What Makes Us Healthy?

João Tiago Aparício
Instituto Superior Técnico
Lisboa, Portugal
joao.aparicio@tecnico.ulisboa.pt
97155

Diana Lopes
Instituto Superior Técnico
Lisboa, Portugal
dianamlopes@tecnico.ulisboa.pt
94121

Jorge Marques
Instituto Superior Técnico
Lisboa, Portugal
jorgemmarques@tecnico.ulisboa.pt
94012

1. 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 more educated on average in Portugal or in the USA? And how are the corresponding rates of suicide 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.

2. 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 Global Health by Esteban Ortiz-Ospina and Max Roser [2], we found a very similar visualization to what we wanted, but it did not have the day to day habits approach. And had more to do with mortality. The focus was on the economic indicator and the generic indicators used were not what we had in mind.

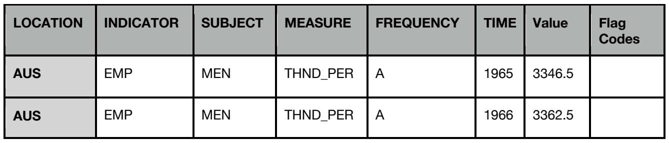
In Health Data [3] we found numerous visualizations and data about health around the world. Here we mainly got to see different ways of showing health related data.

In the OECD [4] 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

All the data used in this project is static. The time span of the visualization is from 2000 to 2017. It has a total of 180 different countries and 1772 data samples with many missing values.
The original data for was on CSV format. We used OECD data. For the following variables: Life expectancy at birth [5], life expectancy at 65 [6] , suicide rates [7], daily smokers [8], alcohol consumption [9], overweight or obese population [10] and deaths from cancer [11], Average wages[12] , employment rate[13], hours worked[14], adult education level [15], social spending [16], gross domestic product (GDP) [17] and life satisfaction [18] and air exposure [19].

Here is an example of how a raw, original dataset extracted from OCDE:



LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value	Flag Codes
AUS	EMP	MEN	THND_PER	A	1965	3346.5	
AUS	EMP	MEN	THND_PER	A	1966	3362.5	

One of the challenges we faced was that it's much easier to work with json than csv in d3. We used some information from some columns such as "MEASURE" to get the relevant data and then got rid of the excess columns. We generated two data files both in json format using data as the one presented:

```
{
  "alcohol": 10.3,
  "average_wages": 35.6,
  "cancer": 201.8,
  "employment_rate": 72.4,
  "happiness": 7.5,
  "life_at_birth": 81.8,
  "life_at_old": 20.4,
  "location": "AUS",
  "obese": "",
  "pollution": 1699.9,
  "smokers": 15.4,
  "social_spending": 6997,
  "suicide": 10.9,
  "time": 2010,
  "hours_worked": 1699.94,
  "gdp": 943228.4205,
  "education": 73.207256
},
```

We had to do many outer joins in order to not loose a lot of important data.

And another with data corresponding to the correlation between each pair of variables each year. These data was totally derived by us.
Here we present a sample:

```
{
  "Time": 2007,
  "Var": "Alcohol",
  "Alcohol": 1,
  "Average Wages": 0.5279264692386093,
  "Cancer": 0.6292916767297805,
  "Employment rate": 0.2995033687080575,
  "Happiness": 0.12690894509802775,
  "LifeAtBirth": 0.2416125432895052,
  "LifeAtOld": 0.3479615347181521,
  "Obese": 0.04173287021184598,
  "Pollution": -0.4315135820552212,
  "Smokers": 0.579468448720404,
  "Social spending": 0.24844971118745038,
  "Suicide": 0.48388640267282473,
  "Hours worked": -0.4315018849668763,
  "GDP": -0.15514749842573902,
  "Education": 0.5419706341451226
},
```

Difficulties

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. This was easily fix, since we only had to transform the csv file into a json file through the use of the outputs in the Data integration tool.

Data chosen had loads of missing values: When we first picked our data, back in checkpoint 1, we didn't check if all the fields were complete. We search for hours on end for other sources of data, but without success. So, we worked on using a lot of outer joins in the data integrator. This leads to some inconsistencies in the animations as the datapoints are literally referring to different countries and have to disappear. We also 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 2007-2017.

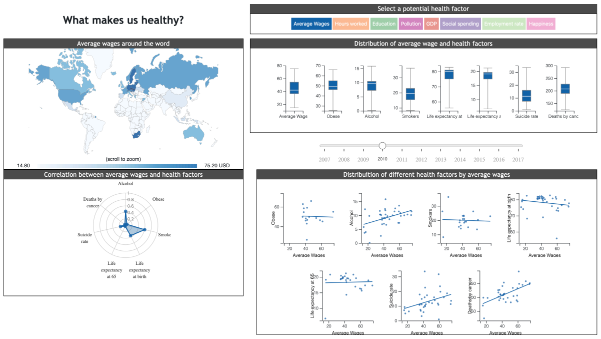
We overcame a big difficulty with the data integrator as there was an unidentified bug with points and comas. But we solved it by correcting all the values with the original dataset using R.

4. 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.

The concretization of this dashboard is the following:


What makes us healthy?
Select a potential health factor
Average Wages
Average wages around the word
(scroll to zoom)
14.80
75.20 USD
Distribution of average wage and health factors
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Correlation between average wages and health factors
Alcohol
Obese
Smoke
Deaths by cancer
Suicide rate
Life expectancy at 65
Life expectancy at birth
Distribution of different health factors by average wages

The dashboard

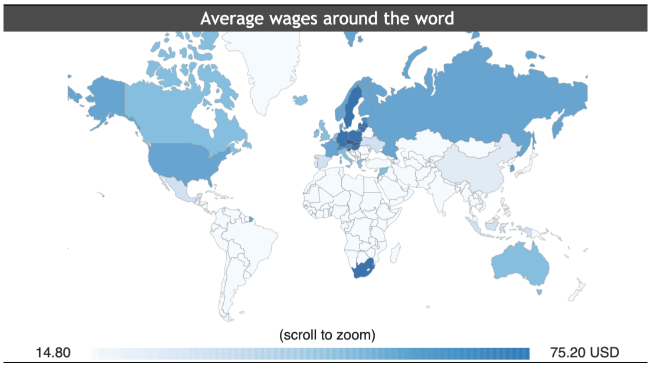
In order, to better understand the idioms and the interactions they have between each other, in the context of this report, each idiom is numbered and described:

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. This will show the data in the box plots and highlight the corresponding points in the scatterplot. This also shows a tooltip showing the actual data and the name of the country selected.

This visualization is important to understand the geographic nature of the data. By knowing the countries location, some data can make more sense as it gives some context. A zoom feature was also added so that the user can see even the smallest countries and to use the space more efficiently.



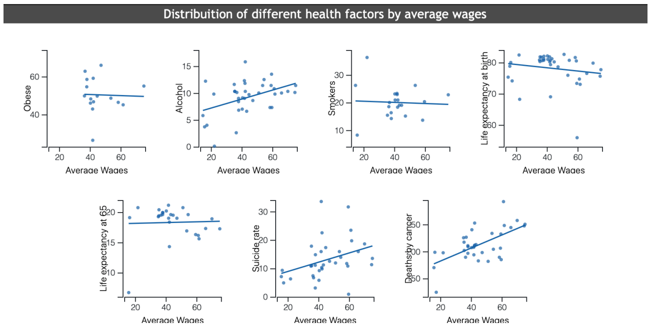
The Choropleth map

2. 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 as well as the data it shows.

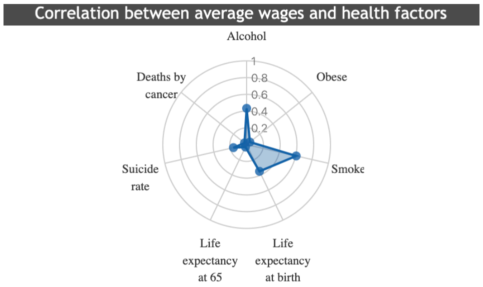
The lines represent the regression between the 2 variables.



The Scatter plots with regression line

3. 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. It is quite important to understand how true our assumptions about the relationships are. By hovering the small circles, the user can see the actual value of the correlation.

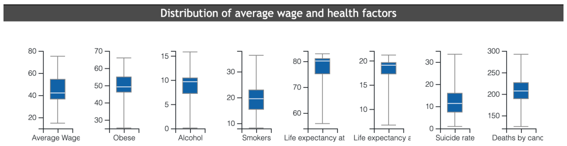


The Star plot

4. 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 the distribution like each year is and who are the outliers. As we can select on the map or on the scatterplot the circles or shape from the corresponding country.

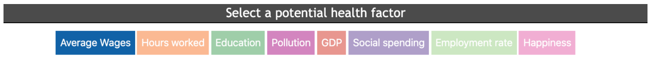
By hovering each box plot, the user can see some statistical values of each distribution.



The Box plots

5. Health influencer selector

You can use this to select the health influencer you want to analyze. The color is more intense as the factor is selected. And you can know what factor you are evaluating at any time because the hole dashboard turns to the color of that factor. Using color for categorical data in this way is not an impediment for people with color blindness, as it is not the only way to understand if the variable is selected at any time.



Health influencer selector

6. Years slider

This can also be seen as a visualization itself; it allows the user to select the year he wants to analyze. Using it can be useful to understand the variations on the data between any 2 years.

Years Slider

4.2. Rationale

1. Choropleth map

The use of a choropleth map was highly debated. As it occupies roughly a quarter of the screen real estate and has a lot of unused space due to the nature of the map chosen. We also hypothesized the use of a semantic map, but because of the lack of many missing values, the map turned out too deformed. We also thought about the use of a tree like representation where countries with a higher value would appear larger, but ultimately, we wanted to **give the user the sense of the geographic positioning** of the data.

Map -> Color: filling each country with the color of the selected variable.
Marks -> Areas: Areas representing each country region.
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.

2. Scatter plot

We chose scatterplots because we wanted the user to get the sense of the real data, but also wanted to **show the tendency** of the data in order **to conferee the sense of relationship** to the visualization. The use of this visualization was due to the fact that it is really easy to show the outliers. We pondered on the use of other visualizations such as the sorted bar charts, but that was implied too many lines and was not understandable, but if we binned the data that was too reductive. Line charts were not an option as countries do not have an order. We also thought about using slope graphs in order to understand the change of the data but that that implied selecting the countries and that did not make sense as the main idea was to establish the regressions and their corresponding strengths, and we did not have enough data to do that.

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.

3. Star plot

The use of this idiom was not immediate. The first idea was to use an ordered bar chart, and that would have worked. But we wanted to have the look of **strength and weakness**

analogy as that is exactly what we are trying to represent. The string correlations and the weak ones.

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.
Channel -> Color: color represents the selected variable.

4. Box plot

Initially we thought about the use of a **violin plot**. But as we showed the prototype to several people, many of them did not understand what was being shown. So we settled for the next best thing, the box plot. We wanted to showcase the distribution of the data, but we did not want a histogram as we want to **showcase some statistical data on the distributions**.

Marks -> Line: represents the median, Q1, Q3, minimum and maximum values.
Channel -> Color: the color of the first violin plot represents the selected variable

5. Health influencer label

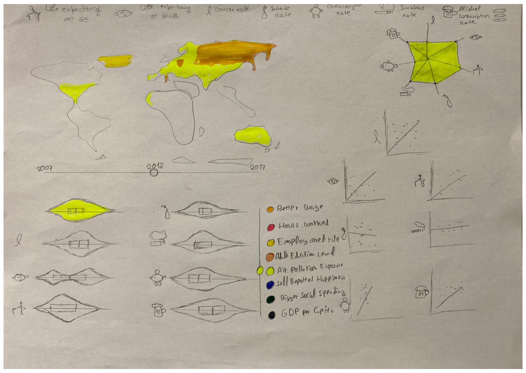
Channel -> Color: represents the relation between the color with each influencer (word on its right)

6. Years slider

Channel -> Position: represents the year being displayed

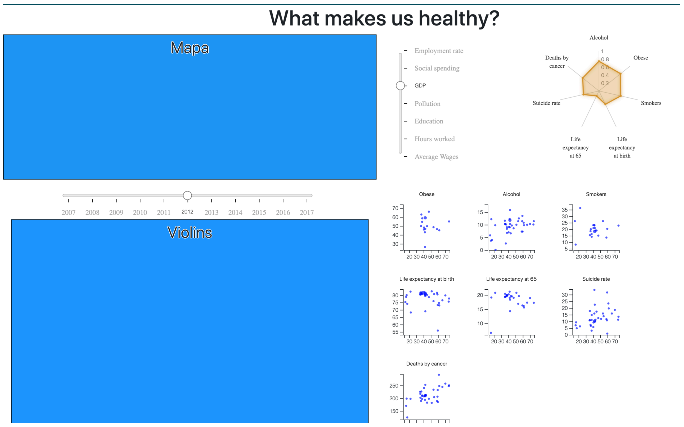
Evolution

Initially, the first sketches were as follows:



The first sketch

The first idea was to represent pictographs instead of text for the description of the axis in each visualization. But as we asked several users, they did not like the idea of having a legend for that and preferred having the text. So we left that idea. We also changed the positioning of the idioms because of the reasons described earlier. In the description of the visualization. We also changed the violin plots for box plots as people did not understood them with much ease.

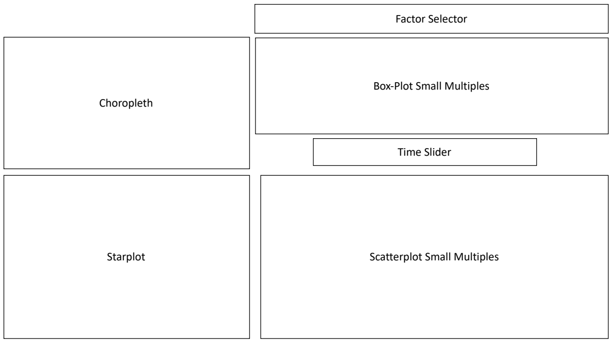


The first functional prototype

We also learned that using a slider to choose different factors did not make sense as there is no order between them. There was no intention of making several factors at the same time because the units are different and so they are not comparable.

For the layout, we positioned the factor selector on the upper left corner because it is a fairly standard place to position your selectors. On the other hand, we positioned the time slider in a more central position because we want the eye of the user to be closer to the center, but not quite in the center. We want to help the user by moving it closer to the boxplots and scatterplots. As these are the most relevant to answer the questions we proposed. We want the user to immediately notice the changes on distribution and relationship year after year. Making it easy to understand what has changed. On the other hand, we put the star plot directly next to the scatterplots as the two idioms are closely related. And the map is right next to box plots as we want the user to immediately notice the circle appearing in the box plot as he or she selects hovers over each country.

The positioning of each element of the dashboard is the following:



The dashboard layout

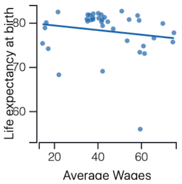
So the final Dashboard is as presented in section 4.1.

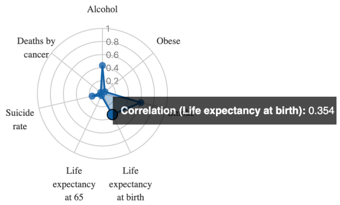
In terms of data complexity and scalability. Our data was quite multidimensional, and so we were obliged to do make choices in terms of the relationships we showed. The idea was to show only the relationships that mattered to the topic chosen. It was hard to deal with the fact that our data had 2 primary keys in a relational perspective. Each datapoint had one year and one country. And so we had to give up the idea of comparing several countries in slopegraphs so that it made sense with the overall idea of the relationships on a global level.

4.3. Demonstrate the Potential

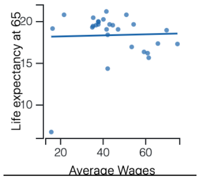
To showcase the **strength of our prototype**, we will be using several questions we proposed in the first checkpoint as well as others we proposed on the following checkpoints as we slowly understood the power of this dashboard.

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



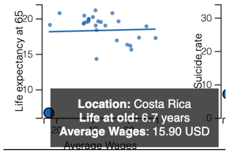


The data shows us that there is in fact a negative coorelation between the average wage and the life epectancy at birth. The coorelation coefitient is given by hovering the circles on the starplot. This is not a particularly high coorelation but is not statisticaly insignificant.



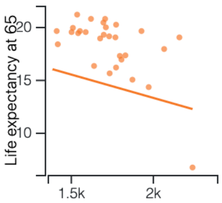
On the other hand, the average wage seems to a slight positive correlation between the life expectancy at 65 and the average wage. But the star plot tells us that that correlation is quite weak.

But that is not the hole story, there is an outlier, who is it?

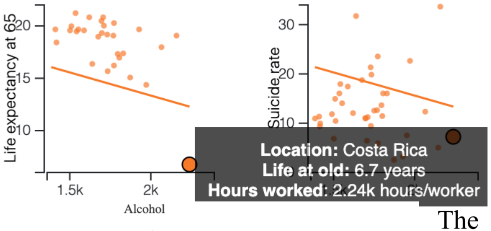


To answer that we can just tooltip the circle. The outlier is Costa Rica.

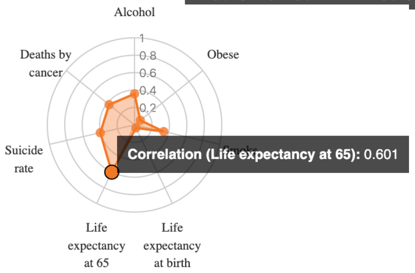
2. **What is the optimal number of hours to work that lead to a healthier life or more life expectancy?**



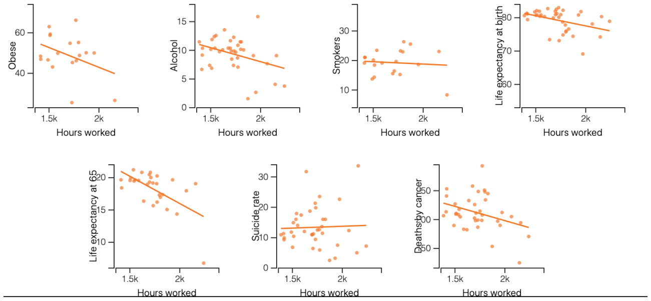
There is a negative relationship between hours worked and life expectancy at 65. This leads us to believe that working the minimum number of hours will lead to the maximum lifespan.



Except in Costa Rica.

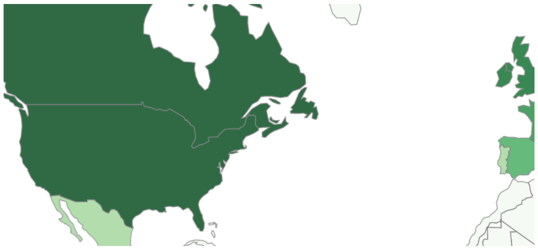


The strength of this correlation is quite high as the correlation coefficient is 0.601.

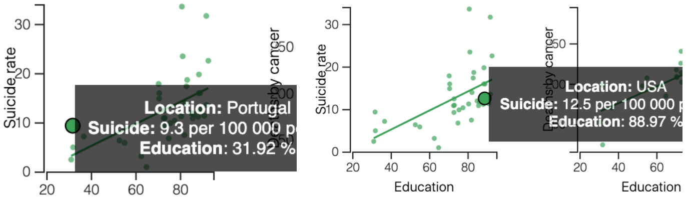


Overall, the data tells us that the relationship between the health variables and the number of hours worked is negative. Meaning that the more hours we work the less healthy we are. But as we have seen in the star plot, not all the correlations are strong.

3. **Are people on average more educated in Portugal or in the USA? And how are the corresponding rates of suicide in each country?**



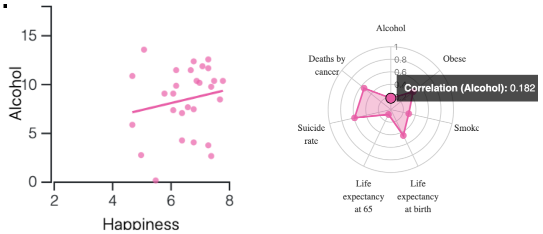
People in the USA are clearly more educated on average than in Portugal.



The suicide rate tends to be strongly correlated with the education. And There is the natural tendency that the USA has a higher suicide rate because of just that.

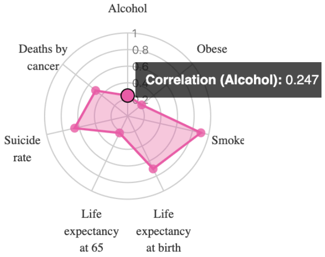
4. **What is the strength of the correlation between Alcohol consumption and Happiness in 2010? And how does it change along time?**

5.



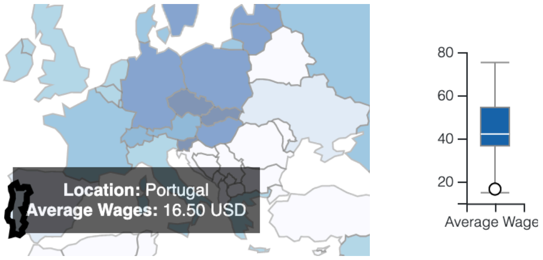
The data shows a positive relationship between alcohol consumption and happiness, even though that correlation is not too strong in 2010, taking the value 0.182.

Fast forwarding to 2017, we see that this correlation gets a little stronger. To take the value of 0.247. But as we can see by the star plot, it is clearly not the most relevant indicator as the others have higher values.



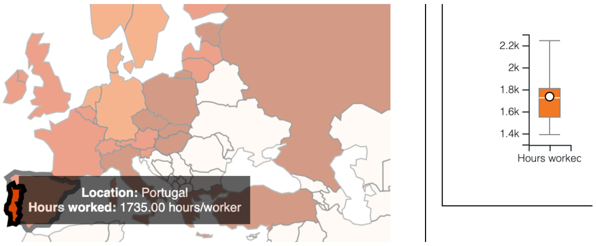
After answering to those questions, we pondered, **how is the average pay in Portugal compared to other countries?**

We zoom on the map and hover on Portugal. Then look at the Average Wage boxplot.



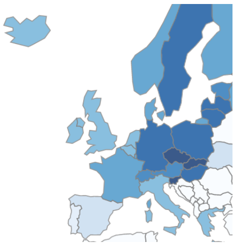
The data tells us that we are payed quite low on average.

But does that have to do with how much time we work in comparison?



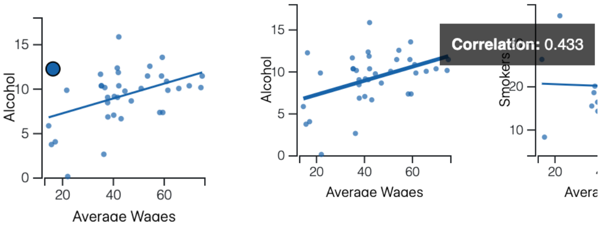
Well, apparently, not, we work more hours than the median.

And how are we in comparison to the other countries in the EU?



Not great. We are one of the lowest earning countries. (This is of course limited by the data available)

Is it because we are unhealthier than the average?



Well, we do tend to drink more than the other countries with the same earning. There seems to be a quite strong relationship between drinking more and earning more.

We also found some **interesting facts** such as:

The **suicide rate is highly correlated with the happiness**, and overall it has the most impact on the suicide rate. Meaning, countries with a low suicide rate have high levels of happiness.

There is a **strong correlation between education and alcohol consumption.**

There is a **strong correlation between the Social spending, GDP and the life expectancy at 65**. So, maybe it is important to have a higher social spending to help the elder live longer.

5. IMPLEMENTATION DETAILS

The links between the views are as follows: every visualization allows for tool tipping to show the relevant data in actual numbers. The map shows Location and selected variable, the boxplot shows statistical information on the distribution, the star plot shows the correlation value, the scatterplot shows the country and the value of both variables being analyzed, the correlation line shows the correlation value.

By hovering in countries in the map, all the datapoints in other idioms from that country are emphasized by means of appearing in the case of the box plot or by being highlighted in the case of the scatterplot. By hovering over a datapoint all the datapoints from that country and the country in the map are highlighted. By hovering on the correlation line, the corresponding dot on the star plot is highlighted and vice versa.

This section enumerates some challenges and explains how we overcame them.

There were some difficulties in terms of the work being divided amongst all the group members.

We also had some difficulties understanding the D3 version 5 documentation. The library in general has a steep learning curve as it was mentioned by the professors. But it gets easier as we dedicate a lot of time to studying it.

We implemented the correlation lines from scratch. It was not trivial.

6 CONCLUSION AND FUTURE WORK

With the development of this project, we conclude that it is a really interesting way to learn about making effective and interactive visualizations with real world data.

We developed a state-of-the-art prototype with real world data about health factors and health influencers. This prototype helps us visualize the data we gathered and processed. And it answers all the questions proposed, as well as some others we were not expecting.

The experience was quite painful at times, but ultimately it was rewarding to develop our very own visualization. The whole process of proposing a visualization, gathering the data, processing it, designing and implementing the visualization helped us learn a lot.

If we had to start over with the knowledge we have now, we probably choose a different theme with completely different types of data. Because that is how we grow and learn more. But assuming we chose this topic, we would do a scientific literature review on the topic. This would be helpful to base some of our initial decisions.

If we had one month and €3000 to enrich the project, we would invest that money in gathering higher quality data for the project. And spend time in allowing for the visualization of several health influencers in each visualization. It would also be interesting to get more health influencers as well as more health data. It would be interesting to have data about more countries. In terms of the visualizations themselves, we would like to develop a semantic zooming approach in order to encode more data in the map. We would also like to explore the approach of the usage of structural equations to understand the causation relationships between the variables. To showcase this, we would need to visualize several key indicators and explore the possibility of encoding stronger filtering mechanisms as we saw on the theoretical lessons. Another interesting thing would be to present the distribution of selected variables o the sliders, this was not possible due to the data configuration we set up focused on relationships, but maybe by filtering by countries instead of by time we could have a equally as interesting result.

7 REFERENCES

[1] Max Roser (2019) - "Life Expectancy". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/life-expectancy' [Online Resource]

[2] Esteban Ortiz-Ospina and Max Roser (2019) - "Global Health". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/health-meta' [Online Resource]

[3] Retrieved 12 16, 2019, from http://www.healthdata.org/results/data-visualizations

[4] Retrieved 12 16, 2019, from https://data.oecd.org/

[5] https://data.oecd.org/healthstat/life-expectancy-at-birth.htm#indicator-chart

[6] https://data.oecd.org/healthstat/life-expectancy-at-65.htm#indicator-chart

[7] https://data.oecd.org/healthstat/suicide-rates.htm#indicator-chart

[8] https://data.oecd.org/healthrisk/daily-smokers.htm#indicator-chart

[9] https://data.oecd.org/healthrisk/alcohol-consumption.htm#indicator-char

[10] https://data.oecd.org/healthrisk/overweight-or-obese-population.htm#indicator-chart

[11] https://data.oecd.org/healthstat/deaths-from-cancer.htm

[12] https://data.oecd.org/earnwage/average-wages.htm

[13] https://data.oecd.org/emp/employment-rate.htm

[14] https://data.oecd.org/emp/hours-worked.htm

[15] https://data.oecd.org/eduatt/adult-education-level.htm

[16] https://data.oecd.org/socialexp/social-spending.htm

[17] https://data.oecd.org/gdp/gross-domestic-product-gdp.htm

[18] https://stats.oecd.org/Index.aspx?DataSetCode=BLI

[19] https://data.oecd.org/air/air-pollution-exposure.htm#indicator-chart