Data visualization project

Bias in Police stops

From 2009 to 2016

Contents

Bias in police stops

1 Motivation

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.

It has become a valuable resource for politicians, scholars, journalists, and campaigners striving to improve policing via data. In our turn, we too want to help ameliorate policing and we will do so in our visualizations by raising awareness on potential racial and gender biases in police stops.

2 Our goal

We want to investigate a potential variation in racial bias in police stops from 2009 to 2016. We also want to explore, in cases such variations exist, if those are constant across different states of America. We wanted to accomplish this by creating a big map of the United States where we would have a preview of the potential bias for each state. In this visualization it would be possible to select two different states to be able to compare them with greater detail.

In the visualization you would be able to:

- Observe the evolution of police stops in each county of each state in the US.
- Observe the evolution of gender and racial bias again independently for each county.
- Explore potential interactions in the features (ethnicity, gender, age ...) for police stops bias
- Establish a profile who would be more likely to be stopped by police.
- Observe the evolution of the police searched of each ethnicity.

Dive in with us in US Police departments

3 First difficulty, data

3.1 Data size

To give you an idea of how massive is the dataset we are using, considering only police stops in California, we have to process more than 30 million rows of data.This impressive quantity of police stops in California amounts for a 6.9 Gb of data. We rapidly realized that loading data for each and every state in the US to pre-process them would be a very tedious task and it was not where we wanted to put our main efforts.

3.2 Data homogeneity

Another problem we faced with Open Policing Stanford dataset is the non-homogeneity of the data. Features which are essential to our project and analysis are completely missing in a big proportion of states and cities, such as *countrabound_found*, a feature we use to compute the hit rate of police stops (ie: a search was conducted after a police stop and contraband was found). This conducted us to focus our study towards larger states and cities where there is a consistently broader choice of features. Note that even in big cities of California such as Los Angeles the *countrabound_found* feature is only available in 4% of police stops. Thus for the map of the number of stops by county, we used data from the California and Texas states patrols, while for the hit rate we used data from three big cities in each state. Again, this is because we needed the data to contain information on if a search was conducted and contraband was found to compute the hit rate.

To simplify our investigation and match the difficulties encountered, we focused our research on California (mostly Democrat) and Texas (mostly Republican).

3.3 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 shape 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* and *countrabound_found*. We added new columns *date_dt*, *relative_arrest* corresponding to number of stops for category A/number of inhabitants of category A, *mean_hit_rate*, *nb_arrest* and *radius*. We also removed some of the data which we couldn't use:

- We dropped the rows for which we didn't have a *county_name*.
- As the data for California ranged from 2009 to 2016 while it ranged from 2006 to 2016 for Texas, we clipped Texas' data such that it matches exactly with California time range.

Data in numbers

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 Illustrating bias

4.1 Ethnicity

Our initial idea was to normalize the number of stops per ethnicity by the total amount of stops in the county. This strategy really didn't gave us the results we expected. Not so surprisingly the number of Black, Hispanic, White and Asian people are extremely unbalanced across the different states of the US. Obtaining 60% of White stops in a county inhabited by 80% of White people gives less information on bias rather than on population. Hence we did not obtain an indicator of racial bias but rather a benchmark of each county ethnicity distribution. To address this problem we set out to find a dataset that would contain the number of inhabitant per county and per ethnicity between 2009 and 2016. Finding such a dataset proved to be an impossible task.

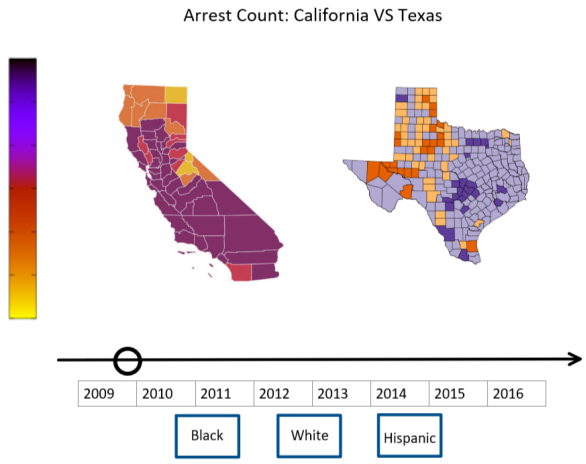

Figure 1: Sketch of number of stops per ethnicity per county

The best solution we found to overcome this major difficulty was using a dataset which contained the proportions of ethnicities per counties in the years 2010 and 2020, **data.census.gov**.

Thus using this dataset we normalized the number of stops for a race in a county by the proportion of people of that race in the county.
To obtain data localized per county we had to focus on the numbers reported by *States patrols*. This type of police exists among other types of patrols (ie:city police departments...), hence the number of stops presented in our analysis represent only a sample of the total amount of police stops that take place in the US, in particular it excludes most cities.

4.2 Gender

We did not have the same issue to illustrate gender bias in police stops as it has been much easier to find a **dataset** with gender proportions per county from kaggle. Further more, this ratio only slightly deviate from the 50% to 50% distribution which makes interpretation of results also much more intuitive.

5 Results

5.1 Hit rates

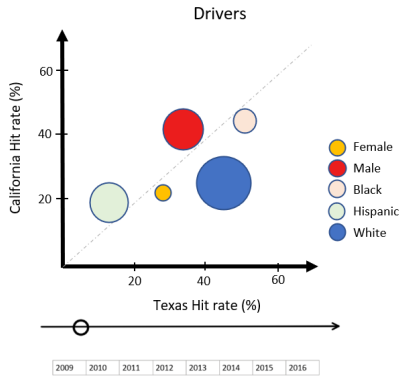

Figure 2: Sketch hit rates per ethnicity

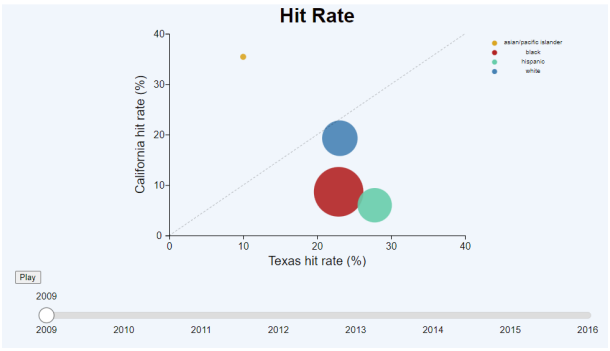

Figure 3: Hit rates final results

The implementation of the hit rate visualization was a total success we were able to show everything that we wanted. The bubble size represent the number of searches per ethnicity, the x-axis and y-axis represent hit rate in Texas and California respectively. The hit rate is defined as the number of searches conducted divided by the number of searches where contraband was found. Thus a lower hit rate suggests that searches are conducted more easily. Hence, a dot centered under the line $y = x$ represents an ethnicity which has a higher hit rate in Texas than in California, meaning that it is more easily searched in California than in Texas. We even added a play button which allows to watch the evolution in time automatically.

5.2 Map View

Figure 4: Clickable map of the US

As part of our extra ideas, we implemented a full map of the US where it is possible to click on two states to display an in depth analysis tool. Again to not focus too much on data processing, we imported only the data to be able analyze California and Texas. However, the processing code and website are already implemented to receive new data. We have made a technical setup on our **github** for anyone who would like to contribute to the project and would like to dig deeper in the analysis.

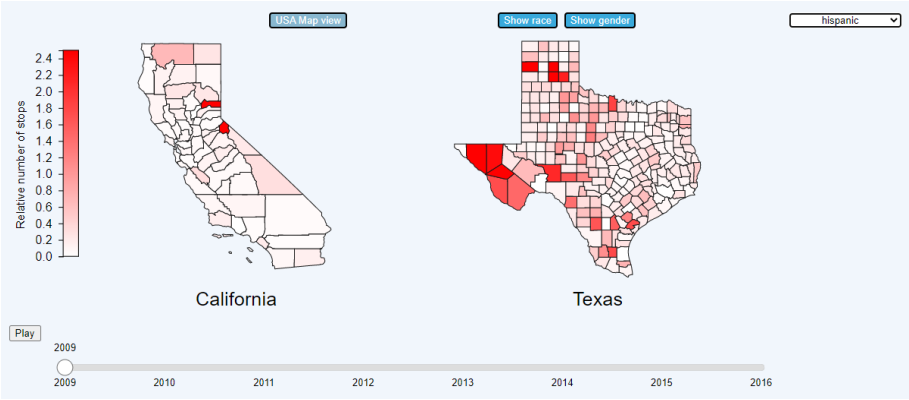

Figure 5: California and Texas comparison map

In this comparison analysis tool it is possible to select whether you want to compare stops based on race or gender and select the ethnicity or gender, respectively, of your choice. Hovering with your mouse will display the county name together with the number of stops for that particular category and the corresponding ratio, **number of stops for category A/number of inhabitants of category A**, per county. When the relative stop ratio is higher for one ethnicity than others, we can consider a bias toward that ethnicity.

Although, we noticed an unwanted pattern in some counties where the relative stop ratio is much higher than 1, due to the number of inhabitants being low compared to the number of tourists or simply travelers in that county. For example, in Sierra and Alpine counties of California we observe very high relative ratios for every category. This is due to the fact that those counties are very famous for their ski stations and mountain area and attract a great number of tourists every year who are naturally not counted in the county population. It is important to note that this doesn't constitute a bias towards any ethnicity since the ratio, although high, is similar for each of them.

Comparing California and Texas, we can see relatively **higher ratios in Texas** overall than in California indicating generally more frequent police stops.

Another observable phenomenon we can observe is the relative ratio of stops near the border with Mexico in South-West Texas is also high for every ethnicity due to control at borders. Indeed relative stops ratios are increasing there for all categories of populations.

For gender however, we observe a constantly **higher relative ratio of stops for males** than females. This remains the case from 2009 to 2016 and is true for both California and Texas.

Finally, when selecting Asian or Black populations, we seem to observe higher ratios for these populations compared to White and Hispanic people. This could represent a bias of police stops towards Asian or Black populations. But, this could also be due to the fact that both of these ethnicities appear in much lesser proportion in the counties they were stopped, provoking a bias in our results due to the arrest of non-inhabitants which inflates the ratio disproportionately.

5.3 Line chart

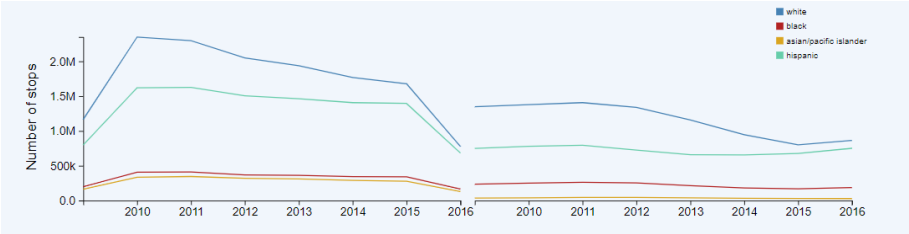

Figure 6: California and Texas line chat comparison

We joined an additional feature to the map visualization which summarizes stops for each population from 2009 to 2016. With the line chart users can get an immediate idea of how stops are distributed among those populations. It is important to note, however, that using the line chart only, can't give any indicator on bias as the number of stops per category is itself biased by the initial proportions of those categories.

6 Peer assessment

Coralie Grobel :
Sliders, website structure, data processing maps, normalizing populations, home page, team page, process book, comparative analysis tool of California and Texas.

Nicolas Delamaide :
USA map, transition to comparative view, data loading and initial pre-processing, hit rate view, line chart, visualizations legends, process book, comparative analysis tool of California and Texas.

Clément Chaffard :
Line chart, button interactions with map, button to go from comparative view to USA map, readme, process book, comparative analysis tool of California and Texas.

7 Resources

7.1 Datasets:

- https://openpolicing.stanford.edu/data
- https://www.data.census.gov
- https://www.kaggle.com/datasets/us_county.csv

Be safe on the road!