Welcome to DATA 151

I'm so glad you're here!

DATA 151: CLASS 3A
INTRODUCTION TO DATA SCIENCE (WITH R)

EXPERIMENTAL DESIGN AND INTRODUCTION TO R

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING

Introduction to Data Science:

- Tuesday:
  - Ch 1: Getting Started with R and R Studio
  - Ch 2: R Basics
- Thursday:
  - Ch 3: Programming basics

HOMEWORK REMINDER

Due this week:

- *HW #2: Practice Problems* ***(due on WISE 9/15)***
- *Project Milestone #0: Communication Plan*
  - ***Due on WISE 9/15***
  - *One submission per group*

EXPERIMENTAL DESIGN

RELATIONSHIPS BETWEEN VARIABLES

Many analyses are motivated by a researcher looking for a relationship between two variables.

Definitions:

- **Response/Dependent variable (Y):** the variable one suspects is affected by the explanatory variable(s).
  - Variable that is of interest to study
- **Explanatory/Independent variable (X):** the variable whose effect one wants to study
  - Is thought to explain or influence the response variable

PRINCIPLES OF EXPERIMENTAL DESIGN

Randomized experiments are build on four principles:

- **1) Control**
  - (verb) Control for lurking variables that might affect the response, most simply by comparing two or more treatments
  - (noun) May also be referred to as the "non-treatment"

- **2) Randomization**
  - Use chance to assign experimental units to treatments

PRINCIPLES OF EXPERIMENTAL DESIGN

- **3) Replication**
  - Use enough experimental units in each group to reduce chance of variation in the results

- **4) Blocking**
  - The arranging of experimental units in groups (blocks) that are known to be similar to one another
  - Blocking factors is typically a source of variability but not the primary interest
  - Common Examples: Space and time…

TYPES OF EXPERIMENTAL DESIGNS

1. Completely Randomized Design
2. Randomized Block Design
3. Matched Pairs Design

COMPLETELY RANDOMIZED DESIGN

- Also known as **CRD**
- The simplest experimental design, in terms of analysis and convenience
- Subjects are randomly assigned to treatments
- Typically done by listing treatment levels and randomly assigning random numbers to each

COMPLETELY RANDOMIZED DESIGN

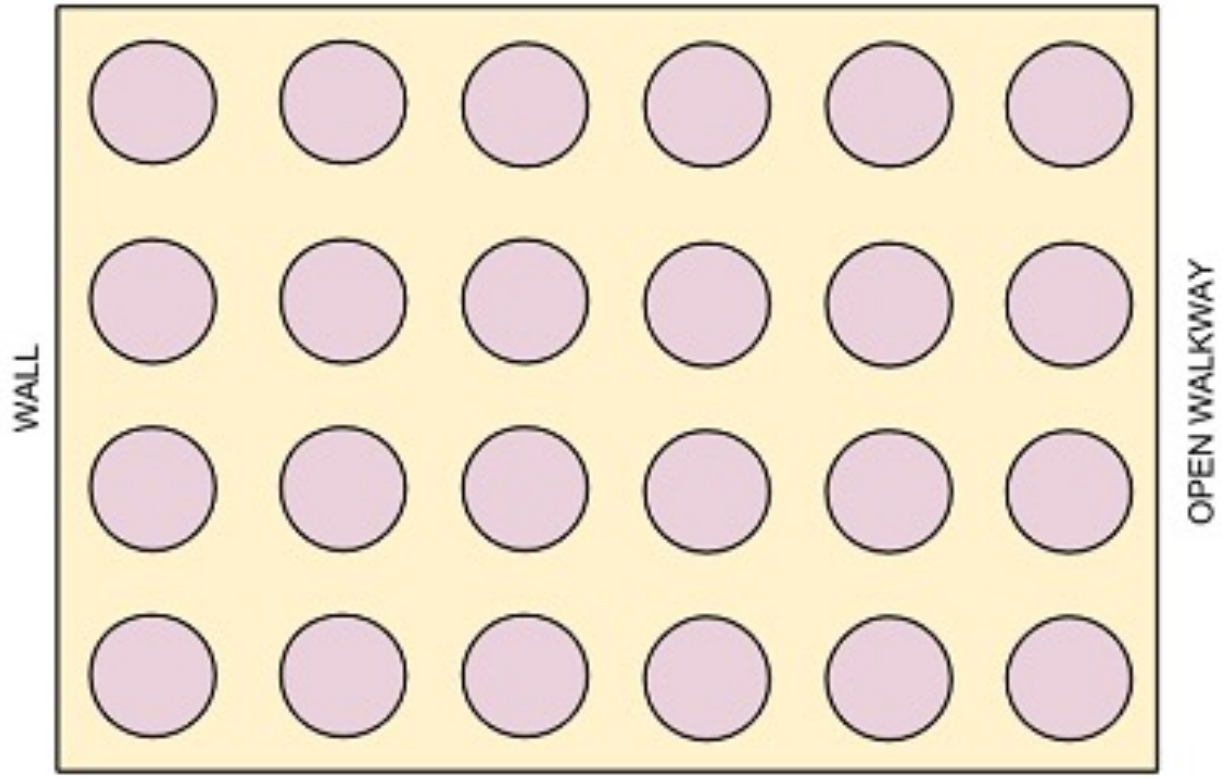
Consider the set up:

- In a greenhouse experiment we want to study a single factors (fertilizer) with 4 levels
- We have enough space for 24 experimental units (a potted plant)
- To maintain balance in the experiment, we will have 6 replications of each treatment

Source: https://newonlinecourses.science.psu.edu/stat502/node/175/

COMPLETELY RANDOMIZED DESIGN

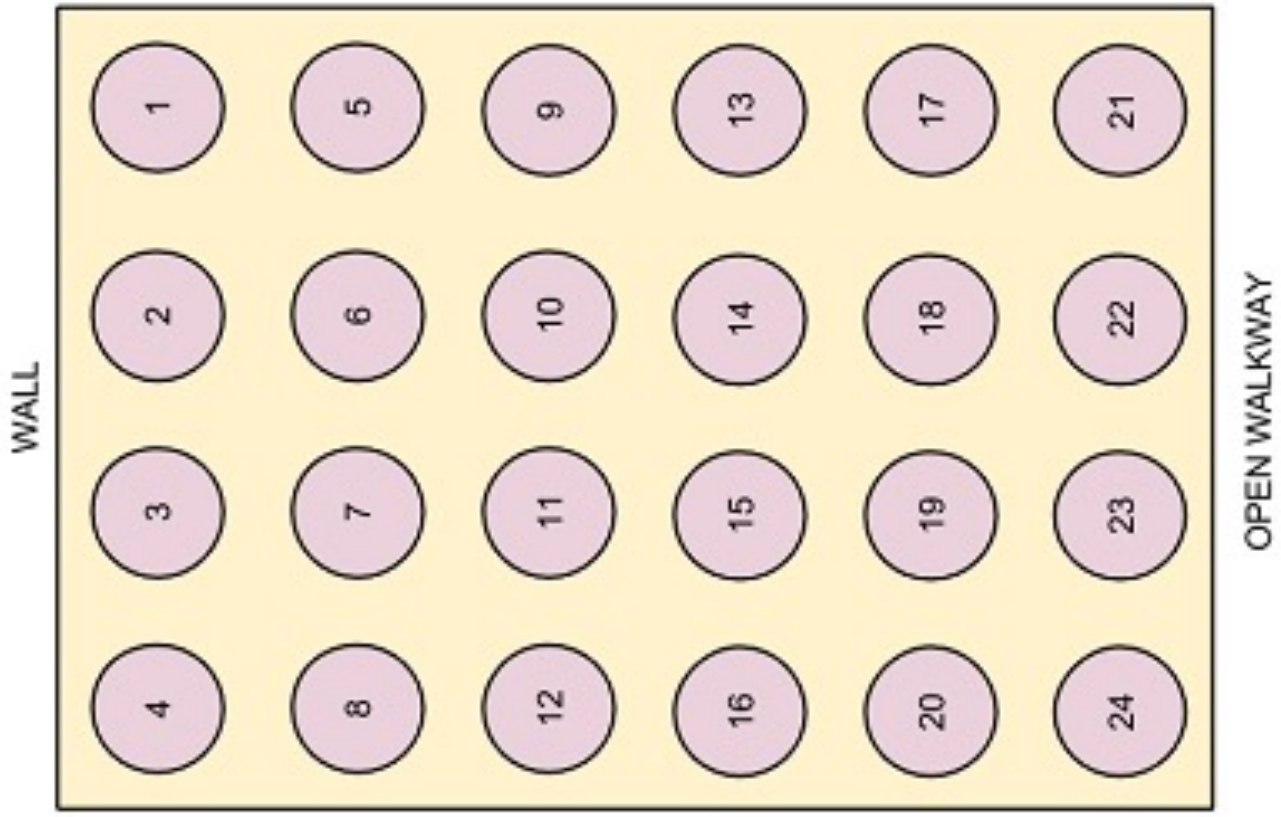
Greenhouse Diagram and bench used for the experiment (viewed from above):



Source: https://newonlinecourses.science.psu.edu/stat502/node/175/

COMPLETELY RANDOMIZED DESIGN

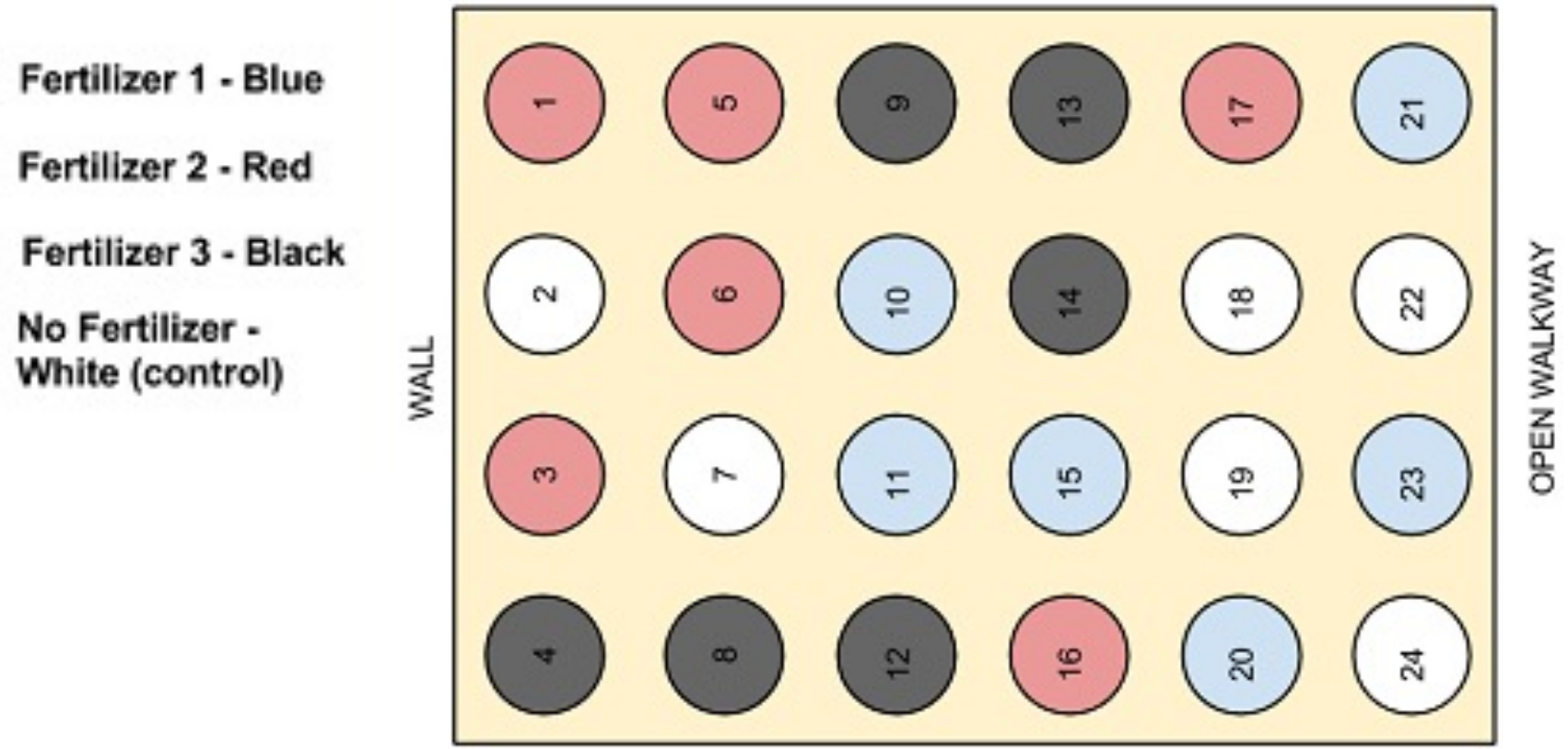
Step 1: Assign it experimental unit a unique id



Source: https://newonlinecourses.science.psu.edu/stat502/node/175/

COMPLETELY RANDOMIZED DESIGN

Step 2: Randomly assign each experimental units to treatments



COMPLETELY RANDOMIZED DESIGN

… but what if there are known nutrient gradients across the bench?

A vanilla CRD will not control for this!

RANDOMIZED (COMPLETE) BLOCK DESIGN

- Also known as **RCBD**
- Variation between blocks is accounted for assigning at least one of each treatment to each block
- Effects of blocks not of interest
- Standard design for agricultural experiments

RANDOMIZED (COMPLETE) BLOCK DESIGN

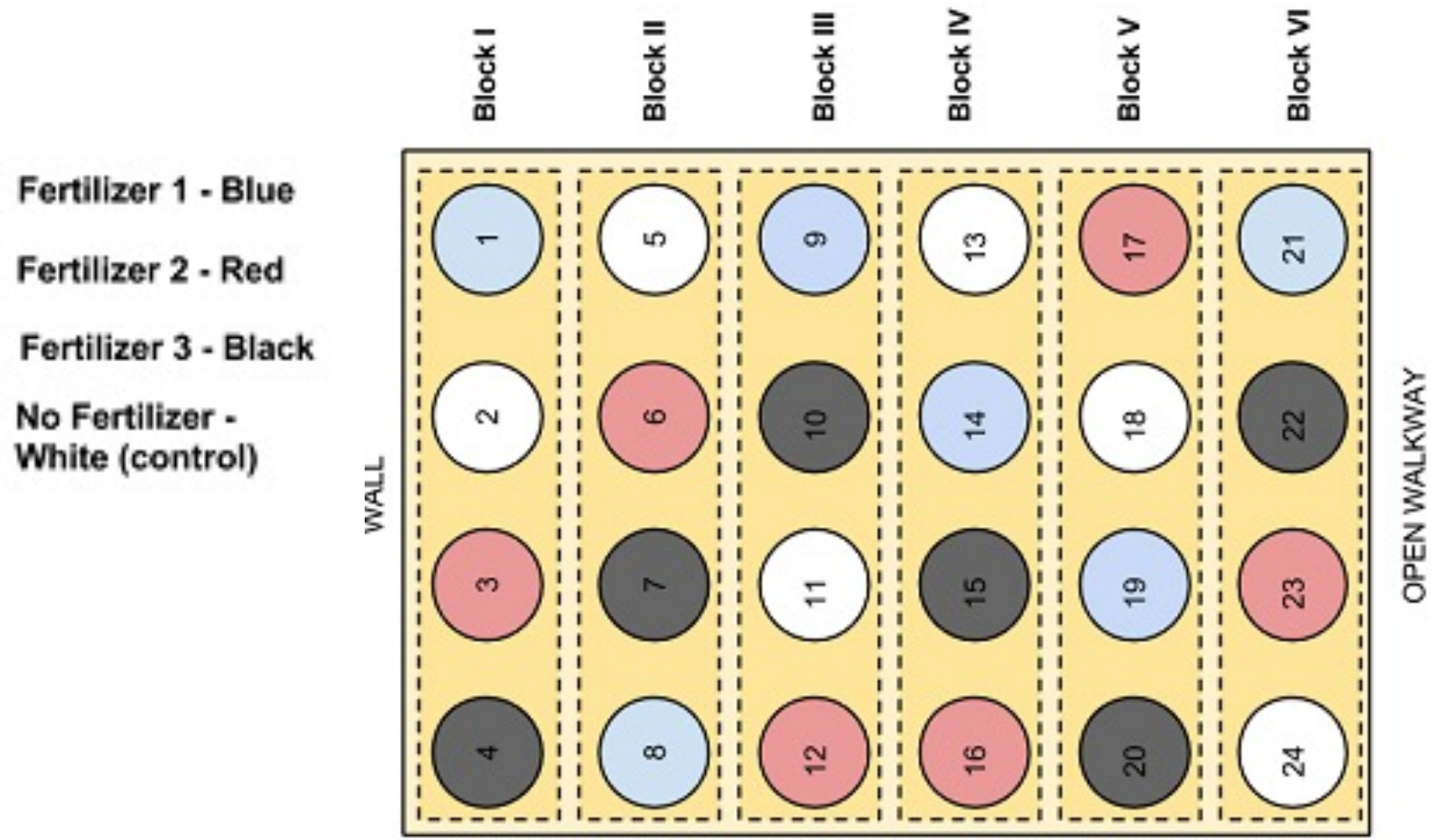
In a block design, the random assignment of experimental units to treatments is carried out within each block

What are the steps in performing a blocked experiment?

1. Form groups (blocks)
   - All individuals within each block should be similar in regard to the lurking variable

2. Within each block, randomly assign experimental units to each treatment.

RANDOMIZED (COMPLETE) BLOCK DESIGN



MATCHED PAIRS (AN EXTENSION OF BLOCKING)

- Big Idea: Create blocks by matching pairs of similar experimental units

- Chance is used to determine which unit in each pair gets each treatment

Ex: **Pre-Post (Before After) Studies**

Data from the same individual is related (treat like a block)

1. Assess baseline
2. Assign treatment
3. Find difference after

WORKSHEET EXAMPLES

WORKSHEET EXAMPLES

Example of a blocked design:

An experiment that showed that high doses of omega-3 fats might be a benefit to people with bipolar disorder involved a control group of subjects who received a placebo. Researchers hoped to design a study with two treatment groups, one taking a high dose of omega-3 fatty acids and the other a placebo. Suppose researchers recognized that some of the participants in the study were very active people who walked a lot or got vigorous exercise several times a week, while others tended to be more sedentary. Design a Blocked Experiment, blocking on activity level.

WORKSHEET EXAMPLES

Example: The Blood Lactate Example - A Matched-Pairs (Before and After) design
The effect of exercise on the amount of lactic acid in the blood was examined by researchers. In a particular study, eight men who were attending a week-long training camp were randomly selected to participate in the study. The blood lactate levels (in mmol/L (millimoles per liter of blood)) were measured before and after playing three games of racquetball for each of the 8 men. Researchers wanted to determine if exercise increased blood lactate levels. Explain why this is an example of a matched-pairs design.

RStudio®

INTRODUCTION TO R

WHAT IS R?

Free and open source programming language created by statisticians as an interactive environment for data analysis

Benefits:

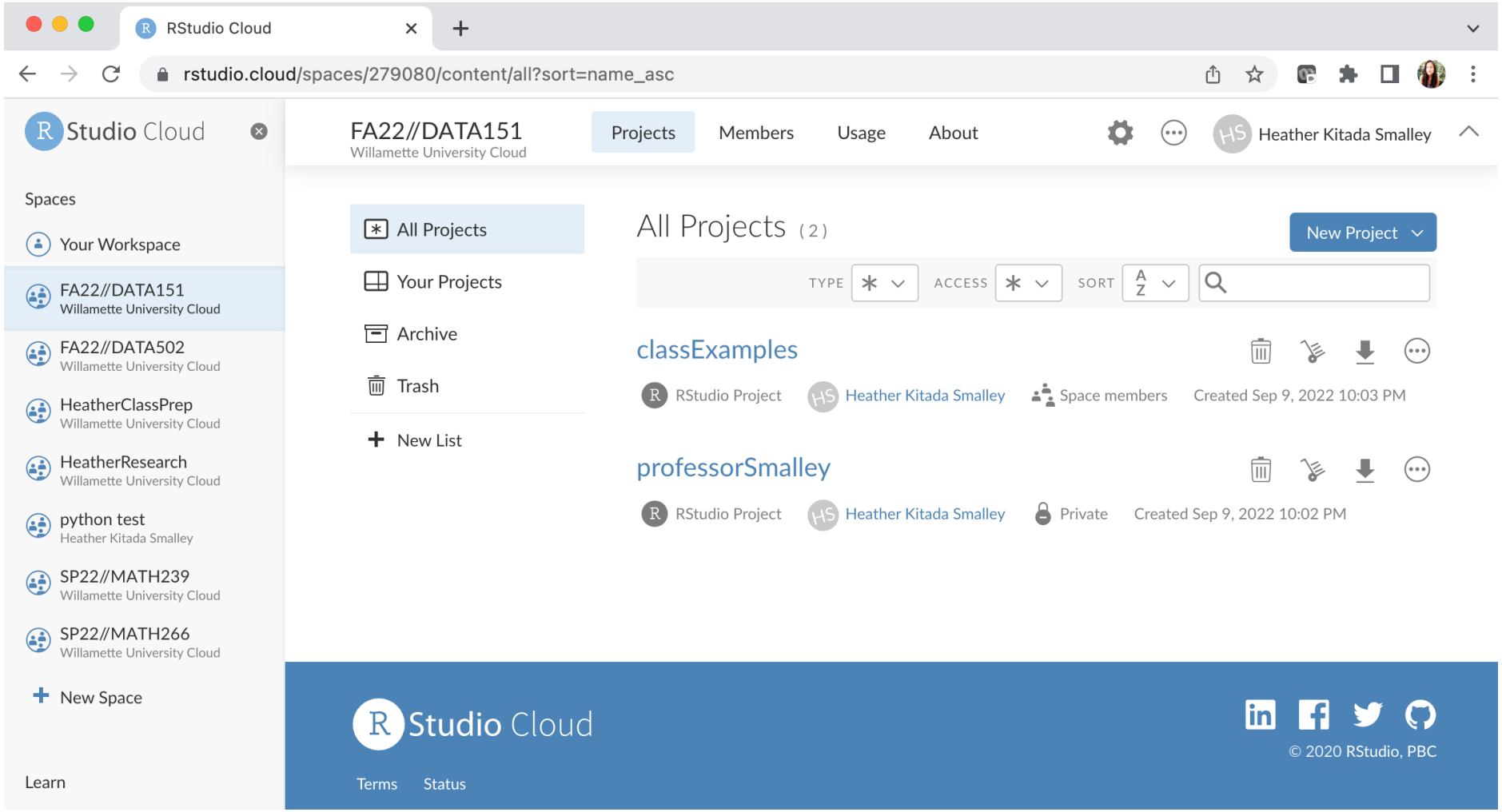
- Relatively quick to learn
- Large community of R users (and online support)
- You can edit and save scripts (rather an point and click)

WHAT IS R STUDIO?

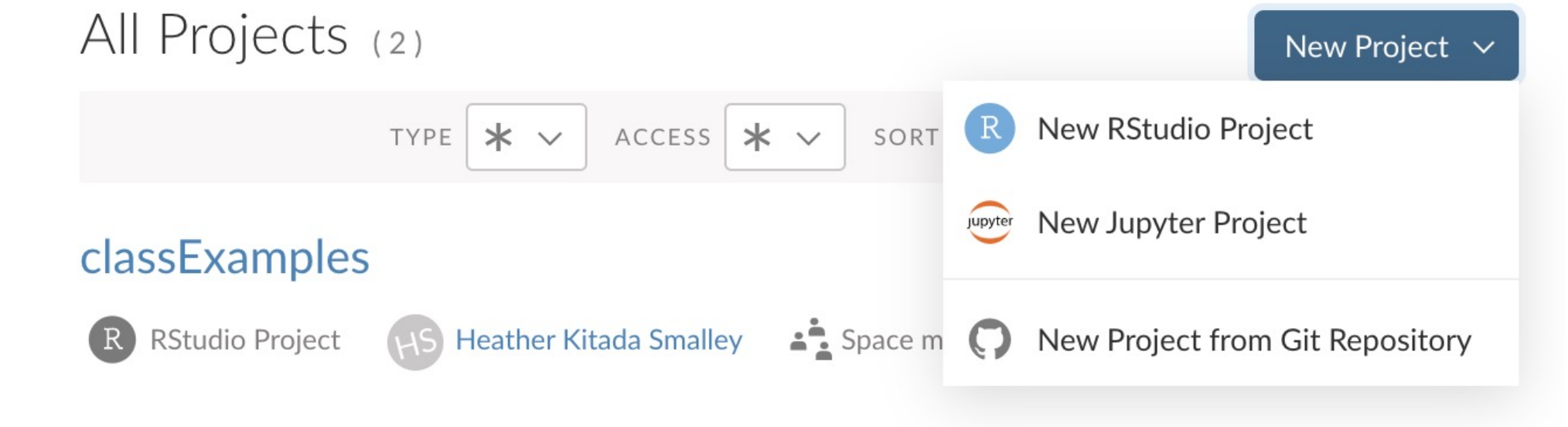
R Studio is an **IDE (Integrated Development Environment)**

R Studio runs on top of R (the programming language and complier) to provide a more aesthetic and organized experience for programming in R

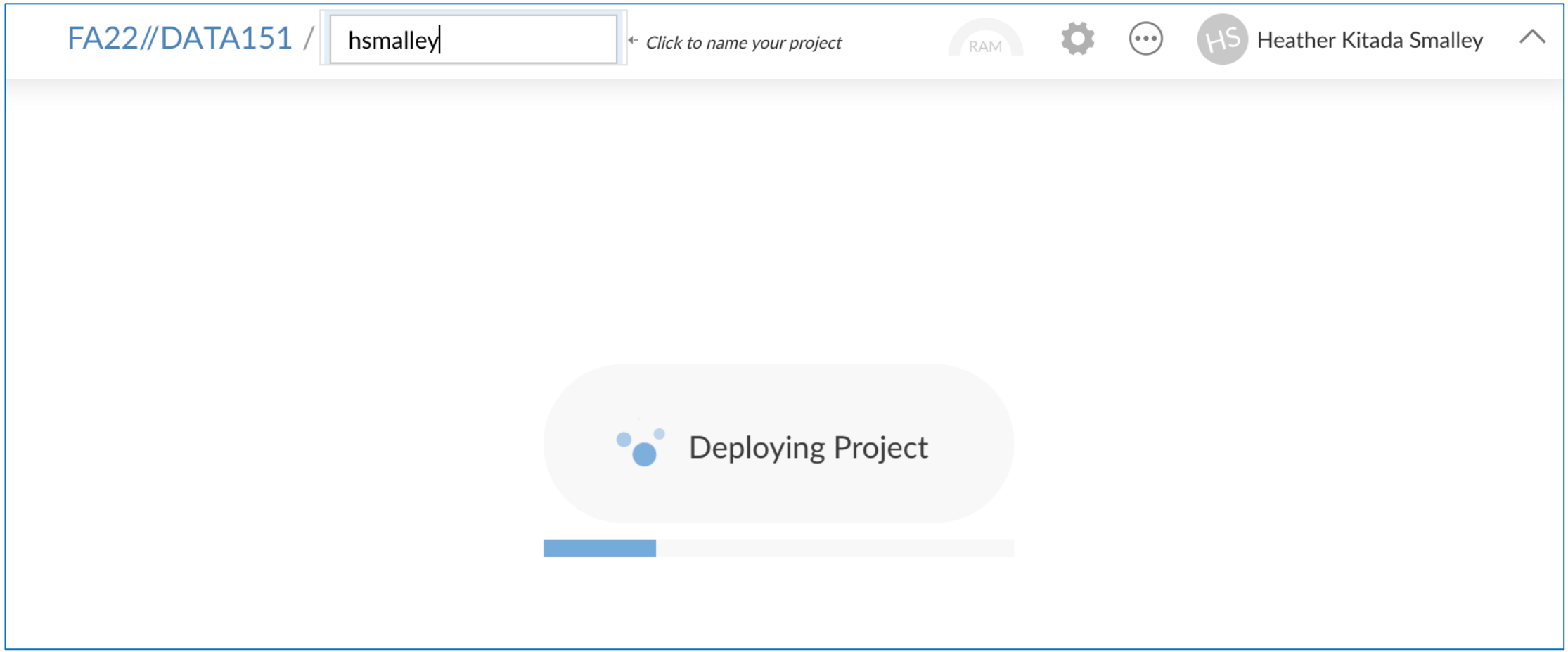
R STUDIO CLOUD



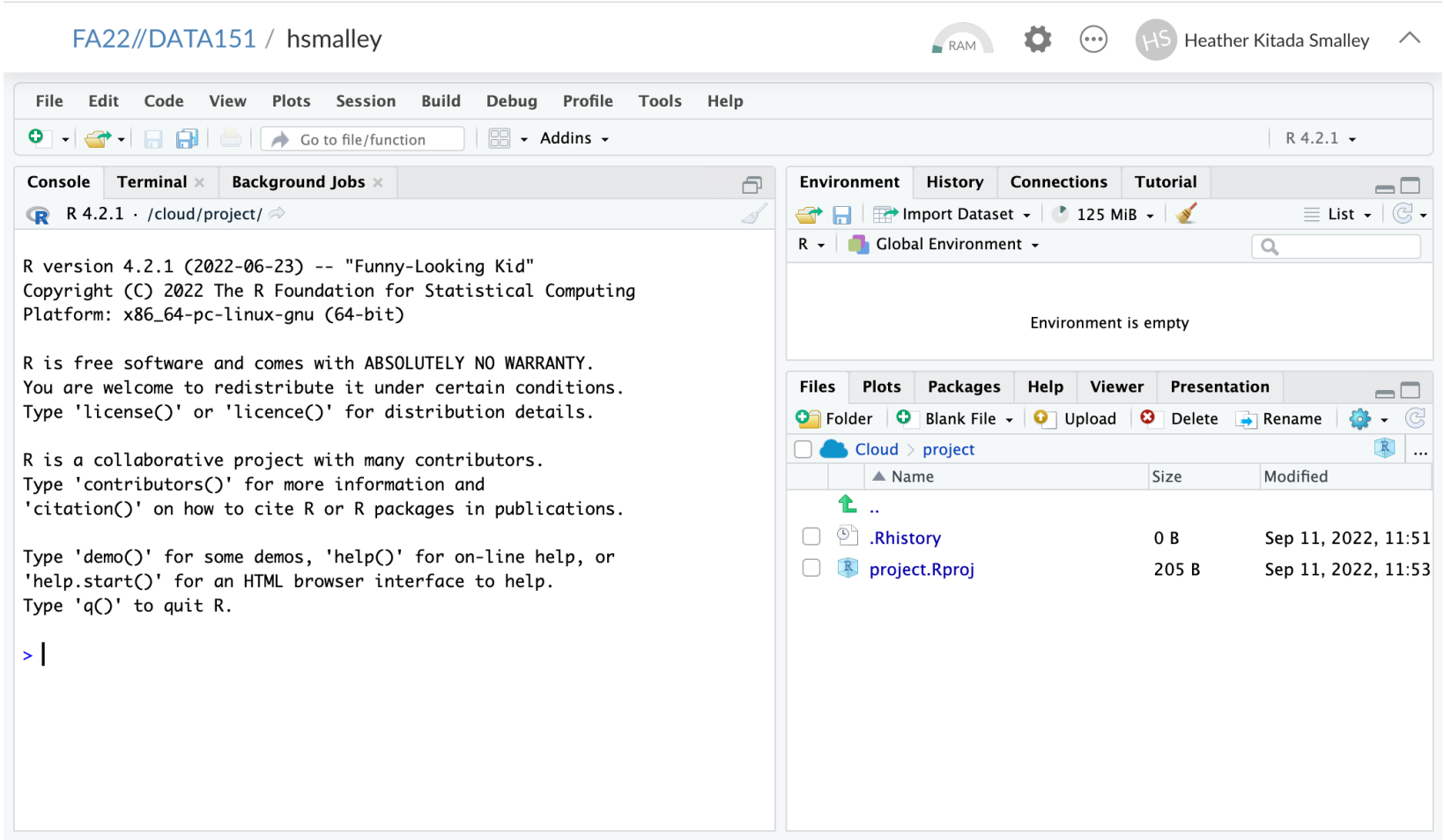
MAKING A PROJECT IN OUR CLASS "SPACE"



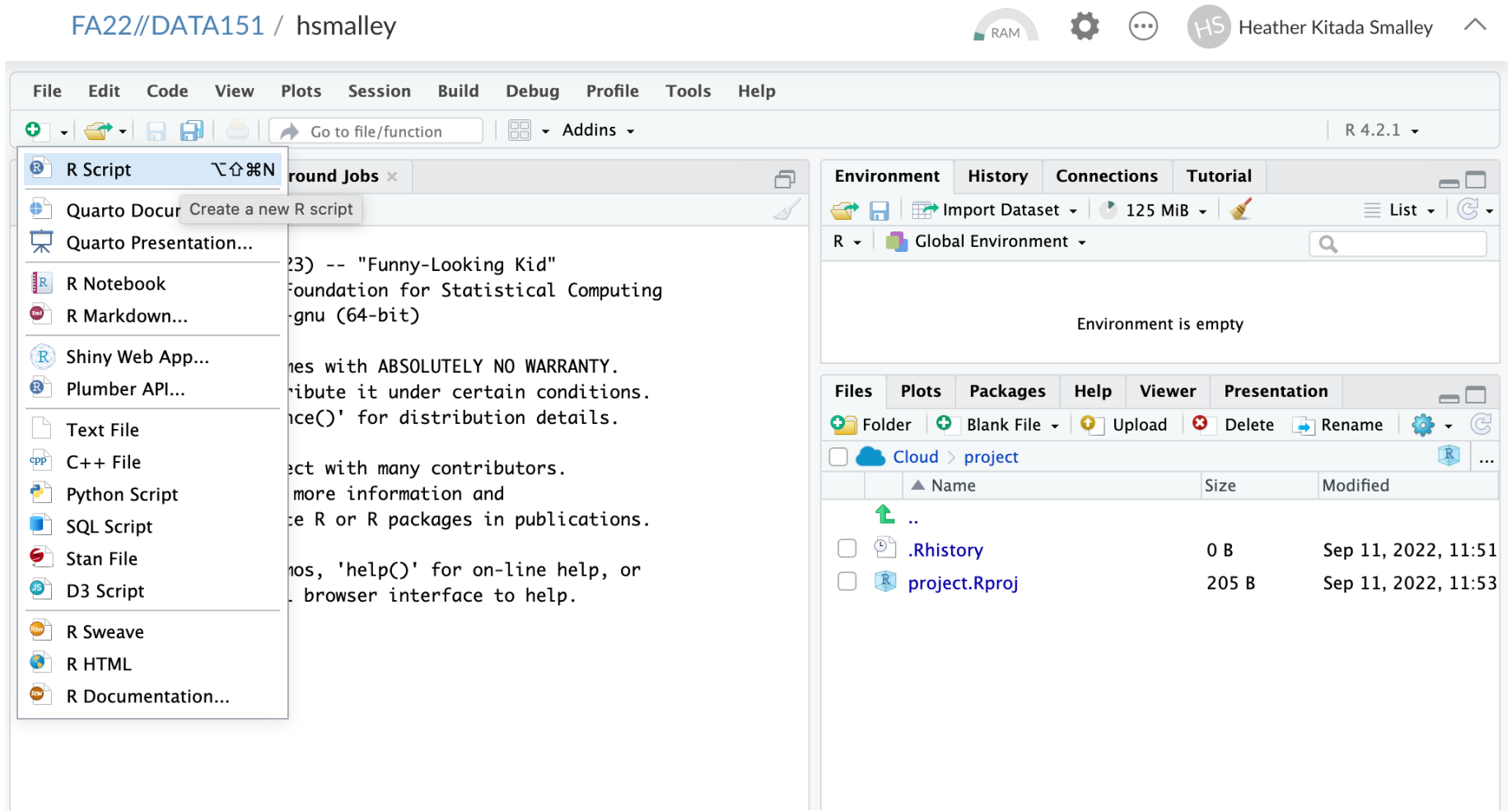
MAKING A PROJECT IN OUR CLASS "SPACE"



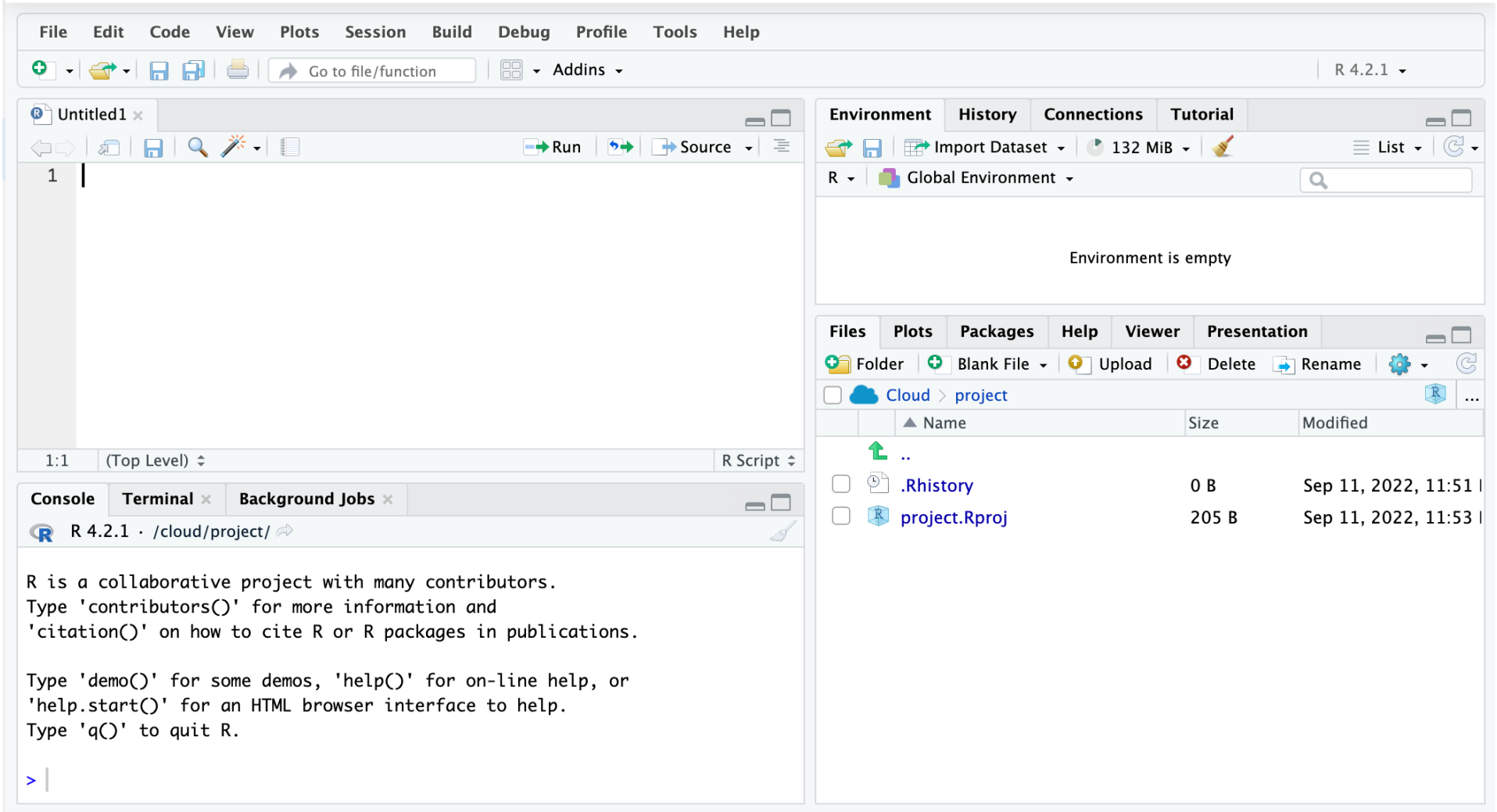
MAKING YOUR FIRST R STUDIO SCRIPT



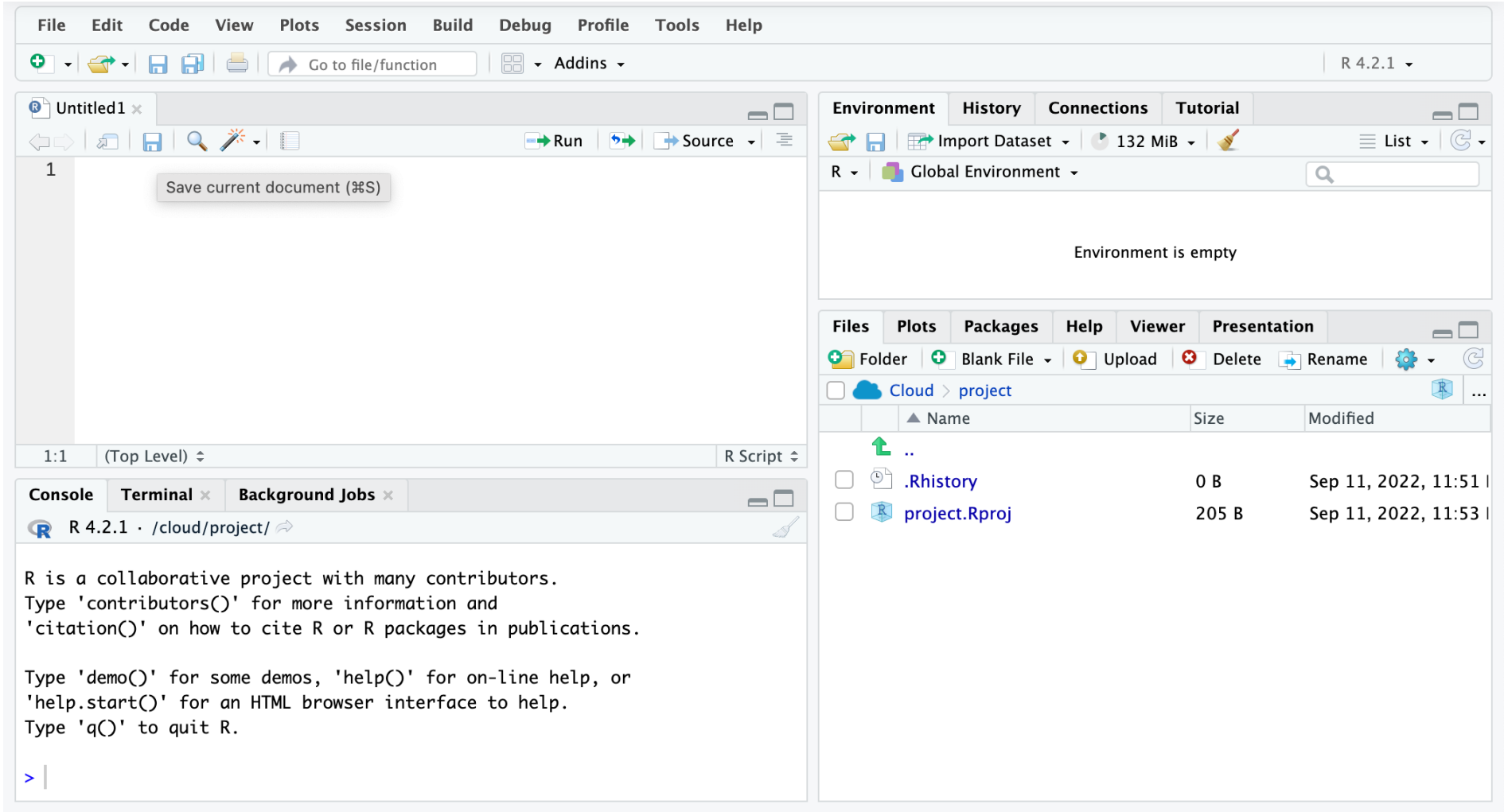
MAKING YOUR FIRST R STUDIO SCRIPT



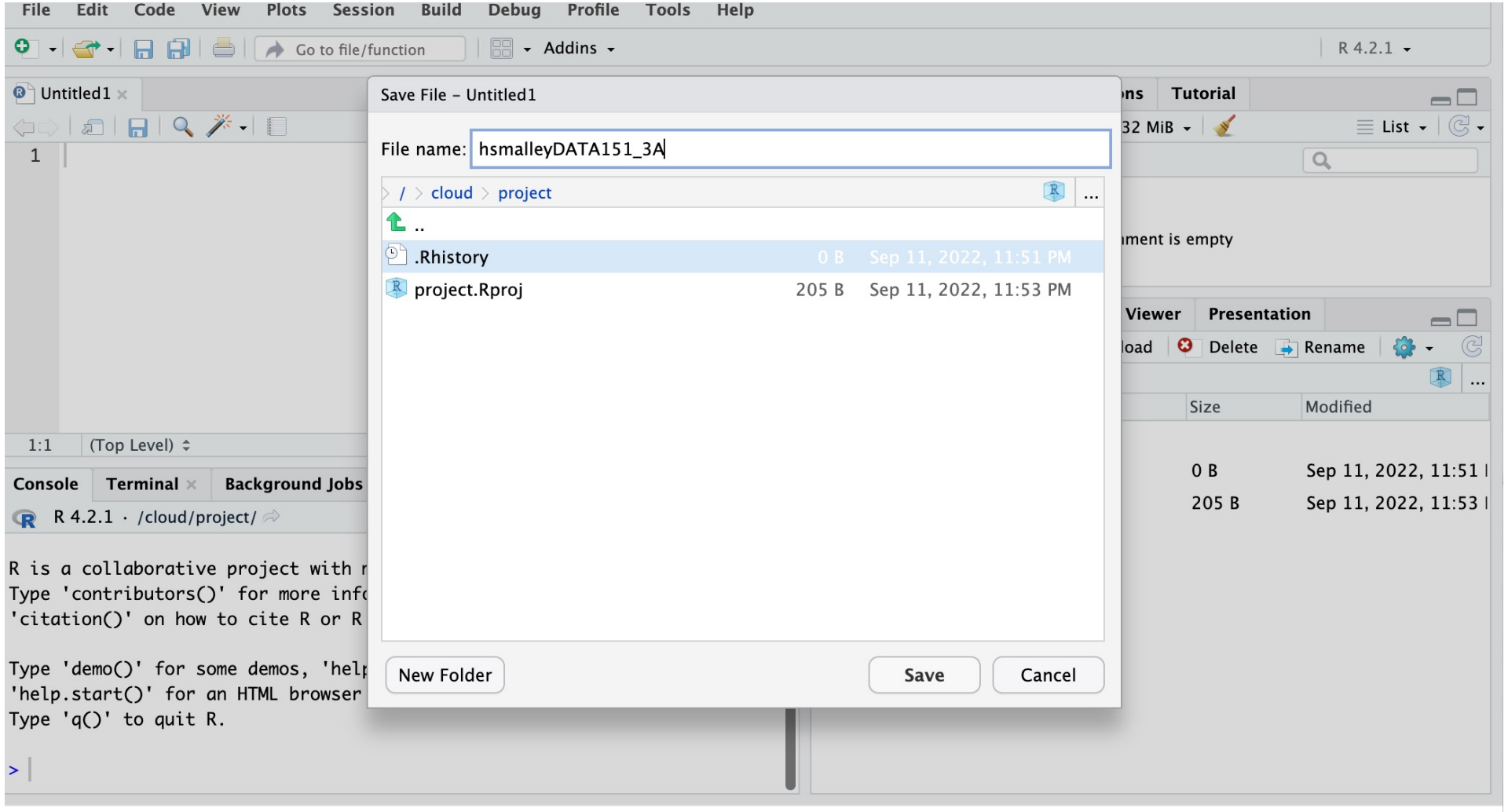
MAKING YOUR FIRST R STUDIO SCRIPT



SAVING AN R SCRIPT

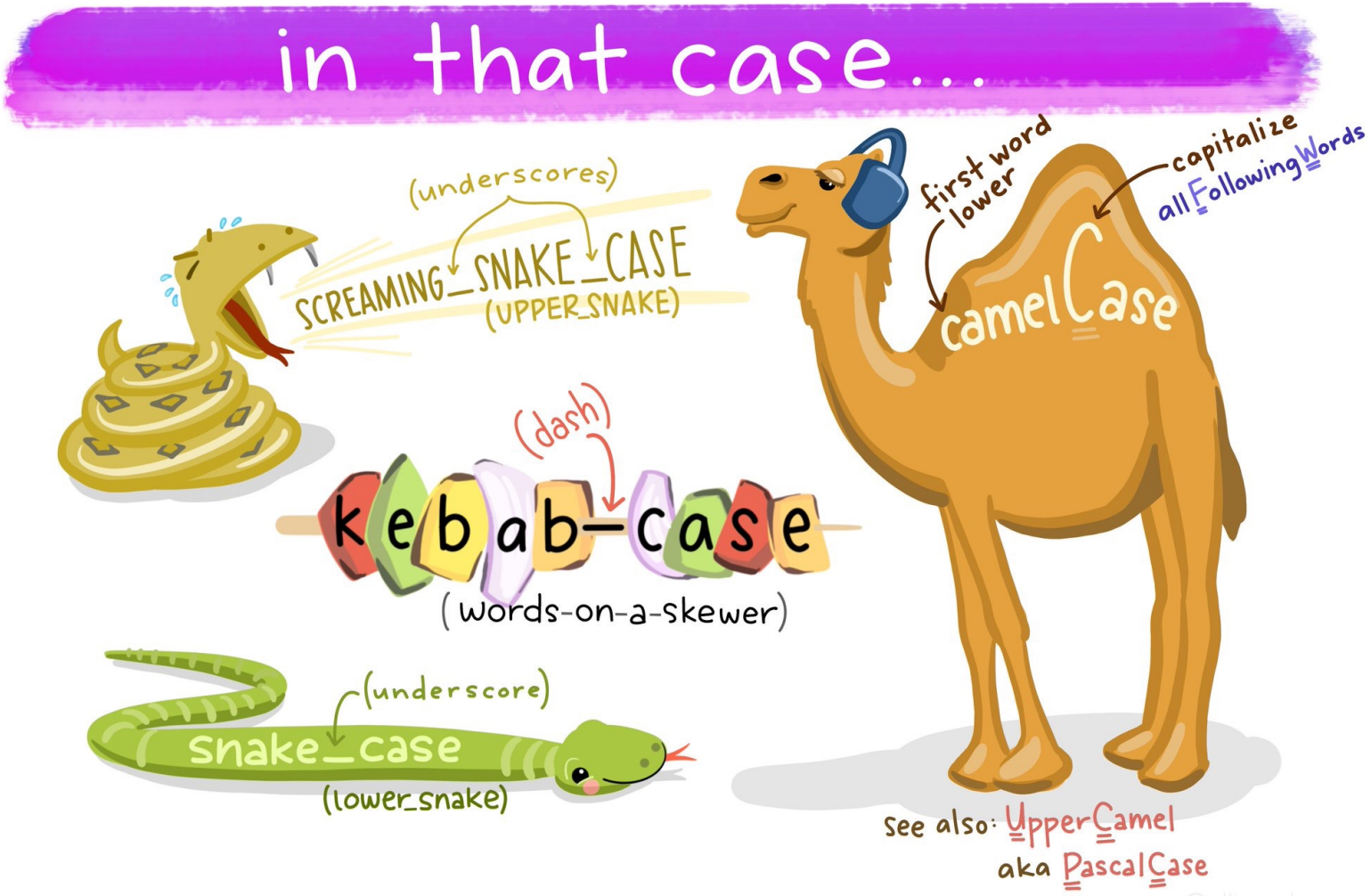


NAMING AN R SCRIPT

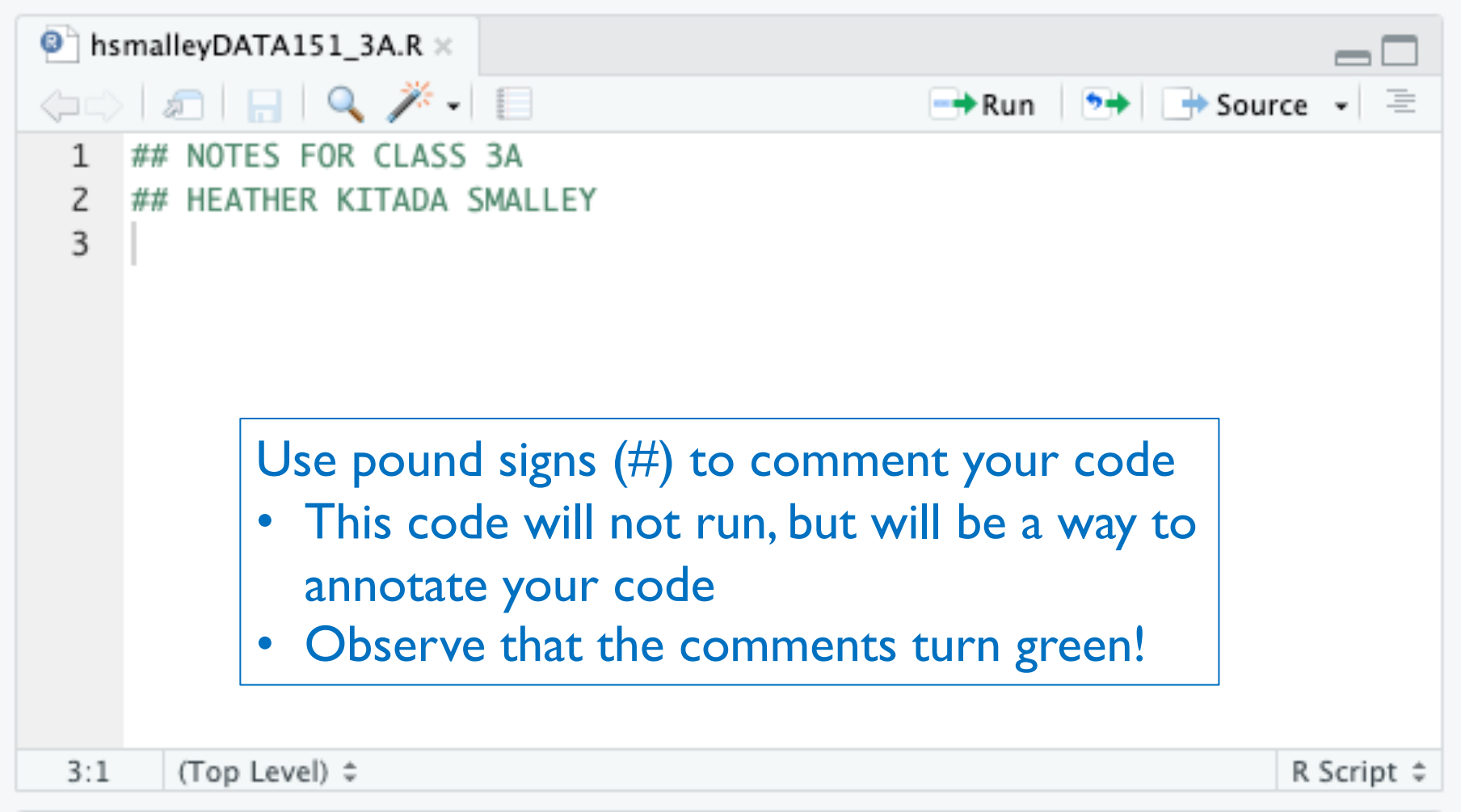


NAMING CONVENTIONS

Allison Horst
@allison_horst



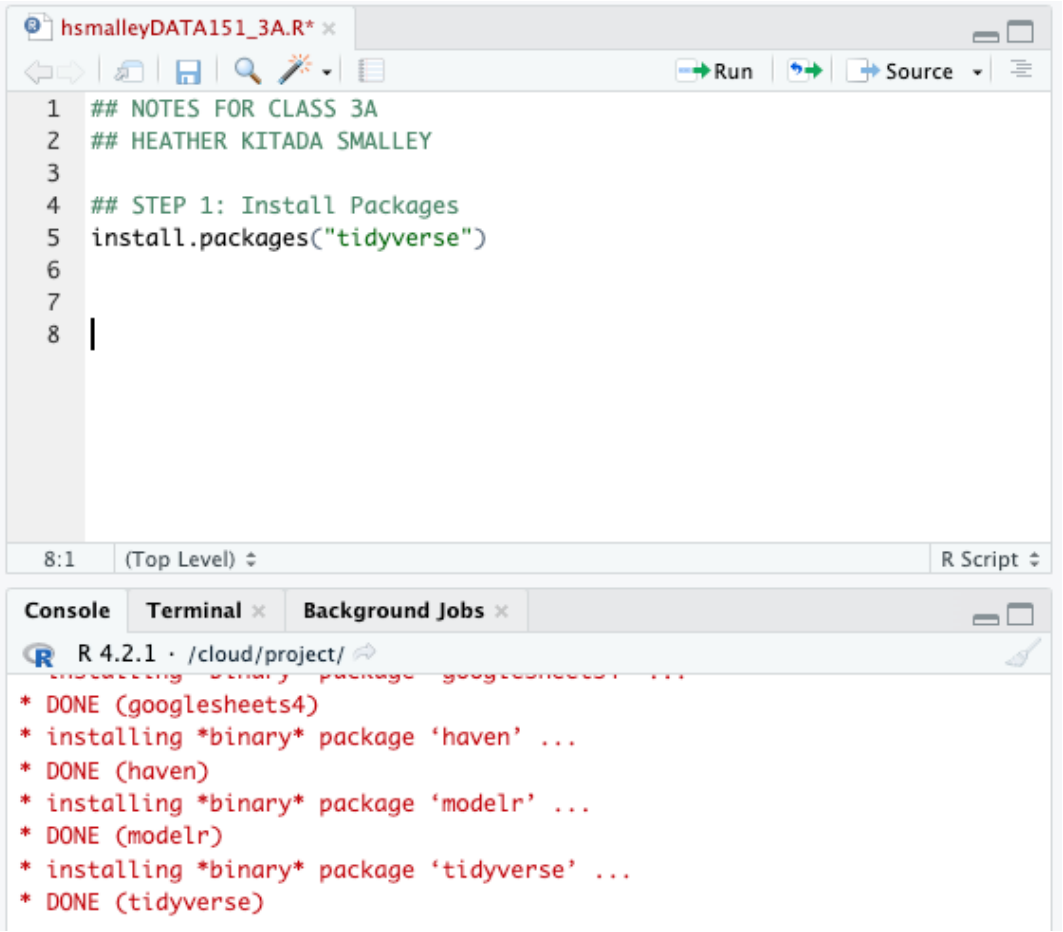
COMMENTING YOUR CODE



Use pound signs (#) to comment your code

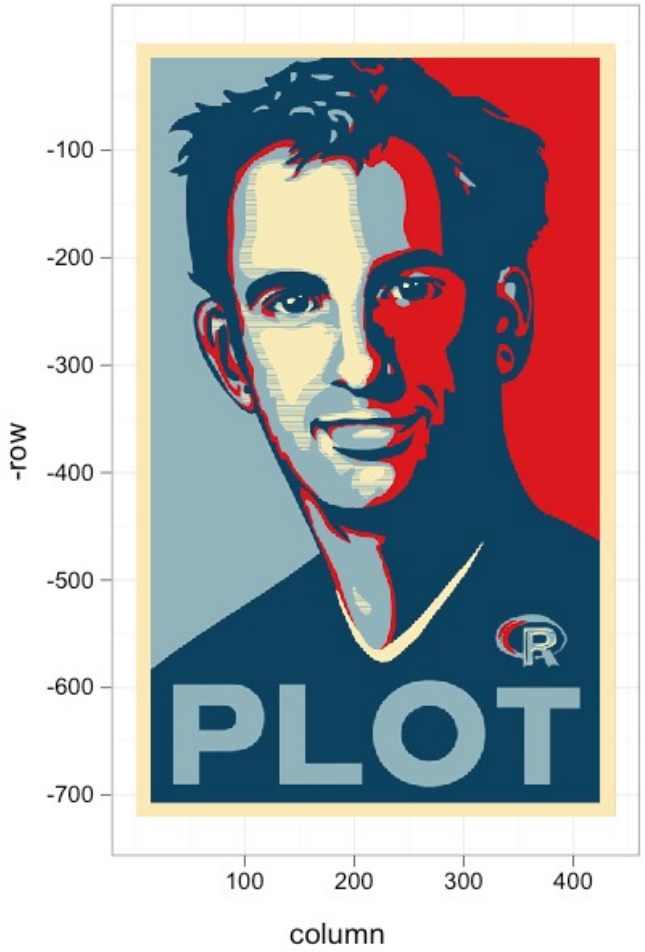
- This code will not run, but will be a way to annotate your code
- Observe that the comments turn green!

INSTALLING PACKAGES



- You should only need to install a package once
- Observe the use of quotes in the `install.packages()` command

HADLEY'S TIDYVERSE

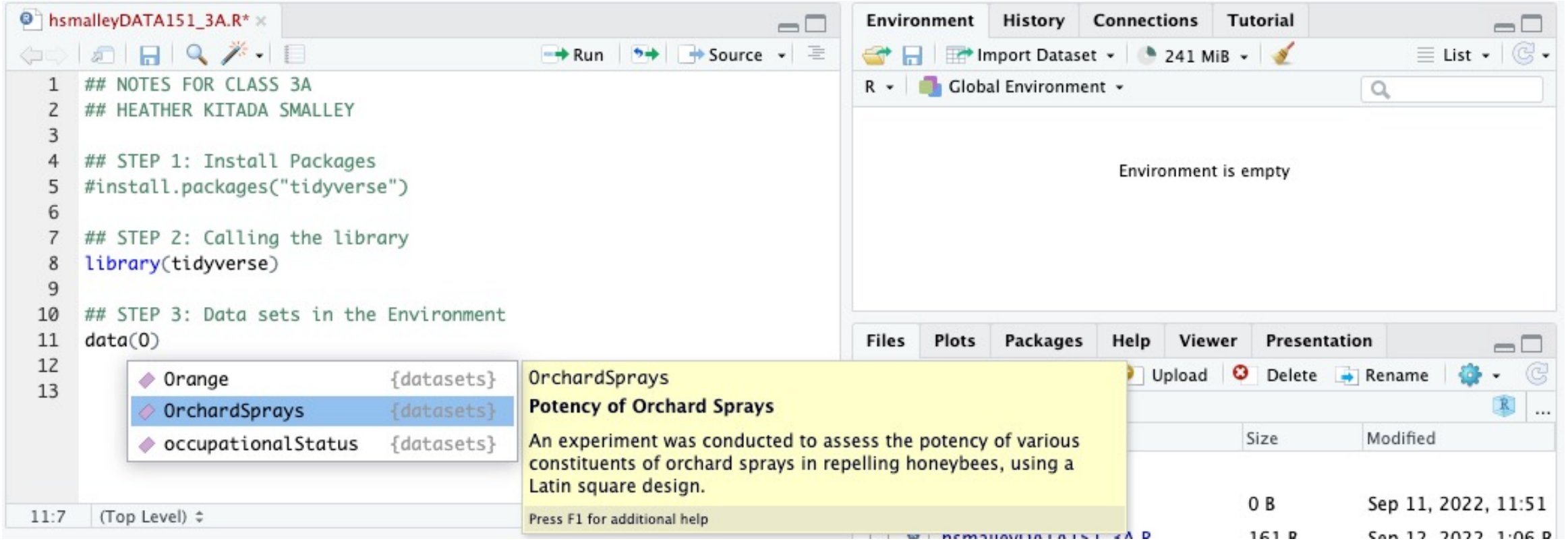


CALLING THE LIBRARY

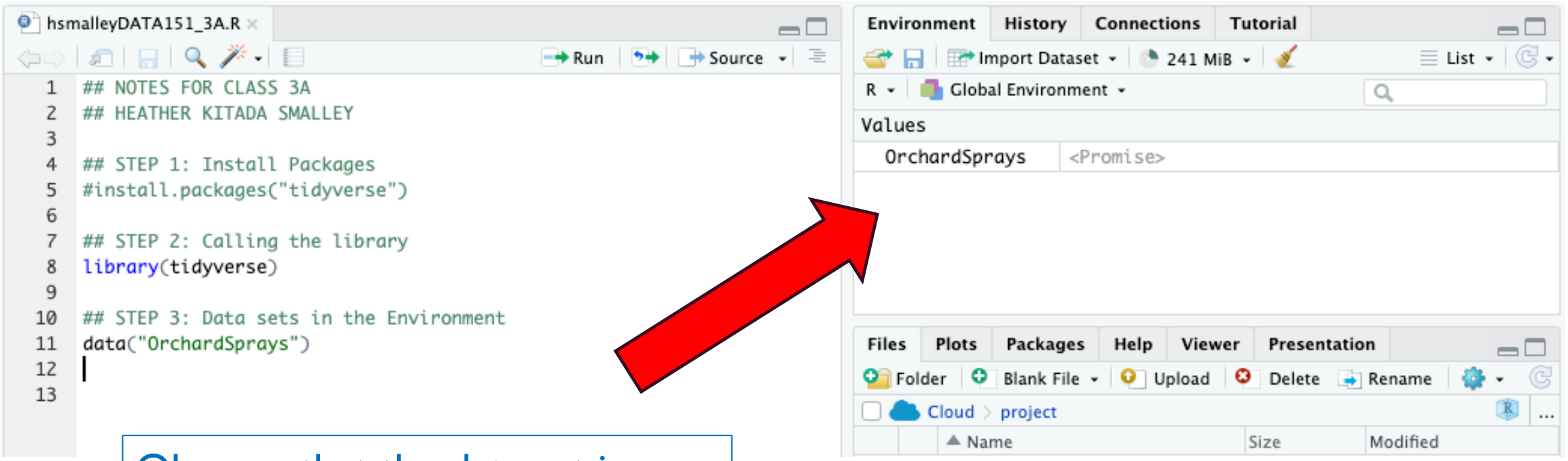
```
> ## STEP 2: Calling the library
> library(tidyverse)
── Attaching packages ─────────────────────────────── tidyverse 1.3.2 ──
✔ ggplot2 3.3.6      ✔ purrr   0.3.4
✔ tibble  3.1.8      ✔ dplyr   1.0.10
✔ tidyr   1.2.1      ✔ stringr 1.4.1
✔ readr   2.1.2      ✔ forcats 0.5.2
── Conflicts ────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
>
```

Note: Conflicts can occur when the library contains functions with the same name as another function

CALLING R BUILT IN DATASETS TO THE ENVIRONMENT

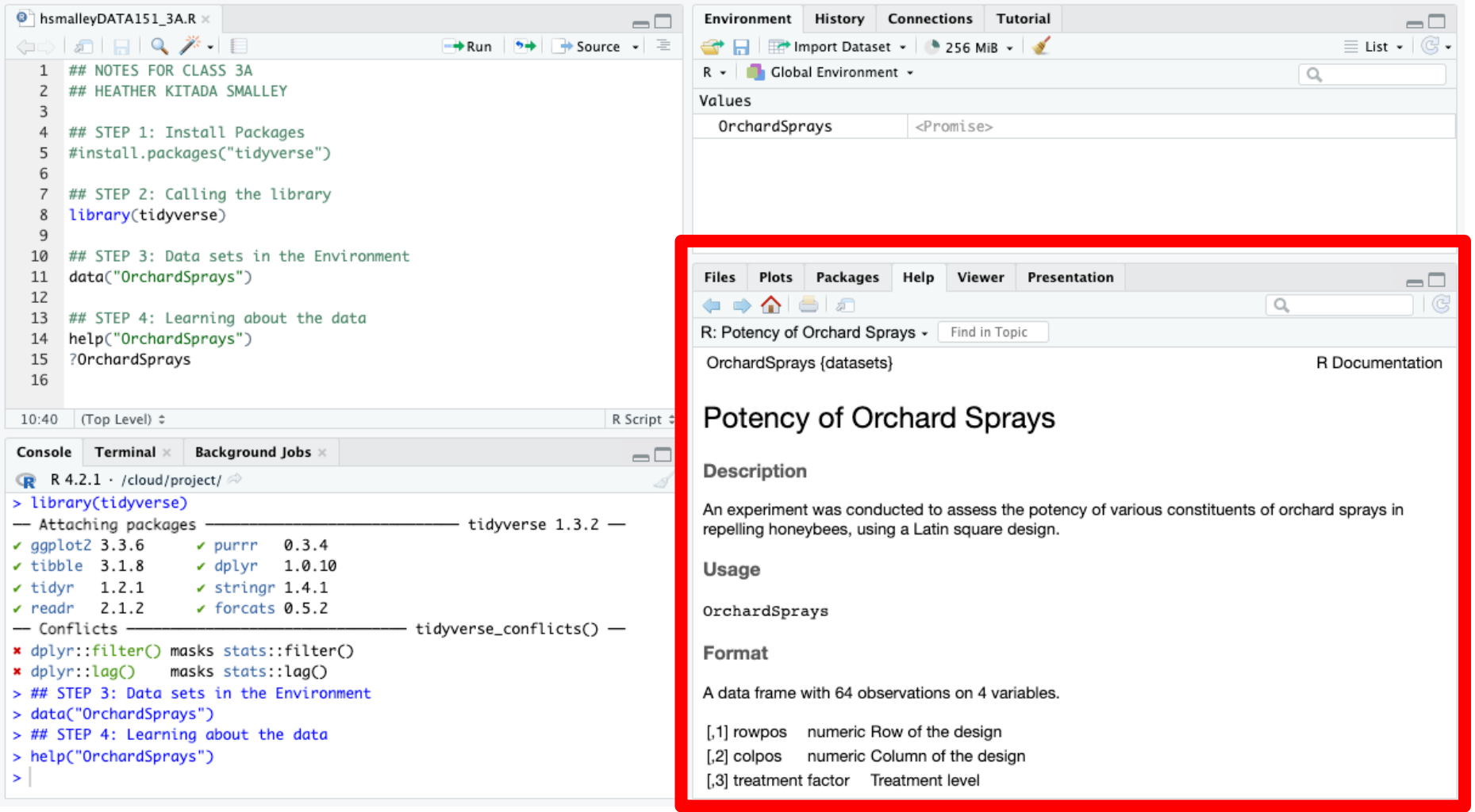


CALLING R BUILT IN DATASETS TO THE ENVIRONMENT



Observe that the data set is now available in the R environment

LEARNING ABOUT THE DATA



LEARNING ABOUT THE DATA

Details

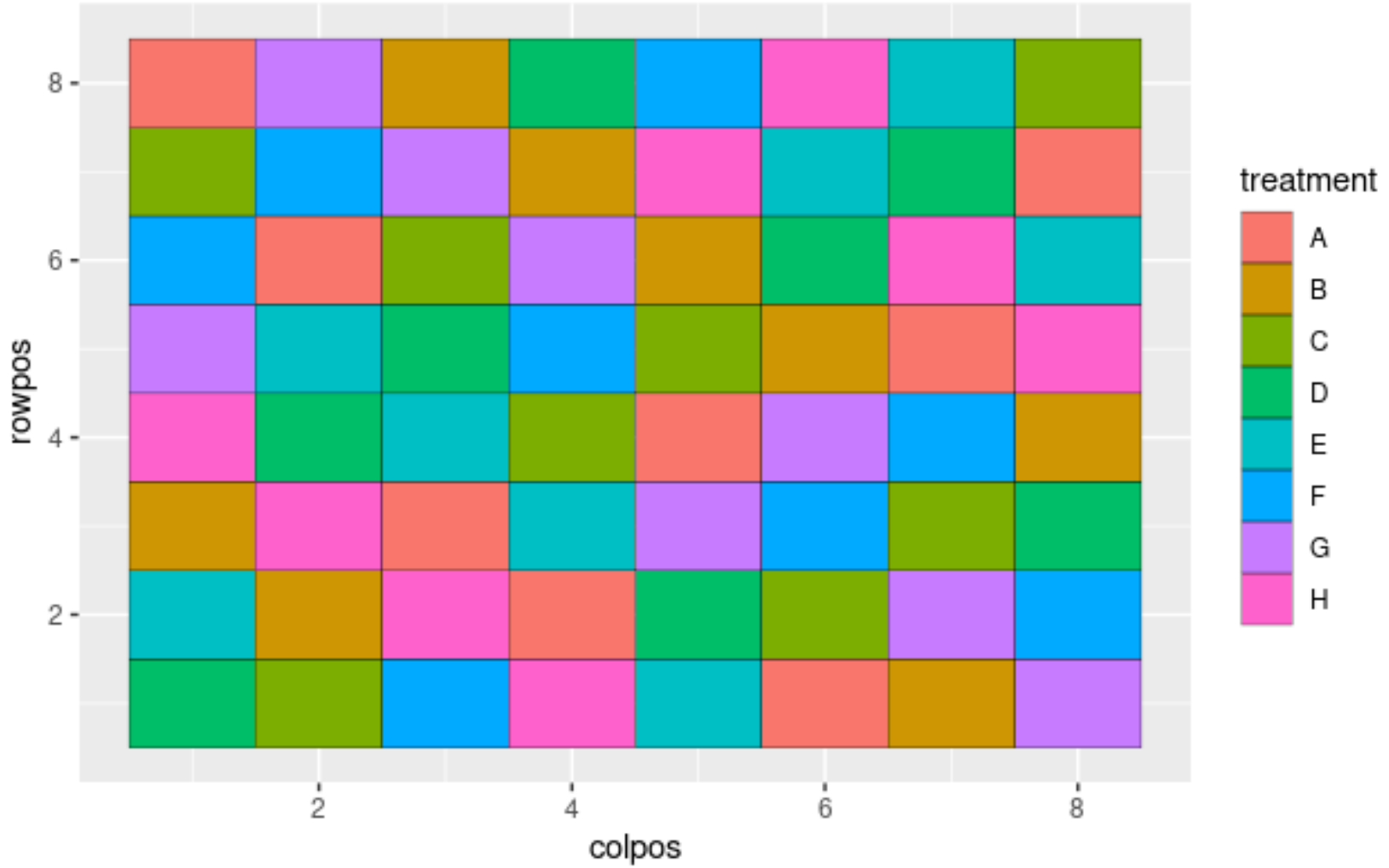
Individual cells of dry comb were filled with measured amounts of lime sulphur emulsion in sucrose solution. Seven different concentrations of lime sulphur ranging from a concentration of 1/100 to 1/1,562,500 in successive factors of 1/5 were used as well as a solution containing no lime sulphur.

The responses for the different solutions were obtained by releasing 100 bees into the chamber for two hours, and then measuring the decrease in volume of the solutions in the various cells.

An 8×8 Latin square design was used and the treatments were coded as follows:

A highest level of lime sulphur
B next highest level of lime sulphur
.
.
.
G lowest level of lime sulphur
H no lime sulphur

WHAT DOES THE EXPERIMENT LOOK LIKE?



WORKSHEET TIME!

Part I: Experimental Design:

1) What are the response and explanatory variables in this study?

2) Are the four principles of a randomized experiment present? Verify each and explain.

1.
2.
3.
4.

If the four principles are met, do you feel comfortable making a cause and effect conclusion?