Lab 2 Work

BSTA 512/612

2024-02-08

Directions

Please turn in your `.html` file on Sakai. Please let me know if you greatly prefer to submit a physical copy.

You can download the `.qmd` file for this lab here.

The rest of this lab's instructions are embedded into the lab activities.

Purpose

The main purpose of this lab is to introduce our dataset, codebook, and variables. We will continue to think about the context of our research question, but our main focus is to become familiar with the data.

Grading

This lab is graded out of 12 points. Nicky will use the following rubric to assign grades.

Rubric

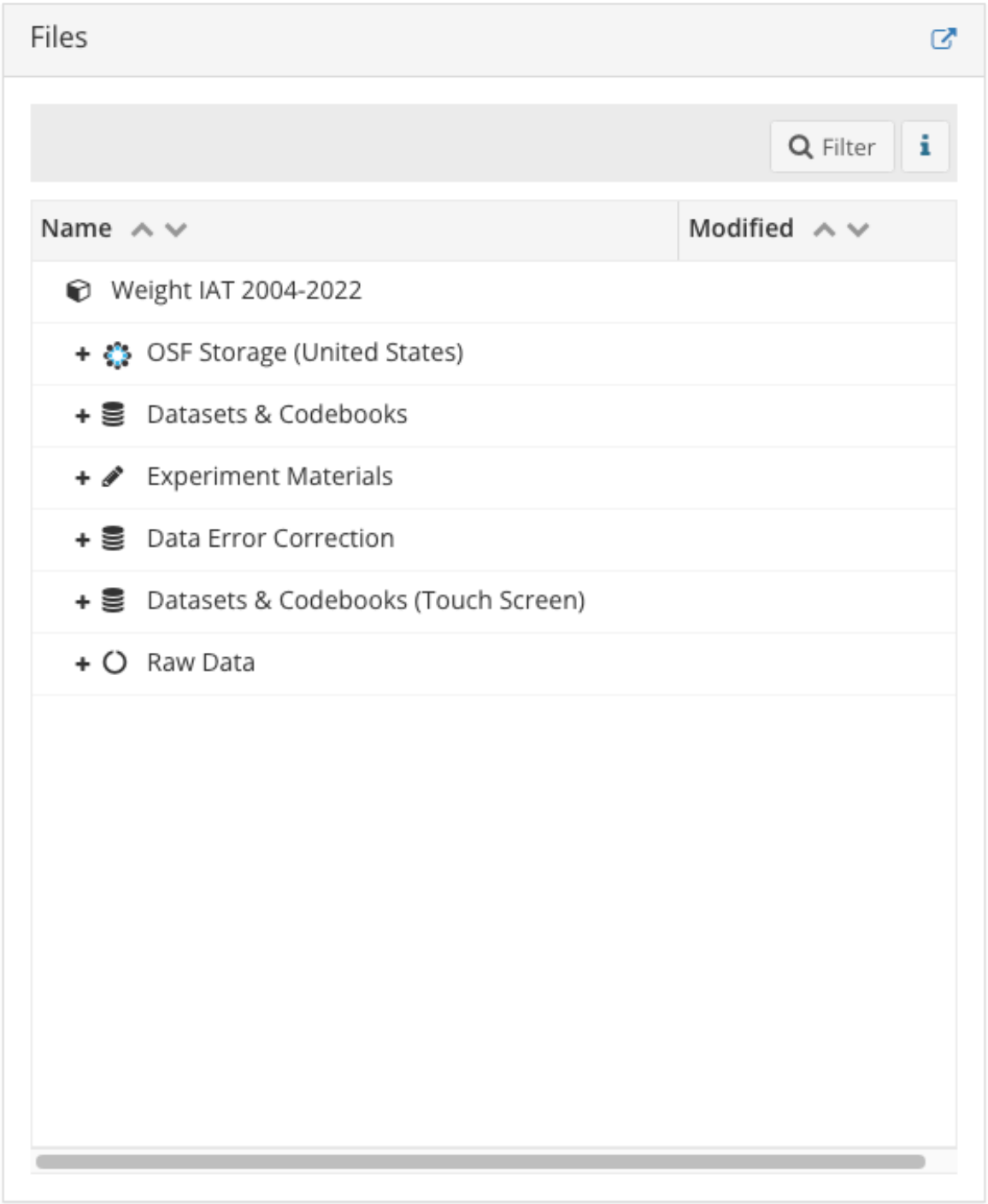
	4 points	3 points	2 points	1 point	0 points
Answers	Answers demonstrate completion and understanding of the needed activity*. Answers are thoughtful and can be easily integrated into the final report.	Answers demonstrate completion and understanding of the needed activity*. Answers are thoughtful, but lack the clarity needed to easily integrate into the final report.	Answers demonstrate completion and minimal understanding of the needed activity*. Answers are fairly thoughtful, but lack connection to the research.	Answers demonstrate completion of needed activities*, although evidently rushed through. Answers seem rushed and with minimal thought.	It is evident that the needed activities* were not completed. Answers seem rushed and without thought.
Formatting	Lab submitted on Sakai with `.html` file. Answers are written in complete sentences with no major grammatical nor spelling errors. With little editing, the answer can be incorporated into the project report.	Lab submitted on Sakai with `.html` file. Answers are written in complete sentences with grammatical or spelling errors. With editing, the answer can be incorporated into the project report.	Lab submitted on Sakai with `.html` file. Answers are written in complete sentences with major grammatical or spelling errors. With major editing, the answer can be incorporated into the project report.	Lab submitted on Sakai with `.html` file. Answers are bulletted or do not use complete sentences.	Lab *not* submitted on Sakai with `.html` file.
Code Reasoning					

Lab activities

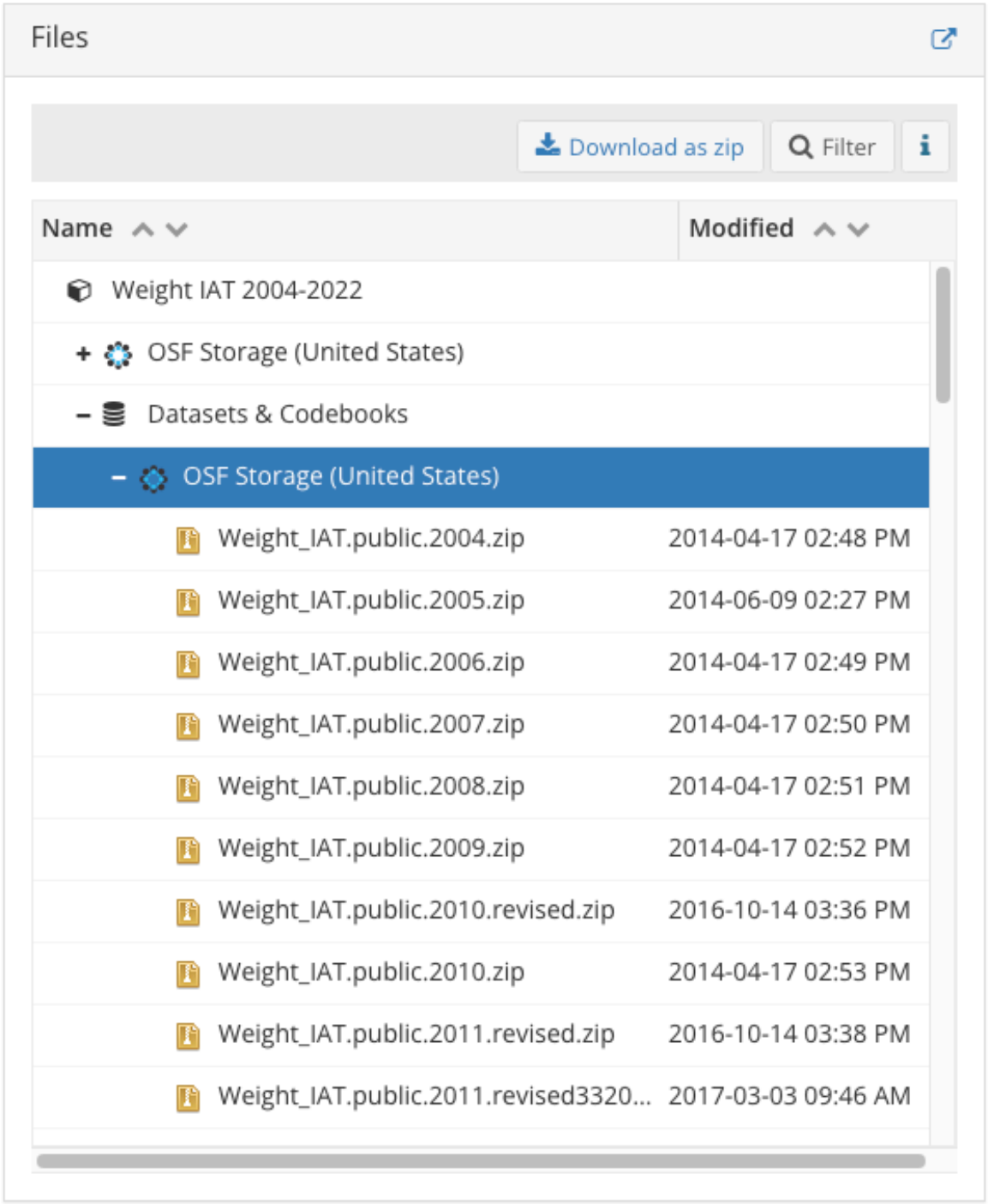
1. Access and download the data

This serves as good practice for accessing data that is online or needs to be downloaded from a collaborator.

Data can be accessed here. Under "Weight IAT 2004-2022" there are several drop down menus:

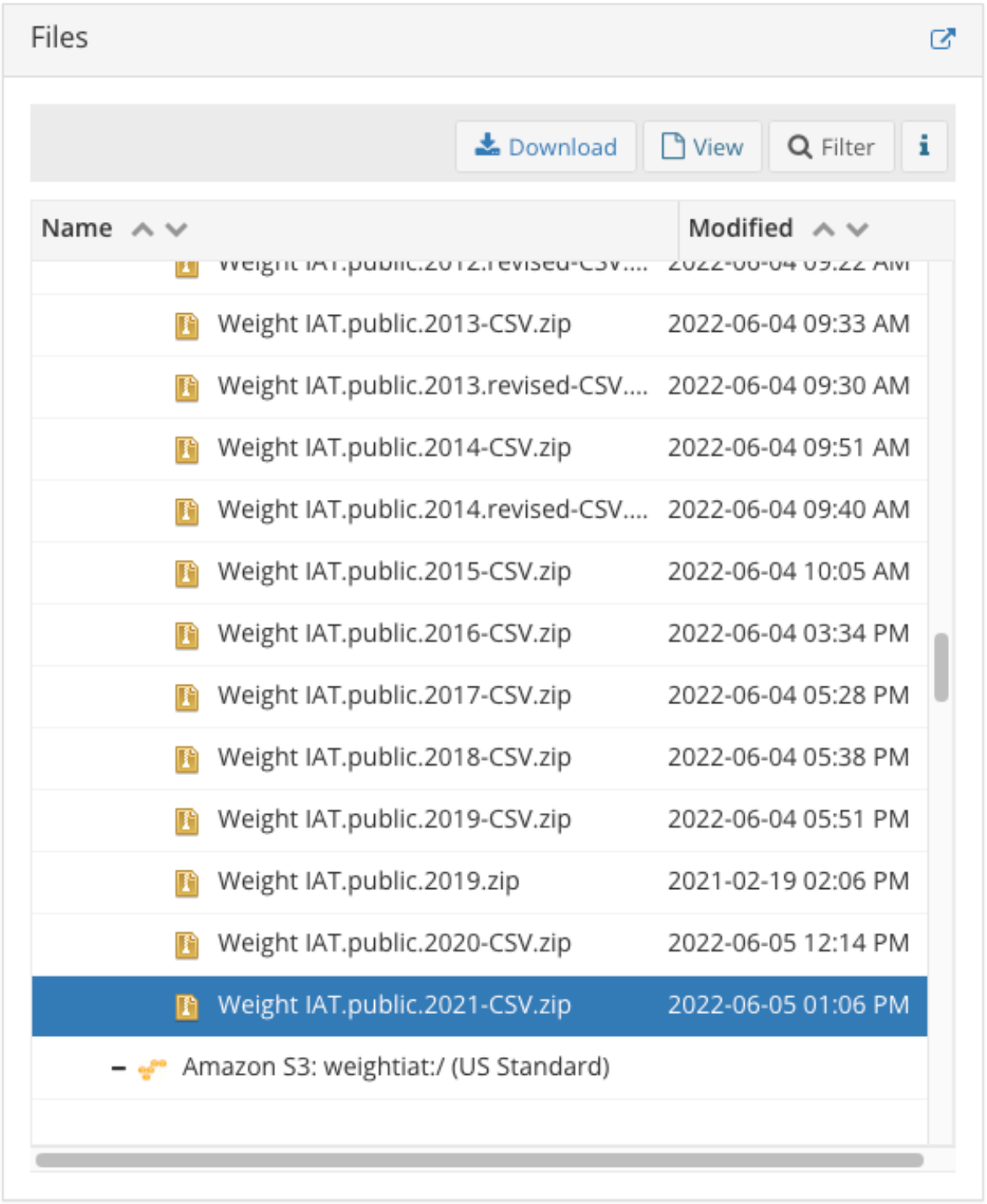


I opened the first "Datasets & Codebooks," then selected "OSF Storage (United States)." Once selected, the "Download as zip" option pops up in the top right part of the Files section.



We will be working with the `Weight_IAT.public.2021.csv` dataset. Please locate the `zip` file called `Weight IAT.public.2021-CSV.zip` . T0 download, you need to click the row of

the zip file, but you can't click the name of the zip file. If a link opens, then you clicked the name. If the row is highlighted blue and clickable "Download" and "View" buttons appear on the top right, then you selected it correctly! (See below image for what it should look like.)



Then click the "Download" button to download! Note that the name does not have an underscore between "Weight" and "IAT." I like to have my datasets named without spaces, so I will

replace the space with an underscore.

For the codebook, perform the same process for the file named: `Weight_IAT_public_2021_codebook.xlsx`

You will need to unzip the actual data.

Move the data to a folder that you can easily access as you work from this document. I like to have a folder named `data` to house my data.

2. Load data and needed packages

First, load the packages that you will need in the remainder of this lab. You can add to this as you need to. At the top of your R code chunk, you can add the following option to repress the messages from the loading packages:

```
{r}
#| message: false
```

```
library(tidyverse)
library(gtsummary)
library(here)
```

Using R, load the data (`csv` file) into this document. Note that this is a `csv` file that we can load with basic R packages. Name your dataset something that feels intuitive to you and will distinguish it from other datasets that you work with.

```
iat_2021_raw = read.csv(file = here("../TA_files/Project/data/Weight_IAT.public.2021.csv")
```

Loading the `csv` file every time you render will take a long time. One way to speed this up is by saving the data as an `rda` file (R data file). Change the following R code to save the `rda` file. You will also need to remove the `#| eval: false` at the top of the code chunk once you have corrected the code. If you are confused on the syntax, don't forget that you can use `?save` for more information.

```
save(<whatever you called the read csv file>, file = "Where you would like to save the fil
```

Check that you have an `rda` file where you save it. Now use `load()` with the file path to load the `rda` data here.

```
load(file = "Where you would like to save the file with its name")
```

At this point, if you think you loaded the file correctly, add `#| eval: false` to the code chunk where you loaded the `csv` file and back to the chunk where you saved the `rda` file.

Take a glimpse at the data to make sure you loaded it correctly. This should be able to run with just the code chunk with your `load()` command.

```
glimpse(iat_2021_raw)
```

```
Rows: 465,886
Columns: 92
$ session_id              <dbl> 2653543637, 2653543649, 2653543656, 2653543718~
$ session_status          <chr> " ", " ", "C", " ", "C", "C", "C", " ", " ", "~
$ study_name              <chr> "Demo.Weight.0004", "Demo.Weight.0004", "Demo.~
$ date                    <chr> "1/1/2021 0:00:49", "1/1/2021 0:02:36", "1/1/2~
$ month                   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ day                     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ year                    <int> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021~
$ hour                    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1~
$ weekday                 <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6~
$ birthmonth              <int> NA, 3, 1, NA, 1, 3, 5, NA, NA, 11, 4, 1, 4, 1,~
$ birthyear               <int> NA, 1975, 1979, NA, 1972, 2002, 1943, NA, NA, ~
$ birthSex                <int> NA, 2, 2, NA, 2, 1, 2, NA, NA, 1, 2, 2, 1, 2, ~
$ genderIdentity          <chr> " ", "[2]", "[2]", " ", "[2]", "[1]", "[2]", "~
$ num_002                 <int> NA, 1, 1, NA, 1, 1, 2, NA, NA, 1, 1, 1, 4, 4, ~
$ ethnicityomb            <int> NA, 2, 2, NA, 2, 3, 2, NA, NA, 1, 2, 2, 2, 2, ~
$ raceomb_002             <int> NA, 6, 6, -999, 6, 5, 6, NA, NA, 6, 7, 5, 6, 5~
$ raceombmulti            <chr> " ", " ", " ", " ", " ", " ", " ", " ", " ", "~
$ D_biep.Thin_Good_all    <dbl> NA, NA, -0.40210378, 0.51834547, 0.67850537, 0~
$ Mn_RT_all_3467          <dbl> NA, NA, 864.4333, 911.1250, 1088.9833, 655.941~
$ N_3467                  <int> NA, NA, 120, 120, 120, 120, 120, NA, NA, 120, ~
$ PCT_error_3467          <dbl> NA, NA, 7.500000, 5.000000, 2.500000, 7.500000~
$ Order                   <int> NA, NA, 2, 1, 1, 2, 1, NA, NA, 2, 2, NA, 2, 1,~
$ Side_Thin_34            <int> NA, NA, 1, 2, 2, 1, 2, NA, NA, 1, 1, NA, 1, 2,~
$ Side_Good_34            <int> NA, NA, 2, 2, 2, 2, 2, NA, NA, 2, 2, NA, 2, 2,~
$ Stimuli                 <int> NA, NA, 3, 3, 3, 3, 3, NA, NA, 3, 3, NA, 3, 3,~
$ pct_300                 <dbl> NA, NA, 0.000000, 0.000000, 0.000000, 8.333333~
$ pct_400                 <dbl> NA, NA, 0.8333333, 0.0000000, 0.0000000, 14.16~
$ pct_2K                  <dbl> NA, NA, 0.8333333, 4.1666667, 7.5000000, 3.333~
$ pct_3K                  <dbl> NA, NA, 0.8333333, 0.0000000, 0.8333333, 1.666~
$ pct_4K                  <dbl> NA, NA, 0.0000000, 0.0000000, 0.0000000, 0.000~
$ att7                    <int> NA, NA, 4, 4, 5, 4, 5, NA, NA, 6, 5, NA, 5, NA~
$ tfat                    <int> NA, NA, 7, NA, 5, 5, 3, NA, NA, 6, 5, NA, 3, N~
$ tthin                   <int> NA, NA, 7, NA, 5, 5, 5, NA, NA, 6, 7, NA, 7, N~
$ comptomost_001          <int> NA, NA, 5, NA, 3, 4, 4, NA, NA, 5, 4, NA, 5, N~
$ controlyou_001          <int> NA, NA, 3, NA, 2, 3, 3, NA, NA, 3, 2, NA, 2, N~
```

```
$ controlother_001        <int> NA, NA, 3, NA, 3, 3, 3, NA, NA, 2, 2, NA, 3, N~
$ easytolose_001          <int> NA, NA, 4, NA, 3, 3, 4, NA, NA, 3, 4, NA, 3, N~
$ iam_001                 <int> NA, NA, 5, NA, 4, 4, 4, NA, NA, 6, 5, NA, 5, N~
$ identfat_001            <int> NA, NA, 3, NA, 2, 3, 3, NA, NA, 2, 1, NA, 2, N~
$ identthin_001           <int> NA, NA, 3, NA, 3, 3, 2, NA, NA, 2, 1, NA, 2, N~
$ important_001           <int> NA, NA, 4, NA, 3, 3, 2, NA, NA, 4, 4, NA, 3, N~
$ mostpref_001            <int> NA, NA, 6, NA, 4, 4, 6, NA, NA, 6, 6, NA, 6, N~
$ othersay_001            <int> NA, NA, 4, NA, 4, 4, 4, NA, NA, 4, 4, NA, 5, N~
$ D_biep.Thin_Good_36     <dbl> NA, NA, -0.65922173, 0.25492457, 0.85758482, 0~
$ D_biep.Thin_Good_47     <dbl> NA, NA, -0.1449858, 0.7817664, 0.4994259, 0.36~
$ Mn_RT_all_3             <dbl> NA, NA, 708.40, 773.70, 978.90, 911.90, 1192.6~
$ Mn_RT_all_4             <dbl> NA, NA, 864.9500, 758.3750, 888.8000, 623.8250~
$ Mn_RT_all_6             <dbl> NA, NA, 890.550, 860.350, 1510.250, 717.200, 2~
$ Mn_RT_all_7             <dbl> NA, NA, 928.875, 1157.975, 1133.575, 529.450, ~
$ SD_all_3                <dbl> NA, NA, 262.0541, 303.0142, 472.6722, 805.6804~
$ SD_all_4                <dbl> NA, NA, 537.2739, 317.2618, 569.8885, 296.7929~
$ SD_all_6                <dbl> NA, NA, 265.6665, 376.0075, 644.9575, 438.7512~
$ SD_all_7                <dbl> NA, NA, 320.9680, 588.8174, 362.0522, 213.2356~
$ N_3                     <int> NA, NA, 20, 20, 20, 20, 20, NA, NA, 20, 20, NA~
$ N_4                     <int> NA, NA, 20, 20, 20, 20, 20, NA, NA, 20, 20, NA~
$ N_5                     <int> NA, NA, 28, 28, 28, 28, 28, NA, NA, 28, 28, NA~
$ N_6                     <int> NA, NA, 20, 20, 20, 20, 20, NA, NA, 20, 20, NA~
$ N_7                     <int> NA, NA, 40, 40, 40, 40, 40, NA, NA, 40, 40, NA~
$ Mn_RT_correct_3         <dbl> NA, NA, 708.4000, 773.7000, 978.9000, 791.0526~
$ Mn_RT_correct_4         <dbl> NA, NA, 708.8235, 745.5385, 809.2051, 601.8421~
$ Mn_RT_correct_6         <dbl> NA, NA, 890.5500, 840.8947, 1487.1053, 661.176~
$ Mn_RT_correct_7         <dbl> NA, NA, 872.3243, 1121.3333, 1114.1282, 501.05~
$ SD_correct_3            <dbl> NA, NA, 262.0541, 303.0142, 472.6722, 613.9011~
$ SD_correct_4            <dbl> NA, NA, 253.1620, 310.7078, 270.6234, 286.9534~
$ SD_correct_6            <dbl> NA, NA, 265.6665, 375.8263, 654.0420, 446.9360~
$ SD_correct_7            <dbl> NA, NA, 258.7414, 597.8491, 344.9726, 185.6936~
$ N_ERROR_3               <int> NA, NA, 0, 0, 0, 1, 0, NA, NA, 1, 0, NA, 2, 0,~
$ N_ERROR_4               <int> NA, NA, 6, 1, 1, 2, 0, NA, NA, 10, 3, NA, 1, 0~
$ N_ERROR_6               <int> NA, NA, 0, 1, 1, 3, 1, NA, NA, 1, 2, NA, 2, 1,~
$ N_ERROR_7               <int> NA, NA, 3, 4, 1, 3, 1, NA, NA, 5, 4, NA, 6, 4,~
$ myweight_002            <int> NA, NA, 24, NA, 20, 32, 19, NA, NA, 34, 18, NA~
$ myheight_002            <int> NA, NA, 36, NA, 34, 38, 33, NA, NA, 33, 30, NA~
$ countrycit_num          <int> NA, 1, 85, NA, 1, 1, 1, NA, NA, 1, 105, 1, 1, ~
$ countryres_num          <int> NA, 1, 85, NA, 1, 1, 1, NA, NA, 1, 105, 1, 1, ~
$ edu                     <int> NA, 7, 7, NA, 9, 4, 11, NA, NA, 5, 11, 5, 13, ~
$ edu_14                  <int> NA, 7, 7, NA, 9, 4, 11, NA, NA, 5, 11, 5, 13, ~
$ occuSelf                <chr> " ", "11-", "43-", " ", "2931", "15-", "2931",~
$ occuSelfDetail          <chr> " ", "1", "43-6000", " ", "29-1000", "15-1000"~
```

```
$ politicalid_7           <int> NA, 6, 4, NA, 5, 4, 6, NA, NA, 5, 4, 5, 6, NA,~
$ STATE                   <chr> " ", "NC", " ", " ", "NY", "NC", " ", " ", " "~
$ CountyNo                <int> NA, 129, NA, NA, 55, 105, NA, NA, NA, 37, NA, ~
$ MSANo                   <int> NA, 48900, NA, NA, 40380, 99032, NA, NA, NA, 3~
$ MSAName                 <chr> " ", "Wilmington, NC MSA", " ", " ", "Rocheste~
$ religion2014            <int> NA, 7, 2, NA, 7, 2, 1, NA, NA, 2, 6, 7, 7, 3, ~
$ religionid              <int> NA, 1, 2, NA, 1, 3, 2, NA, NA, 2, 3, 4, 2, 4, ~
$ iatevaluations001       <int> NA, NA, 4, NA, 3, 3, 3, NA, NA, 4, NA, NA, 3, ~
$ iatevaluations002       <int> NA, NA, 3, NA, 1, 3, 2, NA, NA, 2, NA, NA, 2, ~
$ iatevaluations003       <int> NA, NA, 3, NA, 3, 3, 2, NA, NA, 1, NA, NA, 2, ~
$ broughtwebsite          <chr> " ", " ", "Mention or link at a non-news Inter~
$ user_id                 <int> -1, -1, -1, -1, -1, -1, 11555672, -1, -1, -1, ~
$ previous_session_id     <dbl> NA, 2653543637, NA, 2653543685, NA, NA, 265354~
$ previous_session_schema <chr> " ", "s", " ", "s", " ", " ", "s", " ", "s", "~
```

How many rows and columns are in the dataset? Do you think we will need all these variables for our analysis?

3. Data wrangling

As you go through this process, it is important that you look at the codebook for more information on each variable.

3.1 What's our target population?

As many of you mentioned in Lab 1, individuals taking the IAT test are not necessarily representative of the world population. I want you to articulate the target population that you think our analysis can give information about. To what population can we generalize our analysis results? We can get very specific with this population, but try to restrict your population to 3-5 characteristics.

After you articulate the population, I want to add one more restriction to our population: US residency. The sample includes individuals residing in many different countries. Since we are discussing attitudes and beliefs that is inherently connected to society and culture, I think it is important that we restrict our analysis and discussion to a country that we have some social experience in. **Thus, let's restrict our data to the US only by filtering the variable `countryres` to category 1 (corresponding to the US).**

3.1 Restrict your analysis to 1 outcome and 9 possible covariates/predictors

We are going to restrict our analysis to the single outcome, IAT score, which is named `D_biep.Thin_Good_all`.

We will also restrict our analysis to the following 9 potential variables so our work is a little more manageable.

> ❗ Task
>
> From the following 7 attitudes and beliefs, please select 3 that you think will be the most important variables related to your research question. In 1-2 lines, briefly explain why you chose each variable. This can be informal and bulleted.

(Make sure you chose the variable that is part of your research question!)

1. Self-perception of weight (`iam_001`)
2. Fat group identity (`identfat_001`)
3. Thin group identity (`identthen_001`)
4. Controllability of weight of others (`controlother_001`)
5. Controllability of weight of yourself (`controlyou_001`)
6. Awareness of societal standards (`mostpref_001`)
7. Internalization of societal standards (`important_001`)

We will start our data exploration with the following 4 demographic variables:

1. Age (we need to construct)
2. Race (`raceomb_002` or `raceombmulti`)
3. Ethnicity
4. Sex assigned at birth (`birthSex`)

Please pick 2 additional variables to include in your analysis:

1. Education (`edu`)
2. Gender (`genderIdentity`)
3. Self-reported BMI (through self-reported height and weight)
4. Political identity
5. Religion

I have chosen these variables for a mixture of reasons. For example, I have left out variables about residence and occupation because those variables have hundreds of categories that would be overwhelming in linear regression. For the 4 required demographic variables, I chose age because I really want us to get practice with a continuous variable. I chose race and ethnicity because of the intertwined history of racism and anti-fat bias in Western countries (including the U.S. where most participants reside).

A note of the available variables on race

The dataset has two separate race variables. One has mutually exclusive categories (`raceomb_002`) and the other allows participants to make multiple selections (`raceombmulti`). The former (`raceomb_002`) allows one participant to identify with only one race category.
Important lesson from We All Count about using a multiple selection race question. We can try out all these options!

Finally, I chose sex assigned at birth because adults in 2021 in the US were likely raised in a society where your sex assigned at birth impacted the gender stereotypes that you were raised in, which could impact exposure to diet culture. The reason why I am leaving gender as an optional variable is because the question on gender allows participants to chose multiple options. The binary sex assigned at birth will make our analysis a little easier from a statistics stand point. Unfortunately, we need to balance achievable learning objectives and the most appropriate variable. Since race is a required variable and has a multi-level option, I do not want to overload our analysis with another multi-level variable. This is certainly a limitation in our analysis that we should address in our discussion. I do encourage you to look into gender if the binary sex assigned at birth does not feel right for you. I am happy to help.

diet culture has historically targeted genders in different ways.

A word on self-reported BMI

This variable is rooted in racism and anti-fat bias. The American Medical Association made a few press releases on policies using BMI as a measure, with alternative measures (frankly, just other measures of fatness to use as a diagnostic tool instead of checking true indicators of health). However, I can think of a couple examples where BMI might help us understand some context in this research, so I have left it as an option. Although still self-reported, it might be interesting to see how BMI (which is the closest measurement available in this dataset to an "objective" measure of fatness) is related to individuals' attitudes and beliefs. I am not saying there is anything to the relationship, but it might be worth checking out if you are interested.
I will also say, there are MANY issues, within this dataset, constructing the variable for BMI from height and weight. *If you do not feel strongly about including it, I would suggest you avoid the variable self-reported BMI.* It is not worth bringing in a racist and anti-fat variable into the dataset if you do not have a specific use for it.

If you would like to investigate a variable outside the list, please let me know by emailing or chatting with me.

Task

Using R, select your identified variables from your dataset. Your new dataset should have 11 columns for the 11 variables.

3.1 Manipulating variables that are coded as numeric variables

Many variables in this dataset are coded as numeric values, but have specific categories linking up to the numbers. Using `mutate()` and `cases()` similar to our Data Management lesson, please create a new categorical variable with the specified categories from the codebook. Make sure that you create a variable with a new name! Since some of these variables are ordered categories, we will investigate if it's appropriate to use the numeric or categorical version of the variable.

Example of how I would create new variable for self-perception of weight (`iam_001`):

By looking at the codebook, I see that respondents answer the following question: "Currently, I am:"

- "Very underweight"
- "Moderately underweight"
- "Slightly underweight"
- "Neither underweight nor underweight"
- "Slightly overweight"
- "Moderately overweight"
- "Very overweight"

If I look at the data as is, I see that the variable is numeric.

```
iat_2021_raw %>%
  dplyr::select(iam_001) %>%
  tbl_summary()
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N = 465,886**
iam_001	
1	2,023 (0.6%)

2	7,902 (2.4%)
3	24,399 (7.3%)
4	148,081 (44%)
5	88,566 (27%)
6	43,090 (13%)
7	18,978 (5.7%)
Unknown	132,847

Again, I want to create a varaible with the answers instead of numbers, so I will change transform the variable to include the text:

```
iat_2021_raw = iat_2021_raw %>%
  mutate(iam_001_f = case_match(iam_001,
                                7 ~ "Very overweight",
                                6 ~ "Moderately overweight",
                                5 ~ "Slightly overweight",
                                4 ~ "Neither underweight nor underweight",
                                3 ~ "Slightly underweight",
                                2 ~ "Moderately underweight",
                                1 ~ "Very underweight",
                                .default = NA # to add NA if unknown
                                ))
```

I have called the new variable `iam_001_f` to indicate that the variable is not in factor form. You can also call it something like `iam_001_cat` to indicate the categorical form.

! Task

Identify and list the variables that are coded numerically and correspond to categories. Create a new variable for the categorical/factor version of the variable. If you are using multi-choice categorical variables (might include race, gender), then do not convert the variable yet!

3.2 Converting age from birth date and test date

3.3 Make a new dataset with only complete cases

Quickly make sure that we are not introducing bias by using complete cases

4. Some exploratory data analysis

4.1 Peek at your outcome

This serves as a check to make sure we are all looking at the correct outcome: IAT score. Please plot a histogram of the IAT scores.

4.1 Univariate exploratory data analysis

4.2 Bivariate exploratory data analysis

4.3 Multivariate exploratory data analysis

5. Revisit your research question

Please restate the research question that you proposed in Lab 1. What are your thoughts on the research question now that we looked at the data?

5. Make a Table 1