Never Alone
Speed dating dataset

GROUP Q : Christopher BELINGUIER - Manon CHARVET - Yrieix LEPRINCE - Rémi SANCHEZ - Thai-An THAN TRONG

Table des matières

1. Overall presentation

The topic of our project is to program a tool that helps us exploring a dataset more easily. In particular, we are focusing on what can be learned about the differences between women and men concerning speed dating habits through the study of the Speed Dating dataset.

This dataset is composed of 8,000 observations from speed dating events. It has both match and non-match cases, with answers to several survey questions about how people rate themselves and how they rate others on several dimensions. It is a large and rich dataset with lots of features (almost 200).

The data are saved in the csv format and come from the following link: http://flowingdata.com/2008/02/06/speed-dating-data-attractiveness-sincerity-intelligence-hobbies/

2. Description of the problem

The data visualization tool will be addressed to many different users : from speed dating organizers, who would like to have a review on their events, to dating sites, marriage agencies. All of them look for information on speed dating efficiency, and how relationships work between people. Consequently, the tool should be easily usable, without requiring any complex install. Users would only need to have some basic understanding of web pages, and be able to read a graph, besides the necessity to be already familiar with speed dating concept.

It will be an exploratory analysis tool for this dataset that aims to find insights we could not have detected without any data representation. Users will be able to find answers for their questions and to find new potential problem statements. For instance, they could look for elements that could improve the client's experience, differences between genders in the dating strategy …

3. Description of the data

The dataset represents 8378 observations and 195 attributes, so the dataset size is (8378, 195). Half of the observations are related to female and the other half to male. The dataset contains the observations of 551 candidates (so we can have several observations for one person). The overall percentage of match is about 16%.

The columns are for example: gender, the perception of a candidate on themself and on other candidates, income. Our data are longitudinal so we can track the evolution expectation and impression of users over time.

4.   Brainstorming

The purpose of the second milestone was to brainstorm potential designs. By the end of the brainstorming, we came up with three different design ideas, that we organized into dashboards. Each design represents a point of view from which we intend to analyze the data : from a wave point of view, to a more personal one and a global overview of all success.

5.   Final sketch

For our final design we wanted to have a tool that would allow us to explore the dataset from a number of perspectives. Since some of our visualization ideas were more suited for certain tasks, we thought about splitting our visualisation into different windows.

One option could have been to create a tag for each visualization allowing the user to navigate between different graphs and choose the one with which he intends to explore the data.

We rather chose to create 3 principal tabs, each displaying a dashboard centered around a scale at which to explore the data. The dashboard is composed of different visualizations that make sense together and help the user find insight of the dataset.

The trade-offs of this approach are that :
-   Even if the tabs and the type of analysis that they allow are quite independent, the user might need to navigate between the views using the tabs if they're interested in different analysis. One problem is that they can't see the graph of the first tab next to the graph of the second. To counter this we thought about the ability for the user to navigate between tabs by not just only using a menu but also by interacting with the visualization. For example : when using the visualization for the analysis of a wave, by double clicking on a person on the parallel coordinates we could be redirected to the tab for person analysis for the user considered.
-   It's possible that the user is only interested in one graph and presenting to them a complete dashboard with other graphs that they don't need might disrupt their analysis. That's why we decided to only keep from our 3 initial design only the interesting graphs and only use graphs that allows for different exploration.

For this we needed to select only the interesting part of the 3 initial design and dispose them into different views and remove redundant graphs.

Homepage:
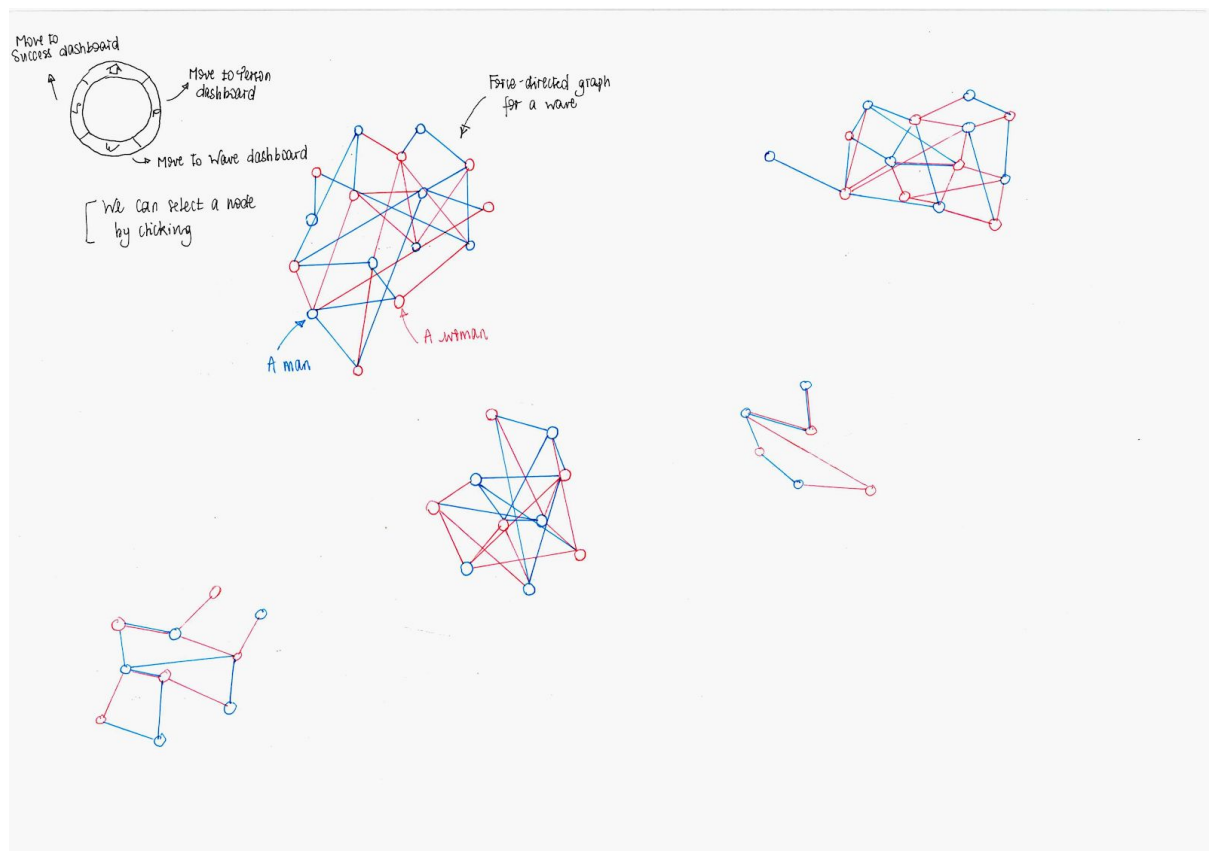
We also decided to add a home page to our tool with a repulsion network graph showing all the participant of all the wave and a menu. You can use the menu to navigate to the three main windows for visualizing data from different perspectives. If you select a node on the graph and chose to navigate to the person or the wave tab it will be filled with the information of the person you selected or with the wave in which she participated. The user will also be able to see some basic information on the graph which could facilitate her future exploration like :
- the size of the wave, represented by the number of nodes in the connected component,
- the gender of a person by the color of the node,
- the decision of the two people by the links between them and their color,
- how much they liked each other by the size and thickness of the link.

We're now going to introduce our 3 different windows for the visualization :

Wave centered tab :

Wave 1: October 16th 2002
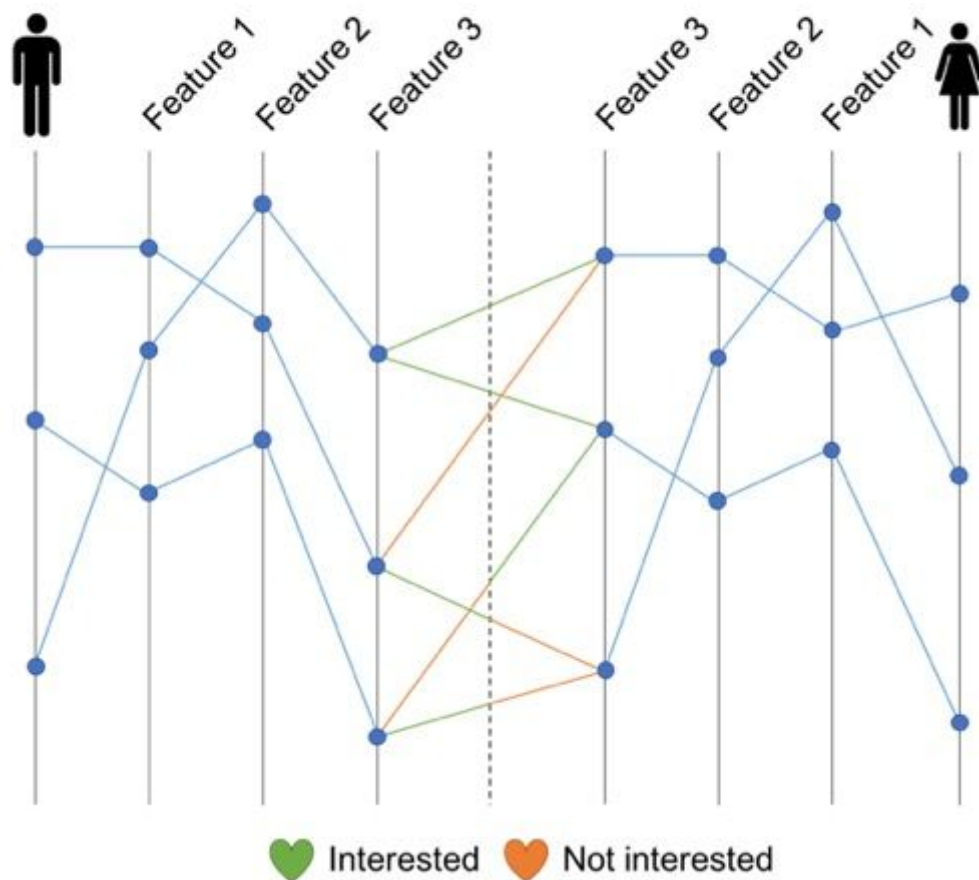
What are we trying to understand from the data ?
The goal of this visualization is to help coach of seduction or speed dating organisator to improve the percentage of match in each wave, by detecting the features that improve the number of match.

How can users interact with the visualization ?
The user can choose the wave, the number of features and the features to plot on the parallel coordinates. Moreover, this visualization is interactive, so we can swap the columns and select a subset of feature.

What are the compromises that we made compared to our initial design ?
We have deleted the hierarchical edge bundling and the personal information from the initial sketch. In fact, the hierarchical edge bundling which allows to check connections between categorical features of two people who have match, is more

focus on success than on wave. Furthermore, the personal information which allows to get more information about a specific person, when we click on it, were replaced by a link to the person centered tab.

Person centered tab :
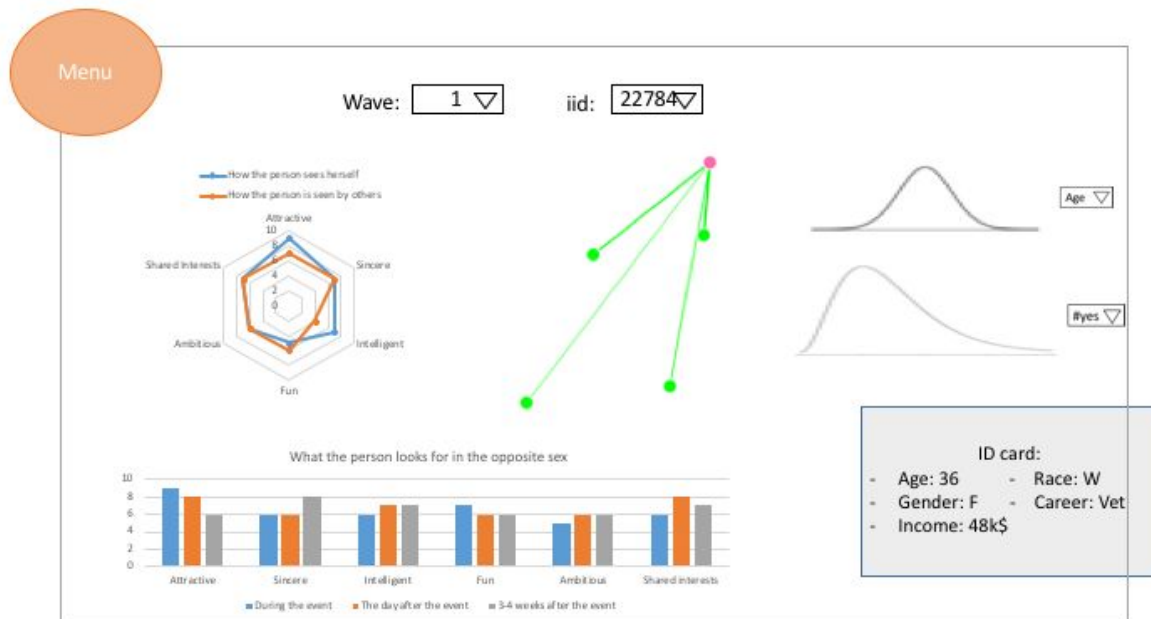
What are we trying to understand from the data ? We look for information regarding the performances of a selected person during his dates as well as we look for aspects that were positive and well perceived by other people or other aspects to improve.

Who is the main type of user ? The main user could be a person that would like to have a review of his dates, or a coach of seduction that present the results to him.

How can he/she interact with the visualization ? For each graph, there are different ways to interact with. For instance we can select the variables to display on histograms, radar chart, by adding a selection.

What are the compromises that we made compared to our initial design ? Instead of adding the force-directed graph that represents the overall wave at the center of the dashboard, we entirely focus on the person's data. The graph is moved to the home page and will still be connected to the Person centered tab : if we select a point on the home page, we can select the person centered tab that will display the data for the selected person.

Success centered tab :
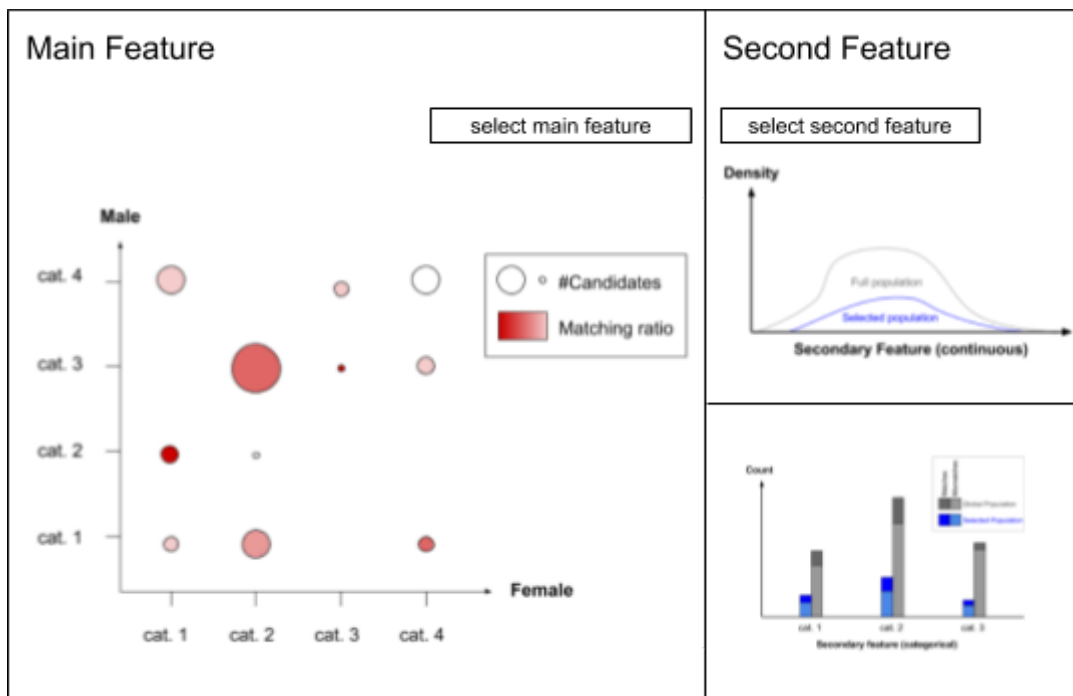
What are we trying to understand from the data ?
This tab aims at exploring the dataset populations by features. It lets the user to choose the features to analyse and display the relationships between males and females. Thanks to its secondary feature analysis it allows to understand more deeply the links between candidates.

Who is the main type of user ?
This tab allows to explore and understand the relationships lying between candidates. So it is a useful tool for people who organizes speed dating sessions, or for seduction coaches.

How can he interact with the visualization ?
On all the visualizations the user can access to the exact values by hovering it.
On the scatter plot, the user can brush data to analyze it more deeply with a second feature.

What are the compromises that we made compared to our initial design ?
The edge bundling graph won't be used in our final dashboard as it can only can draw links if there is a full match (in our single color proposition).
The navigation menu will be available on the top left corner.

6. Final design

As mentioned before, the project is divided in three tabs: Success, Wave and Person. We propose here to show an overview of what is possible to explore thanks to our tool. Let's consider we want to focus our analyses on a specific career men, let's say … teachers!

6.1. Success centred tab

First, we can use the success analysis tab to compare features relationships between females and males. The following 2D scatter plot displays the distribution of dates depending on men and women careers:
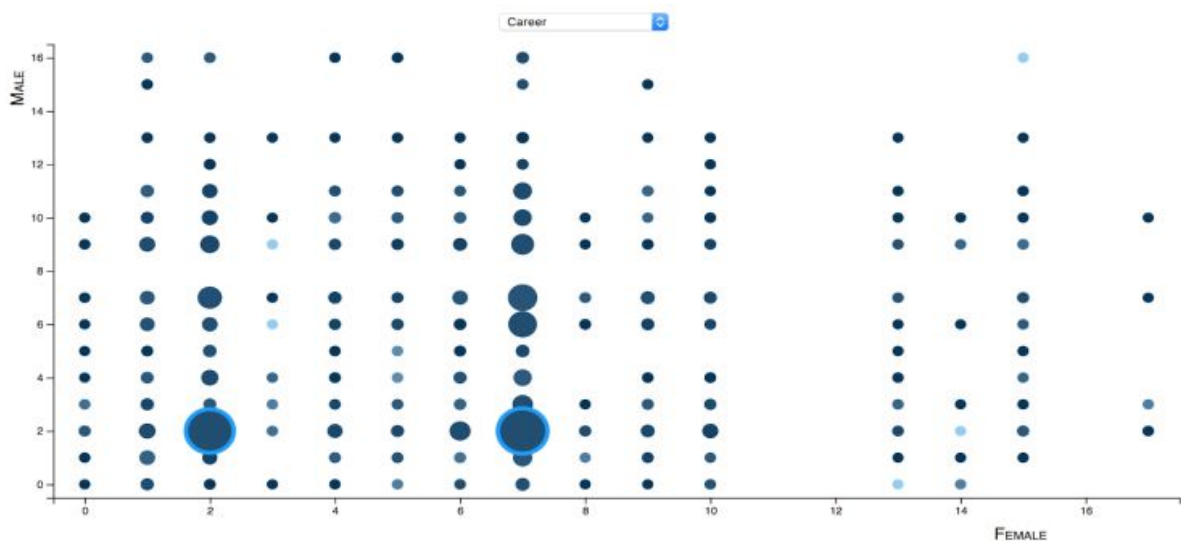
x-axis represents women careers and y-axis represents men axis. Markers color codes for the match/mismatch ratio and the size codes for the number of candidates represented by the marker.

As it can be interesting to explore more deeply the scatter plot selected points, we offer the possibility to analyze the selected candidates on a second feature. We can then focus on teacher men (encoded by 2 in the y-axis) who match with teacher and businesswoman (encoded by 2 and 7 in the x-axis). By selecting bubbles, we can then get additional information thanks to line charts and histograms:
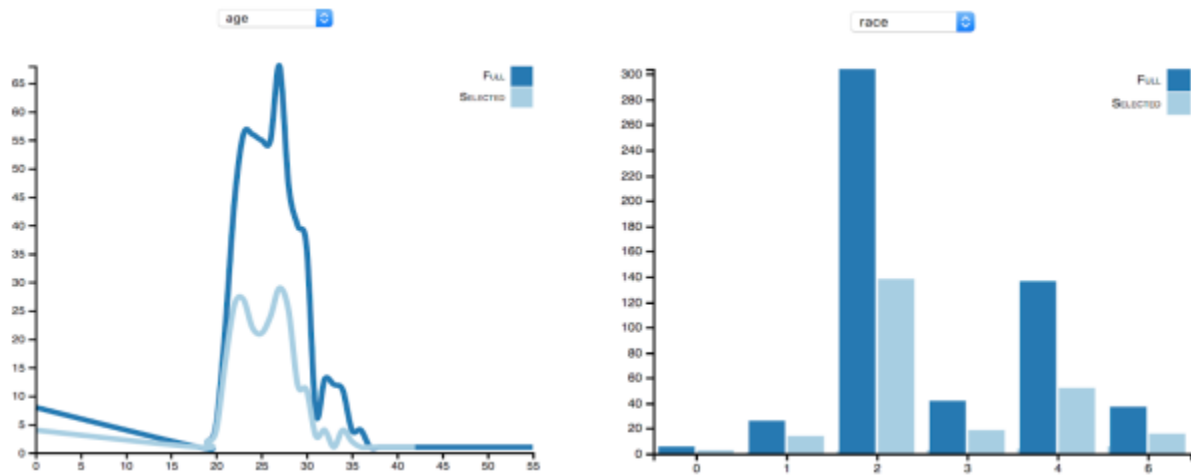
Here, we chose 'age' and 'race' as variables of interests. In our example, people selected in the scatter plot are mostly Europeean/Caucasian-American (encoded by 2 in the histogram) or Asian (encoded by 4 in the histogram) between 20 and 35 years old.

6.2. Wave centred tab

Then, we can then switch on another tab in order to have a global overview of a wave. The goal of this visualization is to help coach of seduction or speed dating organisator to improve the percentage of match in each wave, by detecting the features that improve the number of match.
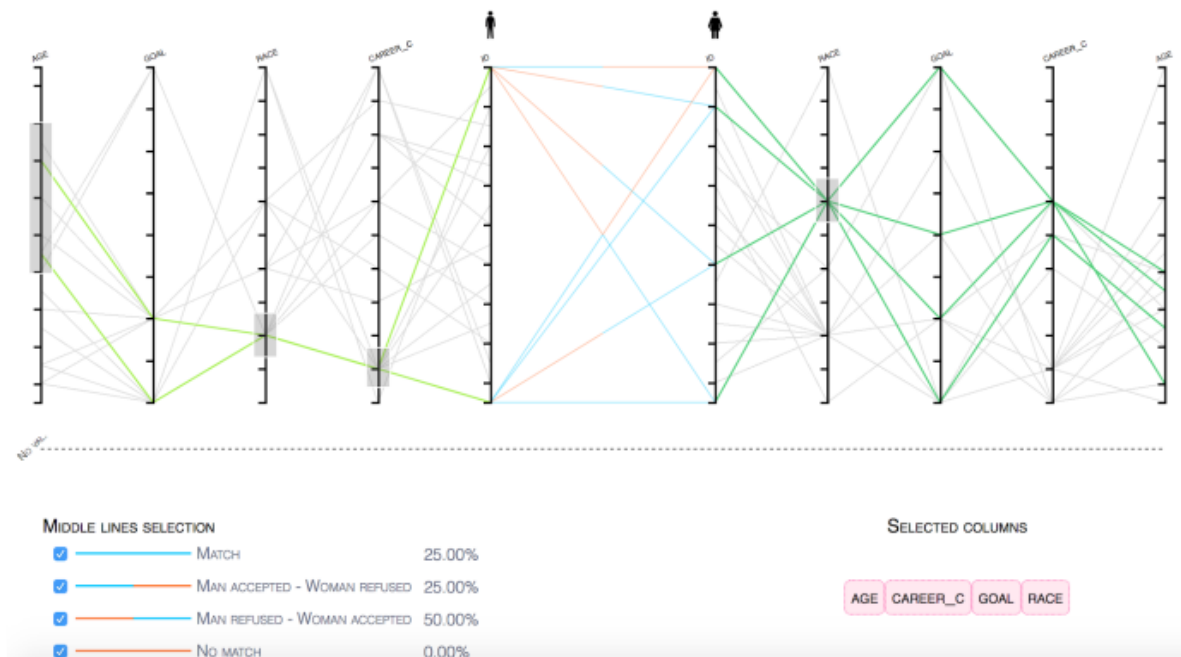
6.3. Person centred tab

Finally, we can get more information about a specific person on the person centered tab. This tab tries to summarize essential information a person would like to know concerning his dates. It could also correspond to a dashboard a coach of seduction would present to a client in order to show him through visualizations and statistics how he performed and what points he could improve next time.

The idea is to show various visualizations to represent different variables.
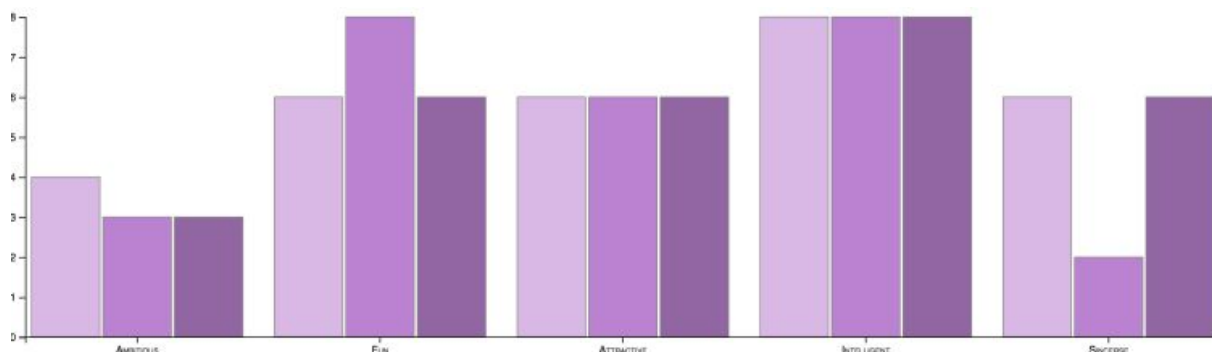
- Intrinsic information compared to the rest of the population during the wave : the age and its position in the age distribution, the income, etc.
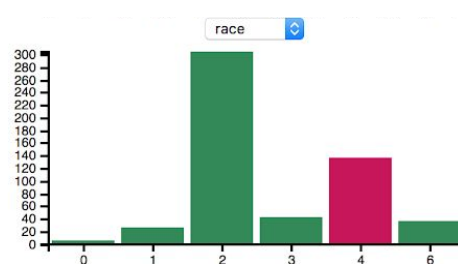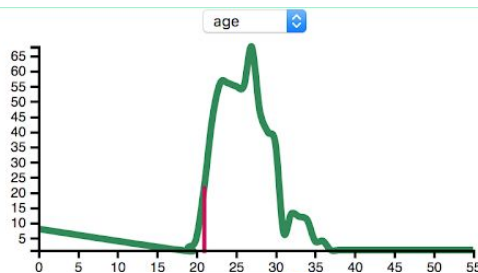
- radar charts to represent the grades a person received on different aspects of his personality.
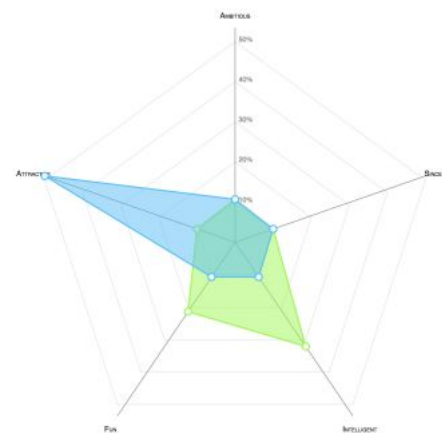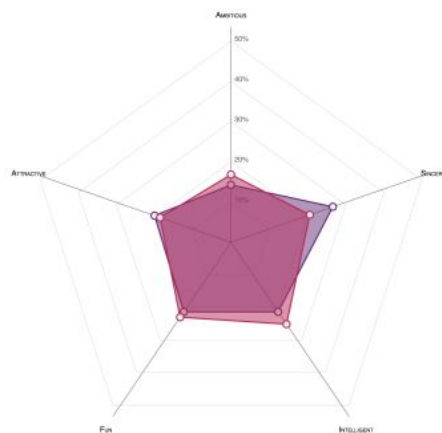
- evolution of grades over time.

7. Git repository

You can get the source code on our Git repository:

https://github.com/yleprince/never-alone