EPFL

COM480 : Data Visualization Project

- The Big Five Personality Test -

Process Book

Tadaviz *group :*
Robin Szymczak
Kenyu Kobayashi
Guilhem Sicard

May 29, 2020

Contents

1 Introduction

In the context of our Data Visualization project, our goal was to make accessible visualizations of a dataset that would give **meaningful insights** to the target audience.

For this purpose, we chose the dataset of the *Big Five Personality Test*, which was taken from *Kaggle* (`https://www.kaggle.com/tunguz/big-five-personality-test#codebook.txt`). This dataset contains over a **million** of answers to a personality test, from people all around the globe.

For each respondent, there are 5 essential variables that are being measured:

1. **EXT** - *Extroversion*
2. **AGR** - *Agreeableness*
3. **CSN** - *Conscientiousness*
4. **EST** - *Neuroticism*
5. **OPN** - *Openness to Experience*

We also have access to other information such as their localisation, their time spent to answer each question.

With our visualization, we wanted to show to what extent cultures can **shape** people's personality. Moreover, we were interested in finding if, in average, people from a given country take more time to answer to questions related to a specific personality trait. The objective was to answer these questions through an intuitive and dynamic **map**.

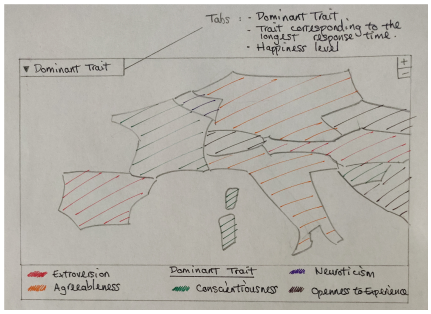
Also, we thought that it would be interesting to visualize the distributions of the answers through a **bar plot**. The purpose of this visualization was to see whether there are some common behaviors between respondents, such as some questions for which the **majority** of them answered positively or negatively.

Our motivation was to answer these questions with simple yet subtle visualization techniques, hoping that we could show interesting facts to the target audience, which generally speaking can be anyone.

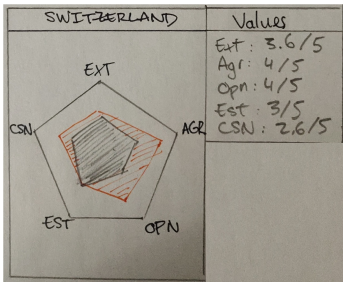
2 Expectations

Our project's principal goal was to show to what extent cultures can **shape** our personality.
For this task, we considered that the **dynamic map** would be best suited. It is a tool that is so well known that people will immediately understand what to do. By choosing some classic visualisation techniques we hope to be more efficient. The data is not complex by what it holds but by the number of possible variations.

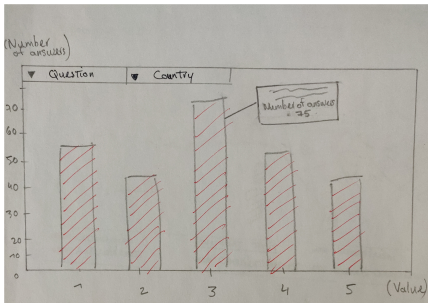
We planned to implement it using *Leaftlet.js*. The map would not only show the dominant personality trait, but also other information such as the happiness level, or the personality trait corresponding to the longest response time.



(a) Sketch of the dynamic map



(b) Sketch of the radar graph



(c) Sketch of the bar plot

Figure 1: Sketches of the visualizations

We planned on giving the users the freedom to zoom in and out as they wish, and to get to choose in the top-left tab between 3 criteria for visualization:

1. The **dominant trait** : By selecting this criteria, we wanted each country to be displayed in the color associated to its dominant personality trait. Moreover, when users hovers over a country, a **radar graph** implemented with *D3.js*, would be displayed to show the distributions of the scores associated to each personality trait:

2. The **trait corresponding to the longest response time** : Each country would be displayed in the color associated to the personality trait for which the questions had the longest response time in average. We hoped that this would answer the question "what are people most preoccupied about?"

3. The **happiness level**: A heatmap of the happiness level across the world from an extra data-set.

Even if maps are great tools to have an overview of the different global score, it is not the best one to explore the answers to each question. For this purpose, we made another goal to create a different aspect of the website to explore the data in a question-oriented fashion.

Again, the complexity of the data isn't due to the data format itself since it is simply represents some answers ranging from 1 to 5. The complexity arises from the 50 different questions and the even greater number of countries. In total, there are 50 thousand different possible graphs. It is therefore very important to keep a very clear graph so that the results of a question for a given country is immediately understood. A lot of effort had to be put in the search tool for the countries as well as the interface to choose the question, to not overload the user with all the different possibilities.

The more we were diving into the project, the more we wanted to do personality test ourselves. We thought that the same desire could emerge from anyone on the website. For this reason, we decided to create our own *custom questionnaire*. The idea falls slightly outside the scope of the course, but we couldn't resist the temptation to inform each user of the country where people are the most similar to him personality-wise, according to our big-five data-set.

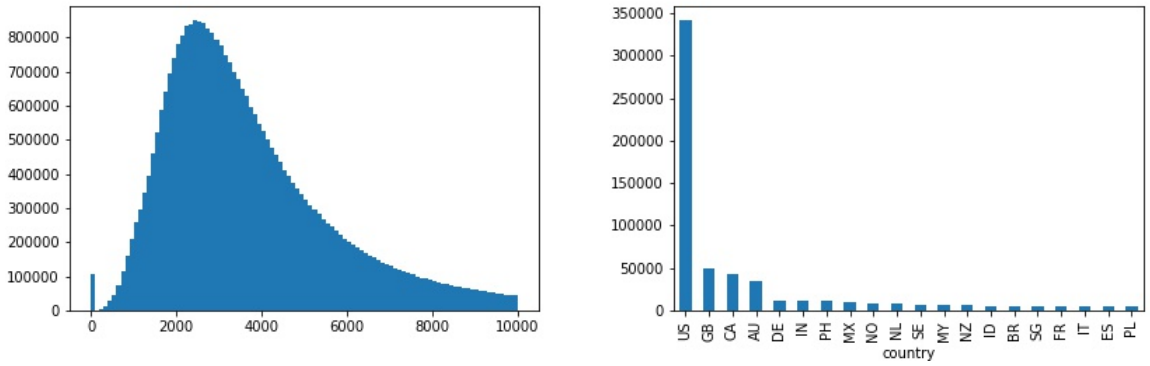
3 Implementation

3.1 Data pre-processing

Before implementing any visualizations, we had, of course, to **pre-process** our dataset.

At the beginning, we had a dataset containing the result of 1 015 341 big five personality tests, conducted from march 2016 to november 2018. We cleaned the dataset by removing **NaNs**, participants when they had **multiple IP adresse**s, **no location**, when they did not answer to more than **90 percent** of the test or when they were **outliers** because of the time they took to answer. Finally, we also discarded the answers from countries which didn't have more than **100 respondents** to the test, since we thought that it wouldn't be representative enough of the country.

We can see on the graph below the distribution of time taken to answer. There were lots of 0. Some of them were due to participant skipping questions, while for others, there was simply no explanation. For this reasons, these were discarded.



(a) Distribution of the variable of the time taken to answer

(b) Number of respondents of the top 20 countries

Figure 2: Plots for pre-processing

3.2 Data wrangling

After the step of pre-processing, we still had to do some data wrangling in order to be able to **use** the data properly. For this purpose, we did the following:

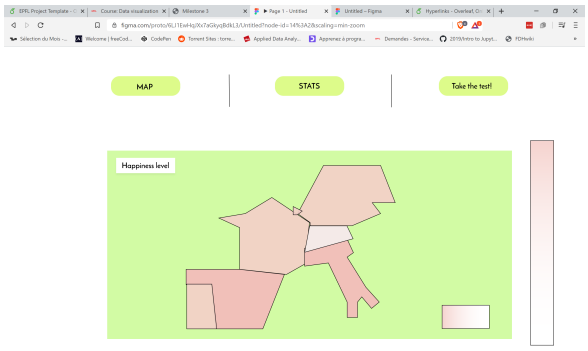
1. Prepare data for our questionary:
   - For each country, compute the mean score for each character trait.
2. Prepare data for our radar graph:
   - Personality traits
     (a) Compute the mean score of each personality trait for each country
     (b) Compute the mean score of each character trait globally
   - Response time
     (a) Compute the mean response time for each character trait for each country
     (b) Compute the mean response time for each character trait globally
3. Prepare data for our core visualization
   - Compute the dominant character traits for each country
   - Compute the character trait associated to the longest response time for each country
   - Export the happiness level of each country

4. Prepare data for our bar plot:
    - For each question, compute the distribution of answers (number people who answered 1, number people who answered 2...) globally, and per country.

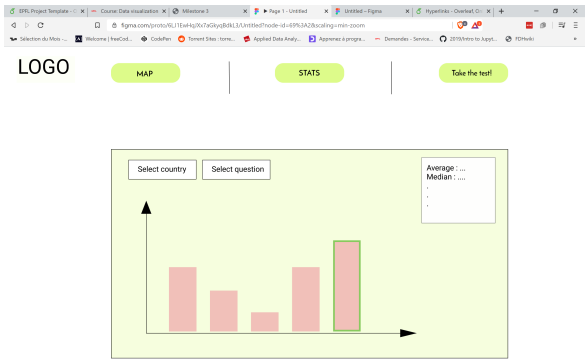
More details regarding the different steps and processes can be found in the *Data wrangling* notebook in the Git repository.

3.3 Prototyping and design choices

In order to have a clear idea of the features we decided to use the prototyping tool *figma*. We sketched the architecture of the website and the target visualisation. The prototype can be found at this link : `https://www.figma.com/proto/6LJ1EwHqJXx7aGkyqBdkL3/Untitled?node-id=52%3A38&scaling=min-zoom`. Some pictures from it will be used as a reference for comparison of how much our design has evolved. It shows a direct insight of our vision in the middle of the ideation process.

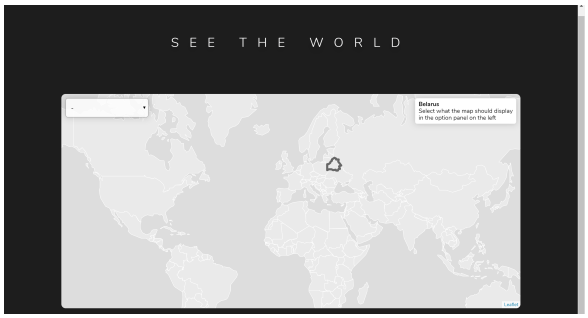


(a) Prototype of the dynamic map

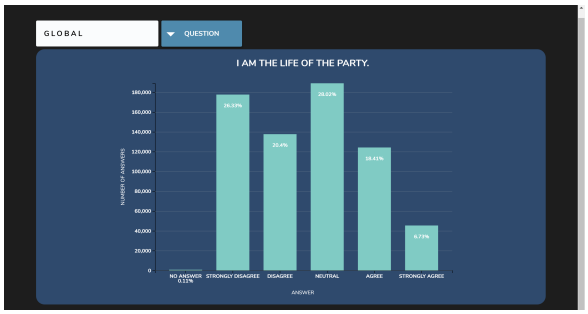


(b) Prototype of the bar plot

Figure 3: Prototypes of the website



(a) Final result of the dynamic map



(b) Final result of the bar plot

Figure 4: Final results

Let us dive into the design choices of developers. First of all, the colors. As an initial idea we chose a watermelon strip color palette with shades of green and a light pink. The color of the water melon was not very soothing. As our data-set is immense, we need a simple yet calming design, made as an invitation to stay. The night-colors shown below seemed perfect. Our website is very exploration focused, and we characterized the mouse as a light-torch in the dark to illuminate all the items that will reflect blue and gold colors.

(a) Prototype colors

(b) Final colors

Figure 5: Color palette of the website

3.4 Dynamic map

3.4.1 Design

We kept a very straightforward color for the map. Our plan was to make it behave exactly like a normal map, and therefore we made it look like a normal map. For this reason, a simple gray and white color is displayed in the idle state to make it appear very familiar. The button on the top left display a "-" on arrival. It directly attracts people's attention to change it, and as they do, the whole map lights up and so will they.

We avoided using the classic green and red palette for the heat-map of the happiness levels as it is hard for colorblind people to decipher. We instead used shades of blue and yellow. We also avoided using green for the principal traits even though clicking on the country would inform the user about which trait we are referring to.

We kept the radar graph visualisation of the different trait as it is a dense representation of a few independent variables. If we had put the answers on a bar instead, it certainly could have led to some confusion. This is because bar plots organize in a meaningful way the values on the x axis, and that in this case, the traits do not share any relation with each others. For all these reasons, the radar graph was a better choice.

3.4.2 Implementation

When we started using the map we had to find some ways to limit the zoom out of the map. Otherwise, it was replicating the world multiple times inconveniently. It was settled by automatically coming back to a specific scroll position when the user would go out of the map. In order to keep the interactivity of the map, we had to, sadly, turn off the interactivity of the radar graph, which originally, showed the values of each score when hovered over. Still, the result is, indeed, compliant to our sketch presented in **Section 2 - Expectations**. Here is the final product of our core visualization:

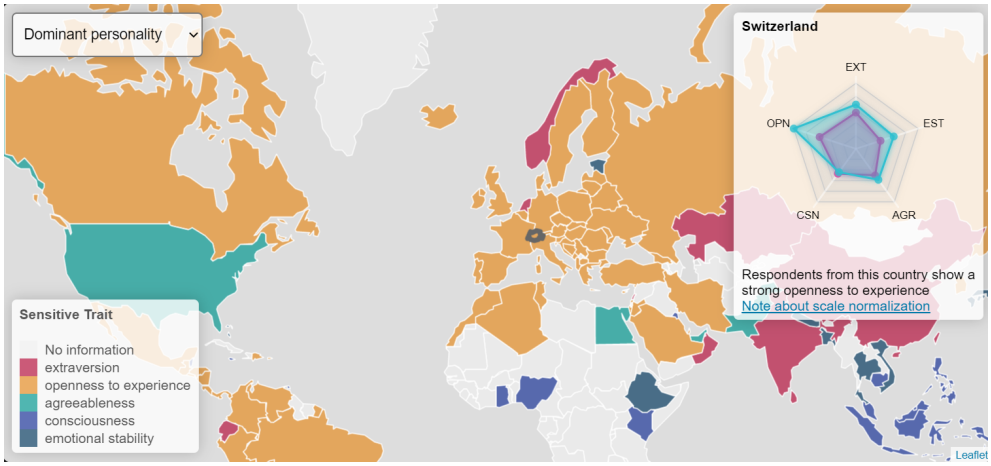


Figure 6: Final Dynamic Map

3.5 Bar plot

3.5.1 Design

As explained in **Section 2 - Expectations**, we wanted a simple and straightforward design for the bar plot. We chose to have 6 entries with explicit names [strongly disagree, disagree, [...], strongly agree] instead of the labels used in the data (1,2,3,4,5).Combined with the full name of the question as the title of the bar plot, it makes the graph **completely independent** so that anyone, regardless of an introduction, can understand the graph in the blink of an eye. The absolute number of respondent is displayed on the left axis and the relative one on the bars. This way, the viewer is not required to do any math and he is invited to quickly navigate through the menu.

For the interface to select the country, the design had to be non-overwhelming as having a list of 100 countries could easily be. To avoid this pitfall, we added a search engine to quickly find a nation of interest and the list is still present for the wanderers. The list of question is made as a tree to avoid the same pitfall. The first level proposes to choose the category of the personality trait, thus dividing the number of questions by five. Once the category of question selected, the second level displays ten boxes. To avoid the sudden appearance of a lot of text, we hid the question name that appear and stayed on and after hovering. The buttons feel very natural and are made user-friendly.

3.5.2 Implementation

We chose to implement our bar plots using *D3.js*, since it provides easy animations and transitions. We successfully implemented the selection of the country and question to visualize the appropriate distribution. It wasn't easy to implement the search bar for countries, and especially, to implement the update of the bar in an animated way when users change the question or the country. To do this, we chose to remove most of the elements present in the SVG on each update, and re-create them using animations. We also delayed some animations to occur after others, in order for everything to be *smooth*.

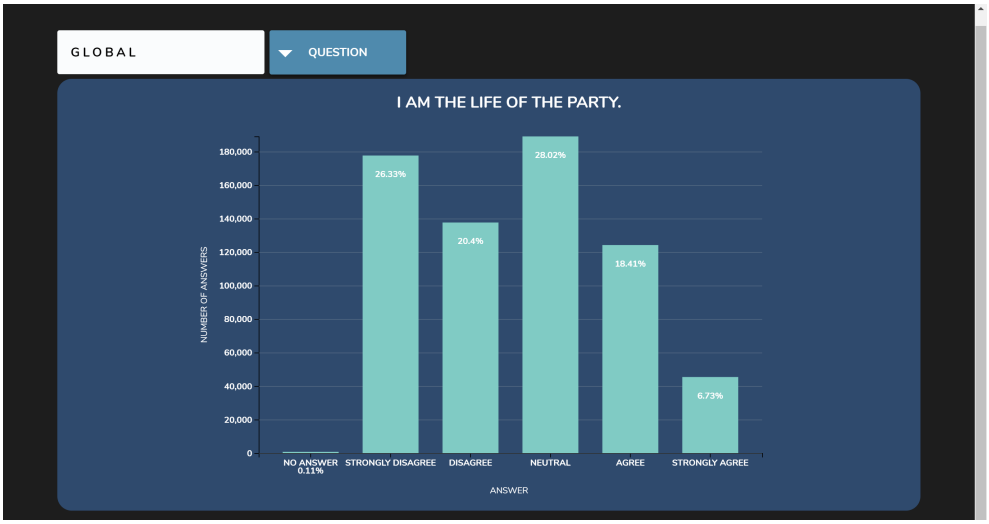


Figure 7: Final Bar plot

3.6 Questionary

Last but not least, we are really happy to have implemented the *Big Five Personality Test* in our website. Once implemented, we realized how much of a user-friendly and fun dimension this functionality had added to our website.

3.6.1 Design

The questionary contains a total of 50 questions. We had to avoid to put them all in one page. Conveniently, these are split into 5 categories. Thus, we divided the questionary into 5 parts. We wanted to avoid the dullness of the usual internet questionary and make it more engaging. This is where the dark colors are interesting, since they give a sens of mystery to the experience that could not have been possible with the white theme.

Moreover, results are displayed in a more childish manner with lots of animation made on purpose to add to the excitement of discovering where the algorithm will place the user in the world. On the last page a "Learn more" button appears underneath the displayed country. This latter redirects to a page with some statistics on a white background. It was very important for us that it did not match the overall look of the website, as this is what you would expect with such type of button, to have gone "too far". On this page you see the detailed results of your test and some Wikipedia explanation of the different traits. The selection of trait is made very efficient with as little animation as possible. Since we are displaying personal information, the page doesn't look too professional to avoid any distrust from the users.

3.6.2 Implementation

We implemented an algorithm that computes the scores of each personality trait, and returns the country which has the smallest L2-distance with respect to each of these scores. These different information are displayed in an animated modal that pops out. We are really happy with this final functionality added to our website, and are glad that we implemented it. The following Figure is a portion of the final product of our questionary:

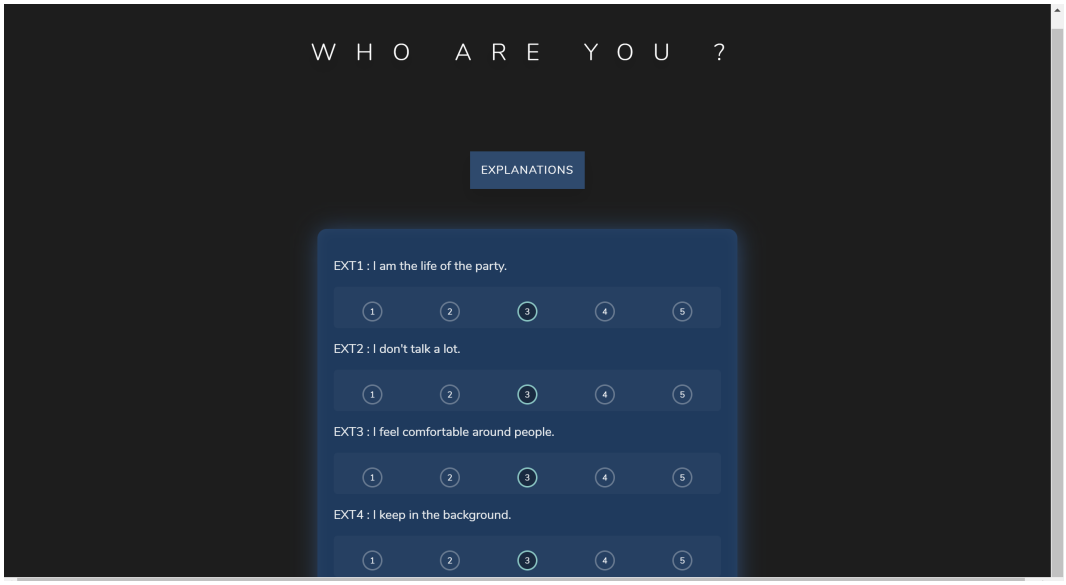


Figure 8: Questionary

4 Peer assessment

To realize this project, since there were **3 main functionalities** to implement, each one of us focused on one of them:

- Guilhem focused on the **dynamic map**:
  - Pre-processing the data,
  - Implementing the map and its interactivity using *Leaflet.js*,
  - Managing the overlays of the map,
  - Implementing the modal displaying the scores for the questionary,
  - Implementing the index page.
- Robin focused on the **questionary** :
  - Wrangling the data in order for it to be used in the questionary,
  - Implementing all the interface regarding the radio buttons,
  - Implementing the storing of answers,
  - Implementing the computation of the scores,
  - Implementing the computation of the closest country,
  - Designing the website, homogenizing the colours and fonts.
- Kenyu focused on the **bar plot** and the **radar graph**:
  - Wrangling the data in order for it to be used in the bar plot the dynamic map,
  - Designing the website, the welcome page, the transitions between pages,
  - Implementing the bar plot and its interactivity,

  - Implementing the search bar and buttons to filter by country and question,
  - Implementing the radar graph used in the dynamic map visualization.

Moreover, throughout the project, each one of us helped the others regularly, regarding the implementation and use of the different data-visualisation libraries. Moreover, we all worked on the design of the website, and are very proud of our work and teammates.

5 Conclusion

Overall, we are very happy and proud with the final version of our website. We managed to implement everything we had planned, with some extra functionalities. However, we still think that more visualisations and animations can be added to our website, to make it even *more enjoyable*. To finish off, here are some prospects of improvement:

- Displaying the average response time on the barplot in addition to the distribution of answers
- Applying a clustering algorithm to the dataset and make some visualizations out of it, to see whether we could find some patterns regarding the personality of people
- Adding smooth transitions between the different visualization pages
- Refine the localisation of the answers to go through the personality difference within a country.