteristics of sex, race, and age, we attempted to create the most robust and explanatory model possible from the data. Below are the calculated AIC and BIC values for each model we considered. The model with the lowest AIC and BIC values was the one we chose, as it had the followed the principle of Occam's Razor the most faithfully (it explained the most in the least complex manner).

Meaning of AIC/BIC here: https://www.sciencedirect.com/topics/medicine-and-dentistry/akaike-informatio
n-criterion#:~:text=Akaike%20information%20criterion%20(AIC)%20(,among%20all%20the%20other%20models.&text=A%20

Logistic regression of purely subject race (BIC, then AIC):

```
## [1] 444902.9
```

```
## [1] 444849.5
```

Logistic regression of subject race and age (no interaction):

```
## [1] 444211.7
```

```
## [1] 444147.6
```

Logistic Regression of Race, Sex, and Age:

```
## [1] 444197.4
```

```
## [1] 444122.6
```

Noting that the best logistic regression included all three variables (despite the fact that sex by itself is not significantly associated with the likelihood of receiving a citation), we tested interaction variables.

Logistic Regression of Race, Sex, Age, and Race * Age:

```
## [1] 444225.5
```

```
## [1] 444107.9
```

BIC value increased, so we eliminated the interaction variable.

Logistic Regression of Race, Sex, Age, and Age * Sex:

```
## [1] 444202.1
```

```
## [1] 444116.6
```

BIC again increased, so we eliminated the interaction variable.

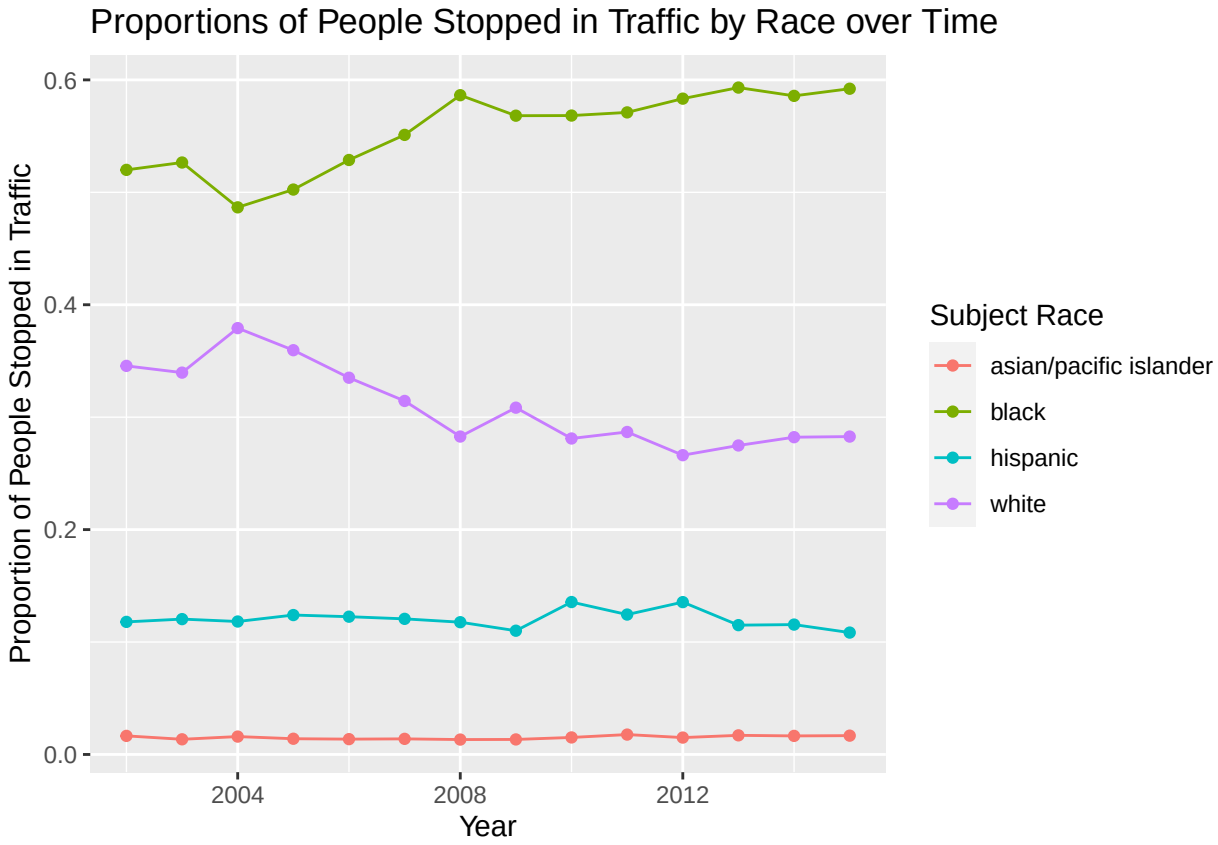
Logistic Regression of Race, Sex, Age, and Race * Sex:

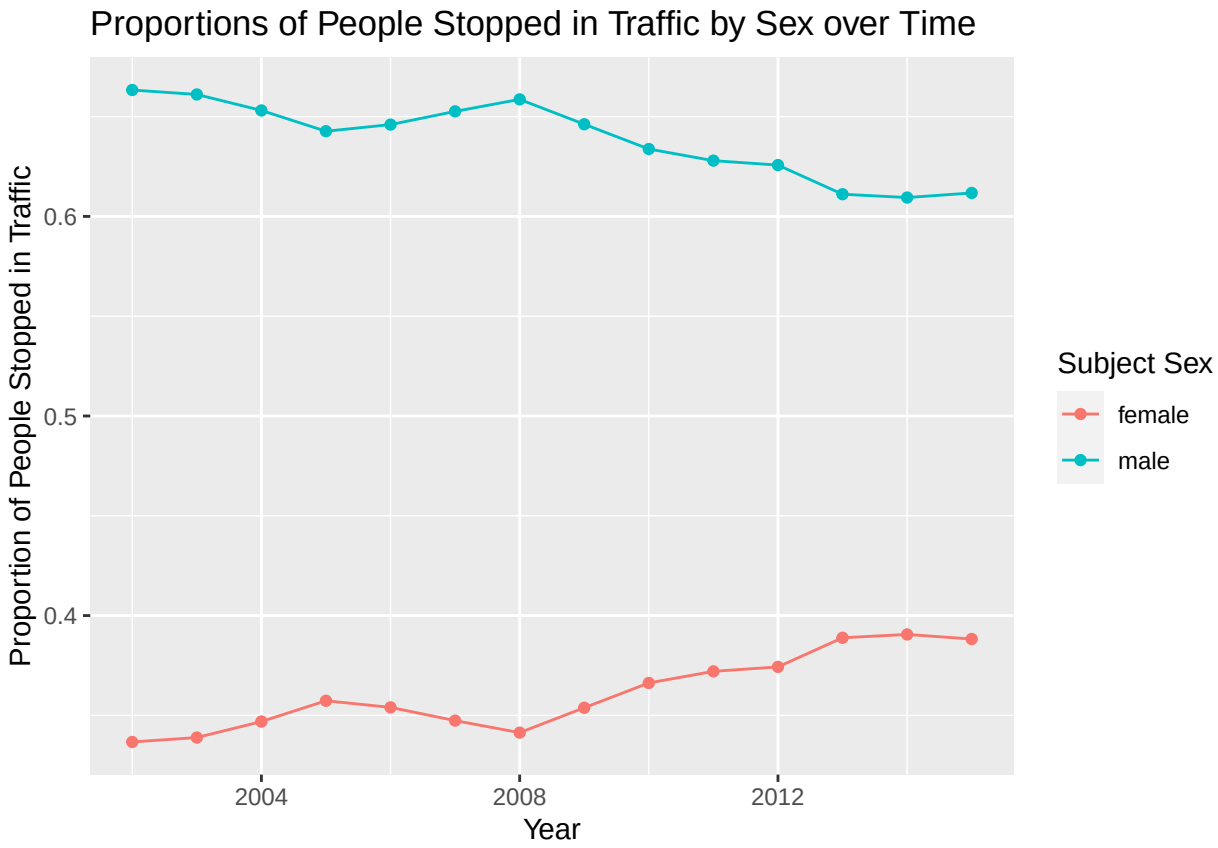
```
## [1] 444182.6
```

```
## [1] 444065
```

We obtained both our lowest AIC value here and a lower BIC value, making this our most robust and explanatory logistic regression model.

Appendix 2


Proportions of People Stopped in Traffic by Race over Time
Proportion of People Stopped in Traffic
0.6
0.4
0.2
0.0
2004
2008
2012
Year
Subject Race
asian/pacific islander
black
hispanic
white


Proportions of People Stopped in Traffic by Sex over Time
Proportion of People Stopped in Traffic
0.6
0.5
0.4
2004
2008
2012
Year
Subject Sex
female
male