An Analysis of Traffic Stops in Durham

NAFTA: Arjun Chadha, Jonny Chang, Sam Fuller, Kishan Gandham

17/11/2021

Introduction and Data

When seeking a dataset to study for our project, we thought to analyze a topic that would be both relevant to all our group members and of social import. More specifically, we mean that we wanted to work with a dataset that might involve us in investigating a social issue or social bias using statistics to confirm or disconfirm our initial understandings. We eventually narrowed in on a dataset that contained information on police traffic stops made in Durham. Not only would analysis of this data be specific and relevant to us, but might also present valuable insight into policing tactics and instances of institutional/systemic bias. Importantly, a statistical investigation in this dataset is grounded in prior literature. Ravi Shroff, a Professor at NYU, used the same dataset to find that across American cities, black drivers are categorically more likely to be stopped than white drivers. Moreover, men are more likely to be traffic stopped, street stopped, and arrested.

Our research question is as follows: Do age, race, and gender affect the frequency, outcomes, and causes of traffic stops in Durham? More granularly, we hypothesize that black, male, and younger drivers are more likely to be searched, frisked, and arrested than other groups of Durham drivers.

Our data is sourced from the Stanford Open Policing Project [1]. This organization has been actively collecting, standardizing, and releasing records from traffic stops made by law enforcement all across the country. Out of the over 200 million records that the organization has released, we have decided to narrow our analysis to just deal with traffic stops made in Durham, as it's the most relevant to us. We are also specifically using the most recent 3 years of data available from the organization (2013 - 2015) to make the data more manageable and feasible to work with in a smaller time frame as well as to ensure our findings/analysis are as relevant and current as possible.

Some of the most relevant variables are evidently subject_age, subject_race, and subject_sex, as these are the identifying attributes that we are using as categories for our analysis and comparisons. Other variables, such as outcome (determining the outcome of the stop, e.g. arrest or citation), search_basis (determining how a search was authorized, e.g. probable cause or consent), and some boolean variables (e.g. search_vehicle, search_person, contraband_drugs) will all be very important and relevant for our analysis on the different types of traffic stops that different groups of people encountered.

[1] https://openpolicing.stanford.edu/

Methodology

Additional variables and data-frames were formulated in order to achieve our research goals.

The "num_stops" data-frame consists of 703 observations of two variables. The first column is comprised of the ID numbers of each Durham PD officer that made one or more traffic stops between the years 2013 and 2015. The second column records the number of stops made by each respective Durham PD officer.

The "stops_by_race" data-frame consists of 6 observations of two variables, subject race and a new variable, "stop_pct_by_race," which reports the percentage of traffic stops that involved subjects of each race.

The "race_analysis" data-frame adds six new variables to the initial data-frame consisting of all Durham traffic stops between 2013 and 2015. These six new variables are each of the six racial demographics accounted for in the data (black, white, hispanic, asian/pacific islander, unknown, and other). If the subject race for an observation aligns with one of the six variables, then that variable is assigned a value of "1" while the others are assigned a value of "0."

Like the prevous data-frame, the "gender_analysis" dataframe adds two new variables to the initial data-frame ("male" and "female"), and operates under the same method of assigning a value of "1" to the variable that matches and a "0" to other variable.

The "black_stops_above_racial_pct" variable indicates with a "yes" or "no" whether or not a Durham PD Officer with greater than 20 stops stopped black residents at a rate higher than that of the black racial demographic percentage in Durham (38.65%).

The "male_stops_above_demographic_pct" variable indicates with a "yes" or "no" whether or not a Durham PD Officer with greater than 20 stops stopped male residents at a rate higher than that of the male demographic percentage in Durham (47.70%).

Regarding the methods we employed in order to evaluate our hypotheses, for race we will break down the percentage of stops made in the dataset by racial category. While this might not be the end-all-be-all to proving that some groups get pulled over at certain rates, this would highlight how the dataset, at face value, has a racial breakdown that indicates that stops differ by race. Next, to identify whether or not certain officers tend to arrest Black drivers at a higher rate, we will find out the percentage of arrests by race for each officer. We will then filter for officers who have a significant number of arrests (using twenty arrests as a meaningful benchmark and see if the breakdown of their Black arrests is higher than the population of Black residents. Lastly, we will conduct a hypothesis test to determine whether the true difference in proportions of Black and White searches being conducted was actually zero. Our statistical analysis here is meaningful as well: we will filter the dataset to include whether or not a certain driver was Black or White and whether or not a search was conducted along with use an observed statistic of the difference in proportions. Methodologically, our race analysis accounts for officers with not enough traffic stops, creates meaningful variables that help facilitate plots of officers who arrest higher than the Durham racial percentage of Black individuals, and allows for a hypothesis test of the difference in proportions.

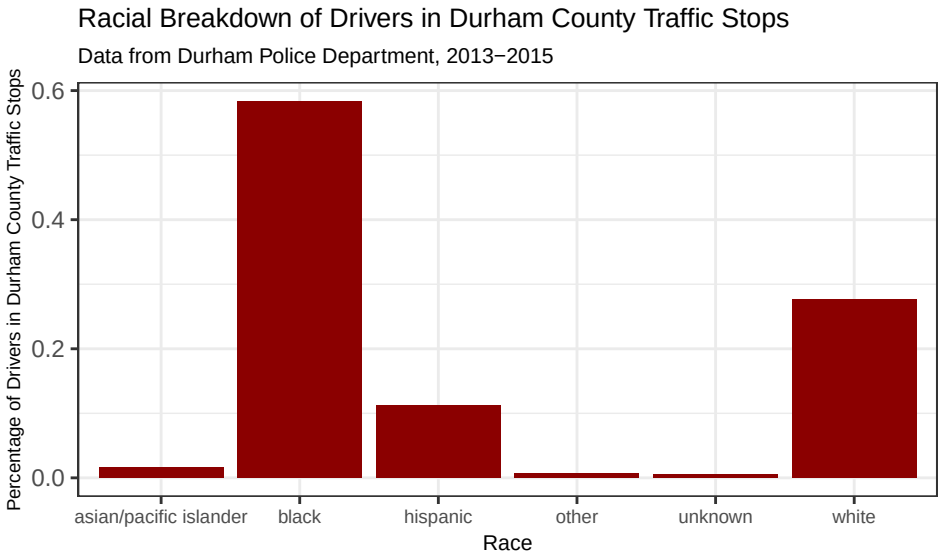
Regarding our analysis for gender and age, we hope to calculate the percentage of individuals of each age above the age of 15 (which is an outlier in the dataset) arrested. With this, we can create a linear regression model that evaluates the relationship between age and the probability of being arrested based on the data. Although more entries would probably lend to more fruitful analysis, it is statistically sound to interpret the linear regression model we hope to create as an evaluation of the likelihood of being arrested given the driver's age. Using a regression to determine this relationship is valid as it helps ascertain if there is a correlation. Concerning gender, we hope to discern the differences in frisks, searches, and arrests based on whether the driver is male or female in the dataset. We hope to use a baseline of twenty arrests with this data and evaluate if there are officers who are arresting men at a greater rate than the population percentage of men in Durham. Lastly, we hope to create a few different plots comparing frisks, arrests, and searches on men to women in Durham. This analysis is statistically meaningful, as it is not useful to use another method of analysis like linear regression makes less sense when you are doing analysis with categories like gender nor is it essential to use a hypothesis test. Creating plots that just compare for searches, arrests, and frisks are comprehensive enough to do a gender analysis.

Results

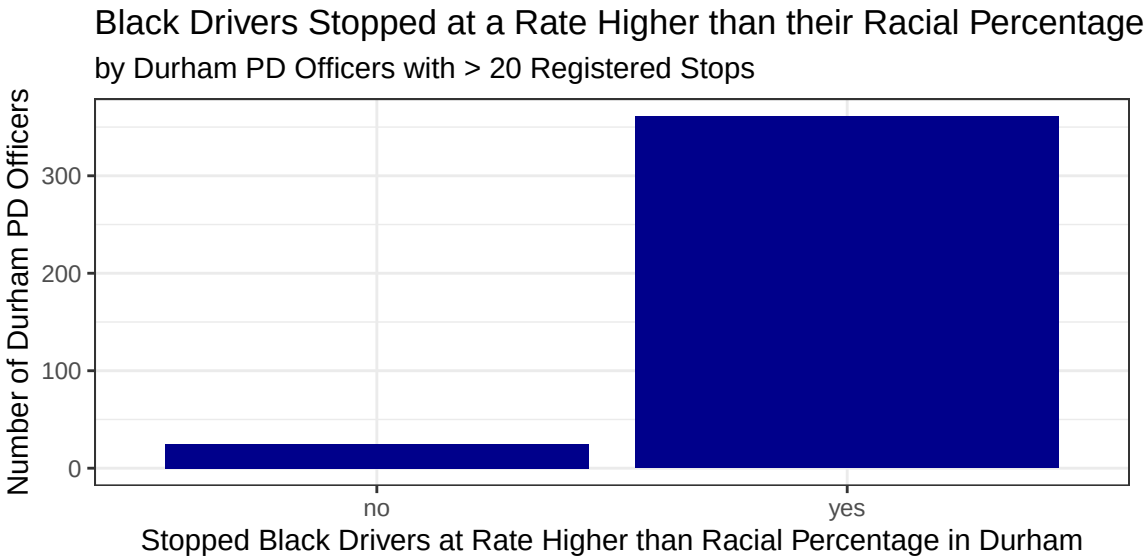
```
## # A tibble: 1 x 1
##   mean_stops
##        <dbl>
## 1       110.
```

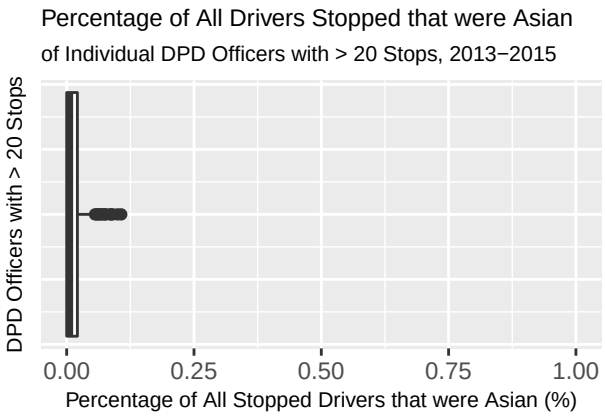
The mean summary statistic of the number of stops per officer is 109.8834.

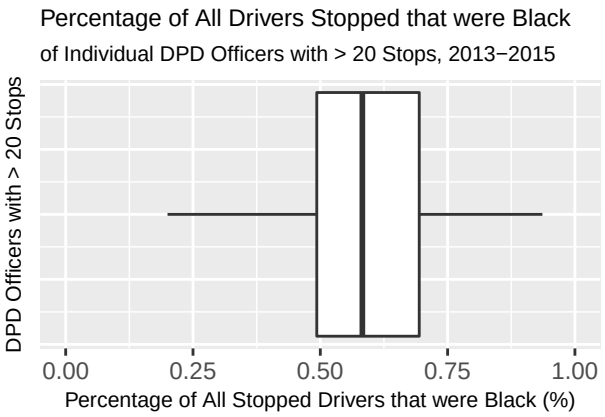
The stop % by race breakdown highlights how Asian/Pacific-Islander drivers make up 1.66% of the traffic stops, Black drivers make up 58.39% of the traffic stops, Hispanic drivers make up 11.19% of the traffic stops, White drivers make up 27.64% of the traffic stops, 'Other' drivers make up 0.64% of traffic stops, and Unknown drivers make up .49% of traffic stops in the dataset.



This plot highlightes % of traffic stop breakdown by race.

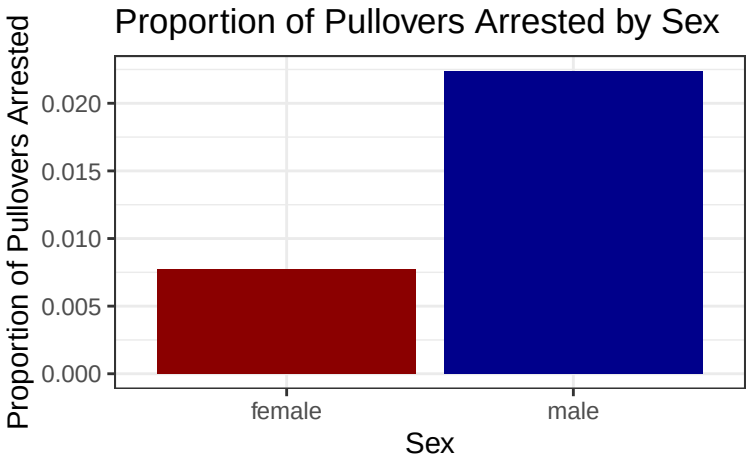


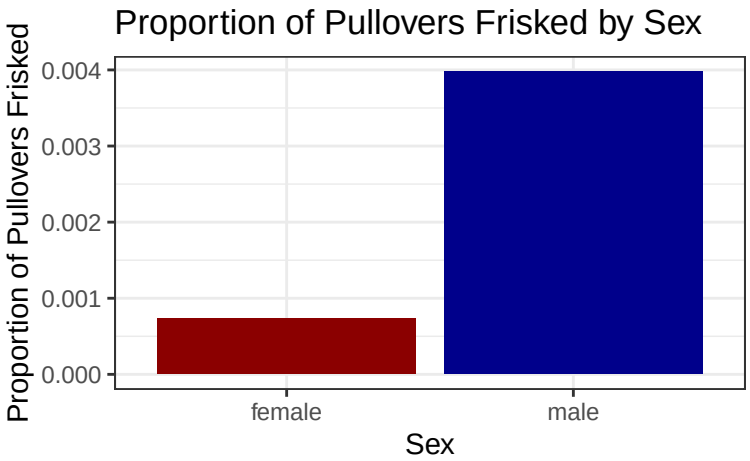

Percentage of All Drivers Stopped that were Asian
of Individual DPD Officers with > 20 Stops, 2013−2015
DPD Officers with > 20 Stops
0.00
0.25
0.50
0.75
1.00
Percentage of All Stopped Drivers that were Asian (%)

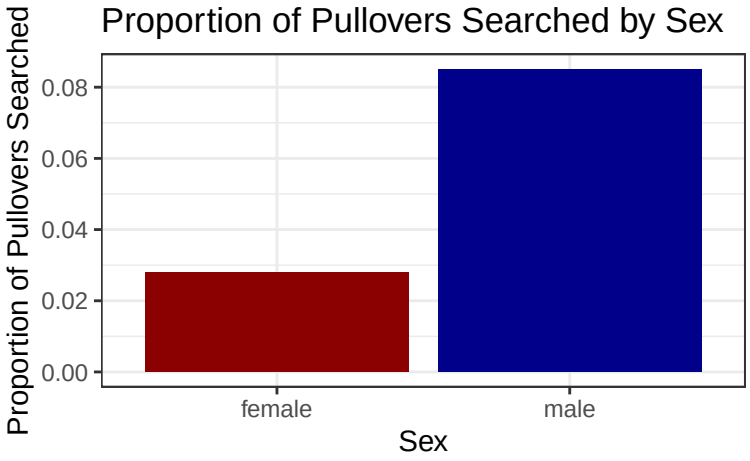

Percentage of All Drivers Stopped that were Black
of Individual DPD Officers with > 20 Stops, 2013−2015
DPD Officers with > 20 Stops
0.00
0.25
0.50
0.75
1.00
Percentage of All Stopped Drivers that were Black (%)


Percentage of All Drivers Stopped that were Hispanic
of Individual DPD Officers with > 20 Stops, 2013−2015
DPD Officers with > 20 Stops
0.00
0.25
0.50
0.75
1.00
Percentage of All Stopped Drivers that were Hispanic (%)


Percentage of All Drivers Stopped that were White
of Individual DPD Officers with > 20 Stops, 2013−2015
DPD Officers with > 20 Stops
0.00
0.25
0.50
0.75
1.00
Percentage of All Stopped Drivers that were White (%)

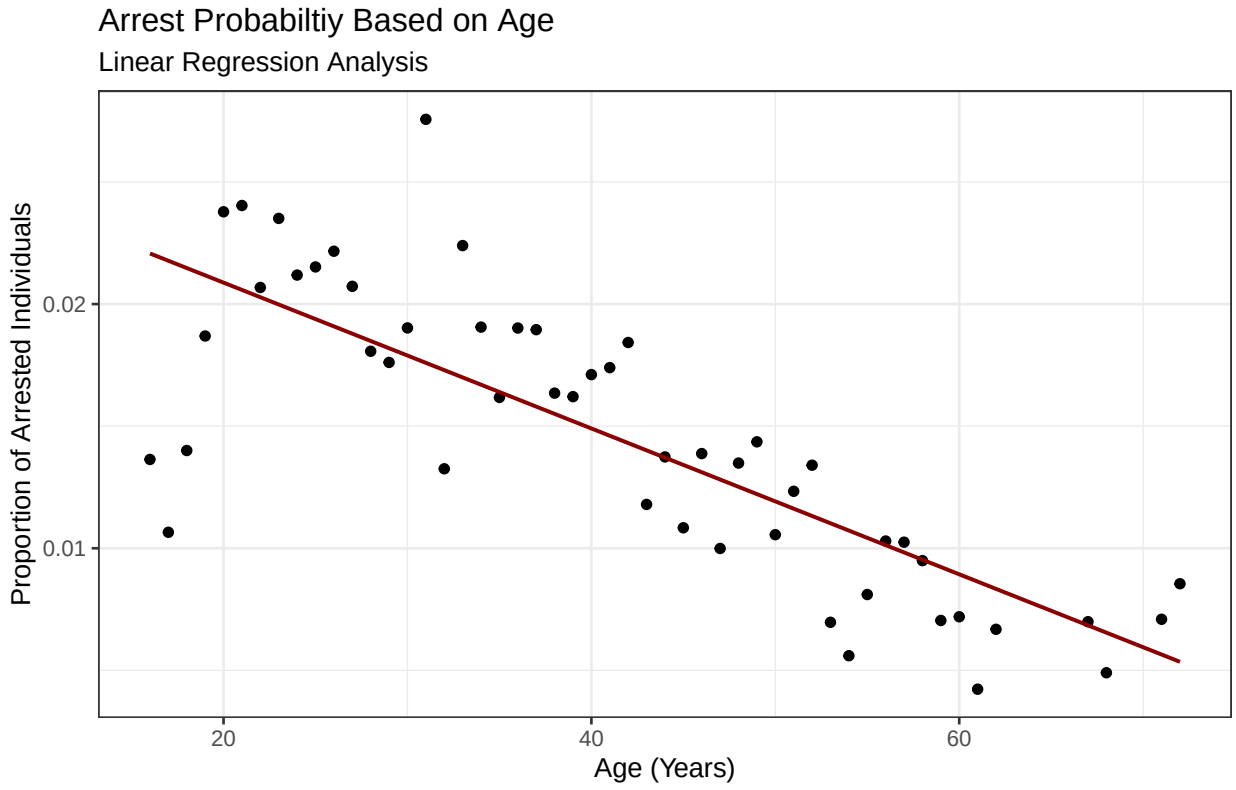

Proportion of Pullovers Arrested by Sex
Proportion of Pullovers Arrested
0.020
0.015
0.010
0.005
0.000
female
male
Sex


Proportion of Pullovers Frisked by Sex
Proportion of Pullovers Frisked
0.004
0.003
0.002
0.001
0.000
female
male
Sex


Proportion of Pullovers Searched by Sex
Proportion of Pullovers Searched
0.08
0.06
0.04
0.02
0.00
female
male
Sex

Age based linear regression:

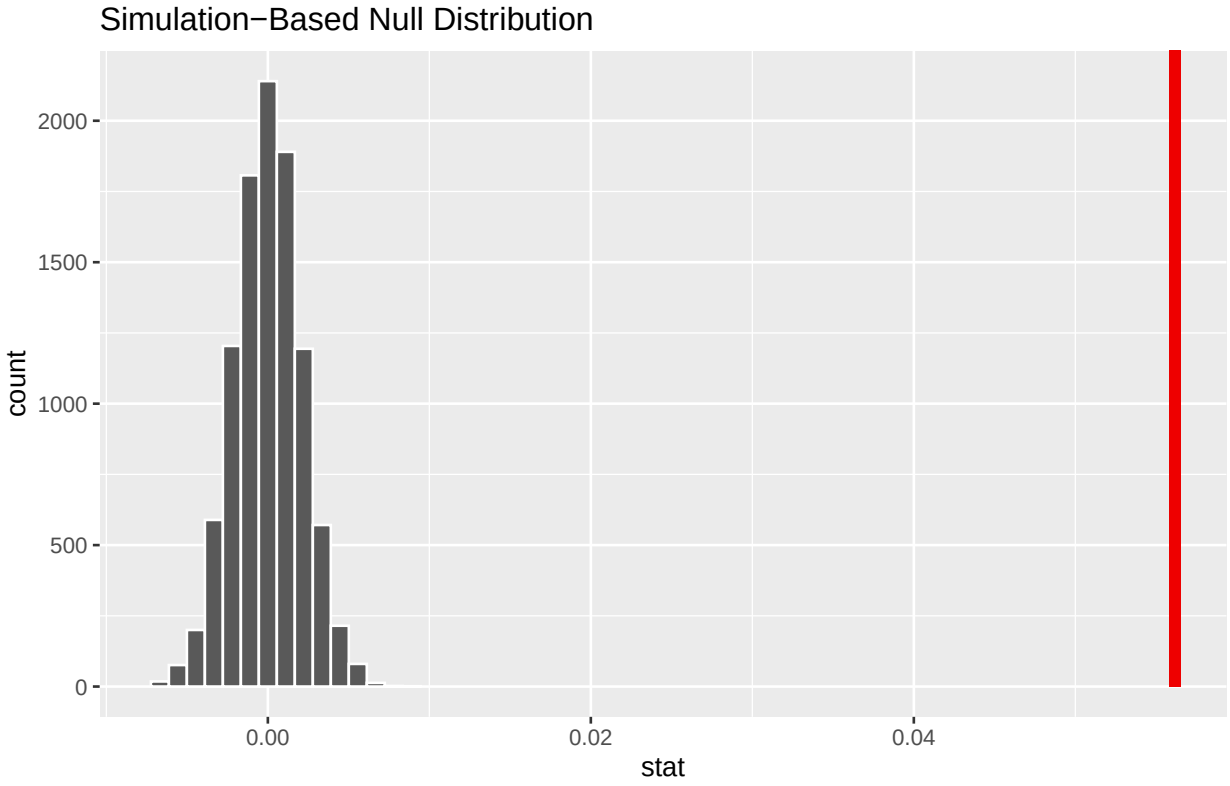
```
## # A tibble: 2 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.0269   0.00145       18.5  8.73e-24
## 2 subject_age -0.000299 0.0000328     -9.10 4.12e-12
```



Race based hypothesis test:

$$H_0 : p_b - p_w = 0 \text{ vs. } H_a : p_b - p_w > 0$$

Null hypothesis: difference in proportions of black and white people that are searched in Durham after being pulled over is equal to zero Alternate hypothesis: difference in proportions of black and white people that are searched in Durham after being pulled over is larger than zero




```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

The chance of observing a sample difference in proportions equal to or larger than 0.0562 is 0. Therefore, we reject the null hypothesis at the 0.001 significance level. The data provides evidence that black drivers in Durham are searched after being pulled over at a disproportionate rate compared to white drivers.

Discussion

From our data, we have learned that Black Drivers are pulled over at higher rates than White Drivers. This is statistically confirmed by our hypothesis test where the chance of observing a sample difference greater than or equal to the observed statistic of 0.0562 was 0. This hypothesis test is illustrative of the fact that the null hypothesis, which presumed no difference between searches of Black and White drivers, ought to be rejected. Furthermore, from the created boxplots that highlight the % of drivers stopped by race, that the median stops for Black drivers is significantly higher than the median stops for White drivers, Hispanic drivers, and Asian/Pacific-Islander drivers (for officers with more than 20 stops). Our 'yes/no' plot of officer stops show that there are are significantly more officers who stop Black drivers more than the Black population average compared to lower than the population average. The first plot is less informative but highlights that the Durham dataset features more data of Black, White, and Hispanic stops than other racial categories.

Regarding gender, our barplots confirm our initial hypothesis that men are arrested more than women. However, the data goes even further to suggest that men are arrested more, frisked more, and searched more than women.

On age, our linear regression suggests that for every year that the driver gets older, the proportion of arrested individuals by -0.000298631 (slope). The intercept suggests that if the subject age is 0, the proportion of arrested individuals predicted to be in the dataset is 0.026850861 which is a meaningless result (intercept). Essentially, we can interpret this resullt as indicative that the older the driver is, the lower the proportion of arrested individuals there are, and concurrently, the chance of being arrested is lower.

It is important to mention certain limitations of our methods. Perhaps most important is our assumption that all races participate in behaviors that would lead to a vehicle being pulled over at equal rates. This, of course, is not necessarily the case. For example, in 2019, Asians accounted for 2.5% of all DUI arrests but made up 5.7% of the U.S. population. Accounting for these differences in crime rates may impact conclusions, and should be factored into future work in order to improve analysis. In addition, much of our data was limited to include only Durham police officers that made greater than 20 traffic stops. However, our conclusions are limited by the fact that whether or not a police officer having made greater than 20 stops is a good enough benchmark to select just police officers that could provide representative data is not definite—this benchmark was one that we felt to be appropriate, but that cannot be said with certainty. Resolving this confidence issue would improve the accuracy of our work. Furthermore, our dataset covers a span of just three years—analyzing data over a longer period of time would provide more accurate and reliable results. Regarding the appropriateness of our statistical analyses, we believe that our methodology was both relevant and suitable to answer the questions that we set out to answer at the start.

Throughout the course of our research, we identified several ideas for future work. One of our ideas was to repeat our methodology with traffic stop data from nearby Orange County, North Carolina, home to Chapel Hill. Relative to Durham County, Orange County has a significantly higher white population and significantly lower minority population across the board. How the racial demographics of an area would affect racial trends in traffic stops is an important question that should be looked into to better solve the issue of racial bias in policing. Another idea we had was to use multivariate linear regression to determine the most important factors contributing to arrests and frisks after being stopped by Durham County Police, enabling us to locate the existence of potential biases and address them in the case that it is warranted.

References

"Driving While Black: Racial Profiling on Our Nation's Highways." n.d. Accessed December 6, 2021. https://www.aclu.org/report/driving-while-black-racial-profiling-our-nations-highways.

"Durham, North Carolina Population 2021." 2021. Accessed December 6, 2021. https://worldpopulationreview.com/us-cities/durham-nc-population.

Horsch, Lauren. 2016. "The Durham City Council Paid $90K for a Study That Found No Institutional Racism at the Police Department." October 12, 2016. https://indyweek.com/news/durham/durham-city-council-paid-90k-study-found-institutional-racism-police-department/.

Stanford University. n.d. "Stanford Sentencing Project Data." THE STANFORD OPEN POLICING PROJECT. Accessed December 6, 2021. https://openpolicing.stanford.edu/data/.

United States Census Bureau. 2019. "QuickFacts: Durham County, North Carolina." Accessed December 5, 2021. https://www.census.gov/quickfacts/durhamcountynorthcarolina.