431 Class 03

Thomas E. Love

2018-09-04

Today's Agenda

1. Describing data numerically and graphically via R Markdown
   - Brief description of the Getting Started with R examples
   - Live demonstration using the Day 1 survey data
2. Assignment 1 - what should you expect?
3. Minute Paper

- Note: We'll discuss the Course Project on Thursday.

Describing Data - Numerically and Graphically

Numerical quantities focus on expected values, graphical summaries on unexpected values.

-- John Tukey

The Getting Started with R document

R Markdown file and resulting PDF describing:

1. A study of chick weights resulting from various feed types
   - Summarizing the distribution of a categorical variable
   - Summarizing the distribution of a continuous variable numerically
   - Graphing the distribution of a continuous variable with a histogram, boxplot and Normal Q-Q plot
2. A study of the growth of orange trees
   - Numerical Summary of two continous variables, and their correlation
   - Scatterplot predicting one continuous variable (outcome) using the other (a predictor)
   - Fitting a Straight-Line model to a scatterplot
   - Stratifying an association on a categorical variable

There are also some tips on getting data into R from Excel.

Getting What You Need

For a complete installation for this course, you will need:

1. **R**, the statistical software (language), which you'll install and then update perhaps once a semester
2. **R Studio**, the integrated development environment we use to make R run more effectively, and to use R Markdown, a language for communicating both code (in R) and text/graphics to the world, which you'll install and then update perhaps once every other month
3. **R Packages**, the "apps" written in the R language to amplify its abilities, and let us do interesting things, which you'll install once, update occasionally but load every time you want to do an analysis (next slide).
4. **Data** for our 431 class, which you don't actually need for Homework 1, but should know how to grab as needed.

R Packages

1. **Installing** an R Package is something you need to do once, ideally. (Unless something breaks.)
    - Our Software Installation document has a detailed list of all packages we want you to have installed.
2. **Updating** an R Package is something you do occasionally. (Usually, when something breaks, but I try for once a day.)
    - Update your packages by visiting the Packages tab on the bottom right panel of R Studio and clicking Update.
3. **Loading** an R Package is something you do with the `library` function, in every R Markdown file you ever write.

The key package for this course that we use in every analysis is actually a suite of packages called the `tidyverse`. To load them, we'll use:

```
library(tidyverse)
```

Doing so actually loads several packages, including `dplyr`, `ggplot2`, `forcats` and more. Details at https://www.tidyverse.org/

The Class 3 document

Do most of the things that are done in the Getting Started with R document, but

1. build an HTML, rather than PDF result
2. build it live in class
3. use our day 1 survey data instead of a pre-packaged R data frame.

- You have the complete pre-class version of the R Markdown file **Class3-pre.Rmd** I will build available to you on the Class 3 Slides page.
- You also have the 2018 day 1 survey data available on the Class 3 Slides page. - Remember to hit **Raw** to see the downloadable version of these files.

Class 3 document tasks

1. Create a new directory on our computer for this work.
2. Download the data called `surveyday1_2018.csv` into that directory.
3. Download the `YOURNAME-hw1.Rmd` template we'll use (designed for HW1) into that directory. (If you like, you can use the `class3-pre.Rmd` version and save some typing, but I'll do things from scratch.)
4. Open a new project in R using the directory in which you have the data and template.
5. Open the template and edit it a bit to personalize.
6. Load the data using a chunk of R code.
7. Begin exploring the data to address four questions.
    1. How do I summarize a multi-categorical variable, like favorite `color`?
    2. How do I summarize a quantitative variable, like `haircut` price?
    3. What is the relationship between age guess and sex?
    4. What is the relationship between `pulse` rate and hours of `sleep`?

To The Live Coding. . .

Assignment 1 (due Friday 2018-09-07 at NOON)

1. Use the `YOURNAME-hw1.Rmd` template to your advantage.
2. Use the Getting Started in R document from our front page to help guide you.
3. The Course Notes contain all the code you might possibly need.
4. Grading will be very light on this assignment compared to later ones.
5. Submit the assignment (two files: R Markdown, plus either HTML or Word files) via canvas.case.edu
6. Apply the 15-minute rule.
    - If you can't solve a problem in 15 minutes, ask for help.
    - You are **absolutely supposed** to use Google and the TAs (and me) to improve your code.

Minute Paper

We will do a minute paper most Tuesdays, and I'll respond in class on most Thursdays.

To get to the Minute Paper form, you'll need to log into Google via CWRU.

The Minute Paper after Class 3 will be available until Wednesday at noon at

INSERT LINK HERE