STAT 515 Final Project Report

I. Why Police Stops Data?

I chose to do my final project using a police stops dataset from the Stanford Open Policing Project (SOPP). I believe the US police has perpetuated a number of systemic problems and continues to discriminate against minorities and those in impoverished communities. Based on knowledge of previous research into police stops, I started this analysis with a few questions to answer. Does race influence police stop outcomes? Does the time of stop – day or night – affect the number of stops more for one race over another? Do more police stops happen in areas with lower socio-economic status? What are the reasons for police stops and do they vary by sex?

The SOPP has compiled a large database of police stop data from cities across the United States on their website. I chose to work with the Nashville, Tennessee dataset because it was more complete with many variables to explore over a long period of time. It technically covers a 10 year span from 2010 to 2019, but as I only use 9 years as I will explain later. This includes over three million observations and forty-four variables with each row representing one police stop. In total it takes up one gigabyte of storage on my computer.

II. Additional Data Sources – Population, Income, Geospatial

To perform some analyses I added three supplemental datasets. When exploring racial components, I used 2018 Nashville census data to scale each race's police stop count by their population. This is a good scaling method, but it's not perfect as not every person stopped by police in Nashville actually lives in the city. To include this data, I copied each race's population from the Nashville census.gov quickfacts table into Excel and loaded with R's *read_excel* function.

I wanted to compare police stops and median income by zip code in a juxtaposed choropleth map. This required downloading a Davidson county shapefile from data.nashville.gov with zip code boundaries. The shapefile can be read in R through the *sf* package, which loads a dataframe with post office names, zip codes, and geometries (a list of polygon type rows with latitude, longitude coordinates to draw zip code boundaries).

I got median income data through many API requests to justicemaps.org. I created a few R functions to accomplish this – one that computes a unique set of latitude and longitudes from the Nashville shapefile *geometry* list and one that loops over this set and makes an API *get* request for each coordinate pair. This returned median income by census tract with latitude and longitude variables that I could join back to the original police stop data.

III. Data Description – Police Stops

Returning to our main police stops dataset, Figure 1 shows a set of sixteen variables explored in this project. There are temporal variables (date, year, month, time), geospatial variables (latitude, longitude), and demographic variables for the person stopped (age, race, sex). Each stop also indicates the violation, or reason for the stop, and the outcome – either a mere warning, a citation, or an arrest. Lastly, the police recorded a few logical (boolean) variables, whether a frisk was performed, a search conducted, contraband found, drugs found, or weapons found. Figure 1 uses the R *glimpse* function to show the variable types as well as the first few observations for each variable. Near the bottom I can see contraband is False for the fifth observation when a search was conducted, while the other rows are missing. I can guess from this pattern (and confirm with a frequency table) that contraband found is only recorded when a search was conducted, and logically this makes sense.

```
Rows: 3,078,116
Columns: 16
$ date              <date> 2010-10-10, 2010-10-10, 2010-10-10, 2010-10-
$ year              <dbl> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 20
$ month             <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 
$ time              <time>       NA, 10:00:00, 10:00:00, 22:00:00, 01:0
$ lat               <dbl> 36.19, 36.16, 36.12, 36.09, 36.18, 36.29, 36
$ lng               <dbl> -86.80, -86.74, -86.90, -86.65, -86.81, -86.7
$ subject_age       <int> 27, 18, 52, 25, 21, 26, 37, 33, 33, 49, 18, 
$ subject_race      <fct> black, white, white, white, black, white, wh
$ subject_sex       <fct> male, male, male, male, male, female, male, m
$ violation         <fct> investigative stop, moving traffic violation
$ outcome           <fct> warning, citation, warning, warning, warning
$ frisk_performed   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL
$ search_conducted  <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FAL
$ contraband_found  <lgl> NA, NA, NA, NA, FALSE, NA, NA, NA, NA, NA, NA
$ contraband_drugs  <lgl> NA, NA, NA, NA, FALSE, NA, NA, NA, NA, NA, NA
$ contraband_weapons <lgl> NA, NA, NA, NA, FALSE, NA, NA, NA, NA, NA, NA
```

Figure 1: Glimpse of Police Stops Data

The majority of the variables are categorical or logical, not continuous, and thus making a histogram or scatterplot matrix is out of the question. Instead I will begin data exploration using base R's *count* function to compute frequencies over a subset of variables. Fortunately, SOPP included a great tutorial on their website that aided this EDA process.

First up is the count of police stops by year (Table 1). Two trends immediately stand out – 2019 has many fewer stops than all other years and stops spike in 2014 then decline each following year. Regarding the former lack of counts, all further analyses will exclude 2019 as it is not a full representative year. I can confirm by looking at counts by year and month as well (Table 2).

year	n
<dbl>	<int>
2010	310622
2011	393248
2012	444146
2013	412695
2014	413114
2015	357261
2016	297248
2017	245565
2018	204217
2019	14235

Table 1: Stops by Year

	1	2	3	4	5	6	7	8	9	10	11	12
2010	28282	25731	29698	28185	19448	25065	25525	26211	27093	27246	26047	22091
2011	29657	27671	35503	33074	36627	33060	31776	37483	32953	32706	30010	32728
2012	43890	38925	39540	38375	40475	38409	33439	35429	33056	35723	35804	31081
2013	41200	35689	37861	34561	35777	35279	34797	32700	30952	33160	30573	30146
2014	41359	30539	39989	39304	37273	35599	30681	30561	32686	33051	31697	30375
2015	37520	26897	32198	34771	33258	34566	28875	28703	27467	26412	23784	22810
2016	29846	27796	29978	26222	24403	26242	24629	24323	23414	20774	19745	19876
2017	23877	21163	21723	22149	20515	19956	22387	21476	18699	19360	17746	16514
2018	25358	20609	22828	21079	18129	17014	17574	17996	14781	13770	10369	4710
2019	5814	4398	4023	0	0	0	0	0	0	0	0	0

Table 2: Stops by Year and Month

Next I examined the break down of police stops by subject race and sex. I see that some sexes and races are missing, unknown, or other. I chose to remove these rows instead of imputing values as there is plenty of data to use in models after their removal.

subject_race	male	female	`NA`
<fct>	<int>	<int>	<int>
asian/pacific islander	26852	14666	150
black	644046	517844	3981
hispanic	116355	47903	556
white	1004390	660698	5785
other	7628	2757	12
unknown	26654	7970	2254
NA	1118	648	84

Table 3: Stops by Race and Sex

With stop counts higher for males than females for every race, I move on to look at stop violations, or the stop reason, by sex. Across every violation category, except child restraint at the very bottom, males are stopped by police more frequently than females. The top two violations – moving traffic and vehicle equipment violations – make up the vast majority of stops, accounting for 2.5 out of 3 million stops.

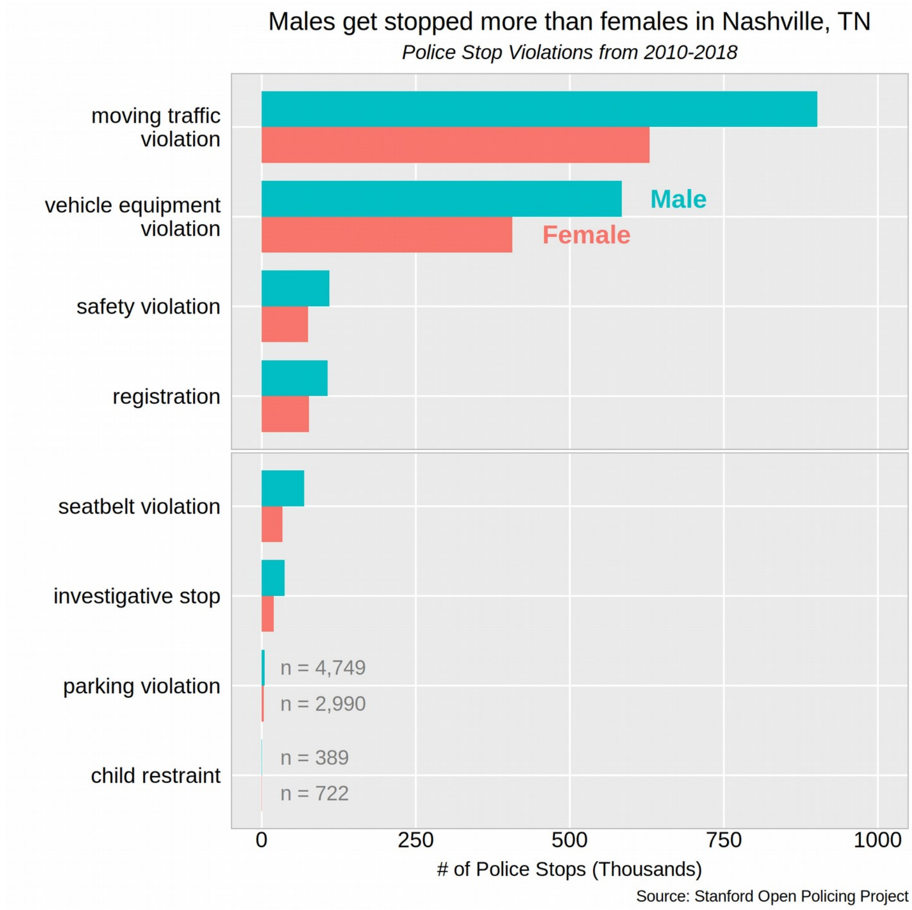

Figure 2: Stops by Violation Type and Sex

By loading the Nashville census data I can scale the stop counts by a race's population and show the number of stops per one thousand persons by race across the three outcomes. I can use this plot in Figure 3 to see if any new patterns emerge regarding outcome discrepancy by race. I abbreviated Asian/Pacific Islander to just Asian to more effectively use plotting space. Across each outcome category – warning, citation, and arrest – black persons are stopped the most per one thousand persons, followed by white, hispanic, and lastly asian.

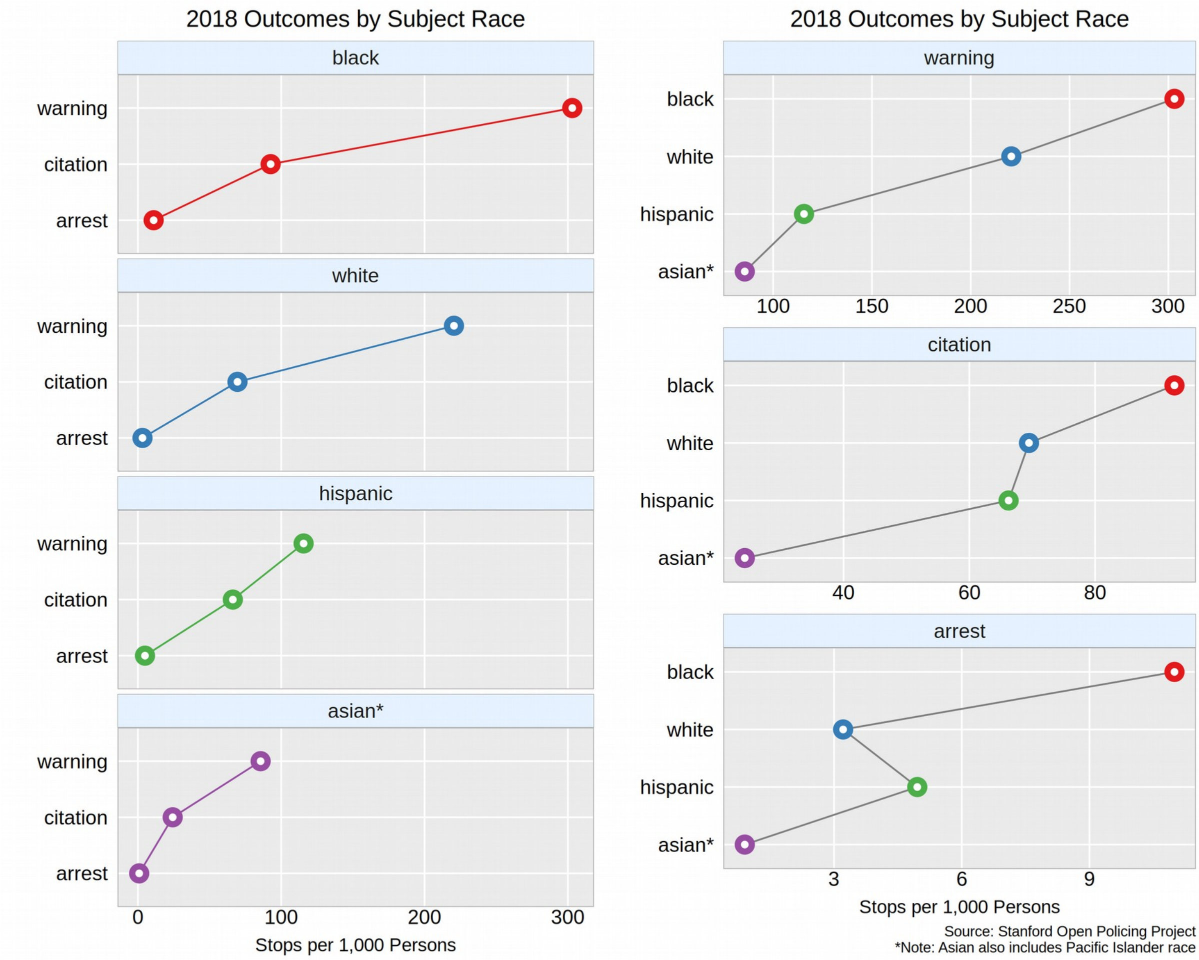

Figure 3: 2018 Stop Outcomes by Race Juxtaposed Plot