

Notes and Materials for Data Computing

Daniel Kaplan

2016-09-07

Contents

1 (PART) Getting Organized	5
2 (PART) Data Infrastructure	7
2.1 Topics	7
3 Untidy data: School enrollments	9
4 Untidy data: Galton’s measurements of height	11
5 (PART) Data Summaries and Graphics	13
6 (PART) Data Verbs	15

This book contains class notes, activities, and projects used in the Data Computing class. Often, there are more activities and projects than we actually used.

The *Data Computing* textbook has 18 chapters. These notes are divided into 8 parts, each of which corresponds to multiple chapters the textbook. This reflects the division of the course itself into eight components.

This is very much a work in progress. It’s not yet complete and what’s here may have many errors. Please do point out the problems so that these notes evolve in a positive way.

Chapter 1

(PART) Getting Organized

The materials from Week 1 will go here.

1. Getting to RStudio
2. Connecting to GitHub
3. RMarkdown
 - Writing a simple RMarkdown document

Chapter 2

(PART) Data Infrastructure

2.1 Topics

- 1) The structure of tabular data
 - cases and variables
 - numerical and categorical variables
 - tidy data
- 2) R Commands
- 3) Files and documents

Chapter 3

Untidy data: School enrollments

The US Census Bureau collects data on many aspects of the population. Data on school enrollments is available [here](#). We’re going to look at one of the data tables they make available:

Table 2: Single Grade of Enrollment and High School Graduation Status for People 3 Years Old and Over, by Sex, Age (Single Years for 3 to 24 Years), Race, and Hispanic Origin: October 2014
XLS or CSV format.

Download one of these files and open it in appropriate software. Or you can view the data on Google Drive [here](#).

1. How many people are represented in this data table?
2. The table is in some ways a graphical visualization of features of school enrollment and age. (Unfocus your eyes and you will see a visual pattern.) What patterns do you see?
3. The table indicates that 74.4% of the people in the table are “not enrolled” in school. Figure out how to calculate this from the numbers in the table. (Hint: You need only look at line 9.)
4. These data are “untidy” in a technical sense. Identify the ways that they are untidy.
5. Some columns contain information that can be calculated from other columns. Look at the data for 4-year olds and identify those that are calculated from other columns. Figure out the minimal set of columns from which the others could be calculated.
6. Imagine that this table was created from a much bigger table in which each case is an individual person in the US.
 - How many cases would there be in that table?
 - What variables would you need so that you could calculate any entry in the “Table 2” provided by the Census Bureau?

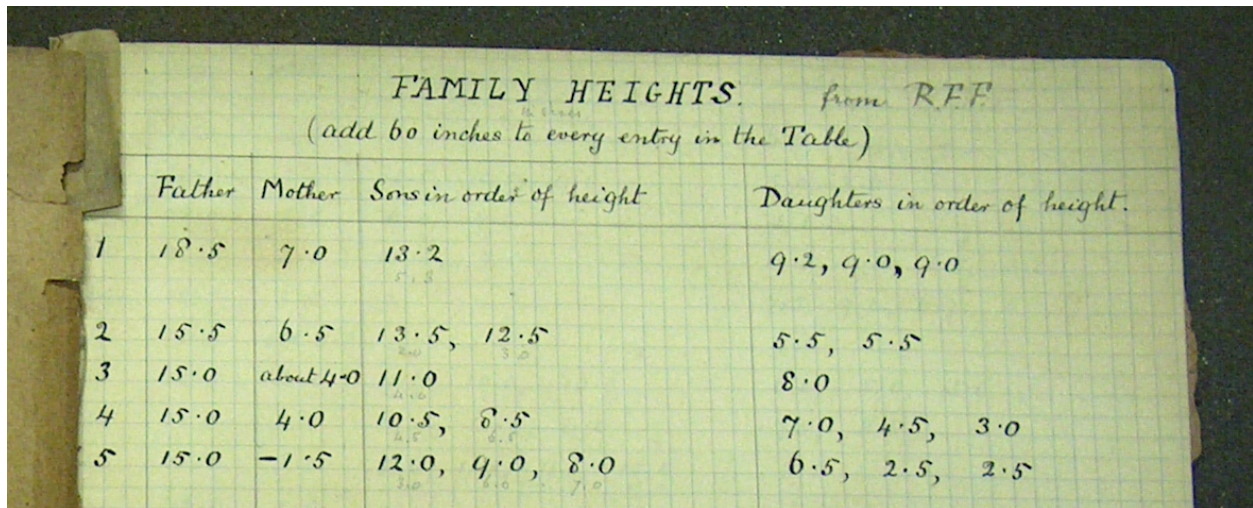
Chapter 4

Untidy data: Galton's measurements of height

In the 1880s, Francis Galton started to make a mathematical theory of evolution. But the basic biology of heritability was not known: nothing about “genes” or DNA, etc.

In order to create a theory, Galton needed a way to measure how traits are inherited from parents. To this end, he visited families in London and measured the heights of the parents and their (adult) children.

Here's part of a page from his lab notebook.



	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

Figure 4.1: A page from Francis Galton's notebook.

Translate this into a tidy form.

- Think about what would be an appropriate “case” for storing this data.
- What variables should there be?
- Fill in a couple of rows of the tidy table that you envision.

Chapter 5

(PART) Data Summaries and Graphics

Chapter 6

(PART) Data Verbs

Bibliography