# Summer R ’23 for the Physician Scientist Training Program

Dominic Bordelon, Research Data Librarian, University Library System, [dbordelon@pitt.edu](mailto:dbordelon@pitt.edu)

Location: Falk Library, 200 Scaife Hall, Library Classroom  
Wednesdays, 4-6pm, 6/7-7/5

## Learning goals

1. Transfer tabular data into and out of R
2. Examine and make sense of data (at the level of descriptive statistics)
3. Visualize data according to what kind of data you have, and what you want to show
4. Manipulate and restructure data so that it can be analyzed (also called “data cleaning”)
5. Write reproducible cleaning tasks, analyses, and plots for sharing with colleagues (and your future self)
6. Utilize R documentation to solve problems and explore possibilities
7. Gain an understanding of the R packages ecosystem

Note: We won’t be doing very much statistical analysis. Rather, I’m giving you the building blocks to confidently handle your data and *prepare* it for statistical analysis. A typical stats course will focus on the mathematical and probabilistic methods, and have you use data that are already well prepared. This mini-course will give you the foundational skills to explore data and manipulate it into an analyzable state.

## Schedule and topics studied

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **#** | **Topics** | **Homework (after class)** |
| 6/7 | 1 | RStudio orientation; using R for simple calculations; loading and saving data; data types and structures; R packages; R notebooks. | Problem set 1.5 |
| 6/14 | 2 | Working with data frames: descriptive statistics; filtering and sorting; selecting and ordering variables; grouping.  Data viz: base R plots (histogram, bar, scatter). | Problem set 2.5 |
| 6/21 | 3 | Data viz: ggplot2 and the grammar of graphics, common plots.  Rendering notebooks to different formats. | Problem set 3.5 |
| 6/28 | 4 | Handling missing data; adding new variables to a data frame; joining tables. Time permitting: working with dates; working with string data.  Data viz: missingness; faceting (small multiples). | Problem set 4.5 |
| 7/5 | 5 | Handling categorical variables (recoding and reordering); pivoting columns and rows; checking a variable for Normality; introductory inference.  Data viz: output options and image formats; annotations; Q-Q plots. |  |

## Expectations

* There are no prerequisites, except that I assume you’ve used Excel (only a little) and know what spreadsheets are. I don’t assume you have taken a statistics course. If you have not taken stats, you will probably learn a little bit of stats here. This mini-course will help your stats and vice versa.
* Prior to our first session, please install:
  + R: <https://cloud.r-project.org/>
  + RStudio Desktop: <https://posit.co/download/rstudio-desktop/>
* Please bring your laptop to each session. If you don’t have a laptop, let me know ASAP to see whether a loaner can be arranged.
* Please make every effort to attend all sessions.
* We will have homework, which I’ll send out on the day after each class. The purpose of this homework is to have you re-engage with the content we’re learning on days when we don’t meet, so that you refresh and strengthen your understanding. Homework is due at the start of next class via email attachment.
* We will use a Discord server for support and sharing. This is optional but recommended (if you have a Discord account). Discord is free at <https://discord.com/>. Our server is called “Research Computing Education at Pitt”; link will be included in the introductory email, or you can ask me for it.
* Participants are welcome to work collaboratively on homework and during class. Of course, you’ll want to make sure you personally understand how it’s all working, since you’ll need to use the skills on your own in the future.
* Please reach out to me via email ([dbordelon@pitt.edu](mailto:dbordelon@pitt.edu)), or on Discord, or book a zoom appointment [here](https://pitt.libcal.com/appointments/research_data_librarian) if you get stuck or have questions. I will also offer feedback cards at the end of each session, aiming to maximize the utility of our time together.
* We will be using realistic data examples from health science domains (e.g., patient data). However, I am not trained as a scientist or analyst in these areas, and we won’t be investigating research questions. Rather, we are learning to use a general tool (think Excel) with examples that will hopefully feel relevant to your studies. Once you are proficient with the tool using example data, you will be better equipped to start applying the tool to real data and questions, whether in coursework or under the guidance of your advisor.
* A note about computer programming: R offers a complete programming language with complex logic and information structures. Although we will write some code—necessary in R—it will be very simple from a programming perspective. We won’t be thinking about algorithms. Our approach to R focuses on *functional data work/tasks*—which are greatly facilitated by the packages we use—rather than on *computer programming* or on *probability/statistics* (alternative approaches).

## Packages we will use

This list is provided for your reference/curiosity. There is no need to install or learn these ahead of time; we will do so together. (This list may be incomplete.)

* tidyverse, which is a collection of packages, most notably:
  + dplyr
  + ggplot2
  + tidyr
  + readr
  + forcats
  + lubridate (potentially)
  + stringr (potentially)
* medicaldata
* naniar

## Resources/links for R

* [R for Data Science](https://r4ds.hadley.nz/) by Wickham, Çetinkaya-Rundel, and Grolemund: probably the most popular resource for learning R, and with good reason. Many of our topics also map easily onto sections of R4DS. Available free online.
* [Big Book of R](https://www.bigbookofr.com/) by Oscar Baruffa is a continually updated online directory of R books on various topics, searchable and organized by topic. Many of the books here are free ebooks (and prices are clearly marked for non-free items). Examples of topic areas are data science, life sciences, and time series analysis.
* [Package cheatsheets](https://posit.co/resources/cheatsheets/) from Posit help you quickly reference the functions in packages you use often. I will bring appropriate printouts to class.
* During Fall and Spring semesters, I run a weekly R & RStudio Drop-In Hour in Hillman Library, as well as periodic workshops. Check event listings at <https://library.pitt.edu/>, <https://calendar.pitt.edu/>, or in the [newsletter](https://confirmsubscription.com/h/t/DF8BAC3FE84F3E2D) from the Library’s Digital Scholarship unit. The [Health Sci Library’s Data Services](https://www.hsls.pitt.edu/data-services) also run many useful workshops ([newsletter](https://subs.hsls.pitt.edu/?p=subscribe&id=3)).