Data Report - Exploring Racial Bias in Traffic Policing: An Analysis of Seven Eastern States of US

Mahfuzur Rahman Chowdhury

Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

1 Introduction

With the aim of investigating whether racial bias influences traffic policing decisions in the United States of America; particularly actions such as stopping, searching and arresting individuals. The author began working with data from the Stanford Open Policing Project [1]. However, the 42 datasets which span 42 states, proved too large to compile and analyze comprehensively. To address this problem, the project was refined to focus on seven states in the Eastern United States, narrowing the scope while maintaining the project's core objectives.

The Stanford Open Policing Project is a national initiative of USA that collects and standardizes data on police interactions; including vehicle and pedestrian stops. It contains over 200 million records [1]. For this project, the author focuses on data from seven states: New Hampshire, Rhode Island, Connecticut, Vermont, Massachusetts, Maryland and Virginia; out of the 42 states data on the project.

2 Datasets: Sources, License and Relevance

In this chapter, the author discusses the sources of the datasets, their licensing and the reasons behind selecting these datasets.

2.1 Sources

All seven datasets used in this project are sourced from the Stanford Open Policing Project [1].

Source Website: `https://openpolicing.stanford.edu/`

Here Table 1 contains all the download links of the seven datasets. In Table 1, it is shown that all seven downloaded files are in the '.zip' format. Each zip file contains a single '.csv' file, which holds the dataset.

2.2 License

One of the primary reasons for using datasets from the Stanford Open Policing Project [1] is their availability under the ODC-BY 1.0 license [2]. According to the license documentation, users are permitted to share and adapt the datasets freely. Additionally, the project webpage specifies that it is acceptable to use these datasets, provided proper attribution is given by citing Pierson et al. [1].

Table 1. Download Links of the Datasets with File Type

State	Download Link	File Type
New Hampshire	https://stacks.stanford.edu/file/druid:yg821jf8611/ yg821jf8611_nh_statewide_2020_04_01.csv.zip	.zip
Rhode Island	https://stacks.stanford.edu/file/druid:yg821jf8611/ yg821jf8611_ri_statewide_2020_04_01.csv.zip	.zip
Connecticut	https://stacks.stanford.edu/file/druid:yg821jf8611/ yg821jf8611_ct_hartford_2020_04_01.csv.zip	.zip
Vermont	https://stacks.stanford.edu/file/druid:yg821jf8611/ yg821jf8611_vt_statewide_2020_04_01.csv.zip	.zip
Massachusetts	https://stacks.stanford.edu/file/druid:yg821jf8611/ yg821jf8611_ma_statewide_2020$_0$4_01.*csv.zip*	.zip
Maryland	https://stacks.stanford.edu/file/druid:yg821jf8611/ yg821jf8611_md_statewide_2020$_0$4_01.*csv.zip*	.zip
Virginia	https://stacks.stanford.edu/file/druid:yg821jf8611/ yg821jf8611_va$_s$*tatewide*$_2$020$_0$4_01.*csv.zip*	.zip

2.3 Relevance

These datasets are highly relevant to the research question, as they include attributes such as the driver's race, the reason for the stop and the outcome of the stop. These features are essential for analyzing whether traffic police decision is influenced by racial bias. The basic hypothesis is that if traffic police in these seven states have a racial bias in general then the data analysis will reveal a larger number of stops involving drivers of certain races without justifiable reasons.

2.4 Pipeline

The author used Python 3[1] as the programming language. The datasets were downloaded using the 'requests' library[2], unzipped with the 'zipfile' library[3], and read into a DataFrame using the 'pandas' library[4]. One of the most challenging tasks was concatenating the seven datasets due to different column names. Although the columns represented the same attributes, their names varied significantly. For example; columns such as subject_race, raw_race, raw_RACE_CDE, and raw_driver_race all referred to the driver's race but were named differently across datasets. As a result, manual mapping was required to standardize the column names before concatenating the datasets. The final combined dataset consists of 22 columns. However not all columns are present in every dataset. During the concatenation process, any missing columns in a dataset were filled with the value '000'. Once the datasets were successfully merged, the combined data was saved in .sqlite format using the SQLAlchemy library[5].

[1] `https://www.python.org/`
[2] `https://requests.readthedocs.io/en/latest/`
[3] `https://docs.python.org/3/library/zipfile.html`
[4] `https://pandas.pydata.org/`
[5] `https://www.sqlalchemy.org/`

3 Result and Limitations

One of the difficult challenges for the author is to work with a large dataset. After managing the seven states datasets, the marged the dataset consists of 13,163,977 rows. The resulting .sqlite file, containing a single table is 1.49 GB in size. This makes data manipulation, cleaning and visualization both time-consuming and computationally expensive.

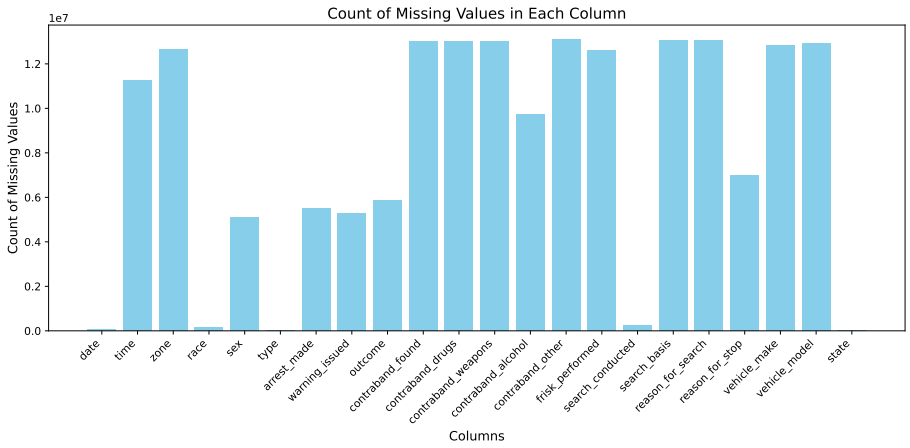

Fig. 1. Column wise Missing Value Count

Figure 1 shows a large number of missing values in the dataset. However, these missing values are primarily found in columns such as contraband_found and contraband_drugs. In most cases, these searches did not occur, which explains the high proportion of missing data in these columns. Additionally, state datasets do not have every column, contributing further increment in missing values.

References

1. Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., Goel, S.: A large-scale analysis of racial disparities in police stops across the United States. Nature Human Behaviour **4**(7), 736–745 (2020). `https://doi.org/10.1038/s41562-020-0858-1`
2. Open Data Commons: Open Data Commons Attribution License (ODC-By) Summary — Open Data Commons: legal tools for open data. `https://opendatacommons.org/licenses/by/summary/`