**Project report - Gun violence**

Group 3 (Marisa, Anshuman)

## **Executive Summary**

Gun violence is known to affect lives in America. This project aimed to determine which states of the US experience frequent mass shootings over a period of time, and the impact it has on the overall death toll. The death toll of teens and children separately were also examined.

In this project, ETL was performed for gun violence data including mass shooting and casualties’ numbers in children and teens since 2014. The ETL process and findings are detailed in this report.

## **Database source choice**

Originally, gun violence dataset was found on Kaggle (<https://www.kaggle.com/datasets/gunviolencearchive/gun-violence-database?select=accidental_deaths_children.csv>). This contained mass shooting data files up to 2016.

When searching for more updated data, the source <https://www.gunviolencearchive.org/charts-and-maps> was found to contain all the csv files for the years in Kaggle, and more updated data files. Hence, csv files were downloaded from this source instead.

## **ETL**

This process was performed using Python programming language in Jupyter Notebook. Csv files were extracted into panda Dataframes.

1. Mass shooting data across the years were concatenated into one single dataframe.
2. Children casualty dataframe was combined from children killed and children injured csv files. Similar with teen casualty dataframe.
3. Duplicates and Null values were trimmed accordingly.
4. Incident\_Date string was converted to date format using pd.to\_datetime in-built function and successfully loaded into the database tables.



**Figure 1** Original ERD

## **Challenges**

## **During transformation:**

Teen\_killed and teen\_injured datasets both had ‘killed’ and ‘injured’ columns. Hence, a record\_type column was added in the combined dataframe (teen\_casualty) to keep track of which csv files particular rows were collected from.

However, upon close examination of the merged file, those rows that had both killed and injured numbers were identical for both teen\_killed and teen\_injured, and became duplicate rows. Rows that had only killed (and 0 injured) were understandably from the \_killed csv file. Rows that had only injured (and 0 killed) were from the \_injured csv file. This means the data were found to be consistent across the 2 csv files, and hence, there is no need for the record\_type column to be kept.

Similar with the children\_killed and children\_injured dataset. Concatenating and removing duplicates were sufficient to ensure the master dataframe (children\_casualty) contained all the data. There was no need to include the record\_type column.

ERD was updated accordingly (**Figure 2**).

Graphical user interface

Description automatically generated with low confidence

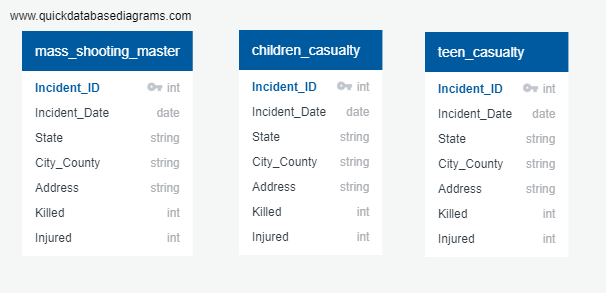
**Figure 2** ERD – version 2

## **During loading:**

Original ERD and ERD version 2 anticipated the relationship between the incident ID from teen\_casualty and children\_casulty and that from mass\_shooting\_master.

However, the schema for foreign keys proved to be erroneous at the Loading step, because some Incident\_IDs from teen/children\_casualty were not found in mass\_shooting\_master. Hence, these relationships between tables were not valid.

ERD was once again modified to remove these relationship (**Figure 3**).



**Figure 3** ERD for gun violence - version 3 (final version)

## **Limitations**

* Scope of data is limited to only from year 2014 onwards.
* The relationships between the tables are not as perfect as originally anticipated.
* All the data files were downloaded from once source for consistency. We did not have external validation as to how accurate and sufficient these datasets are.

## **Research Findings**

## **Key Reports**

1. This report shows the total number of people impacted by mass shooting from year 2014 to 2022 across US.

|  |  |  |
| --- | --- | --- |
| Year | Injured | Killed |
| 2014 | 1086 | 275 |
| 2015 | 1333 | 369 |
| 2016 | 1539 | 453 |
| 2017 | 1807 | 441 |
| 2018 | 1326 | 372 |
| 2019 | 1712 | 465 |
| 2020 | 2535 | 513 |
| 2021 | 2822 | 703 |
| 2022 | 2589 | 652 |

1. Below report shows highest number of children killed from 2012 to 2021. Texas is the state with the highest kill count of children for the past 10 years.

|  |  |  |
| --- | --- | --- |
| Year | State | Kill\_Count |
| 2012 | Connecticut | 28 |
| 2013 | Texas | 9 |
| 2014 | Texas | 28 |
| 2015 | Texas | 32 |
| 2016 | Texas | 46 |
| 2017 | Texas | 61 |
| 2018 | Texas | 31 |
| 2019 | Texas | 35 |
| 2020 | Texas | 49 |
| 2021 | Texas | 41 |
| 2022 | Texas | 65 |

1. Below report shows highest number of **teens** killed from 2021 to 2022. Texas is the state with the highest kill count of children for the past 2 years.

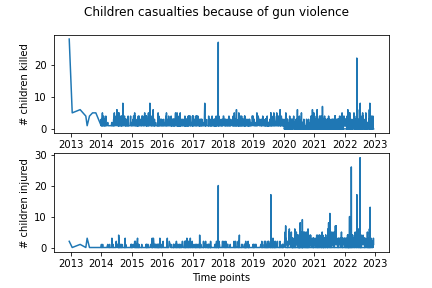
|  |  |  |
| --- | --- | --- |
| Year | State | Kill\_Count |
| 2021 | Texas | 100 |
| 2022 | Texas | 184 |

1. Determine severity of these shooting based on children and teens killed

|  |  |  |
| --- | --- | --- |
| 1 | Total teen killed from 2020-2021 | 2502 |
| 2 | Total teen killed from 2020-2021 | 3563 |
| 3 | Total children killed from 2012-2021 | 3460 |
| 4 | Total children killed from 2012-2021 | 3500 |

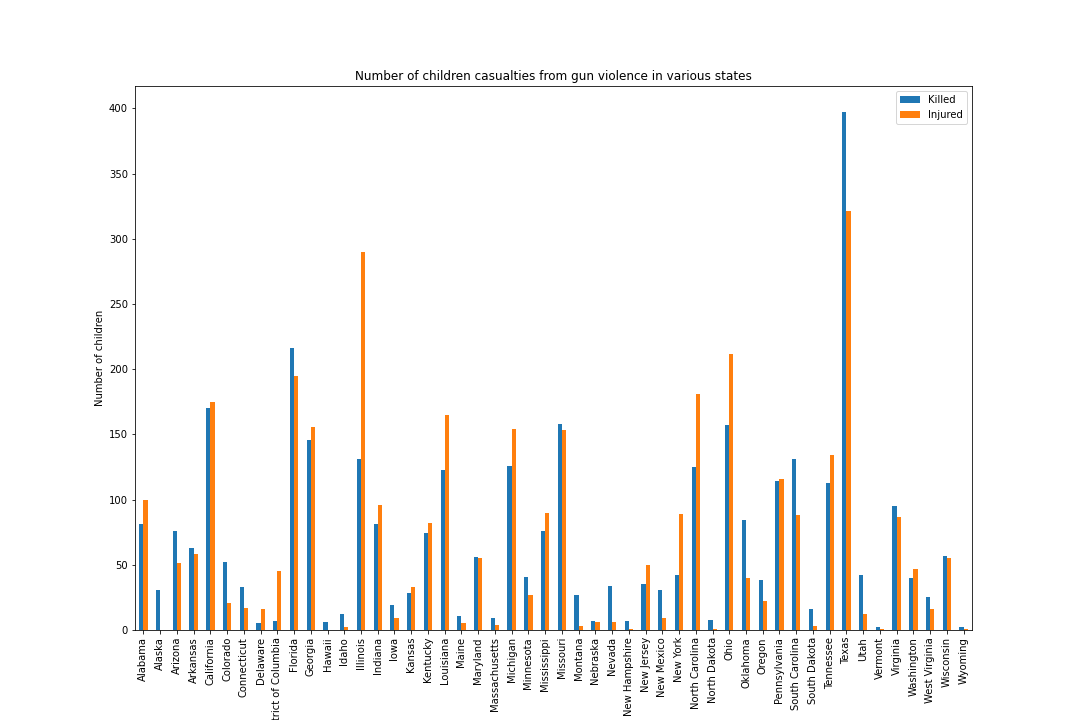
## **Visualization**

* Finding the severity of mass shooting from 2014 to 2019:



Number of children injured due to gun violence seems to be worse in recent years

* Which states have more injured / killed children due to gun violence?

Texas seems to be the worst state in terms of the number of injured/killed children.

Detailed answers to research questions and visualisations can be found in the Jupyter notebook **Gun\_Violence\_ETL\_Main.ipynb** and <https://marduo2022.github.io/Gun-violence>.