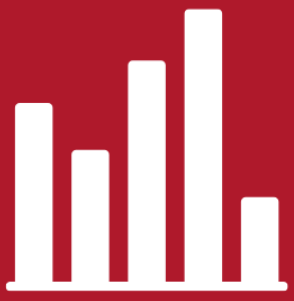
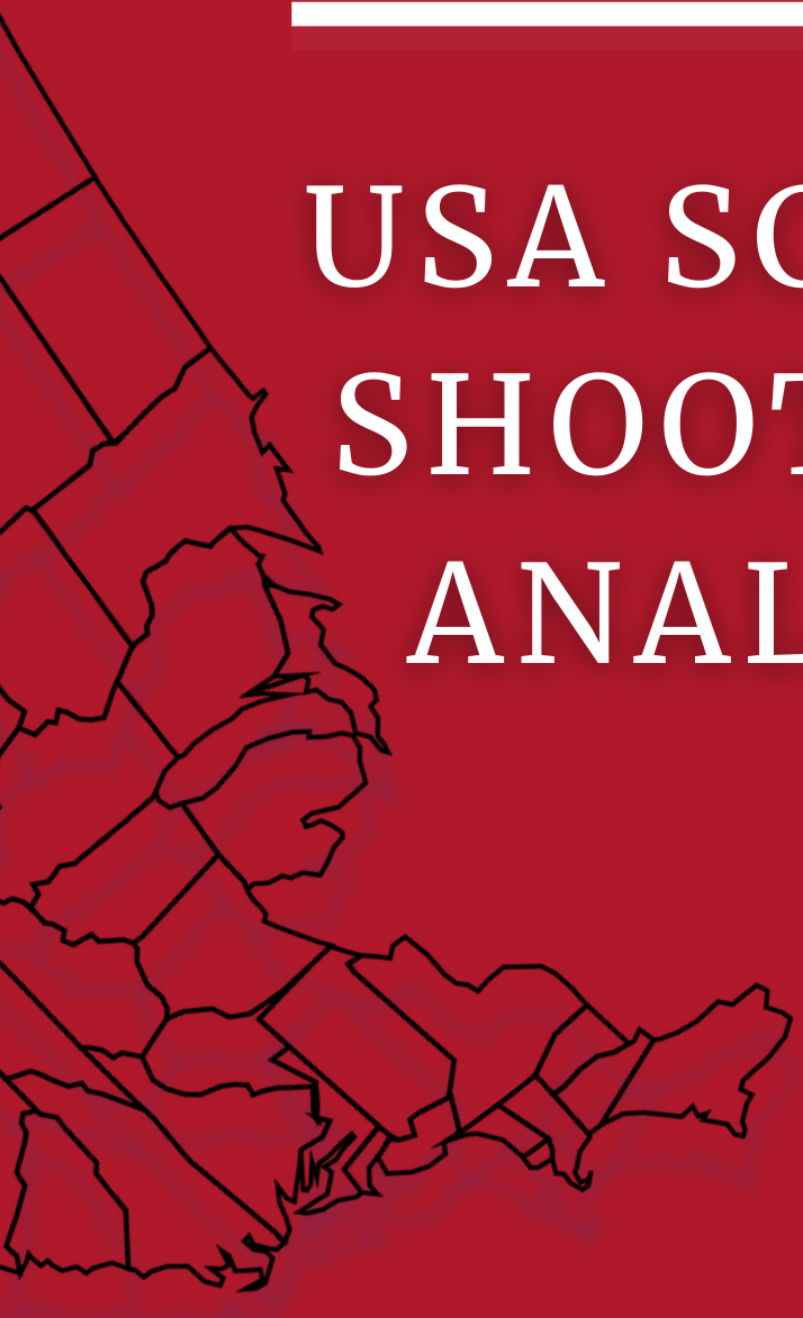
UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI **MATEMATICA E INFORMATICA**

Data Warehouse and Visualization



USA SCHOOL SHOOTINGS ANALYSIS



DANIELE AVOLIO
242423

Contents

Chapter 1

Introduction

This report presents the results of a data warehouse project focused on school shootings in the United States. The project aims to gather, clean, and organize data from a variety of sources in order to better understand the characteristics and dynamics of these incidents. The data sources used in this project include data from 3 different CSV files containing detailed information about individual school shooting incidents, shooters, and victims. The report begins by describing the process of selecting and cleaning the data, followed by a discussion of the design of the data warehouse and the various **ETL** processes used to populate it. The report then presents the results of the data analysis, including 1 or 2 analysis sheets and the dashboard. In the project, the analysis is told using *Tableau stories*. Finally, the report concludes with a summary of the main findings and a discussion of any challenges or limitations encountered during the project.

1.1 The sources

The data warehouse project analyzed school shooting incidents in the United States using data from three CSV files: INCIDENT, SHOOTER, and VICTIM. The INCIDENT file contains information about the shooting incident, including the location and timing of the event. The SHOOTER file includes details about the shooter(s) involved in the incident, and the VICTIM file includes information about the victim(s).

The data for this project was obtained from the K-12 School Shooting Database (K-12 SSDB), which is a comprehensive database of incidents involving guns on school property. The data was downloaded from the K-12 SSDB.

incidentid	String
age	Integer
gender	String
race	String
schoolaffiliation	String
shooteroutcome	String
shooterdied	Boolean
injury	Integer
chargesfiled	Boolean
verdict	String
minorchargedadult	Boolean
criminalhistory	String

Shooter

incidentid	String
race	String
injury	Integer
gender	String
schoolaffiliation	String
age	Integer

Victim

Column name	Data type
Incident_ID	String
Sources	String
Number_News	Integer
Media_Attention	String
Reliability	String
Date	Date
Quarter	String
School	String
City	String
State	String
School_Level	String
Location	String
Location_Type	String
During_School	Boolean
Time_Period	String
First_Shot	Time
Summary	String
Narrative	String
Situation	String
Targets	String
Accomplice	Boolean
Hostages	Boolean
Barricade	Boolean
Officer_Involved	Boolean
Bullied	Boolean
Domestic_Violence	Boolean
Gang_Related	Boolean
Preplanned	Boolean
Shots_Fired	Integer
Active_Shooter_FBI	Boolean

Incident

Figure 1.1: Sources tabeles

1.2 Goals

The goal of this project is to analyze school shootings in the USA in order to understand the trends, patterns, and biases that may be present in these incidents. Specifically, the project aims to analyze the following:

- The times at which shootings occur: This includes analyzing the day of the week, time of day, and season when shootings occur.
- The locations of shootings: This includes analyzing the geographical distribution of shootings, as well as the types of locations where they occur (e.g., schools, colleges, universities).
- The trends over time: This includes analyzing the number of shootings that have occurred each year and identifying any trends or patterns that may be present.
- The gender and race biases: This includes analyzing whether certain genders or races are more likely to be involved in shootings as shooters or victims.
- The connection between shootings and bullying, gang activity, or domestic violence: This includes analyzing whether these factors are present in a significant number of shooting incidents.
- The places where shootings are most likely to occur: This includes identifying the types of locations that are most at risk for shootings and developing strategies to prevent or mitigate these incidents.

Is very important to note that not all of those columns will be used inside the final fact schema. Let's see in the next chapter the **attribute tree**.

Chapter 2

Schemas

2.1 Attribute tree

Starting from the columns I had in the *csv files*, I created an **attribute tree** having as a **root node** a combination of *[Incident, Shooter and Victim] ID*. The main motivation for this is that **we actually have multiple arcs** and using this method we will have only 1 to-1 association between the root node and the dimensions. The first version of the tree is the next:

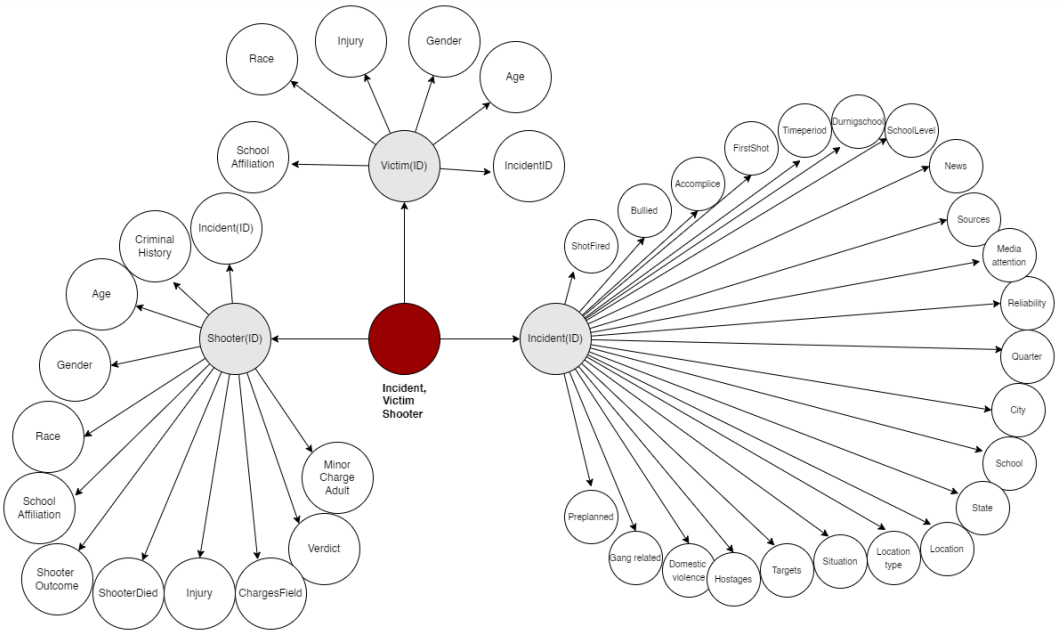


Figure 2.1: Attribute tree

This attribute tree contains a lot of *useless information*, like:

- Sources
- News
- Reliability
- Ecc...

2.2 Cuts on the tree

The next step I performed was to actually remove useless attributes from the dimensions, that in this care are:

- Shooter
- Victim
- Incident

Actually, I created another dimension, that is **Date**, which I took from the **Incident** dimension. Moreover, I created some hierarchies in the dimensions to be able in the future to drill down and roll up on those.

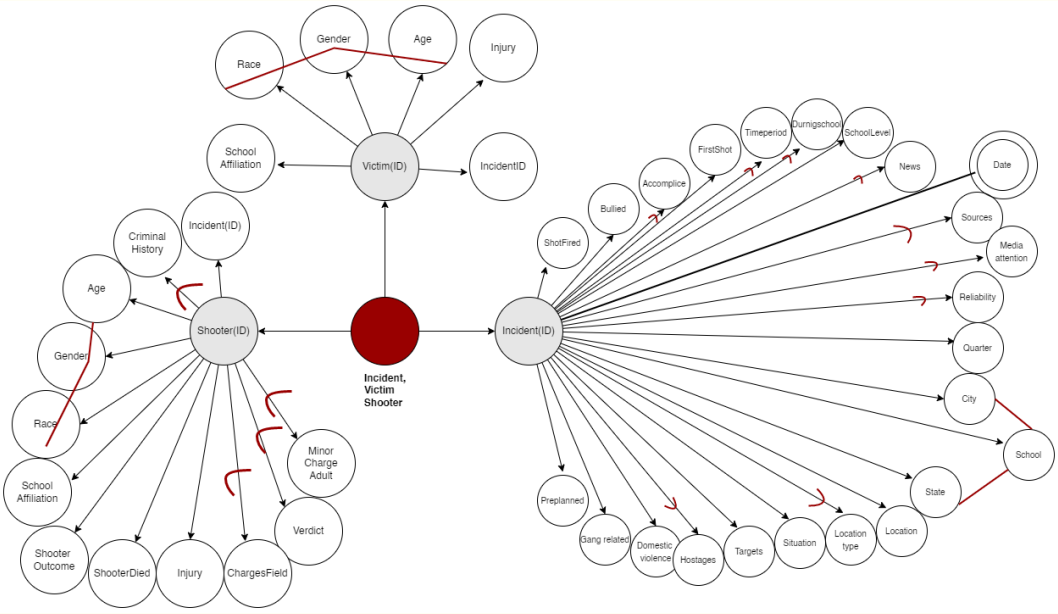


Figure 2.2: Grafting and Pruning

2.3 Edited tree

Finally, the edited tree is the following:

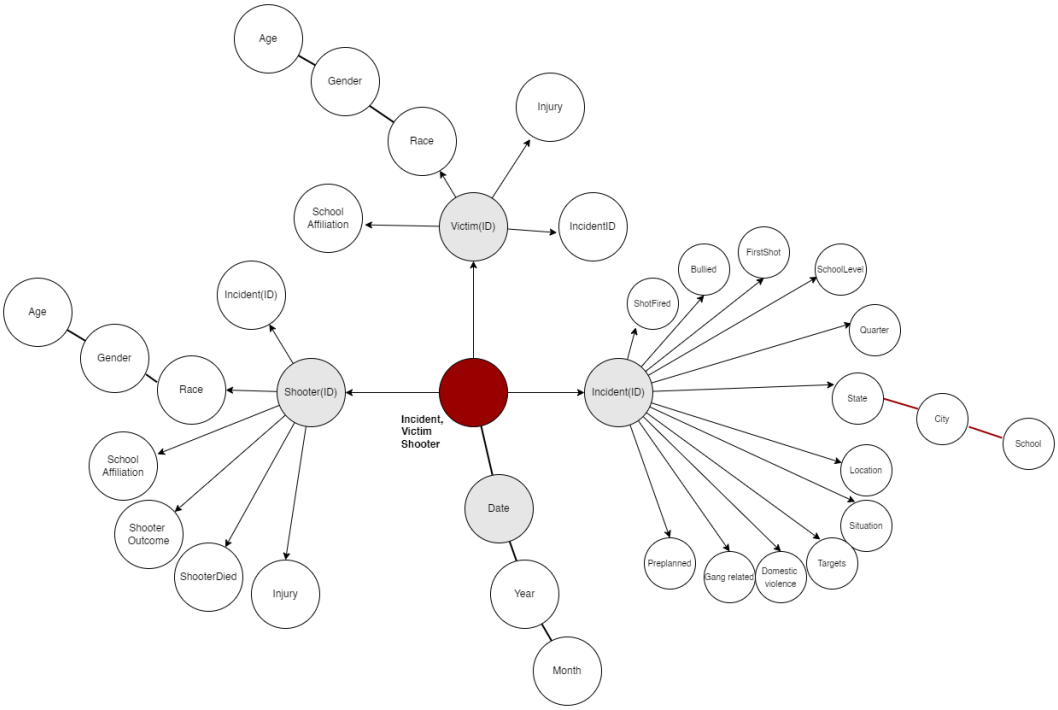


Figure 2.3: Edited tree

So, since this data set is not very rich, finding measures was very tough. In particular, there is very little numerical data, other than shots fired. Moreover, the connection between data is not so good. So, in the end, I decided to use as measures:

- nVictims
- nShooters

Although the measures were not very precise or rich in values, I was able to use Tableau to create dashboards and analysis sheets in the data analysis part that reported fairly accurate and satisfactory results, paying attention to finding important correlations between the attributes of the dimensions.

2.4 Fact schema

From the edited tree, I was able to determine a fact schema for the dimensions, measures, and the fact I want to analyze.

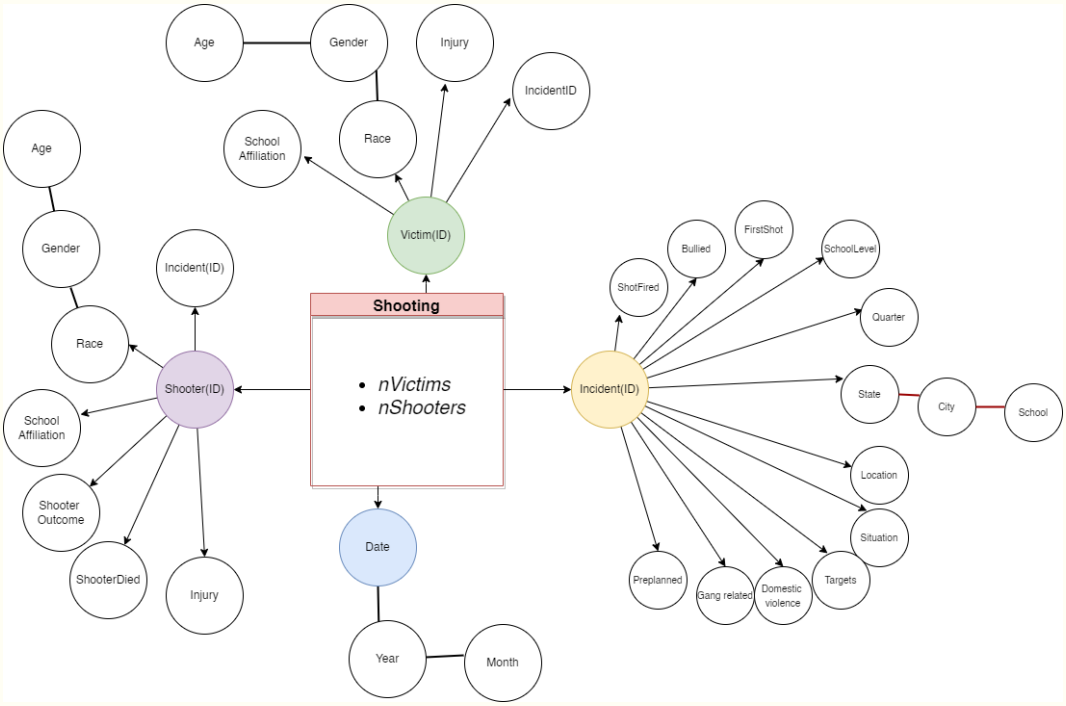


Figure 2.4: Fact schema

2.5 Star schema

Finally, here is the star schema:

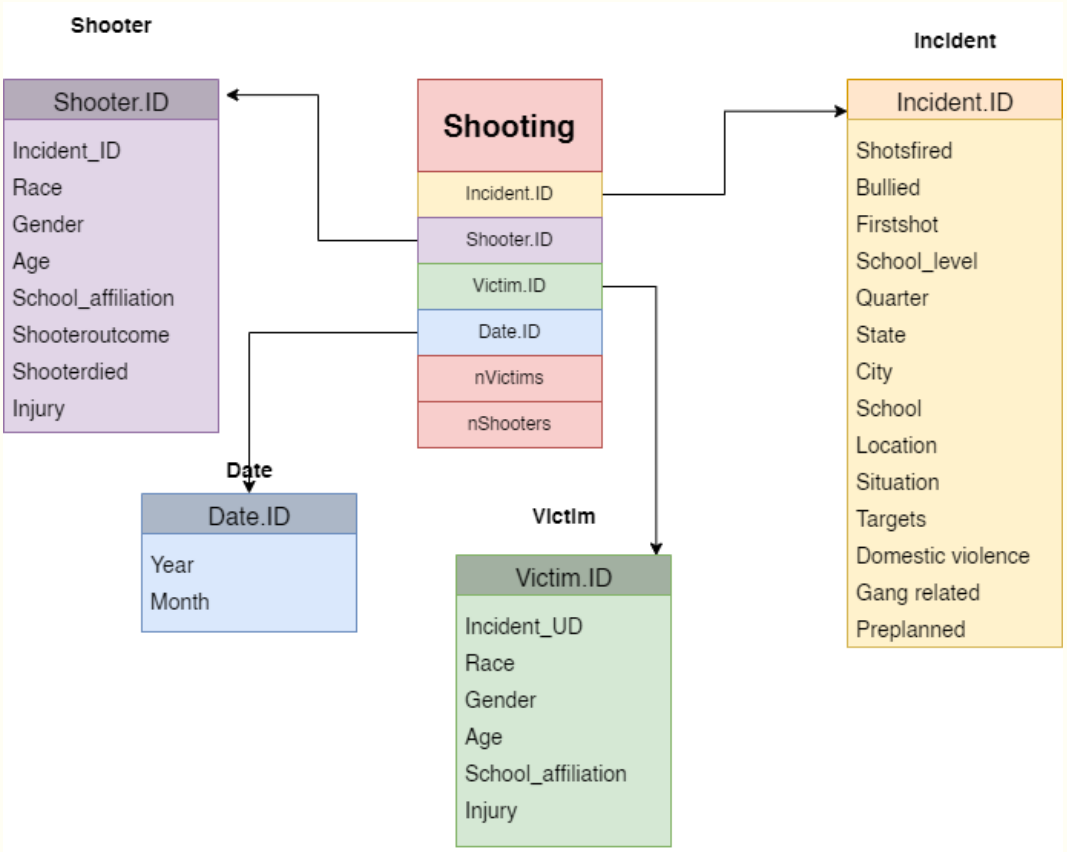


Figure 2.5: Star schema

Chapter 3

Data Cleaning and preprocessing

3.1 Data Quality Assessments

Before going to preprocess the data, let's see the quality of the sources.

```
incident_id in incident.csv:  0.9995147986414362
incident_id in shooter.csv:  0.8980306345733041
incident_id in victim.csv:  0.6029744584545749
Not in cities:  0.039786511402231925
Missing values in incident.csv column Incident_ID :  0.0
Missing values in incident.csv column Sources :  0.0
Missing values in incident.csv column Number_News :  0.6671518680252305
Missing values in incident.csv column Media_Attention :  0.6482290150412421
Missing values in incident.csv column Reliability :  0.0
Missing values in incident.csv column Date :  0.0
Missing values in incident.csv column Quarter :  0.005337214944201844
Missing values in incident.csv column School :  0.00048520135856380397
Missing values in incident.csv column City :  0.0
Missing values in incident.csv column State :  0.0
Missing values in incident.csv column School_Level :  0.011644832605531296
```

```
Missing values in incident.csv column Location :  0.001455604075691412
Missing values in incident.csv column Location_Type :  0.0024260067928190197
Missing values in incident.csv column During_School :  0.010674429888403688
Missing values in incident.csv column Time_Period :  0.07957302280446385
Missing values in incident.csv column First_Shot :  0.2304706453178069
Missing values in incident.csv column Summary :  0.0
Missing values in incident.csv column Narrative :  0.00727802037845706
Missing values in incident.csv column Situation :  0.07763221737020863
Missing values in incident.csv column Targets :  0.14750121300339641
Missing values in incident.csv column Accomplice :  0.14361960213488598
Missing values in incident.csv column Hostages :  0.006307617661329452
Missing values in incident.csv column Barricade :  0.006307617661329452
Missing values in incident.csv column Officer_Involved :  0.003396409509946628
Missing values in incident.csv column Bullied :  0.14944201843765162
Missing values in incident.csv column Domestic_Violence :  0.07520621057738962
Missing values in incident.csv column Gang_Related :  0.20232896652110627
Missing values in incident.csv column Preplanned :  0.043182920912178555
Missing values in incident.csv column Shots_Fired :  0.3197476952935468
Missing values in incident.csv column Active_Shooter_FBI :  0.4721009218825813
Missing values in shooter.csv column incidentid :  0.0
Missing values in shooter.csv column age :  0.18599562363238512
Missing values in shooter.csv column gender :  0.15798687089715535
Missing values in shooter.csv column race :  0.7190371991247265
Missing values in shooter.csv column schoolaffiliation :  0.061269146608315096
Missing values in shooter.csv column shooteroutcome :  0.002188183807439825
Missing values in shooter.csv column shooterdied :  0.003938730853391685
Missing values in shooter.csv column injury :  0.00350109409190372
Missing values in shooter.csv column chargesfiled :  0.6923413566739606
Missing values in shooter.csv column verdict :  0.9203501094091904
Missing values in shooter.csv column minorchargedadult :  0.8936542669584245
Missing values in shooter.csv column criminalhistory :  0.9185995623632385
```

```
Missing values in victim.csv column incidentid :  0.0
Missing values in victim.csv column race :  0.8978338182993857
Missing values in victim.csv column injury :  0.0
Missing values in victim.csv column gender :  0.2276107339152926
Missing values in victim.csv column schoolaffiliation :  0.10410604591011963
Missing values in victim.csv column age :  0.09117361784675072
Not valid dates %: 0.0
Not valid ages for shooter: 0.1864332603938731
Not valid ages for victim: 0.33365664403491757
```

I did not find a parameter to check for Integrity and Consistency, mainly because I had no primary keys for the CSVs, and this will be part of the next steps I illustrated in the report. By the way, there are some columns I dropped mainly because of the low amount of data. So, after this, I decided to fix the problems that are in the ages and dates and made the data valid for the analysis I want to conduct.

The script has been made using Python and Pandas.

```
25  citiesArray = cities['city'].array
26  notIn = 0
27  for city in incident['City']:
28      # check if city is in the cities[city] column
29      if city in citiesArray:
30          continue
31      else:
32          notIn += 1
33
34  print('Not in cities: ', notIn/len(incident['City']))
```

Figure 3.1: Cities reading for Conformity

3.2 Preprocessing steps

To perform this step, I used Pentaho Data Integration.

PDI uses a drag-and-drop design environment and a wide range of built-in connectors and transformers to enable users to extract data from various

sources, transform it according to business requirements, and load it into a target system. It supports a wide variety of data sources, including databases, files, and big data platforms, and can handle structured, semi-structured, and unstructured data.

The first thing I did was work directly from *CSV* files. Later on, I decided to switch from this approach using directly a **database** containing the data I needed. So, I created the tables containing the data which were inside the CSV files. I did this considering a hypothetical situation in which I had to work directly from a database that gets updated daily or weekly, so I decided to convert my **Pentaho** script from a *file input* to a *database input.*

Important note: The primary keys for the tables were not specified inside the CSVs files, and this was very problematic. To solve this issue, I loaded the values inside the database adding a new auto incremental key as the primary key for *victim and shooter.*

3.3 Main challenges during cleaning

The initial thing I did was remove all the useless attributes I didn't need inside all the dimensions, so I used the **Select Value** transformation to keep only the attributes I inserted inside the 4 dimensions.

After this step, I started my research for **abnormal values** and **null values**. This was not very easy, mainly because *abnormal values* are not well indicated. In particular, some strange values like:

- NULL
- Null
- N/A
- None

, were inside a lot of different columns in different ways. So, I got all the unique values for all the columns and created a regex to find all those occurrences and remove them all, changing the value to "Unknown".

Why not remove the rows?

Since the data set is very small and doesn't have a lot of rows by itself, and because the analysis I needed was not mainly on the columns I found with null values, I prefer to keep those rows to have more values to analyze in the general view.

3.3.1 Other data cleaning problems

Other problems were more **format oriented** than **null oriented**. Since I had to work with *ages* both for *Victims* and *Shooters*, I had a lot of problems with values containing *Range of values, Null values, Not numeric values.* Thinking about that, I decided to remove the numerical part of the age to arrange the ages into a **Range of values** using a **Number range**. The ages are divided into these categories:

- From 2 to 10: Child
- From 10 to 13: Young
- From 13 to 18: Teen
- From 18 to 60: Adult
- Over 60: Senior

So basically the main problems for the Victim and Shooter dimension were to find those strange values and remove the null values with the "**Unknown**" value.

Just to have an idea of the flow of transformation:

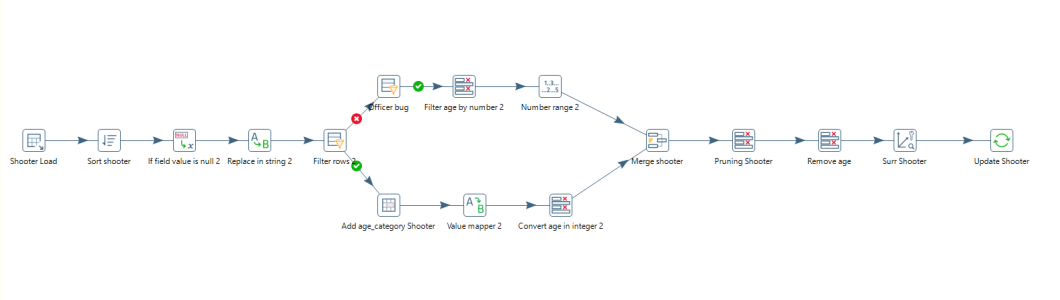


Figure 3.2: Victim processing in Pentaho

3.3.2 Incident cleaning

The incident data source was the most populated one in terms of columns, so of course, was the one that required the majority of cleaning steps. Starting from converting the dates in my preferred format DD/MM/YYYY, then I noticed that the **Shotfired** attribute should have been numerical, and the problem was that some values inside were *number ranges* or categorical values like *A lot, More than 10, Less than 10, Tons*. So, to perform a correct cleaning, I divided the numerical values from the not numerical ones, then what I did was **calculating a mean of shots fired based on the state in which the incidents occurred**, then I used those means only on the rows that contained the not numerical value.

The initial idea was to *calculate a mean for the total bullets*. The idea was discarded because it would have led to a repetition of the same mean with no usage, so the more complex but functional idea was working better.

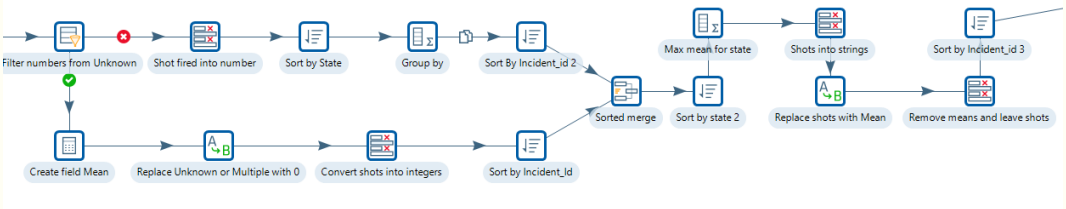


Figure 3.3: Pentaho incident cleaning - Shots problem

3.3.3 Cleaning summary

Here is a list of other cleaning activities I conducted:

1. Sort rows by ID
2. Prune useless attributes
3. Graft needed attributes
4. Substituted null values with "Unknown"
5. Substituted abnormal values with "Unknown"
6. Converted values into correct data type
7. Removed primary keys rows
8. Changed date formats
9. Created not existing fields

3.4 Dimensions creation

The next step was creating the dimensions tables to conduct analysis on. To do it, I created another database in which I inserted the base tables for:

- incident
- date
- shooter
- victim

Is important to specify that the *dimension tables* should have as **primary key** a **surrogate key**, that is different from the primary key of the data source. Surrogate keys are used as an alternative to primary keys for a variety of reasons, including maintaining database integrity, ensuring uniqueness, improving performance, and obscuring sensitive data. So, in the *combination/lookup step*,

I created a surrogate key based on the primary key of the table. All the surrogate keys I created were with auto-increment based on table-maximum +1.

3.4.1 Date creation

To create the **date dimension**, I extracted the date from the incident table, but since the date was never created inside the main data source, the primary key didn't exist. To solve this issue, I used the *Add sequence step* to generate an auto-incremental column for the **Date** table and then generate a surrogate key for it.

3.5 Dimensions loading

After all these steps, I used an **Update step** to update the values inside the dimensions with the correct ones. These steps ensure that every time the transformation is launched, the data inside the dimension tables is updated, so this ensures that every time something changes in the data sources, it will work even in the data warehouse.

3.6 Fact table creation

The last, but not least, step to perform was to create a transformation that reads all the values I needed to create my fact table. To do this, since I needed a combination of 3 different keys to create my main fact, I searched for all the rows that contained the same incident_id and merged them all into one row, extracting only the **surrogate keys** I needed for *victim, shooter, date, and incident.* Then I removed all the useless attributes from the query result, **merged the 4 keys into one *primary key*** and then loaded the table in the data warehouse.

Rows of step: Table output (3703 rows)

#	surr_inc_id	surr_vict_id	surr_shoot_id	surr_date_id	nShooters	nVictims	shooting_id
1	1	1780	1	1	2	2	1-1780-1-1
2	1	1	1	1	2	2	1-1-1-1
3	2	1781	1435	2	1	1	2-1781-1435-2
4	3	2	1436	3	1	1	3-2-1436-3
5	4	3	2	4	1	1	4-3-2-4
6	6	4	4	6	4	4	6-4-4-6
7	6	4	3	6	4	4	6-4-3-6
8	6	5	4	6	4	4	6-5-4-6
9	6	5	3	6	4	4	6-5-3-6
10	7	6	1443	7	48	48	7-6-1443-7
11	7	6	1441	7	48	48	7-6-1441-7
12	7	6	1440	7	48	48	7-6-1440-7
13	7	6	1438	7	48	48	7-6-1438-7
14	7	6	1437	7	48	48	7-6-1437-7
15	7	6	1442	7	48	48	7-6-1442-7
16	7	6	1439	7	48	48	7-6-1439-7
17	7	6	5	7	48	48	7-6-5-7
18	7	1782	1443	7	48	48	7-1782-1443-7
19	7	1782	1441	7	48	48	7-1782-1441-7
20	7	1782	1440	7	48	48	7-1782-1440-7
21	7	1782	1438	7	48	48	7-1782-1438-7
22	7	1782	1437	7	48	48	7-1782-1437-7
23	7	1782	1442	7	48	48	7-1782-1442-7
24	7	1782	1439	7	48	48	7-1782-1439-7
25	7	1782	5	7	48	48	7-1782-5-7
26	7	1783	1443	7	48	48	7-1783-1443-7
27	7	1783	1441	7	48	48	7-1783-1441-7
28	7	1783	1440	7	48	48	7-1783-1440-7
29	7	1783	1438	7	48	48	7-1783-1438-7
30	7	1783	1437	7	48	48	7-1783-1437-7
31	7	1783	1442	7	48	48	7-1783-1442-7
32	7	1783	1439	7	48	48	7-1783-1439-7
33	7	1783	5	7	48	48	7-1783-5-7

Figure 3.4: Output for Fact Table

3.7 The final job

All the transformations above are inserted inside a unique job with the fact table update/creation as the last step.

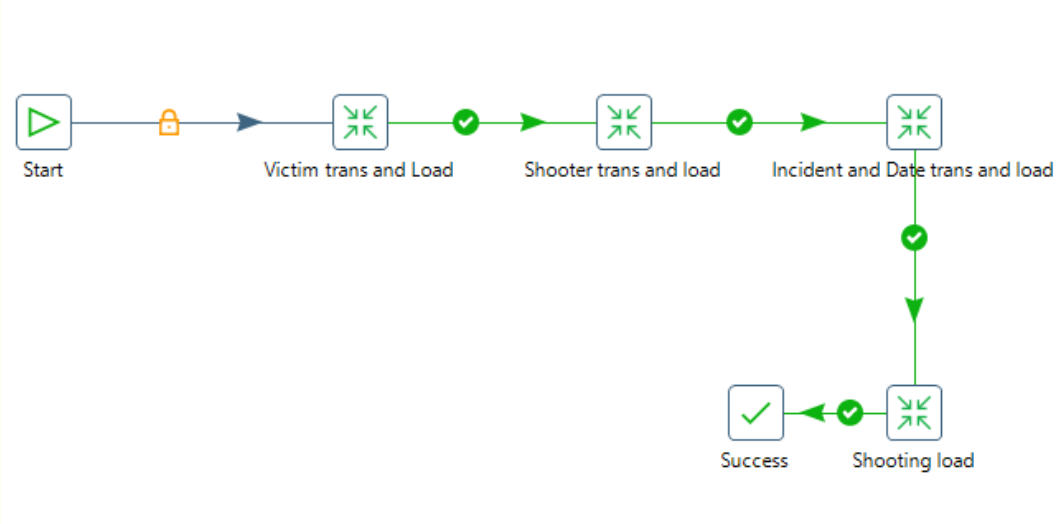


Figure 3.5: Main job

Chapter 4

Data analysis and Tableau

To conduct the analysis, I used Tableau. Tableau is a business intelligence software for creating dashboards and reports that allow users to easily visualize and analyze data. With its drag-and-drop interface, users can create interactive charts and tables.

4.1 Analysis summary

To conduct the analysis I created 4 different dashboards with some self-usable analysis sheets. In particular, the dashboards are not-static and interactive and are the following:

1. General view of Shooting: USA Map with Victims per **State**, Victims during **years** and the **Percentage** of Woman / Men
2. Motivation for School and Gender, showing a tree map of the number of incidents, places, and trends based on the number of Women or Men injured in the incident.
3. Motivation during years. Basically a line trend with the number of incidents and a bar chart for Bullied / Gang-related / Domestic violence.
4. Place, Hour, State, and Quarter. It is the most precise sheet of the analysis and tells the time in which the shooting occurred, the location, the quarter of the year, and the state.

5. Age and situation correlation. A sheet that tells how age influenced the situation of the incident, like the **Child** category has the majority of situations of **Accidental**.

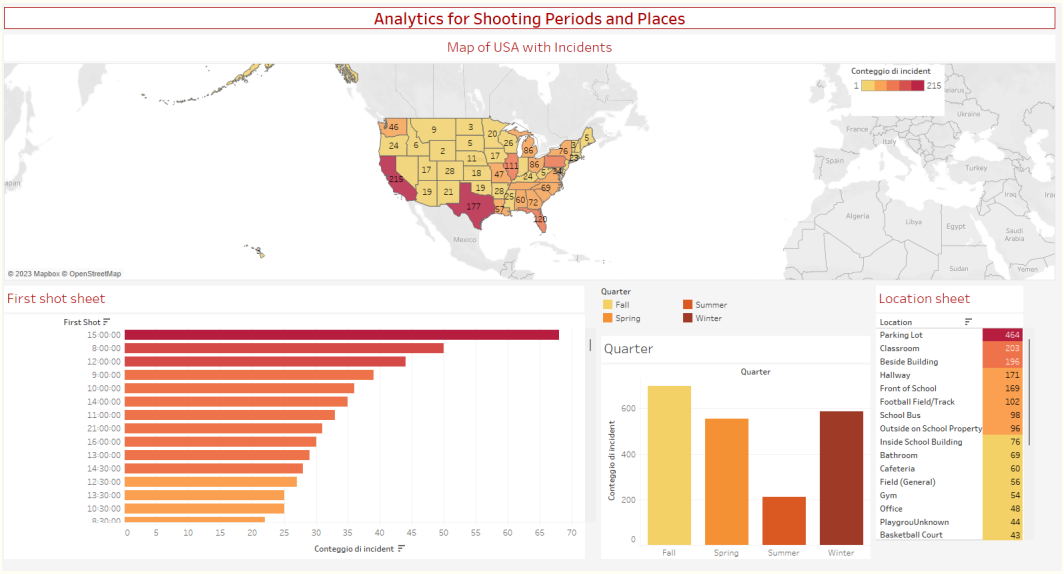


Figure 4.1: Place, Hour, State and Quarter

The analysis sheets are with dashboards inside 2 stories I created inside Tableau, showing how the trends changed over the year, the most dangerous place to hide, the most dangerous state based on the number of shootings, etc...

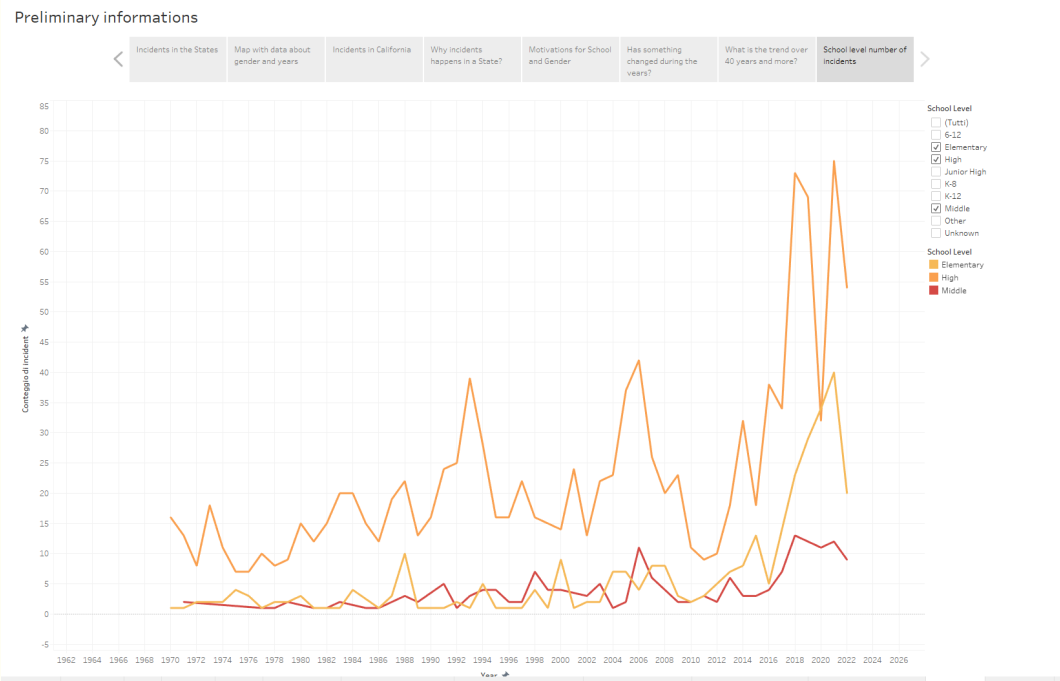


Figure 4.2: Tableau story 1

I am not inserting any outcome of this or any sort of major information. That's because the main focus is to explore the dashboards since the majority of results are possible to extract by comparing some States, comparing years, comparing Dates, etc...

Chapter 5

Conclusions

I want to conclude by saying that I wouldn't ever imagine finding more school shootings in the last 5 years than in 1980-90. A very interesting fact is that the majority of shootings are not caused by any sort of bullied person or any relation with gangs. The majority of them are caused by some *escalation of dispute*, and I expected to find more racist shootings rather than just a few of them.

Probably with a larger data set we would have found records with more occurrences for these reasons. The fact remains that it is puzzling how many shootings occurred for mild motivations.

This proves how much the USA School Shootings problem is growing and needs to be treated with more caution, and this is where the analysis comes in our help. Finding the *places, hours, and schools* in which the shooting occurred, we can find a particular repetition among the data. Working with authorities could activate some *helping courses* for the students, like psychological hours or something like that, mainly in the schools in which a shooting happened or is likely to happen. Moreover, the amount of data for the precise hour in which shootings happened can easily help security place guards in a particular place and at a particular time. This is a very simple and simplified solution, but this is a very simple data set and a very simple data analysis. Having a more precise data set could give more predictive values and technical information about the shooting and future possible shootings.