Code & Data Handling in Empirical Research

Could We Do Better?

Tobias Witter

HUB, TRR 266

Februar 09, 2022

Part 1: An Integrated View on Empirical Research Projects

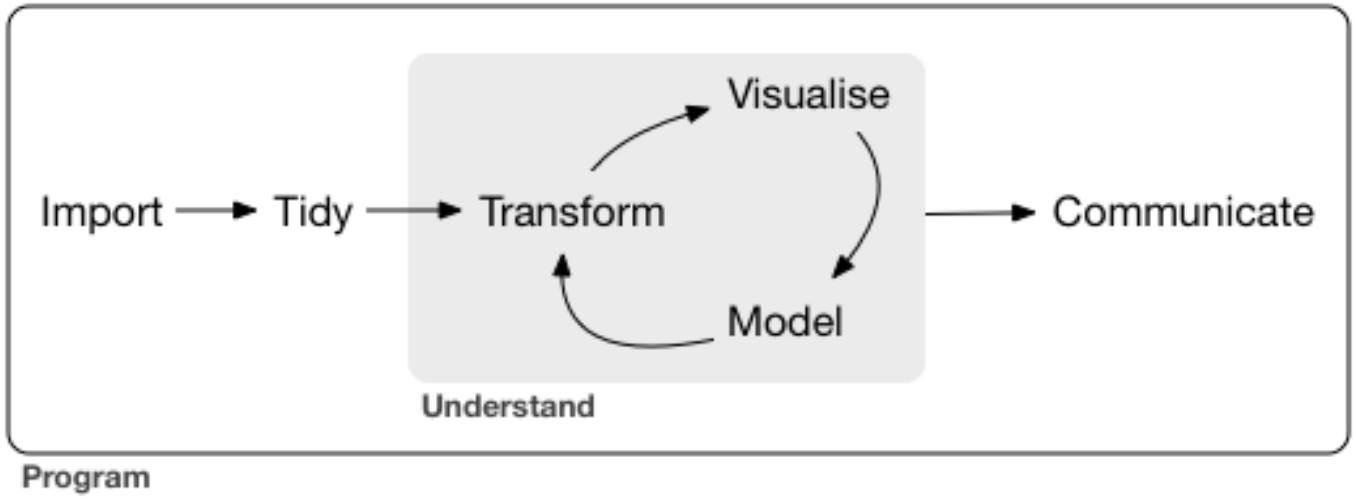
Motivation: Mental Juggling with Research Projects

Empirical Research Projects

1. Retrieve/collect raw data
2. Import raw data
3. Tidy raw data
4. Transform (raw) data
5. Visualize transformed data (= tables, figures, statistics)
6. Model: Explore, describe, causally test for relationships between variables
7. Communicate

Should Empirical Researchers look at Data Scientists and Programmers?

Let's look at advice to Data Scientists says (Wickham and Grolemund 2017):

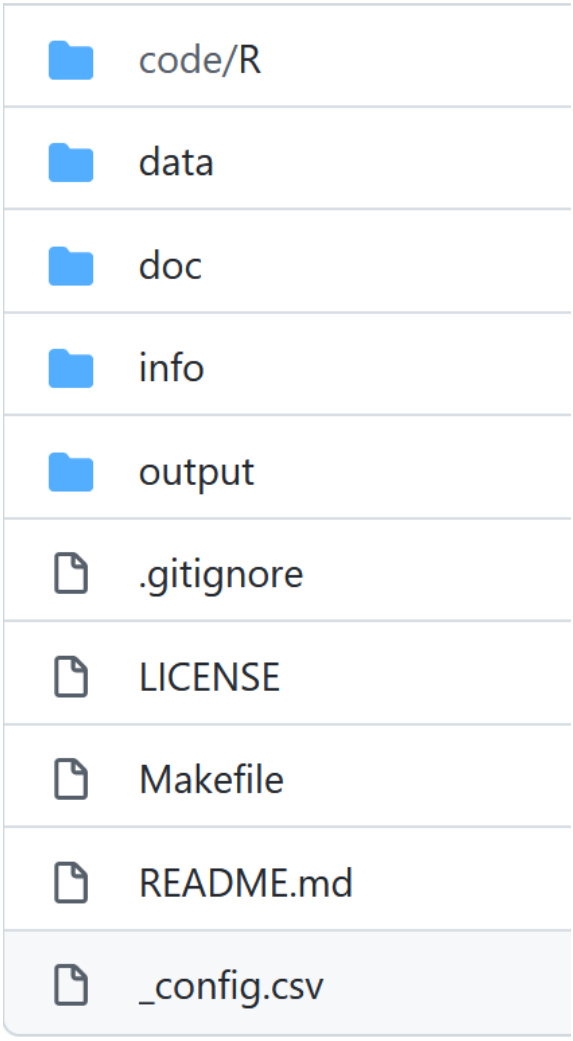


Software Developers as new best buddies?

Python File Structure Tree

```
Project_Name/
├── .git/
├── .vscode/
├── data/
├── debug/
├── docs/
├── etc/
├── include/
│   └── Project_Name/
│       └── public_functions.py
├── lib/
├── src/
│   ├── assets/
│   │   ├── fonts/
│   │   ├── images/
│   │   ├── sounds/
│   │   └── videos/
│   ├── utils/
│   ├── functions_code.py
│   ├── main.py
│   └── private_funtions.py
├── tests/
│   ├── alpha_version/
│   └── beta_version/
├── Python.gitignore
└── ReadMe.md
```

The TRR 266 template:



Software Developers as new best buddies?

- Version Control Systems
  - GitHub
  - Nextcloud

Part 2: Code

Code Organization

- Code automation
- Re-using code
- Product-oriented programming
- Functional instead of object-oriented programming

Part 3: Data

Data Handling

- Automated data retrieval
- Product-orientation
- Have tidy data! (Wickham 2014)

Tidy data

In tidy data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Part 4: Looking at Examples

Code Examples

- ...
- ...

Data Examples

- ...
- ...

Conclusion

- ...
- ...

Resources

- The Python File Structure Tree was taken from AlexDCode (2020).
- The TRR 266 Template for Reproducible Empirical Accounting Research is available from TRR 266 (2021).
- Free data science resources: https://github.com/alastairrushworth/free-data-science
- Read about tidy data here: Wickham (2014)
- Licensing of code and data: Creative Commons (n.d.)

Goals and characteristics of research templates

1. avoid confusion
2. as simple as possible
3. keep your code clean, neat, structured, and clutter free
4. file structure system is modular
5. each folder has explanation
6. more documentation in the folder itself
7. hierarchical tree file organization system
8. standard for small to medium size projects
9. ...

Software Developers as new best buddies?

Do's and Don'ts (adapted) from a large survey of software developers (datree.io 2019):

1. Don't mingle code that is under development with production-ready code
2. Don't commit code as an unrecognized author (always share you identity)
3. Define code owners for faster code reviews (who is responsible?)
4. Don't leak secrets into shared code (protect your passwords)
5. Don't commit dependencies into source control
6. Don't commit local config files into source control
7. Create a meaningful git ignore file
8. Archive dead repositories
9. Lock package version
10. Specify standard package versions
11. Leverage task list
12. Use a branch naming convention
13. Delete stale branches
14. Keep branches up to date
15. Remove inactive GitHub members
16. Enable security alerts

References

AlexDCode. 2020. "Software Development Project Structure: A Template for Different Programming Languages." https://github.com/AlexDCode/Software-Development-Project-Structure.

Creative Commons. n.d. "Share Your Work: The Creative Commons License Generator." https://creativecommons.org/share-your-work/.

datree.io. 2019. "Top GitHub Best Practices Guide for Developers [Expanded Dec 2019]." https://www.datree.io/resources/github-best-practices.

TRR 266. 2021. "The TRR 266 Template for Reproducible Empirical Accounting Research." https://github.com/trr266/treat.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59: 1--23. https://doi.org//url%7Bhttps://doi.org/10.18637/jss.v059.i10%7D.

Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science*. O'Reilly. https://doi.org/https://r4ds.had.co.nz.