Midterm Project

Due: June 14, 2018 @ 10:30am

Instructions

For the midterm, you will be assigned into teams that will conduct an exploratory data analysis using the skills you've developed over the first half of the semester. Teams will be responsible for creating a report that summarizes their exploratory data analysis and for giving an in-class presentation of their results. Each team will motivate their exploration by formulating interesting questions that can be answered with the dataset, which are then answered by wrangling the data into a form that allows you to visualize it. The key word here is **visualize**; all data transformations should be in service of creating visualizations that answer your team's questions. Teams are welcome to supplement their work by performing basic statistics calculations, but it is not a requirement. Teams may also bring in additional data to strengthen the analysis, but please note that any additional data must be documented, which includes describing how you obtained it and how you've integrated it with the main dataset to further your analysis.

The Dataset

All teams will be working with the College Scorecard dataset started by The Obama Administration in September 2015. The dataset is available at https://collegescorecard.ed.gov/data/. **The primary data file that you will be using is labeled as *Most recent data***. The direct link to the dataset is https://ed-public-download.app.cloud.gov/downloads/Most-Recent-Cohorts-All-Data-Elements.csv.

The data code-book 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 code-book 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/.

When 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(
  file = "Most-Recent-Cohorts-All-Data-Elements.csv",
  na = c("NA", "NULL")
)
```

It is highly recommended that, after successfully loading the dataset for the first time, one of the team members compresses the dataset into an rds file and then commits and pushes it to Github.

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!

Submission guidelines

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

- **Your team submission will only be graded if it is on Github and in your team's copy of the midterm project repository. Your team must also submit a Pull Request against the `starting` branch.**
- Your submission must contain the team's RMarkdown file for the final report and a copy of your team's presentation slides. The RMarkdown report file must knit to HTML and PDF without error.
- The report is to have two sections, **Cleaning and tidying the dataset**, and **Exploratory data analysis**, see further down the document for a description.
- The data analysis must be completed using the `tidyverse` tools described in *R for Data Science* and the plots must be generated using `ggplot2`.
- Each team member is expected to formulate one question that he/she then answers in the final report.
- Each team member must contribute substantive commits to the team repository on Github that reflect his/her own work.
- The report represents your team's final results and should only contain the methods used to obtain them. **Do not include questions in the report that you cannot answer.**
- Your R code should be clean and readable. For guidelines, take a look at 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 Markdown text above it that explains the code's purpose and what is being done.
- **Teams must do a final edit on the report so that the final submission has a coherent writing style and structure.** Each student has a different writing style, and it is distracting if the writing style changes from question to question, so select an editor for the team to make it uniform before submission. In addition to enforcing a consistent style, teams should verify that any spelling and grammar errors are fixed.
- The code blocks should look uniform and clean upon knitting. **Code blocks should not run off the side of the page when knitted to PDF!**
- **The report's tone should be professional and should not read like a social media feed or personal blog**. Refrain from editorializing about the project as a whole or about a specific question, as this is not an opinion paper. Do not speculate, instead support your claims and explanations using data and analysis. Avoid self-narration or writing about how you felt or what you were thinking as you complete each question, instead write as if you are constructing a step-by-step tutorial for others to use.
- **Late submissions for the midterm project will not be accepted and your presentation must be given on the scheduled date, no exceptions.**

Presentation guidelines

Your presentation is expected to meet the following guidelines:

- For teams of three, the presentation must be between 8 to 10 minutes in length. For teams of four, the presentation must be between 12 to 14 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. It also shouldn't contain much R code, only include the most important snippets.
- The presentation must include slides that you created using either PowerPoint or Google Slides.
- Each team member must speak during the presentation. Also, while it is understood that each team 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.
- The team should be able to explain the reasoning for taking any particular step during the project.

Grade

The submitted write-up is worth 60% of your midterm project grade and your in-class presentation is worth the remaining 40%. Grading criteria for the written submission will be based on the correctness and readability of your R code, if your write-up 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. Grading criteria for the presentation will be based on falling within the time length, how long each team member speaks during the presentation, whether the spoken content is substantive, and the general quality of how the group presents questions and shows how they arrived at their answers. Finally, your classmates will peer review your presentations using a form, which will be averaged together and factored into the presentation grade.

You will be graded as an individual, even though this is a team project. Any team 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 25% of your class grade.

Cleaning and tidying the dataset section

This dataset is semi-structured 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. Depending on the types of plots you want to create, you may also need to do some data reshaping.

To keep things manageable it is recommended that, after you've written down your questions and decided on the variables you will need for your analysis and visualizations, that you extract those columns using `select()` and save the intermediate dataset to disk, and then load that in order to reduce the size of the dataset. Afterwards you can perform your cleaning operations on the **reduced** dataset.

In your write-up, 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.

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.

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.

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.