Week 1: Jupyterhub, R, Rstudio, and the Tidyverse

Statistical Reasoning and Data Science

Scott Schwartz

May 18, 2021

Statistical Reasoning and Data Science (DS)

Modern Stats+DS is

1. software, programming, and computational tools implementing
2. mathematical and algorithmic data analysis methodologies for
3. facilitating data-driven and evidence-based decision making

1. software, programming, and computational tools

Modern Stats+DS is based on using computational tools

- Programming IS required for DS and applied stats
- Programming isn't a prerequisite: wanting to learn to code is

Class Check Round 1

https://pollev.com/sta (2 questions)

Statistical Reasoning and Data Science (DS)

Modern Stats+DS is

1. software, programming, and computational tools implementing
2. mathematical and algorithmic data analysis methodologies for
3. facilitating data-driven and evidence-based decision making

2. math, algorithms, and data analysis Part I: Structured Learning

When first learning, organized content like our course content is best

Attend (2hr) Lectures	Monday 10:10AM ET and 2:10PM ET	N/A
R/Communication HW	Due before start of tutorial	**15%**
Attend (2hr) Tutorials	Friday 10:10PM ET and 2:10PM ET	**15%**
Study and Review	Slides, Demos, Questions	N/A
Midterm/Final Exams	Concepts, communication, and coding	**45%**
Coures Project	Demonstrate your developed skills	**20%**
Participation	Demonstrate your developed skills	**5%**

Boost your skills with free textbook resources like R for Data Science by Wickham and Grolemund, free learning tools like the DoSS Toolkit from STA130 profs Alexander and Caetano et al., and free online primers, cheatsheets, and courses

2. math, algorithms, and data analysis Part I: Structured Learning RULES

When first learning, organized content like our course content is best

Attending Lectures	isn't mandatory, but they're not recorded	N/A
Attending Tutorials	**No makeups for missed in-person tasks**	15%
R/communication HW	**No late or missed submissions accepted**	15%
Study and Review	not mandatory, but good exam practice	N/A
Final/Midterm Exams	**Makeup exams for excused absence only**	45%
Coures Project	**1-2 day late submissions for special cases**	20%
Participation	**Due by last day of classes**	5%

Boost your skills with free textbook resources like R for Data Science by Wickham and Grolemund, free learning tools like the DoSS Toolkit from STA130 profs Alexander and Caetano et al., and free online primers, cheatsheets, and courses

Special Rules

Highest homework mark replaces lowest
Highest tutorial mark replaces lowest

These are meant to smooth out "small bumps" in your semester

If it's not enough to help, you're experiencing an unusually tough semester and need to see your college registrar immediately

Class Check Round 2

https://pollev.com/sta (7 questions)

2. math, algorithms, and data analysis Part I$^1/_2$: Structured Learning $\rightarrow$ Unstructured Learning

Structured Learning is great whenever you can get it...

When first learning, organized content like our course content is best

but once you start to get the hang of it

you're going to need to troubleshoot and

graduate to the next level of learning...

2. math, algorithms, and data analysis Part II: Unstructured Learning

Modern Stats+DS learning communities propegate knowledge and solutions

- R is ubiquitous in traditional data analysis and statistical communities
  - and increasingly supports machine learning (ML), bioinformatics, etc.
- Python is dominant in nearly all application involving coding, such as DS and ML
  - and increasingly overlaps and supports users traditionally from the R community

2. math, algorithms, and data analysis Part III: Structured Learning $\rightarrow$ Unstructured Learning

So when you get strange error or don't know an answer to something

Ask yourself

1. Have I reviewed the course notes well enough?
2. Is this the kind of problem a google search could fix?
3. Could this have already been answered on the course piazza?
4. Am I SURE this hasn't already been answered on the course piazza?
5. Would discussing my problem in an office hour be more helpful than piazza?
6. Is asking on the course piazza probably the fastest way to get an answer?

Class Check Round 3

https://pollev.com/sta (7 questions)

Statistical Reasoning and Data Science (DS)

Modern Stats+DS is

1. software, programming, and computational tools implementing
2. mathematical and algorithmic data analysis methodologies for
3. facilitating data-driven and evidence-based decision making

3. facilitating data-driven evidence-based work: Jupyterhub

Jupyterhub is a cloud-based service allowing you to run R/Rstudio from any web browser

- You can run R/Rstudio locally as well, and you may need to install Rstudio so you can do this for this class in case of temporary Jupyterhub service outage

From Jupyerhub in a gif-y

Logging in troubleshooting

UNIVERSITY OF TORONTO JUPYTERHUB

3. facilitating data-driven evidence-based work: Rstudio

Rstudio is just a nice GUI IDE program that wraps up R along with a .Rmd code editor and file and data management capabilities

- GUI means "graphical user interface" while IDE means "integrated developers environment", and .Rmd RMarkdown files (based on Markdown markup language) combine text with R code to create .html and .pdf output files and presentation slides (like the ones I'm using here now).

Knitting

Knitting

Knitting

Knitting

Knitting

Knitting

Knitting

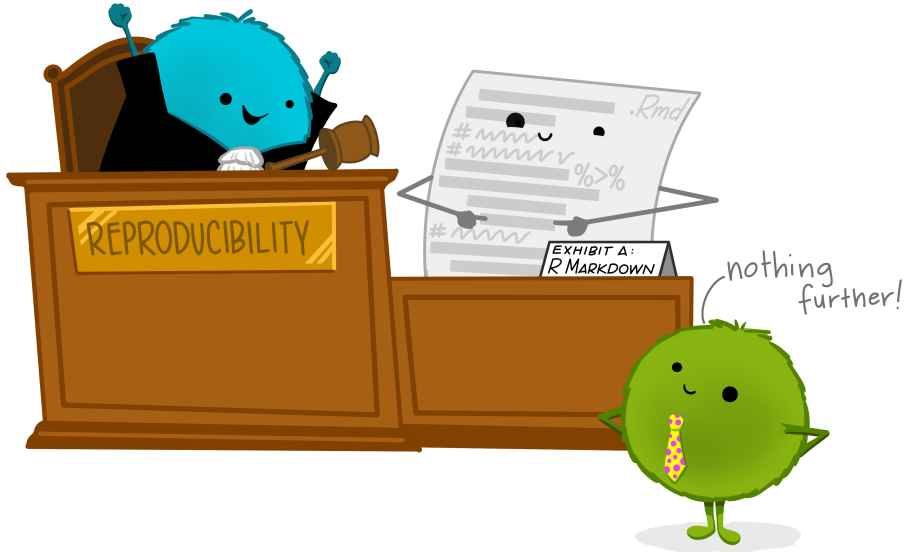
Knitting

Running R code in Rstudio: Knitting

1. ***Source*** [Top Left pane]: Rmarkdown files with code and text are knit
2. ***Console*** [Bottom Left pane]: R code is run to produce output
3. ***Environment*** [Top Right pane]: R variables and data sets are created
4. ***Files*** [Bottom Right pane]: code, outputs, and text combined into output

To run code-chunks use "ctrl-shift-enter" (PC) and "cmd-shift-return" (Mac)

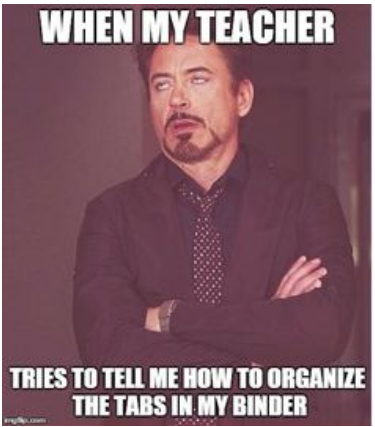
- This is just one of many keyboard (key-binding) shortcuts in Rstudio


REPRODUCIBILITY
.Rmd
#
#
%>%
#
EXHIBIT A:
R MARKDOWN
nothing
further!
@allison_horst

Keeping Rstudio Organized: File Management

From Jupyerhub in a gif-y

folders relocations uploading linking projects files exporting



3. facilitating data-driven evidence-based work: R

R is a data analysis programming console to support the work of statisticians, data scientists, and other data-centric professions

- R is a programming language, but its methods and algorithms are usually built-in or loaded from packages, so most R users don't build algorithms or data types like in C++ or java

R as Illustrated by allison_horst

3. facilitating data-driven evidence-based work: R Packages

Newly developed statistical methods are often made available as R package

This makes the publication manuscripts stronger and increases a methods use

Comprehensive R Archive Network (CRAN) currently has 18586 packages

For those new to R and arriving with interest in DS, the tidyverse is the key set of packages

The tidyverse includes R packages that help facilitate "data wrangling"

Installing Packages, Loading Libraries, and Common Errors

Installing Packages, Loading Libraries, and Common Errors

Installing Packages, Loading Libraries, and Common Errors

- You can't run R code unless its package is loaded as `library(packagename)`
- You can't load packages unless `install.packages("packagename")` is run first
  - Packages installed on Jupyterhub must be re-install for each new session...
- Some pre-installed packages, e.g., `library(tidyverse)`, will always load
  - 'There is no package called...' errors mean `install.packages()` is needed
- But you can't "knit" .Rmd files `install.packages("packagename")` code
  - So you have to frequently comment and uncomment `install.packages()`

Class Check Round 4

https://pollev.com/sta (5 questions)

Self Quiz

Can you log into JupyterHub and open and knit an RMarkdown document in RStudio?

Can you navigate the Code, R Console, and Files panes in the RStudio interface?

Did you see the Environment pane and the Help, Packages, and Tutorial tabs?

Rstudio Demo I

1. Click this jupyterhub repo launcher link
   - If this fails (1) log into RStudio on Jupyterhub, (2) delete the folder, (3) retry
2. Navigate to the intended directory in the (bottom right) files pane
3. Open the intended .Rmd file
4. Follow along / complete the demo
5. Understand Rmd+pdf submission requirements
6. Determine and commit to a file organization scheme

Class Check Round 5

https://pollev.com/sta (6 questions)


R MARKDOWN
we're getting the BAND BACK Together.
TEXT
OUTPUTS!
CODE
HORST '19

Rstudio Demo II

1. Click this jupyterhub repo launcher link

Class Check Bonus Round

Practice Quiz on Quercus (10 questions)

Scratch ideas below will be combined with quiz 1 on quercus.

1. head vs glimpse preference
2. types in columns
3. read/match "and then" pipe statement
4. other things?
5. Speak/write basic code in a human-friendly way, e.g., reading a pipe (%>%) as "and then".
6. Explain what a tibble is and what rules it follows.
7. Types vs Vectors vs Coercion vs Data frames vs read_csv() vs glimpse() and head()