Data Mining Project Report

A.Y. 2023/2024

Funaioli Francesca
Karoui Hamza
Mitola Francesco
Vezzuto Samuele

Contents

Chapter 1

Data Understanding

1.1 Datasets Analysis

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 when the pandas `describe()` method was used. 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.

1.1.1 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.

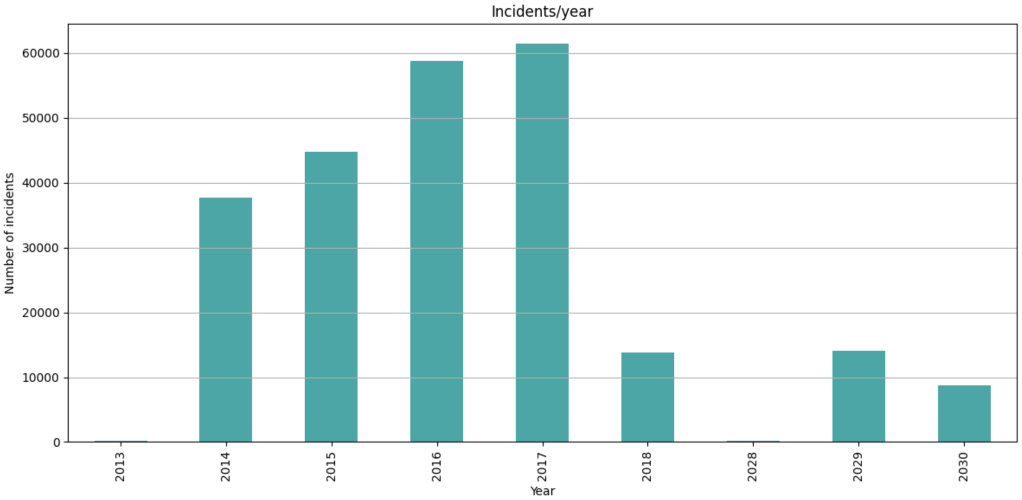


Figure 1.1: Number of incidents per year.

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.

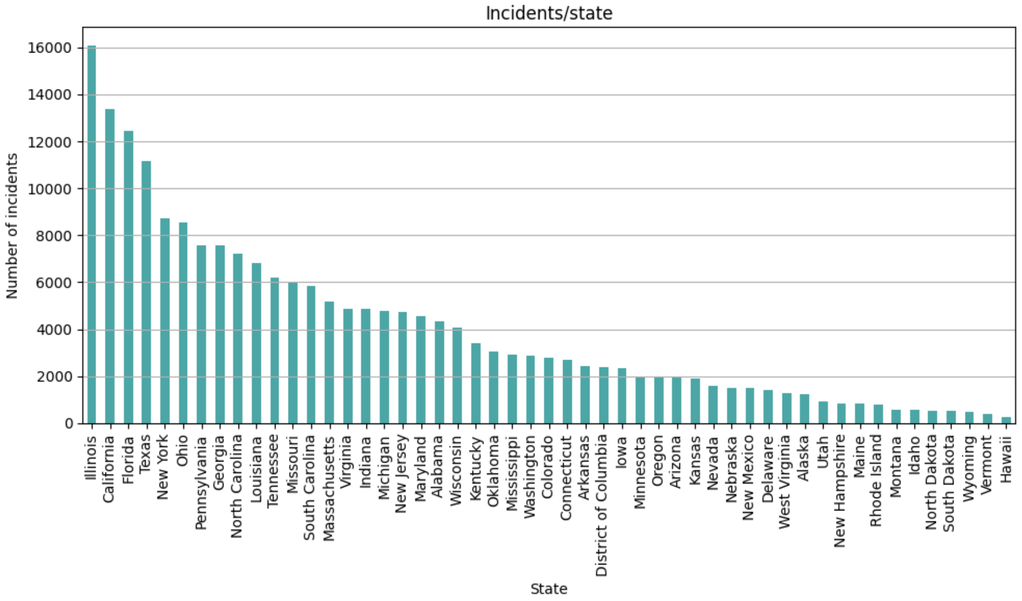


Figure 1.2: Number of incidents per state.

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)". Figure 1.3 shows the number of incidents recorded in each city/county in decreasing order.

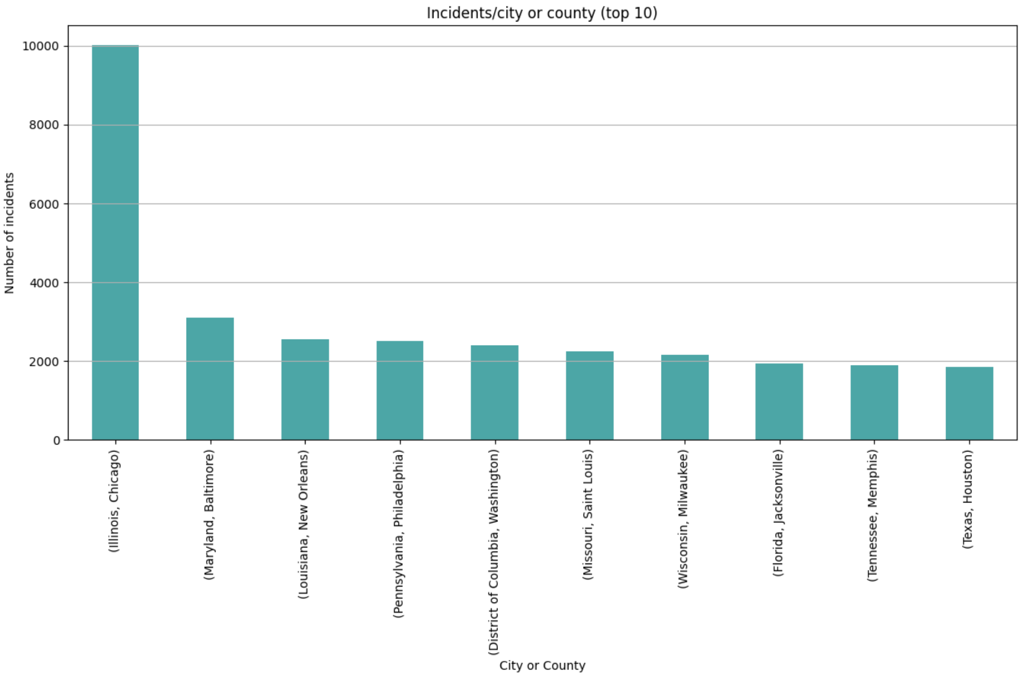


Figure 1.3: Number of incidents per city or county.

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.

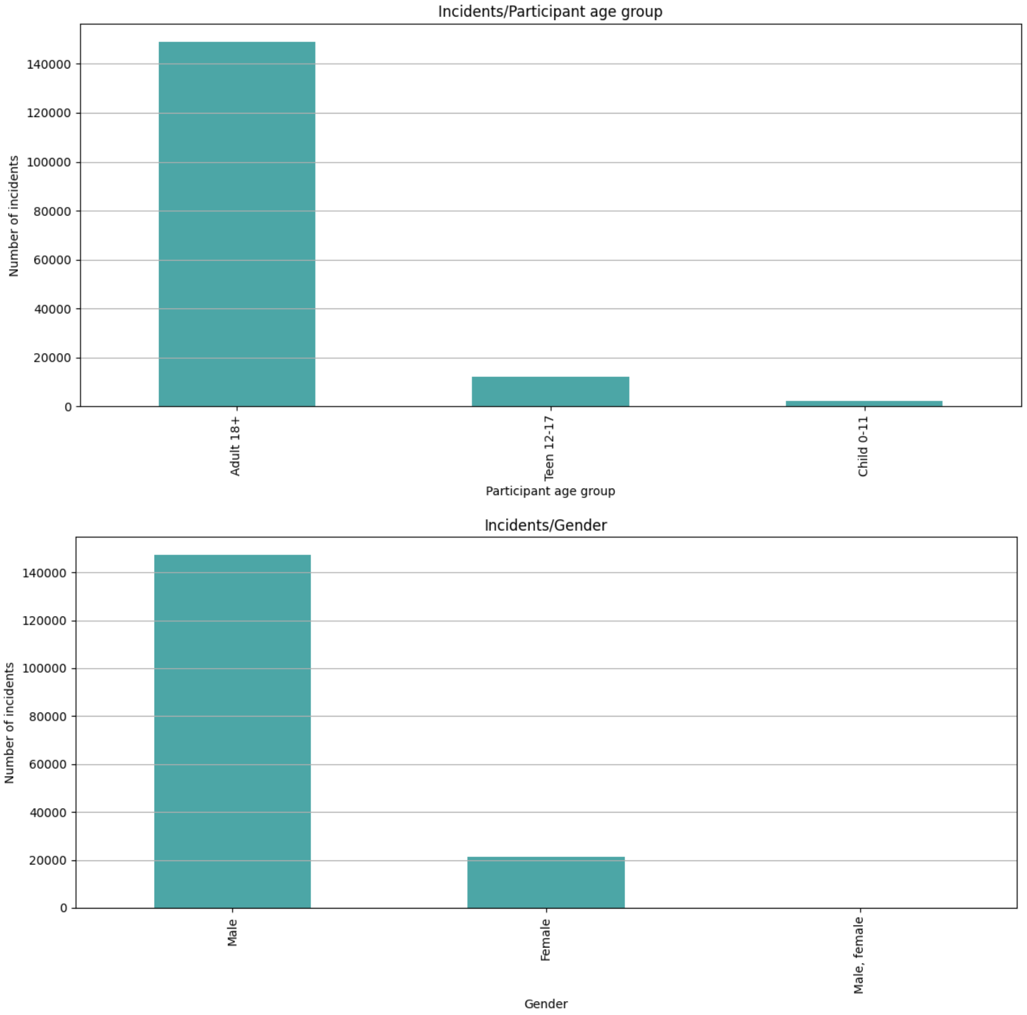


Figure 1.4: Number of incidents per age group (above) and gender (below) of the randomly chosen participant.

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.

1.1.2 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.

1.1.3 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.

1.2 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 will be used to further analyze and relate political party and poverty percentage to each congressional district.

1.3 Distribution Analysis

In order to analyze the features in the dataset, we displayed and examined the distribution of each column. Figure 1.5 shows the plots related to age attributes: as previously said, they 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.

As previously shown in Figure 1.4, the majority of participants recorded in the incidents is comprised of adult (18 years old or older) and male people.

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

We also analyzed the geographical distribution of the incidents over the U.S. territory. Specifically we focused on the number of killed people in each incidents in order to possibly understand the relation between killed people and geographical location. The plots in Figure 1.7 show the individual incidents recorded and the mass shooting events. The incidents that resulted in 0 to 5 killed people are the majority, so the first map does not hold much meaning, given that most of the dots are records that report no killed people. The second plot instead highlights events that are to be considered mass shootings and it shows some major events like, for example, Orlando (Florida) mass shooting in 2016 (50 killed people) and Southerland Springs (Texas) mass shooting (27 killed people).

1.4 Correlation Analysis

The correlation matrix is shown in Figure 1.8. 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.

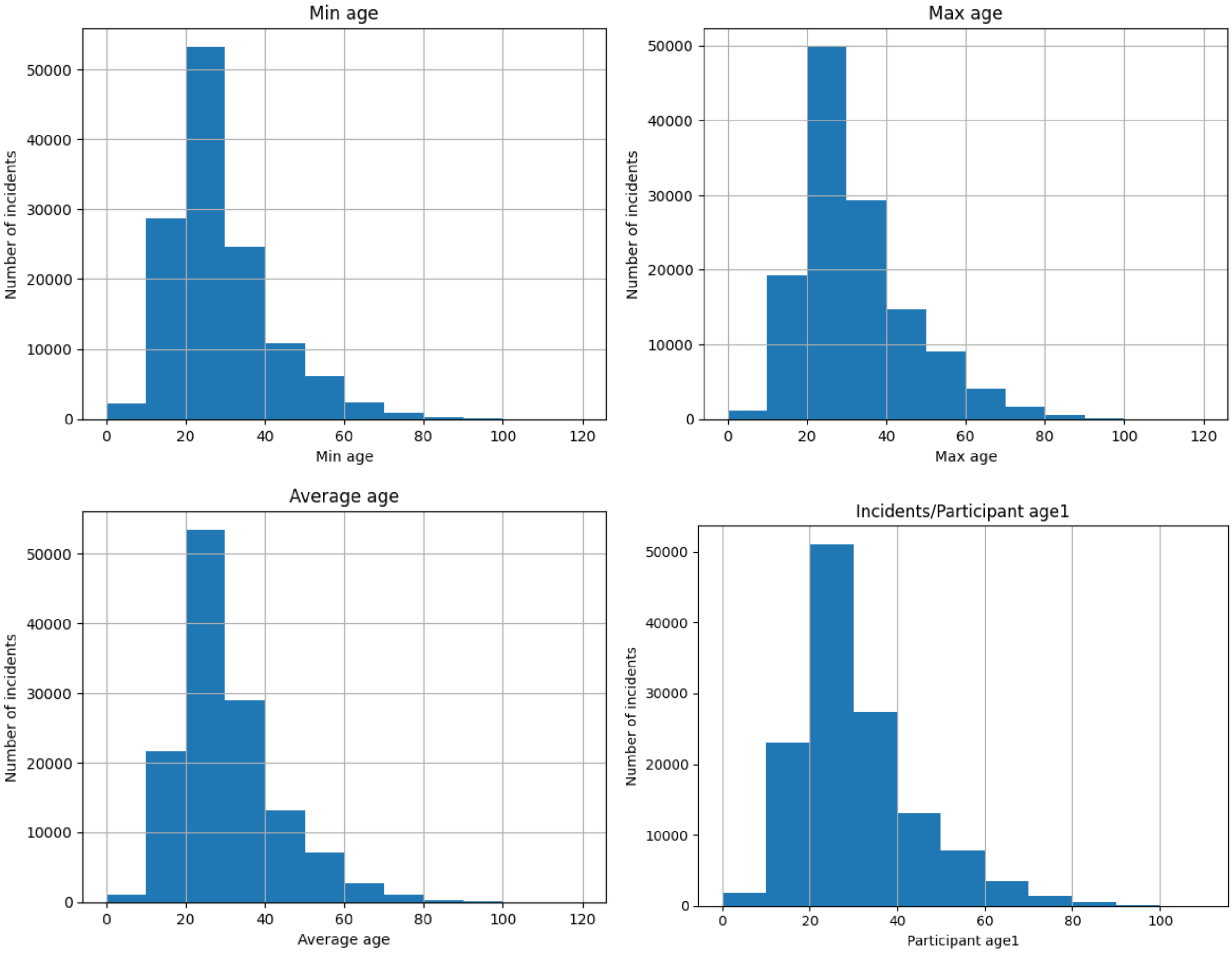


Figure 1.5: Plot of minimum (top left), maximum (top right), average (bottom left) and participant1 (bottom right) age distributions.

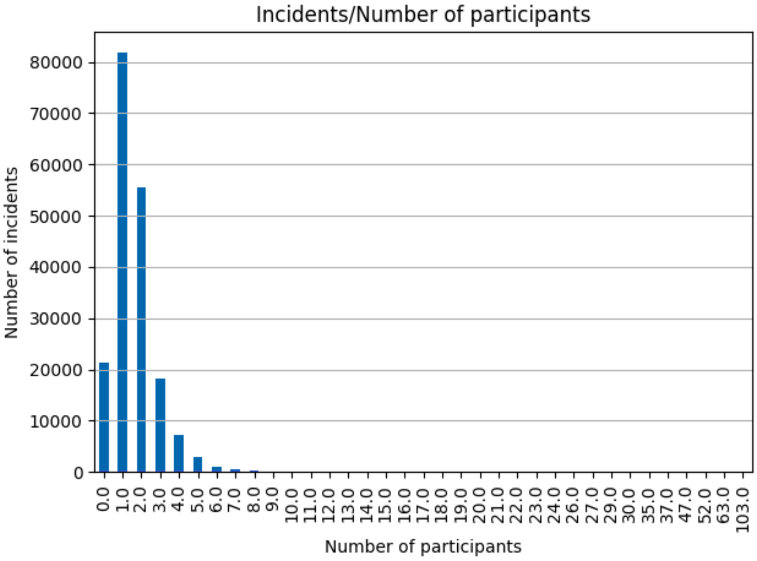


Figure 1.6: Number of incidents per number of people involved.

Figure 1.7: Map representing each incident (left) and mass-shooting incidents (right) recorded, the color used for the representation varies in relation to the number of people killed in the incident. The FBI has not set a minimum number of casualties to qualify an event as a mass shooting, but U.S. statute (the Investigative Assistance for Violent Crimes Act of 2012) defines a “mass killing” as “3 or more killings in a single incident.” (source)

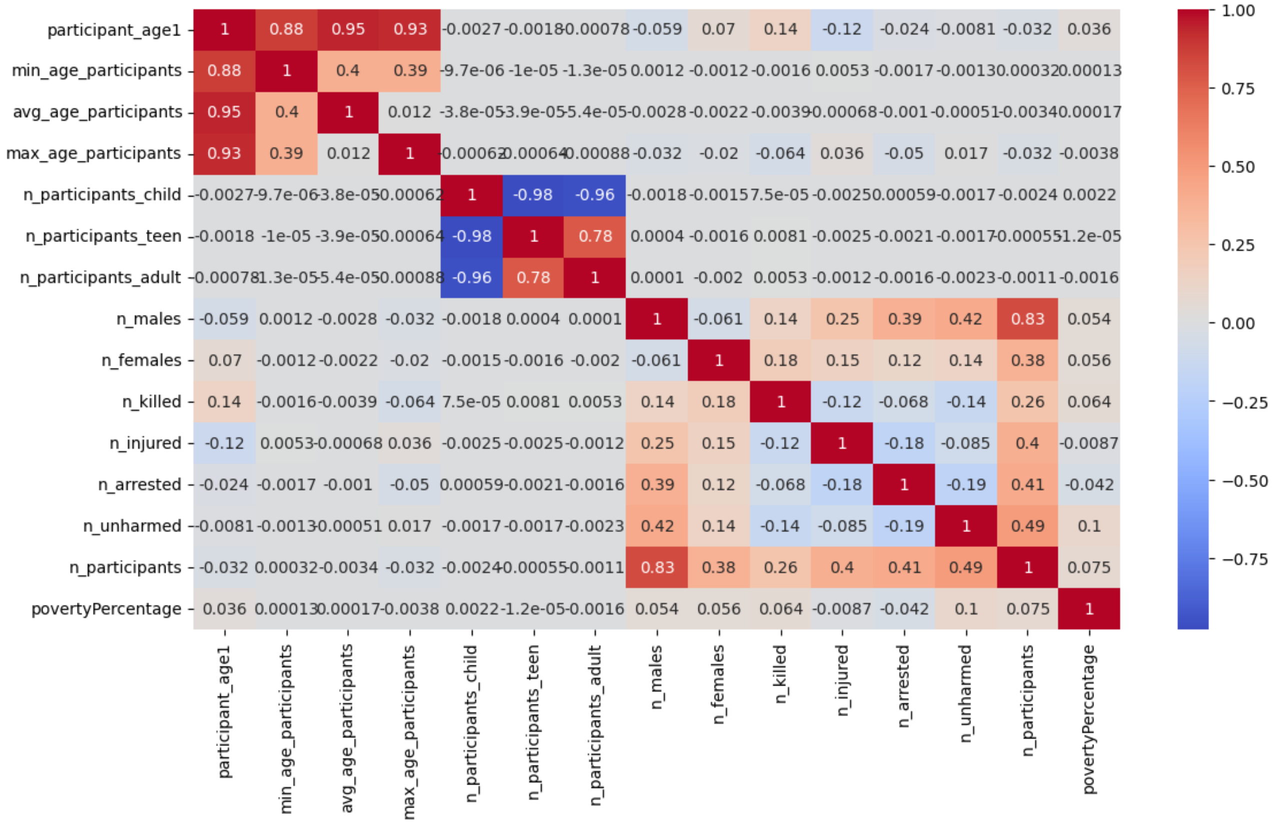


Figure 1.8: Correlation matrix of numerical attributes.

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. Each attribute was casted to the correct data type, in order to be coherent with the semantics of the feature, as shown in Table 2.1.

Duplicate records and records where all null attributes where deleted. For all numerical attributes, negative values were set to NaN.

Date

All records relating to events happened after 2023-10-01 (the date we received the dataset) were considered to be outliers, specifically to be errors in the data, and as such they were dropped. We also dropped all records related to year 2013, as the year was under-represented.

Age attributes

For the attributes *participant_age1*, *min_age_participants*, *max_age_participants* and *avg_age_participants*, values greater than or equal to 120 were considered to be outliers and set to NaN. These three attributes seem to be very correlated, so we consider deleting the columns *min_age_participants* and *max_age_participants* and only keeping *avg_age_participants*. A further motivation for deleting these columns is that *avg_age_participants* can not be inferred by computing the mean of min and max age attributes: when the average age is 0 or null, min and max age are also 0 or null.

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 in a unique way: 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 duplicates. For the rows in which *latitude* and *longitude* are NaN, the missing values were filled using the mean computed for the corresponding *state* and *city_or_county*. This allowed us to replace all missing values for *latitude* and *longitude*.

The columns *state_house_district* and *state_senate_district* were dropped from the dataset, 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* data, because the electoral information provided has the same granularity.

Number of participants attributes

Outliers values found in *n_participants_adult*, *n_participants_teen* and *n_participants_child* features, namely those due to enormous values, were set to NaN. In particular, they were set to NaN when their value exceeded the maximum value found in column *n_participants* (which has no significant large values). Furthermore, in records that had *n_participants* equal to 0, the attributes relating to the number of participants were also set to 0.

In order to keep the dataset coherent, the value in *n_participants* was replaced 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 these sums and/or *n_participants* are not equal. The NaN values were then substituted using the mean of each attribute, grouped by *n_participants*; the few (8) rows for which this replacement was not possible (we were not able to reconstruct the mean) were dropped from the dataset. There were no relevant changes in the distribution of these attributes after the cleaning.

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	year of the incident occurrence
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

We checked whether the number of killed, injured, arrested and unharmed people exceeds the total number of participants in that incident. In case they do, we set them to NaN and then proceeded to fill null values using the mean. There was no major change in the distribution of these feature after this operation.

Incident characteristics

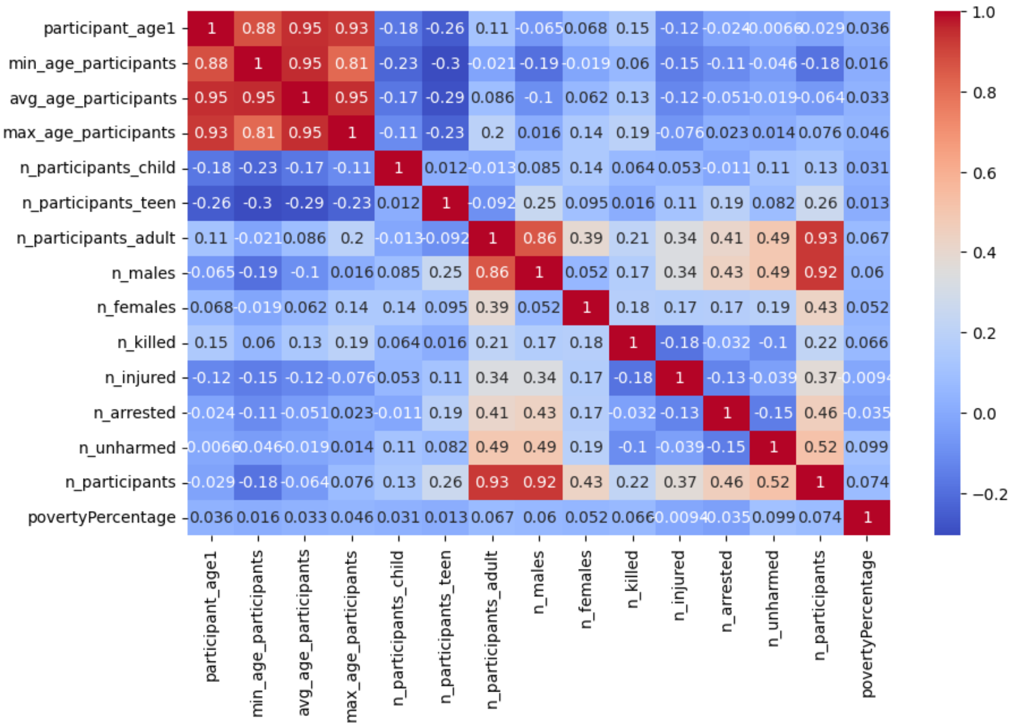
In order to fill the missing values for feature *incident_characteristics1*, records were analyzed by looking at data from *n_killed*, *n_injured* and *notes* attributes. This was done to create subgroups of records, so that incidents characteristics could be reconstructed if missing. For records that did not fit in any of these subgroups, *incident_characteristics1* was filled using the string “Shots Fired - No Injuries”, as they reported no injured or killed people. 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

The correlation matrix, shown in Figure 2.1, has been computed only for numerical attributes. It can be seen that age attributes are highly correlated. Because of this, we decided to drop all of them[1] except for *avg_age_participants*, which is the most correlated (> 90%) to the other attributes and gives us more general informations about all the participants. Given the fact that *participant_age1* was dropped, the attributes *participant_gender1* and *participant_age_group1* become useless because there no longer is any information about the randomly selected person: they were also dropped. The only remaining feature related to the participants age is *avg_age_participants*, which contains some null values. The missing values filled by the mean, by considering the number of adults, teens and children involved in the incident.

It can also be noticed that *n_participants_adult* and *n_males* are highly correlated. Furthermore, they can both be obtained by mean of sums of other attributes[2], so they were dropped.

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



	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	-0.18	-0.26	0.11	-0.065	0.068	0.15	-0.12	-0.024	-0.0066	-0.029	0.036
min_age_participants	0.88	1	0.95	0.81	-0.23	-0.3	-0.021	-0.19	-0.019	0.06	-0.15	-0.11	-0.046	-0.18	0.016
avg_age_participants	0.95	0.95	1	0.95	-0.17	-0.29	0.086	-0.1	0.062	0.13	-0.12	-0.051	-0.019	-0.064	0.033
max_age_participants	0.93	0.81	0.95	1	-0.11	-0.23	0.2	0.016	0.14	0.19	-0.076	0.023	0.014	0.076	0.046
n_participants_child	-0.18	-0.23	-0.17	-0.11	1	0.012	-0.013	0.085	0.14	0.064	0.053	-0.011	0.11	0.13	0.031
n_participants_teen	-0.26	-0.3	-0.29	-0.23	0.012	1	-0.092	0.25	0.095	0.016	0.11	0.19	0.082	0.26	0.013
n_participants_adult	0.11	-0.021	0.086	0.2	-0.013	-0.092	1	0.86	0.39	0.21	0.34	0.41	0.49	0.93	0.067
n_males	-0.065	-0.19	-0.1	0.016	0.085	0.25	0.86	1	0.052	0.17	0.34	0.43	0.49	0.92	0.06
n_females	0.068	-0.019	0.062	0.14	0.14	0.095	0.39	0.052	1	0.18	0.17	0.17	0.19	0.43	0.052
n_killed	0.15	0.06	0.13	0.19	0.064	0.016	0.21	0.17	0.18	1	-0.18	-0.032	-0.1	0.22	0.066
n_injured	-0.12	-0.15	-0.12	-0.076	0.053	0.11	0.34	0.34	0.17	-0.18	1	-0.13	-0.039	0.37	-0.0094
n_arrested	-0.024	-0.11	-0.051	0.023	-0.011	0.19	0.41	0.43	0.17	-0.032	-0.13	1	-0.15	0.46	-0.035
n_unharmed	-0.0066	-0.046	-0.019	0.014	0.11	0.082	0.49	0.49	0.19	-0.1	-0.039	-0.15	1	0.52	0.099
n_participants	-0.029	-0.18	-0.064	0.076	0.13	0.26	0.93	0.92	0.43	0.22	0.37	0.46	0.52	1	0.074
povertyPercentage	0.036	0.016	0.033	0.046	0.031	0.013	0.067	0.06	0.052	0.066	-0.0094	-0.035	0.099	0.074	1

Figure 2.1: Correlation matrix of numerical attributes.

2.2 Definition of indicators

After analyzing the data, we considered the following indicators as being representative of an incident severity:

- *incident_gravity*, number of killed and injured people in an incident,
- *females_rate*, number female people involved in the incident over the total number of participants in that incident,
- *minor_rate*, number minors (younger than 17 years old) involved in the incident over the total number of participants,
- *arrested_rate*, number arrested people in the incident over the total number of participants,
- *survival_rate*, number unharmed people in the incident over the total number of participants,
- *injured_rate*, number injured people in the incident over the total number of participants,

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

[2] In particular, $n\_participants = n\_participants\_adult + n\_participants\_teen + n\_participants\_child$, but also $n\_participants = n\_males + n\_females$. All attributes relating to the number of participants were dropped, with the excepytion of *n_participants*, which conveys the most general information.

- *killed_rate*, number killed people in the incident over the total number of participants,
- *winning_party_percentage*, the number of votes for the winning candidate over the total number of votes in that election.
- *killed_disp_per_district*, number of killed people in the incident over the number of killed people in that same congressional district in the same year,
- *injured_disp_per_district*, number of injured people in the incident over the number of injured people in that same congressional district in the same year,
- *part_disp_per_district*, number of participants in the incident over the number of participants in incidents in that same congressional district in the same year.

The correlation matrix of the chosen indicators is the one in Figure 2.2. Indicators *female_ratio* and *minor_ratio* both have similar behaviors and distributions. For most of the incidents, the rates of survival, arrested, injured and killed people are equal to 0. A lot of incidents involving injuries, arrests or survivors do not involve deaths. Similarly, incidents involving injuries, arrests or deaths do not involve any survivors. Indicators *arrested_rate* and *injured_rate* have a similar behavior. This confirms that these indicators are uncorrelated.

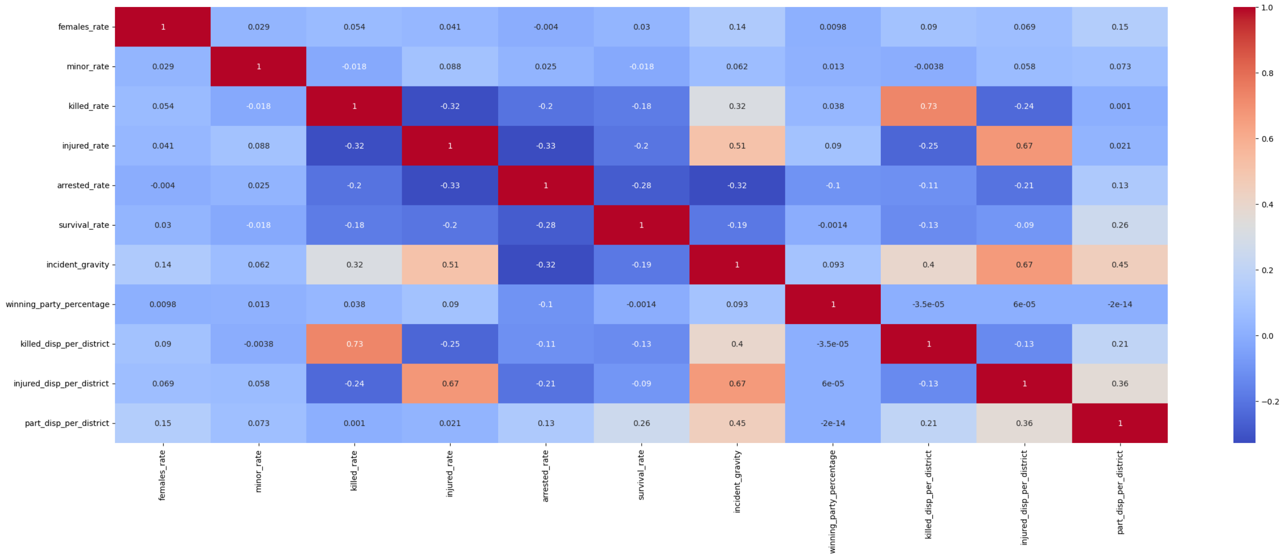


Figure 2.2: Correlation matrix plotted of the newly created indicators

The last three indicators often have value 0, which is also the value of the mean. Over all, the indicators we identified have no missing values; they have some outliers that are coherent with the one in the numerical attributes, so we decided not to clean them. All these indicators provide a good characterization of the incidents and are at most moderately correlated.

Chapter 3

Clustering

3.1 Preprocessing

In order to apply clustering methods, data needs to be prepared through some steps of preprocessing. First of all some new binary columns were added and used to label the PCA results:

- *involve_killing*, its value is 0 if nobody was killed in the incident, it is 1 if at least a person was killed;
- *involve_injury*, its value is 0 if nobody was injured in the incident, it is 1 if at least a person was injured;
- *involve_arrest*, its value is 0 if nobody was arrested in the incident, it is 1 if at least a person was arrested;
- *is_survived*, its value is 0 if nobody survived in the incident, it is 1 if at least a person survived.

The new feature *involve_killing* will also be used in the following classification task. Given that for clustering we will use a distance metric, that only works on numerical attributes, we remove categorical attributes for this task and we only use the indicators listed in the previous section. The features are also normalized: scaling data is useful to avoid some values being too large and prevailing on the others when a clustering algorithm is applied, so we computed a normalization of the numerical values.

The Principal Component Analysis is computed using two components. It yields a 2-dimensional visualization of the multidimensional data considered, shown in Figure 3.1.

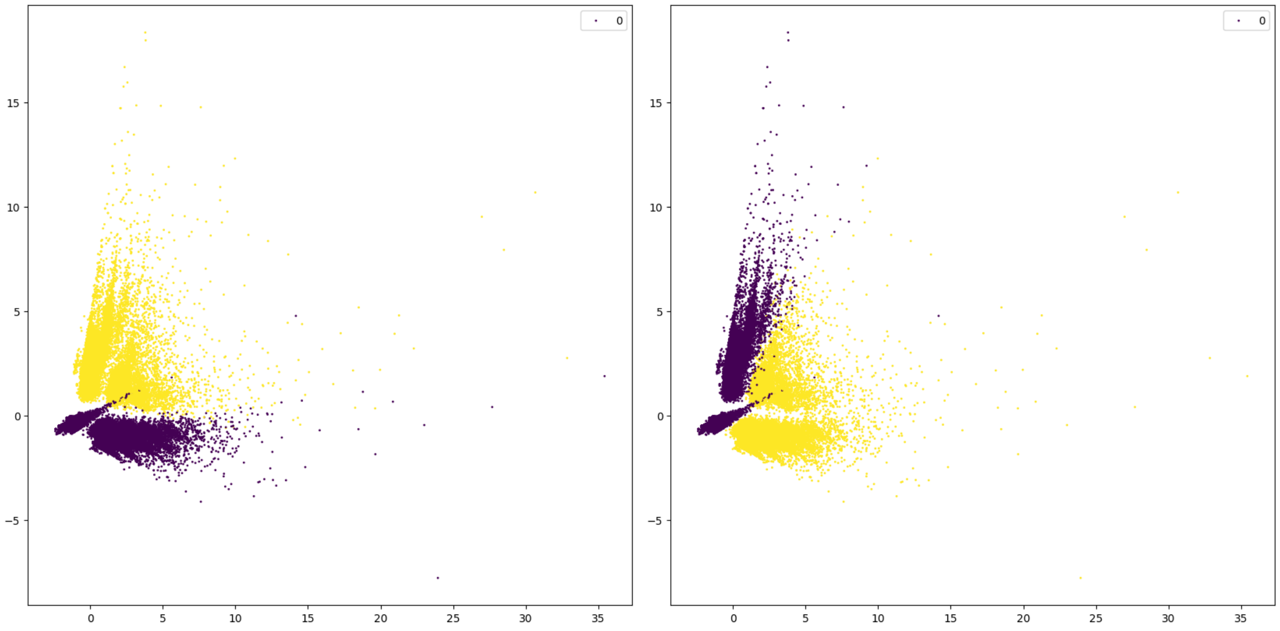


Figure 3.1: PCA, with number of components equal to 2. The color of the points corresponds to the value of *involve_killing* in the left plot and to the value of *involve_injury* in the right plot.

3.2 K-Means

In order to estimate the best value of k to then apply the K-means clustering algorithm, the elbow method was used. The grid search on values of $k \in \{2, 3, 4, 10, 100, 100\}$ yielded the following results, evaluated in terms of the metrics SSE, separation and silhouette. A plot of the SSE and silhouette scores is shown in Figure 3.2:

K	**SSE**	**separation**	**silhouette**
2	1868436.1752223005	1.9881698493281088	0.1938016918423718
3	1562300.2158580532	1.6287560855244605	0.24111879379810464
4	1401333.0002367594	1.435772151754466	0.24457341442342728
10	935031.5401451141	1.4857417427196573	0.27681666821740364
100	315602.0491054469	1.3057093979420389	0.26549526914793403
1000	96365.98745985169	1.2338707526266535	0.2938211523306092

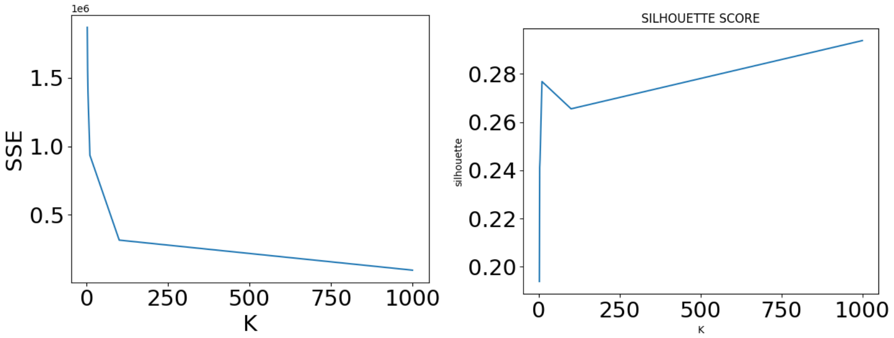


Figure 3.2: Plots of SSE (left) and the silhouette (right) scores.

Both the elbow method and the silhouette score point out that the best value for k is around 3. After fitting the method to our data, the clustering results were plotted on the 2 PCA dimensions, label the points according to the belonging cluster. The plot is shown in Figure 3.3, along with the cluster label distribution.

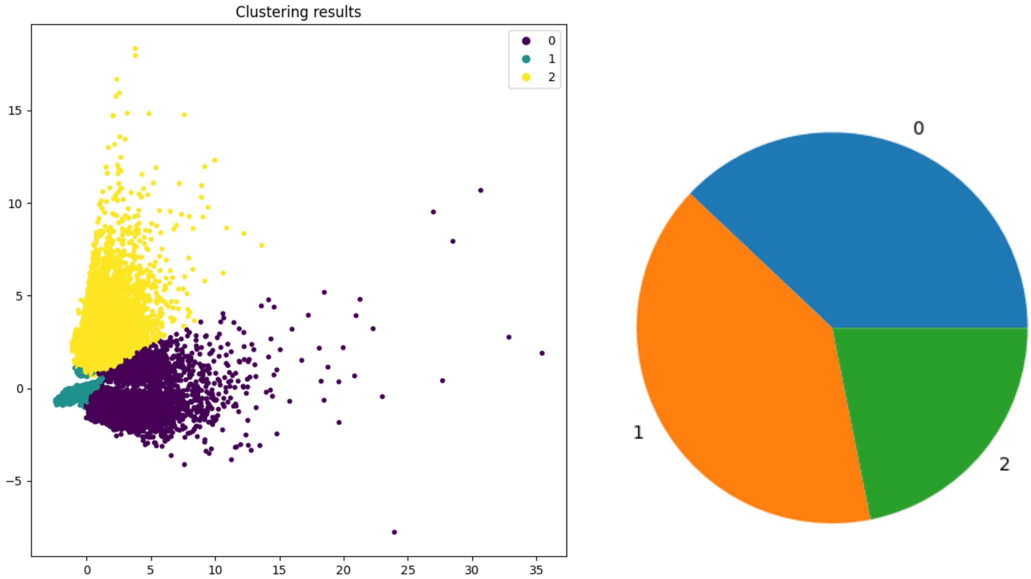


Figure 3.3: Plots of the PCA components after the clustering labeling (left) and the distribution of the cluster labels (right).

Figure 3.4 shows the distribution of values of feature *involve_killing* in the three clusters found. The first two clusters contain a majority of non fatal incidents, while the third cluster only contains incidents involving at least a person killed.

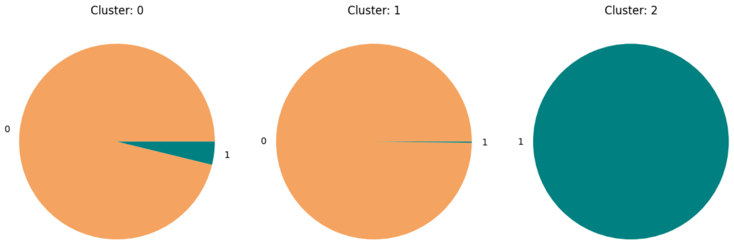


Figure 3.4: Plots of the three clusters found by applying K-means, labeled according to the value of *involve_killing*. Cluster 0 has 68459 incidents for which *involve_killing* is 0, 2714 incidents for which it is 1; Cluster 1 has 75225 incidents for which *involve_killing* is 0, 155 incidents for which it is 1; Cluster 2 contains 40981 incidents for which *involve_killing* is 1.

3.2.1 X-Means

The X-means implementation used is *pyclustering*[1]. The starting parameter for X-means was a maximum number of clusters of 10, repeating 10 times the execution of K-means. Computing the clusters using the X-mean algorithm, the result was composed of 10 clusters. The metrics of silhouette and separation obtained from these clusters were 1.4333188536330668 and 0.2239619283502909 respectively. The distribution of values of feature *involve_killing* per cluster and the distribution of the cluster labels is shown in Figure 3.7. Specifically, the distribution of the cluster labels is uneven.

TODO: *inserire SSE!!*

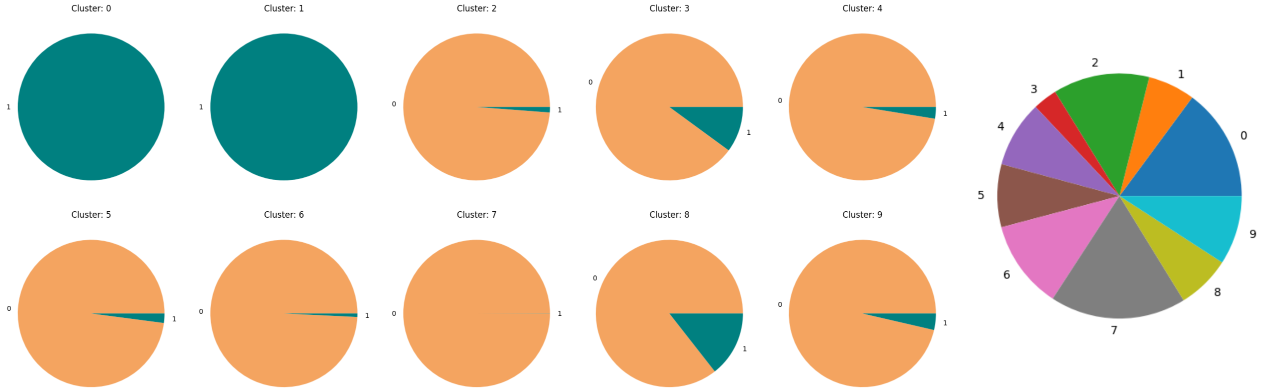


Figure 3.5: Plots of the ten clusters found applying X-means (left), labeled according to the value of *involve_killing*, and the plot of the distribution of the cluster labels (right).

3.3 Dbscan clustering

As per the classification subtask in the project description, only one state was chosen for this clustering task: the state of Illinois, that is the first state per number of gun incidents (15319). Density-based clustering algorithms are used to identify irregularly shaped clusters. In order to estimate the best value for the epsilon (ε) parameter, the elbow method was used: computing the distance to the k nearest neighbors and sorting them in decreasing order resulted in the plot shown in Figure 3.6. The best estimation found for the value of ε is around 1.5. Different values of ε were then explored along with varying values of *min_samples*, resulting in Table 3.1, shown in decreasing order of silhouette score. Therefore, the best values of ε and *min_samples* are 1.6 and 160 respectively.

Applying Dbscan to the dataset using the parameters shown above lead to find 4 clusters. Figure 3.6 also shows the scatter plot of these results on the 2 PCA components defined before, along with the distribution of

[1] Pyclustering: `https://github.com/annoviko/pyclustering/`

the cluster labels. The plots show that all incidents in which there was at least a person killed belong to Cluster 2, while Clusters 0, 1 and 3 only feature non-deadly incidents.

ε	**silhouette**	**DBscore**	**n_clusters**	**min_samples**
1.6	0.202494	001814	5	160
1.6	0.199733	004573	5	170
1.6	0.197965	998093	5	180
1.7	0.180224	643053	3	180
1.7	0.086065	570127	4	160
1.7	0.073308	567566	4	170

Table 3.1: Table showing the search for the best values of ε and *min_samples* parameters.

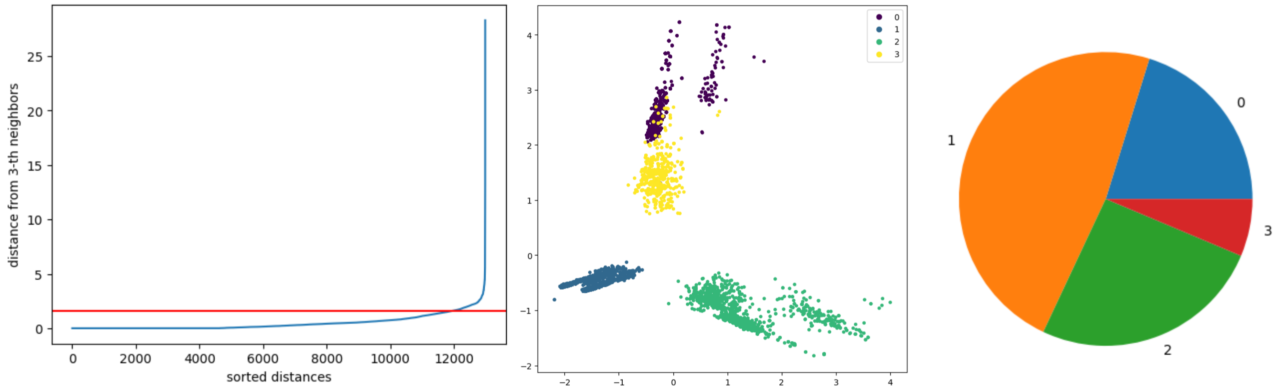


Figure 3.6: Plots of the elbow method result (left), the PCA components after the clutering labeling (middle) and the distribution of the cluster labels (right).

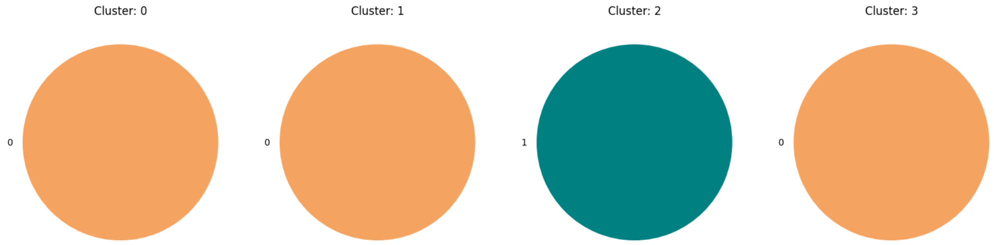


Figure 3.7: Plots of the four clusters found applying Dbscan, labeled according to the value of *involve_killing*.

3.4 Hierarchical clustering

As per the classification subtask in the project description, only one state was chosen for this clustering task: the state of California, that is the second state per number of gun incidents (12980). In order to identify the best metric for distance and the best method, we evaluated different trials. The best results were obtained by using the correlation distance metric[2], the weighted linkage method[3]. By using this final model, the silhouette score is 0.2589355057625014 and the DBscore 1.4509696821004128. The dendrogram displayed in Figure 3.8 was truncated at level 5, so that it can be better observed. The cluster label distribution is not even: the cluster sizes are, in order, 772, 5018, 2697, 513, 3980.

After applying all the methods shown before and analyzing the results obtained, we chose Dbscan as the best performing clustering algorithm, both because of the cluster distribution and the distribution of records reporting at least one dead person in the incident.

[2] Correlation distance between vectors u and v: $1 - \frac{(u-\bar{u})\cdot(v-\bar{v})}{\|(u-\bar{u})\|_2 \|(v-\bar{v})\|_2}$, where $\bar{v}$ is the mean of the elements of vector v.

[3] The weighted method assigns $d(u, v) = \frac{dist(s,v)+dist(t,v)}{2}$, where cluster u was formed with cluster s and t and v is a remaining cluster in the forest.

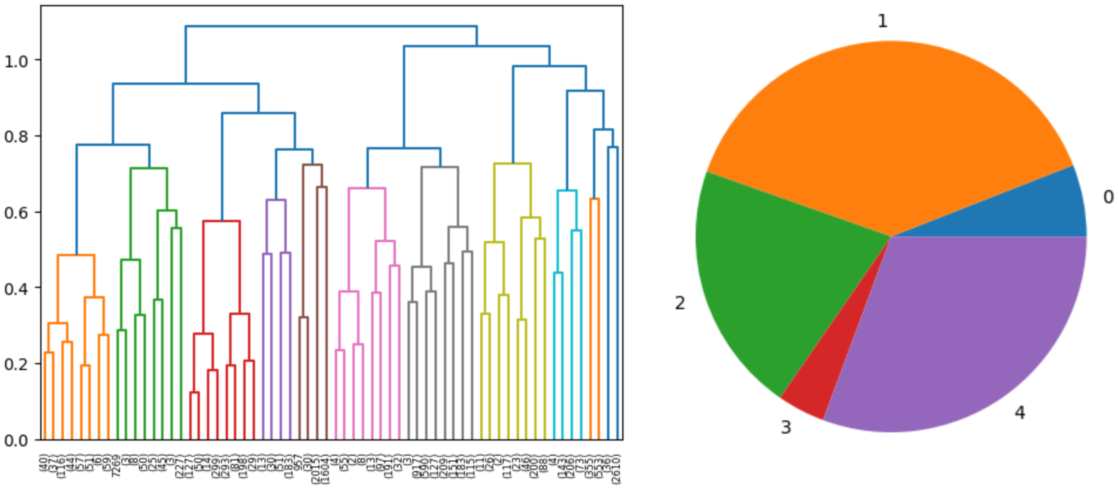


Figure 3.8: Plots of the dendrogram obtained by applying Dbscan (left) and the distribution of the cluster labels (right).

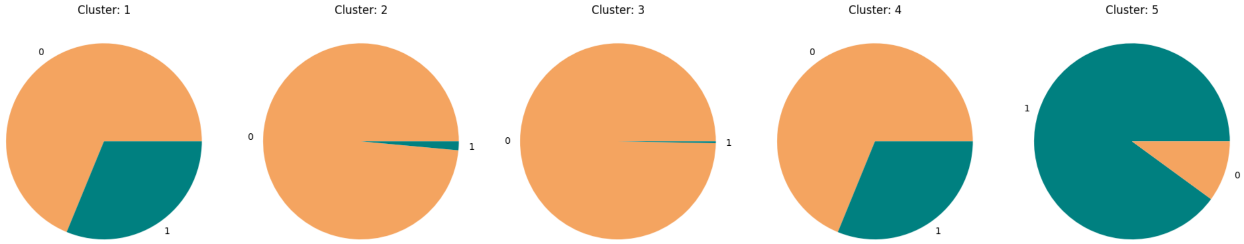


Figure 3.9: Plots of the five clusters found applying agglomerative hierarchical clustering, labeled according to the value of *involve_killing*.

Chapter 4

Classification

Chapter 5

Explanation Analysis