Chapter 4

Data Wrangling Case Study

4.1 Research Question

Romantic relationships provide an interesting realm for observing racial biases, which emerge because people can justify racial prejudices as preferences. A student researcher and her advisor decided to explore this by designing an experiment. The researchers recruited $n = 202$ participants via Amazon Turk. The recruited participants read online dating profiles from a factious dating website that were randomly assigned race information (i.e., Asian, Black, or White).

The researchers were interested in how the participants' interests changed when learning a profile belonged to a Black (but not Asian or White) person, and which groups of participants ranked race information (vs. other aspects of the profile) as most or least influential in increasing their dating interest.

Abundant data indicate that people's social networks are predominantly composed of people of their own race; e.g., Hofstra et al. (2017). Thus, one consequence of online dating is that people are able to meet partners, including those of other races, with whom they may not have otherwise interacted. However, as is the case with more traditional forms of dating, statistics from online dating sites indicate that people are more likely to express interest in people who are similar to them in demographic characteristics, especially race (Hitsch et al., 2010). It has also been observed that White people's dating preferences follow a particular hierarchy such that they are much less likely to date and marry Black individuals than people from other racial groups (Feliciano et al., 2009, 2011).

While these online dating statistics provide a glimpse into how people are behaving in natural dating contexts—they also lack experimental control. Thus, it is difficult to isolate race as the causal factor in these dating outcomes. For example, given the centrality of perceived similarity to both platonic and romantic relationships (Montoya et al., 2008), White people's preferences for same-race romantic partners may reflect attraction to other characteristics that co-vary with race. Thus, the purpose of the present research is to complement ecologically valid statistics from online dating sites with highly controlled experimental designs to isolate the influence of race on online dating interest. In particular, we propose that racial biases may be particularly persistent in the realm of romantic relationships because choosing a romantic partner is a very subjective decision—providing a fertile ground for biases to emerge (Dovidio and Gaertner, 2004). Thus, the researchers hypothesized that independent of profile content, White people would show overt racial biases in dating interest, especially toward Black daters.

While examining these aspects of racism in dating decisions, the researchers also tested an intervention to reduce the observed racial biases. Participants in the treatment group were showed graphs where there was a high percent of people matched on the factitious site are interracial couples – 51% same race, and 49% different race. Participants in the control group were showed graphs where there was a low percentage – 91% same race, and 9% different race. This idea was that participants in the treatment group were be led to believe interracial matches were normal on the factious dating website, and the control group were told the interracial matches on the factious website approximated that of the 2016 US Census.

4.2 Survey

4.2.1 Informed Consent

Overview and Procedure:
You are being asked to take part in a research study. To join the study is voluntary. You may refuse to join, or you may withdraw your consent to be in the study, for any reason, at any time, without penalty. Details about this study are discussed below. It is important for you to understand this information so that you can make an informed choice about being in this research study.

This study examines evaluations of online dating profiles. For that reason, only people who currently identify as "dating" or "single" will be eligible for this study. Your participation in this study will last approximately 10-15 minutes.

Risks and Benefits:
We anticipate few risks in this study. Beyond compensation, you may not benefit personally from being in this study.

Confidentiality:
By completing our study through Amazon Mechanical Turk (MTurk) you are subject to our own standards of confidentiality (the researchers posting this study) as well as those of Amazon.com.

Our standards of confidentiality: The only identifying information that we will collect will be your Mturk worker ID which is linked to your Amazon.com public profile. This ID will only be used to help us assign credit for study completion. Once data collection is complete and you have been paid for your participation we will delete all Mturk IDs from our data files. Thus, only completely de-identified data files will be shared among researchers once data collection is complete and you have been paid for your participation.

Amazon.com standards of confidentiality: Amazon.com has stated that the MTurk platform is NOT meant to support participant anonymity. MTurk worker IDs are linked to Amazon.com public profiles. Amazon.com may disclose information about you as an Mturk worker. Additionally, worker information may be available to others (who submit a request) for tax reporting purposes.

Compensation:
You will receive $0.70 for participating in this study. There are no costs associated with being in the study.

Your Rights:
To join the study is voluntary. You may refuse to join. If you choose to join, you may also withdraw from this study at any time and without penalty.

Contact Information:
All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have questions or concerns, or if you would like to obtain information or offer input, please contact the Institutional Review Board. For more information about the study, please contact the faculty advisor .

4.2.2 Acceptance Criterion

- o I am at least 18 years old and agree to participate in this study.

What is your relationship status?

- o Single
- o Dating
- o In a committed, monogamous relationship

Are you currently interested in dating?

- o Yes
- o Maybe
- o No

Remark: Participants who were in a committed, monogamous relationship or not interested in dating were not welcome to complete the rest of the study. The reason for this is because these participants aren't representative of the population of interest – those people looking for a date.

4.2.3 Welcome

Welcome to the study!

First, we are interested in whether you actually take the time to read instructions. To show that you read the instructions, please ignore the question below about how you are feeling and instead click only the "none of the above" option as your answer. Thank you very much.

Please click the word that best describes how you are currently feeling.

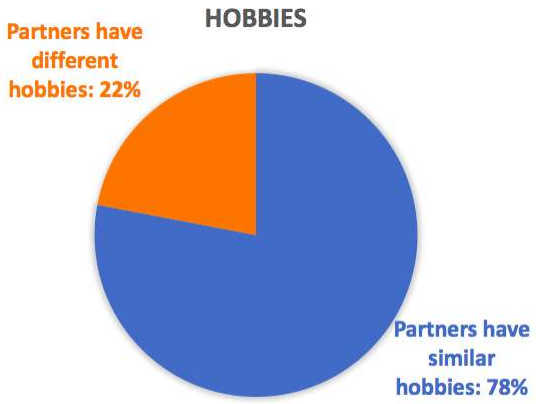
Remark: Questions like these are important for catching bots, or automated survey takers.

This study is being conducted by a team interested in improving the experience of people on a particular dating website (that will remain unnamed for privacy reasons). This dating website is trying to create an algorithm to best match people into successful relationships. To provide some context about the site, on the following pages we will select some information for you to view about those who have used our site to successfully make a romantic match.

On the following pages you will see data (represented in charts) about people who have successfully matched on our dating website. Please examine this data carefully and answer the questions that follow.

4.2.4 Treatment

All participants were presented with the following information about hobbies and and education level.

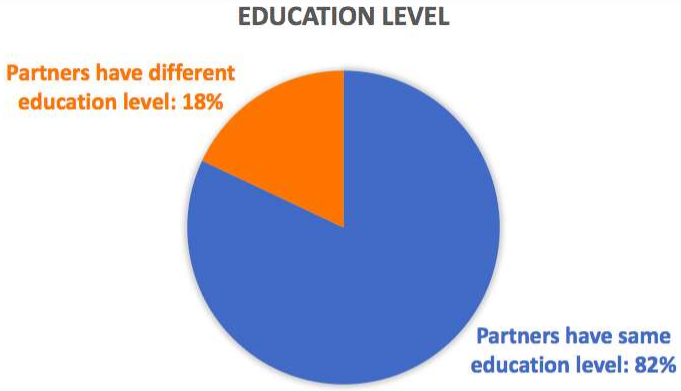


According to the graph above, what percent of people matched on our site have similar hobbies?

22%

78%

5%



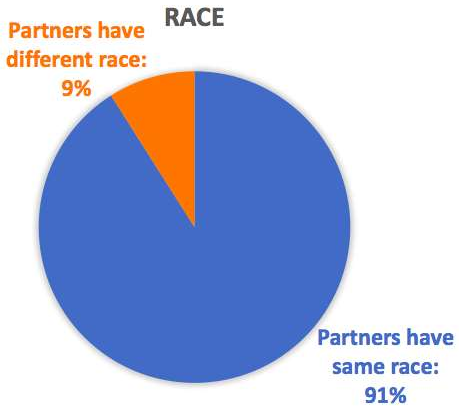
According to the graph above, what percent of people matched on our site have similar education levels?

82%

41%

18%

Participants in the control group were presented with the following information about race.



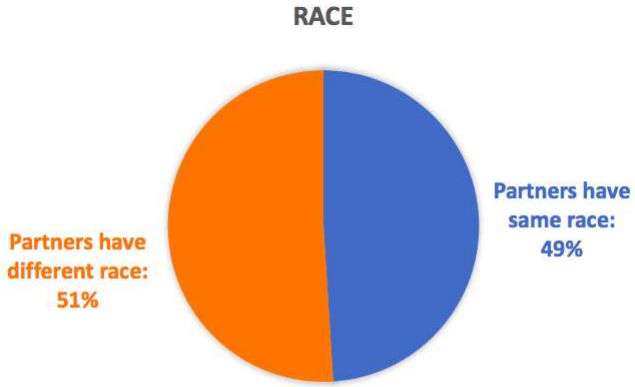
According to the graph above, what percent of people matched on our site are same-race couples?

91%

30%

2%

Participants in the treatment group were presented with the following information about race.



According to the graph above, what percent of people matched on our site are mixed-race couples?

10%

51%

5%

4.2.5 Deciding Preferences

Dating websites and applications vary in how much information is included in people's profiles. Some applications (e.g., Tinder, Bumble) provide very little profile information.

Other websites (e.g., Match.com, eharmoney) include profiles that are much more detailed.

As a part of this research, we are interested in how much information is helpful to users when evaluating potential partners.

On the next several pages you will see real profiles of users of our website (photos omitted). First we will provide you with a version of their profile that is limited in information.

We will then ask you to report your reactions to the profile as honestly as possible, as if you were using this website and considering a date with the person.

Next we will show you the user's FULL profile (which includes a bit more information, but still omitting the photo to protect their identities).

We will again ask you to report your reactions to that profile.

Our hope with this research is to improve our dating website to improve user satisfaction and help people meet long-term matches.

We are particularly interested in how much information is desirable when reviewing potential mates before meeting in person.

First, please report your dating preference below so we can choose profiles from our database that match your preferences. While we know that some people may not care about the gender of a potential partner, for the purposes of this study, please indicate the type of profiles that you would like to examine.

First, please report your dating preference below so we can choose profiles from our database that match your preferences. While we know that some people may not care about the gender of a potential partner, for the purposes of this study, please indicate the type of profiles that you would like to examine.

- o Interested in viewing women's profiles (i.e., interest in dating women)
- o Interested in viewing men's profiles (i.e., interest in dating men)

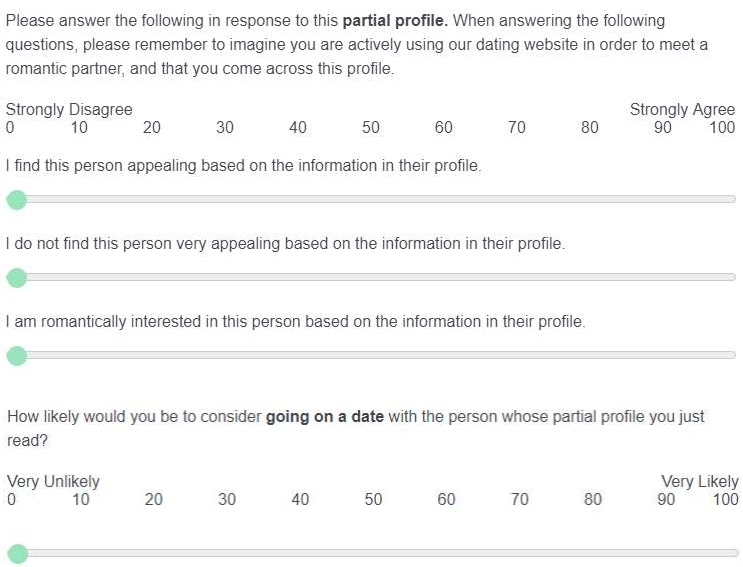
On the next page, you will see an **abbreviated** user profile. Please read the profile carefully, imagine you are using the site to find a potential match, and report your evaluations of the profile.

4.2.6 Measuring Interest in Profile

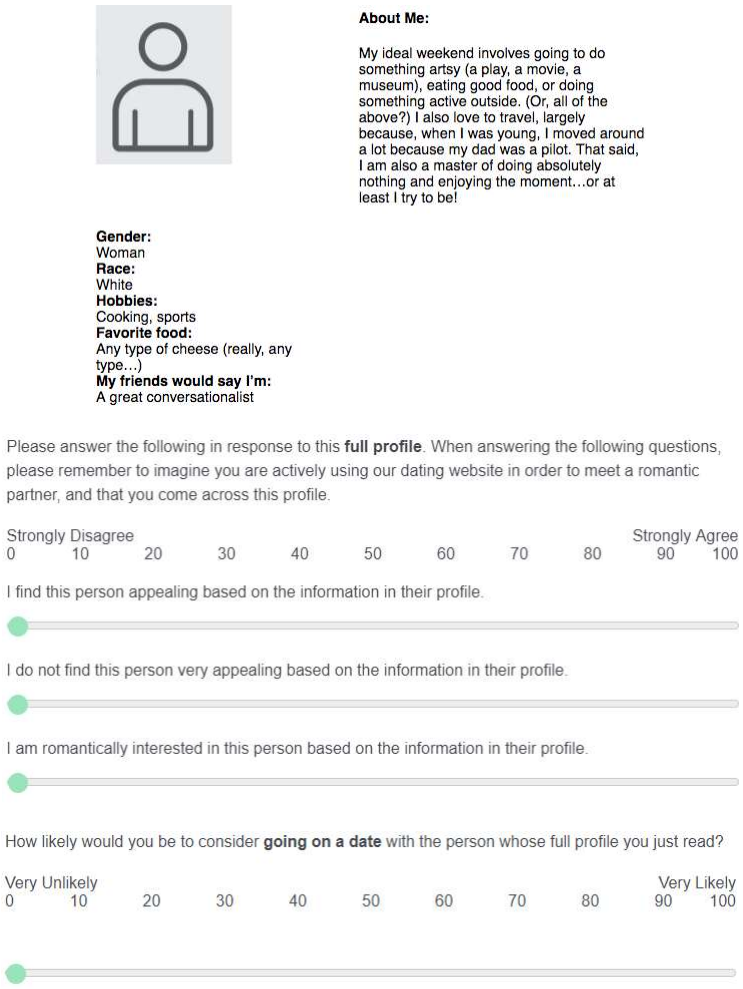
Remark: In this subsection, we provide the flow for evaluating one profile. Each participant evaluates three profiles, one of each race – Asian, Black, and White.

About Me:

My ideal weekend involves going to do something artsy (a play, a movie, a museum), eating good food, or doing something active outside. (Or, all of the above?) I also love to travel, largely because, when I was young, I moved around a lot because my dad was a pilot. That said, I am also a master of doing absolutely nothing and enjoying the moment…or at least I try to be!



On the next page, you will see the **full user profile** (rather than the abbreviated version). Please read the profile carefully, imagine you are using the site to find a potential match, and report your evaluations of the profile.



Please rate the information on the profile in whether you personally felt it was a "Turn On," "Turn Off," or "Neutral/Neither"

Items	Turn On	Turn Off	Neutral/Neither
My ideal weekend involves going to do something artsy (a play, a movie, a museum), eating good food, or doing something active outside.			
(Or, all of the above?)			
I also love to travel, largely because, when I was young, I moved around a lot because my dad was a pilot.			
That said, I am also a master of doing absolutely nothing and enjoying the moment or at least I try to be!			
I'm also a huge fan of watching Netflix shows, trying new foods, and spending time with friends and family.			
Hobbies: Cooking, sports			
Favorite food: Any type of cheese (really, any type)			
Gender: Woman			
Race: White			
My friends would say I'm: A great conversationalist			

Finally, please rank the following components of the profile based on the degree to which they **increased** how attracted you were to the profile from 1 (*most influential at increasing my attraction*) to 6 (*least influential at increasing my attraction*).

About me

Hobbies

Favorite Food

Race

Gender

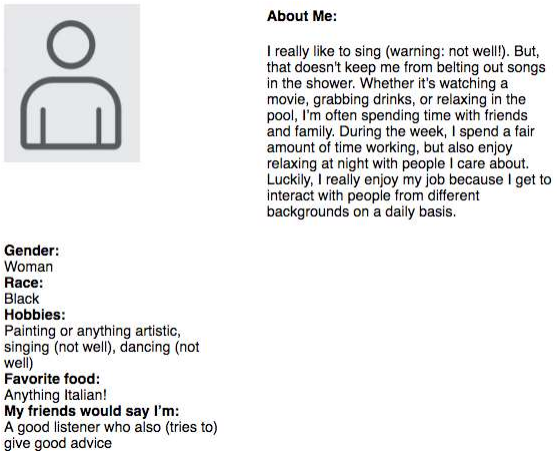
What their friends would say about them

Remark: This flow is repeated twice more. This first profile was for a White woman. The two additional profiles will be for a Asian woman and a Black woman – Qualtrics, the survey software will randomize the order. For reference, the two other profiles, without corresponding questions, for this survey are below.

About Me:

You know that person in your group of friends who is always planning something? Well, that's me. I love planning fun events either for my friends, family, or myself! My perfect weekend would include a hike, followed by a visit to a new local brewery or restaurant. I'm also an avid pie baker and swimmer, but have yet to master doing both at the same time. But you can watch me try! Just kidding, that actually sounds like a bad idea.

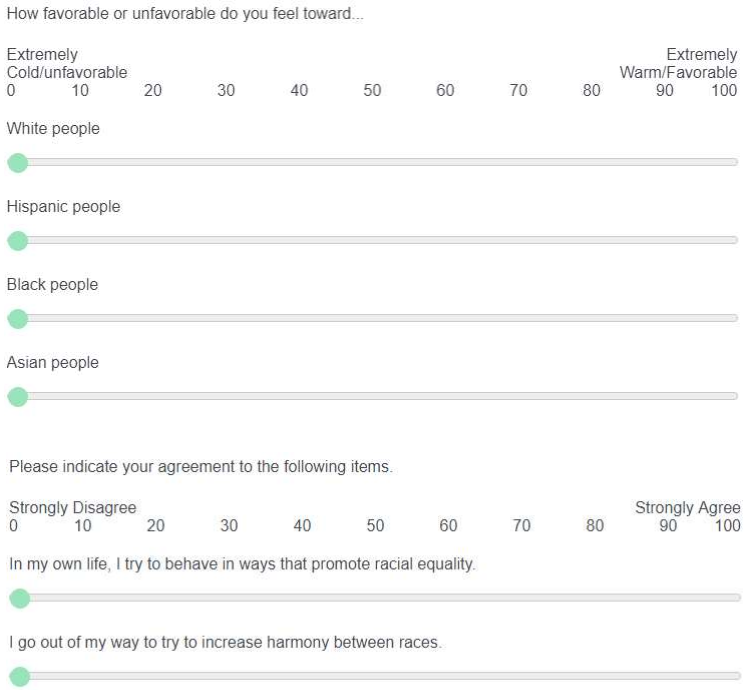
Gender:
Woman
Race:
Asian
Hobbies:
Baking, swimming, and watching comedies
Favorite food:
Anything I haven't tried before—I love new things
My friends would say I'm:
Down-to-earth and sarcastic

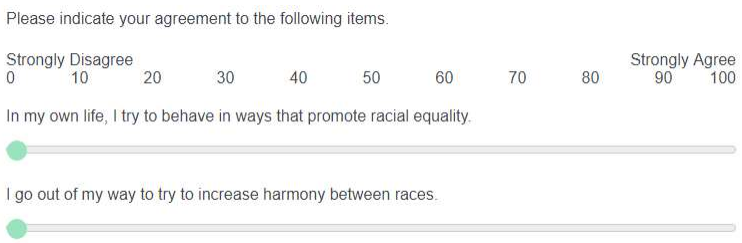


Remark: As participants access the Qualtrics survey, the survey flow is automatically generated. There are five base profiles from which three are selected at random. The race is then randomly assigned to the three selected profiles. This is done to ensure that any differences we observe aren't due to how the profiles are written.

4.2.7 Participant Views on Relationships

Now, we'd like to get your feelings toward certain groups. Below are names of groups. Using the scale below, we'd like you to rate how warm or cold you feel toward each group. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the group and that you don't care too much for that group. You would rate the group at the 50 degree mark if you don't feel particularly warm or cold toward the group.





4.2.8 Participant Views on Race

The study is almost complete. We will now ask you a few more questions about yourself.

The following questions concern various reasons or motivations people might have for trying to respond in nonprejudiced ways toward **Asian** people. Some of the reasons reflect internal–personal motivations, whereas others reflect more external–social motivations. Of course, people may be motivated for both internal and external reasons; we want to emphasize that neither type of motivation is by definition better than the other. In addition, we want to be clear that we are not evaluating you or your individual responses. All of your responses will be completely confidential. We are simply trying to get an idea of the types of motivations that people in general have for responding in nonprejudiced ways. It is important that you respond to each of the questions openly and honestly. Please click on the response that best represents your motivations. There are 5 items total.

I am personally motivated by my beliefs to be non-prejudiced toward Asian people.

Strongly Disagree 0 10 20 30 40 50 60 70 80 90 100 Strongly Agree

Because of my personal values, I believe that using stereotypes about Asian people is wrong.

Strongly Disagree 0 10 20 30 40 50 60 70 80 90 100 Strongly Agree

Being non-prejudiced toward Asian people is important to my self-concept.

Strongly Disagree 0 10 20 30 40 50 60 70 80 90 100 Strongly Agree

Because of today's PC (politically correct) standards I try to appear no prejudiced toward Asian people.

Strongly Disagree 0 10 20 30 40 50 60 70 80 90 100 Strongly Agree

I try to hide any negative thoughts about Asian people in order to avoid negative reactions from others.

Strongly Disagree 0 10 20 30 40 50 60 70 80 90 100 Strongly Agree

If I acted prejudiced toward Asian people, I would be concerned that others would be angry with me.

Strongly Disagree 0 10 20 30 40 50 60 70 80 90 100 Strongly Agree

I attempt to appear nonprejudiced toward Asian people in order to avoid disapproval from others.

Strongly Disagree 0 10 20 30 40 50 60 70 80 90 100 Strongly Agree

I try to act nonpredjudiced toward Asian people because of pressure from others.

Strongly Disagree 0 10 20 30 40 50 60 70 80 90 100 Strongly Agree

Remark: The questions in this subsection are then asked, verbatim, replacing 'Asian' with 'Black.'

4.2.9 Participant Demographics

The study is almost complete. On the following pages we will ask you several final questions.

How old are you

What is your gender?

- o Man
- o Woman
- o Other

What is your ethnicity?

- o Hispanic or Latinx
- o Not Hispanic or Latinx

What is your race?

- o White or Caucasian
- o Black or African American
- o Native American or Pacific Islander
- o Asian
- o Other

Which best describes your sexual orientation?

- o Gay/Lesbian
- o Heterosexual
- o Bisexual
- o Asexual
- o Pansexual
- o Other

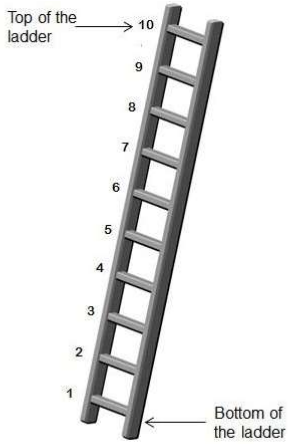
Think of this ladder as representing where people stand in the United States.

At the top of the ladder (rung 10) are the people who are the best off – those who have the most money, the most education and the most respected jobs.

At the bottom of the ladder (rung 1) are the people who are the worst off – those who have the least money, least education, and the least respected jobs or no job.

The higher up you are on this ladder, the closer you are to the people at the very top; the lower you are, the closer you are to people at the very bottom.

Where would you place yourself on this ladder? Click the number of the rung where you think you stand at this time in your life, relative to other people in the United States.



1 Bottom of the ladder	2	3	4	5	6	7	8	9	10 Top of the ladder

What is your yearly household income?

- o Less than $5,000
- o $5,001-$6,999
- o $7,000-$7,499
- o $7,500-$9,999
- o $10,000-$12,499
- o $15,000-$14,999
- o $20,000-$24,999
- o $25,000-$29,999
- o $30,000-$34,999
- o $35,000-$39,999

- o $40,000-$49,999
- o $50,000-$59,999
- o $60,000-$74,999
- o $75,000-$84,999
- o $85,000-$99,999
- o $100,000-$124,999
- o $125,000-$149,999
- o $150,000-$174,999
- o $175,000 or more

What is your highest level of education?

- o Less than a high school degree
- o High school degree
- o Some college (no degree)
- o 2-year college degree
- o 4-year college degree
- o Masters level degree (for example M.S. or M.A)
- o Doctorate level degree (for example PhD, MD, JDS)

Did you take this study on a smartphone, computer, or tablet?

- o smartphone
- o computer
- o tablet
- o none of the above

What do you think the purpose of the current study is?

Did you see any connections between the data about matches made on our website and the profile ratings we asked you to make later in the study?

- o Yes

o No

What did you think was the connection between the data about matches made on our website and the profile ratings we asked you to make later in the study? Please describe in detail, as this will help your data be more meaningful. Thank you.

4.2.10 Debrief

On the next page you will view more information about the purpose of the study.
Your completion code is XXXX.
DEBRIEFING FORM

Data from dating websites indicates that White people tend to prefer dating other White people (Mendelsohn et al., 2014). In the present study, we try to better understand these race-based preferences and also test a potential intervention to reduce these racially-biased dating selections.

If you were in the first sample, you made ratings of dating profiles. In particular, you viewed dating profiles with and without race information included. Some of these dating profiles were described as written by White people, some by Black people, and some by Asian people. These were not real profiles, but instead were created to seem like real profiles. We were interested in how preferences for the profiles would change with and without race information.

If you were in the second sample, you completed the same ratings of the dating profiles. However, before you made these ratings you also read an article about race being biologically determined and fixed, an article about race being fluid and changeable, or a scientific article unrelated to race. These were not real scientific articles. Instead they were intended to shift the way you might think about race. We hypothesized that learning that race is flexible might lead people to be more open to dating people of a different race than their own. For a recent, real scientific article that discusses the genetic basis of race see Yudell et al. (2016).

Our hope with this research is to better understand how race, and racial attitudes, influence the expression of dating preferences.

If you would like to talk with a mental health professional, http://blahtherapy.com/ provides low-cost option for online therapy.

Should you have questions regarding your rights as a participant in research, please contact the Institutional Review Board. Should you have questions about the study, please feel free to contact the researchers working on this project.

4.3 Spreadsheet

The participants see all the questions from Informed Consent, Acceptance Criterion, Welcome, a Treatment, Deciding Preferences, Measuring Interest in Profile (randomized survey questions),

Survey Number	Gender of Interest	Asian Profile	Black Profile	White Profile
1	Female	1	5	4
2	Female	2	1	5
3	Female	3	2	1
4	Female	4	3	2
5	Female	5	4	3
6	Male	1	5	4
7	Male	2	1	5
8	Male	3	2	1
9	Male	4	3	2
10	Male	5	4	3

Table 4.3.1: Randomization of survey flow yields ten possible profiles.

Participant Views on Relationships, Participant Views on Race, Participant Demographics and Debrief.

There are ten surveys that the participants can see through randomization of the five factitious dating profiles to the three races across gender. These surveys are described in Table 4.3.1.

For all sections but Measuring Interest in Profile, all participants see and most answer those questions, so we have observations for most participants in the columns holding those data. Where it gets tricky is that the data for each survey flow got it's own set of columns, because of the way it was coded in Qualtrics. This means that for a given survey there is a set of columns and only participants who were randomized to that survey have observations.

We want to combine these data into rectangular form – meaning, the measurements from each survey are combined into columns of all observations. This takes considerable effort in reading the spreadsheet. This can be done in Microsoft Excel, Google Sheets, LibreOffice Calc, or even RStudio.

4.4 Wrangling the Data in R

4.4.1 Reading the Data in R

To start, we create a data frame with all of the common variables. Below, we name the variables to correspond with their measurement so they'll be easier to ask for later, instead of simply taking the column numbers of interest.

```
> dat.csv<-read.csv("https://cipolli.com/students/data/datingData.csv",header=TRUE,
+                   stringsAsFactors=FALSE)[-1,] #removes questions in second row
> dat.clean<-with(data.frame(#Columns A-J
+                 sub=Subject, IP=V6, Status=V7, StartDate=V8, EndDate=V9,
+                 turkID=turkerId, rstatus=relationshipstatus,
+                 int.dating= interest_dating, attention.check=Att_Check,
+                 treatment=Cond1HS2LS,
+                 #Column Q
+                 int.gender=Pref_1W2M,
```

```
+                 #Columns ASK-AUS
+                 imp.relig=Q65_1, imp.polit=Q65_2, imp.race=Q65_4,
+                 imp.hobbies=Q65_5, most.relig=Q69_1, long.relig=Q69_2,
+                 most.polit=Q69_3, long.polit=Q69_4, most.race=Q69_5,
+                 long.race=Q69_6,most.hobbies=Q69_7,long.hobbies=Q69_8,
+                 FT.white=FT1_1, FT.hispanic=FT1_2, FT.black=FT1_3, FT.asian=FT1_4,
+                 Egal.equity=EgalMotivation_1, Egal.harmony=EgalMotivation_2,
+                 black.non_pred_import=IMS1_1, black.stereotypes_values_ok=IMS2R_1,
+                 black.non_pred_belief=IMS3_1, black.stereotypes_values_wrong=IMS4_1,
+                 black.non_pred_concept=IMS5_1,black.PC=EMS1_1, black.neg_react=EMS2_1,
+                 black.anger=EMS3_1, black.disapprove=EMS4_1,black.pressure=EMS5_1,
+                 asian.non_pred_import=Q698_1, asian.stereotypes_values_ok=Q699_1,
+                 asian.non_pred_belief=Q700_1,asian.stereotypes_values_wrong=Q701_1,
+                 asian.non_pred_concept=Q702_1, asian.PC=Q703_1,asian.neg_react=Q704_1,
+                 asian.anger=Q705_1, asian.disapprove=Q706_1, asian.pressure=Q707_1,
+                 age=age, gender=gender1m, hispanic=hisp, white=race_1, black=race_2,
+                 native.pacific=race_3, asian=race_4, other.race=race_5,
+                 other.race.text=Q399, sexual.orientation=sexo,
+                 sexual.orientation.other.text=other, ladder.self=Q76,
+                 household.income=Q78, highest.edu=Q80, device=Q393,
+                 purpose=Q377, connections.bool=Q379, connections.text=Q381),
+                 data=dat.csv)
```

Remark: We used the `stringsAsFactors=FALSE` so that `R` wouldn't automatically treat columns with characters as factors. The reason we did this is that the first row contains the question text for the data in that column and thus, each column would be treated as a character vector due to automatic cohesion and then as a factor which is undesirable.

Next, we merge the the data about the conditions. Note that these questions were the same, we just varied the graphic for mixed-race matches – now we treat them as the same question. Later, we can ask questions about the control and treatment groups by taking a subset of the data based on the value of `Cond1HS2LS`.

```
> dat.clean$graph.hobbies<-ifelse(dat.csv$Cond1HS2LS==1,dat.csv$simhob,dat.csv$simhob2)
> dat.clean$graph.educ<-ifelse(dat.csv$Cond1HS2LS==1,dat.csv$simed,dat.csv$simed2)
> dat.clean$graph.int<-ifelse(dat.csv$Cond1HS2LS==1,dat.csv$mixedrace,dat.csv$samerace)
```

While this data frame now contains a lot of information, it doesn't contain information key to assessing the research question – namely, the interest in partial and full profiles. This is where the data wrangling gets tricky, making it a perfect case study.

4.4.2 Processing the Survey Flows

The five profiles used to create the ten surveys are below, all specified as an Asian woman. The aspect that makes this so complicated is that the profiles are of different lengths, which changes the number of profile pieces that can be dragged into the 'turn on,' 'turn off,' and 'neutral/neither' boxes. This means that the number of columns differ from survey to survey, so a prescribed column number won't work here either.

About Me:

I was born and raised in Texas, so I absolutely love warm weather. I love going swimming, hanging by the beach, and just about anything that involves water. I have a job that I really enjoy, but I also savor time when I can relax. I'm a pretty good chef and I love cooking and baking for friends and family. My biggest fear is that my oven will break. I'm only half kidding.

Gender:
Woman
Race:
Asian
Hobbies:
Swimming, watching live sports
Favorite food:
Grilled cheese and tomato soup
My friends would say I'm:
A good listener and a very proud new aunt

About Me:

I would describe myself as fun-loving and a big fan of dogs. I have a young black lab who I love going on runs with. Sometimes I bring my dog to work with me, which luckily I am able to do in my job! (Side note: I think it scores me points with my boss, who is also crazy about dogs.) I'm also a huge fan of watching Netflix shows, trying new foods, and spending time with friends and family.

Gender:
Woman
Race:
Asian
Hobbies:
Running, reading mystery books, and watching movies
Favorite food:
Gourmet pizza
My friends would say I'm:
An easy-going person who laughs harder at my own jokes than other people do…

About Me:

My ideal weekend involves going to do something artsy (a play, a movie, a museum), eating good food, or doing something active outside. (Or, all of the above?) I also love to travel, largely because, when I was young, I moved around a lot because my dad was a pilot. That said, I am also a master of doing absolutely nothing and enjoying the moment…or at least I try to be!

Gender:
Woman
Race:
Asian
Hobbies:
Cooking, sports
Favorite food:
Any type of cheese (really, any type…)
My friends would say I'm:
A great conversationalist

About Me:

I really like to sing (warning: not well!). But, that doesn't keep me from belting out songs in the shower. Whether it's watching a movie, grabbing drinks, or relaxing in the pool, I'm often spending time with friends and family. During the week, I spend a fair amount of time working, but also enjoy relaxing at night with people I care about. Luckily, I really enjoy my job because I get to interact with people from different backgrounds on a daily basis.

Gender:
Woman
Race:
Asian
Hobbies:
Painting or anything artistic, singing (not well), dancing (not well)
Favorite food:
Anything Italian!
My friends would say I'm:
A good listener who also (tries to) give good advice

About Me:

You know that person in your group of friends who is always planning something? Well, that's me. I love planning fun events either for my friends, family, or myself! My perfect weekend would include a hike, followed by a visit to a new local brewery or restaurant. I'm also an avid pie baker and swimmer, but have yet to master doing both at the same time. But you can watch me try! Just kidding, that actually sounds like a bad idea.

Gender:
Woman
Race:
Asian
Hobbies:
Baking, swimming, and watching comedies
Favorite food:
Anything I haven't tried before—I love new things
My friends would say I'm:
Down-to-earth and sarcastic

Since all of the data were stored in separate columns, we need to merge them just as we did for the control and treatment data. This is more challenging because there are more and varying amount of columns per survey and there are ten groups. The first step is to create a variable called survey which specifies the survey number; this took some hunting in the spreadsheet.

```
> dat.clean$survey=rep(NA ,nrow(dat.clean))
> dat.clean$survey[which(dat.csv$Q359==1)]=1
> dat.clean$survey[which(dat.csv$Q426==1)]=2
> dat.clean$survey[which(dat.csv$Q492==1)]=3
> dat.clean$survey[which(dat.csv$Q558==1)]=4
> dat.clean$survey[which(dat.csv$Q624==1)]=5
> dat.clean$survey[which(dat.csv$Q82==1)]=6
> dat.clean$survey[which(dat.csv$Q415==1)]=7
> dat.clean$survey[which(dat.csv$Q481==1)]=8
> dat.clean$survey[which(dat.csv$Q547==1)]=9
> dat.clean$survey[which(dat.csv$Q613==1)]=10
```

There are questions in the profile evaluation section – 18 per profile – ignoring, the 'turn on,' 'turn off,' and 'neutral/neither' that don't refer to race or gender. The reason for this decision is that

there are a differing number of items and so accounting for this in a data frame would be difficult. We could, instead, have used a list but the questions not having to race or gender are not important to the hypotheses of interest.

To start, we create empty columns for all of the questions of interest.

```
######################
###Asian Profile
######################
> dat.clean$a.p.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$a.p.not.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$a.p.romantic.int<-rep(NA ,nrow(dat.clean))
> dat.clean$a.p.date<-rep(NA ,nrow(dat.clean))
##################################################
> dat.clean$a.f.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$a.f.not.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$a.f.romantic.int<-rep(NA ,nrow(dat.clean))
> dat.clean$a.f.date<-rep(NA ,nrow(dat.clean))
##################################################
###Ignore turn ons/offs outside of study (there are also varying numbers of these)
> dat.clean$a.onoff.gender.group<-rep(NA ,nrow(dat.clean))
> dat.clean$a.onoff.gender.rank<-rep(NA ,nrow(dat.clean))
> dat.clean$a.onoff.race.group<-rep(NA ,nrow(dat.clean))
> dat.clean$a.onoff.race.rank<-rep(NA ,nrow(dat.clean))
##################################################
> dat.clean$a.rank.about<-rep(NA ,nrow(dat.clean))
> dat.clean$a.rank.hobbie<-rep(NA ,nrow(dat.clean))
> dat.clean$a.rank.food<-rep(NA ,nrow(dat.clean))
> dat.clean$a.rank.friends<-rep(NA ,nrow(dat.clean))
> dat.clean$a.rank.race<-rep(NA ,nrow(dat.clean))
> dat.clean$a.rank.gender<-rep(NA ,nrow(dat.clean))
######################
###Black Profile
######################
> dat.clean$b.p.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$b.p.not.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$b.p.romantic.int<-rep(NA ,nrow(dat.clean))
> dat.clean$b.p.date<-rep(NA ,nrow(dat.clean))
##################################################
> dat.clean$b.f.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$b.f.not.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$b.f.romantic.int<-rep(NA ,nrow(dat.clean))
> dat.clean$b.f.date<-rep(NA ,nrow(dat.clean))
##################################################
###Ignore turn ons/offs outside of study (there are also varying numbers of these)
> dat.clean$b.onoff.gender.group<-rep(NA ,nrow(dat.clean))
> dat.clean$b.onoff.gender.rank<-rep(NA ,nrow(dat.clean))
> dat.clean$b.onoff.race.group<-rep(NA ,nrow(dat.clean))
> dat.clean$b.onoff.race.rank<-rep(NA ,nrow(dat.clean))
##################################################
```

```
> dat.clean$b.rank.about<-rep(NA ,nrow(dat.clean))
> dat.clean$b.rank.hobbie<-rep(NA ,nrow(dat.clean))
> dat.clean$b.rank.food<-rep(NA ,nrow(dat.clean))
> dat.clean$b.rank.friends<-rep(NA ,nrow(dat.clean))
> dat.clean$b.rank.race<-rep(NA ,nrow(dat.clean))
> dat.clean$b.rank.gender<-rep(NA ,nrow(dat.clean))
########################
###White Profile
########################
> dat.clean$w.p.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$w.p.not.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$w.p.romantic.int<-rep(NA ,nrow(dat.clean))
> dat.clean$w.p.date<-rep(NA ,nrow(dat.clean))
####################################################
> dat.clean$w.f.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$w.f.not.appeal<-rep(NA ,nrow(dat.clean))
> dat.clean$w.f.romantic.int<-rep(NA ,nrow(dat.clean))
> dat.clean$w.f.date<-rep(NA ,nrow(dat.clean))
####################################################
###Ignore turn ons/offs outside of study (there are also varying numbers of these)
> dat.clean$w.onoff.gender.group<-rep(NA ,nrow(dat.clean))
> dat.clean$w.onoff.gender.rank<-rep(NA ,nrow(dat.clean))
> dat.clean$w.onoff.race.group<-rep(NA ,nrow(dat.clean))
> dat.clean$w.onoff.race.rank<-rep(NA ,nrow(dat.clean))
####################################################
> dat.clean$w.rank.about<-rep(NA ,nrow(dat.clean))
> dat.clean$w.rank.hobbie<-rep(NA ,nrow(dat.clean))
> dat.clean$w.rank.food<-rep(NA ,nrow(dat.clean))
> dat.clean$w.rank.friends<-rep(NA ,nrow(dat.clean))
> dat.clean$w.rank.race<-rep(NA ,nrow(dat.clean))
> dat.clean$w.rank.gender<-rep(NA ,nrow(dat.clean))
```

We will now have to go into the file and collect the data for the ten surveys. Fortunately, our student researcher thought to label these columns in a reproducible way. The column name begins with "PP" for the partial profile questions and "FP" for the full profile questions. This is followed by the profile number, a letter that denotes race, and a letter that denotes sex; e.g., the survey that randomized profile 3 to a White woman would have "3_WW" following the profile designation. Finally, this is followed by a label denoting the question.

These column names won't be hard to comb through because the pieces are summarized in Table 4.3.1, which outlines the possible surveys. Below is a function that collects the data for each survey based on the information provided. The arguments for the function are

- `d` – the current cleaned data frame
- `d.csv` – the raw data frame
- `survey` – the survey number being processed
- `aperson` – the profile assigned to the Asian profile
- `bperson` – the profile assigned to the Black profile

- `wperson` – the profile assigned to the White profile
- `sex` – the gender of interest

```
> extract.data<-function(d,d.csv,survey,aperson,bperson,wperson,sex){
+   obs<-which(d$survey==survey)
+   ########################
+   ###Asian
+   ########################
+   d$a.p.appeal[obs]<-d.csv[obs,paste("PP",aperson,"_A",sex,"_like_1",sep="")]
+   d$a.p.not.appeal[obs]<-d.csv[obs,paste("PP",aperson,"_A",sex,"_like_2",sep="")]
+   d$a.p.romantic.int[obs]<-d.csv[obs,paste("PP",aperson,"_A",sex,"_like_3",sep="")]
+   d$a.p.date[obs]<-d.csv[obs,paste("PP",aperson,"_A",sex,"_date_2",sep="")]
+   ######################################################
+   d$a.f.appeal[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_like_1",sep="")]
+   d$a.f.not.appeal[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_like_2",sep="")]
+   d$a.f.romantic.int[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_like_3",sep="")]
+   d$a.f.date[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_date_2",sep="")]
+   ######################################################
+   ###Ignore turn ons/offs outside of study
+   d$a.onoff.gender.group[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_sort_14_Group",
+                                 sep="")]
+   d$a.onoff.gender.rank[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_sort_14_Rank",
+                                 sep="")]
+   d$a.onoff.race.group[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_sort_15_Group",
+                                 sep="")]
+   d$a.onoff.race.rank[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_sort_15_Rank",
+                                 sep="")]
+   ######################################################
+   d$a.rank.about[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_rank_1",sep="")]
+   d$a.rank.hobbie[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_rank_2",sep="")]
+   d$a.rank.food[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_rank_4",sep="")]
+   d$a.rank.friends[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_rank_7",sep="")]
+   d$a.rank.race[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_rank_5",sep="")]
+   d$a.rank.gender[obs]<-d.csv[obs,paste("FP",aperson,"_A",sex,"_rank_6",sep="")]
+   ########################
+   ###Black
+   ########################
+   d$b.p.appeal[obs]<-d.csv[obs,paste("PP",bperson,"_B",sex,"_like_1",sep="")]
+   d$b.p.not.appeal[obs]<-d.csv[obs,paste("PP",bperson,"_B",sex,"_like_2",sep="")]
+   d$b.p.romantic.int[obs]<-d.csv[obs,paste("PP",bperson,"_B",sex,"_like_3",sep="")]
+   d$b.p.date[obs]<-d.csv[obs,paste("PP",bperson,"_B",sex,"_date_2",sep="")]
+   ######################################################
+   d$b.f.appeal[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_like_1",sep="")]
+   d$b.f.not.appeal[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_like_2",sep="")]
+   d$b.f.romantic.int[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_like_3",sep="")]
+   d$b.f.date[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_date_2",sep="")]
+   ######################################################
+   ###Ignore turn ons/offs outside of study
+   d$b.onoff.gender.group[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_sort_14_Group",
```

```
+                                   sep="")]
+   d$b.onoff.gender.rank[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_sort_14_Rank",
+                                   sep="")]
+   d$b.onoff.race.group[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_sort_15_Group",
+                                   sep="")]
+   d$b.onoff.race.rank[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_sort_15_Rank",
+                                   sep="")]
+   ####################################################
+   d$b.rank.about[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_rank_1",sep="")]
+   d$b.rank.hobbie[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_rank_2",sep="")]
+   d$b.rank.food[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_rank_4",sep="")]
+   d$b.rank.friends[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_rank_7",sep="")]
+   d$b.rank.race[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_rank_5",sep="")]
+   d$b.rank.gender[obs]<-d.csv[obs,paste("FP",bperson,"_B",sex,"_rank_6",sep="")]
+   ########################
+   ###White
+   ########################
+   d$w.p.appeal[obs]<-d.csv[obs,paste("PP",wperson,"_W",sex,"_like_1",sep="")]
+   d$w.p.not.appeal[obs]<-d.csv[obs,paste("PP",wperson,"_W",sex,"_like_2",sep="")]
+   d$w.p.romantic.int[obs]<-d.csv[obs,paste("PP",wperson,"_W",sex,"_like_3",sep="")]
+   d$w.p.date[obs]<-d.csv[obs,paste("PP",wperson,"_W",sex,"_date_2",sep="")]
+   ####################################################
+   d$w.f.appeal[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_like_1",sep="")]
+   d$w.f.not.appeal[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_like_2",sep="")]
+   d$w.f.romantic.int[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_like_3",sep="")]
+   d$w.f.date[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_date_2",sep="")]
+   ####################################################
+   ###Ignore turn ons/offs outside of study
+   d$w.onoff.gender.group[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_sort_14_Group",
+                                   sep="")]
+   d$w.onoff.gender.rank[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_sort_14_Rank",
+                                   sep="")]
+   d$w.onoff.race.group[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_sort_15_Group",
+                                   sep="")]
+   d$w.onoff.race.rank[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_sort_15_Rank",
+                                   sep="")]
+   ####################################################
+   d$w.rank.about[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_rank_1",sep="")]
+   d$w.rank.hobbie[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_rank_2",sep="")]
+   d$w.rank.food[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_rank_4",sep="")]
+   d$w.rank.friends[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_rank_7",sep="")]
+   d$w.rank.race[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_rank_5",sep="")]
+   d$w.rank.gender[obs]<-d.csv[obs,paste("FP",wperson,"_W",sex,"_rank_6",sep="")]
+   return(d)
+ }
```

Above, we used the `paste()` function to create the column names programmatically based on the details of the survey being processed. The `paste()` function creates a string based on the contents of objects provided, separated by a specified delimiter.

Remark: We note that it would be highly unlikely that anyone would write this code directly, from scratch. What we did was write code that grabs the data for the first survey. Then, we noticed the pattern in column names and worked out a plan to do this programmatically. This is all to say that reading the code will likely not be enough to understand it – we must look through the spreadsheet and consider what the code does – working through how to populate the cleaned data frame with the data from the first survey is a good start.

Now that we have the function, we populate our empty columns with the data from each survey.

```
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=1,
+                 aperson=1,bperson=5,wperson=4,sex="W")
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=2,
+                 aperson=2,bperson=1,wperson=5,sex="W")
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=3,
+                 aperson=3,bperson=2,wperson=1,sex="W")
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=4,
+                 aperson=4,bperson=3,wperson=2,sex="W")
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=5,
+                 aperson=5,bperson=4,wperson=3,sex="W")
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=6,
+                 aperson=1,bperson=5,wperson=4,sex="M")
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=7,
+                 aperson=2,bperson=1,wperson=5,sex="M")
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=8,
+                 aperson=3,bperson=2,wperson=1,sex="M")
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=9,
+                 aperson=4,bperson=3,wperson=2,sex="M")
> dat.clean<-extract.data(d=dat.clean,d.csv=dat.csv,survey=10,
+                 aperson=5,bperson=4,wperson=3,sex="M")
```

We now have a data frame containing 125 columns and 202 rows of data. Admittedly, this took a lot of time and effort to read, think about and write code to wrangle the data into a data frame. It also might have taken less time to simply do this data work in a spreadsheet program using copy and paste, but we needed to practice our basic programming skills from Chapters 0-3.

There are also things that could have made our lives easier. The survey flows could have been coded in a different way to provide a better data file at the end, but we can't go back in time. Another thing that would have saved a little time would be more helpful column names. While the column name structure really helped us pull the data together there were still many columns with names that required investigation into the question being asked. Such planning aspects are crucial to a study or experiment going smoothly.

4.4.3 Checking the Data Types

Checking the data types of the columns, we see that all 125 columns are labeled as character vectors.

```
> table(apply(X=dat.clean,MARGIN=2,FUN=typeof))
character
125
```

Due to the type of data file we are working with, where there at least is some text in each column, we have to work to ensure our data is treated appropriately. While there may be several approaches to this, we aim to practice some of the commands in Chapters 0-3. Below, we list the column names that require character, numeric and factor input and conduct those changes using the `apply()` function to the columns.

```
> character.col<-c("sub","IP","Status","StartDate","EndDate","turkID",
+                  "other.race.text","sexual.orientation.other.text","purpose",
>                  "connections.text")
> numeric.col<-c("imp.relig", "imp.polit", "imp.race","imp.hobbies",
+                "most.relig","long.relig","most.polit","long.polit",
+                "most.race","long.race","most.hobbies","long.hobbies",
+                "FT.white", "FT.hispanic", "FT.black", "FT.asian",
+                "Egal.equity","Egal.harmony","black.non_pred_import",
+                "black.stereotypes_values_ok","black.non_pred_belief",
+                "black.stereotypes_values_wrong","black.non_pred_concept",
+                "black.PC","black.neg_react","black.anger", "black.disapprove",
+                "black.pressure","asian.non_pred_import","asian.stereotypes_values_ok",
+                "asian.non_pred_belief","asian.stereotypes_values_wrong",
+                "asian.non_pred_concept","asian.PC","asian.neg_react","asian.anger",
+                "asian.disapprove","asian.pressure","age","ladder.self","household.income",
+                "a.p.appeal","a.p.not.appeal","a.p.romantic.int","a.p.date",
+                "a.f.appeal","a.f.not.appeal","a.f.romantic.int","a.f.date",
+                "a.onoff.gender.rank","a.onoff.race.rank","a.rank.about",
+                "a.rank.hobbie","a.rank.food","a.rank.friends","a.rank.race",
+                "a.rank.gender",
+                "b.p.appeal","b.p.not.appeal","b.p.romantic.int","b.p.date",
+                "b.f.appeal","b.f.not.appeal","b.f.romantic.int","b.f.date",
+                "b.onoff.gender.rank","b.onoff.race.rank","b.rank.about",
+                "b.rank.hobbie","b.rank.food","b.rank.friends","b.rank.race",
+                "b.rank.gender",
+                "w.p.appeal","w.p.not.appeal","w.p.romantic.int","w.p.date",
+                "w.f.appeal","w.f.not.appeal","w.f.romantic.int","w.f.date",
+                "w.onoff.gender.rank","w.onoff.race.rank","w.rank.about",
+                "w.rank.hobbie","w.rank.food","w.rank.friends","w.rank.race",
+                "w.rank.gender")
> factor.col<-c("rstatus","int.dating","attention.check","treatment",
+               "graph.hobbies","graph.educ","graph.int","int.gender",
+               "gender","hispanic","white","black","native.pacific",
+               "asian","other.race","sexual.orientation","highest.edu",
+               "device","connections.bool","a.onoff.gender.group","a.onoff.race.group",
+               "b.onoff.gender.group","b.onoff.race.group","w.onoff.gender.group",
+               "w.onoff.race.group","survey")
> dat.clean[,numeric.col]<-apply(X=dat.clean[,numeric.col],MARGIN=2,FUN=as.numeric)
> dat.clean[,factor.col]<-apply(X=dat.clean[,factor.col],MARGIN=2,FUN=as.factor)
```

Remark: While we haven't here, we could have specified the factors for each of the factor vectors we specified; e.g., we could have specified that the factor levels for gender are "Man", "Woman", "Other", instead of leaving them specified as the integers 1, 2, and 3, which specify which selection was made by giving the position in the list of possible answers provided. This would require going

back to each question and recording the order of possible solutions; instead, we will do so as we make statements about the data.