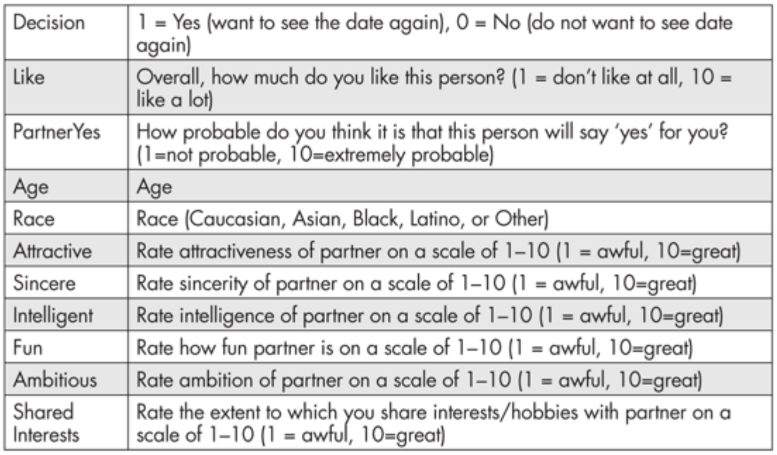
Homework 4: SDGB 7840

Instructor: Prof. Nagaraja

Due: 4/26 in class

In speed dating, participants meet many people, each for a few minutes, and then decide who they would like to see again. The data set you will be working with contains information on speed dating experiments conducted on graduate and professional students. Each person in the experiment met with 10-20 randomly selected people of the opposite sex (only heterosexual pairings) for four minutes. After each speed date, each participant filled out a questionnaire about the other person.

Your goal is to build a model to predict which pairs of daters want to meet each other again (i.e., have a second date). The list of variables are:



Decision	1 = Yes (want to see the date again), 0 = No (do not want to see date again)
Like	Overall, how much do you like this person? (1 = don't like at all, 10 = like a lot)
PartnerYes	How probable do you think it is that this person will say 'yes' for you? (1=not probable, 10=extremely probable)
Age	Age
Race	Race (Caucasian, Asian, Black, Latino, or Other)
Attractive	Rate attractiveness of partner on a scale of 1–10 (1 = awful, 10=great)
Sincere	Rate sincerity of partner on a scale of 1–10 (1 = awful, 10=great)
Intelligent	Rate intelligence of partner on a scale of 1–10 (1 = awful, 10=great)
Fun	Rate how fun partner is on a scale of 1–10 (1 = awful, 10=great)
Ambitious	Rate ambition of partner on a scale of 1–10 (1 = awful, 10=great)
Shared Interests	Rate the extent to which you share interests/hobbies with partner on a scale of 1–10 (1 = awful, 10=great)

We will be using a reduced version of this experimental data with 276 unique male-female date pairs. In the file "SpeedDating.csv", the variables have either "M" for male or "F" for female. For example, "LikeM" refers to the "Like" variable as answered by the male participant (about the female participant). Treat the rating scale variables (such as "PartnerYes", "Attractive", etc.) as numerical variables instead of categorical ones for your analysis.

1. Based on the variable “Decision”, fill out the contingency table below. What percentage of dates ended with both people wanting a second date?

		Decision made by female	
		No	Yes
Decision made by male	No		
	Yes		

2. A second date is planned only if both people within the matched pair want to see each other again. Make a new column in your data set and call it “second.date”. Values in this column should be 0 if there will be no second date, 1 if there will be a second date. Construct a scatterplot for each numerical variable where the male values are on the x-axis and the female values are on the y-axis. Observations in your scatterplot should have a different color (or `pch` value) based on whether or not there will be a second date. Describe what you see. (Note: Jitter your points just for making these plots.)

3. Many of the numerical variables are on rating scales from 1 to 10. Are the responses within these ranges? If not, what should we do these responses? Is there any missing data? If so, how many observations and for which variables?

4. What are the possible race categories in your data set? Is there any missing data? If so, how many observations and what should you do with them? Make a mosaic plot with female and male race. Describe what you see.

5. Use logistic regression to construct a model for “second.date” (i.e., “second.date” should be your response variable). Incorporate the discoveries and decisions you made in questions 2, 3, and 4. Explain the steps you used to determine the best model, include the summary output for your final model only, check your model assumptions, and evaluate your model by running the relevant hypothesis tests. Do not use “Decision” as an explanatory variable.

6. Redo question (1) using only the observations used to fit your final logistic regression model. What is your sample size? Does the number of explanatory variables in your model follow our rule of thumb? Justify your answer.

7. Interpret the slopes in your model. Which explanatory variables increase the probability of a second date? Which ones decrease it? Is this what you expected to find? Justify.

8. Construct an ROC curve and compute the AUC. Determine the best threshold for classifying observations (i.e., second date or no second date) based on the ROC curve. Justify your choice of threshold. For your chosen threshold, compute (a) accuracy, (b) sensitivity, and (c) specificity.