Using Docker Containers in Teaching Statistics and Data Science: How NCShare Can Help

June 13, 215 - 3 pm

Mark McCahill, Senior System Architect, Duke OIT

Elijah Meyer, Postdoctoral Fellow, Duke StatSci …soon to be Elijah Meyer, Assistant Teaching Professor at NC State)

Joan Combs Durso, Undergrad Coordinator, Duke StatSci

Marie Vernet, Scripps Institute, on Adelie Penguins in the Western Antarctic Peninsula:

One of the major hypotheses of the Palmer LTER program is that with the warming of the Peninsula, the amount and extent of winter sea ice has been decreasing, resulting in changes in the breeding success and overall population sizes of the resident penguin species. The penguin species most affected by these changes is the Adélie Penguin, appropriately referred to as the "bellwether of climate change."

How far could you and your students go with access to advanced science capabilities and software environments from day one in your courses?

RStudio Containers for Teaching Statistics & Data Science

Why run RStudio via a web browser?

Problem:

You want to teach Statistics and Data Science

Debugging students' attempted installs of RStudio on their laptops wastes time

You want all students to have a common experience with

- same hardware
- same software and packages installed
- same datasets

Solution: Provide RStudio to students via web browser !

RStudio via a web browser

How does this work?

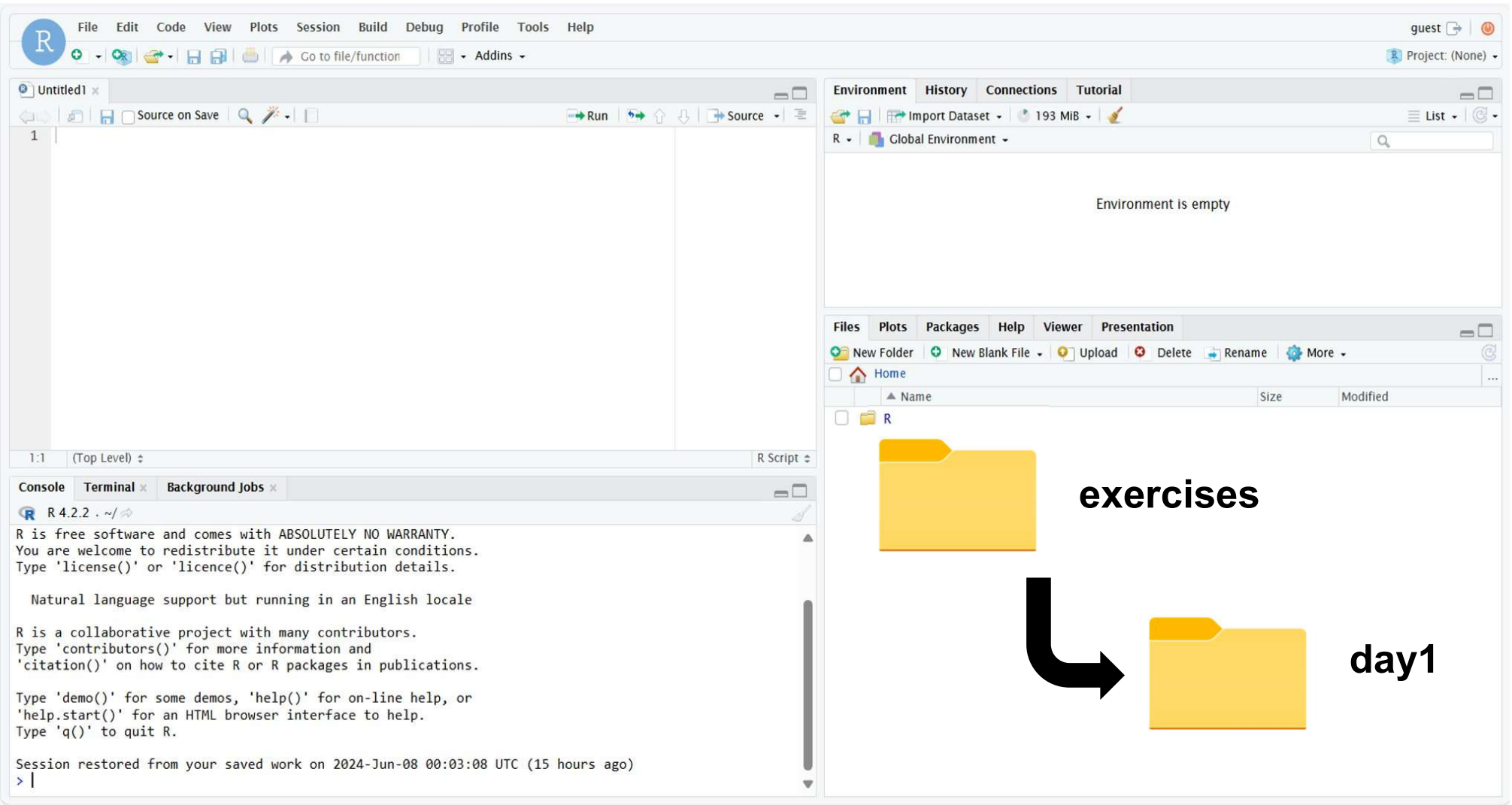
Provide students with course-specific personal instances of RStudio run on a common web server

We package RStudio in a Docker container - a standardized method of creating and automating software environments that can be reproducibly deployed at scale

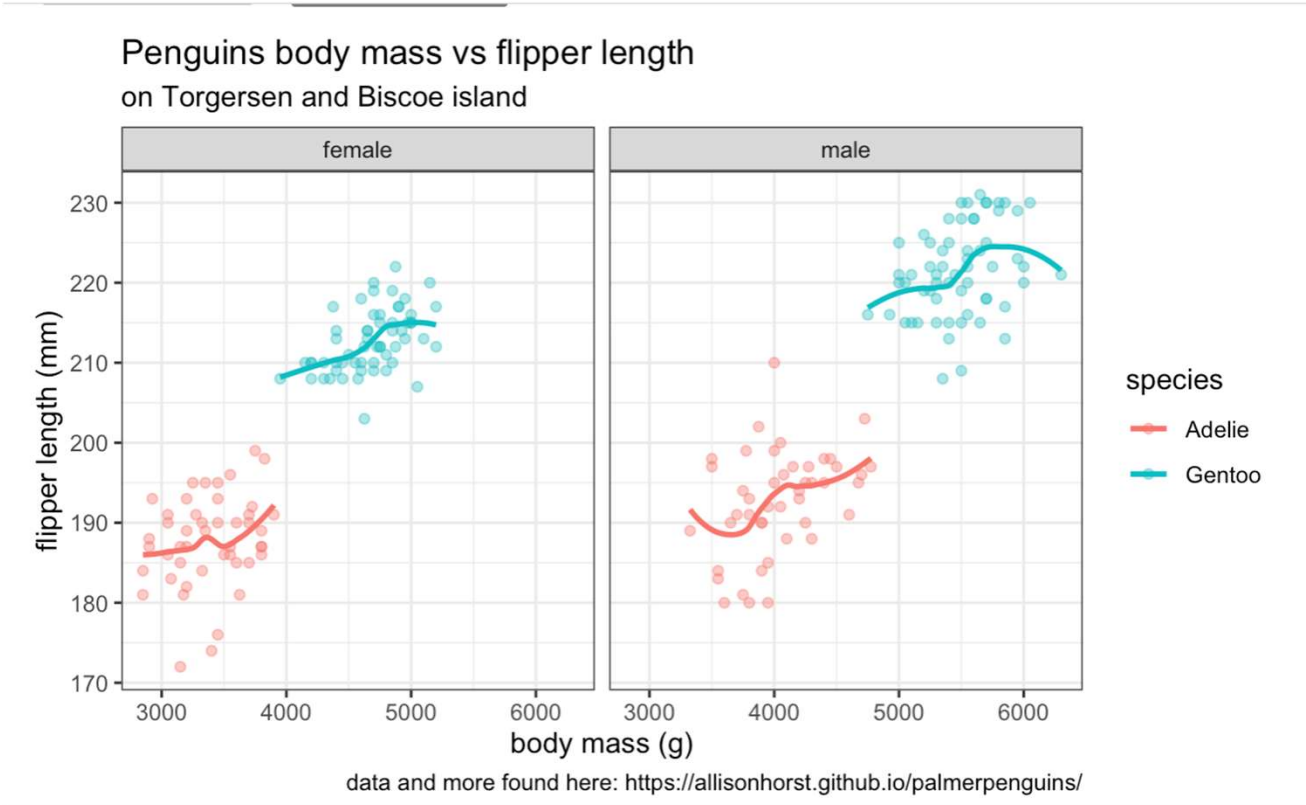
Students log into a reservation system web page

Reservations system assigns students a personal instance of an RStudio container for the course and directs their web browser to the container

RStudio via a web browser: Lecture Day One



Empowering your students to ask and answer questions from the very first day of class, by using R to produce results from Allison Horst's Palmer Penguins R Package, can transform your course experience.



What students say

"I found the Application Exercises (AEs) to be **extremely helpful** in facilitating my learning in this course."

"I really liked the in-class application exercises. The best way to learn how to code is to **actually apply it** and it is useful to go back and reference in-class work. "

"I really appreciated the Application Exercises that we completed in class because they helped me **stay engaged** with the content by putting things into practice right away."

"AE exercises were **useful ways to apply class concepts** vs. Lectures"

"AEs were very helpful. It **kept me focused** during lectures and was a **source of motivation**"

What instructors say

At Davidson College, Compute as a Service resources facilitated by NCShare are being put into practice in an undergraduate Data Science in Society course, designed to allow students to engage with data science methodologies in the context of social justice.

Led by Aubrey Condor, the Duke Professor of the Practice of Data Science at Davidson, this course provides an introduction to methods of data science including visualization, manipulation, programming, and modeling, and integrates an NCShare-funded RStudio container. This container is designed to help streamline logistical hurdles and give students hassle-free access to advanced computing power.

According to Professor Condor, **the platform has allowed her to set up a shared computing environment for the class and students in advance, allowing her students to bypass complex package installation issues and move directly to integrating and practicing the data science methods explored in the class**.

"**When issues arise, we are able to solve them at the container level, and not deal with individual level differences and troubleshooting**," said Professor Condor. "**RStudio makes teaching easier and more streamlined!**"

Where is the RStudio server run?

NCShare *Compute as a Service*

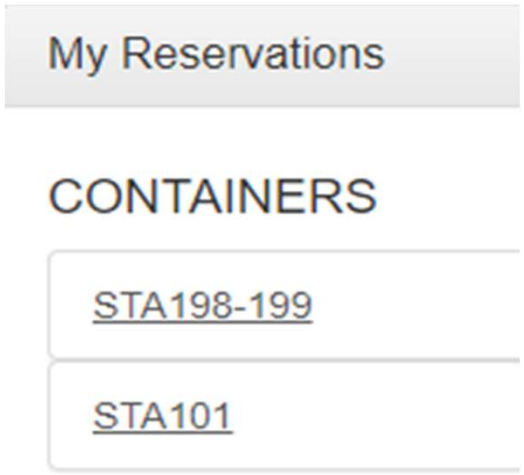
Large institutions such as Duke University have been providing web-based RStudio container access to students for the last 8 years.

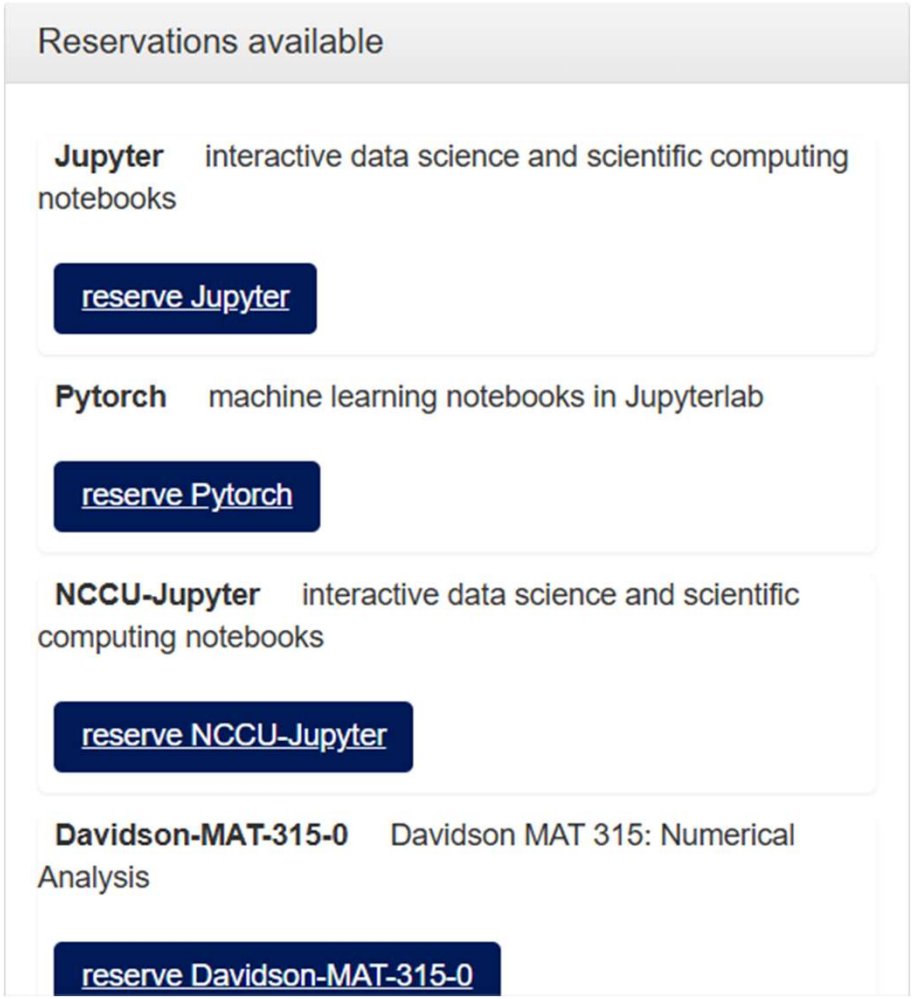
NCShare is an NSF-funded project to provide under-resourced institutions with access to compute environments such as RStudio for coursework and research.

NCShare now supports access from Duke, NCCU, Davidson, Elon, Guilford, Meredith, NC A&T, UNCFSU, UNCW, …

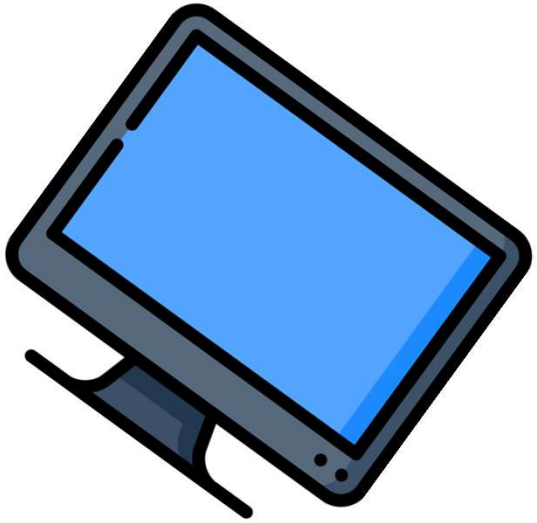
“NCShare Compute as a Service,” is under the direction of Charley Kneifel (Duke), Tracy Futhey (Duke), Joel Faison (North Carolina Central), Mohammad Ahmed (North Carolina Central) and Kevin Davis (Davidson) (NSF OAC – 2201105 PI: Charley Kneifel)

Meet your container: logging in





Day 1 Demo



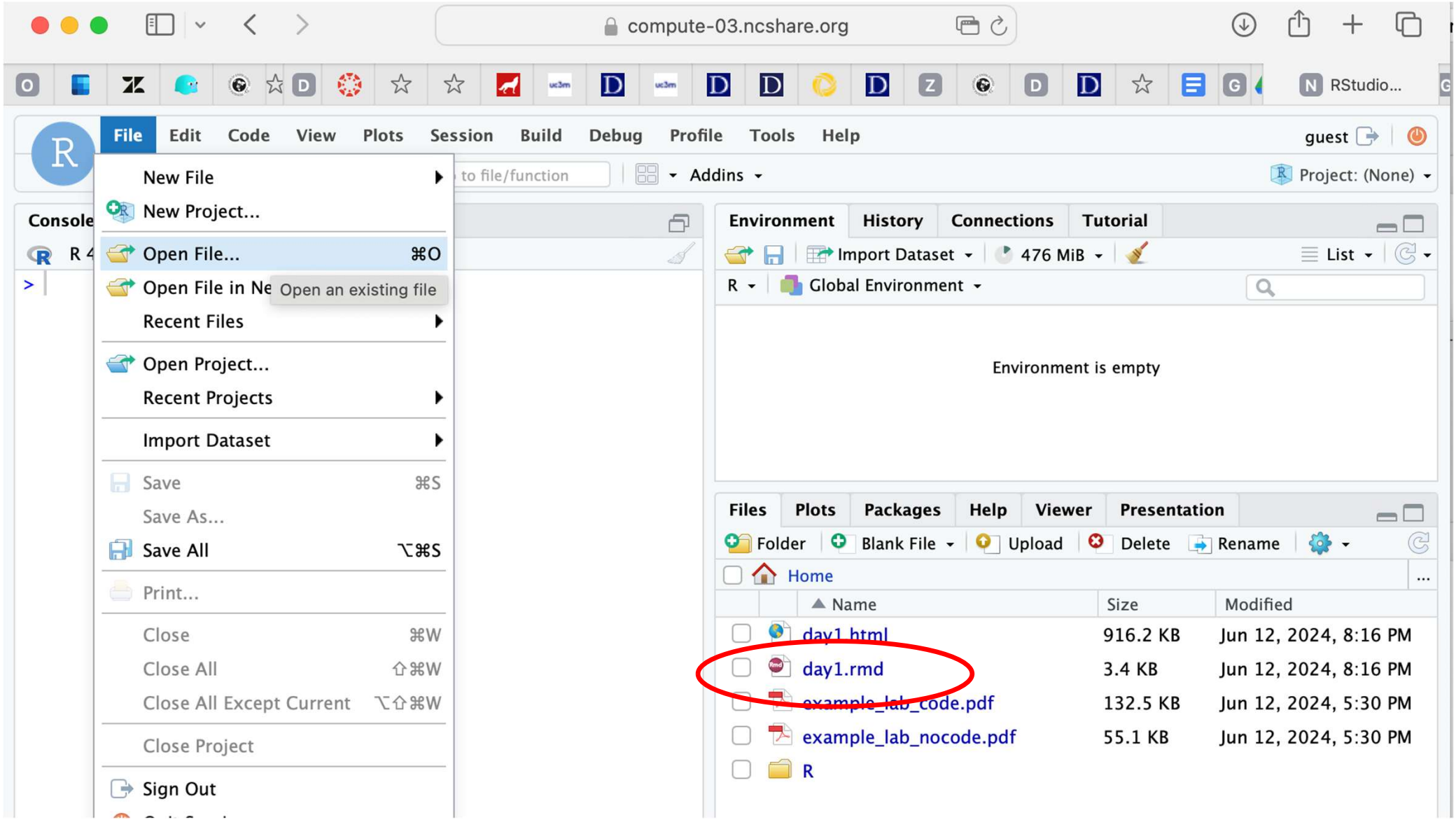
Hands on coding in containers on Day 1

RStudio in a Web Browser

For this hands-on exercise we a bypassing the NCShare container manager reservation system (since not everyone here is from an NCShare school).

These containers will remain active through Friday afternoon.

Please use your guest account url to open your container in your browser!


compute-03.ncshare.org
RStudio...
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Help
guest
Addins
Project: (None)
New File
New Project...
Open File... ⌘O
Open File in Ne
Open an existing file
Recent Files
Open Project...
Recent Projects
Import Dataset
Save ⌘S
Save As...
Save All ⌥⌘S
Print...
Close ⌘W
Close All ⇧⌘W
Close All Except Current ⌥⇧⌘W
Close Project
Sign Out
Console
Environment
History
Connections
Tutorial
Import Dataset
476 MiB
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
Folder
Blank File
Upload
Delete
Rename
Home
Name
Size
Modified
day1.html 916.2 KB Jun 12, 2024, 8:16 PM
day1.rmd 3.4 KB Jun 12, 2024, 8:16 PM
example_lab_code.pdf 132.5 KB Jun 12, 2024, 5:30 PM
example_lab_nocode.pdf 55.1 KB Jun 12, 2024, 5:30 PM
R

```
---
title: "Penguins: Day 1"
author: ""
date: "2024-06-13"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

Packages

```{r echo=FALSE, warning=FALSE, message=FALSE}
library(tidyverse)
library(tidymodels)
library(dplyr)
library(ggplot2)
```

Penguin Exploration

Data were collected and made available by Dr. Kristen Gorman and the Palmer Station,
```

Benefits

- All students have one common experience
- Learn by doing & Learning how to do
- Question variety
- Excited to learn
- Take ownership of their learning
- Trickle-down effect

Where to find examples

Elijah Meyer - STA199 F23-1 Intro to Data Science https://sta199-f23-1.github.io/

Mine Çetinkaya-Rundel – STA199 S24 Intro to Data Science

https://sta199-s24.github.io/

And Lab-0 https://sta199-s24.github.io/labs/lab-0.html

Escalation path if something goes wrong

Where do I start?

1.) usually hyperlocal fix: student restarts their container after clearing cache

2.) student talks with instructor

3.) instructor contacts their local IT support person

4.) IT support person contacts info @ncshare.org

NCShare email: info@ncshare.org web: www.ncshare.org

Getting started with NCShare

Where do I start?

Next steps to host coursework RStudio student containers at NCShare:

1.) name a technical contact at your institution for local support and problem escalation

2.) configure logins so students can use your institution's login service

3.) make arrangements with NCShare for capacity to handle student access

4.) test/verify the course assignment work in the RStudio container

NCShare email: info@ncshare.org web: www.ncshare.org

Questions?