Disability IAT Study: Brief Sample of Data-Cleaning Code

Jenna Harder, MA

## The Data

The data in question are a messy dataset that my collaborators and I downloaded from the Open Science Framework page managed by the Project Implicit team at Harvard: <https://osf.io/y9hiq/>

Project Implicit is a website (<https://www.projectimplicit.net/index.html>) that collects data on the implicit and explicit (read: conscious and unconscious) prejudices of visitors to the site.

We looked at data from the Project Implicit tests of prejudice against people with disabilities (hereafter “disability prejudice”). The data include approximately 700,000 participants who visited the page between April of 2004 and December of 2017.

Note: I worked on this project with my colleagues Victor Keller and William Chopik, both of whom contributed invaluably to all stages of the project. However, as project leader I handled the bulk of the analysis, and the subset presented here is my own work.

In this sample, I walk through how I dealt with one of the simpler issues that came up with cleaning this data: inconsistency in how participants’ ages were recorded.

# Cleaning the Age Variable

This study started out by asking participants their age, and later switched to asking them for their month and year of birth. We want one variable that has everyone’s age in years. Let’s look at this data:

diat$datedate[which(!is.na(diat$age))[length(which(!is.na(diat$age)))]] #the last day on which we have numerical data for age:

## [1] "8/16/2016"

summary(diat$age)

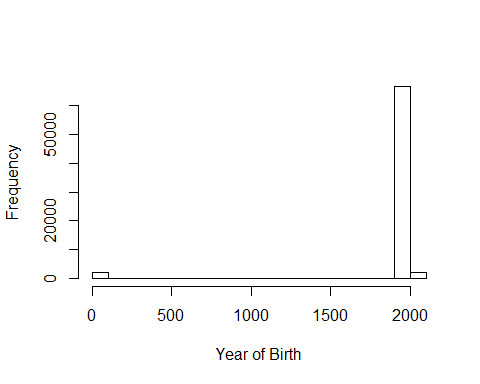
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 10.0 19.0 23.0 28.2 34.0 89.0 73646

summary(diat$birthyear)

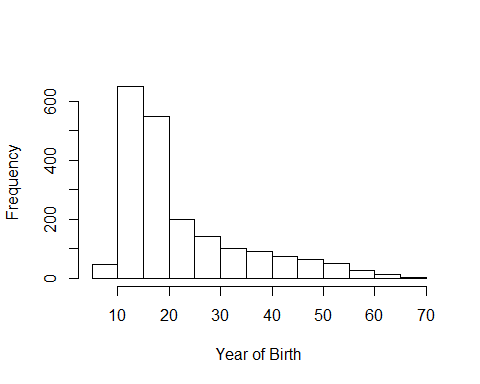
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 6 1978 1992 1932 1996 2009 248123

That birthyear data looks a little weird. It’s unlikely that we had any participants who were born in the year 6 A.D.

A histogram suggests that only a few people put very low numbers:



A second histogram of all “birthyear” values below 1900 shows that all of these are lower than 100.



Perhaps some people entered their age here; perhaps others were putting the month or day of their birth date, or answering at random. It’s impossible to know, and this is just a small fraction of our sample, so we’ll consider these individuals to have missing data for age.

diat$birthyear[which(diat$birthyear < 1900)] <- NA

Since we have birth month/year and current date for all participants after August 2016, we could calculate relatively precise ages for these participants. But to be consistent with the earlier participants, who listed their ages in years, we will estimate integer ages for these later participants. Those who participated after the 15th of their birthmonth in a given year will be assumed to have had their birthday that year.

agerows <- which(!is.na(diat$birthyear))  
for(dude in(agerows)){  
 if(is.na(diat$birthmonth[dude])){  
 if(diat$month[dude]<=6){  
 diat$age[dude] <- diat$year[dude] - diat$birthyear[dude] - 1  
 }  
 else if(diat$month[dude]>6){  
 diat$age[dude] <- diat$year[dude] - diat$birthyear[dude]  
 }  
 }  
 else if(diat$birthmonth[dude] > diat$month[dude]){ #if it's before their birthmonth  
 diat$age[dude] <- diat$year[dude] - diat$birthyear[dude] - 1   
 }  
 else if(diat$birthmonth[dude] < diat$month[dude]){ #if it's after their birthmonth  
 diat$age[dude] <- diat$year[dude] - diat$birthyear[dude]  
 }  
 else if(diat$birthmonth[dude] == diat$month[dude]){ #if it's their birthday month  
 if(diat$day[dude] <= 15){ #if it's before the 15th, assume they haven't had their birthday yet  
 diat$age[dude] <- diat$year[dude] - diat$birthyear[dude]-1  
 }  
 else{  
 diat$age[dude] <- diat$year[dude] - diat$birthyear[dude]  
 }  
 }  
}

Our new age variable looks like this:

summary(diat$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 6.00 20.00 24.00 28.44 35.00 107.00 4747