DATA MINING

REPORT 2: Project development

Team 2 from Grup 11:

Agulló López, Ferran
Alarcón Julián, Samuel
Esquina Muñoz, Daniel
González López, Rubén
Tormos Llorente, Adrián
Volkova Volkova, Alejandra

Date: 29/09/2019

Index

Introduction

In this project, we will be studying a curated sample of a dataset that collects information about the known terrorist incidents that have taken place in the USA in the last ~50 years, with detailed information and variables for each one. All these incidents meet different criteria to be considered as terrorism, with some of them labeled as unclear, but since we didn't see a lot of discrepancy we deemed unnecessary to look further into it and simply considered all the incidents as terrorism, without differences in the criteria that considers them as such.

The reason we chose this topic for the project was that, between all the different datasets we found, this was one of the more complete we had, with enough information, cases and variables to study, and also was interesting enough to extract some conclusions unlike other topics that were more bland or simple, or included concepts we didn't understand and therefore we would have probably ended up finding it too difficult to develop meaningful conclusions with our analysis' results.

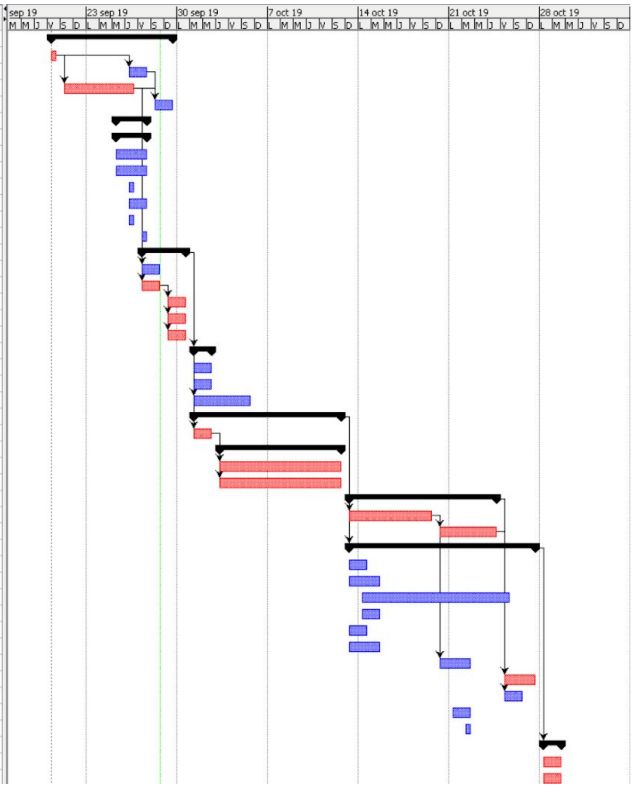
Initial working plan

We've made an initial working plan taking a look at the guidelines and steps for the project. A lot of concepts are unknown to us since they will be introduced in future lessons, so we've tried our best to balance the time span for each one of them in our Gantt chart, and also made sure not to assign a task's initial date before the day on which the according lesson is expected to happen. It's very likely that the real dates won't match with this initial working plan, taking into account what's already mentioned and additional factors that may mess up our progress, like projects from other subjects, as we explain later in the Risk contingency plan section of this report.

Project development timeline with Gantt

To make the Gantt chart we used the free program ProjectLibre, which allowed us to create tasks, assign an initial and final date to them, organize them with subtasks, and set dependencies. Hereunder we've attached an image of the final result, but the complete files (in two different formats, .pod and .xml) are also included in the delivery, in order to look at them in more detail.

	Nombre	Duracion	Inicio	Terminado
1	**Data preparation**	10 days?	**20/09/19 8:00**	**29/09/19 17:00**
2	Data selection	1 day?	20/09/19 8:00	20/09/19 17:00
3	Data merge	2 days?	26/09/19 8:00	27/09/19 17:00
4	Delete excessive or useless information	6 days?	21/09/19 8:00	26/09/19 17:00
5	Basic initial univariate descriptive statistics c	2 days?	28/09/19 8:00	29/09/19 17:00
6	**Writing the 2nd Deliverable**	3 days?	**25/09/19 8:00**	**27/09/19 17:00**
7	**Initial working plan**	3 days?	**25/09/19 8:00**	**27/09/19 17:00**
8	Gantt	3 days?	25/09/19 8:00	27/09/19 17:00
9	Division of tasks	3 days?	25/09/19 8:00	27/09/19 17:00
10	Risk contingency plan	1 day?	26/09/19 8:00	26/09/19 17:00
11	Metadata file	2 days?	26/09/19 8:00	27/09/19 17:00
12	Enumerate steps of the preprocessing proc	1 day?	26/09/19 8:00	26/09/19 17:00
13	List of decisions for each preprocessing step	1 day?	27/09/19 8:00	27/09/19 17:00
14	**Preprocessing**	4 days?	**27/09/19 8:00**	**30/09/19 17:00**
15	Variables declaration in pre-processing tool	2 days?	27/09/19 8:00	28/09/19 17:00
16	Filtering	2 days?	27/09/19 8:00	28/09/19 17:00
17	Outlier detection and correction	2 days?	29/09/19 8:00	30/09/19 17:00
18	Error detection and correction	2 days?	29/09/19 8:00	30/09/19 17:00
19	Missing data treatment	2 days?	29/09/19 8:00	30/09/19 17:00
20	**Statistical descriptive analysis**	2 days	**1/10/19 8:00**	**2/10/19 17:00**
21	Univariate	2 days	1/10/19 8:00	2/10/19 17:00
22	Bivariate	2 days	1/10/19 8:00	2/10/19 17:00
23	Visualization of terrorist attacks on a USA map	5 days?	1/10/19 8:00	5/10/19 17:00
24	**PCA**	12 days?	**1/10/19 8:00**	**12/10/19 17:00**
25	Revise preprocessing	2 days?	1/10/19 8:00	2/10/19 17:00
26	**PCA analysis for numerical variables**	10 days?	**3/10/19 8:00**	**12/10/19 17:00**
27	Scree plot	10 days?	3/10/19 8:00	12/10/19 17:00
28	Factorial map visualization	10 days?	3/10/19 8:00	12/10/19 17:00
29	**Clustering**	12 days	**13/10/19 8:00**	**24/10/19 17:00**
30	Hierarchical Clustering on original data	7 days	13/10/19 8:00	19/10/19 17:00
31	Profiling of clusters	5 days	20/10/19 8:00	24/10/19 17:00
32	**Writing the final Deliverable**	15 days?	**13/10/19 8:00**	**27/10/19 17:00**
33	Write Motivation of the work and general de	2 days	13/10/19 8:00	14/10/19 17:00
34	Write Formal description of Data structure a	3 days?	13/10/19 8:00	15/10/19 17:00
35	Explain the complete Data Mining process pi	12 days?	14/10/19 8:00	25/10/19 17:00
36	Write Detailed description of Preprocessing	2 days?	14/10/19 8:00	15/10/19 17:00
37	Writing Basic statistical descriptive analysis	2 days	13/10/19 8:00	14/10/19 17:00
38	Writing PCA	3 days	13/10/19 8:00	15/10/19 17:00
39	Writing Hierarchical Clustering on original da	3 days	20/10/19 8:00	22/10/19 17:00
40	Writing Profiling of clusters	3 days	25/10/19 8:00	27/10/19 17:00
41	Write Global discussion and general conclusi	2 days	25/10/19 8:00	26/10/19 17:00
42	Write Working plan	2 days	21/10/19 8:00	22/10/19 17:00
43	Finishing up the document	1 day	22/10/19 8:00	22/10/19 17:00
44	**Preparing the oral presentation**	2 days?	**28/10/19 8:00**	**29/10/19 17:00**
45	Creating the slides	2 days?	28/10/19 8:00	29/10/19 17:00
46	Rehearsing	2 days	28/10/19 8:00	29/10/19 17:00



Assignment of tasks

As the guidelines propose, each one of the tasks has at least to team members working on it. Additionally, each task has a coordinator or supervisor, which is represented in the following tables by an uppercase and bold **X**.

- Data preparation

Task	Ferran	Samuel	Daniel	Rubén	Adrián	Alejandra
Data selection	x	x	x	x	**X**	x
Data merge	**X**				x	
Delete excessive or useless information	x		**X**	x		
Basic initial univariate descriptive statistics of raw variables	x			x	**X**	x

- Writing the 2nd Deliverable

Task	Ferran	Samuel	Daniel	Rubén	Adrián	Alejandra
Gantt	x	**X**			x	x
Division of tasks	x	x	x	x	x	**X**
Risk contingency plan		**X**			x	x
Metadata file			x	**X**	x	
Enumerate steps of the preprocessing process that we use		x				**X**
List of decisions for each preprocessing step		x	**X**	x		x

- Preprocessing

Task	Ferran	Samuel	Daniel	Rubén	Adrián	Alejandra
Variables declaration in pre-processing tool	**X**		x	x		

Filtering	**X**		x	x		
Outlier detection and correction	**X**		x	x	x	x
Error detection and correction	**X**		x	x	x	x
Missing data treatment	**X**		x	x	x	

- Statistical descriptive analysis

Task	**Ferran**	**Samuel**	**Daniel**	**Rubén**	**Adrián**	**Alejandra**
Univariate			x	**X**	x	
Bivariate			**X**			x

- Visualization of terrorist attacks on a USA map

Task	**Ferran**	**Samuel**	**Daniel**	**Rubén**	**Adrián**	**Alejandra**
Visualization of terrorist attacks on a USA map	x	**X**				

- PCA

Task	**Ferran**	**Samuel**	**Daniel**	**Rubén**	**Adrián**	**Alejandra**
Revise preprocessing		x		**X**		
Scree plot	**X**	x				
Factorial map visualization				**X**	x	

- Clustering

Task	**Ferran**	**Samuel**	**Daniel**	**Rubén**	**Adrián**	**Alejandra**
Hierarchical Clustering on original data		x			**X**	x
Profiling of clusters			**X**	x		

- Writing the final Deliverable

Task	Ferran	Samuel	Daniel	Rubén	Adrián	Alejandra
Write Motivation of the work and general description of the problem to be analyzed		**X**	x			
Write Formal description of Data structure and metadata		x		**X**		
Explain the complete Data Mining process performed		x			x	**X**
Write Detailed description of Preprocessing and data preparation.	x	**X**				
Writing Basic statistical descriptive analysis			x	**X**	x	
Writing PCA		x			**X**	
Writing Hierarchical Clustering on original data	**X**	x				
Writing Profiling of clusters			x	**X**		
Write Global discussion and general conclusions of the whole work			**X**			x
Write Working plan		**X**			x	
Finishing up the document		x				**X**

- Preparing the oral presentation

Task	Ferran	Samuel	Daniel	Rubén	Adrián	Alejandra
Creating the slides	x	**X**	x	x	x	x
Rehearsing	x	x	x	**X**	x	x

Risk contingency plan

We've thought about the different risks and expected problems that could arise during the development of this project, in order to think in advance about how to prevent them, and also how to manage them if they end up happening anyway.

Risk 1: Bad decisions in preprocessing

- **Description:** As it's expected, we will delete data that we consider unnecessary, but this could give us additional work if we mistakenly delete an attribute that ends up being very important. In that case, we might have to undo a lot of work and redo everything we've done that's affected by those changes. The opposite situation may also happen: we could end up not deleting enough information, and dedicating time and work to study data that's not useful, losing our time in the process.
- **How to prevent:** All the team members must have a clear idea of the scope of the project: which tasks we're gonna perform and what is and isn't necessary. Before making important decisions or taking actions that affect the whole project, a team member must talk about it with all its teammates. If in doubt, consult with the professor what decision should be taken.
- **How to manage:** When facing a problem, decide how to solve taking into account the project goals and the remaining time, consulting the professor if it's necessary. If the outcome is critical to other parts of the project, then dedicate more resources to solve the issue.

Risk 2: Conflicts inside the team

- **Description:** It's possible that bad relationships between team members end up creating conflicts that affect our performance, making the meetings longer and with mixed feelings about how to perform some task, resulting in hard disagreements. Also, any unexpected problem could give additional work and trouble, for instance a team member being apathetic and not finishing their work on time, or an accident leaving them unable to work for a certain period of time.
- **How to prevent:** It's important to keep in mind that we're all human, and have different abilities and opinions. We must be reasonable and respect each other, avoiding unnecessary conflicts. If we realize that a task won't be finished in time, we can ask that task's coordinator in advance, before it's too late, in order to reorganize our resources and prevent future conflicts.
- **How to manage:** Trying to solve the situation through a meeting where all the points of the conflict will be discussed to try to find some middle-ground agreement so that every member is satisfied and can work comfortably.

Risk 3: Wrong task difficulty evaluations

- **Description:** As it's expected for any software project, it's difficult to evaluate how long everything will take, especially since we don't have a lot of experience taking part in this kind of projects, and it's our first experience in data mining.
- **How to prevent:** If there's any doubt, it's always better to be cautious and spend more time and resources with tasks that seem more difficult or unknown to us. If we

overestimate the amount of work it will require, we will just end ahead of planned, but if we underestimate it we'll get into schedule problems.

- **How to manage:** If a task's progress doesn't meet the expectations, find out why it's slower than expected, avoiding stagnation, even changing the team members in charge of it or asking for help if necessary.

Risk 4: Scheduling problems

- **Description:** It's easier to plan what to do each week in advance than actually applying it, since we don't know what will happen in the following weeks. A weekend that we might have destined to develop a certain task could end up very busy due to other subject's work or personal problems, or maybe it was underestimated how long it would take to finish it. This means that it's very likely that some tasks are finished later than scheduled, giving us extra work to reorganize, and stress out.
- **How to prevent:** Take into account the other subjects that other members of the group have, respecting their schedules and other assignments when choosing dates for meetings and tasks, and trying to avoid leaving everything for the last day possible.
- **How to manage:** Trying to always make some progress, bit by bit in a continuous way, and when in doubt always assuming that a task will take more time than expected. If a team member finds themselves in a situation in which they think they won't finish a task before the deadline because of sudden new work from other subjects, they must ask their teammates for help.

Selection of variables and description

Initially, our dataset had 135 variables. It was an excessive amount of information for the scope of this project, so we decided to filter out a lot of them, and in the end even decided to include two additional ones of our own, with external data, in order to be able to study further topics we considered interesting. In this section we explain a few of the motivations that brought us to choose the variables we did, we give a complete description of all the ones that were kept in the end, and an explanation of which ones were deleted and why.

As a summary, the dataset has 28 columns of which 12 are numerical, 10 categorical, 5 binary and the final one is a date.

Selection of variables based on our objectives

We don't know what specific conclusions we'll extract or what we'll discover from our analysis, but we have a few notions of things we want to study and take a look at their results. For example, we want to search if there's any relation between the president's party and the conflicts that take place during the government's ruling, search for relations between the medium used by the criminals and other variables of the incident or if it's irrelevant, study the proportion of injured and dead US citizens compared to people from other nationalities… With those objectives in mind, we have a handful of variables that we believe will give us the exact information we need.

On the other hand, we also want to make a map of the USA to visually represent all the incidents that we have in our data, looking in which zones there's more of them, which zones are more lethal, and so on. To do this, we'll use the Latitude and Longitude variables.

Description of each variable

In the following table, for each one of our variables we've compiled their modalities, type, measuring unit, missing code, measuring procedure, and range. If any of these values doesn't apply (for example, the Range for the city variable), we've instead written a symbol "-". Additionally, we've briefly explained the meaning of each of the variables.

Variable	Modalities	Meaning	Type	Measuring unit	Missing code	Measuring procedure	Range
id	Identifier	Identifier, unique for each row	Num.	-	-	-	[0, 2835]
date	-	Date of the attack	Date	month-day-year	-	-	[1/1/1970, 12/31/2017]
provstate	Every U.S. state (including Puerto Rico,	State where the terrorist attacks were	Cat.	-	Unknown , ""	-	-

	U.S. Virgin Islands)	perpetrated					
city	-	City where the terrorist attacks were perpetrated	Cat.	-	Unknown	-	-
latitude	-	Coordinates of the attack	Num.	degrees	“”	-	[17.97, 64.84]
longitude	-	Coordinates of the attack	Num.	degrees	“”	-	[-157.86, 105.27]
doubterr	-	It indicates whether there is doubt or not about the attack being a terrorist one	Bin.	-	-9	-	-
success	-	It indicates whether the terrorist attack was a success or not	Bin.	-	-	-	-
attacktype1_txt	Assassination, Hijacking, Kidnapping, Barricade Incident , Bombing / Explosion , Armed Assault, Unarmed Assault, Facility / Infrastructure Attack	This field captures the general method of attack and often reflects the broad class of tactics used.	Cat.	-	Unknown	-	-
targtype1_txt	Abortion Related, Airports & Aircraft, Business, Educational Institution, Food or Water Supply, Government (Diplomatic), Government (General), Journalists & Media, Maritime, Military, NGO, Police, Private Citizens & Property, Religious Figures / Institutions, Terrorists/Non-	It captures the general type of target	Cat.	-	Unknown	-	-

	State Militia, Tourists, Transportation, Utilities, Violent Political Party						
natlty1_txt	Every country in the world	This is the nationality of the target that was attacked, and is not necessarily the same as the country in which the incident occurred	Cat.	-	“”	-	-
gname	Attacker group name	Group that carried out the attack	Cat.	-	“”, Unknown	-	-
nperps	-	Total amount of terrorists participating in the incident	Num.	Number of people	“”, -99, NA	Reports from credible sources	[1, 24]
nperpcap	-	Total amount of captured perpetrators	Num.	Number of people	“”, -99, NA	-	[0, 11]
claimed	Yes/1, No/0	Whether a group or person claimed responsibility for the attack	Bin.	-	“”, -9, NA	-	-
claimedmode_txt	Letter, Call (post-incident), Call (pre-incident), E-mail, Note left at scene, Video, “Posted to website, blog, social media”, Personal claim, Other, Unknown	Mode used by the claimant to claim responsibility	Cat.	-	“”, Unknown	-	-
weaptype1_txt	Biological, Chemical, Radiological, Nuclear, Firearms, Explosives, Fake Weapons, Incendiary, Melee, Vehicle, Sabotage Equipment, Other, Unknown	Type of the (main) weapon used in the attack	Cat.	-	“”, Unknown	-	-
nkill	-	Total confirmed fatalities of the attack	Num.	Number of people	“”, NA	-	[0, 190]

nkillus	-	Total confirmed U.S. citizen fatalities of the attack	Num.	Number of people	“”, NA	-	[0, 1360]
nkillter	-	Total confirmed perpetrator fatalities of the attack	Num.	Number of people	“”, NA	-	[0, 5]
nwound	-	Total confirmed non-fatal injuries the attack	Num.	Number of people	“”, NA	-	[0, 851]
nwoundus	-	Total confirmed U.S. citizen non-fatal injuries of the attack	Num.	Number of people	“”, NA	-	[0, 151]
nwoundte	-	Total confirmed perpetrator non-fatal injuries of the attack	Num.	Number of people	“”, NA	-	[0, 36]
propvalue	-	U.S. dollar amount of total damages	Num.	USD (not normalized)	“”, -99	-	[0, 652000000]
INT_IDEO	0 / 1	It indicates whether a perpetrator group attacked a target of a different nationality	Bin.	-	-9, “”	-	-
INT_MISC	0 / 1	It indicates if the attack was miscellaneous international	Bin.	-	-9, “”	Comparison between the location of the attack and the nationality of the target(s) / victim(s)	-
president_party	Republican, Democratic	Governing party of the U.S. at the moment of the attack	Cat.	-	Nan	-	-
state_governor_party	Republican, Democrat, Popular Democratic Party, Domecratic - Farmer - Labor, New Progressive Party	Governing party of the attacked state at the moment it happened	Cat.	-	Nan	-	-

Justification of deleted variables

While taking a look at the chosen dataset, we noticed that some columns either had very little meaning for our research or their occurrence rate was minimal. We have decided to remove all these columns from the dataset in order to make it clearer and also focus it on the matters we are studying. In this section, we've grouped different deleted variables that were removed following the same reasoning, and explained which one it was.

Variables:

- eventid
- iyear
- imonth
- iday
- approxdate

Reasoning:
They were event date related and we exchanged them for a single, fused date column, that gives the same information and is cleaner.

Variables:

- extended
- resolution

Reasoning:
Both this columns referenced acts of terrorism that took more than a day to resolve but there were not enough cases in order to contemplate this issue, so it wasn't worth enough to keep.

Variables:

- country_txt
- region_txt

Reasoning:
In our sample these variables had the same value for every row and therefore it didn't give any additional information. The reason they existed at all in the original dataset was because it contemplated terrorism incidents in the whole world, but since our sample was smaller, the value in this variables became useless.

Variables:

- specificity
- vicinity
- multiple

Reasoning:
These variables were too specific for our project and there were not enough cases in any of them.

Variables:

- location
- summary

- alternative_txt
- motive
- weapdetail
- addnotes

Reasoning:
These were text variables that served no purpose in our study even though they carried additional information about the crime, since we could not develop statistical analysis with them.

Variables:
- crit1
- crit2
- crit3

Reasoning:
We decided that all records of terrorism were going to be conceived as such independently of the criteria because we have at our disposal the column doubtterr, which indicates whether there is doubt or not about the attack being a terrorist one.

Variables:
- suicide
- individual

Reasoning:
There were not enough records to take these variables into account, even though they could have provided interesting information otherwise.

Variables:
- corp1
- target1

Reasoning:
These textual variables did not serve a purpose in our study because they were too specific.

Variables:
- gsubname
- targsubtype1_txt
- weapsubtype1_txt

Reasoning:
These categorical variables were too specific and they had too many values, we considered that the columns gname, targtype1_txt and weaptype1_txt would provide us more than enough information.

Variables:
- guncertain1

Reasoning:
We will assume that in the case that there is a perpetrator group name specified in gname, there is no doubt that they were in fact the perpetrators.

Variables:

- attacktype2_txt
- attacktype3_txt
- targettype2_txt
- targetsubtype2_txt
- corp2
- target2
- natlty2_txt
- targettype3_txt
- targetsubtype3_txt
- corp3
- target3
- natlty3_txt
- gname2
- gsubname2
- gname3
- gsubname3
- guncertain2
- guncertain3
- claim2
- claimmode2_txt
- claim3
- claimmode3_txt
- weaptype2_txt
- weapsubtype2_txt
- weaptype3_txt
- weapsubtype3_txt
- weaptype4_txt
- weapsubtype4_txt

Reasoning:

All these variables, as we can see, are part of a longer series (they all have a “2” or “3”) and existed originally in case extra information was collected, but after further inspection and analysis we deemed them unnecessary, so we will only take into account the main variables (the first one of each series).

Variables:

- compclaim

Reasoning:

This column made no sense in most of the cases as there being more than one claim of an act is a really strange occurrence.

Variables:

- property
- propextent_txt
- propcomment

Reasoning:

These textual and categorical variables referenced the amount of economic damage inflicted in the acts of violence, giving additional insight into it. We have decided to remove them because we already have the column propvalue which properly quantifies it.

Variables:

- ishostkid
- nhostkid
- nhostkidus
- nhours
- ndays
- divert
- kidhijcountry
- ransom
- ransomamt
- ransomamtus
- ransompaid
- ransompaidus
- ransomenote
- hostkidoutcome_txt
- nreleased

Reasoning:

All these variables make reference to different values and attributes of a hostage situation, and although we acknowledge that studying data with a hostage situation would be very interesting, there are not enough records in the dataset in order to take them into account in our study.

Variables:

- scite1
- scite2
- scite3
- dbsource

Reasoning:

These columns carried information about the source of the crime record so it was not focused in our case of study.

Variables:

- INT_LOG
- INT_ANY

Reasoning:

We removed these columns because they carried very similar information to INT_IDEO and INT_MISC. We chose to keep these two just in case something interesting about this data pops up while during the study.

Basic initial univariate descriptive statistics of raw variables

In this section we will study each raw variable in a univariate way. In the R script, which we can comfortably see in a html file called BIUDS_raw_data.html that we attached in the delivery of this report, we can see step by step all the commands we've done to study the variables and the results of their summaries. For some variables, we've also included a graph choosing the most suitable type between two different ones: boxplot or histogram for the numerical ones, and bar or pie chart for the categorical ones. Some variables have been left as textual since they had too many different values to be reasonably considered categorical, so we've described them with some summarized statistics.

The statistics of the numerical variables contain the following information:

- Mean
- Variance and Standard Deviation
- 25% and 75% percentile
- Median
- Skewness
- Kurtosis

The statistics of the categorical variables, on the other hand, contain the following information:

- Count
- Mode

Since the following analysis looked into the raw data, in the different summaries and plots there were some values of NA, unknowns and -99 or -9, because the preprocessing had not been done yet at the moment of the analysis and because of this, some of the skewness and kurtosis functions don't work.

Preprocessing

For this project, we've needed to do a good amount of preprocessing work since there were too many variables, rows, that were irrelevant or didn't have enough instances that used them, misplaced numbers, and so on. A small part of this work was made before importing all the data with R, working directly on a csv file, but the majority of the process has been performed in R, which is reflected in the scripts attached to the delivery, alongside this report.

Enumeration and description of steps

In this section, we explain which steps of the Preprocessing process we've gone through and which ones we haven't, briefly explaining what we did in the first case and why we didn't go through it in the second case.

- Data fusion & merging
    - We briefly went through this step in our data preparation phase, before starting the preprocessing itself, because we got all the data from a single dataset which has enough information, but we decided to add two additional columns that contain information about the political status of the state at the moment of each incident. Specifically, which political party was ruling over the USA and which one was ruling over the specific state where the incident happened.
- Variables declaration in pre-processing tool
    - We also briefly went through this step, deleting redundant categorical variables that were represented in two different formats.
- Visualization and basic descriptive statistics
    - We introduced our data into R to look at the most basic statistics, in order to understand a little more about what we're gonna work with. We've spent a lot of time in this step, and all the results can be found in the attached files.
- Filtering
    - We looked at all the attributes of our data, column by column, and chose which ones to delete (reasoning why). We also deleted all the information about incidents outside the United States of America, therefore bounding our data and reducing the object of study.
- Expert-based variable selection
    - We skipped this step, since we didn't ask for any expert's help because we had more than enough information to begin with.
- Outlier detection and correction
    - We introduced all our data in R and searched there for outliers, using boxplots and summaries to detect them. All the outliers were then deleted, cleaning our data, leaving it inside its most usual interval. As stated earlier in the report, all these results, analysis and reasonings can be found in the attached files.
- Error detection and correction

    - We noticed two different mistakes in our dataset: firstly, since we imported the data in a textual format into excel, part of some rows were displaced after the format transformation (one attribute is missing, and then all the posterior attributes are moved one column to the left). Since we deleted some columns, erasing the attributes we chose not to include in our project, we accidentally deleted the wrong information for those rows, and when we realized it we had to go back to an older version of the dataset to get back the lost data. The second error we detected was that, after manipulating the data, at some point the decimal comma in the values of the Latitude column disappeared for all its rows, so we had to check an older version again in order to get the right numbers back.
- Missing data treatment
    - When introducing our data into R, using its tools we quickly found the different missing values in each column. To decide how to handle each one we reasoned in a case by case basis. All of the reasoning and treatments are shown and explained in the attached files.
- Compositional and multivalued variables
    - We didn't go through this step, because our data already had the compositional variables we needed and therefore we didn't have to create our own.
- Instance selection and Feature selection/dimensionality reduction
    - As we explained earlier, in the Filtering step we reduced our data to the United States instead of the entire planet. We didn't perform any additional tasks in these steps.
- Transformations
    - We've only made a transformation of fusing some variables into a single one: there were some groups of variables consisting of a day, month and year, and we transformed them into a single date type variable. Apart from this, we didn't transform any other variables.
- Creation of new variables
    - We didn't go through this step, since we didn't create any additional variable from previous ones.

Description of the preprocessing made in each variable

As stated earlier, our R script has been made with the objective of being self-explanatory, with a lot of comments alongside the code. All the comments regarding the preprocessing performed to each variable can be found in the Preprocessing.html file that's attached to the delivery, along with the preprocessing itself. Its contents are divided into different categories, and in the second one, titled "Preprocess of each variable", we can find the contents that correspond to this part of the report.

Descriptive statistics of variables that have been modified

In the same way as what we stated in “Description of the preprocessing made in each variable”, we have the explanations that correspond to this category alongside the code that can be found in the BUDS_preprocessed_data.html file that’s attached to the delivery.

The structure of the file is similar to the previous ones: for each variable, we have a small section showing its results with a small analysis or observation of them. Not all variables have been modified, but since we deleted some rows, all of them are affected anyways, so we decided not to exclude any variable from this analysis, even if we’re only showing the small differences in their summary.

Final considerations

After doing all this elaborate preprocessing, we've realized that the dataset was too messy and unfinished, with a lot of variables that lacked values, and mistakes that ranged from very easy to solve (but still annoying, and in sum ended up giving additional work) to quite hard and difficult to decide how to handle. There even were some variables that didn't make much sense, or they should in theory but its values in the whole column didn't.

In the end, if we had known in advance, we would have chosen a different dataset that didn't give us this much trouble, but there was no way we could have predicted it before actually doing the preprocessing task.