Managing and Manipulating Data Using R

Introduction

Karina Salazar

1. Introductions
2. What is R?
3. Why learn R?
4. What is this course about?
5. Course logistics
6. 10 Min Break
7. Create "R project" and directory structure
8. Directories and filepaths

Hello! While you wait for class to begin...

- Navigate to D2L
- On our class D2l site, navigate to *Syllabus & Class Resources*, and download the following:
  - Folder Structure Zip File
- Under *Today's Date* > *Lecture Materials* download the following:
  - Introduction PDF
  - Introduction Problem Set Rmd
- For now... just keep these files in your downloads folder

Introductions

Student introductions

1. Name
2. Pronouns
3. Academic program (and how far along)
4. GA, RA, TA, and/or job?
5. Do you have any experience with R? If not, do you have experience with any other "statistical" software (e.g., SPSS, STATA) or coding languages (e.g., Python, SQL, Java)?
6. Why are you interested in this course?

Karina Salazar, instructor

My start in data management/statistical analysis

- SPSS
  - Evaluated retention programs within institutional research and assessment offices
  - Student-level data on math remediation courses
  - College Academy for Parents, Think Tank, Assessment Institute
- Stata
  - Used loops and user-defined functions to work with national datasets (IPEDS, Survey of Earned Doctorates)

Got sick of the limitations of survey data and/or available data

- No survey asked questions on what I was interested in
  - Universities pledge commitment to access, but enrollments don't tell the whole story
  - Who do they actually recruit?
- We realized "data science" could create data from publicly available data sources
  - Twitter
  - Travel schedules on admissions websites

Recruiting research program and "data science"

- Python
  - web-scraping
  - connecting to Application Program Interfaces (API) (e.g., census data, Twitter, LinkedIn)
  - Natural Language Processing
- R
  - R can do all "data science" tasks Python can
  - R can do all statistical analyses that Stata can (and more!)
  - R has amazing mapping capabilities

What is R?

What is R?

According to the Inter-university consortium for political and social research (ICPSR):

> *R is "an alternative to traditional statistical packages such as SPSS, SAS, and Stata such that it is an extensible, open-source language and computing environment for Windows, Macintosh, UNIX, and Linux platforms. Such software allows for the user to freely distribute, study, change, and improve the software under the Free Software Foundation's GNU General Public License."*

- For more info visit R-project.org

Base R vs. R packages

There are "default" packages that come with R. Some of these include:

- `as.character`
- `print`
- `setwd`

And there are R packages developed and shared by others. Some R packages include:

- `tidyverse`
- `stargazer`
- `foreign`

more about these in later weeks...

Installing and Loading R packages

You only need to install a package once. To install an R package use `install.package()` function.

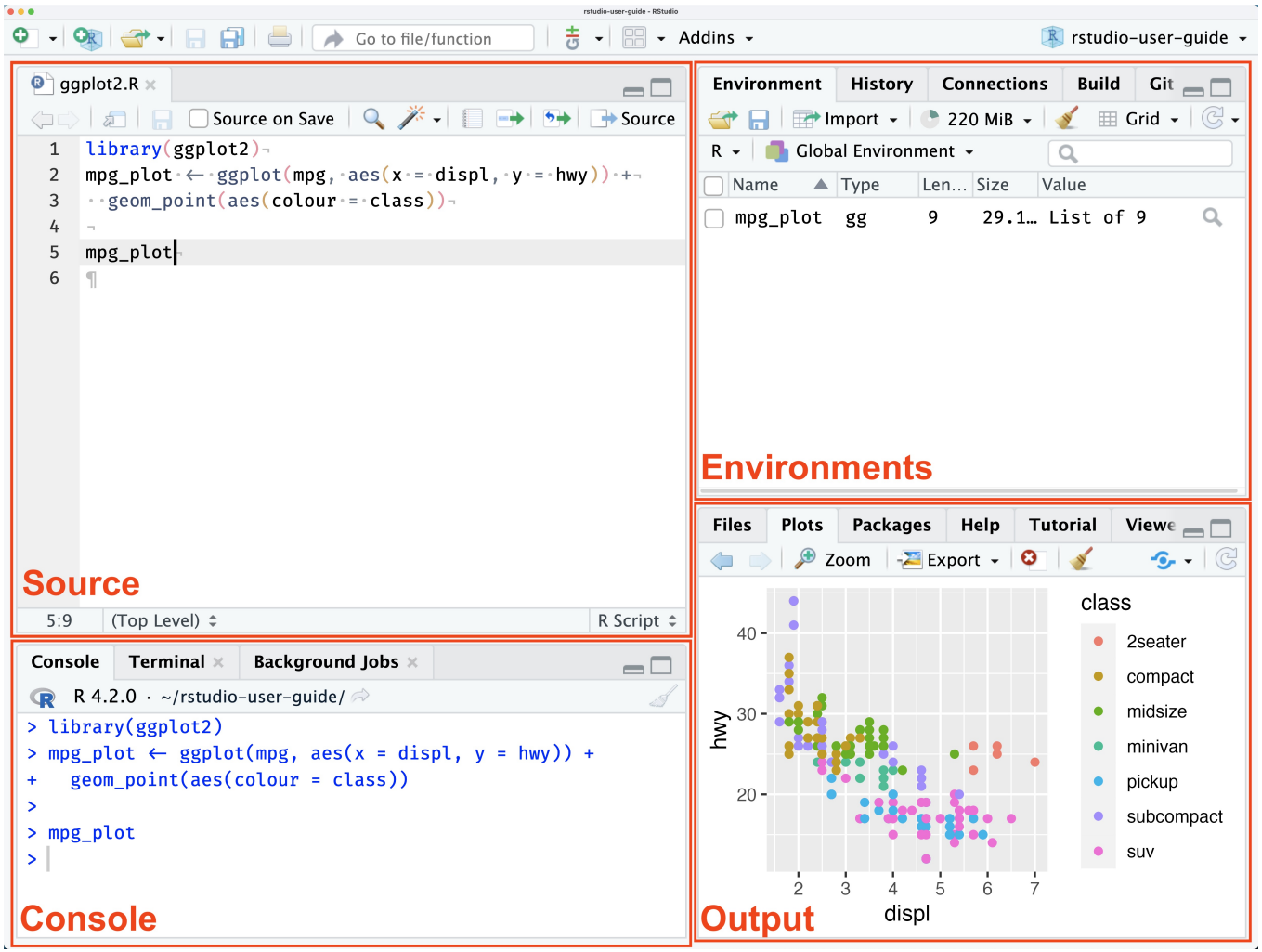
```
#install.packages("tidyverse")
```

However, you need to load a package everytime you plan to use it. To load a package use the `library()` function.

```
library(tidyverse)
#> -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0
#> v dplyr     1.1.4     v readr     2.1.5
#> v forcats   1.0.0     v stringr   1.5.1
#> v ggplot2   3.5.1     v tibble    3.2.1
#> v lubridate 1.9.4     v tidyr     1.3.1
#> v purrr     1.0.2
#> -- Conflicts ------------------------------------------ tidyverse_conflicts()
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag()    masks stats::lag()
#> i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all co
```

RStudio

"RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management."



R Markdown Documents

R Markdown produces dynamic output formats in html, pdf, MS Word, dashboards, Beamer (i.e., power point) presentations, etc.

- We will be using R Markdown for lectures and homework assignments
- These files names end with `.Rmd`
- Show ABOR Literature Review Example

R Scripts

R Scripts are simply a text file containing all the same commands that you would enter on the command line of R. The "text" you can include in these files are in the form of comments.

- We will be using R Markdown for some homework assignments.
- These files names end with `.R`
- Create our first R Script!

Why learn R?

Why learn R?

How we have used R+RStudio+RMarkdown in our research team

- Stuff traditional statistical software (e.g., SPSS, Stata) can do
  - Data manipulation, creating analysis datasets
  - Descriptive statistics and statistical models
  - Graphs
- Stuff traditional statistical software cannot do
  - Static policy reports
  - Static presentations
    - All lectures for this class written in RMarkdown
  - Interactive presentations
  - Interactive map
    - CFPB
    - Karina's recruitment redlining map
  - Interactive dashboards
  - Interactive graphs

Some of the other stuff R can create/do:

- Websites; journals; books; web-scraping; network analysis; machine learning/artificial intelligence

What is this course about?

What is data management?

- All the stuff you have to do to create analysis datasets that are ready to analyze:
  - Collect data
  - Read/import data into statistical programming language
  - Clean data
  - Integrate data from multiple sources (e.g, join/merge, append)
  - Change organizational structure of data so it is suitable for analysis
  - Create "analysis variables" from "input variables"
  - Make sure that you have created analysis variables correctly

Why I don't call this class "R for data science"

Data management and manipulation is the building blocks of data science!

- Data Science implies doing "fancy" things like mapping, network analysis, web-scraping, etc.
- But if you don't know how to clean data, these "fancy" analyses and visualizations will be impossible to execute
- "80% of data science is data cleaning"
- The skills you learn in this data management class are foundational to data science tasks! (and a prerequisite to taking my data science seminar next semester)

Who is this class for?

This class is for anyone who wants to work with data, that is people who want to be:

- Researchers working with survey data and doing traditional statistical analyses
- Researchers who want to do "data science" oriented research involving mapping, NLP, connecting to APIs
- Analysts working at think tanks or non-profits that work with large federal or state datasets

Course logistics

Course logistics

- follow the syllabus

Course Structure

- The first week of the course:
  - Focus on getting everyone comfortable with R
  - We will likely take up the full class time
    - Troubleshoot errors in-class, together
    - Covering material in "traditional" lecture style
- Week 2, the class incorporates an asynchronous component
  - **Asynchronous BEFORE WEEKLY MEETINGS**
    - Short readings from textbooks
    - Watch ~30 min lecture overview recording by instructor
    - Read/Execute Code in lecture "slides"
  - **Synchronous 1-1.5 HOUR WEEKLY WORKSHOP CLASS**
    - Collectively or in pairs, we will work through class materials or the weekly problem set together
    - Instructor will move from group to group helping and answering questions
    - Students will need to complete and submit the problem set every Wednesday after the synchronous meeting...

10 Min Break

Create “R project” and directory structure

What is an R project? Why are you doing this?

What is an "R project"?

- Helps you keep all files for a project in one place
- When you open an R project, the file-path of your current working directory is automatically set to the file-path of your R-project

Why am I asking you to create R project and download a specific directory structure?

- I want you to be able to run the .Rmd and .R files for each lecture on your own computer
- Sometimes these .Rmd and .R files point to certain sub-folders
- If you create R project and create directory structure I recommend, you will be able to run .Rmd and .R files from your own computer without making any changes to file-paths!
- This process allows us all to work off our individual computers but to "start" in the same working directory (R Project) and be able to navigate to other "shared" folders (same folder structure)

Follow these steps to create "R project" and directory structure

1. In your downloads folder, you'll have a zip file:
   - Unzip the folder: this is a shell of the file directory you should use for this class
   - Move it to your preferred location (e.g, documents, desktop, dropbox, etc)
2. You should also have one .Rmd and one PDF file in your downloads folder

- Move the **introduction.pdf** file into *"HED696C_Rclass/modules/introduction"*
- Move the **introduction_ps.Rmd** file into *"HED696C_Rclass/problemsets/introduction"*

3. In RStudio, click on "File" » "New Project" » "Existing Directory" » Navigate to the HED696C_Rclass folder » Create Directory

After you follow these steps

- You can add any additional sub-folders you want to the "rclass" folder
  - e.g., "syllabus", "resources"
- You can add any additional files you want to the sub-directory folders you unzipped
  - e.g., in "HED696C_Rclass/modules/module1" you might add an additional document of notes you took

Directories and filepaths

Working directory

(Current) Working directory

- The folder/directory in which you are currently working
- This is where R "automatically" looks for files
- Files located in your current working directory can be accessed without specifying a filepath because R automatically looks in this folder

Function `getwd()` shows current working directory

```
getwd()
#> [1] "/Users/karinasalazar/Library/CloudStorage/Dropbox/HED696C_RClass/modules
```

Command `list.files()` lists all files located in working directory

```
getwd()
#> [1] "/Users/karinasalazar/Library/CloudStorage/Dropbox/HED696C_RClass/modules
list.files()
#>  [1] "introduction.log"          "introduction.pdf"
#>  [3] "introduction.Rmd"          "introduction.tex"
#>  [5] "old"                       "pane_layout_23.jpeg"
#>  [7] "pane_layout.png"           "problemset_intro.pdf"
#>  [9] "problemset_intro.Rmd"      "problemset1_solutions.pdf"
#> [11] "problemset1_solutions.Rmd" "problemset1.pdf"
#> [13] "problemset1.Rmd"           "rticles.png"
#> [15] "sf.png"                    "shiny.png"
```

Working directory, "Code chunks" vs. "console" and "R scripts"

When you run **code chunks** in RMarkdown files (.Rmd), the working directory is set to the filepath where the .Rmd file is stored

```
getwd()
#> [1] "/Users/karinasalazar/Library/CloudStorage/Dropbox/HED696C_RClass/modules
list.files()
#>  [1] "introduction.log"           "introduction.pdf"
#>  [3] "introduction.Rmd"           "introduction.tex"
#>  [5] "old"                        "pane_layout_23.jpeg"
#>  [7] "pane_layout.png"            "problemset_intro.pdf"
#>  [9] "problemset_intro.Rmd"       "problemset1_solutions.pdf"
#> [11] "problemset1_solutions.Rmd" "problemset1.pdf"
#> [13] "problemset1.Rmd"           "rticles.png"
#> [15] "sf.png"                    "shiny.png"
```

When you run code from the **R Console** or an **R Script**, the working directory is. . . .

Command `getwd()` shows current working directory

```
getwd()
#> [1] "/Users/karinasalazar/Library/CloudStorage/Dropbox/HED696C_RClass/modules
```

Absolute vs. relative filepath

Absolute file path: The absolute file path is the complete list of directories needed to locate a file or folder.

```
setwd("Users/Karina/rclass/modules/module2")
```

Relative file path: The relative file path is the path relative to your current location/directory. Assuming your current working directory is in the "lecture2" folder and you want to change your directory to the data folder, your relative file path would look something like this:

```
setwd("../../data")
```

`File path shortcuts`

Key	Description
~	tilde is a shortcut for user's home directory (mine is my name)
../	moves up a level
../../	moves up two levels

Install TinyTex

- Why am I asking you to do this?
  - You will need to install LaTeX (lah-tech or lay-tech) on your computer to create pdf documents in R Markdown files (.Rmd)
  - You do not need to know how to use LaTeX. LaTeX is used in the background to compile pdf documents for you.
  - Here is a helpful article on creating PDf reports using R, R Markdown, LaTeX, and knitr.
- Instructions for installing tinytex
  - Here is a helpful link to install tinytex
    1. Open up RStudio
    2. In the "console" paste the following and hit return(enter): **install.packages('tinytex')**
    3. Once the package is installed, paste the following code in the "console" and hit return(enter): **tinytex::install_tinytex()**

Intro Problem Set and Knit

- Open the *introduction_ps.Rmd*
- Let's knit to pdf!