Lecture 03.
R Basics (3) Data Wrangling (1)

POLI3148. Data Science in Politics and Public Administration

Dr. Haohan Chen HKU-PPA

Review: R Basics (1) (2)

- R & Rstudio intro
- R project setup
- Get help from online forums and Large Language Models
- Data type
- Data structure
  - Vector
  - Matrix, Array, List (briefly mentioned in lectures, only care about List for now)
  - Data Frame basics

Today

- R Basics
  - Rmarkdown for reproducible data science report
  - Git and Github
- Data Wrangling with tidyverse

Data Frame additional notes

(Douglas 3.3, 3.4, 3.5, 3.6)

- **Traditional** way to handle data frames
- Read for your general information
- We will **<u>not use them much</u>** going forward because we have a better tool
- But don't be surprised if you find online resources that does data processing in R in these ways

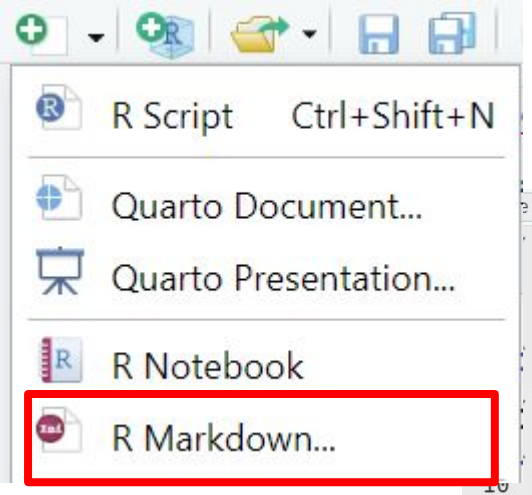
R Basics (3)

- **Rmarkdown** for reproducible report
- **Git and Github** for version control, sharing, and collaboration

Rmarkdown: Motivation

- Seamless integration:
  - Coding
  - Visualization of results
  - Writing: Analysis, interpretation, conclusion
- Reproducible outputs

Rmarkdown: Make our first Rmarkdown document



Visual

Outline

```
---
title: "First R Reproducible Report"
output: html_document
date: "2023-09-20"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```
```

Try Rmarkdown

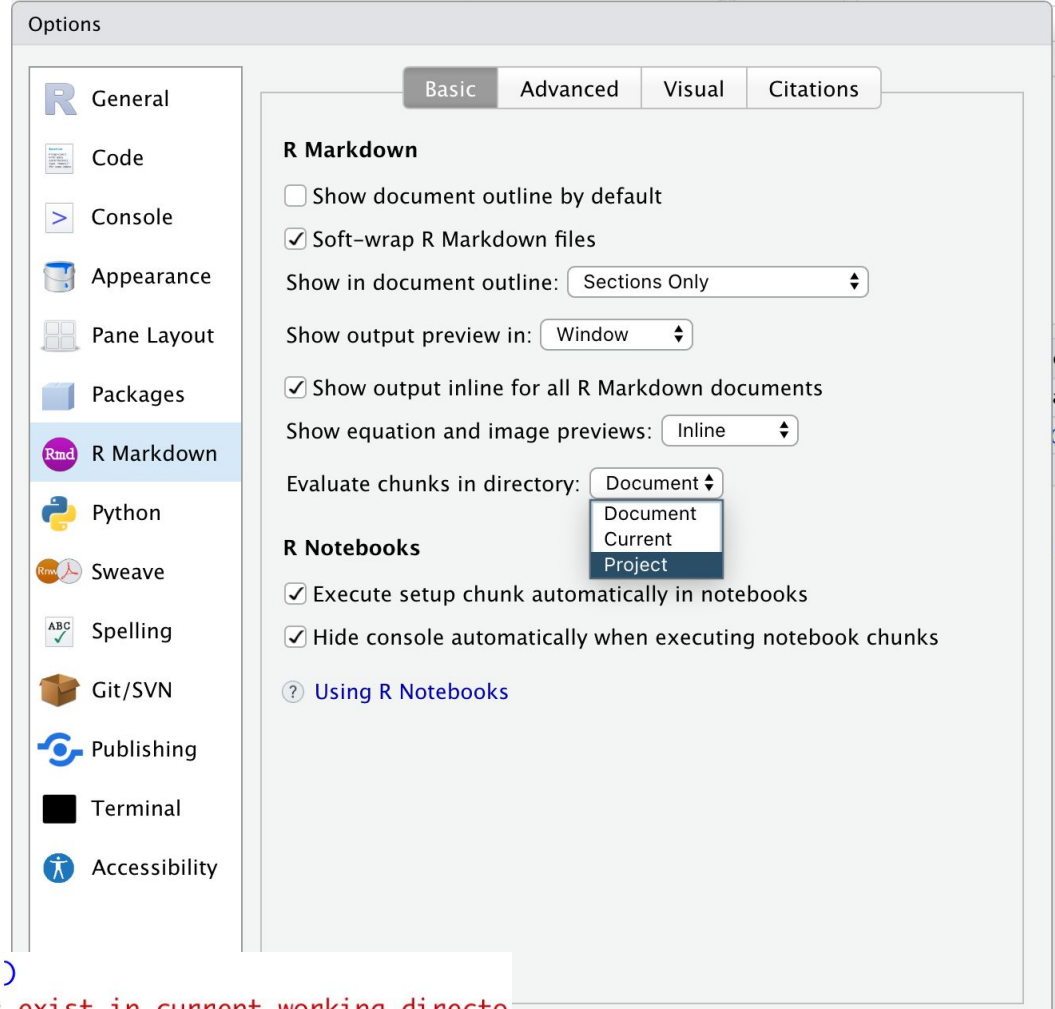
- YAML header
- Text
- Code chunk
- Inline R code
- Figures as output
- Tables as output
- Output formats
  - PDF
  - HTML
  - Word

Configuration tip: Use the “Visual” editor

But go back the check the “source” from time to time (especially when you need to delete things)

Configuration tip

Evaluate chunks in the “Project” directory




```
> d <- read_csv("data/raw_data/vdem/vdem_1999_2022.csv")
Error: 'data/raw_data/vdem/vdem_1999_2022.csv' does not exist in current working directo
ry ('/Users/haohanchen/Library/CloudStorage/GoogleDrive-haohanch@gmail.com/My Drive/TEAC
```

Check it out after class

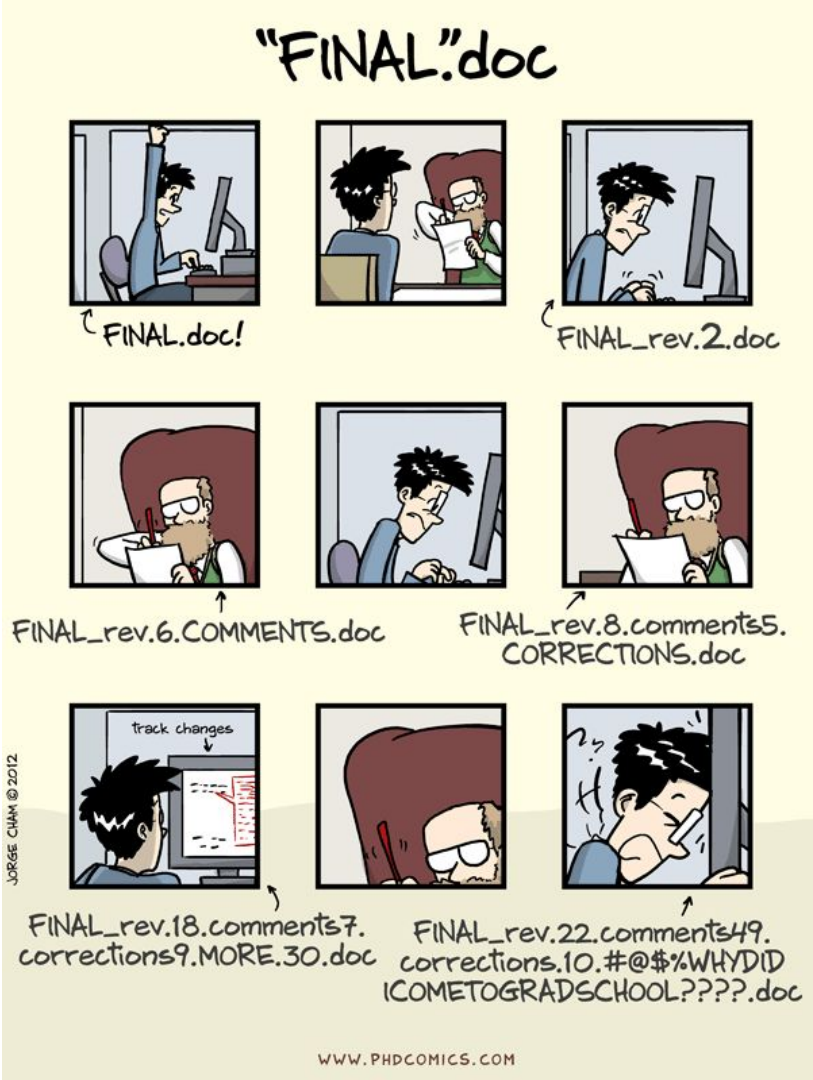
Rmarkdown Cheatsheet

Link: https://rstudio.github.io/cheatsheets/rmarkdown.pdf

There is nothing magical. Learn to use Rmarkdown like learning Microsoft Word.

Git and Github: Motivation

- Version control
- Collaboration
- Distribution



Use Github

- Register a GitHub account
- Download GitHub desktop
- Clone the course’s GitHub repo to your computer

Clone the GitHub repo of my in-class demo code to your local storage

In-class exercise 1:
Starting your DaSPPA portfolio

- Make a new Github repo
  - Include a README.md
  - For .gitignore, choose “R”
- In Github Desktop, clone the Repo to your local storage
- In the cloned local repo, create a folder named “Lecture_3_RBasics_Data_Wrangling”
- In the cloned local repo, setup the “data” “report” “script” folder
- Copy the Rmarkdown script we created together to your “report” folder
- In Rstudio, compile your Rmarkdown script
- In Github Desktop, *commit* and *push* your change online
- In Moodle, submit the **link** to your GitHub repo

Data Wrangling

- tidyverse intro
- Overview and resources
- Basic data wrangling
  - Import data
  - Select variables/ columns
  - Rename variables/ columns
  - Filter rows

Tidyverse https://www.tidyverse.org/

Packages Blog Learn Help Contribute

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")
```

Take a look at the cheat sheets

Resources: https://rstudio.github.io/cheatsheets

Learning Objectives:
You will be able to import and wrangle with data using functions in a selected set of cheat sheets

readr, dplyr, tidyr, lubridate, stringr

Use the printed copies of cheat sheets frequently :)

Our case for in-class demo: V-Dem data

V-Dem
VARIETIES OF DEMOCRACY