1.Welcome to PSTAT 188!

Transfer exploration seminar: Statistics and Data Science

Dr. Uma Ravat

PSTAT 188

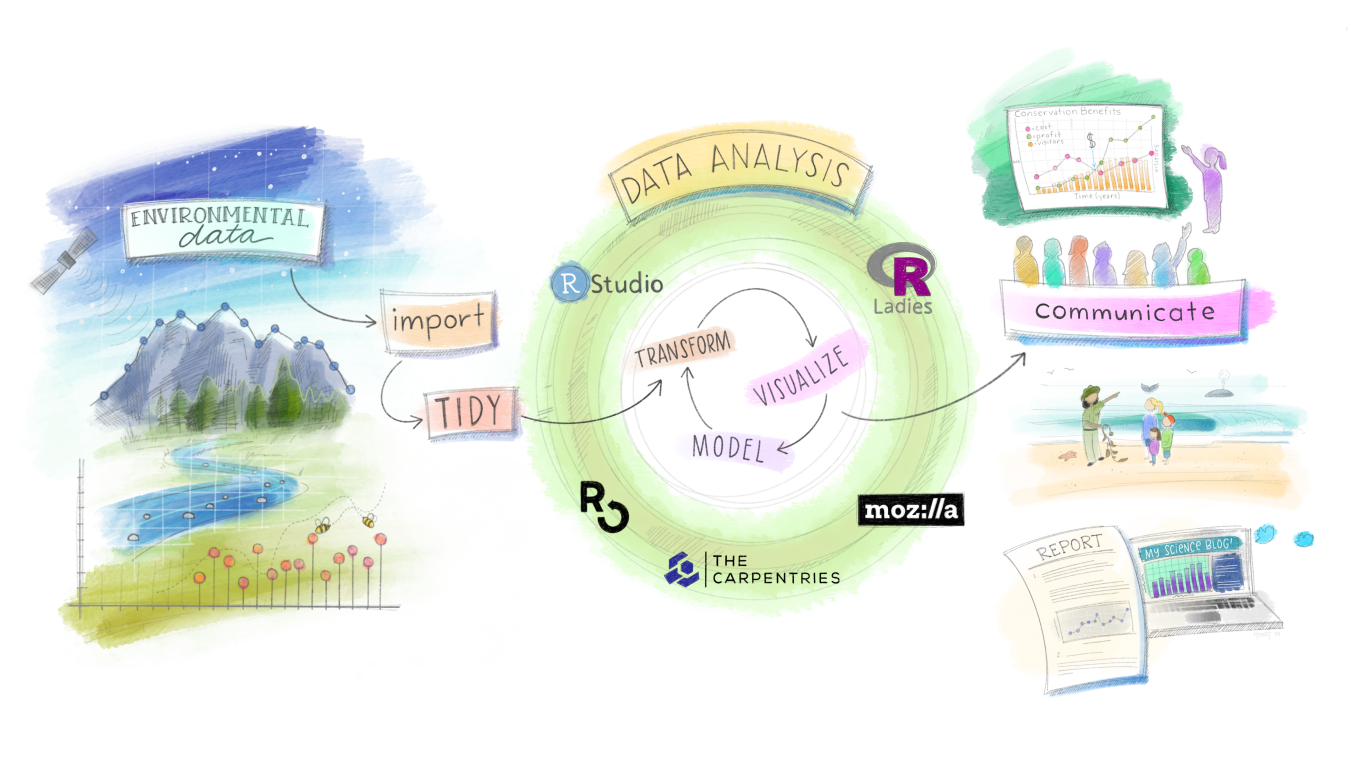
0.1 Welcome!

Plan for today

1. What is Data Science?
2. Your Turn at Data Science
3. **(R Toolkit)** Rstudio, Rmarkdown

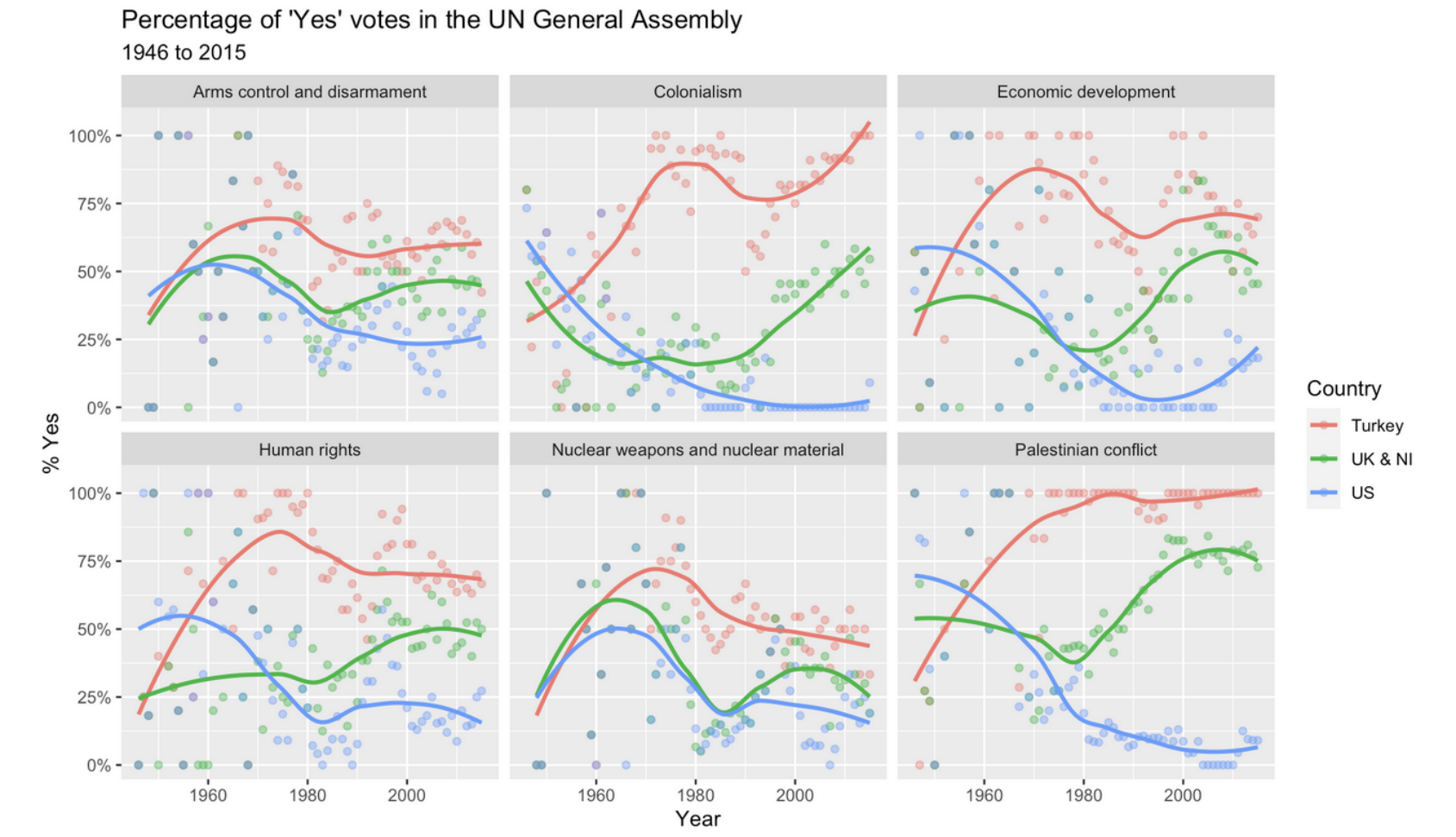
What is Data Science?

- Data science is an exciting discipline that allows you to look at raw data and transform it into understanding, insight, and knowledge that you can act on.



Your data science tasks result in...

1. A Visualiztion



2. Report/Website

UN Votes

Your name here

2022-08-24

Introduction

How do various countries vote in the United Nations General Assembly, how have their voting patterns evolved throughout time, and how similarly or differently do they view certain issues? Answering these questions (at a high level) is the focus of this analysis.

Packages

We will use the **tidyverse**, **lubridate**, and **scales** packages for the data wrangling and visualization, and the **DT** package for interactive display of tabular output.

```
library(tidyverse)
library(lubridate)
library(scales)
library(DT)
```

Data

The data we're using originally come from the **unvotes** package, but it's been modified a bit (by joining the various data frames provided in the package) to help you get started with the analysis.

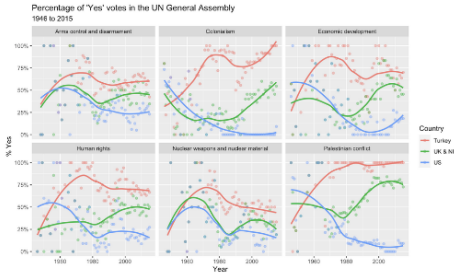
```
unvotes <- read_rds("data/unvotes.rds")
```

UN voting patterns

Let's create a data visualization that displays how the voting record of the UK & NI changed over time on a variety of issues, and compares it to two other countries: US and Turkey.

We can easily change which countries are being plotted by changing which countries the code above `filter`s for. Note that the country name should be spelled and capitalized exactly the same way as it appears in the data. See the Appendix for a list of the countries in the data.

```
unvotes %>%
  filter(country %in% c("UK & NI", "US", "Turkey")) %>%
  mutate(year = year(date)) %>%
  group_by(country, year, issue) %>%
  summarize(percent_yes = mean(vote == "yes")) %>%
  ggplot(mapping = aes(x = year, y = percent_yes, color = country)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~issue) +
  scale_y_continuous(labels = percent) +
  labs(
    title = "Percentage of 'Yes' votes in the UN General Assembly",
    subtitle = "1946 to 2015",
    y = "% Yes",
    x = "Year",
    color = "Country"
  )
```



References

0. Mine Çetinkaya-Rundel, Introduction to Data science course
1. David Robinson (2017). unvotes: United Nations General Assembly Voting Data. R package version 0.2.0.
2. Erik Voeten "Data and Analyses of Voting in the UN General Assembly" Routledge Handbook of International Organization, edited by Bob Reinalda (published May 27, 2013).
3. Much of the analysis has been modeled on the examples presented in the unvotes package vignette.

Appendix

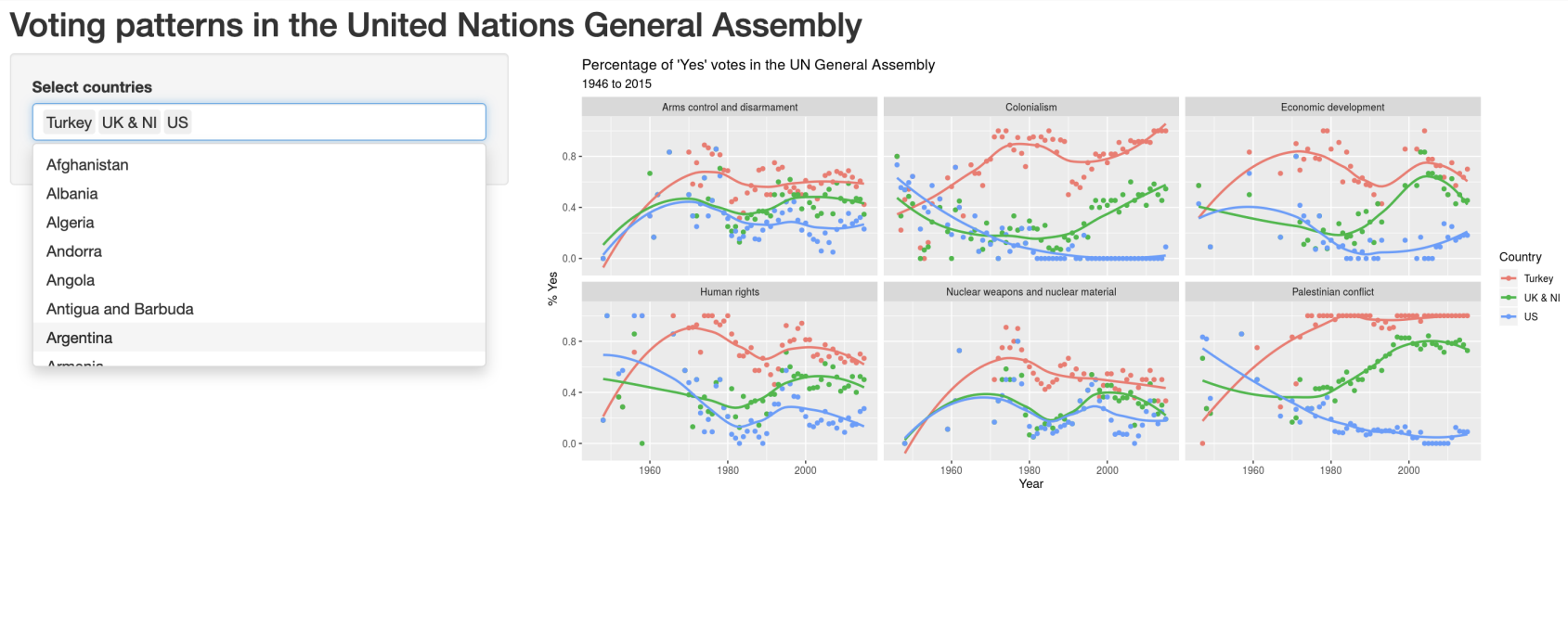
Below is a list of countries in the dataset:

Show 10 **entries** **Search:**

	country
1	Afghanistan
2	Albania
3	Algeria
4	Andorra
5	Angola
6	Antigua and Barbuda
7	Argentina
8	Armenia
9	Australia
10	Austria

Showing 1 to 10 of 199 entries Previous 1 2 3 4 5 … 20 Next

3. a data product: RShiny App



Data Science tasks require

1. **Computer Programming**
    - A computer program consists of instructions that tell the computer what to do (with some data).
    - PSTAT 10
2. **Statistical Analysis**
    - the **science** of collecting, organizing, exploring, interpreting, and presenting **data** to uncover patterns, trends, making predictions based on the data.
    - Statistical Analysis is grounded in Probability Theory
3. **Probability**
    - is the study of uncertainty or randomness and its consequences in the world around us.
    - PSTAT 120A.

- In one part of this course, you will be using the `Python` programming language to do Data Science tasks.
- The other part of this course, will be an introduction to data science using `R` programming language and a reviwe of probability fundamentals

`R` Course Objectives

1. Learn basics of reproducible research using `Rmarkdown`.
2. Learn basics of `R` programming language for Data science.
3. Learn introductory Statistics and Probability for Data science using `R`

Student learning outcomes

By the end of the `R` portion, students will

- Be comfortable using `Rmarkdown` to write reproducible reports.
- Have experience with introductory programming and data analysis tasks in `R`,
- Have a basic understanding of the fundamental concepts in probability theory.

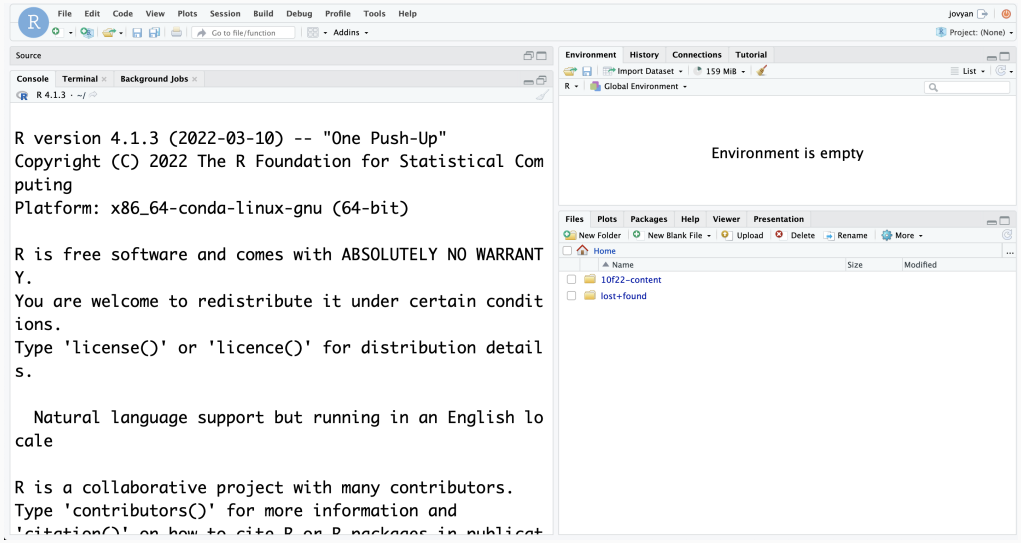
To succeed . . .

- Arrive on time
- Stay engaged
  - Avoid distractions (phone, email, social media, messaging apps turned off)
  - Actively participate
  - Ask questions
- Keep your files organized.
- Take half an hour after each class to review the materials and revise your notes.
- Take fifteen minutes before next class to skim material for upcoming class.

0.2 YOUR TURN 1 (10 minutes)

0.2 YOUR TURN 1 (10 minutes)

1. Login to our course Rstudio server using your **ucsb credentials** (link on canvas)



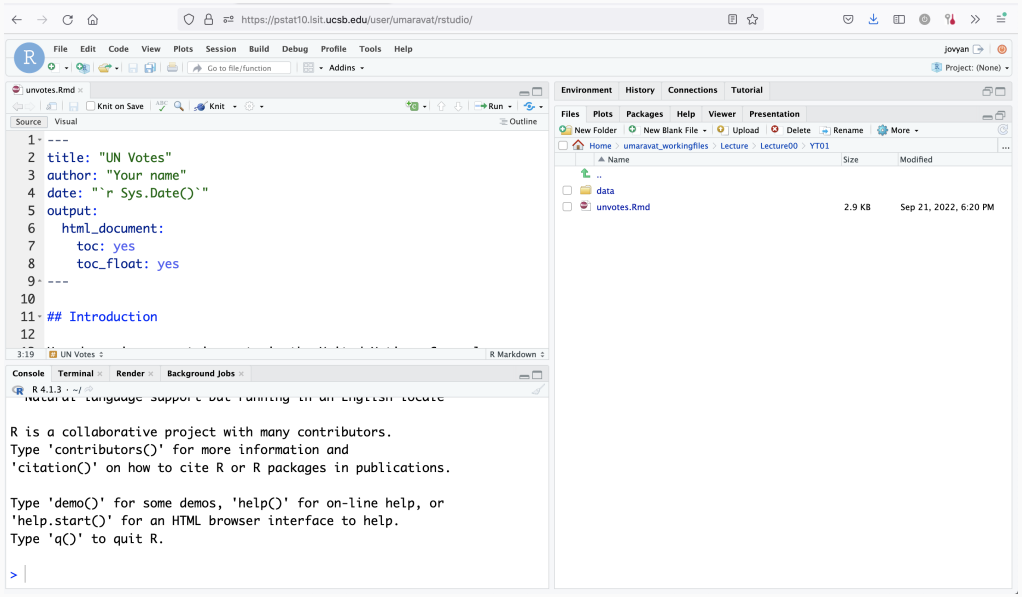
Setup My Rstudio Correctly

.Rmd

2. In the `Files pane` (bottom right corner), navigate to

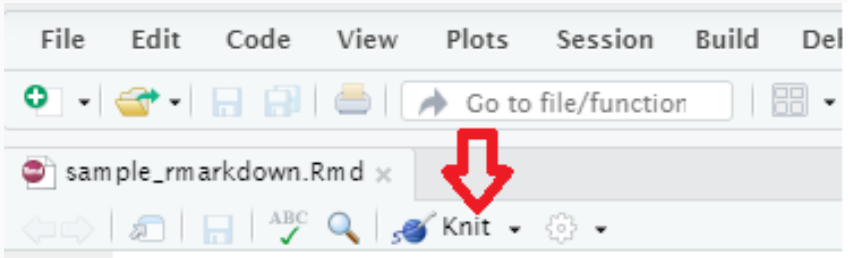
yournetid_workingfiles -> Lecture00 -> 3YT01

and spot the file called `unvotes.Rmd`.



Knitting

5. Open it and click “Knit”.



6. Go back to the file `unvotes.Rmd` and change your name on top (in the `yaml` – we'll talk about what this means later) and knit again.

7. Change the country names to those you're interested in. Spelling and capitalization must match how the countries appear in the data, so take a peek at the Appendix to confirm spelling.

8. Knit again. Voila, your first data visualization!

You should now be familiar with the computing environment we will use in this course.

0.3 R Markdown documents (and newer quarto documents)

0.3 R Markdown documents (and newer quarto documents)

- with `.Rmd` extension
- **rmarkdown** is an `R` package
  - write code and prose in **reproducible** computational documents

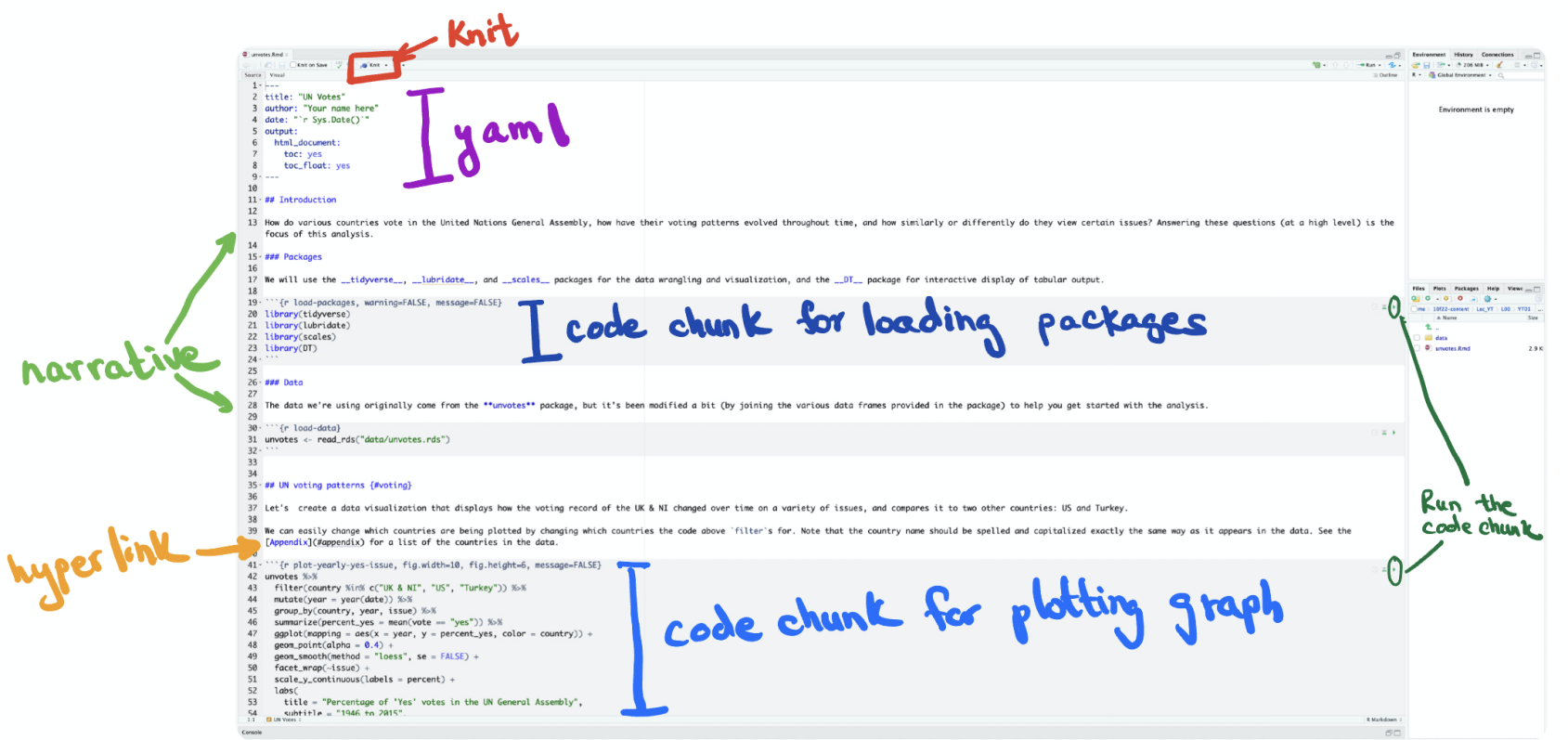
rmarkdown.rstudio.com

Take a look at the **Rmarkdown gallery**

R Markdown syntax

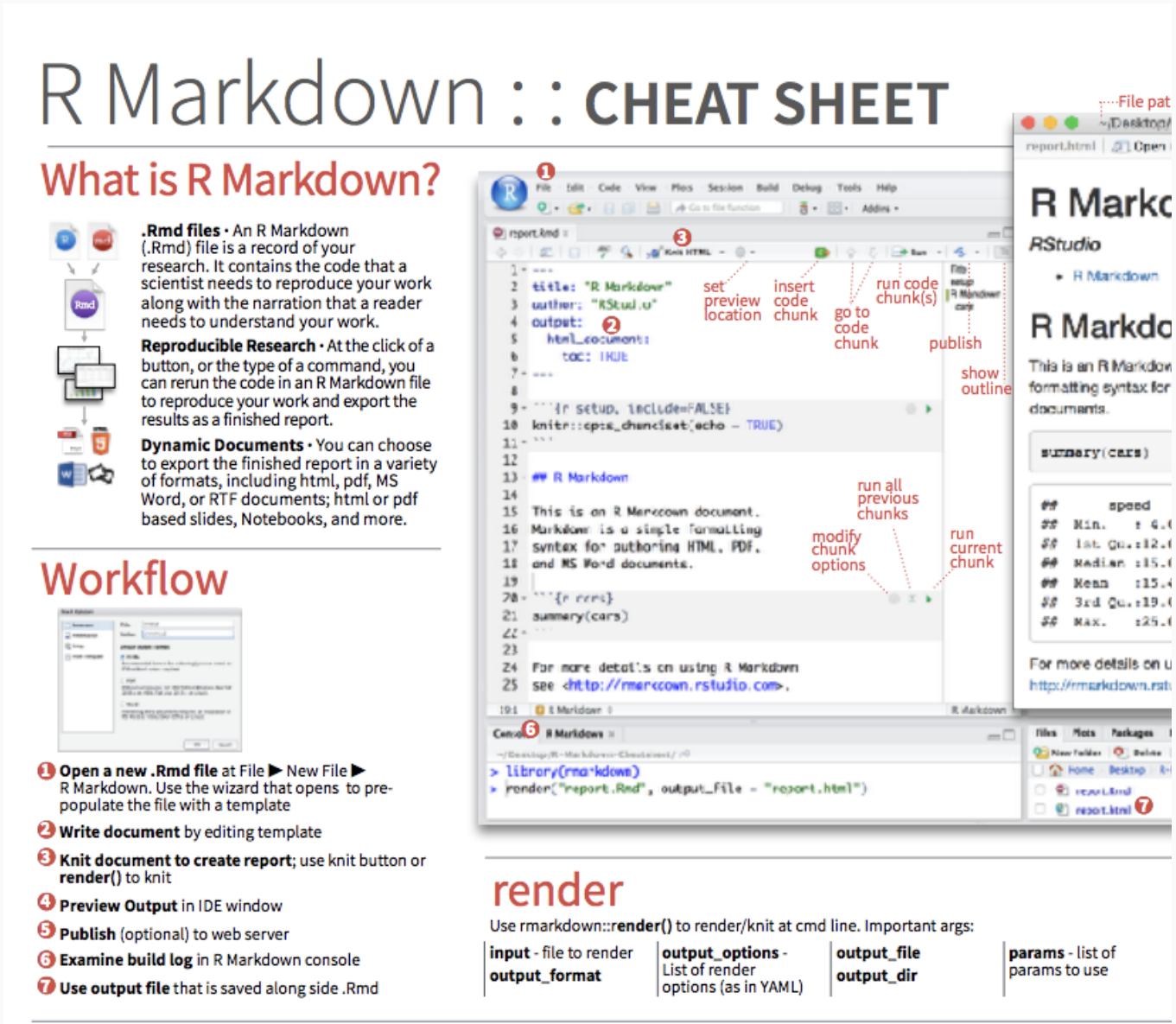
- Code goes in code chunks, defined by three backticks
- narrative goes outside of chunks
- Simple markdown syntax for text

Tour: R Markdown



R Markdown help in RStudio

R Markdown Cheat Sheet `Help -> Cheatsheets`



R Markdown : : CHEAT SHEET

What is R Markdown?

.Rmd files · An R Markdown (.Rmd) file is a record of your research. It contains the code that a scientist needs to reproduce your work along with the narration that a reader needs to understand your work.

Reproducible Research · At the click of a button, or the type of a command, you can rerun the code in an R Markdown file to reproduce your work and export the results as a finished report.

Dynamic Documents · You can choose to export the finished report in a variety of formats, including html, pdf, MS Word, or RTF documents; html or pdf based slides, Notebooks, and more.

Workflow

❶ **Open a new .Rmd file** at File ▶ New File ▶ R Markdown. Use the wizard that opens to pre-populate the file with a template

❷ **Write document** by editing template

❸ **Knit document to create report**; use knit button or **render()** to knit

❹ **Preview Output** in IDE window

❺ **Publish** (optional) to web server

❻ **Examine build log** in R Markdown console

❼ **Use output file** that is saved along side .Rmd

set preview location · insert code chunk · run code chunk(s) · go to code chunk · publish · show outline · run all previous chunks · modify chunk options · run current chunk

```
---
title: "R Markdown"
author: "RStudio"
output:
  html_document:
    toc: TRUE
---

```{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.

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

For more details on using R Markdown
see <http://rmarkdown.rstudio.com>.
```

```
> library(rmarkdown)
> render("report.Rmd", output_file = "report.html")
```

render

Use rmarkdown::**render()** to render/knit at cmd line. Important args:

input - file to render **output_format**	**output_options** - List of render options (as in YAML)	**output_file** **output_dir**	**params** - list of params to use

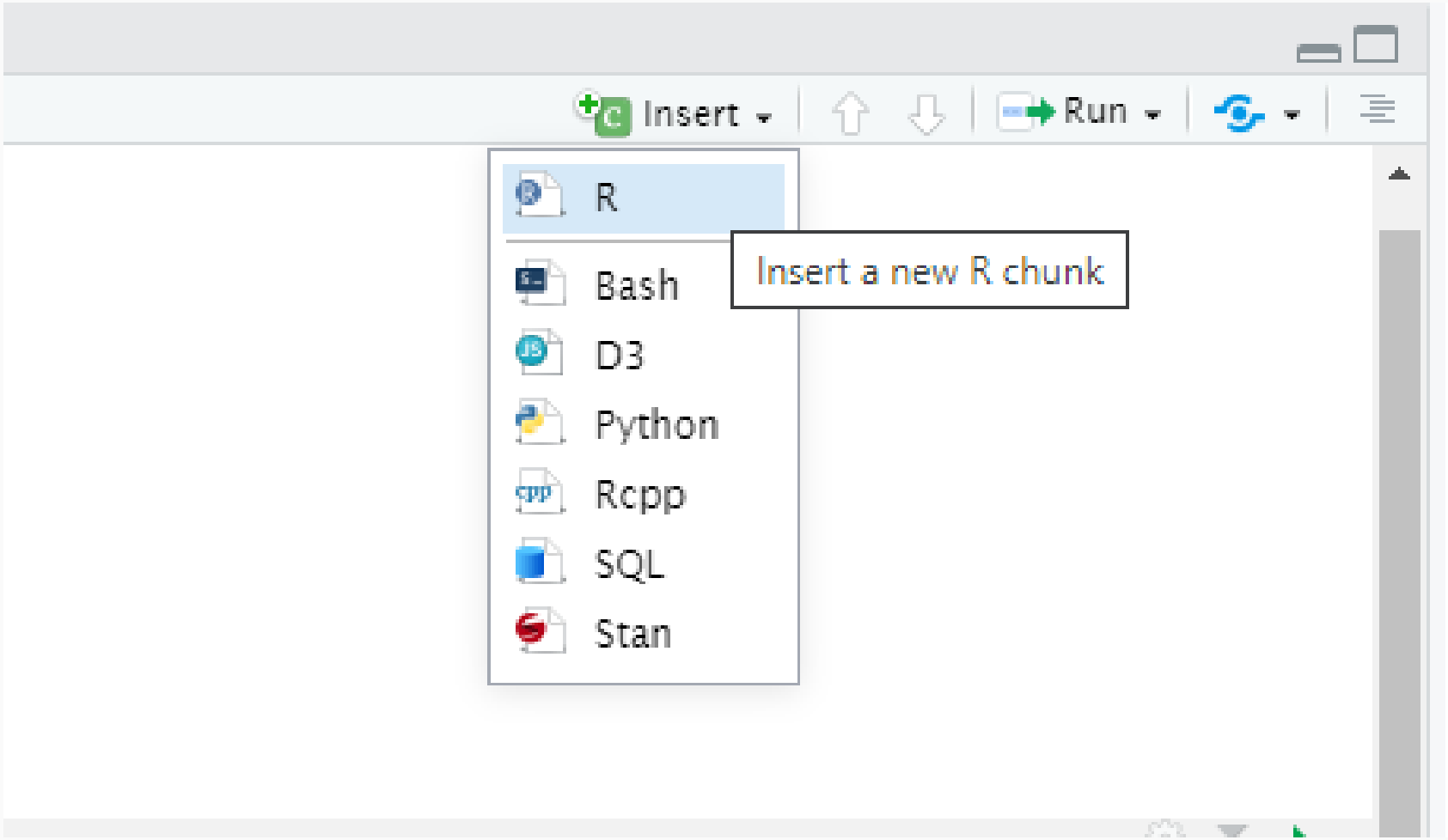
How will we use `R Markdown`?

- You'll always have a template R Markdown document to start with.
- You can also create your own .Rmd file (`File -> New File -> R markdown .. -> Knit` and modify text or code as necessary.)

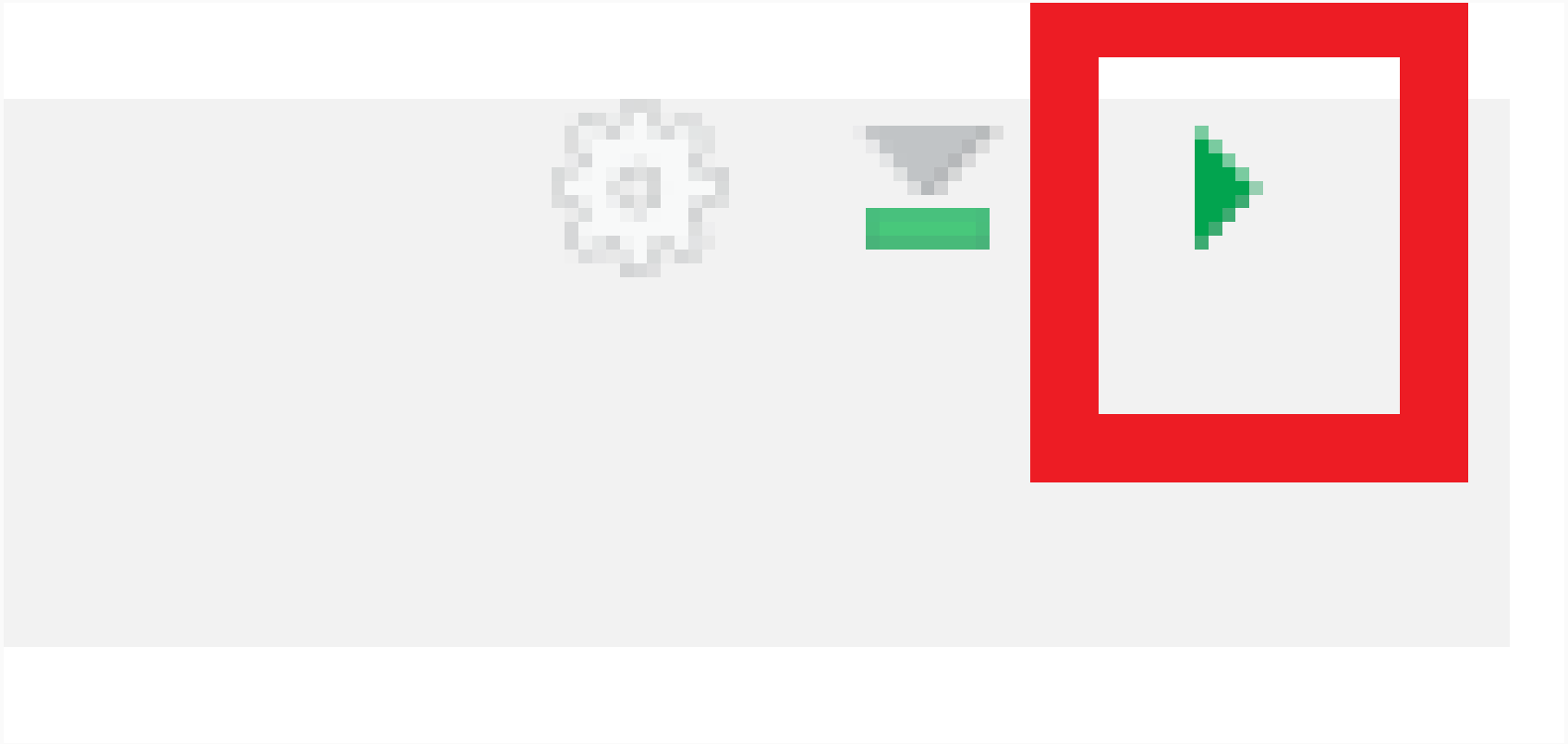
(Optional) Your Turn: Rmarkdown basics

- Open `yournetid_workingfiles/L00/P01_optional/ -> P01_netid.Rmd`
- This is optional but highly recommended.

Demo - Adding Chunks



Demo - Run Code



- Run entire chunk using the “green triangle”
- Run any part(one or more lines) of your code by selecting it an pressing Ctrl+Enter / Cmd+Enter

Lecture 0 Summary

- Core elements of Data Science project life-cycle
  - Programming
  - Statistics
  - Probability
- Accessing Rstudio server instance for the course
- Created a Data Science project report for UN votes.
- Rmarkdown essentials.(Complete it but do not turn it in.)

Post Lecture 0 to-do for you

- Read syllabus carefully
- Note down important dates,
- Get familiar with Course site on Canvas
- Visit Office hours
  - Get help with lecture material if you struggled in lecture today.
  - Practice will make it perfect for you!

Have a great start to the course! See you next lecture!