Introduction

Berkeley, California is known as one of the cities most progressive police departments in the US. However, recent reports have said that the city is better than most, there are still racial disparities found in recent crime data from 2010-2015. This report will analyze more recent, publicly available stop data, as well as attempt to consider newly RIPA (Racial Identity and Profiling Act) compliant data that's been collected since late 2020, to see if these racial disparities still exist in more recent interactions with the public.

- How much does prior perception of race play a role in the result of a stop, particularly if the citizen being stopped is a person of color?
- Can environmental factors like income, residential demographics, and/or amount of previous police activity affect and/or reduce how many stops occur in an area? Do these vary by neighborhood, or are they largely consistent throughout the city?
- Can RIPA-Compliant data reveal more nuanced relationships between features of a stop? Can and/or will more detailed data echo previous discoveries about race and law enforcement?

Background

In 2018, the Center for Police Equity released a report on the Berkeley Police Department saying that there were racial disparities in arrest rates between white and nonwhite people of Berkeley. Their study found that black and hispanic citizens were substantially more likely to be stopped than white citizens, and more likely to be searched as well. Along other lines however, the study also found that if a search was conducted, black and hispanic citizens were less likely than white citizens to actually be arrested. Overall statistics showed that black citizens were substantially more likely to be arrested than white citizens regardless of if a search was conducted, and that despite representing less of Berkeley's population, were also more likely to to be subjected to use of force. This analysis aims to use available stop data to see if some of these trends are still in effect: specifically arrest rates post stop, and arrest/use of force rates by population.

The report also mentioned that neighborhood factors like distance from the university, neighborhood racial composition, and previous crime rates all affected stops by census tract. This analysis will aim to consider this as well.

The Data and Methods

The data used in this project was collected by the city of Berkeley Police Department from 2015-2022, and downloaded in March of 2022 from its publicly available open access data portal.[1] The data comes in two formats: RIPA[2] and Non-RIPA Compliant data. All data since October of 2020 is RIPA-Compliant and as a result is more rich. Non-Ripa Compliant data has been kept to attempt to comprehensively represent the city, but some variables have been changed to match with new RIPA terminology, and certain models may vary in observation size due to lacking shared features between the two datasets. This is unfortunate, but some key assumptions can still be tested regardless of the differences in data.

Crime and arrest data is also not directly available through the city, and will be left out of this analysis. While it is still beneficial to look at stops to analyze police activity, it's important to clarify the difference between stops and arrests, and admit that while comprehensive, the presently available stops data do not paint the whole picture.

[1] Berkeley PD - Stop Data (Jan 26, 2015 to Sep 30, 2020)
^ Berkeley PD - Stop Data (October 1, 2020 - Present)
[2] Racial and Identity Profiling Act (RIPA)

Arrest is the main dependent variable examined in this study. It's studied in two ways: 1) with a few logistic regression models that compare the likelihood of arrest or no arrest given certain conditions and 2) a multilinear regression looking at the number of arrests per census tract in Berkeley and how we can reduce this. Alternative dependent variables include *noactions* and *warning*, although these aren't given as much focus.

Longstop and/or *duration of stop* are also in focus as a dependent variable, but only apply for the data from 2020 and later. RIPA data mandates that the length of the stop be recorded. In the case of this study, and stops over 270 minutes, or 4.5 hrs were excluded. In the context of this study, this classification variable essentially asks "based on certain conditions, what are the chances a person has a longer stop than people usually expect?"

Forceused is another dependent variable, although given the complexity of the coding[3] for actions taken (*example: 1|5|8|16|18|20|21*), it's unclear how reliable of a dependent variable this is. It has been simply classified as a yes/no variable depending on if action 5

Perceived Race or Ethnicity represents the race of the person stopped. This variable is a combination of "race" from the non-ripa dataset and "perceived race" from the ripa dataset.

Race Perceived Prior to Stop is a binary variable representing (1) the subject's race was perceived by the officer before the stop and (0) the subject's race was not perceived before the stop. This data, while insightful, is only available for the models based on the RIPA-compliant data. For the purpose of this analysis, single and double race categories have been simplified into "White", "Black/African American", "Hispanic/Latino", "Asian", and "Other". Any perceived racial categories with more than 3 mentioned races were classified as "Other". The 2017 report

[3]

on Berkeley PD specifically cited concerns with enforcement towards black and hispanic populations, so these categories are focused on most here.

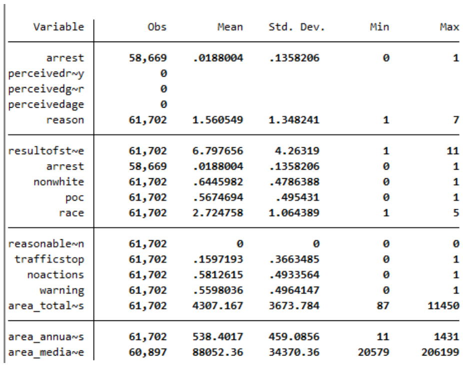
Whitepop, aapop, na_aipop, hawaiian, and mixed2 all represent specific estimated counts of racial demographics per Berkeley census tract in 2020 based on data from the US Census.[4] These are useful in the multilinear regression as well as in the mapping portion of the analysis.
Totalpop is the estimated total population per census tract from the same census data and city information from geolocation maps.[5]
Nonwhitecomp is a simple calculation of the estimated proportion of nonwhite residents (all categories except white) of total residents per census tract. This is a broad representation but is acceptable for creating simple variables regarding “whiter” neighborhoods where stops would supposedly occur less. *Nonwhite* and *poc* variables are also generated later in analysis during logistic regression.

Far and *Distancefromcal* represent another variable mentioned in the earlier study, distance from the university. Although we later find that this is at best curvilinear, it’s still worth including to see if it is still statistically significant.

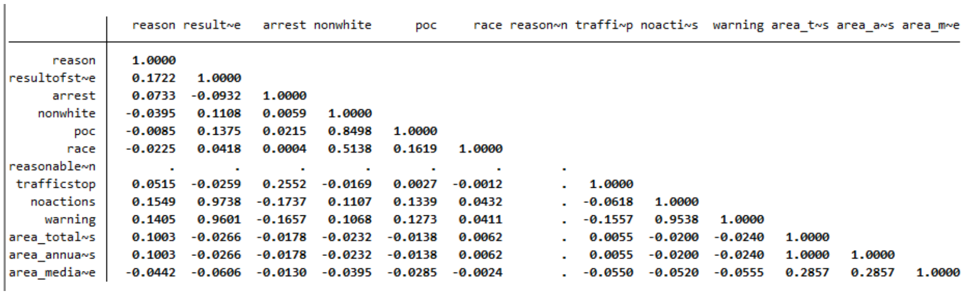
[4] American Community Survey, B02001 RACE

[5] (Census Tract Polygons 2010) Census tract polygons built from US Census Bureau 2010 decennial data for the City’s redistricting process

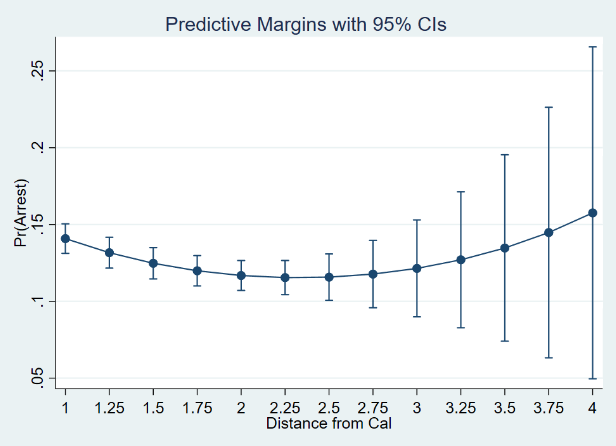
Results

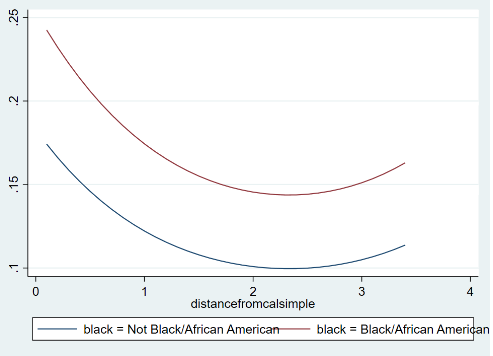


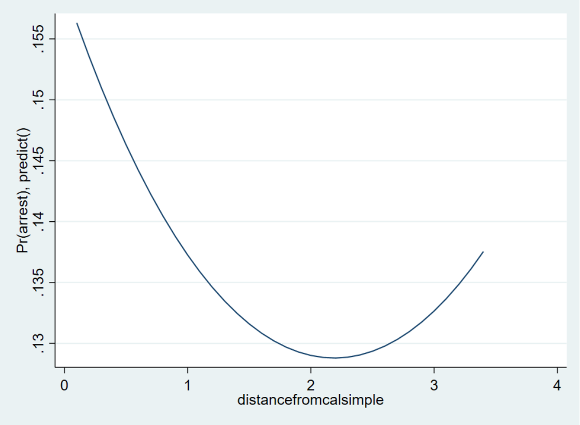
```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      arrest |     58,669    .0188004    .1358206          0          1
perceivedr~y |          0
perceivedg~r |          0
perceivedage |          0
      reason |     61,702    1.560549    1.348241          1          7
-------------+---------------------------------------------------------
resultofst~e |     61,702    6.797656     4.26319          1         11
      arrest |     58,669    .0188004    .1358206          0          1
    nonwhite |     61,702    .6445982    .4786388          0          1
         poc |     61,702    .5674694     .495431          0          1
        race |     61,702    2.724758    1.064389          1          5
-------------+---------------------------------------------------------
reasonable~n |     61,702           0           0          0          0
 trafficstop |     61,702    .1597193    .3663485          0          1
   noactions |     61,702    .5812615    .4933564          0          1
     warning |     61,702    .5598036    .4964147          0          1
area_total~s |     61,702    4307.167    3673.784         87      11450
-------------+---------------------------------------------------------
area_annua~s |     61,702    538.4017    459.0856         11       1431
area_media~e |     60,897    88052.36    34370.36      20579     206199
```



```
             |   reason result~e   arrest nonwhite      poc     race reason~n traffi~p noacti~s  warning area_t~s area_a~s area_m~e
-------------+---------------------------------------------------------------------------------------------------------------------
      reason |   1.0000
resultofst~e |   0.1722   1.0000
      arrest |   0.0733  -0.0932   1.0000
    nonwhite |  -0.0395   0.1108   0.0059   1.0000
         poc |  -0.0085   0.1375   0.0215   0.8498   1.0000
        race |  -0.0225   0.0418   0.0004   0.5138   0.1619   1.0000
reasonable~n |        .        .        .        .        .        .        .
 trafficstop |   0.0515  -0.0259   0.2552  -0.0169   0.0027  -0.0012        .   1.0000
   noactions |   0.1549   0.9738  -0.1737   0.1107   0.1339   0.0432        .  -0.0618   1.0000
     warning |   0.1405   0.9601  -0.1657   0.1068   0.1273   0.0411        .  -0.1557   0.9538   1.0000
area_total~s |   0.1003  -0.0266  -0.0178  -0.0232  -0.0138   0.0062        .   0.0055  -0.0200  -0.0240   1.0000
area_annua~s |   0.1003  -0.0266  -0.0178  -0.0232  -0.0138   0.0062        .   0.0055  -0.0200  -0.0240   1.0000   1.0000
area_media~e |  -0.0442  -0.0606  -0.0130  -0.0395  -0.0285  -0.0024        .  -0.0550  -0.0520  -0.0555   0.2857   0.2857   1.0000
```







More results are on the way. I'm building separate models from what data I can put together from the small dataset as well as the large, combined datasets. Sorry will finish over weekend and bring in next week. Very excited about this project and happy with my progress so far.

Discussion