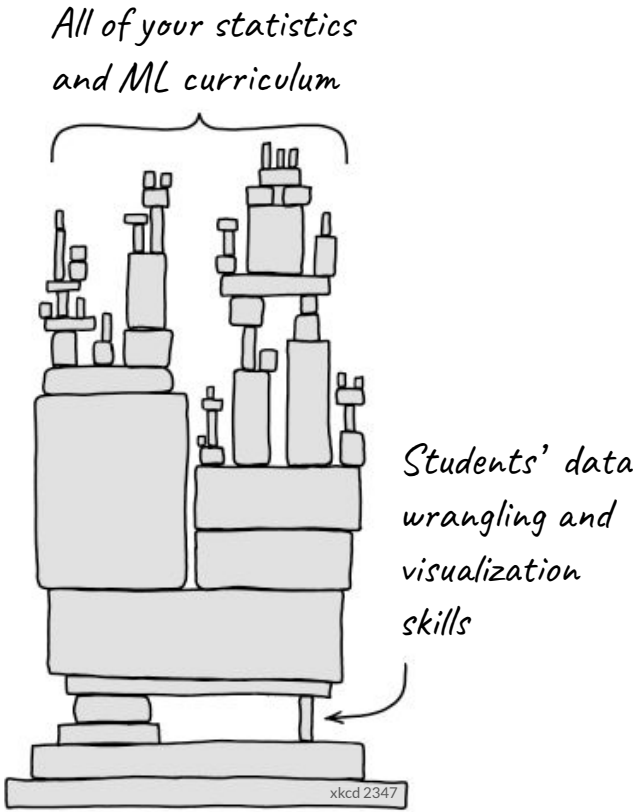
Coding with challenges

- **an effective path to learn data analysis**

Linda Hartman
Centre for Mathematical Sciences
Lund University

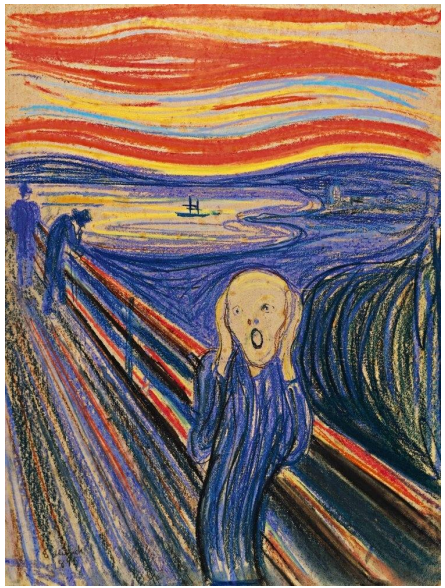
https://bit.ly/challenges-wcs

These slides →



Meet Johan

- Master Program in Energy sciences
- Asks for help with master project
  - 1000 apartments
  - **hourly** measurements of energy consumption for **5 years**
  - pre, during and post-COVID
- Excel
- Data wrangling
  - Manually removing erroneously entered values
  - Manually deleting apartments with >25 missing values
  - Needs to reshape the data, uses pivot tables
- Analysis
  - Needs to do basic modeling and report conclusions
  - Only knows Excel
  - Took Programming course in Java



Meet me

Linda Hartman
Senior Lecturer, Mathematical Statistics
Excellent Teaching Practitioner

and my teaching…

- **Applied statistics in science, engineering and medicine**
- **Machine learning**

Data Literacy

- Reproducible data analysis of authentic data

150

PhD 2007

Senior lecturer 2022

MEDICINSKA FAKULTETEN

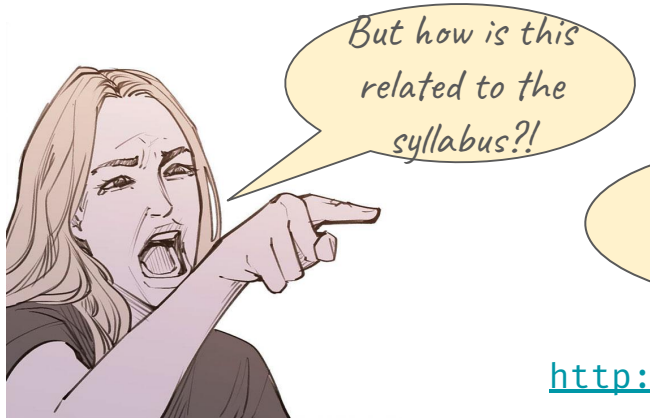
The Challenge

QUICK INTRO TO
R/PYTHON

The annual Sisyphus lecture

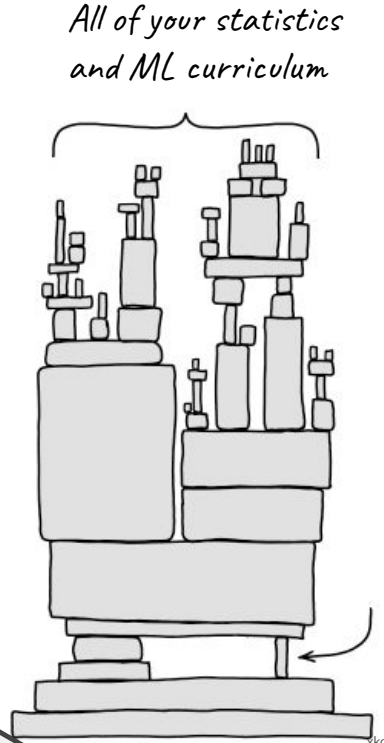
The tragedy of the commons

The weakest link



Let them eat cake (first)!

http://bit.ly/let-eat-cake



Course initiatives

- **Statistical learning: Data analysis and visualization**
  Data Literacy + ML basics (Regression + Classification)
  - R: tidyverse + caret + quarto
  - Bachelors: 124 students (6 hp)
  - Masters: 20 + 20 + 5 PhD students (7.5 hp)
  - Since 2024

- coursera **MOOC - Data Literacy with Python**
  - Software stack: uv, JupyterLab, plotnine, polars, quarto.
  - Fall 2026 (?)
  - Interactive?

- **Data Literacy with Python 3 hp**
  - General entry requirements
  - Data visualization + wrangling + reproducible reports (quarto)
  - MOOC + supervised practicals + assignment
  - First instance November 2025

Inspire the curiosity

Traditional Intro to R

```
1  num <- 2.2
2  class(num)
3  ## [1] "numeric"
4
5  char <- "hello"
6  class(char)
7  ## [1] "character"
8
9  logi <- TRUE
10 class(logi)
11 ## [1] "logical"
```

Intro to R - Chapter 3

Data analysis course

Creating the first plot

⚠ Do people in rich countries live longer than people in poor countries? What does the relationship between GDP per capita and Life expectancy look like? Is this relationship linear? Non-linear?

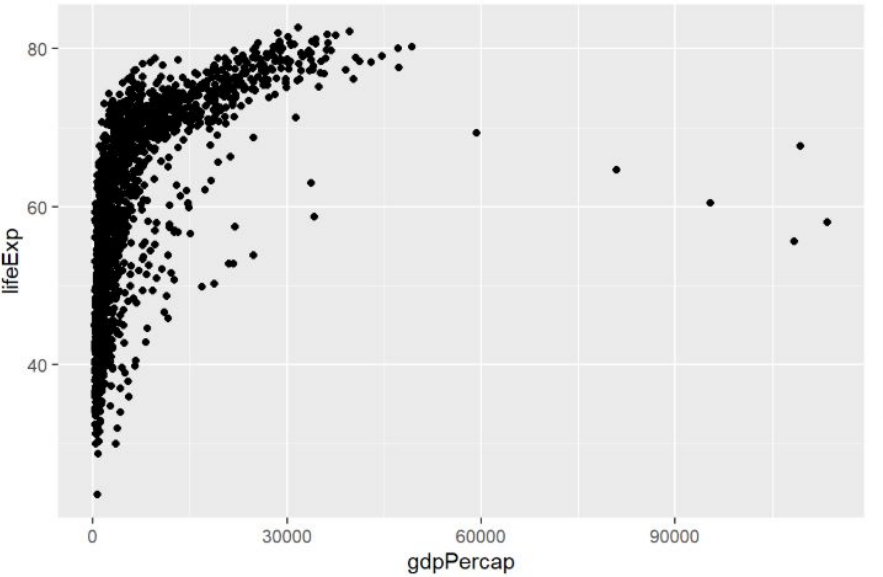
```
#
ggplot(data = gapminder) +
  geom_point(mapping = aes(x = gdpPercap, y = lifeExp))
```

First set of code in first practical

Active learning

1. Instruction

```
ggplot(data = gapminder) +
  geom_point(mapping = aes(x = gdpPercap, y = lifeExp))
```



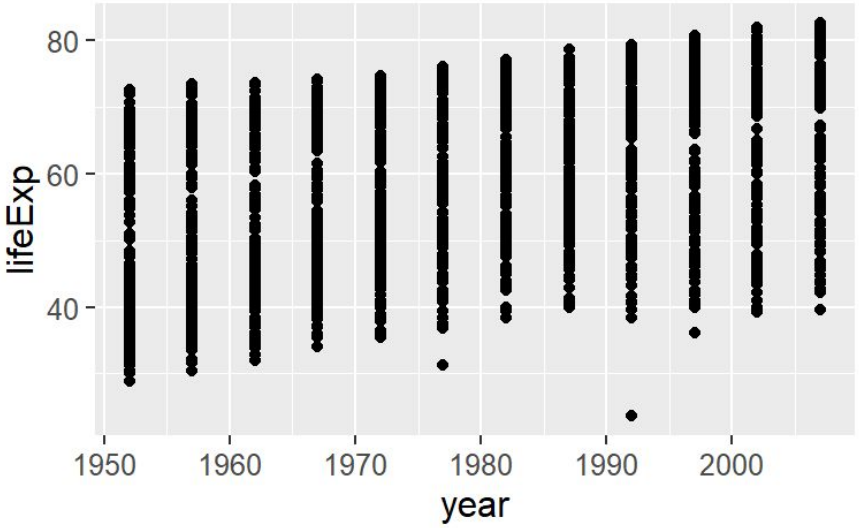
2. Frequent challenges

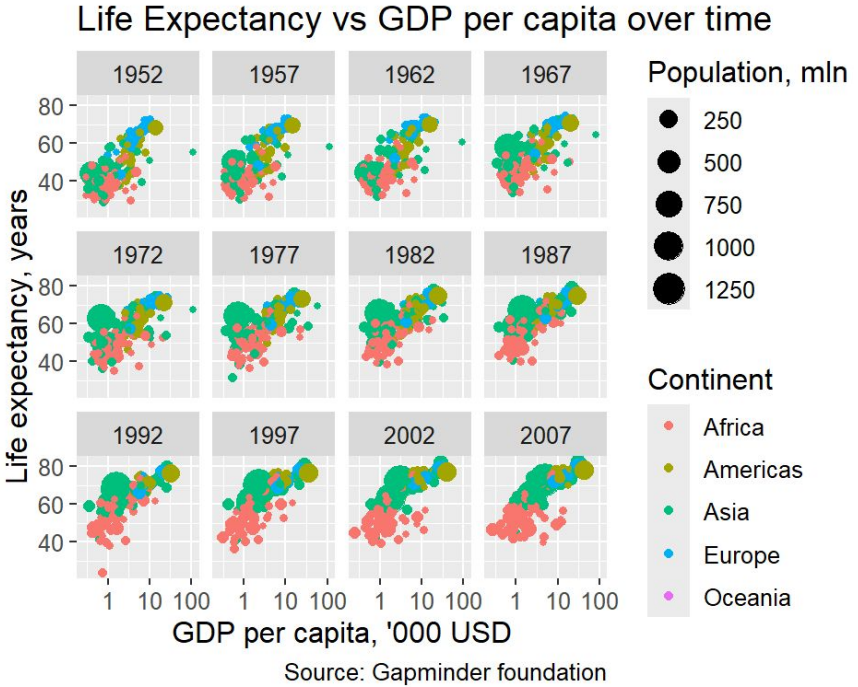
Challenge 1.

Assignment | Solution after class

- How did Life expectancy change over time? What do you observe? *Hint: the* `gapminder` *dataset has a column called* `year`*, which should appear on the x-axis.*
- See if you can visualize Life expectancy by continent. Which continent tends to have higher life expectancy? Which continent has highest spread in life expectancy values?

From zero to publication quality graph in first lab





The Solution

- The Achilles Heel of modern scientists.
- Click-along practicals, anyone?

Where do we start?

- **Pre-built** environment (Posit Cloud): data, libraries, first Quarto report etc.
- Aim for **curiosity and intuitions**, small steps in the zone of proximal development
- “Never hesitate to sacrifice truth for clarity”. (c) Greg Wilson, Rule 6, TTT
- Copy-paste-modify. Think aloud. Apprenticeship model. Helpers.

Unintimidate them!

- Empowered to take on (almost) any data and start visualizing.
- Learning to learn. Teach them how to get unstuck.

Student importing research data from Excel

Bing image creator, DALL-E, Microsoft Designer

Blue kryptonite skillset

- ★ **Data visualization**
- ★ **Subsetting and Aggregation**
- ★ **Pivots and Joins**
- ★ **Literate programming**
- ★ **Version control***

- Simple, yet expressive grammar of data visualization. Few will work at FT, everyone will write a thesis.
- Subsetting and sampling is important for big data.
- Right expectations for post-join data size.
- …

Bing image creator, DALL-E, Microsoft Designer

Rest of Data Analysis Course

FMSF86/90 course (6.0/7.5 ECTS)

- Data wrangling & visualization
- Train, Test & Cross-validation
- Regression
  - Shrinkage models
  - Reg Trees, Random Forest
- Classification
  - Shrinkage models
  - Random Forest, Boosted Trees

~125 Bachelor, 40 Master, 5 PhD students, 10 weeks

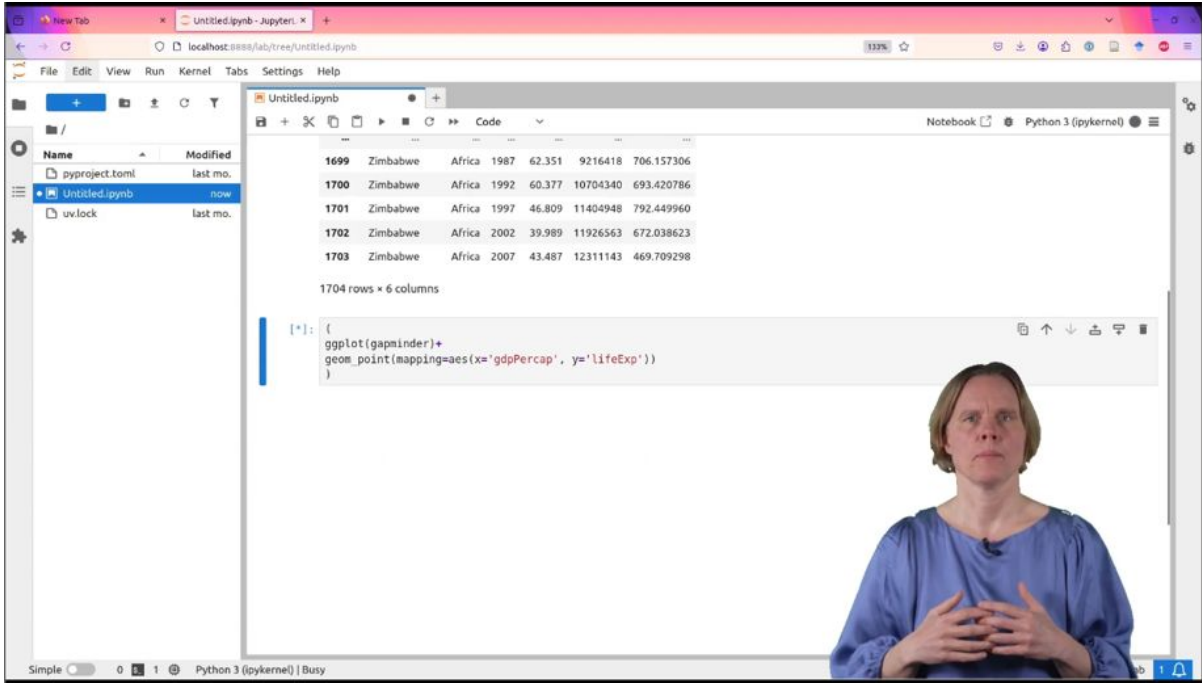
- Ames2 on Kaggle (private, in-class competition)
  - Started familiarizing students with it in practicals.
  - Beat the benchmark
  - Sweet and savory prizes for top-3 winners
- 3 (+1) Assignments
- Peer review

Bing image creator, DALL-E, Microsoft Designer

Data Literacy with Python 3 hp

- Data visualization + wrangling + reproducible reports (quarto)
- First course in programming = NO prerequisites (+ all faculties)

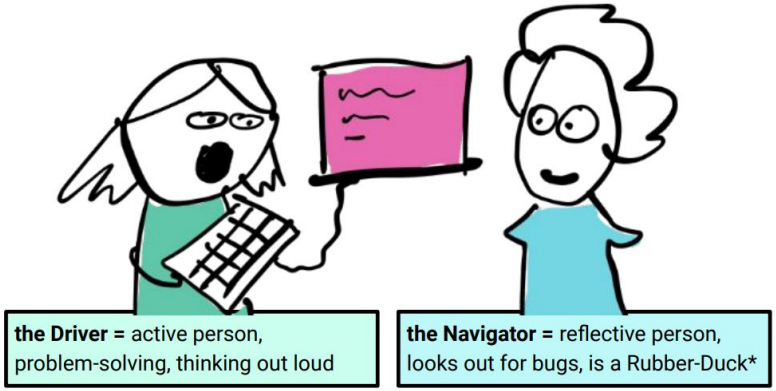
- MOOC videos as preparation to Practicals
  - Mimics Live-coding
  - Interspersed with challenges
- Practicals with Challenges
  - Pair-programming



Pair-programming

https://pairprogramming.ed.ac.uk/

Two Roles in Pair Programming and How It Works

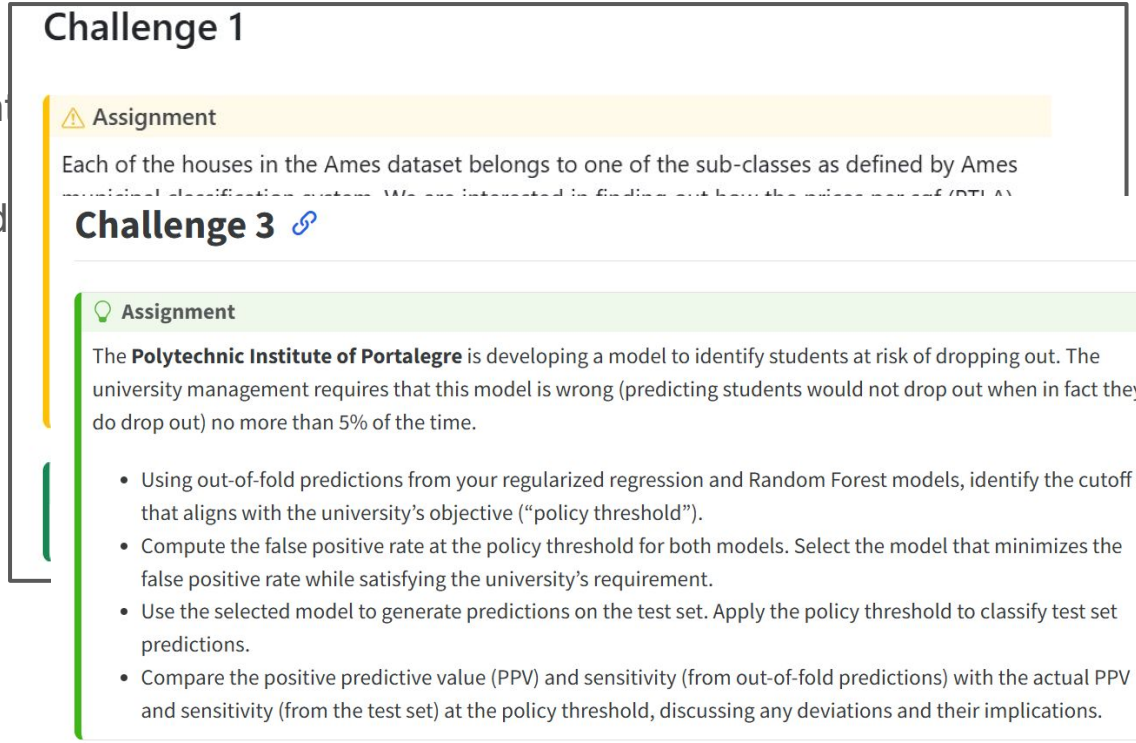


They switch roles every 10 minutes

I was skeptical about the practicals and the pair programming but they turned out to be the highlight of the course. I was nervous that everyone would be math and programming geniuses so I was pleasantly surprised to find out that the group was quite diverse and that we were all new to the task. The atmosphere in the practicals was really good with open-minded and friendly teachers/aides and a lot of concentration in the pairs programming together. Getting this

Assignments: Consolidate - and assess the skills

- Assignments
  - 3-4 challenges per assignment
  - Submit a reproducible report
  - Correct, relevant, documented and comprehensive thinking



Peer-review

- Struggled with AI and ping-pong submissions
- Decided to lean in to code sharing and demand it
- Grade the review first and final assignment second.
- Leverage peer pressure, smoke out (mis)use of AI
- If failed in final submission: New assignment

AI

- AI Statement:
  OK to use AI, but reference what you used it for!
- Keep to course code or explain why
- Discussion of using AI to aid thinking, not avoid thinking
- Peer-review instructions: look out for misuse of AI

Bing image creator, DALL-E, Microsoft Designer

Things we wish we did

- Going deeper in Data Literacy: Ethics, data quality, bias
- Self-paced exercises (with feedback) as preparation or reinforcement
- Stop separating lectures from practicals
- Video code alongs as preparation + pair programming in class for later parts of Data analysis course
- More formative assessments activating students' theoretical knowledge prior to practical session
- In-class exam (rewarding)!

Thanks!



Dmytro Perepolkin

https://bit.ly/challenges-wcs
linda.hartman@matstat.lu.se

Cake Images https://nomsmagazine.com/most-challenging-cakes/

References

- **Mine Çetinkaya-Rundel**
  https://mine-cr.com/ eg.
  - **Çetinkaya-Rundel, M., & Ellison, V. (2020). A Fresh Look at Introductory Data Science.**
  - **http://bit.ly/let-eat-cake**
- **Greg Wilson**
  Teaching Tech Together
  https://teachtogether.tech/
- **Pair programming:**
  https://pairprogramming.ed.ac.uk/