PSTAT 100 Project Report: An Analysis of a Speed Dating Dataset

Cynthia Shim, Anum Damani, Annie Huang

Author contributions

Cynthia Shim studied the data documentation and prepared the data semantics/description as well as analysis.
Anum Damani described the background of the dataset and worked on the exploratory analysis.
Annie Huang prepared the tidied dataset as well as various graphs and analyses.

Abstract

We are interested in exploring various variables that may or may not be correlated to a participant's match rate in a Speed Dating experiment. We utilize visualization tools such as a heat map to guide us into a more thorough investigation to find correlation values through regression analysis, as well as various plots to show our findings. We were able to see that a participant's match rate was positively correlated with their expected number of matches by the end of the event.

0. Background

This project is an exploration of a Speed Dating dataset of subjects who met potential partners through an experiment designed, not only to facilitate matches but also, to provide insight into the dating preferences and possible correlation of variables that lead to a successful pairing. The raw Speed Dating dataset contains 8,378 observations that include participants, variables such as age and race that describe the subjects' matched and unmatched partners, as well as answers to survey questions that ask participants to rate themselves and others based on different criteria. This dataset relates to the areas of psychology, sociology, and human interaction. The motivation behind collecting data of this nature is to hopefully gain insight into the kinds of external and internal factors that lead to a romantic match, as well as seeing how these factors affect a participant's expected level of happiness from speed dating.

This data was collected by Columbia Business School professors Ray Fisman and Sheena Lyengar and it has been used in Ray Fisman's publication "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment" in The Quarterly Journal of Economics, May 2006. To provide more context on how the experiment was conducted, subjects were asked to meet a designated number of potential partners (between nine and twenty one, depending on the night of the event) for four minutes each and were given the opportunity to either match with or reject their partners of the evening. If participants were interested in seeing their matched partners again (meaning that both parties selected "yes" for wanting to meet in the future), they were asked to fill out an online questionnaire of survey questions which would then release an email informing them of their matches and their contact information upon completion.

1. Data Description

Participants of this study were recruited through mass email and fliers posted around Columbia University campus as well as advertised by research assistants. In other words, subjects of this study participated voluntarily and the data collected through surveys were also collected through voluntary participation, with the incentive of receiving matches. Participants were invited to different dates that spanned approximately one and a half years (10/16/2002-04/07/2004), and it may be important to note that certain dates had specific variations, for example, Nov 20th, 2002 had all undergraduate student participants.

We believe that the relevant population for this dataset would be heterosexual single adults who are looking to find matches through speed dating. Since the experiment was publicly posted and people within the general area opted to participate due to personal interest, the relevant population for this study would likely be people who are also interested in this method of dating as well, and while we do not know why, this study and Ray Fisman's publication focused only on heterosexual relationships, that also places a limitation on who this study may be relevant to. The sampling frame in this study are people who are around the Columbia Business School area (New York) and our sample is the adults around this area who registered for this event and filled out survey responses to contribute to the data collection. As of now, we believe that our data had a good sampling design as the scope of inference is broad and because there is a wide range of ages, interests, and career paths in our sample, we believe the conclusions we will come to based on this dataset will be representative of the relevant population we have defined.

In the raw Speed Dating dataset, the observational unit was individual meetings between participants. In an effort to tidy the data and make things more clear to analyze, we condensed the dataset to be grouped by participants' ID numbers, so some of the variables were modified to become averages of a single participants' total meeting experiences. As a result, the new observational unit for the dataset that we are using is individual participants in the Speed Dating experiment. It may also be important to note that most of the variables in this dataset are discrete variables (only one non-discrete variable was given, int_corr, which we took the average of for each participant to create a variable called "Int Corr Avg"), and we created another non-discrete variable called "Match Rate".

Variable Descriptions

Variable Name	*Variable Description*	*Type*	*Units of measurement*
Unique ID	Unique ID of Participant	Numeric	Unique Subject Number
Age	Age of Participant	Numeric	years of age
Gender	Gender of Participant	Numeric	0=Female, 1=Male

Field	Field; the field in which the speed dating participant works	Numeric	1=Law; 2=Math; 3=Social Science, Psychologist;
			4=Medical Science, Pharmaceuticals, and Bio Tech;
			5=Engineering;
			6= English/Creative Writing/ Journalism;
			7=History/Religion/Philosophy;
			8=Business/Econ/Finance; 9= Education, Academia;

			10= Biological Sciences/Chemistry/Physics; 11= Social Work; 12=Undergrad/undecided; 13=Political Science/International Affairs; 14=Film; 15=Fine Arts/Arts Administration; 16=Languages;17=Architecture; 18=Other
exphappy	How happy do you expect to be with the people you meet during speed dating?	Numeric	Scale from 1 to 10 (where 1 is the least happy and 10 is the most happy)
expnum	Out of the 20 people you meet, how many do you think will be interested in dating you?	Numeric	Scale from 1 to 20 (which represents the number of people)
Goal	What is your primary goal in participating in speed dating?	Numeric	1=Seemed like a fun night out; 2=To meet new people; 3=To get a date; 4=Looking for a serious relationship; 5=To say I did it; 6=Other
Race	What is your race?	Numeric	1=Black/African American; 2= European/Caucasian-American; 3=Latino/Hispanic American; 4=Asian/Pacific Islander/Asian-American; 5= Native American; 6=Other
Race Importance	How important is it to you that your partner is of the same race?	Numeric	Scale from 1 to 10 (where 1 is means not important and 10 means very important)
Religion Importance	How important is it to you that your partner is of the same religion?	Numeric	Scale from 1 to 10 (where 1 is means not important and 10 means very important)
Income	Median household income based on zipcode using the Census Bureau website	Numeric	U.S. Dollars
Yes Match	Sum of Yes matches participant received in total	Numeric	Number of Total Yes Matches (Match=1)
No Match	Sum of No matches participant received in total	Numeric	Number of Total Yes Matches (Match=0)
Total Meetings	Count of total number of meetings participant had	Numeric	Number of Total Yes Matches + Number of Total No Matches
Match Rate	Rate of Matches	Numeric	Number of Total Yes Matches / Total Number of Meetings
Int Corr Avg	Correlation between participant's and partners ratings of interest in Time1	Numeric	Correlation measurement

Aims

We are interested in analyzing the factors that affect a participant's match rate from speed dating, specifically exploring which variables seem to be highly correlated, if any. After conducting our initial data explorations to become more familiar with our dataset (as seen in the appendix), we decided to create a heat map to check for correlations. Most notably, we saw that race importance and religion importance seemed to be positively correlated as did the expected number of matches and match rate. We were particularly interested in exploring further whether one’s expected number of matches was correlated with match rate, so after conducting a simple linear regression on “Match Rate” and “Int Corr Avg” (which was mainly motivated by both variables being continuous), we conducted multiple linear regression analyses with “expnum”, which we found had relatively high correlation! We continued our search with other covariates such as gender and expected happiness from speed dating (variable “exphappy”), but we found that both exhibited very low values of R^2.

We finished off our investigation curious about two more variables, age and race, and through various plots we found that there did not seem to be an overall trend in match rate and age, taking into account possible dips in our line plot occurring due to variations in our sample quantity. However, there did seem to be a possible trend in median match rate and race.

2. Methods and Results

Heatmap of quantitative variable correlations

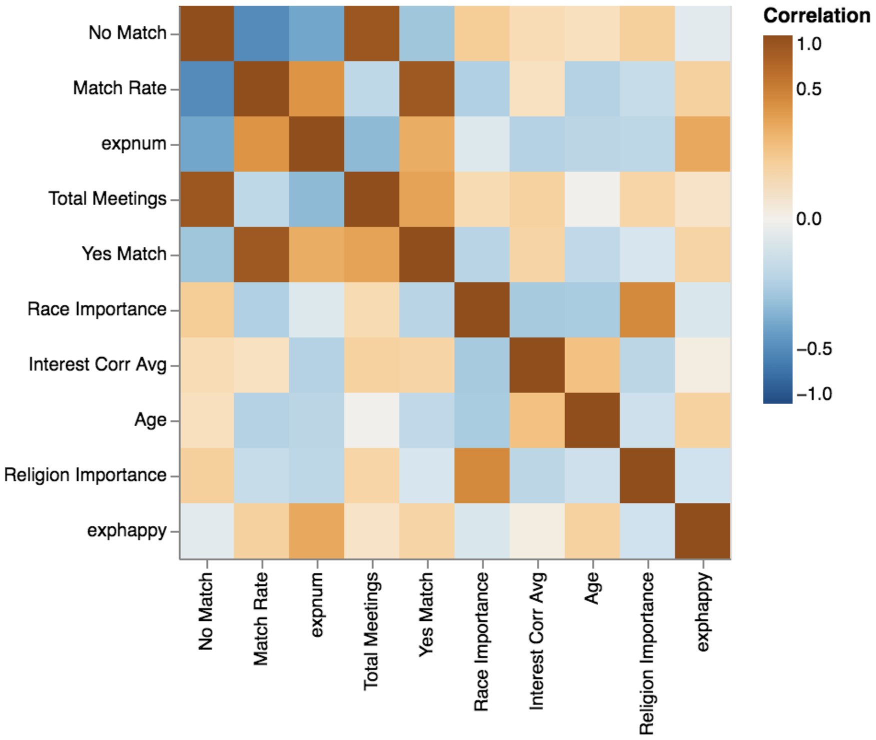


Figure 1: Heat map displaying the correlation between variables, with the blue-orange color gradient over the extent (-1,1) where zero correlation is white. Cells that are darker orange suggest a high positive correlation while cells that are a darker blue suggest a high negative correlation.

In Figure 1, we noticed that race importance and religion importance seem to be very positively correlated as it is denoted by a slightly darker orange shade. We can also see that “match rate” and expected number of matches (“expnum”) seem to be positively correlated as denoted by a slightly less dark orange shade, which, since we are interested in the factors that affect match rate, we found especially fascinating. We decided that one possible, interesting way to continue on our exploration on the correlation of various variables and “match rate” would be to conduct regression analyses, including one on the expected number of matches, variable “expnum”.

Since “int corr avg” and “match rate” are two continuous variables and the heat map suggests that these two variables are slightly, positively correlated, we decided to conduct a simple linear regression on these variables.

Regression analysis of match rate, interest correlation averages, and expected number of mutual matches

We began by conducting simple linear regression and multiple linear regression on the expected number of matches that a participant expects to receive by the end of the speed dating event.

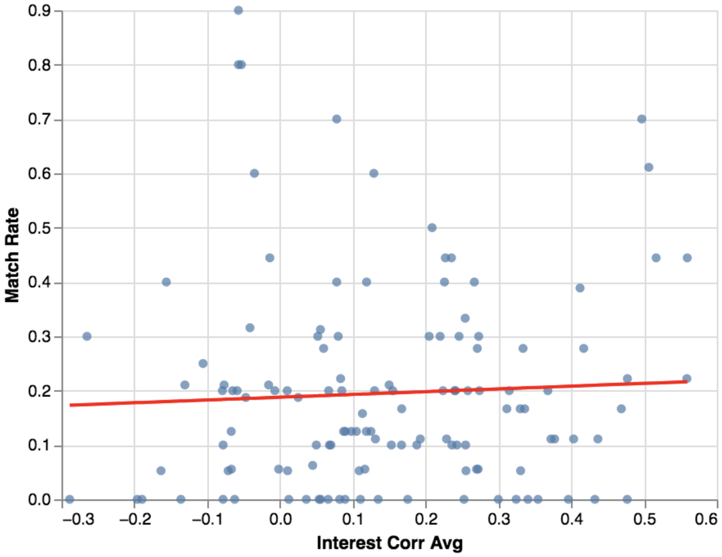


Figure 2: Simple Linear Regression line plot of fitted values against interest correlation average. The red line represents the fitted values. The match rate is on the y-axis and the interest correlation average is on the x-axis.

In Figure 2, we found that among the people in the sample, a one-unit increase in the interest correlation average is associated with a 5.098% increase in rate of matches. For the simple linear regression, the R^2value was approximately 0.00251146, and this is significantly low. We performed multiple linear regression in order to determine if the fit considerably improves.

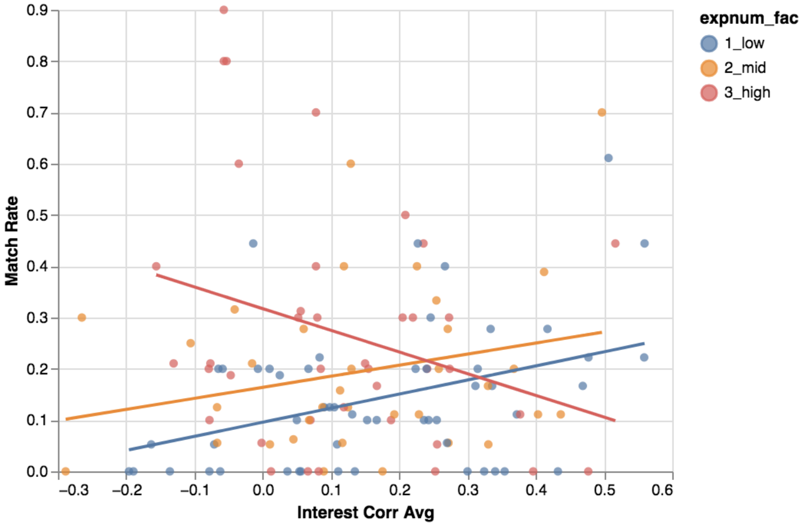


Figure 3: Multiple linear regression line plot of fitted values (y-axis) against interest correlation average (x-axis) with the expnum factor mapped to the color aesthetic. In this plot, blue points indicate a high level of expected matches from speed dating, yellow points indicate a mid-level of matches expected from speed dating, and red points indicate a low level of matches expected from speed dating.

We created the expnum_fac factor by generating three evenly-split levels based on the values of the expnum variable in the dataset. We chose to add the expnum_fac (which is categorical) to the model. We found that the R^2 for the multiple linear regression model is 0.168047. Therefore, since the R^2 value for the multiple linear regression model is higher than the value for the simple linear regression model, we concluded that the fit considerably improved.

For the 1_low and 2_mid expnum_fac groups that participated in the study, the multiple linear regression model shows that there is a positive correlation between interest correlation average and match rate; in other words, as interest correlation average increases, the match rate increases for the participants who expected to have a low or medium number of matches during the event. However, for the 3_high expnum_fac group, the multiple linear regression model explains that there is a strong, negative correlation between interest correlation average and match rate; as interest correlation average increases, the match rate decreases for the participants who expected to have a high number of matches during the event. There is a stronger, positive relationship between interest correlation average and match rate for the 1_low group than the 2_mid group.

Regression analysis of match rate, interest correlation averages, and expectation of happiness

We continued our analysis by conducting simple linear regression and multiple linear regression on the expected level of the happiness of the participants with relation to the partners they meet during the speed dating event.

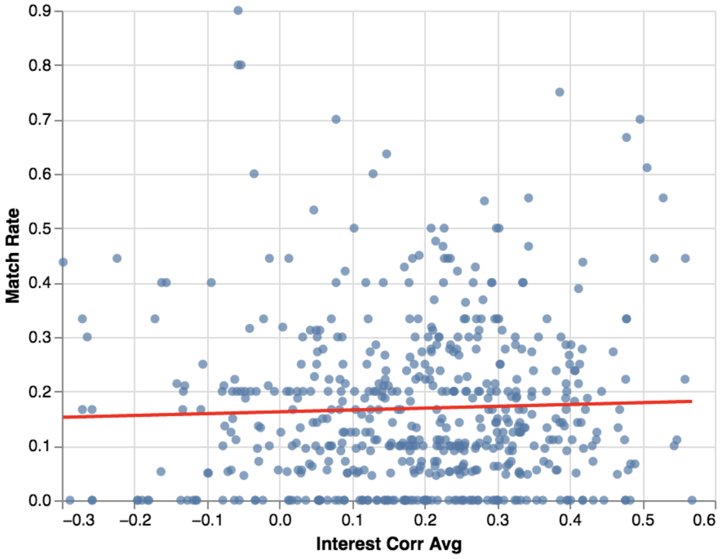


Figure 4: Simple Linear Regression line plot of fitted values against interest correlation average. The red line represents the fitted values. The match rate is on the y-axis and the interest correlation average is on the x-axis.

Based on Figure 4, we see the fitted line does not go straight through all of the data points, and therefore, does not capture the overall trend well. Among the people in the sample, a one-unit increase in the interest correlation average is associated with a 3.364774% increase in rate of matches. For the simple linear regression, we found that the R^2value was approximately 0.0013931, which is very low. We performed multiple linear regression to see if the fit considerably improves.

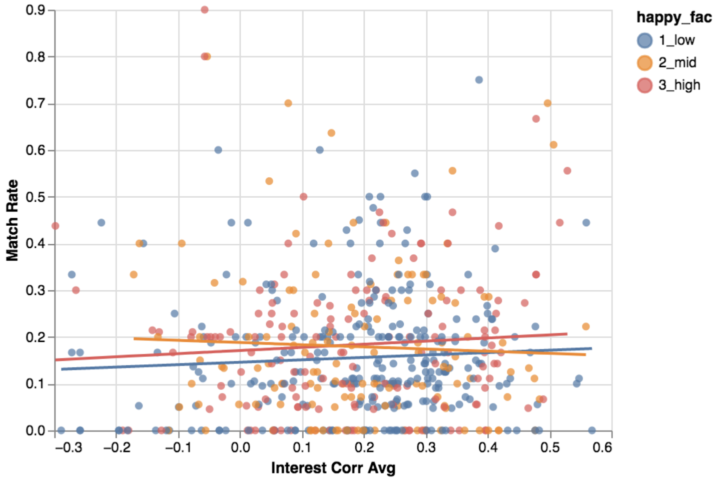


Figure 5: Multiple linear regression line plot of fitted values (y-axis) against interest correlation average (x-axis) with the exphappy factor mapped to the color aesthetic. In this plot, blue points indicate a high level of happiness expected from speed dating, yellow points indicate a mid-level of happiness expected from speed dating, and red points indicate a low level of happiness expected from speed dating.

In order to perform multiple linear regression, we chose to add the happy_fac (which is categorical) to the model. We created the happy_fac factor by generating three evenly-split levels based on the values of the exphappy variable. We found that the $R^2$for the multiple linear regression model is 0.0111217. Therefore, since the R^2 value for the multiple linear regression model is lower than the value for the simple linear regression model, we concluded that the fit did not considerably improve.

For the 1_low and 3_high happy_fac groups that participated in the study, the multiple linear regression model (Figure 5) suggests that there is a slightly positive correlation between interest correlation average and match rate; for the groups of participants that expected a low and high level of happiness from the speed dating event, as interest correlation average increases, the match rate slightly increases. However, for the 2_mid happy_fac groups that participated in the study, the multiple linear regression model suggests that there is a slightly negative correlation between interest correlation average and match rate; for the group of participants who expected a mid-level of happiness from the speed dating event, as interest correlation average increases, match rate slightly decreases. Therefore, we conclude that there is little to no correlation between interest correlation average and match rate based on the expected happiness level of the participants in relation to the people they meet during the speed dating event.

Regression analysis of match rate, interest correlation averages, and gender

We decided to perform simple linear regression and multiple linear regression on the gender of the speed dating participants.

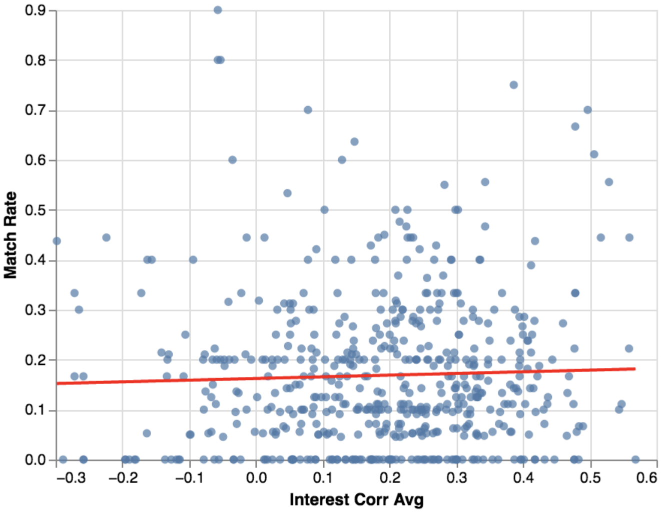


Figure 6: Simple Linear Regression line plot of fitted values against interest correlation average. The red line represents the fitted values. The match rate is on the y-axis and the interest correlation average is on the x-axis.

Figure 6 reveals that the fitted line does not go straight through all of the data points, and therefore, does not capture the overall trend well. Among the people in the sample, a one-unit increase in the interest correlation average is associated with a 3.349% increase in rate of matches. For the simple linear regression, we found that the R^2 value was low; it was approximately 0.00138. Since the simple linear regression R^2 suggests that the fit is not good, we performed multiple linear regression to see if the fit considerably improves.

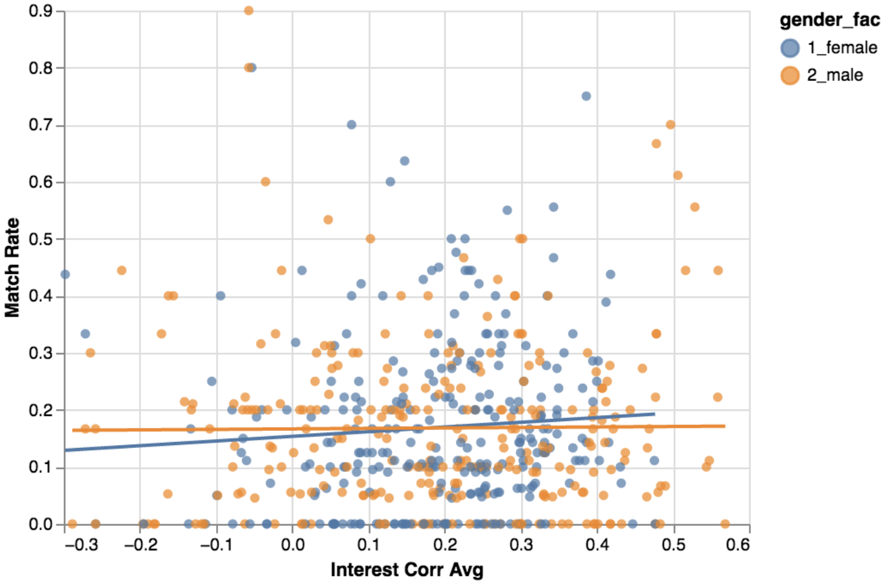
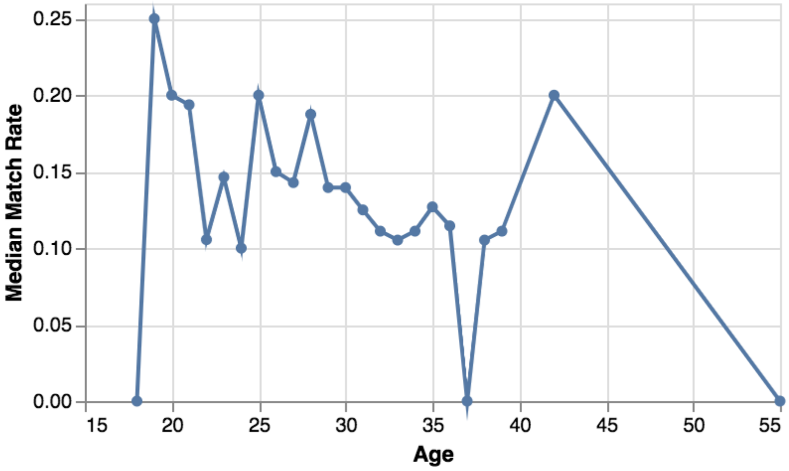


Figure 7: Multiple linear regression line plot of fitted values (y-axis) against interest correlation average (x-axis) with the gender factor mapped to the color aesthetic. In this plot, blue points represent the female speed dating participants and the yellow points represent the male speed dating participants.

We found that the R^2 for the multiple linear regression model is 0.00288. Therefore, since the R^2 value for the multiple linear regression model is slightly higher than the value for the simple linear regression model, the fit slightly improved. However, a R^2 value of 0.00288 is still very low, so there is no significant correlation between interest correlation average and match rate based on gender.

Figure 7 shows that for the female and male speed dating participants in this event, the correlation between interest correlation average and match rate is slightly positive. It is important to note that there is a stronger, positive correlation between interest correlation average and match rate for the female speed dating participants than the male speed dating participants. This suggests that, for the female speed dating participants in this sample, as interest correlation average increases, the match rate increases. However, for the male speed dating participants in this sample, as interest correlation average increases, the match rate slightly increases (but it is basically an insignificant amount).

Measure of median match rate by age



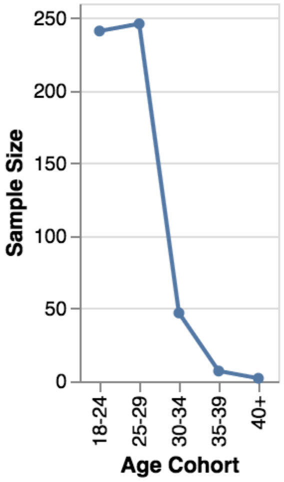
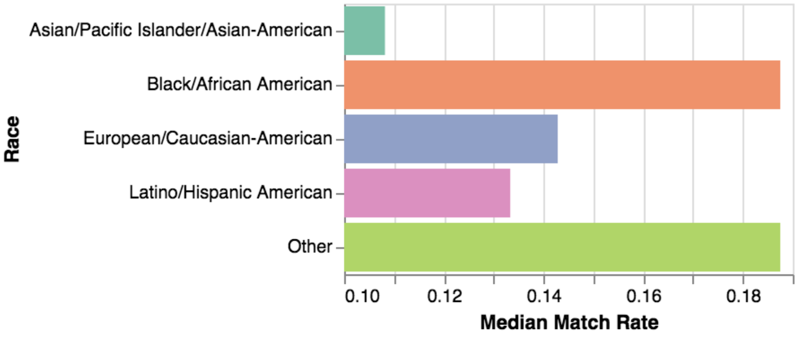


Figure 8a: The median match rates (y-axis) of each age of the participants (x-axis), ranging from 15 to the maximum age of 55.

Figure 8b: The sample sizes (y-axis) of the age cohorts of the participants (x-axis), which range from the minimum age of 18 to the oldest ages of 40+.

In looking at Figure 8a, we see a slight downwards trend in the median match rate as age increases, with the exception of an age at around 42. However, as we look at Figure 8b, we see that the sample sizes for age groups 30 years and higher are drastically lower than the age groups between 18 and 29. Therefore, while there is ample evidence for the slight downwards median match rate trend in age for 15-29, it is likely that the median match rate for the older age groups is biased because of the small sample sizes.

Measure of median match rate by race



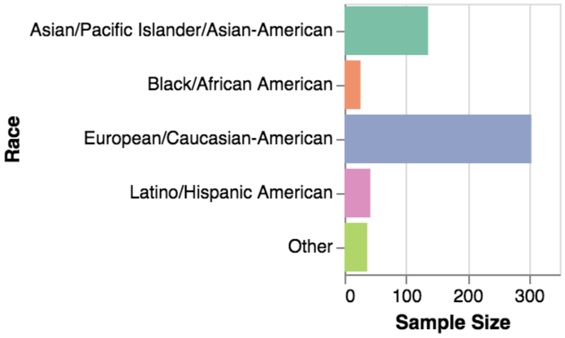


Figure 9a: The median match rates (y-axis) of each race group of the participants (x-axis).

Figure 9b: The sample sizes (y-axis) of the race groups of the participants (x-axis).

In looking at the medians of match rates by race, it seems that participants that identified as Black/African Americans or did not identify with a race group listed had the highest median match rates, whereas Asian/Pacific islander/Asian-American had the lowest median match rate. Although the distribution of participants across the races was not even, there were at least 25 participants in each group.

3. Discussion

In conclusion, our analysis of the speed dating dataset revealed that there is some positive correlation between the rate of mutual matches participants received and the expected number of matches a participant expected to receive. This is interesting because it suggests that participants who are confident in their ability to find a match were more likely to actually do so. Because we were unable to find high correlation between match rate and other variables, including expected level of happiness and gender, we decided to continue our investigation by creating line plots and bar plots between "Match Rate" and "Age" and "Median Match Rate" and "Race". In addition, participants who identified as Asian/Pacific Islander/Asian-American had significantly lower rates of Median Match Rate while participants who identified as Black/African American and Other had higher rates of Median Match Rate.

Without overspeculating, we believe that our data suggests that one's perceived attractiveness and likeability (which are qualities we believe participants may have taken into account when rating their expected number of matches) can play a role in the actual outcome in one's match rate. Of course, it would be naive to believe that this is the only variable that has a relatively higher level of correlation in regards to match rate that exists as there are infinite reasons as to why people choose the romantic partners that they do, but it is interesting to see one's own confidence show statistical evidence of higher match rates. If we were to continue this investigation, it may be interesting to investigate further about the effects of race on match rate. It may also be interesting to conduct analysis based on each individual meeting between participants, instead of focusing on the overall dating outcomes for participants as a whole.

---

4. Appendix

Initial Exploration

In order to familiarize ourselves with the Speed Dating Dataset, we created multiple plots to visualize the data, a few of which we have included here. We were most interested in seeing if there was an even spread of participants across various variables.

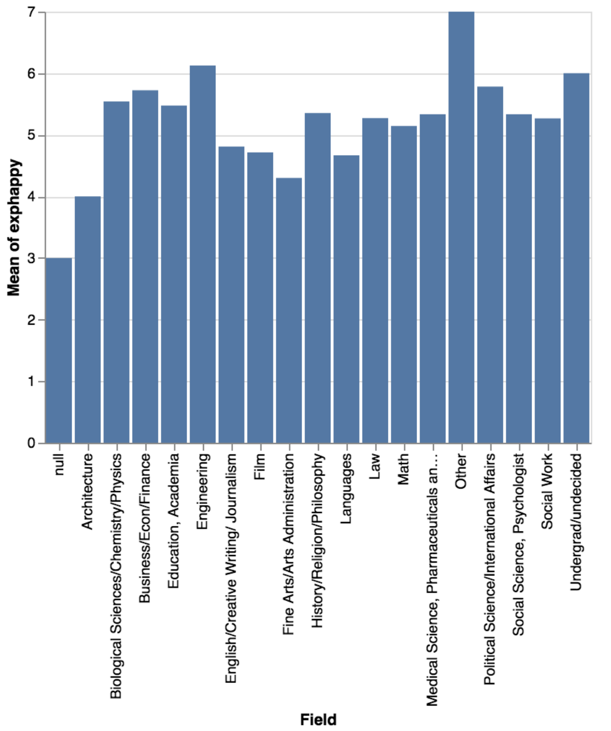


Figure 10: Bar Plot of Exhappy over Field (of participant's work). It appears that participants working in fields outside of the given options had the highest expectations of happiness from speed dating.

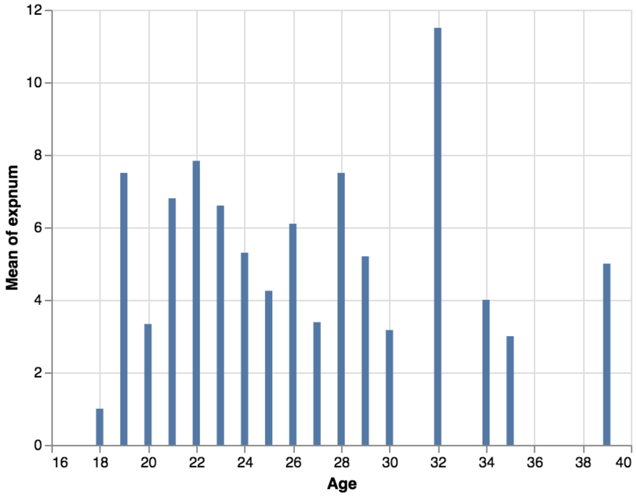


Figure 11: Bar Plot of Mean of Expected Number of Matches over Age. It appears that participants of the age 32 expected the most number of matches.

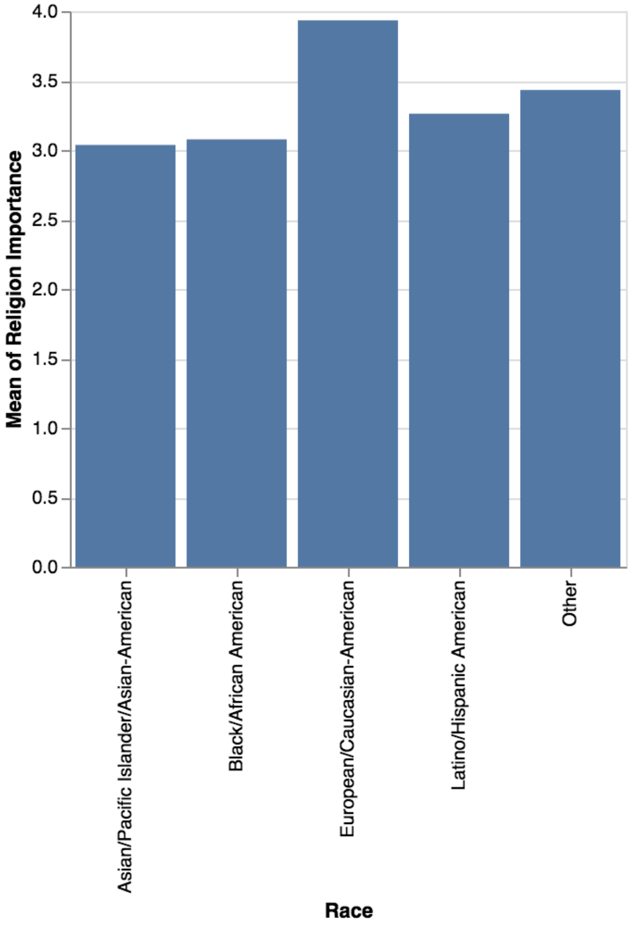


Figure 12: Bar Plot of Race and Mean of Religion Importance. It appears that European/Caucasian American participants as a group placed the highest level of importance on religion.

5. References

1. Fisman, R., Iyengar, S. S., Kamenica, E., & Simonson, I. (2006). Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment. *The Quarterly Journal of Economics*, *121*(2), 673–697. https://doi.org/10.1162/qjec.2006.121.2.673

   Link: https://academiccommons.columbia.edu/doi/10.7916/D8DR31Q9/download