Project report

AUTHOR Rui Zhang and Yilan Guo

In this project, we visualize and analyze the police shooting records, hoping to gain some critical insights out of it. We will walk you through every aspects of our project and provide you with a structured introduction on our analysis. Hopefully, you will understand our project better via reading this report.

***Section I: About the Data***

* ***How to get access to the data ?***

The dataset is collected by The Washington Post and it records the fatal shooting by the police since January 1st, 2015. Anyone who are interested in such information can get the data via the link [***https://github.com/washingtonpost/data-police-shootings***](https://github.com/washingtonpost/data-police-shootings)***.***

* ***Why we pick this dataset?***

This dataset contains information that could be valuable. The data is stored in a CVS format and come with 17 attributes as well as 6000 rows. According to The Washington post, the data set is far more comprehensive and complete than what were officially released by the police. The most attributes of our data are categorical, which should be preprocess before any visualization. One can also utilize the Geospatial and time series data to further explore and analyze the pattern.

***Section 2: Visualization techniques***

We make three types of visualization, each of which will be explained separately. After that, several tricks that facilitate the effectiveness of visualization will also be introduced.

* ***Line Chart***

The reason of why we choose line chart is that we want show trends and change or shift of trend. The first interesting question we raised before doing the analysis is that does the total shooting case surge after the outbreak of the pandemic. We decide to answer this question via making a line plot where the horizontal axis represents the date, and the vertical axis shows the total number of fatal shooting. We also plot the total shooting number as time elapsed for different races. The finding of these will be discussed in later section.

* ***Stacked Bar Chart***

This will be our main tool to analyze the dataset since the most attributes of the dataset are categorical. The advantage of such chart is easy to understand and interpret. The readers can be having a better understanding of the composition of our data. However, the disadvantage of such chart is that when two proportions of certain population are close, it is hard for the readers to discern the difference. This is especially cumbersome when we try to show the readers the subtle difference of two populations. But the possible solution to this is to amplify the difference in another bar chart.

* ***Geospatial Analysis***