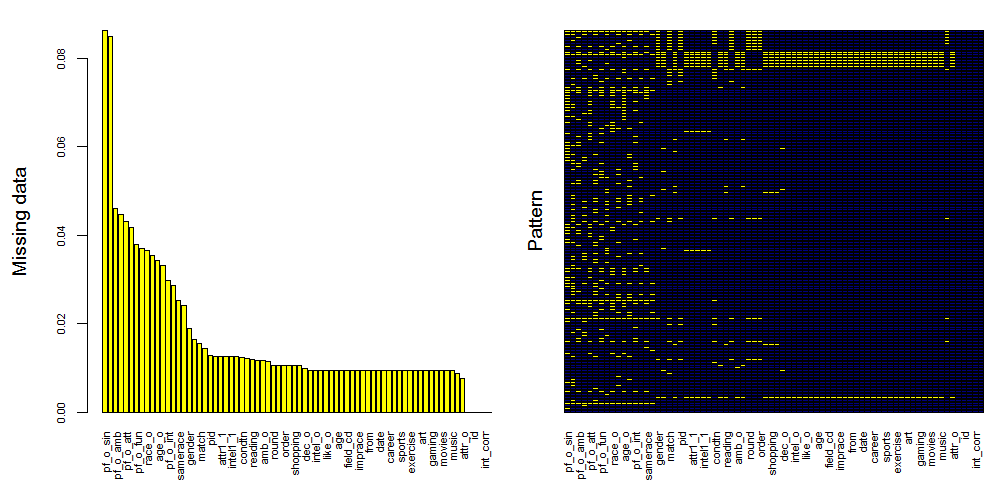
1. **Loading Data**

After loading my data into R, I converted all categorical variables into factors. I had to carefully read the info sheet to see the meaning behind each variable. To load in SAT scores, I had to remove commas form the string before converting to an integer.

1. **Cleaning Data**

Then, I started my cleaning process. First, I removed all categories with greater than about 10% NA’s. This reduced the number of columns I have by about half. Next, I performed imputation using the mice package. The mice package uses different methodologies depending on if the category is binary, categorical, or numerical. When imputing my data set, mice chose to use pmm imputation on the numerical data, and polyreg imputation on the categorical variables. This completely removed the NA’s.

Also, It mice assumes that the NA’s are randomly distributed throughout the data. However, from the mice\_plot I made, this doesn’t seem to be completely true. In some waves, collections of categories are missing. So we have two options with this: remove all rows with large missing values, or remove columns where missing values are clustered. I decided to go with the second option because those categories didn’t interest me as much for what I wanted to study.



In the right plot, notice the cluster of yellow in the right, indicating a nonrandom distribution of NA’s

1. **Adding new attributes**

My goal was to determine attributes that can predict a match, so I decided to add some new variables that I thought might be effective. First, I combined age\_o and age into age\_diff, which is a column representing the difference in age between the participant and the date. Next, I added a variable called “confidence” which was the sum of how highly the participant rated his/herself in each of the 5 attributes. Finally, I added an attribute called “similarity”. This variable is the distance in rated attribute values between the participants.

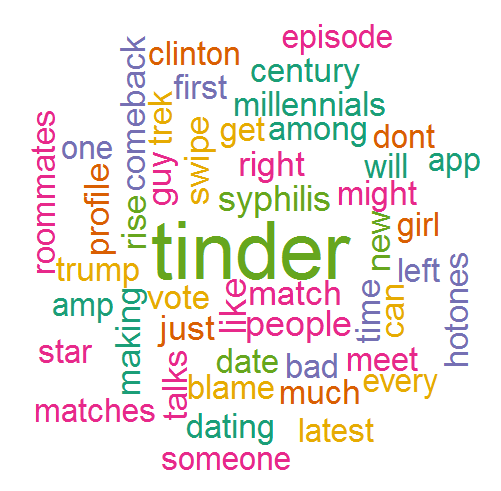
1. **Normalizing the data**

Weighting was a little tricky, because data wasn’t taken consistently between waves. Some attributes where rated on a scale of 1-10 for some waves, and in other waves participants had 100 total points to distribute between the collection of 5 attributes. To combine data collected using both methodologies, I found the sum of the points in the 5 categories for each row. Then, I divided the value in each category by the point sum I found. This way, the attributes are always between 0 and 1 even though the attributes were collected in different ways in certain waves. This is not perfect, but I thought it would be close enough for my purposes. Normalizing the other variables was easier, because I just had to loop over all other attributes, see if the attribute is numeric, and if so divide all values in that column by the maximum value of the column.

1. **Analysis**

I compared the feature importance for determining a match for both males and females and compared. I only included variables that I thought might influence the match, but didn’t include the variable “like” since that would be too strong to be interesting. I noticed that my similarity attribute was the most important of the attributes tested. I also noticed that a participant’s attractiveness was more important for females, while confidence was more important for males. More discussion will come from the slides.

I also performed some sentiment analysis from twitter based on the common dating app “tinder” specifically, I wanted to see the difference in how two different places (Los Angeles and New York) felt about the app. The results from the analysis will be in my slides as well.



AN example wordcloud using tweets from New York City