|  |  |
| --- | --- |
| **CSDE 502 2021**  **Assignment 2**  ***Data structures and tidyverse***  **Instructor: Phil Hurvitz**  [**phurvitz@uw.edu**](mailto:phurvitz@uw.edu) | **My Name:**  **My UWNetID:** |

**Due Date: 2021-01-21 09:00**

**Instructions:**

1. Fill out your name and UWNetID at the top of this page.
2. Add your answers to this document.
3. Use the "00Answers" Word style for your answers so they will be clearly discernible from the questions.
4. When you are completed with your work, create a PDF from the Word document.
5. Name your PDF document with the pattern UWNetID\_HW\_n.pdf where UWNetID is your UWNetID and n is the week number when the homework was assigned. For example, the first assignment by me would be named *phurvitz\_HW\_01.pdf*. Upload your completed document to Canvas.
6. Upload any specified data or code files to Canvas. You should be uploading this document (completed) and an HTML file (output from a rendered Rmd file).

**Explanation:**

The questions here are to help you understand R data structures, to explore some of the tidyverse functions, and to practice editing R Markdown for generating informative outputs.

Feel free to use any reference materials at your disposal to answer these questions (i.e., do not consider the course materials a complete set of reference materials).

**Questions:**

1. Is as.logical(NA) a vector? What is its length?
2. Is as.logical(NULL) a vector? What is its length?
3. The following code creates an object whose value is NA:

x <- NA

* 1. What type of data structure is x?
  2. What data type is x?
  3. How would you create an NA object with data type integer?

1. Explain how a data frame in R is a type of R list.
2. Use the R Markdown file generated at the end of last week’s class session as an example/template for presenting the following results. The output HTML file should contain all of the code that generated the output as well as the requested output. Use tidyverse functions as much as possible to generate readable code. Include commentary as necessary for interpretation of your methods and/or results.

Using the built-in iris data frame:

* 1. Present three tables that have all of the original variables, one for each species of iris.
  2. Present a table with only sepal length and sepal width for spp. *virginica* that also has columns indicating whether the sepal length and sepal width are greater than the mean for this species.

Use the general health of respondents, question “S3Q1 GENERAL HEALTH-W1” from the Add Health table we used in class:

* 1. Create a new variable that classifies health, stratified at the break between “(2) Very good” and better versus “(3) Good” and worse. Make sure you explicitly handle missing or unknown values.
  2. Tabulate (count and percent) this new variable for all respondents.
  3. Tabulate (count and percent) this new variable for those who self-identified as White versus those who self-identified as African American.