



# ENVIRONMENTAL DATA ANALYTICS

## WEEK 3 – M2 – CODING BASICS AND REPRODUCIBILITY

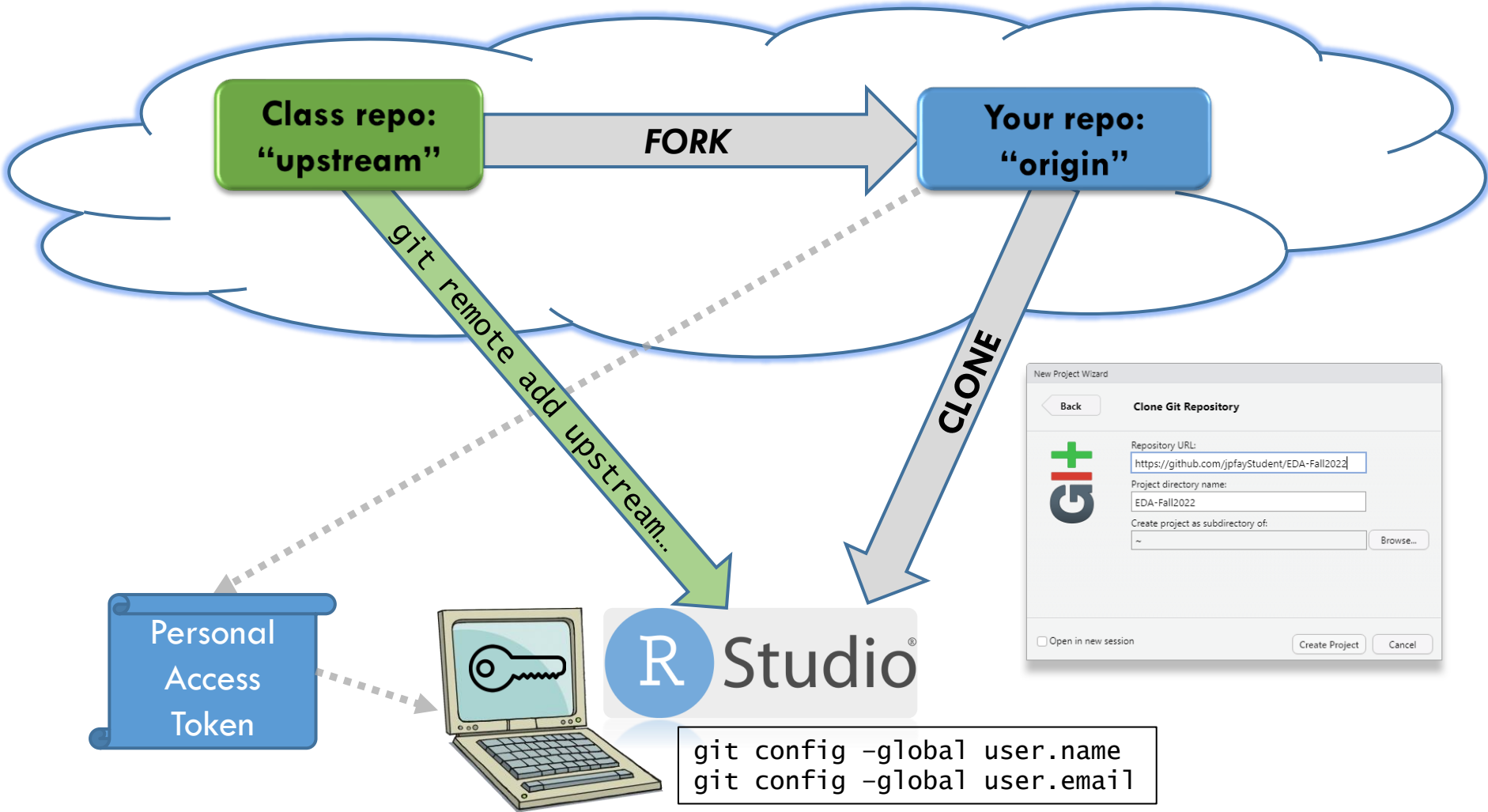
Fall 2023

Nicholas School of the Environment - Duke University

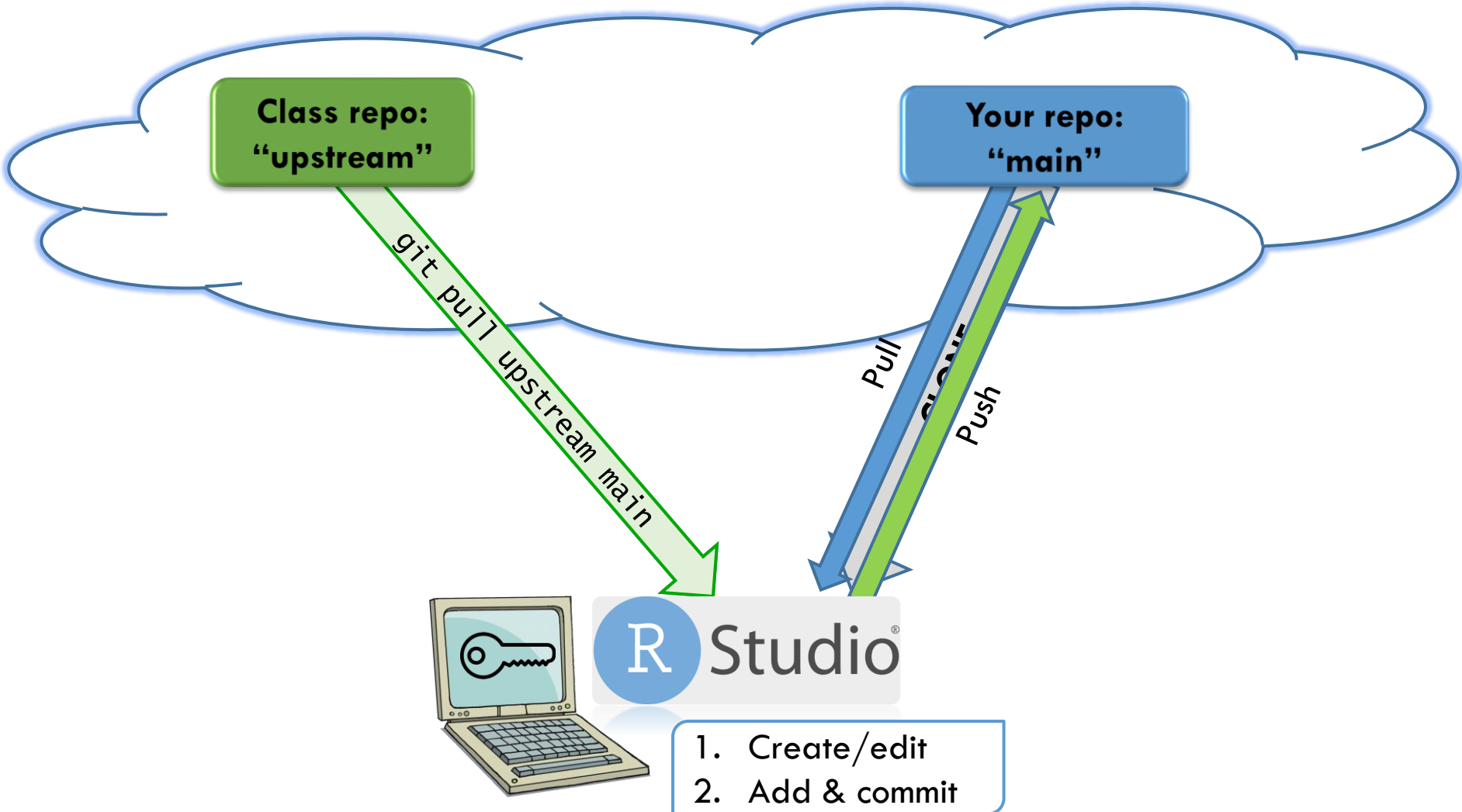
# Week 3 - Agenda

- Questions – M1 & A1
  - ▣ Git/GitHub...
  - ▣ What is “Data Analytics”
- Overview M2
  - ▣ Questions – M2 videos
- Intro to Data frame in R
- A2 posted and due on Thursday @ 5pm

# Explaining Git/GitHub: *Setup*



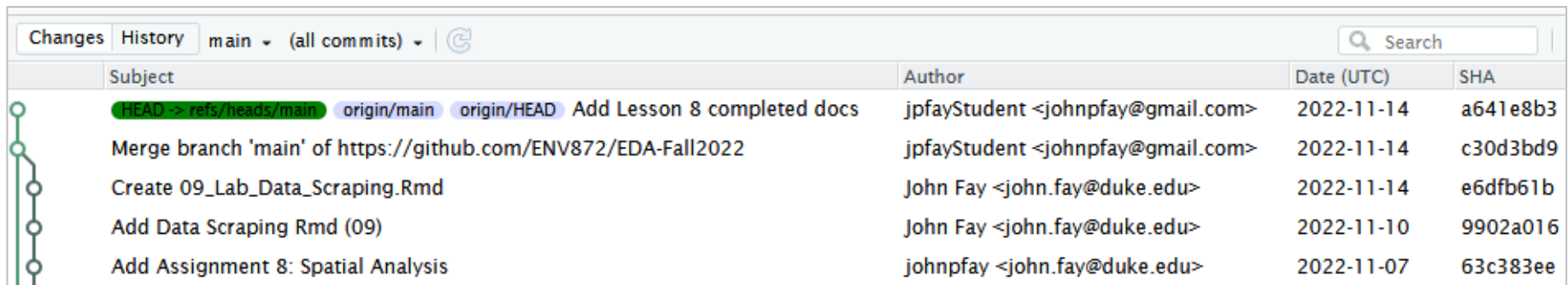
# Explaining Git/GitHub: *Use*



# Explaining Git/GitHub: *Workflow*

<https://git-school.github.io/visualizing-git>

- Single repository...
- Remotes repositories...
- Upstream changes...
- Explaining merges & divergent branches



The screenshot shows a Git commit history table with the following columns: Subject, Author, Date (UTC), and SHA. The table lists five commits, with the top one being a merge of the 'main' branch from a remote repository. The commit subjects are: 'HEAD -> refs/heads/main Merge branch 'main' of https://github.com/ENV872/EDA-Fall2022', 'Create 09\_Lab\_Data\_Scraping.Rmd', 'Add Data Scraping Rmd (09)', and 'Add Assignment 8: Spatial Analysis'. The authors are listed as 'jpfayStudent <johnpfay@gmail.com>' and 'John Fay <john.fay@duke.edu>'. The dates range from 2022-11-07 to 2022-11-14. The SHA values are a641e8b3, c30d3bd9, e6dfb61b, 9902a016, and 63c383ee.

Subject	Author	Date (UTC)	SHA
HEAD -> refs/heads/main Merge branch 'main' of https://github.com/ENV872/EDA-Fall2022	jpfayStudent <johnpfay@gmail.com>	2022-11-14	a641e8b3
Create 09_Lab_Data_Scraping.Rmd	John Fay <john.fay@duke.edu>	2022-11-14	c30d3bd9
Add Data Scraping Rmd (09)	John Fay <john.fay@duke.edu>	2022-11-14	e6dfb61b
Add Assignment 8: Spatial Analysis	John Fay <john.fay@duke.edu>	2022-11-10	9902a016
	johnpfay <john.fay@duke.edu>	2022-11-07	63c383ee

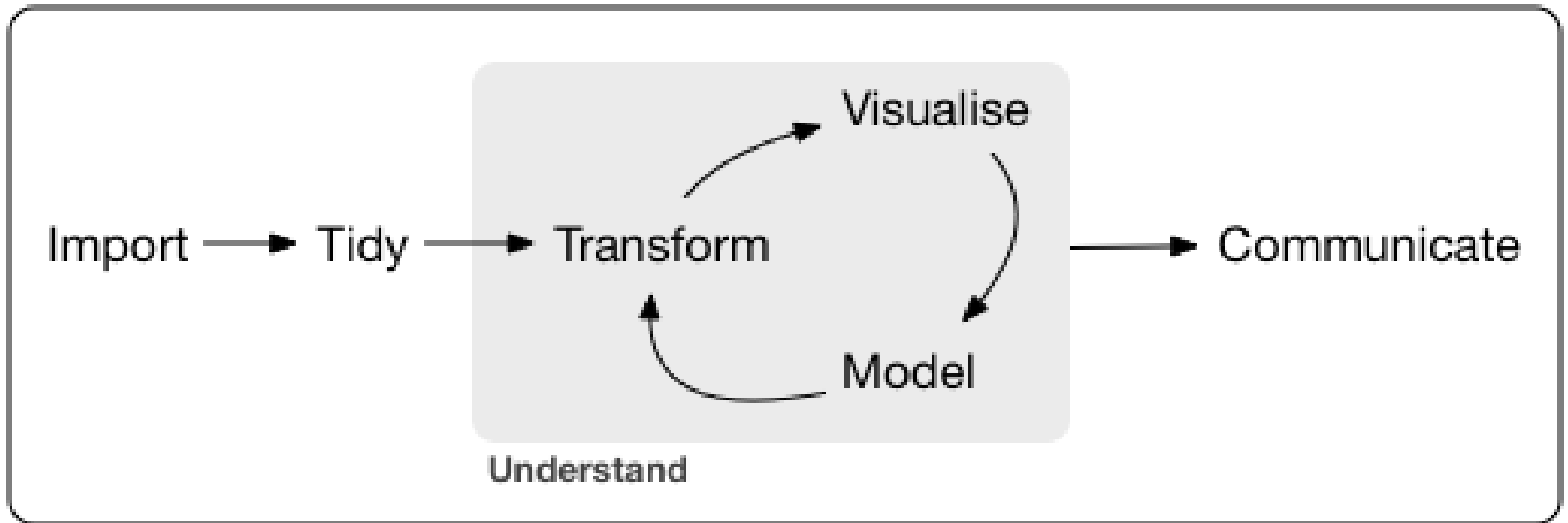
# Explaining Git/GitHub: *Terms*

- Repository
- Git vs GitHub
- Fork vs Clone
- Staging vs Committing
- SHA vs Commit message
- Push vs Pull
- `git pull` vs `git pull upstream main`
- Merge

# If Git goes totally sideways...

- Make a back up (e.g. zip) of you project folder
- Close RStudio project and rename project folder
- Create a new project linked to your forked repo
- Link to upstream remote, as before
- Copy over any missing items from your renamed folder to your newly cloned repository

# Q&A: What is Data Analytics



<https://vita.had.co.nz/papers/tidy-data.pdf>



# Tidy Data

## Tidy data

1. Each variable is in its own column
2. Each observation is in its own row
3. Each value is in its own cell

country	year	cases	population
Afghanistan	2000	35	1995071
Afghanistan	2000	366	20045360
Brazil	1999	3737	17204362
Brazil	2000	8188	17404698
China	1999	21258	127015272
China	2000	21266	12801583

variables

country	year	cases	population
Afghanistan	2000	35	1995071
Afghanistan	2000	366	20045360
Brazil	1999	3737	17204362
Brazil	2000	8188	17404698
China	1999	21258	127015272
China	2000	21266	12801583

observations

country	year	cases	population
Afghanistan	2000	35	1995071
Afghanistan	2000	366	20045360
Brazil	1999	3737	17204362
Brazil	2000	8188	17404698
China	1999	21258	127015272
China	2000	21266	12801583

values

<http://garrettgman.github.io/tidying/>

# Tidy Data

	Site_ID	State Code	County Code	Site Number	Month1	Month2	Month3	Month4	Month5	Month6	Month7	Month8	Month9	Month10	Month11	Month12
1	010030010	01	003	0010	7.366667	7.211111	8.1625	6.93	6.736364	6.85	8.36	8.24	6.78	6.444444	8.43	5.86
2	010270001	01	027	0001	5.77	5.144444	7.875	6.9625	6.85	6.975	7.65	8.59	6.5125	6.388889	6.91	6.07
3	010491003	01	049	1003	8.07	6.228571	7.12	7.7625	7.472727	7.24	10.18	9.38	6.925	6.8	7.688889	7.55
4	010550010	01	055	0010	8.31	7.571429	8.54	8.91	7.890909	7.425	10.133333	9.5	7.77	10.05	<Null>	7.923183
5	010730023	01	073	0023	9.580169	12.099422	11.653191	12.961716	11.564921	10.522225	12.110792	14.396602	13.992682	13.834073	11.500682	12.58887
6	010731005	01	073	1005	7.46	6.75	9.76	7.12	7.48	6.88	8.6	11.84	8.32	8.52	10.28	4.84
7	010731010	01	073	1010	6.94	6.328571	9.09	7.84	7.94	6.98	8.89	9.6875	8.57	9.491667	8.48	4.49
8	010732003	01	073	2003	7.586667	8.272727	9.493333	8.4	8.04	8.01875	9.773333	9.038462	8.386667	9.689474	9.993333	7.22

	Site_ID	State Code	County Code	Site Number	Month	Mean_SO2_ppm
1	010030010	01	003	0010	1	7.366667
2	010030010	01	003	0010	2	7.211111
3	010030010	01	003	0010	3	8.1625
4	010030010	01	003	0010	4	6.93
5	010030010	01	003	0010	5	6.736364
6	010030010	01	003	0010	6	6.85
7	010030010	01	003	0010	7	8.36
8	010030010	01	003	0010	8	8.24

# Basics of Data Analytics

- Understand what it means to “tidy” data
- Differentiate “primary” and “secondary” data
- Differentiate “qualitative” and “quantitative” data
- Identify different file types used in data analytics and discuss why some formats are better than others in terms of transparency and reproducibility
- Describe the various data structures used in data analytics and what each are used for: *Vectors, matrices, arrays, data frames, lists*
- Understand the difference between *R* and *RStudio*
- Become familiar with the typical layout of an *RStudio session*

A horizontal decorative bar at the top of the slide, consisting of an orange square on the left and a blue rectangle on the right.

# Reproducibility & Coding Basics

# Reproducibility

- Raw data are always separated from processed data; Raw datasets are NEVER changed
- Cleaning/transformations done through coding, not by editing (e.g., within Excel)
- Edits are documented by well-commented code
- Open-source formats are used wherever possible

*Often, you'll spend the majority in the data processing phase (cleaning & wrangling data)*

# Navigating RStudio

The screenshot displays the RStudio interface with the following components:

- Source Editor:** Shows a markdown file with the following content:

```
1 ---
2 title: "2: Reproducibility and Coding Basics"
3 author: "Environmental Data Analytics | John Fay and Luana Lima | Developed by Kateri Salk"
4 date: "Spring 2022"
5 output: pdf_document
6 geometry: margin=2.54cm
7 editor_options:
8   chunk_output_type: console
9 ---
10
11 ## Objectives
12
13 1. Discuss the benefits and approach for reproducible data analysis
14 2. Perform simple operations using R coding syntax
15 3. Call and create functions in R
16
17 ## Reproducible Data Analysis
18
19 ### Fundamentals of reproducibility
20
21 **Reproducibility**: when someone else (e.g., future self) can obtain the same
22 outcomes from the same dataset and analysis
23
24 * Raw data are always separate from processed data
25 * Link data transformations with a reproducible pipeline
26 * Raw datasets NEVER changed
27 * Cleaning/transformations done through coding, not by editing within Excel
```
- Console:** Shows the execution of R code:

```
R 4.1.2 - V:/Environmental_Data_Analytics_2022/
> meal1 <- recipe4(4); meal1
[1] 2
> meal2 <- recipe4(2); meal2
[1] 4
> meal3 <- recipe5(3); meal3
[1] 3
>
> recipe6 <- function(x){
+   ifelse(x<3, x*2, x/2) #log_exp, if TRUE, if FALSE
+ }
>
> meal4 <- recipe6(4); meal4
[1] 2
> meal5 <- recipe6(2); meal5
[1] 4
>
> # Chunk 12
> ??seq
> |
```
- Environment Pane:** Shows the current environment with the following objects:

Object	Type	Value
doublecomplex...	List of 2	
doublesimplem...	List of 2	
long_name_for_...	int	11
meal1	int	2
meal2	int	4
meal3	int	3
meal4	int	2
meal5	int	4
simplemeal	int	10
ten_sequence	int [1:10]	1 2 3 4 5 6 7 8 9 10
x	int	12
- Files Pane:** Shows the file structure of the project:

Name	Size	Modified
..		
.gitignore	636 B	Jan 13, 2022, 9:16 AM
.rhistory	0 B	Jan 13, 2022, 10:00 AM
Assignments		
Environmental_Data_Analytics_2022...	218 B	Jan 14, 2022, 8:26 AM
Lessons		
LICENSE	35 KB	Jan 13, 2022, 9:16 AM
README.md	150 B	Jan 13, 2022, 9:17 AM
Resources		

# R Configuration tips

Options dialog, Basic tab. The left sidebar shows categories: General, Code, Console, Appearance, Pane Layout, Packages, R Markdown, Python, Sweave, Spelling, Git/SVN, Publishing, Terminal, and Accessibility. The main area is titled 'Basic' and contains the following sections:

- R Sessions**
  - R version: [64-bit] C:\Program Files\R\R-4.3.1 (Change...)
  - Default working directory (when not in a project): ~ (Browse...)
  - Restore most recently opened project at startup
  - Restore previously open source documents at startup
- Workspace** (highlighted with an orange box)
  - Restore .RData into workspace at startup
  - Save workspace to .RData on exit: Never (dropdown)
- History**
  - Always save history (even when not saving .RData)
  - Remove duplicate entries in history
- Other**
  - Wrap around when navigating to previous/next tab
  - Automatically notify me of updates to RStudio
  - Send automated crash reports to RStudio

Options dialog, Basic tab. The left sidebar shows categories: General, Code, Console, Appearance, Pane Layout, Packages, R Markdown, Python, Sweave, Spelling, Git/SVN, Publishing, Terminal, and Accessibility. The main area is titled 'Basic' and contains the following sections:

- R Markdown**
  - Show document outline by default
  - Soft-wrap R Markdown files
  - Show in document outline: Sections Only (dropdown)
  - Show output preview in: Window (dropdown)
  - Show output inline for all R Markdown documents
  - Show equation and image previews: Inline (dropdown)
  - Evaluate chunks in directory: Project (dropdown, highlighted with an orange box)
- R Notebooks**
  - Execute setup chunk automatically in notebooks
  - Hide console automatically when executing notebook chunks
  - [? Using R Notebooks](#)

# Coding Basics: R-Markdown



- ▣ Creating an RMarkdown document
- ▣ What is the **header portion** called? What is its purpose?
- ▣ What is the difference between **text/markdown** and **code**
- ▣ Creating code cells/code chunks; components of a code cell
- ▣ Knitting - reports
- ▣ Folding code, contents



# Coding Basics: Elements of R

- **Values**

- 25, “Environment”, False, July 4, 1776

- **Objects** - A “container” that holds a value or values

- `mpg <- 42`

- `my_dog <- "Rover"`

- `colors <- c("Red", "Blue", "Green")`

- **Functions** – instructions applied to values/object

- `mean(2, 5, 19, 20)`

# Review

- Which of these are numbers?

11

“11”

“eleven”

eleven

---

number

---

value (string)

---

object

# Coding Basics



- Console vs scripts
- Running commands (console | script)
- Variable assignments & naming strategies
- Comments
- Functions
  - ▣ Structure
  - ▣ Invoking existing functions
  - ▣ Creating new functions

# Coding Basics: Data Structures

## Vectors

- 1-dimensional sequence of data elements
- All items have of the same data type (e.g. int, chr)

## Matrices

- 2-dimensional sequence of data elements
- Allows matrix multiplication and other linear algebra operations

## Arrays

- Includes vectors and matrices, includes  $> 2$  dimensions

# Coding Basics: Data Structures (cont'd)

## Data Frames

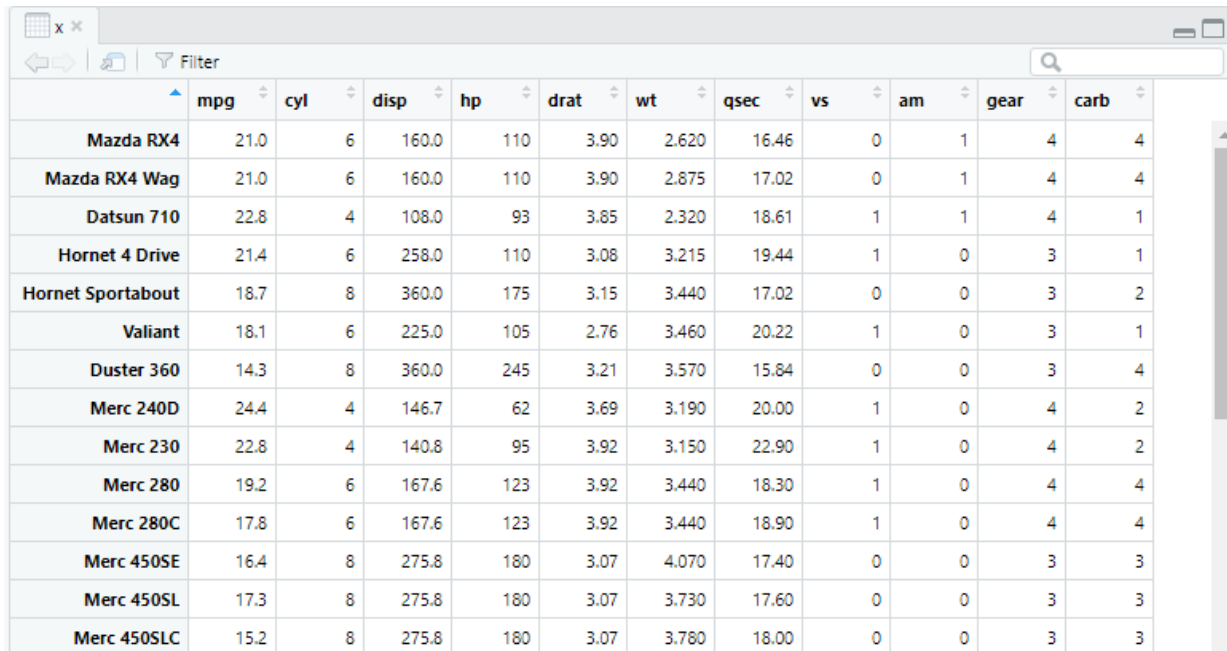
- Stores data in 2 dimensions: rows and columns
- Each column is a “named vector” (same data type)

## Lists

- Ordered collection of objects (mixed data types)

# Coding Basics: Data Frames

- What does a dataframe consist of?
- Difference between *dataframe* and a *matrix*
- Advantages of a dataframe?



A screenshot of a data frame table with 13 columns and 16 rows. The columns are labeled: mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb. The rows represent different car models and their specifications.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3

# Coding Basics (Hands on)



- Coding with dataframes
  - ▣ Referencing columns
  - ▣ Creating dataframes
  - ▣ Properties of dataframes
  - ▣ Extracting data from dataframes

# Up next: M3

## More on coding basics/Data exploration:

- Loading data into R
  - ▣ Data types; dealing with pesky dates
- Exploring data with R: structure and values
- *Visually* exploring data with plots
  - ▣ Bar plots, histograms, scatterplots, etc.



# The class “rhythm”

Each week = 1 module = {recordings + exercise + assignment}

- Recordings:

- ▣ Watch recordings prior to class

- In class:

- ▣ Ask questions about recordings and about assignment

- ▣ Group exercises to re-inforce concepts (+ some new concepts)

- Assignments:

- ▣ Made available after class discussion

- A01 made available after class session last week; A02 available now

- ▣ Due on Thursday @ 5pm