Data Science for Research and Institutional Intelligence

Topics to be covered

- General overview of data science and analytics
- Descriptive analytics
- Diagnostics analytics
- Predictive and prescriptive analytics
- Machine Learning + QGIS Integration
- Model Deployment using R Shiny

Pre-Webinar Survey

- Take a pre-webinar survey

https://docs.google.com/forms/d/e/1FAIpQLScg_N2_jee-z7fYjbnlw8r_XexIpsuKMJrQujsHJG3a9J8wSg/viewform?usp=dialog

Installation of software

https://drive.google.com/drive/folders/1OKihGHcjUbJAFlvviZwGD4EciGt8pBum?usp=share_link

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

What is Data Science?

- **Data Science** is **a data-driven, interdisciplinary field** focused on extracting **knowledge and insights** from **data.**

- At its core is **analytics,** powers this process by transforming raw data into actionable insights.

  - Descriptive analytics
  - Diagnostics analytics
  - Predictive analytics
  - Prescriptive analytics/Generative artificial Intelligence (AI)

What is Data Science?

- To achieve this, **data science** draws on **multiple disciplines**, including:
  - **Statistics**, for understanding data patterns and distributions
  - **Computer Science**, to manage, process, and scale data workflows.
  - **Mathematics**, to model relationships and optimize outcomes.
  - **Artificial Intelligence & Machine Learning**, to automate insight generation and prediction.
  - **Domain Expertise**, to contextualize data within specific fields like agriculture, health, finance, and more.

Requirements for Data Science

- Strong foundation in quantitative skills (e.g., statistics, mathematics, econometrics, biometry, etc)
- Proficiency in tools and programming languages (e.g., R, RStudio, Python)
- Ability to work with domain experts for contextual understanding
- Problem-solving and analytical thinking
- Data manipulation, visualization, and modeling skills

Tools for this webinar

- R and R-Studio
- QGIS
- R Shiny

Analytics

What is Analytics?

- Analytics allows us to derive value from data by answering four key questions:
  - What happened?
  - Why did it happen?
  - What's is happening now/What might happen?
  - What might happen in the future?
- In other words, analytics is a process by which you take raw material (data) and convert it into relevant insights that can inform organization, improve performance and guide strategy.

Benefits of Analytics

- Transforms raw data into **actionable insights**
- **Helps organizations:**
  - Optimize resources
  - Reduce inefficiencies
  - Improve performance
  - Forecast future outcomes

Descriptive Analytics – What Happened?

- Descriptive analytics summarizes historical data to understand what has happened over time.
- In agriculture, this could involve tracking yield trends, average soil nutrient levels, rainfall patterns, or pest incidences across seasons.
- These summaries help stakeholders identify patterns and assess performance.

Examples

- Monitor historical yield performance
- Summarize rainfall by region
- Report fertilizer or pesticide usage

Crop recommendation data

1. chickpea
2. coffee
3. cotton
4. maize
5. potato
6. rice
7. sugarcane
8. wheat

Crop	Nitrogen (N)	Phophorous (P)	Patassium (K)	temperature	humidity	pH	rainfall	yield_kg_ha	yield_category
maize	75	32.3	130.6	25.7	69.5	6.84	171.5	6062	High
cotton	78.9	48.8	105.7	28.9	33.3	6	63.6	1789	High
rice	101.3	58.8	142.5	31.3	74.9	5.69	156.6	5697	High
coffee	108.8	26.6	100.9	25.5	89.3	5.79	126.4	1579	Low
cotton	70.5	32.2	90.8	23.8	32.8	6.83	92.4	1930	High
coffee	98.7	34.9	102.2	25.7	88.9	6.56	144.4	1675	High
chickpea	43.6	26.9	43.7	26.9	35.6	6.17	67.4	1726	Low
maize	115	35.4	87.4	18.5	73.3	6.48	198.3	6558	High
cotton	50.7	43.9	96.4	27.7	32.8	7.28	71.6	1750	High
wheat	97	25.1	93.9	18.9	54	6.78	128.8	4465	Low
chickpea	48.7	39.2	79.9	21	42.2	6.27	86	1780	High
maize	121.1	40.1	134.6	25	75.5	6.18	139.9	6004	Low
rice	99	58.1	100.4	28.2	76.4	6.35	174.7	5788	High
sugarcane	127.2	34.9	185.2	31.1	84.6	6.79	279.5	71783	Low
maize	73.5	36	112.4	22.1	75.3	6.47	172.2	7024	High
rice	86.5	46.9	99.7	29.9	88.2	6.12	166.2	5423	Low
sugarcane	130.6	32.6	100.1	24.5	86.8	6.97	147.5	79779	High
coffee	94.5	24.9	110.1	21	78.6	5.41	180	1579	Low
rice	119.7	42.3	134.9	30	80.2	6.17	175	5422	Low
cotton	91	24.5	127.3	24.5	36	7.22	113	1746	High
rice	106.6	29.1	118.8	22.8	90.7	5.72	263.7	4802	Low
cotton	106.1	34	73.8	27.3	41.8	5.76	53	1868	High
sugarcane	129.8	34.5	183.3	24.1	72.7	6.83	200.1	84071	High
sugarcane	139.4	36.7	114.7	32.3	84.7	6.81	251.1	78879	High
cotton	60.6	40.3	119.2	22.6	43.5	7.28	100	1821	High
maize	109.4	26.1	80.3	18.6	68.2	7.1	108.6	6188	Low

Descriptive analytics - What happened

- From the crop recommendation data:
  - ➢ Average fertilizer applied for N, P, ad K.
  - ➢ Average yield for each crop

	Crop	N
1	chickpea	43.13333
2	coffee	85.40667
3	cotton	81.91333
4	maize	98.57667
5	potato	94.82667
6	rice	105.45000
7	sugarcane	120.00333
8	wheat	87.10667

	Crop	P
1	chickpea	30.79333
2	coffee	33.13667
3	cotton	32.97667
4	maize	38.92000
5	potato	39.21333
6	rice	39.29000
7	sugarcane	46.65000
8	wheat	35.53000

	Crop	K
1	chickpea	52.60000
2	coffee	103.29333
3	cotton	101.63667
4	maize	110.60333
5	potato	145.35000
6	rice	113.57000
7	sugarcane	152.79000
8	wheat	87.99333

	Crop	yield_kg_ha
1	chickpea	1767.900
2	coffee	1599.600
3	cotton	1774.033
4	maize	6450.967
5	potato	27302.267
6	rice	5469.233
7	sugarcane	77003.267
8	wheat	4368.467

Diagnostic Analytics – Why Did It Happen?

- Diagnostic analytics seeks to uncover the reasons behind certain trends or anomalies.
- For instance, if maize yield dropped in a particular year.
- Diagnostic analytics can help determine whether the cause was due to inadequate rainfall, pest infestation, or soil nutrient deficiency.

Examples

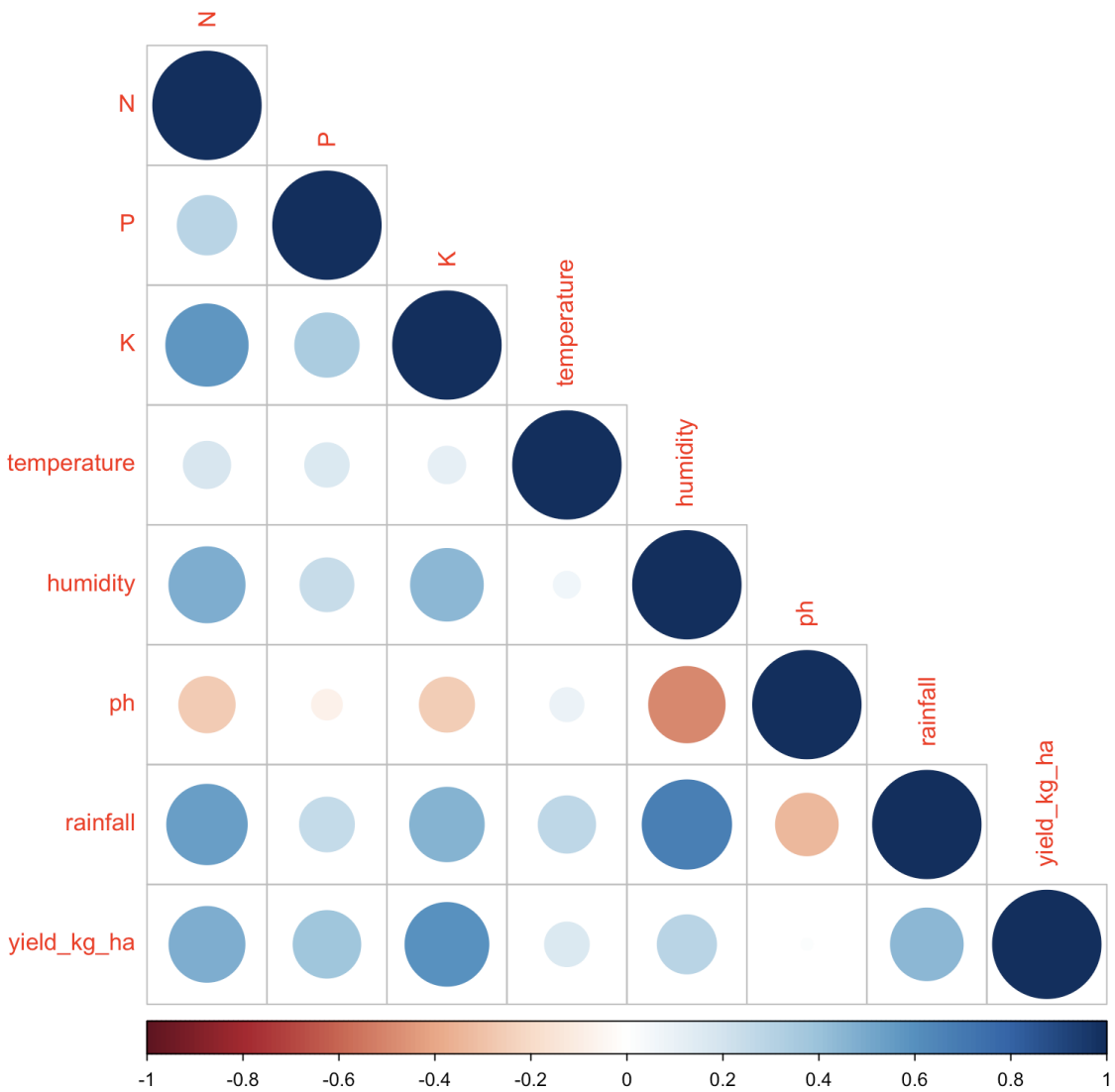
- Identify causes of yield drops
- Detect relationships between soil and crop health
- Report fertilizer or pesticide application

Diagnostics - Analytics

- Correlation plot of variables:

- Positive correlation:
- **Yield** with rainfall, N. P, K

- Negative correlation:
- **Yield** with pH



Predictive Analytics – What Might Happen?

- Predictive analytics uses statistical models and machine learning algorithms to forecast outcomes.
- In agriculture, this includes predicting yields, forecasting disease outbreaks, or anticipating market demand.
- It relies on historical and real-time data to generate insights.

Examples

- Forecast crop production
- Predict pest or disease spread
- Anticipate future input requirements

Predictive Analytics – What Might Happen?

Historical data

Crop	N	P	K	temperature	humidity	ph	rainfall	yield_category
maize	75	32.3	130.6	25.7	69.5	6.84	171.5	High
cotton	78.9	48.8	105.7	28.9	33.3	6	63.6	High
rice	101.3	58.8	142.5	31.3	74.9	5.69	156.6	High
coffee	108.8	26.6	100.9	25.5	89.3	5.79	126.4	Low
cotton	70.5	32.2	90.8	23.8	32.8	6.83	92.4	High
coffee	98.7	34.9	102.2	25.7	88.9	6.56	144.4	High
chickpea	43.6	26.9	43.7	26.9	35.6	6.17	67.4	Low
maize	115	35.4	87.4	18.5	73.3	6.48	198.3	High
cotton	50.7	43.9	96.4	27.7	32.8	7.28	71.6	High
wheat	97	25.1	93.9	18.9	54	6.78	128.8	Low
chickpea	48.7	39.2	79.9	21	42.2	6.27	86	High
maize	121.1	40.1	134.6	25	75.5	6.18	139.9	Low
rice	99	58.1	100.4	28.2	76.4	6.35	174.7	High
sugarcane	127.2	34.9	185.2	31.1	84.6	6.79	279.5	Low
maize	73.5	36	112.4	22.1	75.3	6.47	172.2	High
rice	86.5	46.9	99.7	29.9	88.2	6.12	166.2	Low
sugarcane	130.6	32.6	100.1	24.5	86.8	6.97	147.5	High
coffee	94.5	24.9	110.1	21	78.6	5.41	180	Low
rice	119.7	42.3	134.9	30	80.2	6.17	175	Low
cotton	91	24.5	127.3	24.5	36	7.22	113	High
rice	106.6	29.1	118.8	22.8	90.7	5.72	263.7	Low
cotton	106.1	34	73.8	27.3	41.8	5.76	53	High
sugarcane	129.8	34.5	183.3	24.1	72.7	6.83	200.1	High
sugarcane	139.4	36.7	114.7	32.3	84.7	6.81	251.1	High

Use historical data to model yield category: High or Low
Using machine learning

Predictive Analytics – What Might Happen?

New or test data

Crop	N	P	K	temperature	humidity	ph	rainfall
maize	109.4	26.1	80.3	18.6	68.2	7.1	108.6
maize	81.5	25.7	135.6	23.4	59.2	6.7	80.9
coffee	84.1	33.7	93.2	20.9	71.7	5.59	130.1
cotton	62.7	34.1	111.3	34.3	41.4	6.53	99.3
chickpea	25.5	42.6	30.5	19.9	30.9	6.22	89.6
maize	93.6	27.4	125.3	23.2	64.1	5.8	101.7
chickpea	33.2	42.9	32.4	22.6	51.3	6.3	43.8
rice	88.4	28	80.5	31.4	86.6	5.82	267.1
chickpea	33	42.3	56.5	25.4	47.7	7.31	64
cotton	66.7	26.1	73	29.3	44.1	6.02	43.6
maize	81	54.4	107.5	27.4	69.1	6.64	97.4

Predict yield category

yield_category

Predictive Analytics – What Might Happen?

New or test data

Crop	N	P	K	temperature	humidity	ph	rainfall
maize	109.4	26.1	80.3	18.6	68.2	7.1	108.6
maize	81.5	25.7	135.6	23.4	59.2	6.7	80.9
coffee	84.1	33.7	93.2	20.9	71.7	5.59	130.1
cotton	62.7	34.1	111.3	34.3	41.4	6.53	99.3
chickpea	25.5	42.6	30.5	19.9	30.9	6.22	89.6
maize	93.6	27.4	125.3	23.2	64.1	5.8	101.7
chickpea	33.2	42.9	32.4	22.6	51.3	6.3	43.8
rice	88.4	28	80.5	31.4	86.6	5.82	267.1
chickpea	33	42.3	56.5	25.4	47.7	7.31	64
cotton	66.7	26.1	73	29.3	44.1	6.02	43.6
maize	81	54.4	107.5	27.4	69.1	6.64	97.4

Predict

Reference yield_category	Predicted yield_category
High	High
High	Low
Low	Low
High	Low
Low	High
High	High
Low	Low
Low	Low
High	High
Low	High
High	Low

```
Confusion Matrix and Statistics

          Reference
Prediction High Low
      High   23   2
      Low    10  37

               Accuracy : 0.8333
```

Prescriptive Analytics – What Should Be Done?

- Prescriptive analytics goes beyond prediction to recommend actions that should be taken to achieve desired outcomes.

- These could include recommending specific:

  - Fertilizer applications
  - irrigation schedules
  - optimal crop rotations

Examples

- Forecast crop production
- Predict pest or disease spread
- Anticipate future input requirements

Prescriptive analytics

- Prescriptive analytics helps to plan and budget for resources needed to realize certain amount of yield.

- E.g: If we need at least 120 Kg of nitrogen, 60 Kg of phosphorus, and 80 Kg of potassium, and we want to buy them as cheaply as possible, then we need a list of prices from different sources.

- We then use price optimization models to choose the best price for the fertilizers.

Generative AI – What Should Be Done?

- **Generative AI (GenAI)** is a type of new artificial intelligence that can create new content: such as text, images, audio, code, or simulations

- Instead of only analyzing existing data (as with traditional AI), GenAI generates new possibilities, making it useful for creativity, prediction, and problem-solving.

- It complements traditional analytics by expanding possibilities for training AI models and exploring future scenarios

Key Features of GenAI

- **Learning from examples:** Trains on large datasets to understand patterns, relationships, and styles.

- **Creating novel outputs:** Generates unique text, images, sounds, or models from learned patterns.

- **Powered by advanced models:** Often uses deep learning methods such as transformers (e.g., GPT (Generative pre-trained transformer), LLaMA) or diffusion models.

Uses of GenAI in Agriculture

- **Crop Disease Diagnosis:** Generate diseased crop images for AI training and give farmers AI-based treatment advice.
- **Climate & Yield Simulation:** Create synthetic climate scenarios and forecast yields under different farming strategies.
- **Agricultural Content Creation:** Produce personalized crop guides and farmer training materials in local languages.
- **Smart Farm Planning:** Develop optimized crop rotation and fertilizer schedules using soil and historical data.
- **Synthetic Data Generation:** Produce realistic satellite or drone imagery to improve crop monitoring systems.

Institutional applications of Generative AI

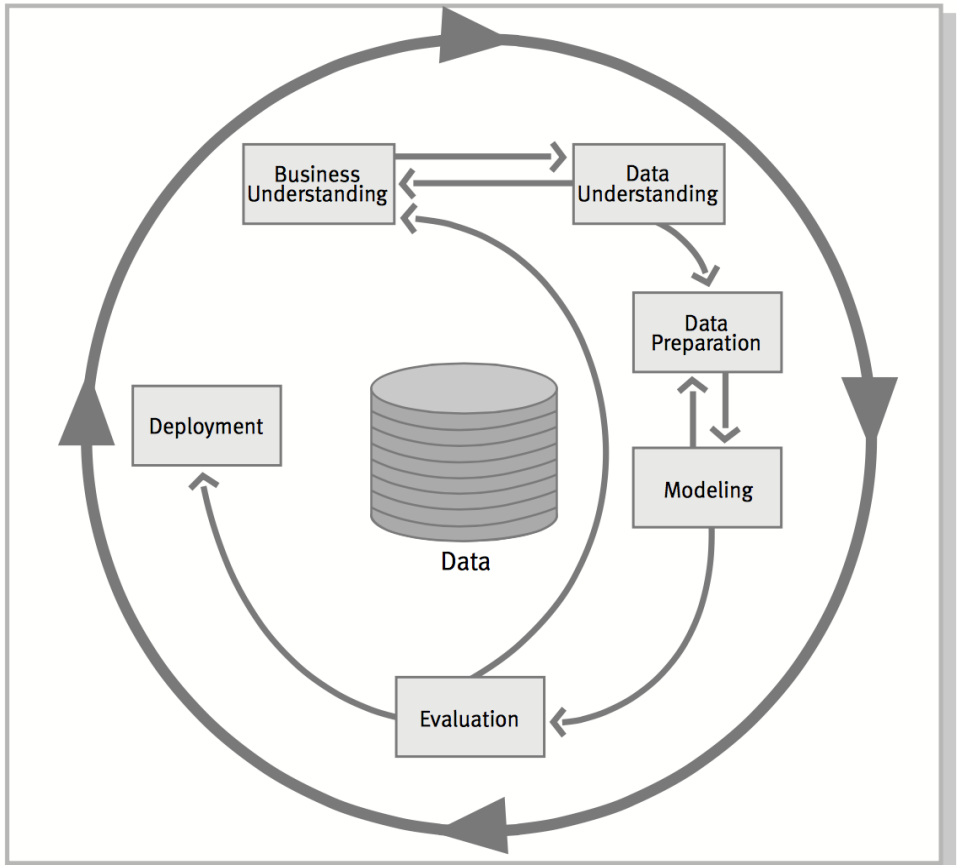
- **Education & Research:** Assist in literature reviews, content creation, and training.
- **Healthcare:** Support clinical decision-making and patient education.
- **Public Policy:** Draft policy briefs, simulate outcomes, and analyze large datasets.
- **Business Operations:** Automate reporting, customer service, and content generation.
- **Creative Industries:** Produce prototypes, media, and design concepts.

Ethical Considerations of Generative AI

- **Transparency and Explainability:** Clearly communicate AI decision-making processes.
- **Bias and Fairness:** Ensure datasets and algorithms are free from discriminatory bias.
- **Privacy and Data Protection:** Safeguard personal and sensitive information.
- **Accountability:** Define responsibility for AI-generated outputs.
- **Responsible Use:** Prevent misuse in misinformation, plagiarism, or harmful content.

Analytics Life Cycle

Analytics Lifecyle



1. **Business/Research Understanding Phase**
2. **Data Understanding Phase**
3. **Data Preparation Phase**
4. **Modeling Phase**
5. **Evaluation Phase**
6. **Deployment Phase**

According to Cross Industry Standard Process for Data Mining (CRISP-DM), a given data mining project has a life cycle consisting of six phases, as illustrated above

1.Business/Research Understanding Phase

Start by saying clearly what you want to achieve and what you need: explain it in terms of your work or department.

Turn those goals into a data problem: figure out what information you need to collect or analyze to reach your goal.

Make a first plan: outline the steps you will take to get the data, study it, and find answers.

1. Business/Research Understanding Phase

Objective: Support decisions that increase yield and reduce costs.

Questions:

Descriptive: What happened to yield, rainfall, and inputs over time/regions?

Diagnostic: Why do yields differ (soil, weather, pests, inputs)?

Predictive: Can we predict the crop label from conditions?

Prescriptive: What nutrient additions meet targets at minimum cost?

2. Data Understanding Phase

- **Start by gathering your data:** bring together all the information you need.
- **Look through it and explore:** get to know what is in the data and spot any useful trends or surprises.
- **Check if the data is good quality:** make sure it's accurate, complete, and reliable.
- **If you want, pick out the most interesting parts:** focus on the pieces that might reveal useful actions or decisions.

Data

year	month	lon	lat	NDVI	Rainfall_mm	Temp_C	Humidity_pc	Elevation_m	Pesticide_kg	SoilType	LandCover	PopDensity_	LivestockDe	DistToWater	HealthAcces	CropYieldCla	MalariaRisk	BrucellosisR	MalariaCase	BrucellosisC	Outbreak_M	Outbreak_Brucellosis
2025	2	39.5717535	0.38372333	-0.0874588	1212.44289	25.0836092	92.8771162	964.182901	63.5083257	Loamy	Urban	106.303517	133.448507	28.2269898	16.4368577	High	High	High	4	4	0	0
2024	4	36.2891147	-3.89903	-0.440396	1206.07587	27.2495964	65.8780663	542.071976	44.3144715	Loamy	Cropland	134.62017	10.318365	6.64847059	10.9612604	High	Medium	Medium	7	2	0	0
2024	7	35.8148116	1.38028384	0.8657833	1450.13521	24.60198	35.8677569	1050.9322	97.4277691	Sandy	Cropland	288.455051	40.5507916	21.4096691	5.29484183	High	Medium	Medium	8	3	0	0
2024	12	38.4105182	4.46477695	-0.3712973	885.808383	21.4403955	86.2809391	2259.13514	47.7580942	Clay	Grassland	446.798984	53.0290089	0.8801142	4.18011919	High	Medium	High	4	2	0	0
2025	5	39.7557518	2.42457603	0.81942932	699.30832	19.2713015	67.4256546	512.406664	51.6082819	Loamy	Cropland	493.315677	3.00567652	4.6026391	1.03984152	High	High	Medium	17	1	1	0
2025	1	37.3848517	0.66396702	-0.9131638	953.528521	23.6173135	75.5746922	146.931071	43.4841678	Loamy	Forest	170.476558	15.5106858	4.07856795	16.453529	Medium	Medium	Low	7	2	0	0
2025	8	41.8461136	-3.5762827	0.41423012	835.001589	23.9176249	75.2305217	1008.26068	16.1309191	Sandy	Forest	172.664158	73.1812141	7.79274478	28.5357197	High	High	High	11	5	1	1
2025	1	39.4786379	1.79850667	-0.0322219	948.630928	23.7057253	69.4483485	801.952077	44.7369324	Clay	Grassland	114.233755	54.6533322	25.9111934	14.3127033	High	Medium	Medium	4	5	0	1
2024	2	37.8474552	1.45051081	-0.1115579	789.956202	19.4084613	72.8501395	404.776677	54.6407394	Loamy	Urban	76.7939256	46.333084	1.89564511	26.2761038	High	Medium	Medium	3	1	0	0
2024	1	37.1369402	-4.0581258	-0.9273533	1113.57601	27.2233954	68.5814101	1623.58374	80.0184524	Clay	Cropland	246.250729	59.9106316	7.42918037	0.20763205	High	Medium	High	4	2	0	0
2025	10	36.7454241	2.63069372	-0.9186336	1102.59657	27.2178892	48.4042114	744.725541	59.8391426	Sandy	Cropland	358.279764	44.2017995	2.23360065	24.0191062	Medium	Medium	Medium	7	2	0	0
2024	2	39.8323977	0.16844932	-0.3344928	994.615482	21.5301445	70.500983	700.549873	46.2123563	Sandy	Cropland	670.747064	73.2153543	24.638321	7.03086233	High	Medium	High	9	4	0	0
2024	10	37.508578	-0.6671908	0.89423908	1062.3163	26.4179463	80.3585826	1582.95543	51.1988639	Clay	Urban	282.841545	15.1243088	7.43541117	18.1047648	High	High	Medium	9	0	0	0
2025	4	34.4774232	2.59368456	0.23531995	971.589855	26.8584926	61.8491538	719.618841	56.5468073	Loamy	Cropland	134.223516	48.8627857	14.4050271	3.71945994	High	High	High	17	5	1	1
2024	8	37.184354	-0.795877	-0.2622503	1150.40673	24.1153412	73.2695333	952.614328	42.4206788	Loamy	Forest	707.250826	64.6177361	10.8296631	15.5585238	High	Medium	High	4	4	0	0
2024	2	39.9039633	-0.1707635	0.22395408	1013.92815	21.1545041	71.2938372	476.724712	63.6447403	Sandy	Water	445.416027	43.5470945	26.7173314	27.6710831	High	Medium	Medium	14	0	1	0
2024	4	35.4599338	-2.8653404	-0.5877369	946.910875	28.8401446	92.234781	270.674956	82.5017632	Loamy	Grassland	102.944732	27.1088946	14.05853	23.262695	High	Medium	Medium	10	1	0	0
2024	11	35.4036141	-1.6081299	-0.6698671	1185.91689	25.1427643	61.2087455	636.226772	42.7370252	Loamy	Urban	314.933598	98.6908663	4.7458974	20.8048343	High	Medium	High	4	3	0	0
2024	8	38.252411	3.10979697	-0.2763655	1252.16425	23.2468876	60.6199376	223.446094	74.1506053	Loamy	Cropland	322.345287	49.5404345	12.9130704	20.9645032	High	Medium	Medium	11	1	1	0
2025	2	38.2546207	-2.8178663	0.7267067	1267.49077	26.3185362	63.2460239	2431.86683	67.5265261	Loamy	Forest	575.890648	74.4987671	12.1617102	9.01955165	High	Medium	High	4	4	0	0
2025	6	39.0752077	-0.7443805	0.01880346	801.945062	28.8666255	63.4641339	1689.59414	60.7448972	Loamy	Forest	143.601227	70.9259196	0.30518706	22.4035024	High	High	High	12	4	1	0
2024	2	40.7954544	4.40131057	-0.406197	893.110558	22.5085196	30.9107343	886.535873	52.2385124	Loamy	Grassland	402.974812	65.921417	10.5026903	14.5947015	High	Low	High	7	4	0	0

(1) year, (2) month, (3) lon, (4) lat, (5) NDVI, (6) Rainfall_mm, (7) Temp_C, (8) Humidity_pct, (9) Elevation_m, (10) Pesticide_kg, (11) SoilType, (12) LandCover, (13) PopDensity_km2, (14) LivestockDensity_km2, (15) DistToWater_km, (16) HealthAccess_km, (17) CropYieldClass, (18) MalariaRisk, (19) BrucellosisRisk, (20) MalariaCases, (21) BrucellosisCases, (22) Outbreak_Malaria, (23) Outbreak_Brucellosis

2. Data Understanding Phase

```
      year          month             lon              lat              NDVI          Rainfall_mm       Temp_C
 Min.   :2024   Min.   : 1.000   Min.   :34.02   Min.   :-4.46821   Min.   :-0.999836   Min.   : 442.4   Min.   :15.62
 1st Qu.:2024   1st Qu.: 3.000   1st Qu.:36.44   1st Qu.:-2.36467   1st Qu.:-0.460975   1st Qu.: 881.3   1st Qu.:21.54
 Median :2025   Median : 6.000   Median :38.17   Median :-0.32136   Median :-0.024135   Median :1015.2   Median :23.57
 Mean   :2025   Mean   : 6.315   Mean   :38.03   Mean   :-0.08687   Mean   :-0.005262   Mean   :1014.9   Mean   :23.68
 3rd Qu.:2025   3rd Qu.: 9.000   3rd Qu.:39.75   3rd Qu.: 2.32602   3rd Qu.: 0.472000   3rd Qu.:1149.7   3rd Qu.:26.12
 Max.   :2025   Max.   :12.000   Max.   :41.96   Max.   : 4.46478   Max.   : 0.997837   Max.   :1512.0   Max.   :34.71
  Humidity_pct    Elevation_m      Pesticide_kg      SoilType          LandCover          PopDensity_km2
 Min.   :20.00   Min.   :  86.72   Min.   : 2.953   Length:200         Length:200         Min.   :  2.253
 1st Qu.:54.37   1st Qu.: 414.91   1st Qu.:37.982   Class :character   Class :character   1st Qu.:147.537
 Median :65.28   Median : 723.01   Median :49.881   Mode  :character   Mode  :character   Median :274.991
 Mean   :63.86   Mean   : 760.40   Mean   :50.179                                         Mean   :300.401
 3rd Qu.:73.31   3rd Qu.: 982.05   3rd Qu.:63.542                                         3rd Qu.:405.102
 Max.   :97.37   Max.   :2431.87   Max.   :97.428                                         Max.   :993.569
 LivestockDensity_km2 DistToWater_km      HealthAccess_km    CropYieldClass     MalariaRisk        BrucellosisRisk
 Min.   :  2.91       Min.   :2.376e-04   Min.   : 0.04266   Length:200         Length:200         Length:200
 1st Qu.: 29.79       1st Qu.:4.986e+00   1st Qu.: 8.41576   Class :character   Class :character   Class :character
 Median : 50.20       Median :1.046e+01   Median :15.71029   Mode  :character   Mode  :character   Mode  :character
 Mean   : 58.75       Mean   :1.076e+01   Mean   :15.51728
 3rd Qu.: 75.42       3rd Qu.:1.577e+01   3rd Qu.:21.28566
 Max.   :254.81       Max.   :2.862e+01   Max.   :44.14289
  MalariaCases    BrucellosisCases Outbreak_Malaria Outbreak_Brucellosis
 Min.   : 1.000   Min.   : 0.00    Min.   :0.000    Min.   :0.000
 1st Qu.: 5.000   1st Qu.: 2.00    1st Qu.:0.000    1st Qu.:0.000
 Median : 7.000   Median : 3.00    Median :0.000    Median :0.000
 Mean   : 7.505   Mean   : 3.03    Mean   :0.165    Mean   :0.215
 3rd Qu.:10.000   3rd Qu.: 4.00    3rd Qu.:0.000    3rd Qu.:0.000
 Max.   :17.000   Max.   :10.00    Max.   :1.000    Max.   :1.000
>
```

3. Data Preparation Phase

- **Start with the messy, raw data:** this is the unprocessed information you collected.
- **Choose the records and details you actually need:** only keep what's useful for your analysis.
- **Change or adjust some details if needed:** for example, combine columns, change formats, or calculate new values.
- **Clean it up:** remove errors, fix missing values, and make sure everything is in good shape for the next steps.

4. Modeling Phase

- **Pick the right method to build your model:** choose the approach that fits your problem.
- **Adjust the settings to get the best results:** fine-tune the model so it works well.
- **Try more than one method if needed:** sometimes comparing different approaches gives better answers.
- **Go back and fix the data if required:** if a method needs the data in a certain format, adjust the data and try again.

5. Evaluation Phase

- **Check how good the model is:** see if it works well and gives reliable results.
- **Make sure it meets your original goals:** confirm it answers the problem you set out to solve.
- **See if anything important was missed:** check if the model ignored any key part of the business or research problem.
- **Decide what to do with the results:** choose whether to start using the model or make more changes first.

6. Deployment Phase

Building the model is not the end: you still need to put it to use.

Simple use: for example, create a report with the results.

Bigger use: run the same kind of analysis in another team or department.

Interactive use: build an app (like an R Shiny dashboard) so people can explore and use the model results easily.

In business: the customer might take the model and use it themselves.

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.

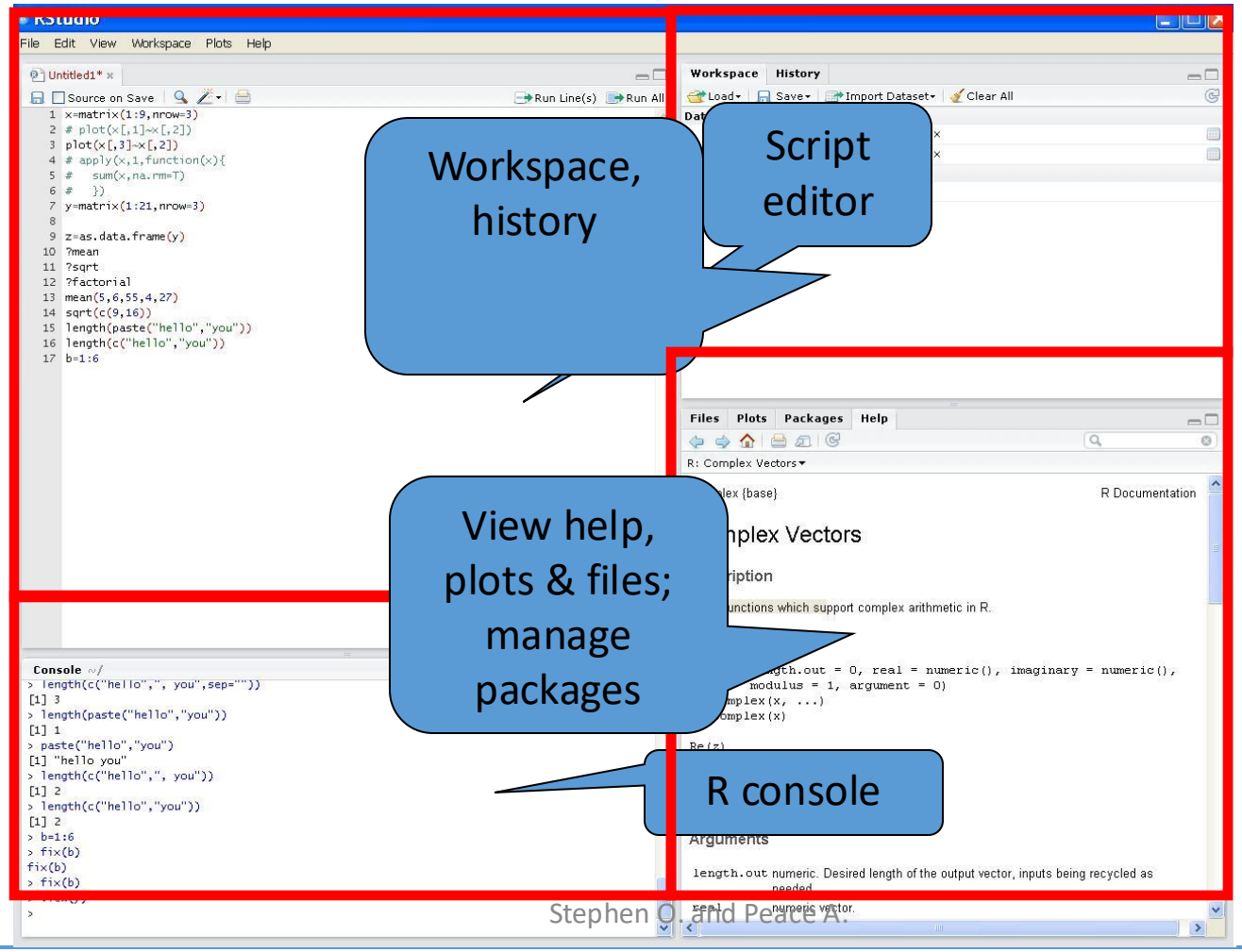
- R Studio is available in **open source** and **commercial** editions and runs on the desktop (Windows, Mac, and Linux).

What is R Studio ?

- More information:

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

Using RStudio



Naming objects in R

You can enter a command at the command prompt in a console (>).

- To write a comment, use #

```
# This is my first comment
```

- Results are stored in an object using the assignment operator: (<-) or the equal character (=).

```
# The code below will create an object named Test with a value of 2.
Test <- 2
Test = 2

#The code below will combine different values to form one object.
Test1 <- c(1,2,3,4,5)

#To print (show) the items in an object, just enter the name of the object
Test
```

Naming objects in R

- Object names cannot contain `strange' symbols like !, +, -.

# The code below will work because the object name does not contain strange symbols.
Test <- c(1,2,3,4,5)

The code below will not work because the object name contains strange symbols. There will be an error message.
Test! <-c(1,2,3,4,5,6)

```
> Test! <-c(1,2,3,4,5,6)
Error: unexpected '!' in "Test!"
```

- Object names can contain a dot (.) and an underscore (_).

#The code below will work because variable names can contain a dot(.).
Temp.1 <- c("Second", "Third")

#The code below will work because variable names can contain an underscore(_).
Temp_1 <- c("Second", "Third")

Naming objects in R

- Object names can contain a number but cannot start with a number.(E.g., Test_1, not 1Test).

#The code below will work because object names can contain numbers.

```
Test_1 <- c(1,2,3,4,5,6,7)
```

#The code below will work because object names can contain numbers.

```
Te1st <- c(1,2,3,4,5,6,7)
```

#The code below will not work because object names cannot start with numbers. An error message will be displayed.

```
1Test <- c(1,2,3,4,5)
```

```
> 1Test <- c(1,2,3,4,5)
Error: unexpected symbol in "1Test"
```

R output

Naming objects in R

- Object names are case sensitive.(e.g (X and x) (temp.1 and Temp.1) are different)

#The code below will create an object Temp.1 that contains strings.

```
Temp.1<- c("Second", "Third")
```

#The code below will view the object Temp.1 because the object name is the same.

```
Temp.1
```

```
> Temp.1
[1] "Second" "Third"
```

R output

#The code below will not view the object Temp.1 because the variable name temp.1 contains lower case . This is considered different and will display an error message.

```
temp.1
```

```
> temp.1
Error: object 'temp.1' not found
```

R output

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:

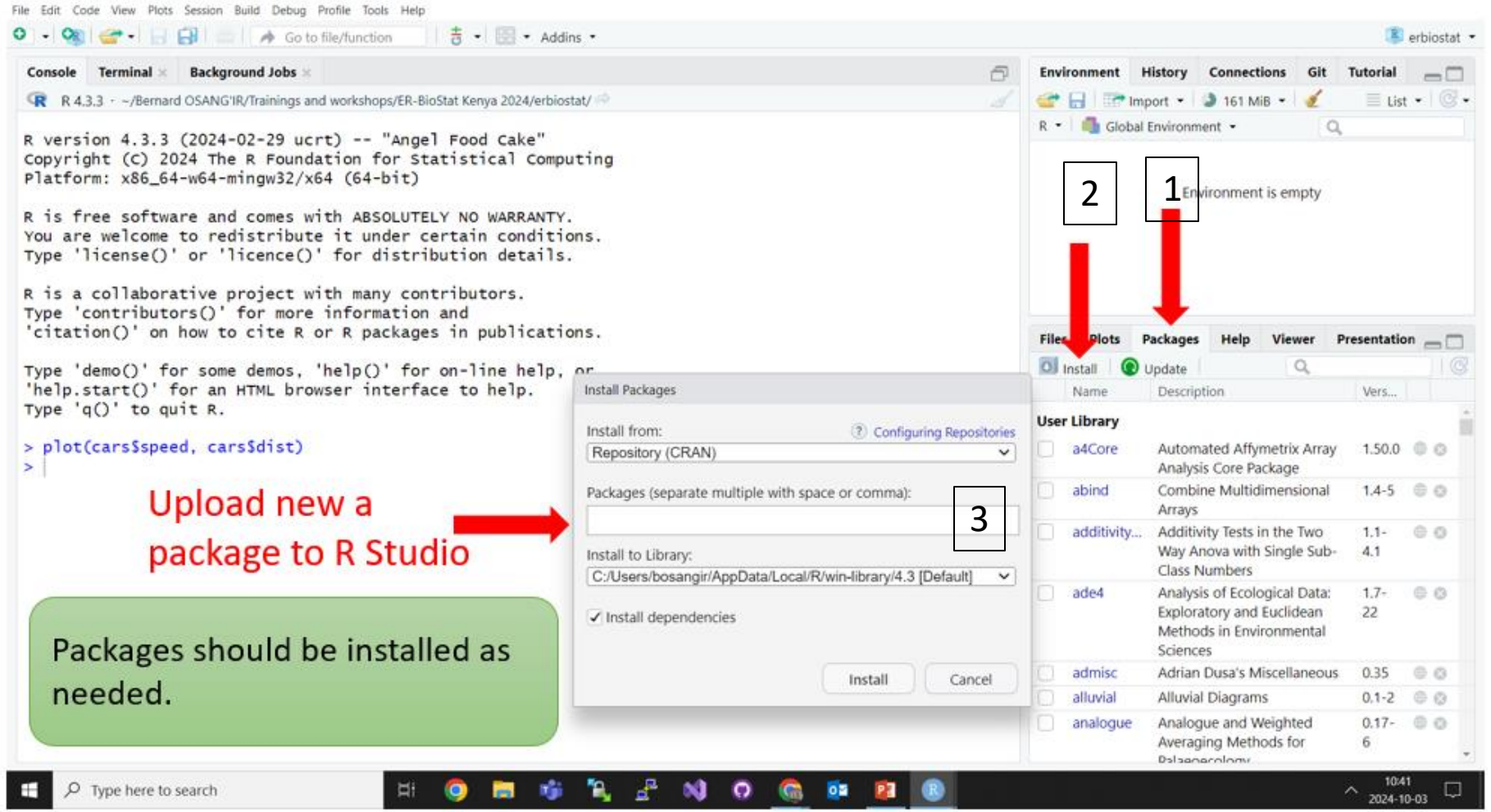
  install.packages ("dplyr ")

- 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(dplyr)

R studio: packages



Activity

- Install the following packages:
  1) dplyr
  2) readxl
  3) writexl
  4) haven
- Load the libraries of all the above packages

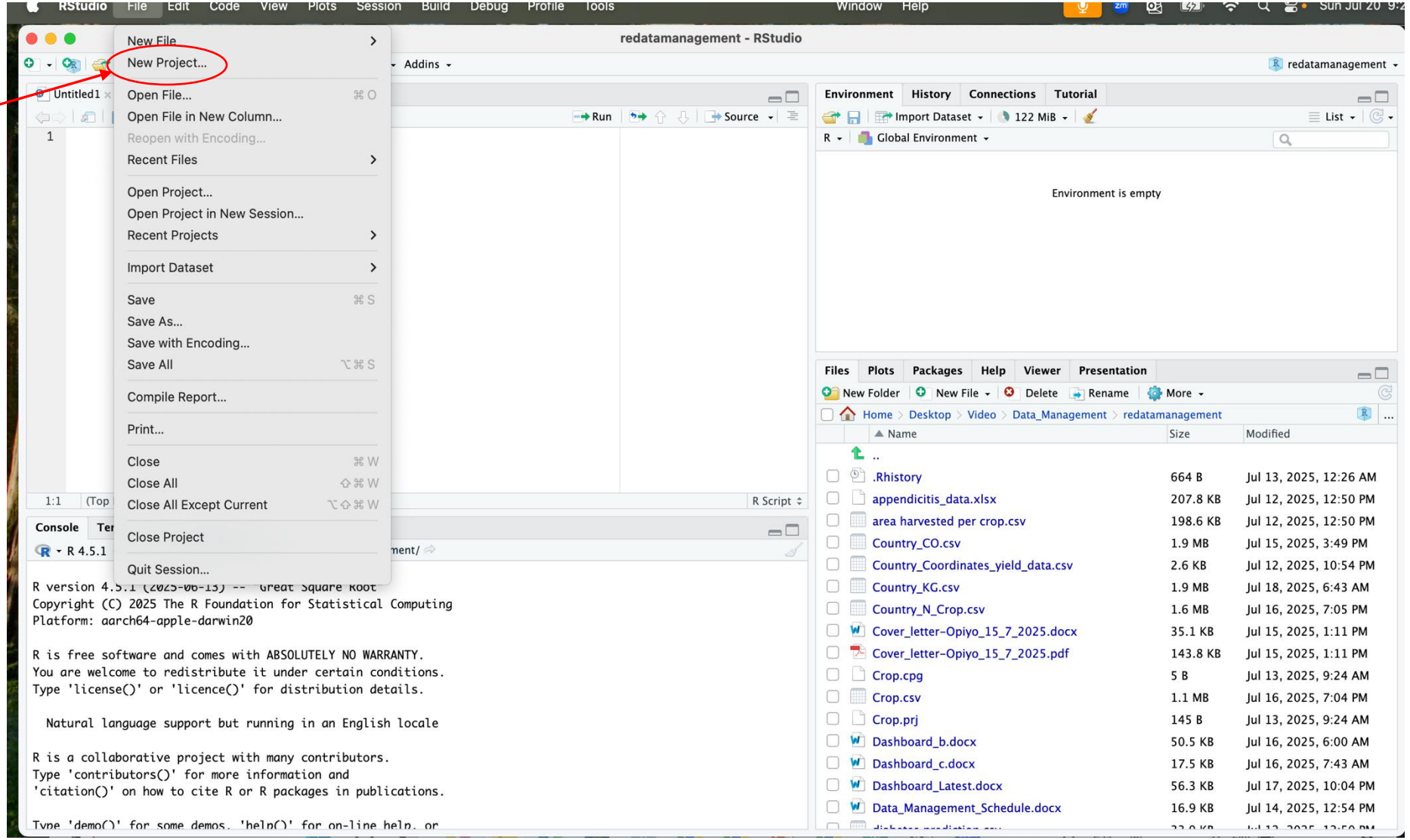
R projects

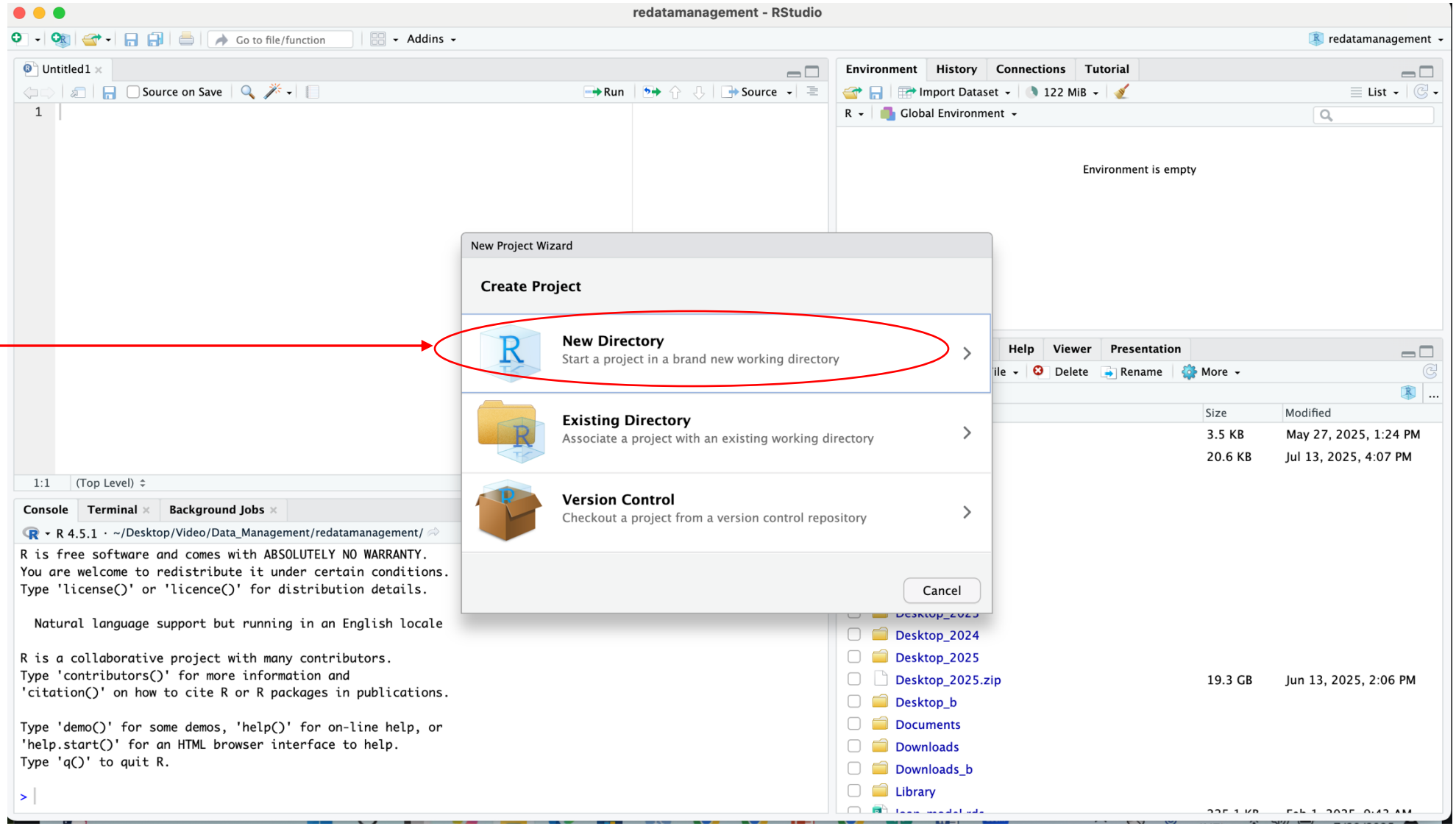
- An R project helps to organize the workflow.
- An R script stores R code for analysis.
- To create a project for this training.

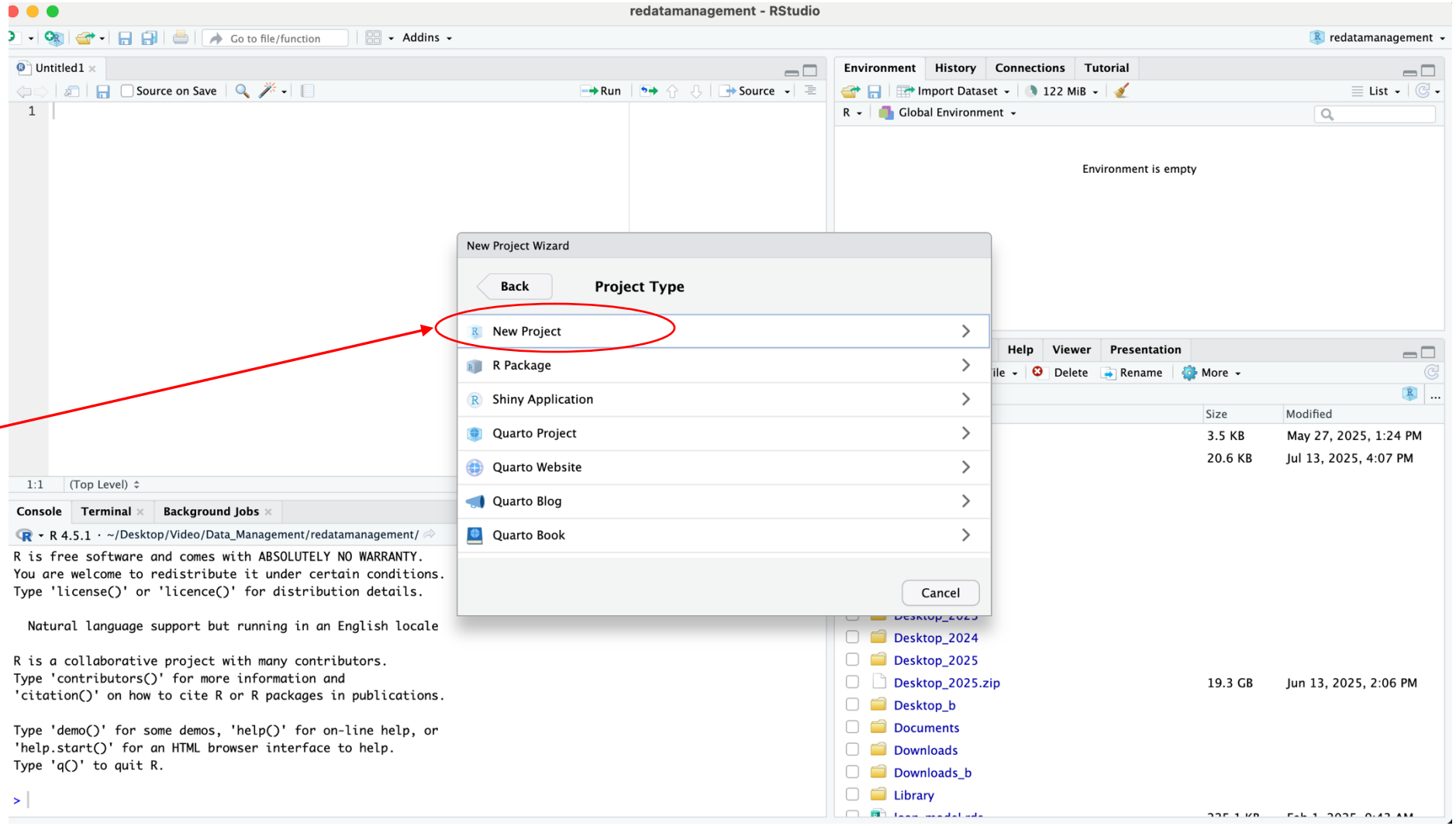
❖Exercise: Create an R project for this webinar series

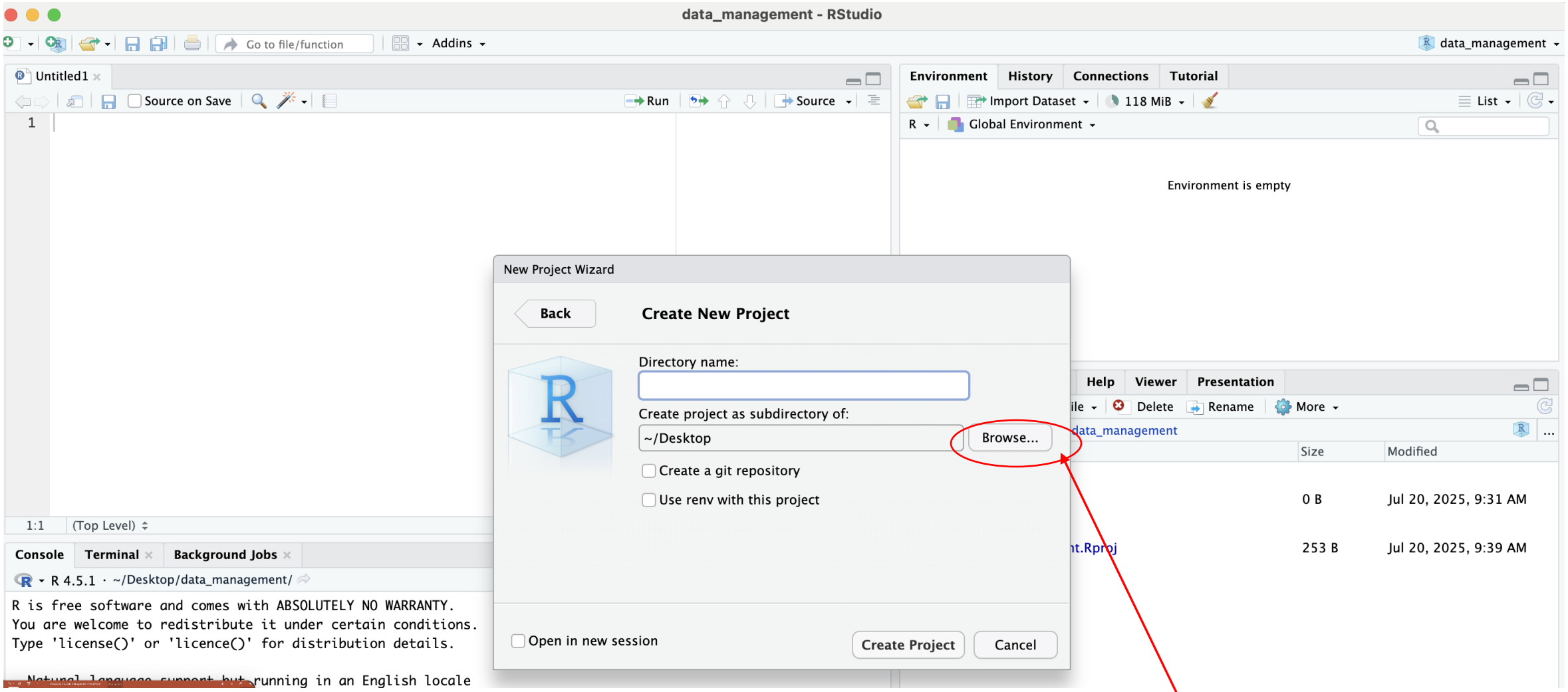
❖Demonstration using the R-Studio

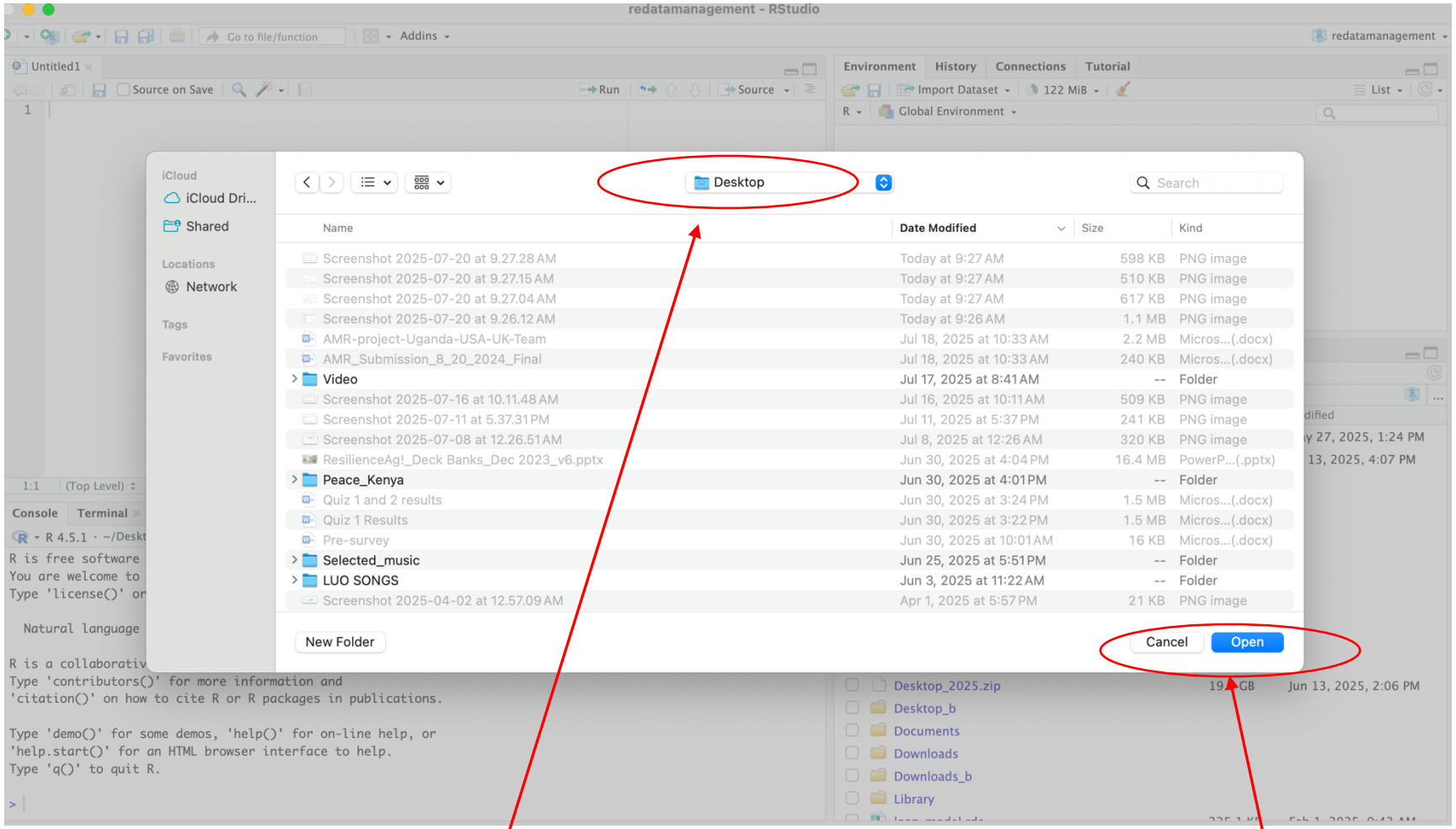
R projects

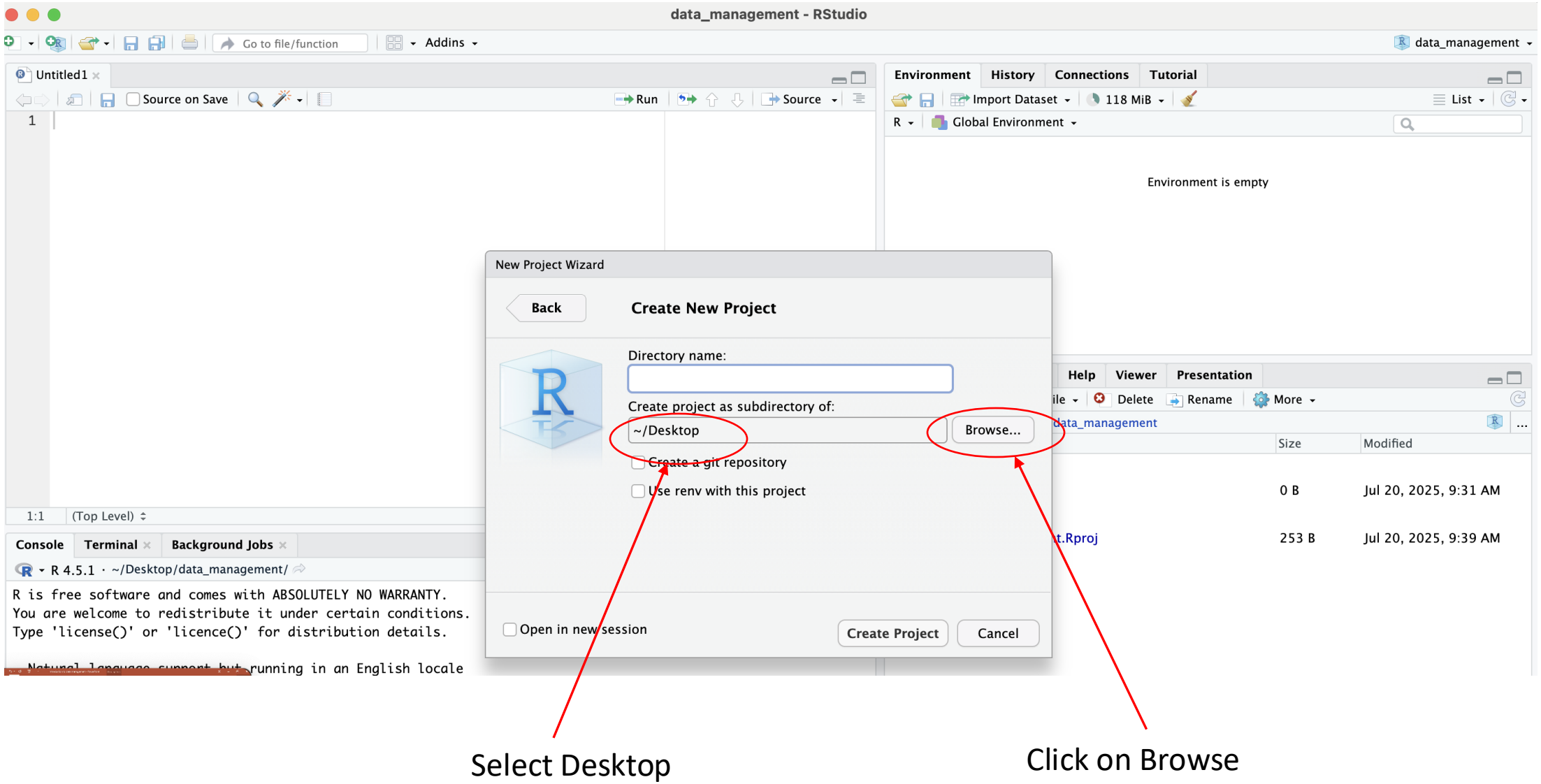

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
redatamanagement - RStudio
New File
New Project...
Open File... ⌘O
Open File in New Column...
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save ⌘S
Save As...
Save with Encoding...
Save All ⌥⌘S
Compile Report...
Print...
Close ⌘W
Close All ⇧⌘W
Close All Except Current ⌥⇧⌘W
Close Project
Quit Session...
Addins
redatamanagement
Untitled1
Run
Source
1
1:1
(Top
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.
Type 'demo()' for some demos, 'help()' for on-line help, or
Environment History Connections Tutorial
Import Dataset
122 MiB
List
R
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
R
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
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


redatamanagement - RStudio
Go to file/function
Addins
redatamanagement
Untitled1
Source on Save
Run
Source
1
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.
>
Environment
History
Connections
Tutorial
Import Dataset
122 MiB
List
R
Global Environment
Environment is empty
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
New Project Wizard
Back
Project Type
New Project
R Package
Shiny Application
Quarto Project
Quarto Website
Quarto Blog
Quarto Book
Cancel

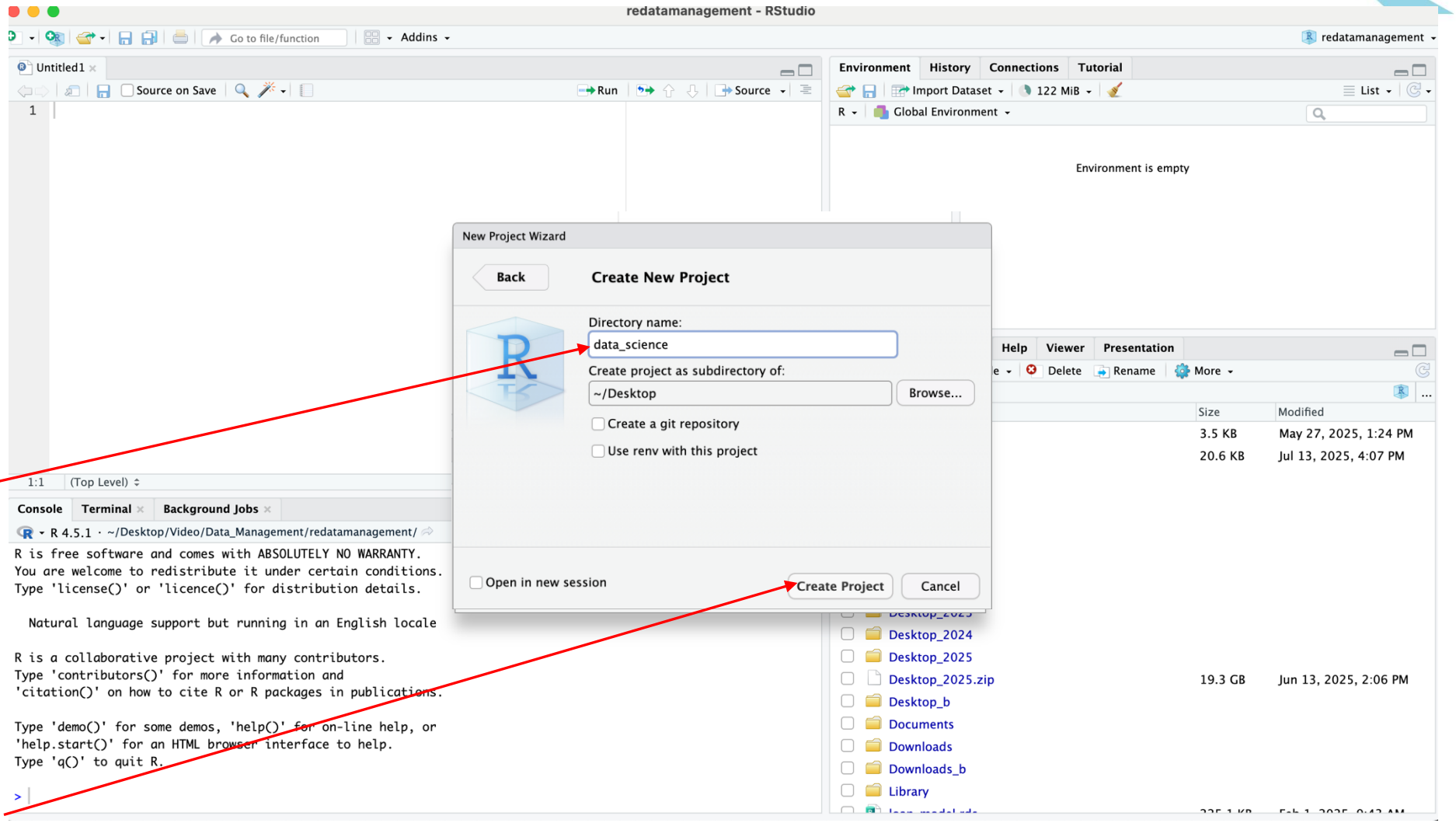

data_management - RStudio
Go to file/function
Addins
data_management
Untitled1
Source on Save
Run
Source
1
Environment
History
Connections
Tutorial
Import Dataset
118 MiB
List
R
Global Environment
Environment is empty
New Project Wizard
Back
Create New Project
Directory name:
Create project as subdirectory 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.
Natural language support but running in an English locale


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

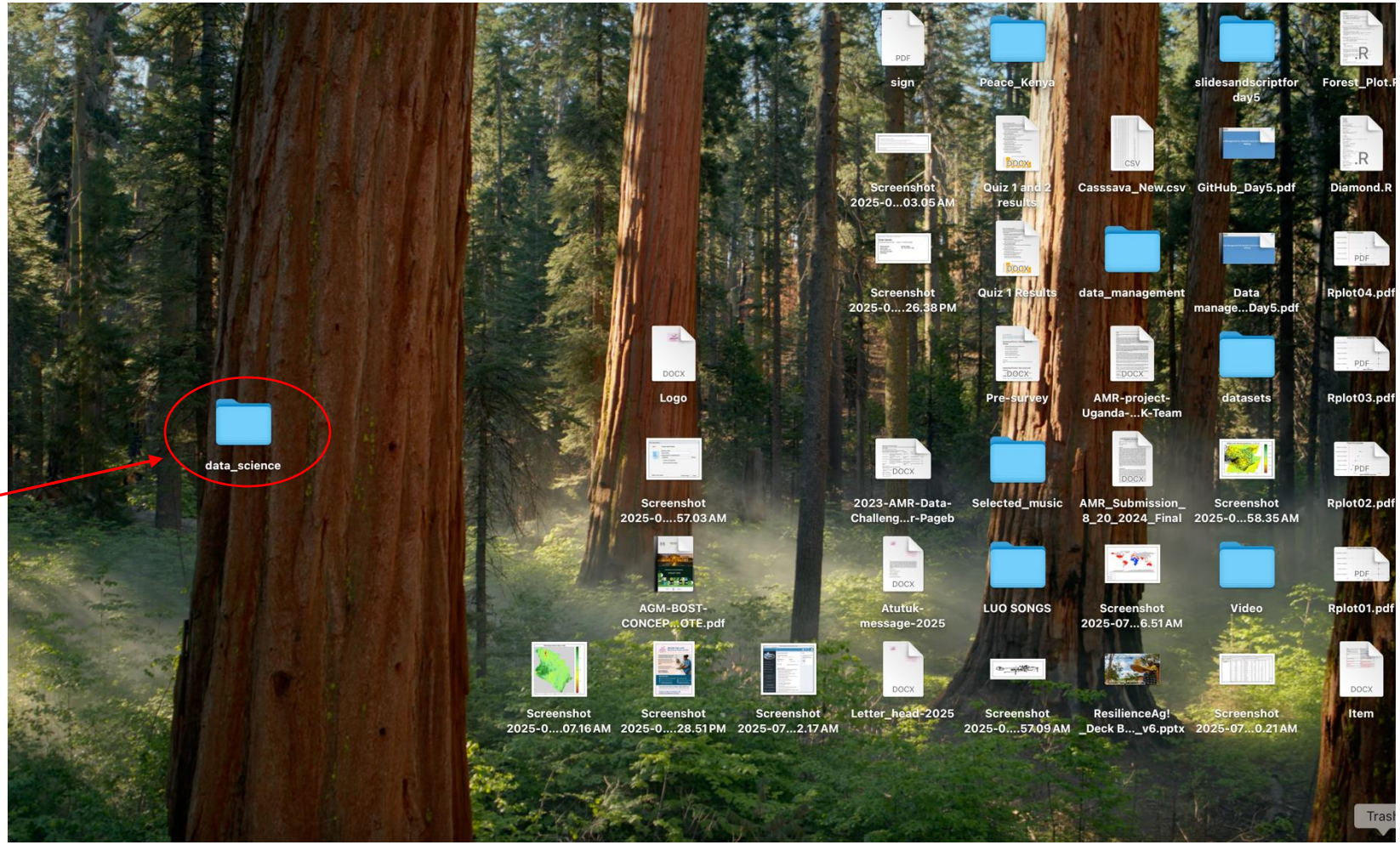

data_management - RStudio
Go to file/function
Addins
data_management
Untitled1
Source on Save
Run
Source
Environment
History
Connections
Tutorial
Import Dataset
118 MiB
List
R
Global Environment
Environment is empty
New Project Wizard
Back
Create New Project
Directory name:
Create project as subdirectory 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_science

2. Click on Create Project

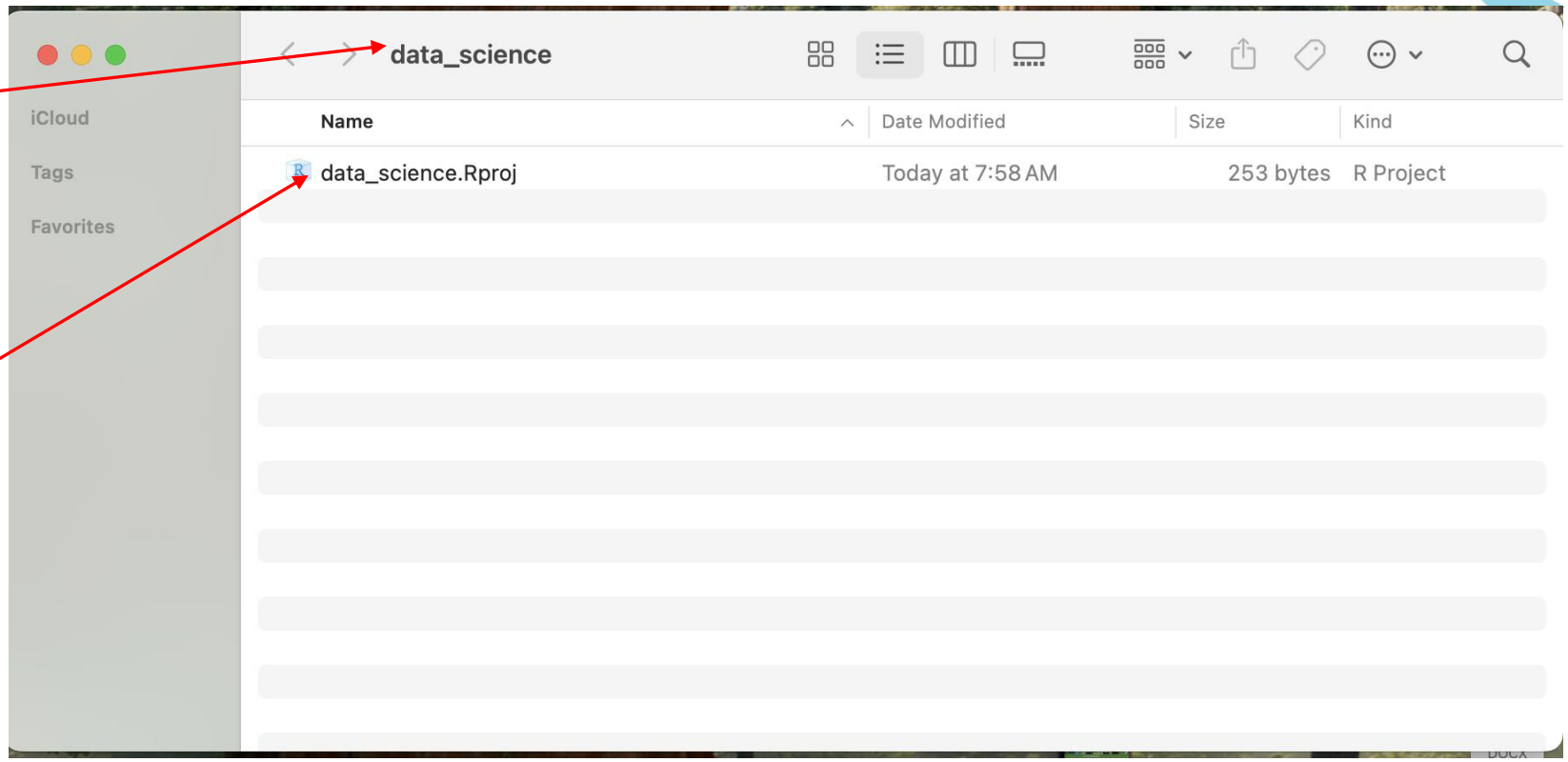


data_science folder on a desktop



Open data_science folder

data_science project



Download three datasets from google drive

- **Download the following datasets:**

1) crop_recommendation.csv

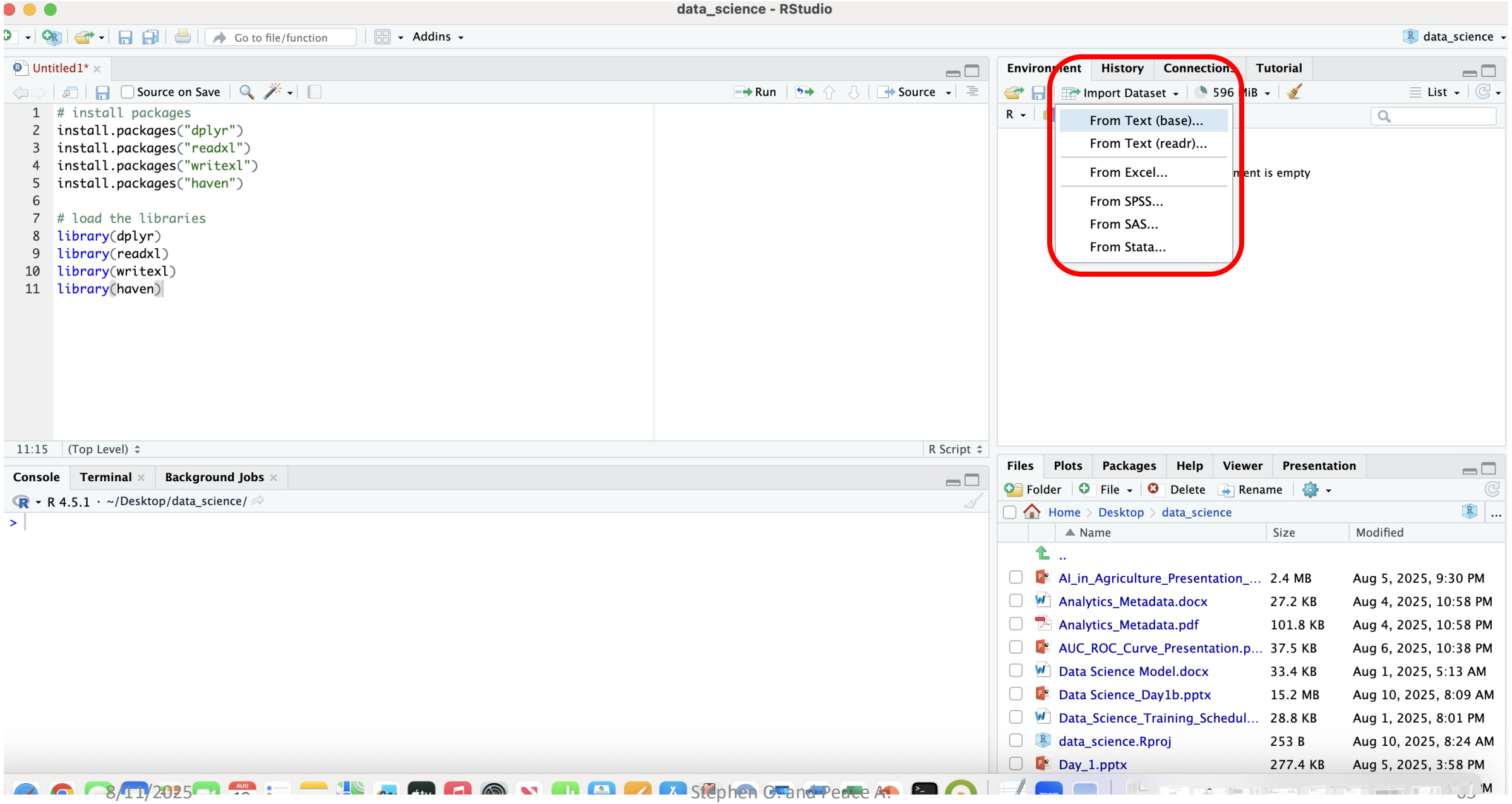
2) Kenya_agri_disease_spatial.csv , Kenya_agri_disease_spatial.xls, Kenya_agri_disease_spatial.dta, Kenya_agri_disease_spatial.sas, Kenya_agri_disease_spatial.sav

3) crop_recommendation.xlsx

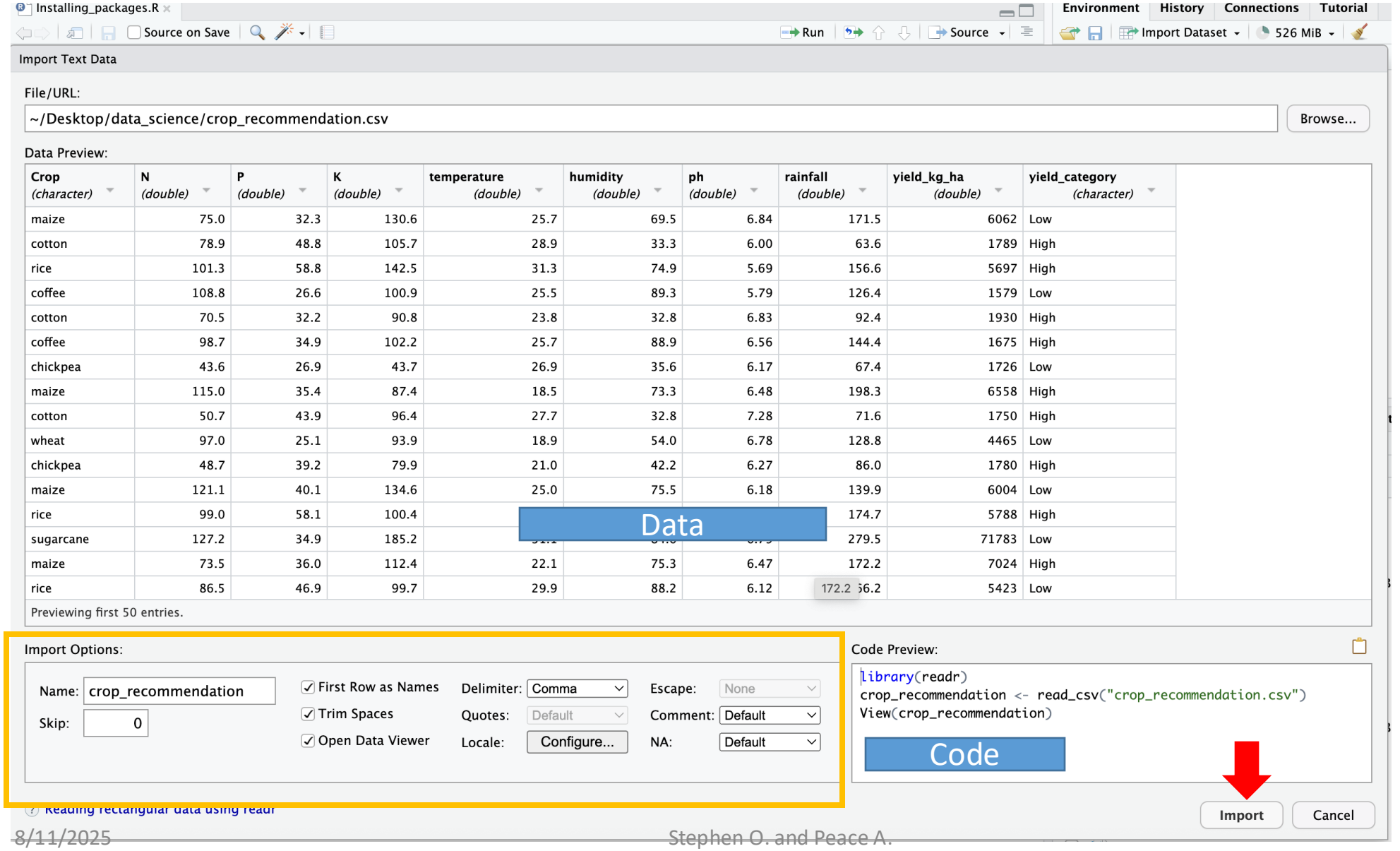
from the following link:
https://drive.google.com/drive/folders/1OKihGHcjUbJAFlvviZwGD4EciGt8pBum?usp=sharing

- Store in the datasets in data_science project folder that you created

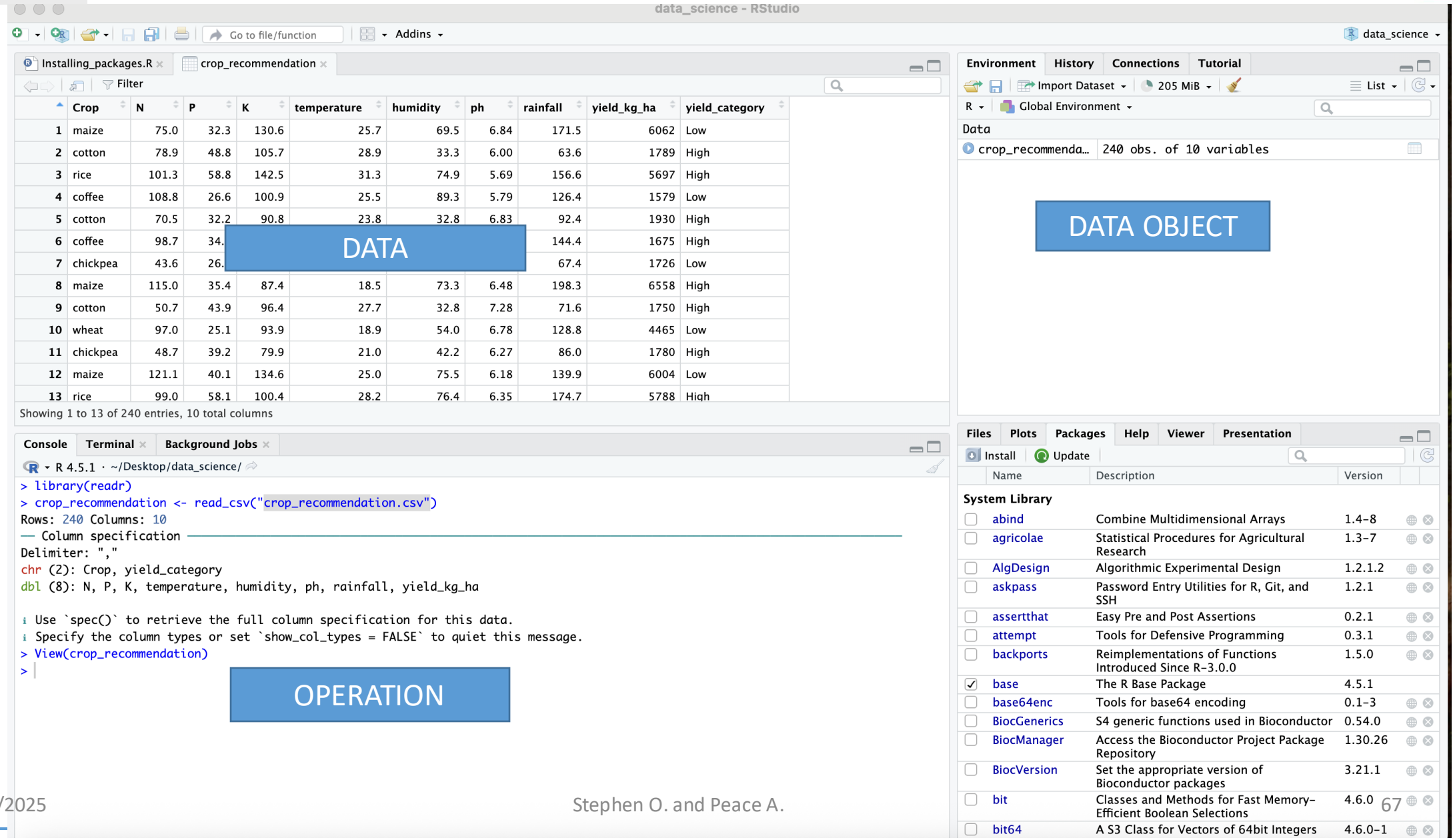
Load data in RStudio



Load data into R



Load data into R



	Crop	N	P	K	temperature	humidity	ph	rainfall	yield_kg_ha	yield_category
1	maize	75.0	32.3	130.6	25.7	69.5	6.84	171.5	6062	Low
2	cotton	78.9	48.8	105.7	28.9	33.3	6.00	63.6	1789	High
3	rice	101.3	58.8	142.5	31.3	74.9	5.69	156.6	5697	High
4	coffee	108.8	26.6	100.9	25.5	89.3	5.79	126.4	1579	Low
5	cotton	70.5	32.2	90.8	23.8	32.8	6.83	92.4	1930	High
6	coffee	98.7	34.					144.4	1675	High
7	chickpea	43.6	26.					67.4	1726	Low
8	maize	115.0	35.4	87.4	18.5	73.3	6.48	198.3	6558	High
9	cotton	50.7	43.9	96.4	27.7	32.8	7.28	71.6	1750	High
10	wheat	97.0	25.1	93.9	18.9	54.0	6.78	128.8	4465	Low
11	chickpea	48.7	39.2	79.9	21.0	42.2	6.27	86.0	1780	High
12	maize	121.1	40.1	134.6	25.0	75.5	6.18	139.9	6004	Low
13	rice	99.0	58.1	100.4	28.2	76.4	6.35	174.7	5788	High

Importing other data into R

Importing data into R

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

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

Exporting data out of R

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:

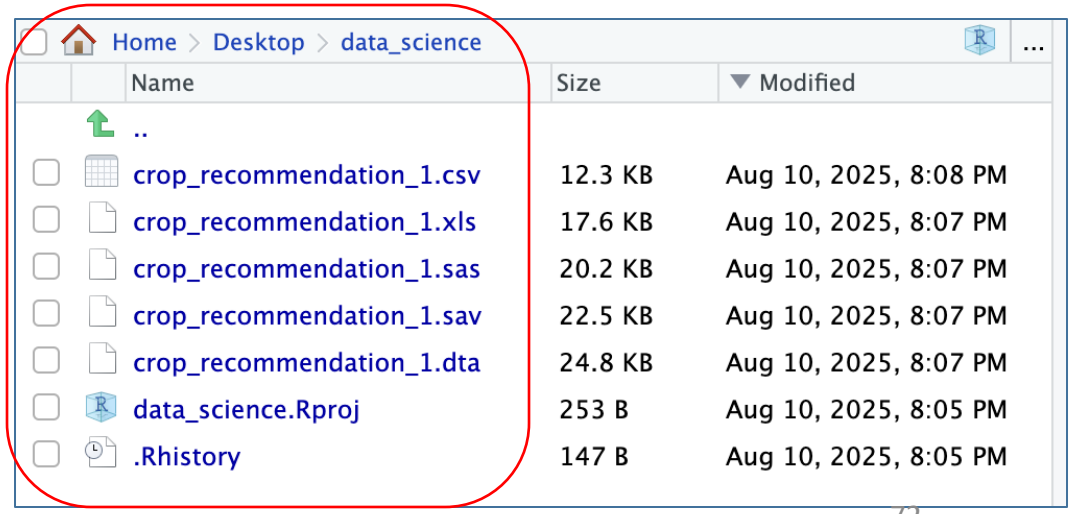
- Comma-separated values
- Excel files
- STATA files
- SPSS files
- SAS files

R uses the write() function to export data to other

formats as below:

```
library(readr)
library(haven)
library(writexl)

write_csv(crop_recommendation,"crop_recommendation_1.csv")
write_xlsx(crop_recommendation,"crop_recommendation_1.xls")
write_dta(crop_recommendation,"crop_recommendation_1.dta")
write_sav(crop_recommendation,"crop_recommendation_1.sav")
write_xpt(crop_recommendation, "crop_recommendation_1.sas")
```



Activity

https://docs.google.com/forms/d/e/1FAIpQLSezFfi4xZFsah-0idjZPzj8sZPaSBprLM5B4yliC2gHwBG26A/viewform?usp=header