Introduction to Data Science

Welcome!

T. ARNOLD

Introduction to Data Science

Today we are going to get all of the administrative details dealt with. Here is a quick outline:

1. syllabus
2. course content
3. introductions
4. install course materials

There should be plenty of time for questions throughout the class.

1. Syllabus

2. Course Content

Data Science?

This semester we will learn and practice a series of methods for **organizing**, **collecting**, **visualizing**, **manipulating**, and **exploring** different kinds of data. We are focused on the creation and application of **methods**, rather than theoretical or foundational questions.

This is not a mathematics course, nor will it resemble a traditional introductory statistics class. We will spend the entire semester writing code to apply data science concepts.

Programming

There are several different programming languages for data science. By far the two most popular are R and Python.

We will be using R this semester but will learn a version that is easily adapted to other languages such as Python.

No specific experience with R or Python is expected!

An Example

If you are not familiar with the kinds of tasks that are common in data science and exploratory data analysis, an example can be very helpful. Here is a slightly dated but still the best concise example I know of:

https://www.youtube.com/watch?v=Z8t4k0Q8e8Y

3. Introductions

About Me

- From New England: born in Maine, school in MA, ME, CT
- Moved to Richmond in 2016
- Research on large text and image datasets in linguistics and cultural studies

Distant Viewing Lab

NATIONAL ENDOWMENT FOR THE HUMANITIES

About Me

- Lots of industry experience in DS:
  - IBM (Healthcare)
  - Travelers (Insurance)
  - DARPA (social media)
  - AT&T (location analytics)
  - Telperian (pharmaceuticals)

About Me

I have a dog named Roux. He is often in my office; please come say hello!

You?

Now, it’s your turn to tell me a little about yourself:

https://forms.gle/vJsiikcnPoD6FTvdA

4. Course Setup

Installing Course Software

We need to install three different components for this semester:

1. The R Programming Language
2. The RStudio IDE
3. A set of R Packages and data

All of these components are open-source and available for all modern operating systems. You may have trouble, however, if you have an older OS and have not updated it recently.

Even if you already have R installed, I suggest doing a fresh update for the semester.

1. Installing the R Language

To install R, go to https://cran.r-project.org/ and select your operating system. Then:

macOS => click on R-4.2.1.pkg and follow instructions
Windows => click on **base** followed by "Download R"

For Linux, either install from source or use your favorite package manager.

You need to install R before anything else, but we will never actually open it directly. So feel free to remove and shortcuts or links that are created during installation.

2. Installing RStudio

To install RStudio, follow the following link and download either the dmg (macOS) or exe (Windows):

https://www.rstudio.com/products/rstudio/download/#download

Note: On macOS, you need to drag the RStudio icon into your Applications directory after downloading.

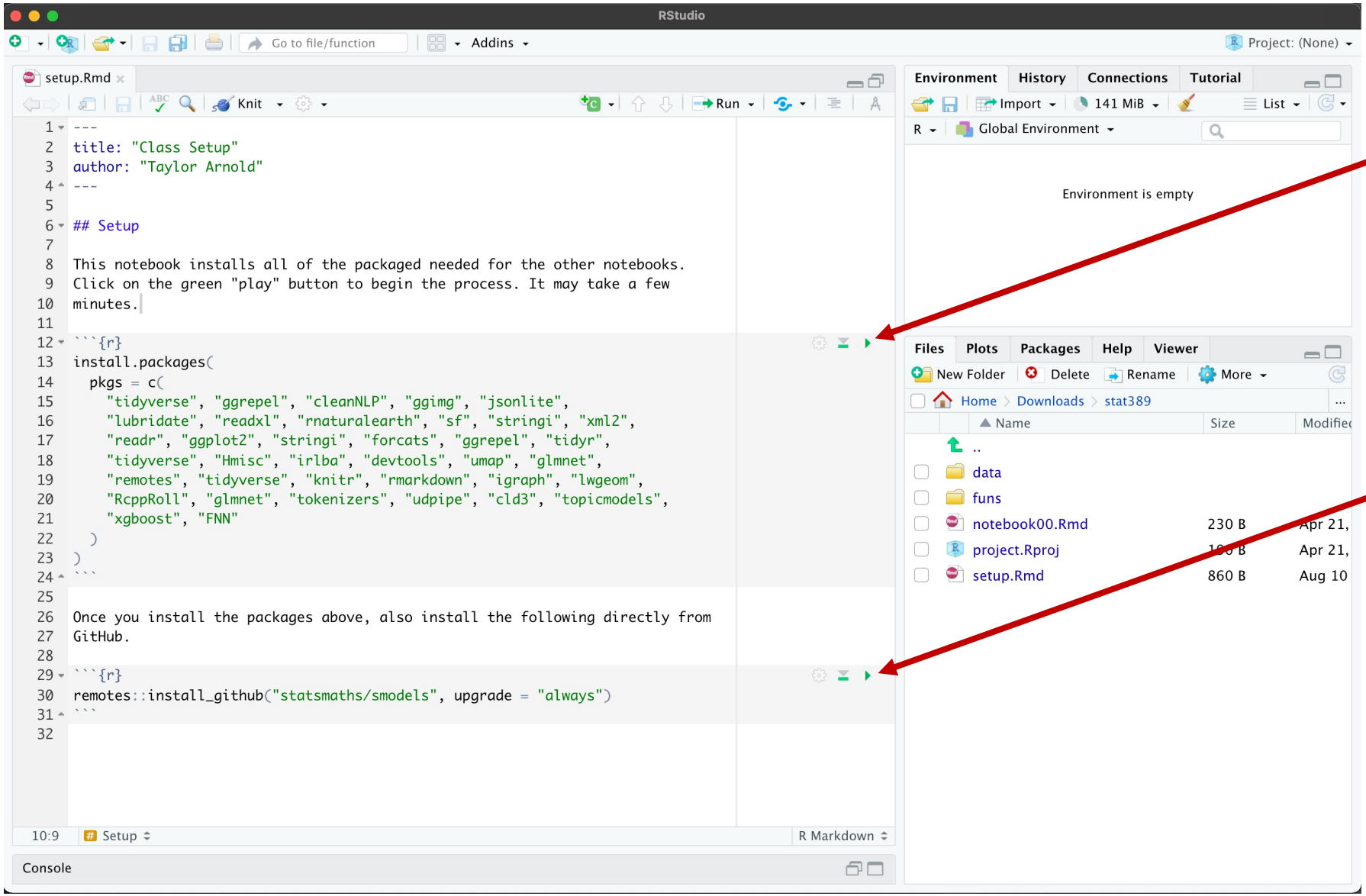
3. Install R Packages

Finally, download and unzip the "materials.zip" file from the class website. We will use this directory throughout the semester, so put it somewhere you will remember it.

Then, open the setup.Rmd file in RStudio, and click the green play buttons (see next slide).

Make sure to put this folder somewhere for the semester. You will need it!

3. Install R Packages



Click here

Then, click here

For the Semester

You should plan on bringing a laptop with a working version of R, RStudio, and all of the installed packages to each class meeting. If that is or becomes an issue, just let me know and we will find a solution.

Note that if you are having computer issues, particularly during an exam, it is always possible to use the lab computers in Jepson as a back-up. They have R and RStudio installed, but not all of the class R packages. Simply start from Step 3 in these notes before getting started.