# Stat 159: Final Project, Due Friday Dec-02

November 3, 2016


### Abstract

The final project for Stat 159 consists of performing a virtual consulting data analysis project using data from College Scorecard. The underlying goal is to use all the computational tools and apply all the concepts covered in the course to create a computationally reproducible analysis. You will have to form groups of four members, and work together to complete the project.


## College Scorecard

In this project, you will work with data from *College Scorecard*:

https://collegescorecard.ed.gov/

*College Scorecard* is developed by the U.S. Department of Education (under Obama's Administration) to provide "key indicators about the cost and value of institutions across the country to help students choose a school that is well-suited to meet their needs, priced affordably, and is consistent with their educational and career goals". Read more in this press release.

You could also watch President Obama's video: https://www.youtube.com/watch?v=Tpaj9Sm7i7I

The sources of data behind *College Scorecard* are available at:

https://collegescorecard.ed.gov/data/

In that web site you will find the following documents:

- Policy Paper:

https://collegescorecard.ed.gov/assets/BetterInformationForBetterCollegeChoiceAndInstitutionalPerformance.pdf

- Technical Paper:

https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf

- Full Data Documentation:

https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf

- Data Dictionary (`xlsx` file):

https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary-09-26-2016.xlsx

- API with documentation available at:

http://api.data.gov/ed/collegescorecard

## Motivation

According to the various documents in the collegescorecard.ed.gov/data website, the available data:

- provide reliable and unbiased information about college performance.
- can be used to measure the performance of postsecondary education institutions in terms of the access they provide, the level of affordability, and the outcomes of students.
- provide new tools for higher education officials and policymakers concerned with how well institutions are providing access to diverse groups of students.
- be used by stakeholders to better assess the extent to which institutions are promoting the success of all their students and target support where necessary.
- will help institutions to benchmark and improve their performance.

### Consulting project for a virtual client

The final project is very open-ended. To help you come up with a research question, as well as to narrow down the analytical component, pretend that you are doing a consulting project for one of the following client profiles:

- An organization advocating for undocumented students not qualified for governmental financial aid is about to start its fund raising campaign. They would like to identify potential "bad apple" colleges in which it is most likely that the students it advises won't be able to finish their education.
- A group of policy makers that wants to assess the performance of publicly funded schools to decide whether to allocate grant money. They are interested in increasing equity and graduation rates to serve more underserved populations.
- A group of administrators trying to make their school more competitive, wanting to compare themselves to similar schools in demographics and population centers, in terms of diversity, and graduation rates.
- A credit institution that provides financial aid to students. The managers are interested in expanding their customer base but they would also like that most of the loans be payed back. Should they give credit to students attending any college? Should they give credit to students attending a university in which most students tend to find jobs after graduating?
- A non-profit NGO providing career and academic guidance that is interested in supporting education of minority students. They want to identify "best value" schools: quality programs that costs the least.
- A philantropic foundation that wants to know under what circumstances they could give college scholarships to students based on their interests, school preferences, SAT scores, and performance of intended school.
- A newspaper or magazine that asks your help to write a piece of data journalism about whether public spending in higher education is effective, and which will offer an "X-ray" of the educational landscape in the US or in a given state.
- The CEO of a biotech startup is looking for candidates. She is interested in diversifying the workforce in regards to women in STEM (Science, Technology, Engineering, and Math).

Where should the startup focus their recruitment and outreach for maximum impact?

- The head of a marketing team from a for-profit university wants to expand reach and increase enrollment (thus profit). Who should they target and what strategy would you recommend them?

Choose one client among the suggested profiles; or come up with a different client profile and develop an ad-hoc analysis for them. Regardless of the type of client profile that you choose, you should NOT work on a project that mimics the service provided by https://collegescorecard.ed.gov/. In other words, do not perform an analysis that gives recommendations to a student (or her/his family) wanting to attend college.

You can choose to focus on one or more of the following aspects:

- Exploratory Data Analysis
- Hypothesis testing
- Model Building
- Simulations

### Main deliverables

The ultimate deliverables consist of:

- Report (summary report)
- Slides
- Shiny App(s)
- And obviously the github repo of the project

Pretty much all the final materials from the project (report, slides, shiny app, images, derived data sets, intermediate and final results), should be prepared and designed having in mind the target audience formed by the virtual client.

### Proposal

Please prepare a first proposal (pdf) by Thursday Nov-10: write a document describing the chosen client profile and what type of analysis you'll perform. Be as much specific and detailed as possible. You should be able to provide a roadmap of the analysis workflow: data preparation, data exploration, analysis components, and applied methods. Your proposal should also include a preliminary version of the file structure.

### Computational Reproducibility

Keep in mind that the major underlying goal of the project is Computational Reproducibility (i.e. replicability). We, the instructor and the GSI (or any other user) should be able to clone or fork your github repository, follow the instructions in your README file and, by executing the series of indicated commands, obtain the exact same results (derived data sets, images, summaries, tables, report, slides) that you generated.

## Team Work

It is important that you brainstorm together and come up with a list of chores for each member. For instance, one person may be in charged of developing the shiny app. Another person could be responsible of being the chief editor of the report. A third person could be in charge of the slides.

The division of tasks does not imply that only one person does one thing. Everybody should be able to review the code of the shiny app, and if there is a bug or a suggestion, either: create an **issue** to the person in responsible for the app, or make a **pull request** to fix the bug, or directly modify the code.

You (i.e. the entire team) are responsible for code peer review. You should also known which person does what. We (instructor and GSI) may ask you who was in charge of the data collection, or who was responsible for the data cleaning, or who performed the exploratory data analysis. If you fail to answer correctly, your entire team will lose points.

Another suggestion is to come up with a style-guide that helps the entire team achieve consistency along the project: naming functions, naming files, naming directories, naming commits, etc.

## Requirements

- First written proposal (pdf) by Thursday Nov-10: write a small report describing the chosen client profile and what type of analysis you'll perform.
- Complete submission by Dec-02
- Use a public independent github repo
- Use of git branches
- Main `README` file containing:
  - title of project
  - description of the project's structure (primary directories and main files)
  - instructions for other users about how to reproduce your project
  - list of Make commands for phony targets
  - information about licenses: one creative commons license for media-content https://creativecommons.org/choose/, one open-source license for code http://choosealicense.com/
  - authors names (of both members)
- If you perform simulations or any other computation that involves calling random generator functions, please use random seeds.
- Report written in **LaTeX** with an `.Rnw` file, and compiled in PDF format. Do NOT use an Rmd file.
- Slides using one of the slide frameworks:
  - ioslies: http://rmarkdown.rstudio.com/ioslides_presentation_format.html
  - beamer: http://rmarkdown.rstudio.com/beamer_presentation_format.html
  - slidy: http://rmarkdown.rstudio.com/slidy_presentation_format.html
  - reveal.js: http://rmarkdown.rstudio.com/revealjs_presentation_format.html
- Your `Makefile` should include:
  - declaration of variables
  - use of Make automatic variables
  - comments for rules, targets or dependencies that need further description

  - all required phony targets (you must come up with a list of phony targets)
- Session Information in `session-info.txt`:
  - R's session information `devtools::session_info()`, with versions of all the used R packages
  - git version
  - latex version
  - pandoc version
  - Make version
  - version of other software tools you use

In addition to not meeting the previous requirements, points will be deducted for:

- having inconsistent styleguides
- writing bad (e.g. uninformative, dummy) commit messages
- use of absolute filepath names (this breaks reproducibility)
- hard coding values in the `Rnw` file(s) or slide's `Rmd` file.