**PRESENTATION**

* Do it over-what would you do differently?
* Discuss restarting
* Misunderstood data – regression
* Data- thought it was clean and coded
  + WRONG
    - 8378 entries
    - 195 features
  + This took me the longest time – deciphering every feature b/c all were labeled the same just with a different reference # at the end to indicate when in the survey / experiment the question had been administered and/or who (subject or partner)
    - Had to sort though every feature to figure out which would be relevant to my question 🡪 issue there was I couldn’t figure out the approach I wanted to take on my question so instead of making up my mind, I decided to try and keep all the day (bad idea)
* Started by pulling everything I thought would be interesting; had a lot of questions about the data
* After some observations, decided I wanted to make a condensed dataset per iid
* Wrote functions to condense data
* NaN value issues
  + Had to refine my subsets; only use data to answer specific question

Further exploration – look at gender perceptions; how does that effect dating outcomes?

### Does one’s perception of their gender generalizations differ from their own evaluations of what’s important when it comes to selecting mates?

\*\*Question 2B\*\*: \*Does this differ from self-evaluations (Q3)? E.g.: do men rate ‘attractiveness’ as less important for their own dating choices but more important for other men?\*

\*\*Hypothesis\*\*: men will rate ‘attractiveness’ as less important for their own dating choices but more important for other men’s decisions when choosing a partner.

Other question:

Look at probability rating and score; e.g. if person gave high score but thought the prob of other person choosing them was low, how that affects decision (y/n); ultimately predicting match

NOTE:

for sums of # of matches, I put an idea in the answer above. but assigning the resulting value to a column such as df.num\_matches = \* should work. then you can do linear regression to predict how many matches they'd get, or if you normalize it 0-1 then you could also do logistic regression or even classification (where you'd be classifying if they got 'a lot' of matches vs not)

for viewing the data, you can always do .head(a) or .tail(b) where a and b are optional, but specify the # of rows returned - so the top a rows or the bottom b rows. there are also scatter plots of an independent variable vs the outcome variable, and histograms of any single variable. compared to describe(), I think histograms are probably what you're looking for. df.variable.plot(kind='hist')

exploratory analysis

data exploration and analysis (visuals)

scatter w/ 2 variables (color code)

MODELING

Random forests (determine which features are important)

* using those, use logistic regression

KNN

Accuracy tests

False positive vs. true negative

E.g. better to get false negative than a false positive: would rather think you aren't getting a match and then get one instead of thinking you are getting a match and then not get one