EPFL

COM480 : Data Visualization Project

- The Big Five Personality Test -

Process Book

Tadaviz *group :*
Robin Szymczak
Kenyu Kobayashi
Guilhem Sicard

May 28, 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*

For each respondent, we also have access to other information such as their localisation, and their time spent to answer each of the questions.

With our visualization, we wanted to show to what extent cultures can **shape** our personality. If people's personality are shaped by the culture of the country there are living at, then that would also mean that each country has a **dominant** personality trait. Moreover, we wanted to see if, in average, people from a given country take more time to answer to questions from a given category of personality trait. We wanted to answer these questions by making visualizations through a dynamic **map**.

Also, we wanted to visualize the distribution of the answers (from 1 to 5) for each question, to see whether there are some questions for which the **majority** of people answered positively or negatively. In order to do this, we thought that a **barplot** would be the best match.

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. Thus, we wanted to study, for each country, which personality trait is dominant, and make an **interactive** visualization which could convey this information the most easily.

For this task, we considered that the dynamic map would be the best suited. We wanted this latter to be our **core visualization**, and to be interactive. We hoped to implement it using *Leaftlet.js*. We wanted our map not only to show the dominant personality trait, but also other information such as the happiness level of countries or their trait corresponding to the longest response time.

Here is a sketch of the visualization we hoped to achieve:

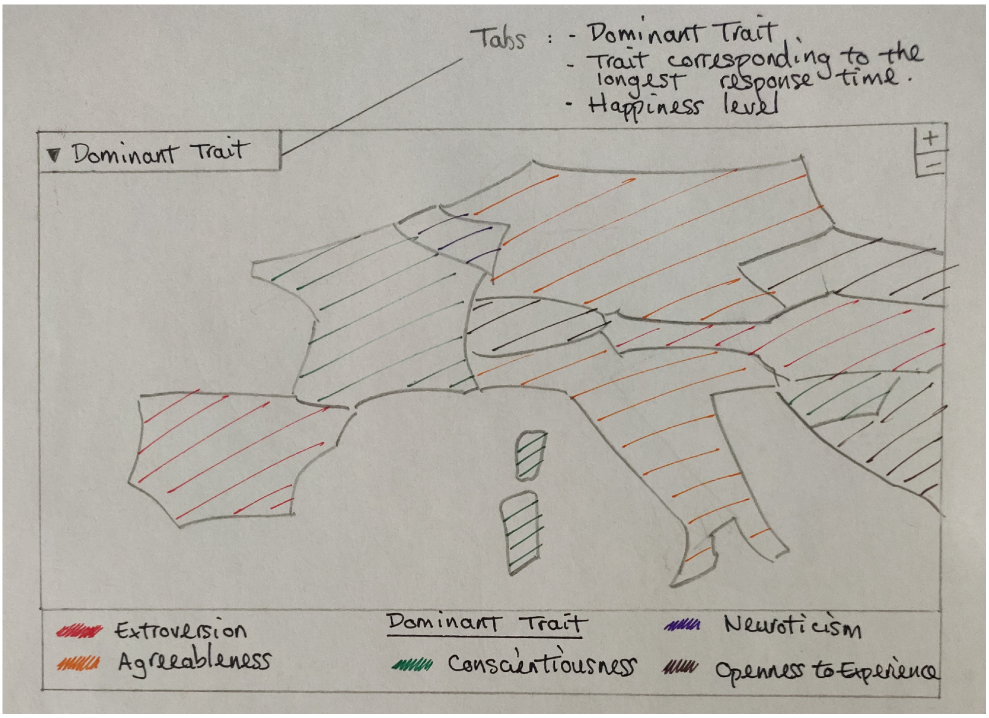


Figure 1: Sketch of the dynamic map

We wanted its users to be able 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, we wanted a **radar graph**, to be displayed to show the distributions of the scores associated to each personality trait:

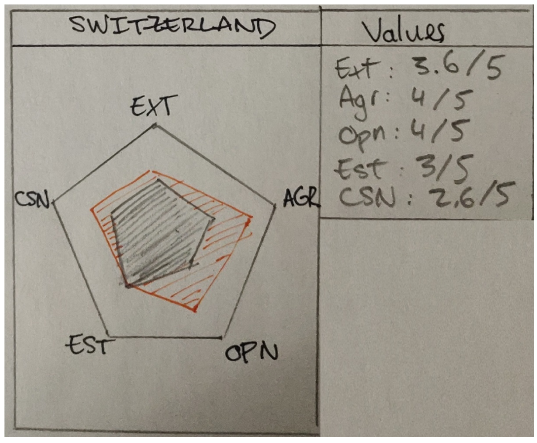


Figure 2: Sketch of the radar graph

We wanted to implement this latter using *D3.js*. **TODO : explain?**

2. The **trait corresponding to the longest response time** : By selecting this criteria, we wanted each country to be displayed in the color associated to the personality trait for which the questions had the longest response time in average. We thought that this criterion would be interesting to visualize as it could have been interesting to see if people from a given country are in average more "preoccupied" by questions for a given personality trait.

3. The **happiness level**: We planed on using an extra dataset, which could give us the happiness level of each country. We thought of visualizing this latter using a heat map, which could indicate whether a happiness level of a given country is high or not. For this visualization, we wanted the labels to indicate the happiness level for given colors.

Additionally to this visualization, we also wanted to make an interactive bar plot to visualize the **distribution of the answers** to a given question, for a given country. We wanted the users to have the freedom to choose these two parameters, and visualize the associated distribution. We thought that visualization could provide some interesting insights about whether there are some big disparities in answers for some questions. We wanted to implement this using *D3.js*, or libraries providing inbuilt interactive bar plots.

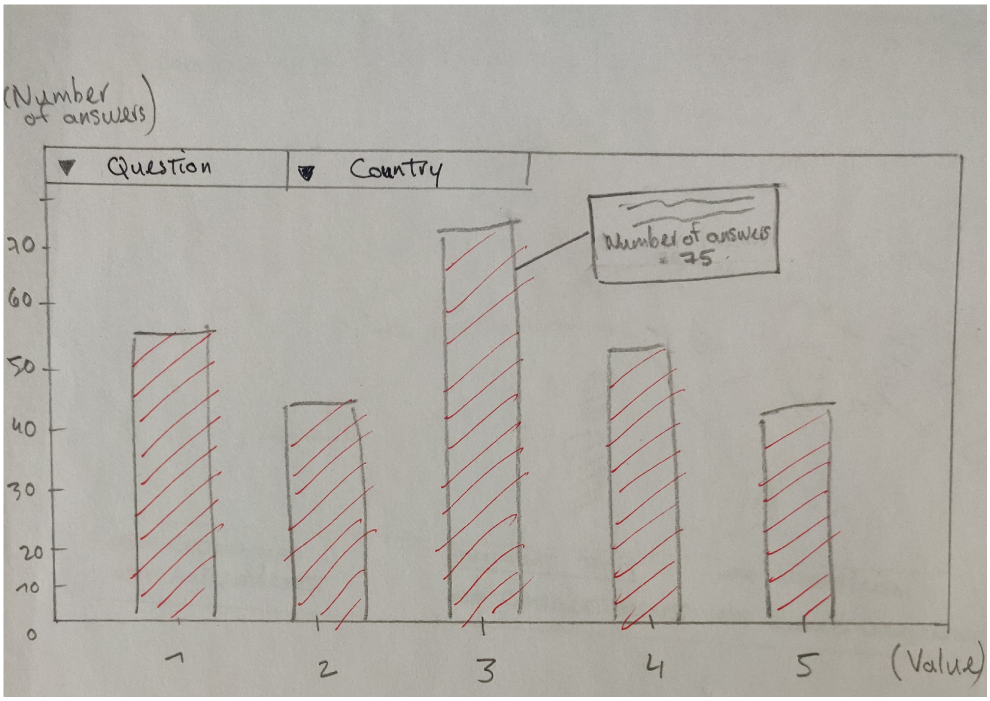


Figure 3: Sketch of the bar plot

In the context of this project, all these visualizations were required to be hosted in a website, to be implemented as well. For this reason, later on, we had an idea of implementing an extra functionality allowing users to **do the test** themselves, on our website. By taking into account their answers and making the necessary computations, we wanted this functionality to **reveal** which country best suits them according to the dataset.

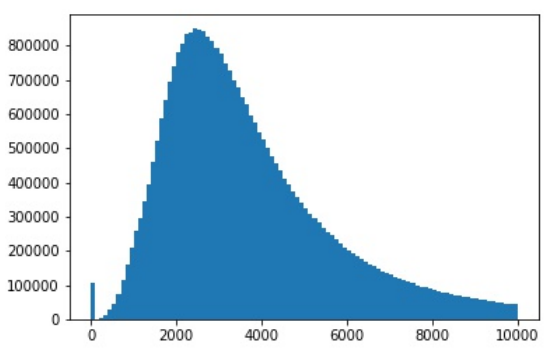
3 Implementation

3.1 Data pre-processing

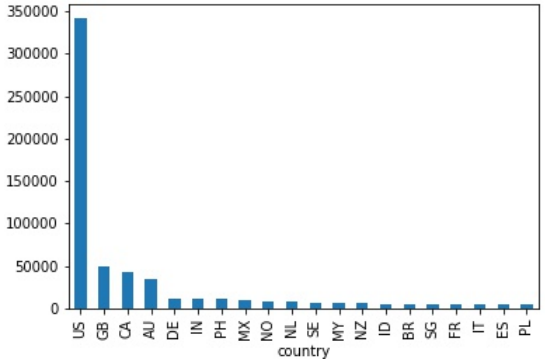
Before implementing any visualizations, we had, of course, to **pre-process** our dataset.

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 with respect to countries (Top 20 countries)

Figure 4: 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 Dynamic map implementation

As planned, our dynamic map was implemented using *Leaflet.js*. Moreover, additionally to implementing visualizations for each of the criterions that we had planned, we also implemented visualizations regarding the number of respondents with respect to each country. We also implemented the radar graph as we had planned, using *D3.js*. In order to keep the interactivity of the map, we had, sadly, to 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 - Expectations**. Here is the final product of our core visualization:

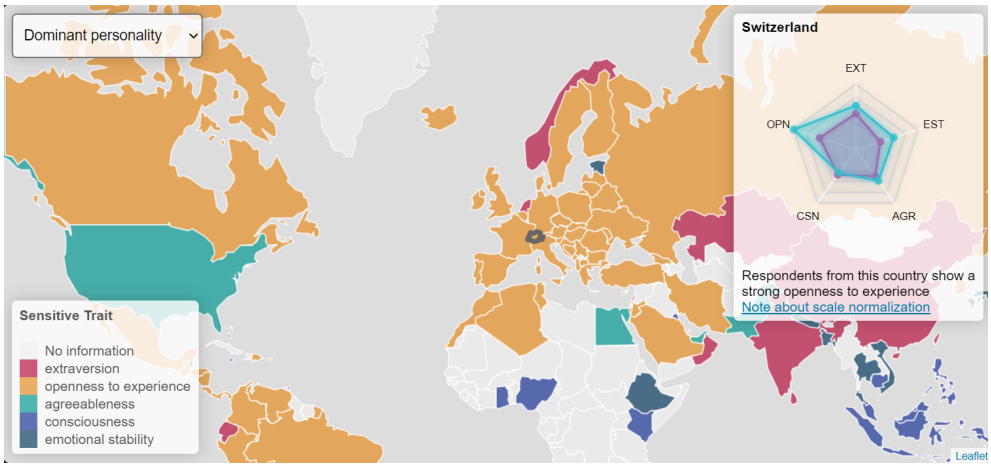


Figure 5: Final Dynamic Map

3.4 Bar plot implementation

We chose to implement our bar plots using *D3.js*, since it provides easy animations and transitions for the interactiveness. 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. We are very happy with our final product of this visualization, as even if it was challenging, we managed to produce everything we had planned in the beginning. Here is the final product of the bar plot visualization:

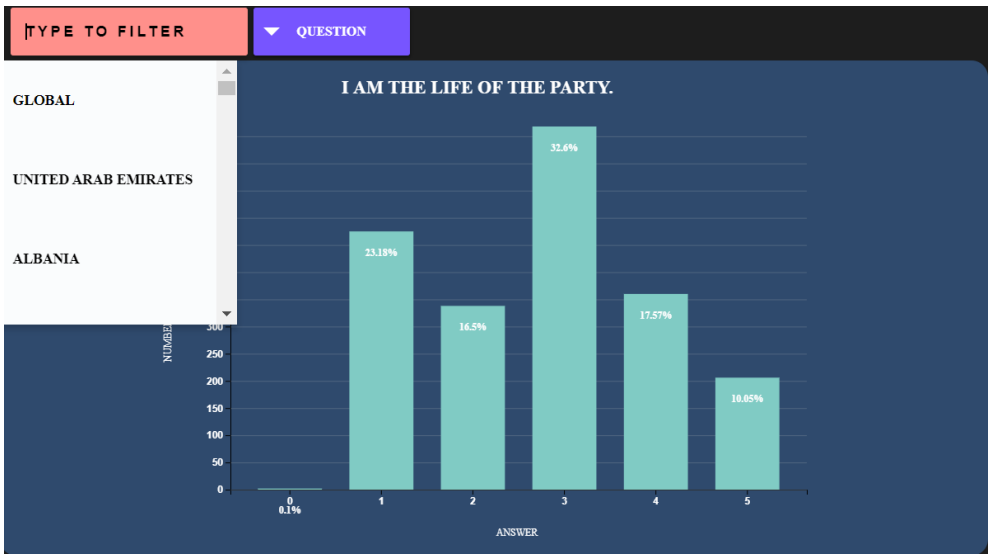


Figure 6: Final Bar plot

3.5 Questionary implementation

Finally, 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 dimension this functionality had added to our website. It was quite hard to design it in a nice-looking way, but still, we managed to take some inspirations from some other questionaries online. As planned, our questionary allows the user to take the test. Moreover, when the user submits his answers, we implemented an algorithm to calculate the scores of each personality traits, and compute the country which has the smallest L2-distance with respect to each of these scores. These different informations are displayed in an animated modal that pops out in the screen of the user. We are really happy with this final functionality added to our website, and are glad that we implemented it. Here is a portion of the final product of our questionary :

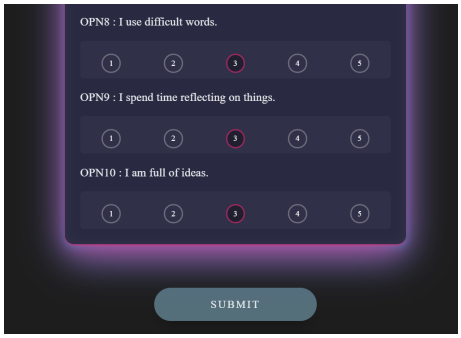


Figure 7: Questionary

4 Peer assessment

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

- Guilhem focused on the Dynamic map:
  - Pre-processing the data,
  - Implementating the map using *Leaflet.js*,
  - Implementating the modal displaying the scores for the questionary,
  - Implementating the index page.
- Robin focused on the questionary :
  - Wrangling the data in order for it to be used in the questionary,
  - Implementating all the interface regarding the radio buttons,
  - Implementating the storing of answers,
  - Implementating the computation of the scores,
  - Implementating 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,
  - Implementating the bar plot and its interactiveness,
  - Implementing the search bar and buttons to filter by country and question,
  - Implementating the radar graph used in the dynamic map visualization.

Moreover, throughout the project, each one of us helped the others regularly, regarding the implementation of some animations or interfaces such as buttons. Moreover, we all worked on the design of this website, that we all find magnificient.

5 Conclusion

Overall, we are very happy 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 interactiveness and animations can be added to our website, to make it more user friendly. 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 more animations to the questionary page
- Adding smooth transitions between the different visualization pages