**Speed Dating Data Analysis Report**

## **Data Exploring**

Using pandas functions we look at the general structure of our data (size, columns, missing data, outliers)

## **Structuring and cleaning Data**

Remove variables that will not be included in the analysis

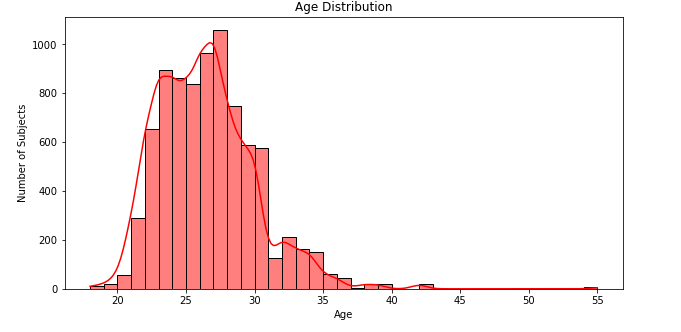
Add variables that will be used later in the analysis

Remove or replace NAN values (executed when needed) because it useless to look at all NAN values

## **Data Analysis**

## Age Distribution (in which category participants' age fall into)

1. assign all NaN values in the age column to the average age
2. generate a histogram using **seaborn.histplot** to show the age distribution of the participants

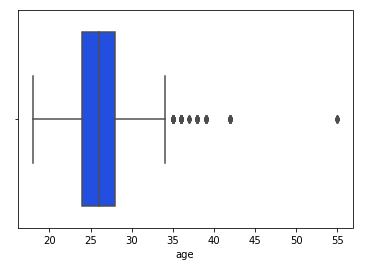


**Observation**

* Most participants were aged 22-31 years (81.9%)
* Older adults are barely interested in speed dating (only 5.8%)

Note: percentages are computed using **pandas.qcut**

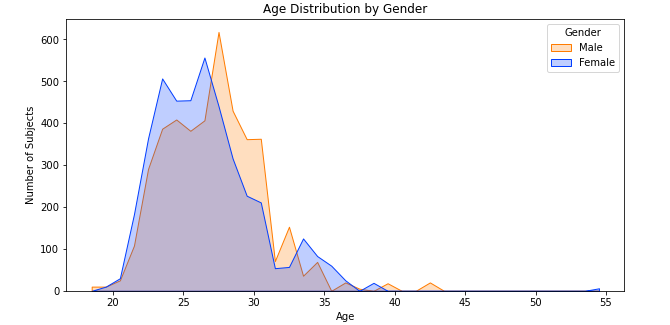
1. Generate a box plot using **seaborn.boxplot** to show the spread and centers of age distribution



**Observation**

* the box plot confirms the results of the histogram.
* There is no great variability of age among participants
* Several outliers are observed beyond 34 years old, which confirms the fact that older adults are not really interested in speed dating

1. Generate a histogram using **seaborn.histplot** to show the age distribution by gender

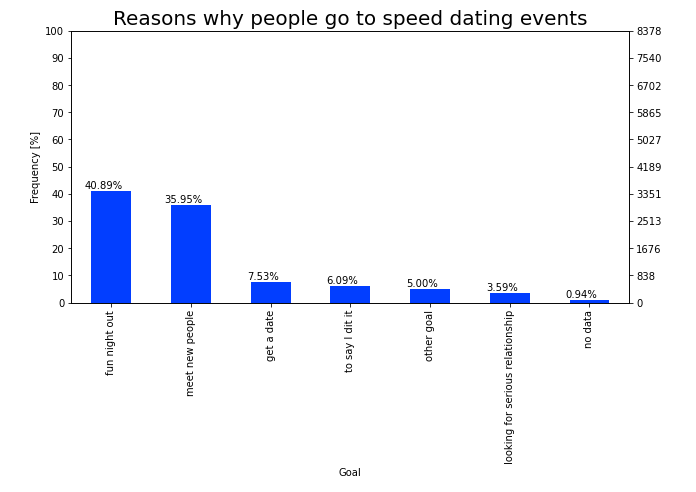


**Observation**

In young adulthood, a higher proportion of female participants were aged 20-27 than males. This pattern was repeated for middle age females (34–37 years). However, a higher proportion of male participants were aged (28-33 years) compared to females.

## Goal Distribution (What is the primary goal of subjects for participating in Speed Dating)

1. change the goal variables to include string instead of integers while keeping missing data as **no data**
2. generate a bar chart using **seaborn.countplot** to show the subjects’ reasons for participating in speed dating events with twin axes using **Axes.twinx()**



**observation**

* Most people go to speed dating events to have fun night out (40,89%) or meet new people (35,95%).
* Getting a date is actually the primary goal for only 7,53 % of participants.

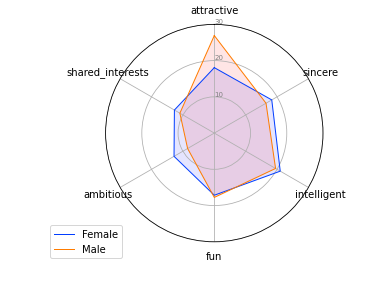
## Attributes that participants look for in their partners

1. Select relevant data (iid, gender, 6 attributes at the registration phase)
2. drop redundant data as we have unique iid
3. add a new column to sum up the rates of the different attributes

**Note**:

* the sum 0 means that all entries are missing so we drop those observations
* the data is not normalised as different scales are used as mentioned in the data description, hence we should normalize the data in order to get significative results

1. generate a radar /spider chart to find out which attributes that participants really care about and if there is a difference of behaviour between male and female participants



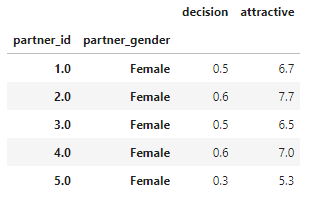
We can see that there is a great difference between what male and female participants are looking for.

* For male participants, attractiveness is given a lot more weight, and the ambitiousness or if they have any shared interset are ranked not as high.
* For females, the points are more evenly distributed across all of the attributes, with intelligence ranked slightly higher compared to others.
* There is no particular difference between male and female participants for some attributes like fun and intelligence

Men are looking for attractive women, and are less concerned with a woman’s abmition and shared interests. On the other hand, women are looking for a well-rounded male and value intelligence in a man.

## Which of the 6 Attributes Actually Influence Decision Making

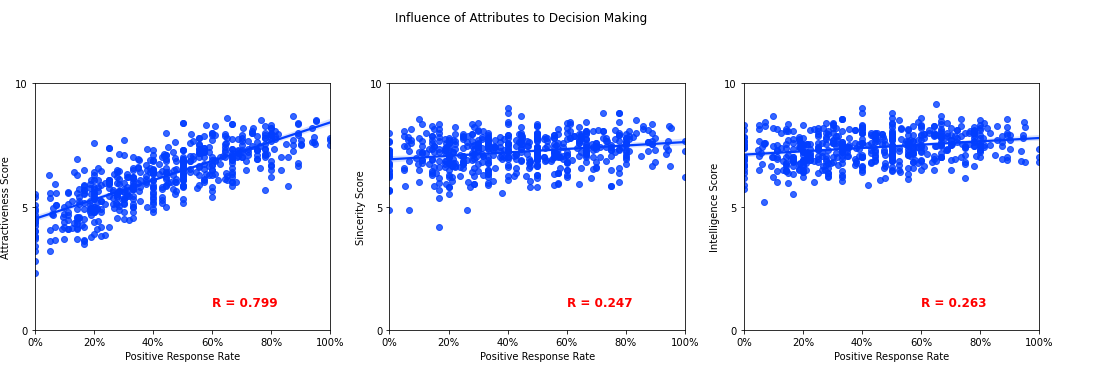
1. Select relevant data (iid, gender, 6 attributes at the speed dating event)
2. drop nan values for the variable pid (attributes are evaluated for a specific partner)
3. drop observations where all attributes are not rated because the amount of missing data for those variables is important

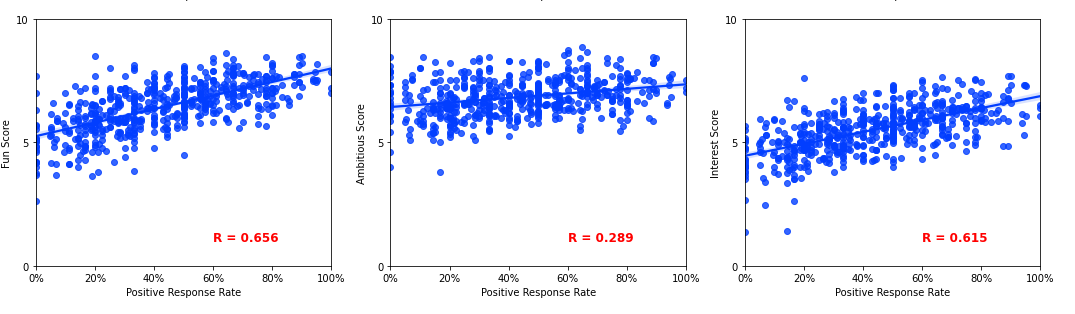


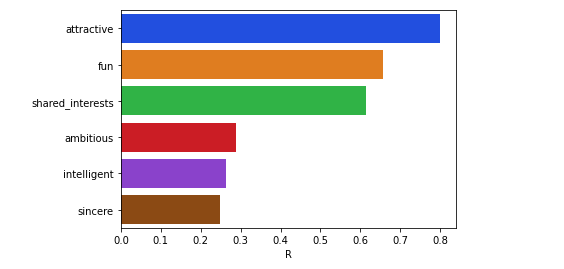
the female subject (partner\_id == 1.0) as a partner had 0.5 of yes(s) in average and rated as attractive at 6.7 in average as well (from 1-10)

**are the yes(s) he got related to his attractiveness attribute 🡪 Correlation**

1. generate a **seaborn.regplot** and **plotly.subplots** for each attribute to see the correlation between decision made and the different attributes of the partner
2. compute Pearson Product-Moment Correlation Coefficient (R score)



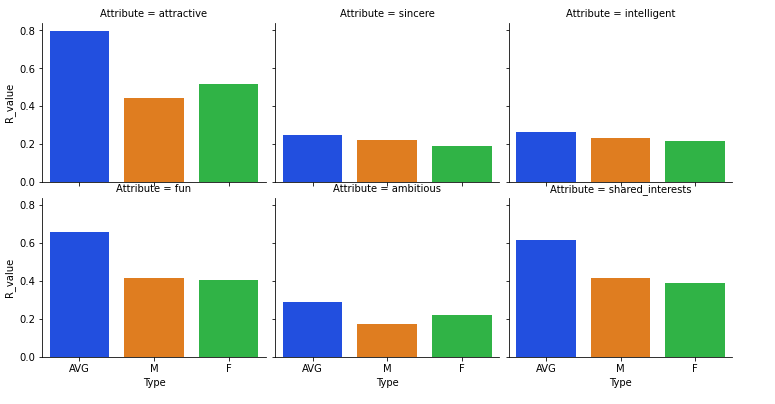




**Observation**

* looking at the graph and correlation score, the most important attribute is attractiveness of the individual, followed by how fun the individual is and if the two people have any shared interest.
* In contrast, being sincere, being intelligent, and being ambitious would not necessarily get that person a higher positive response rate.

Is there any difference between male and female participants



**Observation**

* An attractive female would win more positive responses compared to an attractive male , that confirms that man do really care about attractiveness
* A sincere male would win slightly more positive responses compared to a sincere female. Women care more for sincerity
* An intelligent male would win more positive responses compared to an intelligent female.
* No significant difference between male and female on fun scores.
* An ambitious female is highly desirable by males
* Females are more likely to give out a positive Reponses if the male has shared interest with her. In contrast, males are not so likely to give out a positive response even if the female has shared interest with him.