2022

Reproducible Practice in Taming the Wild Data

Presented in Toronto Workshop on Reproducibility

Dewi Amaliah
Monash University

Acknowledgment

This work is done in collaboration with:

- Dianne Cook (Monash University)
- Emi Tanaka (Monash University)
- Nicholas Tierney (Telethon Kids Institute), and
- Kate Hyde (Monash University)

Traits of Ideal Data

Data should satisfies 3R (Kim, Ismay, Chunn, 2018):

1. **Rich** enough to answer meaningful questions;
2. **Real** enough to ensure the existence of a context;
3. **Realistic** enough to convey that pre-processing is often needed;

On another side:

> For learning and teaching of statistics/data science-purposes, or textbook's data, the prerequisite to analyze the data should ideally be minimum (Cobb, 2015 in Kim, Ismay, Chunn, 2018).

Tame vs Wild Data

Wild data

The 3rd R of ideal data.

Tame data

Minimum prerequisite.

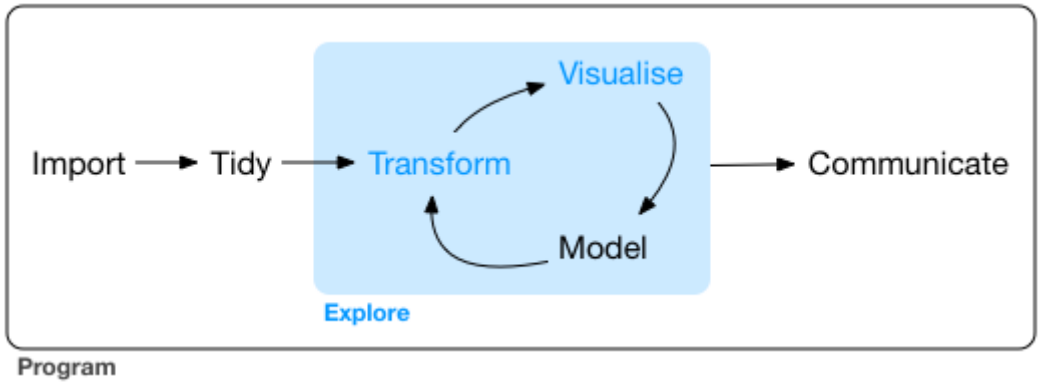
🎯 The Goal

Dataset that is rich, real, tidy, clean.

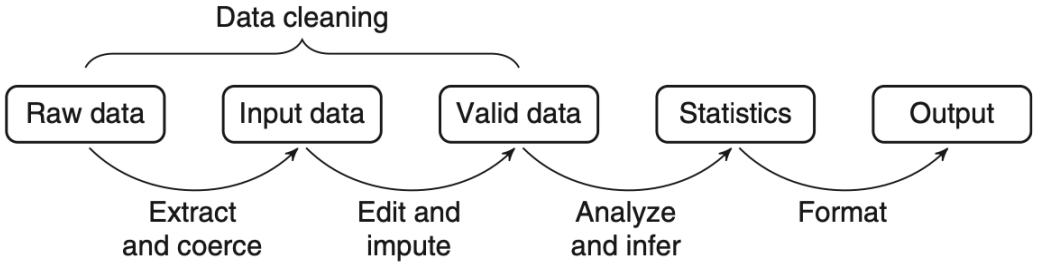
Tame data

Taming the Wild Data: Works to Do

Data science pipeline (Wickham & Grolemund, 2017)

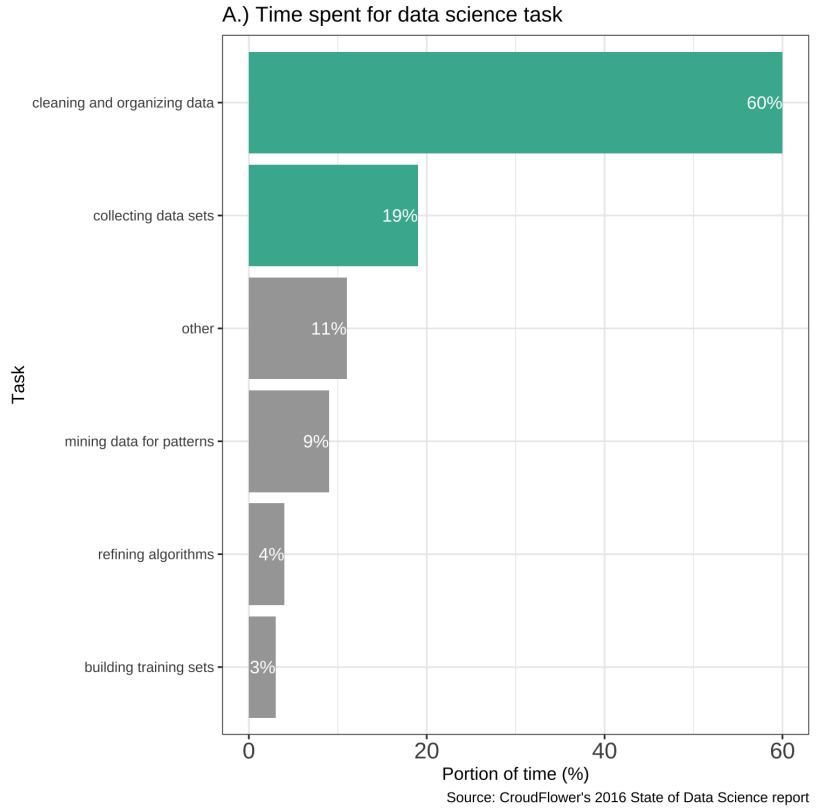


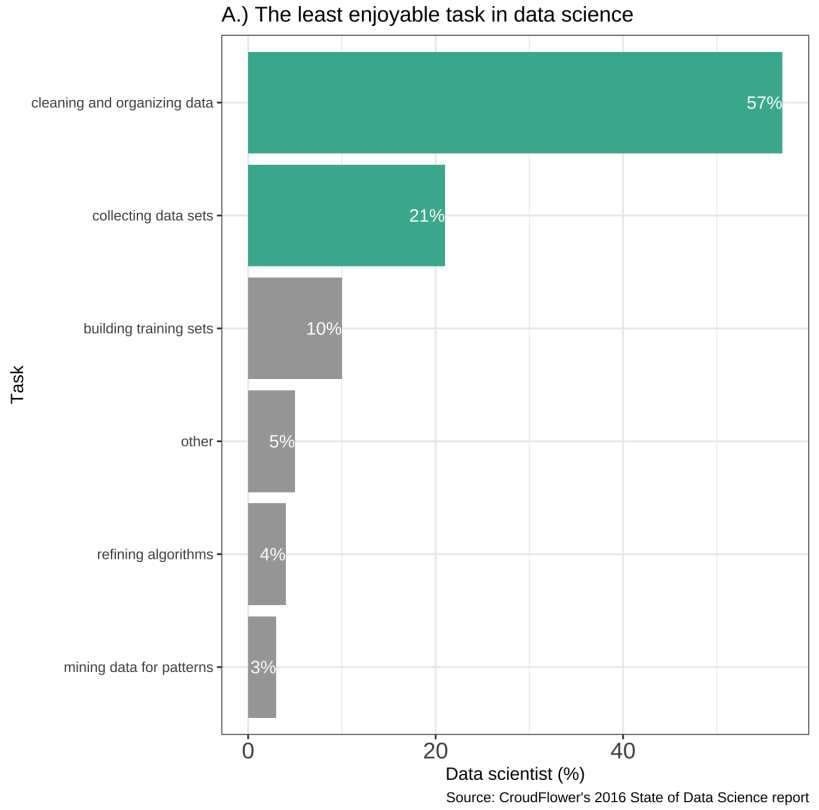
Statistical value chain (van der Loo & de Jonge, 2018)



What is the Problem?

Most of the time is spent for data preparation, yet it is the least enjoyable task.





Another Problem

"Data cleaning and preparation are often neglected or disorganized and decision made during these steps are often unreported." (Huebner, Vach, le Cessie, 2016)

Lack of Reproducibility

Case Study: yowie

Stands for **Y**ears **o**f **W**ages to **I**nvestigate and **E**xplore

An R package contains longitudinal wages datasets and several demographic variables of the National Longitudinal Survey of Youth 1979 cohort. The period covered is 1979-2018.

The aim is to provide longitudinal data that is suitable for learning and teaching longitudinal data with **reproducibility and transparency emphasized**.

3 datasets in the package:

- Demographic data of the NLSY79 cohort.
- Longitudinal wages data of the NLSY79 cohort.
- Longitudinal wages data of high school dropouts subset of the NLSY79 cohort (the refreshed version of Singer & Willet's (2003); This data is only from 1979 - 1994)

Why wages data from NLSY79?

- NLSY79 covers various variable and has been used in research in various discipline (Pergamit at al., 2001).
- The survey is carefully designed with high retention rates so it suitable for life course research (Pergamit et al. 2001; Cooksey 2017).
- The data has been widely used in textbook, including in Singer and Willet (2003). It has been used in teaching of exploratory data analysis. This data is also used as example data in `brolgar` (Tierney, Cook & Prvan, 2020)

Variables included in the original data

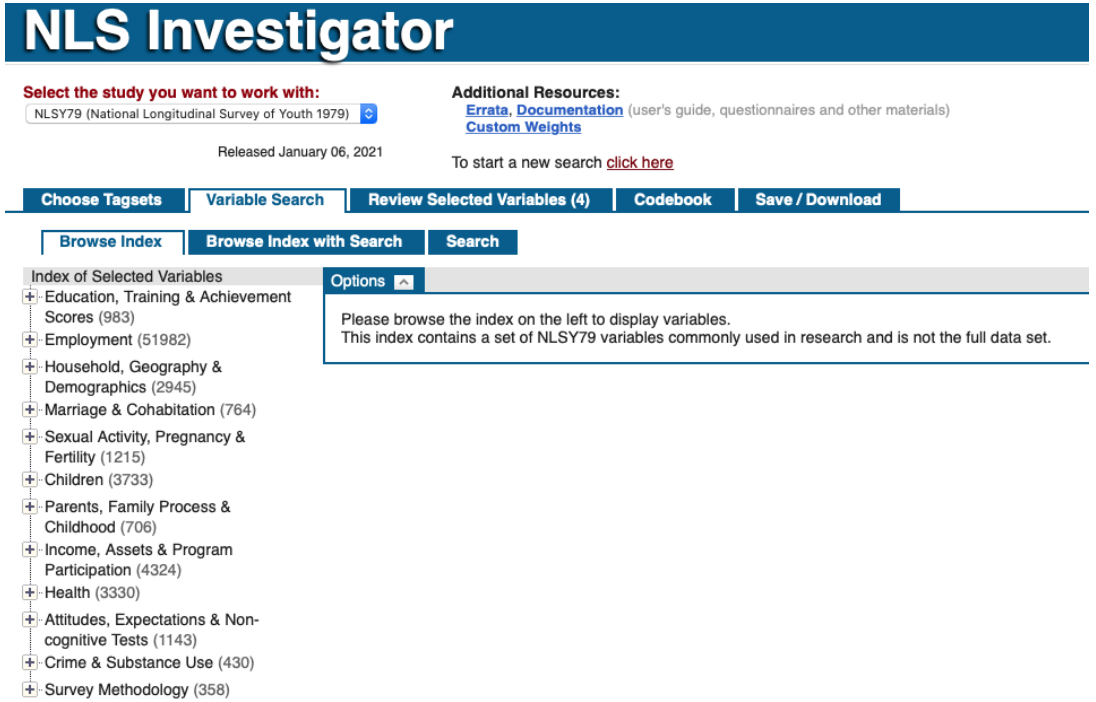
The original data (Singer and Willet, 2003) covers the variables of ID of the data, wages, work experience, GED status, highest grade completed, dummy variable of whether the ID belongs to Black or Hispanic group, and the unemployment rate.

Variables included in the refreshed data

The same variables with the original data (except the unemployment rate) with some other variables to enrich the analysis (e.g., gender, year when start working) and to improve transparency (whether the data is imputed or not).

Getting the Data

The data is publicly available in the NLS Investigator website.



😊 We can create a tagset which contains the variables name -> 👍 for the reproducible workflow.

😕 Some variables in the original data are not explicitly available in the database -> need 🕵️ work.

Tidy the Data

The downloaded NLSY79 data is not tidy -> The 3rd "R" in the criteria of ideal data.

```
'data.frame':   12686 obs. of  742 variables:
 $ CASEID_1979                              : int  1 2 3 4 5 6 7 8 ...
 $ HRP1_1979                                : int  328 385 365 NA 310 NA NA NA ...
 $ HRP2_1979                                : int  NA NA NA NA 375 NA NA NA ...
 $ HRP3_1979                                : int  NA NA 275 NA NA NA NA NA ...
 $ HRP4_1979                                : int  NA NA NA NA NA 250 NA NA ...
 $ HRP5_1979                                : int  NA NA NA NA NA NA NA NA ...
 $ HRP1_1980                                : int  NA 457 397 NA 333 275 300 394 ...
 $ HRP2_1980                                : int  NA NA 367 NA NA NA NA NA ...
 $ HRP3_1980                                : int  NA NA 380 NA NA NA 290 NA ...
 $ HRP4_1980                                : int  NA NA NA NA NA NA NA NA ...
  [list output truncated]
```

- Done by using `tidyverse` (Wickham, Averick, et al. 2019) to pivot longer the data, rename the variables, and change the data type with appropriate data type.
- The wages data frame is saved in `tsibble` (Wang, Cook, Hyndman, 2020), a data frame class that is suitable for temporal data.

🕵️ On the Quest for `experience`

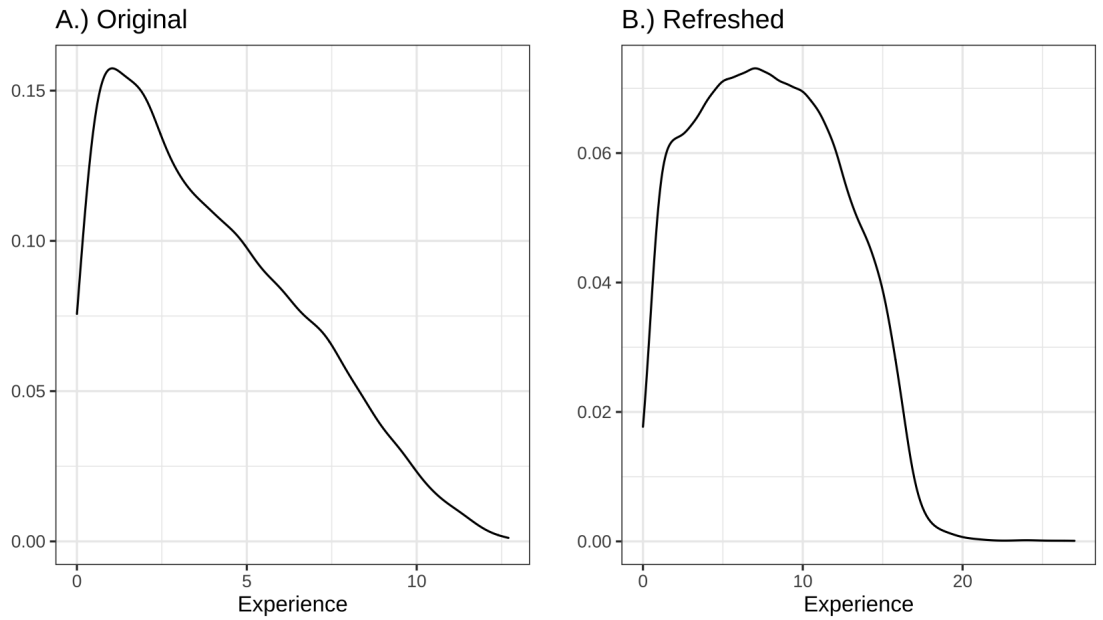
Problems | Comparison | Try it again

- Singer and Willet (2003) use `experience` as the time index.
- This variable is not explicitly available in the database **->** calculated variable.
- There is no code or explicit explanation available in the book on how `experience` is calculated from the raw data.
- The only explanation available is the length of time (in years) since entering the labor force, with t_0 for each subject starting on their first day at work.

🕵️ On the Quest for `experience`

Problems | **Comparison** | Try it again

- Recreate the variable based on the definition available, calculated from the year of individual started working (available in the database).

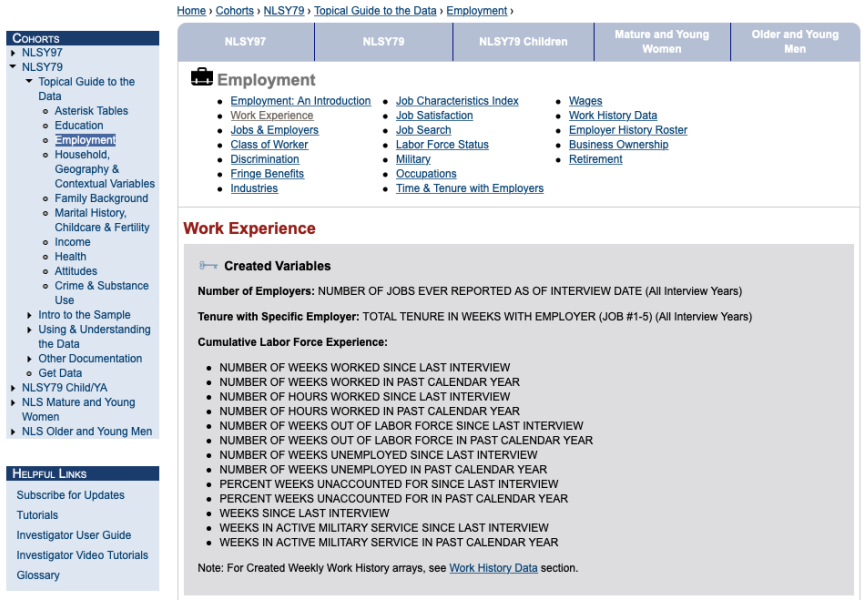


✋ Does not seems match.

🕵 On the Quest for `experience`

Problems | Comparison | **Try it again**

From the topical guide of NLSY79 data we find:



COHORTS
- NLSY97
- NLSY79
  - Topical Guide to the Data
    - Asterisk Tables
    - Education
    - Employment
    - Household, Geography & Contextual Variables
    - Family Background
    - Marital History, Childcare & Fertility
    - Income
    - Health
    - Attitudes
    - Crime & Substance Use
  - Intro to the Sample
  - Using & Understanding the Data
  - Other Documentation
  - Get Data
- NLSY79 Child/YA
- NLS Mature and Young Women
- NLS Older and Young Men

HELPFUL LINKS
Subscribe for Updates
Tutorials
Investigator User Guide
Investigator Video Tutorials
Glossary

Home › Cohorts › NLSY79 › Topical Guide to the Data › Employment ›

NLSY97 | NLSY79 | NLSY79 Children | Mature and Young Women | Older and Young Men

Employment

- Employment: An Introduction
- Work Experience
- Jobs & Employers
- Class of Worker
- Discrimination
- Fringe Benefits
- Industries
- Job Characteristics Index
- Job Satisfaction
- Job Search
- Labor Force Status
- Military
- Occupations
- Time & Tenure with Employers
- Wages
- Work History Data
- Employer History Roster
- Business Ownership
- Retirement

Work Experience

Created Variables

Number of Employers: NUMBER OF JOBS EVER REPORTED AS OF INTERVIEW DATE (All Interview Years)

Tenure with Specific Employer: TOTAL TENURE IN WEEKS WITH EMPLOYER (JOB #1-5) (All Interview Years)

Cumulative Labor Force Experience:

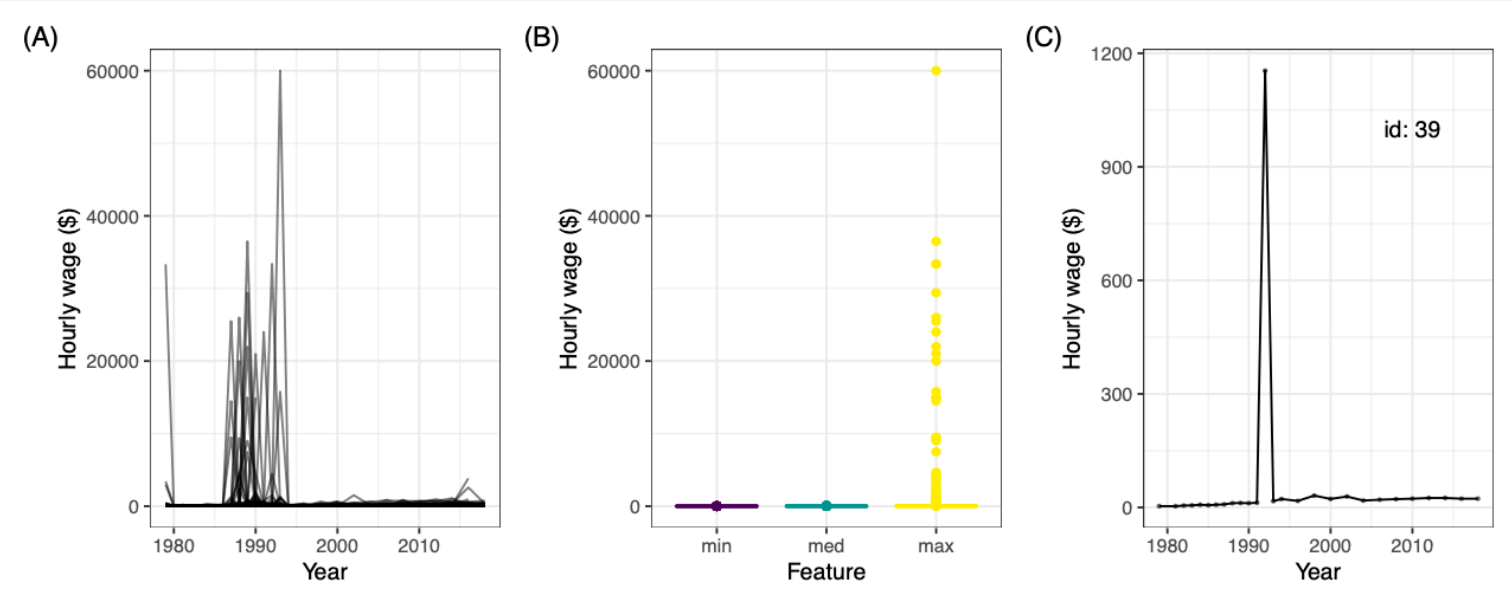
- NUMBER OF WEEKS WORKED SINCE LAST INTERVIEW
- NUMBER OF WEEKS WORKED IN PAST CALENDAR YEAR
- NUMBER OF HOURS WORKED SINCE LAST INTERVIEW
- NUMBER OF HOURS WORKED IN PAST CALENDAR YEAR
- NUMBER OF WEEKS OUT OF LABOR FORCE SINCE LAST INTERVIEW
- NUMBER OF WEEKS OUT OF LABOR FORCE IN PAST CALENDAR YEAR
- NUMBER OF WEEKS UNEMPLOYED SINCE LAST INTERVIEW
- NUMBER OF WEEKS UNEMPLOYED IN PAST CALENDAR YEAR
- PERCENT WEEKS UNACCOUNTED FOR SINCE LAST INTERVIEW
- PERCENT WEEKS UNACCOUNTED FOR IN PAST CALENDAR YEAR
- WEEKS SINCE LAST INTERVIEW
- WEEKS IN ACTIVE MILITARY SERVICE SINCE LAST INTERVIEW
- WEEKS IN ACTIVE MILITARY SERVICE IN PAST CALENDAR YEAR

Note: For Created Weekly Work History arrays, see Work History Data section.

Finally, we calculated `experience` using `number of weeks worked since the last interview`.

Impute Anomalies

Summary | Sample | Treat the anomalies | The result

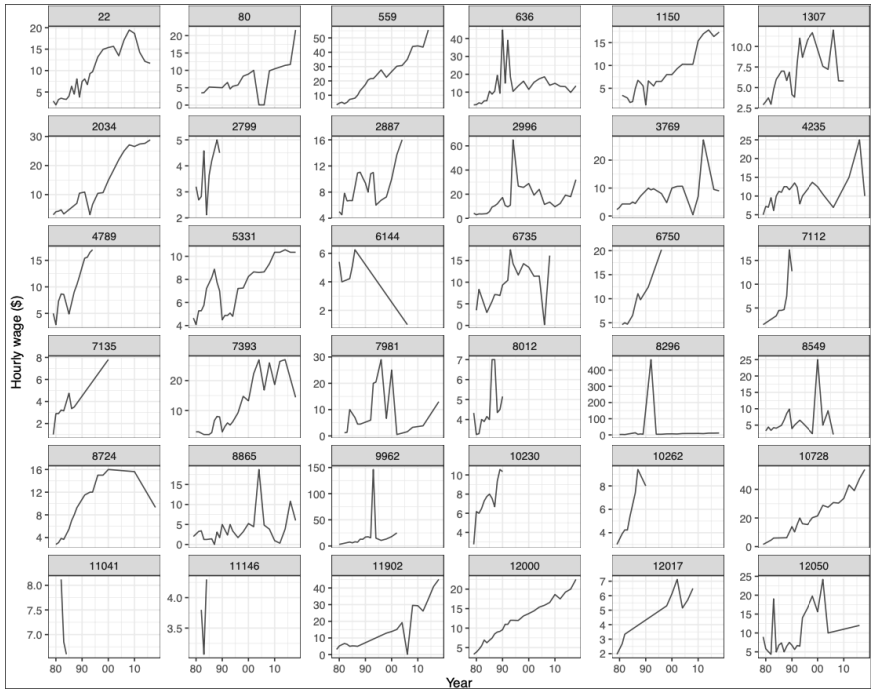


⚠ Extremely high wages observed.

⚠ Some IDs, for example ID=39, only experience high wages in 1 survey year 🤔

Impute Anomalies

Summary | Sample | Treat the anomalies | The result



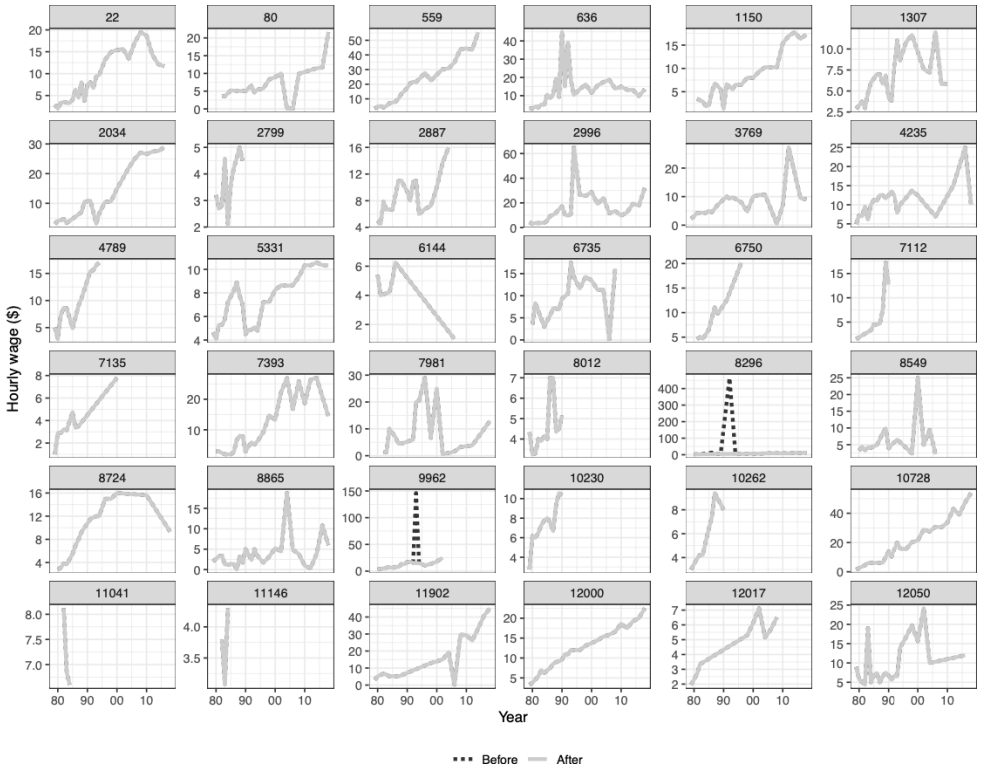
- ID 11041, 11146, 10262 only participated in few years of survey.
- ID 8296, 9962, possibly have error in their wage.

Impute Anomalies

- Using robust linear model (`rlm` function from `MASS` package (Venables and Ripley 2002)) with `wage` and `year` as response and predictor, respectively.
- We build the model for each individual using nest and map function from `tidyr` (Wickham 2020) and `purrr` (Henry and Wickham 2020).
- Each observation has weight -> this is used as a threshold to decide whether the observation is anomalies or not.
- A thereshold of 0.12 was chosen to maintain the variability of the data.

Impute Anomalies

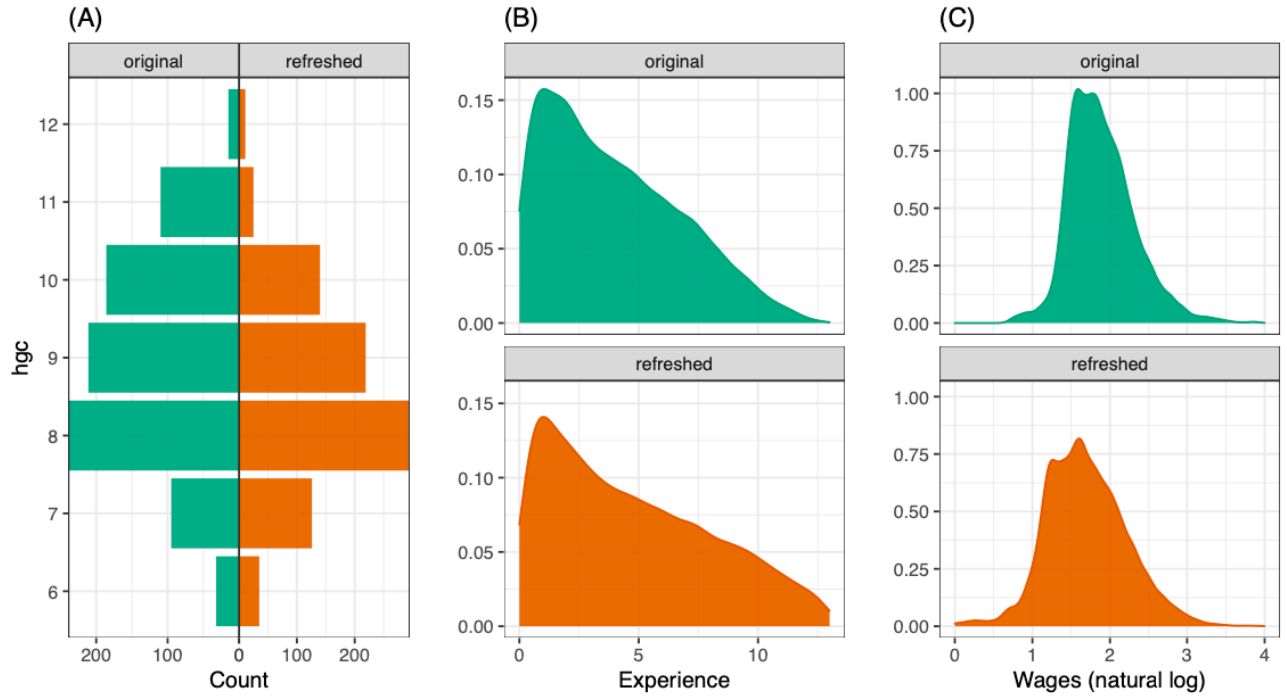
Summary | Sample | Treat the anomalies | **The result**



The extreme spikes, corresponding to anomalies, have been imputed for individuals 8296 and 9962 with their RLM's predicted value.

Comparison with Singer & Willet's (2003)

- Comparison of the subset of dropouts cohort in refreshed and the original data.
- We cannot get the exact same dropouts as in the original data because the criteria of dropouts is not clearly articulated.



Reproducible Aspects in `yowie`

Code and Documentation | shiny app | 👰 Next plan

- We made all of the cleaning and pre-processing codes and the variable tagsets file available here.
- Also in the package's vignette (not updated yet).
- It can be used as data cleaning example

Reproducible Aspects in `yowie`

Code and Documentation | shiny app | 👰 Next plan

- We create a `shiny` (Chang, et.al., 2020) app to simulate different threshold for anomalies treatment.
- Can be found here

Reproducible Aspects in `yowie`

Code and Documentation | shiny app | 👩🏻‍💻 Next plan

- Add function to package to do inflation adjustment on the wages data.
- Put the data into .csv format so it will not only be isolated in R.
- Create the metadata of the datasets.
- Deposit it in place like Zenodo.

Takeaways and Summary

- This case study has shown that documentation (codes and metadata) is really essential to ensure reproducibility.
- Several difficulties encountered in the absence of it:
  - Deciding the variables to be downloaded.
  - Calculating the `experience`.
  - Subsetting the high school dropouts.
  - Matching the original and the refreshed data.
- A better validation rule is suggested to be applied for the data providers in the data entry stage.

Thankyou!

The slide and it's code can be found in https://github.com/Dewi-Amaliah/TWR-2022.

This made with `xaringan` (Xie, 2019) in `rmarkdown` (Xie, Dervieux, Riederer, 2020)

References

Aden-Buie, Garrick (2020). xaringanthemer: Custom 'xaringan' CSS Themes. R package version 0.3.0. https://CRAN.R-project.org/package=xaringanthemer

Bureau of Labor Statistics, U.S. Department of Labor. 2021a. “National Longitudinal Survey of Youth 1979 Cohort, 1979-2016 (Rounds 1-28).” Produced and distributed by the Center for Human Resource Research (CHRR), The Ohio State University. Columbus, OH, through https://www.nlsinfo.org/bibliography-citing-nls-data.

Cooksey, Elizabeth C. 2017. “Using the National Longitudinal Surveys of Youth (Nlsy) to Conduct Life Course Analyses.” In Handbook of Life Course Health Development, edited by Richard M. Lerner Neal Halfon Christoper B. Forrest, 561–77. Cham: Springer. https: //doi.org/https://doi.org/10.1007/978-3-319-47143-3_23.

CrowdFlower. (2016). 2016 Data Science Report. https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Henry, Lionel and Hadley Wickham. (2020). purrr: Functional Programming Tools. R package version 0.3.4. https://CRAN.R-project.org/package=purrr

Huebner, Marianne, Werner Vach, and Saskia le Cessie. (2016). “A Systematic Ap- proach to Initial Data Analysis Is Good Research Practice.” The Journal of Thoracic and Cardiovascular Surgery 151 (1): 25–27.

References

Grolemund, G., & Wickham, H. (2017). R for Data Science. O'Reilly Media.

Kim, Albert Y.; Ismay, Chester; and Chunn, Jennifer, "The fivethirtyeight R package: 'Tame Data' Principles for Introductory Statistics and Data Science Courses" (2018). Mathematics and Statistics: Faculty Publications, Smith College, Northampton, MA. https://scholarworks.smith.edu/mth_facpubs/47

Pergamit, Michael R., Charles R. Pierret, Donna S. Rothstein, and Jonathan R. Veum. (2001). "Data Watch: The National Longitudinal Surveys." The Journal of Economic Perspectives 15 (2): 239–53.

Singer, Judith D, and John B Willett. 2003. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford u.a: Oxford Univ. Pr.

Tierney, N. J., Cook, D., & Prvan, T. (2020). brolgar: An R package to BRowse Over Longitudinal Data Graphically and Analytically in R. In arXiv [stat.AP]. arXiv. http://arxiv.org/abs/2012.01619

van der Loo, Mark P. J., and Edwin de Jonge. (2021). "Data Validation Infrastructure for R." Journal of Statistical Software 97 (10): 1–31. https://doi.org/10.18637/jss. v097.i10.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

References

Wang, E, D Cook, and RJ Hyndman (2020). A new tidy data structure to support exploration and modeling of temporal data, Journal of Computational and Graphical Statistics, 29:3, 466-478, doi:10.1080/10618600.2019.1695624.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham, Hadley. (2021). tidyr: Tidy Messy Data. R package version 1.1.3. https://CRAN.R-project.org/package=tidyr

Xie, Yihui. (2020). xaringan: Presentation Ninja. R package version 0.17. https://CRAN.R-project.org/package=xaringan

Xie, Yihui and Christophe Dervieux and Emily Riederer (2020). R Markdown Cookbook. Chapman and Hall/CRC. ISBN 9780367563837. URL https://bookdown.org/yihui/rmarkdown-cookbook.