ETL Project

**Pre-Processing**

This election year, gun violence remains a bi-partisan issue for debate. I have been hired to analyze gun violence data to address whether gun control or gun education is the more viable solution.

The following table illustrates the observations and actions taken to ensure a clean data set.

| **Pre-process Step** | **Data Need** | **Observation** | **Action** |
| --- | --- | --- | --- |
|  | 1000+ rows of data needed | 501 rows for each set (three sets to be used) | None taken. |
|  | Unique identifier | Incident ID is the unique identifier for two of the data sets. When the third set is incorporated, the unique identifier will need to be determined. | None taken. |
|  | Verified formats of alphanumeric data and availability of all rows | Date, City, State and Address data | Ensured each row contained usable data. Some N/As were found for Address data. I determined that a street address may not be relevant to the study if City and State were available. Some Addresses were also listed as approximations (e.g. 10th block of Veneer Avenue). These approximated Addresses will also be eliminated. |
|  | Values for mass and accidental killings | # killed provided | None. This addresses the original hypothesis. |
|  | Values for mass and accidental killings | # injured provided | Study expanded to include these as well. This additional data was beyond the original scope, but my consensus was to retain it for relevance to the study. |
|  | Investigate | *\*The data set documentation indicates that a mass shooting is categorized as any crime involving 4 or more victims shot. The shooter is not included in this count.* | What categorizes a crime as a mass shooting? |
|  | Investigate | Research will be done to confirm the definition of accidental shooting. | What categorizes a crime as an accidental shooting? |
|  | Investigate | I have the data for the following information, but it didn’t specifically fit into the scope of this project. | Which year seems to have the least shootings? |
|  | Investigate | TBD | What are the top 10 states for mass shootings? Accidental shootings? |
|  | Investigate | I have the data for the following information, but it didn’t specifically fit into the scope of this project. | Do any addresses have repeat offenses? |

**Extraction**

I used 3 different datasets from the public platform Kaggle which lead us to the Gun Violence Archive website. The data in the three files included the following information:

* Accidental death
* Mass Shootings
* Accidental injuries

The fields of interest include the following:

* Incident date
* State
* City/county
* Number killed
* Number injured

\*Including data from 2018/2019 years.

The following sources for our datasets used:

<https://www.kaggle.com/gunviolencearchive/gun-violence-database>

<https://www.gunviolencearchive.org/reports>

<https://www.gunviolencearchive.org/mass-shooting>

<https://www.gunviolencearchive.org/accidental-deaths>

**Transformation**

In order to transform the public data and use it in our study I performed the following:

* Used Pandas functions in Jupyter Notebook to load all three CSV files.
* Reviewed the files and transformed into data frames
* Removed the operator’s column and the address column due to missing information which was not relevant to the focus of this study.
* Identified duplicates by doing an inner merge on the incident id column across all three data sets.
* Created queries to address our hypothesis by grouping the data by state and getting the sum of the number of people killed and the number of people injured. I sorted the data in descending order so I could visually see which state had the highest numbers.

**Accidental Death**

A screenshot of a cell phone

Description automatically generated

**Accidental Death Injury**

A screenshot of a cell phone

Description automatically generated

**Mass Shooting**

A screenshot of a cell phone

Description automatically generated

**Load**

After I pulled in the CSV files and loaded them into the data frames, I did an initial connection to the Postgres database using PG admin to store my original clean data sets. I used the quick database website to create the initial table schema that got loaded into the Postgres database that generated the first set of tables. After running the queries and created the new tables with only the relevant information I reconnected to the database and generated additional tables for the data frames.

**Postgres Database:**

**A screenshot of a computer

Description automatically generated**

**Mass Shootings**

A screenshot of a cell phone

Description automatically generated

**Accidental Injuries**

A screenshot of a cell phone

Description automatically generated

**Accidental Deaths**

A screenshot of a cell phone

Description automatically generated

* *See Jupyter notebook for Gun\_Violence-Final.ipynb for more information*

**Summary**

There were some limitations to my findings due to the data available. However, I was able to address all hypothesis question in my initial project proposal below:

H0: There is no difference in deaths of mass shootings as compared to accidental shootings.

* Accidental deaths
* Mass shootings

I took some additional steps by specifically investigating the finding of mass shootings /accidental shootings by state. In conclusion, I found that deaths resulting from mass shootings outnumbered deaths resulting from accidental shootings. Thus, I recommend the following:

More gun control in the following states while these states have the highest number killed by mass shootings.

1. California
2. Texas
3. Illinois
4. Pennsylvania
5. Ohio

I recommend more gun education/training in the following states Which involved states with the highest number of deaths from accidental shootings.

1. Texas
2. Georgia
3. Alabama
4. Ohio
5. Mississippi