Data Handling: Import, Cleaning and Visualization

Wrap up, Q&A, Exam info, Feedback

Aurélien Sallin, PhD

Updates

Decentral exam

Tonight, 16:15, 01-113, BYO + cord

Plan for today

1. Recap
2. Exam Info
3. Q&A
4. Suggested Improvements
5. Practical case with real data
6. Happy Holidays! 🎄

Recap Visualisation

Data visualization

Two ways: display data through **tables** or **graphs**.

Depends on the purpose.

Grammar of Graphics/`ggplot2`

- The **`ggplot2`** package (Wickham 2016).
- … an implementation of Leland Wilkinson's 'Grammar of Graphics'.

ggplot2 basics

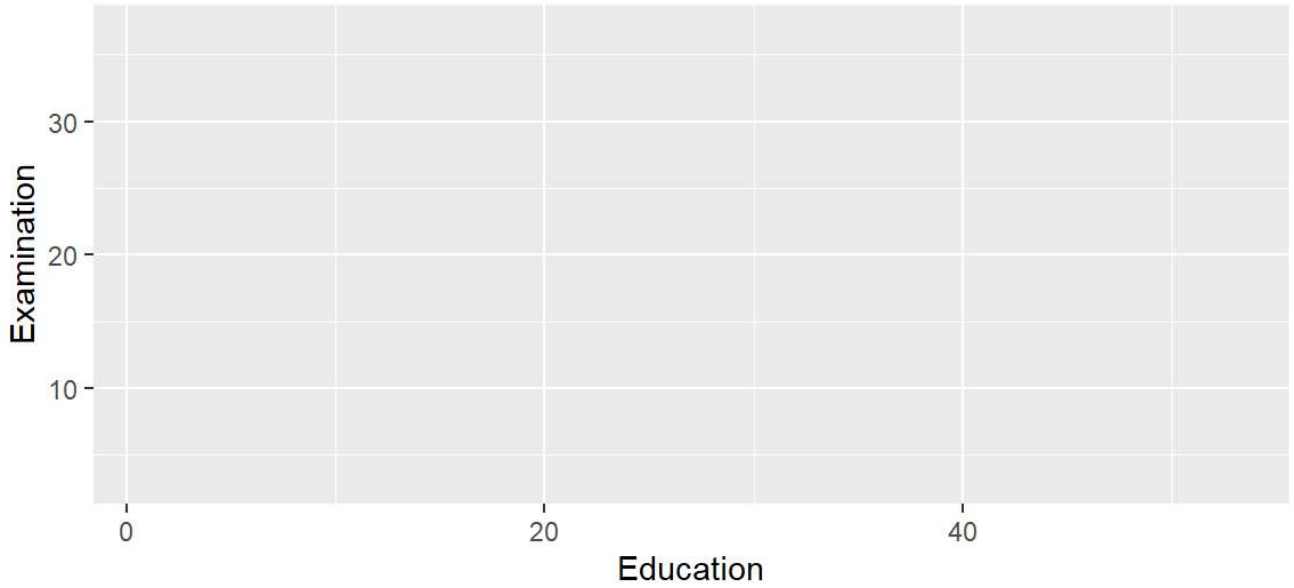
Using **ggplot2** to generate a basic plot in R is quite simple. Three key points:

1. The data must be stored in a **data.frame**/**tibble** (in tidy format!).
2. The starting point of a plot is always the function **ggplot()**.
3. The first line of plot code declares the data and the 'aesthetics' (e.g., which variables are mapped to the x-/y-axes):

```
ggplot(data = my_dataframe, aes(x= xvar, y= yvar))
```

ggplot2: building plots layer by layer

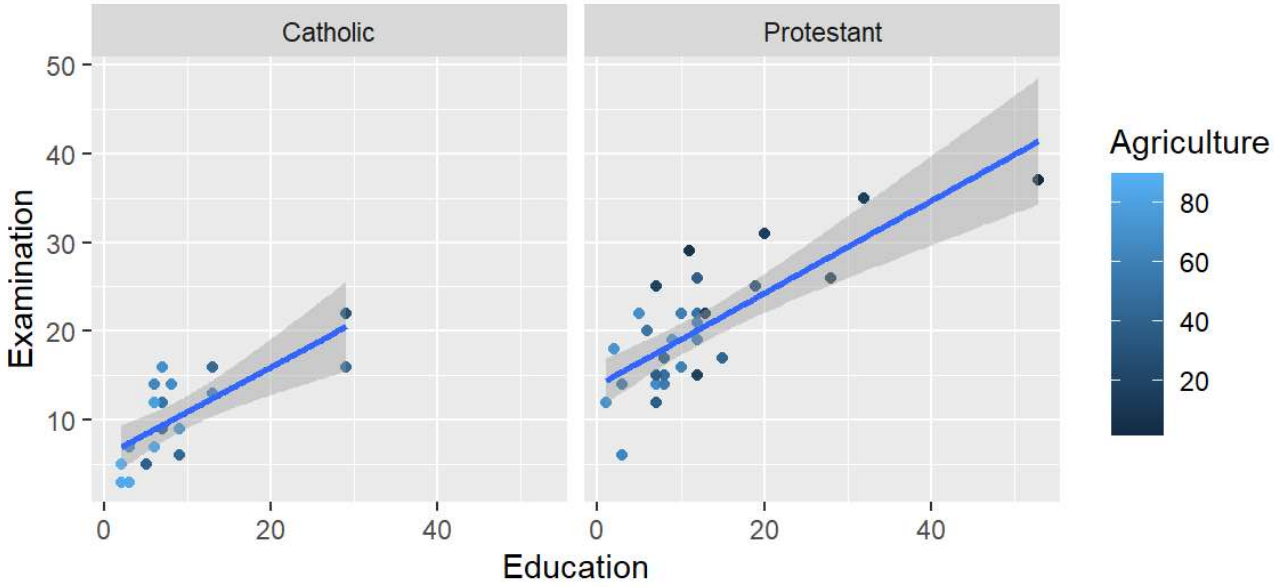
```
ggplot(data = swiss, aes(x = Education, y = Examination))
```



ggplot2: building plots layer by layer

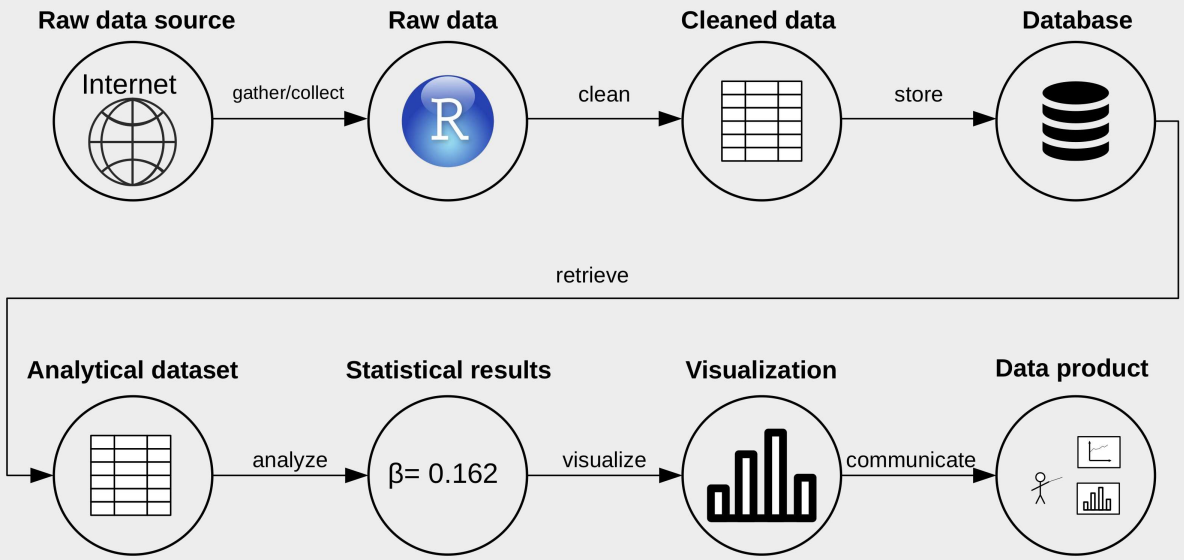
```
ggplot(data = swiss, aes(x = Education, y = Examination)) +
  geom_point(aes(color = Agriculture)) +
  geom_smooth(method = 'lm') +
  facet_wrap(~Religion)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Wrap up

Data (science) pipeline



Wrap up: 'Theory'/Conceptual part

- Understand the very basics of how computers process data.
  - Binary code.
  - Representation of binary code as text (encodings/standards!).

Wrap up: 'Theory'/Conceptual part

- Understand the very basics of how computers process data.
  - Binary code.
  - Representation of binary code as text (encodings/standards!).
  - Text files to store data: special characters (comma, semicolon, etc.) define the structure (following a standard!)
    - CSV: two-dimensional/table-like structure
    - JSON/XML: hierarchical data (high-dimensional)

Wrap up: 'Theory'/Conceptual part

- Understand the very basics of how computers process data.
  - Binary code.
  - Representation of binary code as text (encodings/standards!).
  - Text files to store data: special characters (comma, semicolon, etc.) define the structure (following a standard!)
    - CSV: two-dimensional/table-like structure
    - JSON/XML: hierarchical data (high-dimensional)
  - Data in text file: how is data structured when stored on the hard disk (mass storage device).
  - Data structures in R (objects): How data is structured/represented when loaded into RAM (via a 'parser')

Wrap up: Applied part: import, cleaning, analysis/visualisation.

- How to get from the data source to the final data product.
- `tidyverse`: tools to help you with every part of the data pipeline.

Wrap up: Applied part: import, cleaning, analysis/visualisation.

- How to get from the data source to the final data product.
- `tidyverse`: tools to help you with every part of the data pipeline.
- How to import data into R? What to do if the parsing fails?
- How to clean/prepare data in R? Aim: tidy data set (rows:observations, columns:variables).
- How to filter for specific observations? How to select a set of variables/columns?
- How to modify/add variable?
- How to compute basic summary statistics?
- How to visualize raw data and basic statistics?

Exam

‘Code questions’

- ❗ Exactly same style and structure as quizzes and mock exam ❗
- We do not invent wrong function parameters or misspell function names etc to mislead you!
- ‘Passive’ knowledge of key R syntax and most important functions of core lecture contents are important!

‘Code questions’

- Work step-by-step through the code example. What happens in each parenthesis, on each line?
- From the inner part to the outer part, from top to bottom.
- Use pencil/paper to keep track of values, data structures (how does a data-frame look like?), classes.
- Importantly, if a code-example question refers to specific functions, you can always assume that the corresponding package is also loaded.

Question types

- MC questions type A (multiple correct)
- MC questions type B (only one correct)
- T/F
- Essay/open questions

Q&A

Questions?

Course Evaluation

Course Evaluation

Ongoing! See link on our course's StudyNet/Canvas page.

Suggested Improvements

Improvements

- **Learning Progress**
  - For some high, for some low (dsf program). Suggestions?

- **More quizzes**?
  - 😵

Improvements

- **More programming**
  - "We don't really need programming for the exam, so we don't learn the skill as much as we could."
  - "More R Programming during the course and more material on Canvas for Straight–Up R programming"
  - **Suggestions?** (programming tasks? term project?)

- **Exercises**
  - "Confusing" and not related enough to the course.
  - My idea: course complement (scrape your own webpage, build your own Markdown/Shiny/Streamlit app, work with an API)

Real case data

Real case data

In recent years, and especially in the post-COVID time, the psychological health of the Swiss population has deteriorated. This is especially true for people in the younger age groups, especially for women. For more information, see OBSAN 2023 (Obsan_03_2023_BERICHT.pdf (admin.ch)).

Real case data

You are interested in understanding the financial consequences of deteriorating psychological health. You will use health insurance claim data from SWICA, a leading Swiss health insurance.

Real case data

You receive two datasets containing aggregated data:

- **“stamm” data**: this dataset contains basic information about the population insured in the base mandatory health plan at SWICA between 2020 and 2021.

- **“medi” data**: this dataset contains the aggregated number of psychological consultations and related costs.

Use these data to answer the following questions.

Note: these data are confidential. They cannot be used outside of this classroom. They have been modified to preserve confidentiality.

Final Remarks

Final Remarks

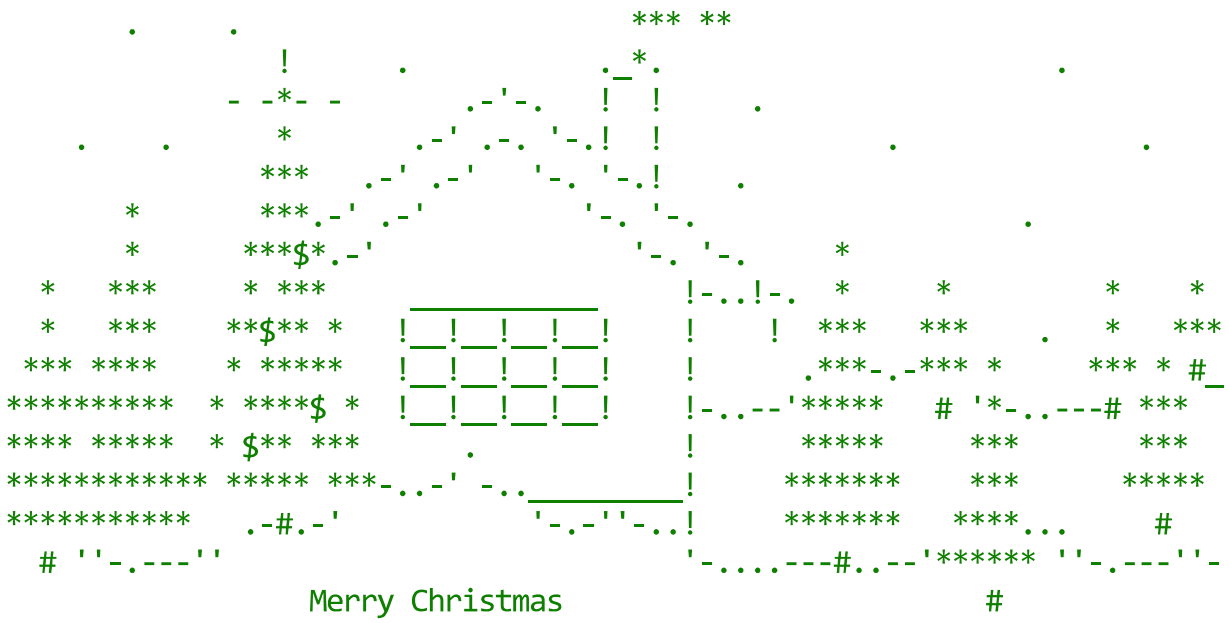
- I had a lot of fun 🙏
- I'll be happy to keep in touch (LinkedIn)
- All the best for your exams! 👍

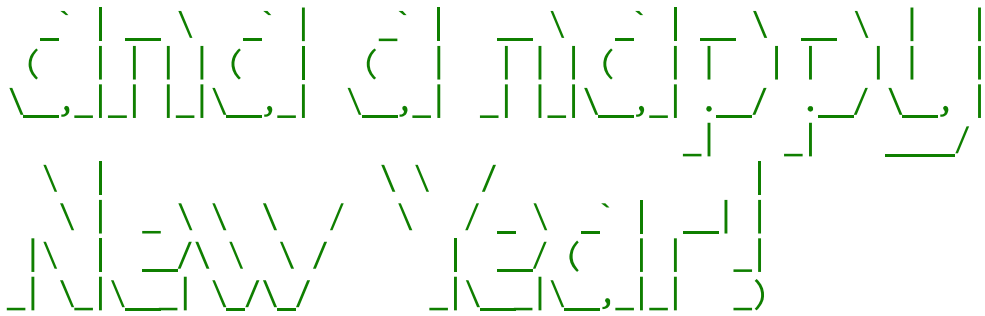
Final Remarks

- I had a lot of fun 🙏
- I'll be happy to keep in touch (LinkedIn)
- All the best for your exams! 👍
- All the best for your studies and careers!

Final Remarks

- I had a lot of fun 🙏
- I'll be happy to keep in touch (LinkedIn)
- All the best for your exams! 👍
- All the best for your studies and careers, **and finally, of course, ...**


Merry Christmas



References

Wickham, Hadley. 2016. **Ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York. http://ggplot2.org.