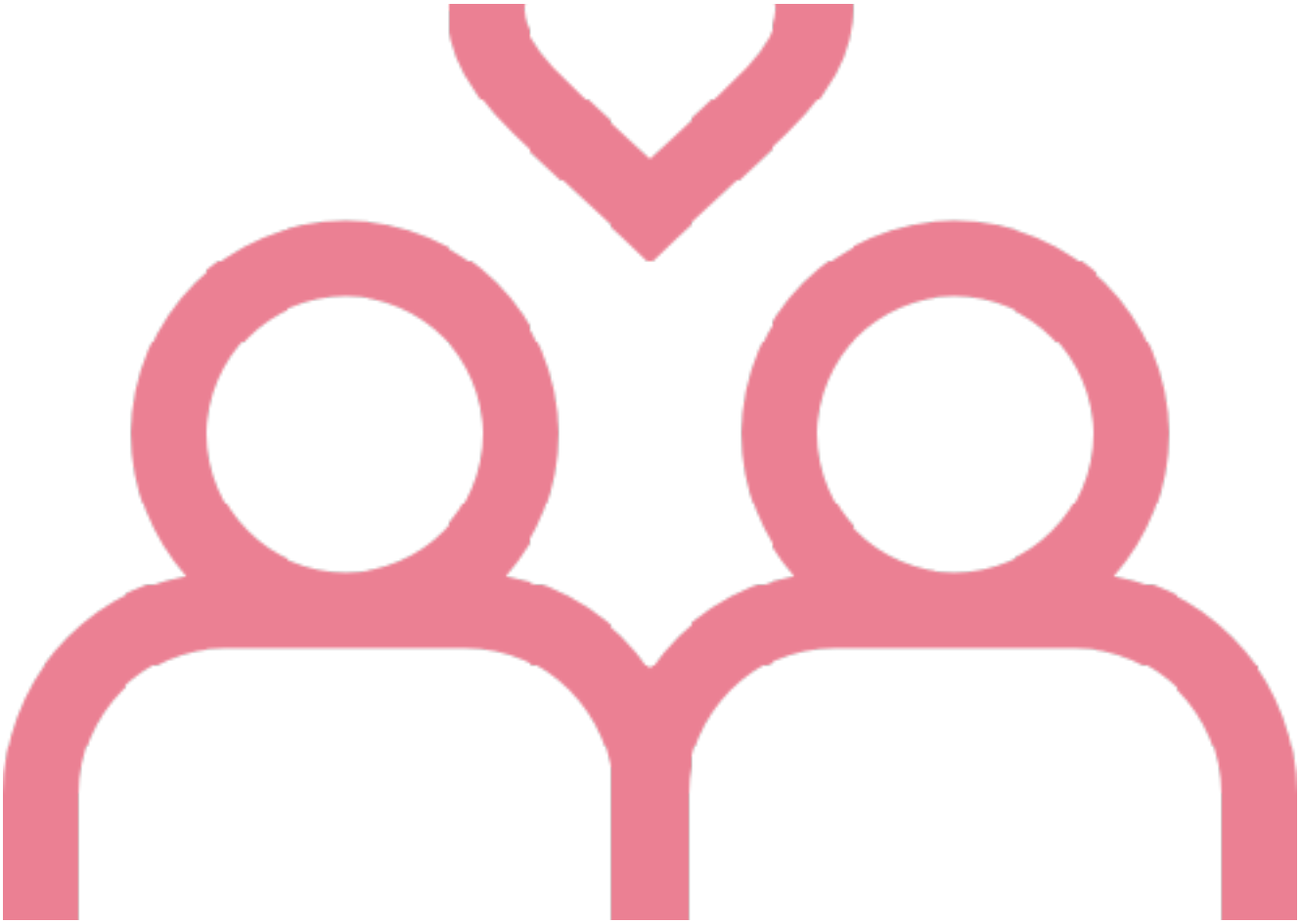
Attributes of Dating

What drives a decision for a 2nd date?

Tais Pancier - May 20, 2018



Introduction

Today's generations are looking for soul mates and have more opportunities than ever to find them. The $2.4 billion online-dating industry has exploded in the past few years with the arrival of dozens of mobile apps and dating events. Facebook has recently announced that they will be entering this market.

Part of this market, Speed dating is a formalized matchmaking process whose purpose is to encourage singles to meet new potential partners in a very short period of time through scheduled events. Men and women are rotated to meet each other over a series of short "dates" usually lasting from three to eight minutes depending on the organization running the event.

Most speed dating events match people at random, and participants will meet different "types" that they might not normally talk to in a club. According to the New York Times, participants in speed dating experience an average of 2 in 10 or 3 in 10 matches. Online dating participants, in contrast, only find a compatible match with 1 in 100 or fewer of the profiles they study.

There is an opportunity of using an algorithm to predict if people will like each other and use it to pre-match Speed dating participants and possibly increase the average of matches, which is the purpose of this project.

Dataset

The dataset is a result of a Speed Dating experiment, in which participants engage in four-minute conversations to determine whether or not they are interested in meeting each other again. If both people "accept," then each is subsequently provided with the other's contact information.

The participants were drawn from students in graduate and professional schools at Columbia University.

Upon checking in, each participant was given a clipboard with a scorecard, a pen, and a name tag on which only his or her ID number was written. The scorecard was divided into columns in which participants indicated the ID number of each person they met. Participants would then circle "yes" or "no" under the ID number to indicate whether they would like to see the other person again. Beneath the Yes/No decision was a listing of the six attributes on which the participant was to rate his or her partner: Attractive, Sincere; Intelligent; Fun; Ambitious; Shared Interests.

The morning after the Speed Dating event, participants were sent an e-mail requesting that they complete the follow-up online questionnaire. Ninety-one percent (51 percent female, 49 percent male) of the Speed Dating participants completed this follow-up questionnaire in order to obtain their matches. Upon receipt of their follow-up questionnaire responses, participants were sent an e-mail informing them of their match results.

The dataset consists of one csv file obtained in Kaggle:

https://www.kaggle.com/annavictoria/speed-dating-experiment/

The file contains the participants' gender and several sets of the 6 attributes (Attractive, Sincere; Intelligent; Fun; Ambitious; Shared Interests) recorded as a scale. At sign up, the participants were asked to rate the attributes that were most important for them in a partner and to assess how they would evaluate themselves in each of these attributes. The night of the event, they were asked to evaluate each partner based on these attributes. The dataset also contains a column that indicates how much a participant liked the people they met in a scale from 1 to 10 and a indicator of the decision for a second date. If both participants in the date indicated Yes for a second date, there's a match. Each row represents a speed date between 2 participants.

Data Wrangling

The raw dataset had 8378 rows and 95 columns. The Dataset after cleaning has 8378 columns, 69 columns.

The following steps were performed to clean the dataset:

- identified columns that wouldn't be used in the analysis and could be dropped (26 in total)
- changed 17 columns to categorical to allow proper summary analysis
- Attributes scale was different depending on the date/time of event. These columns were normalized by dividing each rating value by the sum of the related ratings (row-wise)
- Outliers:
    - 1 rating record was higher than the max of the scale for the column. As it was 1 record and just 1 point higher than the top of the scale, no special treatment was performed. This value was normalized as well.
- Missing values:
    - 19 columns with missing values were dropped because they were not relevant for the analysis
    - pid column (partner id): missing data found based on another id field and date/time of event
    - Attribute rating columns: missing data represented attributes that were not important for participants on each dating or they were not sure. NaN replaced with zero to allow correlation analysis
    - NaN in columns that might not be important for the analysis were left as NaN and will be addressed if these columns become necessary

Exploratory Data Analysis

The following steps were performed as part of the initial analysis:

1) Summary statistics for overall data and by gender:

- for both men and women the attribute ‘intelligence' has the highest mean and median
- Attribute ‘attractive' has higher mean for men than for women
- Attribute ‘ambitious' has higher mean for women than for men

	attr	sinc	intel	fun	amb
count	4042.000000	4042.000000	4042.000000	4042.000000	4042.000000
mean	0.158134	0.189935	0.198478	0.161111	0.170878
std	0.053920	0.054617	0.048718	0.046401	0.061353
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.133333	0.166667	0.173913	0.142857	0.153846
50%	0.157895	0.181818	0.190476	0.166667	0.173913
75%	0.179487	0.208333	0.216216	0.183673	0.200000
max	1.000000	0.909091	0.692308	0.421053	0.500000

Ratings the night of even given by female participants

	attr	sinc	intel	fun	amb
count	4088.000000	4088.000000	4088.000000	4088.000000	4088.000000
mean	0.171085	0.188514	0.189433	0.165598	0.158910
std	0.060274	0.047091	0.043728	0.042950	0.053378
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.145833	0.166667	0.166667	0.150000	0.145833
50%	0.166667	0.182746	0.184211	0.166667	0.166667
75%	0.190476	0.205882	0.205882	0.184211	0.183673
max	1.000000	0.666667	0.538462	0.562500	0.500000

Ratings the night of even given by male participants

2) Correlation between attribute ratings given by each participant to their partners (continuous from 0 to 10) and how much they like the partners (ordinal from 1 to 10):

- Attractive and like: r = 0.132
- Sincere and like: r = -0.159
- Intelligent and like: r = -0.260
- Fun and like: r = 0.197
- Ambitious and like: r = -0.174
- Shared interests and like: 0.2322 (higher positive correlation)

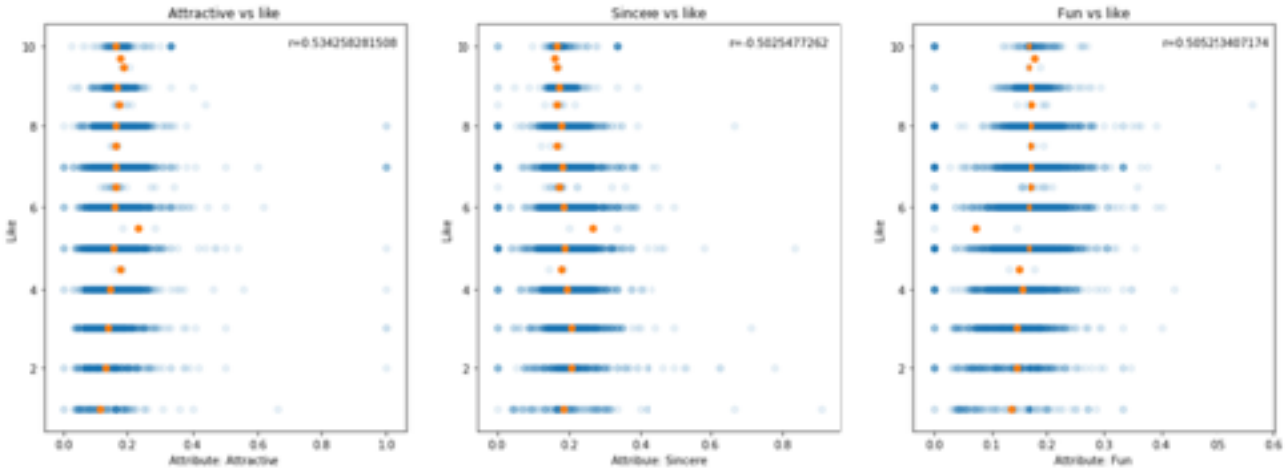
These correlations may indicate that the attributes that most impact whether a participant likes their partner are shared interests, fun and attractive.

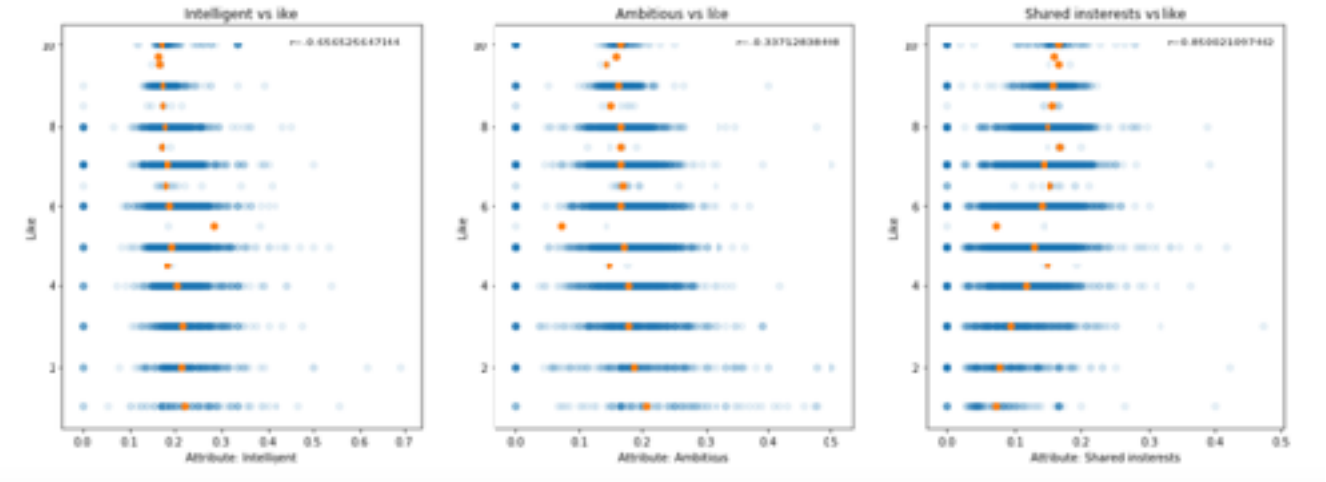
3) Correlation between median of attribute ratings given by each participant to their partners (continuous from 0 to 10) at each like level (ordinal from 1 to 10):

- Attractive median and like: r = 0.534
- Sincere median and like: r = -0.502
- Intelligent median and like: r = -0.656
- Fun median and like: r = 0.505
- Ambitious median and like: r = -0.337
- Shared interest median and like: r = 0.850

In general, without considering the decision for a second date, the numbers above show stronger correlations for shared interests, fun and attractive.

It shows a negative correlation for sincere, intelligent and ambitious. It may indicate that these attributes are not correctly perceived in a speed dating (5 min), may be interpreted as “trying too hard” to be sincere, intelligent or ambitious.



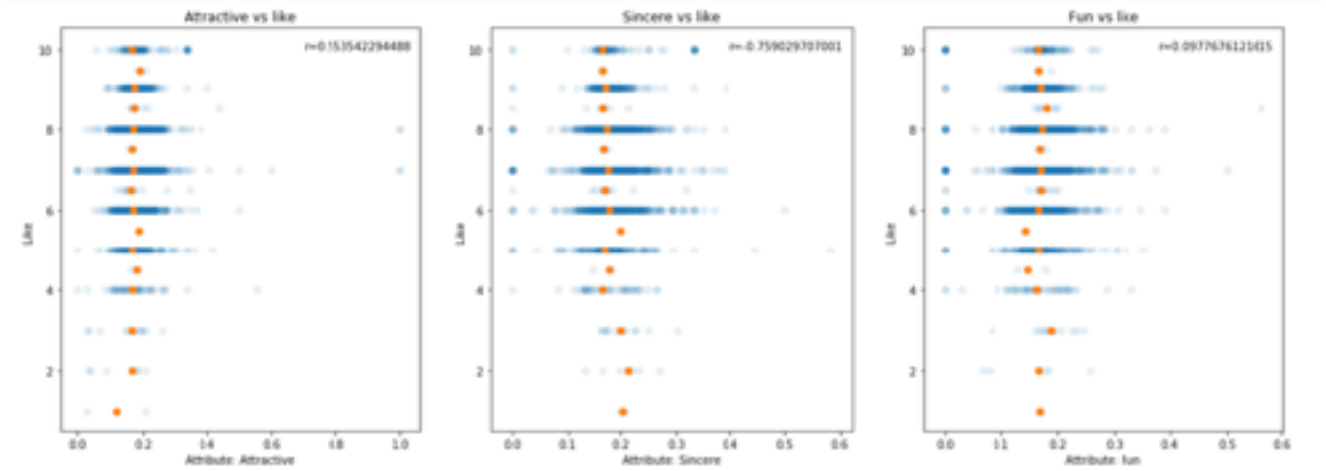


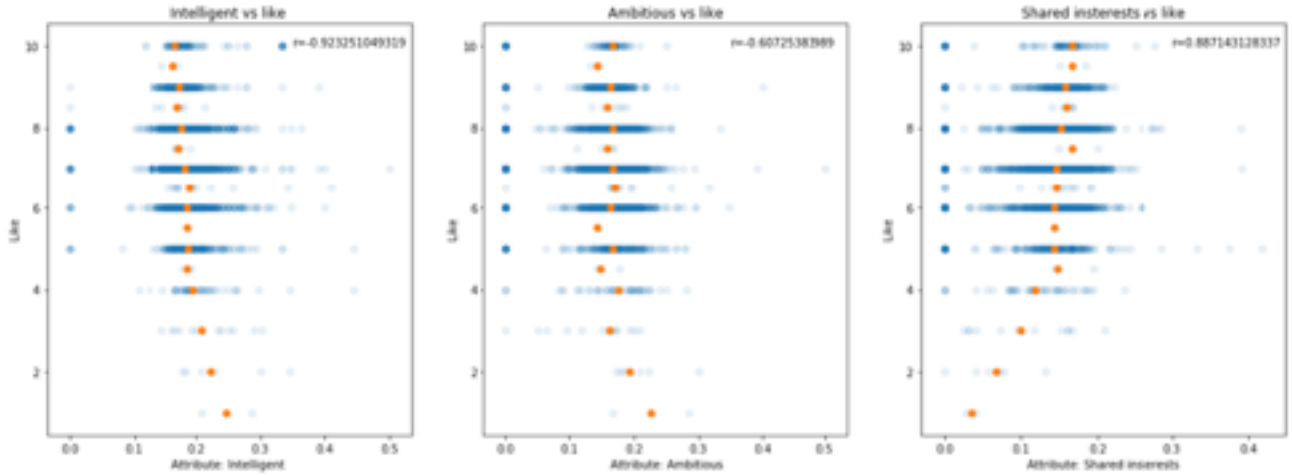
Attributes vs Like (orange: attribute median for each like datapoint)

4) Correlation between median of attribute ratings given by each participant to their partners (continuous from 0 to 10) at each like level (ordinal from 1 to 10) for partners that said yes to a 2nd date:

- Attractive median and like: r = 0.553
- Sincere median and like: r = -0.759
- Intelligent median and like: r = -0.923
- Fun median and like: r = 0.097
- Ambitious median and like: r = -0.607
- Shared interest median and like: r = 0.887

For participants that decided on a 2nd date, there is a stronger correlation for shared interests and attractive. Fun correlation is weak for participants who decided on a 2nd date.





Attributes vs Like when decision 'yes' (orange: attribute median for each 'like' datapoint)

5) Correlation between attributes rated based on importance for participants at sign-up and ratings given to partners the night of the event:

	Decision = Yes	Decision = No
Attractive	0.086546249763918573	-0.00051471268108420061
Sincere	0.020917505254446113	-0.039554598538752421
Intelligent	-0.0030664664717438223	0.0055873166593916343
Fun	0.049377289530955416	-0.0057500830536163497
Ambitious	0.097102323614510239	0.07300848738938924
Shared Interests	-0.012589961511008526	-0.0004194855264972844

Decision 'Yes' group: Weak correlation between attributes that were ranked based on importance at sign up and attributes' ranking the night of even for participants that said yes to a second date. It may indicate that the decision for a 2nd date is not driven by how correlated it is how the partner was evaluated vs attributes that are most important.

Decision 'No' group: All but intelligence attribute have negative correlation, which is expected for the group with decision = no. Weak correlation between attributes that were

ranked based on importance at sign up and attributes ranking the night of even for participants that said no to a second date.

6) Analysis between self-assessment and ratings received by partner the night of the event:

	attr3_s	sinc3_s	intel3_s	fun3_s	amb3_s
count	3518.000000	3518.000000	3518.000000	3518.000000	3518.000000
mean	0.091278	0.104728	0.106778	0.097079	0.097294
std	0.093961	0.107857	0.108881	0.100130	0.101337
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.186047	0.210526	0.210526	0.200000	0.200000
max	0.281250	0.307692	0.333333	0.290323	0.333333

Summary statistics of how participants perceive themselves for Decision Yes

	pf_o_att	pf_o_sin	pf_o_int	pf_o_fun	pf_o_amb
count	3518.000000	3518.000000	3518.000000	3518.000000	3518.000000
mean	0.227519	0.165922	0.200293	0.174836	0.107590
std	0.138729	0.071696	0.072030	0.064658	0.062776
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.150000	0.125000	0.173100	0.150000	0.050000
50%	0.200000	0.179500	0.200000	0.181836	0.100000
75%	0.250000	0.200000	0.222222	0.200000	0.150000
max	1.000000	0.600000	0.500000	0.500000	0.358108

Summary statistics of how participants are perceived by their partners for Decision Yes

	attr3_s	sinc3_s	intel3_s	fun3_s	amb3_s
count	4629.000000	4629.000000	4629.000000	4629.000000	4629.000000
mean	0.088412	0.097906	0.100823	0.093916	0.091831
std	0.095391	0.105859	0.108146	0.101353	0.100769
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.187500	0.208333	0.205882	0.200000	0.200000
max	0.281250	0.307692	0.333333	0.290323	0.333333

Summary statistics of how participants perceive themselves for Decision No

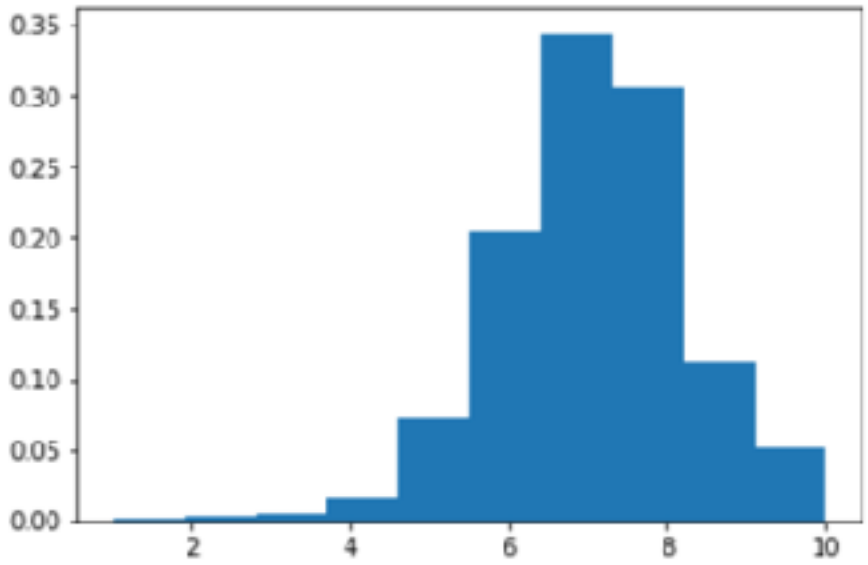
	pf_o_att	pf_o_sin	pf_o_int	pf_o_fun	pf_o_amb
count	4629.000000	4629.000000	4629.000000	4629.000000	4629.000000
mean	0.218194	0.177146	0.201322	0.170772	0.104174
std	0.116682	0.072222	0.069449	0.062077	0.059327
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.150000	0.150000	0.173100	0.150000	0.050000
50%	0.200000	0.190000	0.200000	0.173900	0.100000
75%	0.250000	0.200000	0.238100	0.200000	0.150000
max	1.000000	0.600000	0.500000	0.500000	0.358108

Summary statistics of how participants are perceived by partners for Decision No

In general participants give a lower rating for themselves in comparison to rating given by their partners in the event. The means of all self-assessed ratings are higher for participants who received a 'Yes' from their partners than for participants who received a 'No'. It may indicate that the higher a participant values themselves, most likely it is to have a hair assessment from their partner and a second date.

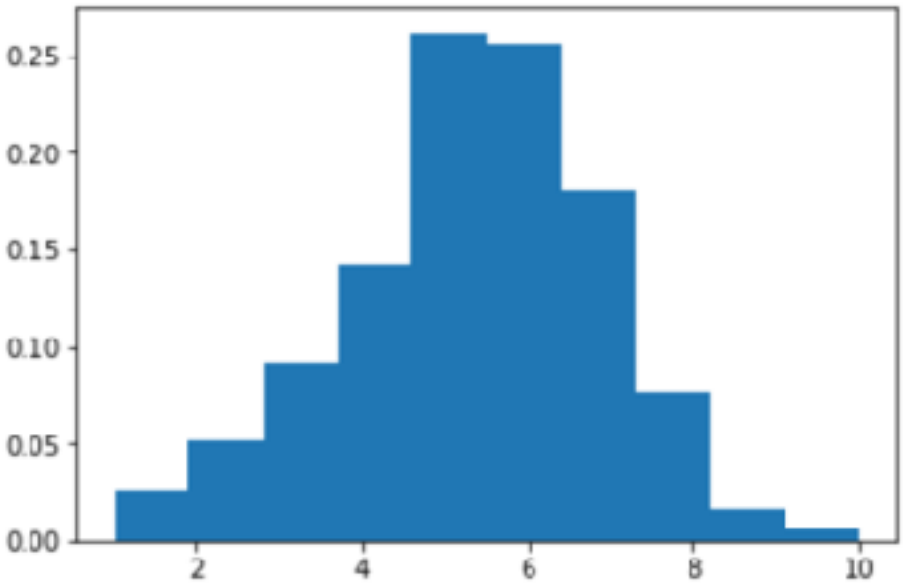
7) Like scale vs decision:

Mean of Like for participants that had decision equal to yes (decision for a 2nd date) was **7.22**.



Histogram of 'Like' when decision is Yes

Mean of Like for participants that had decision equal to No was **5.32**.



Histogram of ‘Like' when decision is No

Mean of ‘Like' is higher for participants that decided for a 2nd date compared to the ones that said No for a 2nd date. The histograms above also show that Like needs to be considerably high for someone to decide for a second date.

8) Distance between what is important for each partner at sign up:

- When there is match (both participants said yes to a 2nd date), the distance between how important is an attribute for them is smaller than when there is not a match, except for attribute 'Fun'.
- It may indicate that participants that said yes to a second date are more similar in what they value than participants who said no to a 2nd date.
- Some participants said No to a 2nd date even liking the other participant (Like > 8 in histogram).

	Decision = Yes	Decision = No
Attractive	0.0002086859994138933	0.001191432646415069
Sincere	5.946176012950125e-05	9.015316846107116e-05
Intelligent	0.00014342963366103986	0.0004077175102896651
Fun	0.00010765879640289726	5.5173271308189316e-05
Ambitious	4.639330212091883e-06	4.4306781203773666e-05

Euclidean distance of attributes ranked as important by both participants at sign-up

Insights

- Attributes that mostly impact like: Shared interests, Fun and Attractive
- Attributes that mostly impact like when the decision for a 2nd date is ‘Yes’: Shared interests and Attractive
- In general participants give a lower rating for themselves in comparison to the rating given by their partners in the event. The average of all self-assessed ratings are higher for participants who received a ‘Yes’ from their partners than for participants who received a ‘No’. It may indicate that the higher a participant values themselves, most likely it is to have a higher rating from their partner and a second date.
- The distances between attributes scaled as important for each participant at sign up are smaller for participants that said ‘Yes' to a 2nd date, except for attribute ‘Fun’.

Files in Github

Dataset	Speed Dating Data.csv
Data Wrangling	Speed Dating Dataset - Data Wrangling.ipynb
EDA	Speed Dating Dataset - EDA.ipynb