Mental Health Analyzer

Visual Analytics 2021/22

Felice Massotti 1699768
Daniela Moretti 1708855

INTRODUCTION

Mental disorders are health conditions involving changes in emotion, thinking or behavior. These are very common, every year the number of cases increases, fortunately they are treatable and improvement is possible. In our project, we analyze the correlation between healthcare resource and some socio-demographic status to find common patterns and behaviors of mental health disorders in different countries of the world, in order to give the possibility to anybody to intervene in a specific and accurate way. In particular, we focus on 9 mental health disorders. With regard to support public health professionals, we need to find some visual analytic solutions that are easy to be analyzed and understood. For this purpose we realized 6 types of visualizations:

- **Disorders mapping:** to get an initial view of the distribution of these disorders;
- **Parallel coordinates**: useful to find patterns and trends, analyzing disorders for each country and also to confront the values over time;
- **Bar chart**: for comparing the population categories(ages and genre);
- **Line chart**: for having a complete view in the years of the GDP values;
- **Scatter Plot**: useful to find possible clusters between the countries.

TABLE OF CONTENTS

01 Related Works

[1] *An Empirical Study of Chronic Diseases in the United States: A Visual Analytics Approach to Public Health:* from this paper we take the main inspiration for the dataset, in particular for the demographic characteristic like the population and the gender. For the visualization, we have used the choropleth map but in different countries of the world and in a more interactive way. And we take from them the inspiration to use the bar charts for the gender and the ages.

[2] A *framework for identifying similarities among countries to improve cross-national comparisons of health systems*: from here we got the idea of taking the population divided by three different age groups. As them, we compute the clusters and show them in the map.

[3] *Spatial analysis of COVID-19 and socio-economic factors in Sri Lanka*: we saw how they use K-Means to compute the clusters and it also suggests to consider some parameters related to the economy, like the GDP.

[4] *Visual Analytics in Effects of Gross Domestic Product to Human Immunodeficiency Virus Using Tableau*: we got the idea to how represent the capital health expenditure of GDP in relation with the health values.

[5] *Big Data Visual Analytics with Parallel Coordinates*: it suggests to color the line of the parallel coordinates according the colors of the cluster.

02 Dataset

The datasets of our project can be divided into two macro-categories:

- **Mental health data:** the csv *Number with a mental or neurodevelopmental disorder by type* taken on the Our World in Data website. Inside it, there are the numbers of the population of each country in the world suffering of a certain mental disorder. The categories analyzed are the nine shown in the figure below.
- **Socio-Demographic data:**
  - *Genre:* It contains the percentages of men and women with any mental health or development disability disorder for each country (Our World in Data website)
  - *Age of population and current health expenditure (% GDP):* taken by World Bank Data website.

After analyzing all of them, we noticed that some categories were missing a lot of data, which make accurate analysis impossible, so we have decided to remove some of them.
So, we decided to focus on the period that goes from 2010 to 2019 and exclude some countries (grey in the map).
The final dataset contains 16 columns and in each row, we find a country and a particular year, for a total of 174 countries.

02 Dataset

A small part of preprocessing was done in the analytical part, where for the calculation of the clusters, we decided to remove the column concerning the population (redundant) and the percentage representing the male and female gender was transformed into the exact number of people suffering of mental disorders .

Genre Dataset

Entity	Code	Year	Prevalence - Mental	Prevalence - Mental
Afghanistan	AFG	2010	16.9662887	17.79714853
Afghanistan	AFG	2011	16.88636398	17.73028509
Afghanistan	AFG	2012	16.80756124	17.68164291
Afghanistan	AFG	2013	16.74080658	17.63762032

Ages and GDP Dataset

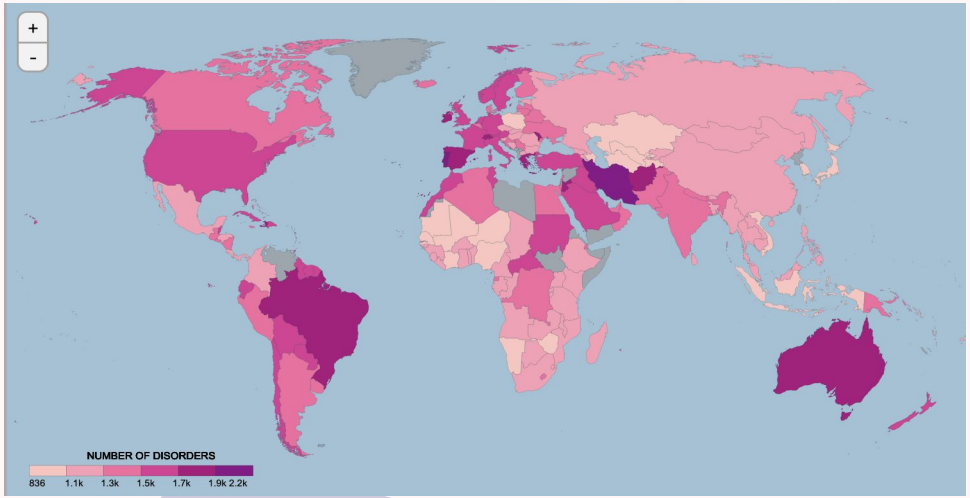
Year	Code	Name	Current health exper	Population, total	Population ages 15-(	Population ages 65 a	Population ages 00-1
2010	AFG	Afghanistan	1315839195	29185511	14444996	679477	14061038
2011	AFG	Afghanistan	1320251998	30117411	15041824	709103	14366484
2012	AFG	Afghanistan	1373040989	31161378	15743958	743602	14673818
2013	AFG	Afghanistan	1616793039	32269592	16519812	780991	14968789

Mental Disorder Dataset

Entity	Code	Year	Depressive	Anxiety	Bipolar	Eating	Schizophrenia	Attention	Conduct	intellectualDisability	Autism
Afghanistan	AFG	2010	1025342.743	1220428.929	157159.8504	30559.81382	41696.592	352855.4018	259433.4595	1435306.873	89471.96582
Afghanistan	AFG	2011	1071820.071	1276328.293	164688.5498	32607.98363	43611.81019	366302.7842	267299.0395	1451963.801	92924.54721
Afghanistan	AFG	2012	1119166.732	1335187.569	172222.8072	34627.64153	45663.48173	379388.156	274450.3668	1464608.547	96356.30432
Afghanistan	AFG	2013	1165521.341	1394572.289	179580.9725	36735.93502	47742.40154	392237.8945	281401.0686	1476663.841	99763.54396

Visualizations and Analytics

Disorders mapping 3.1



A GeoJson file was used to create the map with the different paths for each country.

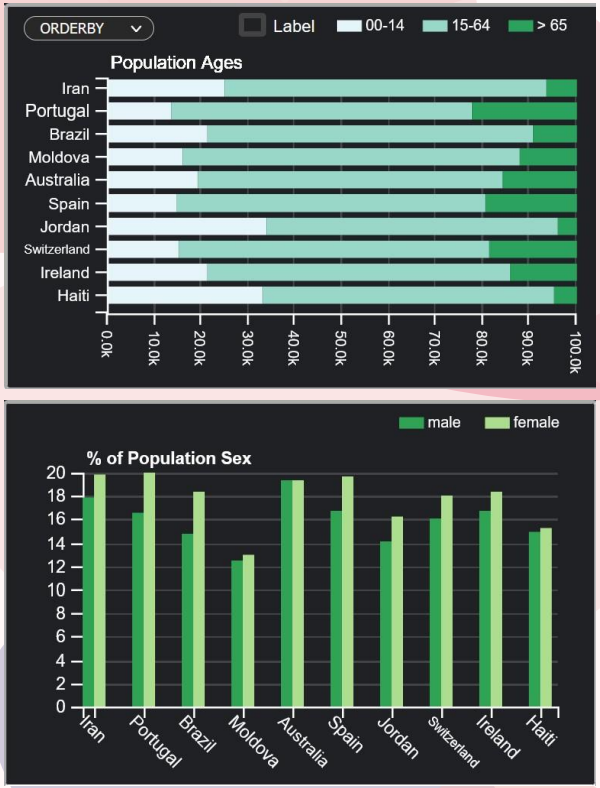
It is called a choropleth map where regions, states or geographical areas are coloured, using different intensities, based on the aggregation of some data that fall within certain ranges of values. In our case, the data that we will take into consideration are those of mental disorders (by selecting them from the menu)

We decided to take the shade that reaches purple because this is the colour used for the World Mental Health Day.

The legend explains how the different shades of colour fall into specific ranges. If the calculation of the data is on *Total population*, we will have an additional element with a stroke yellow dashed that represents the outliers of the dataset.

Bar chart

3.2



The aim of bar charts is to represent data regarding the three age ranges and the male/female percentages. This type of visualization is usually used to compare different categories (a max of 10/11 countries).

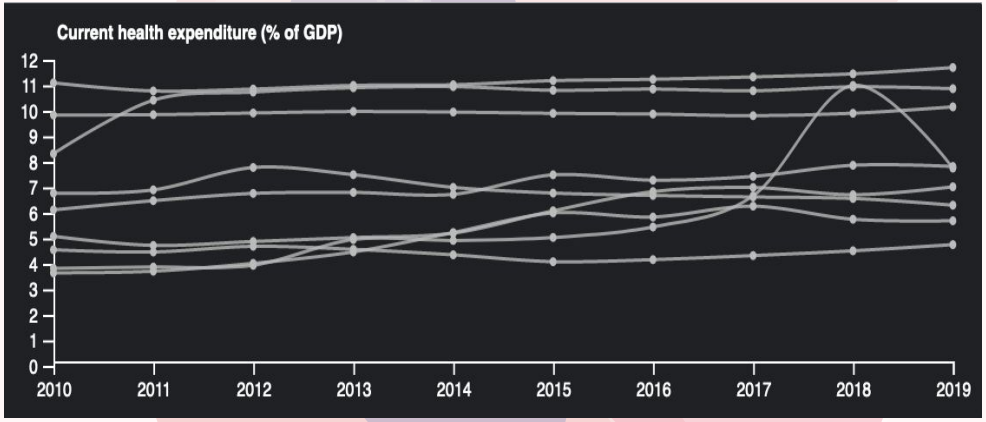
- *Horizontal stacked bar charts*: used for the ages of the population. In the case of 100k inhabitants a labels will be shown, in the other case a tooltip is implemented
- *Vertical multi-series bar charts*: used for the male/female percentages of the population affected by mental disorders.

For all the information concerning the demographic part, we have chosen the green color.

3.3 Line chart

A line chart is used very often to show information that changes over time. Inside, it is possible to find lines that will represent the evolution in the range from 2010 to 2019 of the level of current health expenditure expressed as a percentage of GDP for the countries that will be analyzing at that particular moment.

We choose for this visualization a grey color in order not to exceed and avoid to create confusion with the others chart.



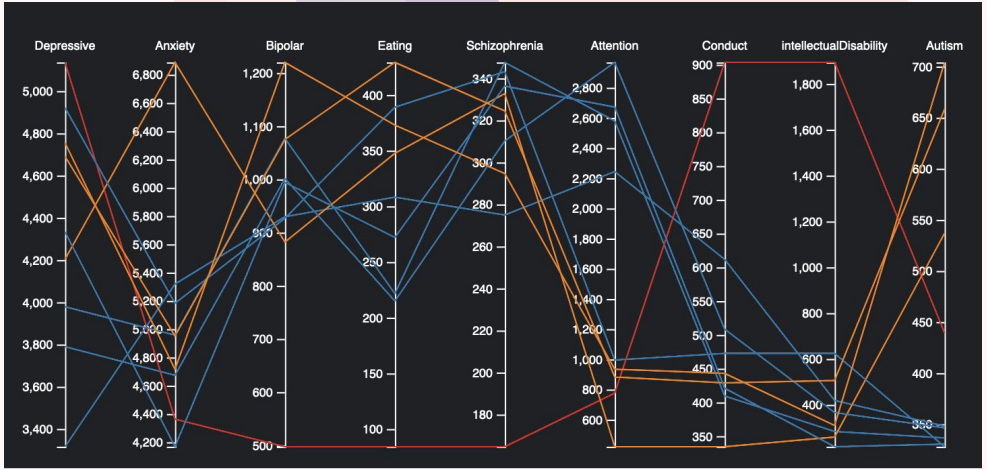
3.4 Parallel coordinates

The parallel coordinates is used for plotting multivariate, numerical data.

They are ideal for comparing many variables together and seeing the relationships between them.

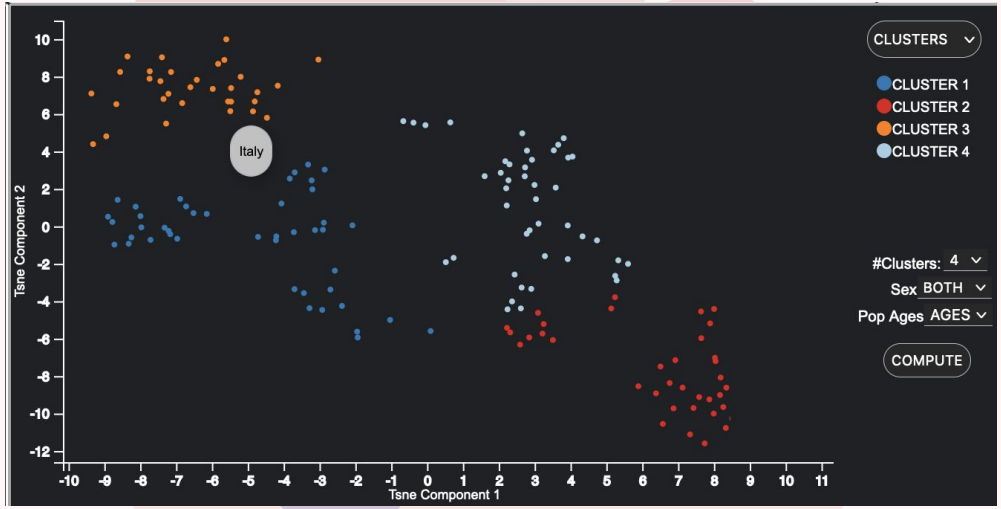
We have used it to represent the disorders.

Brushing highlights allows to impose a range of value for each disorders, highlighting only the countries that respect this constriction.



Scatter plot

3.5



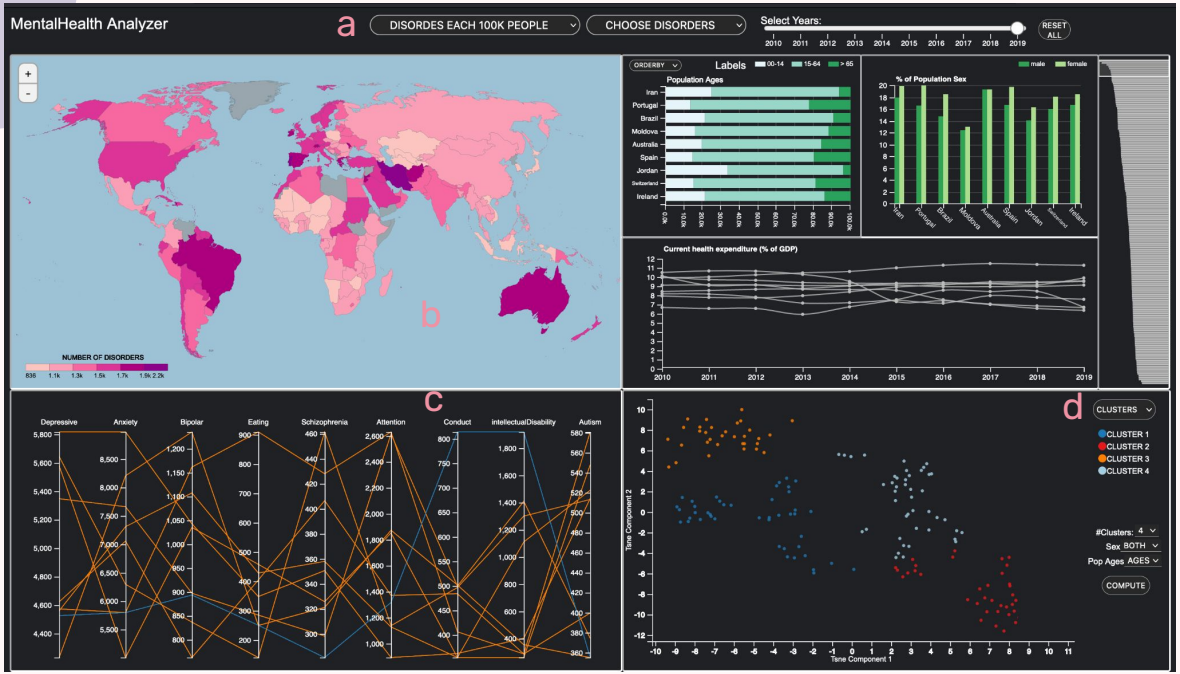
Scatterplot uses a collection of points placed with Cartesian coordinates to display the values of two variables.

It is used to detect if a relationship or correlation between the two variables exist.

With K-means algorithm is possible to compute clusters. Through selection options we can choose disorders, years, genre and age range and updating the clusters according to the selected values.

3.6 Visualizations coordination and interaction

- All the visualizations are fully linked and self adaptive.



- Menu: allows to choose different values from selection options.(a)
- Map: it's possible to add manually countries to each visualization by clicking on them(b)
- Parallel Coordinates: when the brushing is applied all the other visualization are updated.(c)
- Scatter Plot: when zoom on a specific cluster all the visualizations show only the countries of the selected cluster.(d)

3.7 Analytics

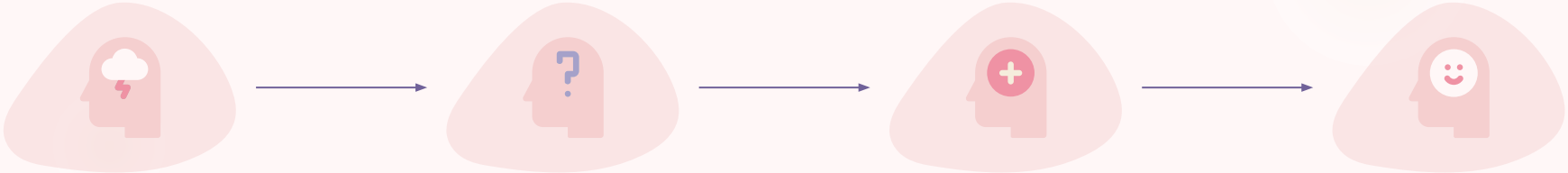
In the analytic part, clusters are created that group different countries to find similarities and differences that can help the user analysis.
This whole part was created on PythonAnywhere to develop a back-end service that through a POST and GET calls, return the data elaborated.

Given the large amount of data, it is done initially a dimensionality reduction through t-SNE and then applies K-Means .

T-distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear method that attempt to preserve the local neighbors by trying to optimize a stress function. All this allows to have a greater separation between the different samples and to create more defined clusters.

For the creation of the clusters is been used the well known K-Means algorithm, but instead of the classic algorithm we have used **K-Means++**. The main difference between the two is the initial choice of centroids (that is not randomic).
The disadvantage of this algorithm is that, we need to choose previously the number of clusters.

Live Demo



05

Conclusion

Our project allows to :

- Analyze the correlation between countries finding some possible patterns;
- Analyze the situation in a specific country, individuating trend between the different disorders both from the spatial point of view and from the temporal one.

Future Works

- When the site Our World Data will update the dataset with the last 2 years, it will possible to analyze how the situation is changed after the Covid pandemic;
- New characteristics could be considered: for example with the specific number of cases for each age groups and the educational level.

THANKS FOR THE ATTENTION