Aggregated Datset

Cholyeon Cho

1 Definition of Police Violence

The introduced aggregated dataset contains any death happening between the years 2015 to 2021 that is inflicted by an on-duty or off-duty officer where the officer fatally shoots the victim. The dataset is called aggregated because it's an aggregation of the observations inside three fatal police datasets: Fatal Encounters, Mapping Police Violence, Washington Post.

2 Notes on Methodology

The three fatal police datasets have different definitions of police violence. Fatal Encounters includes any death that happened in the presence of on-duty or off-duty officers, Mapping Police Violence includes any death that happened where on-duty or off-duty officers directly inflicted fatal force to the victim, and Washington Post includes any death that happened where an on-duty officer shot the victim to death.

These three datasets were then filtered down. The Fatal Encounters dataset was filtered down on fatal intent, and gunshot. The Mapping Police Violence dataset was filtered down on gunshot.

A locally developed algorithm was used to compare every possible pair of combination of observations among the datasets. These observations were considered to be identical if a combined score based on name, age, gender, race, date, and location exceeded a certain pre-defined threshold. More specifically, for a certain pair of observations between two different datasets, if the criteria were considered to be identical with some level of tolerance for error, then a score was added up for that particular pair. The name criterion was compared using the fuzzy name matching algorithm using Levenshtein distance. Location variables were imputed - if the observation were missing longitude and latitude but had an address - then compared by longitude and latitude. Age and date was simply numerically compared. Age, race, and gender were compared string to string.

All observations that were not caught by at least one dataset was manually reviewed. Any observation that did not fit into the definition of on-duty or off-duty fatal shooting were excluded from the dataset. Each observation were labeled as on-duty or off-duty. Moreover, those observations that did not exactly

match the definition of police violence but were obscure enough not to exclude were labeled as in the grey area.

For each observation, if the observation is included in the Fatal Encounters, the variables in the aggregated dataset are drawn from the corresponding variables from Fatal Encounters. For example, if a victim is named Charles Darnell Baker Jr. in Fatal Encounters and Charles Baker in Mapping Police Violence, the aggregated dataset draws on Fatal Encounters and enlists the name of the victim as Charles Darnell Baker Jr. If the observation is not in Fatal Encounters but in Mapping Police Violence, the variables in the aggregated dataset are drawn from the corresponding variables from Mapping Police Violence. If the observation was only caught by Washington Post, the variables in the aggregated dataset are drawn from the Washington Post.

All variables that were used by the algorithm, variables that were produced during the manual review, and additional helpful variables were included in the aggregated dataset. The additional helpful variables include the unique identifiers of the observations for each of the three datasets. These identifiers can be used to link with the corresponding dataset to pull various other covariates for further analysis. Additionally, the link to the corresponding media report was attached to the aggregated dataset. Originally, there were no reference to the links for the Washington Post dataset. However, these sources were individually found included in the dataset.

3 Cautionary Notes

The fatal police datasets often get updated to include observations from the past. This might incur discrepancies between the aggregated dataset and the fatal police datasets.

Specific variables such as cause of death of the victim, agency of the officer cannot be found in the aggregated dataset. However, these variables can be imputed by using the unique identifiers of each of the fatal datasets, that include such information and more.

Marking for obscurity is highly subjective. The criteria of inclusion for these obscure observations in any of the three datasets were highly subjective as well. In general, obscure observations were allowed into the aggregated dataset when the police shot the victim and the victim died of some obscure reason. However, the level of obscurity might change if these datasets specify the reason of inclusions or exclusions of these obscure observations.

Any miscategorization by human error in the initial filtering phrase - filtering Fatal Encounters and Mapping Police Violence on gunshot and fatal intent - was unaccounted for.

4 Codebook

1. **Name**: Name of the victim in the fatal police shooting.

2. **Age**: Age of the victim
3. **Gender**: Gender of the victim
4. **Race**: Race of the victim
5. **Date**: Date the shooting took place. If not available, the date of the article.
6. **State**: The state the shooting took place .
7. **Longitude, Latitude**: The longitude and latitude where the shooting took place.
8. **Link**: Link to the articles referenced in the fatal shooting datasets. Washington Post did not provide links to the article presented. Those articles inside observations found only in Washington Post are those found manually.
9. **id_fatalenc**: Unique identifiers to observations in the Fatal Encounters dataset.
10. **id_mpv**: Unique identifiers to observations in the Mapping Police Violence dataset.
11. **id_wpost**: Unique identifiers to the Washington Post dataset.
12. **off_duty**: True if the shooting was done by an off-duty officer. Observations caught by all three datasets were assumed to be committed by on-duty officers. Moreover, observations where an officer used gun for domestic violence was labeled as True even if the description does not explicitly mention the officer as off-duty.
13. **grey_area**: True if the observation was obscure, False if the observations perfectly matched the definition of on-duty or off-duty fatal police shootings. An example of an observation that is obscure is when a police shoots at the victim, but the cause of death of the victim is obscure.