Speed Dating Project Midterm Report

Marlene Berke and Anna Xu

I. The Data

The dataset we are using for this project was found on Kaggle.com. The data is on 552 participants during 21 speed dating events from 2002-2004. Participants who attended the same speed dating event (and thus filled out all the same surveys) are called “waves.” During the speed dating event, participants were asked to fill out questionnaires about themselves and their partner. Each speed dating participant rates themselves on a scale from 1 to 10 on five key attributes: Attractiveness, Sincerity, Intelligence, Fun, and Ambition. They also rate their date those same five key attributes, plus Shared Interests. They then rate how much they liked their partner (1-10 scale), and they give a yes-or-no decision as to whether they would like to see that partner again. Furthermore, participants filled out surveys about what they look for in the opposite sex, what they think the opposite sex looks for in their gender. Some waves answered these questions at different time intervals before and after the speed dating event.

The dataset also gives us demographic data on each participant, like their race, age, hometown, job, income, and hobbies. In the original dataset, each row of the 8378 rows is an interaction between two participants. There are 551 participants total, and 195 features on each interaction (from demographic information about the participants to their assessment of their own attractiveness a couple of weeks after the event).

Not all of the waves were asked the same questions at the same time intervals. When a particular participant or wave did not answer a particular question, the is an NA. To explore a particular feature, we dropped the interactions for which there were NAs. After omitting all the rows that had NAs in any of the attributes or demographic information that we care about, we were left with 6740 interactions.

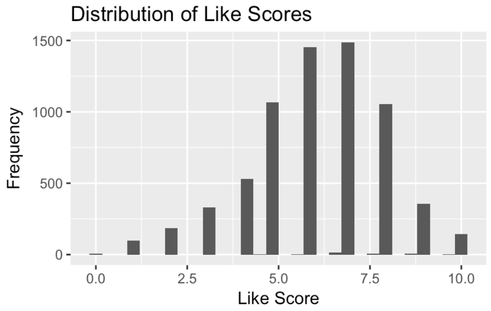

Figure 1 - The scores for liking (1-10) look normally distributed. They have a mean around 6.14 and a standard deviation of 1.84.

Practical note: from here on out, we will use second person. Although less formal, it’s the clearest way of distinguishing between the individual of interest and his/her partner. “You” will refer to the individual of interest, and “your partner” will refer to your partner.

II. Preliminary Analyses

Question 1: How does how you score someone on Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests, along with demographic background and their liking of you, impact how much you like them?

We found that your assessment of your partner’s attributes showed a strong positive trend with liking. In retrospect, this seems obvious: if you like someone, then you probably think that they are sincere, intelligent, fun, attractive, etc. And if you think that they are sincere, intelligent, fun, attractive, etc., then you probably like them. The correlation matrix shows the high correlation between attribute rating and liking. It also shows a small (0.12) positive correlation between your liking of your partner and your partner’s liking of you. In a linear regression of liking against same race, whether the partners were of a same race was a significant factor ($p<0.001$). The other factors, like age difference, didn’t seem to matter.

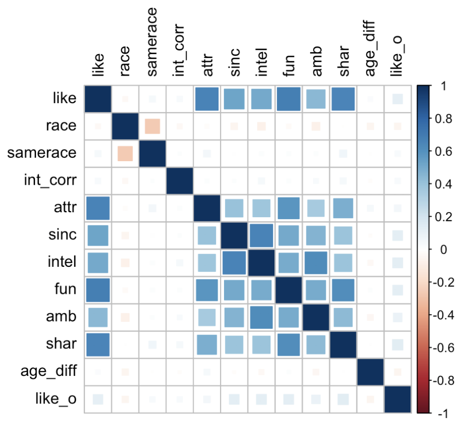

Figure 2 - Correlation Matrix of all the features.

Question 2: Can your demographic information predict your self-awareness? In particular, how does a person's perception of his/herself differ from potential partners' perceptions of him/her? Do people of a particular background have more distorted self-images? Can a person accurately predict how others will perceive him/her?

To do this, we created a self-awareness score for each person by calculating the mean difference between other people's rating of you and your rating of yourself on the 5 key attributes vs how their partners rated them on the 5 key attributes. A negative self-awareness score mean that you overrate yourself, and a positive means you underrate yourself. A score of 0 mean you're perfectly accurate. So, on average, people rated themselves a point higher on each attribute than other people rated them.

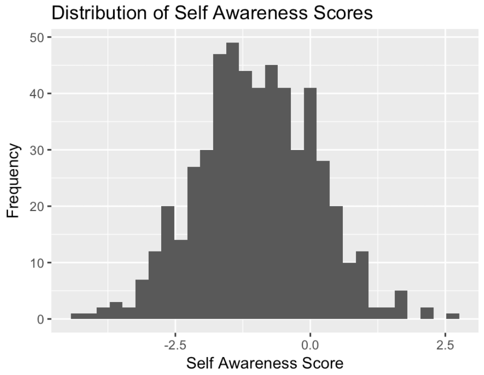

Figure 3 - The Self Awareness Scores look normally distributed. The mean self awareness score was -1.071934 and standard deviation was 1.13.

We explored whether demographic information like age, gender, race, and career field could predict self-awareness / veridicality of self-image. We found that Latinos and Asians had significantly more accurate self-awareness (according to a linear regression, $p<0.0244$ and 0.0133 respectively), or overestimated themselves less than other groups did. Gender, age, and profession/major had no significant effects.

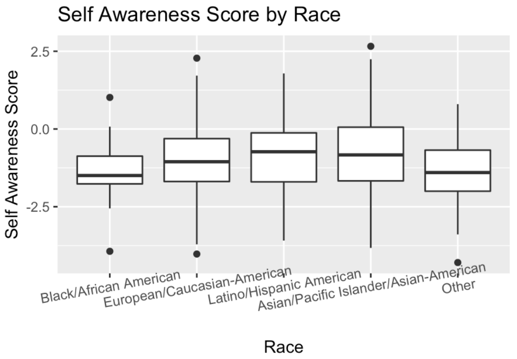

Figure 4 - Self Awareness Scores are lowest for Blacks and Others, meaning they scored themselves higher than their partners scored them.

Participants were also asked how they expected that others would rate them on the 5 key attributes. We calculated an accuracy score comparing how you expected to be evaluated with how you were evaluated, using the same metrics as we used to make the self-awareness score. We call this the expected rating score. For the 313 people who answered both questions, the mean self-awareness score was -1.039492 and the mean expected ratings score was -0.8586617. A paired t-test ($p<10e-05$) showed a significant difference between how you see yourself and how you expected others to see you: on average, you think that people will rate you a bit lower than you would rate yourself. In other words, you partially correct for your overrating of yourself, but not entirely.

Question 3: Who is popular? Who is unpopular? What are the demographics of each group?

Demographics of all participants	Female	Male
Black/African American	16	10
European/Caucasian-American	140	160
Latino/Hispanic American	24	17
Asian/Pacific Islander/Asian-American	71	64
Other	16	21

Table 1 - Demographic breakdown of the 539 participants in the study who answered all the relevant questions. 55.7% of participants were White and 25.0% were Asian.

We looked at the 20 most popular people and 20 least popular people. The 20 most popular people were those who received an average “liking” score of 7.65 or higher (see Table 2). Of those people, only 1/20 were Asian, despite that 25% of speed dating participants were Asian (see Table 1). Strikingly, the distribution of women among the top twenty more closely follows the proportions expected based on race, while white men were heavily prefered over any other race.

20 Most Popular	Female	Male
Black/African American	1	0
European/Caucasian-American	4	8
Latino/Hispanic American	2	0
Asian/Pacific Islander/Asian-American	0	1
Other	2	1

Table 2 - Demographic breakdown of the 20 most popular

The 20 least popular people received an average liking score below 4.57. Of the 12 least popular men, 7 were Asian, while only 4 were White. This difference is so striking because more than twice as many White males as Asian males participated. The trend is less pronounced for females. When lumping the genders together, 10 of the 20 least popular speed daters were Asian, while only 8 were White. It looks like race highly predicts how much a participant is liked by partners, benefiting White males and disadvantaging Asian males.

20 Least Popular	Female	Male
Black/African American	1	0
European/Caucasian-American	4	4
Latino/Hispanic American	0	0
Asian/Pacific Islander/Asian-American	3	7
Other	0	1

Table 3 - Demographic breakdown of the 20 least popular

III. Next Steps

Our nexts steps are to randomly select 7 out of the 21 dating “waves” to leave as a test set and work solely on the training set. On our training set, we plan to fit two linear models to answer the two questions posed above. One model will attempt to predict how much you will like your partner and the other will predict your overall popularity. For both of these questions, we will develop a few variations of the linear model and use 5-fold cross validation to choose one that gives the smallest MSE, helping us to avoid overfitting. We will test how effective our model is by showing how accurately it performs on our test set, and computing a test error rate.

We are also hoping to incorporate some of the other techniques we will learn in class such as using logistic regression to predict whether or not you would like to see your partner again and K-means clustering to see if there are “types” of speed daters.