Università degli Studi di Milano Bicocca

Scuola di Scienze

Dipartimento di Informatica

Analysis and comparison of the reasons and causes of death in the various countries of the world over the years with specific interest in the case of alcohol and drugs

June 2023

Amendolia Marzia -881368- m.amendolia@campus.unimib.it
Di Gifico Gianluca- 887742- g.digifico1@campus.unimib.it
Macrì Silvia - 886653- s.macri10@campus.unimib.it
Stoffa Giacomo - 830159 - g.stoffa1@campus.unimib.it

Abstract

The project is developed with the aim of creating two different types of visualizations and graphical interfaces relating to the causes of death in different countries over the years. The first interface compares all the diseases available in the dataset, finding correlations, more frequent causes of deaths, analyzing the data during the years from 1990 to 2019. The second dashboard, specifically, uses data taken from a second data source which is closely related to the first. It compares the risks of substance use as a drug tobacco and alcohol use based on different countries by comparing the ages of the subjects and the different types of drugs used

Indice

1. Introduction

Dataset

Around 56 million people die each year.
Two data sources were used to construct the following dataset.
The first dataset was taken from kaggle at the following LINK.
This Dataset contains the causes of death and how the causes of death changed over time between different countries and world regions.
7273 rows for 261 countries for 29 years divided into 31 diseases / causes of death from 1990 to 2022. The diseases available in the dataset are the following:

```
'meningitis',
"alzheimer's_diesease",
"parkinson's_disease",
'nutritional_deficiency',
'malaria',
'drowning',
'interpersonal_violence',
'maternal_disorders',
'hiv/aids',
'drug_use_disorders',
'tuberculosis',
'cardiovascular_diseases',
'lower_respiratory_infections',
'neonatal_disorders',
'alcohol_use_disorders',
'self_harm',
'exposure_to_forces_of_nature',
'diarrheal_diseases',
'environmental_heat_and_cold_exposure',
'neoplasms',
'conflict_and_terrorism',
'diabetes_mellitus',
'chronic_kidney_disease',
'poisonings',
'protein_energy_malnutrition',
'terrorism',
'road_injuries',
'chronic_respiratory_diseases',
'chronic_liver_diseases',
'digestive_diseases',
'fire_heat_hot_substance',
'acute_hepatitis']
```

The second data source comes from kaggle like the first and is made up of different sub-categories of datasets. You can see it at the following LINK

Of which we decided to use only 3 of our interest for what we had in mind to do:

- `substances-risk-factor-vs-direct-deaths.csv`
- `deaths-substance-disorders.csv`
- `deaths-substance-disorders-age.csv`

Respectively the first csv shows the risks of using substances such as alcohol, tobacco and drugs in the influence of the causes of death.
The second csv called "deaths-substance-disorders" shows which are the most frequent types of substances, while the third refers to the number of deaths from substance use divided by 5 age groups.
All 3 of these data are divided equally to the first data source so 261 countries from 1990 to 2019.

Goals

The project aims to create visualizations that best express the differentiation of causes of death over the years and in different countries.
We also decided to focus on the specific case of the main problems of alcoholism and drug use to see the risk of death they cause and what are the main age groups subject to this addiction, also specifically comparing the type of drug and the country
The purpose of this project is also to be able to transmit information by visualizing the data by showing and avoiding correlations that can be noted such as the different types of disease linked to the geographical position and the reference year.

2. Exploration

Dataset

The main dataset has the following initial structure:

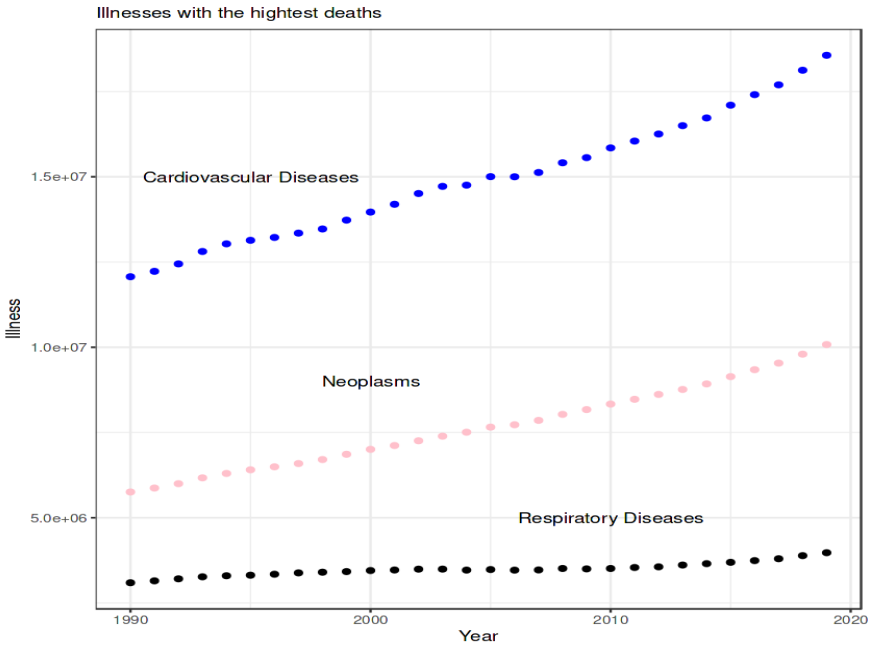
- **Country**;
- **Code** ;
- **Year**.

All diseases available for each column:

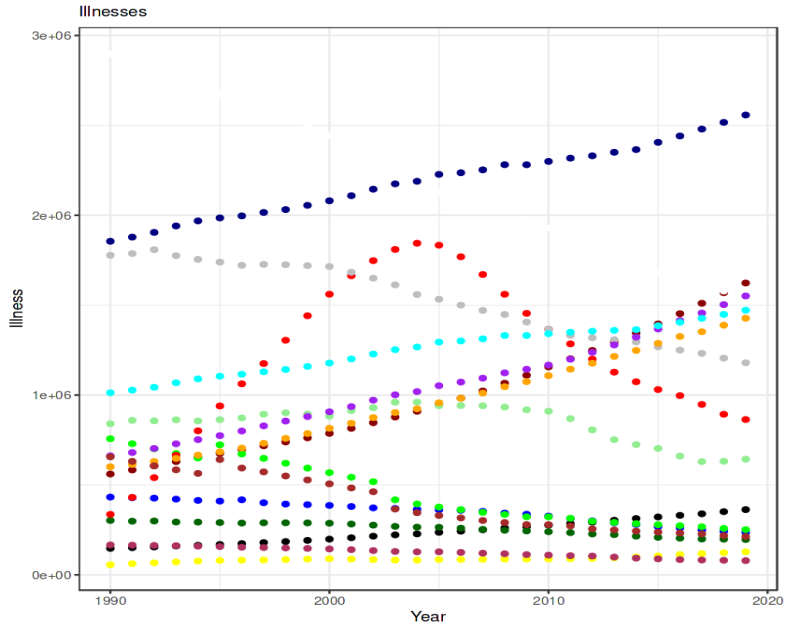
	country	code	year	meningitis	alzheimer's_diesease	parkinson's_disease	nutritional_deficiency	malaria	drowning	interpersonal_violence	...	chronic_kidney_disease	p
0	Afghanistan	AFG	2007	2933.0	1402.0	450.0	2488.0	393.0	2127.0	3657.0	...	4490.0	
1	Afghanistan	AFG	2008	2731.0	1424.0	455.0	2277.0	255.0	1973.0	3785.0	...	4534.0	
2	Afghanistan	AFG	2009	2460.0	1449.0	460.0	2040.0	239.0	1852.0	3874.0	...	4597.0	
3	Afghanistan	AFG	2011	2327.0	1508.0	473.0	1846.0	390.0	1775.0	4170.0	...	4785.0	
4	Afghanistan	AFG	2012	2254.0	1544.0	482.0	1705.0	94.0	1716.0	4245.0	...	4846.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
7268	Zimbabwe	ZWE	2015	1439.0	754.0	215.0	3019.0	2518.0	770.0	1302.0	...	2108.0	
7269	Zimbabwe	ZWE	2016	1457.0	767.0	219.0	3056.0	2050.0	801.0	1342.0	...	2160.0	
7270	Zimbabwe	ZWE	2017	1460.0	781.0	223.0	2990.0	2116.0	818.0	1363.0	...	2196.0	
7271	Zimbabwe	ZWE	2018	1450.0	795.0	227.0	2918.0	2088.0	825.0	1396.0	...	2240.0	
7272	Zimbabwe	ZWE	2019	1450.0	812.0	232.0	2884.0	2068.0	827.0	1434.0	...	2292.0	

7273 rows × 35 columns

Example of the trend of the 3 main disease categories with the highest number of deaths such as Cardiovascular diseases , neoplasms and Respiratory diseases, as follows:



Compare to other diseases:



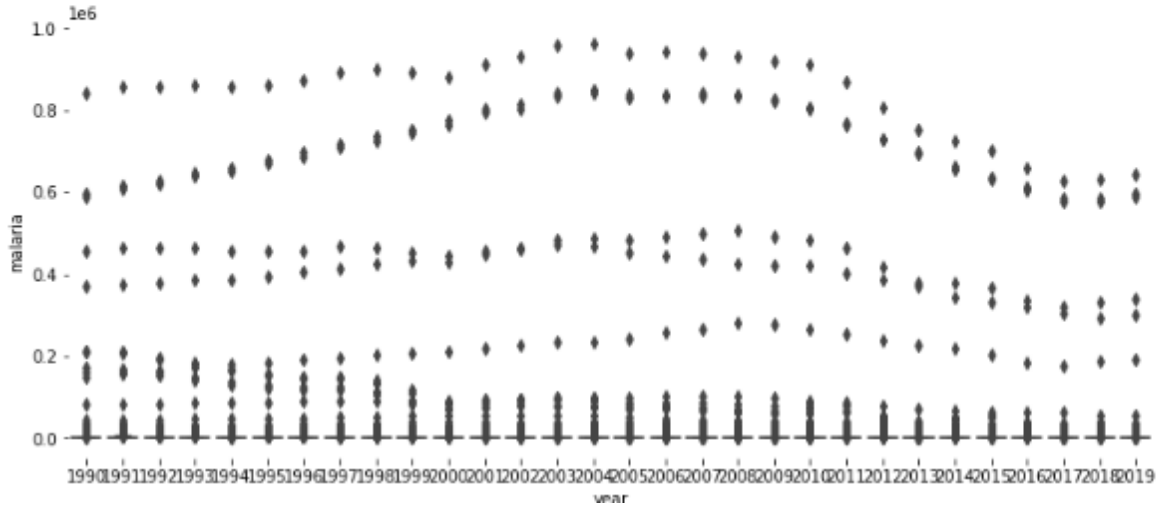
Missing values

Some fields have missing values but we can consider them not very influential considering the size of the dataset (considering a total of 7273 fields)

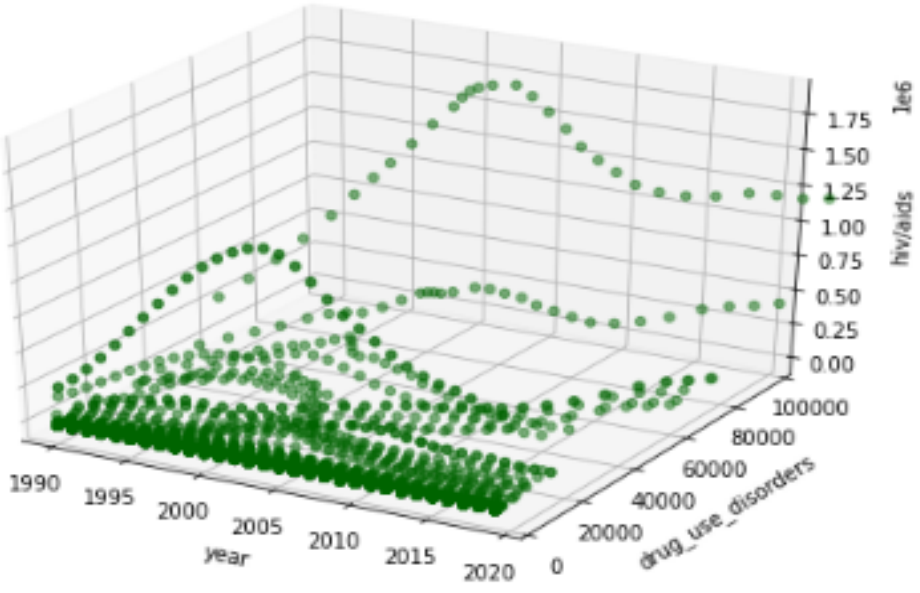
```
Data columns (total 35 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   country                               7273 non-null   object
 1   code                                  6206 non-null   object
 2   year                                  7273 non-null   int64
 3   meningitis                            6840 non-null   float64
 4   alzheimer's_diesease                  6840 non-null   float64
 5   parkinson's_disease                   6840 non-null   float64
 6   nutritional_deficiency                6840 non-null   float64
 7   malaria                               6840 non-null   float64
 8   drowning                              6840 non-null   float64
 9   interpersonal_violence                6840 non-null   float64
 10  maternal_disorders                    6840 non-null   float64
 11  hiv/aids                              6840 non-null   float64
 12  drug_use_disorders                    6840 non-null   float64
 13  tuberculosis                          6840 non-null   float64
 14  cardiovascular_diseases               6840 non-null   float64
 15  lower_respiratory_infections          6840 non-null   float64
 16  neonatal_disorders                    6840 non-null   float64
 17  alcohol_use_disorders                 6840 non-null   float64
 18  self_harm                             6840 non-null   float64
 19  exposure_to_forces_of_nature          6840 non-null   float64
 20  diarrheal_diseases                    6840 non-null   float64
 21  environmental_heat_and_cold_exposure  6840 non-null   float64
 22  neoplasms                             6840 non-null   float64
 23  conflict_and_terrorism                6840 non-null   float64
 24  diabetes_mellitus                     6840 non-null   float64
 25  chronic_kidney_disease                6840 non-null   float64
 26  poisonings                            6840 non-null   float64
 27  protein_energy_malnutrition           6840 non-null   float64
 28  terrorism                             2891 non-null   float64
 29  road_injuries                         6840 non-null   float64
 30  chronic_respiratory_diseases          6840 non-null   float64
 31  chronic_liver_diseases                6840 non-null   float64
 32  digestive_diseases                    6840 non-null   float64
 33  fire_heat_hot_substance               6840 non-null   float64
 34  acute_hepatitis                       6840 non-null   float64
dtypes: float64(32), int64(1), object(2)
```

Some examples of distribution (Malaria):

```
f, axe = plt.subplots(1, 1,figsize=(12.18,5))
sns.boxplot(x=df['year'],y=df['malaria'], ax=axe)
sns.despine(left=True, bottom=True)
axe.yaxis.tick_left()
axe.set(xlabel='year', ylabel='malaria');
```



here we can see an overview of a 3D graph that correlates 3 categories such as year,drug use disorders and hiv/aids:



In the second data source, there are the following categories:

- `substances-risk-factor-vs-direct-deaths.csv`

Deaths - Drug use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Alcohol use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Tobacco - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Drug use - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Alcohol use - Sex: Both - Age: All Ages (Number)
93	72	9723	174	356
102	75	9918	188	364
118	80	10386	211	376
132	85	10992	232	389
142	88	11466	247	399
...	...	...	...	...
104	48	9862	1068	4854
110	49	9974	1042	4915
115	50	10060	1007	4992
121	51	10162	969	5044
127	53	10317	963	5156

6840 rows x 8 columns

Divided into different categories:

```
Data columns (total 8 columns):
 #   Column                                                                Non-Null Count  Dtype
---  ------                                                                --------------  -----
 0   Entity                                                                6840 non-null   object
 1   Code                                                                  6150 non-null   object
 2   Year                                                                  6840 non-null   int64
 3   Deaths - Drug use disorders - Sex: Both - Age: All Ages (Number)      6840 non-null   int64
 4   Deaths - Alcohol use disorders - Sex: Both - Age: All Ages (Number)   6840 non-null   int64
 5   Deaths - Cause: All causes - Risk: Tobacco - Sex: Both - Age: All Ages (Number)   6840 non-null   int64
 6   Deaths - Cause: All causes - Risk: Drug use - Sex: Both - Age: All Ages (Number)   6840 non-null   int64
 7   Deaths - Cause: All causes - Risk: Alcohol use - Sex: Both - Age: All Ages (Number)  6840 non-null   int64
dtypes: int64(6), object(2)
```

mainly considering the risks associated with the use of substances such as alcohol, drugs and tobacco.
Comparing the number of annual deaths from substance use with the number of deaths caused by substance use but not necessarily deaths from substance use.

- `deaths-substance-disorders.csv`

Deaths - Cocaine use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Drug use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Opioid use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Alcohol use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Other drug use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Amphetamine use disorders - Sex: Both - Age: All Ages (Number)
3	93	73	72	15	2
3	102	80	75	16	2
4	118	93	80	19	3
4	132	104	85	21	3
5	142	112	88	22	3
...	...	...	...	...	...
10	104	77	48	10	7
11	110	81	49	10	7
12	115	85	50	11	8
13	121	88	51	11	8
14	127	92	53	12	9

6840 rows x 9 columns

Here in this dataset the deaths divided by type of substance are always considered for each year and for different countries.

- Drug use disorders (total number)
- Cocaine use disorders
- Opioid use disorders
- Amphetamine use disorders
- Other drugs use disorders
- Alcohol use disorders

	Year	Deaths - Cocaine use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Drug use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Opioid use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Alcohol use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Other drug use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Amphetamine use disorders - Sex: Both - Age: All Ages (Number)
Year	1.000000	0.091072	0.042918	0.048641	0.013359	0.018563	-0.012951
Deaths - Cocaine use disorders - Sex: Both - Age: All Ages (Number)	0.091072	1.000000	0.916492	0.932282	0.696854	0.790128	0.690392
Deaths - Drug use disorders - Sex: Both - Age: All Ages (Number)	0.042918	0.916492	1.000000	0.996634	0.858435	0.962323	0.896207
Deaths - Opioid use disorders - Sex: Both - Age: All Ages (Number)	0.048641	0.932282	0.996634	1.000000	0.863061	0.939705	0.859051
Deaths - Alcohol use disorders - Sex: Both - Age: All Ages (Number)	0.013359	0.696854	0.858435	0.863061	1.000000	0.858989	0.743191
Deaths - Other drug use disorders - Sex: Both - Age: All Ages (Number)	0.018563	0.790128	0.962323	0.939705	0.858989	1.000000	0.965454
Deaths - Amphetamine use disorders - Sex: Both - Age: All Ages (Number)	-0.012951	0.690392	0.896207	0.859051	0.743191	0.965454	1.000000

- deaths-substance-disorders-age.csv

Deaths - Substance use disorders - Sex: Both - Age: 70+ years (Number)	Deaths - Substance use disorders - Sex: Both - Age: 50-69 years (Number)	Deaths - Substance use disorders - Sex: Both - Age: 15-49 years (Number)	Deaths - Substance use disorders - Sex: Both - Age: Under 5 (Number)	Deaths - Substance use disorders - Sex: Both - Age: 5-14 years (Number)
13	51	101	0	0
14	52	111	0	0
14	53	130	0	0
15	55	148	0	0
15	57	158	0	0
...	...	...	...	...
10	70	71	0	0
10	74	75	0	0
11	77	78	0	0
11	80	81	0	0
12	84	84	0	0

6840 x 8 columns

Same as the previous dataset, I consider the total number of deaths for each country and for each year and divide it into 5 age categories:

- under 5 years
- 5-14 years
- 15-49 years
- 50-69 years
- 70+ years

3. Data Preprocessing

Dataset 1

As regards the primary dataset that we have seen previously it was decided ,for convenience with the use of the tableau software, to arrange the order of the diseases.
Going from a structure in which there was only one column for each disease to having a single category called "disease" containing all the diseases with their relative value relative to the deaths next to.

	country	code	year	disease	value
0	Afghanistan	AFG	2007	meningitis	2933.0
1	Afghanistan	AFG	2008	meningitis	2731.0
2	Afghanistan	AFG	2009	meningitis	2460.0
3	Afghanistan	AFG	2011	meningitis	2327.0
4	Afghanistan	AFG	2012	meningitis	2254.0
...	...	...	...	...	...
232731	Zimbabwe	ZWE	2015	acute_hepatitis	146.0
232732	Zimbabwe	ZWE	2016	acute_hepatitis	146.0
232733	Zimbabwe	ZWE	2017	acute_hepatitis	144.0
232734	Zimbabwe	ZWE	2018	acute_hepatitis	139.0
232735	Zimbabwe	ZWE	2019	acute_hepatitis	136.0

232736 rows × 5 columns

Going from 7273 rows x 35 columns to 232736 rows x 5 columns

With the aim of having a single category of diseases to make the best use of the possibility of showing the correlations between the various diseases and not considering only one disease at time.

Dataset 2

For the second data source, the different files were assembled into a single one, in order to have a single dataset to compare the risks of drug use and abuse based on several aspects. Combined together obtaining the following characteristics:

			disease	value	Deaths - Drug use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Alcohol use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Tobacco - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Drug use - Sex: Both - Age: All Ages (Number)	Deaths - Cause: All causes - Risk: Alcohol use - Sex: Both - Age: All Ages (Number)	Deaths - Cocaine use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Opioid use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Other drug use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Amphetamine use disorders - Sex: Both - Age: All Ages (Number)	Deaths - Substance use disorders - Sex: Both - Age: 70+ years (Number)	Deaths - Substance use disorders - Sex: Both - Age: 50-69 years (Number)	Deaths - Substance use disorders - Sex: Both - Age: 15-49 years (Number)	Deaths - Substance use disorders - Sex: Both - Age: Under 5 (Number)	Deaths - Substance use disorders - Sex: Both - Age: 5-14 years (Number)
country	code	year																
Afghanistan	AFG	1990	drug_use_disorders	93.0	93.0	72.0	9723.0	174.0	356.0	3.0	73.0	15.0	2.0	13.0	51.0	101.0	0.0	0.0
		1990	alcohol_use_disorders	72.0	93.0	72.0	9723.0	174.0	356.0	3.0	73.0	15.0	2.0	13.0	51.0	101.0	0.0	0.0
		1991	drug_use_disorders	102.0	102.0	75.0	9918.0	188.0	364.0	3.0	80.0	16.0	2.0	14.0	52.0	111.0	0.0	0.0
		1991	alcohol_use_disorders	75.0	102.0	75.0	9918.0	188.0	364.0	3.0	80.0	16.0	2.0	14.0	52.0	111.0	0.0	0.0
		1992	drug_use_disorders	118.0	118.0	80.0	10386.0	211.0	376.0	4.0	93.0	19.0	3.0	14.0	53.0	130.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Zimbabwe	ZWE	2017	alcohol_use_disorders	50.0	115.0	50.0	10060.0	1007.0	4992.0	12.0	85.0	11.0	8.0	11.0	77.0	78.0	0.0	0.0
		2018	drug_use_disorders	121.0	121.0	51.0	10162.0	969.0	5044.0	13.0	88.0	11.0	8.0	11.0	80.0	81.0	0.0	0.0
		2018	alcohol_use_disorders	51.0	121.0	51.0	10162.0	969.0	5044.0	13.0	88.0	11.0	8.0	11.0	80.0	81.0	0.0	0.0
		2019	drug_use_disorders	127.0	127.0	53.0	10317.0	963.0	5156.0	14.0	92.0	12.0	9.0	12.0	84.0	84.0	0.0	0.0
		2019	alcohol_use_disorders	53.0	127.0	53.0	10317.0	963.0	5156.0	14.0	92.0	12.0	9.0	12.0	84.0	84.0	0.0	0.0

14546 rows × 16 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14546 entries, 0 to 14545
Data columns (total 19 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   country                                       14546 non-null  object
 1   code                                          12412 non-null  object
 2   year                                          14546 non-null  int64
 3   disease                                       14546 non-null  object
 4   value                                         13680 non-null  float64
 5   Deaths_drug_use                               13680 non-null  float64
 6   Deaths_alcohol_use                            13680 non-null  float64
 7   Deaths - Cause: All causes - Risk: Tobacco    13680 non-null  float64
 8   Deaths - Cause: All causes - Risk: Drug use   13680 non-null  float64
 9   Deaths - Cause: All causes - Risk: Alcohol use  13680 non-null  float64
 10  Cocaine_deaths_drug_use_disorders             13680 non-null  float64
 11  Opioid_deaths_drug_use_disorders              13680 non-null  float64
 12  Other-drug_deaths_drug_use_disorders          13680 non-null  float64
 13  Amphetamine_deaths_drug_use_disorders         13680 non-null  float64
 14  Deaths_substance_age_70+_years                13680 non-null  float64
 15  Deaths_substance_age_50-69_years              13680 non-null  float64
 16  Deaths_substance_age_15-49_years              13680 non-null  float64
 17  Deaths_substance_age_under-5_years            13680 non-null  float64
 18  Deaths_substance_age_5-14_years               13680 non-null  float64
dtypes: float64(15), int64(1), object(3)
memory usage: 2.1+ MB
```

4. Methodology

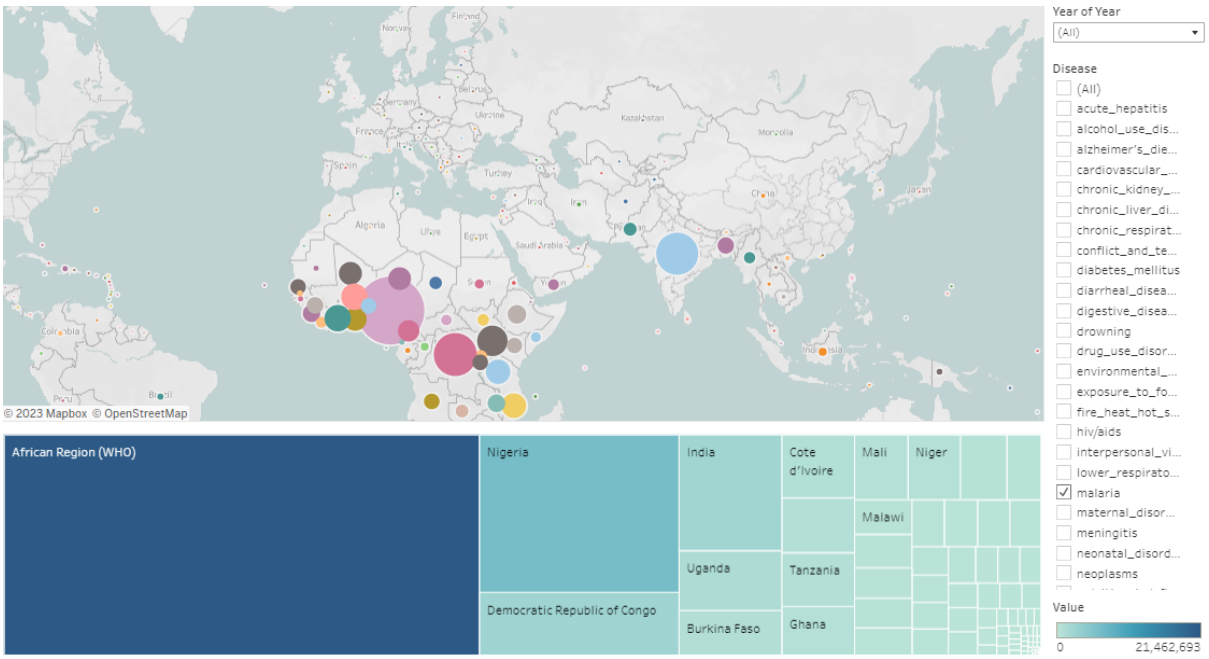
For the visualizations we decided to create two groups with 4 visualizations for each dashboard, in order to give more information and in the clearest way possible.

Dashboard 1: Causes of death form 1990 to 2019 by type and country

The first dashboard relates to the first dataset, therefore with all the general diseases, and can be consulted at the following LINK.

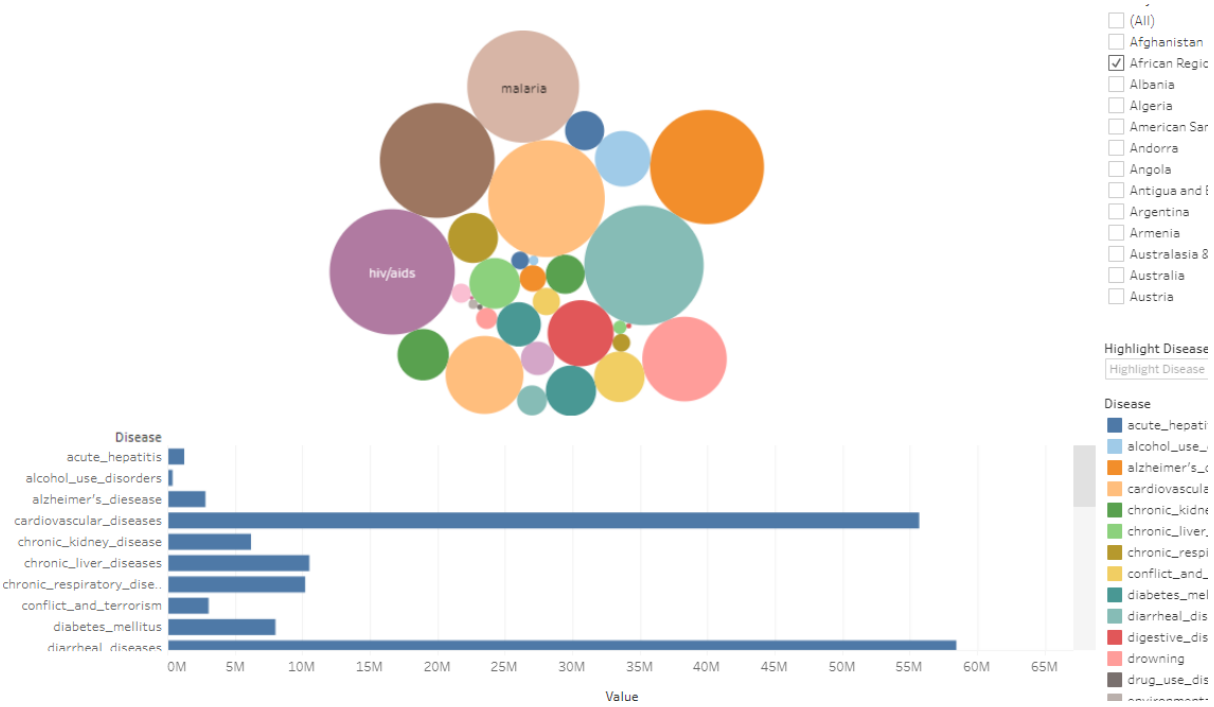
1: Distribution of deaths by type of disease

In this visualization it is possible to see through the size of the circle, the number of deaths for one or more years and for one or more diseases (example case of Malaria).



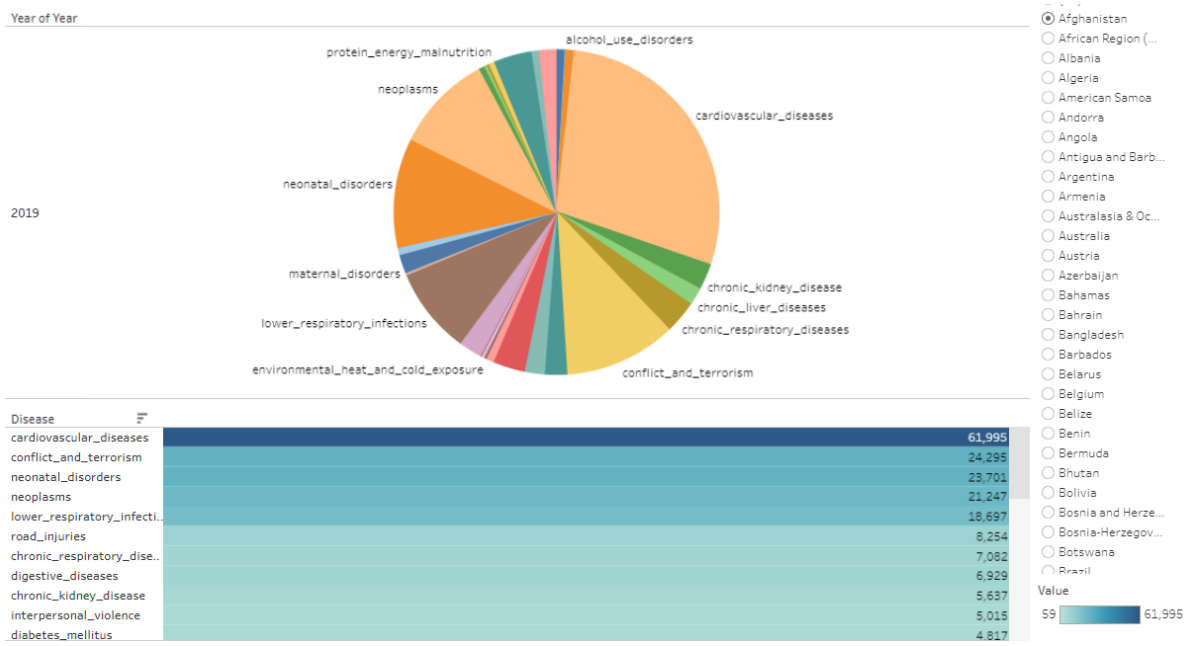
2: Distribution of deaths by type of disease

In this view you can select one or more years and one or more countries and see the major disease that caused deaths.



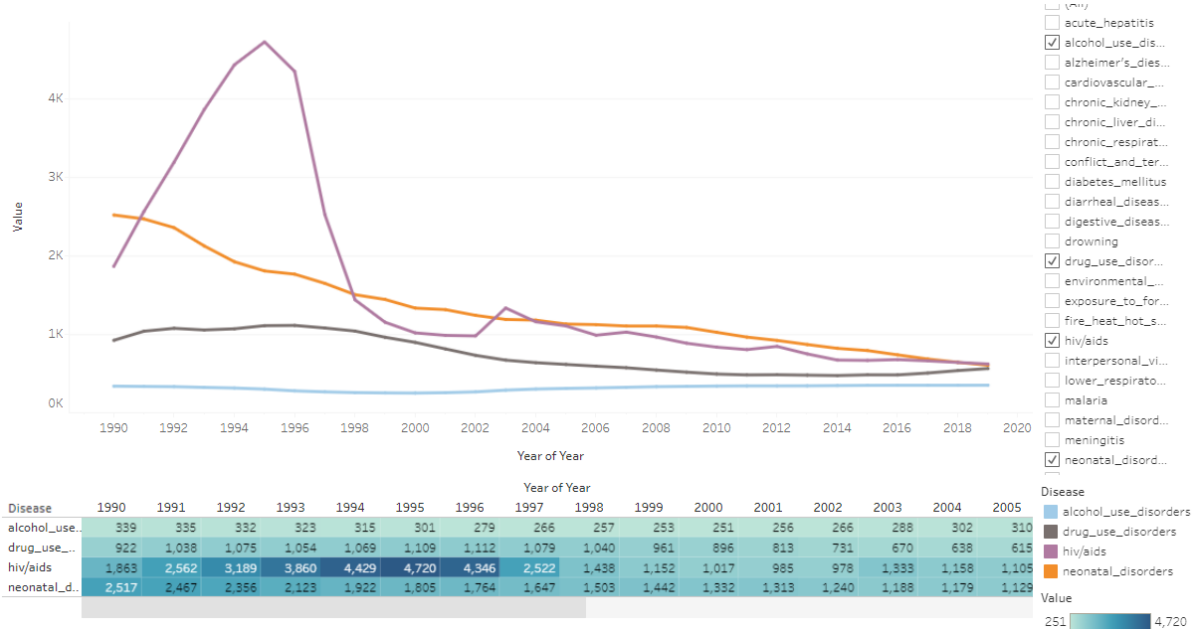
3: Distribution of deaths by country in a single year

In this view you can see the number of deaths for each country in a single year. Graphically with the pie chart above and numerically with the table below. The case of Afghanistan is highlighted looking at conflict and terrorism in 2019.



4: DIsease progression over the years

In this view it is possible to see the trend of one or more diseases over the years for one or more selected countries, thus making comparisons between them. Let's look at the case of Italy and of diseases like hiv/aids which peaked in the 90s and then stabilized. We then see drug and alcohol use disorders which we will see better in the next part.

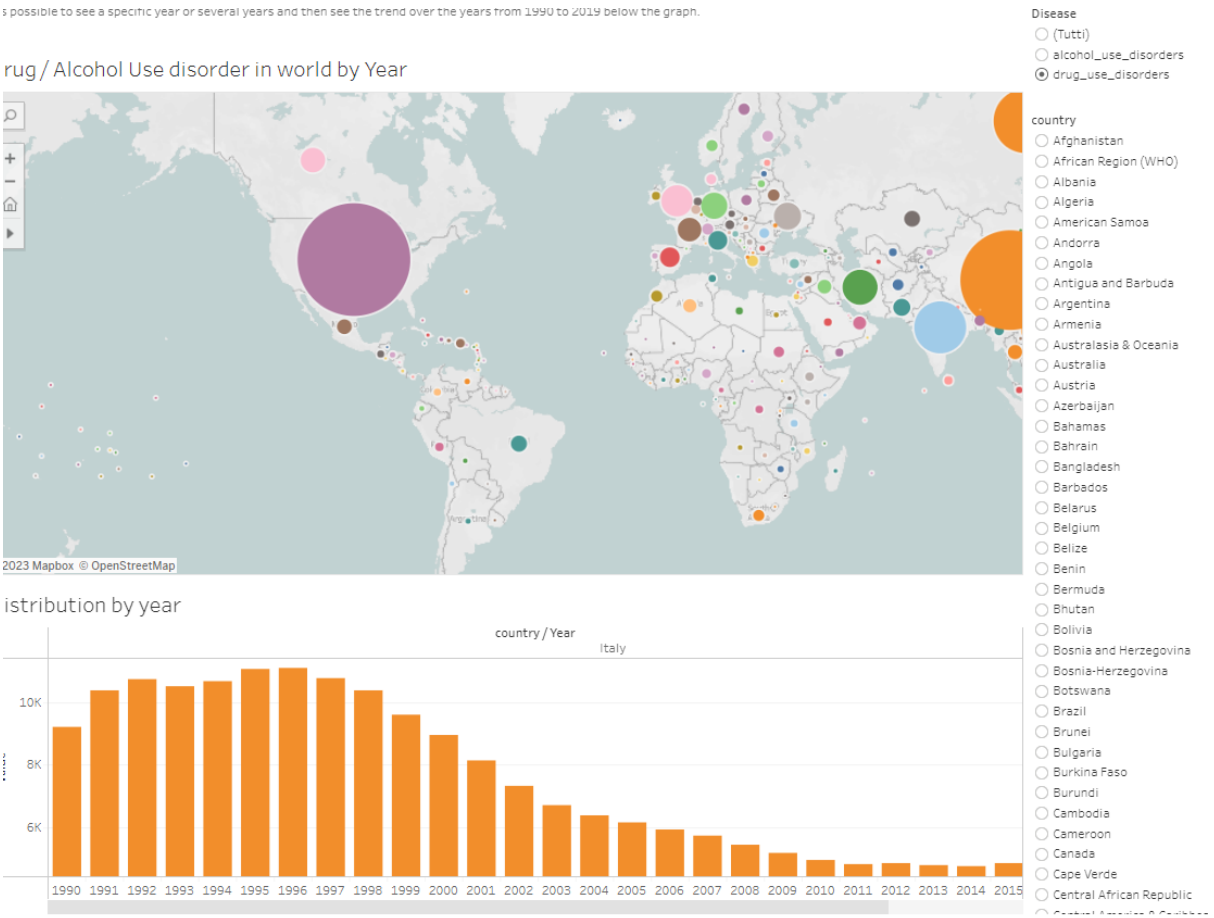


Dashboard 2: Drug and alcohol use disorders

The second dashboard, on the other hand, is specific to the second dataset, so it analyzes the use and problems related to drugs and alcohol.
It can be viewed at the following LINK.

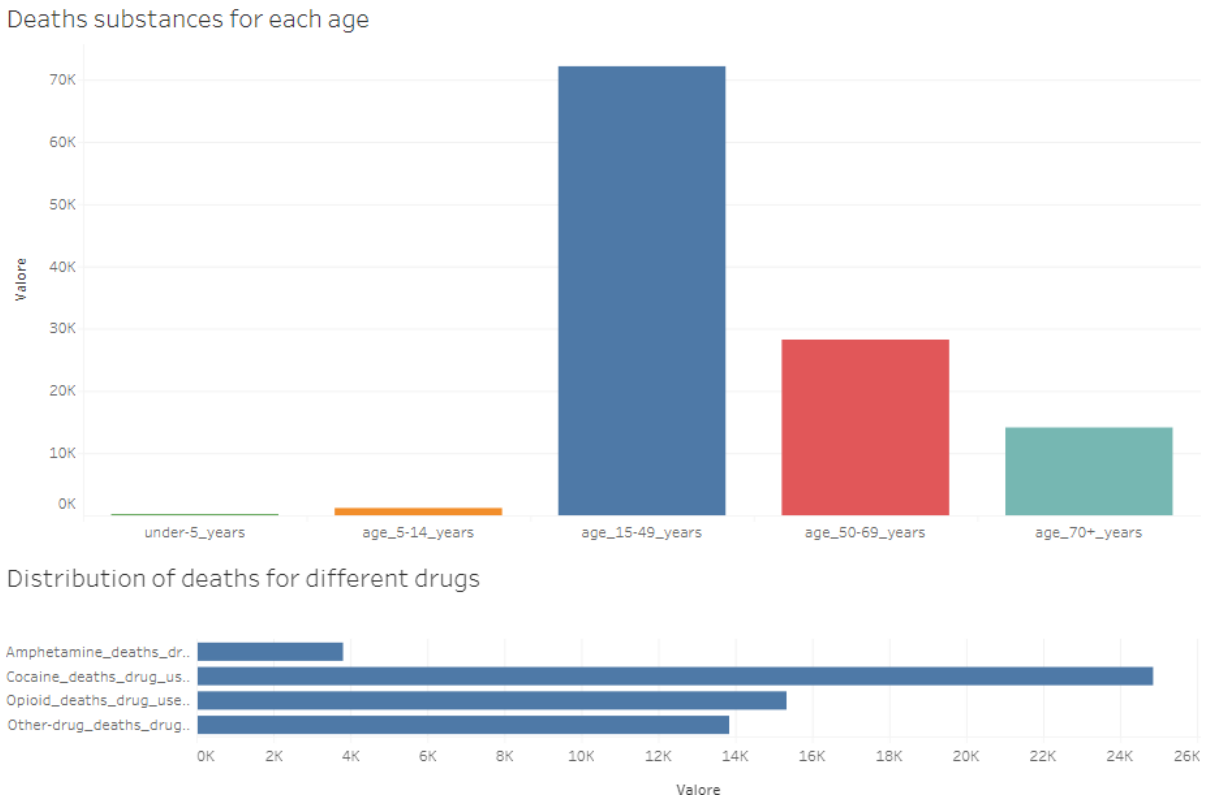
1: Drugs and Alcohol use deaths in world by years

Specifically, in this category of views we are going to analyze and examine the distribution of the number of deaths from alcohol and drugs. In this specification it is possible to graphically see for each category which state is most affected by this problem. It is possible to see a specific year or several years and then discover the trend over the years from 1990 to 2019 below the graph.



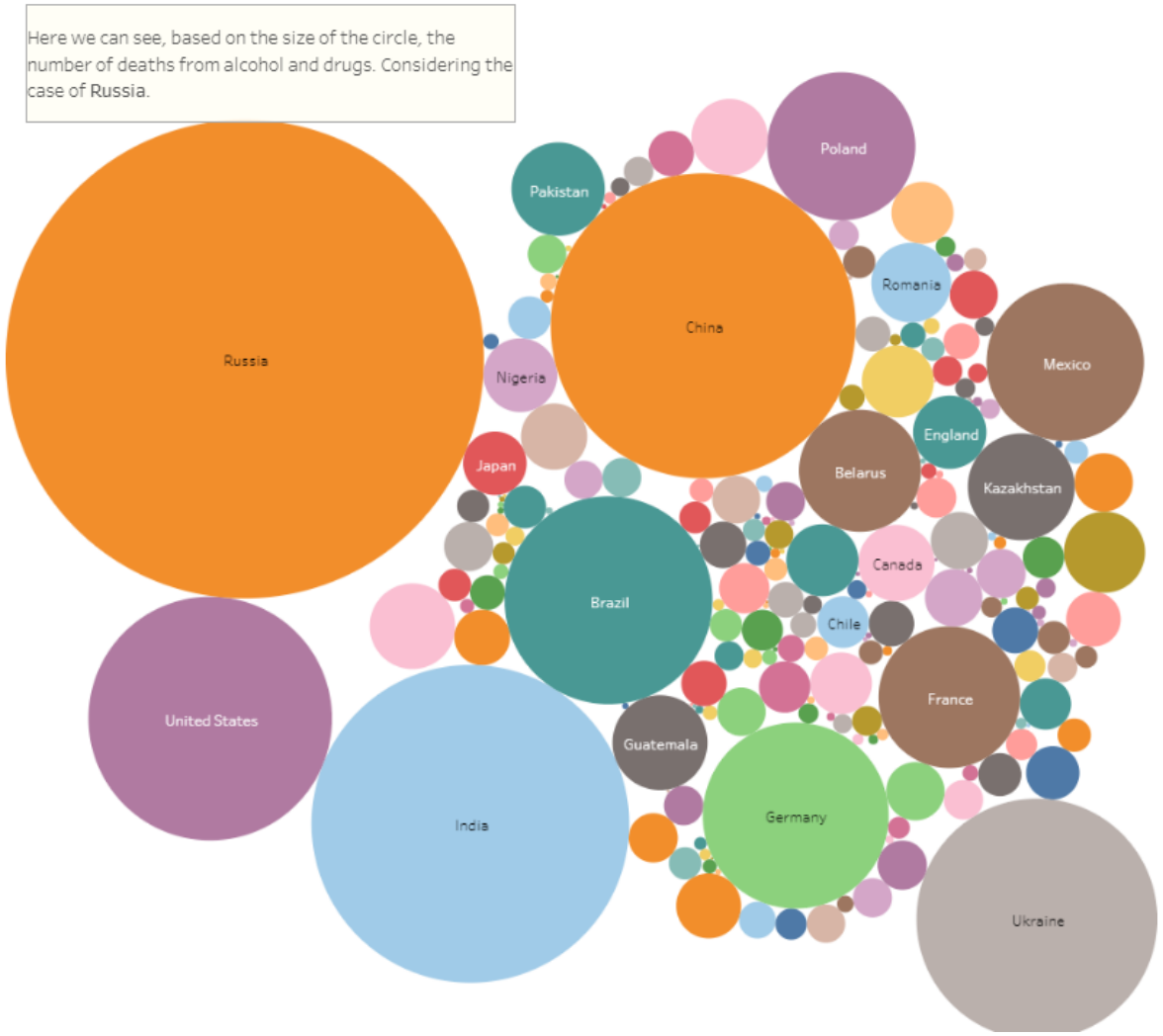
2: Distribution of deaths for each type of drugs and different ages

Here, in this specific view you can see the distribution of deaths by age and type of substance. Specifically, we have analyzed the famous case of cocaine in Colombia (it is also possible to select even more years).



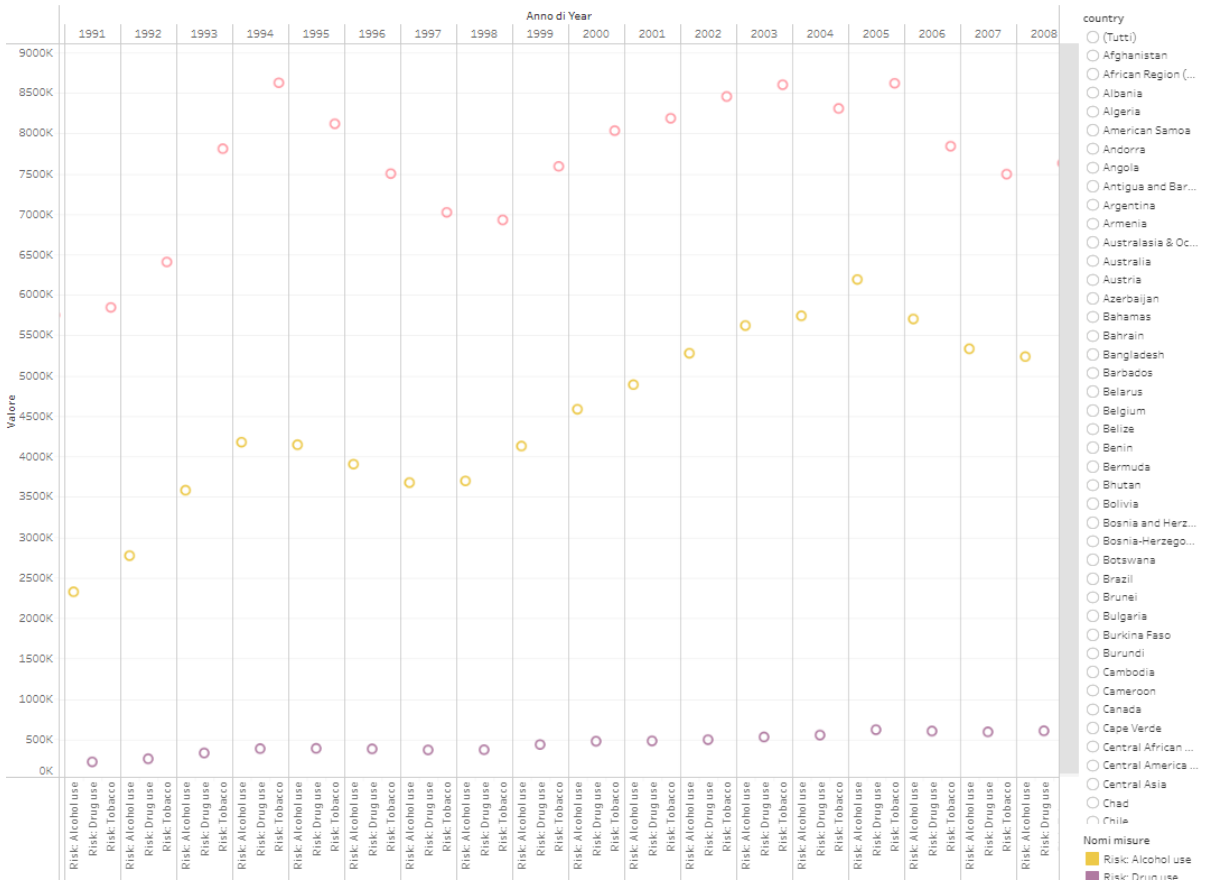
3: Countries with highest number of deaths cause alcohol and drugs

Here we viewed, based on the size of the circle, the number of deaths from alcohol and drugs considering the case of Russia.



4: Risks of death using drugs,alcohol or tobacco during years

This type of interactive graph, in some ways, can be considered more interesting because it compares the types of deaths or diseases that have been caused or negatively influenced by the use of substances such as drugs, alcohol or tobacco. Analyze the actual risks and how these have changed over time.



5. Evaluation

Evaluation Phase

To assess the quality of our Data Visualization, we adopted three methodologies:

1) Heuristic evaluation to identify general issues with the data visualization.
1) User testing to evaluate individual graphs through specifically designed tasks.
2) Psychometric questionnaire to assess the overall quality of the data visualization.

5.1 Heuristic Evaluation

The heuristic evaluation involved five users who were asked to interact freely with the data visualization and verbally express (think aloud protocol) their thoughts during the interaction, as well as describe and/or explain their actions. Through the think aloud protocol, it was possible to identify the issues with the entire data visualization.

Initially, the colors did not allow for clear distinction between similar-sized results, as there was insufficient contrast between the colors.

The Tableau font was not well-received, and it was noted that the size was too small, making the description of the visualization not immediately noticeable. Therefore, a decision was made to change the font and dimensions.

Difficulties were encountered in zooming and navigating within the graph.

Some visualizations did not highlight the labels of the captions.
In the last graph of the second group of visualizations, the legend was smaller.
Some sizes, represented by circles, were not immediately distinguishable to the naked eye.
Some filters were useful but self-referential, creating confusion.
Overall, after this evaluation phase, it was deemed appropriate to standardize the presentation by making the slides more similar to each other. Color and font changes were implemented.

5.2 User test

In this evaluation phase, we conducted user tests using the task analysis method. We designed nine tasks that we administered to eight users within a specially created setting, free from external distractions.
The table with the tasks is reported below.

Task	**Objective**	**Procedure**
1) Identify the top 3 countries with the highest number of Alzheimer's deaths (First graph)	Interact with the graph and use the disease filter to select the disease	Use the filter to select the disease and observe the changes in the graph
2)Identify the leading causes of death in 1992 for a selected country (Second graph)	Use the year and cause of death filters to select the desired year	Use the filters to select the year and observe the changes in the graph
3)Determine the number of deaths due to cardiovascular problems in Italy in the year 2000 (Third graph)	Select the disease using the filter and carefully observe the graph, focusing on the year	Use the filter to select the disease and observe the underlying graph to obtain the precise number
4)Compare the number of deaths between the years 1990 and 2005 in the United States due to alcohol (First graph, alcohol/drug group)	Select the specific country and use the underlying graph to compare the number of deaths for those years	Simultaneously use filters to select the country and the disease. Interact with the underlying graph to obtain the precise number
5)Identify the age group in which the highest number of deaths occurred in Argentina in the year 2012 (Second graph)	Select the country and the specified year, then observe the graph	Select the country and the specified year, then observe the graph
5.1)What is the cause of death that resulted in the highest number of victims? (Second graph)	Look at the underlying graph to determine the cause of death	Observe the underlying graph to answer this question
6)Determine the number of deaths due to Alcohol and Drug use, paying attention to the country that suffered the most (Third graph)	Interact with the graph to understand which country may have been most affected, and then state the number	Observe the graph and base your answer on the size of the circle. Hover over the circle to view the precise number
7)Identify the cause of death that resulted in the highest number of victims in Cuba (Fourth graph)	Observe the graph and the colors of the circles	Reason about the meaning of the colors

7.1) Identify the precise number of deaths in the year with the highest number of victims in Cuba (Fourth graph)	Interact with the graph and move the cursor to view the number	Observe the graph and hover over the exact point to obtain the precise number of deaths

Results

TASK EXECUTION TIME

To assess the usability of the data visualization, we recorded the execution times for each task. The table below shows the times expressed in seconds.

ID	TASK 1	TASK 2	TASK 3	TASK 4	TASK 5	TASK 5.1	TASK 6	TASK 7	TASK 7.1
1	20	21	15	1.10	17	7	23	25	32
2	15	1.07	20	24	38	7	24	35	10
3	50	48	35	1.20	30	7	30	38	30
4	37	50	40	40	29	7	17	27	50
5	23	50	20	50	20	7	17	25	9
6	40	40	24	59	25	9	25	20	20
7	30	50	20	40	20	10	20	20	20
8	30	45	22	35	30	12	11	50	10

ADD MEAN AND MEDIAN

Regarding the execution times of our tasks, we can therefore state that some tasks can be performed very quickly by all the users we tested, while others differ in execution, probably due to personal differences in interpreting the graphs. Based on these results, we believe that no task created execution issues, and for this reason, our infographic is accessible to all users.

General remarks

During the completion of the tasks, we noticed the following:

Users do not spend much time reading the descriptions of the visualizations that provide context to the graphs.
In all the graphs, filters are seen and used without any issues, with ease and immediacy.
To answer the questions, users make use of all the graphs present in each visualization.

Detected Problems

TASK 1

Two users noticed the first graph more because the circles had a greater impact.

TASK 2

There were issues with selecting the country filter, as users had to remove the previous selection to choose a new one.
The second graph was primarily used to find the desired result.
Users easily found and used the filter to select the years.

TASK 3

The second graph was noticed first and described as easier to use due to the use of a gradient scale, which quickly drew attention to the required data.
The first graph was described as more challenging to read due to the numerous colors, requiring more time to distinguish and making it harder to find the requested data.

TASK 4

The second graph, which contained all the years, was not noticed. Users often directly used the first graph (the map), which was frequently described as more intuitive for finding the desired data. In some cases, users manually selected the years by changing them in the filter.

TASK 5

Users directly selected the year filter and used the second graph.
The graph where the answer is displayed upon hovering was not used.

TASK 5.1

No problems found.

TASK 6

No problems found.

TASK 7

Users encountered issues with the small and similarly colored legend.
The graph was not very clear due to the use of small circles, which were not very visible initially.

TASK 7.1

Users did not immediately realize that it was a scrollable graph and not displayed on a single page.

ERRORS

To evaluate the usability of the data visualization, we also recorded the error rate for each task. The results are presented in the following table, where:

- **0 = No errors**
- **1 = Error**

ID	TASK 1	TASK 2	TASK 3	TASK 4	TASK 5	TASK 5.1	TASK 6	TASK 7	TASK 7.1
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	1
7	0	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0

The table highlights how there were difficulties in executing some tasks, but users were still able to complete them after receiving a suggestion. We want to clarify that these errors were primarily due to inexperience with using Tableau and the provided interactive interface, rather than the graphs themselves. Despite these suggestions, no task was

completely incorrect, allowing us to state that the data visualization we developed is usable by users.

5.3 Psychometric questionnaire (LINK)

The psychometric questionnaire was administered to 13 users. The questionnaire consists of three sections:

1. Demographic section: This section includes questions about the user's:

a. Gender
b. Age
c. Education level
d. English language proficiency: to investigate if it could be a factor of difficulty in understanding the graphs, as they were written in English.
e. Perceived self-efficacy in understanding graphs: to assess its possible influence on the approach to the graphs.

2. Section for evaluating the quality of individual graphs: In this part, for each screen of our data visualization, we asked users to evaluate using the Cabitza-Locoro scale, testing:

a. Utility
b. Clarity
c. Informativeness
d. Beauty
e. Overall quality

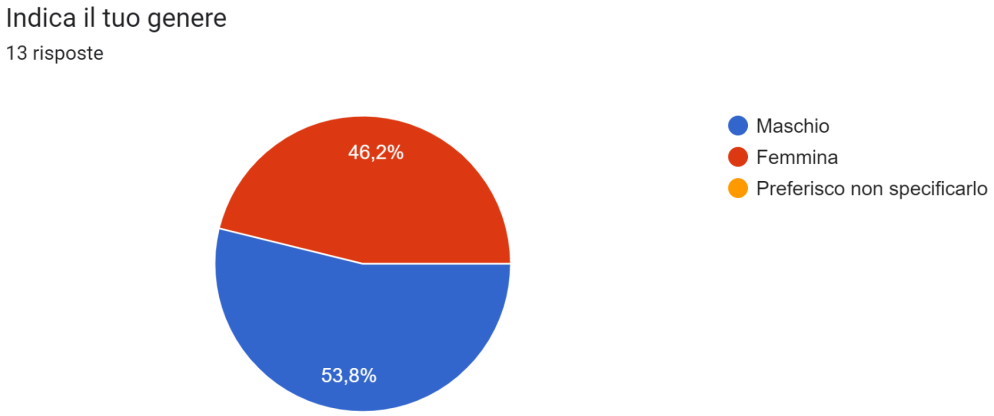
3. Section for evaluating the quality of the entire data visualization: Using the same scale as in point 2, we asked users to provide an overall evaluation of the entire data visualization.

Results of questionnaire

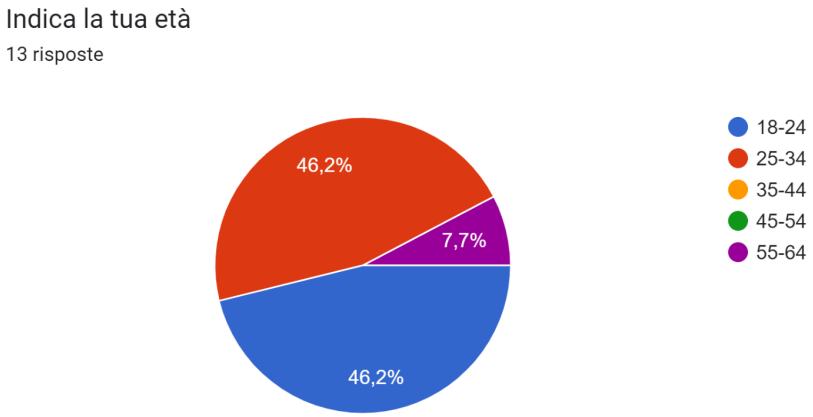
DEMOGRAPHIC QUESTIONS

The users who responded to our questionnaire were slightly more male:

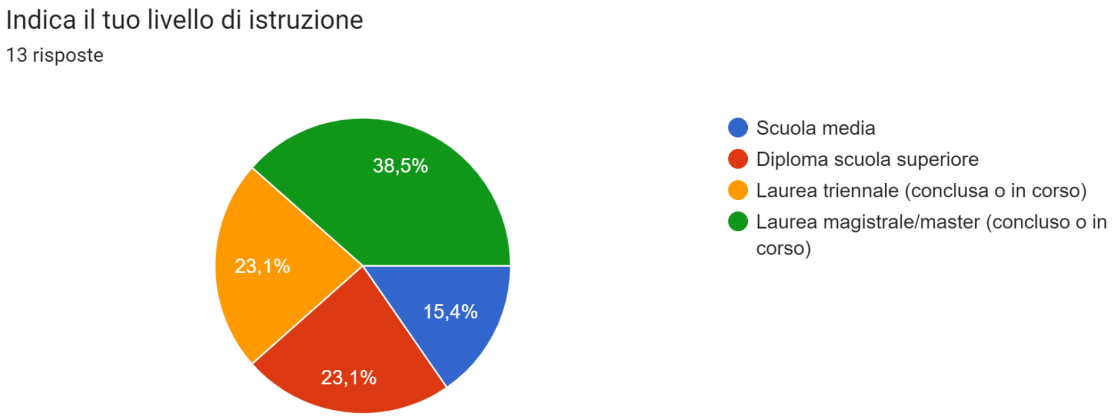
Male: 54%
Female: 46%



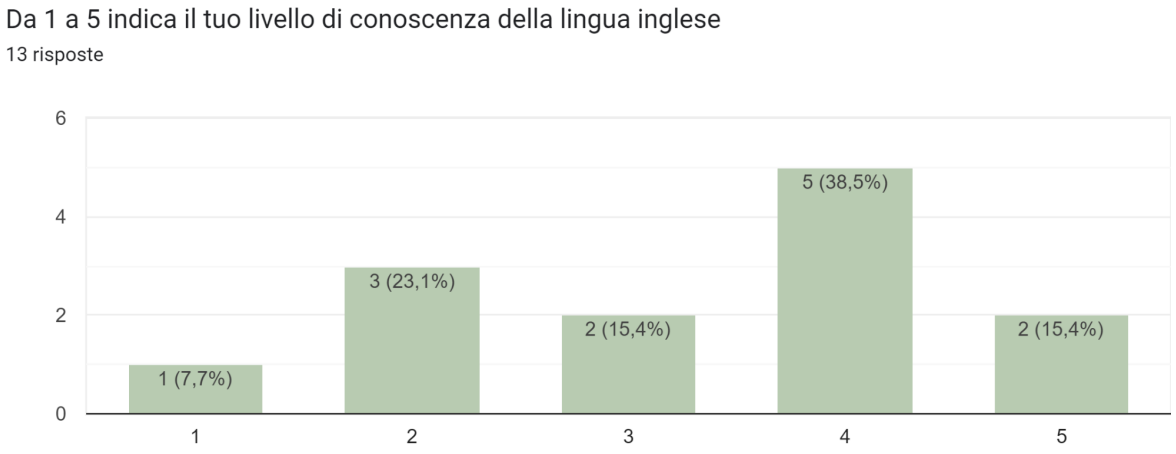
The age of our users corresponds to the target audience for which the data visualization was designed.

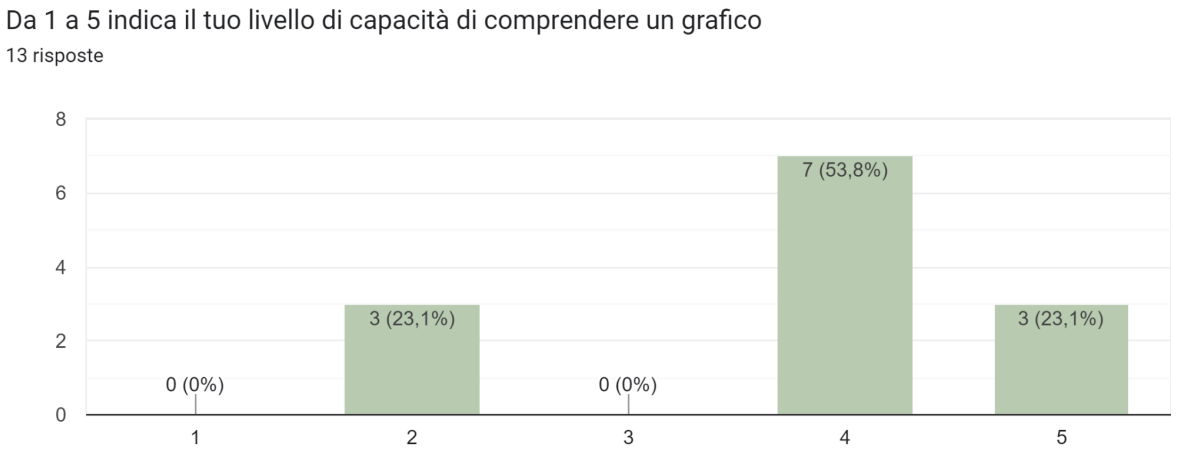


The education level of the users who responded to the questionnaire varies from a middle school diploma to a master's degree or master's degree, providing us with a wide and meaningful range for evaluation.



Similarly, regarding the level of English proficiency and perceived self-efficacy in understanding the graphs, we have subjects ranging from low to high levels. This allows us to obtain a realistic and comprehensive evaluation for all types of users who may come across our infographic.

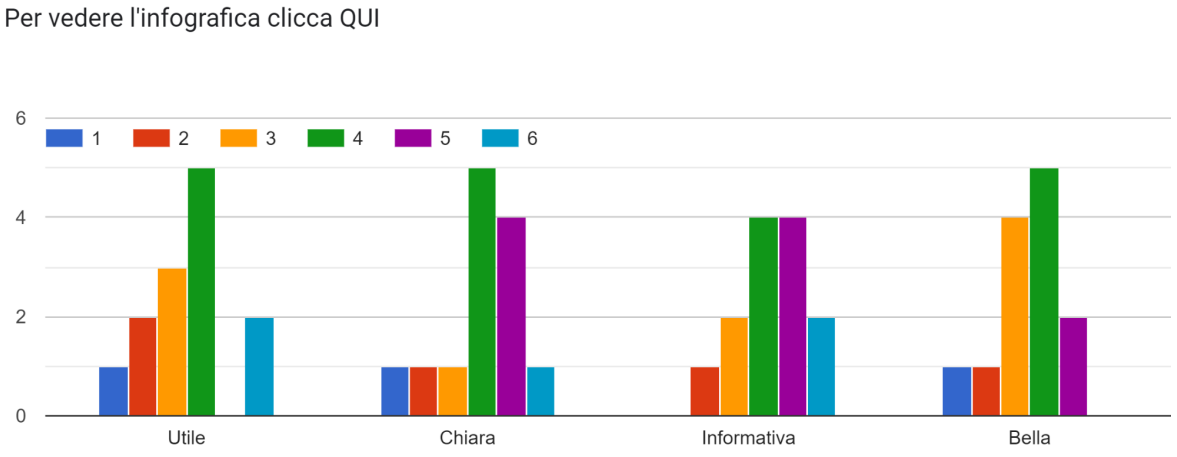




EVALUATION OF INDIVIDUAL GRAPH QUALITY:

Graph 1

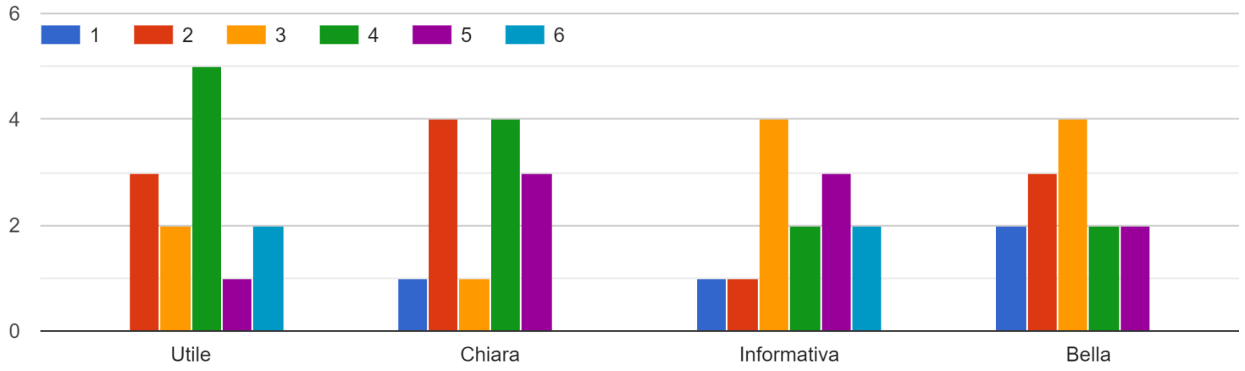
From the results, it can be observed that the examined graph was perceived as very clear, informative, and visually appealing by the majority of participants, consistently receiving scores between 5 and 6. Regarding the utility of the graph, although the result is above average, it may be worth considering revising the aesthetic aspect, as there were several evaluations corresponding to a score of 4.



Graph 2

From this second set of responses, we notice that the majority of users rated the clarity and utility of the second graph positively. However, for the rest of the questions, we observe a series of responses with lower values.

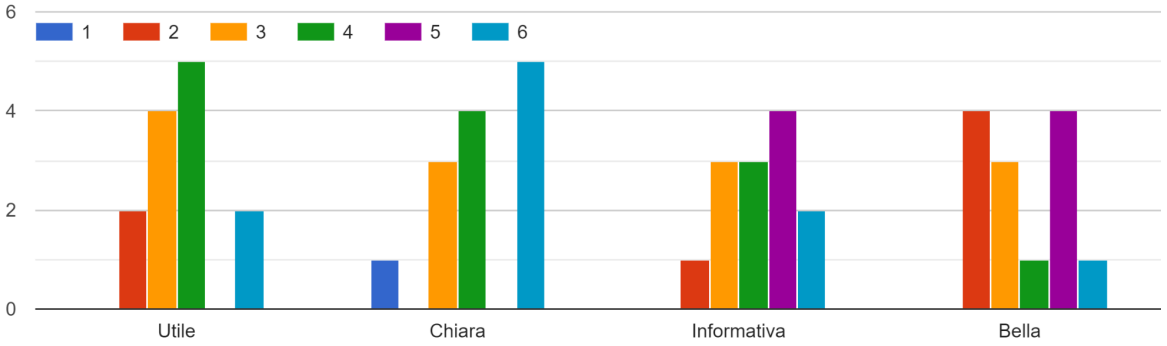
Per vedere l'infografica clicca QUI



Graph 3

From the results of this graph, it can be observed that it is very clear and informative. However, in terms of beauty, the result is below average, as there are more ratings of 3 or lower.

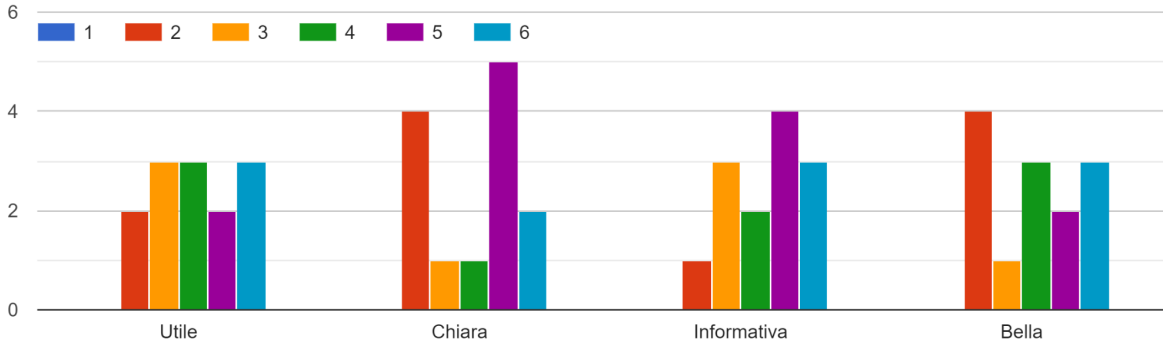
Per vedere l'infografica clicca QUI



Graph 4

From the results of this graph, it can be observed that it is very clear and informative. However, in terms of utility, the ratings are evenly distributed, which suggests that it may not have been immediately clear to all users.

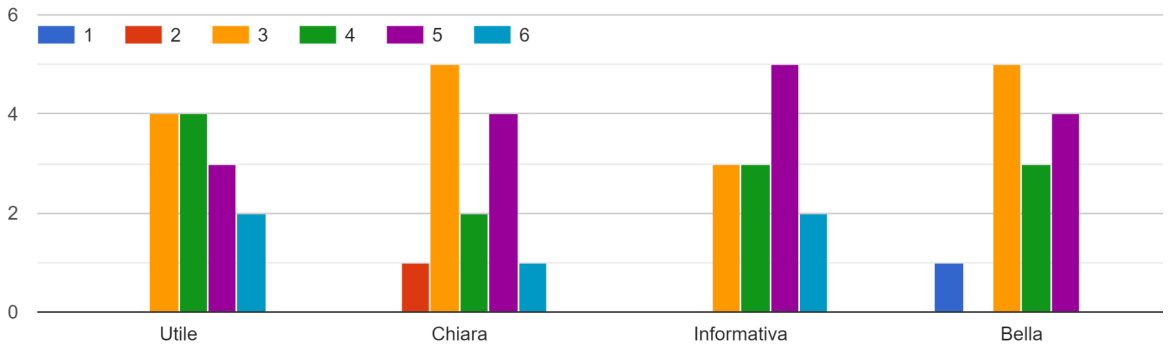
Per vedere l'infografica clicca QUI



Graph 1 (Alcohol and Drug)

From the results of this graph, it can be observed that it is very useful, informative, and quite visually appealing.

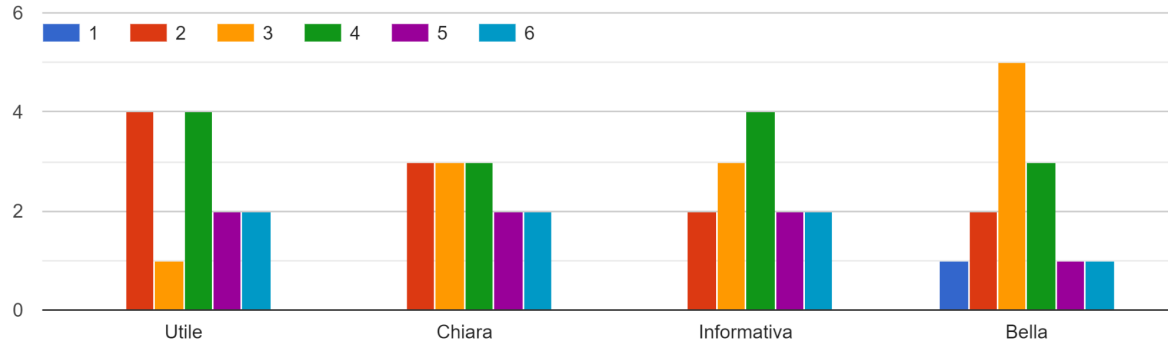
Per vedere l'infografica clicca QUI



Graph 2 (Alcohol and Drug)

From this second set of responses, we notice that the majority of users rated the utility and informativeness of the second graph positively. However, for the utility aspect, we observe slightly lower values.

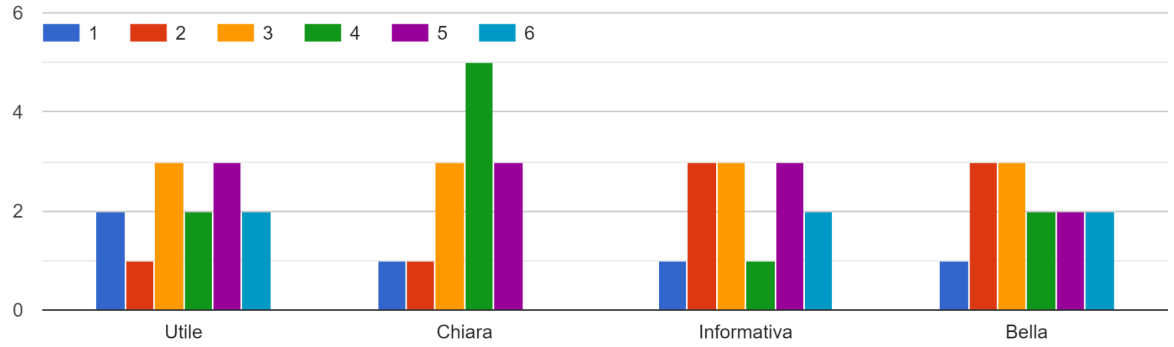
Per vedere l'infografica clicca QUI



Graph 3 (Alcohol and Drug)

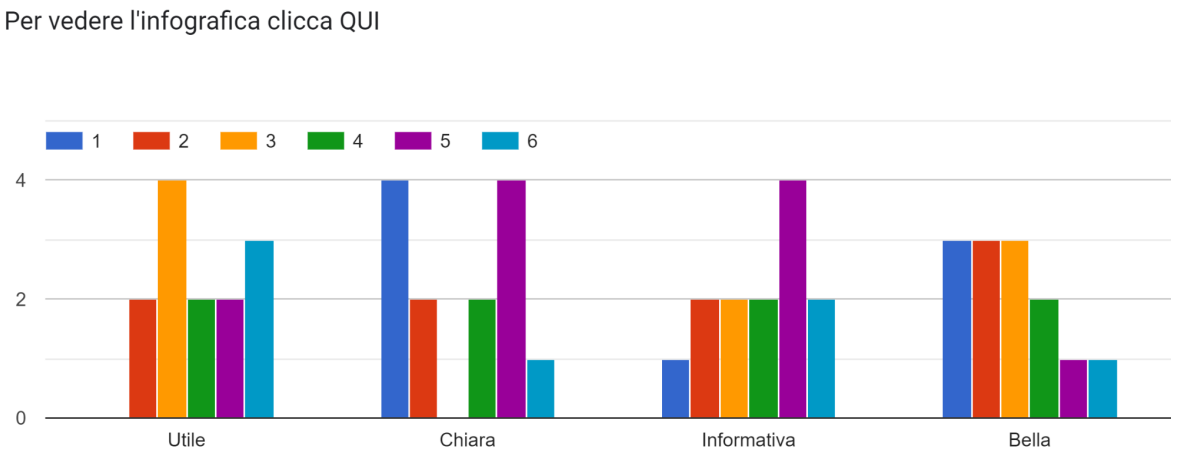
This graph has some issues regarding beauty and informativeness, but it is very clear.

Per vedere l'infografica clicca QUI

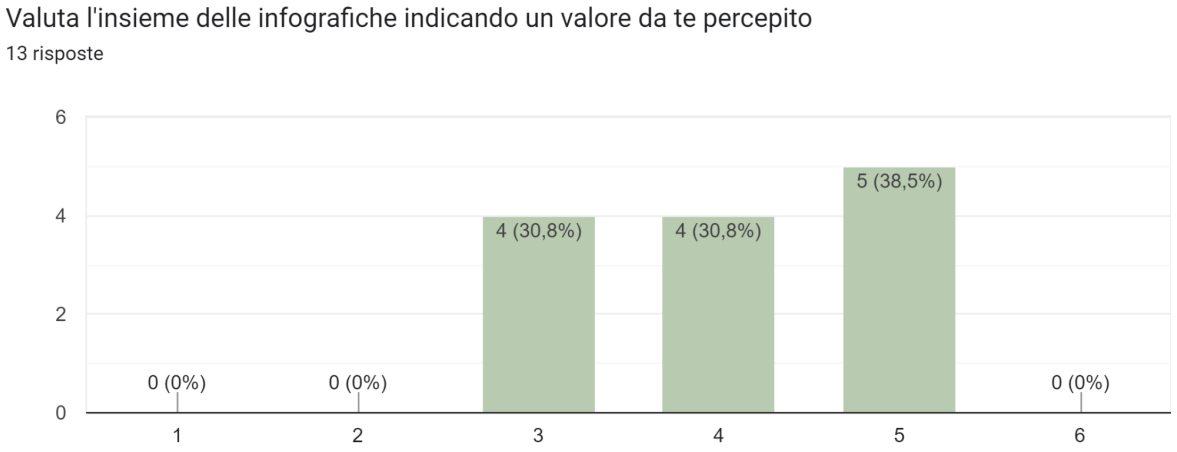


Graph 4 (Alcohol and Drug)

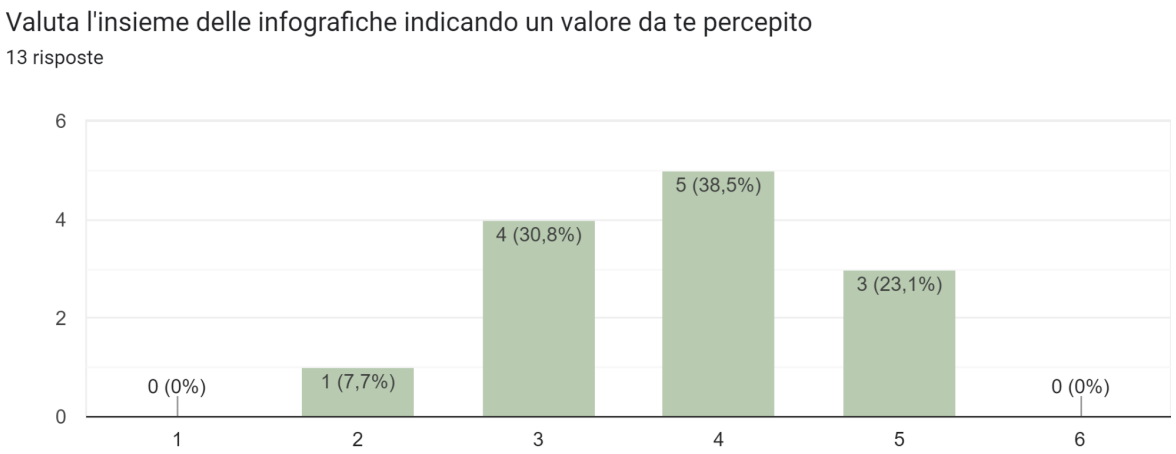
This graph appears to be the least appreciated based on the emerging results. The ratings for beauty and informativeness are mostly 3 or lower. However, in terms of utility and clarity, the ratings are more evenly distributed.



First group of infographics



Second group of infographics



Conclusions

Based on the psychometric questionnaire, we can conclude that our infographics are considered average. Some graphs are rated higher in terms of beauty, clarity, utility, and informativeness than others. However, we believe that individual differences play a key role in assigning scores, which is why this evaluation involves a relatively large number of users.

6. Future considerations and developments

Certainly one of the main future developments that we could implement in this project would be the integration of our data with external data such as sources of information like news or health data that allow us to better understand the correlation of disease trends.To make the dataset more informative and useful, and surely always updated for the new years and for new diseases that are perhaps less frequent but at the same time important.

Another feature that we could have developed is certainly the possibility of focus better on a single disease in order to study its trend and geographical diffusion as much as possible in order to be able to better understand the trend and distribution, correcting errors

reported by users in the user test phase to make interactive graphs more informative and easily understandable.

7. Bibliography

[1] *Matplotlib.pyplot#* (no date) *matplotlib.pyplot - Matplotlib 3.5.3 documentation*. Available at this LI

[2] Hrterhrter (2022) *Drugs use*, *Kaggle*. Available at: https://www.kaggle.com/datasets/programmerrdai/drugs-use?select=deaths-substance-disorders-age.csv

[3] Pant, M. (2022) *World deaths and causes (1990 - 2019)*, *Kaggle*. Available at: https://www.kaggle.com/datasets/madhurpant/world-deaths-and-causes-1990-2019

[4] Ritchie, H. and Roser, M. (2019) *Drug use*, *Our World in Data*. Available at: https://ourworldindata.org/drug-use