Data Mining Project Report

A.Y. 2023/2024

Funaioli Francesca
Karoui Hamza
Mitola Francesco
Vezzuto Samuele

Contents

Chapter 1

Data Understanding

The data understanding phase aims to prepare data for the following data mining tasks and to gain informations on the general properties of our data. This phase of data analysis is focused on an identification of missing values, outliers, duplicates and distribution analysis. The three csv files, `incidents.csv`, `povertyByStateYear.csv` and `year_state_district_house.csv` were imported and analyzed separately. The first preliminary checks comprised looking for null values and checking the data types of the attributes, before a more in depth analysis of each column of each dataset. By computing the statistics of the numerical values of each dataset, we noticed a wrong value in attribute *participant_age_1* (values of 311). We also found out that some of the columns considered contained non-numerical values, as they were not displayed. The value of 0 for attribute *participant_age_1* occurs when there are infants involved and therefore it is not to be considered a wrong value. In a similar way, *n_participants* has value 0 if the incident had no victims or no shots where fired, but also in records with most of the attributes set to null.

1.0.1 Incidents

Date

Each *date* value was converted to a datetime object, after checking that there where no null or NaT values. The plot in Figure 1.1, representing the number of incidents per year, shows an increasing trend in the number of crimes over the years. It also shows a lack of records related to years 2013 and 2018: there are only 253 incidents happened in 2013, hence they will be removed during data preparation. As for the year 2018, there is only data relating to the first three months of the year in the dataset, so year 2018 will be analyzed on its own.

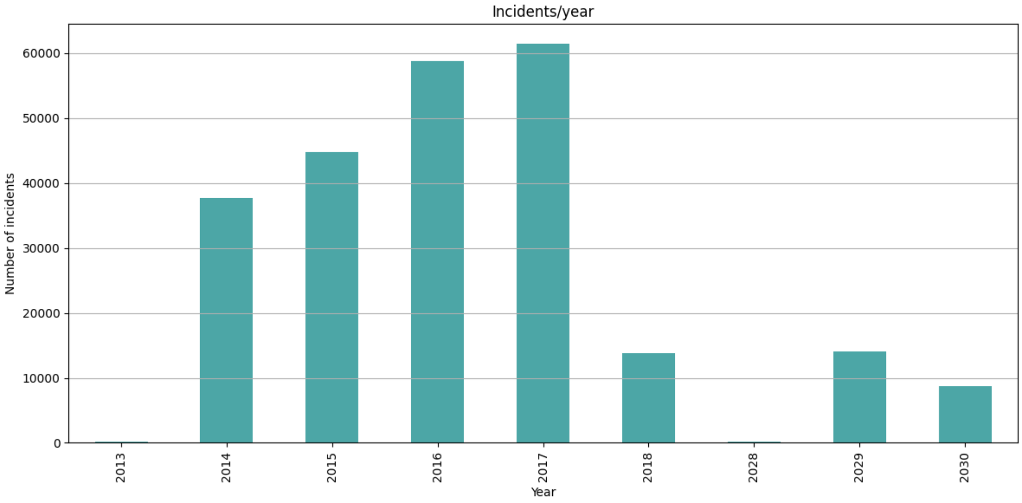


Figure 1.1: Number of incidents per year.

Geographical information

The *state* attribute contains no null values and it has 51 unique values: the 50 states of the United States and the District of Columbia, which we will consider as a special state. Figure 1.2 shows the number of incidents recorded in each state in decreasing order.

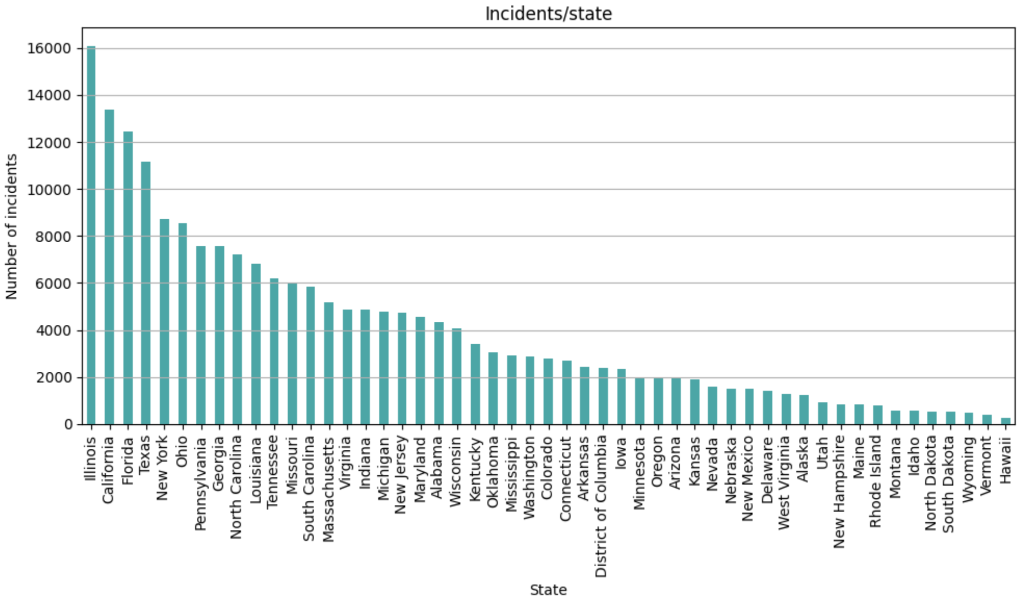


Figure 1.2: Number of incidents per state.

The *city_or_county* attribute has no null values, but it often contains additional informations about the suburb or the neighborhood in brackets, that provide a more precise location for the incident, e.g. “Minneapolis (Brooklyn Center)”. We consider removing this kind of information in the next phase. There are also instances containing the substring “(county)” at the end, e.g. “Orange (county)”, that also appears as “Orange”: these instances will be substituted using just the name of the city/county. Figure 1.3 shows the number of incidents recorded in each city/county in decreasing order.

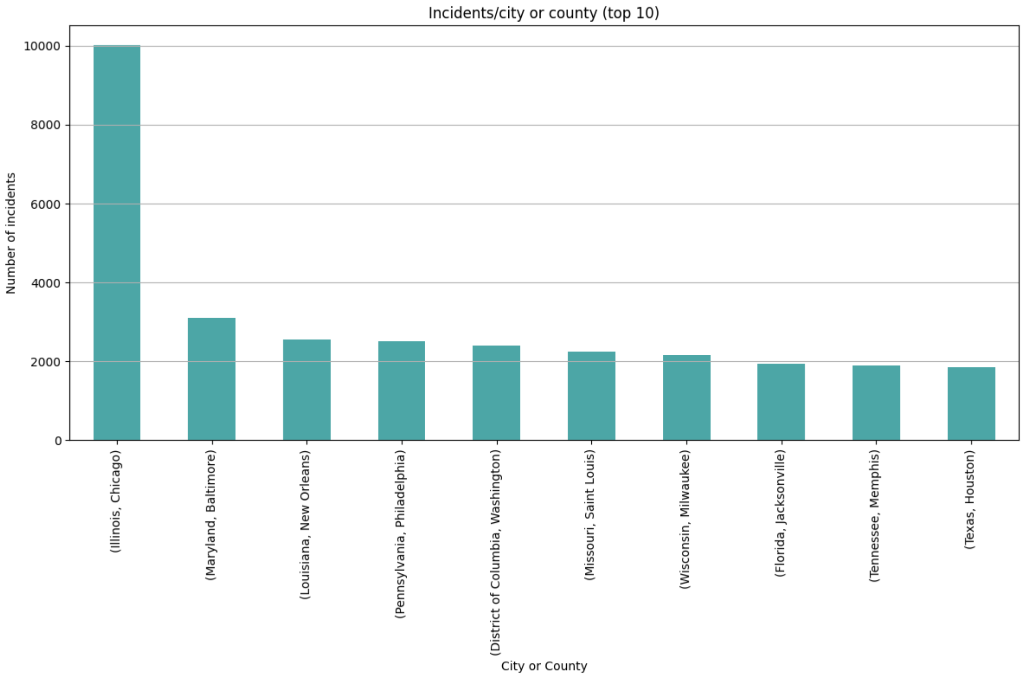


Figure 1.3: Number of incidents per city or county.

The *address* attribute contains some null values, but we believe it does not hold any statistical value, given that more specific information about the exact location of an incident is found by using the geographical coordinates. Furthermore, there are only 6020 records for which address information can not be inferred using *latitude* and *longitude* attributes, so we will remove this column.

The *latitude* and *longitude* attributes contain some null values. We also noticed some outliers by drawing empirical box boundaries of the United States: there are some incidents recorded outside of the U.S. that will be removed in the next phase.

Attributes *congressional_district*, *state_house_district* and *state_senate_district* contain some null values. We also noticed that most of the incidents happened in the state of Illinois, by plotting the top 10 incidents for each of these attributes.

Age and gender information

The *participant_age1* attribute contains some outliers, mostly being values of type string and values that are too large to be the age of a person. There are also some outliers if this attribute is compared to the corresponding value reported in the *participant_age_group1* field. As shown in Figure 1.4, most of the participants are adult males. The attributes *participant_age1*, *min_age_participants*, *max_age_participants* and *avg_age_participants* all have similar distributions.

The attributes *n_participants_child*, *n_participants_teen* and *n_participants_adult* all present the same issues: they all contain outliers given by non-numerical strings, very large or negative numbers.

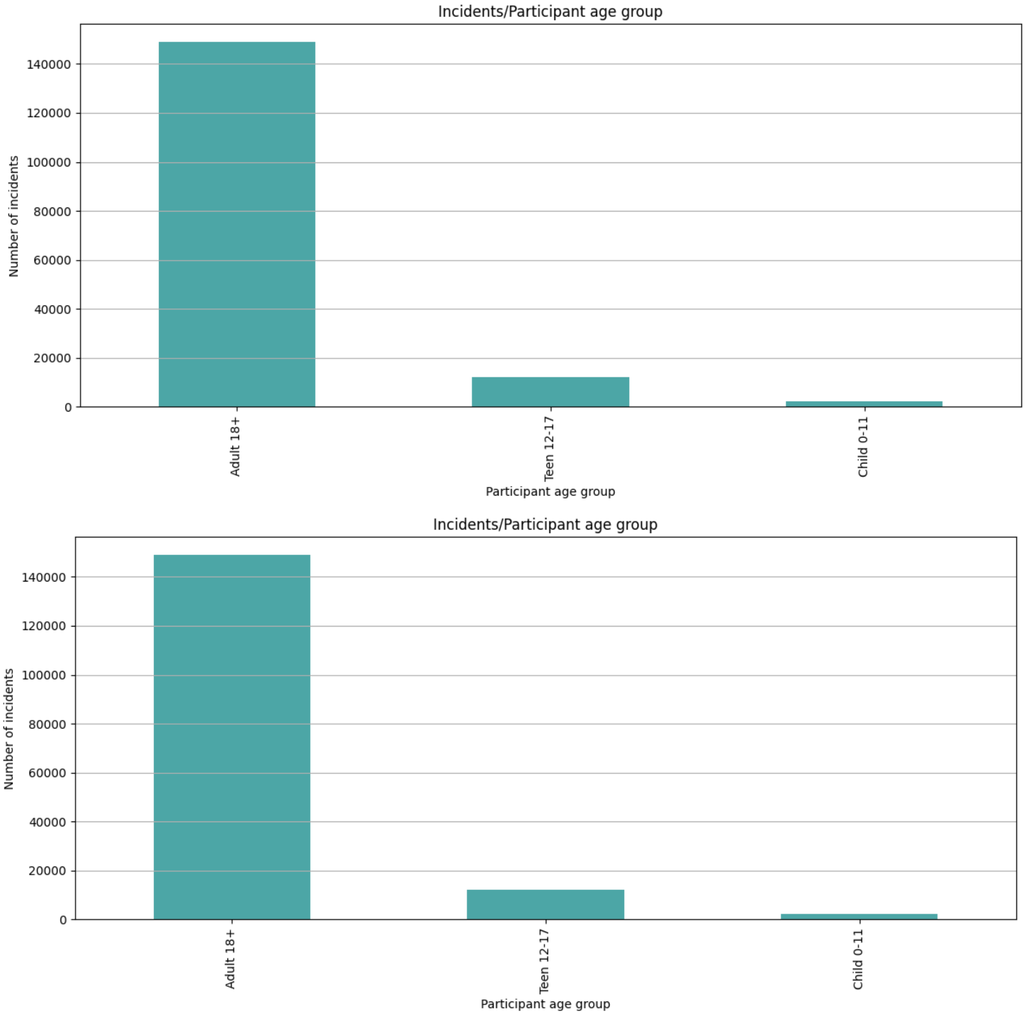


Figure 1.4: Number of incidents per age group (above) and gender (below) of participant1.

Number of involved people

The majority of the incidents only involve between 0 and 5 people, with almost no killed, injured, unharmed or arrested people.

Notes and incident characteristics

We consider these attributes to hold no statistical value.

1.0.2 Poverty by state

The *state* attribute contains 52 unique values: 51 of them are the same as the states in the incidents dataset, the remaining one is labeled "United States" and contains the average of the whole country. We consider using the average to possibly fill the missing values in the following phase.

There are no *povertyPercentage* values for the year 2012, but we are only interested in relating this information to the incidents dataset, which only contains relevant incidents in the range of years 2013-2018.

1.0.3 Year state district house

This dataset contains no null values. We will only consider data in the range of years 2013-2018 for integrating this data with the incidents dataset.

1.1 Data Integration

We created an additional column called *total_votes_for_state* in the year-state-house-district dataset: this column contains the total number of votes for each state and for each year. We merged the incidents dataset with the poverty dataset using the attributes *state* and *year*. We then merged the resulting dataset with the remaining one using the attributes *state*, *year* and *congressional_district*. During the data integration process, records containing incidents set outside the U.S. were automatically deleted, resulting in a dataset that has no outliers in attributes *latitude* and *longitude*.

The dataset obtained by data integration can be used to further analyze and relate political party and poverty percentage to each congressional district.

TODO *si potrebbero aggiungere degli screen dei plot delle mappe, quella degli shooting e quella dei map shooting. in questo momento però non riesco a fare gli screen*

1.2 Distribution and Correlation Analysis

The plots in Figure 1.5 show that, as previously said, age attributes all have very similar distributions.

From Figure 1.6 we can see that most of the incidents involve very few participants. Specifically, the majority of them only involve between 0 and 3 participants, with incidents having only one person involved being the most common ones.

The correlation matrix is shown in Figure 1.7. At first glance, the only noteworthy correlations are those between the *participant_age1* feature of the randomly taken person and the attributes *min_age_participants*, *max_age_participants* and *age_age_participants*. This correlation is confirmed by the fact that the majority of the incidents only involve 1 or 2 participants. The correlation matrix will be computed again once outliers in the numerical attributes, which compromise the correlation calculation, have been eliminated.

We also observed that the number of males involved has a high correlation (0.83) with the number of participants because on average, as seen above, incidents tend to have a much more higher number of males participants than females participants.

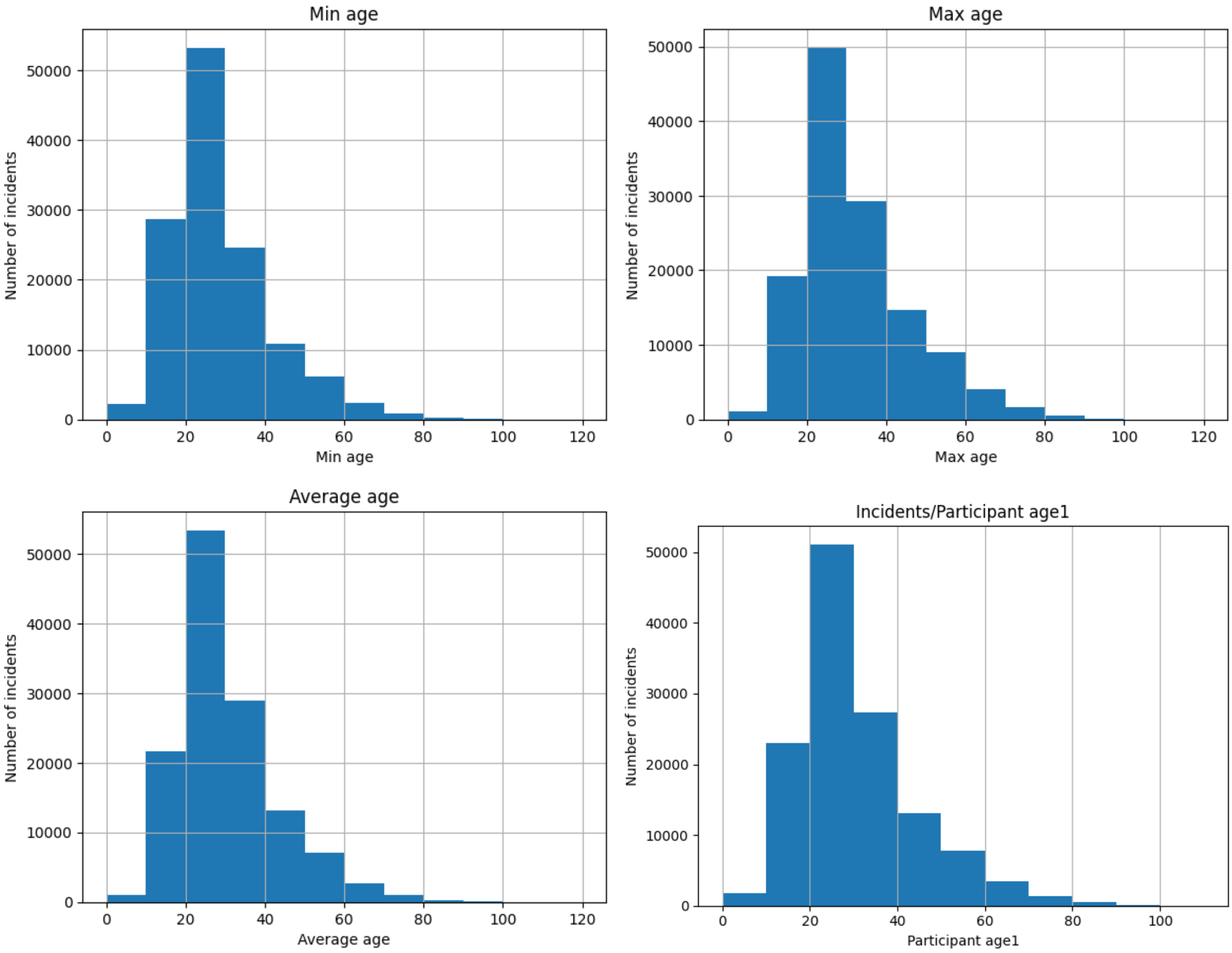


Figure 1.5: Plot of minimum (top left), maximum (top right), average (bottom left) and participant1 (bottom right) age distributions.

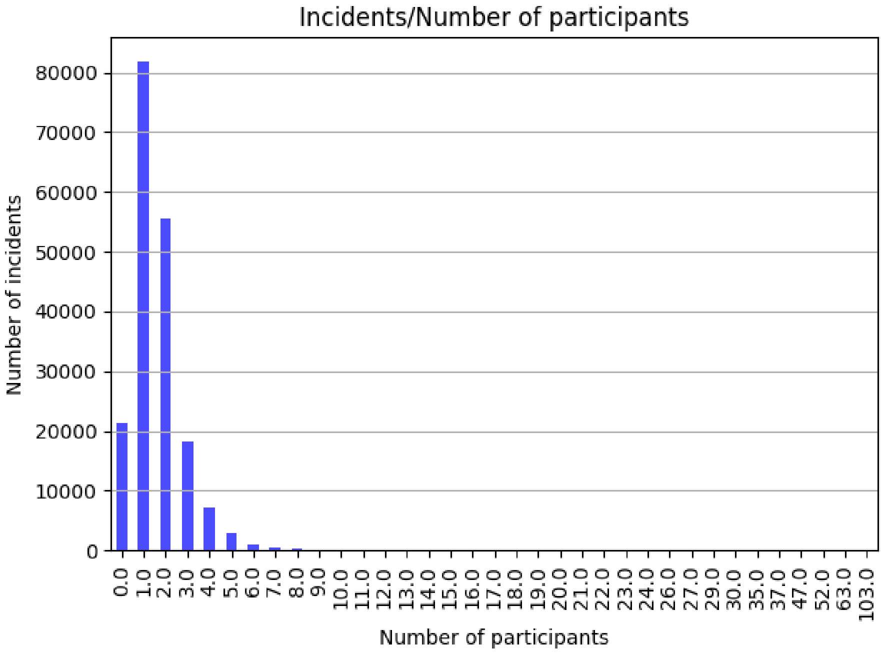
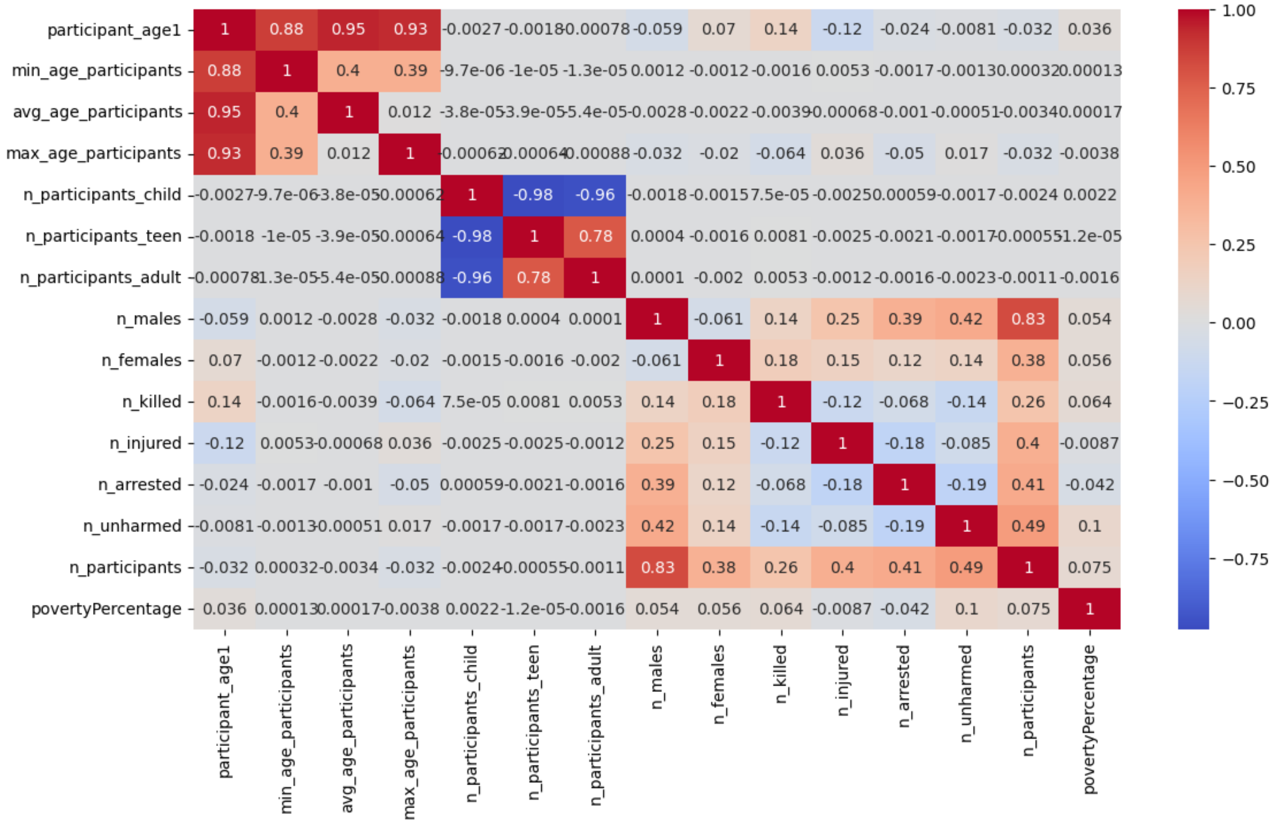


Figure 1.6: Number of incidents per number of people involved.



	participant_age1	min_age_participants	avg_age_participants	max_age_participants	n_participants_child	n_participants_teen	n_participants_adult	n_males	n_females	n_killed	n_injured	n_arrested	n_unharmed	n_participants	povertyPercentage
participant_age1	1	0.88	0.95	0.93	-0.0027	-0.0018	-0.00078	-0.059	0.07	0.14	-0.12	-0.024	-0.0081	-0.032	0.036
min_age_participants	0.88	1	0.4	0.39	-9.7e-06	-1e-05	-1.3e-05	0.0012	-0.0012	-0.0016	0.0053	-0.0017	-0.0013	0.00032	0.00013
avg_age_participants	0.95	0.4	1	0.012	-3.8e-05	-3.9e-05	-5.4e-05	-0.0028	-0.0022	-0.0039	-0.00068	-0.001	-0.00051	-0.0034	0.00017
max_age_participants	0.93	0.39	0.012	1	-0.00062	0.00064	0.00088	-0.032	-0.02	-0.064	0.036	-0.05	0.017	-0.032	-0.0038
n_participants_child	-0.0027	-9.7e-06	-3.8e-05	-0.00062	1	-0.98	-0.96	-0.0018	-0.0015	7.5e-05	-0.0025	0.00059	-0.0017	-0.0024	0.0022
n_participants_teen	-0.0018	-1e-05	-3.9e-05	0.00064	-0.98	1	0.78	0.0004	-0.0016	0.0081	-0.0025	-0.0021	-0.0017	-0.00055	-1.2e-05
n_participants_adult	-0.00078	-1.3e-05	-5.4e-05	0.00088	-0.96	0.78	1	0.0001	-0.002	0.0053	-0.0012	-0.0016	-0.0023	-0.0011	-0.0016
n_males	-0.059	0.0012	-0.0028	-0.032	-0.0018	0.0004	0.0001	1	-0.061	0.14	0.25	0.39	0.42	0.83	0.054
n_females	0.07	-0.0012	-0.0022	-0.02	-0.0015	-0.0016	-0.002	-0.061	1	0.18	0.15	0.12	0.14	0.38	0.056
n_killed	0.14	-0.0016	-0.0039	-0.064	7.5e-05	0.0081	0.0053	0.14	0.18	1	-0.12	-0.068	-0.14	0.26	0.064
n_injured	-0.12	0.0053	-0.00068	0.036	-0.0025	-0.0025	-0.0012	0.25	0.15	-0.12	1	-0.18	-0.085	0.4	-0.0087
n_arrested	-0.024	-0.0017	-0.001	-0.05	0.00059	-0.0021	-0.0016	0.39	0.12	-0.068	-0.18	1	-0.19	0.41	-0.042
n_unharmed	-0.0081	-0.0013	-0.00051	0.017	-0.0017	-0.0017	-0.0023	0.42	0.14	-0.14	-0.085	-0.19	1	0.49	0.1
n_participants	-0.032	0.00032	-0.0034	-0.032	-0.0024	-0.00055	-0.0011	0.83	0.38	0.26	0.4	0.41	0.49	1	0.075
povertyPercentage	0.036	0.00013	0.00017	-0.0038	0.0022	-1.2e-05	-0.0016	0.054	0.056	0.064	-0.0087	-0.042	0.1	0.075	1

Figure 1.7: Correlation matrix of numerical attributes.

Chapter 2

Data Preparation

The data preparation phase uses the information gained in the previous phase to select records, manage outliers and missing values and improve data quality. We started by changing the data types of the attributes as shown in Table 2.1.

We then removed negative values by setting them to NaN.

Age attributes

For the attributes *participant_age1*, *min_age_participants*, *max_age_participants* and *avg_age_participants*, we considered values ≥ 120 as outliers and we set them to NaN. These three attributes seemed to be very correlated, so we consider deleting the columns *min_age_participants* and *max_age_participants* and only keeping *avg_age_participants.*

Date

We considered all dates after 2023-10-01 (the date we received the dataset) to be outliers, in particular errors in the data. We dropped all records related to year 2013, as the year was under-represented (only 242 records).

Geographical attributes

The records with coordinates outside U.S. (other that null values) were automatically deleted after the data integration. We consider the triple $<$*date,latitude,longitude*$>$ to be a key identifying an incident: we assume that there are no incidents happening on the same day in the exact same geographic coordinates. Hence, we decided to eliminate the records in which these 3 values are duplicate.

For the rows in which *latitude* and *longitude* are NaN, we filled the missing values using the mean computed for the respective *state* and *city_or_county.*

We decided to drop the columns *state_house_district* and *state_senate_district*, given that they represent further subdivisions of the US territory that are not pertinent to our analysis. In fact, we are only interested in the *congressional_district*, because the electoral information has the same granularity.

Number of participants' attributes

We checked whether the number of killed, injured, arrested and unharmed people exceeded the total number of participants in that incident. Since, *n_arrested* and *n_killed* were the only two attributes with null values, we set them to 0 in case *n_participants* was 0; we filled the remaining null values using the mean.

We set to NaN the outliers found in *n_participants_adult*, *n_participants_teen* and *n_participants_child* found during data understanding. The outliers considered in this step were the values larger than the maximum values of *n_participants.* We also set to zero the attributes *n_participants_adult*, *n_participants_teen*, *n_participants_child*, *n_males* and *n_females* when *n_participants* is 0. We replace the value in *n_participants*

Feature Name	Initial Type	Cast Type	Description
date	object	Datetime64	date of incident occurrence
state	object	String	state where incident took place
city_or_county	object	String	city or county where incident took place
address	object	String	address where incident took place
latitude	float64	float64	latitude of the incident
longitude	float64	float64	longitude of the incident
congressional_district	int64	Int64	congressional district where the incident took place
state_house_district	int64	Int64	state house district
state_senate_district	float64	Int64	state senate district where the incident took place
participant_age1	float64	Int64	exact age of one (randomly chosen) participant in the incident
participant_age_group1	object	String	exact age group of one (randomly chosen) participant in the incident
participant_gender1	object	String	exact gender of one (randomly chosen) participant in the incident
min_age_participants	object	Int64	minimum age of the participants in the incident
avg_age_participants	object	float64	average age of the participants in the incident
max_age_participants	object	Int64	maximum age of the participants in the incident
n_participants_child	object	Int64	number of child participants 0-11
n_participants_teen	object	Int64	number of teen participants 12-17
n_participants_adult	object	Int64	number of adult participants (18 +)
n_males	float64	Int64	number of males participants
n_females	float64	Int64	number of females participants
n_killed	int64	Int64	number of people killed
n_injured	int64	Int64	number of people injured
n_arrested	float64	Int64	number of arrested participants
n_unharmed	float64	Int64	number of unharmed participants
n_participants	float64	Int64	number of participants in the incident
notes	object	String	additional notes about the incident
incident_characteristics1	object	String	incident characteristics
incident_characteristics2	object	String	incident characteristics
year	int64	Int64	
povertyPercentage	float64	float64	poverty percentage for the corresponding state and year
party	object	String	winning party for the corresponding congressional_district in the state, in the corresponding year
candidateVotes	int64	Int64	number of votes obtained by the winning party in the corresponding election
totalVotes	int64	Int64	total number of votes for the corresponding election
total_votes_for_state	int64	Int64	total number of votes for each year and for each state

Table 2.1: Features of the merged dataset

with the sum *n_males* + *n_females*, if this sum is equal to (*n_participants_adult* + *n_participants_teen* + *n_participants_child*) and different from *n_participants*. Viceversa, we set to NaN these attributes in the rows where the sums and/or *n_participants* are not equal. We then substituted NaN values using the mean of each attribute and dropped the few rows for which we were not able to reconstruct the mean.

Incident characteristics

We dropped the column *incident_characteristics2* given that it has 40% of null values and does not add meaningful details for our analysis.

2.1 Correlation analysis

We plotted the correlation matrix (Figure 2.1) only for the numerical attributes. We noticed that age attributes are highly correlated: we decided to drop all of them[1] except for *avg_age_participants*, which is the most correlated to the other attributes and gives us more general informations about all the participants. Given the fact that we dropped *participant_age1*, then the attributes *participant_gender1* and *participant_age_group1* become useless, so we dropped them. To fill NaN values in *avg_age_participants* we used the mean of each grouping on *n_participants*.

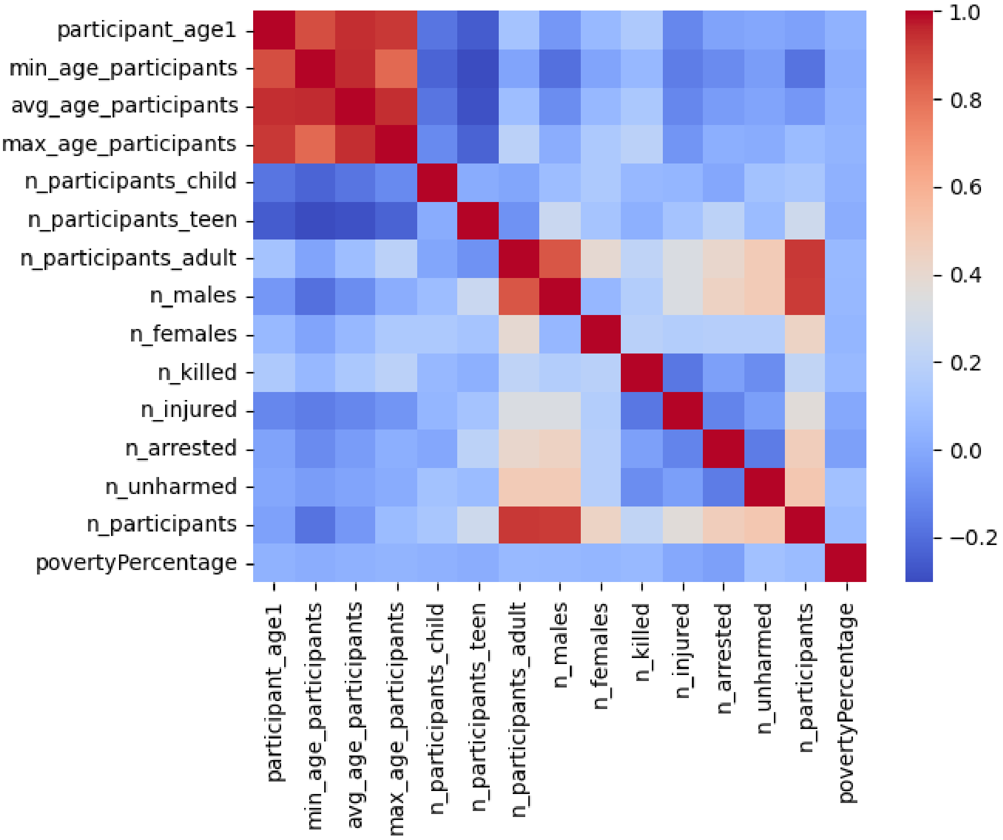


Figure 2.1: Correlation matrix plotted of the numerical attributes

2.2 Definition of indicators

We computed the following indicators:

[1] We dropped *participant_age1*, *min_age_participants* and *max_age_participants*.

- *males_percentage_per_city* (*females_percentage_per_city*), the number of males (females) involved in an incident over the total number of males (females) involved in incidents in the same city over the same time period;
- *killed_percentage_per_district* (*injured_percentage_per_district*, *arrested_percentage_per_district*, *unharmed_percentage_per_district*), the number of killed (injured, arrested, unharmed) people in an incident over the total number of people killed (injured, arrested, unharmed) in incidents in that same congressional district over the same time period;
- *killed_percentage_per_incident*, the number of killed people in each incident over the total number of participants in that same incident;
- *unharmed_percentage*, the number of unharmed people in the incident over the average of unharmed people in all the incidents in the same time period;
- *arrest_percentage*, the number of arrested people over the total number of participants in each incident;
- *killed_rate_per_state* (*injured_rate_per_state*, *arrested_rate_per_state*, *unharmed_rate_per_state*), the total number of people killed (injured, arrested, unharmed) per date and state over the total number of people killed (injured, arrested, unharmed) in that same date;
- *age_entropy_per_state*, the entropy of the *avg_age_participants* grouped for date and state;
- *winning_party_percentage*, the number of votes of the winning candidate over the total number of votes for that election.

We then plotted the correlation matrix (Figure 2.2) for the indicators defined above.

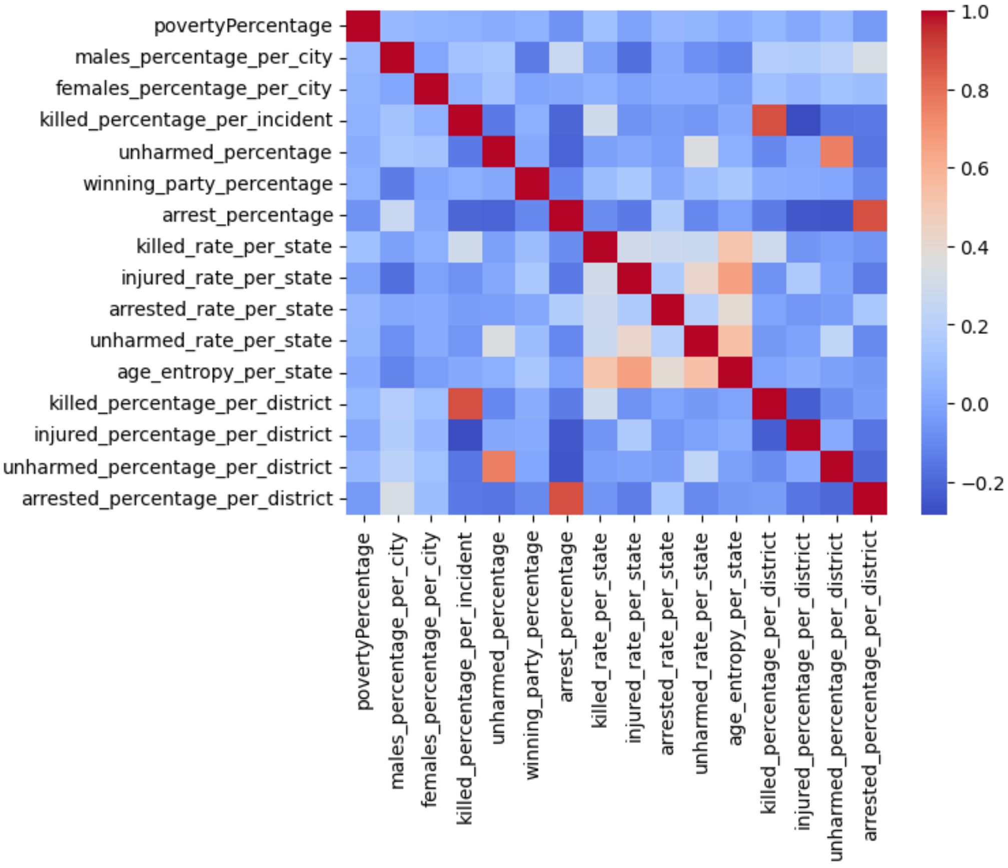


Figure 2.2: Correlation matrix plotted of the newly created indicators

We also decided to drop the columns *year* (just a result of data integration), *address* and *notes*.

Chapter 3

Clustering

In order to prepare data for applying the clustering algorithms, we did some steps of preprocessing. First of all we added a binary column *involve_killing* that has value 0 nobody was killed in the incident, and it has value 1 if there was at least a person killed in the incident. We applied a normalization to numerical values. The we computed the PCA with 2 components and got the visualization showed in Figure 3.1.

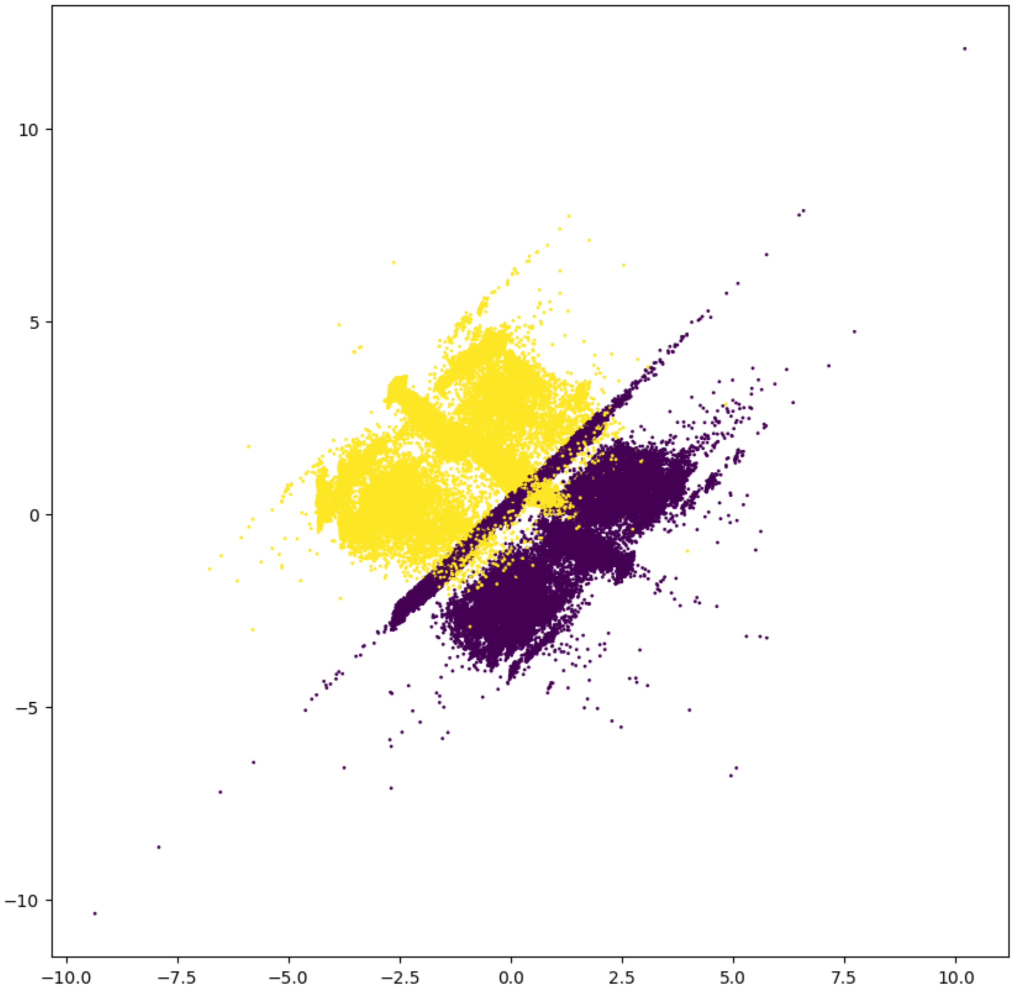


Figure 3.1: Principal component analysis, computed on 2 components. The color of the points corresponds to the value of *involve_killed*.

3.1 K-Means

3.1.1 X-Means

3.2 Hierarchical clustering

3.3 Dbscan clustering