# R warmup assignment

The examples in this assignment are meant to cover the basics of R syntax by comparing it to what you know/expect from your primary statistical software.

## Instructions

Make sure you have installed R. You may wish to install RStudio as well. If it’s been awhile, you may want (or need) to update what version you’re using.

You will use the “Sample Dataset 2019” dataset, which is a variation of the “Sample Dataset 2014” data used in most of the SAS and SPSS tutorials. You can download the files from [libguides](https://libguides.library.kent.edu/workshop-intro-r):

* CSV: [Sample Dataset 2019.csv](https://libguides.library.kent.edu/ld.php?content_id=60015812)
* SPSS: [Sample Dataset 2019.sav](https://libguides.library.kent.edu/ld.php?content_id=64412583)
* Excel: [Sample Dataset 2019.xlsx](https://libguides.library.kent.edu/ld.php?content_id=68263783)

Except where indicated, use base R functions to solve the following problems. You can install packages using the command install.packages("packagename"). If you encounter failures while attempting to install, load, or use an R package, and you have a relatively newer Windows laptop with OneDrive enabled, you may need to go through an extra configuration step to get your R packages to work as expected.

**Do not use attach() to load/refer to a dataframe.** (This is an all-the-time recommendation, not just for this exercise.)

## Questions

1. What version of R are you using?
   1. Run the function sessionInfo() in the console.
   2. Run the function Sys.info() in the console.
   3. Run the function devtools::session\_info() in the console (you may need to install package devtools).
2. Import CSV or TXT data using read.csv() or read.table(), respectively.
   1. After importing the data frame, run it through the functions str() and summary(). What are these showing you?
      1. *Caution: The answer to this question will be different if you have installed R version 4.0 or later. Version 4.0 changed the base R read functions’ default behavior with respect to importing character variables; specifically, it is a change to related to the stringsAsFactors argument. (*[*Developer blog post about this change*](https://developer.r-project.org/Blog/public/2020/02/16/stringsasfactors/)*)*
   2. How many missing values are there for variable State?
3. Import CSV or TXT data using readr::read\_csv. (Note: leave stringsAsFactors=TRUE)
   1. If you have not already installed the readr package, do that first.
   2. Load the readr package. Which function did you use to do this? *(Note: There is more than one answer that will work, but stylistically there is an answer that is preferred.)*
   3. After importing the data frame, run it through the functions str() and summary(). What are these showing you?
   4. How do the results of str() and summary() compare to the data frame you imported with read.csv?
   5. How many missing values are there for variable State?
4. Import the SPSS \*.sav data file using haven::read\_spss().
   1. If you have not already installed the haven package, do that first. *(Note: Package haven contains functions for importing SAS, SPSS, and Stata data files into R. Notably, it can import Stata \*.dta files created using Stata 13 and later.)*
   2. After importing the data frame, run it through the functions str() and summary(). What are these showing you?
   3. How do the results of str() and summary() compare to the data frame you imported with read.csv and readr::read\_csv()?
5. Import the Excel \*.xlsx data file using readxl::read\_excel().
   1. If you have not already installed the haven package, do that first. *(Note: Package readxl contains functions for reading xls and xlsx files into R, but is not capable of writing data to Excel file formats. While there are R packages capable of writing to Excel files, these packages typically require Java. The process of installing and configuring Java can be especially difficult on Mac laptops.)*
   2. After importing the data frame, run it through the functions str() and summary(). What are these showing you?
   3. How do the results of str() and summary() compare to the data frame you imported with read.csv, readr::read\_csv(), and haven::read\_spss()?

*For the rest of the questions, use the dataframe created using read.csv().*

1. Create a frequency table for variable State.
   1. Create a frequency table that explicitly shows the number of missing values.
   2. How would you add marginal frequencies (i.e., row/column totals) to the table?
   3. How would you display the proportions for this table? (Note: you do not need to have the frequencies and proportions printed side-by-side to answer this question)
2. Create a frequency table for variable Rank.
   1. Create a frequency table for the variable Rank. Notice that it prints the data values 1, 2, 3, 4.
   2. Like SAS, SPSS, and Stata, R has the ability to assign labels to numeric values (R treats these as a special type(s) of variable) and treat the resulting variables as nominal or ordinal. What base R function(s) do this? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   3. After finding and applying the base R function(s) to assign labels to the numeric values, rerun the frequency table.
3. What proportion of students are overweight (BMI > 25)?
   1. Compute BMI from weight (in pounds) and height (in inches) as a new variable in your dataframe.
   2. Compute a new variable categorizing the numeric BMI values into ranges (can use the CDC guidelines to determine cutpoints) as a new variable in your dataframe.
   3. Tabulate the new variable from step B and get proportions.
4. Obtain descriptive statistics (mean/sd/confidence intervals, possibly by groups, and Pearson correlations; OR frequency tables/crosstabs/proportions) for the following sets/pairs of variables. (Note: The appropriate responses will not be ones that have all NA values)
   1. Height and Weight
   2. English, Reading, Math, Science
   3. Mile and Athlete Y/N
   4. State and LiveOnCampus
   5. LiveOnCampus, HowCommute, CommuteTime
5. What is the Cronbach’s alpha for the extraversion scale? (*Note: There isn’t a Cronbach’s alpha function in base R, as far as I know; there is one in package* ***psych***)
   1. Check if any of the items need to be reverse-coded before doing this.
   2. After obtaining the Cronbach’s alpha value, compute a new variable containing the subjects’ composite scores by averaging their responses to the Likert items. *Note: Simply using the mean() function alone will not achieve this.*
6. If you do a linear regression of SleepTime (Y) on StudyTime (X), are the assumptions of linear regression met?
   1. What is the base R function to run a linear regression?
   2. How can you print the “classic” table of regression parameter estimates, t-statistics, p-values from this linear regression?
   3. How can you plot the residuals from this linear regression?
   4. Re-do the linear regression after mean-centering the independent variable.
7. If you do a linear regression of SleepTime (Y) on StudyTime and Rank (X), are the assumptions of linear regression met?
   1. Run the model with Rank as a numeric variable. What terms are in the model?
   2. Run the model with Rank as an “ordered” variable. What terms are in the model?
   3. Run the model with Rank as a “factor” variable. What terms are in the model?
   4. What are the differences between these 3 approaches?
   5. Re-do the regression after group-mean centering variable StudyTime about the means of the class rank groups.
8. (Optional) Answer these questions in an RMarkdown or Quarto document, so that the questions, code, output, and write-up/interpretation are all in one place.