

Visualizing Worldwide Contrasts – G03

Bernardo Quinteiro

Instituto Superior Técnico
Lisbon, Portugal

bernardo.quinteiro@tecnico.ulisboa.pt

Diogo Lopes

Instituto Superior Técnico
Lisbon, Portugal

diogo.andre.fulgencio.lopes@tecnico.ulisboa.pt

Gonçalo Mateus

Instituto Superior Técnico
Lisbon, Portugal

goncalo.filipe.mateus@tecnico.ulisboa.pt

ABSTRACT

In this paper, we describe the process of implementing an interface that allows the visualization and critical analysis of different statistics of countries worldwide. This will allow the user to come to their own conclusions on which country, continent or region is best or worse and how all of the statistics provided are intertwined with each other in a way that can help understand societal, historical and contextual disparities. This interface could be used for curiosity purposes, research purposes or even by users that intend to weigh their options on visiting or moving to a new country, for example, therefore, seeing this as a useful tool in making said decisions.

INTRODUCTION

It is widely understood that across the world, cultural and historical differences led to the shaping of extremely different civilizations, with unique values, structures and people. However, there is one thing that every single person on Earth has in common: their pursuit of happiness. Therefore, with this project, we aimed at understanding, firstly, how happiness differs from country to country, region to region and continent to continent, to understand where are the biggest discrepancies and where are the most uniform regions. In a secondary phase, we wanted to understand what caused these patterns, by analysing several indicators, such as GDP per capita, freedom or temperature. Finally, we also intended to analyse how correlated all of these statistics were, for example, if countries with high levels of corruption would normally not have a great freedom score, and so on.

Although several tools were available to visualize all of the data we intended to use, these tools were all directed to a single type of data. Alas, our project aimed at being unique for giving the user the opportunity of visualizing and comparing all of those datasets in one interface, to be able to develop a greater insight as to how the previously mentioned cultural and historical disparities mould these societies and how each model fares against others around.

Alas, we settled on 6 main questions that we would have to answer with this project:

- **Question 1:** “What makes a country happy?” - As aforementioned, we intended to perceive what are the characteristics of a country that are the most intertwined with its happiness levels, i.e., which statistics have high values where the happiness

values are also high and that have low values where the happiness is low.

- **Question 2:** “Is corruption directly associated with lack of safeness?” - This question aimed at understanding the correlation between a people’s perception of corruption and how safe it is to live in a country.
- **Question 3:** “Does being rich mean one is happier?” - With this question, we aimed at understanding whether or not indicators of wealth had a high correlation with a country’s happiness values.
- **Question 4:** “Is a country’s wealth a synonym of good healthcare?” - Similar to the last question, this question’s purpose was to visualize whether or not a country’s wealth indicators were associated with the quality of healthcare provided.
- **Question 5:** “How does my country fare against other countries?” - With this question, we intended to provide the user with a way of making an in-depth comparison of countries selected by them that would allow them to get to conclusions as to what disparities there are between them and why these disparities exist.
- **Question 6:** “Which are the best continents and the ones with the most disparities?” - Finally, we also aimed at understanding how these disparities functioned within and between continents, so that we could further understand and hypothesize the causes of said disparities or similarities.

REALTED WORK

While we were looking for information, we found an interactive map where we can visualize world happiness [5]. However, we couldn’t find any visualization where we could compare several factors in the same map making that one of the most important advantages of our visualization.

To design the charts from our project we looked up some examples present in the d3.js gallery [6] and decided which ones we should use to represent every aspect of our project.

Finally, to better understand how we could display more than two variables in the same graph, we learned how to work

with parallel charts [4,7] because we believe that those graphs are the best way to represent all the 7 variables that we wanted to display.

THE DATA

The data used in this project is composed by 7 datasets from the [World Population Review](#) [1] website, with these being a 2021 [Corruption Perception Index](#) that lists countries based upon their perceived levels of public sector corruption, a 2021 [Human Freedom Index](#) that encompasses both personal and economic freedom of countries around the world, a 2022 [GDP per capita](#) in order to better analyze the country's economic conditions, a 2021 [World Happiness Report](#) that lists countries based on their happiness, a 2020 [Healthcare Rank](#) that takes into consideration the various systems humans rely upon to help maintain their personal health, a 2022 [Global Peace Index](#) that ranks the safest and most peaceful countries in the world and a 2021 [Average Temperature](#) by world country. Apart from these datasets, we also used another one that [matched the countries to their respective continents](#).

We came across some issues while trying to find good datasets, since at first, we found many incomplete datasets or incompatible because many countries were missing from one to the other. Apart from this, we had to make sure the data made sense, since we stumbled upon some datasets which data did not seem very accurate. Afterwards, we found a website with all the information we needed and this way the chosen datasets would be more correlated, meaning this could help reduce the need for fixing missing values or incompatibility issues. Finally, since our first main goal was to define what attributes influence a country's happiness, we had to do some research in order to find the datasets that best suited our objective, help us take the best conclusions and answer at least the desired questions.

Merge Data

All the data treatment and dataset rearrangements were performed using the Python library Pandas.

Firstly, we merged the 8 previously mentioned datasets and right away stumbled upon some issues. There were some country names that varied between datasets, so this was our first fix in order to fully complete the merges. Afterwards, we noticed there were some missing values for the different attributes, which could be solved in other situations with a mean of the neighboring rows, but in this case, it's not applicable since these could happen to be not close countries and even countries that share borders can have completely different attributes, being an example of this South and North Korea. Therefore, in order to solve this issue, we did some research and estimated these missing values based on other datasets, where we found other values to be similar to ours, trying to be as accurate as possible.

Scalability Issues

We also had some scalability issues, since some datasets had different sizes, what increased the number of missing values that had to be estimated, countries that had to not be considered and values that could not be estimated. Therefore, after some more cleaning and rearrangements, we were able to obtain an alphabetically organized, almost complete dataset with 9 columns (countryName, corruption, freedom, gdppc, happiness, healthcareRank, safeness, averageTemperature, continent) and 185 rows, only missing some values that could not be estimated.

Derived Measures

At this moment, we were now able to proceed to define derived measures that would be useful for better analysis and conclusions. This way, we decided to do 20 derived measures, being these a world country rank of each attribute (used in the world rank table and Spearman correlation), a continent country rank of each attribute (used in the span chart to rank countries by continent and attribute) and the continent average of each attribute (also used in the span chart).

Post-Checkpoint II

Afterwards, upon beginning the implementation of the visualizations, we came across some errors in the dataset, for example missing ranks ("41st, 43rd", missing 42nd) and we were able to estimate some more values in order to have a more complete dataset. The main issue in these situations was that any change in the original dataset meant there was a need to recalculate the 20 derived measures in order for the data to be coherent, which was a bit frustrating and happened sometimes when we least expected.

VISUALIZATION

Overall Description

The final solution is composed of 5 visualizations that are linked to give the user an interactive experience, making analysis and conclusions easier. We tried to make the visualizations simple and appealing, so any user prompted with the interface would be able to interact without any prior knowledge and without the need to memorize how each idiom works.

The chosen visualizations provide knowledge of how some countries' attributes are correlated and how influential they can be, while also being able to compare up to 3 countries simultaneously or even compare these with continent status. This way, it's possible to answer a full range of questions, even further than the previously defined ones.

System Overview

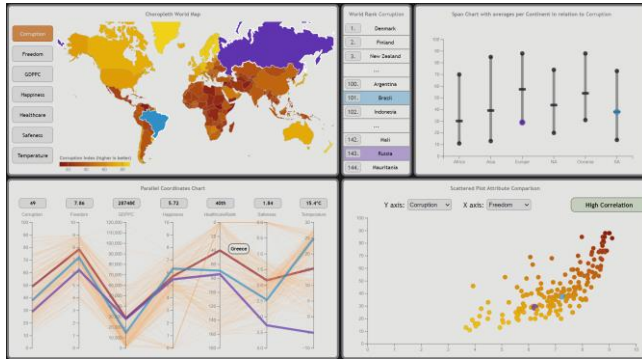


Figure 1. Interface Layout

The interface is composed of 5 linked visualizations, the top ones being a Choropleth world map with filters, a world rank table and a customized span chart, while at the bottom it's possible to find a parallel coordinates chart and a scattered dot chart. All the visualizations are linked, meaning some changes on one will affect all the remaining. We decided to use the bottom half for the previously defined charts, since these represent all the countries simultaneously, meaning a larger width would be better for visualization.

The main input across all the visualizations is the left and right click. This action lets the user select 2 countries, with the left-click country being highlighted in purple, while the right-click one in light blue across all the idioms. These colors were selected since they contrasted with the map color scale.

The filter present in the Choropleth World map is used to modify the color scale of the map itself, sort the world rank table and change the customized span chart y-axis.

Choropleth World Map

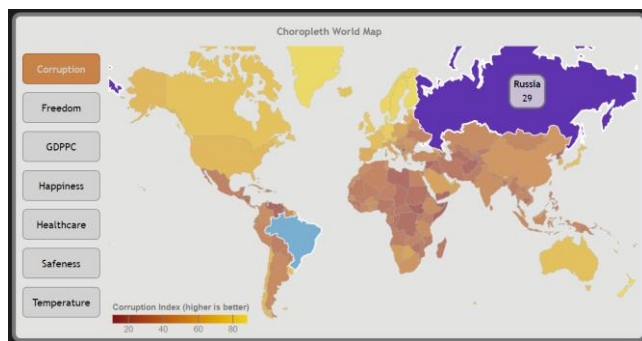


Figure 2. Choropleth World Map

We decided to use a Choropleth World map since it helps visualize information tied to geography, and also compares and analyzes data from across locations. This way, it's easier to visualize how the filtered attribute is spread across the map and if neighboring countries are similar. There is also a legend that helps better understand the idiom.

In order to interact with this visualization, the user can select one of the filters on the left to modify the color scale of the map and legend. Furthermore, by hovering over a country, its name and the value of the selected filter are shown on a tooltip. Apart from this, the user is able to choose two countries by left/right-clicking on them, which will be highlighted with the corresponding color, and modified on the further explained idioms. Finally, it's also possible to zoom and pan (with boundaries) to better see the countries or perform a more accurate selection.

World Rank Table

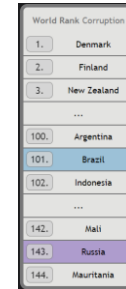


Figure 3. World Rank Table

This visualization is a world rank table based on the selected filter in the Choropleth world map. It sorts countries based on the world rank-derived measures for each attribute and always represents the top3 and the previous and following country of the currently selected ones. This helps the user know which are the best countries for a specific attribute, while also understanding where the selected countries rank and what are their closest matches.

The selected countries are highlighted by the corresponding color and other countries can be selected by left/right-clicking on their name on the rank table.

Customized Span Chart

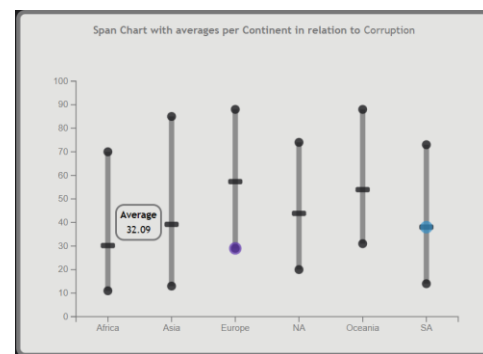


Figure 4. Customized Span Chart

We decided to use a span chart, since these are ideal to compare ranges between a minimum and maximum value, which, in this case, helps the user understand what's the country with the highest and lowest attribute value from each continent or even from all continents. This chart is customized due to the addition of the average of each continent for a better comparison and circles representing the

currently selected countries. This customization is further explained in the “Custom Visualization” section.

The x-axis contains each continent, with NA and SA, North America and South America, respectively. The y-axis values are based on the previously selected filter in the Choropleth map, having each attribute different scales (Freedom 0-10, Corruption 0-100, ...). For each continent, the lowest circle represents the country with the lowest/worst attribute value, while the highest circle represents the country with the highest/best attribute value. The rectangle represents the average value for each continent, which is very useful to better understand if a country is above/below average in relation to its continent and if there is a big discrepancy from a country to the average continent value, which happens when the average is further to one of the ends than to the other one. The countries selected in the Choropleth map are also shown in this plot by the corresponding selection colors. When hovering any of the circles or averages a tooltip will be displayed with the country name and attribute value or average value, respectively.

It is also possible to right/left-click the circles represented in this chart in order to modify the currently selected countries.

Parallel Coordinates Chart

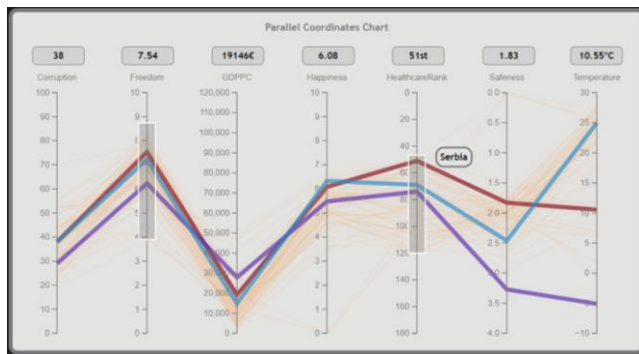


Figure 5. Parallel Coordinates Chart

We decided to use a Parallel Coordinates chart since these are used for plotting multivariate numerical data and are ideal for comparing many variables together and seeing the relationships between them.

In this chart, the currently selected countries are represented with the corresponding colors and when hovering the mouse over any other line it's possible to have a third comparison that is highlighted with the color red.

The country that is being hovered has its attributes displayed over each axis and its name on a tooltip. At first, we implemented axis reordering but we did not leave it in the final version, in order to simplify the representation of the values over each axis and since the benefits of this reordering are already fulfilled by the further explained scattered plot.

We also implemented brushing that can be performed in each axis in order to filter values, this way being able to perform more specific comparisons.

Furthermore, it's also possible to right/left-click on any of the represented lines in order to modify the currently selected countries.

Scattered Dot Chart

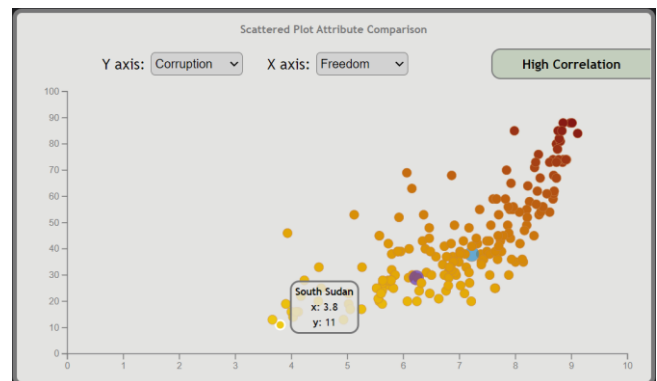


Figure 6. Scattered Dot Chart

We used a Scattered Plot since it is used to observe and show relationships between two numeric variables. This way, the user is not only able to check the x and y-axis attribute value but can also verify if there is any correlation between the two selected attributes.

The attributes to be compared can be selected from the selection boxes and then the scattered plot will be updated taking these into consideration. The color scheme used is the same as the Choropleth World map, this way keeping consistency all over the idioms.

By hovering over a “dot”, the country name and x and y-axis values will be displayed on a tooltip, this way the reader can also do a comparison of countries based on the selected attributes.

Furthermore, using the world ranking-derived measure, we also calculated the Spearman correlation, in order to understand if the two selected attributes had any correlation or not. Firstly, we were using the Pearson correlation, but afterwards, we changed since the Spearman correlation can also evaluate monotonic relationships. We defined that a low correlation value would be < 0.55 , a medium correlation value between 0.55 and 0.70 and a high correlation value above these previous numbers. The type of correlation is then displayed on the top right corner of the chart and highlighted with red, yellow or green depending on the achieved correlation. This way the user can take conclusions on how each attribute is related to the other and how these can influence a country.

Furthermore, the currently selected countries are highlighted by the corresponding colors and by right/left clicking any of the dots, the currently selected countries can be modified.

Rationale

We decided to use these visualization techniques with the previously mentioned characteristics, since they are simple

and any user prompted with the interface would be able to interact without any prior knowledge and without the need to memorize how each idiom works.

In relation to the visual encodings, the channels used for the Choropleth World map were a color scale, that modifies the appearance of the map depending on the selected filter, and the position in the map of the world countries. For the world rank table, the positions represent the rank of the top3 countries and the neighbors of the currently selected countries. Furthermore, in the span chart the y positions represent the value of the filtered attribute in the Choropleth map and the x positions represent the continent of each span. In regard to the Parallel coordinates chart, the y position represents the value of each attribute scale present in the x positions. Finally, in relation to the scattered plot, the y and x positions represent the chosen attributes from the selection boxes on top and the color scale is updated to represent worse/better countries from the current selection. For all the idioms, there is a color channel that highlights the currently selected countries by right/left click.

Taking in consideration these choices, we were able to come up with a design for our first sketch:

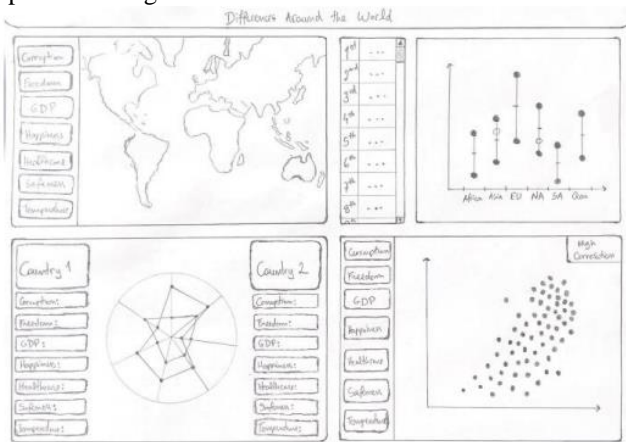


Figure 7. First Sketch

In the first sketch and first prototype (Figure 8 below) we had a Radar Chart instead of a Parallel Coordinates chart. This Radar Chart displayed all the attributes from the two currently selected countries, which made it easier to compare on what each country was better than the other.

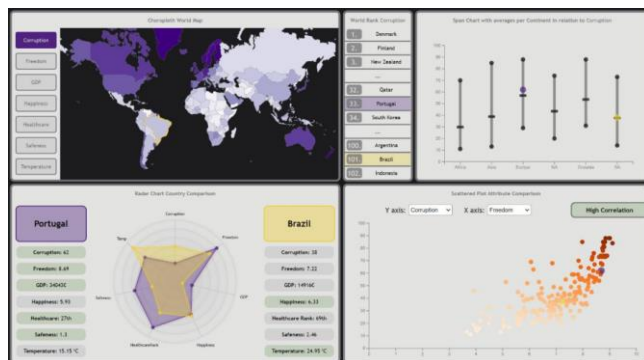


Figure 8. First Prototype

It's possible to see in Figure 8 that the Radar Chart was implemented with the colors corresponding to the highlighted countries. The values under each country name were also highlighted in green in case they represented a better value than the other country.

From the sketch to the first prototype, it's also possible to see that we chose to use selection boxes for the filters in the scattered plot instead of individual buttons for each filter. We decided to do this modification, since this method was already being used in the Choropleth map and could cause some confusion to if these filters had any correlation. Furthermore, by using the selection boxes instead, we were able to increase the width of the plot, which is better to represent large amounts of data.



Figure 9. Final Version

From the first prototype to the final version, we decided to use a Parallel Chart instead of the Radar Chart, since it is better for comparing many variables together and seeing the relationships between them. This way, instead of only being able to compare the two currently selected countries, we are now able to simultaneously compare all the countries in the dataset, or more specifically, the currently selected countries and a third one by hovering on one of the lines of the chart. The brushing functionality is also very useful to filter the data shown in this chart, giving this way a new tool for the user to achieve further insights.

Another change from the first prototype to the final version is the color scheme. The purple colors were not ideal to visualize data in the Choropleth map and even worse having a left click selection be purple (the same color as the map). This way, we opted for a yellow-red color scheme, which made comparisons in the Choropleth map easier and we chose blue and purple for the country selection since these colors were able to contrast with the map colors. The country selection on the first prototype was only modifying the color of the country's borders, whether on the final version the whole country changes color which makes it easier to visualize specifically in smaller countries. A legend to the Choropleth map was also added and the borders of all containers across the interface was reduced, which gave it a cleaner look. The axis colors and font size and colors were also verified in order to guarantee it was consistent across the interface.

Custom Visualization

As mentioned previously, our custom visualization is a span chart, which are ideal to compare ranges between a minimum and maximum value, but are not perfect. A problem with span charts is that they focus the reader on only the extreme values and give no information on the data points between the minimum and maximum values, so we decided to customize the span chart by also displaying the average values of each continent calculated by the derived measures mentioned previously and also display the currently selected countries on the respective position.

The attributes on the x axis are nominal positions that represent each continent, being NA and SA, North America and South America respectively. The attributes on the y axis are ratio positions (including the derived measured averages), except “Healthcare” that is a rank.

The maximum and minimum value for the filtered attribute is a country represented by a circle in the extreme positions of the span, while the rectangle in the “middle” of the span represents the average value for said attribute for each continent.

These modifications made it possible to take more insights from this type of chart that are very useful to be able to answer some of the previously mentioned questions, such as “How does my country fare against other countries?” and “Which are the best continents and the ones with the most disparities?”. This is possible, since now the user can compare the average values of each continent, verify if there is a big discrepancy within the continent (if the average is closer to one end than the other), check if the selected countries attributes are above/below average in some other continent or even if they are higher/under the extreme values of each continent.

Other customization that was thought was to illustrate the Spearman correlation line on the scattered plot but this wouldn’t add any value to the information, since we already had a container characterizing the correlation on high/medium/low.

This way, we think the customization on the span chart was the one that provided a biggest change on the number of insights that could be taken from an idiom, even further than the needs for the defined questions.

Demonstrate the Potential

With the final version of the interface implemented, we are now able to answer the previously defined questions:

Question 1: “What makes a country happy?”

This is the most generic question, since it does not have a “correct” answer but instead is the result of the combination of multiple factors. In the case of this interface, there are 7 factors that can contribute for a country happiness, these being the corruption index, freedom, gdppc, healthcare, safeness and temperature.

In order to answer this question, the user can select the Happiness filter in the y axis selection box of the scattered plot and go through all the other possible attribute selections for the x axis.

Each different combination will modify the scattered plot and display how these two selected attributes are related according to the Spearman correlation.

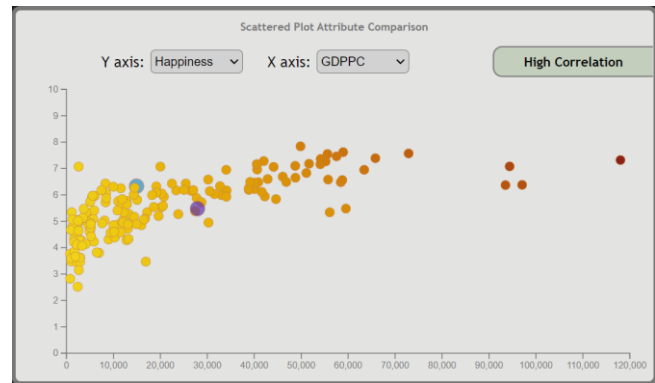


Figure 10. Scattered Plot to answer Question 1 and 3

From the steps mentioned before, we chose “Happiness” as the y axis and in order to simultaneously answer Question 3 (“Does being rich mean one is happier?”) we chose “GDPPC” as the x axis. We can verify in the top right corner these attributes have a “High Correlation”, meaning a higher gdppc is one of the factors that can influence the happiness of a country.

Another way to answer Question 1 could be by selecting the “Happiness” filter in the Choropleth map section and then right/left clicking the country with the highest value from the world rank table or from the customized span chart. This way, this country is added to the currently selected countries and in the Parallel Coordinates chart it’s possible to verify which attributes are also high, meaning they are characteristics of a happy country.

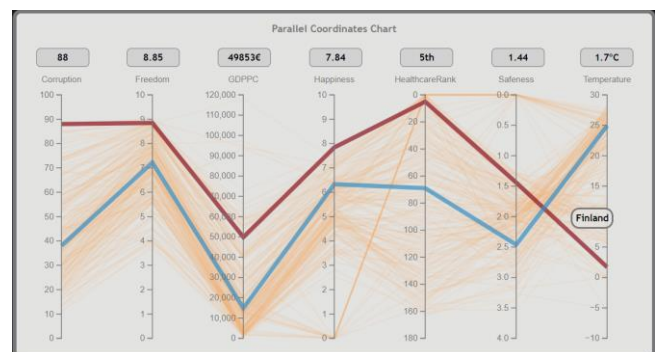


Figure 11. Parallel Coordinates chart to answer Question 1

Following the steps previously mentioned, we selected Finland since it’s the country with the highest happiness value and then in the Parallel Coordinates chart we hovered

its line in order to show all the country's attributes (the line turned red and a tooltip with the country name is displayed).

By analyzing this plot, it's possible to see Finland is also one of the least corrupt (high corruption index value), one of the freest and one of the best ranked countries in healthcare, from what is possible to conclude that these attributes can also contribute to a country's happiness.

Question 5: "How does my country fare against other countries?"

This question can be answered in different ways, but at first, the user shall select "their" country from the Choropleth world map, either with the right or left click. Now it's possible to compare the selected country with any other country on all the remaining idioms.



Figure 12. World Rank Corruption to answer Question 5

In the world rank table, it's possible to see how the selected country faces against other in a ranking system of the currently selected filter ("Corruption" in the example from Figure 12). In case the selected country was Brazil, it's possible to see it's a country considered corrupt, since ranks 101st in the world rank table, being more corrupt than Argentina but less than Indonesia.

This question can also be answered in any of the other idioms:

- In the customized span chart, the selected country appears highlighted with the corresponding color and it's possible to perform a comparison with the extreme countries of each continent and average values, for the selected filter;
- In the parallel coordinates chart, it's possible to compare the multiple attributes with any other country from the dataset;
- In the scattered plot, the user can compare 2 attributes from the selected country with any other country from the dataset, making it possible to verify how "their" country freedom and happiness face against other countries.

Question 6: "Which are the best continents and the ones with the most disparities?"

In order to answer this question, the user can select a filter on the Choropleth World Map (corruption, freedom, gdppc,

happiness, healthcareRank, safeness or temperature) and then by looking at the customized span chart it's possible to take some insights.

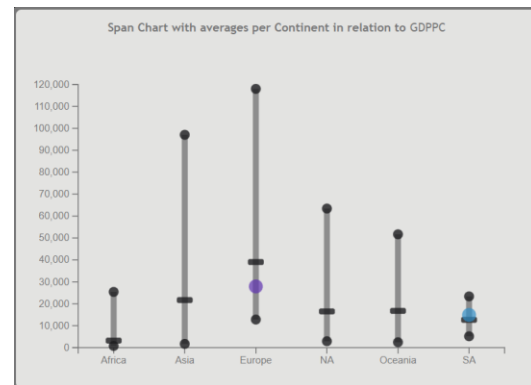


Figure 13. Customized Span Chart to answer Question 6

From Figure 13, it's possible to see the chosen filter from the Choropleth World Map was GDPPC (title of the chart). By analyzing this chart, it can be concluded that the highest gdppc country is from Europe, while the lowest is from Africa (these countries can be known by hovering the extreme circles on each span, where a tooltip will be displayed). Furthermore, the lowest gdppc country in Europe is above the average value in Africa and the highest in this continent is under the Europe's average. Apart from this, it's also possible to conclude about disparities, since these are represented when the average value is not centered in the span. For example, the average value in Africa is very close to the minimum value, meaning the country with the highest value in this continent is a disparity and does not represent a reality across the rest of the continent. In Asia, Europe, North America and Oceania it's also possible to verify this discrepancy from the average value to the highest one.

Unexpected Insights

When developing the visualizations, we came across some golden unexpected insights, being one of them the discrepancy that exists within continents specially when using the gdppc filter, meaning very rich countries are surrounded by poor ones. For example, Singapore has a gdppc of 97057€ when the average of the Asian continent is almost 24000€.

Furthermore, when analyzing the correlations on the scattered plot, we came across interesting insights, for example a high correlation when comparing gdppc with happiness and healthcare and that least corrupt countries tend to be safer and freer. Temperature is the attribute that did not have any correlation with the others, meaning people do not need a paradisiac climate or snowy environment to be happy. Apart from this, on average Europe is the least corrupt, freer, richest, happiest, better healthcare and safest continent.

IMPLEMENTATION DETAILS

The first main challenge was to complete the dataset and rearrange it since there was missing data and some country names were not the same from one dataset to the other.

Afterwards, the topojson file also had different country names, so we had to fix this to stay consistent and it had some countries that are part of other, for example countries like French Guyana were associated with France and Greenland to Denmark.

A small problem that still persists in the final version is that the dataset has more countries than the file used by topojson, meaning some countries will not “react” when clicked on the idioms since these are not present in both datasets.

The color scale was not easy to decide since we always needed two colors to be highlighted, so in the final prototype we decided to use a yellow-red color scale instead of the purplish one on the first prototype.

The parallel chart was a last-minute modification as well, since we thought it would represent in a better way multivariate data than the previously implemented Radar chart. This chart has a brush in order to filter data on the corresponding axis and it works by just displaying the lines between an intersection of selected thresholds.

The views are in separated files but are connected through an update function. In the case of the Choropleth map, when another filter is selected, the update function is triggered and the map color-scale is changed, such as the order of the world rank table and the y axis of the customized span chart. The same way works with the currently selected countries that are a shared array between all views, and, when modified, all idioms are updated with smooth transitions, this way displaying the correct currently selected countries with their respective highlighted color.

The visual attributes are defined in a CSS file named styles.css in the ./css folder, the charts JavaScript files are inside the ./charts directory (also containing the previously implemented radar chart file), the topojson and d3 modules are in the ./d3modules folder, the datasets csv files can be found in the ./data directory and in the ./cleaningDataPython are some of the scripts used for cleaning and rearranging the datasets. A JavaScript file to create the Choropleth map legend [2] and a Spearman correlation algorithm [3] were also used.

We took inspiration from online examples for the idioms, except for the customized span chart which was an interesting idea for the data in question. All the idioms had to be customized and modified for our desired implementation, applied transitions and linking between the views.

CONCLUSION AND FUTURE WORK

In conclusion, we believe that the final solution produced completes every task we set out to complete, by answering every question posed and doing so in an intuitive and interactive way. Therefore, we can also conclude that this visualization would suit the interests of the possible users we defined. If we were to start over or have more time to continue the development of this visualization, we would've, firstly, liked to develop some visualization that would've accomplished what we hoped the radar chart would do, but for multiple selected countries, rather than just two. Finally, we would have also wanted to further incorporate more datasets that would allow for even more thorough analysis and have intra-region comparative measures, similar to what we did with continents (for example, comparing only Mediterranean countries).

REFERENCES

1. 2022 World Population by Country. (n.d.). <https://worldpopulationreview.com/>
2. Bostock, M. (2021, November 25). Color Legend. Observable. <https://observablehq.com/@d3/color-legend>
3. GeeksforGeeks. (2022, August 3). Program for Spearman's Rank Correlation. <https://www.geeksforgeeks.org/program-spearman-rank-correlation/>
4. Gemignani, Z. (2021, September 15). Better Know a Visualization: Understanding Parallel Coordinates Charts — Juice Analytics. Data Analytics and Visualization Made Easy - Juice Analytics. <https://www.juiceanalytics.com/writing/writing/parallel-coordinates>
5. Hamer, H. (2020, May 12). Visualizing World Happiness With An Interactive Map. Towards Data Science. <https://towardsdatascience.com/visualizing-world-happiness-with-an-interactive-map-e33ebbaa3936>
6. Holtz, Y. (n.d.). The D3 Graph Gallery – Simple charts made in d3.js. <https://d3-graph-gallery.com/index.html>
7. Parallel Coordinates Plot - Learn about this chart and tools. (n.d.). https://datavizcatalogue.com/methods/parallel_coordinates.html