What Makes Us Healthy?

| João Tiago Aparício  Instituto Superior Técnico  Lisboa, Portugal  joao.aparicio@tecnico.ulisboa.pt | Diana Lopes  Instituto Superior Técnico  Lisboa, Portugal  dianamlopes@tecnico.ulisboa.pt | Jorge Marques  Instituto Superior Técnico  Lisboa, Portugal  jorgemmarques@tecnico.ulisboa.pt |
| --- | --- | --- |

# INTRODUCTION

Human health quality has been on the rise since humans started learning more about how our habits and environment impact health. The understanding of how theses factors relate to health is quite valuable as we may use this information to make decisions form policies to daily habits. We did not find any tool to visualize this kind of data.

This project was developed with the objective of discovering which of the selected habits influences a set of health indicators, in each country each year and how they evolve. We propose an artifact that supports the visualization of the relationship between several habits and health metrics. For this, we are going to focus on what can, potentially, affect our health. Even though the study is focused on OECD countries, the datasets we used had data about many more countries. So, we have a worldwide scope.

In terms of definitions, we will be using several indicators to define both the factors and health. To define health, we use: the percentage of the population aged 15 and older, that are overweight or obese population, smokes daily, that consumes alcohol daily, the number of deaths by cancer, suicide rates, life expectancy at birth and at age 65.

On the other hand, the variables used to correlate to the health will be work-related: based on wages, employment rates and hours worked; air pollution exposure; adult education level, the number of adults that concluded below upper secondary education; the country’s social spending per capita and GDP per capita (Gross domestic product per capita).   
We hope to find what factors are more relevant in terms of what constitutes a healthy life. This visualization is entitled “What makes us healthy?” as it aims to answer just that.

At first, we had 6 questions we wanted our visualization to answer:

1. Does a better wage mean a healthier life or a longer life expectancy?

2. What is the optimal number of hours to work that lead to a healthier life or more life expectancy?

3. What is the relationship between, more people working and being healthier and live more?

4. How does adult education influence our health?

5. How does the Air and GHG emissions affect our health? (Based on life expectancy and suicide.)

6. Does a bigger social spending in general influences people to live more and suicide, smoke and drink less?

But we added several others because our visualization has the power to answer many more questions than the ones proposed. A few examples are:

1. Does a bigger social spending or a higher GDP in general influences people to live more and suicide, smoke and drink less?

2. How does self-reported happiness correlates to Alcohol consumption, Smoking habits and Suicide rates?

3. Are people making more money on average in Portugal or in the USA? And how are the corresponding rates of obesity in each country?

4 What is the strength of the correlation between Alcohol consumption and Happiness in 2010? And how does it change along time?

How these questions are answered is better explained in the **4.3. Demonstrate the Potential** section. The idea is to answer the most questions with our visualization.

We also added 1 question (7) in order to have a more complete visualization.

The tool developed allows for the visualization of each factor described in terms of, distribution, country and relationship between each health variable, as well as the relationships strength. It is quite intuitive in the way we have interactions that help relate each visualization medium to one another. We can tool-tip everything to get the concrete actual data it is meant to show.

# Related work

We noticed there are a lot of visualization and articles about health factors around the world in the internet. Mostly about life expectancy. However, not many try to explain or show if these metrics are related to the habits in each country in a quantitative approach.

Next, we enumerate some websites from where we got inspiration for the project.

In Life Expectancy by Mark Roser [1], we found this article interesting because it mentions the evolution of life expectancy around the world along the years. But it doesn’t relate this evolution with people habits. So, we we’re inspired to do more. On top of this, it has some well put together visualization from where we got some inspiration. Mainly for our choropleth map.

In Health Data [2] we found numerous visualizations and data about health around the world. Here we mainly got to see different ways of showing health related data.

In the OECD [3] website we got most of our data. The reason why we chose to include this website in this section was because, on top of all the data available in this website, there we also found some visualizations of said data. So, we were not only getting the data but also an idea on how to represent it on our visualization. Overall, we consider this website the main contributor and source of inspiration of our work.

# 3. The data

# Visualization

## 4.1. Description

Our visualization has 4 idioms. All of them have interaction between each other. The data on these changes accordingly to the time slider, and the factor selector.

In figure 1 it’s possible to see the layout of the visualization.

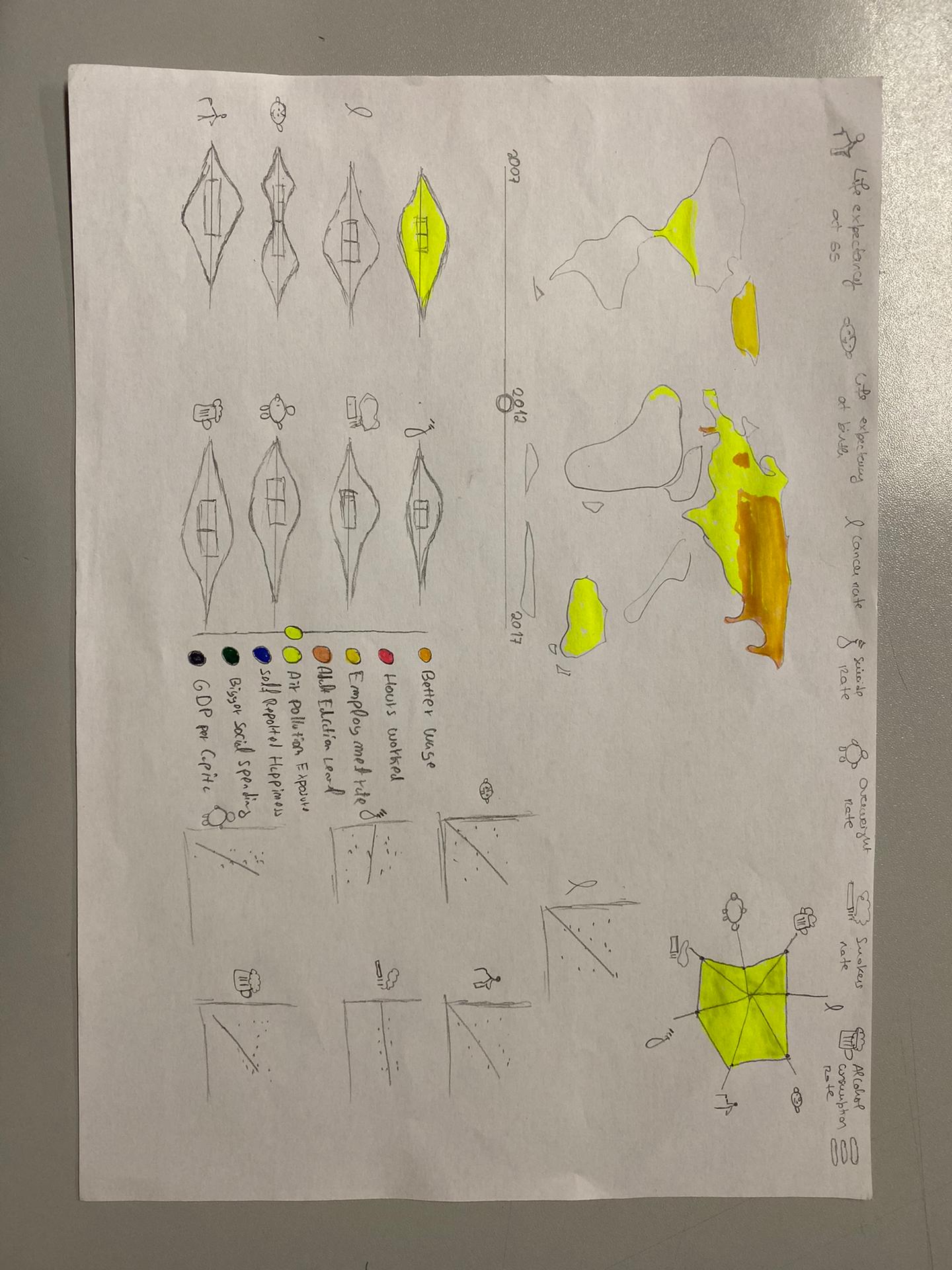


Figure 1

In order, to better understand the idioms and the interactions they have between each other, in the context of this report, each idiom is numbered and described:

1. Choropleth map

This map will change the color based on the health influencers selected and the year selected.

The user can then use the map as a slicer, where he can click a country to highlight the data being shown to that country.

1. Scatter plot with regression line

We’ll have 1 scatter plot per each variable of health being analyzed. This means we have 7 different scatter plots as small multiples. On the x axis we have the health factor and on the y axis we have the habit selected on the slicer described above. The point of the country selected on the choropleth map will have higher luminance so the user can identify the country selected and compare it to the rest of the countries. You can also click or hover over a dot to see what country it is.

The lines represent the regression between the 2 variables.

1. Star plot

In this plot we’ll have the correlation coefficient between each health variable and the variable selected. This is used to understand how strong the correlation is between each pair of variables.

1. Box plot

There are 8 box plots, in the form of small multiples. Each box plot will show the distribution of each health variable and the variable selected. This visualization is important to understand what is the distribution like each year and who are the outliers.

1. Health influencer slider

You can use this to select the health influencer you want to analyze.

1. Health influencer label and Pictogram caption

This serves as a label for the icons close to the plots, so the user can understand which data is in the plots. The Pictogram caption associates the pictograms with their meaning.

1. Years slider

This allows the user to select the year he wants to analyze.

## 4.2. Rationale

1. Choropleth map

Map -> Color: filling each country with the color of the selected variable.

Channel -> Color -> hue: a higher value for the hue represents a higher value of the selected variable.

Channel -> Color -> lightness: if a point from a country is selected on a scatter plot, the lightness goes up on the map.

1. Scatter plot

Marks -> Point: represents a pair (health variable, health influencer)

Marks -> Line: represents the linear approximation of the scatterplot between the 2 variables.

Channel -> Position: represents the position of the point in a Cartesian axis.

Pictograph -> Represents the respective health variable being correlated with health influencer

1. Star plot

Marks -> Point: represents the correlation coefficient normalized between 0 and 1.

Channel -> Color -> hue: The color of the area between the points in the start chart represents the variable being analyzed.

Pictograph -> Represents the respective health variable being correlated with health influencer

Channel -> Color: color represents the selected variable.

1. Violin plot

Pictograph -> Represents the respective health variable

Curve -> Represents the distribution of each of the health variables and the selected health influencer

Size -> Is proportional to cardinality countries of the value interval

Channel -> Color: the color of the first violin plot represents the selected variable

1. Health influencer slider

Channel -> Position: represents the health influencer being displayed

Channel -> Color: represents the health influencer being displayed

1. Health influencer label

Channel -> Position: represents the year being displayed

Channel -> Color: represents the relation between the color with each influencer (word on its right)

1. Years slider

Channel -> Position: represents the year being displayed  
Channel -> Color: represents the relation between the color with each influencer (word on its right)

1. Health influencer caption

Channel -> Color: represents the relation between the color with each influencer (word on its right)

## 4.3. Demonstrate the Potential

# 5. Implementation details

Completing this project was no easy task. We faced multiple challenges along the way. This section enumerates these challenges and explains how we overcame them.

1. Format the data is stored

At first all our data was in the csv format. We quickly realized it’s much easier to work with json than with csv when using d3. So, we found ourselves with a “useless” data format. This was easily fix, since we only had to transform the csv file into a json file through the use a python script.

1. Data chosen at first was not complete

When we first picked our data, back in checkpoint 1, we didn’t check if all the fields were complete. We later realized that it was not, as is usually expected. This is one of the main lessons learned from doing this project.

To address this problem, we took 2 steps:

* 1. We limited the interval of years being analyzed to the interval we had more data. In this case we limited the data to the years between 2012-2018.
  2. How did we fix missing spots?

1. Data chosen at first had almost 0 values

This problem is similar to the previous one. Why did we choose to split 1 problem into 2? Because, while some data first collected had some missing values, some had almost no values at all. From our point of view this is a completely different problem that required a different solution. So, it’s deserving of it’s on section.

In this case, the data first chosen for the values of self-reported happiness had almost 0 values. So, we had to find a solution. We opted for a simple one, and simply decided to find another source with similar kind of data. More specifically we changed the OECD data source for a …

1. Choropleth map

# 6. Conclusion and Future Work

# References

|  |  |
| --- | --- |
| [1] | Max Roser (2019) - "Life Expectancy". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/life-expectancy' [Online Resource] |
| [2] | [Online]. Available: http://www.healthdata.org/results/data-visualizations. [Accessed 16 12 2019]. |
| [3] | [Online]. Available: https://data.oecd.org/. [Accessed 16 12 2019]. |