R Programming for Research

Colorado State University, ERHS 535

Brooke Anderson, Rachel Severson, and Nicholas Good

2019-08-26

Contents

Online course book, ERHS 535 **5**

Course information **1**
0.1 Course overview 1
0.2 Time and place 1
0.3 Detailed schedule 1
0.4 Grading 2
0.5 Course set-up 6
0.6 Coursebook 6

A Appendix A: Vocabulary **9**
A.1 Quiz 1—R Preliminaries (Updated for 2018) 9
A.2 Quiz 2—Entering / cleaning data #1 (Updated for 2018) 10
A.3 Quiz 3 (Updated for 2018) 11
A.4 Quiz 4 (Updated for 2018) 12

B Appendix B: Homework **13**
B.1 Homework #1 13
B.2 Homework #2 16

Online course book, ERHS 535

This is the online book for Colorado State University's *R Programming for Research* courses (ERHS 535, ERHS 581A3, and ERHS 581A4).

This book includes course information, course notes, links to download pdfs of lecture slides, in-course exercises, homework assignments, and vocabulary lists for quizzes for this course.

> ""Give someone a program, you frustrate them for a day; teach them how to program, you frustrate them for a lifetime."—David Leinweber"

Course information

Download a pdf of the lecture slides covering this topic.

0.1 Course overview

This document provides the course notes for Colorado State University's **R Programming for Research** courses (ERHS 535, ERHS 581A3, and ERHS 581A4). The courses offer in-depth instruction on data collection, data management, programming, and visualization, using data examples relevant to data-intensive research.

0.2 Time and place

Students for ERHS 535, ERHS 581A3, and ERHS 581A4 will meet together. Students in ERHS 535 will meet for the entire semester, completing a three-credit course. Students in ERHS 581A3 will meet for the first five weeks of the semester, completing a one-credit course. Students in ERHS 581A4 will meet from the sixth week to the final week of the semester, completing a two-credit course.

For the first five weeks of class, the course meets in the first-floor classroom of the Military Sciences building on Mondays and Wednesdays, 10:00 am–11:50 am. For the remaining weeks, the course meets in Room 120 of the Environmental Health Building on Mondays and Wednesdays, 10:00 am–12:00 pm.

Exceptions to these meeting times are:

- There will be no meeting on Labor Day (Monday, Sept. 2).
- There are no course meetings the week of Thanksgiving (week of Nov. 25).
- Office hours will be 10:00–11:00 AM on Fridays in EH 120.

0.3 Detailed schedule

Here is a more detailed view of the schedule for this course for Fall 2019:

Week	Class dates	Level	Lecture content	Graded items
1	Aug. 26, 28	Preliminary	R Preliminaries	
2	Sept. 4	Basic	Entering and cleaning data	Quiz (W)
3	Sept. 9, 11	Basic	Exploring data	Quiz (W), HW #1 (F)
4	Sept. 16, 18	Basic	Reporting data results	Quiz (W)
5	Sept. 23, 25	Basic	Reproducible Research	Quiz (W), HW #2 (F)
6	Sept. 30, Oct. 2	Intermediate	Entering and cleaning data	Quiz (W)
7	Oct. 7, 9	Intermediate	Exploring data	Quiz (W)
8	Oct. 14, 16	Intermediate	Reporting data results	Quiz (W), HW #3 (W)
9	Oct. 21, 23	Intermediate	Reproducible Research	Quiz (W)
10	Oct. 28, 30	Advanced	Entering and cleaning data	Quiz (W), HW #4 (F)
11	Nov. 4, 6	Advanced	Exploring data	
12	Nov. 11, 13	Advanced	Exploring data (mapping)	HW #5 (F)
13	Nov. 18, 20	Advanced	Reporting data results	
14	Dec. 2, 4	Advanced	Reproducible Research	HW #6 (F)
15	Dec. 9, 11	Advanced	Continuing in R	Project draft (M)
16	Week of Dec. 16		Group presentations	Final project

Students in ERHS 581A3 will be in weeks 1–5 of this schedule. Students in ERHS 581A4 will be in weeks 6–16 of this schedule.

0.4 Grading

0.4.1 Grading for ERHS 535

For ERHS 535, course grades will be determined by the following five components:

Assessment component	Percent of grade
Final group project	30
Weekly in-class quizzes, weeks 2-10	25
Six homework assignments	25
Attendance and class participation	10
Weekly in-course group exercises	10

0.4.2 Grading for ERHS 581A3

For ERHS 581A3, course grades will be determined by the following four components:

Assessment component	Percent of grade
Weekly in-class quizzes, weeks 2-5	40
Two homework assignments	30
Attendance and class participation	10
Weekly in-course group exercises	20

0.4.3 Grading for ERHS 581A4

For ERHS 581A4, course grades will be determined by the following five components:

Assessment component	Percent of grade
Final group project	30
Weekly in-class quizzes, weeks 1–5 (weeks 6–10 of the semester)	25
Four homework assignments	30
Attendance and class participation	5
Weekly in-course group exercises	10

0.4.4 Attendance and class participation

Because so much of the learning for this class is through interactive work in class, it is critical that you come to class.

If you are in **ERHS 535**, out of a possible 10 points for class attendance, you will get:

- **10 points** if you miss two or fewer classes
- **8 points** if you miss three classes
- **6 points** if you miss four classes
- **4 points** if you miss five classes
- **2 points** if you miss six classes
- **0 points** if you miss seven or more classes

If you are in **ERHS 581A3** or **ERHS581A4**, out of a possible 10 points for class attendance, you will get:

- **10 points** if you miss one or fewer classes
- **8 points** if you miss two classes
- **6 points** if you miss three classes
- **4 points** if you miss four classes
- **2 points** if you miss five classes
- **0 points** if you miss six or more classes

Exceptions:

- Attendance on the first day of class (Aug. 26) will not be counted.
- If you miss classes for "University-sanctioned" activities. These can include attending a conference, travel to collect data for your dissertation), For these absences, you must provide a signed letter from your research adviser. For more details, see CSU's Academic Policies on Course Attendance.
- If you have to miss class for a serious medical issue (e.g., operation, sickness severe enough to require a doctor's visit), the absence will be excused if you bring in a note from a doctor of other medical professional giving the date you missed and that it was for a serious medical issue.

For an absence to be excused, you must email me a copy of the letter by 5:00 pm the Friday afternoon of the week of the class you missed.

0.4.5 Weekly in-course group exercises

Part of each class will be spent doing in-course group exercises. As long as you are in class and participate in these exercises, you will get full credit for this component.

If you miss a class, to get credit towards this component of your grade, you will need to turn a few paragraphs describing what was covered in the exercise and what you learned. To get credit for this, you must submit it to me by email by 5:00 pm the Friday afternoon of the week of the class you missed.

All in-class exercises are included in the online course book at the end of the chapter on the associated material.

0.4.6 In-class quizzes

There will be weekly in-course quizzes for weeks 2–10 of the course. Students in ERHS 535 will take all these quizzes. Students in ERHS 581A3 will take quizzes in weeks 2–5. Students in ERHS 581A4 will take quizzes in weeks 6–10.

Each quiz will have at least 10 questions. Typically, a quiz will have more questions, usually 12–15 questions. The grading of the quizzes is structures so that you can get full credit for the quiz portion of the grade without getting 100% of quiz questions right. Instead, if you get ten questions right per quiz on average, you will get full credit for the quiz portion of the grade.

Once you reach the maximum possible points on quizzes, you can continue to take the quizzes for practice, or you can choose to skip any following quizzes.

Quiz questions will be multiple choice, matching, or very short answers. The "Vocabulary" appendix of our online book has the list of material for which you will be responsible for this quiz. Most of the functions and concepts will have been covered in class, but some may not. You are responsible for going through the list and, if there are things you don't know or remember from class, learning them. To do this, you can use help functions in R, Google, StackOverflow, books on R, ask a friend, and any other resource you can find. The final version of the Vocabulary list you will be responsible will be posted by the Wednesday evening before each quiz. In general, using R frequently in your research or other coursework will help you to prepare and do well on these quizzes.

Except in very unusual situations, the only time you will be able to make up a quiz is during office hours of the same week when you missed the quiz. Note that you can still get full credit on your total possible quiz points if you miss a class, but it means you will have to work harder and get more questions right for days you are in class.

0.4.6.1 Quiz grade calculations for ERHS 535

For students in ERHS 581A3, the **nine quizzes** in weeks 2–10 count for **25 points** of the final grade. The final quiz total for students in ERHS 535 will be calculated as:

$$\text{Quiz grade} = 25 * \frac{\text{Number of correct quiz answers}}{90}$$

0.4.6.2 Quiz grade calculations for ERHS 581A3

For students in ERHS 581A3, the **four quizzes** in weeks 2–5 count for **40 points** of the final grade. The final quiz total for students in ERHS 581A3 will be calculated as:

$$\text{Quiz grade} = 40 * \frac{\text{Number of correct quiz answers}}{40}$$

0.4.6.3 Quiz grade calculations for ERHS 581A4

For students in ERHS 581A4, the **five quizzes** in weeks 6–10 count for **25 points** of the final grade. The final quiz total for students in ERHS 581A3 will be calculated as:

$$\text{Quiz grade} = 25 * \frac{\text{Number of correct quiz answers}}{50}$$

0.4.7 Homework

There will be homework assignments due every two to three weeks during the course, starting the third week of the course (see the detailed schedule in the online course book for exact due dates).

The first two homeworks (HWs #1 and # 2) should be done individually. For later homeworks, you may be given the option to work in small groups of approximately three students.

Homeworks will be graded for correctness, but some partial credit will be given for questions you try but fail to answer correctly. Some of the exercises will not have "correct" answers, but instead will be graded on completeness.

For later homeworks, a subset of the full set of questions will be selected for which I will do a detailed grading of the code itself, with substantial feedback on coding. All other questions in the homework will be graded for completeness and based on the final answer produced.

Homework is due to me by email by 5:00 pm on the Friday it is due. Your grade will be reduced by 10 points for each day it is late, and will receive no credit if it is late by over a week.

0.4.8 Final group project

For the final project, you will work in small groups (3–4 people) on an R programming challenge. The final grade will be based on the resulting R software, as well as on a short group presentation and written report describing your work. You will be given a lot of in-class time during the last third of the semester to work with your group on this project, and you will also need to spend some time working outside of class to complete the project. More details on this project will be provided later in the semester.

0.5 Course set-up

Please download and install the latest version of R and RStudio (Desktop version, Open Source edition) installed. Both are free for anyone to download.

Students in ERHS 535 and ERHS 581A4 will also need to download and install a version of LaTeX (MikTeX for Windows and MacTeX for Macs). They will also need to download and install git software and create a GitHub account.

Here are useful links for this set-up:

- R: https://cran.r-project.org
- RStudio: https://www.rstudio.com/products/rstudio/#Desktop
- Install MikTeX: https://miktex.org/ (only ERHS 535 / 581A4 with Windows)
- Install MacTeX: http://www.tug.org/mactex/ (only ERHS 535 / 581A4 with Macs)
- Install git: https://git-scm.com/downloads (only ERHS 535 / 581A4)
- Sign-up for a GitHub account: https://github.com (only ERHS 535 / 581A4)

0.6 Coursebook

This coursebook will serve as the only required textbook for this course. I am still in the process of editing and adding to this book, so content may change somewhat over the semester (particularly for later weeks, which is currently in a rawer draft than the beginning of the book). We typically cover about a chapter of the book each week of the course.

This coursebook includes:

- Links to the slides presented in class for each topic
- In-course exercises, typically including links to the data used in the exercise
- An appendix with homework assignments
- A list of vocabulary and concepts that should be mastered for each quiz

If you find any typos or bugs, or if you have any suggestions for how the book can be improved, feel free to post it on the book's GitHub Issues page.

This book was developed using Yihui Xie's wonderful bookdown framework. The book is built using code that combines R code, data, and text to create a book for which R code and examples can be re-executed every time the book is re-built, which helps identify bugs and broken code examples quickly. The online book is hosted using GitHub's free GitHub Pages. All material for this book is available and can be explored at the book's GitHub repository.

0.6.1 Other helpful books (not required)

The best book to supplement the coursebook and lectures for this course is R for Data Science, by Garrett Grolemund and Hadley Wickham. The entire book is freely available online through the same format at the coursebook. You can also purchase a paper version of the book (published by O'Reilly) through Amazon, Barnes & Noble, etc., for around $40.

This book is an excellent and up-to-date reference by some of the best R programmers in the world.

There are a number of other useful books available on general R programming, including:

- R for Dummies
- R Cookbook
- R Graphics Cookbook
- Roger Peng's Leanpub books
- Various books on bookdown.org

The R programming language is used extensively within certain fields, including statistics and bioinformatics. If you are using R for a specific type of analysis, you will be able to find many books with advice on using R for both general and specific statistical analysis, including many available in print or online through the CSU library.

Appendix A

Appendix A: Vocabulary

You will be responsible for knowing the following functions and vocabulary for the weekly quizzes.

A.1 Quiz 1—R Preliminaries (Updated for 2018)

- Grading policies for the course
- Course requirements / policies for in-class quizzes
- Open source software
- "free as in beer" versus "free as in speech"
- Difference between R and RStudio
- R packages
- CRAN
- Installing packages
- `install.packages()`
- Loading a package
- `library()`
- `package::function()` notation
- Types of package documentation: vignettes and helpfiles
- `vignette()`, option `package =`
- Accessing a function's helpfile using ?
- R objects and object names
- "gets arrow": <−
- = vs. <− for object assignment
- Rules and style guidelines for naming objects
- Tab completion
- `ls()`
- Vectors
- `c()`
- Two of the basic classes of vectors: character and numeric

- `class()`
- Square bracket indexing for vectors: `[...]`
- Dataframes
- `data_frame()`
- Square bracket indexing for dataframes: `[..., ...]`
- `read_csv`, option `skip =`
- `str()`
- `summary()`
- `dim()`
- `ncol()`
- `nrow()`
- Nate Silver
- FiveThirtyEight
- R session
- R scripts
- Working at the console versus working from an R script
- # comment character
- `NA` for missing values
- $ to get a column from a dataframe
- `paste()`, option `sep =`
- `paste0()`

A.2 Quiz 2—Entering / cleaning data #1 (Updated for 2018)

- What kinds of data can be read into R?
- delimited files (csv, tsv)
- fixed width files
- delimiter
- `read_delim`, options `delim =`, `skip =`, `n_max =`, `col_names =`
- `read_fwf`
- `read_csv`, options `skip =`, `n_max =`, `col_names =`
- `readxl` package and its `read_excel()` function
- `haven` package and its `read_sas()` function
- `NA`
- Computer directory structure
- working directory
- `getwd()`
- `list.files()`
- relative pathnames
- absolute pathnames
- shorthand for pathnames: `.`, `..`, `../data`, etc.
- Reading in data from either a local or online flat file
- `paste()`, option `sep =`

- `paste0()`
- How to read flat files of data that are online directly into R
- `dplyr` package
- `rename()`
- Why you might want to rename column names (e.g., uppercase, long, unusual characters)
- `select()`
- `slice()`
- `mutate()`
- `filter()`
- `arrange()`, including with `desc()`
- Main types of vector classes in R: character, numeric, factor, date, logical
- `lubridate` functions, include `ymd`, `ymd_hm`, `mdy`, `wday`, and `mday`
- %>%, advantages of piping
- Common logical operators in R (==, `!=`, `%in%`, `is.na()`, `&`, `|`)

A.3 Quiz 3 (Updated for 2018)

- `data()` (with and without the name of a dataset as an option)
- `library()` (with and without an argument in the parentheses)
- logical vectors, including running `sum` on a logical vector
- What the bang operator (`!`) does to a logical operator
- The tidyverse
- `range()`
- `min()`
- `max()`
- `mean()`
- `median()`
- `table()`
- `cor()`, both for two variables in a dataframe, and to get the correlation matrix for several variables in a dataframe
- `summary()`, as applied to: different classes of vectors (numeric, factor, logical) and dataframes
- What to do if you want to apply a summary statistic function to a vector with missing values (you do not need to know every option name for all the functions, just know that you would need to include an option like `na.rm=` or `use=`, and that you can use the help file for a function to figure out the option call for that function).
- The following about object-oriented programming: In R, it means that some functions, like `summary()`, will do different things depending on what type of object you call it on.
- `summarize()`
- Special functions to use with `summarize()`: `n()`, `n_distinct()`, `first()`, `last()`
- Using `group_by()` before using `summarize()`
- The three basic elements of a `ggplot` plot: data, aesthetics, and geoms

- `aes` function and common aesthetics, including `color`, `shape`, `x`, `y`, `alpha`, `size`, and `fill`
- Mapping an aesthetic to a column in the data versus setting it to a constant value
- Some common geoms: `geom_histogram`, `geom_points`, `geom_lines`, `geom_boxplot()`
- The difference between "statistical" geoms (e.g., `geom_boxplot`, `geom_smooth`) and "non-statistical" (e.g., `geom_point`, `geom_line`)
- Common additions to `ggplot` objects: `ggtitle`, `labs`, `xlim`, `ylim`, `expand_limits`

A.4 Quiz 4 (Updated for 2018)

- Guidelines for good graphics
- Data density / data-to-ink ratio
- Small multiples
- Edward Tufte
- Hadley Wickham
- Where to put the + in ggplot statements to avoid problems (ends of lines instead of starts of new lines)
- Can you save a ggplot object as an R object that you can reference later? If so, how would you add elements on to that object? How would you print it when you were ready to print the graph to your RStudio graphics window?
- `geom_hline()`, `geom_vline()`
- `geom_text()`
- `facet_grid()`, `facet_wrap()`
- `grid.arrange()` from the `gridExtra` package
- `ggthemes` package, including `theme_few()` and `theme_tufte()`
- Setting point color for `geom_point()` both as a constant (all points red) and as a way to show the level of a factor for each observation
- `size`, `alpha`, `color`
- Re-naming and re-ordering factors
- **Note:** If you read this and find and bring in an example of a "small multiples" graph (from a newspaper, a website, an academic paper), you can get one extra point on this quiz

Appendix B

Appendix B: Homework

This section provides the homework assignments for the course.

B.1 Homework #1

Due date: Sept. 11 by 5:00 pm

For this assignment, you will submit the assignment to me **by email** by the due date. You should include three files in your submission:

1. A Word document with seven paragraphs. Each paragraph should be headed with the name of one swirl lesson and the body of the paragraph should describe that lesson and what you learned from it.

For this homework assignment, you'll be working through a few swirl lessons that are relevant to the material we've covered so far. Swirl is a platform that helps you learn R **in** R—you can complete the lessons right in your R console.

Depending on your familiarity with R, you can either work through seven lessons of your choice in the `R Programming: The basics of programming in R` and `Getting and Cleaning Data` courses (suggested lessons are listed further below) (**Option #1**), or you can work through seven lessons of your choice taken from any number of swirl's available courses (**Option #2**) .

For each lesson completed, please write a few sentences that cover: 1. A summary of the topic(s) covered in that lesson, and 2. The most interesting thing that you learned from that lesson. Turn in a hardcopy of this (with your first and last name at the top) during class on the due date.

To begin, you'll first need to install the swirl package:

```
install.packages("swirl")
```

If you've never run `swirl()` before, you will be prompted to install a course. You can do that with the `install_course` function. For example, to install the `R Programming` course, you would run:

```
library(swirl)
install_course("R Programming")
```

Once you've installed a course, every time you enter the swirl environment with `swirl()`, `R Progamming` should show up as a course option to select. You can enter `R Programming` to start lessons in that course by typing the number in fromt of it when you run `swirl()`.

Once you have at least one course installed, you call the `swirl()` function to enter the interactive platform in RStudio. The console will take you through a few prompts: you'll give swirl a name to call you, and take a look at some commands that are useful in the swirl environment. Those commands are listed further below.

```
library(swirl)
swirl()
```

After calling `swirl()`, you may be prompted to clear your workspace variables by running `rm=(list=ls())`. Running this code will clear any variables you already have saved in your global environment. While swirl recommends that you do this, it's not necessary.

Some of these lessons complement online courses through Coursera, so sometimes you will be asked after you complete a lesson if you want to report your results to Coursera. You should select "No" for that option each time.

B.1.1 Option 1

For **Option 1** of this homework, you will need to work through seven of the 15 available lessons in the `R Programming` course. Here are some suggestions for particularly uesful lessons that you could choose (the lesson number within the course is in parentheses):

R Programming course:

- Basic Building Blocks (1)
- Sequences of Numbers (3)
- Vectors (4)
- Missing Values (5)
- Subsetting Vectors (6)

- Logic (8)
- Looking at Data (12)
- Dates and Times (14)

Getting and Cleaning Data course:

- Manipulating Data with dplyr (1)
- Grouping and Chaining with dplyr (2)
- Dates and Times with lubridate (4)

Each lesson should take at most 10–15 minutes, but some are much shorter. You can complete the lessons in any order you want, but you may find it easiest to start with the lowest-numbered lessons and work your way up, in the order we've listed the lessons here.

You'll be able to get started on some of these lessons after your first day in class ("Basic Building Blocks"", for example), but others cover topics that we'll get to in weeks 2 and 3. Whether or not we've covered a swirl topic in class, you should be able to successfully work through the lesson. At the end of each lesson, you may be asked if you would like to receive credit for completing this course on Coursera.org. Always choose"no" for this option.

Again, you'll need to compose and turn in a few sentences for each lesson. Make sure to include a summary of what each lesson was about, and the most interesting thing about that lesson.

B.1.2 Option 2

If you're already somewhat familiar with R, you might want to choose your seven lessons from other swirl courses instead of or in addition to those available in the `R Programming` and `Getting and Cleaning Data` courses.

Check out the list of available Swirl Courses to see which ones you would like to install and check out available lessons for. For example, to choose a lesson in the `Exploratory Data Analysis` course, you would run:

```
library(swirl)
install_course("Exploratory Data Analysis")
swirl()
```

After entering the `Exploratory Data Analysis` course, you could choose from any one of its available lessons.

In your written summary for each lesson (again, a few sentences that cover a summary of the lesson and the most interesting thing you learned), make sure to specify which course each lesson you completed was from.

B.1.3 Special swirl commands

In the swirl environment, knowing about the following commands will be helpful:

- Within each lesson, the prompt `...` indicates that you should hit Enter to move on to the next section.
- `skip()`: skip the current question.
- `play()`: temporarily exit swirl. It can be useful during a swirl lesson to play around in the R console to try things out.
- `nxt()`: regain swirl's attention after `play()`ing around in the console.
- `main()`: return to swirl's main menu.
- `bye()`: exit swirl. Swirl will save your progress if you exit in the middle of a lesson. You can also hit the Esc. key to exit. (To re-enter swirl, run `swirl()`. In a new R session you will have to first load the swirl library: `library(swirl)`.)

B.1.3.1 For fun

While they aren't required for class, you should consider trying out some other swirl lessons later in the course. You can look through the course directory to see what other courses and lessons are available. For the first part of our course, you might find the "Exploratory Data Analysis" course helpful. If you would like to learn more about using R for statistical analysis, you might find the "Regression Models" course helpful.

B.2 Homework #2

Due date: Sept. 27 by 5:00 p.m.

For this assignment, you will submit the assignment to me **by email** by the due date. You should include three files in your submission:

1. The final rendered Word document (rendered from an RMarkdown file).
2. The original RMarkdown file used to create that final document.
3. The dataset you used for the assignment.

For this assignment, start by picking a dataset either from your own research or something interesting available online (if you're struggling to find something, check out Five Thirty Eight's GitHub data repository). You will then use this dataset to practice what you've learned with R so far. This will also be a chance, if you're using a dataset from your own research for me to get an idea of how you might be planning to use R in future research.

Very important note: Some research datasets have privacy constraints. This includes any datasets collected from human subjects, but can also include other datasets. For some projects, the principal investigator may prefer to keep the data private until publication of results. Before using a dataset for this project, please confirm with your research advisor that there are no constraints on the data. I do not plan to make the results of this assignment public, but I also do not want to us to be emailing back and forth a dataset with any constraints.

Using your dataset, create an RMarkdown document with the content listed below. Each section I've listed below should be included in the RMarkdown document as its own section, with a section header created using Markdown code.

- **Section 1: Description of the data**: Describe the dataset you are using, both in terms of the **content** (what is this data measuring? how was it collected? what kinds of research questions are you hoping to use it to answer?) and in terms of its **format** (what type of file is it saved in? what if it is in a flat file, is it fixed width or delimited? if it is delimited, what is the delimiter? if it is binary, what is the program that would normally be used to open it?).
- **Section 2: Reading the data into R**: Include code that reads the data into R and assigns it to a dataframe object that you can use later in the document. Explain in the text which R function you used to read in the data (e.g., `read_csv`) and which package it came from (if it was not a base R function). If there were any special options you needed to use (e.g., `skip` to skip some rows without data), list those and explain why you used them. Next, include some code to clean the data (e.g., rename columns, convert any dates into a "Date" format). You can filter to certain rows if you would like, but do **not** filter out missing values, as we'll want to learn more about those later.
- **Section 3: Characteristics of the data**: Describe the dataframe you just read in. How many rows does it have? How many columns? (Use inline code in the RMarkdown document to put this information into a sentence that reads "This dataframe has … rows and … columns.".) What are the names of the columns? What does each row measure (i.e., what is the unit of observation)? Include a table (using Markdown directly or `kable`) explaining what each column measures. This table should have three columns: (1) the column name in the R dataframe;

(2) a very brief description of what each column measures; and (3) the units (if any) of the measurement in the column.

- **Section 4: Summary statistics**: Pick three columns of the dataframe. Use the `summarize` function to get the following summaries of these columns: (1) minimum value; (2) maximum value; (3) mean value; (4) number of missing values. If there are missing values, make sure you use the appropriate options in summarizing these values to exclude those when calculating the minimum, maximum, and mean. Assign the result of this `summarize` call to a new R object, and print it out, so these summaries show up in your final, rendered Word document.
- **Section 5: Visualizing the data**: Create two plots of your dataframe. One should use a "statistical" geom (e.g., histogram, bar chart, boxplot) and one a "non-statistical" geom (e.g., scatterplot, line plot for time series). Explain why these plots help you learn more about this data and about the interesting research questions you're hoping to explore with the data. Be sure to customize the final size of each plot in the Word document using the `fig.width` and `fig.height` commands. For each plot, also be sure to customize the x- and y-axis labels. Finally, explain how each plot is following at least two of the principles of "good graphics" covered in week 4 of the course (Chapter 4 of the book)—if necessary, use `ggplot` functions and options to make the plots comply with some of these principles.