Managing Demographic Big Data Analysis in R

Jon Minton, University of Glasgow

Introduction

- Welcome!
- Data Used:
  - Human Fertility Database (HFD): Approx 0.3 Gb. `http://www.mortality.org/`
  - Human Mortality Database (HMD): Approx 1.3 Gb. `http://www.humanfertility.org`
- 'Old Big Data'

Motivations

- Going beyond summary statistics
- Process Automation
  - Data Input
  - Data Tidying
  - Data Exploration
  - Output production
- Rapid Exploration (lowering the cost of curiosity)

Process automation

- Human Error
- Scalability
- Getting from data to value more quickly

Key stages in the data-to-value chain

1. Inputting 'raw' data **(Automatable)**
2. Producing 'tidy' data **(Automatable)**
3. Initial exploratory analyses
4. Producing summary statistics and visualisations for each of the inputs **(Automatable)**
5. Producing final results and outputs

- Note on difficulty: Unfortunately, unlike a game, the first steps can be the most challenging

HFD and HMD as case studies

- Parallel case studies: HFD, HMD
- Processes
  1. **Initial data tidying and harvesting.**
  2. **Exploratory analyses.**
  3. **Automated output generation.**
- Tools:
  - `purrr`: Aids functional programming processes in R
  - `%>%`: The pipe operator
  - `tidyverse`: packages that fit together
  - `RStudio projects`: A magic suitcase

The general approach:

- *do first, learn later*
- Mind tools:
  1. Tidy Data
  2. Code Piping with %>%
  3. `map`, `nest`, and `unnest`

Process today

- Exercise-based: Work as fast as you want to *but no faster*
- Ask me for help throughout
  - *Impromptu breaks if many people ask the same questions*
- Do, do again (HMD/HFD), learn, *(internalise)*
- Code freely available:
  `https://github.com/JonMinton/Comp_Soc_Sci_Course`
- **The Exploratory Buffer**: Be curious!

Have fun!