Week 1: Introduction

*Who are we? Why are we here? What does it all mean???*

Spoiler alert: I won’t answer that last question, at least not today. I won’t answer it primarily because I don’t know yet, but that’s the whole point, right?. We’re *learning.* By the end of this series, maybe together we can get close to discovering what it’s all about. At least when it comes to data science with R. You’re on your own for the rest of it.

So who are we? In the broadest sense, we’re R newbies! Within that, we all might be coming in with different experiences and skills. Some of you may be researchers, looking to supplement your data analysis with new tools. Others might be looking to pivot into data science as a career. Still others might just be curious. Some of you might have previous programming or data analysis experience, some might not. It’s all okay, everyone is welcome here! I’ll probably get into a bit more of my background as we go along, but for now, just know that I’m brand new to R just like you. My goal for this series is to create a kind of journal of my experiences learning R, with the idea that we might ask similar questions and have similar struggles. Think of this kind of like your study group, with the book we’re following as our wise teacher.

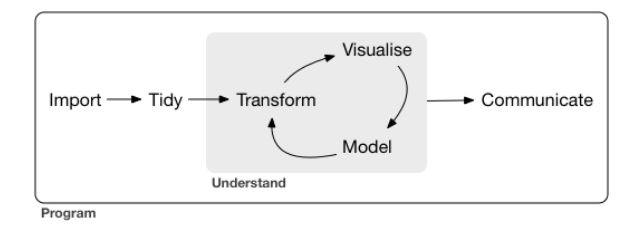
Which brings me to the all-important Details. The book we will be following for this series is called *R for Data Science* by Hadley Wickham and Garrett Grolemund. This book is 100% FREE (how often can you say that anymore?) and can be found at <https://r4ds.had.co.nz/>. We will be exploring the book together and I’ll be hitting the high points in my blog posts, but I will not be copying it word-for-word because 1) that would be wrong and 2) I don’t need to, because the book is FREE. Go check it out if you want to follow along. There is a physical copy you can order from amazon, but as the online version can be updated much more readily than a physical book I recommend you stick with the online version, so that we’ll all be working with the same material.

Ready to get started? Let’s do it!

*The What, How, and Why: Not Necessarily In That Order*

If you’re still here, you probably have some idea about why you might want to learn more about data science in general, and R in particular, but I’ll go into why *I’m* interested. I am a technician in a neuroscience research lab, and so I work with data all the time. For most of my years as a scientist, I never once thought about data science, even though I am a *scientist* who often works with *data*. Why? It really comes down to the fact that we (usually) work with small data. What do I mean? I mean that our experiments usually generate data that can relatively easily be recorded on a spreadsheet, manually cleaned up, copied and pasted into a statistics program, and the figures generated from there. However, while there might be some pros to this approach (you see the data at every step, it requires very minimal technical know-how), there are certainly cons as well. With all of that hands-on time comes many more opportunities for error, and as many of you may be aware, this manual process does not scale well at all.

Even with small data sets, many of the same principles and processes can apply. Have I convinced you? If not, that’s okay. As we go through we’ll be practicing on small data. Hopefully you’ll see that, even though it seems like a lot of work to put in to learn a different way to do a thing you already know how to do, the effort will save you time in the long run, with the added benefit of scalability.



<https://r4ds.had.co.nz/introduction.html>

I pulled the image above from the introduction to *R for Data Science* (I’ll abbreviate this going forward as *R4DS*). It’s a simple schematic for the processes involved in data science. It’s a nice way to think about the tasks and tools involved. I’ll go through each step briefly:

* **Import:** We’ve got to get our data into the program we’re using before we can manipulate it. For those of us used to using statistical programs, I imagine that this step corresponds most closely with throwing all of our data into a spreadsheet in whatever form we get it; it may be messy, but at least it’s all in one place.
* **Wrangle:** This is the umbrella term for the process of tidying and transforming our data.
  + **Tidy:** You may also hear “data tidying” referred to as “data cleaning.” The idea is that before we can do anything else our data need to follow a consistent set of rules. For “tabular” or “rectangular” data, they are as follows:

1. Each variable is one column
2. Each observation is a row.

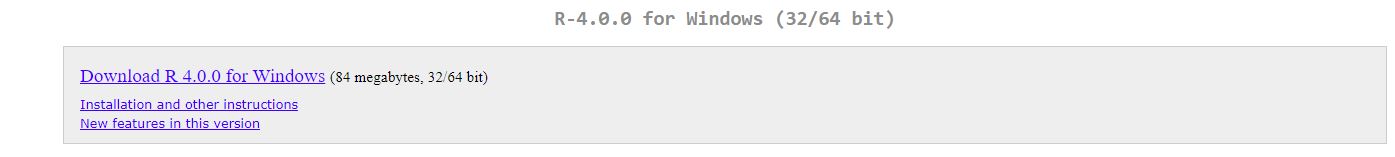
Making our data follow these rules early in the process will apparently save us time later.

* + **Transform:** This is the step where we will “zoom in” on the data we actually want to work with. This might include a number of tasks, such as narrowing the data to just the subset we’re interested in (looking at a specific timeframe or zip code, for example), calculating new variables based on our newly tidy data (the example used in *R4DS* is speed calculated from distance and time), and/or calculated summary statistics (mean, median, etc.). This is kind of like transferring data to a statistics program, where it has to be in a certain format and then it can run its calculations.
* **Visualization**: Once you have all of the nice tidy data that you want, one of the fun parts you can do is make a visualization- an actual visual representation of your data. This can take a lot of forms: bar charts, line graphs, word clouds… there are tons of options! Visualization is key for creating scientific figures (this is where visualization ion feeds into **communication**), but they are also invaluable for *understanding* your data and guiding your future questions.
* **Modeling**: In this step, you can use computation to answer the questions that are raised by the previous steps. As *R4DS* explains, although visualization and modeling are both used to understand the data, modeling scales better than visualization; computers are generally cheaper than people, believe it or not, and it isn’t dependent on humans for interpretation.
* **Communication**: Once you’ve got an understanding of your data, you want to share what you’ve learned, right?

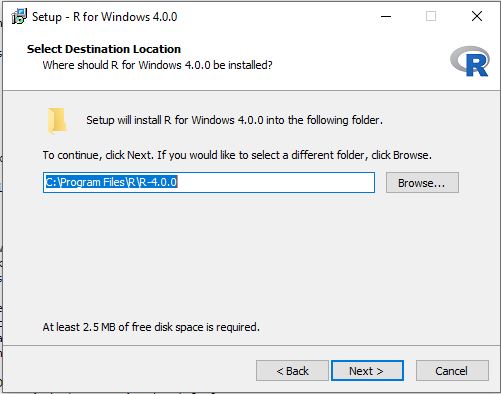
*R4DS* starts their instruction with data visualization. Why not start at the beginning? They claim that visualization is more fun and less frustrating than importing and wrangling data. I know when I think of my work with data that it is always cooler to see the results than to organize everything, so it isn’t surprising to me. But before we start anything, we need to do a few things: download R, download RStudio, and install a few packages. If you already have these installed, you can skip these steps (though take a look and make sure you have all of the packages we’ll be using).

Downloading and Installing R: A Simple Step-by-Step

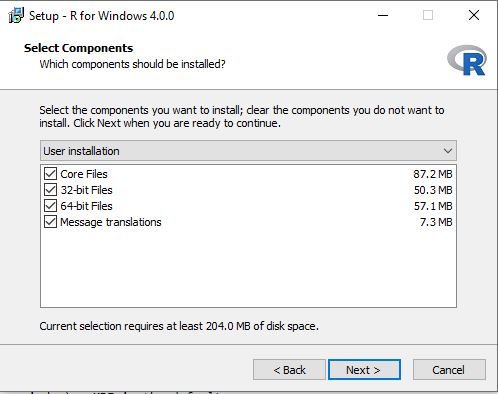
* Go to <https://cloud.r-project.org/>
* Link on the link for the download you want, depending on whether you have a Mac or a PC that runs Windows.
* If you have never installed R before, click the base link. If you have… you probably know better than I do what you should do.
* Hit the big download link at the top of the page!



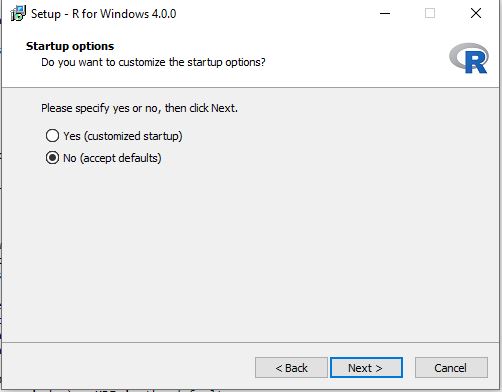
* Next you need to select a destination for the download… I just went with the default destination, but you can put it wherever you’d like.



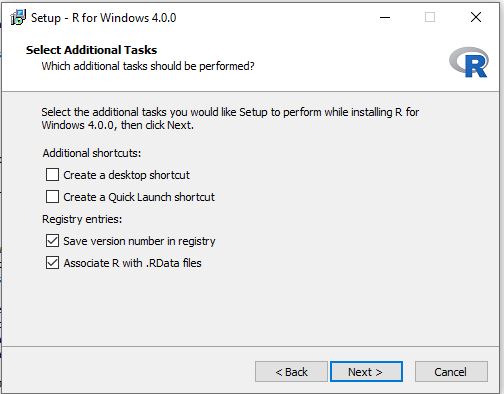
* Next you’ll select which components you want to be installed… Go with all of them so that you’re sure not to be missing anything. That’s what I did anyway.



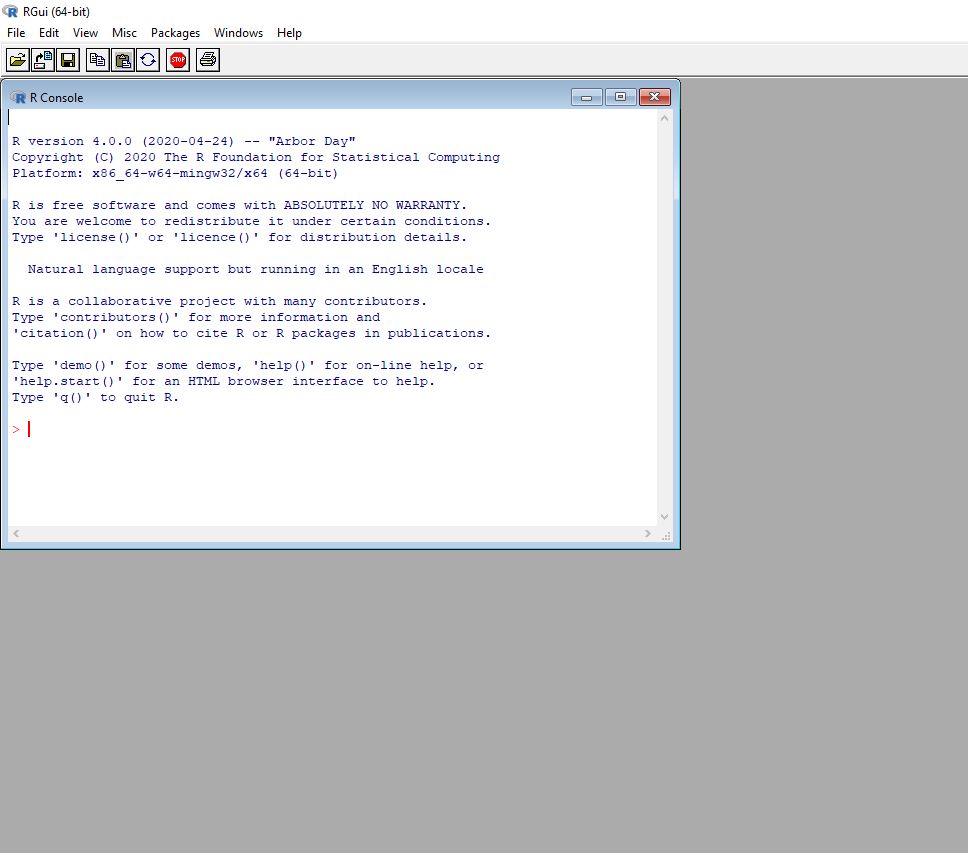
* You’ll be asked for startup options next. Just select No go with the defaults.



* Now it’ll ask you about naming your startup menu folder. Let’s keep it simple yes? The default is “R” and that sounded good to me. Sensing a pattern? You can also put it in another folder if you really want to.
* Next you’ll select additional tasks. I chose not to create desktop or Quick Launch shortcuts, but you can if you want to. I would recommend selecting the boxes for saving the version number and associating R with .RData files. Those things just sound important.

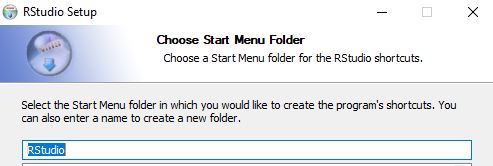


* You should see an installation progress bar, followed by a “Setup Complete” window of some kind. Hooray! If you decide to check and make sure everything is all good like I did, you can find your installation wherever you decided to save it and open it. If all is well, you should get a window that looks like this:



Downloading and Installing RStudio

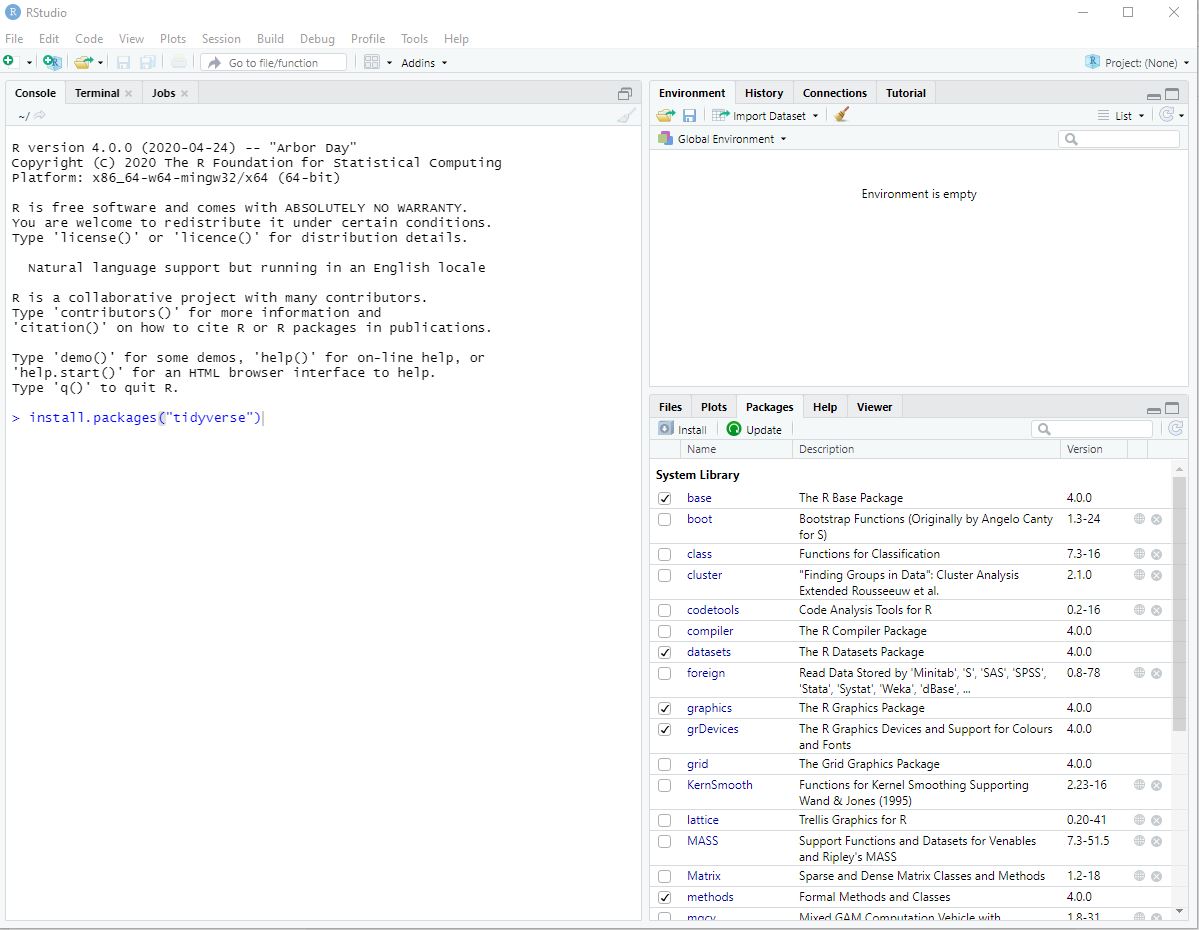
RStudio is the *integrated development environment* (or IDE) that we’ll be using to write R code. While an IDE is not necessary for coding, it’s nice because you’ll generally have some built-in tools that will help you along the way. They contain a text editor, debugging tools, and other neat stuff.

* Go to <http://www.rstudio.com/download>.
  + Note: Some of you, especially if you’ve worked with Python at all, might have Anaconda installed and have seen that you can install R and RStudio there.I don’t recommend it… mostly because I couldn’t get it to work. Let’s stick to the instructions in *R4DS*, shall we?
* Choose the version you want (might I suggest going with one of the FREE versions?). I went with the desktop version.
* Hit that blue download button!
* The website will recommend a download for your system. That’s the one you should probably just go with, but I would just double check that your system meets the requirements.
* You should see another Setup wizard. Hit “next.”
* Like with R you’ll have to select a destination folder for RStudio. Put it wherever, but I’d recommend putting it in the same location as R so that RStudio will be able to easily “find” R on your machine. If you just go with the defaults you should be good.
* As with R, you need to set up a start menu folder for RStudio. Default is good. 
* Once RStudio is finished installing, find it and open it- you’ll need it to install some packages.

Installing the Tidyverse and Other Fun Stuff

* The R installation we’ve just downloaded does not contain all of the functionality we will use as we work through *R4DS*. The tidyverse is actually a group of packages that all work together and, as *R4DS* describes it, share a “common philosophy” of working in R. It is easy to install, with our first line of code!

install.packages(“tidyverse”)



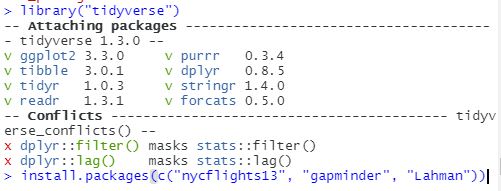
That’s it. Seriously. It may take a bit, so be patient. You will know it is finished when you see the file path and the blue “>”.



* Next we need to load the tidyverse, with another simple line of code:

library(tidyverse)

* Lastly, there are three more packages to install and load: nycflights13, gapminder, and Lahman. I have no idea what these packages do- we’ll find out together!



* Oh, one last thing. I got a warning when installing the tidyverse and the other packages that looked like this:



I don’t think it’s a problem at this point… but if it becomes an issue in the future we can go back to CRAN where we downloaded R and get Rtools if we need to.

That’s it, we’re all ready to get started! Next we’ll start making some plots and charts.