CSDE 502 2021 Winter **Assignment 6** ***Add Health data; Variable creation*** **Instructor: Phil Hurvitz** **phurvitz@uw.edu**	**My Name: David Coomes** **My UWNetID: dcoomes**

Due Date: 2021-02-18 09:00 AM

Instructions:

1. Fill in your name and UWNetID above.
2. Put answers to the questions on this document, using the “00Answers” Word style so your answers are clearly distinguished from the questions.
3. Create a PDF file from this document.
4. Create a **single zip** file including this document as a PDF file, along with the RDS file and R code file.
5. Upload the **single zip file** to Canvas.

Explanation:

For this assignment, you will be perusing some of the documentation for the Add Health Wave 1 data set. You will use the documentation to make some updates to a data frame containing some of the Add Health data, and then save the data frame as an RDS file. You will update a metadata table that partially describes the data set and changes you made to the variable names and variable labels.

To open a Stata version 13 file in R there are two main options:

1. Use `haven::read_dta()`. To access variable labels in R use `labelled::foreign_to_labelled()`. To update variable labels, use the `labelled::var_label()` function.
2. Use `readstata13::read.dta13()`. Variable labels for this format are available, e.g., for a data frame named `dat` as `attributes(dat)$var.labels`. This is a vector of text strings that can be updated by assigning a new value to the specified element, e.g., `attributes(dat)$var.labels[1] <- “foo”`.

To save the RDS file, use the base function `saveRDS()`.

Here is a base R code snippet that will rename a single variable:

```
colnames(data_frame)[grep("^original_variable_name$", colnames(data_frame))]
<- new_variable_name
```

The `grep()` function finds the position of the named variable in the list of variables in the data frame. The characters `^` and `$` are regular expressions to specify the start and end of the string to be matched (assuring that the pattern does not match multiple similar variable names).

It is much simpler with tidyverse and magrittr:

```
data_frame %<>% rename(new_variable_name = old_variable_name)
```

Additional hint for dealing with PDF documentation:

1. Use `pdfgrep` (should be available in a Linux or Mac package manager; for Windows, search for a version or use Cygwin).
2. Use the R `pdftools` package. This could be used in a loop over each PDF file to create a data frame with the name of the PDF file, page number, and text of each page. The str_match() function could be used to identify the file name and page number where specific text strings occur. For a minimal example, this shows that the string “h1gi1m” is found on page 1 of INH01PUB.PDF. Conversion of the PDF file’s text to lowercase simplifies the matching:

```
> x <- pdftools::pdf_text(pdf = "INH01PUB.PDF")
> str_match(string = x %>% str_to_lower(), pattern = "h1gi1m")
      [,1]
 [1,] "h1gi1m"
 [2,] NA
 [3,] NA
 [4,] NA
 [5,] NA
 [6,] NA
 [7,] NA
 [8,] NA
 [9,] NA
[10,] NA
[11,] NA
[12,] NA
[13,] NA
[14,] NA
[15,] NA
```

Questions:

1. Explore the Add Health website (http://www.cpc.unc.edu/projects/addhealth) and answer the following questions (making sure to cite as necessary):
    1.1. What was the sampling frame for this study?

The sampling frame for the Add Health study was all high schools included in the Quality Education Database (QED). High school was defined as schools with an 11th grade and more than 30 students.[1]

    1.2. What were the three kinds of respondents at Wave I?

The core sample included a stratified random sample from the selected schools. They also included all enrolled students (who agreed to participate) in two large schools and 14 small schools.[1]

1.3. What was the instrument with the largest sample size?

The instrument with the largest size is the core sample.[1]

1.4. Is it possible for a respondent to be in Wave III without being in Wave II?

Yes, for Wave III they attempted to follow up with everyone from Wave I so a respondent could have been left out, or chosen not to participate in Wave II but included in Wave III.[1]

1.5. What is the time span of the Add Health data collection (all waves)?

The Add Health data was collected between 1994 – 2009.[1]

1.6. What is the difference between the public and the restricted-use Add Health data?

The public use data contains only a subsample of the core sample questionnaire. The restricted-use data contains data for the entire core sample as well as additional data including obesity, neighborhood environment, genetics, disposition, political context, and alcohol density.

1.7. Describe a research question that you might be able to answer using the Add Health dataset.

What is the difference in obesity for young adults who lived in rural areas during their adolescence compared to those that lived in urban areas during their adolescence?

2. Download the public-use Add Health documentation at https://canvas.uw.edu/courses/1434040/files. Answer the following questions:

2.1. In what pdf document is the documentation for the race items for the Wave I In-Home questionnaire?

INH01PUB.PDF

2.2. How many respondents were of Hispanic/Latino origin?

743

2.3. What is the "Knowledge Quiz" in the Wave I In-Home questionnaire?

The “Knowledge Quiz” was designed to examine a respondent’s knowledge about pregnancy and birth control.

2.4. What is the unique identifier for the In-home data?

The unique identifier is “aid”

3. Download the Stata 13 format file AHwave1_v1.dta (http://staff.washington.edu/phurvitz/csde502_winter_2021/data/AHwave1_v1.dta).

3.1. Fill in the grey missing cells in Table 1 below based on the data and/or documentation. Optimally, use the documentation to familiarize yourself with the structure of the code books.

3.2. Using questions 6 and 8 in INH01PUB.PDF, create a new variable named "race" that uses recoded values (white = 1; black/African American = 2; American Indian = 3; Asian/Pacific Islander = 4; other = 5; unknown/missing = 9).

3.3. Rename the variables, and update variable labels using Table 1 as a guide and save the data frame as the file as **AHwave1_v2.rds**. Use a single R code file for your edits to the data file.

3.4. Update the status in Table 1 as needed.

Table 1: Codebook for variables from Add Health Wave 1 data

new variable name	original variable name	status*	data type	values	new variable label	codebook file name
aid	aid	unchanged	text	8 digit string	unique case (student) identifier	SECTAPUB.PDF
imonth	imonth	unchanged	integer	1 4 to 12	month interview completed	SECTAPUB.PDF
iday	iday	unchanged	integer	1 - 31	day interview completed	SECTAPUB.PDF
iyear	iyear	unchanged	integer	94, 95	Year interview completed	SECTAPUB.PDF
bio_sex	bio_sex	unchanged	integer	1,2,6	interviewer confirmed sex	SECTAPUB.PDF
bmonth	h1gi1m	renamed	integer	1-12, 96	birth month	INH01PUB.PDF
byear	h1gi1y	renamed	integer	74-83,96	birth year	INH01PUB.PDF
hispanic	h1gi4	renamed	integer	0=No 1=Yes 6=Refused 8=Don’t know	Hispanic/Latino	INH01PUB.PDF
white	h1gi6a	renamed	integer	0 = not marked 1 = marked 6 = refused 8 = don't know	race white	INH01PUB.PDF
black	h1gi6b	renamed	integer	0=No 1=Yes 6=Refused 8=Don’t know	race black or African American	INH01PUB.PDF
AI	h1gi6c	renamed	integer	0=No 1=Yes 6=Refused 8=Don’t know	race American Indian or Native American	INH01PUB.PDF
asian	h1gi6d	renamed	integer	0=No 1=Yes 6=Refused 8=Don’t know	race Asian or Pacific Islander	INH01PUB.PDF
raceother	h1gi6e	renamed	integer	0=No 1=Yes 6=Refused 8=Don’t know	race other	INH01PUB.PDF
onerace	h1gi8	renamed	integer	1=White 2=Black/African American 3=American Indian/Native American	one category best describes racial background	INH01PUB.PDF

				4=Asian/Pacific Islander 5=Other 6=Refused 7=Legitimate skip 8=Don’t know 9=Not applicable		
observedrace	h1gi9	renamed	integer	1=White 2=Black or African American 3=American Indian or Native American 4=Asian or Pacific Islander 5=Other 6=Refused 8=Don’t know	interviewer observed race	INH01PUB.PDF
health	h1gh1	renamed	integer	1=excellent 2=very good 3=good 4=fair 5=poor 6=refused 8=don’t know	how is your health	INH03PUB.PDF
race	not applicable	derived	integer	1=White 2=Black or African American 3=American Indian or Native American 4=Asian or Pacific Islander 5=Other 9=unknown/missing	race recoded as white; black/African American; American Indian; Asian/Pacific Islander; other; unknown/missing	NA

*status categories: unchanged, renamed, missing defined, derived

References

[1] Harris, K. M. (2013). The add health study: Design and accomplishments. *Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill*, 1-22.