Racial Disparities in Traffic Stops/Citations

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

2020-11-13

Introduction and Data

The US incarcerates more people than any other country, and people of color make up a disproportionate percent of the prison population . Police funding has grown significantly over the past four decades, and overpolicing in communities of color is a serious issue. In Michelle Alexander's book, The New Jim Crow: Mass Incarceration in the Age of Colorblindness, she discusses the rise of incarceration rates for black and brown people in the US. Alexander cites the War on Drugs as one of the biggest causes of contemporary mass incarceration, and she uses police pretext stops as an example. In pretext stops, cops can pull over a "suspicious" driver on the pretext of a very minor traffic violation (e.g. turning on red, going over the speed limit) and then do a drug sweep of the car, which may result in an arrest for drug-related charges. According to a Pew Research Center survey, "black adults are about five times as likely as whites to say they've been unfairly stopped by police because of their race or ethnicity." Given this background information on the criminal justice system, we would like to investigate if similar elements are at play in Durham, North Carolina.

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" (Stanford Policing Project). Our dataset, the "Durham" dataset in the Stanford Open Policing Project, records each observation as an individual police stop recorded in Durham between December 2001 and December 2015, and has 29 variables and 323147 observations. 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 as markers for comparison. To match our census information, we will filter the Stanford Policing Project dataset to only include data from Durham County, exclude races that are inputted as "unknown" or "NA," and exclude ages that are inputted as "NA." In doing so, a total of 2881 observations are removed from the dataset, a value that should not substantially alter the results with 323147 observations remaining.

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 receiving a citation upon being stopped in traffic.

Variables

Of the 29 variables, relevant categorical variables in the data set include `subject_race`, which describes the race of the subject involved in the traffic stop, and `subject_sex`, which describes the sex 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

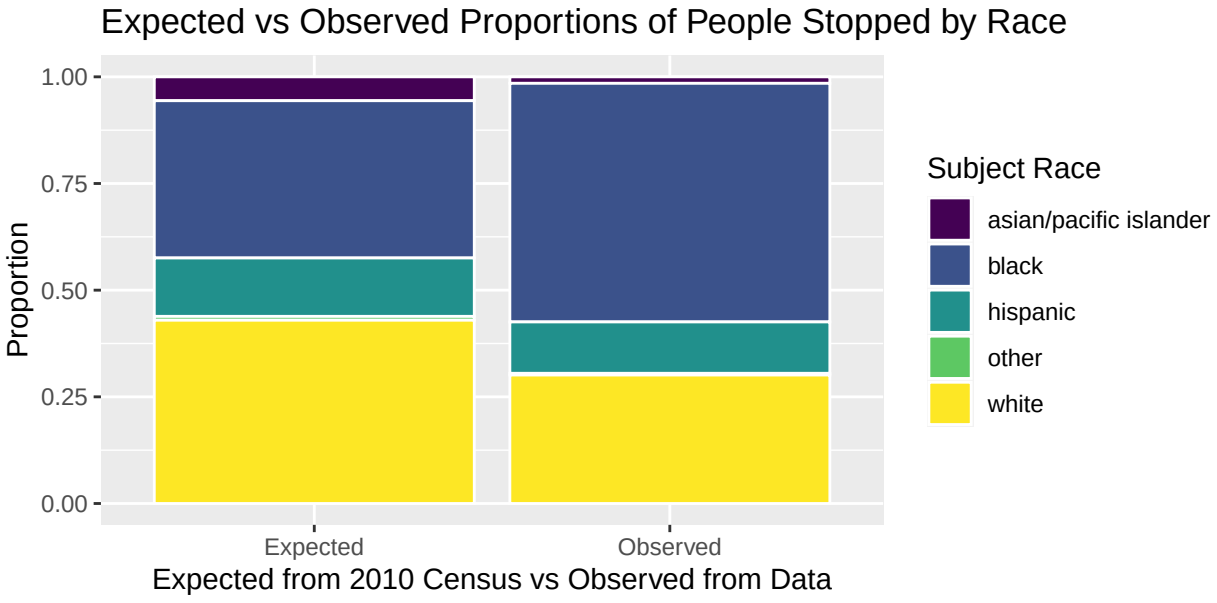
The variables we use to address the research question are `subject_race`, `subject_sex`, `subject_age`, and `citation_issued`.

To begin with, we calculate summary statistics for stop rate based on proportions by race in 2010:

subject_race	n	population	stop_rate
asian/pacific islander	471	15120	0.031
black	17691	99630	0.178
hispanic	4220	36990	0.114
other	133	2430	0.055
white	8749	116100	0.075

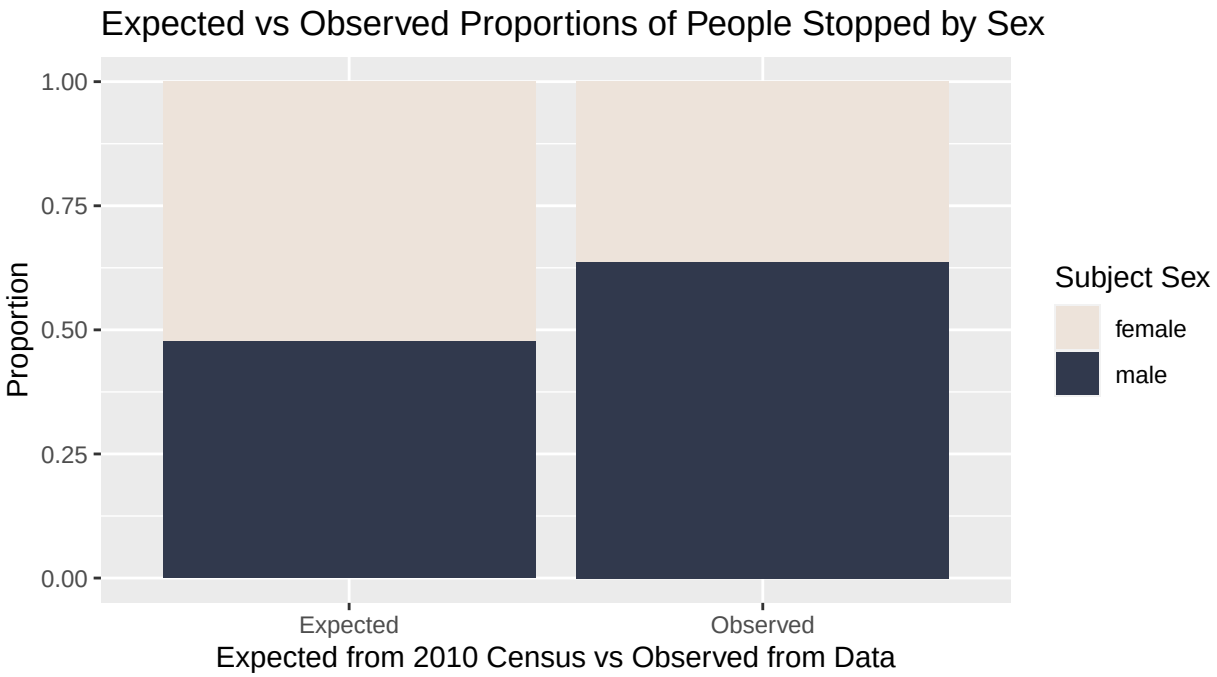
From the table, it appears that black people have the highest stop rate among races in Durham County in 2010. This rate is roughly 2.5 times higher than the rate that white people get stopped in Durham County.

To further visuzlie this trend, we use census data of population proportions by race and compare the expected proportion of people stopped by race to the observed proportion of people stopped by race in Durham County.



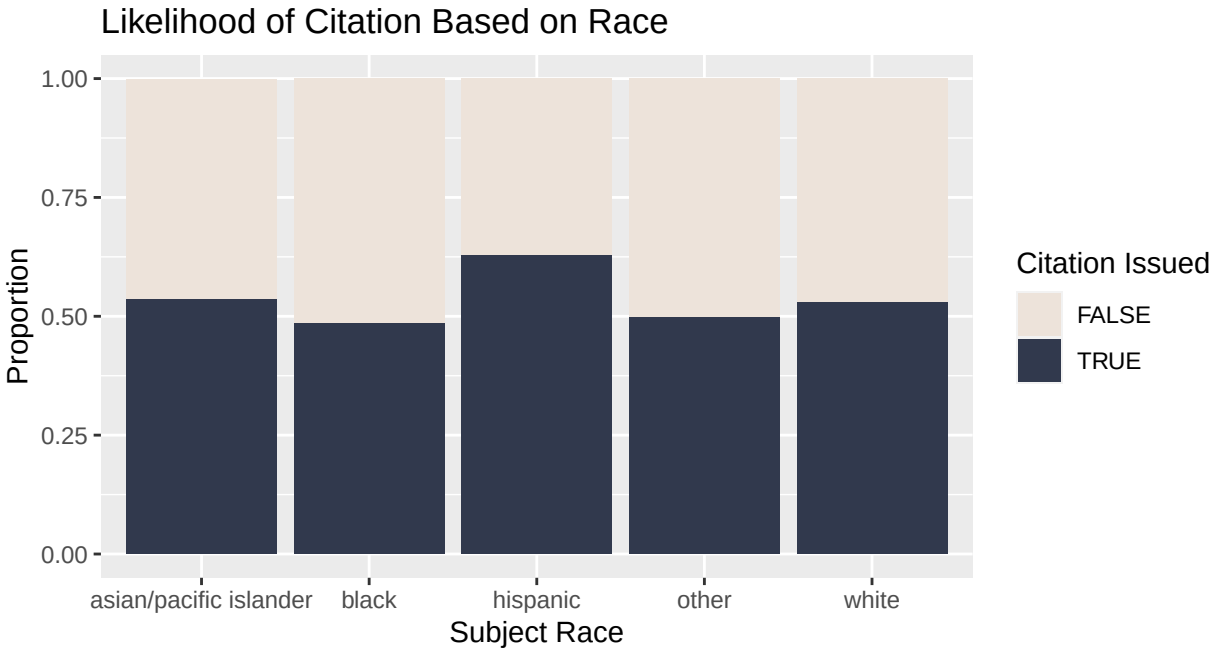
The chart provides some initial evidence for the hypothesis, indicating that a significantly greater proportion of black people were stopped compared to what was expected based on the proportion of black people in Durham County in the 2010 census. Within the "Results" section, we will run statistical tests to determine if the differences between expected and observed proportions were significant.

Next, we use census data to compare the expected proportion of people stopped by sex to the observed proportion of people stopped by sex in Durham.



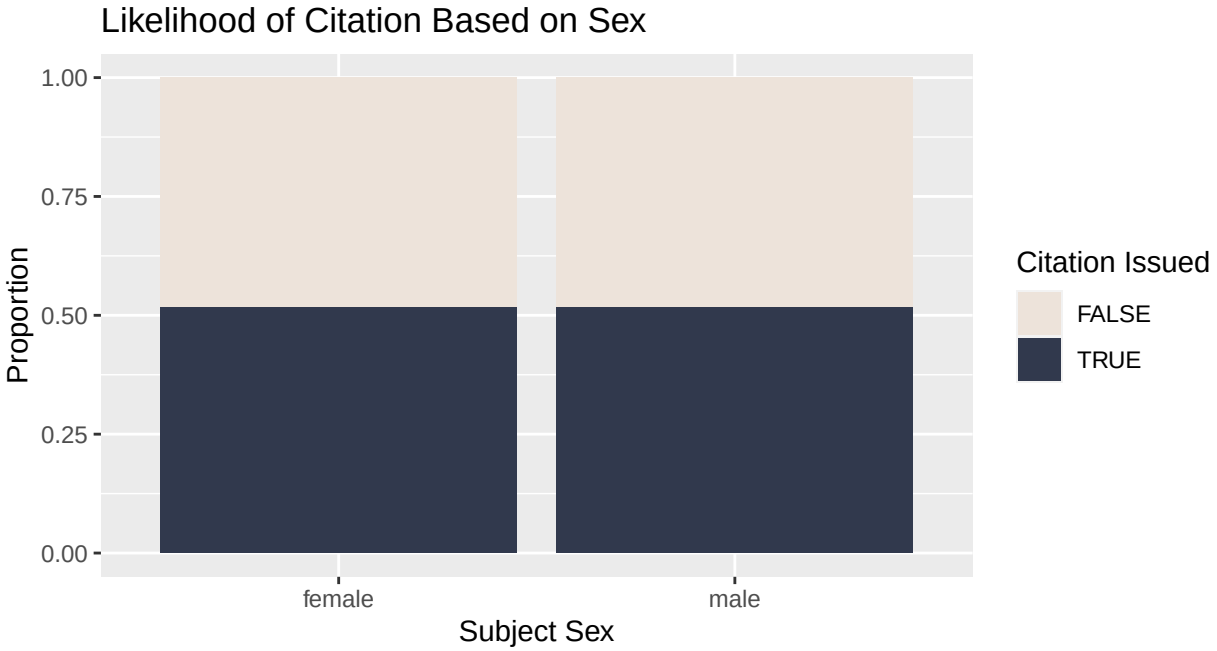
According to the chart, a significantly greater proportion of males were stopped compared to what was expected based on the proportion of males in Durham County in the 2010 census, providing some initial evidence against our hypothesis. Thus, both studied demographics initially appear to have disproportionately more people of a certain sex or race stopped compared to the 2010 census proportions.

To initially explore our second question, we visualize a segmented bar graph with the proportion of citations based on race below.



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

Lastly, we visualize a segmented bar graph with the proportion of citations based on sex below.



Based on this chart, the proportion of citations issued for females and males appears to be roughly equal. Thus, based on our exploratory data analysis, it initially appears that race influences the likelihood of receiving a citation, while sex does not appear to be associated.

To answer the first part of our research question on the relationship between race and sex and the likelihood of being stopped in traffic by police, we conducted a hypothesis test using a simulation-based approach. We created null distributions for the proportion of stopped drivers separately based on race and on sex, and then compared these proportions with the census data on proportions of race and sex in Durham to test if there was any statistical difference.

To answer the second part of our research question, on how race and sex are related to receiving a citation after a traffic stop, we utilized the chi-squared test. The chi-squared test helps us find evidence for an association or lack of association between race and citation issued (the first chi-squared test) and sex and citation issued (the second chi-squared test) when race, sex, and citation issued are categorical variables.

Within each chi-squared test, we compared observed versus the expected counts of a citation being issued based on race or sex if each H_0 were true. If these total differences are "large enough," then we would reject the null hypothesis. We performed each chi-squared test at the $\alpha = 0.05$ significance level.

If we find that any two variables are associated in a significant manner, we will then engage in a logistic regression model to investigate the nature of the association, using the methods described in Appendix 1 to select the best model.

Results

First Research Question (Demographic Factors and Likelihood of Being Stopped)

Exploring Race: We first investigated the research question regarding if race and sex demographics were proportionately (according to census proportions) stopped in traffic. Our exploratory data analysis for the true proportion of black people stopped in traffic indicated that black people appeared to be stopped at a disproportionately higher rate 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
```

Conclusion: 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.

Exploring Sex: To investigate the second element of our first 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 on the true proportion of female drivers stopped in traffic 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
```

Conclusion: 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 Durham population.

Second Research Question (Demographic Factors and Likelihood of Citation)

Exploring Race: We then investigated the second element of our research question–the relationship between race and citation issued or sex and citation issued–and conducted a series of chi-squared 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.

Conclusion: 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.

Exploring Sex: 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.

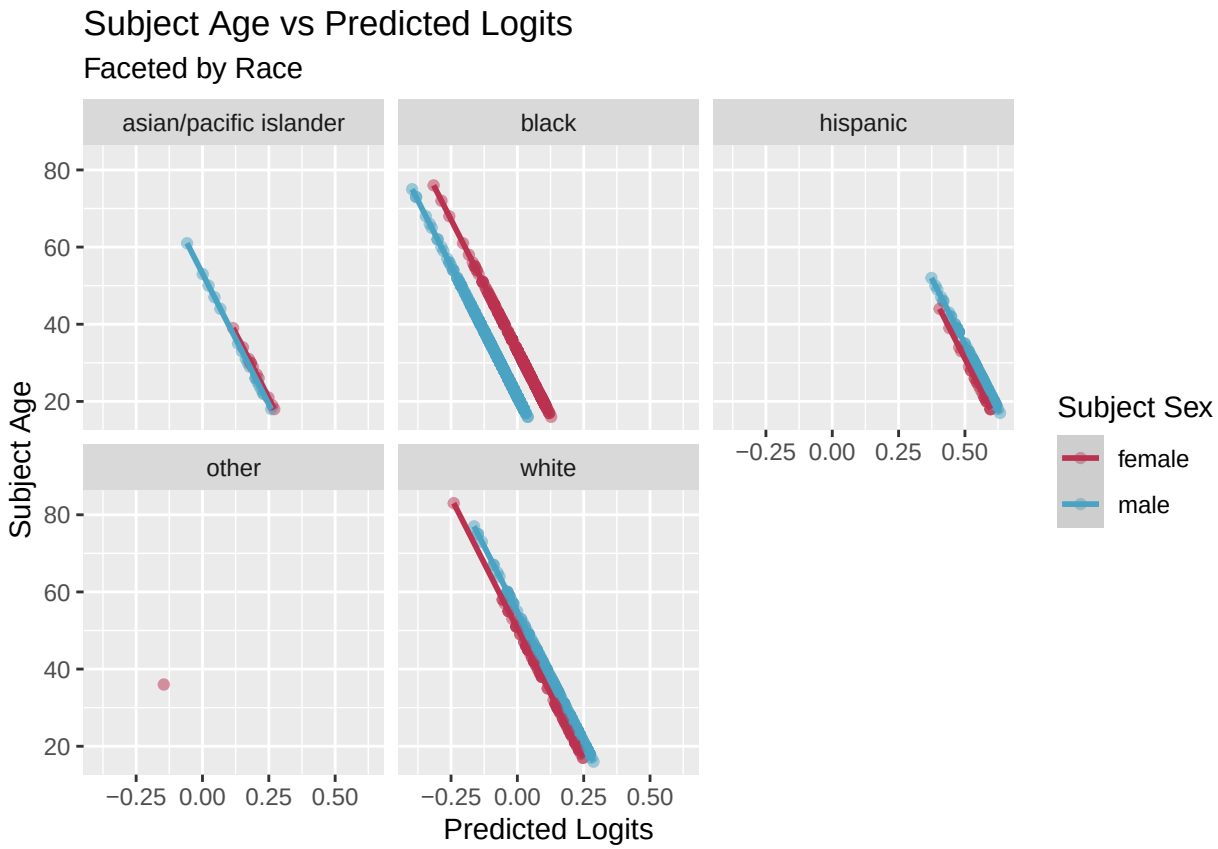
Conclusion: Since our p-value of 0.981 is greater than our α of 0.05, we fail to reject the null hypothesis. The data does not provide sufficient evidence to indicate that sex and the likelihood of receiving a citation upon being stopped are associated.

Logistic Regression Model for Likelihood of Citation

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.

term	estimate	std.error	statistic	p.value
(Intercept)	0.373	0.015	25.484	0.000
subject_raceasian/pacific islander	0.031	0.048	0.634	0.526
subject_raceblack	-0.127	0.013	-9.899	0.000
subject_racehispanic	0.357	0.025	14.458	0.000
subject_raceother	-0.253	0.105	-2.402	0.016
subject_age	-0.007	0.000	-26.459	0.000
subject_sexmale	0.032	0.013	2.443	0.015
subject_raceasian/pacific islander:subject_sexmale	-0.043	0.061	-0.715	0.474
subject_raceblack:subject_sexmale	-0.121	0.016	-7.364	0.000
subject_racehispanic:subject_sexmale	-0.004	0.029	-0.136	0.892
subject_raceother:subject_sexmale	0.151	0.125	1.206	0.228

Using the above table outputted by the logistic regression model, we can create four separate equations for the predicted log-odds of receiving a citation:

$\widehat{CitationLogOdds}$ = 0.373 + 0.003 * subject_raceasian/pacific islander - 0.007 * age + 0.032 * male - 0.043 * subject_raceasian/pacific islander * male.

$\widehat{CitationLogOdds}$ = 0.373 - 0.127 * subject_raceblack - 0.007 * age + 0.032 * male - 0.121 * subject_raceblack * male.

$\widehat{CitationLogOdds}$ = 0.373 + 0.357 * subject_racehispanic - 0.007 * age + 0.032 * male - 0.004 * subject_racehispanic * male.

$\widehat{CitationLogOdds}$ = 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 $\neq 0$).
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 significant at the 1% level (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

Through our research, contrary to our original hypothesis, we have learned that males are more likely to be stopped than females in Durham County 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 is in line with previous research that males tend to take more risks when driving than females.

Consistent with our original hypothesis, we have also learned that black people are disproportionately more likely to be stopped for a traffic violation. This stop rate is both statistically significant and practically relevant, with black people being stopped at a 2.5 times higher rate than white people. Additionally, we found that Hispanics were the most likely to receive a citation upon being stopped. Unfortunately, our observational data cannot create a causal association between a person's race and the likelihood of being stopped or receiving a citation. However, our data is consistent with the discussion on racial profiling and pretext stops in Michelle Alexander's The New Jim Crow: Mass Incarceration in the Age of Colorblindness, which indicates that black people are more likely to be stopped in traffic and that race does play a role in traffic penalties. Nonetheless, we only have evidence that suggests statistically significant differences in proportion of black drivers being stopped (as opposed to their population percentage). In addition, our robust logistic regression model indicates that age does have a statistically significant association with the likelihood of receiving a citation, with the predicted log-odds decreasing as age increases in line with our original hypothesis.

Appendix 2 further explores the trends of the relationship between traffic stops, sex, and race by time. Overall, these visualizations show that policing trends by race have not dramatically changed over the time period from 2002 to 2015, which may indicate the enduring relevance of our conclusions today.

For most races, there is very little change in data between citations being given by sex. However, the odds that a black female will receive a citation are significantly greater than their male counterparts. This could possibly show unspoken discrimination against black females, an idea we believe necessitates further statistical analysis. A statistical exploration of stop rates of black females compared to black males and white females could illuminate the unique intersectional experiences faced by black females within America.

A key limitation of our data analysis was that the Stanford Policing Project separated races differently than the census data. Specifically, the Stanford Policing Project data on race did not account for biracial people, people of other races, and did not detail what category white, black, or Asian Hispanics would fall under. 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.

The chi-squared test is another limitation that hinders our analysis. The chi-squared test only tells us whether or not a relationship exists between two variables, it does not tell us anything about the nature of the relationship between the variables.

One of our biggest limitations is that 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 cannot isolate causation.

References

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

Pew Research Study referenced in Introduction: https://www.pewresearch.org/fact-tank/2020/06/03/10-things-we-know-about-race-and-policing-in-the-u-s/

Statistic for more black people incarcerated: (https://www.sentencingproject.org/criminal-justice-facts/)

Stanford Policing Project: (https://openpolicing.stanford.edu/)

2010 Durham Census: (https://www.census.gov/quickfacts/fact/table/durhamcountynorthcarolina/RHI625219#RHI625219)

Study that Males are Riskier: https://www.sciencedirect.com/science/article/pii/S1369847814001727

Code: https://stackoverflow.com/questions/23479512/stratified-random-sampling-from-data-frame http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

Michelle Alexander's New Jim Crow

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).

Logistic regression of Race (BIC, then AIC):

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

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

Logistic regression of Race and Age:

```
## [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.

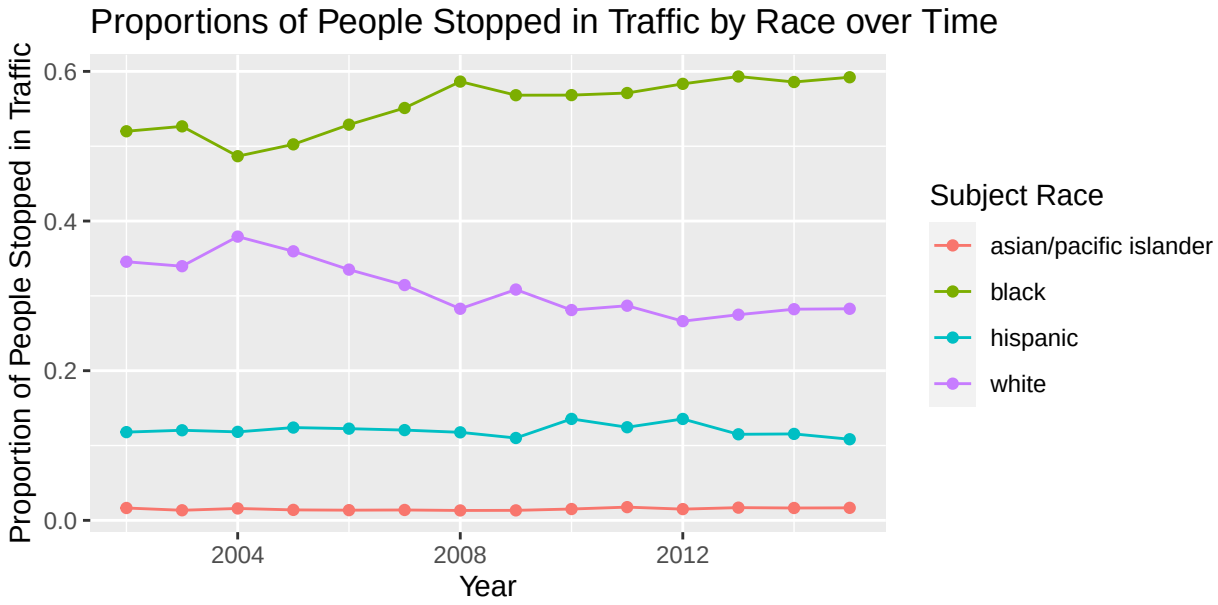
Logistic Regression of Race, Sex, Age, and Race * Sex:

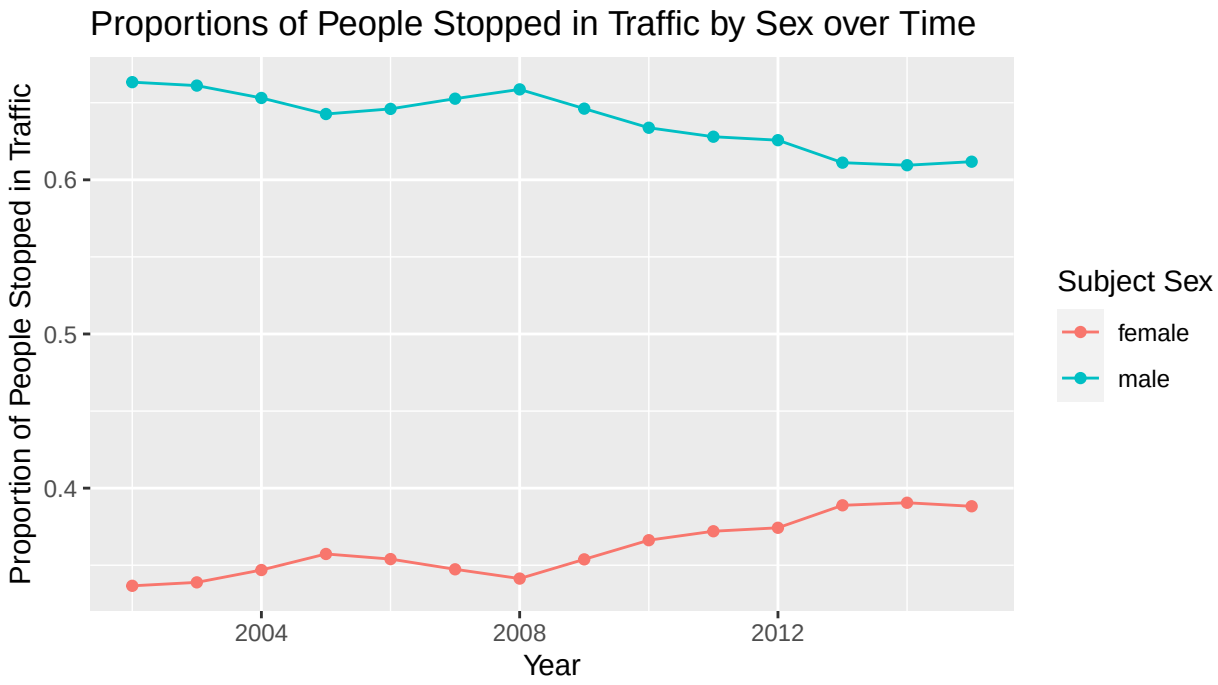
```
## [1] 444182.6
```

```
## [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 Sex over Time
Proportion of People Stopped in Traffic
0.6
0.5
0.4
2004
2008
2012
Year
Subject Sex
female
male