**Abstract**

The project we worked on was undertaken out of a conviction to further the cause of social justice and equity within our society. There is a distinct lack of applications of data mining and data analysis skills in the domain of social sciences due to the logical disparity within the two disciplines of study. The project’s focus was to draw attention to facts embodied in the datasets from a non-political and scientific perspective. The datasets analyzed were the first section of a 3 part series published by user JohnM on the Kaggle website [1]. The datasets were sourced by the author from multiple places such as the Washington post. While the topics discussed have a huge amount of politics surrounding them, our purpose was to highlight the nature of the problem from a data analysis perspective as opposed to one that just blames a specific section of the political spectrum. Over the course of the project we used various data visualization techniques, performed a FAMD analysis and performed some form natural language processing to generate word clouds from the provided datasets. There was a huge volume of data that was analyzed as a part of the project and gave us some new insights on where the problem is prevalent in its most grievous forms. A notable inference we gained from the dataset was the prevalence of gun related violence that the police and victims of state violence are subjected to. Other important conclusions gained were the discrepancy in the reported number of crimes, people arrested, and people killed in police encounters by each police precinct. The discrepancy in the data available about the victims of state violence when compared to the data collected on policemen, i.e. state agents, is also a notable observation in itself.

**Introduction**

The problem of innocents being killed by state agents in violent encounters has been a problem that has long prevailed in American society. This problem has been around for a long time and routinely results in social disruptions every decade or so. We tried to analyze the data regarding state agents, their death records, records regarding the victims of state violence, records of fatal encounters and records of reported shooting statistics to understand the problem domain from a systemic perspective. Instead of taking a perspective that seeks to blame state officials and people’s explicit and implicit biases, this project seeks to address the root problems that might be enabling and promoting the prevalence of the current problems within the state system. We tried to draw conclusions from the data, impute and remove the missing values from the provided dataset and tried to do a topical analysis on the text provided in the dataset about journalistic articles.

**Technical Details**

Since the work we undertook needed a wider application in comparison to a conventional, deeper approach, we chose to divide the work vertically by delegating datasets to each team member. Most of the time spent on the project was dedicated to data wrangling. Data imputation and dataset sanitation were non-trivial tasks due to the sensitive nature of the data that we were dealing with. Due to how we split the datasets, we shall go over the steps for processing and detail them as processes performed on each dataset.

The dataset contained within the ‘police\_deaths\_538.csv’ file contained the least amount of missing data.

Chart

Description automatically generated

**Figure 1:** A heatmap showing the number of missing values in the dataset. The yellow (lighter) values here would symbolize missing data while the teal (darker) values symbolize the amount of data that is present in the dataset.

# References

|  |  |
| --- | --- |
| [1] | JohnM, "Police Violence & Racial Equity - Part 1 of 3 | Kaggle," 2020. [Online]. Available: https://www.kaggle.com/datasets/jpmiller/police-violence-in-the-us. [Accessed 25 March 2022]. |