# Step 1: Remove all the unnecessary variables

The data set started out with 8,378 rows and 195 columns. This high number of columns (variables) concerned me. I knew that not all of these needed to be included, so I did some exploratory analysis and removed the unnecessary variables to get down to 49.

**Some of the variables were repetitive.** For example, the data asked the same questions multiple times throughout the experiment (before going on dates, halfway through going on dates, after the dates, etc.). After taking a brief look at these variables, I didn’t see much insight into asking these questions so many times, so I removed them.

**There were also some variables that seemed irrelevant.**  These included “position” of the dater and what “group” the dater was in, and these didn’t seem to have much an effect on the process. I removed these as well.

**Some variables were too varied to be able to gain insight from.** These included things like “field of study” and the “zip code” you grew up in. The sample size was too small for these, so they had to be removed.

**Finally, some of the variables had significant missing data.** If the variable was going to cause me to lose a large portion of my data, I removed it. This included many of the survey questions such as “what do you think the opposite sex is most interested in”. It also included a couple of the rating variables, such as “rate your ambition” and “rate the level of shared interests”.

# Step 2: Filter the rows

I began with 8,378 rows of data, but I filtered out data for a few reasons. I **removed any dates with null values**. I had already gotten rid of the attributes with significant missing data so that getting rid of nulls wouldn’t hurt me too much in terms of lost data.

I also **removed any dates that involved one person who was 55 years old**. This person was 13 years older than the next oldest person, and was an outlier in terms of age. Removing this person only lost about 6 dates.

I **removed any dates that involved a person who went on less than 5 dates**. The reason I did this was because I will go on to make new variables that are based on average ratings from the participants’ partners. Because of this, I wanted each person to have at least a sample size of 5 to get an accurate measure.

In the end I was left with 7,328 rows of data.

# Step 3: Sort the variables

Now that I had the variables I wanted, the next step was to sort them into groups. I separated each variable into one of the following groups:

* **IDs:** This consists of two variables for identification of the two participants: the primary and the partner.
* **Important Facts:** These were facts about both participants. Attributes here included things like: gender, age, and race.
* **Interests:** These are the primary’s ratings of 1-10 for a collection of interests including things like: sports, dining, clubbing, and art.
* **Preferences:** These are a collection of the primary’s preferences including things like: important of religion, how often do you go out, and how often do you date.

*The following 3 categories each have ratings on attributes of: Attractiveness, Sincerity, Intelligence, Fun, and Decision (yes or no to seeing the person again).*

* **Primary Desires:** Primary rates what they look for in the opposite sex
* **Primary on Partner:** How the primary rated their partner
* **Partner on Primary:** How the partner rated the primary
* **Match:** Did both participants decide “Yes” they want to see the other person again?