Git/GitHub setup, Tidy Data & Elements of RStudio

Code

Instructor: Soumya Mukherjee

Content credit: Dr. Matthew Beckman and Olivia Beck

July 5, 2023

Desktop R/RStudio Install

- R and RStudio Desktop
- make sure this is working *right away*
- a word about packages
- install `DataComputing` and `tidyverse` packages (and some others) by copying the following code into an R file and running it.

Hide

```
install.packages("devtools")
library(devtools)
devtools::install_github("DataComputing/DataComputing")
install.packages(c("tidyverse","knitr","rmarkdown","rpart","shiny","manipulate"))
install.packages(c("mosaic","mosaicData","NHANES"))
```

Git & GitHub setup

- Troubleshoot any problems with GitHub profile creation
- Install git if not already available
    - In the Terminal in RStudio, type the following commands:

Hide

```
which git
git --version
```

- If `git` is installed, the result will indicate a where the software is located and the version you have available.

- If `git` has not yet been installed, go to https://git-scm.com/downloads (https://git-scm.com/downloads) and select the version which is compatible with the OS of your computer (e.g. Windows/Mac/Linux/Solaris).
- Notes for Windows:
    - When asked about “Adjusting your PATH environment”, make sure to select “Git from the command line and also from 3rd-party software”. Otherwise, we believe it is good to accept the defaults.
    - Note that RStudio for Windows prefers for `git` to be installed below C:/Program Files and this appears to be the default. This implies, for example, that the Git executable on my Windows system is found at C:/Program Files/Git/bin/git.exe. Unless you have specific reasons to otherwise, follow this convention.
- Notes for MacOS:
    - Go to the Terminal in RStudio and enter one of these commands to elicit an offer to install developer command line tools, then accept the offer and click on “install”:

Hide

```
git --version
git config
```

- For additional support corresponding to various operating systems see https://happygitwithr.com/install-git.html (https://happygitwithr.com/install-git.html).
- We will be using `git` almost entirely within RStudio.

RStudio Tour

- Windows, panes, and tabs in RStudio.
- Markdown / RMarkdown / **R Notebooks**
    - Opening an Rmd file for editing.
    - Saving Rmd files
    - Compiling Rmd to HTML (or PDF or MS Word)
    - Documents, slides, webpages
    - Difference between RMarkdown & R Notebook documents
- **All STAT 184 files should be R NOTEBOOKS submitted as HTML unless otherwise stated**
    - Rmd header should say: `output: html_notebook`
    - HTML document will have a button that says “Code” in top right corner

RMarkdown is good for …

- Headings, lists
- (beautiful) mathematics (e.g. $t = \frac{\bar{x} - \mu_0}{SE}$)
- Links, images
- R code chunks for embedding data analysis & visualization along with your sentences and paragraphs describing the analysis

Posit generously provides “Cheat Sheets” to get people off and running with these and other tools. Here’s a link to several of them (https://posit.co/resources/cheatsheets/), including RMarkdown, RStudio, and other topics we’ll hit in this course.

Tidy Data

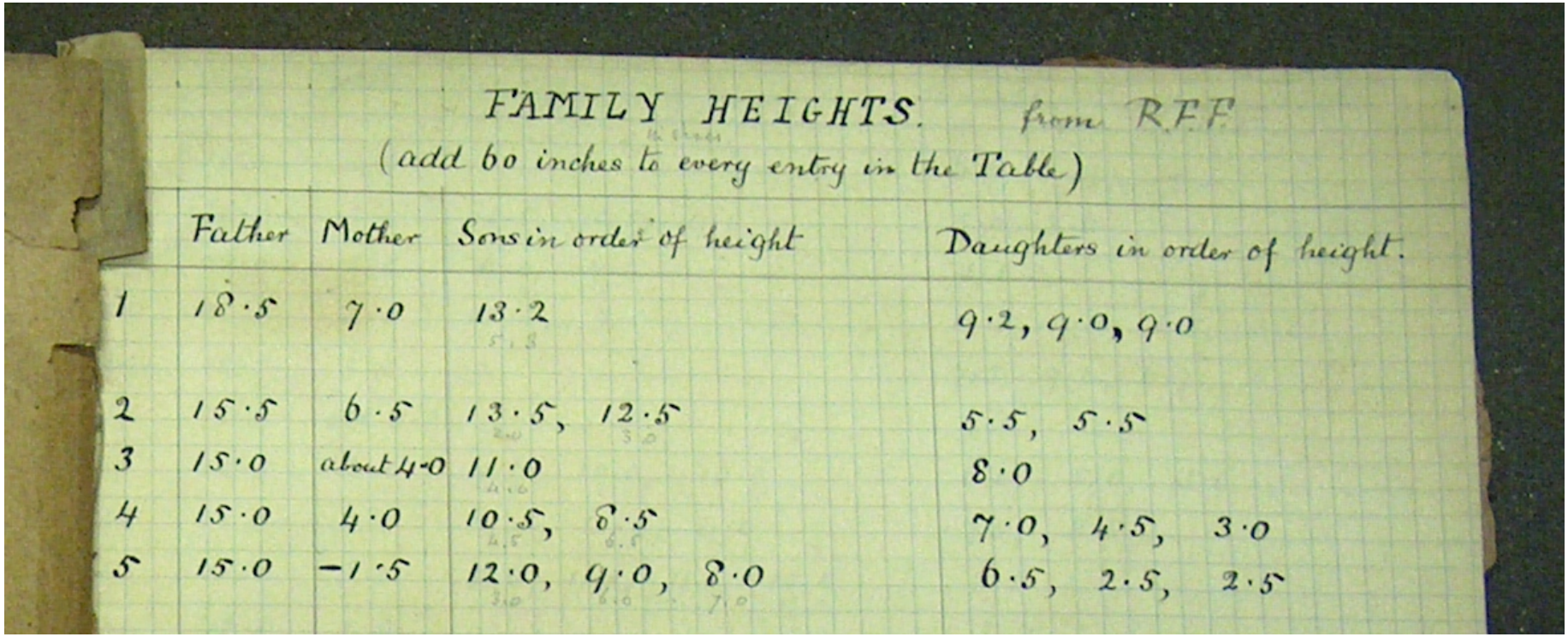
- Key ideas: Cases, Variables, rows, columns, quantitative, categorical
- How should we define **case**?
- How do we identify **variables**?
- Can data be presented in more than one way such that each is still **tidy** by definition?
    - If No, why not?
    - If Yes, how can we tell which is the correct form of **tidy** data?

Galton Data

In the 1880s, Francis Galton started to make a mathematical theory of evolution.

Here’s part of a page from his lab notebook. Things to think abou here:

- What might he investigate with these data (e.g., **Research Question**)?
- Are these data **tidy** according to our definition?
- What are the **cases**?
- What are the **variables**?
- How many **rows** of data should the result have?
- How many **columns** of data should the result have? What is the data type of each column?
- What are some additional variables (not yet shown) that might be of interest? How would you recommend showing that information in the data table?



FAMILY HEIGHTS. from R.F.F.

(add 60 inches to every entry in the Table)

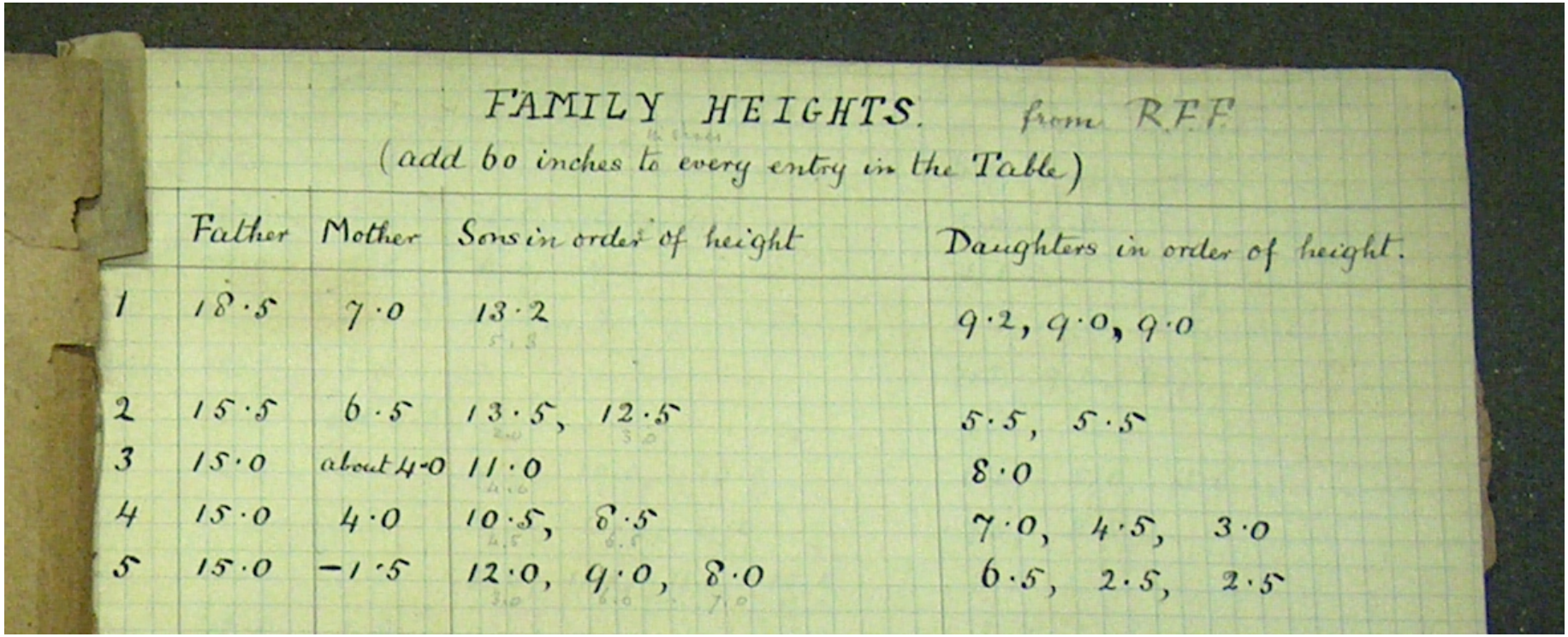
	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5

A page from Francis Galton's notebook.

Activity: Put these into tidy form.

- As a team, you will put this data set into "tidy" form.
- **See Canvas for details**
  - View-only source data is provided
  - use any software you like
  - must submit a CSV to Canvas
- Tip: **Sketch things out on paper before you do anything in the computer**

Table 1: **Galton's Height measurements data**



FAMILY HEIGHTS. from R.F.F.

(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5

A page from Francis Galton's notebook.

Assignment before next lecture tomorrow (July 6)

- Tidy Data assignment using the Galton/HeightMeasurement dataset
- DC Ebook Exercises from Chapter 1
- GitHub Profile creation (ungraded)