Data Management for Research and Institutional Decision Making

Peace Aber & Stephen Opiyo

Topics to be covered

- General overview of data management
- Installation of tools for data management, i.e, R, RStudio, QGIS, and GitHub
- Data set description and metadata
- Using RStudio for data management
- Using RStudio for exploratory data analysis
- Using QGIS for the management and visualization of spatial data
- Using GitHub for storing and sharing data

General overview of data management

- Data management is the practice of collecting, organizing, protecting, and storing an organization's data so it can be used effectively for analysis and decision-making.

- It involves a wide range of activities, from ensuring data quality to establishing data governance policies.

- Data management maximizes the value of data as a strategic asset.

Key aspects of Data Management

- **Data Collection:** Gathering data from various sources, both internal and external to the organization.
- **Data Storage:** Storing data securely and efficiently, often across different platforms and locations (cloud, on-premises).
- **Data Organization:** Structuring and organizing data in a way that makes it accessible and usable for analysis.
- **Data Quality:** Ensuring the accuracy, completeness, and consistency of data.

Key aspects of Data Management

- **Data Security:** Protecting data from unauthorized access, breaches, and loss.
- **Data Governance:** Establishing policies and procedures for data management, including data ownership, access control, and compliance.
- **Data Analysis:** Utilizing data to gain insights and support decision-making.
- **Data Lifecycle Management:** Managing data throughout its entire lifecycle, from creation to archiving or destruction.

Tools for data management

- R and R-Studio
- QGIS
- GitHub

Installation of software

https://drive.google.com/drive/folders/10ZlJjxE1rlL9Dc3kdjvng-5dIbLPOIC9

To install R and R-Studio: Open the install_R_Rstudio.pdf and follow the instructions.

To install QGIS: Open the install_QGIS.pdf and follow the instructions

Data description

Dataset name: Kisumu_main_vendor.csv and YSM Animal Health Data.csv

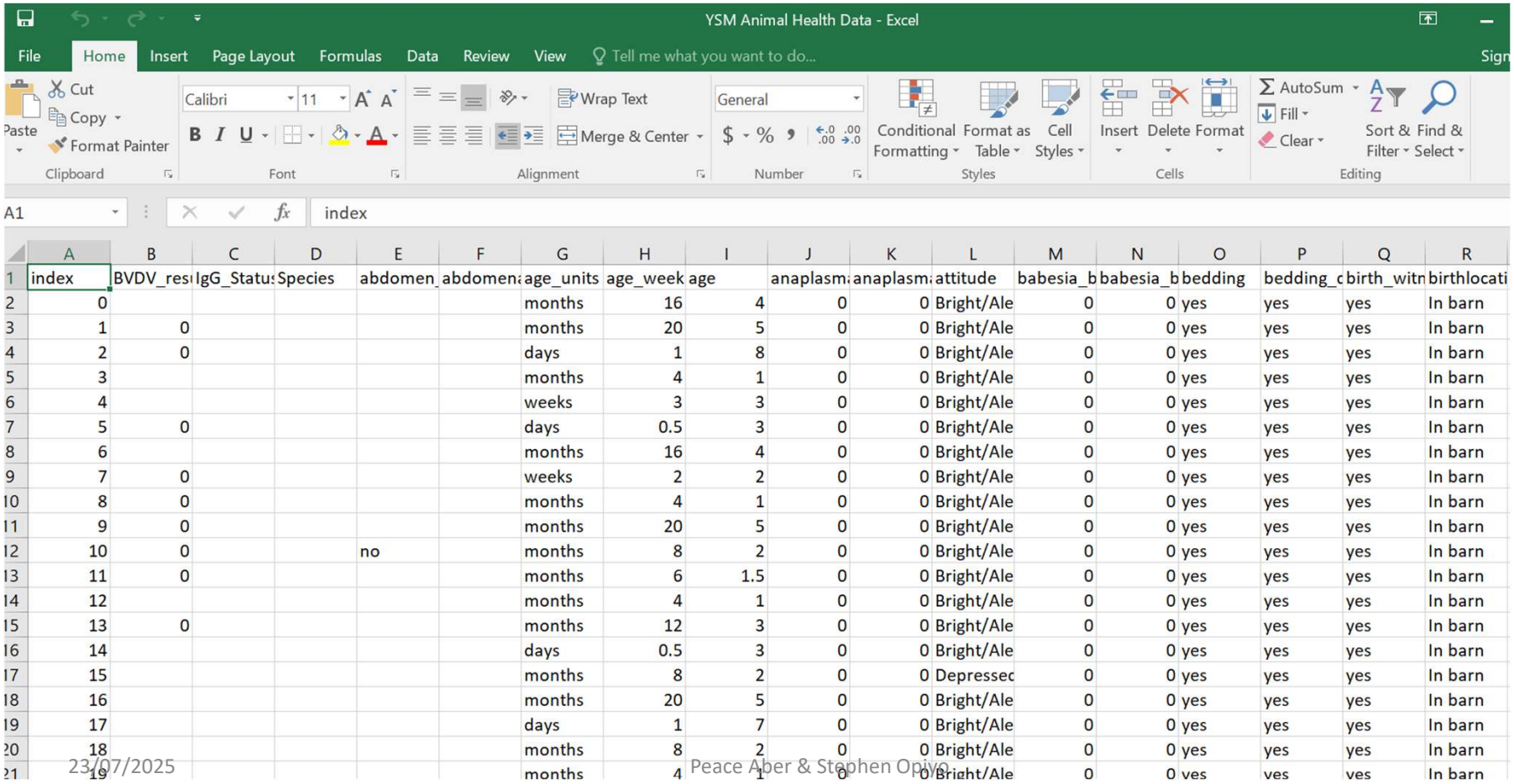
kisumu_main_vendor - Excel

A1 | start_time

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	start_time	end_time	survey_date	survey_tin	consent	region	country	county	subcounty	ward	slum	cu	location_t	vendor_ty	gender
2	52:46.9	08:38.3	2/22/2022	09:53:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
3	08:43.2	18:58.1	2/22/2022	10:09:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Kiosk	Multiple females
4	30:04.8	27:52.5	2/22/2022	10:30:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Multiple females
5	47:24.0	29:48.3	2/22/2022	10:47:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
6	03:59.6	10:24.0	2/22/2022	11:04:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Kiosk	Multiple females
7	41:52.6	31:26.7	2/22/2022	11:42:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	4	Peri-Urbar	Stalls/Tab	Male
8	51:48.3	57:51.1	2/22/2022	11:54:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	both male and fem
9	05:47.7	16:40.6	2/22/2022	12:05:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
10	19:37.7	31:05.2	2/22/2022	12:20:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Kiosk	Female
11	37:49.0	48:48.8	2/22/2022	12:37:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Kiosk	Female
12	51:30.0	57:52.4	2/22/2022	12:51:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
13	05:41.1	15:05.4	2/22/2022	13:05:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
14	24:03.5	33:58.4	2/22/2022	13:24:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Male
15	33:48.4	40:28.5	2/22/2022	13:34:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
16	41:44.7	47:33.2	2/22/2022	13:42:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Male
17	51:57.9	36:32.7	2/22/2022	13:52:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
18	01:52.4	37:26.5	2/22/2022	14:02:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
19	16:01.9	22:14.2	2/22/2022	14:16:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
20	[illegible]	26:19.2	2/22/2022	14:23:00.C	1	East Africa	Kenya	[illegible]	[illegible]	[illegible]	Manyatta	3	Peri-Urbar	Stalls/Tab	Female
21	26:27.9	31:40.1	2/22/2022	14:26:00.C	1	East Africa	Kenya	Kisumu	Kisumu Ea	Manyatta	Manyatta	3	Peri-Urbar	Stalls/Tab	Female

Data description

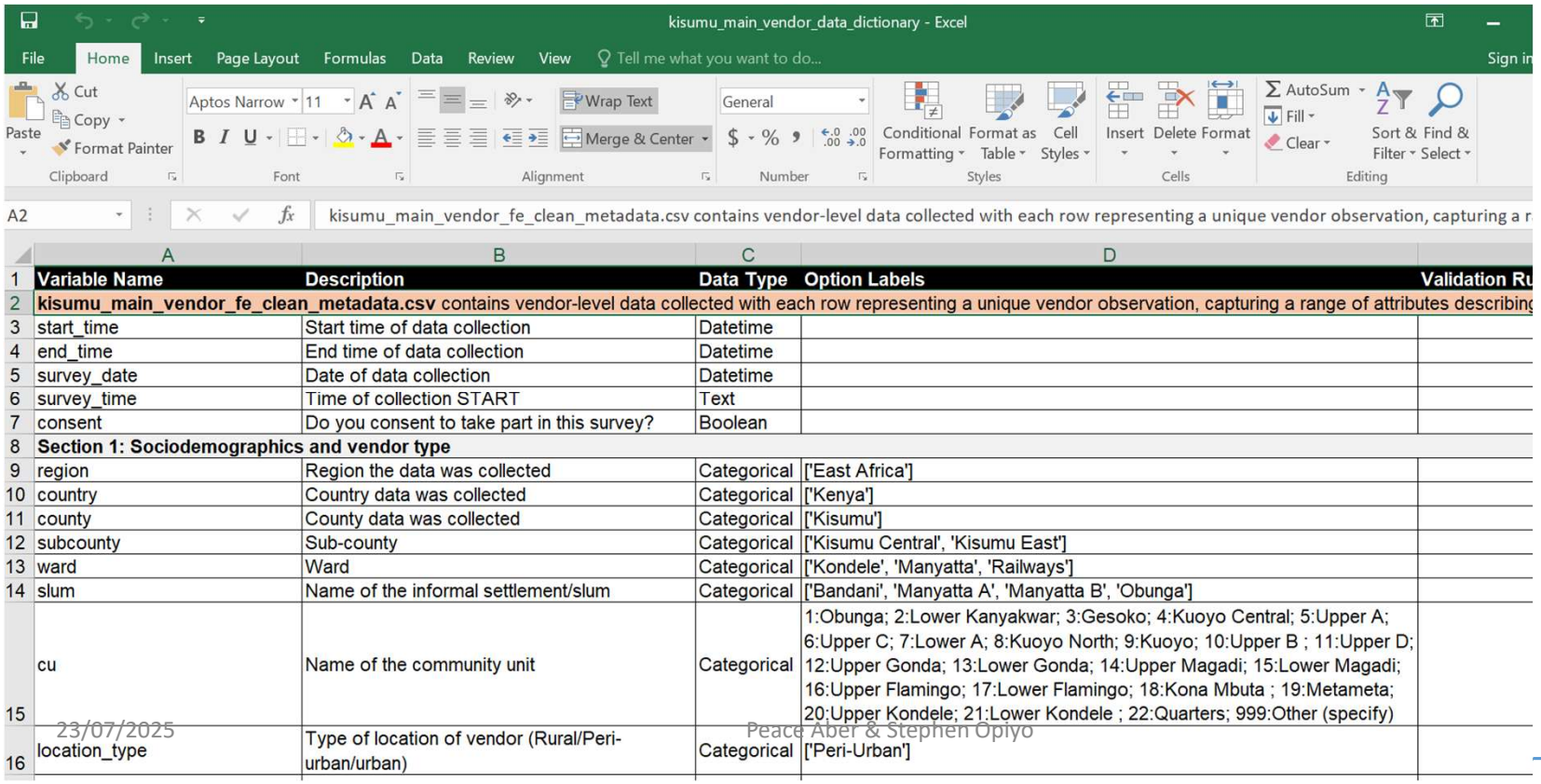
Dataset name: Kisumu_main_vendor.csv and YSM Animal Health Data.csv



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	index	BVDV_res	IgG_Status	Species	abdomen_	abdomen	age_units	age_week	age	anaplasm	anaplasm	attitude	babesia_b	babesia_b	bedding	bedding_c	birth_witn	birthlocati
2	0						months	16	4	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
3	1	0					months	20	5	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
4	2	0					days	1	8	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
5	3						months	4	1	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
6	4						weeks	3	3	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
7	5	0					days	0.5	3	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
8	6						months	16	4	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
9	7	0					weeks	2	2	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
10	8	0					months	4	1	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
11	9	0					months	20	5	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
12	10	0			no		months	8	2	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
13	11	0					months	6	1.5	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
14	12						months	4	1	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
15	13	0					months	12	3	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
16	14						days	0.5	3	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
17	15						months	8	2	0	0	Depressed	0	0	yes	yes	yes	In barn
18	16						months	20	5	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
19	17						days	1	7	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
20	18						months	8	2	0	0	Bright/Ale	0	0	yes	yes	yes	In barn
21	19						months	4	1	0	0	Bright/Ale	0	0	yes	yes	yes	In barn

Metadata

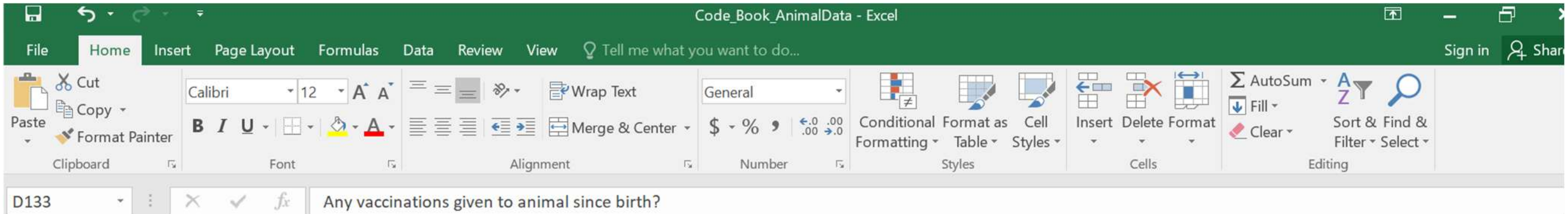
- Metadata is information that describes data that has been collected.
- It comes in many formats e.g, Excel, Word,….etc
- Examples: Kisumu_main_vendor_data_dictionary and Code_Book_AnimalData stored in excel sheets



	A	B	C	D
1	**Variable Name**	**Description**	**Data Type**	**Option Labels**
2	**kisumu_main_vendor_fe_clean_metadata.csv** contains vendor-level data collected with each row representing a unique vendor observation, capturing a range of attributes describing			
3	start_time	Start time of data collection	Datetime	
4	end_time	End time of data collection	Datetime	
5	survey_date	Date of data collection	Datetime	
6	survey_time	Time of collection START	Text	
7	consent	Do you consent to take part in this survey?	Boolean	
8	**Section 1: Sociodemographics and vendor type**			
9	region	Region the data was collected	Categorical	['East Africa']
10	country	Country data was collected	Categorical	['Kenya']
11	county	County data was collected	Categorical	['Kisumu']
12	subcounty	Sub-county	Categorical	['Kisumu Central', 'Kisumu East']
13	ward	Ward	Categorical	['Kondele', 'Manyatta', 'Railways']
14	slum	Name of the informal settlement/slum	Categorical	['Bandani', 'Manyatta A', 'Manyatta B', 'Obunga']
15	cu	Name of the community unit	Categorical	1:Obunga; 2:Lower Kanyakwar; 3:Gesoko; 4:Kuoyo Central; 5:Upper A; 6:Upper C; 7:Lower A; 8:Kuoyo North; 9:Kuoyo; 10:Upper B ; 11:Upper D; 12:Upper Gonda; 13:Lower Gonda; 14:Upper Magadi; 15:Lower Magadi; 16:Upper Flamingo; 17:Lower Flamingo; 18:Kona Mbuta ; 19:Metameta; 20:Upper Kondele; 21:Lower Kondele ; 22:Quarters; 999:Other (specify)
16	location_type	Type of location of vendor (Rural/Peri-urban/urban)	Categorical	['Peri-Urban']

Metadata

- Metadata is information that describes data that has been collected.
- It comes in many formats e.g, Excel, Word,….etc
- Examples: Kisumu_main_vendor_data_dictionary and Code_Book_AnimalData stored in excel sheets



	Section	Dataset Column	Survey_Questi	Question/Description	Levels
2	n/a	index	n/a	Response identification number	
3	Laboratory Form	BVDV_result	n/a	Result of BVDV test	0 (negative); 1 (positive
4	Laboratory Form	IgG_Status	n/a	Immunoglobulin status	adequate; failure; parti
5	n/a	Species	n/a	Livestock species type	bovine; goat kid; lamb
6	Physical Exam Form	abdomen_pingsounds	8	Abdominal Auscultation Rumen: Are there any ping sounds?	0 (no), 1 (yes)
7	Physical Exam Form	abdomenauscultation_notes	9	Abdominal Auscultation Notes: Notes on abdominal auscultation	
8	Young Stock Enrollment Form	age	2	Age	
9	Young Stock Enrollment Form	age_units	2	Units of time	days, weeks, months
10	n/a	age_weeks	n/a	Age in weeks	
11	Laboratory Form	anaplasma_centrale_result	n/a	Result of anaplasma centrale	0 (negative); 1 (positive
12	Laboratory Form	anaplasma_marginale_result	n/a	Result of anaplasma marginale	0 (negative); 1 (positive
13	Physical Exam Form	attitude	1	Attitude	Bright/Alert; Depressed
14	Laboratory Form	babesia_bigemina_result	n/a	Result of babesia bigemina	0 (negative); 1 (positive
15	Laboratory Form	babesia_bovis_result	n/a	Result of babesia bovis	1 (negative); 1 (positive
16	Young Stock Enrollment Form	bedding	36	Is the animal housed on bedding?	Yes; No
17	Young Stock Enrollment Form	bedding_dryandclean	37	Does the area where the animal lies down look clean and dry?	Yes; No
18	Young Stock Enrollment Form	birth_witness	8	Birth witnessed	Yes; No

Introduction to R

What is R?

- R is a free programming language that is used for data management, data analysis and data visualization.
- R easily integrates with other programing languages.
- R has a vast community of users.

Introduction to R Studio

What is R Studio ?

- R Studio is an integrated development environment (IDE) for R.
  - It includes a console, syntax-highlighting editor that supports direct code execution.
  - Tools for plotting, history, debugging and workspace management.

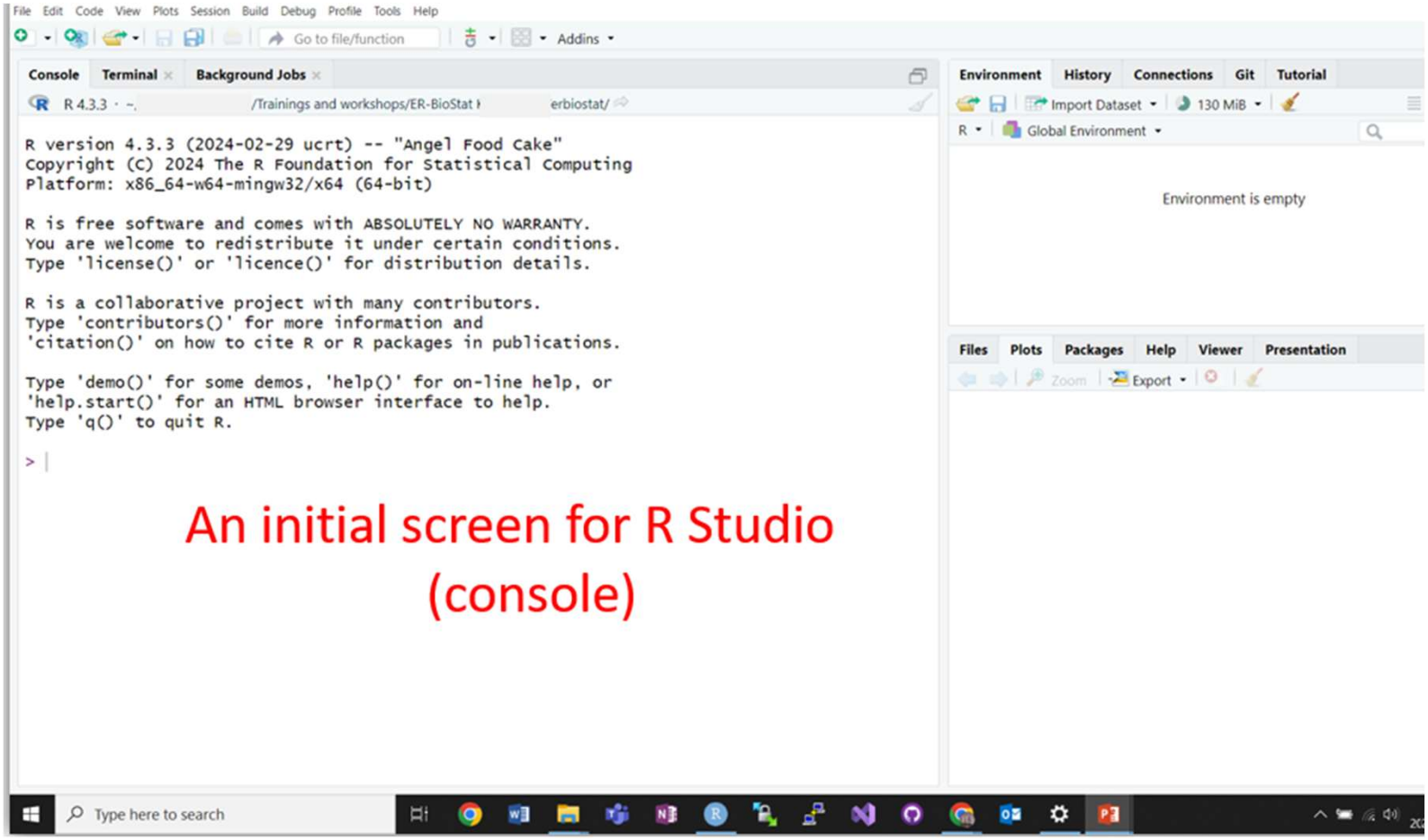
- R Studio is available in **open source** and **commercial** editions and runs on the desktop (Windows, Mac, and Linux).

What is R Studio ?

- More information:

  https://rstudio.com/products/rstudio/

Rstudio

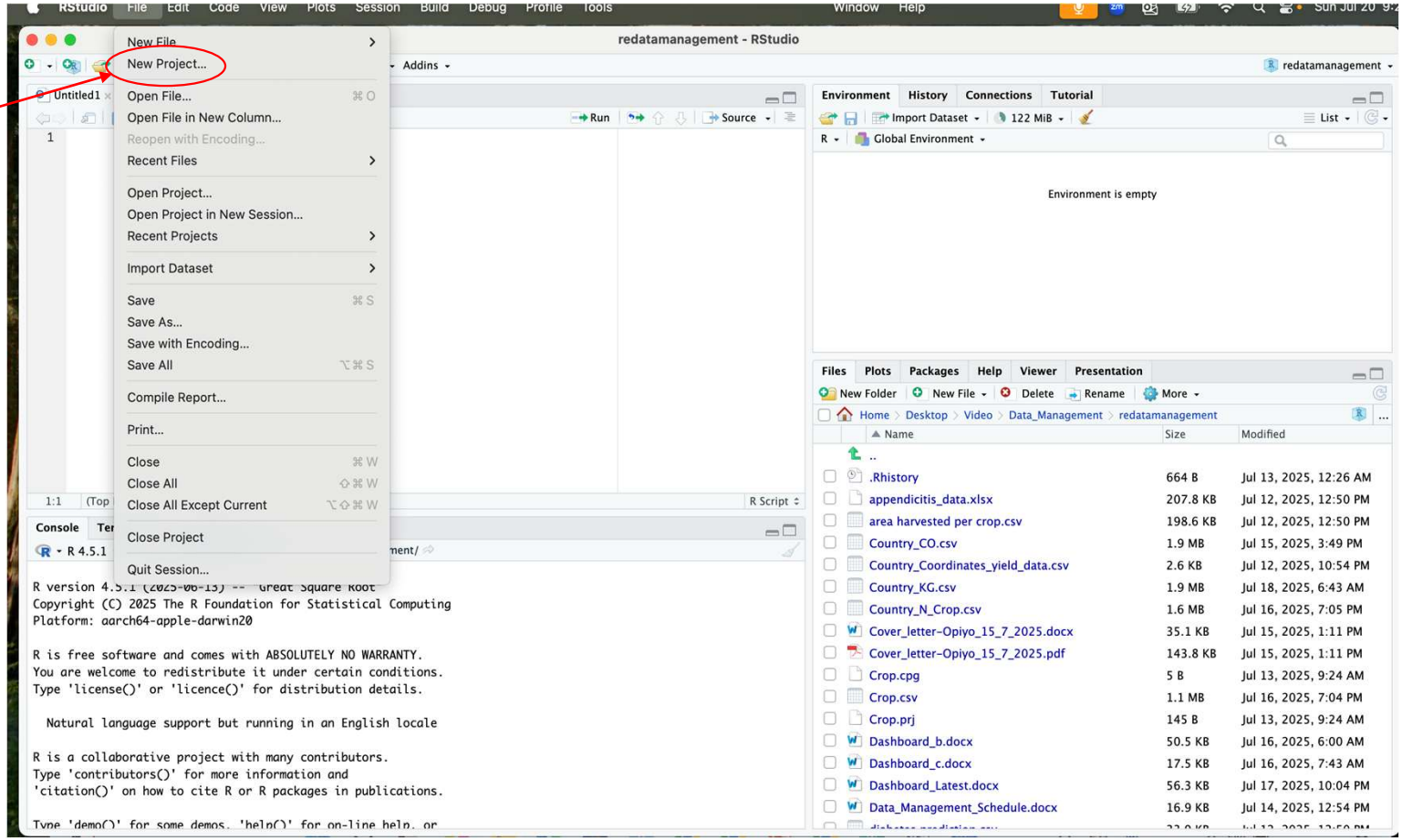

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Console
Terminal
Background Jobs
R 4.3.3 · ~.
/Trainings and workshops/ER-BioStat
erbiostat/
R version 4.3.3 (2024-02-29 ucrt) -- "Angel Food Cake"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
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.
>
Environment
History
Connections
Git
Tutorial
Import Dataset
130 MiB
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Type here to search
An initial screen for R Studio (console)

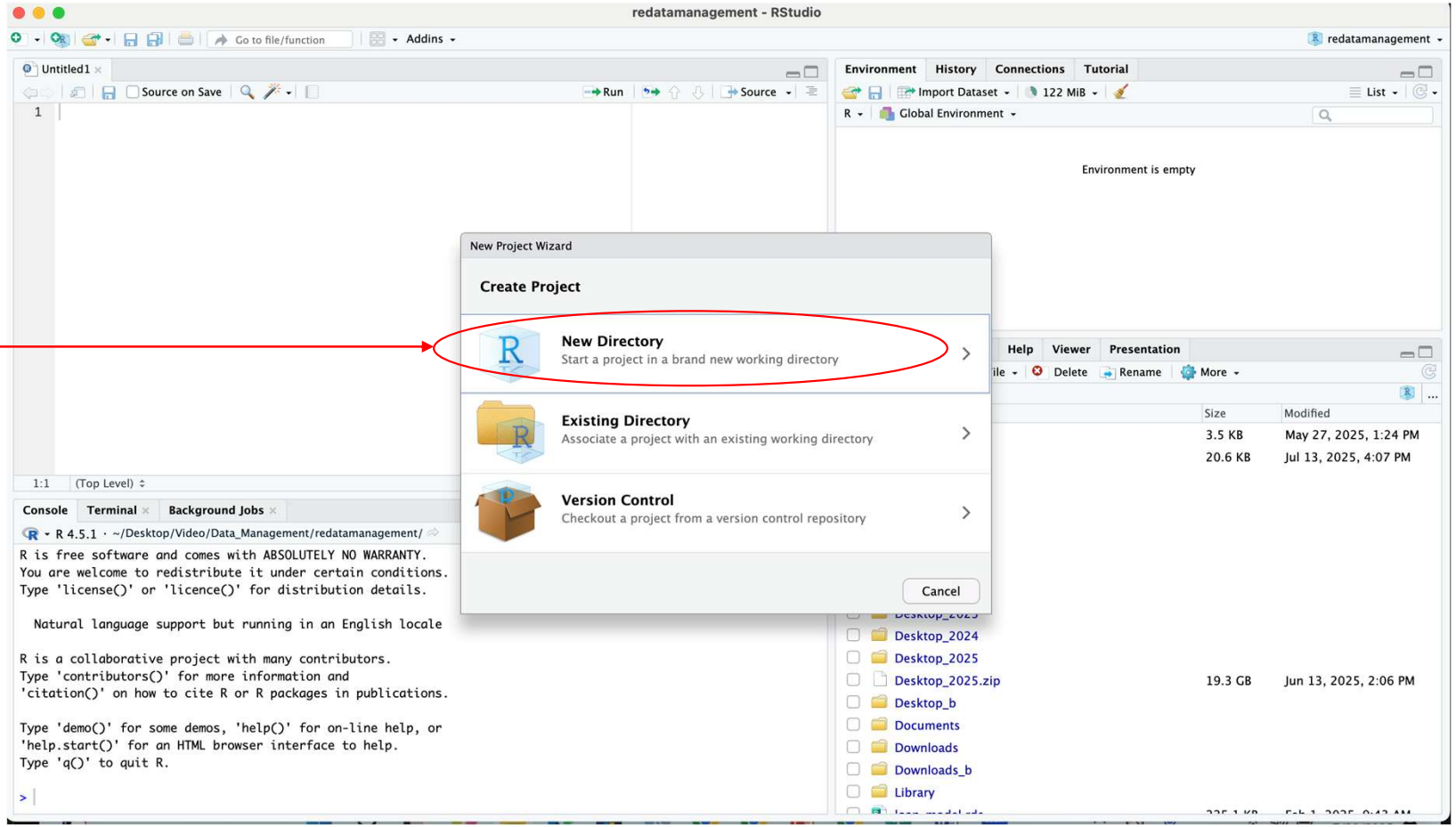
R projects

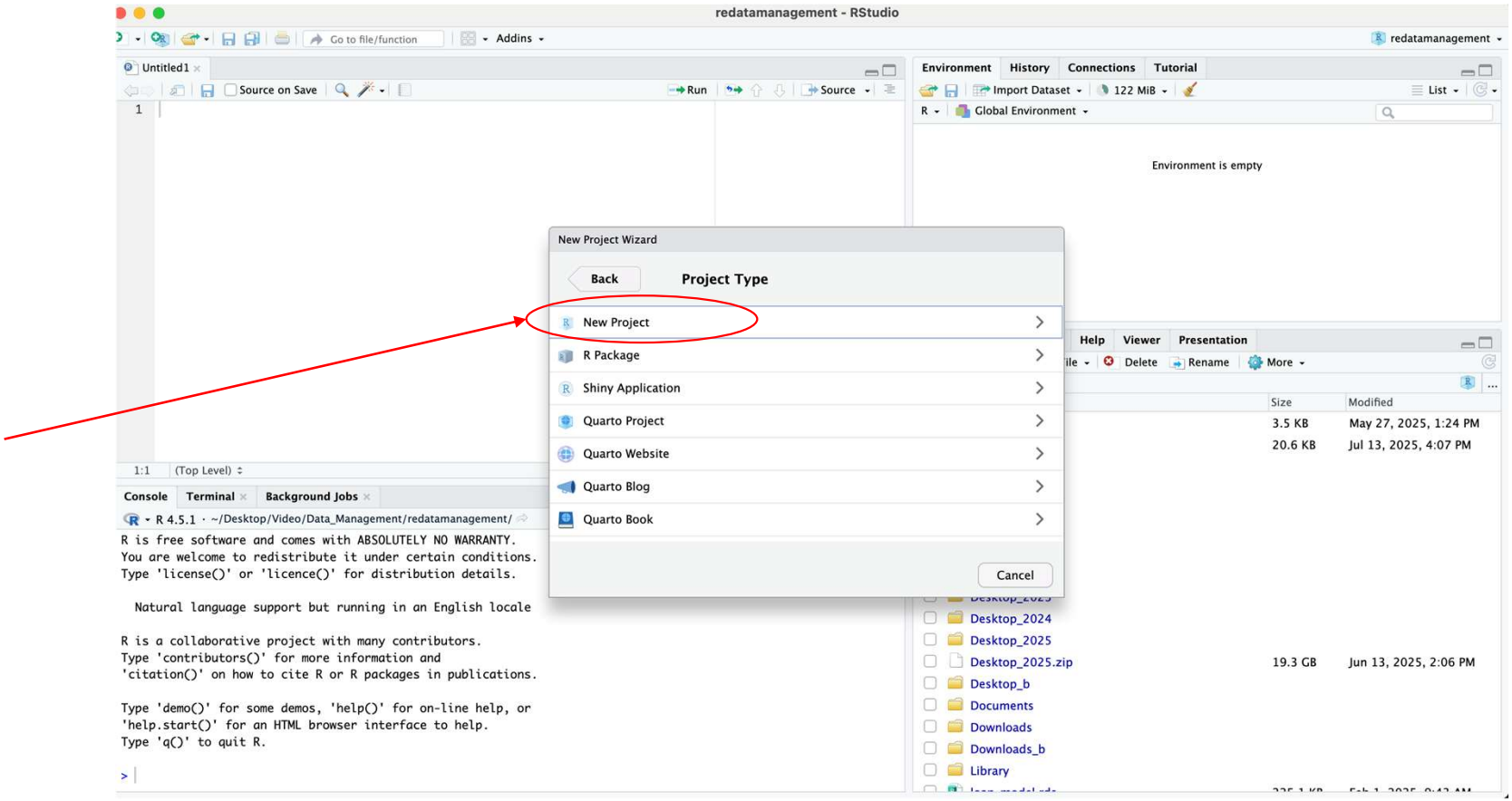
- An R project helps to organize the workflow.
- An R script stores R code for analysis.
- To create a project for this training.

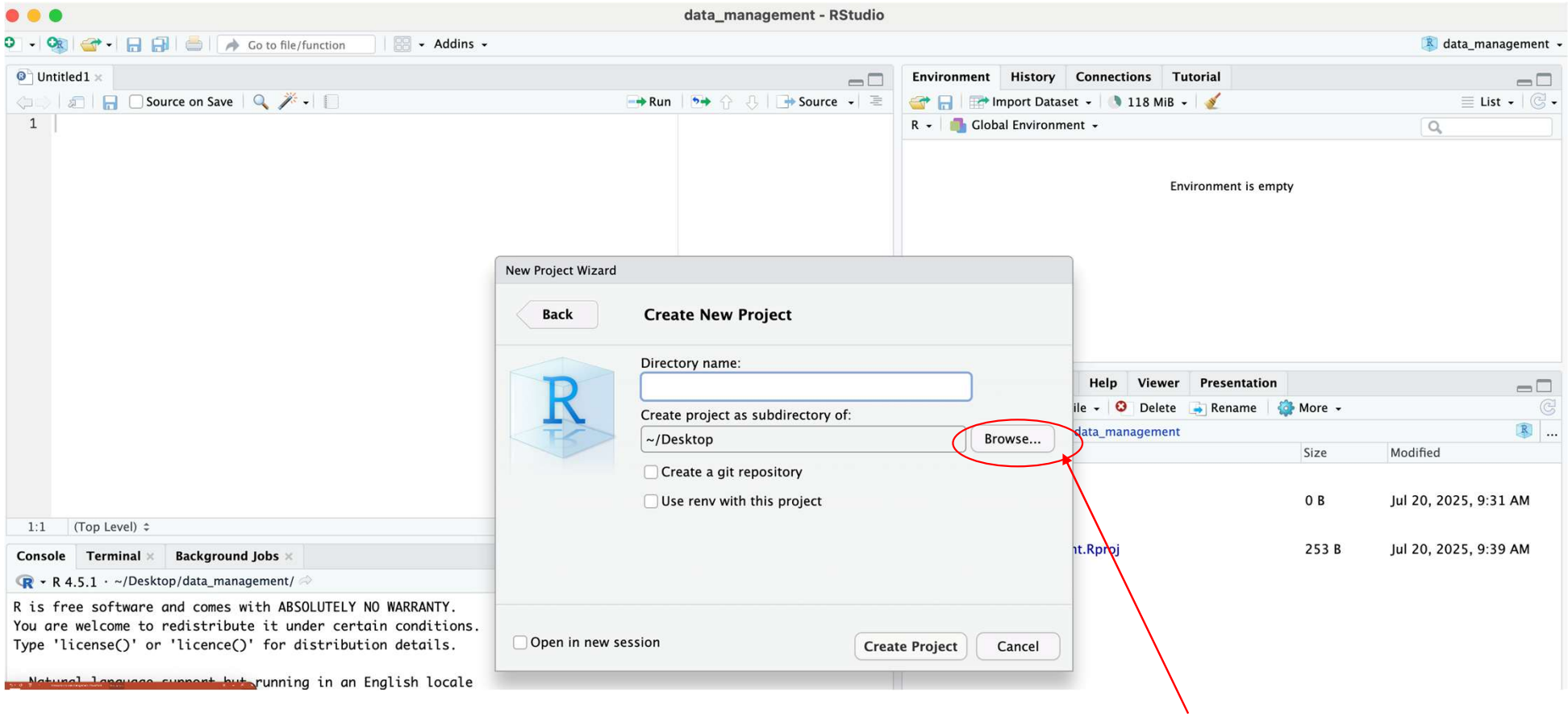
❖Exercise: Create an R project for this webinar series

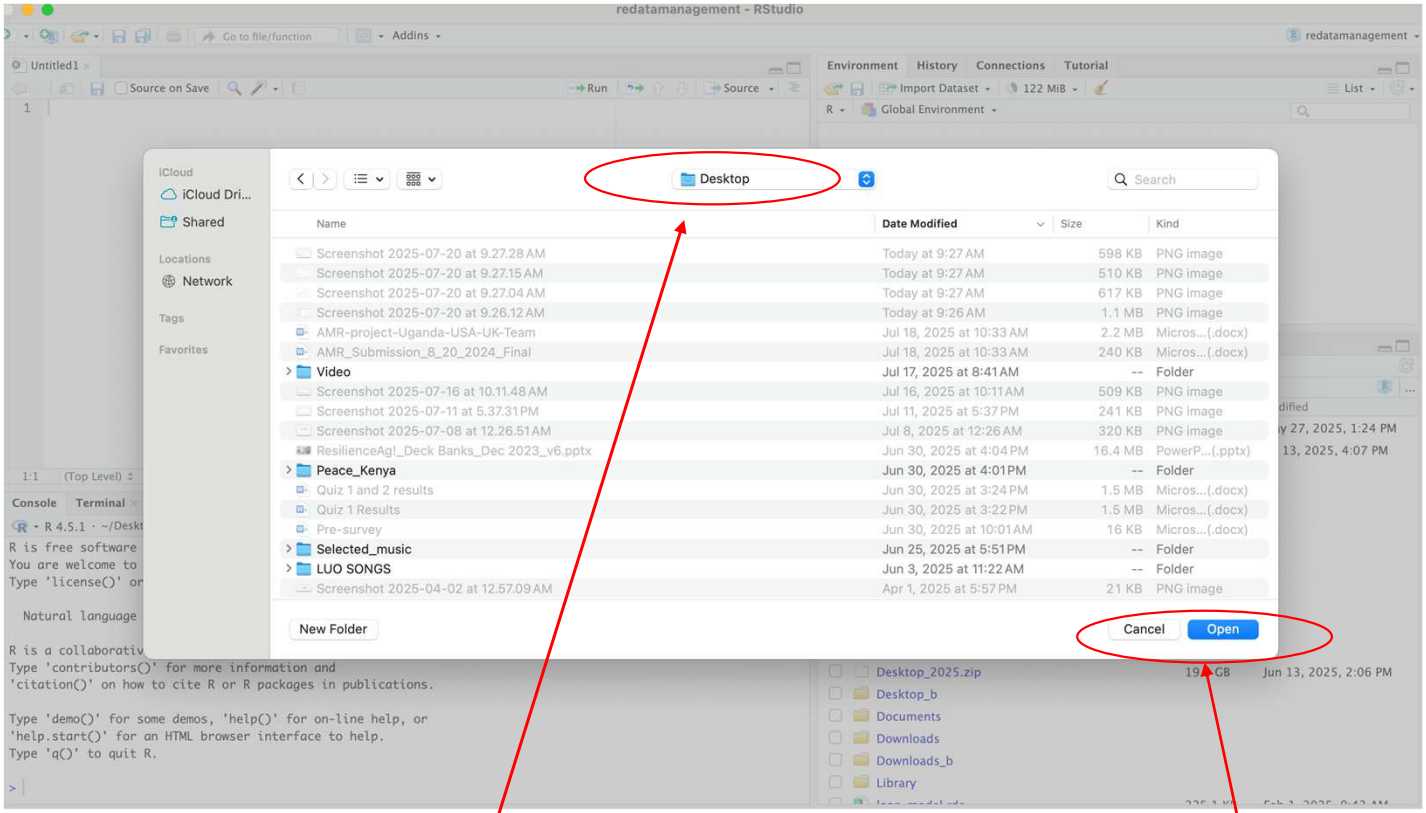
❖Demonstration using the R-Studio

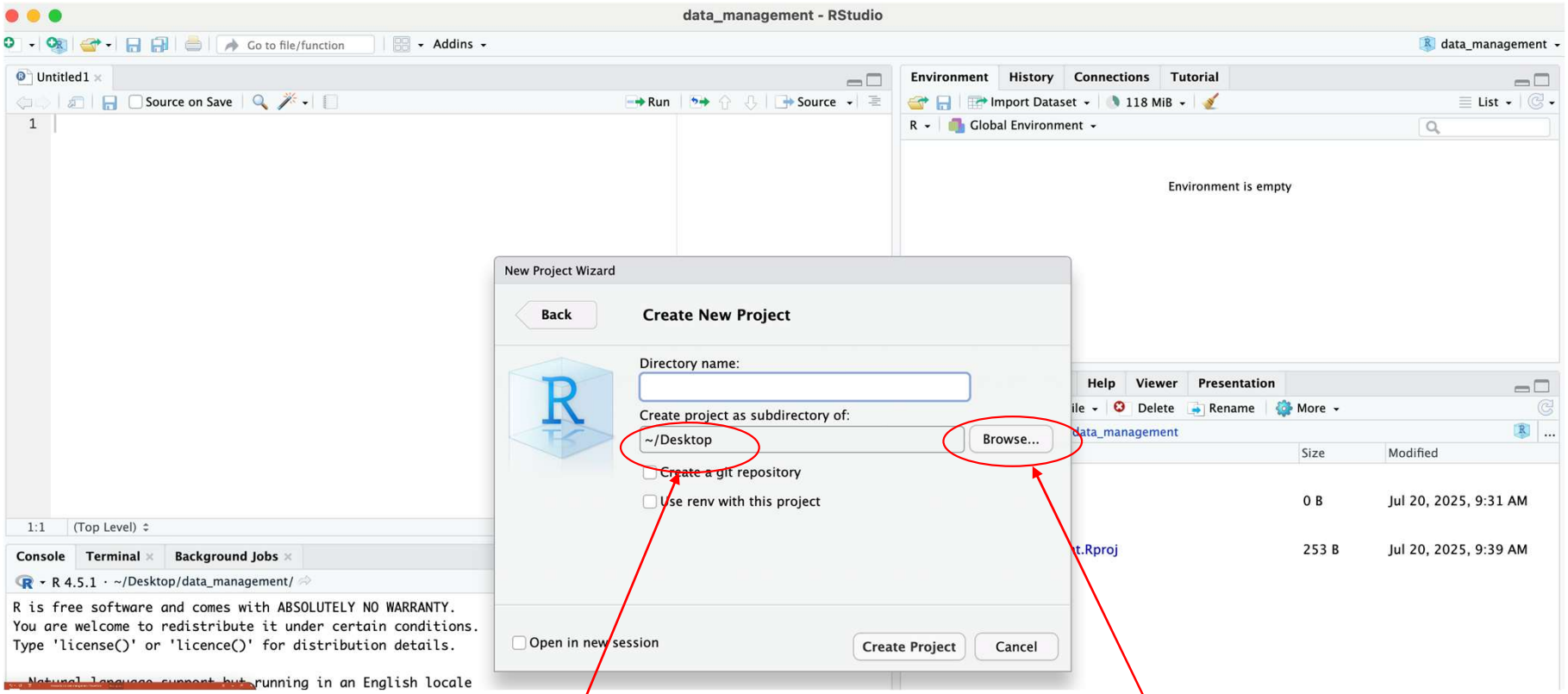

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
redatamanagement - RStudio
New File
New Project...
Open File... ⌘O
Open File in New Column...
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save ⌘S
Save As...
Save with Encoding...
Save All ⌥⌘S
Compile Report...
Print...
Close ⌘W
Close All ⇧⌘W
Close All Except Current ⌥⇧⌘W
Close Project
Quit Session...
Untitled1
Addins
Run
Source
R Script
Console
R 4.5.1
R version 4.5.1 (2025-06-13) -- "Great Square Root"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
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.
Environment History Connections Tutorial
Import Dataset 122 MiB
Global Environment
Environment is empty
Files Plots Packages Help Viewer Presentation
New Folder New File Delete Rename More
Home Desktop Video Data_Management redatamanagement
Name Size Modified
.Rhistory 664 B Jul 13, 2025, 12:26 AM
appendicitis_data.xlsx 207.8 KB Jul 12, 2025, 12:50 PM
area harvested per crop.csv 198.6 KB Jul 12, 2025, 12:50 PM
Country_CO.csv 1.9 MB Jul 15, 2025, 3:49 PM
Country_Coordinates_yield_data.csv 2.6 KB Jul 12, 2025, 10:54 PM
Country_KG.csv 1.9 MB Jul 18, 2025, 6:43 AM
Country_N_Crop.csv 1.6 MB Jul 16, 2025, 7:05 PM
Cover_letter-Opiyo_15_7_2025.docx 35.1 KB Jul 15, 2025, 1:11 PM
Cover_letter-Opiyo_15_7_2025.pdf 143.8 KB Jul 15, 2025, 1:11 PM
Crop.cpg 5 B Jul 13, 2025, 9:24 AM
Crop.csv 1.1 MB Jul 16, 2025, 7:04 PM
Crop.prj 145 B Jul 13, 2025, 9:24 AM
Dashboard_b.docx 50.5 KB Jul 16, 2025, 6:00 AM
Dashboard_c.docx 17.5 KB Jul 16, 2025, 7:43 AM
Dashboard_Latest.docx 56.3 KB Jul 17, 2025, 10:04 PM
Data_Management_Schedule.docx 16.9 KB Jul 14, 2025, 12:54 PM


redatamanagement - RStudio
Go to file/function
Addins
redatamanagement
Untitled1
Source on Save
Run
Source
Environment
History
Connections
Tutorial
Import Dataset
122 MiB
List
Global Environment
Environment is empty
New Project Wizard
Create Project
New Directory
Start a project in a brand new working directory
Existing Directory
Associate a project with an existing working directory
Version Control
Checkout a project from a version control repository
Cancel
Help
Viewer
Presentation
Delete
Rename
More
Size
Modified
3.5 KB
May 27, 2025, 1:24 PM
20.6 KB
Jul 13, 2025, 4:07 PM
Desktop_2024
Desktop_2025
Desktop_2025.zip
19.3 GB
Jun 13, 2025, 2:06 PM
Desktop_b
Documents
Downloads
Downloads_b
Library
1:1
(Top Level)
Console
Terminal
Background Jobs
R 4.5.1 · ~/Desktop/Video/Data_Management/redatamanagement/
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.


redatamanagement - RStudio
Go to file/function
Addins
redatamanagement
Untitled1
Source on Save
Run
Source
Environment
History
Connections
Tutorial
Import Dataset
122 MiB
List
R
Global Environment
Environment is empty
New Project Wizard
Back
Project Type
New Project
R Package
Shiny Application
Quarto Project
Quarto Website
Quarto Blog
Quarto Book
Cancel
Help
Viewer
Presentation
Delete
Rename
More
Size
Modified
3.5 KB
May 27, 2025, 1:24 PM
20.6 KB
Jul 13, 2025, 4:07 PM
1:1
(Top Level)
Console
Terminal
Background Jobs
R 4.5.1 · ~/Desktop/Video/Data_Management/redatamanagement/
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.
>
Desktop_2024
Desktop_2025
Desktop_2025.zip
19.3 GB
Jun 13, 2025, 2:06 PM
Desktop_b
Documents
Downloads
Downloads_b
Library

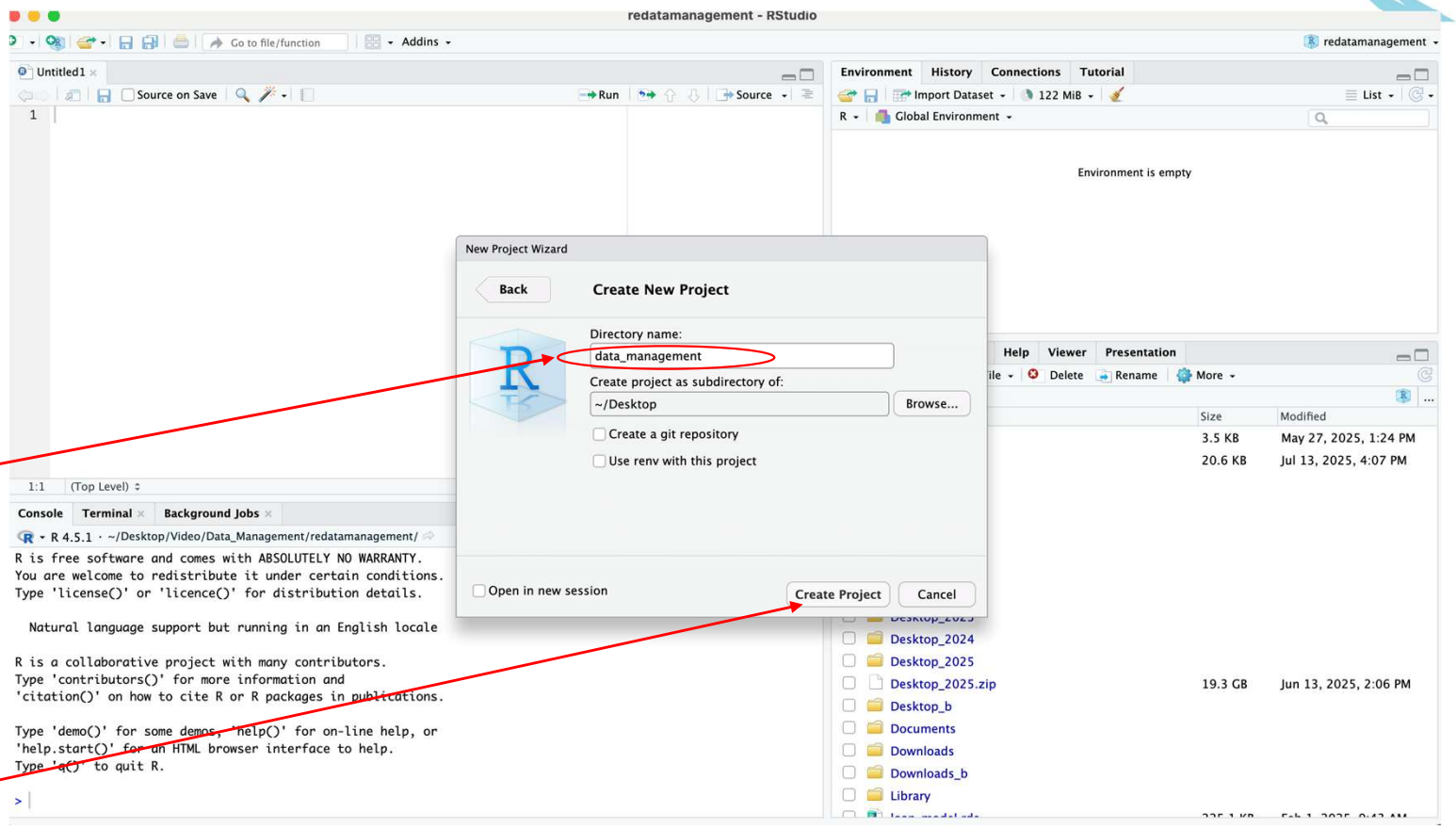

data_management - RStudio
Go to file/function
Addins
data_management
Untitled1
Source on Save
Run
Source
Environment
History
Connections
Tutorial
Import Dataset
118 MiB
List
R
Global Environment
Environment is empty
New Project Wizard
Back
Create New Project
Directory name:
Create project as subdirectory of:
~/Desktop
Browse...
Create a git repository
Use renv with this project
Open in new session
Create Project
Cancel
Help
Viewer
Presentation
Delete
Rename
More
data_management
Size
Modified
0 B
Jul 20, 2025, 9:31 AM
253 B
Jul 20, 2025, 9:39 AM
1:1
(Top Level)
Console
Terminal
Background Jobs
R 4.5.1 · ~/Desktop/data_management/
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.
running in an English locale


redatamanagement - RStudio
Desktop
Search
Name
Date Modified
Size
Kind
New Folder
Cancel
Open
1. Select Desktop
2. Open

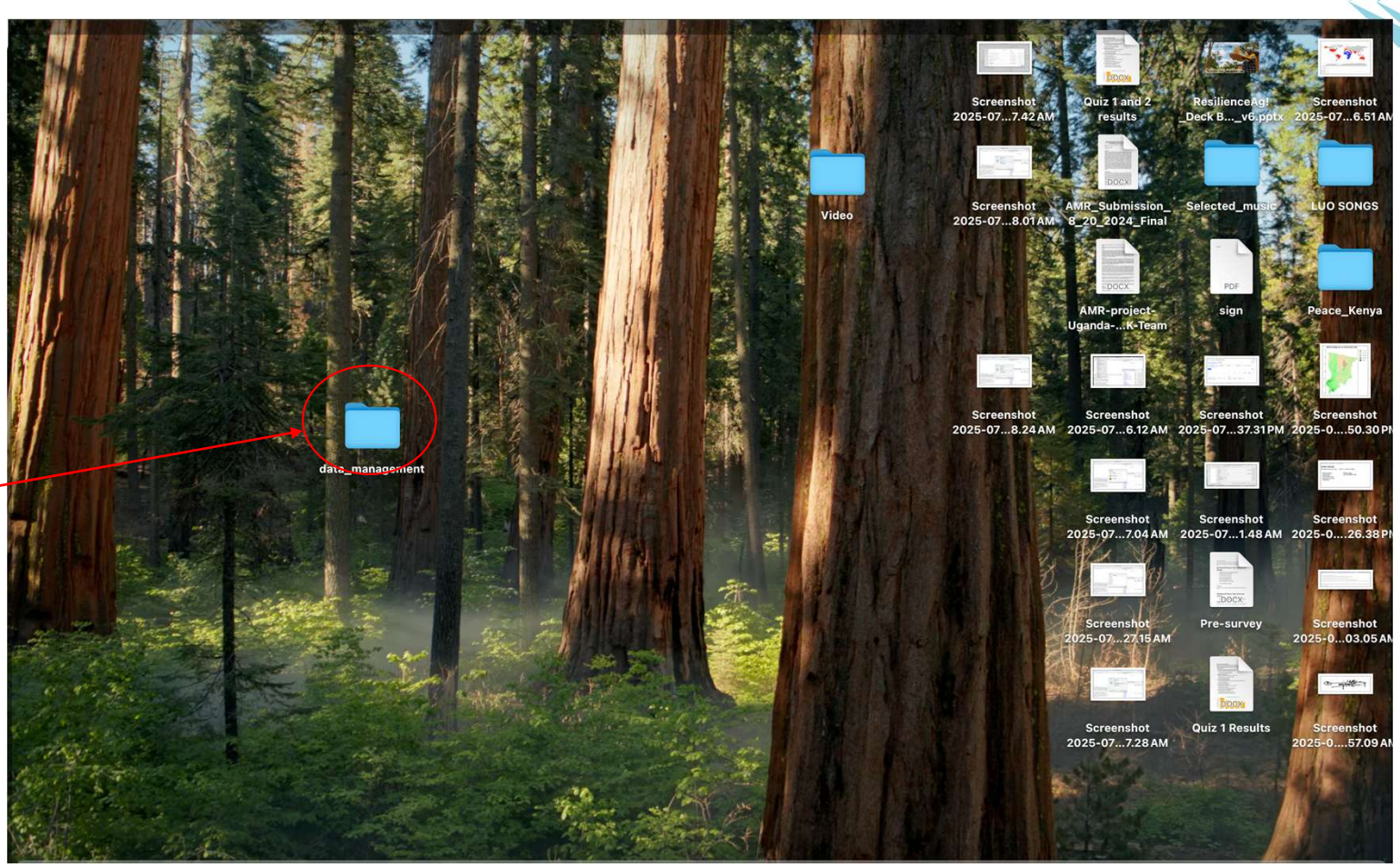

data_management - RStudio
Go to file/function
Addins
data_management
Untitled1
Source on Save
Run
Source
Environment
History
Connections
Tutorial
Import Dataset
118 MiB
List
R
Global Environment
Environment is empty
New Project Wizard
Back
Create New Project
Directory name:
Create project as subdirectory of:
~/Desktop
Browse...
Create a git repository
Use renv with this project
Open in new session
Create Project
Cancel
Help
Viewer
Presentation
Delete
Rename
More
data_management
Size
Modified
0 B
Jul 20, 2025, 9:31 AM
253 B
Jul 20, 2025, 9:39 AM
1:1
(Top Level)
Console
Terminal
Background Jobs
R 4.5.1 · ~/Desktop/data_management/
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.
running in an English locale
Select Desktop
Click on Browse

1. Type data_management

2. Click on Create Project

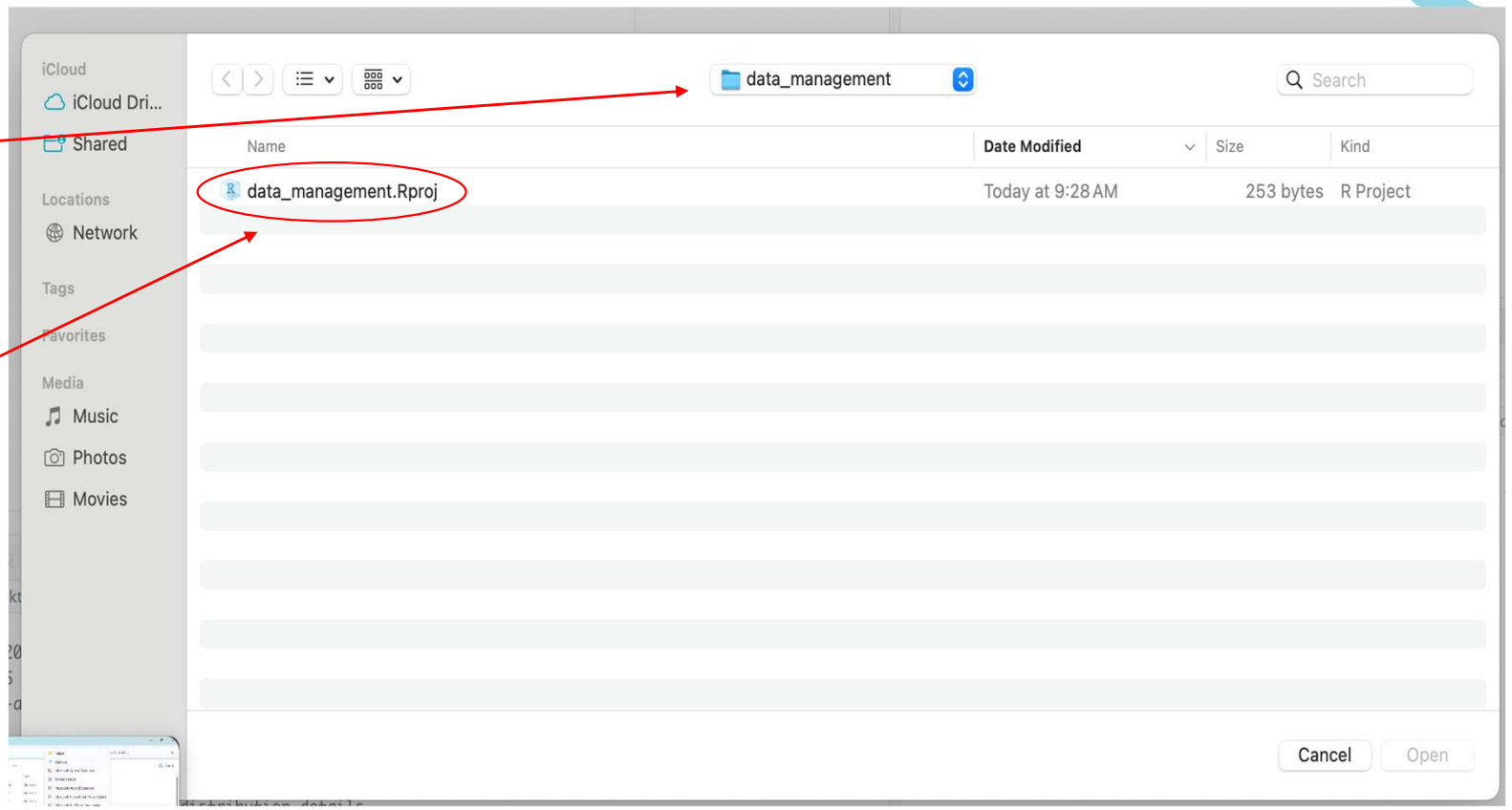


data_management folder on a desktop



Open
data_management folder

data_management



R-Syntax

- Language that R uses to execute tasks such as data management and analysis
- R syntax is case sensitive: i.e data is not the same as Data
- R syntax is stored in R scripts
- Comments can be added to scripts for better explanations
- Comments are added using the # symbol e.g

  #This is my first comment

R objects

- R objects are used to store data within R.
- R objects are identified by unique names, e.g, data, mydata
- Object names should not contain spaces e.g, use my_data instead of my data
- Objects are saved into R using the assignment operator: <-

  Example: data <- a

The above will create an object named data that contains the letter a.

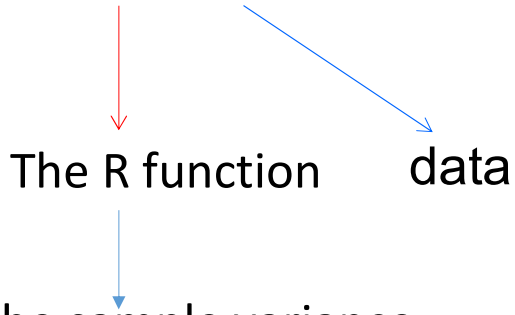
R functions

- A procedure that was programmed in R that uses data to produce output.
- R functions perform specific tasks

  Example:

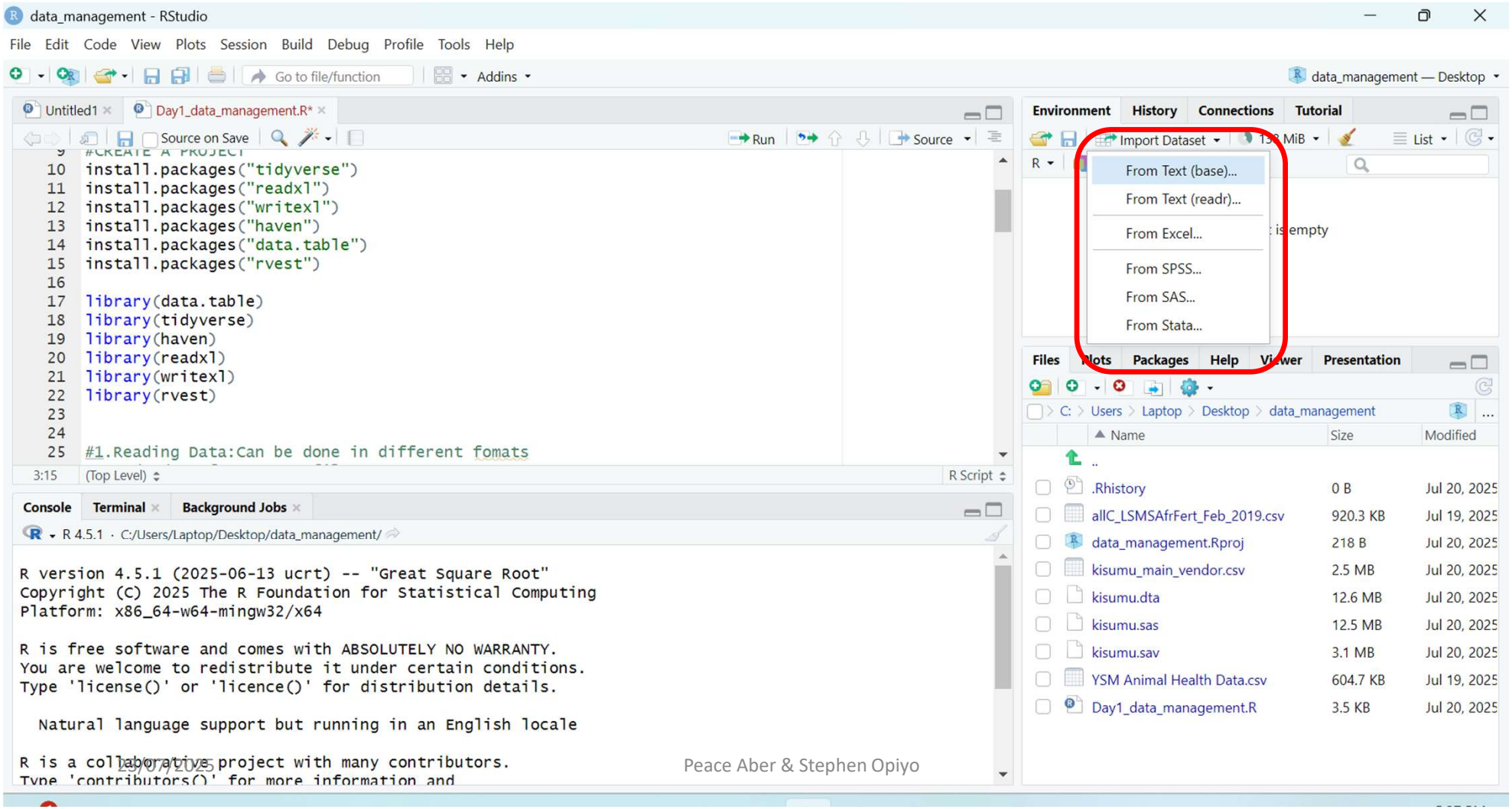
```
function(data)
```

```
> var(x)
```

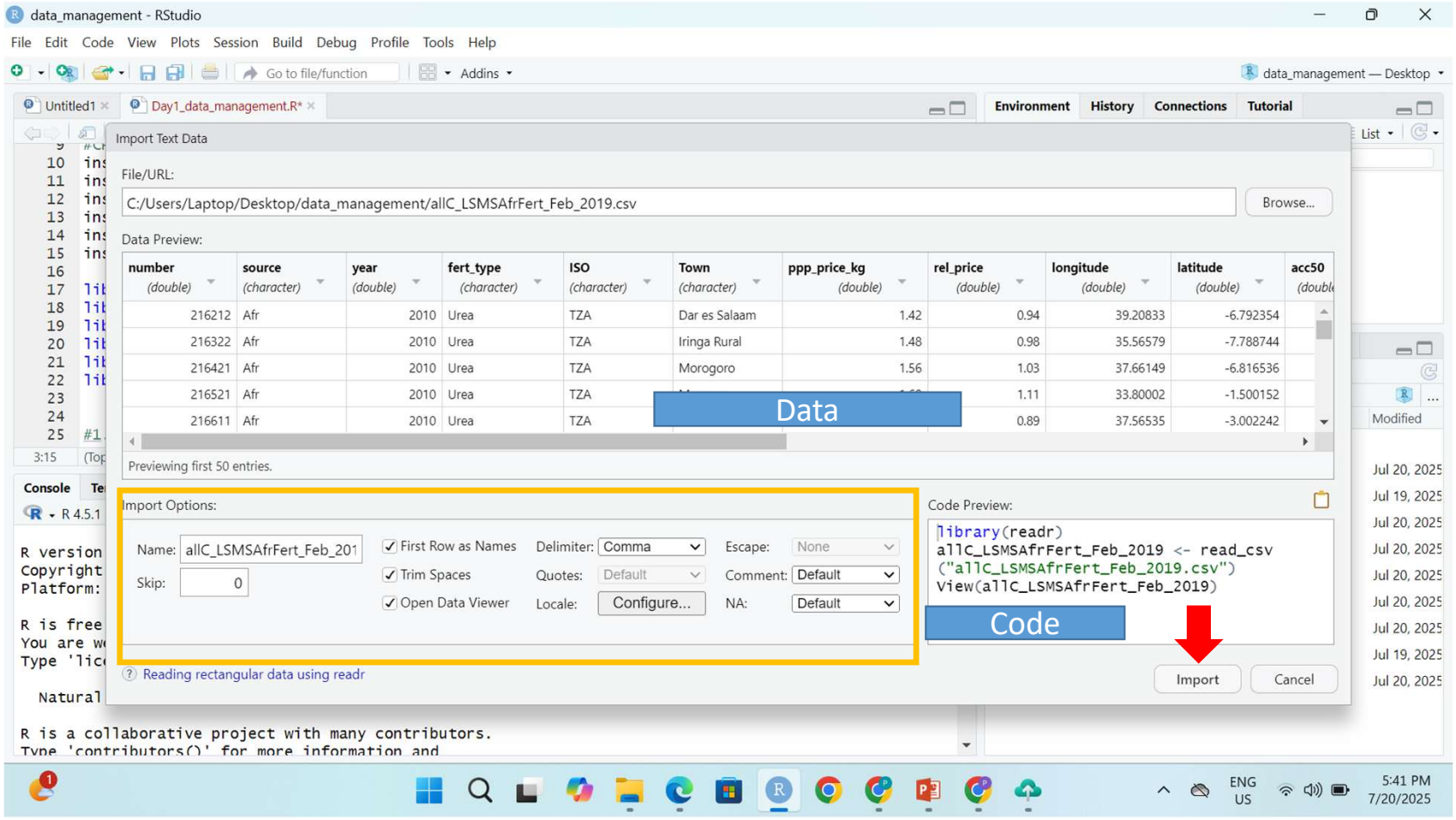


Calculates the sample variance.

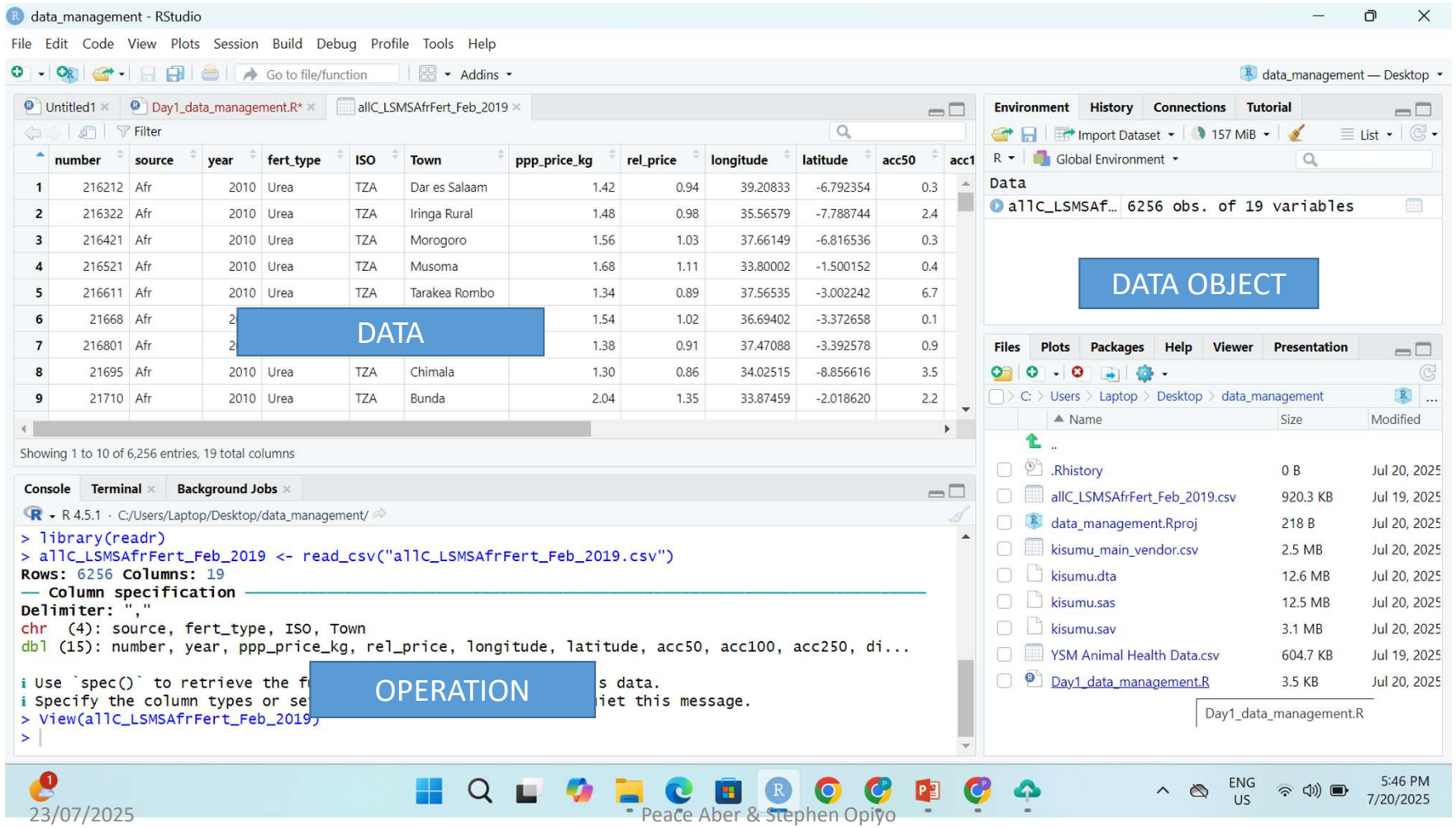
Load data in RStudio



Load data into R



Load data into R



R packages

- Packages are groups of functions that perform specific tasks in R, such as data management and data analysis.
- Packages are installed whenever they are needed.
- To install a package in R, use the **install.packages()** function as follows:

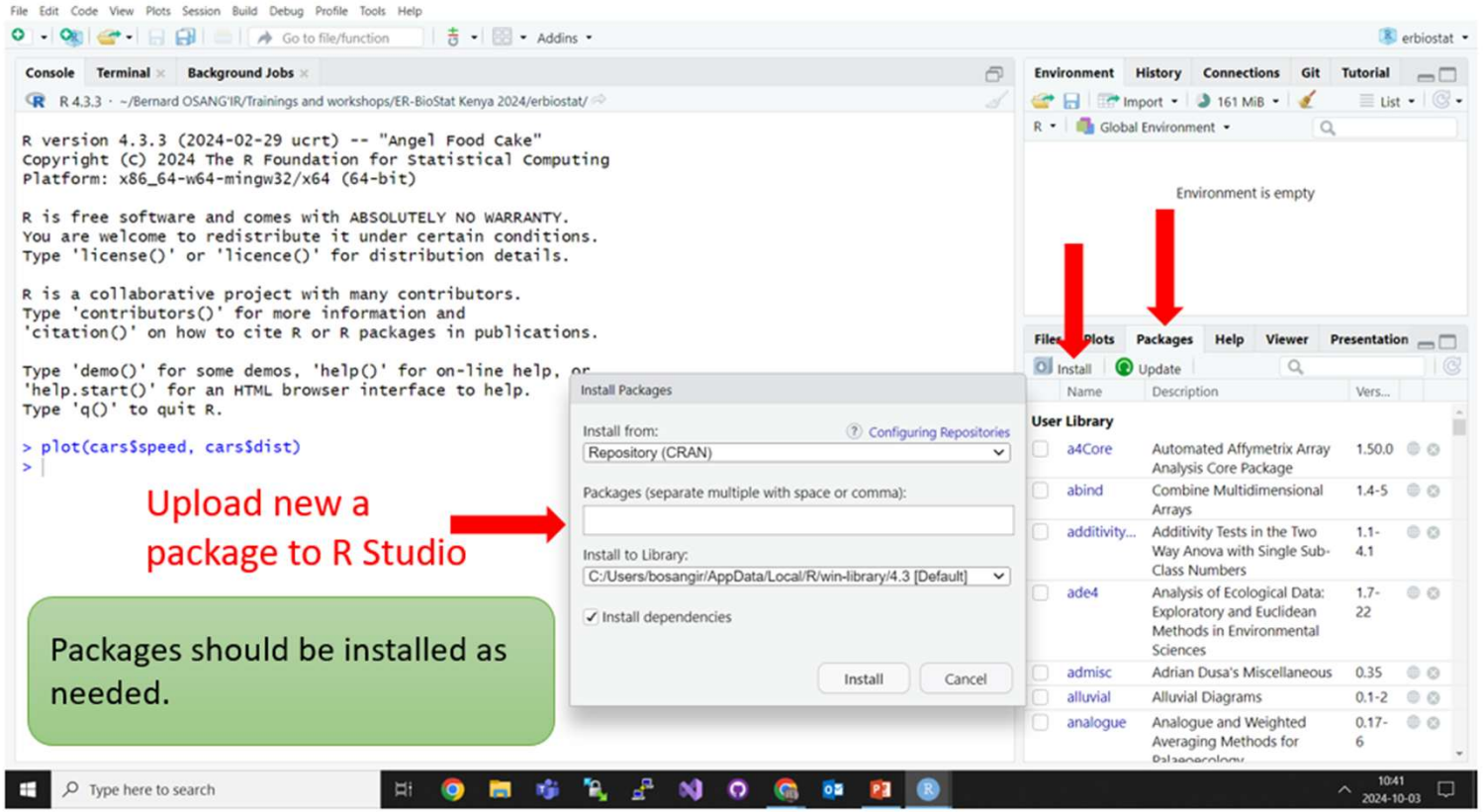
```
install.packages ("tidyverse")
```

- Packages are installed once in Rstudio, but the library must be loaded in every session.

To load a package, use the library() function as follows.

```
library(tidyverse)
```

R studio: packages



Part 1

Importing data into R

Importing data into R

- R works with various file formats of data e.g CSV, EXCEL, STATA, SAS, SPSS, TEXT.
- R can import data from the web.
- One does not need the software used to create the file in order to use the dataset.
- Different functions import different file formats.

❖Demonstration using R script

Data structure inspection

- Data inspection involves using functions that help to understand the data:

➢View data

➢Checking the data structure

➢Check the dimensions of the data

➢View snapshot of the data, etc…

❖ Demonstration using R script

Filtering rows in data sets

filter() selects the rows of a data frame that meet a column citeria.

Example: Create a new dataset containing male vendors

```
Male <- filter(kisumu_main_vendor, gender == "Male")
```

Example2: Create a new dataset containing animals aged less than 12.

```
data <- filter(Animal_Data, age < 12)
```

Cols: « ‹ 1 - 50 › »

ward	slum	cu	location_type	vendor_type	gender
Manyatta	Manyatta B	4	Peri-Urban	Stalls/Tabletop	Male
Manyatta	Manyatta B	3	Peri-Urban	Stalls/Tabletop	Male
Manyatta	Manyatta B	3	Peri-Urban	Stalls/Tabletop	Male
Manyatta	Manyatta B	3	Peri-Urban	Kiosk	Male
Manyatta	Manyatta B	4	Peri-Urban	Kiosk	Male
Manyatta	Manyatta B	3	Peri-Urban	Kiosk	Male
Manyatta	Manyatta B	3	Peri-Urban	Kiosk	Male
Kondele	Manyatta A	NA	Peri-Urban	Kiosk	Male
Manyatta	Manyatta B	NA	Peri-Urban	Mordern restaurant/ Fast food outlet	Male
Manyatta	Manyatta B	3	Peri-Urban	Kiosk	Male
Manyatta	Manyatta B	3	Peri-Urban	Kiosk	Male

abdomenauscultation_notes	age_units	age_weeks	age
NA	months	16.0	4.0
NA	months	20.0	5.0
NA	days	1.0	8.0
NA	months	4.0	1.0
NA	weeks	3.0	3.0
NA	days	0.5	3.0
NA	months	16.0	4.0
NA	weeks	2.0	2.0
NA	months	4.0	1.0
NA	months	20.0	5.0
NA	months	8.0	2.0

arrange

arrange() orders the rows of a data frame by the values of selected columns.

arrange() sorts values in ascending order or descending order.

Example: Sort the age in ascending order

```
data1 <- arrange(Animal_Data, age)
```

Sort age in ascending order

```
data2 <- arrange(Animal_Data, (-age) )
```

abdomenauscultation_notes	age_units	age_weeks	age
NA	months	0.12	0.03
NA	months	0.40	0.10
NA	months	1.00	0.25
NA	months	1.00	0.25
NA	months	1.00	0.25
NA	months	1.00	0.25
NA	months	1.00	0.25
NA	months	1.00	0.25
NA	months	1.00	0.25
NA	months	1.00	0.25

abdomenauscultation_notes	age_units	age_weeks	age	anaplasma_ce
NA	days	4.000000	28	
NA	weeks	24.000000	24	
NA	weeks	24.000000	24	
NA	weeks	24.000000	24	
NA	weeks	24.000000	24	
NA	months	96.000000	24	
NA	months	96.000000	24	
NA	months	96.000000	24	
NA	months	96.000000	24	
NA	days	3.000000	21	
NA	days	3.000000	21	

Select columns

select() is used to select columns that you want to **retain**.

Example: Select age, age_weeks,age_units, then store them.

```
data3 <- select(Animal_Data,age,age_weeks,age_units)
#deselect/remove columns from the data set
data5 <- select(kisumu_main_vendor,-subcounty)
```

Filter

	age	age_weeks	age_units
1	4.0	16.0	months
2	5.0	20.0	months
3	8.0	1.0	days
4	1.0	4.0	months
5	3.0	3.0	weeks
6	3.0	0.5	days
7	4.0	16.0	months
8	2.0	2.0	weeks
9	1.0	4.0	months
10	5.0	20.0	months
11	2.0	8.0	months
12	1.5	6.0	months

mutate

mutate() creates new variables in the data set

It adds the new variable to existing data
#Example: creates a new variable "new_age" by squaring the "age" variable
new_age <- mutate(Animal_Data, new_age = age^2)

weaned_hay_rainyseason	wheezes	year	new_age
NA	no	1	16.00
NA	no	1	25.00
NA	no	1	64.00
NA	no	1	1.00
NA	no	1	9.00
NA	no	1	9.00
NA	no	1	16.00
NA	no	1	4.00
NA	no	1	1.00
NA	no	1	25.00
NA	no	1	4.00

##mutate
logage <- mutate(Animal_Data, logage = log(age))

logage | data3 | kisumu1

ed_hay_dryseason	weaned_hay_rainyseason	wheezes	year	logage
	NA	no	1	1.3862944
	NA	no	1	1.6094379
	NA	no	1	2.0794415
	NA	no	1	0.0000000
	NA	no	1	1.0986123
	NA	no	1	1.0986123
	NA	no	1	1.3862944
	NA	no	1	0.6931472
	NA	no	1	0.0000000
	NA	no	1	1.6094379
	NA	no	1	0.6931472

rename

rename() creates new names for variables in the data

Example: rename region to location

```
data7 <- rename(kisumu1, location = region)
```

survey_date	survey_time	consent	region	country	county	subcounty	ward
2/22/2022	09:53:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man
2/22/2022	10:09:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man
2/22/2022	10:30:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man
2/22/2022	10:47:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man
2/22/2022	11:04:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man
2/22/2022	11:42:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man
2/22/2022	11:54:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man
2/22/2022	12:05:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man
2/22/2022	12:20:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man
2/22/2022	12:37:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East	Man

survey_date	survey_time	consent	location	country	county	subcounty
2/22/2022	09:53:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	10:09:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	10:30:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	10:47:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	11:04:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	11:42:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	11:54:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	12:05:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	12:20:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	12:37:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East
2/22/2022	12:51:00.000+03:00	1	East Africa	Kenya	Kisumu	Kisumu East

relocate

relocate() gives a new order to column names

Example: relocate the slum, ward and subcounty from their original position

kisumu1 <- relocate(kisumu1, slum, ward,subcounty)

Filter Cols: « < 1 - 50 > »

	slum	ward	subcounty	start_time	end_time	survey_date	survey_time	consent	region	country
1	Manyatta B	Manyatta	Kisumu East	52:46:00	08:38:00	2/22/2022	09:53:00.000+03:00	1	East Africa	Kenya
2	Manyatta B	Manyatta	Kisumu East	08:43:00	18:58:00	2/22/2022	10:09:00.000+03:00	1	East Africa	Kenya
3	Manyatta B	Manyatta	Kisumu East	30:04:00	27:52:00	2/22/2022	10:30:00.000+03:00	1	East Africa	Kenya
4	Manyatta B	Manyatta	Kisumu East	47:24:00	29:48:00	2/22/2022	10:47:00.000+03:00	1	East Africa	Kenya
5	Manyatta B	Manyatta	Kisumu East	03:59:00	10:24:00	2/22/2022	11:04:00.000+03:00	1	East Africa	Kenya
6	Manyatta B	Manyatta	Kisumu East	41:52:00	31:26:00	2/22/2022	11:42:00.000+03:00	1	East Africa	Kenya
7	Manyatta B	Manyatta	Kisumu East	51:48:00	57:51:00	2/22/2022	11:54:00.000+03:00	1	East Africa	Kenya
8	Manyatta B	Manyatta	Kisumu East	05:47:00	16:40:00	2/22/2022	12:05:00.000+03:00	1	East Africa	Kenya
9	Manyatta B	Manyatta	Kisumu East	19:37:00	31:05:00	2/22/2022	12:20:00.000+03:00	1	East Africa	Kenya
10	Manyatta B	Manyatta	Kisumu East	37:49:00	48:48:00	2/22/2022	12:37:00.000+03:00	1	East Africa	Kenya
11	Manyatta B	Manyatta	Kisumu East	51:30:00	57:52:00	2/22/2022	12:51:00.000+03:00	1	East Africa	Kenya

Exporting/writing data out R

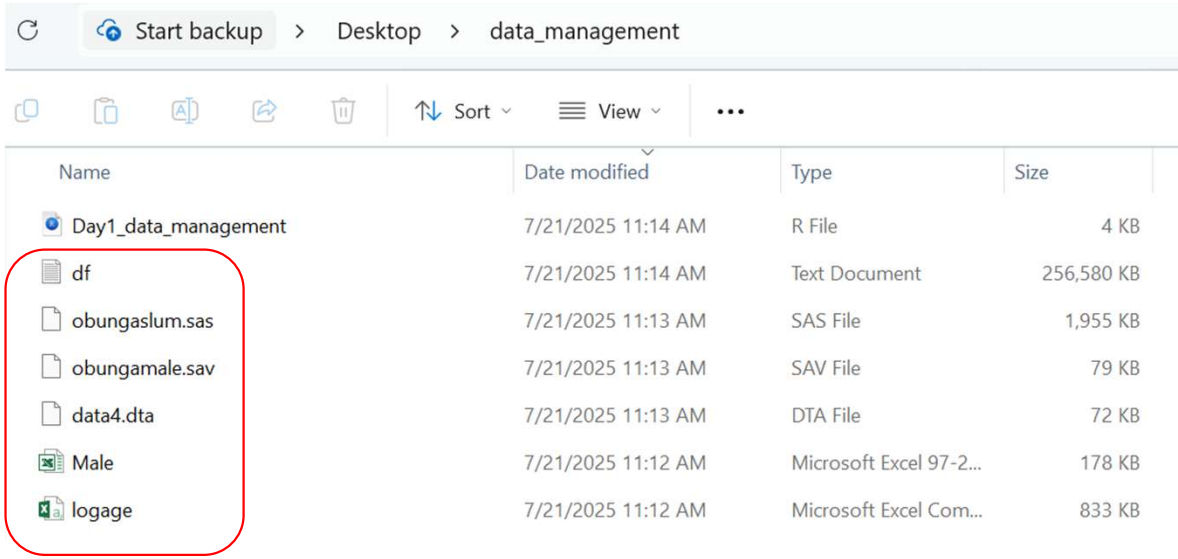
Data sets in R can be exported out of RStudio to different file formats, ready for sharing.

R supports data exportation in a wide variety of file formats below:

- Comma-separated values
- Excel files
- SPSS files
- SAS files
- Text files

R uses the write() function to export data to other formats as below:

```
write_csv(logage,"logage.csv")
write_xlsx(Male,"Male.xls")
write_dta(data4,"data4.dta")
write_sav(obungamale,"obungamale.sav")
write_xpt(obungaslum, "obungaslum.sas")
write.table(df,"df.txt")
```



Data summary and missingness

- Data summary gives a comprehensive understanding of large datasets and variables.
- Categorical data can be summarized using frequency distribution tables.
- Frequency distribution tables are obtained using the table function: table()
- Missingness can be detected using the table function: table()
- Descriptive summaries are obtained using the summary function: summary()

❖Demonstration in RStudio

Distribution of data

- Understanding of distribution of numeric data is key.

- It involves testing for normality of key variables

- Distributions can be assessed in two ways:

➢ Visually using graphs such as the histogram

➢ Performing a statistical test of normality such as Shapiro's wilk test.

Histogram and test for normality

- Histograms are generated using the hist() function:

  hist(Animal_Data$age)

- The Shapiro's test for normality is performed using the shapiro.test()

Example: shapiro.test(Animal_Data$age)

```
        Shapiro-Wilk normality test

data:  Animal_Data$dam_milkyield_liters
W = 0.76177, p-value < 2.2e-16
```

P-value

❖Demonstration in RStudio

Outlier detection

- Outliers are values that are considered out of range in a given numerical variable.

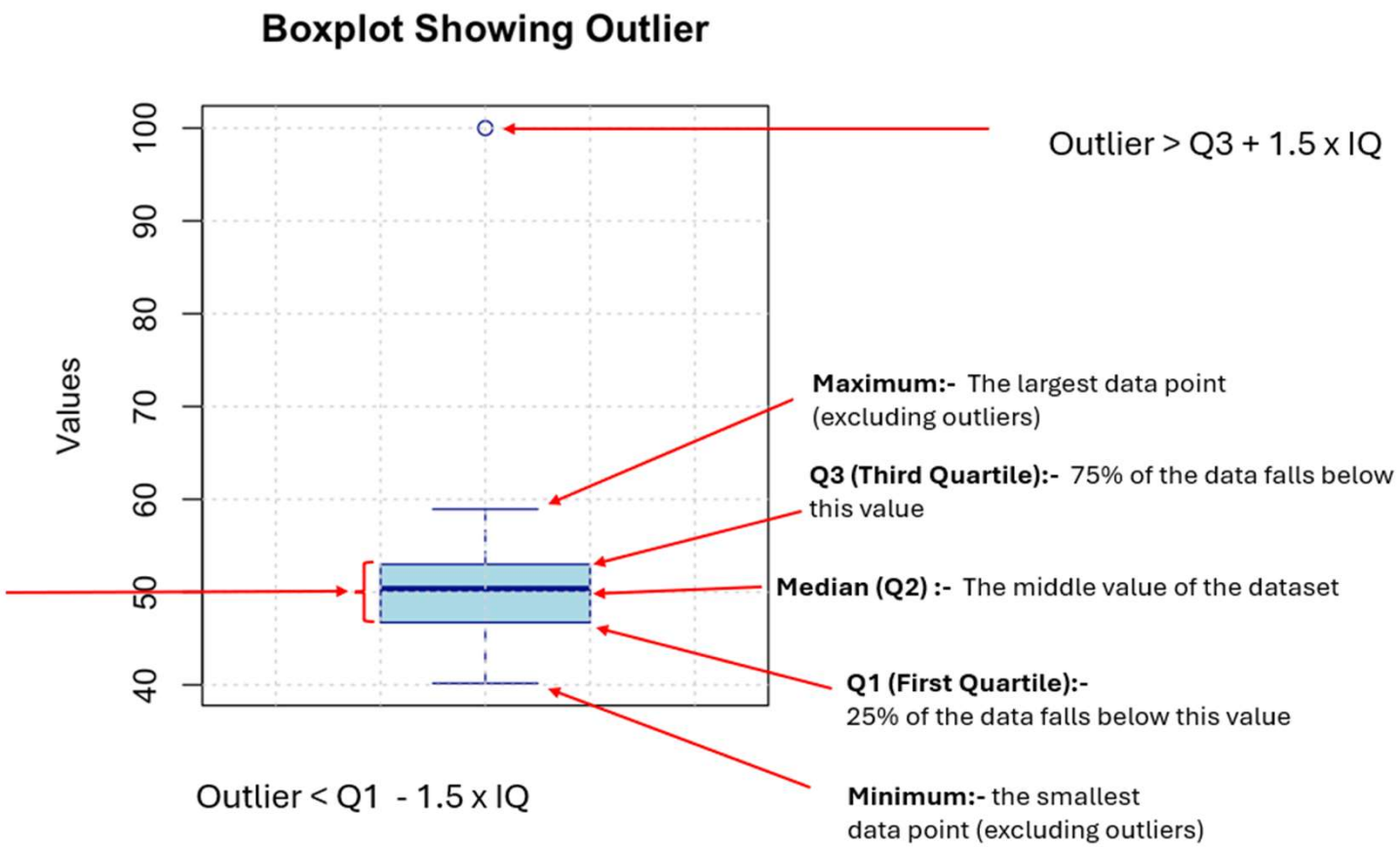
- Outliers affect data modelling and can cause drastic changes in results.

- Outliers can be quickly detected visually using box plots.

Box plots

- Box plots are figures that contain important information about the summary of numeric variable.
- The box represents the interquartile range (IQR), from Q1 (25th percentile) to Q3 (75th percentile).
- The line inside the box is the median.
- Whiskers extend up to 1.5 × IQR from Q1 and Q3.
- Points beyond the whiskers are considered outliers.
- Minimum:- the smallest data point (excluding outliers)
- Maximum:- The largest data point (excluding outliers)

Box plots showing outlier



Box plot demonstration in RStudio

Activity

https://docs.google.com/forms/d/e/1FAIpQLSeu8TKnEMO0UjdiZNsFLIXI4Mz8AhbCahJ72nc2Ij9Z21zF4w/viewform?usp=dialog

Link to WhatsApp group

https://chat.whatsapp.com/GO9nLQK0zD25OMany4I01I?mode=r_c