CSDE 502 Winter 2020
Assignment 7
Introduction to Add Health
Instructor: Phil Hurvitz
phurvitz@uw.edu

Due Date: 2021-02-25 09:00

Explanation:

The main purpose of this assignment is for you to get practice working with variables and creating a completely reproducible work flow, and being able to share that work flow with a collaborator. Although the source data set for this assignment is from a publicly accessible source consisting of non-identifiable data, a similar work flow could be generated using sensitive data shared by a small group of collaborators on a shared drive.

For this assignment, create an R Markdown file that will generate an HTML document that relies on no resources from your local computer. The best way to assure this will be to create the R Markdown file on one computer and then render it on another computer. For example, you could create the R Markdown file on your local computer, then copy the file to the CSDE terminal server and render it there, or *vice versa*.

The idea is that you should be able to pass the code to any number of people with a working copy of R (hopefully at the same version or more recent than the one you are developing on), and with the exception of needing to install any necessary packages, everyone should be able to render the HTML and create the same output document (other than perhaps the creation date).

Start by creating a basic Rmd file and pushing it to your github.com repository. As you create any meaningful edits, commit and push changes.

Focus on these specific guiding principles:

- Every document you create, unless specifically requested not to, should identify the author and the date of creation or modification.
- Look at the document with "fresh eyes." Imagine that the document was written by someone else and you are the recipient. What would you like to see and read? Or imagine this is a document from you to your committee. What would you like them to see and read?
- The document should be structured around a narrative, rather than simply being a collection of R code and outputs.
- The HTML should not contain any warnings or messages that detract from the document's readability. See code chunk options (https://rmarkdown.rstudio.com/lesson-3.html) if you do not understand this instruction.
- Any tables or figures should be captioned with automatic numbering and properly cross-referenced.
    - Tables including numeric values should have an appropriate number of significant digits past the decimal point.

    - Tables should be nicely formatted for readability.
    - Figures should be scaled so that they are clearly readable.
        - Pay attention to font sizes, appropriate rotation of axis labels, crowding of text.
        - If you use `ggplot()`, this may make things easier.
- Any inline values in the narrative should be derived from `r code` statements rather than hard-coded values.
- Any bibliographic references and citations should be generated automatically (although including them is optional).
- Complete source code should be included as an appendix. Do not just include the R code—if you have included inline expressions, those will not show up in the rendered HTML without the Rmd code itself.

Instructions:

The workflow for this assignment should include the following tasks, with an explanation of each task's intention and result in the narrative:

1. Generate a new data frame from the full public Add Health data set (http://staff.washington.edu/phurvitz/csde502_winter_2021/data/21600-0001-Data.dta.zip) that consists of a subset of at least four of the original columns. ***Do not use any of the variables used as examples in Lesson 7***.
    a. At least one of the variables should be able to stratify the respondents into meaningful analytic groups.
    b. The data frame should have a "label" attribute that provides a brief but informative description of the table.
    c. The columns should be formatted as factor variables with proper value labels and ordering if applicable.
    d. The columns should have informative attributes as you see fit.
2. The code should save the data frame as an RDS file in the $TEMP location, which can be specified in R as `Sys.getenv("TEMP")`.
3. Create some frequency tables:
    a. Create a frequency table from each variable, using both counts and percentages.
    b. Create a frequency table based on at least two variables, also with counts and percentages.
4. Create at least one graph from data in the data frame.

Deliverables:

Include the URL to your Rmd file on github.com. Anyone should be able to download the Rmd file and render an HTML file from your Rmd file.

The URL to my Rmd file is: `<paste in your URL here>`_________________________

Convert this Word document to a PDF file, and then create a zip file including the PDF file. Upload the zip file to the course's Canvas site (https://canvas.uw.edu/courses/1434040), Assignment 7. Only upload a single zip file containing this document. Do not include in the zip file your Rmd source or HTML output—the Rmd needs to come from github.com and the HTML needs to be generated by running the Rmd code.