Midterm Project

Due: October 18, 2017 @ 3:00pm

Instructions

For the midterm, you will be assigned into groups that are then tasked with performing a small-scale exploratory data analysis using the skills you've developed over the first half of the semester. All groups in the class will be working on the same dataset, the College Scorecard dataset started by The Obama Administration in September 2015. You will explore this dataset by formulating questions within your group, writing them down, and then answering each one by wrangling the data into a form that allows you to visualize it. The key word here is **visualize**; your data transformations should be in service of creating visualizations that answer each of your formulated questions. Beyond creating visualizations, you may also supplement your analysis by performing some basic statistics calculations if it's within your group's comfort level (this is not a requirement). You may also bring in additional data if you want, but any additional data must be documented, which includes exactly how you obtained it and how you're able to combine it with the main dataset to further your analysis.

This is a large dataset that is over 100 megabytes in size and contains millions of individual cells. As such, there is no one right way to approach this project. There are many different avenues that you can take, so have fun with it! As this is a large dataset with millions of individual cells, each group is likely to showcase a completely different aspect of the dataset.

Submission guidelines

Your Midterm Project submission is expected to meet the following guidelines:

- **Your group submission will only be graded if it is on Github and in your group's copy of the midterm project repository. Your group must also submit a Pull Request against the `starting` branch.**
- The writeup and analysis must be completed in a RMarkdown file that must compile without error.
- The data analysis must be completed using the `tidyverse` tools described in *R for Data Science* and the plots must be generated using `ggplot2`.
- Your R code should be clean and readable. It is worth consulting an R style guide during the editing phase. Two good sources are Google's R Style Guide and Hadley Wickham's R Style Guide.
- Your work is to be documented using Markdown blocks. Each block of R code should have a Markdown block above it that explains its purpose and what is being done.
- As with other assignments, you are expected to write using complete sentences and to explain your reasoning. Your submission should have a coherent writing style and structure, which means that the project should not read like a long question-and-answer worksheet or like it was written by several people and then pasted together.
- The project writeup should have two sections, **Cleaning and tidying the dataset**, and **Exploratory data analysis**, see below for a description.
- As detailed in the syllabus, *late submissions for the midterm project will not be accepted* and your presentation **must** be given on the scheduled date, no exceptions.
- Your submission should contain the group's report in RMarkdown form, which should contain a direct link for downloading the dataset, and a copy of your group's presentation slides (powerpoint or RMarkdown file that outputs to the `ioslides` format).
- Do not include questions that you were unable to answer in your report.
- Each group member is expected to make **at least one** substantial commit (or a collection of smaller commits) to the group repository on Github. Each member's commits should be his/her own work.

Presentation guidelines

Your presentation is expected to meet the following guidelines:

- The presentation must be between 7 to 10 minutes in length.
- The presentation summarizes key aspects of your report, such as what your questions are, what you needed to find that would answer them, and the choices you made that led to answers. The presentation should **not** be a laundry list of everything you tried to do that didn't work.
- The presentation must include slides, either in powerpoint format or an RMarkdown file that compiles to the `ioslides` format.
- Each team member must speak during the presentation. Also, while it is understood that each group member will offer different contributions, every team member should be able to speak independently about the steps taken in your project and answer basic questions. It is not in your benefit to use a "divide and conquer" strategy if you never share your findings with your group afterward!
- The group should be able to explain the reasoning for taking any particular step during the project.

Grade

The submitted writeup is worth 70% of your midterm project grade and your in-class presentation is worth the remaining 30%. Grading criteria for the written submission will be based on the correctness and readibility of your R code, if your writeup is structured, coherent, and has proper spelling and follows standard rules of grammar, and the general quality of how you answer each of your presented questions. You will also submit individual evaluations of your group members contributions after submitting your project, which will factor into the writeup grade. Grading criteria for the presentation will be based on falling within the time length, if you speak for long enough during the presentation (a minimum of 1–2 minutes a person is sufficient), and the quality of how you present questions and how you determined their answers. In addition, your classmates will rate your presentation quality, which will factor into the presentation grade.

You will be graded as an individual, even though this is a group project. Any group members that are judged to have not sufficiently contributed to the final product will have their grade penalized.

As stated in the class syllabus, this project is worth 20% of your class grade.

The Dataset

The College Scorecard dataset that you'll be using is available at https://collegescorecard.ed.gov/data/. The primary data file that you will be using is labeled as *Most recent data* on the above website. The direct link to the dataset is https://ed-public-download.app.cloud.gov/downloads/Most-Recent-Cohorts-All-Data-Elements.csv.

The data codebook is available at https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary.xlsx, which describes all of the variables that are in the dataset. You will have to look through the codebook to understand the meaning of the variables, and this should be your starting point before you start running an analysis on the dataset.

For further information about the dataset, consult the documentation pages at https://collegescorecard.ed.gov/data/documentation/.

Finally, loading the dataset, you will need to specify values that will count as `NA` entries, otherwise it will give you errors. The simplest way to load without errors is to run

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

For your convenience, this command is included in the setup block in your group's project template.

Cleaning and tidying the dataset section

This dataset is semistructured and at least somewhat clean, but it is likely that you will have to perform a small amount of cleaning and/or tidying. A simple example are cell entries containing "PrivacySuppressed", which may reside in columns that otherwise contain numerical data. You may need to fix the data type for some columns after dealing with the "PrivacySuppressed" entries. Also, some of the columns don't follow the "each variable must have its own column" rule, so you will likely have to do some data reshaping before starting your analysis.

Attempting to reshape the full dataset so that it is tidy would be an involved task. To keep things manageable, it is recommended that, after you've written down your questions and decided on the variables you will be analyzing and visualizing, that you extract those columns using `select()` in order to reduce the size of the dataset. Assign the reduced dataset to a different variable (don't overwrite the original dataset). Afterwards you can perform your tidying and cleaning operations on the **reduced** dataset.

In your writeup, the data cleaning and tidying section should include documentation of your procedure. This would be in the form of code blocks with explanations for why the cleaning/tidying is necessary (for example, cite the tidy data rule that you're addressing).

Exploratory data analysis section

Your group will construct and answer 3 questions about the dataset in this section. Each question must involve one or more visualizations, so bare minimum, your report will contain three figures. 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 visualisation* (`ggplot2`). Running visualizations on different columns "out of the box" without any kind of filtering or grouping/summarizing is not sufficient for this project.

In your writeup, you should document your question then document your procedure for answering it via code blocks and plain text. After you obtain your final result in the form of a visualization, **be sure to interpret it for the reader**. 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.