Achievement 6.1

Data Source

U.S. Gun Violence Records 2014-2021 from Kaggle

https://www.kaggle.com/datasets/konivat/us-gun-violence-archive-2014

- The data set I will be using for the following achievement is the U.S. Gun Violence Records from 2014 to 2021.

- The source is an external data source collected by organization, Gun Violence Archive. This organization is an online archive that collects gun violence data in the United States from 7,500 law enforcements, media, government and commercial sources daily. The data is collected through automatic queries and manual research from the sources mentioned previously and then verified by initial researchers and by secondary validation process.

- The data consists of gun shooting incidents in the United States categorized by state, city or county and address with number of individuals killed or injured.

- I chose this data because I am consistently hearing about gun shooting within the country and I wanted to know if there is an increasing number of shooting throughout the years. Also, I wanted to know if there are specific regions with more incidents and if the trend that I am expecting to see actually shows through the data .

**Data wrangling/cleaning and consistency check**

- The data has 3391 rows and 7 columns. After removing missing values, there was 3389 rows and 7 columns. No duplicates were seen.
- Incident date was separated into 3 columns to separate month, day and year. The final result showed 3389 rows and 10 columns

| Variables | Problem | Solution |
|---|---|---|
| Incident Id | Incident id was set as integer | Changed to string because the number is only used for identification purpose |
| Address | 5 missing values (NaN) | Not removed because specific address does not have to be known for analysis |
| Injured | 2 missing values (NaN) and datatype was float | Rows removed due to missing value. Datatype was changed to integer after removing missing values |

| Variables | Time Variant/ Time Invariant | Structured/ unstructured | Qualitative/ Quantitative | qualitative: nominal/ordinal quantitative: discrete/continuous |
|---|---|---|---|---|
| Incident id | Time invariant | structured | Qualitative | Ordinal |
| Incident Date | Time invariant | Structured | Qualitative | Ordinal |
| State | Time invariant | Structured | Qualitative | Nominal |
| City or County | Time invariant | Structured | Qualitative | Nominal |
| Address | Time invariant | Structured | Qualitative | Nominal |
| # Killed | Time variant | Structured | Quantitative | Discrete |
| # Injured | Time variant | Structured | Quantitative | Discrete |

**Data Profile**

**Descriptive analysis**

| Variables | Mean | Minimum | Maximum |
|---|---|---|---|
| # Killed | 1.05 | 0 | 59 |
| # Injured | 4.18 | 0 | 441 |

**Data Limitation and Ethics**

- The data is limited because it does not include any gun shooting incidents that has been missed and unreported. However, the collected data seems to be a good representative because it is collected through thousands of sources and has been double validated. Also, I do not see any ethical issue regarding the data because no demographic information of individuals in the incidents are noted so no personal information is leaked.

**Questions to Explore**

- Is there a specific region with higher number of gun violence incidents?
- Is there an increase in number of incidents throughout the year?
- Is there any correlation between time of the year and number of incidents?
- Is there any trend in number of incidents and region throughout the year?
- Which state has the highest number of cases?
- Are more populated cities, such as Los Angeles, New York, Chicago, have higher number of incidents?