DATA WAREHOUSE

Analysis of school shootings in the U.S.

Giorgio Andronico
227815

Sommario

1 Introduction *2*

1.1 Background *2*

1.2 Record structure 2

1.3 Goals of the project 4

2 Schemas creation *5*

2.1 Entity-relationship diagram *5*

2.2 Attribute tree 6

3 Data Preparation *15*

3.1 Data quality *15*

3.2 PDI Transformations 16

4 Data Visualization *20*

4.1.1 Shootings and casualties by day of week *20*

4.1.2 Shootings and casualties by year *21*

4.1.3 Casualties by state *22*

4.1.4 Casualties by city *23*

4.1.5 Shootings and casualties by shooting type *24*

4.1.6 Shootings by school grades *25*

4.1.7 Shootings and casualties by school type *26*

4.1.8 Shootings and casualties by school ethnicity *27*

4.1.9 Shootings and casualties by presence of resource officer *28*

4.1.10 Shootings by shooter gender and ethnicity *29*

4.1.11 Shooter age *30*

4.2 Dashboards 31

4.2.1 Dashboard: School shootings by Time and Type *31*

4.2.2 Dashboard: Shootings by Location and Shooters *32*

4.2.3 Dashboard: Shootings by School Features *33*

1 Introduction

1.1 Background

This dataset was found in one of the main sources of dataset to analyze, following the link provided in the course. This dataset was collected over time by *The Washington Post* and details every shooting that took place in both private and public schools since the Columbine massacre up until today. Specifically, the data outlines how many people have been killed or injured, what weapon was used for the shooting, the age and race of the shooter, and other important quantities worth analyzing. It also looks at where the shootings have taken place, and is regularly updated; indeed, the latest shooting recorded is the one that took place in *Uvalde, Texas* in May of 2022.

1.2 Record structure

The *school-shootings-data.csv* collects 339 shooting events identified by the unique field **uid**, which is a surrogate key. There are several other attributes, 49 to be exact, which are summarized below. Only the most important informations are reported below; the full schema is available here.

Field Name	Description
uid	Unique identifier
nces_school_id	National Center for Education Statistics unique school ID
school_name	Name of school
nces_district_id	National Center for Education Statistics unique district ID
district_name	Name of school district
date	Date of shooting
school_year	School year of shooting
year	Year of shooting
time	Approximate time of shooting
day_of_week	Day of week of shooting
city	City where school is located
state	State where school is located
school_type	Type of school (public or private)
enrollment	Enrollment at school at time of shooting
killed	Number killed in shooting (excludes shooter)

injured	Number injured in shooting (excludes shooter)
casualties	Number killed and injured in shooting (excludes shooter)
shooting_type	Type of shooting
age_shooter1	Age of first shooter
gender_shooter1	Gender of first shooter
race_ethnicity_shooter1	Race or ethnicity of first shooter
shooter_relationship1	First shooter's relationship to school
shooter_deceased1	Flag indicating whether first shooter died in shooting
deceased_notes1	If first shooter deceased, how first shooter died
age_shooter2	Age of second shooter
gender_shooter2	Gender of second shooter
race_ethnicity_shooter2	Race or ethnicity of second shooter
shooter_relationship2	Second shooter's relationship to school
shooter_deceased2	Flag indicating whether second shooter died in shooting
deceased_notes2	If second shooter deceased, how first shooter died
white	Enrollment of white students at time of shooting
black	Enrollment of black students at time of shooting
hispanic	Enrollment of Hispanic students at time of shooting
asian	Enrollment of Asian students at time of shooting
american_indian_alaska_native	Enrollment of American Indian and Alaskan native students at time of shooting
hawaiian_native_pacific_islander	Enrollment of Hawaiian native and Pacific islander students at time of shooting (unavailable prior to 2009)
two_or_more	Enrollment of students of two or more races at time of shooting (unavailable prior to 2009)
resource_officer	Flag indicating presence of school resource officer or security guard on school grounds at time of shooting
weapon	Weapon(s) used in shooting
weapon_source	Where shooter acquired weapon(s) used in shooting
lat	Latitude of school
long	Longitude of school
staffing	Full-time equivalent teachers at school at time of shooting
low_grade	Lowest grade-level offered by school
high_grade	Highest grade-level offered at time of shooting
lunch	Number of students at school eligible to receive a free or reduced-price lunch
county	County name where school is located
state_fips	Two-digit state Federal Information Processing Standards code
county_fips	Five-digit county Federal Information Processing Standards code
ulocale	National Center for Education Statistics urban-centric locale code

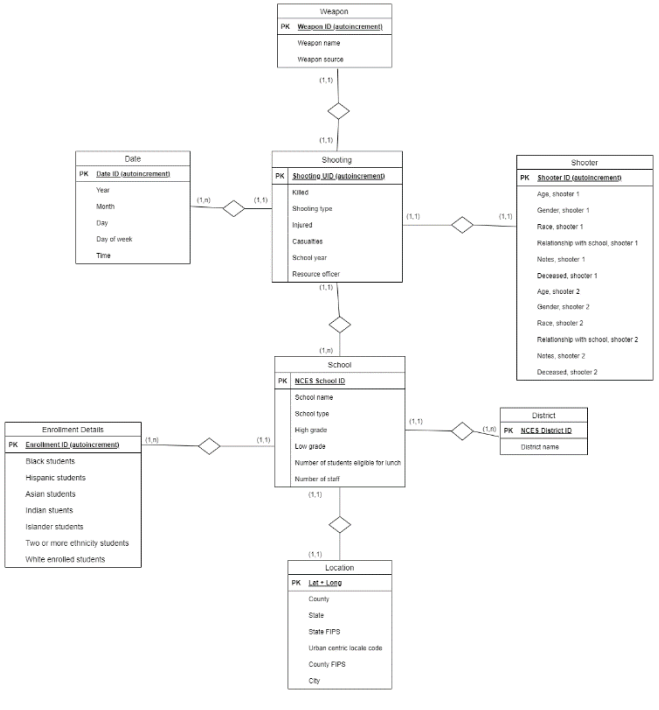
1.3 Goals of the project

The main objective of this brief study is to create a small data warehouse, with only one source, which is the comma-separated value file indicated here. Once an **Operational data store (ODS)** has been created, an entity-relationship schema will be outlined, and from this an attribute tree will be built. From here we will outline the main dimensions and the fact to pick. After building this tree, a set of analyses will be picked for later, and based on this, the tree will be pruned. Finally, a fact schema will be built, along with a snowflake schema; this schema will be created on a **PostgreSQL** instance, and data will be cleaned in the process via **Pentaho Data Integration**. The last step will be to plot graphical analyses using **Tableau**.

2 Schemas creation

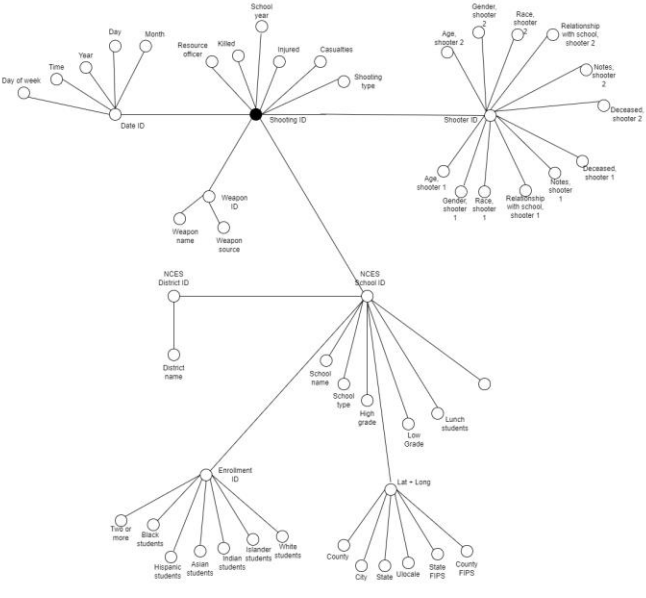
2.1 Entity-relationship diagram

This was constructed looking at the original file and creating surrogate autoincrement keys where necessary.



2.2 Attribute tree

This was constructed using the usual recursive method starting from the entity relationship schema.



2.3 Choosing analyses to perform

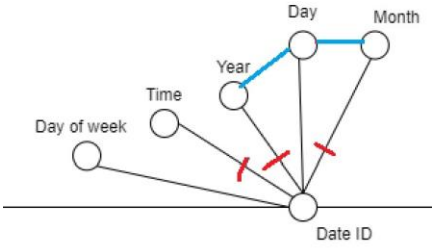
Here some analysis to be performed are chosen. This step needs to be taken now as it is needed to prune the attribute tree. The analyses chosen are:

- Average age of shooters
- Frequent gender and race ethnicity of shooters
- Geographic areas where there were more school shootings
- Geographic areas where there were more casualties
- Number of casualties by years
- Number of casualties by day of week
- Number of shootings by school type, public or private
- Number of shootings by school grade
- Number of shootings by shooting type
- Number of shootings based on the mixed ethnicity of schools
- Possible relation between a resource officer on-site and the casualties

2.4 Attribute tree pruning

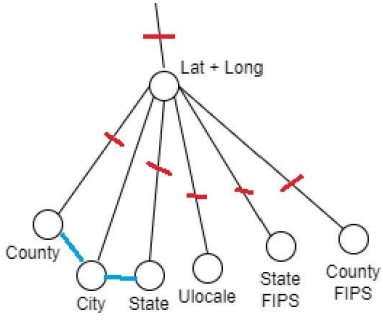
Based on the chosen analyses, many attributes can be done away with, and many more can be merged via hierarchies.

2.4.1 Date pruning and hierarchy creation



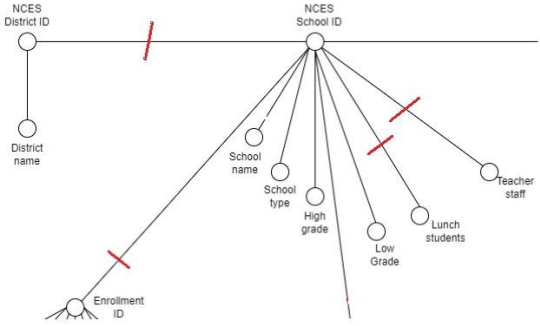
Here, the **Date** dimension was pruned from the attributes **Time**, as it's not relevant for what we want to do. **Year, Day** and **Month** can be seen as a hierarchy, s.t. Day is followed by Month and the latter is followed by Year.

2.4.2 Location pruning and hierarchy generation



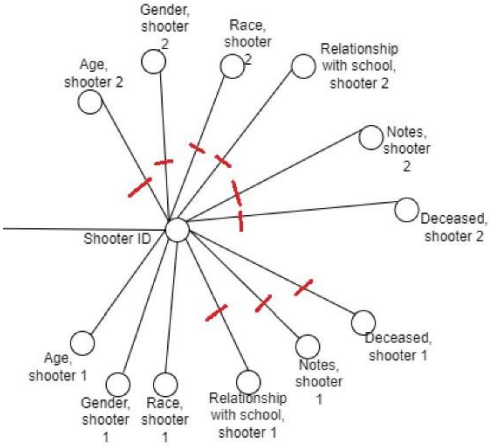
Regarding the school location, the interest is focused on **City, County** and **State**, as it's easy to see that it forms a hierarchy. **Latitude** and **Longitude** fields are calculable by plotting software, and the *Federal Information Processing Standards* (FIPS) code are not really intuitive to show on a world map, as they are just like zip codes: not really needed. Same goes for the *National Center for Education Statistics urban-centric locale code* (**ulocale**), which serves no purpose.

2.4.3 School pruning



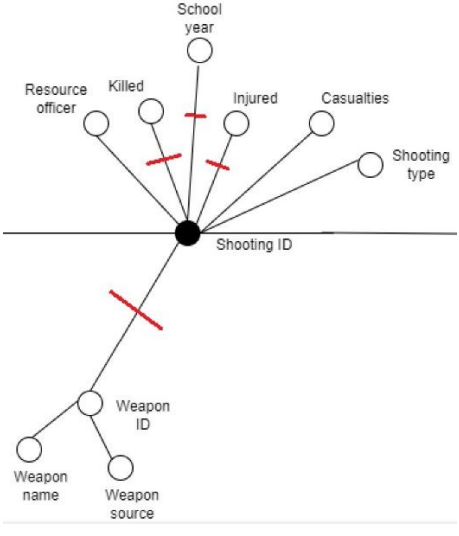
The *National Center for Education Statistics* District ID serves no purpose for us, as we do only per-school analyses, not per-district. However, the **District Name** attribute is kept, as later in the Data Preparation phase, we will use it together with **School Name** as a key for the School entity. The number of teachers (**Teacher Staff**) is not of relevance, as we're not taking into consideration the size of the school (this could be a good way to measure it).
Same goes for the number of students eligible for free or reduced-price lunch (**Lunch students**): in the analyses, victims are victims, without differentiation.
All details about **Enrollment** are done away with and substituted with a **calculated attribute**: **Ethnicity.** Essentially, if one single ethnicity (such as white) takes up 80% of the total enrolled students, that school is labeled as having *homogeneous* ethnicity, *heterogeneous* otherwise.

2.4.4 Shooting pruning



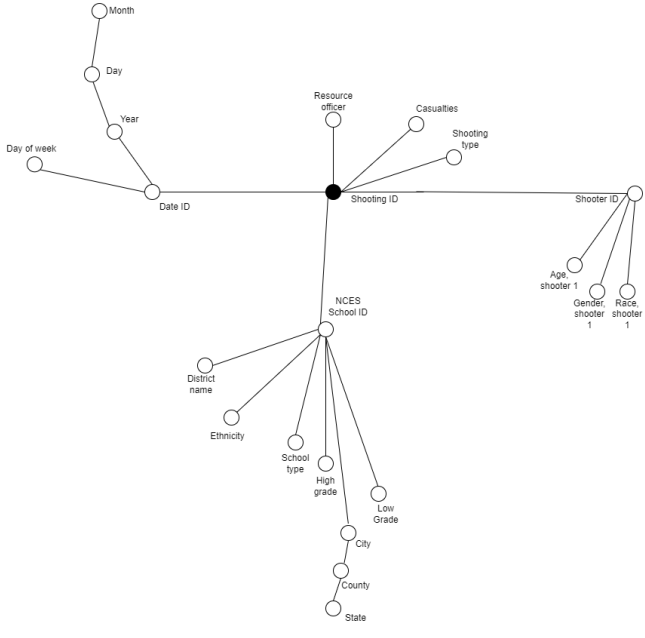
The vast majority of shootings detailed in the data does not contain information about the second shooter: in Data Quality terms, most of these columns are *null*. Thus, we prune these attributes, simply because most of the times, there is only one shooter on site. The analyses do not take into account whether the shooter dies or not, thus the **Deceased** attribute is pruned. Same goes for the **Relationship with school** and **Notes** column, also because the data relating to those columns is very sparse and would make for inaccurate analyses.

2.4.5 Shooting pruning

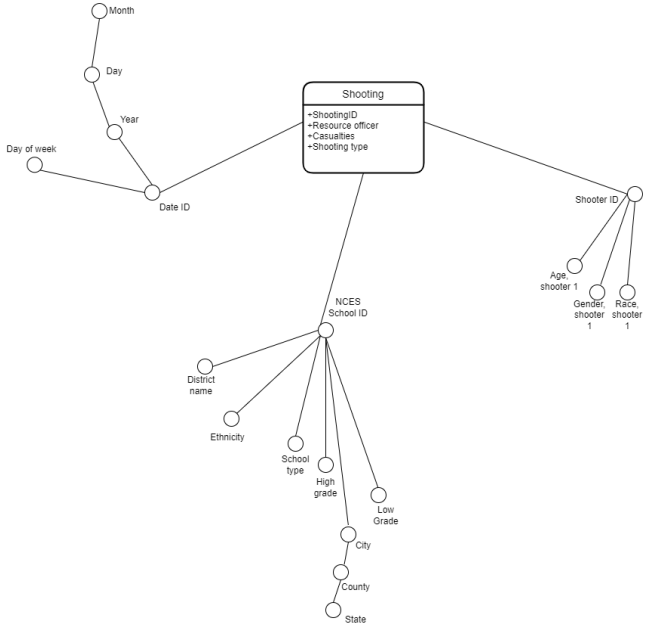


Finally, regarding what is considered as **Fact** (that is, the **Shooting**), non-relevant attributes are removed. Weapon is not interesting for the scope of the project, and also, it's not really a measurable thing. We're not interested in both **killed** and **injured** counts, just their sums: **casualties**.

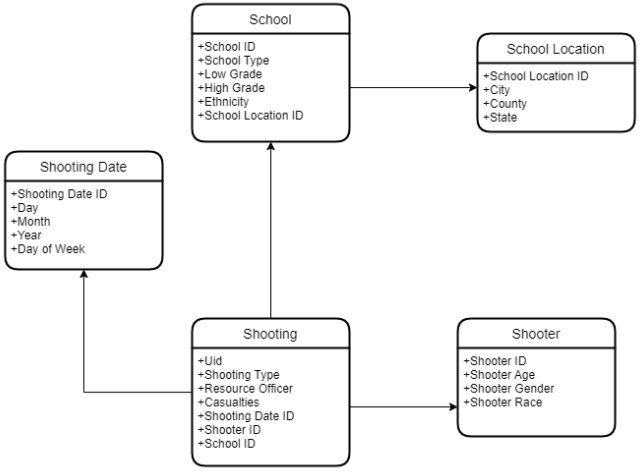
2.4.6 Complete attribute tree



Same tree but with fact nested in:



2.5: Snowflake schema



3 Data Preparation

3.1 Data quality

The dataset is not particularly dirty, most likely because it is quite small, just a few hundred of records. Also, *The Washington Post* takes great care of how it reports information as outlined in the original article. However, a few small problems can be outlined:

- More often than not, gender and ethnicity information are missing from the shooter details
- The *record_layout.csv* details how there can only be two genders detailed: male and female. However, a handful of records report a 'b' or 'h' gender: this is assumed to be noise, dirty data.
- Same goes for the *shooter_ethnicity* column: only valid data are reported to be 'a' (for Asian), 'b' (for black), 'ai' (for American Indian), 'h' (for Hispanic) and 'w' (for white). However, values of 'n' and 'm' are present for a handful of records.
- At times, whitespace may be present in a record. For example, ' m' instead of 'm' for the *shooter_ethnicity* column.

Finally, most columns regarding the second shooter are null. But this is not a matter of data quality, simply most shootings involved just one shooter. Either way, the points noted above are taken care during the operations performed on PDI.

3.2 PDI Transformations

Below all the transformations created in PDI are detailed: table creation, null value cleaning, feature engineering, etc.
Only the most relevant steps are explained.

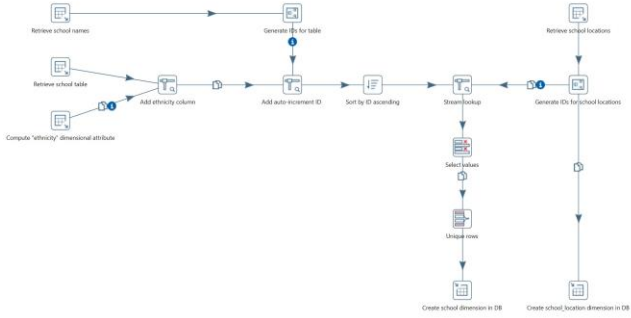
3.2.1 Main Job

- **Create ODS:** creates the Operational Data Store.
- **Creation of "School" dimensions**: creates the School and School Location tables.
- **Creation of "Date" dimension**: creates the Shooting Date table.
- **Creation of "Shooting" fact**: creates the Shooting table and sets foreign key values.

3.2.2 Create ODS

- **CSV file input:** reads the *school_shootings_data.csv* file.
- **Useless attribute removal**: removes the attributes pruned from the tree.
- **Create ODS**: creates the *school_shooting* table in the *ods* database.

3.2.3 Create school dimensions

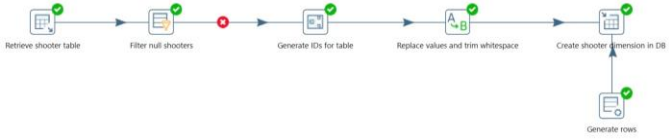


- **Retrieve school names:** select all distinct pairs of *school_name* and *district_name*. This is needed to generate the auto-increment ID field for the *school* table *in* **Add auto-increment ID**.
- **Retrieve school locations:** select all distinct pairs of *city* and *county* and *state*. This is needed to generate the auto-increment ID field for the *school_location* table in **Generate IDs for school locations**.
- **Retrieve school table**: retrieves all the fields related to the school dimension.
- **Compute "ethnicity" dimensional attribute**: computes the attribute as specified above.
- **Stream lookup:** enters the correct value for the *school_location_id* foreign key based on the *<city, county,state>* tuple.
- **Select values**: removes attributes that are not part of the school dimension itself, but have been used to computer other attributes. For example, all of the enrollment numbers have been used to compute *ethnicity*, while *school_name* and *district_name* have been used to create the *school_id* field.
- **Unique rows**: removes duplicate schools (as the PK constraint on *school_id* would not be respected).
- Creation of dimension tables in *reconciled* database.

3.2.4 Create date dimension

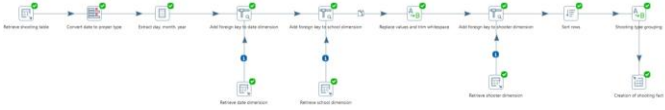
- **Retrieve school table**: retrieves all distinct *<event_date, day_of_week>* couples.
  Convert date to proper type: parses *event_date* as an actual date data type. This is needed by **Extract day, month, year** to compute fields from the *date* field.
- **Remove unneeded column:** deletes *event_date*.
- **Creation** of dimension table in *reconciled* database.

3.2.5 Create shooter dimension



- **Retrieve school table**: retrieves all distinct *<age, race, gender>* tuples from the ODS.
  Filter null shooters: filters rows where all three fields are null. This is why **Generate rows** is needed: it creates a row where *shooter_id* equals -1, indicating an unknown shooter.
- **Replace values and trim whitespace:** trims whitespace where needed, and translates names like ‘a’ to ‘Asian’ or ‘w’ to ‘White’, just to have “friendlier” labels to give Tableau.
- **Creation** of dimension table in *reconciled* database.

3.2.6 Create shooting fact

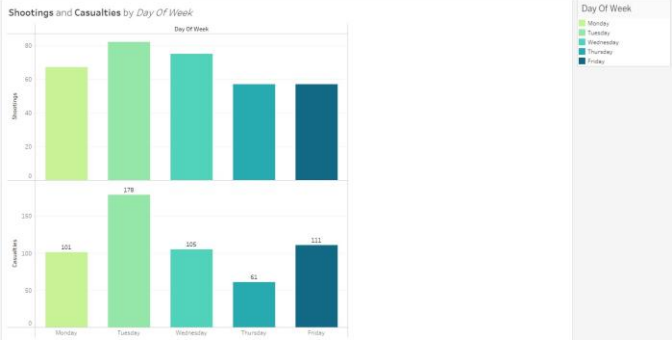


- **Retrieve school table**: retrieves all relevant columns from the ODS needed to create the fact.
- **Add foreign key to date dimension**: fills FK *shooting_date_id* with the correct ID, using *<day_of_week, month, day, year>* to lookup.
- **Add foreign key to school dimension**: fills FK *school_id* with the correct ID, using *<school_name, district_name>* to lookup.
- **Add foreign key to shooter dimension**: fills FK *shooter_id* with the correct ID, using *<age, gender, ethnicity>* to lookup.
    - **Note**: when all three fields are null, *shooter_id* is assigned to -1.
- **Shooting type grouping**: group similar classes of shooting, such as 'suicide' and 'attempted suicide'.
- **Creation of shooting fact**: creates the table in the *reconciled* database.

4 Data Visualization

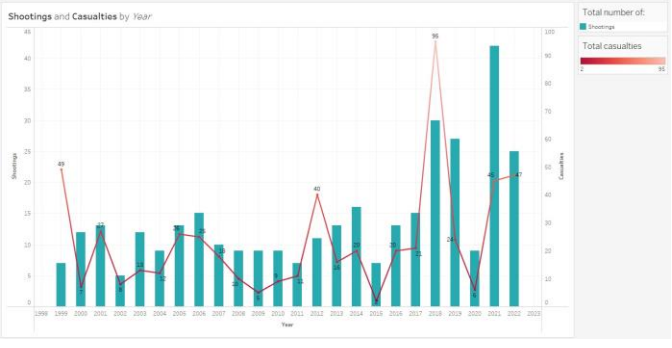
This final step concerns dashboards and plots created in **Tableau**, a data visualization software. It is configured to read directly from the *reconciled* PostgreSQL database.

4.1.1 Shootings and casualties by day of week



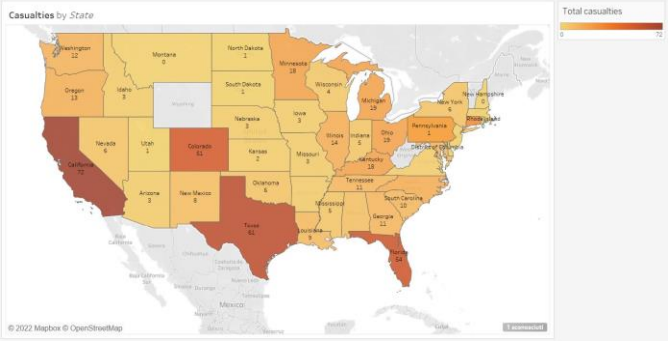
From this visualization, no real trend emerges. Shootings seem to happen a bit more during Tuesdays and Wednesdays, i.e., during the middle of the week. However, the margin w.r.t. Monday etc., is quite small: about a 15 shooting difference. Tuesday, however, seems to be the day where most casualties were counted.

4.1.2 Shootings and casualties by year



Shootings seem to be on an upward trend, with the highest value being counted in 2021 and the fourth lowest in 2020. This may be related to the COVID-19 pandemic, as schools were closed for most of 2020. Unsurprisingly, the year with the most shootings also saw the most casualties.

4.1.3 Casualties by state

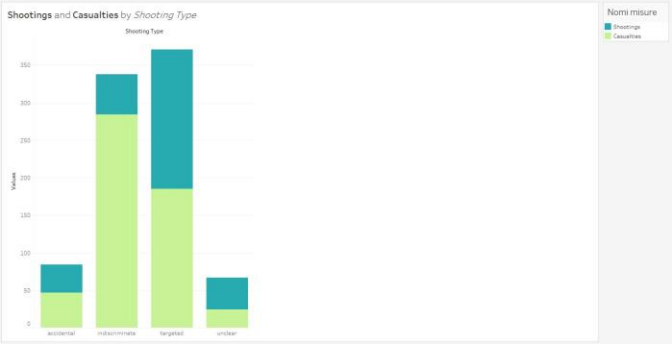


Visualized above are the states where casualties of shootings lay. Northwestern states seem to register lower rates, while southern states get higher rates.

4.1.4 Casualties by city

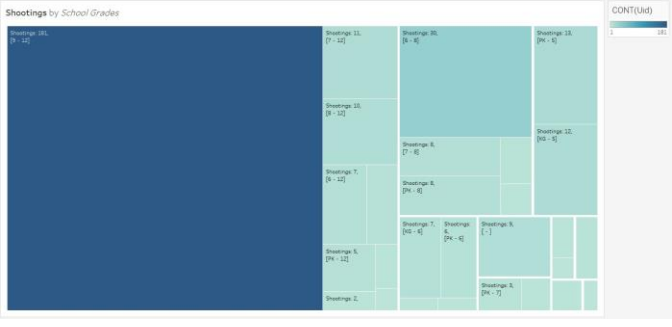
The more a dot is red and big, the more shootings took place there. It's possible to see that eastern states register a larger number of shootings, while critical points westward include Los Angeles. Note that this visualization does not take in account casualty count.

4.1.5 Shootings and casualties by shooting type



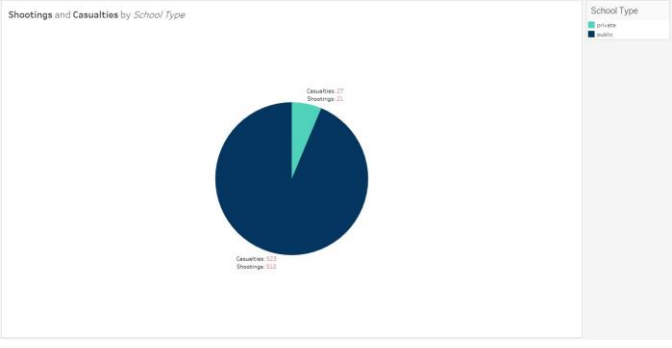
Most shootings appear to be either indiscriminate or targeted. Also, when the shooting is indiscriminate, the number of casualties appears to be higher. A small number of shootings is accidental: indeed, there is a case of accidental shooting where the age of the shooter is just six years old, to name an example.

4.1.6 Shootings by school grades



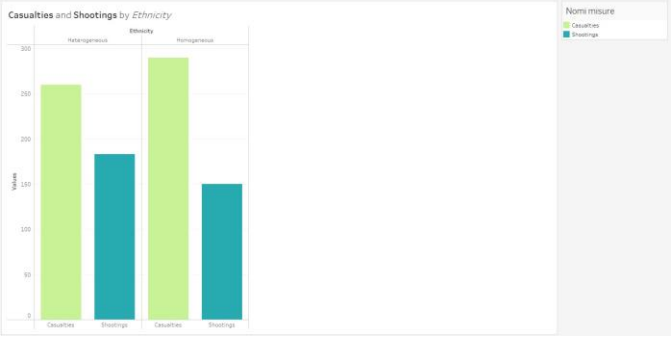
The vast majority of shootings appear in schools where the lowest grade is 9 and highest is 12: these are high schools.

4.1.7 Shootings and casualties by school type



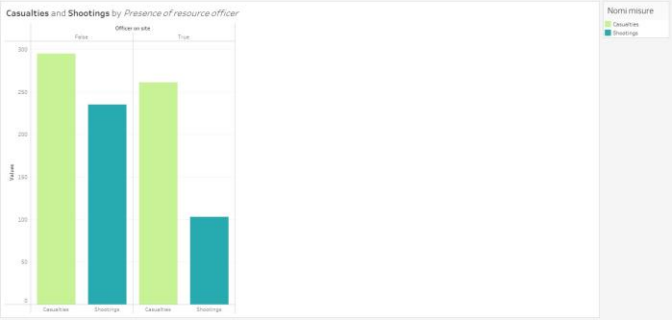
Most shootings take place in public schools. This is simply due to the fact that public schools in the U.S. vastly outnumber private schools.

4.1.8 Shootings and casualties by school ethnicity



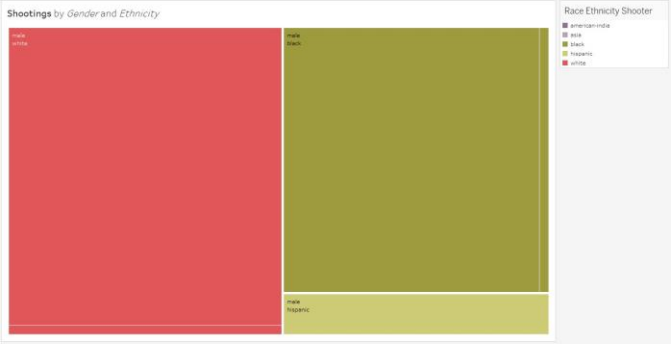
School ethnicity seems to have no effect on shootings: the distribution looks thes same both on the left and the right.

4.1.9 Shootings and casualties by presence of resource officer



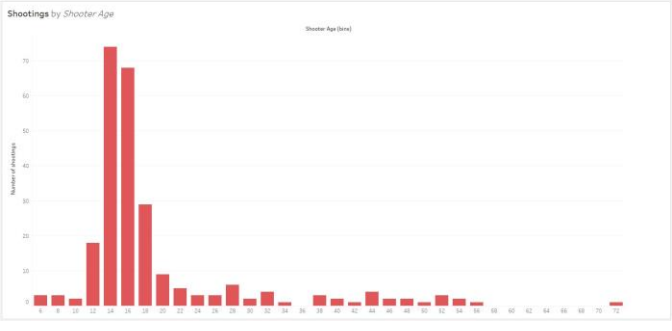
The presence of a police officer greatly decreases the chance of a shooting happening, most likely due to the shooter knowing that they're going to encounter at least some armed resistance.

4.1.10 Shootings by shooter gender and ethnicity



White males do the majority of the shootings, followed by black and Hispanic males. There are a handful of instances of female shooters, accounting for a total of six.

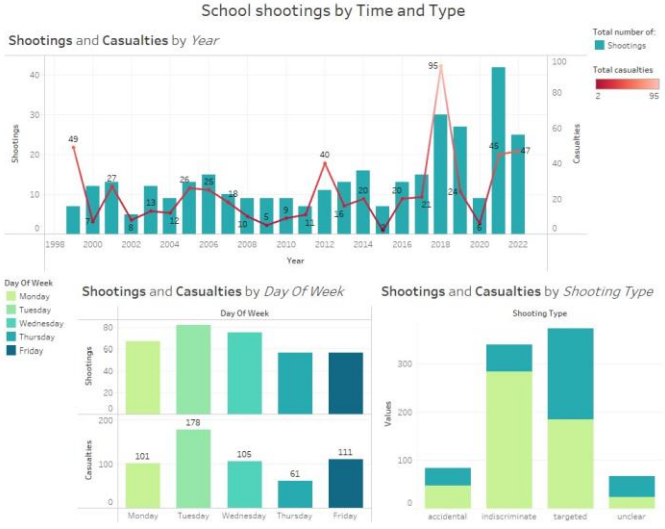
4.1.11 Shooter age



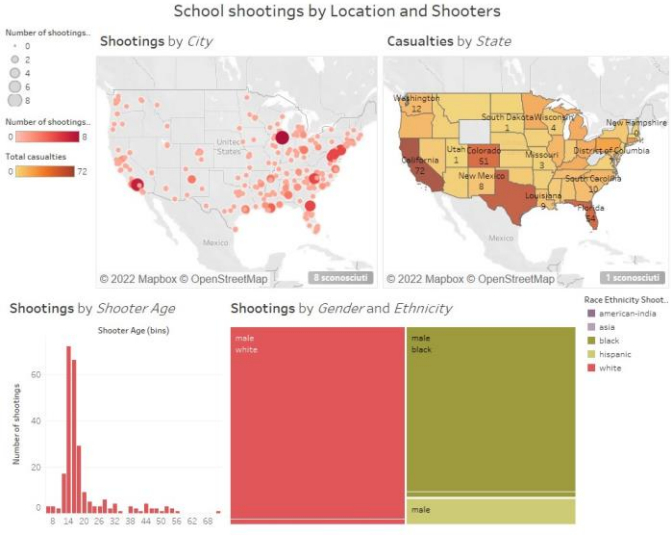
Most of the shootings are committed by shooters aged from 14 to 18: this is due to the fact that these shootings are committed by students of the schools, which by nature rarely exceed 18 in age.

4.2 Dashboards

4.2.1 Dashboard: School shootings by Time and Type



4.2.2 Dashboard: Shootings by Location and Shooters



4.2.3 Dashboard: Shootings by School Features