Data Mining report

Funaioli Francesca
Karoui Hamza
Francesco Mitola
Vezzuto Samuele

November 25, 2023

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. We started by importing the three csv files, `incidents.csv`, `povertyByStateYear.csv` and `year_state_district_house.csv`. We did some preliminary checks, for example we looked for attributes containing null values, and then we started analysing each column of each dataset.

1.0.1 Incidents

Date

We converted each *date* value to a datetime object: there where no null values or NaT values. We then plotted the distribution of the dates over the years: we noticed that the dates ranged from year 2013 to year 2018 and from year 2028 to year 2030. Year 2013 only contains 253 incidents so we consider removing these few rows; we also decided to delete incidents recorded in years from 2028 to 2030, because they are to happen after the date the dataset was provided.

State

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.

City or county

The *city_or_county* attribute has no null values; we noticed it contains additional informations about the suburb or the neighborhood in brackets, e.g. "Minneapolis (Brooklyn Center)". We consider removing this informations in the following phases.

Address

The *address* attribute contains some null values, but we believe it does not hold any statistical value, given that a more specific information about the exact location of the incident is found by using the geographical coordinates.

Latitude and longitude

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 we will remove in the next phase.

Congressional, State House and State Senate district

These attributes 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 attributes

The *participant_age1* attribute contains some outliers wrt the corresponding value reported in the *participant_age_group1*. There also some outliers, e.g. maximum value of *participant_age1* is 311. We noticed that 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 some outliers given by non-numerical strings, very large numbers or negative numbers.

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*.

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* 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.

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

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*.

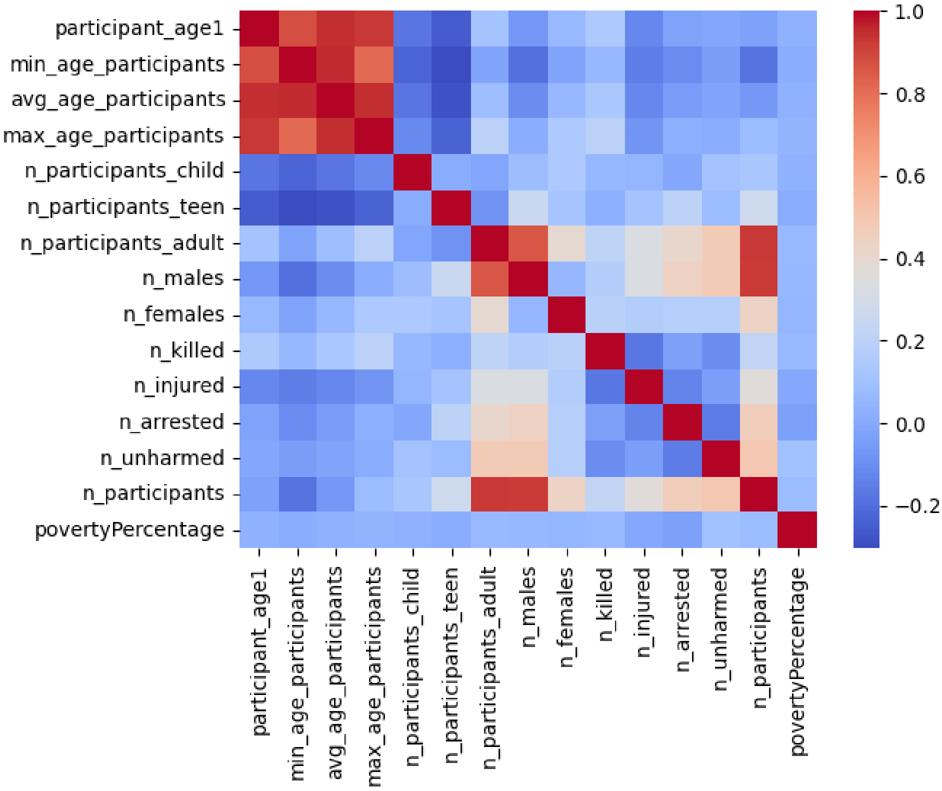

Figure 2.1: Correlation matrix plotted of the numerical attributes

2.2 Definition of indicators

We computed the following indicators:

- *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;

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

- *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.

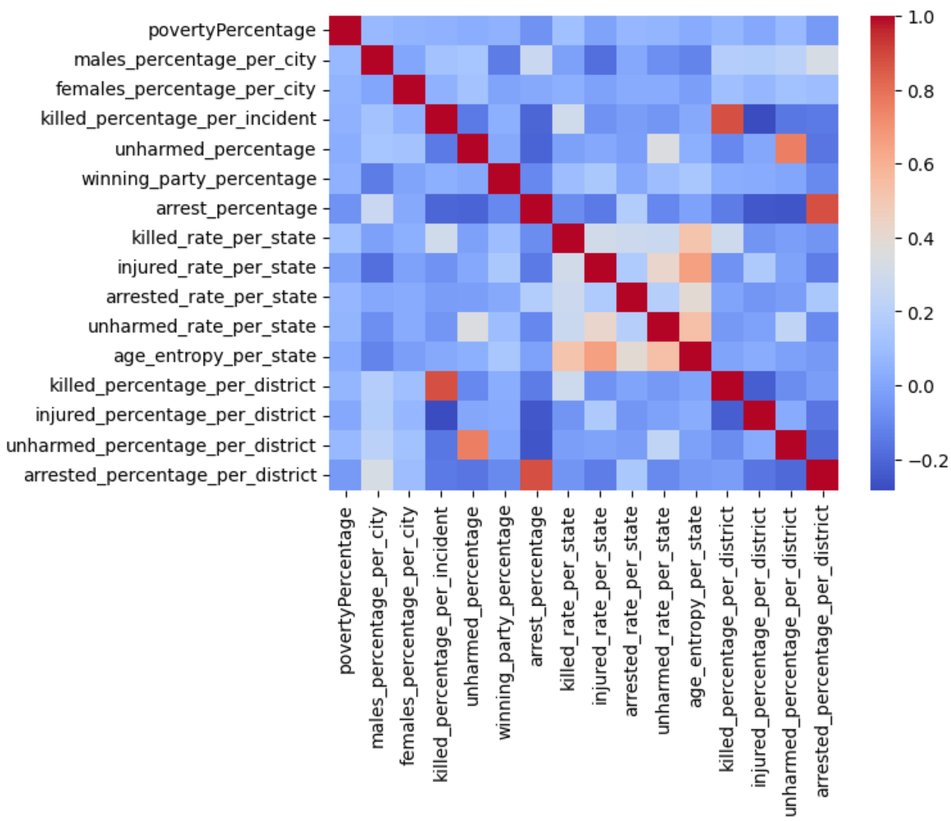

Figure 2.2: Correlation matrix plotted of the newly created indicators

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.

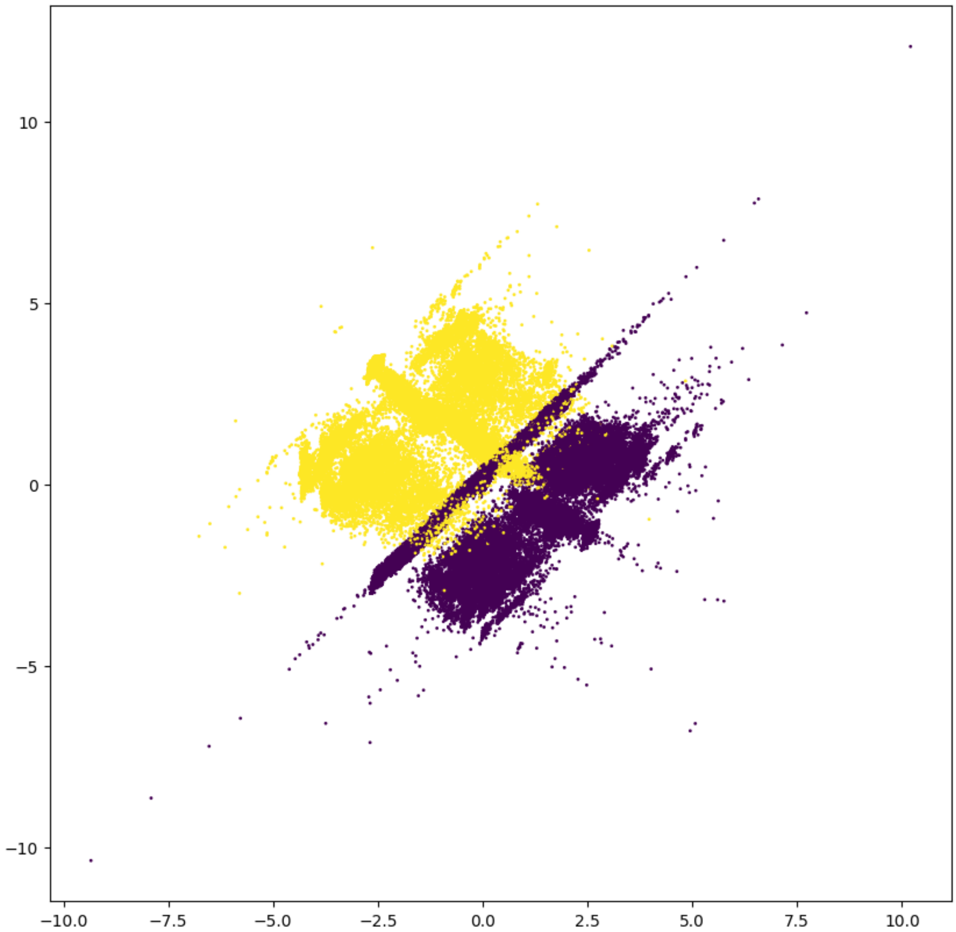

Figure 3.1: Principal component analysis, computed on 2 components. The color of the points corresponds to the value of *involve_killed*.

3.1 K-Means

3.1.1 X-Means

3.2 Hierarchical clustering

3.3 Dbscan clustering