SOC-GA 2332 Intro to Stats Lab 1

Di Zhou

NYU Sociology

2/5/2021

Lab & Assignment Plan

▶ **Goals of the lab**
  ▶ Technical skills for statistical analysis & result presentation: R, LATEX
  ▶ Review concepts taught in class and operationalize them in R

▶ **(Tentative) Plan of the lab**
  ▶ Three sessions (11:00 to 11:55; 12:00 to 12:50; 1:00 to 1:40)
  ▶ I will stop for questions every 10 to 15 minutes.

▶ **(Tentative) Plan of Assignments**

Assignment	Release Date	Due Date
1	2/8	2/26
2	3/1	3/19
3	3/22	4/9
4	4/5	4/30

Communication

- ▶ **Office hour**
  - ▶ Friday 2:30 - 3:30 pm
  - ▶ Or by appointment (di.zhou@nyu.edu)

- ▶ **Slack Workspace**
  - ▶ You can ask questions about assignments and final projects using our Slack chat group.
  - ▶ One key advantage of Slack is its built-in code-sharing feature. Follow this guide to share code snippets on Slack.
  - ▶ Click to join the **2021 Stats** Slack group

- ▶ **Lab Materials**
  - ▶ Lab materials will be uploaded to stats lab's github repository before lab. In future labs, you should **download all the materials in the respective lab folder before coming to lab**.

Questions?

Logistics	Outline	The Standard Workflow	Basics of R & Rstudio	Tidy Data	LaTeX & Overleaf
000	●	○	00000000000	000000	00000

Outline

1. The standard workflow with data
2. Basics of R & Rstudio
3. Tidy data & `Tidyverse`, including how to use `Tidyverse` for descriptive statistics and plots
4. Basics of LaTeX & Overleaf

Working with Data: The Standard Workflow
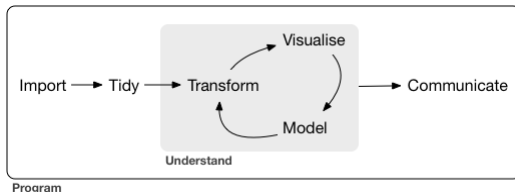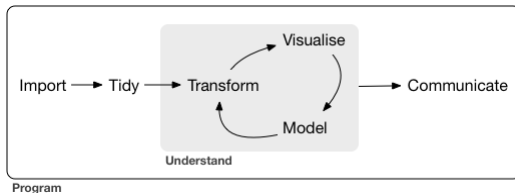
▶ A typical data science project looks like this :

Figure 1: The Standard Workflow (Wickham & Grolemund 2017)

▶ In practice, we spend a LOT of time just to clean the data so that it is ready for descriptive analysis and modeling exercise. For example, dealing with missing values, incoherent variable codings across survey years, outliers, inflations, etc.

Working with Data: The Standard Workflow

▶ A typical data science project looks like this :

Figure 1: The Standard Workflow (Wickham & Grolemund 2017)

▶ In practice, we spend a LOT of time just to clean the data so that it is ready for descriptive analysis and modeling exercise. For example, dealing with missing values, incoherent variable codings across survey years, outliers, inflations, etc.

R & Rstudio: Intro

- ▶ **R** is a free programming language commonly used for statistical programming and graphics. Download & install: https://cloud.r-project.org/

- ▶ **Rstudio** is an IDE (Integrated Development Environment) for R. It's an application that enables you to write, run, and save your R code and programming outputs. Download & install: https://rstudio.com/products/rstudio/download/

Logistics
○○○

Outline
○

The Standard Workflow
○

Basics of R & Rstudio
○●○○○○○○○○○

Tidy Data
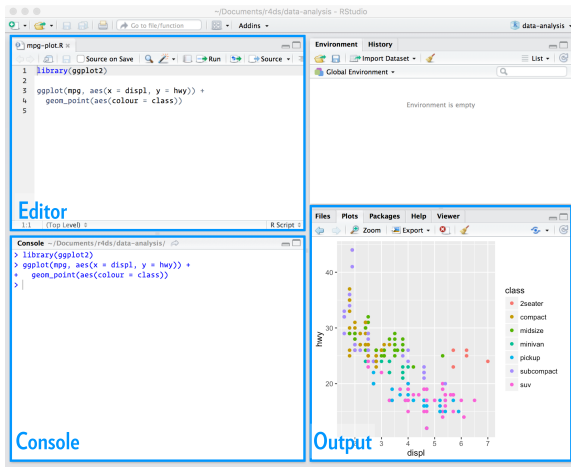○○○○○○

LATEX & Overleaf
○○○○○

R & Rstudio: Layout

Figure 2: Layout of Rstudio (Wickham & Grolemund 2017)

R & Rstudio: Best Practice Step by Step

▶ Every time when you start a new project or assignment, create an individual folder for that task.

▶ Open **Rstudio**, click **File**, then **New Project...**, then choose your project folder, name the project and save it.

▶ Now open that **.Rproj** file, and start coding.

▶ When ending the programming process, save your session's workspace in a **.RData** file, so that when you open **.Rproj** again, you start exactly from where you stopped (see Figure 4).

▶ I recommend that you create subfolders within the project folder to organize your files, for example, a folder for **data**, one for **graphs**, etc. (see Figure 3)

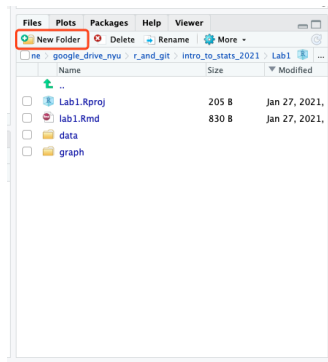R & Rstudio: Best Practice Step by Step

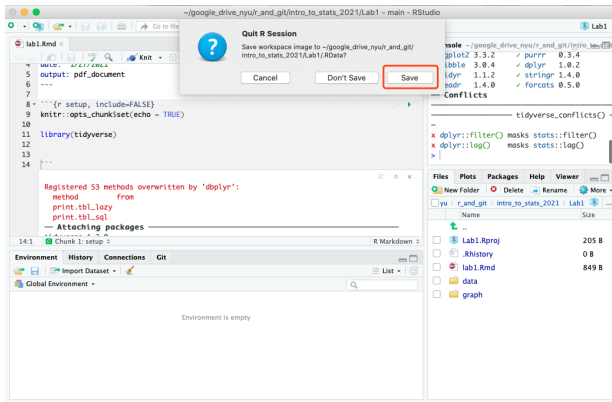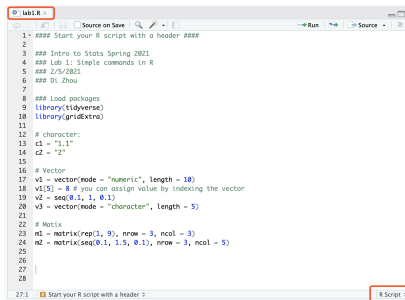Figure 3: Shortcut for create new folders in Rstudio interface

Logistics
○○○

Outline
○

The Standard Workflow
○

Basics of R & Rstudio
○○○○○●○○○○○○

Tidy Data
○○○○○○

LATEX & Overleaf
○○○○○

R & Rstudio: Best Practice Step by Step

Figure 4: Save your session data before you exit Rstudio

Logistics	Outline	The Standard Workflow	Basics of R & Rstudio	Tidy Data	LATEX & Overleaf
000	0	0	00000●00000	000000	00000

R & Rstudio: R Script vs. R Markdown

▶ **R Script** is a simple code script document. The output of R
script cannot be saved within the script.

Figure 5: R Script

R & Rstudio: R Script vs. R Markdown

▶ **R Markdown** is a simple formatting syntax for authoring
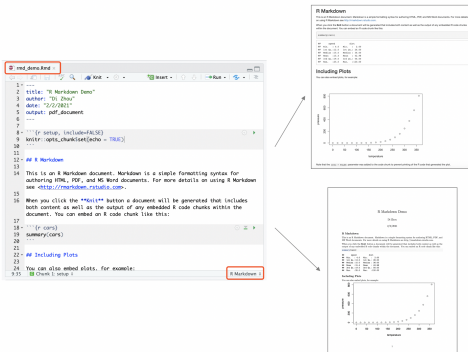 HTML, PDF, and even Microsoft Word documents.

Figure 6: R Markdown can be rendered to HTML and PDF

R & Rstudio: R Script vs. R Markdown

- ▶ Different from R Script, **R Markdown** allows users to **present both their code and the code's output (tables, plots, etc.) in a single document**, usually by "knitting" (rendering) an R Markdown to a HTML or PDF file.

- ▶ R Markdown allows you to divide your code into sections, which helps you **better organize** your code.

- ▶ R Markdown also allows you to type **mathematical equations** efficiently.

- ▶ If you are coding for assignments, **use R Markdown**. If you are coding for simpler tasks, you can use R script.

Questions?

R & Rstudio: Demo

Demo

▶ How to create a new project

▶ Layout of Rstudio

▶ Coding in console

▶ Coding in R Script

▶ Installing and using packages

▶ Coding in R Markdown

▶ Debugging in R

▶ Types of variables and data structures in R, functions in R

▶ Typing equations in R Markdown

▶ Knitting R Markdown to HTML or PDF

Questions?

Tidy Data

▶ Before doing analysis to your data, look carefully at your dataframe, and ask: is it a "tidy" data set?

▶ In a tidy data set:
  1. Each **unit of observation** is saved in its own **row**.
  2. Each **variable** is saved in its own **column**.
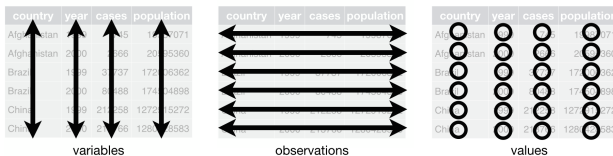 3. Each value must have its own cell.

Figure 7: Tidy Data (Wickham & Grolemund 2017)

Tidy Data: Are they tidy? Why?

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Figure 9: Data Frame B

country	year	type	count
Afghanistan	1999	cases	745
Afghanistan	2000	cases	2666
Brazil	1999	cases	37737
Brazil	2000	cases	80488
China	1999	cases	212258
China	2000	cases	213766
Afghanistan	1999	population	19987071
Afghanistan	2000	population	20595360
Brazil	1999	population	172006362
Brazil	2000	population	174504898
China	1999	population	1272915272
China	2000	population	1280428583

country	type	1999	2000
Afghanistan	cases	745	2666
Afghanistan	population	19987071	20595360
Brazil	cases	37737	80488
Brazil	population	172006362	174504898
China	cases	212258	213766
China	population	1272915272	1280428583

Figure 8: Data Frame A

Figure 10: Data Frame C

Tidy Data

▶ Tidy data is a foundation for data transformation in R using the `tidyverse` package collection.

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

▶ To learn `tidyverse` and coding in R more systematically, I strongly recommend the free online book by Wickham & Grolemund (2017): R for Data Science. (A lot of the examples used in this lab is drawn from the book.)

Questions?

Tidy Data & `Tidyverse`: **Demo**

Demo

▶ Import .csv file

▶ Browse data in R

▶ Basic `Tidyverse` command

▶ "Pipe" in `Tidyverse` coding

▶ Make untidy data tidy

▶ Summarise & group data

▶ Plot using `ggplot2`

▶ Exercise

Questions?

Logistics	Outline	The Standard Workflow	Basics of R & Rstudio	Tidy Data	LaTeX & Overleaf
000	0	0	00000000000	000000	●0000

LaTeX

- ▶ LaTeX is a typesetting language for creating documents, including manuscripts and presentations.
- ▶ It provides typesetting syntax that allows user to quickly tackle complicated typesetting tasks, such as inputting equations, cross-referencing, creating tables and bibliographies.
- ▶ You don't have to learn coding LaTeX from scratch. There are a wide variety of templates ready to use on **Overleaf**.

Overleaf

- ▶ **Overleaf** is a collaborative cloud-based LaTeX editor.
- ▶ **Overleaf** offers comprehensive guidance on how to create and edit LaTeX documents.
- ▶ My suggestion for you to use LaTeX for the final project is to start from **the default template in Overleaf**, and search and learn the specifics when you come across them –such as how to insert tables, how to fit tables to page, insert equations, how to change the font size, how to organize citations, etc.
- ▶ However, if you want to learn the basics and try creating a LaTeX document from scratch, you can learn it within 30 minutes [according to the author(s) of the guide].

Questions?

LaTeX & Overleaf: Demo

Demo

▶ How to create documents in Overleaf

▶ Resources for using Overleaf & LaTeX

▶ Mathematical equations using LaTeX

▶ Exercise

Questions?