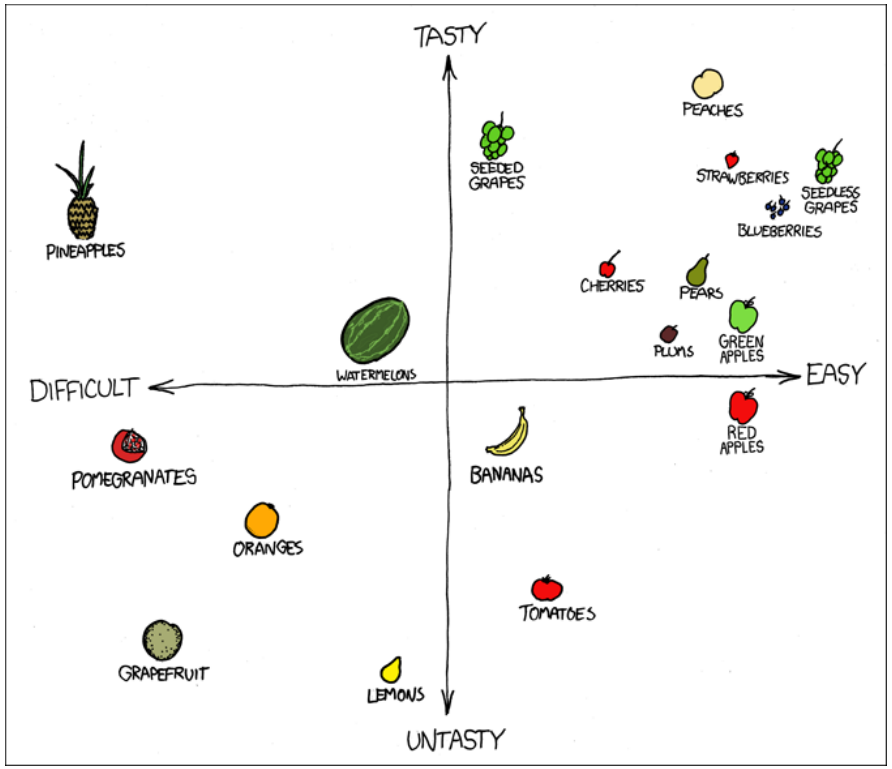
Introduction to Data Display and Analysis



Overview

Introductions

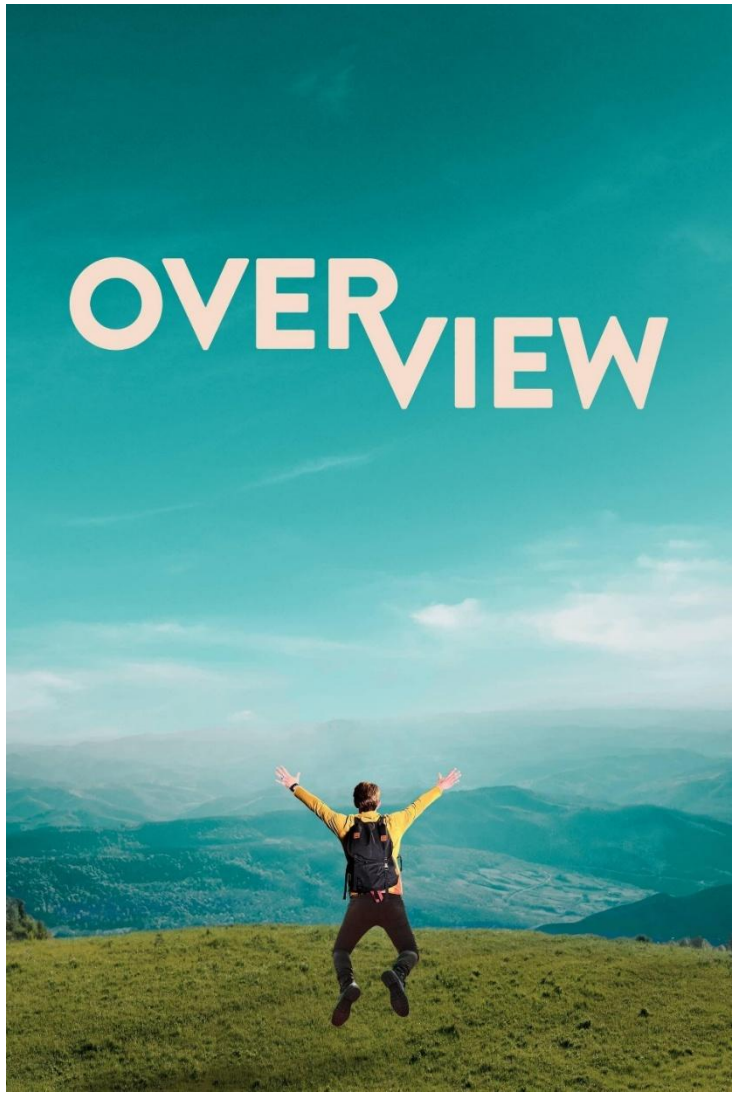
Overview of the class and logistics

Introduction to R

- R as a calculator
- Dat types, functions, and vectors
- Packages and the SDS111 package

If there is time

- Quarto
- Data Frames



Introductions and course logistics

Contact information and office hours

Contact email: ethan.meyers@yale.edu

- Please only send me emails for questions that only pertain specifically to you
- For questions about course content, please use Ed Discussions

Note: I'm going to be a little distracted by a particular neural network…

Office hours:

- Tuesdays and Thursdays, 10:45-11:45

Teaching Assistants

Undergraduate Learning Assistants (ULA)

- Sonam Wangchuk: sonam.wangchuk@yale.edu

Office hours

- Sundays and Mondays 8-10pm, location WLH 009
- Possible additional office hours if requested

Introductions

Let’s do some quick introductions

Please say:

- Your name
- Where you are coming from
- Why you are interested in this class
- Anything else you would like to share with your group

What is this course about?

Course objectives

The objective of this course is to learn how to extract insights from data and convey these insights to others

In particular, you will learn how to use the R programming language to visualize, summarize, and clean data

- The emphasis will be on practical data manipulation and interpretation

The intended audience is students planning to start in S&DS 1000 and S&DS 1230 or other introductory statistics courses at Yale during the upcoming school year

Course plan

1	July 1-3	Basics of the R programming language
2	July 8-10	Data visualization
3	July 15-17	Descriptive statistics and data manipulation
4	July 22-24	Interactive graphics
5	July 29-31	Final exam and class debriefing

Examples of questions we might look at…

R Basics: How much money does Elon Musk make every second?

Data visualization: What industry have most billionaires made their money in?

Data summarization: which airlines have the longest flight delays?

Scrollytelling: Create a webpage that walks you through insights extracted from data

WARNING: There will be a lot of bad jokes

Class logistics

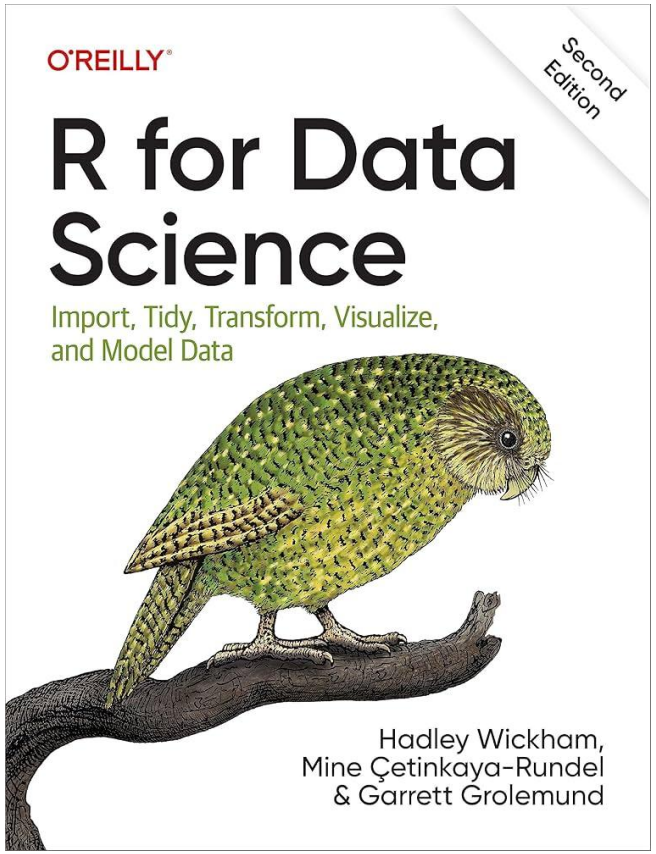
Class time 9-10:15am Tuesdays and Thursdays

- New content introduced, questions answered

Canvas website:
https://yale.instructure.com/courses/108275

The course textbook is R for Data Science by Wickham, Cetinkaya-Rundel and Grolemund

- You should have been given a copy of this book
- We will only cover a small portion of this book, but it will be a useful resource if you want to learn more



Office hours

My office hours are after class from 10:45-11:45am

- i.e., Tuesday and Thursday, 10:45-11:45am in WLH 211

For questions about content covered in class, best to first ask on Ed Discussion

- Class participation grade based on questions and answers on **Ed Discussion**

Workshops

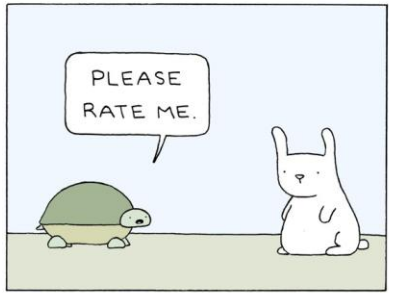
Sonam will run workshops where you will gain additional practice with the course material

Workshops will be on Tuesdays and Thursdays in CTL 118A

Workshop assignments

- **3-4pm**: Mayeesha, Seamus, Valerie, Faizah, Sabrina, Camila, Ida
- **4-5pm**: Rocio, Evan, Ivan, Naila, Patrick, Wendy

Assignments and grades

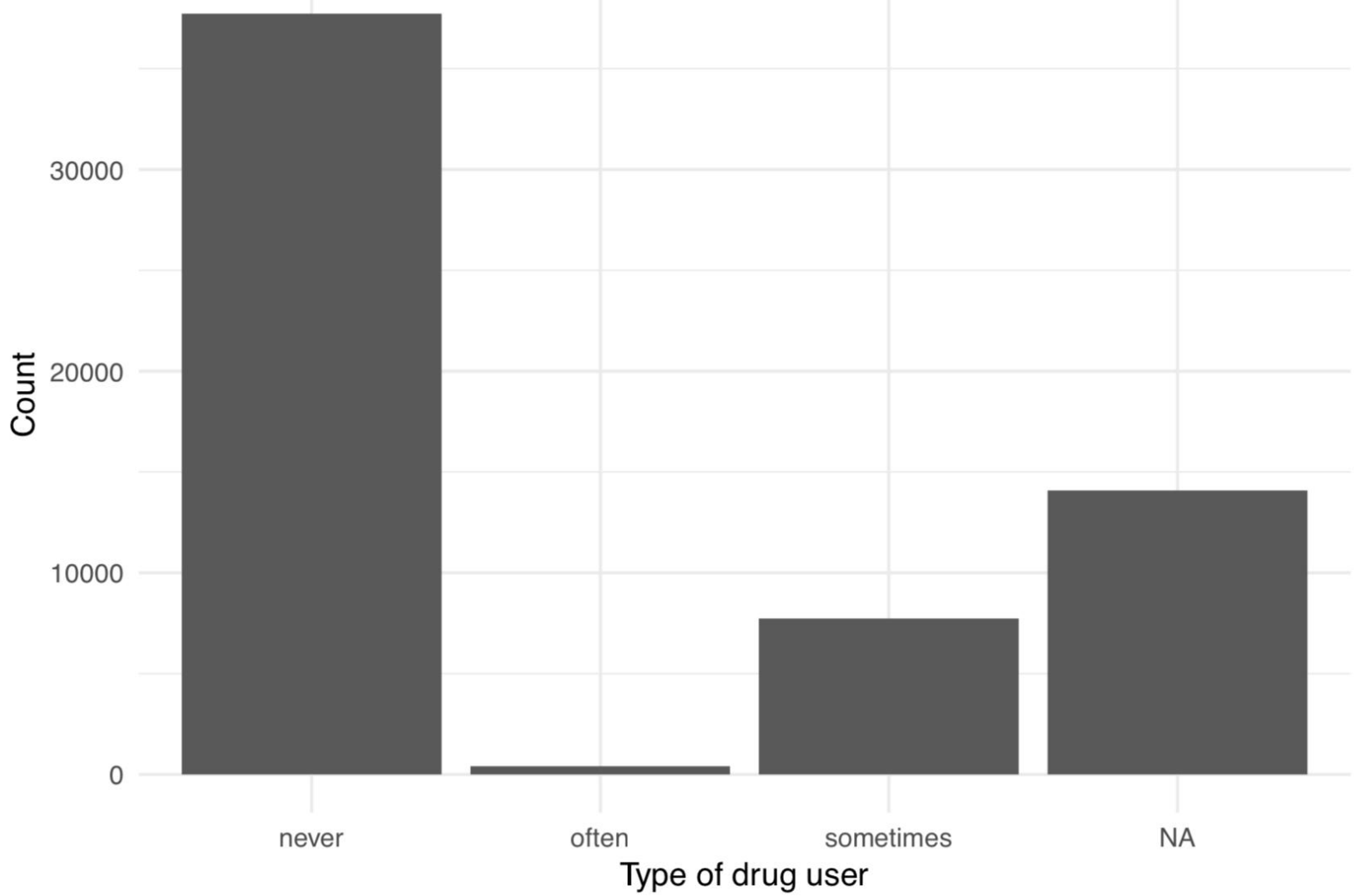


1. Homework problem sets (60%)
   - Analyze data using R and explain results
   - Weekly: 4 total

Homework policies

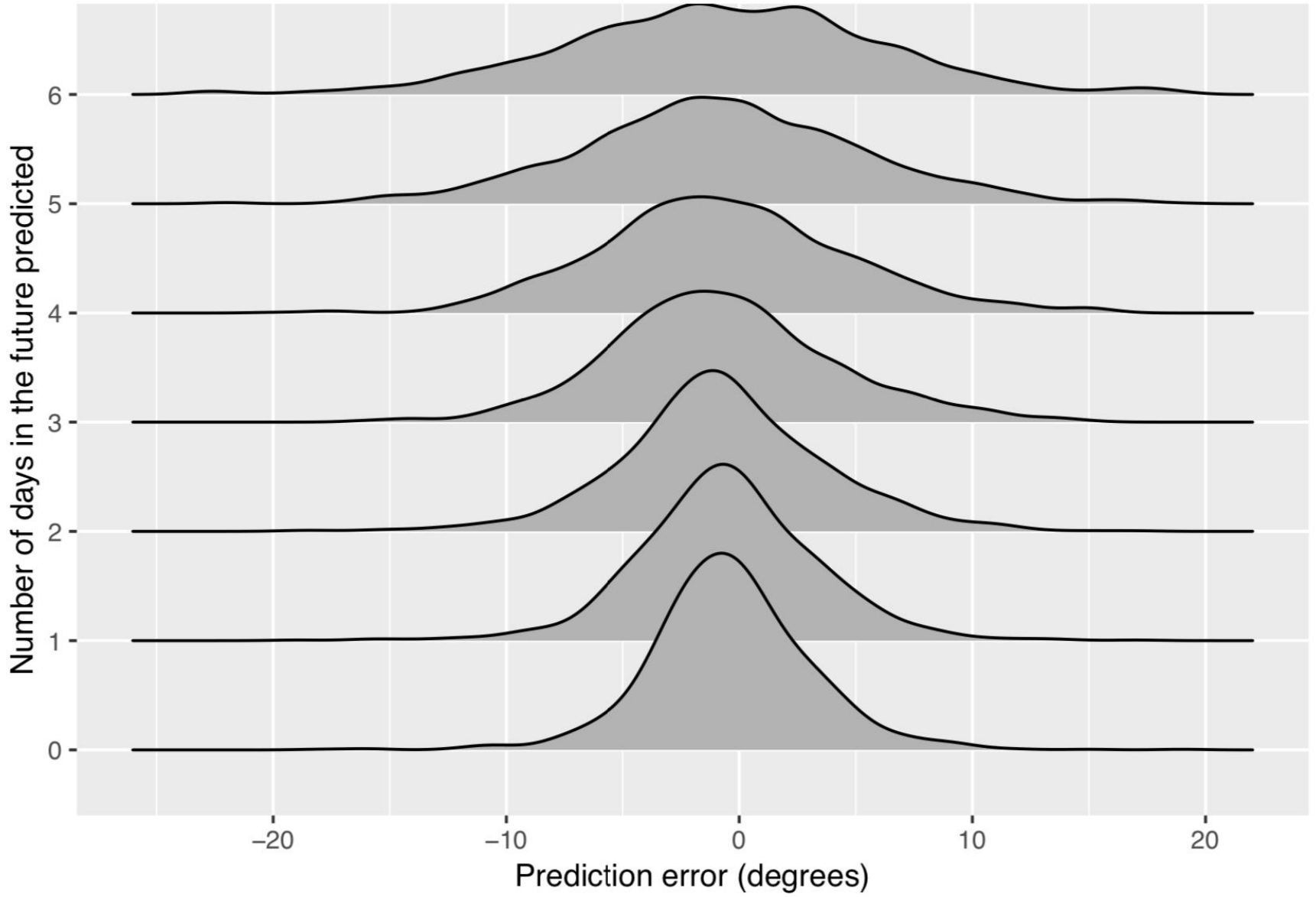
- You may discuss questions with other but the work you turn in must be your own
- Homework assigned on Tuesdays and are due at 11pm on Mondays
  - (with a 59 minute grace period)
- Late homework (90%) credit if turned before class on Tuesday (before 9am)
  - For any other extension a Dean's Extension is needed

Example homework assignment piece




```
# Bonus: create a pie chart of the self reported frequency of
# drug use and make it look good!
profiles %>% count(drugs) %>% filter(!is.na(drugs)) %>% ggplot(aes(x = "",
    y = n, fill = drugs)) + geom_col(width = 1) + coord_polar(theta = "y") +
    theme_minimal() + theme(axis.title.x = element_blank(), axis.text.x = element_blank(),
    axis.ticks.x = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()) + xlab("")
```

Example homework assignment piece



Answers: Personally I like the joy plot best here because it most clearly shows how the distribution becomes more spread out for predictions made further in the future (although all three plots do a reasonable job of showing this).

Final exam

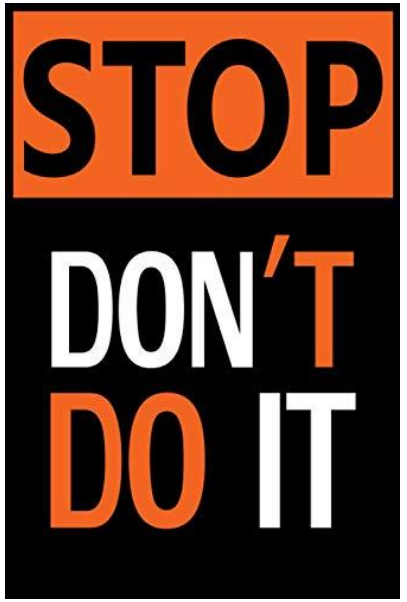
2. Final Exam (38%)
   - There will be a final exam at the end of the course
   - The exam will be a paper exam during regular class time
   - We will also have an exam/course debriefing on the last day of class

3. Participation (2%)
   - Asking and answering questions on Ed Discussion

Academic honesty

Plagiarism/cheating

- Yale's Academic Integrity Statement



You are allowed to talk with others about the homework, but the work you turn in must be your own

- Do not share answers
- Do not copy answers off the Internet

ChatGPT and other LLMs

You can use as a reference

- E.g., "What does the max() function do?"
  - i.e., ok to use it like Google/Stack Overflow

Do not use it to answer full questions

- i.e., do not type a homework question in chatGPT

To be an efficient data analyst, it's important to be fluent with the material

- And if you don't learn the material, you will be in a lot of trouble on the exams

Class background survey

In order for me to get to know you and to better adjust the class to your interests, please fill out the class background survey on canvas

- Under the Quizzes link on the left on Canvas


ANY QUESTIONS?

R Basics

Was everyone able to log into the YCRC RStudio server?

- If not, an alternative is to install R and RStudio on your own computer

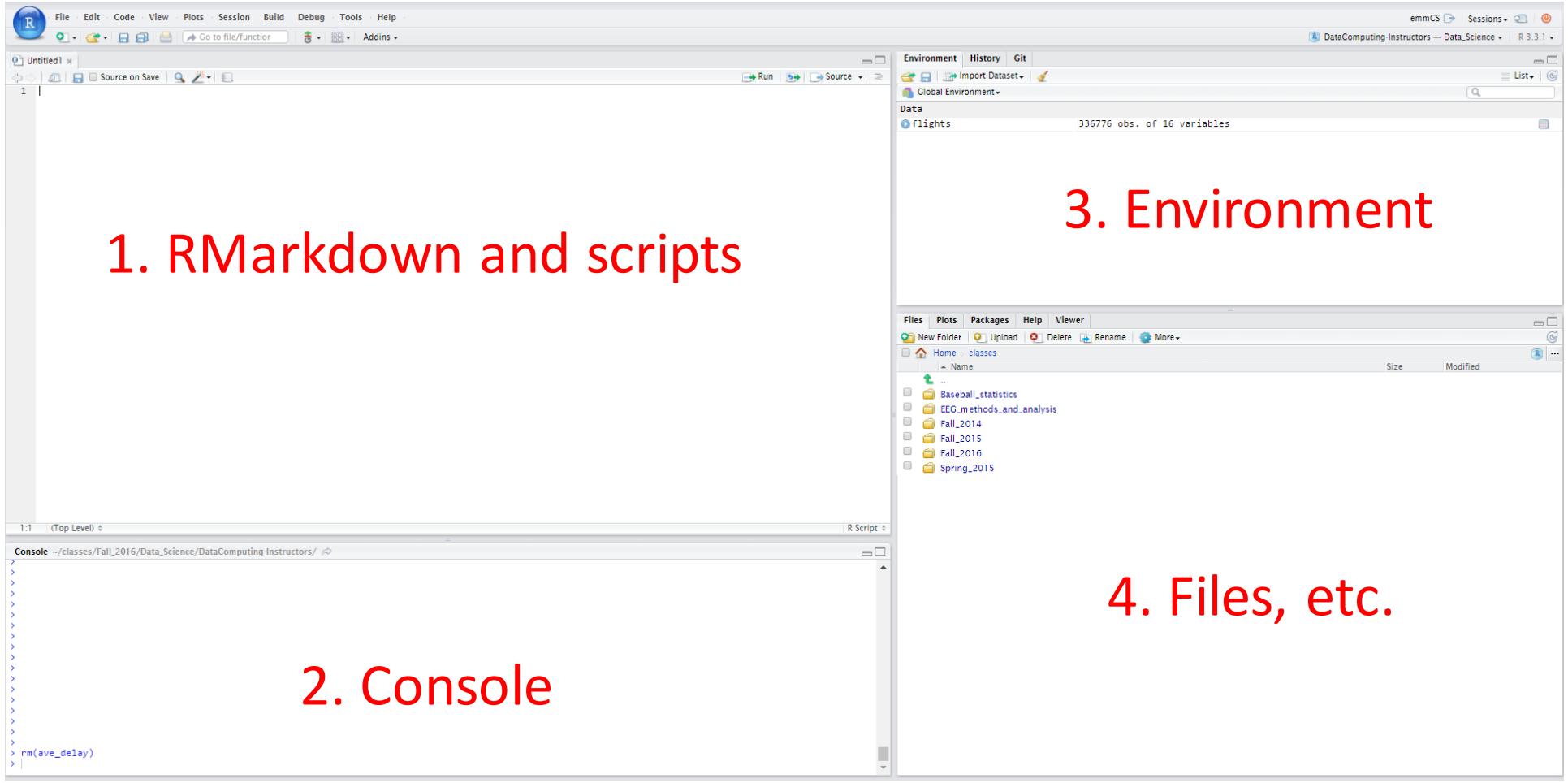
Let's take a 2 minute break and open R Studio and follow along…

R and R Studio

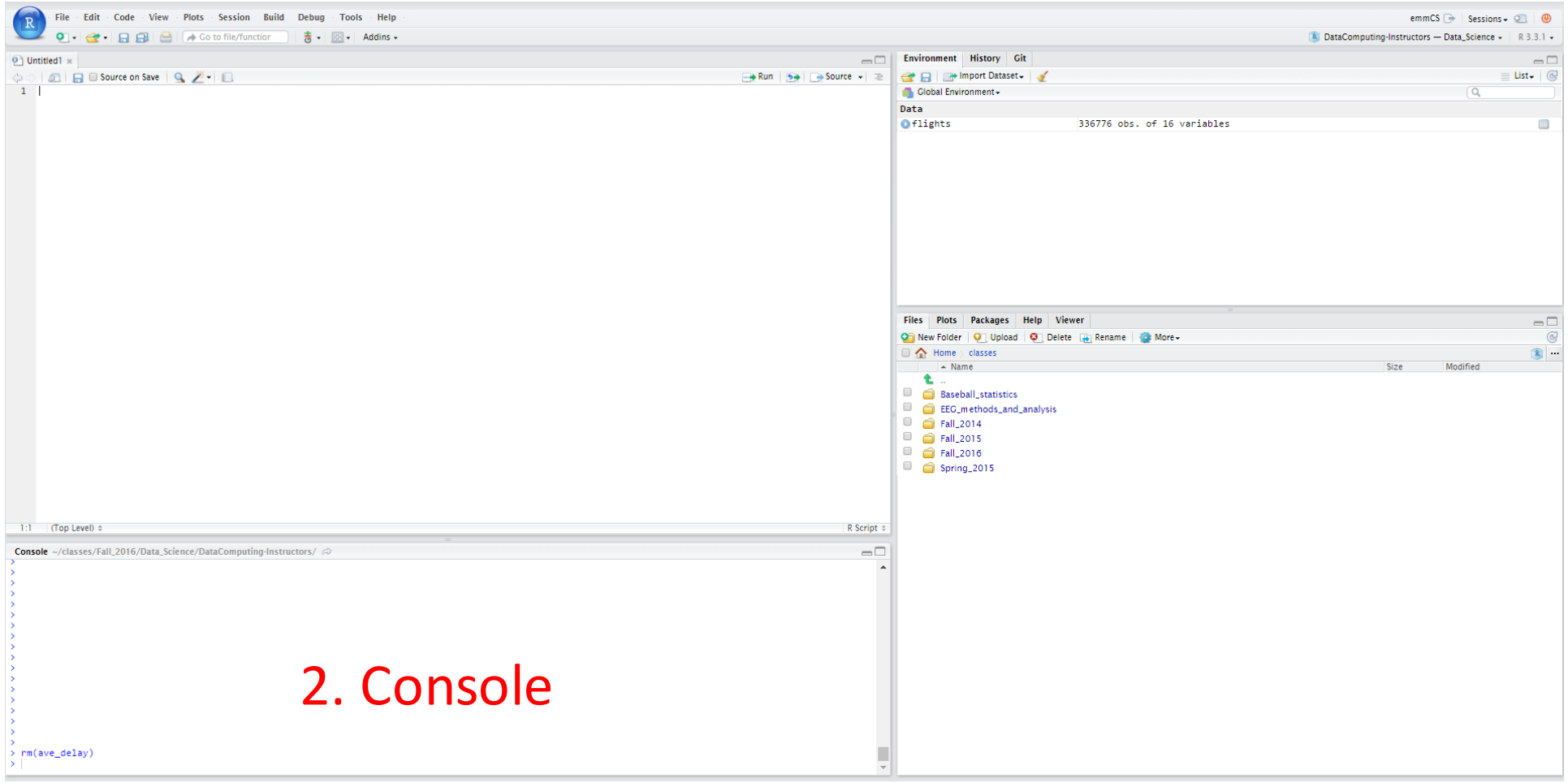
R: Engine

RStudio: Dashboard

RStudio layout



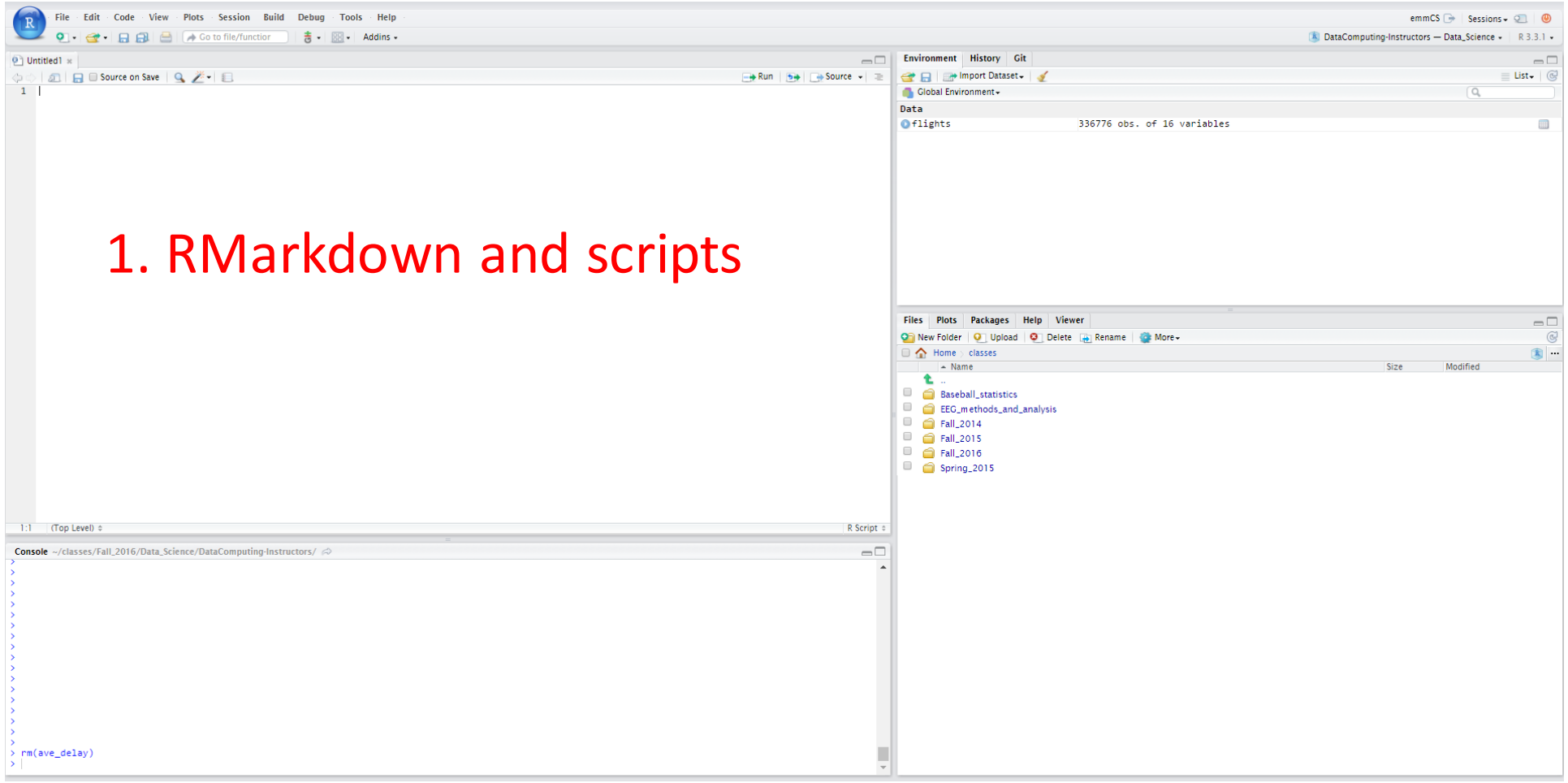
RStudio layout



R as a calculator

```
> 2 + 3
> 7 * 5
```

RStudio layout



Create a new script

File -> New File -> R Script

Save the script with a reasonable name, e.g., week1_notes.R

R Basics

Arithmetic:

```
> 2 + 2
> 7 * 5
```

Assignment of values to ***objects***:

```
> a <- 4
> b <- 7
> z <- a + b
> z
[1] 11
```

Number journey…

Number journey

```
> a <- 7
> b <- 52
> d <- a * b
> d
[1]  364
```

Character strings and Booleans

```
> a <- 7
> s <- "s is a terrible name for an object"
> b <- TRUE

> class(a)
[1] numeric

> class(s)
[1] character
```

Functions

Functions use parenthesis: functionName(x)

```
> sqrt(49)
> tolower("DATA is AWESOME!")
```

To get help

```
> ? sqrt
```

One can add comments to your code

```
> sqrt(49)    # this takes the square root of 49
```

Vectors

Vectors are ordered sequences of numbers or letters

The c() function is used to create vectors

```
> v <- c(5, 232, 5, 543)
> s <- c("statistics", "data", "science", "fun")
```

One can access elements of a vector using square brackets []

```
> s[4]        # what will the answer be?
```

Vectors continued

One can also apply functions to vectors

```
> z <- 2:10
> sqrt(z)
> min(z)
> which.min(z)
```

We can also add and subtract vectors of the same length

```
> v1 <- c(2, 4, 8)
> v2 <- c(1, 2, 3)
> v1 + v2
```

Questions?

R packages

R packages

Packages add additional functionality to R

We will use many additional packages in this class

- ggplot2, dplyr, tidyr, etc.

There is a class specific package (SDS111) I wrote that you can use to download homework and other files

- All class materials are also on GitHub: https://github.com/emeyers/SDS111

SDS111 package

If you are using R and Rstudio on your own computer, instructions to install the SDS111 package are at: https://github.com/emeyers/SDS111

- Sonam and I can help you with this too during office hours

If you are using RStudio on the YCRC cluster, the SDS111 package is already installed

To use the functions in the package you can use

```
library(SDS111)
download_homework(1)
```

For next class

If you have not done so already

1. Fill out class survey on Canvas under the Quizzes link

2. Optional
   - Install R/Rstudio, and the SDS111 package on your own computer

Questions?

If there is time…

Quarto

Quarto

Quarto (.qmd files) allow you to embed written descriptions, R code and the output of that code into a nice looking document

Creates a way to do reproducible research!

Qaurto

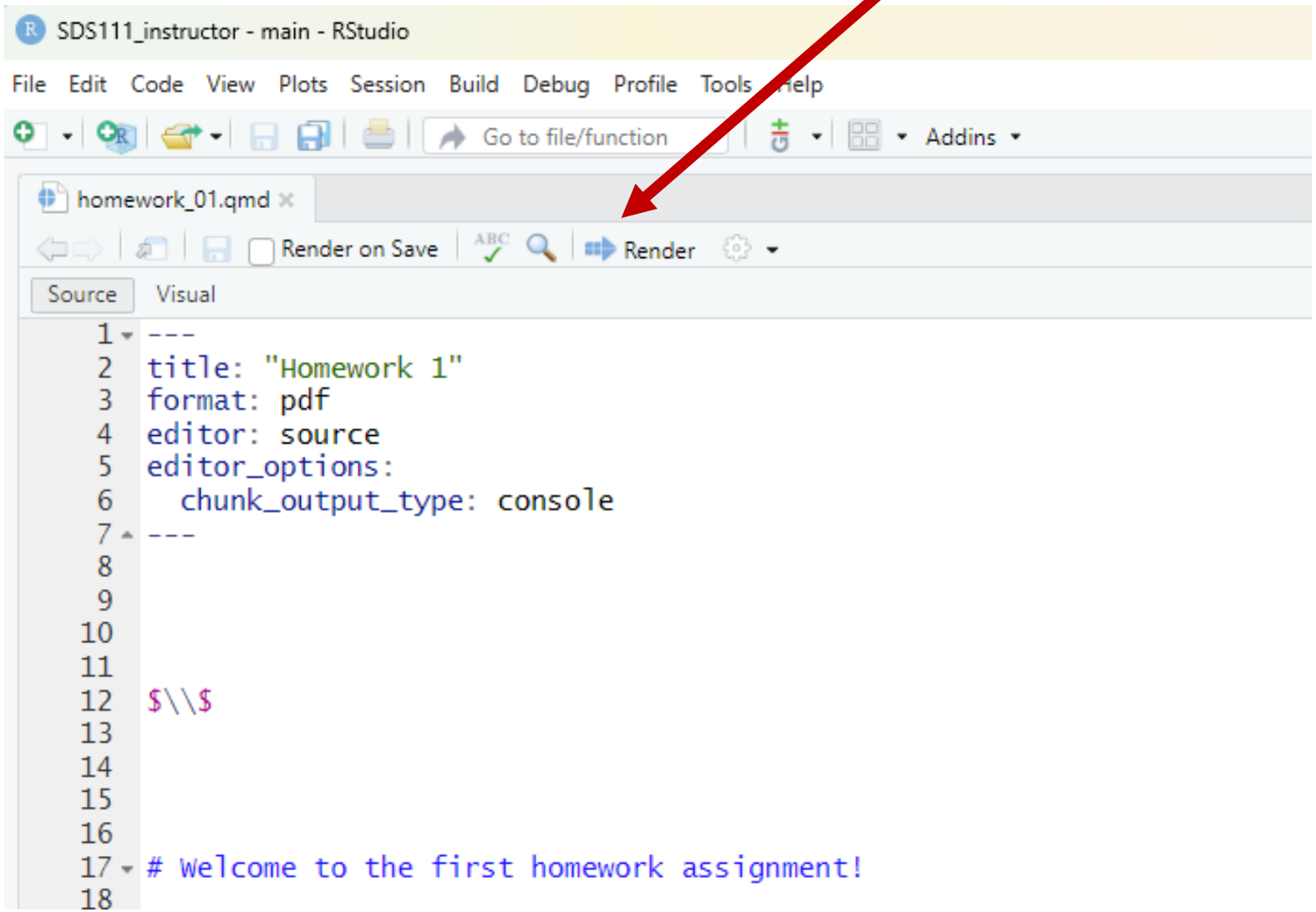
Everything in R chunks is executed as code:

```{r}
    # this is a comment
    # the following code will be executed
    2 + 3
```

Everything outside R chunks appears as text

Render to a pdf

Turn in a pdf or html document with your solutions to Canvas



Quarto

Note: When you render a Quarto document, your Quarto document **does not have access to variables in the global environment**, but instead have their own environment.

Why is this a good thing???

Formatting in Quarto

We can add formatting to text outside the code chunks

Examples:

```
## Level 2 header
**bold**
![](https://statistics.yale.edu/sites/default/files/logo2.png)
```

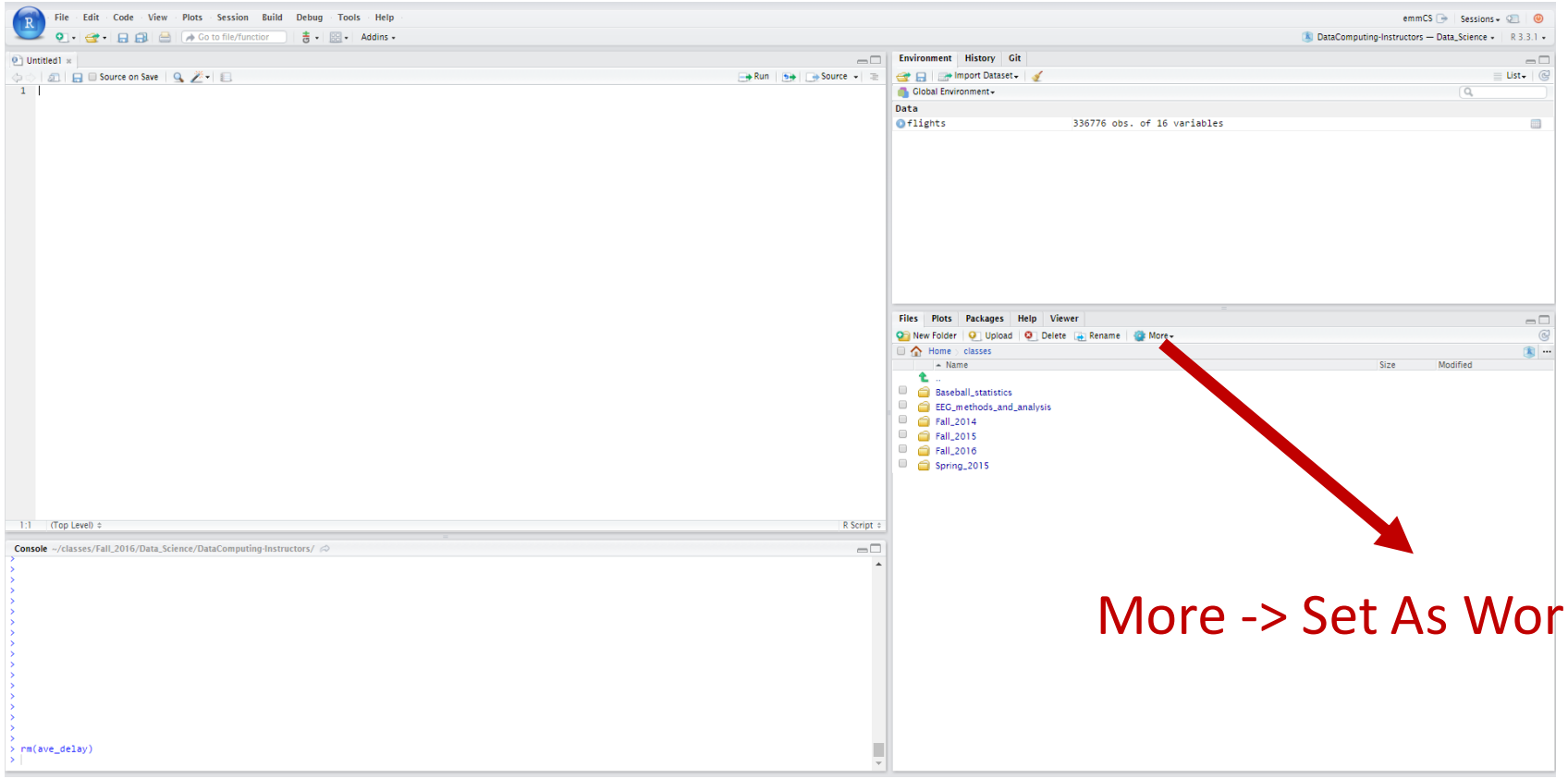
To repeat: avoid hard to debug code!

Only change a few lines at a time and then render your document to make sure everything is working!

If you document isn't rendering:

- **For code chunks:** use the # symbol to comment out code until you can find the line of code that is giving the error message
- **Outside of code chunk:** cut out part of the document until it renders and then paste it back

Setting your working directory



More -> Set As Working Directory

1. In the files tab, navigate to the directory that contains the .qmd file
2. Click More -> Set As Working Directory

Practice!

I highly recommend reviewing everything and experimenting with the code so that you feel well prepared for next class

The workshops with Sonam should also be very helpful!

You can start on the first homework using:

```
SDS111::download_homework(1)
```