# Project IGR204

*Speed dating*

Antoine Boulat | Simon Delarue | Mohammed El Yaagoubi | Mathias Nourry | Lingli Zhan

MS BDG / IA

## I. Choice of dataset

We choose the dataset Speed Dating. This dataset is composed by columns and rows that are contained in Comma Separated Values format.
This dataset can help us to answer some of the following questions :

1. Evolution of the criteria of women/men according to their age
2. Analysis of proportions of matches between men/woman according to criteria
3. Reproduction analysis (are people looking for the same profiles as they are)
4. The bias of the financial situation (will rich people attach themselves to different profiles?)
5. Impact of money, is the financial situation crowding out the other criteria
6. Does the match rate increase with the profile completion rate?
7. ...and other questions might come after studying more deeply the dataset

**Context:**
A sociologist seeks to study the criteria of human interactions and more specifically how do couples come together and the alteration of these criteria by societal biases (money impact, social reproduction …). This study must be very broad in scope because it will appear in a press article.

**What are readers background ?**
We assume readers with a wide variety of profiles.
So, the aim of this article is to provide general information as well as a comprehensive overview.

**What is the objective of the sociologist ?**
His goal is to analyze human behavior, confront the clichés and hence help readers to develop a more defined understanding of nowadays gender expectations.
And why not even made them think about the implications of all this for our society.

**Are visualization tools used by the sociologist aimed primarily at *exploring* or *communicating* the data?**
They will be used at exploring potential correlations in the data. The purpose of this approach is to analyse interactions between groups of people having in mind all the above questions.
Then, they will be used by the sociologist to communicate his findings to readers, especially make the understanding easier for non-specialists.

## II. Description of dataset

The dataset contains 8379 entries for 195 variables. Yet, data is missing for almost 26% of the whole dataset (or is non relevant to be filled - for example if the candidates had to choose between specific items to fill).
In the data, we have got information about candidates from all around the world, like gender, age, background (studies) but also about what they expect from the speed-dating meeting, i.e their goal. Finally, for each candidate we have the answers about questionnaires that were given to them, regarding their feelings about themselves, the attributes they put in their

scorecards about the candidate they met and feeling about the event. Most of this information is already encoded as numerical values. A smaller part of the variables are still qualitative.

We have 551 unique candidates for the whole dataset, 49.94% female and 50.06% male. Data has been gathered on 21 waves of speed-dating.

**Graphiques**

1. Evolution of the criteria of women/men according to their age :
   a. **drill-down approach** : users has the choice between several interactions to explore the data
   b. differentiate the categories thanks to the filter buttons: income, sex, etc.
   c. temporal slider
   d. Graph with moving balls

2. Analysis the rating between men/woman according to criteria
   **a. martini glass approach**
   b. radar charts
   c. difference between group expectations / what groups think others expect

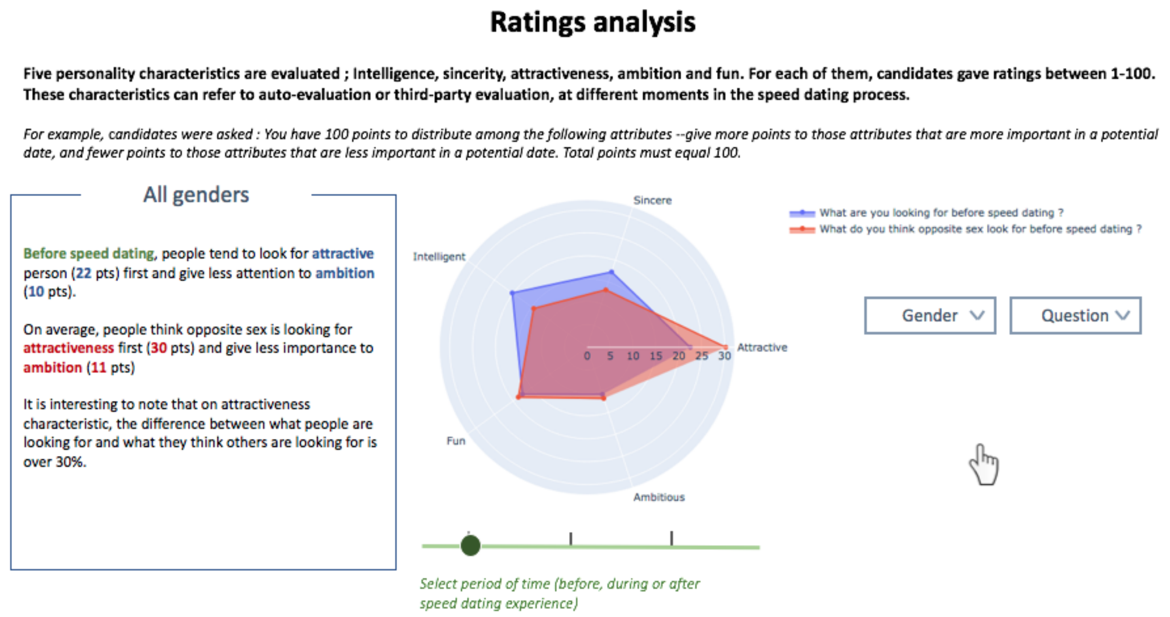
3. Which criterias are strongly correlated to the match rate ?
   **a. Interactive slideshow**
   b. more classic graphs (Altair) ; scatter plot etc
   c. study of the match rate according to several criteria

## Sketche N°1

It is quite usual that - when asked to judge and rate ourselves on subjective items such as *attractiveness* or *fun* - our own perception is different from the one proposed by a third-party. Yet, is this hypothesis just a feeling or can we measure this divergence ?
This visualization proposes to give insight to answer this question, by providing the user with an interactive analysis based on ratings fulfilled by candidates before, during and after the speed dating exercise.

For this study, we propose a **martini-glass** based approach, meaning that the user is given a global idea about the answer, with a small analysis provided, and is then invited to interact with the tool in order to find answers to more precise questions that could eventually arise during the first part.

The general overview of the visualization is the one following



**Ratings analysis**

**Five personality characteristics are evaluated ; Intelligence, sincerity, attractiveness, ambition and fun. For each of them, candidates gave ratings between 1-100. These characteristics can refer to auto-evaluation or third-party evaluation, at different moments in the speed dating process.**

*For example, candidates were asked : You have 100 points to distribute among the following attributes --give more points to those attributes that are more important in a potential date, and fewer points to those attributes that are less important in a potential date. Total points must equal 100.*

All genders

**Before speed dating**, people tend to look for **attractive** person (**22** pts) first and give less attention to **ambition** (**10** pts).

On average, people think opposite sex is looking for **attractiveness** first (**30** pts) and give less importance to **ambition** (**11** pts)

It is interesting to note that on attractiveness characteristic, the difference between what people are looking for and what they think others are looking for is over 30%.

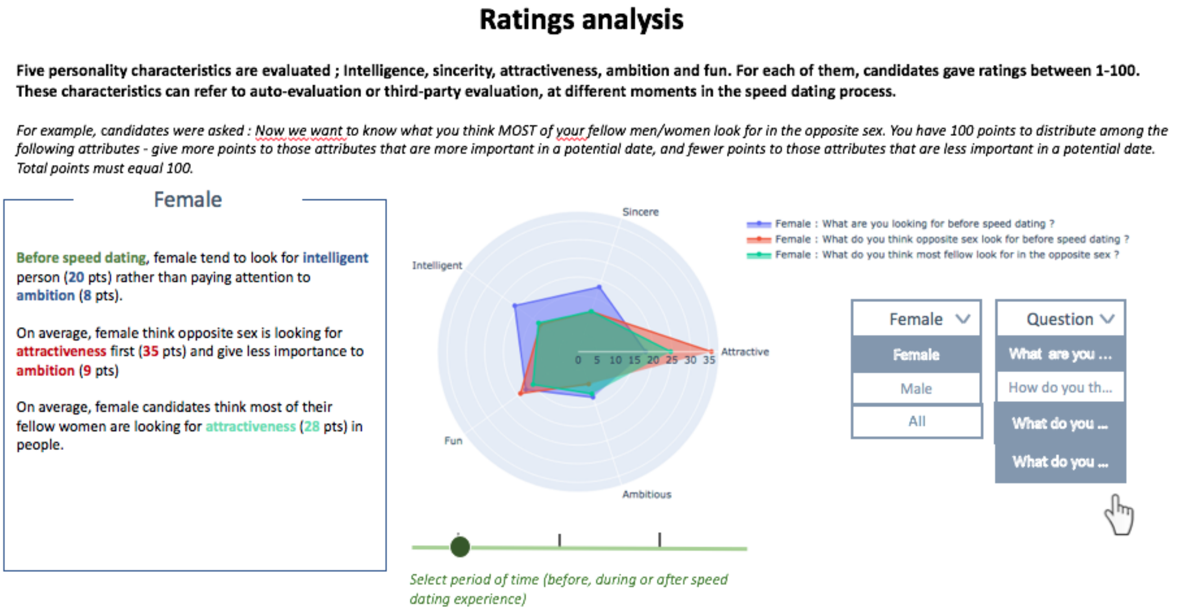
*Select period of time (before, during or after speed dating experience)*

Sincere
Intelligent
Attractive
Fun
Ambitious
0 5 10 15 20 25 30
What are you looking for before speed dating ?
What do you think opposite sex look for before speed dating ?
Gender
Question

- First, we give the user a small textual **description** about the context, allowing him to understand the visualization under the title. Also, a brief example of the rating process during speed dating is shown, in order to make very clear what candidates had to go through.
- Then the user has its attention catched by the **radar-chart** in the middle of the page. Each branch of this chart refers to a characteristic mentioned in the brief description, which helps to quickly understand the way it displays information.
  The choice of the shape itself is not random. It perfectly fits the data ; indeed, ratings are static values for each question/candidate, and it gives the user the ability to get a global view on different ratings at the same time.
  The **legend** of the chart gives the full question asked to candidates to be sure that the user understands exactly the meaning of the ratings.
- Under this chart, the user can see a **timeslider**, inviting him to select the moment where the questions were asked to the candidates. A small description under the timeslider is provided to help the user.

- As the user gets a better understanding of data, he can move along the dashboard and see - on the left of the radar chart - **information** displayed in a rectangle, under the title “All genders”. This title implies that it will be possible to filter data on gender. Under this title, details about ratings are given : **match on content** is used to refer to the legend of the chart and the time slider and suggest a dynamic update according to the user's choices. The astonishing element is detailed to the user, and helps him to further analyse the chart.
- Finally, after giving the user a global understanding on rating data - he can move on the right of the chart, where two buttons invite him to interact with the visualization. These buttons let him filter the data on different characteristics :
    - gender
    - questions asked to the candidates

Let’s have a look at what the visualization would look like when the user selects elements in the button’s lists.



Ratings analysis

**Five personality characteristics are evaluated ; Intelligence, sincerity, attractiveness, ambition and fun. For each of them, candidates gave ratings between 1-100. These characteristics can refer to auto-evaluation or third-party evaluation, at different moments in the speed dating process.**

*For example, candidates were asked : Now we want to know what you think MOST of your fellow men/women look for in the opposite sex. You have 100 points to distribute among the following attributes - give more points to those attributes that are more important in a potential date, and fewer points to those attributes that are less important in a potential date. Total points must equal 100.*

Female

Before speed dating, female tend to look for intelligent person (20 pts) rather than paying attention to ambition (8 pts).

On average, female think opposite sex is looking for attractiveness first (35 pts) and give less importance to ambition (9 pts)

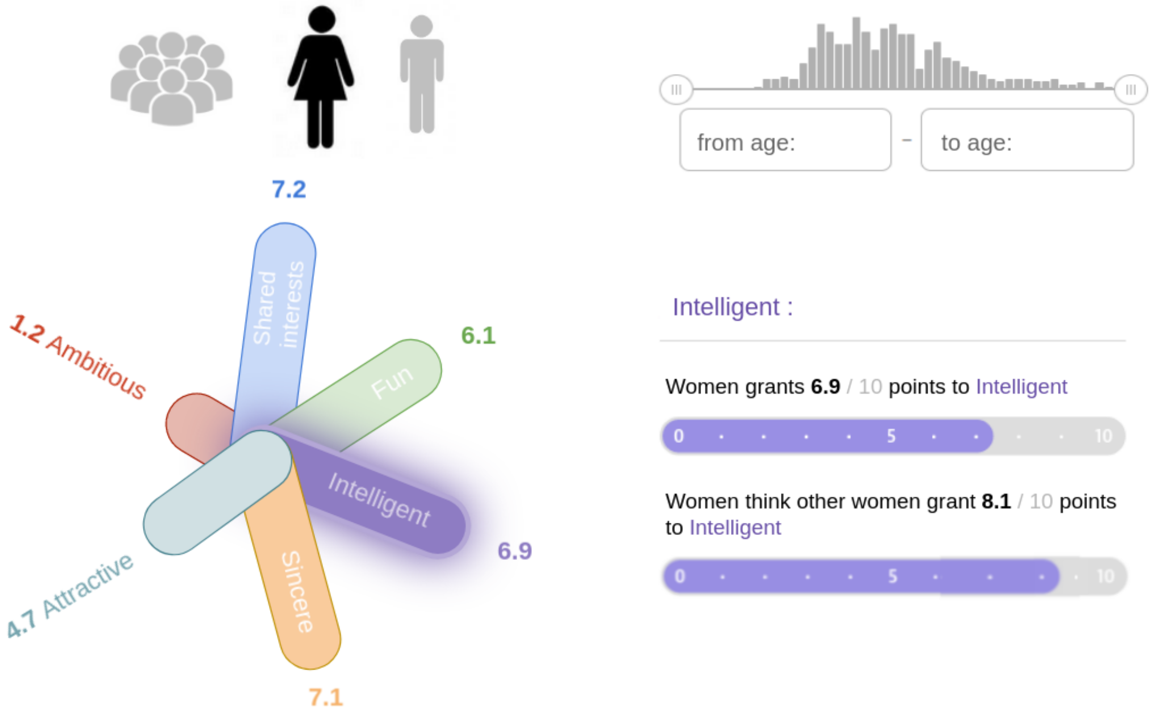
On average, female candidates think most of their fellow women are looking for attractiveness (28 pts) in people.

Female : What are you looking for before speed dating ?
Female : What do you think opposite sex look for before speed dating ?
Female : What do you think most fellow look for in the opposite sex ?

*Select period of time (before, during or after speed dating experience)*

- The title and description under it remain the same, in order to constantly remind the user with global information and context.
- The radar chart and its legend are updated with filtered data ; in this example, we are looking at ratings fulfilled by women, on 3 different questions. Note that the user should have the ability to select multiple questions.
- The title of the left rectangle is updated according to the selection. Also, the detailed information is updated to match selected data. Yet, no more conclusions about data are given to the user, only facts.

## Sketche N°2



Intelligent :

Women grants **6.9** / 10 points to Intelligent

Women think other women grant **8.1** / 10 points to Intelligent

From sketche 1, for those who are interested to have a closer look on the survey, all the information from participant, this option is available by clicking on :

Click here to access more information on participant profiles.

These datas may be of interest to a more specialised audience such as sociologists, that's the reason why it is on a separate webpage, and also because of a change of approach.

After the first part using the **martini-glass** based approach, this section, as the user is an expert, **drill-down approach** is preferred.

Having studied the difference betwwen what people are looking for in the opposite sex compare to what people think the fellow men/women are looking for in the opposite sex, it could be interesting to analyse if the gap could elvoved with age or any difference between gender.

- On the upper left, the user is able to select the interested group:
  The gender: female, male or the general population.

- On the upper right, the ‘range slide” can be added, thus allowing a deeper analysis. The associated histogram gives a good insight for evaluating the statistical confidence level ( the confidence level is higher when a large sample is selected).

Once the population is selected, the user can visualize a "flower diagram" allowing him to appreciate the importance of different criteria during the search for partners during the speed-dating. By clicking on one of these criteria, a ticket appears allowing the user to observe the difference between the importance of this criterion, for the selected population, in the choice of a partner in comparison with what they consider the importance of this criterion to the member of the same gender.

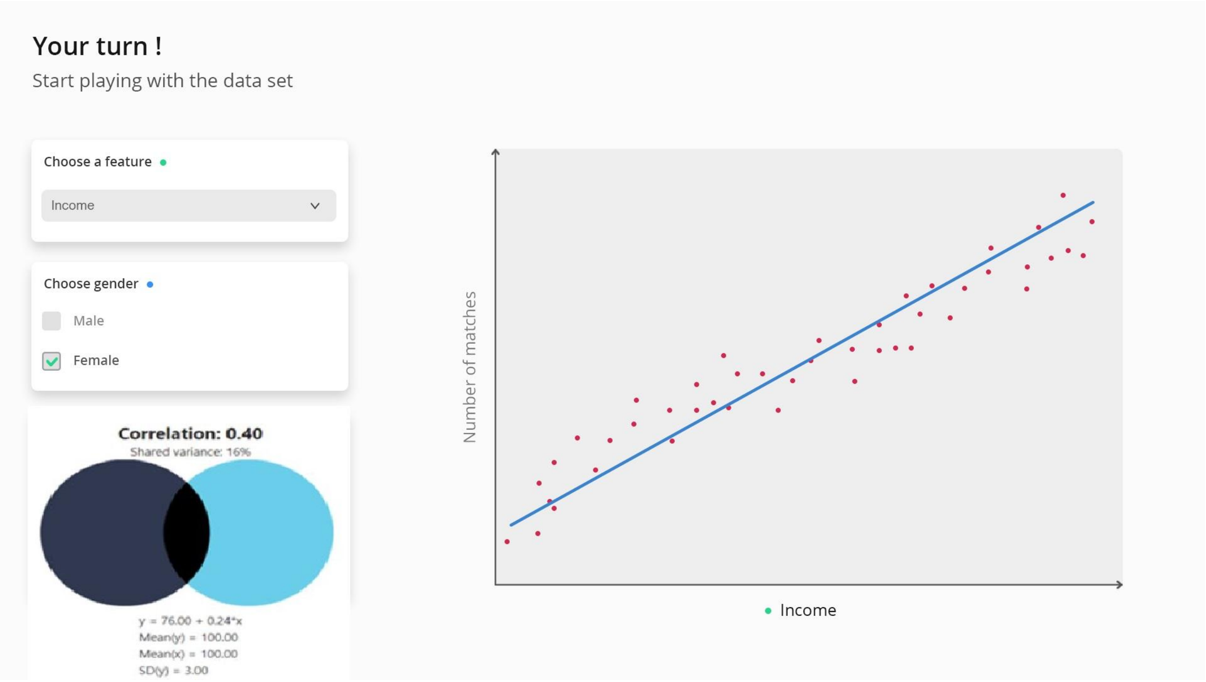
## Sketche N°3

We all want to know the secret to success in the dating world. Every self-proclaimed love guru swears to know this mysterious formula that makes people desirable to others. In the age of data, a great hunch would be to use the speed dating data with over 8,000 observations of matches and non-matches, with answers to survey questions about how people rate themselves and how they rate others on several dimensions. And that's exactly what we did!

This data visualization gives answers to this question. We suggest an interactive slideshow structure that follows a typical slideshow format, but incorporates interaction mid-narrative within the confines of each slide.

In this context, the tool would be used by sociologists : on the first slide, there would be a presentation of the tool, then a scatter plot with features ranked by their capacity to explain the number of match (income, then correlation between participant"s and partner"s ratings of interests -int_corr-...). Users will have the possibility to plot the regression line and then obtain some statistics indicators (p-value, R2).

We can use this graph to analyze the correlation between the different features and the match ratio among the candidates. The y-axis represents the number of matches while the x-axis represents the feature to be selected. The scrolling menu on the left allows the user to select the feature. The features are ranked by determination coefficient. Higher is the feature, higher this feature is correlated to match rate.



Your turn !

Start playing with the data set

Choose a feature ●

Income

Choose gender ●

□ Male

☑ Female

Correlation: 0.40

Shared variance: 16%

y = 76.00 + 0.24*x

Mean(y) = 100.00

Mean(x) = 100.00

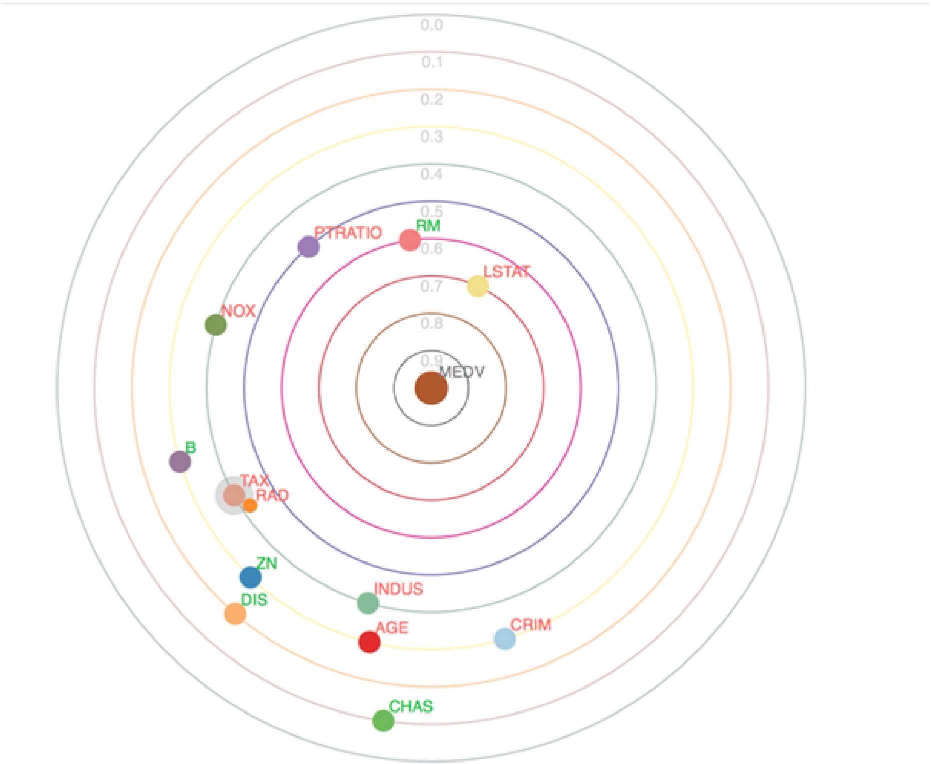
SD(y) = 3.00

Each bubble will represent a unique id (=unique candidate). The place of the bubble on the graph will depend on the number of matches and the value of the feature at hand. For instance, if the chosen feature is income, the bubbles will more likely be placed very close to the line and the line will form an angle close to 90° with the x-axis which proves that there's a strong correlation between income and the match ratio.

We can also have two analyses based on gender and therefore have a widget that enables the user to choose the study group (men or women).

We can also generate a correlation matrix and explore it thanks to a solar correlation map. The solar correlation map is designed for a dual purpose—it addresses:

- the visual representation of the correlation of each input variable (other features besides gender), to the output variable (the match ration)
- the intercorrelation of the input variables (Intercorrelation is the correlation between explanatory variables. Adding many variables, where one suffices, conjures up the curse of dimensionality).


MEDV
PTRATIO
RM
LSTAT
NOX
B
TAX
RAD
ZN
DIS
INDUS
AGE
CRIM
CHAS