Section A01 CDS 101 *Introduction to Computational and Data Sciences*

Class 11: Introduction to the Midterm Project dataset

June 5, 2018

General

Announcements

- Homework 2 due on June 6th @ 11:59pm: http://summer18.cds101.com/assignments/homework-2/

Midterm project instructions

"Fast facts" for Midterm Project

- Assigned into teams that will use Github to collaborate on writing an RMarkdown report (60% of Midterm Project grade) and giving a powerpoint presentation (40% of Midterm Project grade)

"Fast facts" for Midterm Project

- Assigned into teams that will use Github to collaborate on writing an RMarkdown report (60% of Midterm Project grade) and giving a powerpoint presentation (40% of Midterm Project grade)
- Completed in groups, but graded individually

"Fast facts" for Midterm Project

- Assigned into teams that will use Github to collaborate on writing an RMarkdown report (60% of Midterm Project grade) and giving a powerpoint presentation (40% of Midterm Project grade)
- Completed in groups, but graded individually
    - Each student's contribution to the group will be inferred from the Github commit history and his/her relative level of participation in the group's private Slack channel.

"Fast facts" for Midterm Project

- Assigned into teams that will use Github to collaborate on writing an RMarkdown report (60% of Midterm Project grade) and giving a powerpoint presentation (40% of Midterm Project grade)
- Completed in groups, but graded individually
    - Each student's contribution to the group will be inferred from the Github commit history and his/her relative level of participation in the group's private Slack channel.
- The report consists of two sections: **Cleaning and tidying the dataset** and **Exploratory data analysis**

"Fast facts" for Midterm Project

- Assigned into teams that will use Github to collaborate on writing an RMarkdown report (60% of Midterm Project grade) and giving a powerpoint presentation (40% of Midterm Project grade)

- Completed in groups, but graded individually

  - Each student's contribution to the group will be inferred from the Github commit history and his/her relative level of participation in the group's private Slack channel.

- The report consists of two sections: **Cleaning and tidying the dataset** and **Exploratory data analysis**

- Each team member is responsible for formulating one question about the dataset that he/she then answers in the final report.

"Fast facts" for Midterm Project

- Assigned into teams that will use Github to collaborate on writing an RMarkdown report (60% of Midterm Project grade) and giving a powerpoint presentation (40% of Midterm Project grade)
- Completed in groups, but graded individually
    - Each student's contribution to the group will be inferred from the Github commit history and his/her relative level of participation in the group's private Slack channel.
- The report consists of two sections: **Cleaning and tidying the dataset** and **Exploratory data analysis**
- Each team member is responsible for formulating one question about the dataset that he/she then answers in the final report.
- The report should be well-formatted and not have obvious errors such as code blocks that are too wide or pictures with illegible labels or unknown acronyms.

"Fast facts" for Midterm Project

- This is an evidence-based formal report that should be written using a professional tone and fully edited before final submission

"Fast facts" for Midterm Project

- This is an evidence-based formal report that should be written using a professional tone and fully edited before final submission

- The presentation length must be between 8 to 10 minutes in length for teams of three and between 12 to 14 minutes in length for teams of four

"Fast facts" for Midterm Project

- This is an evidence-based formal report that should be written using a professional tone and fully edited before final submission

- The presentation length must be between 8 to 10 minutes in length for teams of three and between 12 to 14 minutes in length for teams of four

- The presentation is a **summary** of your report that states what your questions are, what you needed to find in the dataset to answer them, key steps in getting your answers using R, and your answer or conclusion for each question

"Fast facts" for Midterm Project

- This is an evidence-based formal report that should be written using a professional tone and fully edited before final submission
- The presentation length must be between 8 to 10 minutes in length for teams of three and between 12 to 14 minutes in length for teams of four
- The presentation is a **summary** of your report that states what your questions are, what you needed to find in the dataset to answer them, key steps in getting your answers using R, and your answer or conclusion for each question
- During the presentation, each team member must speak and the speaking time of each student should be approximately equal

Cleaning and tidying the dataset section

- Dataset is semi-structured and somewhat clean
- May still require a small amount of cleaning and reshaping
- Cell entries containing `"PrivacySuppressed"`: need to decide how to handle these, one option is to replace them all with `NA`
- Depending on the plots you want to create, you may need to do some data reshaping
- **After** you've written down your questions and decided on the variables needed to answer them, use `select()` to reduce the size of the dataset

Exploratory data analysis section

Each team member will construct and answer 1 question about the dataset in this section. So, for example, a team of three members will have 3 questions in total. Each question must involve one or more visualizations. **Your questions must be about comparing relationships between two or more variables in the dataset**, which can include how a variable is distributed across several different categories. In addition, answering the question must require that you make use of both *data transformation* (dplyr) and *data visualization* (ggplot2). Your question cannot just be a simple filter query or visualizations of different columns “out of the box” without any kind of subsetting or grouping. Answering the question must require data aggregation or identifying a trend between 2 or more variables.

Exploratory data analysis section

The instructor will conference with all groups a few days before the report and presentation are due. At the conference each group must be able to state all the questions they will be addressing. There should also be some kind of justification for why you're asking each question. For example, if you are looking up what fraction of the student body are women at just two schools, you need to explain why this is interesting to know and why this comparison is meaningful. The instructor reserves the right to veto any questions that do not meet the outlined criteria or that cannot be appropriately justified.

Exploratory data analysis section

In the report, each question should be clearly stated and followed by the procedure used to answer it. The procedure takes the form of both code blocks and plain text. Then, after you obtain your final result in the form of a visualization, **be sure to interpret it for the reader**. For example, if it's a distribution, what is it's shape and center? If it's a scatter plot, what is the trend of the points? After analyzing the various outputs, synthesize it and provide a formal answer to your stated question.

Github with groups

Dealing with merge conflicts

Demonstration of what causes a merge conflict, and how to resolve it in RStudio Server.

Using branches to collaborate

Demonstration of how to create branches in RStudio Server and how to use **Pull Requests** to merge in each group member's contribution to the midterm report

Importing the midterm project dataset into RStudio Server

Download the dataset

- The dataset for the midterm project is larger than what you've seen in the homework assignments and in-class examples
- **Not recommended:** download the 141 MB data file to your computer and then try to upload it to RStudio Server through the web browser
- **Recommended:** download the file directly using R
- The following code will download the CSV file directly into your Midterm project folder on RStudio Server:

```
download.file(
  url = "https://ed-public-download.app.cloud.gov/downloads/Most-Recent-Cohorts-All-Data-Elements.csv",
  destfile = "Most-Recent-Cohorts-All-Data-Elements.csv"
)
```

Read the CSV file into RStudio

- After you download the CSV file into your project folder on RStudio Server, the dataaset can be imported using `read_csv()`:

```
college <- read_csv(
  file = "Most-Recent-Cohorts-All-Data-Elements.csv",
  na = combine("NA", "NULL")
)
```

- This will take a few seconds to fully import into RStudio.

Important note!

*The data file is **much** too large to commit directly to Github!*

Compress the dataset

- In general, "best practices" state that datasets should not be put in repos unless they've very small (files less than 1 MB in size, for example)

- For our convenience, we will relax that rule a little bit and allow ourselves to commit and push the dataset in compressed form

- Use this code to compress and write your file in the `rds` format:

```
college %>%
  write_rds(
    path = "Most-Recent-Cohorts-All-Data-Elements.rds",
    compress = "gz"
  )
```