Data Science and Data Analytics

Getting started

Julian Amon, PhD
Charlotte Fresenius Privatuniversität

March 14, 2025

Course design

Welcome to Data Science and Data Analytics!

- This is **B-GV-12 Data Science and Data Analytics**.
- The goal of this course is to teach you…
  - about the nature and role of data science in a data-driven world.
  - about the data science workflow.
  - to use open-source software to analyse, visualize and model data from various sources.
  - about different machine learning algorithms.
  - to implement your own data science projects.
  - to communicate the results of your analyses effectively.
  - and much more…

Lectures

- With a few exceptions, lectures will be held on **Fridays from 13:30 to 16:30**.
- Please regularly check your course schedule to not miss any lectures.
- Lectures will consist of theory and practice discussed with the help of **slides** as well as **live coding sessions**.
- You are highly encouraged to **actively participate**!
- Exercises to practice what you have learned will be provided (but not graded).

Grading

- There will be **no** final exam in this course!
- Instead, grading will be based on **group projects**:
  - Teams of 4 - 5 students will initiate and design a small data science project autonomously.
  - Each group will:
    - identify a business case mimicking a real-world research problem and an accompanying available data set.
    - formulate research questions on the basis of the chosen data set.
    - perform analyses using the concepts and methods learned throughout this course.

Grading

- Grading will therefore be based on the following components:
  - **Group project report** (5-10 pages per group member): **65 %**
  - **Group presentation** ($\leq$ 20 mins/group and $\geq$ 2 mins/group member): **25 %**
  - **Peer review: 10 %**
- In line with the usual grading scheme, grades will be given as follows:

Percentage	Grade
95 - 100 %	1,0
90 - 94 %	1,3
85 - 89 %	1,7
80 - 84 %	2,0
75 - 79 %	2,3
70 - 74 %	2,7

Percentage	Grade
65 - 69 %	3,0
60 - 64 %	3,3
55 - 59 %	3,7
50 - 54 %	4,0
below 50 %	5,0

Grading: Group project

- The **structure** of the group project (reflected in report and presentation) should be something like this:
  - Introduction / Motivation and research question
  - Data (sources, description, statistics, visualizations, …)
  - Models and model evaluation
  - Results
  - Discussion and comments
- Aspects that will influence your grade will be: the originality of the question, understanding of the business case, data and methods, correctness of application, thoroughness of evaluation, creativity and quality of report and presentation (both verbal and visual)
- Deductions will be made for purely AI-generated contributions.

Grading: Group project

Choice of topic: while you are completely free in your choice of topic in the group, here are some areas of suggestion:

- Finance / Economics / Marketing
- Text analysis
- Entertainment (in particular: movies and music)
- Social network analysis
- Social sciences

Sources for data sets: while you are again free also in your choice of data set, good places to get you started are:

- Statistik Austria
- Kaggle
- UCI Machine learning repository
- World bank
- EU
- …

Caution

When selecting a data set, make sure, you are **allowed** to use this data for the purposes of your project! When selecting from the sources given, this should generally be ensured.

Grading: Peer review

- After final project presentations, each **individual** student will be asked to write a **peer review** of one of the other groups' projects.
- Each student will be randomly assigned **two projects**, out of which **one** should be reviewed (based on their presentation only).
- Evaluation is based on the **quality of review** that you write, not on the feedback that your project receives.
- The review should be **max 250 words** answering the following questions:
  - Briefly describe the topic, research questions and the employed methods.
  - State and briefly explain two positive comments about the work.
  - State and briefly explain two improvement suggestions.

Important

In your peer review, focus on **content**, not on the formatting or quality of the slides, for instance.

Schedule

March

14th: Getting started, Introduction to Data Science (DS)

21st: The essentials of R programming

31st: The DS workflow – Part I: Import, Tidy and Transform

April

4th/11th: The DS workflow – Part II: Visualize

30th: The DS workflow – Part III: Model

May

6th/9th/16th/23rd: The DS workflow – Part III: Model

28th: The DS workflow – Part IV: Communicate

June

6th: Buffer session

13th: Final presentations of the group projects

Deadlines

- **21st March**: organize in teams, send team members via e-mail
- **12th June**: Send presentations via e-mail
- **27th June**: Hand in group project reports and peer reviews

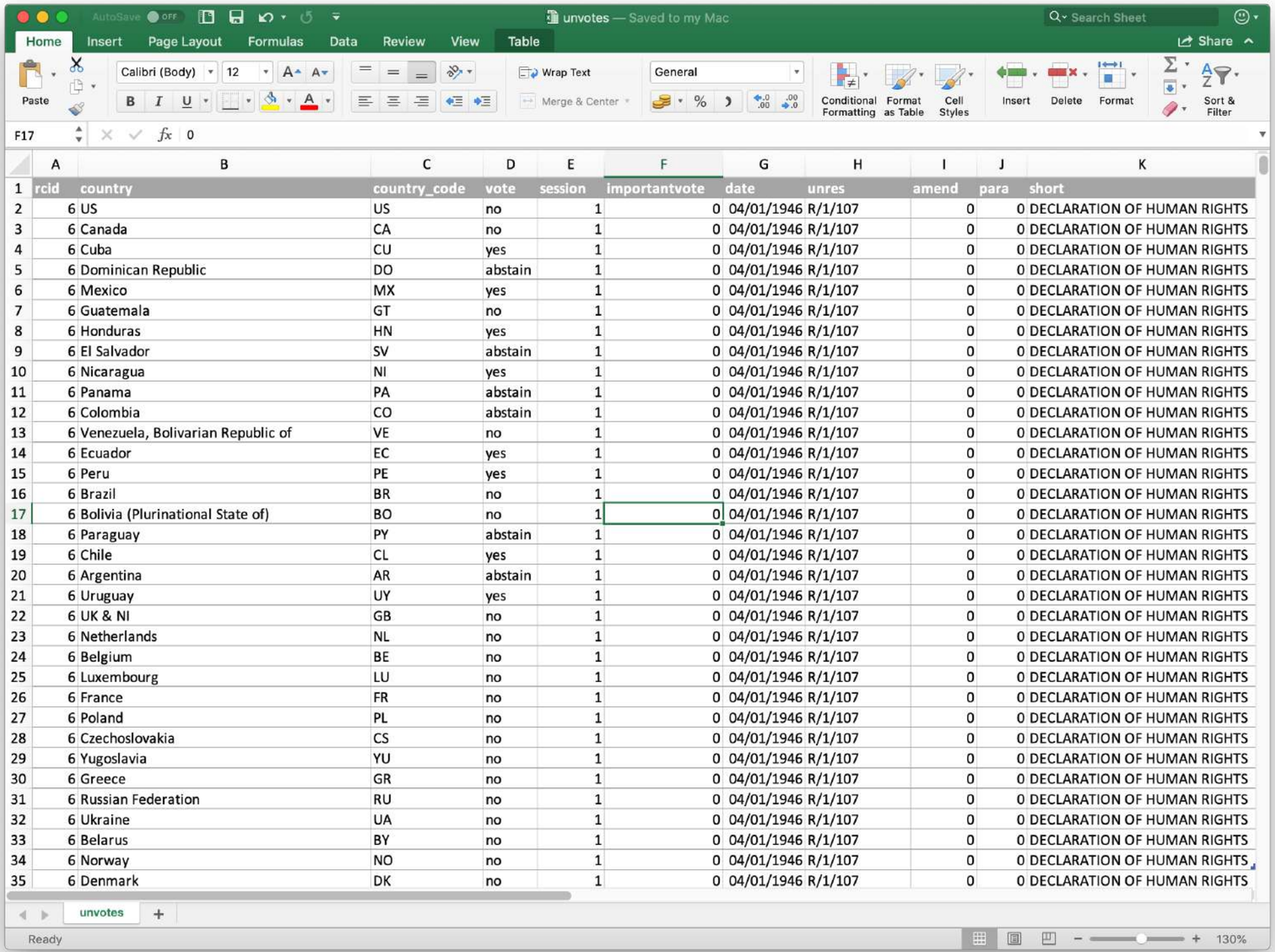
Questions and contact

- Any questions?

- Contact:
  - Anytime via e-mail: julianamonphd@gmail.com

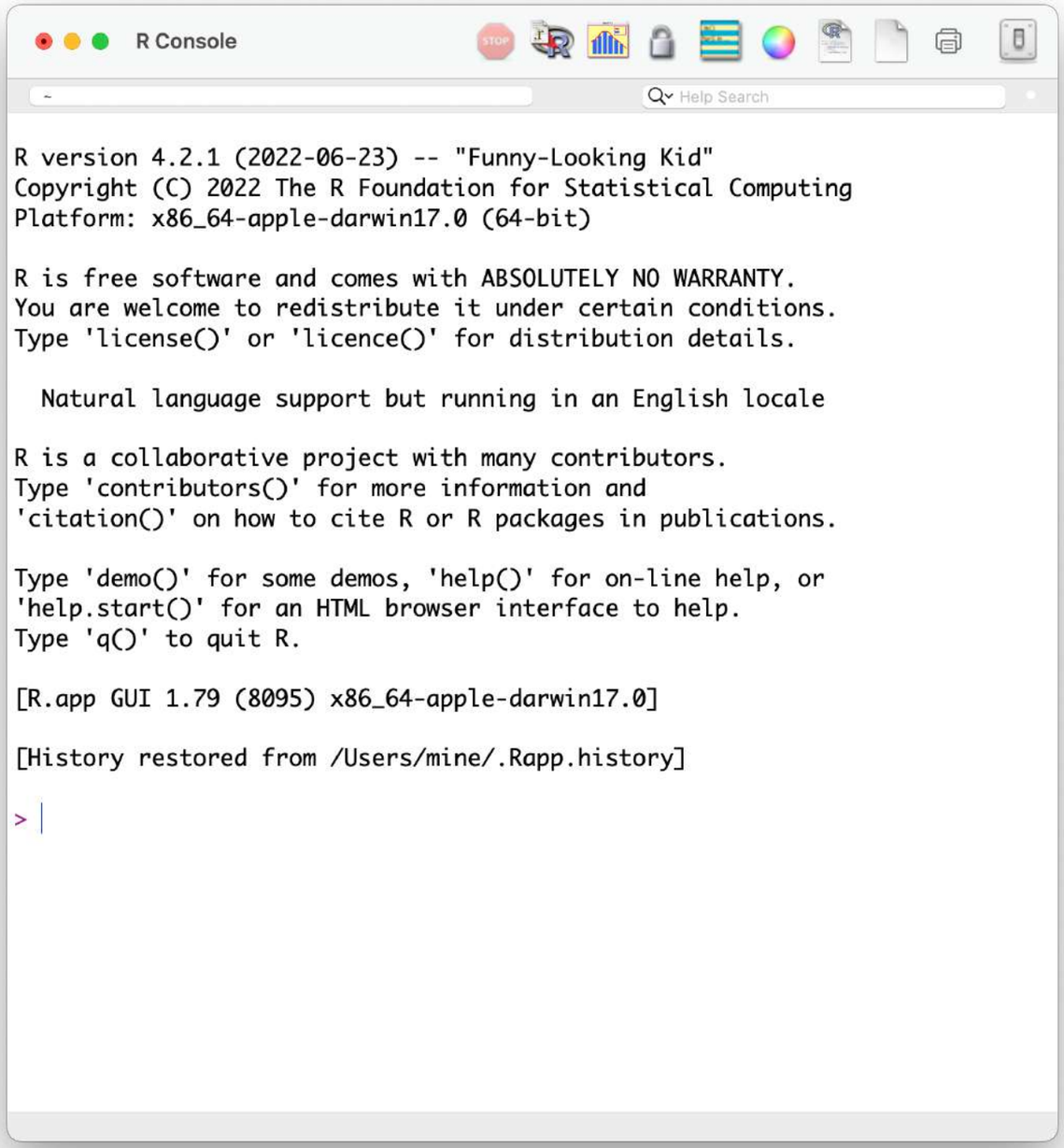
Course toolkit

Software – Excel? ❌

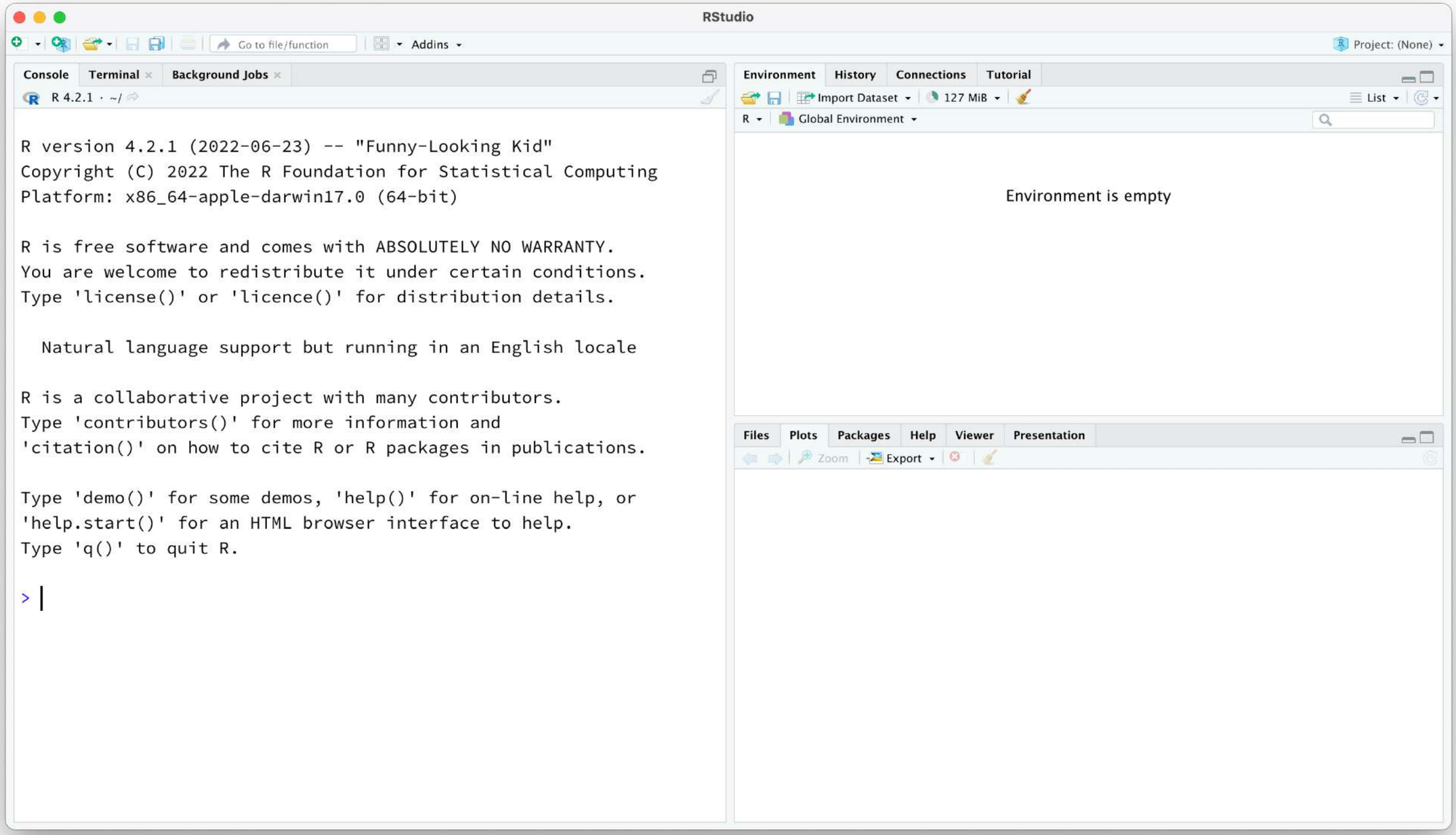


rcid	country	country_code	vote	session	importantvote	date	unres	amend	para	short
6	US	US	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Canada	CA	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Cuba	CU	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Dominican Republic	DO	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Mexico	MX	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Guatemala	GT	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Honduras	HN	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	El Salvador	SV	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Nicaragua	NI	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Panama	PA	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Colombia	CO	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Venezuela, Bolivarian Republic of	VE	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Ecuador	EC	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Peru	PE	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Brazil	BR	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Bolivia (Plurinational State of)	BO	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Paraguay	PY	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Chile	CL	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Argentina	AR	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Uruguay	UY	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	UK & NI	GB	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Netherlands	NL	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Belgium	BE	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Luxembourg	LU	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	France	FR	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Poland	PL	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Czechoslovakia	CS	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Yugoslavia	YU	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Greece	GR	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Russian Federation	RU	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Ukraine	UA	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Belarus	BY	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Norway	NO	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	Denmark	DK	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS

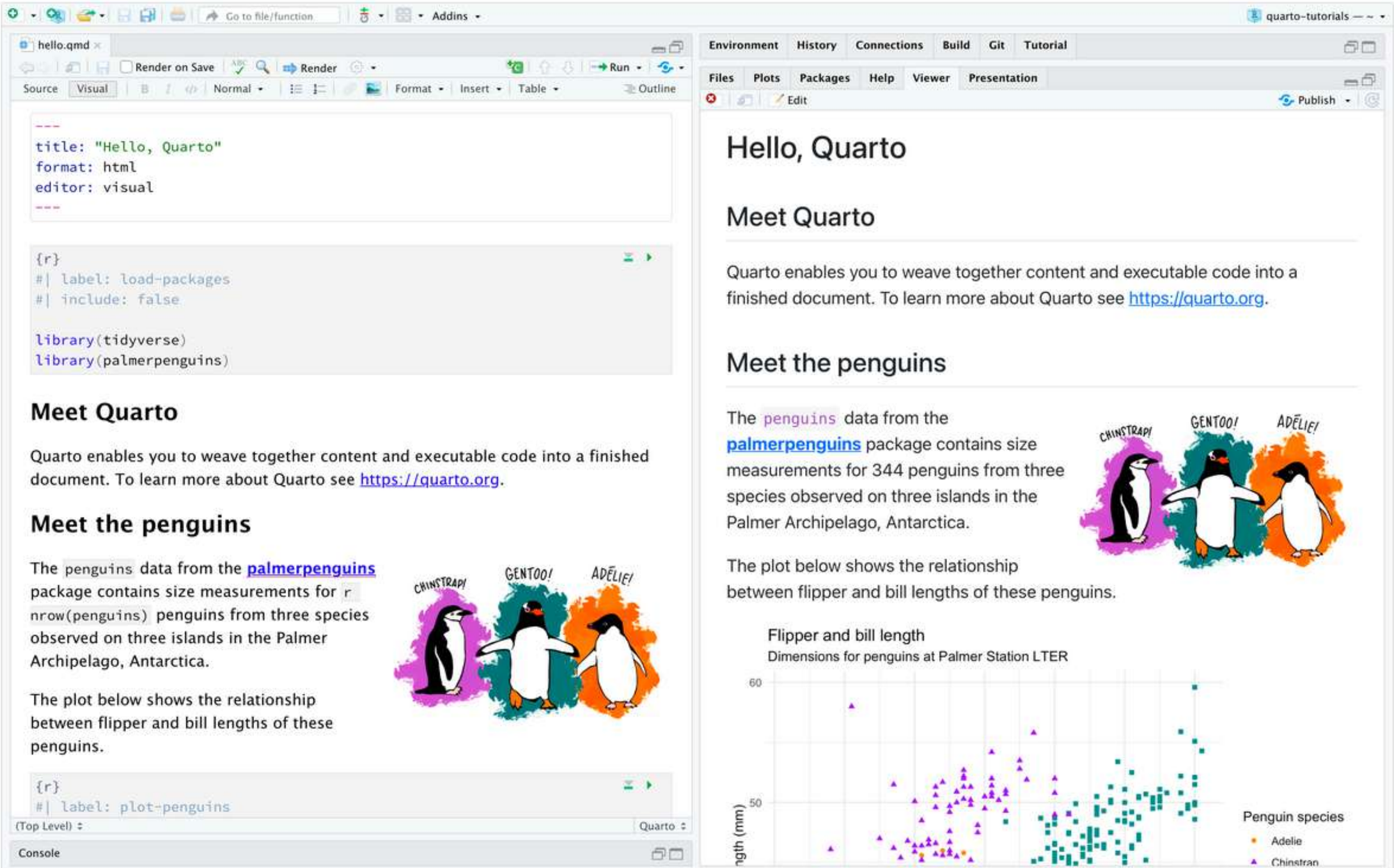
Software – R ✅


R Console
Help Search
R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid"
Copyright (C) 2022 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.
[R.app GUI 1.79 (8095) x86_64-apple-darwin17.0]
[History restored from /Users/mine/.Rapp.history]
>

Software – RStudio ✅



Software – Quarto ✅



Software

- Modern data science is unthinkable without **computer programming**: typically, either Python or R is used.
- For the purposes of this course, we will use:
  - The open-source statistical programming language R.
  - A bespoke integrated development environment (IDE) for R called RStudio.
  - An authoring framework for creating beautiful reports, presentations, web sites, etc., combining text, code, results and visualizations, called Quarto.
- Until next time, therefore please
  - either **install** R, RStudio and Quarto on your laptop (recommended) or
  - register for a free account at Posit Cloud.

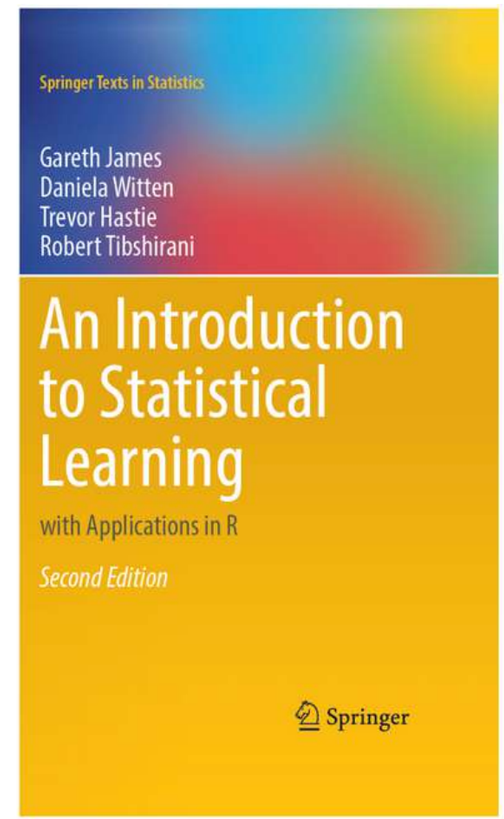
Resources

- **Primarily**: slides and exercises provided
- However, for a deeper dive and additional materials, I recommend:

For R programming

Excellent courses from Harvard Professor Rafael Irizarry, available for free here and here.

For machine learning



Excellent book, available for free here.

Resources – How about AI?

- With the large-scale adoption of AI tools like **ChatGPT**, the way data scientists work is rapidly changing.
- This course therefore **actively encourages** the use of AI tools for R programming. Here are some guidelines:
  - Use ChatGPT for **programming**, not for writing the project report.
  - Do **not** just copy-paste code generated by ChatGPT. Run it line-by-line, try to understand and edit as needed.
  - Engineer your prompts until the response starts to look like code you are learning in this course.
  - If the response is not correct, ask for a correction.
- With the arrival of AI, programming is becoming ever more **accessible**, but the need for people like you who actually **understand** the code they are running, is also increasing.

Resources – How about AI?