ANTH 572S: Multivariate Statistics

Laure Spake

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Why statistics?

Statistics is the science/art/magic of learning insights from data

Statisticians collect, analyze, interpret data and communicate results of analyses

We use statistics every time we summarize or interpret data, and every time we use data to help us make a decision

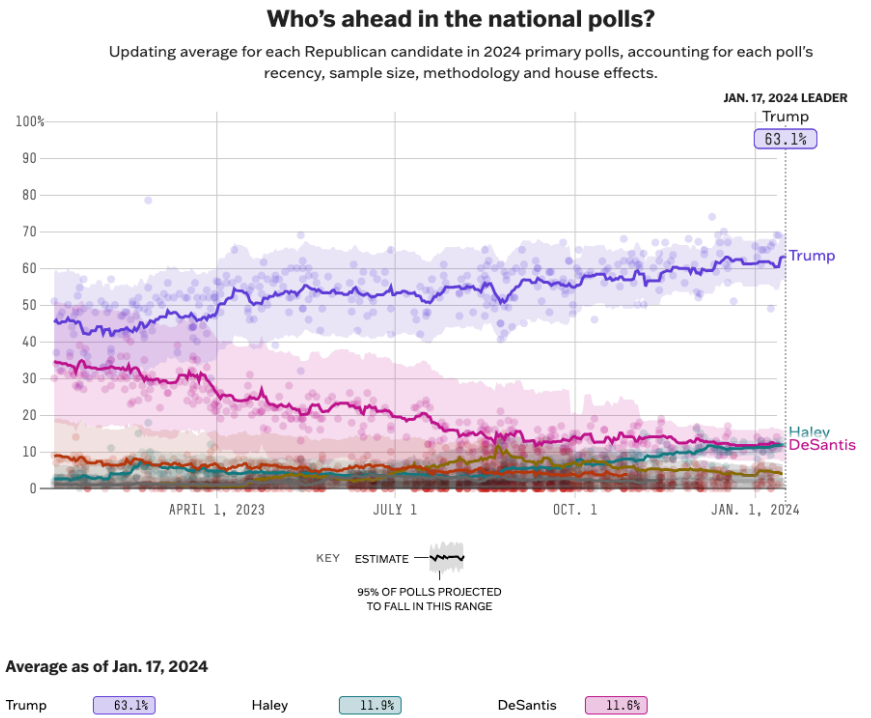
Data and statistics in a modern world

In academia and the workplace, massive datasets are constantly being generated and leveraged for insights

As the amount of data we collect increases, and we want to make more salient insights, we need to learn to use basic programming skills

Political Polling

Polls rely on samples of people from which data is collected, analyzed, and then extrapolated onto the voting population of the USA



What is this course?

This is a statistics class, but the purpose of this course is not to teach you statistics…

but rather how to think about data, shape it into analyzable form, and understand both the data and statistics mean

Learning goals

You will be introduced to:

- The R programming language
- Data management and wrangling
- Summary of basic statistics
- Data reduction
- And a lot more…

The goal of this course is to get you comfortable thinking, using, and interpreting data, while getting you up to speed with scientific programming principles

In this course, we will focus on the data analysis pipeline (and some R programming) because:

1. These teach critical thinking skills, which will serve you in any career path
2. These are more interesting than equations
3. Once you have mastered these skills, you can pick up specific statistical analyses

Course Outline

Schedule

- Weeks 1-3: Basics of R, data visualization, data wrangling
- Weeks 4-10: Review of fundamental concepts in statistics
- Week 12-14: Advanced concepts

Evaluation

- Assignments - 30% - 5 throughout the term
- Quiz - 20% - One quiz, testing core concepts in the course
- Term project - 50% - Analyze real-world data and report results

Instructor

Dr. Laure Spake
Biological anthropologist

Email: lspake@binghamton.edu
Office hours: Monday 3:00pm - 5:00pm, Science 1 Rm 219

About office hours

Please use them! I am here for you.

This is an *intensive* and *fast-paced* course. If you need help, or you are struggling to grasp a concept, please come see me early

There is access to computing resources during my office hours.

Course delivery

We will be using two main spaces to deliver this course:

1. *Brightspace* - this is where I will post documents such as lecture slides, the syllabus, review materials for the quiz, and instructions for the term project

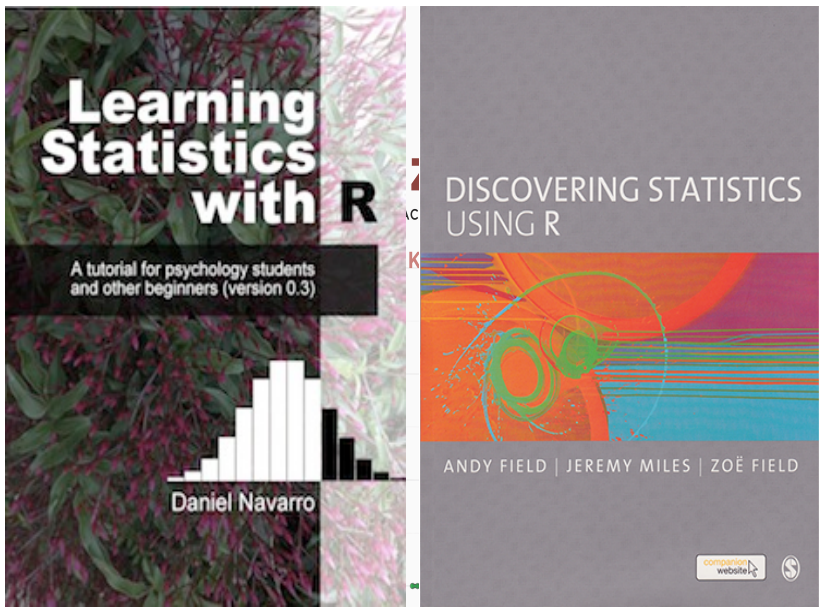
2. *Posit Cloud* - this is an integrated environment that allows you to use R and RStudio (more on this later) in your browser without needing to download software on your computer. You will use this to retrieve and complete in-class exercises, assignments, and the term project.

Note on classes over the semester

Your experience in this course will be 100% better if you do come to class and lab.

However, if for whatever reason you have to miss class, know that all materials you need to catch up are always posted on Brightspace/Posit Cloud. My slides and exercises are wordy, and you can always come to office hours to ask questions

Your textbooks



Sharing and reusing code

- There is lots of code available to you on the internet, and unless an assignment specifically says not to, you may use it to help write your own code
- If you do use code from an online source, please cite it at the end of your documents
- Any code used but not cited is considered plagiarism (CC-BY licenses)

Sharing and reusing code

- You are welcome to work together to solve any of the assignments or exercises, but please submit your own assignment/exercise files
- You may not copy or make direct use of code from another student
- You may not copy or make direct use of code from another group for the term project

Syllabus finalization

The syllabus has several topics designated as *TBC* - let's take a minute to finalize those now

Introduction to the R computing environment

Learning objectives

At the end of this lesson you will:

- Differentiate between R and RStudio.
- Understand and navigate the different components of RStudio.

What is R?

- **R** is a programming language that is based off S, a statistical programming language developed in the 1970-90s
- **R** was created in 1991 and released to the public under a free and open license in 2000
- **R** is fully **open source**, meaning that anyone can access and modify the source code
- **R** has industry-leading graphics capability, which makes it super popular for creating scientific graphics

Because R is open-source, developers are able to create and share packages (add-ons) to facilitate complicated statistical procedures

From a scientific computing perspective, this is super useful!

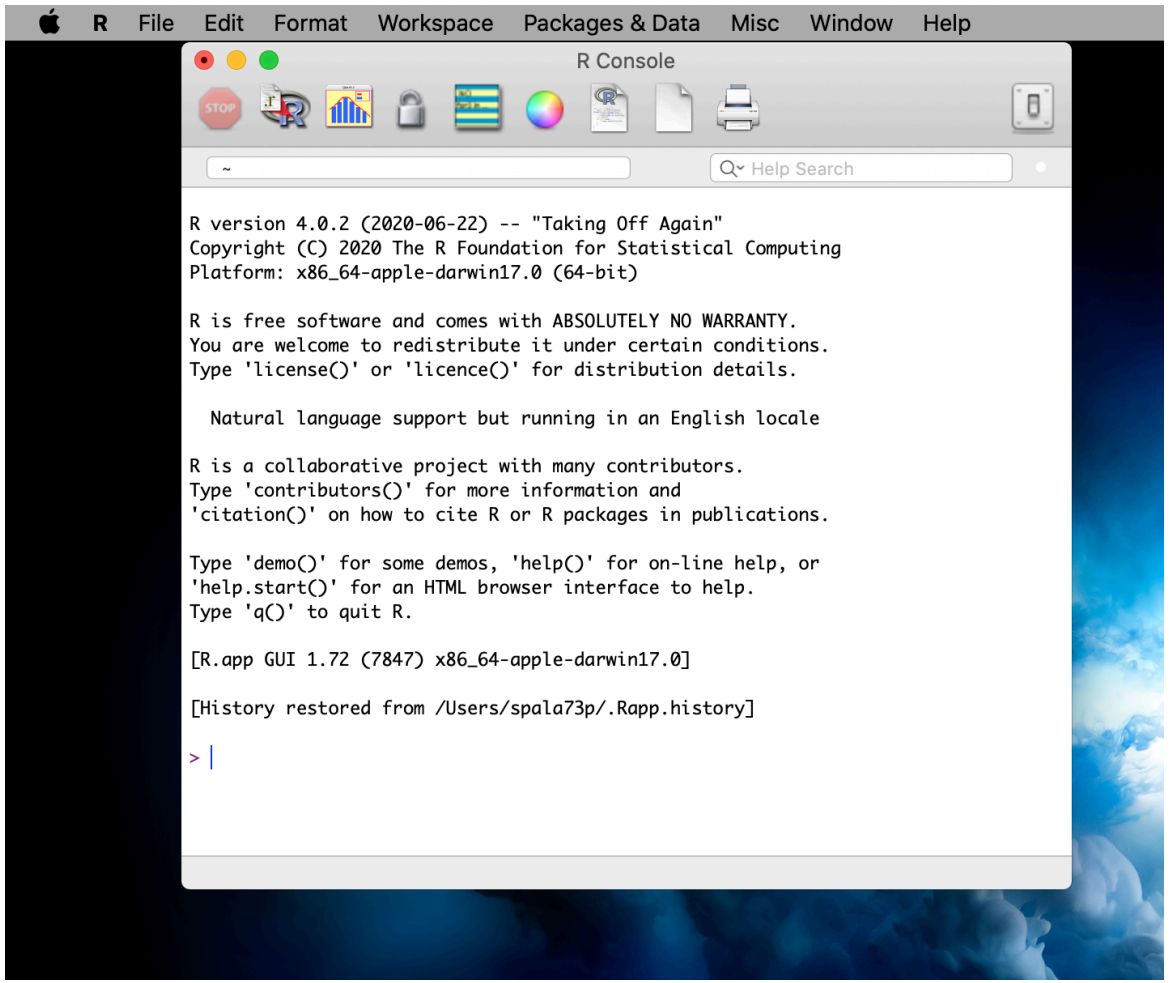
The R system

There are two “components” of R:

- R itself, also known as *base R*, is distributed freely on CRAN
- Packages that extend functionalities:
  - Some are available on CRAN (10,000+)
  - Many more are in development or not reviewed by CRAN and available on personal websites or GitHub (number unknown)

Working in R

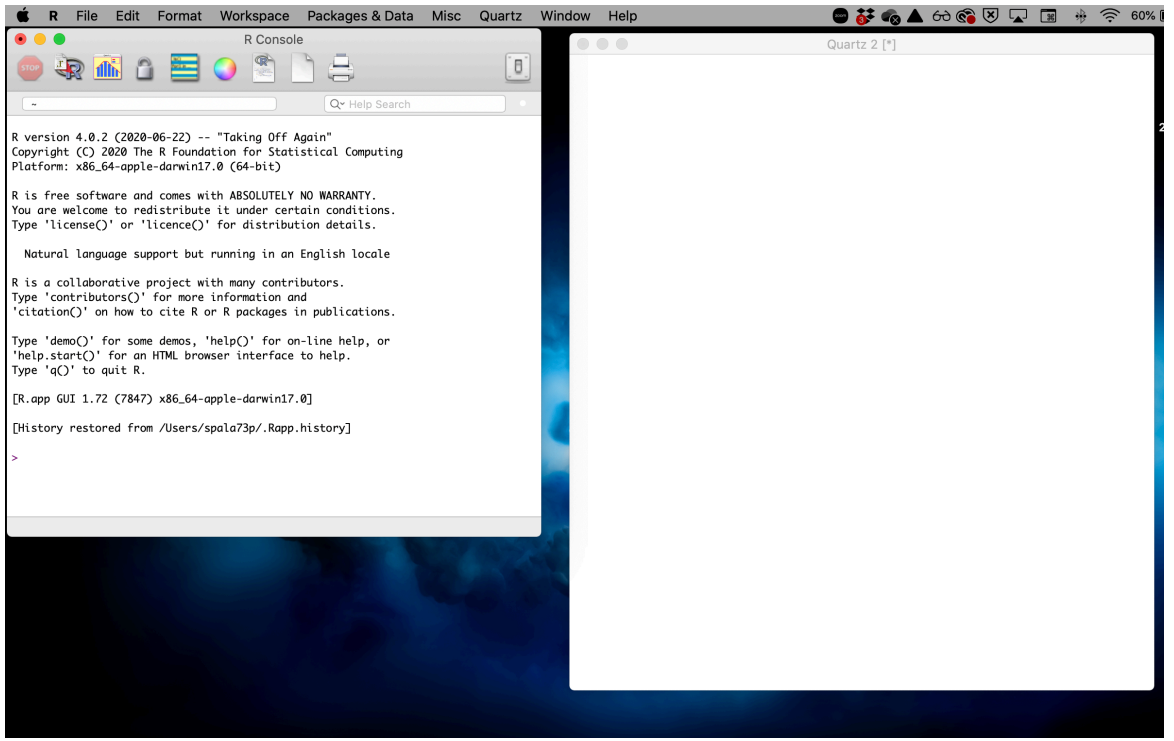
Base R has a very bare interface



Working in R

Base R has a very bare interface

If you wanted to work with plots, you'd need to open a separate window

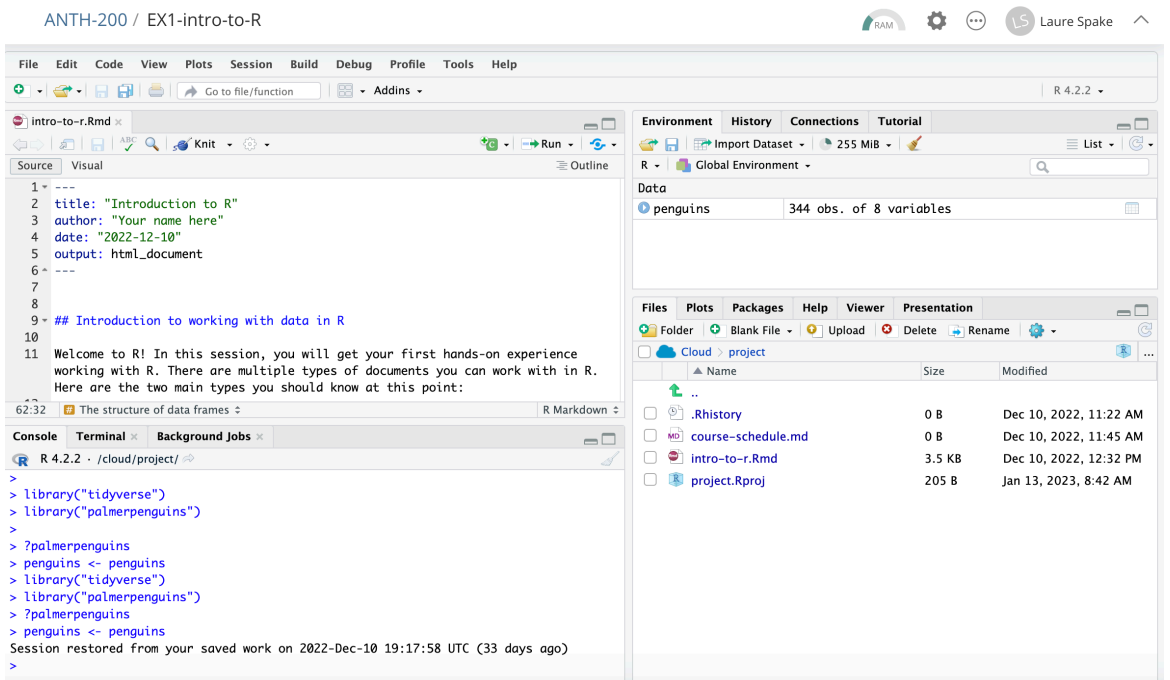


Working in RStudio

For this reason, most people who work in R work in a text editor.

The most popular is RStudio.

This is what we'll be using.

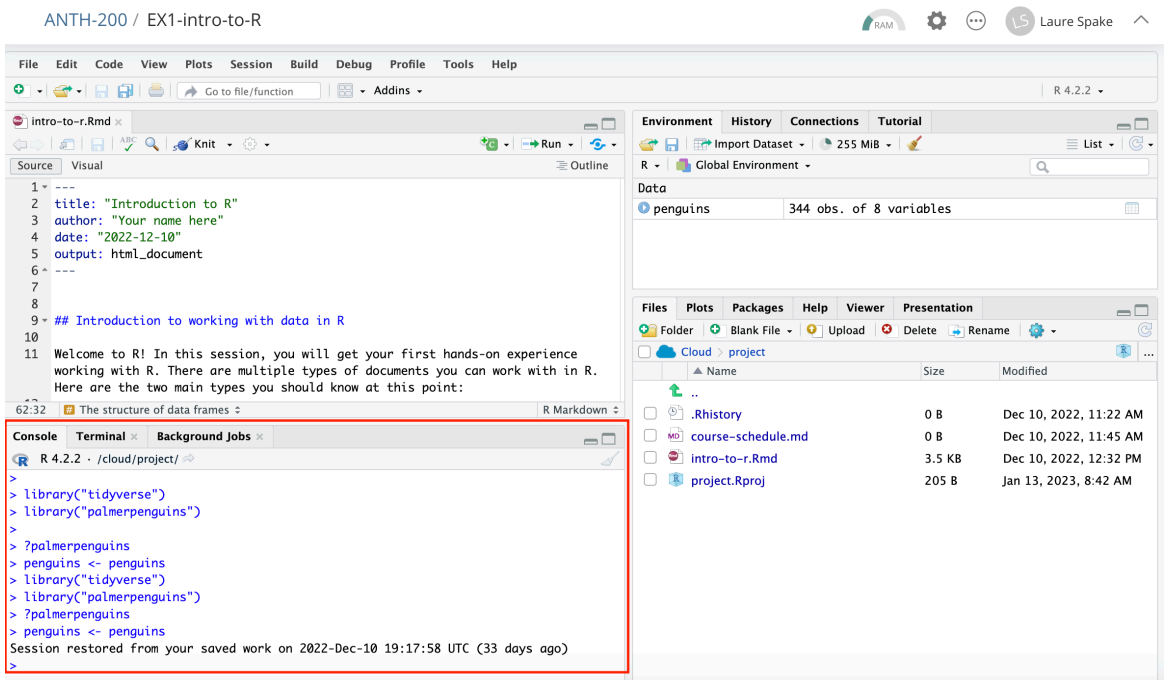


Getting started with RStudio in the cloud

https://posit.cloud/spaces/541752/join?access_code=FEg-W10feqS6598Gy9MoREz1QVKNhyZAaup7m2XE - Last Modified Aug 22, 2024 6:52 PM

The console

This is where all your computations happen



The console

```
1  5 + 2
```

```
[1] 7
```

The console

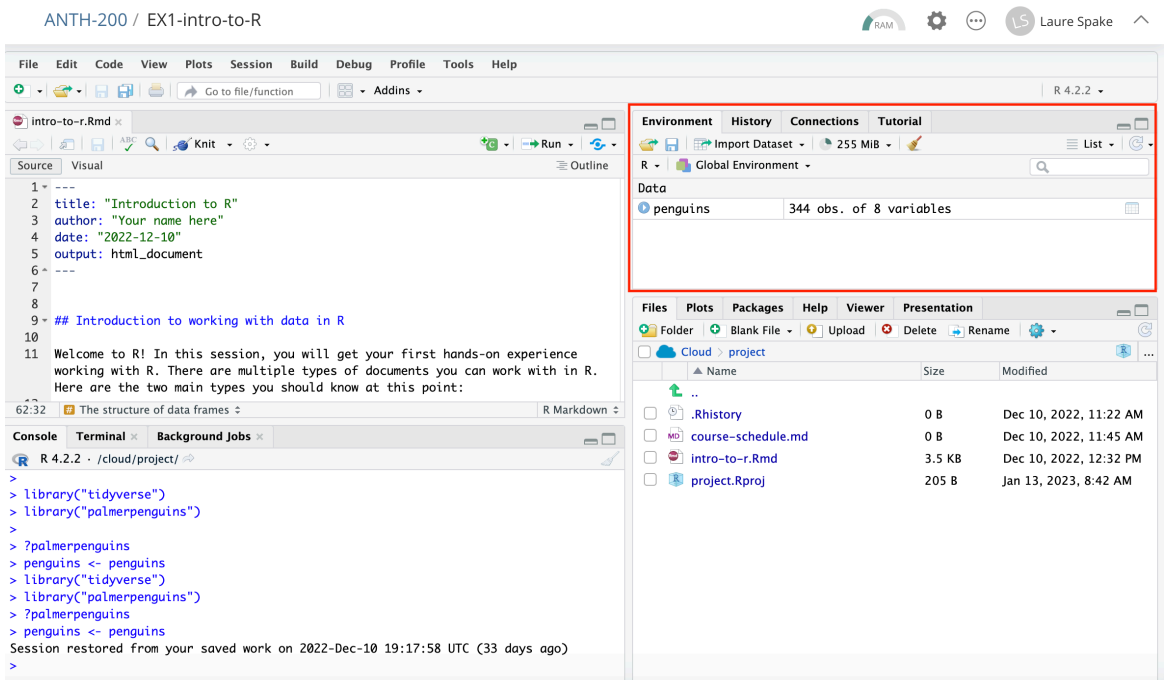
```
1 data(mtcars)
2 head(mtcars)
```

```
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

The enviroment

This is where all the objects available to you are displayed.

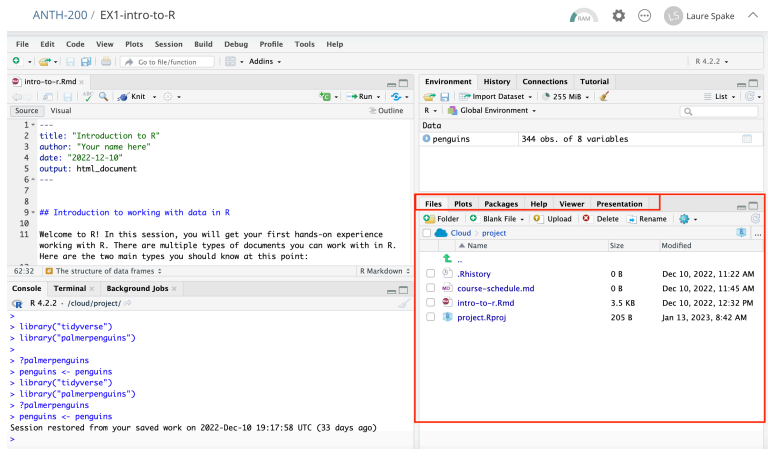
These objects can be datasets, lists, vectors, models, etc (more on this later)



The viewer

The viewer pane has many sub-panels - you can use it to flip between seeing your files and your plots.

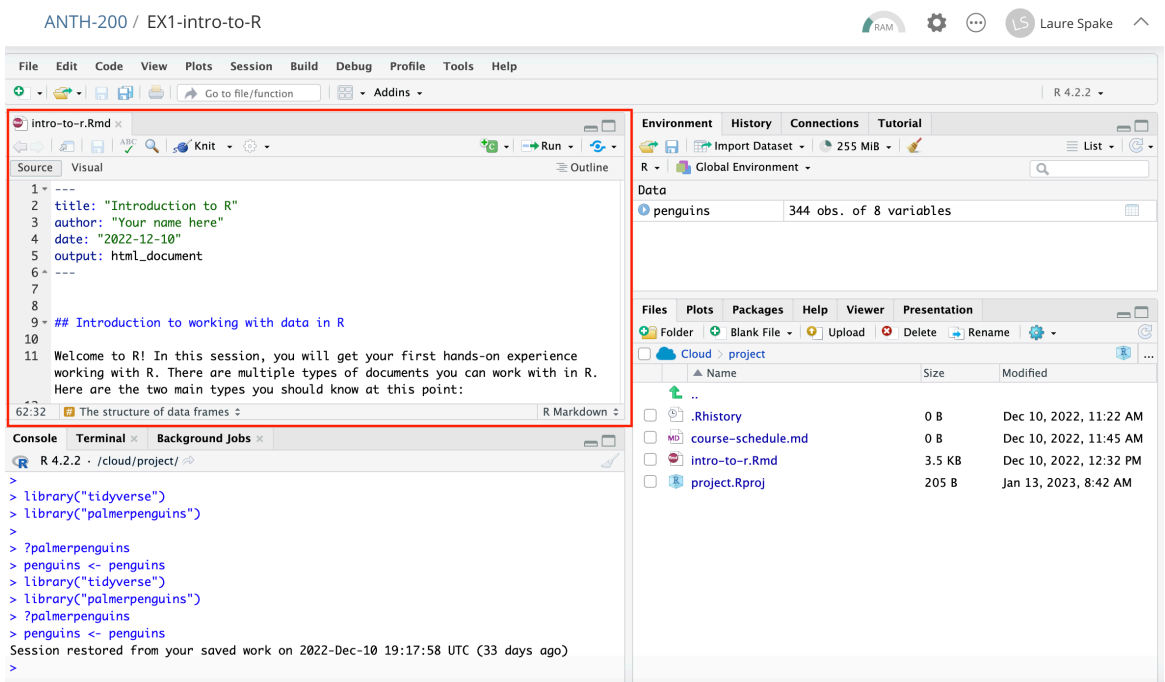
You can also use it to see which pacakges you have installed and access the documentation for each package/function you use.



The script(s)

When you write code that you want to save, you can develop it in a script.

There are multiple types of scripting files - more on that later.



Summarizing

Learning objectives

At the end of this lesson you will:

- Differentiate between R and RStudio.
- Understand and navigate the different components of RStudio.

Break!

Get a stretch, we start working in R right afterwards.

Attribution

Content for parts of this lesson was adapted from:

- Stephanie Hicks' Intro to R
- Mine Cetinkaya's Data Science in a Box
- Nicholas Tierney's Introduction to Data Analysis
- RD Peng's Biostatistics Lectures