

Data Science Basics in R

Day 2: Data management and version control

Goals for today

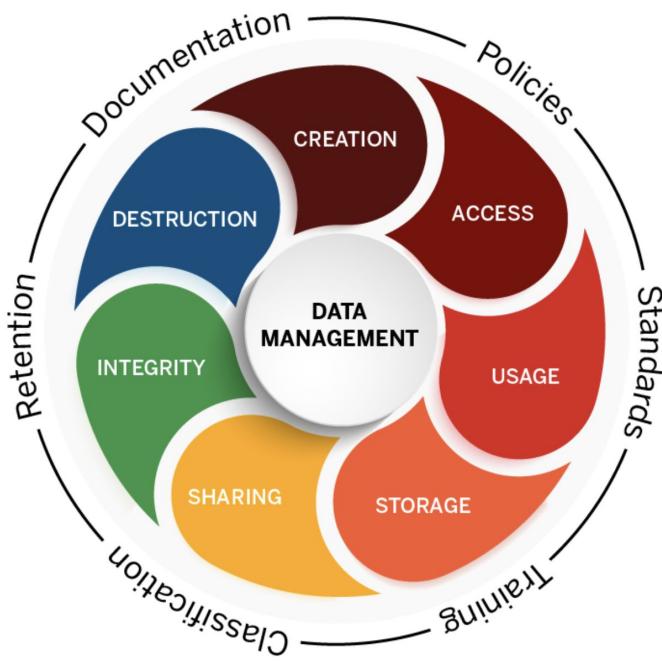
- Understand the foundations of data management
- Learn best practices for documentation
- Load and clean your first dataset in R
 - learn about data structures
 - filter data to excluding missing values
 - explore a new dataset with summary statistics and plots
- Upload your code to GitHub

What is data management?

Data management

Data management is the practice of creating, maintaining, documenting, and securing data.

It includes data cleaning, among many other tasks.



Data access

- Where do your data “live”?
 - How can you access them?
 - Who else can access them?
 - For how long will they be available?
- What format are they in?
- Where is the documentation?
- How, and with whom, can you share data, code, and/or results?
- Identifying other discrepancies...

Data management

Data cleaning tasks depending on what types of data you're working with and where you got them!

- Identifying and addressing missing values
- Fixing “weird” characters
- Combining multiple datasets
- Deduplicating lists
- Identifying other discrepancies...

Real-life examples data entry and coding

Global approaches to tackling antimicrobial resistance: a comprehensive analysis of water, sanitation and hygiene policies

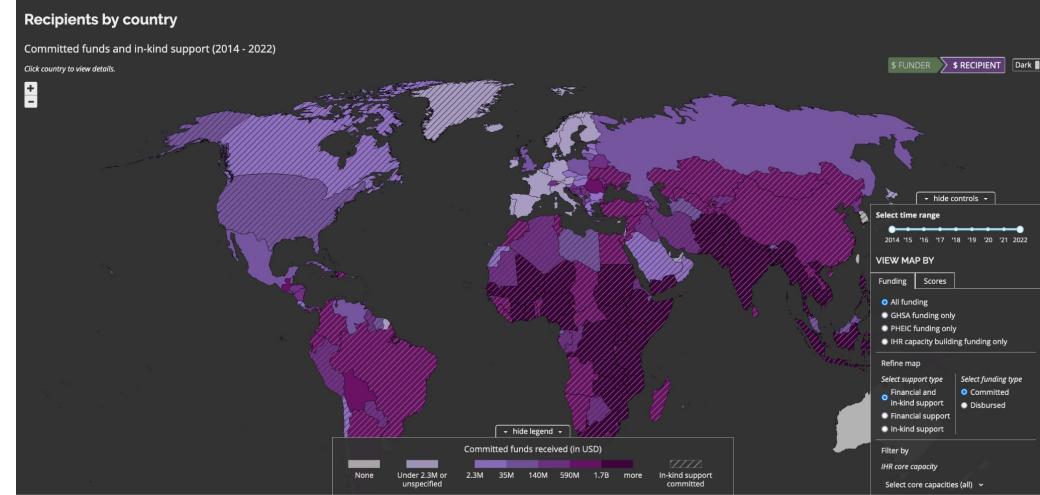
Ciara M Weets  , Rebecca Katz

A	B	C	D	E	F	G	H	
1	Country	Topic	Status justification	Color	Level	Government	Voting	Subtopic
2	Argentina	Sewerage	There is policy that requires a competent agency within the government to regulate sewerage systems for the consumption of human waste, however, in Argentina, sanitation services are regulated at the provincial level.	Option 1	Subnational	Federal	Presidential	Sewerage
3	Australia	Water quality standards	There is policy that mandates a competent environmental agency within the governing body to control and set standards for water quality, including drinking-water quality, environmental water quality, and/or recreational water quality. This happens at the provincial level, rather than the national level in Australia.	Option 1	Subnational	Federal	Parliamentary	Water Quality Standards
4	Australia	Water quality monitoring	There is policy that grants a competent environmental agency within the governing body with a mandate to monitor residues in water sources to prevent contamination that could cover contamination with antimicrobials. This happens at the provincial level, rather than the national level in Australia.	Option 1	Subnational	Federal	Parliamentary	Water Quality Monitoring
5	Australia	Pollutant disposal in water sources	There is policy that mandates a competent environmental agency within the governing body to regulate the disposal of pollutants in freshwater sources. This happens at the provincial level, rather than the national level in Australia.	Option 1	Subnational	Federal	Parliamentary	Effluent Wastewater Disposal
6	Australia	Effluent wastewater disposal	There is policy that mandates a competent environmental agency within the governing body to regulate the disposal of effluent wastewater, however, in Australia, this regulation takes place at the subnational level.	Option 1	Subnational	Federal	Parliamentary	Pollutant Disposal in water sources
7	Australia	Sewerage	There is policy that requires a competent agency within the government to regulate sewerage systems for the consumption of human waste, however, in Australia, sanitation services are regulated at the provincial level.	Option 1	Subnational	Federal	Parliamentary	Medical Waste Disposal
8	Australia	Medical waste disposal	There is policy that requires the government to regulate the disposal of medical waste from healthcare facilities, however, in Australia, this regulation takes place at the subnational level.	Option 1	Subnational	Federal	Parliamentary	Total

Real-life examples data aggregation



Data source	Description
BWC Working Papers	Occasionally, Member States of the BWC submit national papers describing capacity building efforts for health security. Data relevant from these papers are included in GHS Tracking.
CEPI Progress Reports	The Coalition for Epidemic Preparedness Innovations (CEPI) is a partnership created in 2017 to bring together global stakeholders to develop vaccines with the goal of stopping future pandemics. During the COVID-19 pandemic that began in 2019, it operated as a facilitator in the COVAX Marketplace. CEPI announces funding calls for vaccine candidate development and coordinates partnerships to develop specific vaccines as needed. CEPI releases annual progress reports that include actual and planned donor funding amounts by year. More information about CEPI is available at https://cepi.net/ .
Ebola Recovery Tracking Initiative	The Ebola Recovery Tracking Initiative tracks official development assistance towards Ebola recovery efforts in Guinea, Liberia, Sierra Leone, and the Mano River Union. The initiative is a partnership between the governments of Guinea, Liberia and Sierra Leone, the United Nations Office of the Secretary-General's Special Adviser on Community Based Medicine and Lessons from Haiti, and the United Nations Development Programme (UNDP). The Ebola Recovery Tracking Initiative is available online at: https://ebolarecovery.org/ .



tracking.GHScosting.org

Real-life examples reformatting & cleaning

Methods

- Approximately 54% of PLM patients contribute free-text to the platform through user bios, forum conversations, or as annotations associated with structured data.
- Posts contributed by PTSD patients were pre-processed to remove HTML and stopwords, and to map emojis to text-based descriptions.
- Latent Dirichlet allocation (LDA), a form of probabilistic topic modeling, was performed to identify topics discussed among patients.
- Model parameters were selected on the basis of perplexity as measured on a 10% holdout group.

The Patient Voice Includes Emojis:

A case study in the use of probabilistic topic modeling to characterize patient conversations in an online community of PTSD patients

Eaneff, S

Fig 3A. "Sleep" Topic

Top 3 emojis



Sample post*

"Nightmares kicked my ass last night. Got woken up by a nightmare and stayed up for a long time feeling really freaked out."

tonight good early awake hope
staying feel nap morning normal
feeling asleep work till tomorrow
bedtime sleeping body
apnea today hour days
nights bed nightmares
bad dream bad back yesterday fall
rest long time couple ready make
tired energy late staying
hours slept woke day sleeping
night sleep sleeping time have taking
sleepy stay sleeping by evening
trouble sleeping insomnia afternoon
stay sleeping by evening afternoon

Fig 3B. "School/Work" Topic

Top 3 emojis



Sample post*

"This semester, I'm taking four classes for the first time in a long time. I've quit my factory job, which was too physically demanding, and decided to focus on finishing school."

learn taking today break
position math training volunteer @ hope
teaching hard classes make degree study
end years part working learning balance
jobs back team starting career
great day physics year interview
week head program full kids
physics students teacher days things
head program full kids
program project start hours
teacher days things
full kids days things
team project start student
days things team project
work started fun
homework meeting community
research
high weeks
due research
long

Documenting your data, code, and results

Documentation

- Throughout your analysis, documentation might be the single most important, and also the single most under-appreciated step of the process.
- Documentation is helpful for :
 - **current you:** writing out simple, clear, documentation helps you think clearly
 - **future you:** you'll inevitably forget why you did certain things, even if you can see your old code
 - **others:** hopefully other people will also use your work, and they need to know about it

Documentation

Different materials, and different types of analysis benefit from different types of documentation.

- **Datasets:** Data dictionaries, rules to generate datasets
- **Code:** Comments in code, README files
- **Analysis/Results:** Methods documentation, in a word document, a google document, or in github

Documenting datasets:

Data dictionaries

The most common way to document a dataset is a data dictionary. At a minimum, data dictionaries:

- Define the columns (“fields”) of a dataset
- Specify what a row corresponds to (e.g., one country)
- Give context about what the data are, including where they come from and how often they are updated

Documenting datasets:

Data dictionaries

Data dictionaries come in lots of formats and flavors. Work with the type that makes the most sense for who you are trying to communicate with.

Documenting datasets:

Example data dictionary: USGS Landsat Sata

Acquisition Date

- *Field Definition:* The year, month and day that the scene was acquired.

Format:

YYYY/MM/DD

Acquisition Quality

- *Field Definition:* Acquisition Quality is a value, expressed as a single digit number, based on (1) errors encountered during archive processing; and/or (2) visible artifacts in the data when manually inspected.

Values:

9 = Excellent (no quality issues or errors detected)

7–8 = Good (minor quality issues and/or errors detected)

Bias Parameter File Name OLI

- *Field Definition:* The Bias Parameter File used for processing OLI.

Documenting datasets:

Example data dictionary: US Bureau of Labor Statistics

2022 ATUS Data Dictionary: Public ATUS Interview Data			
Name	Description		File
TEABSRSN	Edited: what was the main reason you were absent from your job last week?		Respondent File
	Edited Universe: TELFS = 2		
	Valid Entries:	1	On layoff (temporary or indefinite)
		2	Slack work/business conditions
		3	Waiting for a new job to begin
		4	Vacation/personal days
		5	Own illness/injury/medical problems
		6	Childcare problems
		7	Other family/personal obligation
		8	Maternity/paternity leave
		9	Labor dispute
		10	Weather affected job
		11	School/training
		12	Civic/military duty
		13	Does not work in the business
		14	Other

US Bureau of Labor Statistics. American Time Use Survey. Available online: <https://www.bls.gov/tus/dictionaries.htm>

Documenting datasets:

Example data dictionary: WHO Case Reporting Form

A	B	C	D	E	F
1 Tx section	Variable name	Short label EN	Short label FR	Description	Data type Format
2 Reporting	report_date	Date (DD/MM/YYYY)*: ____/____/____	Date*: ____/____/____	Date of reporting	Date DD/MM/YYYY
3 Reporting	report_country	reporting country	Pays*:	Country/territory/national boundary within which the case currently/usually resides. If the case was tested for COVID-19, indicate the country where the test was performed.	String Free text
4 Reporting	report_test_reason	Why the case was tested for COVID-19		Why the case was tested for COVID-19	Coded Coded variables
5 Reporting	report_test_reason_other	Other reason the case was tested for COVID-19		Any other reason the case was tested for COVID-19	String Free text
6 Patient information	patinfo_ID	country case ID*: _____	Numéro d'identification unique*: _____	Unique case identification number (used in country)	String Free text
7 Patient information	patinfo_ageonset	or Age: _____	ou Age*: _____	Age in units on the date of illness onset.	Numeric ###
8 Patient information	patinfo_ageonsetunit	unit of age	Unité de l'âge	Years or months or days	Coded Coded variables
9 Patient information	patinfo_sex	Sex at birth: Male Female	Sexe à la naissance*: Homme Femme	Biological sex. That is the biological differential characteristics (chromosomes, hormonal profile)	Coded Coded variables
10 Patient information	patinfo_idadmin0	where the case was diagnosed, admin level 0 (country)		Administrative level 0: Country where the case was diagnosed.	String Free text
11 Patient information	patinfo_idadmin1	identified Admin Level 1 (province):	Niveau admin 1*(préfecture):	Administrative level 1: First sub-national boundary (e.g. province, state, territory prefecture)	String Free text
12 Patient information	patinfo_resadmin0	place of residence admin level 0		Administrative level 0: Country within which the case's currently/usually resides.	String Free text
13 Clinical status	Lab_date1	Date of first laboratory confirmation	Date prélevement1	Date of first laboratory confirmation	Date DD/MM/YYYY
14 Clinical status	patcourse_dateonset	Date of onset of first symptoms: ____/____/____	Date de début des premiers symptômes: ____/____/____	Date of first appearance of the signs or symptoms of the illness/disease.	Date DD/MM/YYYY
15 Clinical status	patcourse_asymp	Patient asymptomatic at time of specimen collection	Patient asymptomatique	Is the case asymptomatic?	Coded Coded variables
16 Clinical status	Comcond_present	Does the patient have any underlying conditions?		Does the patient have any underlying conditions?	Coded Coded variables
17 Clinical status	Comcond_preg	Pregnancy	Grossesse	Is the patient pregnant?	Coded Coded variables
18 Clinical status	Comcond_pregt	Trimester of pregnancy	Trimestre de grossesse		Coded Coded variables
19 Clinical status	Comcond_partum	Post-partum (<6 weeks)	Post-partum (<6 semaines)	Is the patient in the post partum period defined as less than 6 weeks after delivery date	Coded Coded variables
20 Clinical status	Comcond_immu	Immunodeficiency including HIV	Immunodéficience, y compris le VIH	Is the patient an acquired immunodeficiency (HIV) or is the patient treated with drugs that suppress the immune system?	Coded Coded variables
21 Clinical status	Comcond_cardi	Cardiovascular disease including hypertension		any cardiovascular disease	Coded Coded variables
22 Clinical status	Comcond_diabetes	Diabetes			Coded Coded variables
23 Clinical status	Comcond_liver	Liver disease		any liver diseases	Coded Coded variables
24 Clinical status	Comcond_renal	Renal disease		any renal diseases	Coded Coded variables
25 Clinical status	Comcond_neuro	Chronic neurological or neuromuscular disease			Coded Coded variables
26 Clinical status	Comcond_malig	Malignancy	Malignité		Coded Coded variables
27 Clinical status	Comcond_lung	Chronic lung disease	Maladie aiguë ou chronique associée:		Coded Coded variables
28 Clinical status	Comcond_other	Other, specify	Autre spécifier	Describe other underlying conditions and comorbidity	String Free text
29 Clinical status	patcourse_admit	admission to hospital?:	Hospitalisation* ?:	Was the case hospitalized, admitted to a hospital or other health facility as an inpatient?	Coded Coded variables
30 Clinical status	patcourse_presHCF	For this episode, date first admitted in hospital: Pour cet épisode, quelle est la date à laquelle le cas a été admis dans une autre structure de santé.	Date the case was first admitted to any health facility.	Date	DD/MM/YYYY

WHO Data Dictionary for Case-Based Reporting form. Available online:
<https://www.who.int/publications/m/item/data-dictionary-for-case-based-reporting-form>

Documenting datasets: Example data dictionary: Game of Thrones Survival

Background

The *Game of Thrones* mortality and survival dataset (hereinafter “the dataset”) was created by Dr Reidar P. Lystad and Dr Benjamin T. Brown and used in the following original research article, which was published in the journal *Injury Epidemiology* in December 2018:

Lystad RP, Brown BT. “Death is certain, the time is not”: mortality and survival in *Game of Thrones*. *Injury Epidemiology* 2018; 5: 44.

The version of the dataset used in the original research article included data from *Game of Thrones* Seasons 1–7 only, whereas the present version of the dataset includes data from *Game of Thrones* Seasons 1–8.

The dataset comprises two separate datasets: (1) a character dataset and (2) an episode dataset. The character dataset contains 359 observations (i.e. characters) and 35 variables, including information about sociodemographics, exposures, and mortality. The episode dataset contains 73 observations (i.e. episodes) and 8 variables, including information about episode running time. Please note that the character dataset only includes “important” characters. As per the original research article:

An important character was defined as any individual who fulfilled each of the following criteria: human; listed in either the opening or closing credits; appeared on screen during current events (i.e. excluding flashbacks); and was not already deceased when first appearing on screen. Additional non-credited characters were included if they interacted with another character in a way that was either crucial to the storyline or character development. Having a speaking role was not an essential requirement because some characters were unable to speak for medical reasons (e.g. acquired brain injury and non-elective glossectomy).

Data Dictionary

Version 2.0

Variables in *Game of Thrones* character dataset

<i>id</i>
Variable number: 01
Description: Unique three-digit character identification number
Representation class: Identifier
Data type: String

<i>name</i>
Variable number: 02
Description: Name of character
Representation class: Identifier
Data type: String

Lystad RP, Brown BT. “Death is certain, the time is not”: mortality and survival in *Game of Thrones*. *Injury Epidemiology* 2018;5:44.

Data Dictionary accessed via Figshare: https://figshare.com/articles/dataset/Game_of_Thrones_mortality_and_survival_dataset/8259680

Documenting datasets:

Example data dictionary template

<p>[optional: add logo here]</p> <p>This data dictionary is intended to document the dataset(s) accessible at [URL]</p> <p>The data dictionary was last updated on [date]</p> <p>If you have any questions, please reach out to [name] at [email or other contact information]</p>				
Data Dictionary - Tables/Datasets				
Table	Description	Resolution(s)	Update frequency	References or other notes
example: measles_cases	example: table with information on country-level data on measles caseload, per year	example: one row per IHR member state per year	example: one-time (not updated)	example: teaching dataset

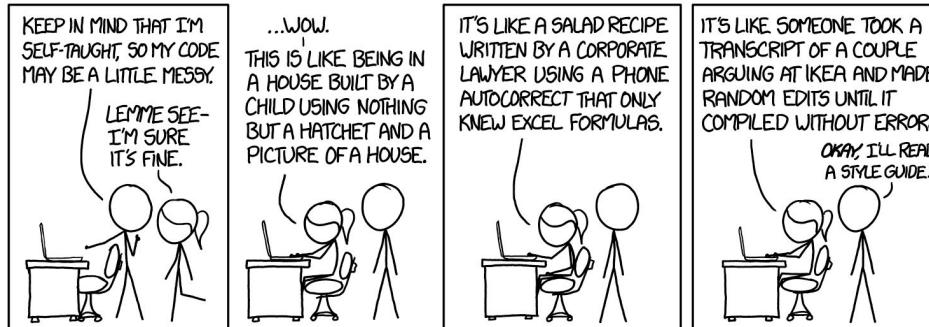
Data Science Basics in R Data Dictionary Template. Steph Eaneff.

Available online: <https://github.com/seaneff/data-science-basics-2024/blob/main/extras/Data%20dictionary%20template.xlsx>

Documenting code:

comments

- As we start writing code, it's helpful to leave "notes" for yourself and others reminding yourself what you did
- Comments can also help break code into sections



Comic by XKCD

Documenting code: comments

```
## #####
## Setup #####
#####

## Load libraries
library(dplyr) ## reshape, reformat, recode data: https://dplyr.tidyverse.org/reference/recode.html
library(ggplot2) ## for plotting: https://ggplot2.tidyverse.org/
library(scales) ## for commas on axes of plots
library(treemap) ## for treemap visual

## #####
## Read in data #####
#####

line_items <- read_excel("calculator-tool/jee3_costing_worksheet.xlsx",
                         sheet = "Line items (JEE 3)")

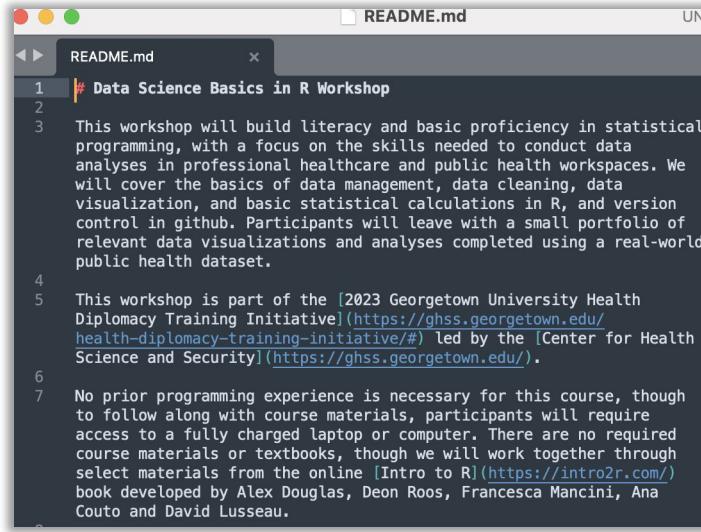
unit_costs <- read_excel("calculator-tool/jee3_costing_worksheet.xlsx",
                         sheet = "Unit costs")

multipliers <- read_excel("calculator-tool/jee3_costing_worksheet.xlsx",
                         sheet = "Multipliers")
```

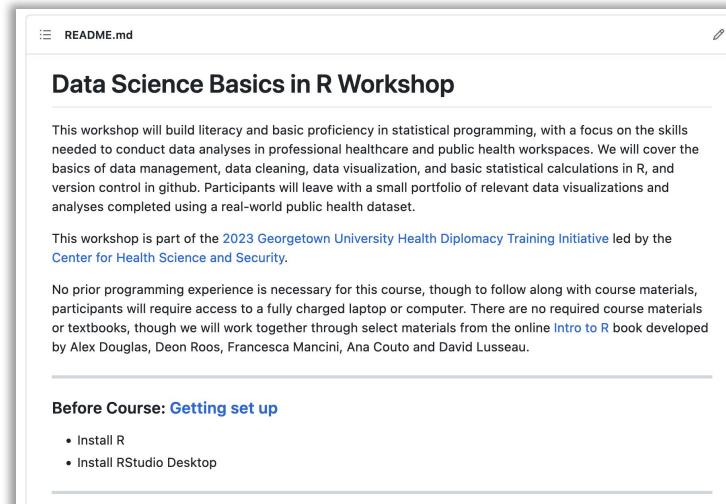
Documenting code:

README files

- README files summarize the contents of a specific folder
- Common in github, rendered using Markdown



```
README.md
1 # Data Science Basics in R Workshop
2
3 This workshop will build literacy and basic proficiency in statistical
4 programming, with a focus on the skills needed to conduct data
5 analyses in professional healthcare and public health workspaces. We
6 will cover the basics of data management, data cleaning, data
7 visualization, and basic statistical calculations in R, and version
control in github. Participants will leave with a small portfolio of
relevant data visualizations and analyses completed using a real-world
public health dataset.
8
9 This workshop is part of the [2023 Georgetown University Health
10 Diplomacy Training Initiative](https://ghss.georgetown.edu/health-diplomacy-training-initiative/) led by the [Center for Health
11 Science and Security](https://ghss.georgetown.edu/).
12
13 No prior programming experience is necessary for this course, though
14 to follow along with course materials, participants will require
15 access to a fully charged laptop or computer. There are no required
16 course materials or textbooks, though we will work together through
17 select materials from the online [Intro to R](https://intro2r.com/)
18 book developed by Alex Douglas, Deon Roos, Francesca Mancini, Ana
19 Couto and David Lusseau.
```



Data Science Basics in R Workshop

This workshop will build literacy and basic proficiency in statistical programming, with a focus on the skills needed to conduct data analyses in professional healthcare and public health workspaces. We will cover the basics of data management, data cleaning, data visualization, and basic statistical calculations in R, and version control in github. Participants will leave with a small portfolio of relevant data visualizations and analyses completed using a real-world public health dataset.

This workshop is part of the [2023 Georgetown University Health Diplomacy Training Initiative](#) led by the [Center for Health Science and Security](#).

No prior programming experience is necessary for this course, though to follow along with course materials, participants will require access to a fully charged laptop or computer. There are no required course materials or textbooks, though we will work together through select materials from the online [Intro to R](#) book developed by Alex Douglas, Deon Roos, Francesca Mancini, Ana Couto and David Lusseau.

Before Course: Getting set up

- Install R
- Install RStudio Desktop

Reproducible results

We won't cover it in detail in this workshop, but R is a great way to ensure that your code, and your results, are well documented and reproducible

- Use a single version of a **clean dataset**
- Document and version your **data cleaning decisions**
- Store your **code, results, & documentation** in one spot

Reproducible results

Code versioning in github

The screenshot shows a GitHub user profile for `seaneff`. The main navigation bar includes `Pull requests`, `Issues`, `Codespaces`, `Marketplace`, and `Explore`. A modal window titled "Join GitHub Global Campus!" is open, describing the program's purpose of providing practical industry knowledge through access to tools, events, learning resources, and a student community. It features several cards: "Follow your Expert", "Breaking into tech: internship edition with Helen Huang", "Level up your code with TwilioQuest", "Learning by teaching for your community - Cassidy Williams", "Popular offers you have not claimed", "Curated Experiences with popular offers", "Virtual event kit", "Connect your local Expert", "View projects at our gallery", "Learn more about an event", "Watch a Campus TV episode", "Web meet 202", and "Web Meet". A large green button at the bottom right of the modal says "Join Global Campus". Below the modal, the "For you" feed shows a message: "Welcome to the new feed! We're updating the cards and ranking all the time, so check back regularly. At first, you might need to follow some people or star some repositories to get started." The sidebar on the left lists "Top Repositories" and "Recent activity". The main content area also includes sections for "Your teams" and "Explore repositories".

Latest changes

- 22 minutes ago GitHub Actions: All Actions will run on Node16 instead of Node12 by default
- 1 hour ago GitHub Actions: You can now disable repo level self-hosted runners in an Enterprise and Organization
- 20 hours ago Dependabot version updates now supports pnpm
- Yesterday GitHub Advanced Security trial now available on GitHub Enterprise Cloud
View changelog →

Explore repositories

zulip / zulip
Zulip server and web application. Open-source team chat that helps teams stay productive and focused.
18k Python

Ebazhanov / linkedin-skill-assessments-quizzes
Full reference of LinkedIn answers 2023 for skill assessments (aws-lambda, rest-api, javascript, react, git, html, jquery, mongodb, java, Go, python, machine-learning, power-point) linkedin excel ...
25.2k Python

Course datasets

Measles

MEASLES AND RUBELLA STRATEGIC FRAMEWORK **2021–2030**



WHO. Measles and rubella strategic framework: 2021–2030.

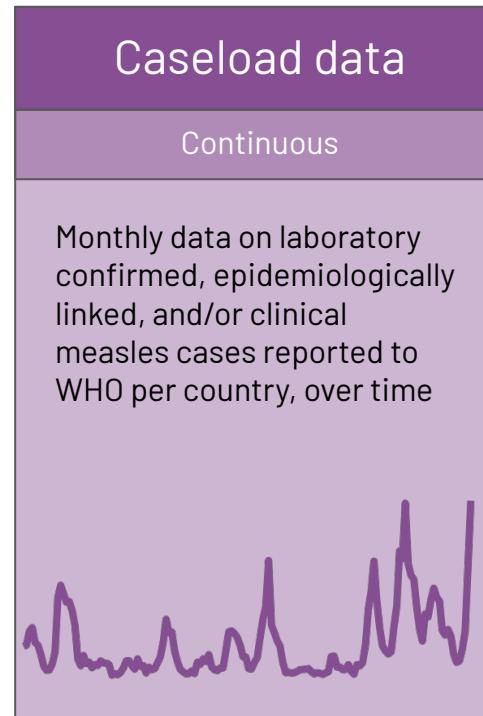
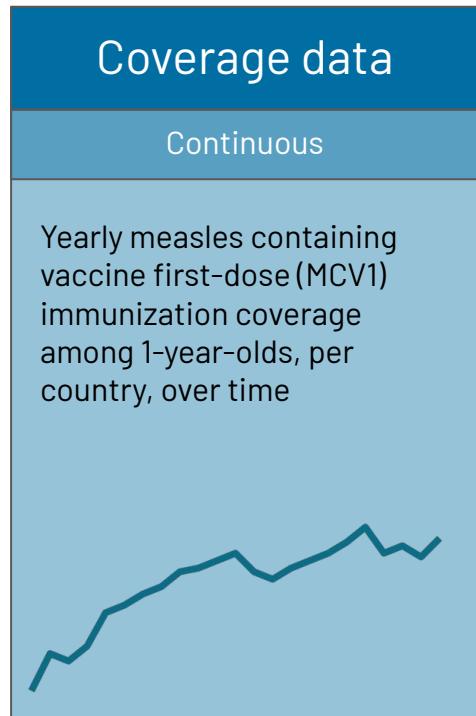
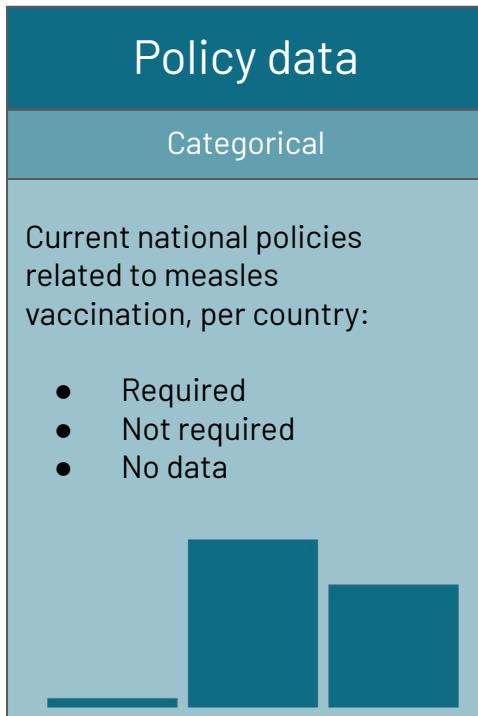
<https://www.who.int/publications/i/item/measles-and-rubella-strategic-framework-2021-2030>

Course datasets: Vaccination

For this course, we'll primarily use global-scale datasets related to vaccine-preventable diseases.

- **policy data:** national policies related to vaccination
- **coverage data:** % of relevant populations vaccinated
- **caseload data:** information on the number of cases per year

Course datasets: Vaccination



Course datasets: Vaccination

seaneff / data-science-basics-2024

Type to search

Code Issues Pull requests Actions Projects Security Insights Settings

Files

main Go to file

beforeclass-download-R course-datasets

- Data dictionary.xlsx
- countries.tsv
- measles_cases.tsv
- measles_vaccine_policy.tsv**

day1 day2 day3 extras .gitignore README.md Syllabus Data Science Basics 202...

measles_vaccine_policy.tsv

seaneff initial policy data -- measles only 3f236bb · now History

Preview Code Blame 198 lines (198 loc) · 6.58 KB Code 55% faster with GitHub Copilot

Raw

Search this file

	country_name	iso_code	measles_vaccine_policy
1	Afghanistan	AFG	not required
2	Albania	ALB	required
3	Algeria	DZA	not required
4	Andorra	AND	required
5	Angola	AGO	not required
6	Antigua and Barbuda	ATG	not required
7	Argentina	ARG	required
8	Armenia	ARM	not required
9	Australia	AUS	required
10	Austria	AUT	not required
11	Azerbaijan	AZE	not required

Course datasets: Vaccination

seaneff / data-science-basics-2024

Code Issues Pull requests Actions Projects Security Insights Settings

Files

main Go to file

beforeclass-download-R course-datasets

- Data dictionary.xlsx
- countries.tsv
- measles_cases.tsv
- measles_vaccine_policy.tsv**

day1 day2 day3 extras .gitignore README.md Syllabus Data Science Basics 202...

data-science-basics-2024 / course-datasets / measles_vaccine_policy.tsv

seaneff initial policy data -- measles only 3f236bb · now History

Preview Code Blame 198 lines (198 loc) · 6.58 KB Code 55% faster with GitHub Copilot

Search this file

	country_name	iso_code	measles_vaccine_policy
1	Afghanistan	AFG	not required
2	Albania	ALB	required
3	Algeria	DZA	not required
4	Andorra	AND	required
5	Angola	AGO	not required
6	Antigua and Barbuda	ATG	not required
7	Argentina	ARG	required
8	Armenia	ARM	not required
9	Australia	AUS	required
10	Austria	AUT	not required
11	Azerbaijan	AZE	not required

Raw   

Download the file

Your turn!

Explore the data dictionary for this course at:

bit.ly/data-dictionary-gt2024



This data dictionary is intended to document the dataset(s) accessible at <https://github.com/seaneff/data-science-basics-2024>.
The data dictionary was last updated on May 24, 2024.
If you have any questions, please reach out to Steph Eaneff at sde31@georgetown.edu.

Data Dictionary - Tables/Datasets				
Table	Description	Resolution(s)	Update frequency	References or other notes
countries.tsv	Country-level information, including measles policy and vaccine coverage information, for World Health Assembly member states. Table also includes relevant information about the country including population size, region, and income designation	one row per World Health Assembly member state	one-time (not updated)	World Health Organization. (10 February, 2023). Countries overview. https://www.who.int/countries/ World Bank. (29 March, 2024). World Bank Country and Lending Groups. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups World Bank. (29 March, 2024). Population estimates and projections, 2024. https://databank.worldbank.org/source/population-estimates-and-projections# Manually added missing ISO codes for Republic of Kiribati and Tonga
measles_cases.tsv	Country-level information, per month, on laboratory confirmed, epidemiologically linked, and/or clinical measles cases reported to the World Health Organization	one row per World Health Assembly member state per month	one-time (not updated)	World Health Organization. (22 March, 2024). Distribution of Measles Cases by Country and Month. https://www.who.int/teams/immunization-vaccines-and-biologicals/immunization-analysis-and-insights/surveillance/monitoring/provisional-monthly-measles-and-rubella-data

Loading and cleaning datasets in R

Raw data files

We're going to talk about structured, flat files today, but there are many types of data that R can work with!

- **Excel:** proprietary file format (Microsoft), but common (.xls files)
- **CSV:** comma-separated value files
- **TSV:** tab-separated value files
- **RData:** data saved from within your R environment
- **SPSS or SAS:** data from other statistical programming tools
- **Shapefiles:** geospatial or mapping data

Functions to load data

- **Excel:** `read_excel("countries.xls")`
- **CSV:** `read.csv("countries.csv")`
- **TSV:** `read.delim("countries.tsv")`
- **RData:** `load("countries.RData")`
- **SPSS or SAS:** use the 'foreign' package
- **Shapefiles:** use the 'sf' or other geospatial data packages

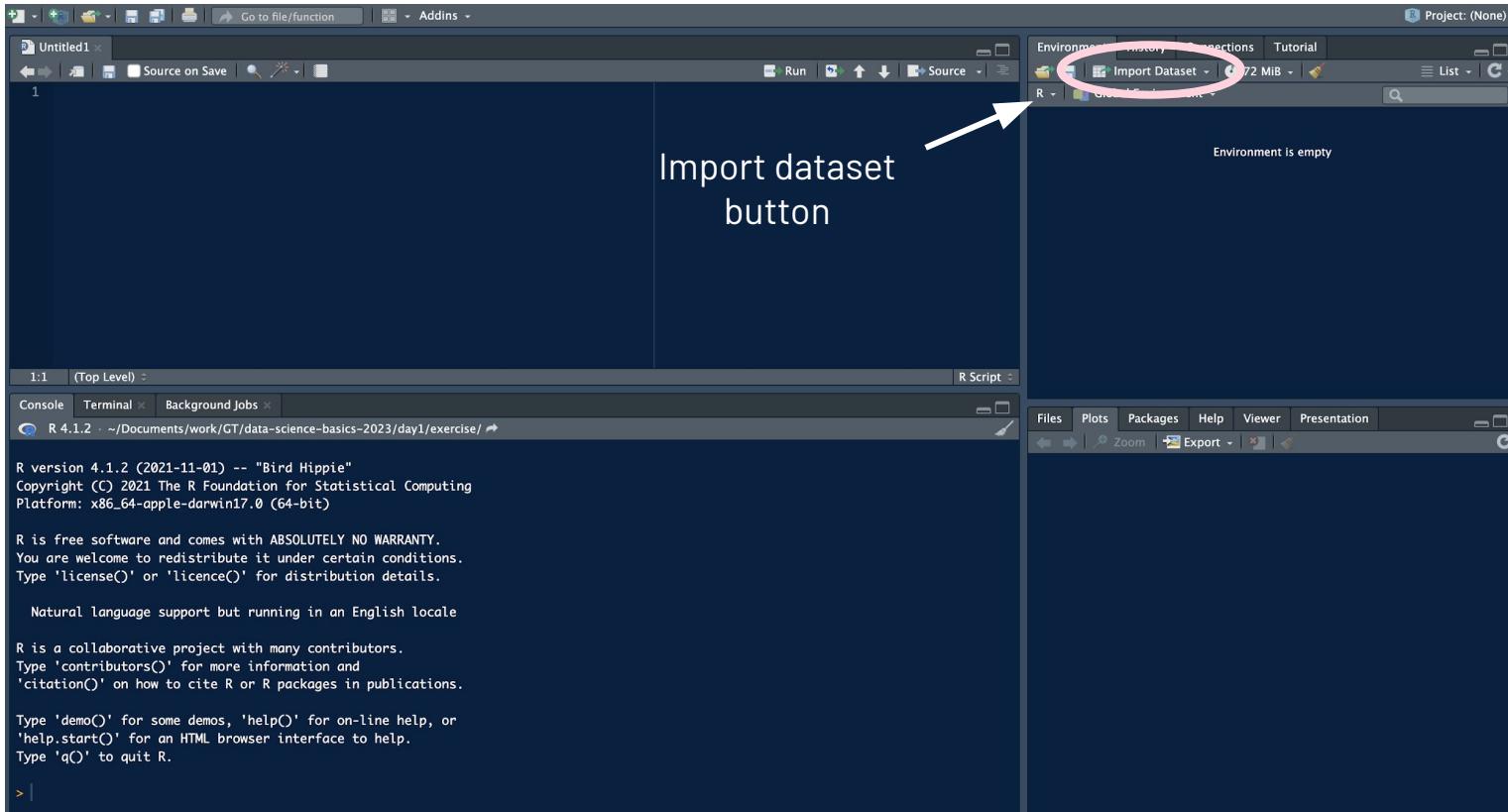
A note on Excel files and libraries in R

To use the function `read_excel()`, we need to have installed the package `readxl` into R. You (usually) only need to install it once, and then you will need to re-load it every new R session

- **Installing:** `install.packages("readxl")`
- **Loading:** `library("readxl")`

If you are working with RStudio and using the workflow in the next slides, you will automatically be prompted to do these things if they are required

Loading data with RStudio



Loading data with RStudio

Several default options of data types, including Excel.
Here, I'll select the first option, using base R

R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

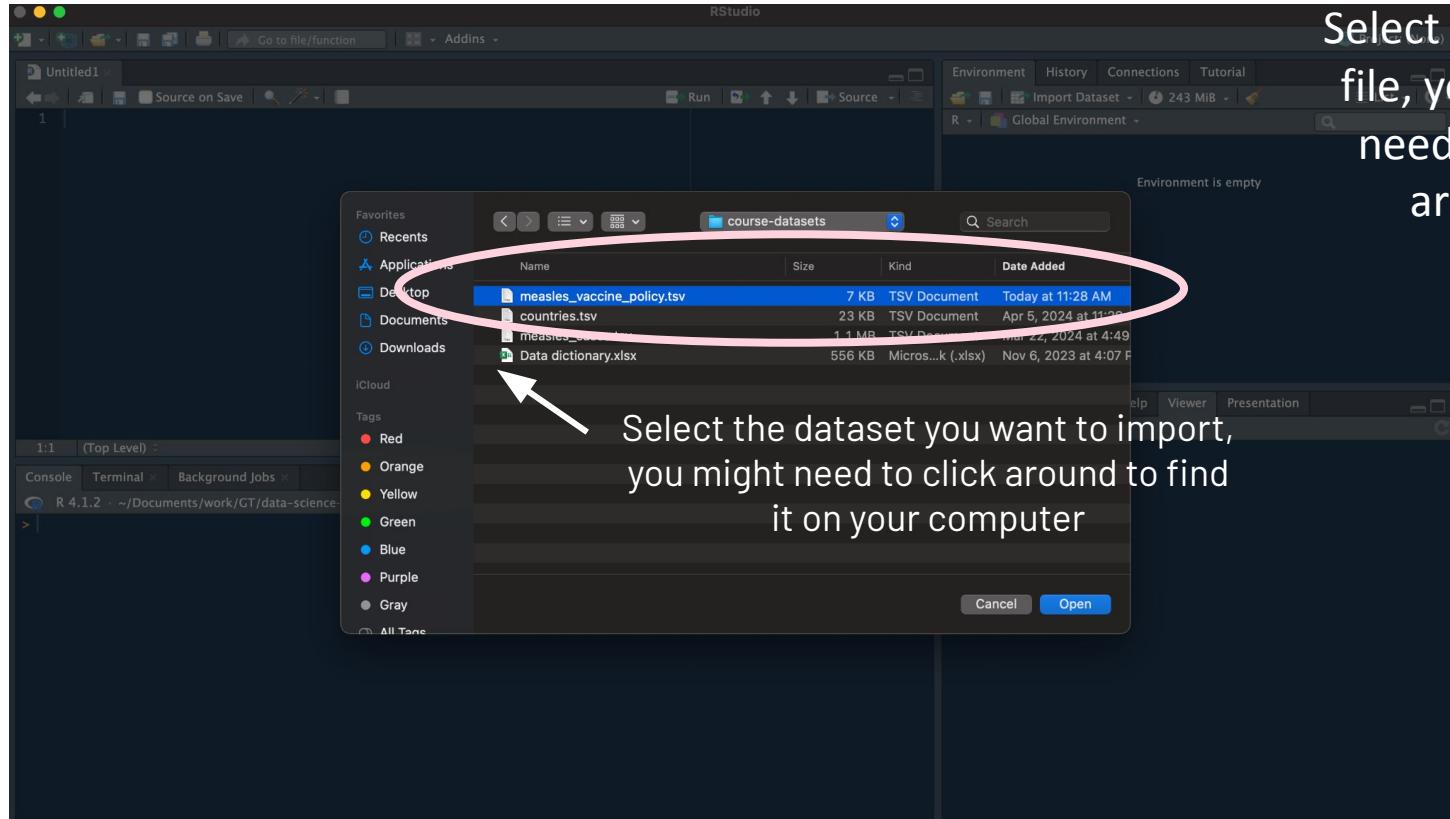
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

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

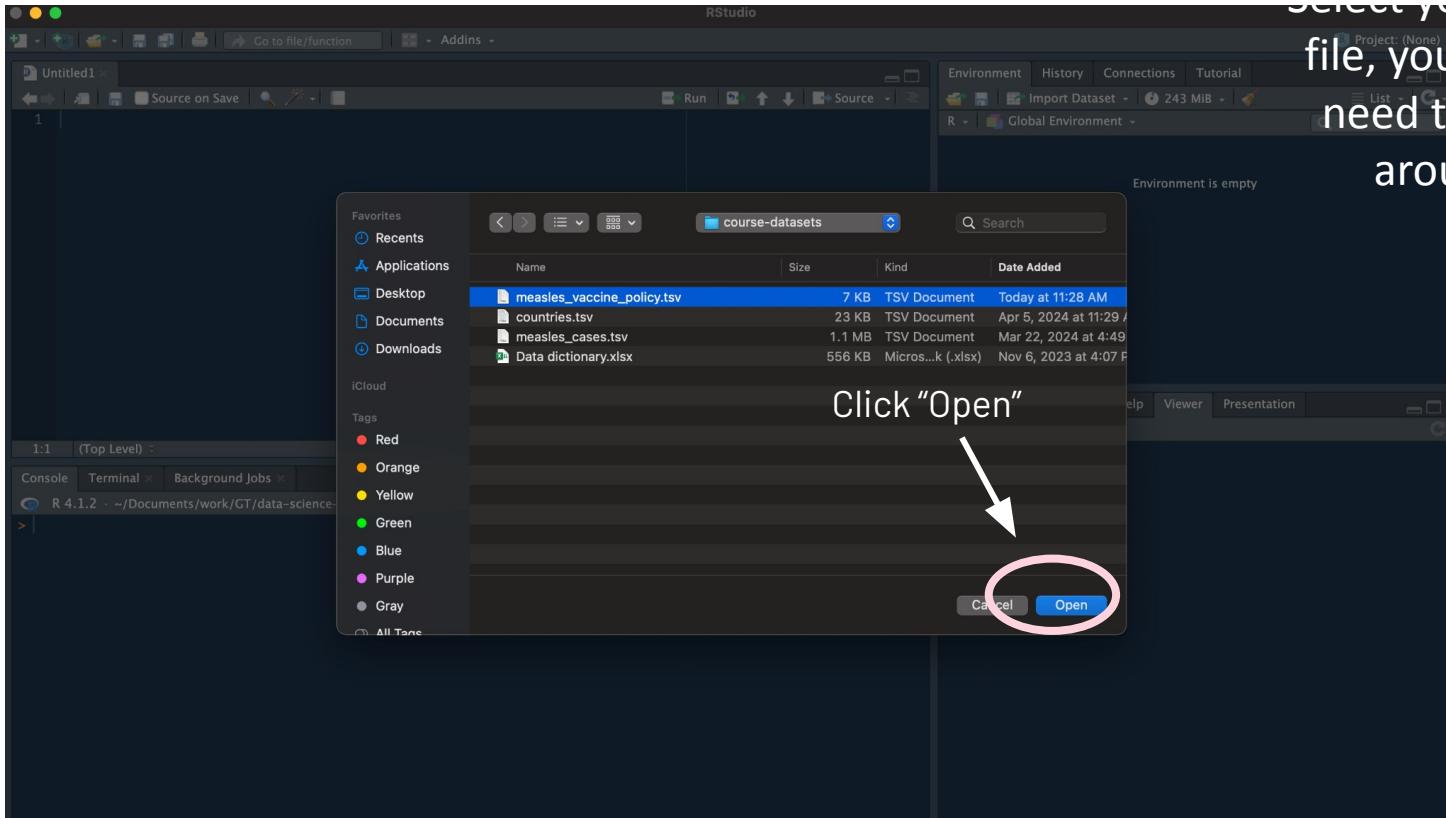
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Loading data with RStudio

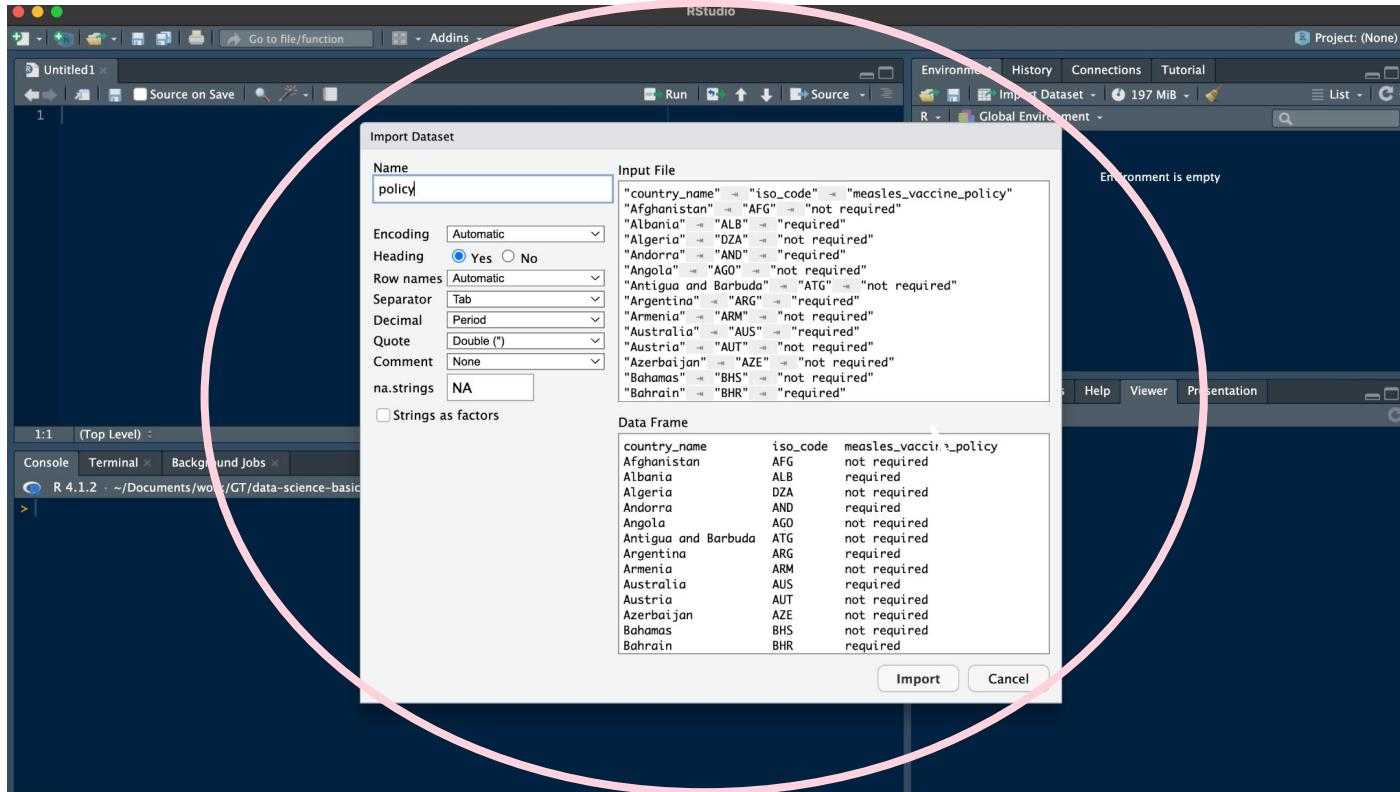


Select your file, you
need to click around

Loading data with RStudio



Loading data with RStudio



Loading data with RStudio

The screenshot shows the RStudio interface with the 'Import Dataset' dialog box open. The 'Name' field is set to 'policy'. The 'Input File' section displays a preview of the data, which includes columns for 'country_name', 'iso_code', and 'measles_vaccine_policy'. The 'Data Frame' section shows the same data in a tabular format. A red circle highlights the 'Import' button at the bottom right of the dialog box. The RStudio environment shows an empty 'Environment' pane.

Import Dataset

Name: policy

Encoding: Automatic

Heading: Yes

Row names: Automatic

Separator: Tab

Decimal: Period

Quote: Double ("")

Comment: None

na.strings: NA

Strings as factors

Input File

```
"country_name" -> "iso_code" -> "measles_vaccine_policy"
"Afghanistan" -> "AFG" -> "not required"
"Albania" -> "ALB" -> "required"
"Algeria" -> "DZA" -> "not required"
"Andorra" -> "AND" -> "required"
"Angola" -> "AGO" -> "not required"
"Antigua and Barbuda" -> "ATG" -> "not required"
"Argentina" -> "ARG" -> "required"
"Armenia" -> "ARM" -> "not required"
"Australia" -> "AUS" -> "required"
"Austria" -> "AUT" -> "not required"
"Azerbaijan" -> "AZE" -> "not required"
"Bahamas" -> "BHS" -> "not required"
"Bahrain" -> "BHR" -> "required"
```

Data Frame

country_name	iso_code	measles_vaccine_policy
Afghanistan	AFG	not required
Albania	ALB	required
Algeria	DZA	not required
Andorra	AND	required
Angola	AGO	not required
Antigua and Barbuda	ATG	not required
Argentina	ARG	required
Armenia	ARM	not required
Australia	AUS	required
Austria	AUT	not required
Azerbaijan	AZE	not required
Bahamas	BHS	not required
Bahrain	BHR	required

Import Cancel

Environment is empty

Click "Import"

Loading data with RStudio

A screenshot of the RStudio interface. On the left, the Data View shows a data frame with three columns: 'country_name', 'iso_code', and 'measles_vaccine_policy'. The data includes rows for Afghanistan, Albania, Algeria, Andorra, Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, and Bahrain. A pink circle highlights this view. On the right, the Global Environment pane shows the 'policy' object with 197 observations and 3 variables. Below it, a text box says 'There's the data!'. At the bottom, the Console pane displays the R code used to load the data:

```
R 4.1.2 · ~/Documents/work/GT/data-science-basics-2024/ ↵
> policy <- read.delim("~/Documents/work/GT/data-science-basics-2024/course-datasets/measles_vaccine_policy.tsv")
> View(policy)
> |
```

Loading data with RStudio

The screenshot shows the RStudio interface with the following components:

- Data View:** Displays a data frame titled "policy" with columns: "country_name", "iso_code", and "measles_vaccine_policy". The data includes rows for Afghanistan, Albania, Algeria, Andorra, Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, and Bahamas.
- Environment View:** Shows the "policy" data frame listed under "Global Environment". It indicates 197 observations and 3 variables.
- Console View:** Shows the R code used to load the data:

```
R 4.1.2 · ~/Documents/work/GT/data-science-basics-2024/ ↵
> policy <- read.delim("~/Documents/work/GT/data-science-basics-2024/course-datasets/measles_vaccine_policy.csv")
```

A red oval highlights the first line of code, and a white arrow points from it to the explanatory text on the right.
- Text Overlay:** A blue box contains the text: "R Studio wrote and logged some code for us here, which loaded the data"

Loading data with RStudio

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Shows the title "RStudio" and a "Project: (None)" dropdown.
- Data View:** A data grid titled "policy" showing 12 rows of data. The columns are "country_name", "iso_code", and "measles_vaccine_policy". The data includes entries for Afghanistan, Albania, Algeria, Andorra, Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, and Bahamas.
- Environment View:** Shows the dataset "policy" with 197 observations and 3 variables.
- Console View:** Displays R code and its output. The code reads a CSV file into a data frame and then calls the "View" command to print the data frame to the console. The output shows the first 12 rows of the "policy" data frame.
- Message:** An annotation with a red circle and arrow points to the "View(policy)" line in the console, with the text "The 'View' command printed our data above".

country_name	iso_code	measles_vaccine_policy
Afghanistan	AFG	not required
Albania	ALB	required
Algeria	DZA	not required
Andorra	AND	required
Angola	AGO	not required
Antigua and Barbuda	ATG	not required
Argentina	ARG	required
Armenia	ARM	not required
Australia	AUS	required
Austria	AUT	not required
Azerbaijan	AZE	not required
Bahamas	BHS	not required

```
R 4.1.2 · ~/Documents/work/GT/data-science-basics-2024/ ↵
> policy <- read_delim("~/Documents/work/GT/data-science-basics-2024/course-datasets/measles_vaccine_policy.tsv")
> View(policy)
```

The "View" command printed our data above

Loading data with RStudio

The screenshot shows the RStudio interface with the following components:

- Data Browser:** On the left, a table titled "policy" displays data for 12 countries. The columns are "country_name", "iso_code", and "measles_vaccine_policy". The data shows various vaccination requirements.
- Data View:** On the right, the "Data" view shows the "policy" dataset with 197 observations and 3 variables. A pink circle highlights this area.
- Console:** At the bottom left, the console shows R code used to load the data:

```
R 4.1.2 · ~/Documents/work/GT/data-science-basics-2024/ ↵
> policy <- read.delim("~/Documents/work/GT/data-science-basics-2024/course-datasets/measles_vaccine_policy.tsv")
> View(policy)
> |
```
- Global Environment:** The top right pane shows the global environment with the "policy" dataset loaded.

We can see a list of all of our loaded datasets here

country_name	iso_code	measles_vaccine_policy
Afghanistan	AFG	not required
Albania	ALB	required
Algeria	DZA	not required
Andorra	AND	required
Angola	AGO	not required
Antigua and Barbuda	ATG	not required
Argentina	ARG	required
Armenia	ARM	not required
Australia	AUS	required
Austria	AUT	not required
Azerbaijan	AZE	not required
Bahamas	BHS	not required

Loading data directly from a website

- Go to bit.ly/measles-policy-data
- Copy URL
- Read data into R



```
policy_data <-  
read.delim("https://github.com/seaneff/data-science-basics-2024/blob/  
main/course-datasets/measles_vaccine_policy.tsv")
```

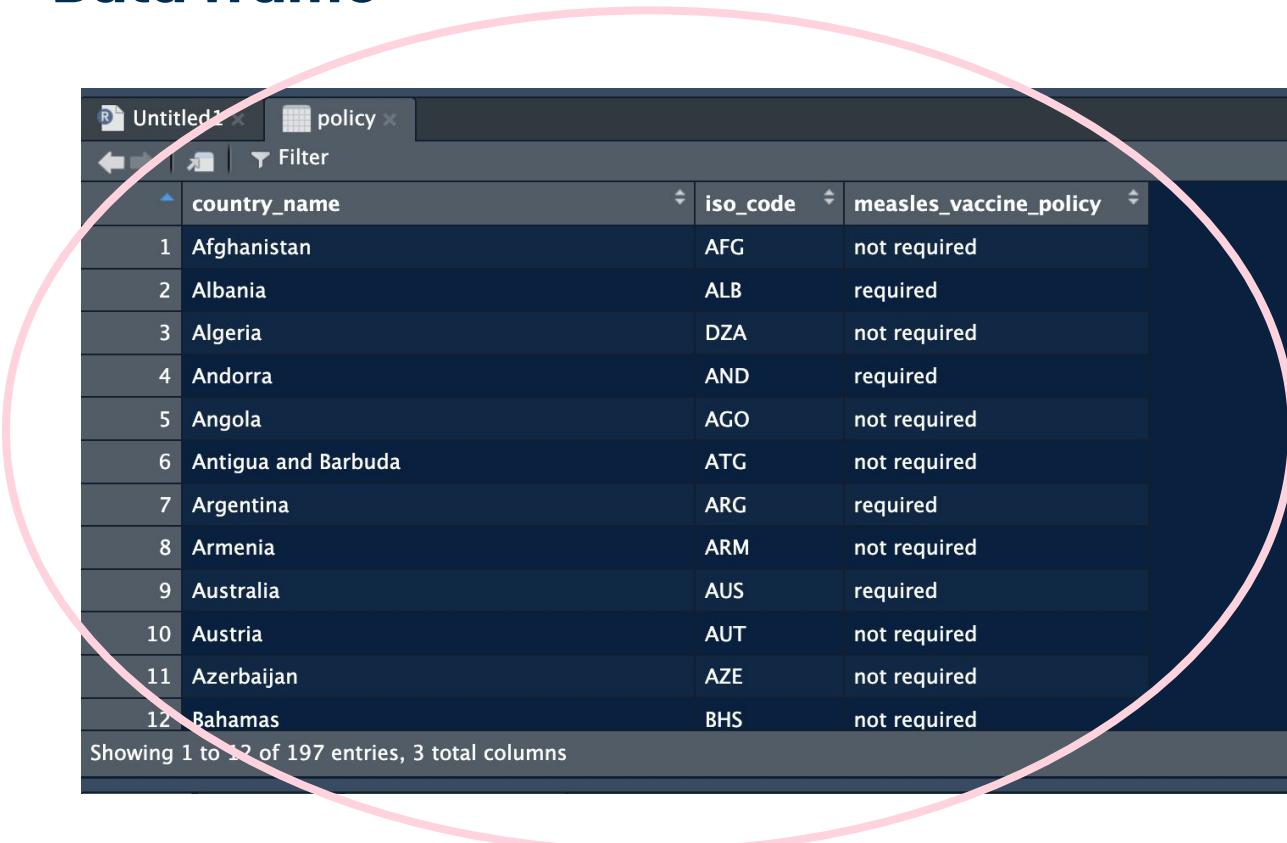
Data structures in R

Data Structures in R

Yesterday we discussed vectors, today, we're going to talk about other types of data structures in R. There are a few different types, but the ones we'll talk about this week are:

- **vectors**: group of data elements of a certain type (1D)
- **matrix**: group of data elements of a certain type (2D)
- **data frame**: group of data elements with different types

Data frame



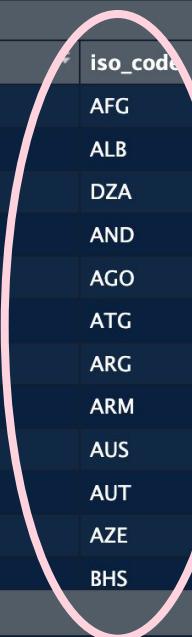
The screenshot shows the RStudio environment with a data frame titled "policy". The window title bar says "Untitled1" and "policy". The main area displays a table with three columns: "country_name", "iso_code", and "measles_vaccine_policy". The data consists of 12 rows, indexed from 1 to 12, showing the measles vaccination policy for various countries. A pink circle highlights the entire data frame window.

	country_name	iso_code	measles_vaccine_policy
1	Afghanistan	AFG	not required
2	Albania	ALB	required
3	Algeria	DZA	not required
4	Andorra	AND	required
5	Angola	AGO	not required
6	Antigua and Barbuda	ATG	not required
7	Argentina	ARG	required
8	Armenia	ARM	not required
9	Australia	AUS	required
10	Austria	AUT	not required
11	Azerbaijan	AZE	not required
12	Bahamas	BHS	not required

Showing 1 to 12 of 197 entries, 3 total columns

Vector

R Untitled1 × policy × Filter

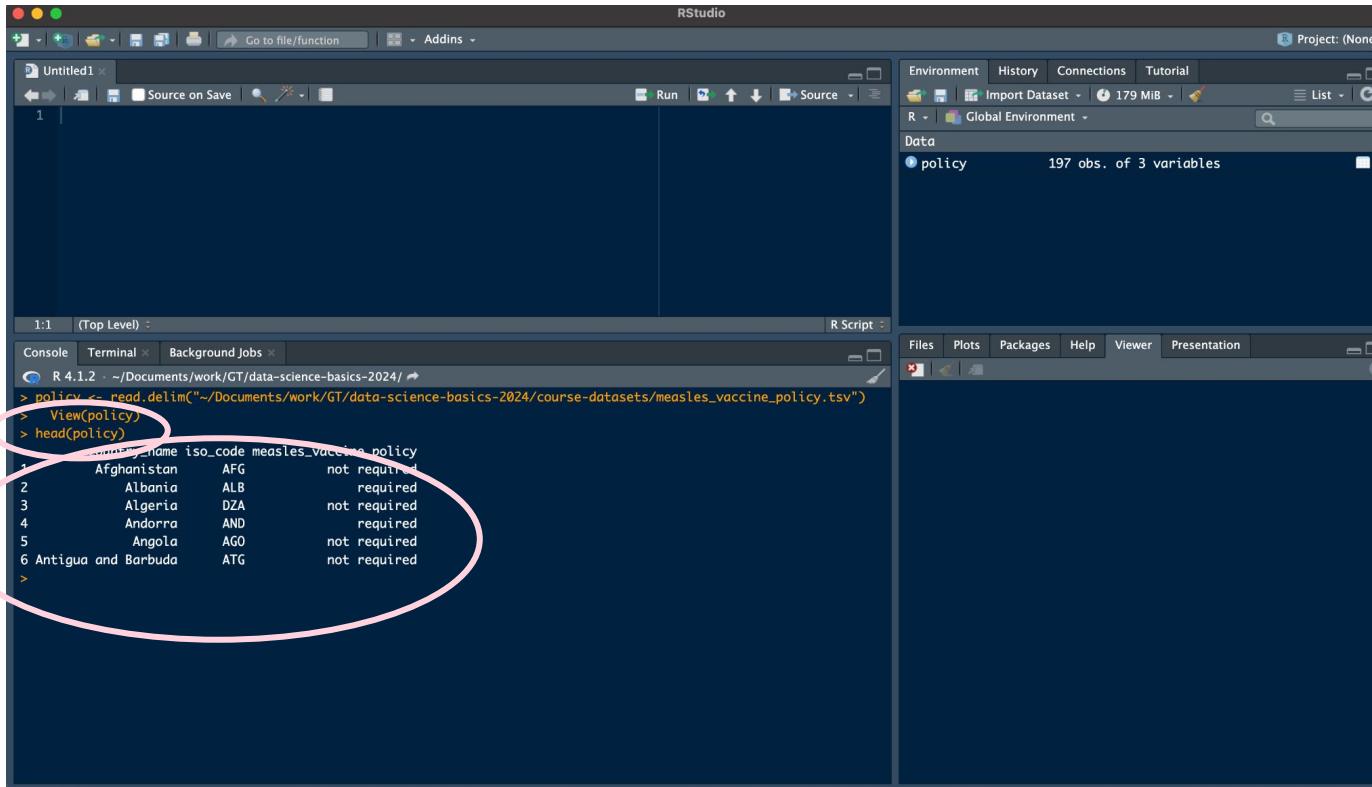


	country_name	iso_code	measles_vaccine_policy
1	Afghanistan	AFG	not required
2	Albania	ALB	required
3	Algeria	DZA	not required
4	Andorra	AND	required
5	Angola	AGO	not required
6	Antigua and Barbuda	ATG	not required
7	Argentina	ARG	required
8	Armenia	ARM	not required
9	Australia	AUS	required
10	Austria	AUT	not required
11	Azerbaijan	AZE	not required
12	Bahamas	BHS	not required

Showing 1 to 12 of 197 entries, 3 total columns

Exploring your dataset

`head()` function



A screenshot of the RStudio interface demonstrating the use of the `head()` function. The left pane shows the R Script console with the following code and its output:

```
R 4.1.2 · ~/Documents/work/GT/data-science-basics-2024/ ↵
> policy <- read.delim("~/Documents/work/GT/data-science-basics-2024/course-datasets/measles_vaccine_policy.tsv")
> View(policy)
> head(policy)
```

The output displays the first six rows of a dataset named `policy`, which contains columns for country name, ISO code, and measles vaccine policy requirement. A pink oval highlights the first six rows of the output.

	country_name	iso_code	measles_vaccine_policy
1	Afghanistan	AFG	not required
2	Albania	ALB	required
3	Algeria	DZA	not required
4	Andorra	AND	required
5	Angola	AGO	not required
6	Antigua and Barbuda	ATG	not required

Exploring your dataset

View() function

The screenshot shows the RStudio interface with the following components:

- Environment pane:** Shows a data frame named "policy" with 197 observations and 3 variables.
- Data pane:** Displays the same information about the "policy" data frame.
- Console pane:** Shows the R code used to load the data and the resulting output. The output is circled in red, highlighting the first few rows of the data frame.
- Code area:** Shows the R code being typed, including the highlighted line "View(policy)".

The data frame "policy" contains the following data:

	country_name	iso_code	measles_vaccine_policy
1	Afghanistan	AFG	not required
2	Albania	ALB	required
3	Algeria	DZA	not required
4	Andorra	AND	required
5	Angola	AGO	not required
6	Antigua and Barbuda	ATG	not required
7	Argentina	ARG	required

Exploring your dataset

Dimensions and field names

```
dim(policy)
```

```
## [1] 197    3
```

```
names(policy)
```

```
## [1] "country_name"           "iso_code"                 "measles_vaccine_policy"
```

Exploring your dataset

Use `anyNA()` to check for missing values

```
anyNA(policy)
```

```
## [1] FALSE
```

```
anyNA(policy$country_name)
```

```
## [1] FALSE
```

Accessing subsets of data

selecting specific columns

```
policy$country_name
```

```
## [1] "Afghanistan"  
## [2] "Albania"  
## [3] "Algeria"  
## [4] "Andorra"  
## [5] "Angola"  
## [6] "Antigua and Barbuda"  
## [7] "Argentina"  
## [8] "Armenia"  
## [9] "Australia"  
## [10] "Austria"  
## [11] "Azerbaijan"  
## [12] "Bahamas"  
## [13] "Bahrain"  
## [14] "Bangladesh"  
## [15] "Barbados"  
## [16] "Belarus"  
## [17] "Belgium"  
## [18] "Belize"  
## [19] "Benin"  
## [20] "Bhutan"  
## [21] "Bolivia (Plurinational State of)"  
## [22] "Bosnia and Herzegovina"  
## [23] "Botswana"
```

Your turn!

Select a different column from the policy dataset

Accessing subsets of data

selecting specific rows

```
policy[which(policy$country_name == "Angola"),]
```

```
##   country_name iso_code measles_vaccine_policy
## 5      Angola      AGO        not required
```

```
library(dplyr)
```

```
policy %>%
  filter(country_name == "Angola")
```

```
##   country_name iso_code measles_vaccine_policy
## 1      Angola      AGO        not required
```

Your turn!

Select a different row from the policy dataset

Accessing subsets of data

counts of categories with table()

```
table(policy$measles_vaccine_policy)
```

```
##  
##      no data not required      required  
##            2              113            82
```

Let's switch to R

Your turn!

Go to the course github, and copy-paste the code from the
day2/ folder into the top left corner of RStudio

Live demo

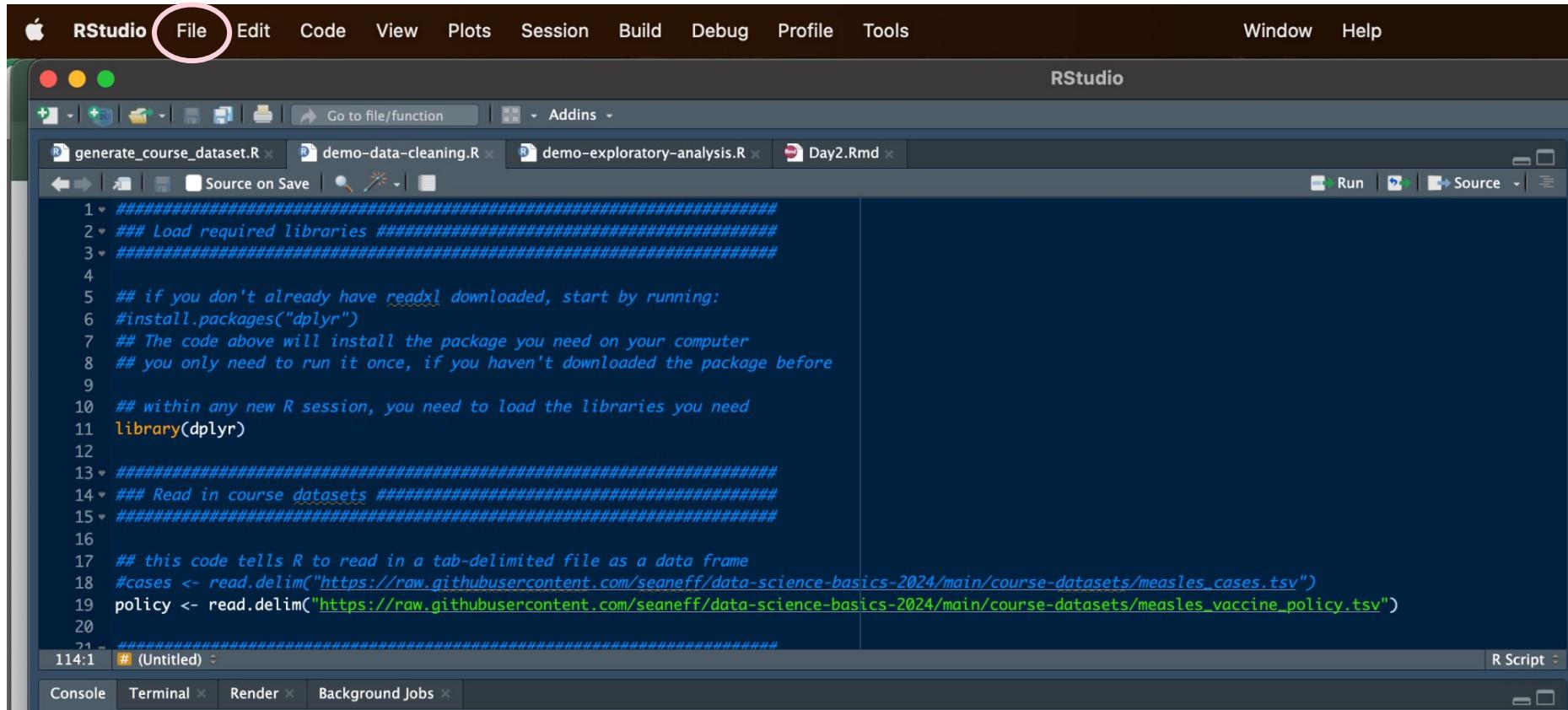
Your turn!

Pick a dataset (or a set of datasets) you'd like to work with today and tomorrow for your worked examples.

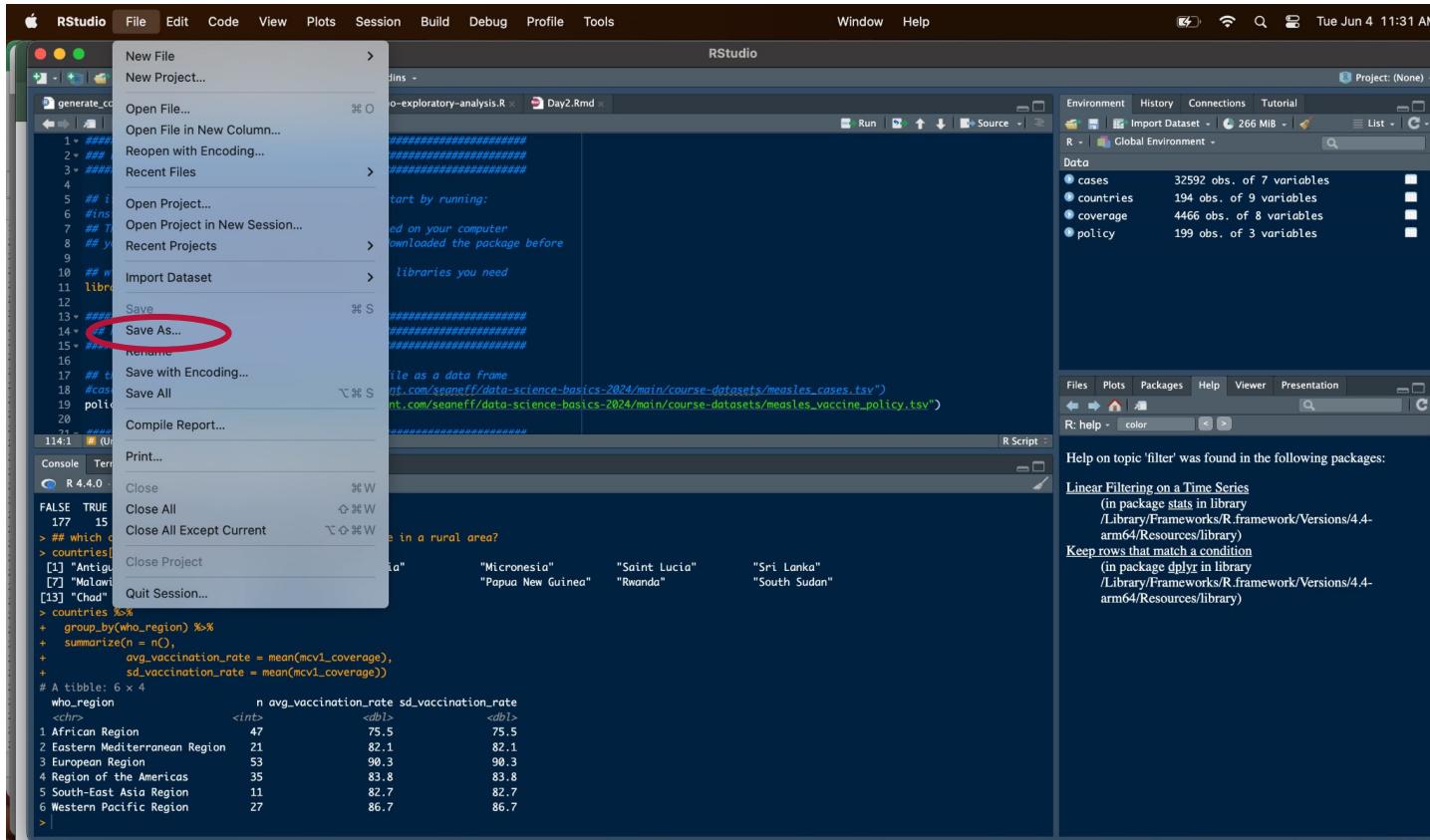
In the next 15 minutes, please draft 5-10 research or operational questions that you could explore using the dataset you selected. You can work alone or with a small group.

Add your code from today to github

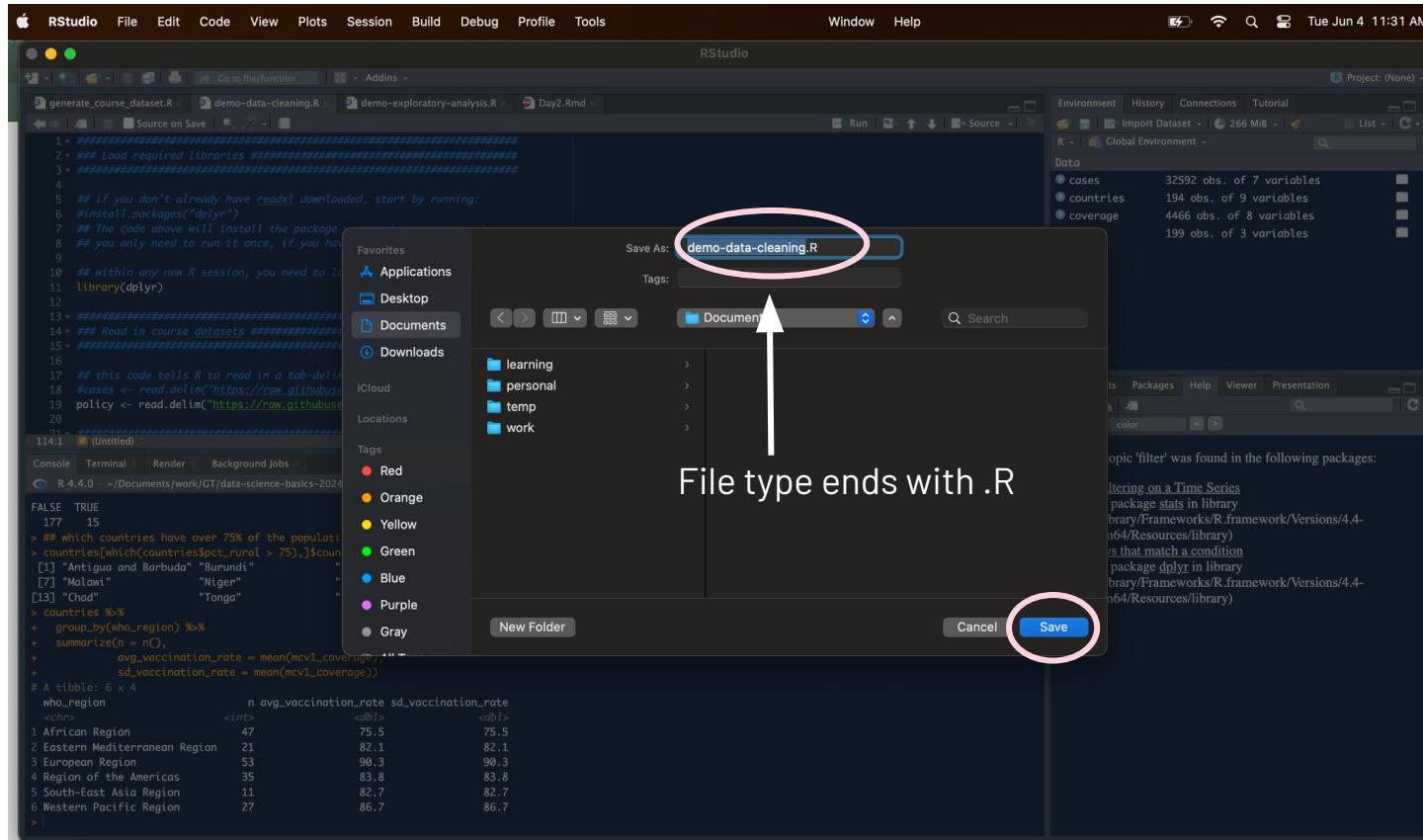
Save your R code



Save your R code



Save your R code



Upload your code to GitHub

The screenshot shows a GitHub repository page for 'data-science-basics-2024'. The repository is public and has 2 branches and 0 tags. The main branch has 48 commits. The repository is described as 'In progress materials for the data science basics course in 2024'. It has 0 forks and 1 star.

Code

Issues

Pull requests

Actions

Projects

Security

Insights

Settings

Unpin

Unwatch 1

Fork 0

Starred 1

main

2 Branches 0 Tags

Go to file

Add file

Code

seaneff update gitignore to add day 2 materials

8c7a75d · 4 minutes ago 48 Commits

course-datasets initial commit of energy dataset 19 hours ago

day1 fix day 1 slides 19 hours ago

day2 update gitignore to add day 2 materials 4 minutes ago

download-R reorganize and show only active days of course yesterday

About

In progress materials for the data science basics course in 2024

Readme

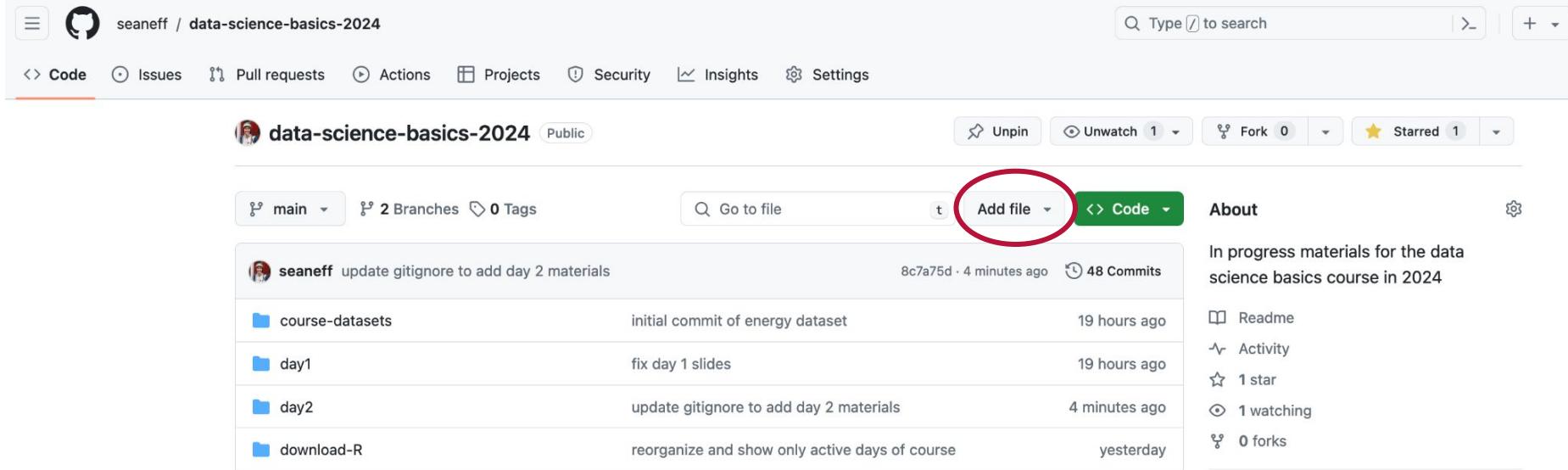
Activity

1 star

1 watching

0 forks

Upload your code to GitHub



The screenshot shows a GitHub repository page for 'seaneff / data-science-basics-2024'. The repository is public and has 2 branches and 0 tags. The main branch is selected. The repository has 48 commits. A commit by 'seaneff' from 4 minutes ago updated .gitignore to add day 2 materials. Other commits include initial commit of energy dataset, fixing day 1 slides, updating .gitignore for day 2 materials, and reorganizing course materials. The 'About' section notes that it's in progress and for the data science basics course in 2024. It has 1 star, 1 watching, and 0 forks.

Code Issues Pull requests Actions Projects Security Insights Settings

data-science-basics-2024 Public

main 2 Branches 0 Tags

Go to file Add file <> Code

seaneff update gitignore to add day 2 materials 8c7a75d · 4 minutes ago 48 Commits

course-datasets initial commit of energy dataset 19 hours ago

day1 fix day 1 slides 19 hours ago

day2 update gitignore to add day 2 materials 4 minutes ago

download-R reorganize and show only active days of course yesterday

About

In progress materials for the data science basics course in 2024

Readme Activity 1 star 1 watching 0 forks

Upload your code to GitHub

The screenshot shows a GitHub repository page for 'seaneff / data-science-basics-2024'. The repository is public and contains 2 branches and 0 tags. The commit history is listed, with the most recent commit being a merge from 'seaneff' that updated '.gitignore' to add day 2 materials. A context menu is open over this commit, with the 'Upload files' option circled in red.

Code Issues Pull requests Actions Projects Security Insights Settings

data-science-basics-2024 Public

Unpin Unwatch 1 Fork 0 Starred 1

main 2 Branches 0 Tags

Go to file Add file Code

seaneff update gitignore to add day 2 materials
course-datasets initial commit of energy dataset
day1 fix day 1 slides
day2 update gitignore to add day 2 materials
download-R reorganize and show only active days of course

+ Create new file
Upload files

About

In progress materials for the data science basics course in 2024

Readme Activity 1 star 1 watching 0 forks

Upload your code to GitHub

The screenshot shows a GitHub repository page for 'seaneff / data-science-basics-2024'. The 'Code' tab is selected. Below the navigation bar, there's a search bar and a header with the repository name. The main area displays a large button with a red oval highlighting the 'Drag files here to add them to your repository' text and a file icon. A red arrow points from the right side of the image towards this button, with the text 'Drag and drop or click to upload your .R file' written next to it. At the bottom, a modal window titled 'Commit changes' is open, containing fields for adding files via upload and an optional extended description, and two radio button options for committing directly to the main branch or creating a new branch.

seaneff / data-science-basics-2024

Type to search

Code Issues Pull requests Actions Projects Security Insights Settings

data-science-basics-2024 /

Drag files here to add them to your repository
Or [choose your files](#)

Commit changes

Add files via upload

Add an optional extended description...

Commit directly to the `main` branch.

Create a new branch for this commit and start a pull request. [Learn more about pull requests.](#)

Commit changes Cancel

Drag and drop or click to upload your .R file

Upload your code to GitHub

The screenshot shows a GitHub repository page for "seaneff / data-science-basics-2024". The "Code" tab is selected. Below the navigation bar, there is a search bar and a file upload area with the placeholder "Drag files here to add them to your repository" and a "choose your files" button. A red oval highlights the "Commit changes" modal, which contains fields for "Add files via upload" and "Add an optional extended description...". It also includes two radio button options: one selected for "Commit directly to the main branch." and another for "Create a new branch for this commit and start a pull request." A red arrow points from the text "Add some notes describing what the code does" to the "extended description" input field.

seaneff / data-science-basics-2024

Type to search

Code Issues Pull requests Actions Projects Security Insights Settings

data-science-basics-2024 /

Drag files here to add them to your repository
Or [choose your files](#)

Commit changes

Add files via upload

Add an optional extended description...

Commit directly to the `main` branch.

Create a new branch for this commit and start a pull request. [Learn more about pull requests.](#)

Commit changes Cancel

Add some notes describing what the code does

Upload your code to GitHub

The screenshot shows a GitHub repository page for "seaneff / data-science-basics-2024". The "Code" tab is selected. Below the navigation bar, there is a search bar and a file upload area with the placeholder "Drag files here to add them to your repository" and a "choose your files" button. A modal window titled "Commit changes" is open, containing fields for "Add files via upload" and "Add an optional extended description...". At the bottom of the modal, there are two radio button options: one selected for "Commit directly to the main branch." and another for "Create a new branch for this commit and start a pull request." A red arrow points to the "Commit changes" button, which is highlighted with a red circle.

data-science-basics-2024 /

Drag files here to add them to your repository
Or [choose your files](#)

Commit changes

Add files via upload

Add an optional extended description...

Commit directly to the `main` branch.

Create a new branch for this commit and start a pull request. [Learn more about pull requests.](#)

Commit changes **Cancel**

Commit your changes

OPTIONAL homework

explore some repositories on github

- Lots of people put code on github for lots of different types of projects
- Spend some time exploring open-source projects on github.
 - What did you find? What types of projects?
 - How were the repositories organized?
 - What surprised you?

OPTIONAL homework

explore some repositories on github

- Some examples of repositories on github to explore
 - [Plague data analysis](#) - Dr. Colin Carlson
 - [A color palette for R](#) - Dr. Karthik Ram
 - [Ebola data](#) - Dr. Caitlin Rivers
 - [Amazing data visualizations](#) - Dr. Cara Thompson

Recap from today

Data management and version control

- Understand the foundations of data management
- Learn best practices for documentation
- Load and clean your first dataset in R
 - learn about data structures
 - filter data to excluding missing values
 - explore a new dataset with summary statistics and plots
- Upload your code to GitHub

Plan for tomorrow

Exploratory data analysis

- What is exploratory data analysis (EDA)?
- Calculate basic descriptive statistics in R
- Explore different strategies for data visualization
- Make a repository on github
- Build your first data visualizations in R

Thank you!

See you tomorrow.

Please come with a fully charged laptop.