0.Welcome to PSTAT 10!

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

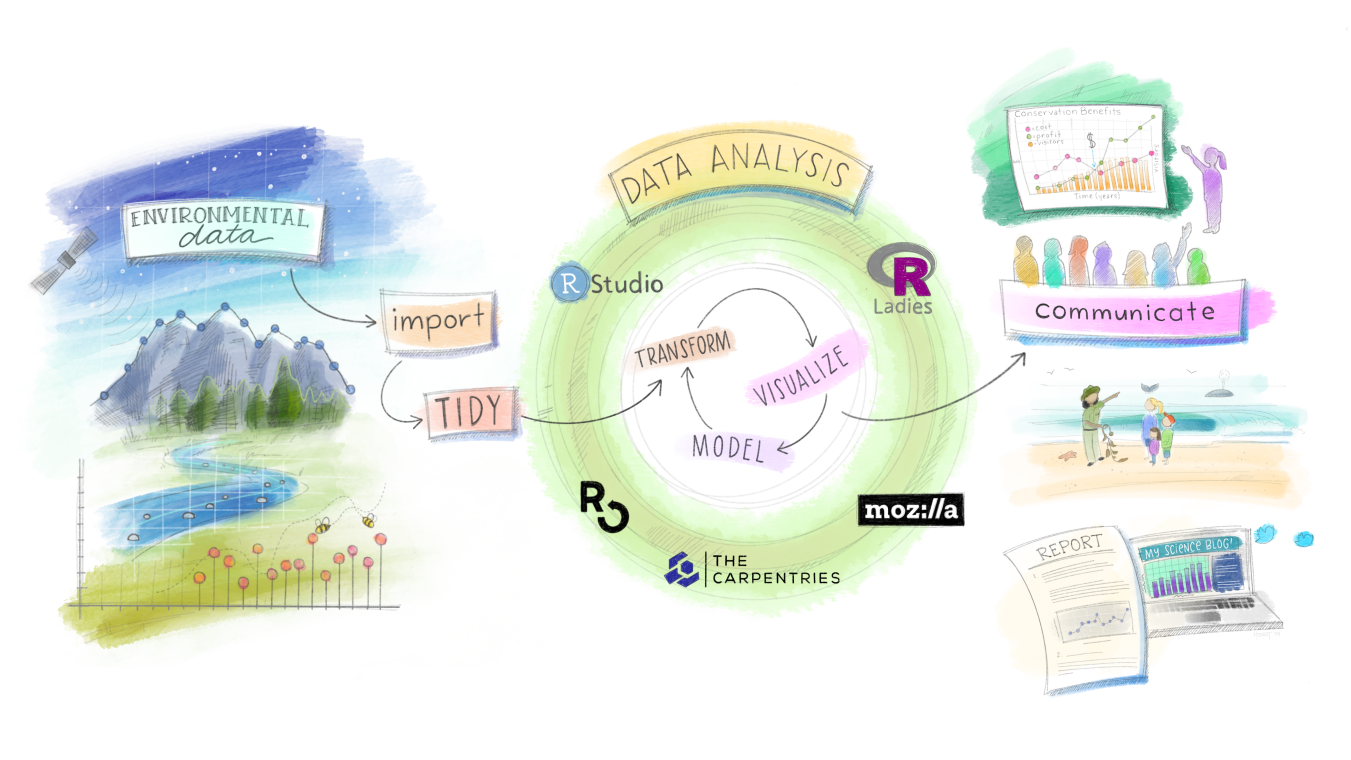
0.1 Welcome!

Plan for today

1. What is Data Science?
2. Your Turn at Data Science
3. **Course Overview** Read syllabus at length after lecture
4. **(Course Toolkit)** Rmarkdown - finish in section 1, homework 1

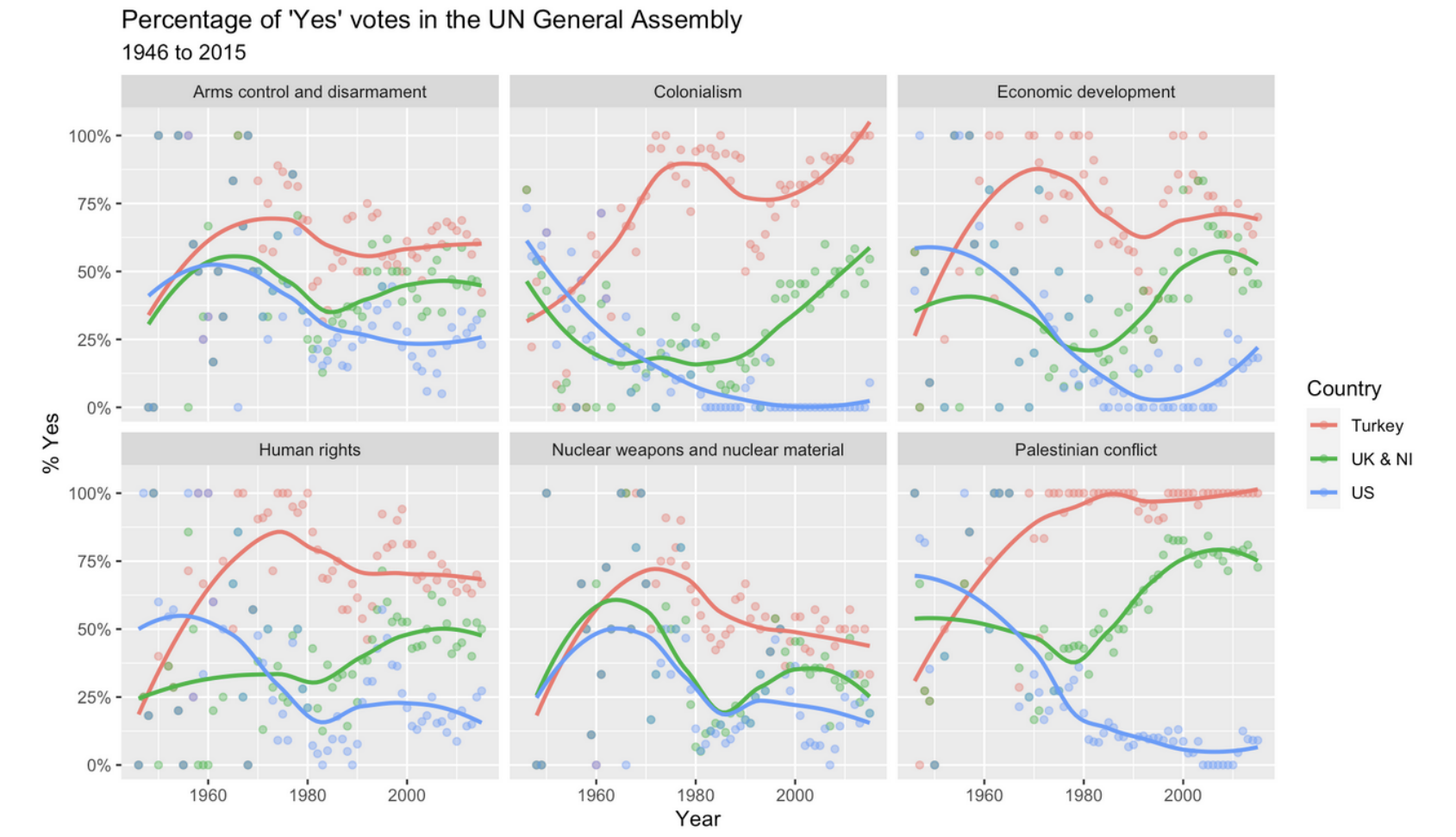
What is Data Science?

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



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.

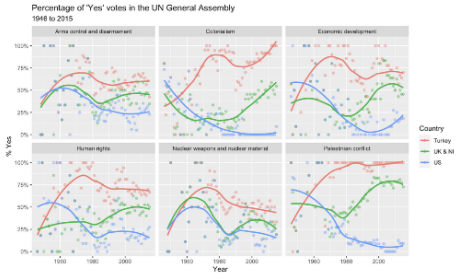
```
unvotes <- read_rds("data/unvotes.rds")
```

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

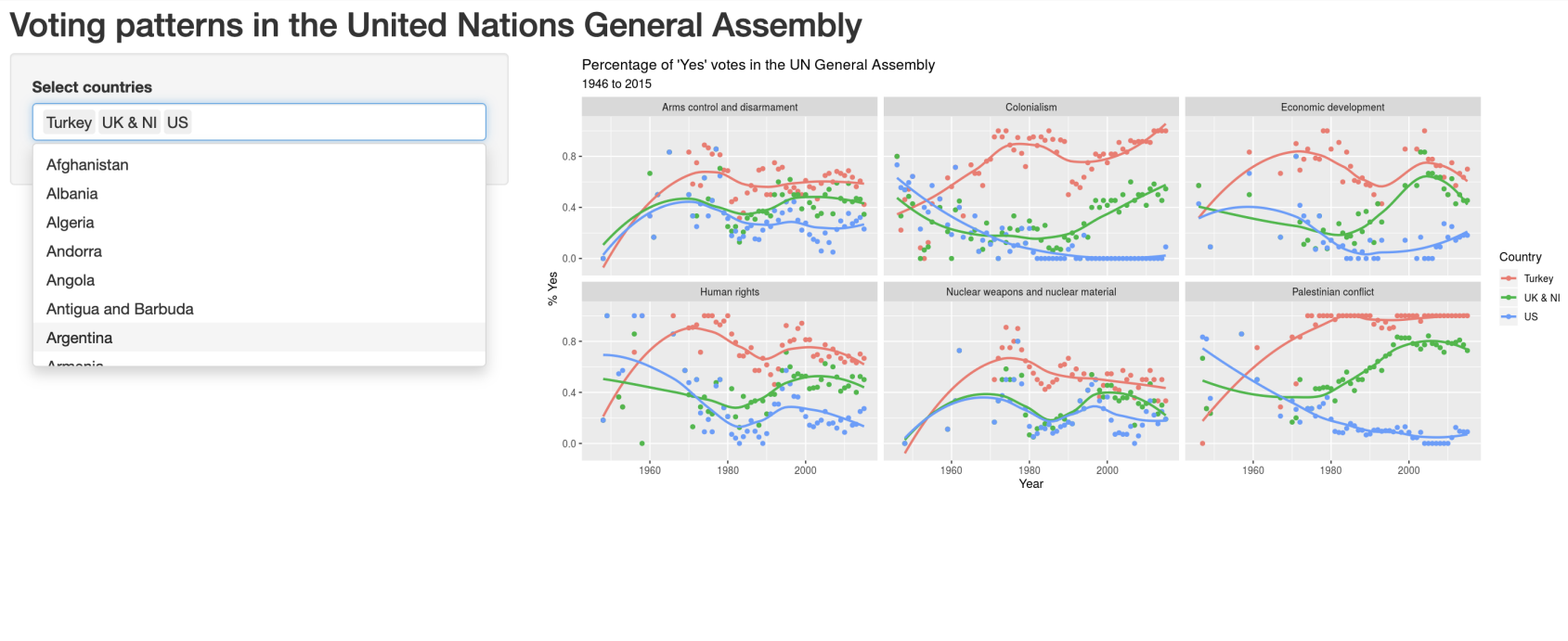
Below is a list of countries in the dataset:

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).
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. (for large amount of data) **Working with Databases** using **SQL** (Structured Query Language) to communicate with all the data stored in a database.

- This course is an introduction to these components of data science using the software `R`

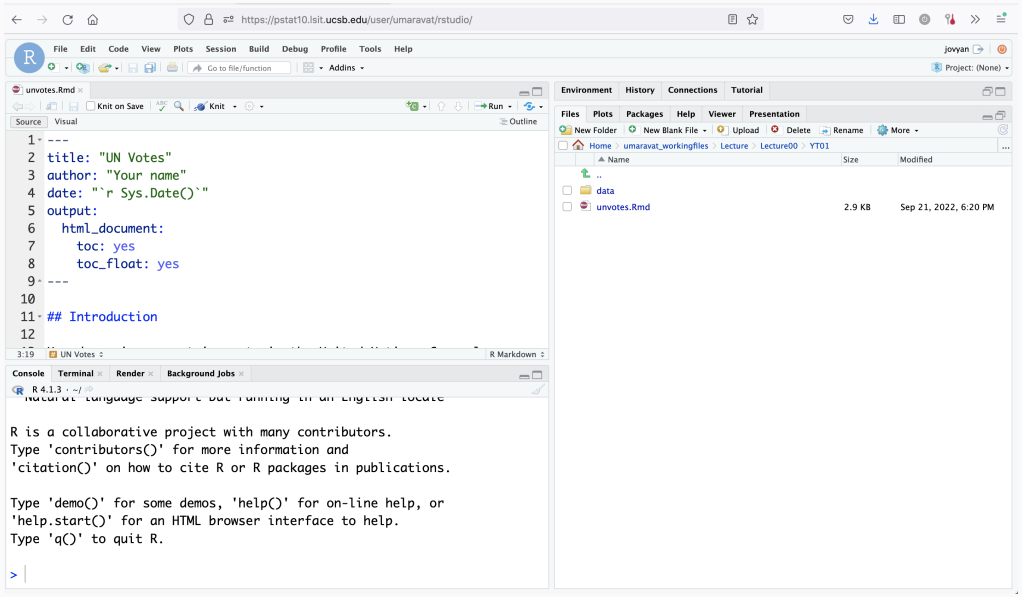
0.2 YOUR TURN 1 (10 minutes)

0.2 YOUR TURN 1 (10 minutes)

1. Login to our PSTAT 10 Rstudio computing environment using your **ucsb credentials** at the Rstudio Server link on left navigtion menu on Canvas

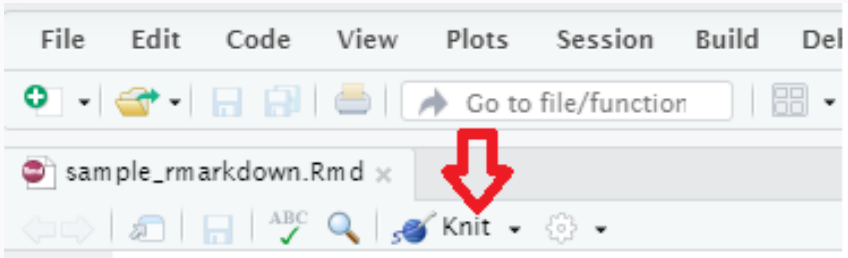
.Rmd

2. In the `Files pane` (bottom right corner), navigate to **yournetid_workingfiles** -> Lecture -> Lecture00 -> YT01 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.

Review Your Turn 1

0.3 Course and syllabus overview

Course Objectives

Using the R programming language to introduce the basics of data science

Part 1. Learn the basics of programming using the R programming language.

Part 2. Learn Introductory Statistics and Probability for Data science using R

Part 3. Learn Relational Databases and SQL by interfacing with R

Student learning outcomes

By the end of the course, students should have a solid foundation in

- programming with R,
- a basic understanding of probability and simulations and
- the ability to work with databases using SQLite.

This course provides a stepping stone for further studies in data science and related fields.

Administrative items

1. No crashing - wait till you are officially enrolled
2. No section switching
3. On Canvas, see syllabus and read about office hours, email policy, other polices.

Course success

Show up, don't fall behind.

1. Syllabus **Read it**
2. Lectures **Show up and participate**
3. Labs **Don't miss**
4. Homeworks **Do them on time, come to Office Hours for help**
5. Homework Reflection Surveys **Don't miss**
6. Quizzes **Schedule a time to take it in Week 2,3,5,7,10**
7. Final exam **Don't miss**
8. Office hours are helpful - over 20 hours each week. Make use of them!

Don't share my work with anyone or on sites like Coursehero, Chegg, AI etc.

Preferred order of Methods of communication

1. **Office Hours**
   - Come and get real time answers to your questions or just say hi!
   - ULA office hours
   - TA office hours
   - instructor office hours
   - HW clinics
   - over 20 hours each week. **Please make use of these help hours**

3. Email for **truly private matters** and follow **email policy**
   - Grade discrepancies do not need an email - Fill in the form
   - your TA is your first point of contact for any issues regarding anything else about section, homework, quiz
   - Read syllabus for DSP accomodations and conflicts before contacting Instructor during office hours for any DSP accommodation and conflicts
   - instructor for private matters that you do not want any other

Collaboration and sharing your/my work

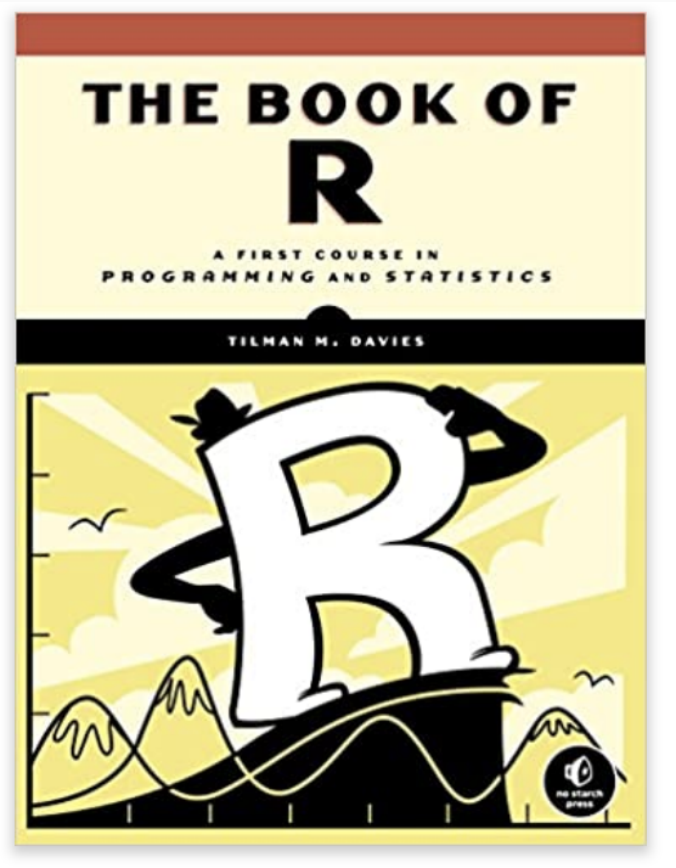
Asking questions and helping others is encouraged

- Discussing course topics with other is welcome!
- Discuss with peers sitting next to you.
- Ask questions, answer questions in class
- Don't judge people or questions
- share ideas liberally.

Limits of collaboration

- Don’t share solutions with each other
- You should never see or have possession of anyone else’s solutions, including from previous versions of this course
- See syllabus about Heuristics to keep in mind while sharing

TEXTBOOK



You do not need a textbook for this class. Slides will be provided and starter .Rmd will be provided.

Title: The Book of R: A First course in Programming and Statistics

Author: Tilman M. Davies

Publisher: No Starch Press

Summary of Technology/Platforms

Personal PC, Mac or Laptop, Phone : To access course sites

Canvas Course site: Canvas for submitting assignments, grades, reviewing solutions

Rstudio Instance: for code files for lecture, lab, homework

Any of the many Office hours: for questions

Email: Use sparingly for truly private matters for lab TA, headTA or instructor

Do not email course staff separately for the same issue - Use Emails' CC feature while emailing multiple course staff about the same issue.

Most importantly...

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

0.4 R Markdown documents

0.4 R Markdown documents

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

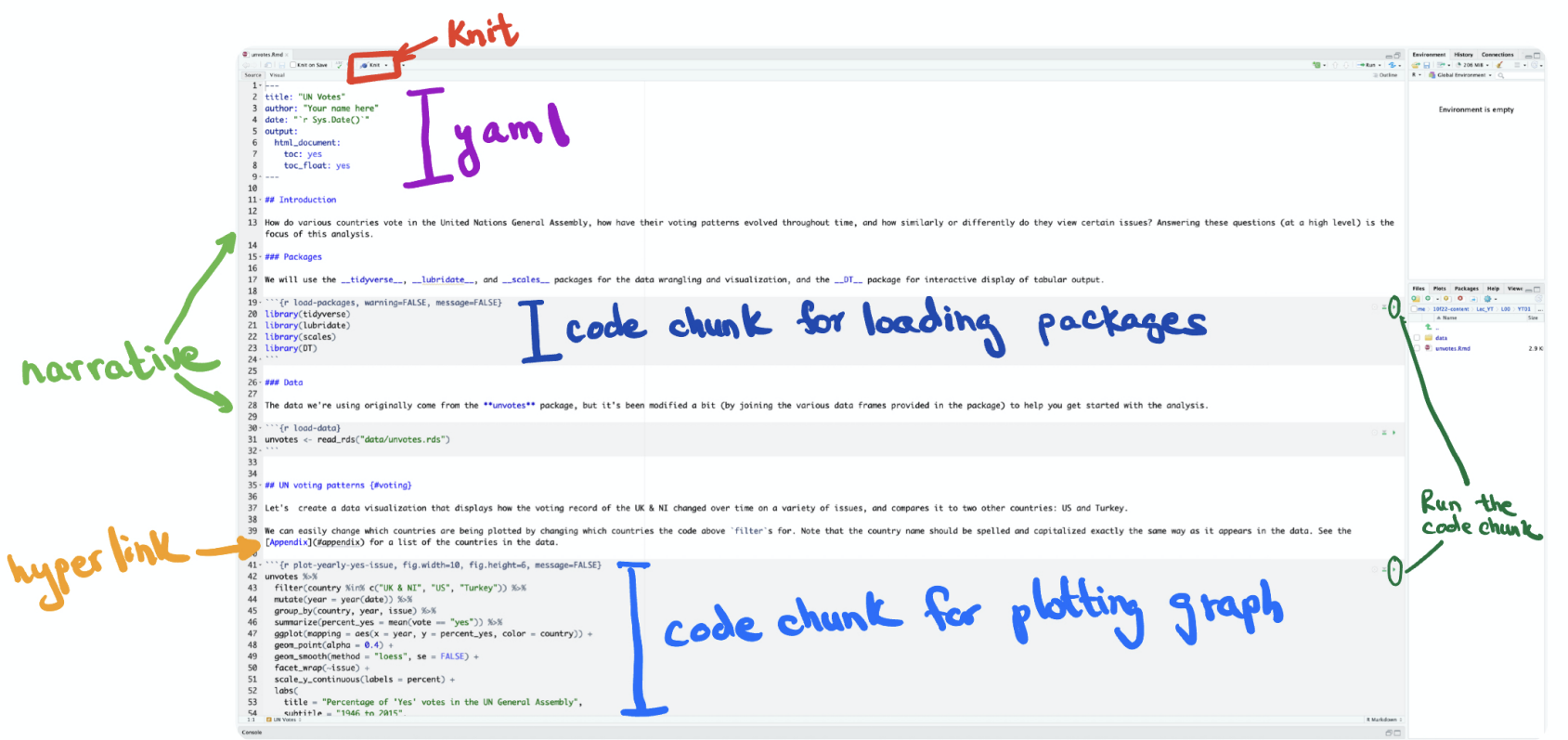
rmarkdown.rstudio.com

Take a look at the **Rmarkdown gallery**

R Markdown syntax

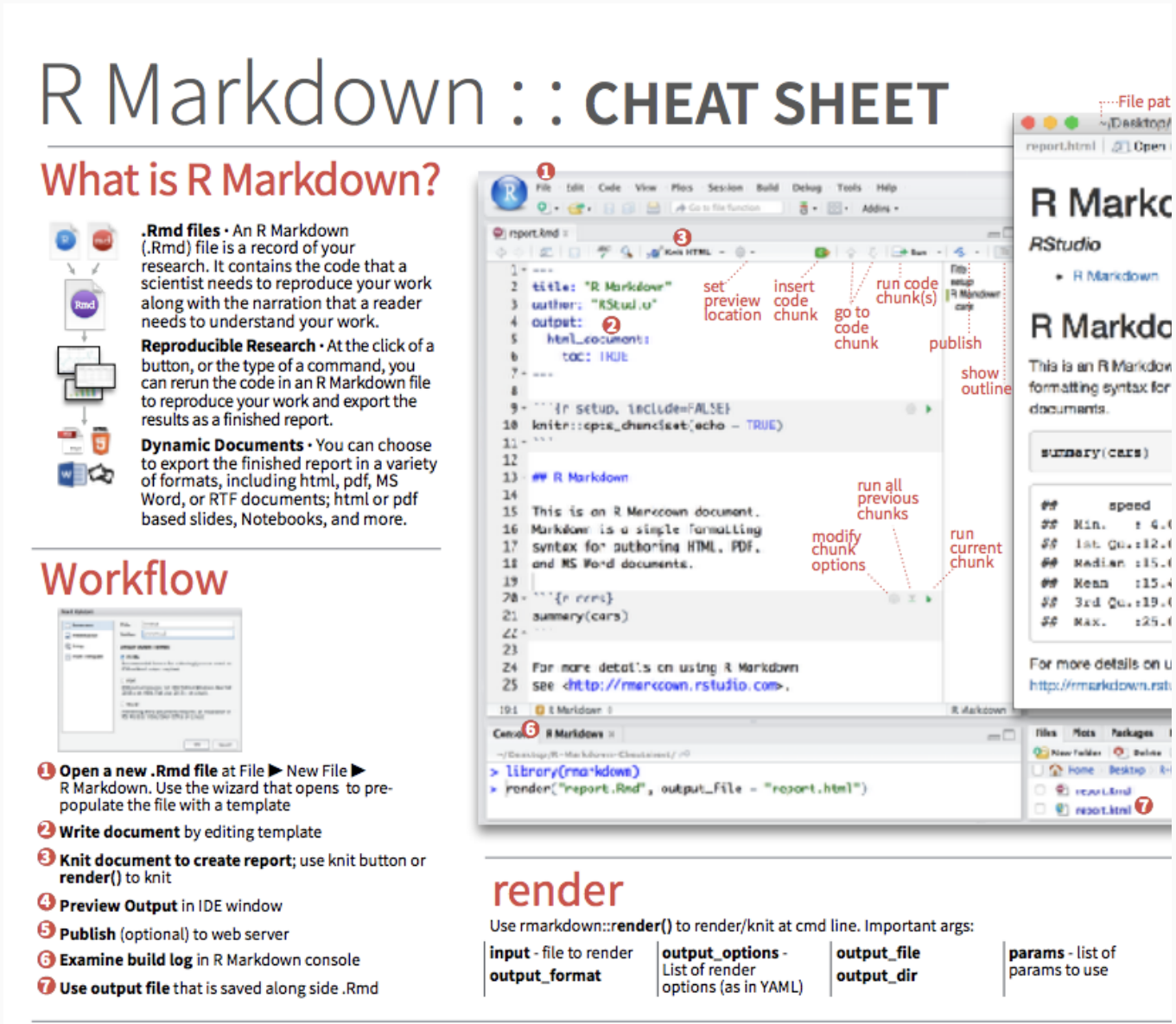
- Code goes in code chunks, defined by three backticks
- narrative goes outside of chunks
- Simple markdown syntax for text

Tour: R Markdown



R Markdown help in RStudio

R Markdown Cheat Sheet `Help -> Cheatsheets`



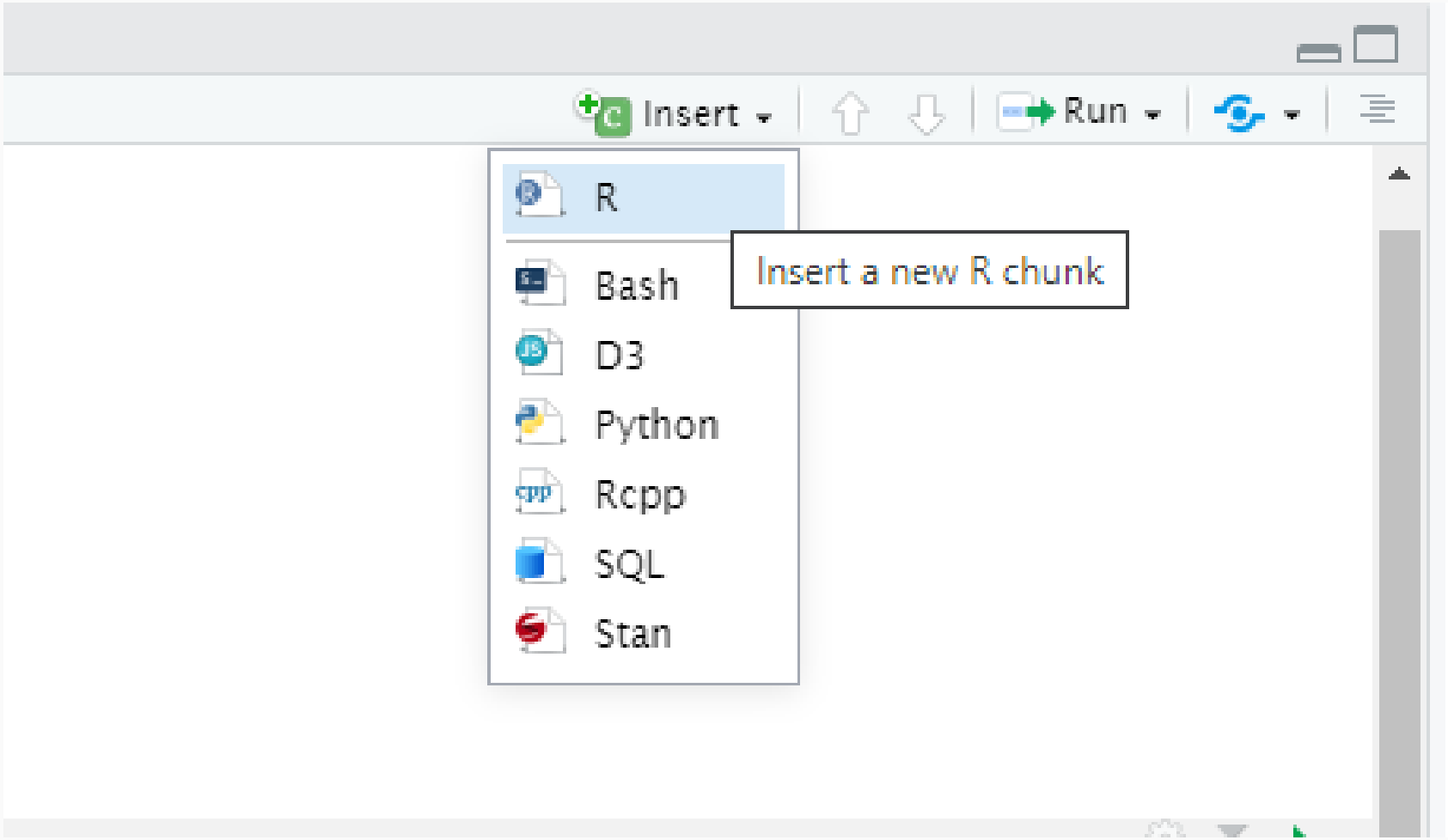
How will we use R Markdown?

- Every worksheet / homework etc. is an R Markdown document
- You'll always have a template R Markdown document to start with
- The amount of scaffolding in the template will decrease over the quarter
- You can also create your own .Rmd file (File -> New File -> R markdown .. -> Knit and modify text or code as necessary.)

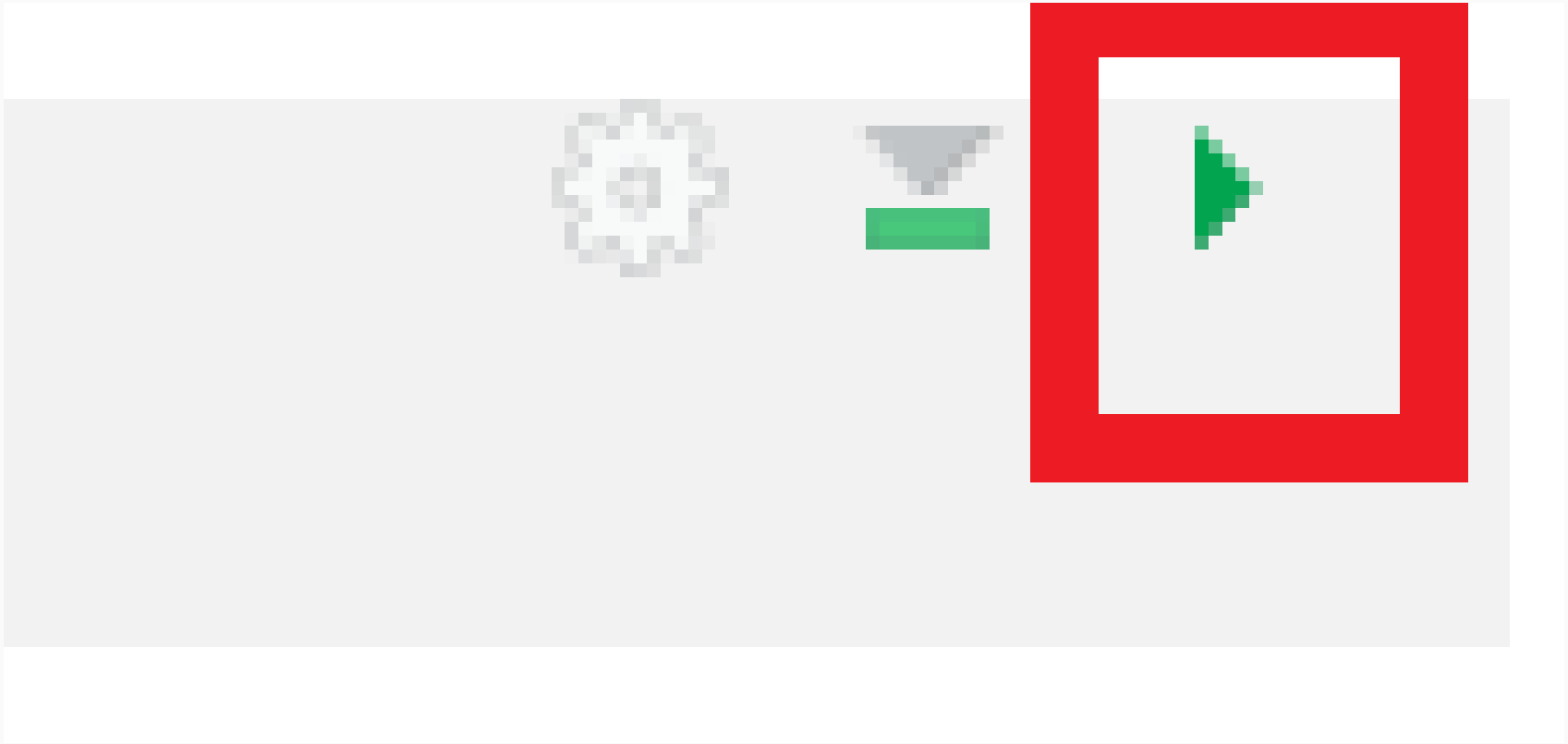
Your Turn: Rmarkdown basics

- Open `yournetid_workingfiles/Homework/HW01/ -> Homework01.Rmd`
- You will start here, continue in Section, complete it and submit it as your first homework on Wednesday on canvas.
- HW Submission link will be released in Week 1 module on Canvas
- Use office hours to practice, review ask questions about today's lecture and these exercises.

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

Summary: Introduction

- Core elements of Data Science project life-cylcle
  - Programming
  - Statistics and Probability
  - Databases
- Accessing Rstudio instance for the course
- created a Data Science project report for UN votes.
- Course overview and Brief Syllabus walk through
- Rmarkdown essentials.(Complete it in section 1)

Post Lecture 0 to-do for you

- Read syllabus carefully
- Note down important dates, final exam
- Get familiar with Course site on Canvas
- Go to both Sections each week and ask questions when you are stuck.
- Complete Homework 1 and submit on time.
- 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 quarter! See you next lecture!