Information Visualization
CHECKPOINT II: Data cleaning and processing

G03-A

1. Initial Dataset

Composed by static tables of the Wikipedia related to the WW2.

Sample:

```
Germany: 157,621 total casualties (c. 49,000 dead)
1,236 aircraft lost[5][9]
795 tanks destroyed[10]
Italy: 6,029
Total: 163,650 casualties
```

2. Selected/Derived Data

The selected data were the battles' names, its location, its winner and the side of each country, its casualties, its dates and the commanders.

There were no derived measures since the group had to gather all the data by hand and create the final dataset. The data that the group took from the tables were the only ones who had interest to answer the proposed questions, since they are about the number of casualties and the commanders for each side of the war.

3. Data abstraction

Items: battles' commanders.

Composed by three static files:

Commander_Allies.csv and *Commander_Axis.csv*: table containing the name and country of an Ally or Axis commander, respectively, that participated in a certain battle.

```
Battle;CommandersAlliesCountry;CommandersAlliesName;
Battle of Britain;United Kingdom;Winston Churchill;
```

- Attributes:
  - Battle, CommandersXCountry and CommandersXName: nominal, where X is his side.
- Meaning:
  - Battle: battle's name;
  - CommanderXCountry: commander's country and name, respectively, where X is his side.

Items: battles.

Date_Casualties_Victory.csv: table containing the battles, the location, the dates, casualties for each side and the winner.

```
MainBattle;Battle;StartDate;EndDate;TotalDays;CasualtiesAllies;CasualtiesAxis;Location;
Victory
Battle of Britain;Battle of Britain;10/07/40;01/11/40;111;91964;4303;London;Allies
```

- Attributes:
  - MainBattle, Battle, Location and Victory: nominal;
  - StartDate and EndDate: quantitative, sequential, hierarchical;
  - TotalDays, CasualtiesAllies and CasualtiesAxis: quantitative, ratio.
- Meaning:
  - MainBattle, Battle and Location: main battle's name, battle's name and location, respectively;
  - StartDate and EndDate: battle's starting and ending date;
  - CasualtiesX: casualties' numbers where X is the side;
  - Victory: the side who won the battle.

4. Dataset processing

The problems were that there was no or insufficient data regarding casualties. The first problem was fixed by deleting the battle from the dataset. The second was fixed by inputting a value that made sense according to the data found (e.g. "unknown, but higher", "unknown, but significant").

5. Mapping (Data sample / Questions)

- Question:
  - How did the casualties vary throughout the war?
- How to provide answers:
  - Use the dates from each battle and the casualties of each side.
- Question:
  - Which was the biggest battle?
- How to provide answers:
  - Use the battles' names and casualties of each side.