S&DS 625 Introduction

August 30, 2017

Computing and Reproducible Research

This document was produced in RStudio using R Markdown, choosing the PDF option for the document. You can work without RStudio, but RStudio makes many things like this wonderfully easy. And R Markdown itself is pretty simple (far simpler than full-blown LaTeX). You'll need packages `knitr` and `rmarkdown`, which are bundled in with RStudio.
You'll also want LaTeX on your system, though you won't use it directly. In Windows, this is called `MiKTeX`. The **full, complete distribution** is required (and is quite large, about 2 GB – don't try to download it during class, please). The `MiKTeX` installer will really try to encourage you to install the basic version. It won't work. On the Mac, you'll need `MacTeX`.
It's an easy installation.

Recommendation: start by installing R (www.r-project.org) and either `MiKTeX` or `MacTeX`. Then install RStudio.

Some Advice

Work together on computing stuff like this! There are lots of little computing "gotchas" that can be annoying, but fighting through them and succeeding is important. When you graduate and get a real job (or head to graduate school or whatever), being able to solve problems and be computationally self-sufficient will be valuable. Yes, we're here to study statistics (or data analysis, we would prefer to say).

Markdown?

Looking for information about Markdown? Try http://daringfireball.net/projects/markdown/ or a more concise "cheat sheet" at https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf.

LaTeX?

Let's call this essential if you are doing a PhD – you'll use it for your dissertation and for 95% of the papers you write. But for many it is probably overkill and R Markdown and R scripts will be sufficient. As a side note, you can include many LaTeX things in R Markdown documents.

Toy Example

Let's try working with a dataset. This is a code chunk, which is processed and nicely displayed with the output in the resulting PDF:

```
x <- read.csv("https://www.dropbox.com/s/3qlh23gz41vz908/Diving2000.csv?raw=1",
              as.is=TRUE)
dim(x)
```

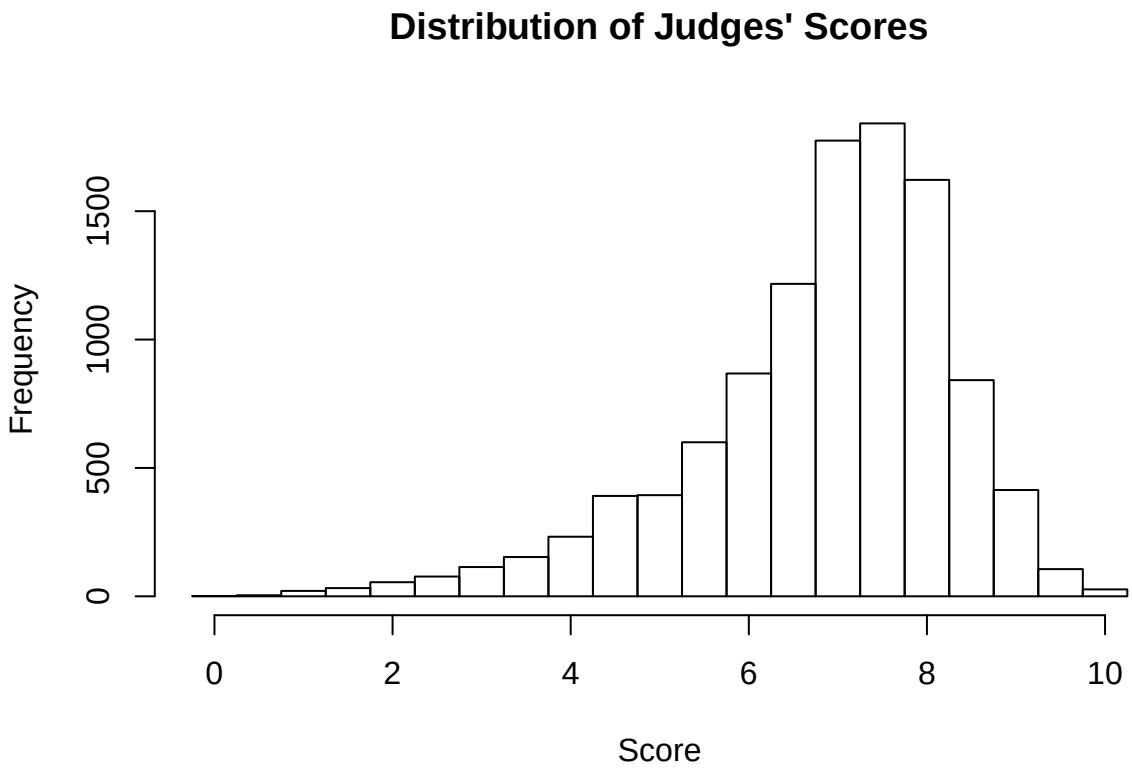
```
## [1] 10787    10
```

```
mean(x$JScore)
```

```
## [1] 6.832576
```

And R Markdown also makes it easy to integrate graphics. (Note that only one graphical display should appear in any one code chunk.)

```
hist(x$JScore, xlab="Score",
     main="Distribution of Judges' Scores",
     breaks=seq(-0.25, 10.25, by=0.5))
```



Reproducible?

This document is entirely ***reproducible*** as long as you have the original R Markdown file, `1-1_Intro.Rmd`. You should be able to reproduce this document exactly (or with perhaps superficial formatting differences), including all the analysis and the plot. Using LaTeX involves a similar workflow that includes code with the document, but is (potentially) more complicated (and hence potentially much more sophisticated).

In this course...

The choice of computing environment (plain vanilla R GUI vs. RStudio) is left up to you. I will be picky on coding style and efficiency. Bad code leads to mistakes and wasted time. Communication is also important, as this is a course about data analysis... so effectively, neatly, and concisely presenting your work will be an important component of the course.

TODAY

- Syllabus
- Hoops data scrape example
- Basic New Haven real estate property record toy scrape and challenge
- Homework for Friday (because it is effectively a Monday)

Homework Due Friday

I will provide a link to the VisionAppraisal files, zipped. Please download from this link and unpack the zip file. That is, please don't scrape the data directly from the web site – it generates lots of traffic (more than 1.5 GB per batch) and may cause problems.

For Friday, write a script that will parse files associated with pid 1 through 1000 (when available) and creates a data frame with three columns (and 1000 rows):

- parcel id (call it `pid`),
- raw address (call it `location`)
- year built (call it `yearbuilt`)

You'll have a row for each property – if there is no information (or no raw file) there should still be a row for it in the data set (with the `pid` but `NA` values for the other variables). Example: for parcel number 600, the `pid` would be 600, the `location` would be "108 HYDE ST", and the `yearbuilt` would be 1910.

Submit your script as "hw1.R" by Friday 11AM onto Canvas. Indentation and quality of code matters. The comments at the top of the script should acknowledge any collaboration (working together is fine but your script should be your own work and you should understand everything that you do).