Example Data Wrangling File with output to RDA

Instructions - Wrangled dataset (Thursday, March 31)

The (reproducible!) code to create the dataset(s) should be saved in an R or Rmd file within your group repository and named "data-wrangling." I will be running the code, so make sure your repository is organized (so I can easily find the "data-wrangling" file) and be sure your code is reproducible, readable, organized logically, and documented with informative comments.

At this stage, your data should be in the format needed to create the visualizations and summaries you'll present in the Shiny app. You want to have little or no wrangling code within your Shiny app program, so it's best to save your wrangled datasets as permanent files that you can load in at the top of the Shiny app program.

You may have more than one wrangled dataset! For instance, if there are some outputs in your Shiny app that require the data to be in long format and other outputs in your Shiny app that require the data to be in wide format, then you'll want to create both a long format dataset and a wide format dataset in your "data-wrangling" file.

Include a note at the start of the data wrangling file to indicate each person's contribution to the wrangling process.

How to Proceed

Create a .Rmd or .R script file that your group will use for data wrangling. Do all your wrangling there and output a final version of the dataset (or versions, if you need several) to .csv or other convenient format (example below).

The file should explain where ALL data comes from. That is, include your references.

Example:

Group ProfDemo's Wrangling File (.Rmd example)

Our data sources are: LIST (including references).

From the first source, we . . . .

```
# Code + documentation
```

We had to scrape from the second source and needed to save . . .

```
# Code + documentation
```

Finally, from the third source, we had to. . .

```
# Code + documentation
```

Then we joined the files together. Our final datasets are: Blah and Blah2.

```
# Code + documentation

# output final datasets to convenient format
```

If your initial datasets are LARGE (> 100 MB)

Files larger than 100 MB cannot be hosted on git. Here's how to proceed in that case:

1. Host the files somewhere like Dropbox in a shared folder for the group. Remember to grant me access as well.
2. ONE member should read in the files into R and then output another format that CAN be hosted on Git. This part is likely to use an unreproducible path - that's fine, but it's why you have to describe the process clearly in the data wrangling file. I recommend that the group member who does this part NOT be working on R via the R server (due to space limitations). If you do this part on the server, delete the large input files afterward.
3. That format can then be read in for the start of the wrangling process (with everything reproducible from here on).
4. This process should be included in your data wrangling file as part of the description. In the .Rmd, you can show the code with eval = FALSE, or leave a separate R script file for this initial step that is clearly marked that it will not run without the external files.

The typical file format that will work for this is .rda files - these are R's natural data files. They can be much smaller than the corresponding .csv.

An example is below. This example uses the College Scorecard data that you saw a tiny subset of in lab. The entire data set is too big.

Creating a .rda

Downloaded the most recent institution-level data file from https://collegescorecard.ed.gov/data/

The associated data dictionary is from: https://collegescorecard.ed.gov/data/documentation/

The file is too big to publish to github. So, we want to read in to R and save the .rda/.rds file to use. You need to make the files accessible in some way - perhaps use Dropbox.

```r
# read large CSV into R
# This path is allowed to be non-reproducible
# (meaning someone else would have to edit to use it)
# since you have to host the data not in Git to do this
# This example shows this from my Desktop
library(readr)
CollegeScorecardMostRecent <- read_csv("C:/Users/awagaman/Desktop/CollegeScorecardMostRecent.csv",
                                       na = "NULL")
```

Now we save this to .rda - I saved it to my Github folder directly from here (because the .Rmd was in that directory), but you could save it anywhere and then move it into your repo.

```r
# save to .rda
save(CollegeScorecardMostRecent, file = "Data Analysis/college.rda")
```

Then, to load it later to continue your data wrangling (probably in a new file), we just use the load function:

```r
# load the .rda
load("Data Analysis/college.rda")
```

This was in a subfolder called Data Analysis, relative to where the .Rmd was. Again, the idea is the college.rda file is MUCH smaller than the original .csv.