

# **1.Welcome to PSTAT 188!**

Transfer exploration seminar: Statistics and Data Science

---

Dr. Uma Ravat

PSTAT 188

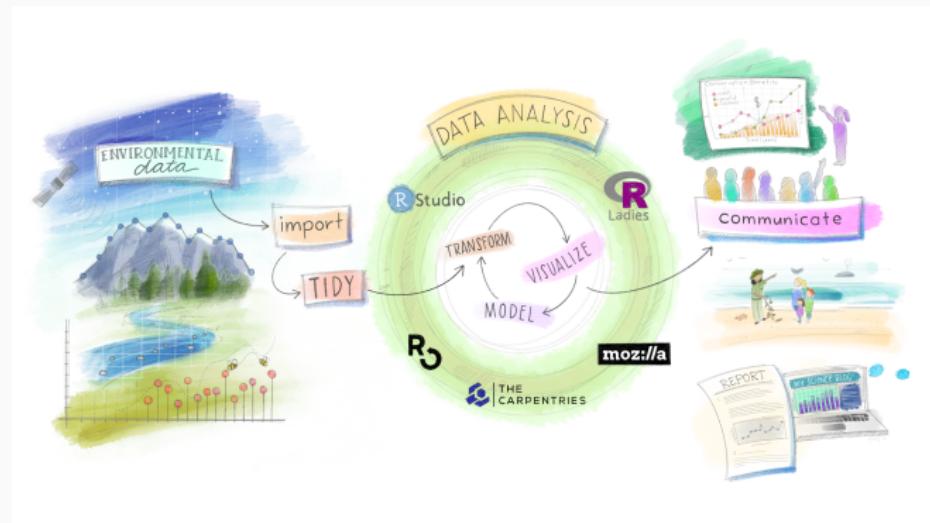
## 0.1 Welcome!

### Plan for today

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

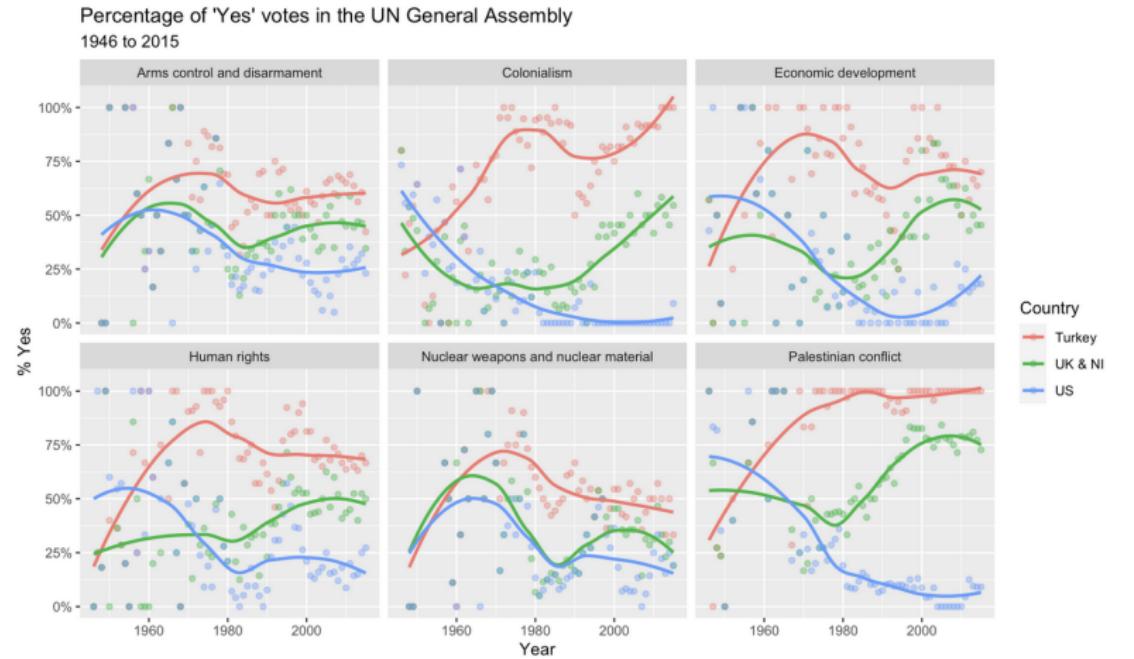
# What is Data Science?

- 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 Visualization



## 2. Report/Website

**Introduction**

U.S. election patterns

Appendix

---

**U.N. 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/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`, `shiny`, and `shinyWidgets` packages for data cleaning and visualization, and the `gapminder` package for geographical data.

```
library(tidyverse)
library(shiny)
library(shinyWidgets)
library(gapminder)
```

### Data

The data we're using originally came from the `anewdemos` 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.

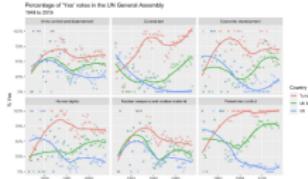
```
library("readr")
library("anewdemos", "gapminder")
```

## U.N. voting patterns

The following code shows the voting record of the U.N.G.A. changed over time on a variety of issues, and compared to its two other counterparts, UN and Turkey.

```
# We can use the gapminder package to change which countries the code shows. UN for a list of countries, and Turkey for a specific country should be selected and capitalized exactly as they appear in the data. See the Anycards for a list of countries.
```

```
countries %>
  filter(general == "UN", id == "ID", name == "Turkey") %>
  mutate(general = "UN", year = as.Date(year), percent_1st = percent_1st / 100,
        percent_2nd = percent_2nd / 100, percent_3rd = percent_3rd / 100)
  # calculate the sum of the three percentages
  # calculate the total number of votes cast
  # calculate the percentage of "Yes" votes in the UN General assembly
  # calculate the same for Turkey
  # calculate the difference between the two
  # calculate the difference between the two
  # calculate the difference between the two
```



### References

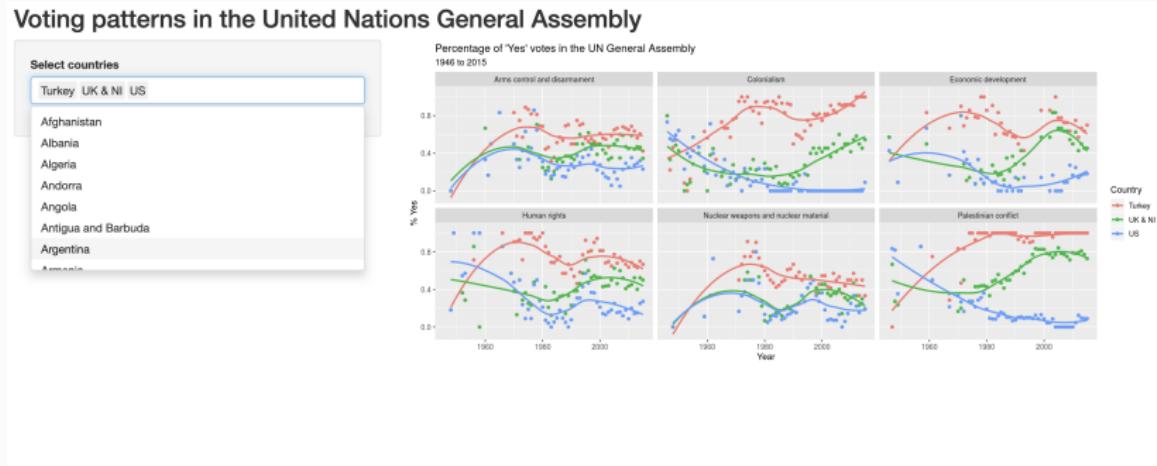
6. *Wine-Datagroup.Rundf*, Introduction to Data science course  
1. David Robinson (2011). *UNIVote*: United Nations General Assembly Voting Data. R package version 0.2.3.  
2. Eric Weisstein. "Data and Analyses of Voting in the UN General Assembly". Routledge Handbooks of International Organizations, edited by Bob Reiterman (published May 27, 2008).  
3. Much of the analysis has been modified on the exercises presented in the *univote* package vignettes.

## Appendix

Below is a list of countries in the dataset.

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

### 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 review 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 . . .

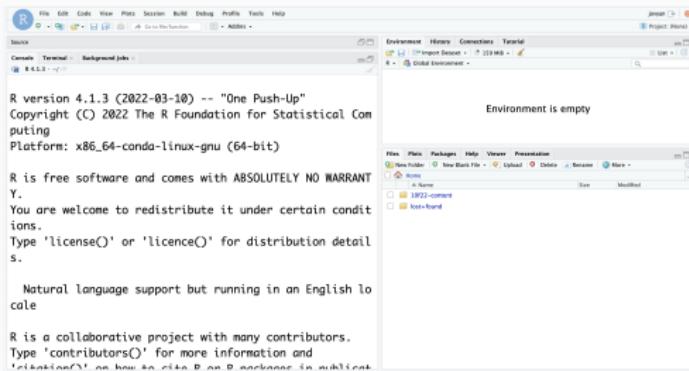
- Arrive on time
- Stay engaged
  - Avoid distractions (phone, email, social media, messaging apps turned off)
  - Actively participate
  - Ask questions
- Keep your files organized.
- Take half an hour after each class to review the materials and revise your notes.
- Take fifteen minutes before next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



The screenshot shows the RStudio interface. The left pane is the R Console, displaying the R startup message and basic information about the R version and platform. The right pane is the Environment pane, which is currently empty. Below the main panes is the file browser, showing a directory structure with two files: '19P22-comes' and 'test.html'.

```
R version 4.1.3 (2022-03-10) -- "One Push-Up"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-conda-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
Y.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'trivia()' on how to cite R or R packages. To add your
```

## 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.

The screenshot shows the RStudio interface. The top navigation bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a gear icon for Preferences. Below the menu bar is a toolbar with icons for Source, Visual, Knit or Save, and other functions. The main workspace is divided into several panes: a Source pane containing R code, a Console pane showing R startup messages, a Terminal pane, a Render pane, and a Background Jobs pane. The bottom right corner features a 'Files' pane titled 'Environment' which displays a file tree. The tree shows a 'unvotes.Rmd' file under '3YT01' in the 'Lecture' folder, which is itself under 'workingfiles' in the 'Lecture00' project. The 'unvotes.Rmd' file has a size of 2.9 kB and was modified on Sep 21, 2022, at 6:28 PM.

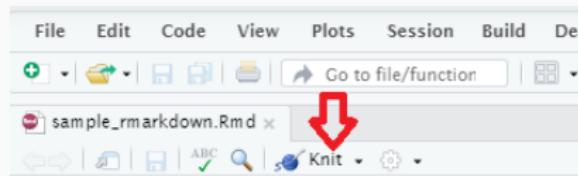
```
1: ---
2: title: "UN Votes"
3: author: "Your name"
4: date: "r Sys.Date()"
5: output:
6:   html_document:
7:     toc: yes
8:     toc_float: yes
9: ---
10: # Introduction
11:
12:
```

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

# 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](https://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

**Knit**

Iyaml

**narrative**

**hyperlink** →

I code chunk for loading packages

I code chunk for plotting graph

Run the code chunk

```
1 #> source('yaml.R')  
2  
3 #> title("UN Votes")  
4 #> subtitle("Your name here")  
5 #> date = Sys.Date() -  
6 #> output: yaml  
7 #> toc: true  
8 #> toc_depth: 2  
9  
10  
11 ## Introduction  
12  
13 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.  
14  
15 Load Packages  
16  
17 We will use the tidyverse, shinydashboard, and shiny packages for the data wrangling and visualisation, and the grid package for interactive display of tabular output.  
18  
19 

```
(r load.packages, warning=FALSE, message=FALSE)
```

  
20 library(tidyverse)  
21 library(shiny)  
22 library(shinydashboard)  
23 library(grid)  
24  
25 ## Load Data  
26  
27 # The data we're using originally came from the munitor 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.  
28  
29 

```
(r load.data)
```

  
30 munitor <- readRDS("data/unvotes.rds")  
31  
32  
33 ## UN Voting Patterns (Meeting)  
34  
35 Let's create a data visualization that displays how the voting record of the UK & NL changed over time on a variety of issues, and compares it to two other countries: US and Turkey.  
36  
37 We can easily check which countries are being plotted by checking which countries the code above filters for. Note that the country name should be spelled and capitalised exactly the same way as it appears in the data. See the Dependencies for a list of the countries in the data.  
38  
39 

```
(r plot.party.votes.issue, filter=cb, filg.height=50, filg.message=FALSE)
```

  
40 unvotes %>%  
41 filter(country %in% c("UK", "NL", "Turkey")) %>%  
42 mutate(year = year(date)) %>%  
43 group_by(country, year, issue) %>%  
44 summarise(percent = sum(vote == "yes")) %>%  
45 ggplot(mapping = aes(x = year, y = percent, just = "center")) +  
46 geom_line(mapping = aes(x = year, y = percent), size = 1) +  
47 geom_smooth(mapping = aes(x = year, se = FALSE) +  
48 facet_wrap(~issue) +  
49 theme_minimal() +  
50 labs(y = "Percentage of 'Yes' votes in the UN General Assembly",  
51 x = "Year")
```

# 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



- ① Open a new .Rmd file at File ► New File ► R Markdown. Use the wizard that opens to pre-populate the file with a template
- ② Write document by editing template
- ③ Knit document to create report; use knit button or render() to knit
- ④ Preview Output in IDE window
- ⑤ Publish (optional) to web server
- ⑥ Examine build log in R Markdown console
- ⑦ Use output file that is saved along side .Rmd

The screenshot shows the RStudio interface. In the top-left, the R Markdown editor displays a code block starting with `# R Markdown` and a summary table for the `cars` dataset. In the top-right, the R console shows the command `library(rmarkdown)` and its execution. Below the editor, the file browser shows a folder named 'report.Rmd' with several sub-folders like 'report.html' and 'report.Rmd'. A red box highlights the 'File path' entry in the file browser.

## 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

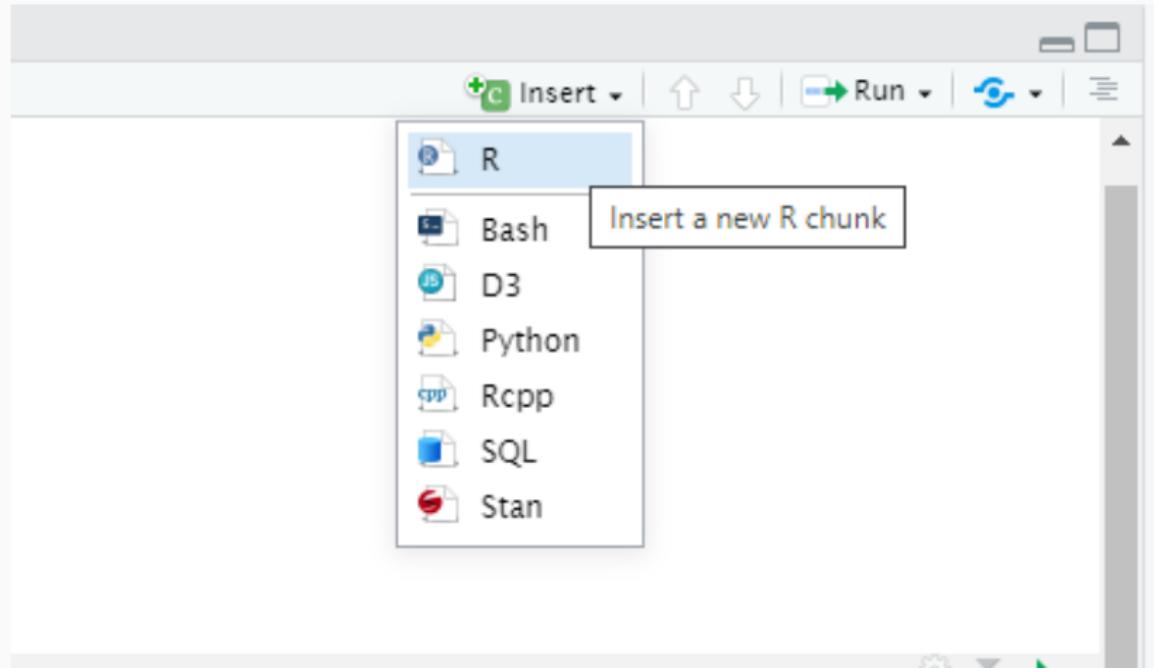
## How will we use R Markdown?

- 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.)

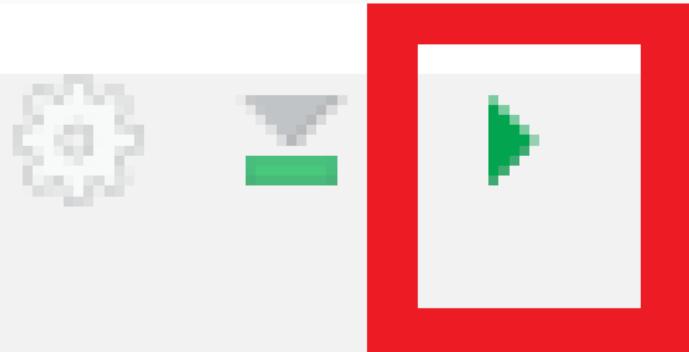
## (Optional) Your Turn: Rmarkdown basics

- 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 and 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!