Project of Data visualization

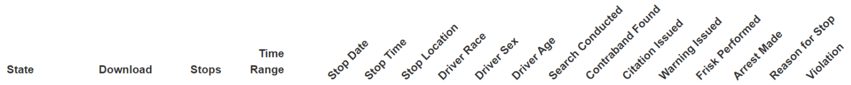
1 Dataset

Stanford Open Policing Project, 100 million traffic stops in the USA.

More than 20 million Americans are stopped for traffic violations each year, making it one of the most common ways for the public to interface with the police. However, there has never been a comprehensive national repository containing information on these interactions. To alter that, the Open Policing Stanford dataset collaborated with Big Local News to collect a dataset of over 200 million traffic stops done in dozens of cities and states throughout the US, making it the largest such endeavor to date.

In early 2020, Nature Human Behaviour released a large-scale examination of approximately 100 million of these traffic stop records.

It has become a valuable resource for politicians, scholars, journalists, and campaigners striving to improve policing via data.



State	Download	Stops	Time Range	Stop Date	Stop Time	Stop Location	Driver Race	Driver Sex	Driver Age	Search Conducted	Contraband Found	Citation Issued	Warning Issued	Frisk Performed	Arrest Made	Reason for Stop	Violation

Figure 1 – Open Policing Stanford dataset features

The main features of the dataset we plan to use are the *date*, *county_name*, *subject_race* and *subject_sex*. We will also use other features in our visualization such as *search_conducted*, *contraband_found* and *arrest_made*. Since the dataset is collection of smaller datasets representing police stops made by state patrols or city police departments, we chose to use the data from the California and Texas state patrol. One of the features (*contraband_found*) is present in only 4.1% of the California state patrol dataset). Though our main goal will be to use *date*, *county_name*, *subject_race* and *subject_sex*, if we were to need *countraband_found* we would scale this part of the visualization to one city in each state since this feature is more prevalent in the city police departments datasets.

2 Problematic

2.1 Visualization scope

We want to investigate a potential variation in racial bias in police stops from 2009 to 2016. We also want to explore, in case such variations exist, if those are constant across different states of America. To simplify our investigation, we will focus our research on California (mostly Democrat) and Texas (mostly Republican).

With our chronological dataset, we would like to visualize :

- Profiles who are most likely of getting stopped by police
- Geographical representation by county of stops in California and Texas
- Interactive evolution of bias across time
- Potential other biases like gender and age

- Explore potential interaction in the features (ethnicity, gender, age ..) for police stops bias
- Measure Hit rate (the percent of searches that find contraband)

2.2 Motivation and target audience

Our goal is to make a visualization where anyone can find answers on all these questions in an interactive way. This study could be very valuable as a sensitization tool showcasing how such biases evolved across time and might be different across states considering their main political party. It could also be very useful for the formation of new police officers, highlighting the danger that those biases represent.

3 Exploratory data Analysis

3.1 Pre-processing

Most of the data was already pre-processed before being made available by the Stanford Open Policing Project. However, we still made some pre-processing to remove some NaN values or change some of the data for easier manipulation later on.

First, we only use the following rows from the data : *date*, *county_name*, *subject_race*, *subject_sex*, *search_conducted*, *search_basis* and *outcome*. We added a new column *date_dt* where the date is represented in the Datetime format for easier manipulation. For the rest of the pre-processing, we did the following changes :

- For rows where *search_conducted* was equal to $False$, the *search_basis* was equal to NaN. We changed it so it is equal to *no_search* instead.
- We replaced the NaN values in the outcome column by the value *unknown*.
- We dropped the rows for which we didn't have a *county_name*.

We didn't need to drop other rows as the data was complete. Though, as the data for California ranged from 2009 to 2016 while for Texas it ranged from 2006 to 2016, we removed the data for Texas that was from before 2009.

3.2 Data analysis

Since one of the goals of the visualization is to look at the evolution of the number of stops (or other metrics) during the time period $2009 - 2016$ we looked at how many stops there are per year and location. The results can be found Table 1.

	California	Texas
2009	2 494 343	2 441 270
2010	5 060 321	2 524 704
2011	5 047 002	2 587 556
2012	4 588 924	2 435 070
2013	4 415 601	2 133 954
2014	4 135 931	1 878 458
2015	4 027 143	1 745 353
2016	1 914 268	1 831 975
Total	31 683 532	17 578 339

TABLE 1 – Number of stops by year for each state.

One of the second goals would be to visualize the evolution of the possible disparities in traffic stops, based on age and gender. Table 2 summarizes the distribution of those attributes for the two states we selected.

		California	Texas
Sex	Male	22 100 300	17 254 281
	Female	9 583 138	7 971 522
Race	White	14 024 423	14 440 205
	Hispanic	10 499 508	7 286 405
	Black	2 601 967	2 488 330
	Asian / Pacific islander	2 173 708	377 323
	Other	2 383 926	36 142
	Unknown	0	597 398
	Searches	1 073 027	543 466

TABLE 2 – Distribution of the driver's sex and race for the two states.

4 Related work

Multiple articles have been written based on this dataset. In the article of Pierson et al. [1], they wanted to show the racial disparity in policing in the United States. To do so, they compared the stops and searches between race and they found evidence that the bar for searching black and Hispanic drivers was lower than that for searching white drivers.

Other studies have been conducted on the hit rate of all searches. They wanted to find a way to rigorously assess the racial bias. Contrary to the first article, they did not want to compare the rate at which whites and minorities are treated favourably but the success rate of the decision to search [2].

This approach is innovative in many ways. Firstly, we will look at the evolution of bias across time. Indeed, a lot of studies have already been done to show racial bias in the United States, still none included a temporal analysis. With our approach, we will differentiate social bias across time. Secondly, we will include the sex and age of the subjects arrested. Finally, we will focus on two USA states having different political opinions, one being mostly Republican and the other having a majority of Democrats.

For the map of each state, we will do a visualization like this Interactive map of the world with a slider. Though we will also breakdown the two states we investigate into their respective counties. Also for the hit rate we will do a plot similar to [1] (Fig. 3).

This dataset was also used in a project for the course ADA by one person in the group, with the report that can be found here. Though the data used was from city police departments mostly and a plot similar to Fig. 3 of [1] was made but using the race of the officer on top of the race of the driver.

Références

[1] E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, *et al.*, "A large-scale analysis of racial disparities in police stops across the united states," *Nature human behaviour*, vol. 4, no. 7, pp. 736–745, 2020.

[2] C. Simoiu, S. Corbett-Davies, and S. Goel, "The problem of infra-marginality in outcome tests for discrimination," *The Annals of Applied Statistics*, vol. 11, no. 3, pp. 1193–1216, 2017.