

GUN VIOLENCE IN THE US, 2013-2018

DATA ANALYSIS PROJECT



I. DATA SOURCE

Dataset: “ Gun Violence Data, 2013-2018”, from Kaggle.

URL: <https://www.kaggle.com/datasets/jameslko/gun-violence-data>

Cause for Selection: The major cause for choosing this data set is the alarming rise of gun violence in the US which is not only disturbing for people living in the country but also affects country’s image on the global platform.

Data Summary: The data is collected from a non-profit organization named Gun Violence Archive. The organization publishes an online archive of gun violence incidents collected from over 7,500 law enforcement, media, government and commercial sources daily in an effort to provide near-real time data about the results of gun violence. GVA is an independent data collection and research group with no affiliation with any advocacy organization.

Data Sourcing and Collection: The data is primarily sourced from an external data source owned by Gun Violence Archive and then further compiled by third party. Since, the data is not posted directly from the government authorities, one can expect it to be not completely trustworthy. However, since the authors and the organization have publicly published there data collection and publication methodology, one can consider to it produce note-worthy insights.

The data is collected from the period 2013-2018. Thus, it will be good for the historical analysis rather than presenting the present day scenario.

II. DATA PROFILE

Please refer to the Excel file for data consistency check, wrangling steps, and data understanding steps.

III. DATA LIMITATION AND ETHICS

As mentioned above data is not up-to-date, thus will not present current data situation of gun violence in the US. However, it is good for historical analysis and thus help in having a basis understanding about the severity of the issue.

Ethical Concerns:

Collection Bias: There are three stages of data collection in my data set. In spite of all the efforts, there are chances that the data collection process might have a certain level of error at human level or machine level. This could be one of the major reason the dataset contained ample missing data rows.

Cultural Bias: The data lacks demographic information like race and ethnicity, and the factor influence statistics on gun violence. This highlights a sense of cultural bias because the stats show that 38% of White Americans own a gun, compared with smaller shares of Black (24%), Hispanic (20%) and Asian (10%) Americans. Link. In order to remove this cultural bias, one can use another data sets which contains the number of gun owned by particular race and compare the stats.

Privacy concerns: The data set contains personal identifiable information (PII), which raises an issue of data privacy. Publishing PII information like name, age, and addresses of the participants on public platform is risky and thus threatens privacy concern of both the victims and suspects family members.

Measurement Bias: The qualification of subject matter experts are not available in the data set, thus this project will assume a certain level of measurement bias could be present in the final publishing of the dataset.

IV. PROJECT RESEARCH QUESTIONS

- Which state sees the highest number of gun violence?
- What is the growth rate of gun violence in the US from 2013-2018?
- Who are the victims of this gun violence?
- What kind of guns are frequently involved in the violence?
- Is there any gender based difference among the suspects involved in gun violence?
- Is there any relation between the suspect and victim of the gun violence?
- What are the top 5 states and city or county that saw the high rate of gun violence in the US?