**Speed Dating Data Cleaning Log**

**Data Description:** Each row in the data contains information about the dater, their partner, and questions asked by the dater and their partner. Answers to questions up to those ending in “1\_1” (questions asked to the dater about their preferences) have been kept, and the rest have been kept.

* Dropped idg, condtn, wave, round, position, positin1 and partner – no use for these.
* Filtered out rows where pid was blank
* Imputed blank int\_corr rows with the mean value of the column
* Blank age\_o and race\_o values
  + Looked up missing values with the pid, iid, age and race fields
* Imputed blank age fields with the mean value
* Imputed blank race fields with the mode value
* Used xlookup to convert race and race\_o codes to text
* Imputed partner preferences (pf\_o\_x) with mean values
* Converted dec\_o to text
* Imputed missing partner perceptions (attr\_o, sinc\_o, intel\_o, fun\_o, amb\_o, and shar\_o) with mean values
* Dropped mn\_sat
  + Too many blanks tot impute
* Dropped undergrad
  + Too many blanks to impute
* Dropped tuition
  + Too many blanks to imputed
* Imputed imprace with mean
* Imputed imprelig with mean
* From column
  + Note: the survey question is unclear – I assume it refers to place of birth
    - Further
  + Removed non-geographic answers
    - Ex. “world citizen,” universities, employers
  + Filtered answers giving multiple locatioins
    - Ex. “CT, FL, TN”
  + Removed unclear answers
    - Ex. Berkeley, Siberia, Yugoslavia, Europe
  + Replaced values relevant country (Beijing becomes China, MA becomes United States, etc.)
    - Treated Hong Kong, Puerto Rico, and Taiwan as countries
  + Impute blanks with the mode, United States
* Zip Code
  + Added an unknown value for zip codes with more or less than five digits or countries outside of the US
* Dropped income – only two values after full cleaning
* Code values: goal, date, go\_out, career\_c, and field\_cd
  + Replaced integers with associated text values, imputing missing values with the mode
  + Dropped the provided string answers. Too much inconsistency.
  + Dropped the codes
* Interests (sports, tvsports, exercise, dining, museums, art, hiking, gaming, clubbing, reading, tv, theater, movies, concerts, music, shopping, and yoga)
  + Imputed missing values with mean
* Imputed expnum with mean
* Imputed exphappy with mean
* The features ending in 1\_1
  + Imputed values with mean
  + Note some looked imputed already – participants were given a budget of 100 points to spread across these questions, and certain responses are non-integer numbers like 15.56, 13.33 in the raw data. These are unusual but not dropped.
* Delete unwanted fields
* Create date\_id
  + Unique identifier for each date consisting of the concatenation of iid & pid