# Homework Assignment 1

STA 141A

Due Saturday, April 21st by midnight

### Description

In this assignment, you will analyze a subset of the U.S. Department of Education's College Scorecard Data[1]. This dataset combines demographic and economic information for all 4-year colleges in the U.S. in 2013. Each row corresponds to one college campus. A description of all features in this dataset is included at the end of this document.

The dataset is available on Canvas as the file `college_scorecard_2013.rds`.

### Questions

Use R to find answers to all of the following questions (that is, don't do any by hand or by point-and-click). Save your code in an R script. Try to complete at least one every day until the assignment is due.

1. How many observations are recorded in the dataset? How many colleges are recorded?
2. How many features are there? How many of these are categorical? How many are discrete? Are there any other kinds of features in this dataset?
3. How many missing values are in the dataset? Which feature has the most missing values? Are there any patterns?
4. Are there more public colleges or private colleges recorded? For each of these, what are the proportions of highest degree awarded? Display this information in one graph and comment on what you see.
5. What is the average undergraduate population? What is the median? What are the deciles? Display these statistics and the distribution graphically. Do you notice anything unusual?
6. Compare tuition graphically in the 5 most populous states. Discuss conclusions you can draw from your results.
7. For the following questions, use code to justify your answer:

   **Part a.** What is the `name` of the university with the largest value of `avg_sat`?

   **Part b.** Does the university with the largest amount of `undergrad_pop` have open admissions?

   **Part c.** List the zip code of the *public* university with the smallest value of `avg_family_inc`.

   **Part d.** Does the university you found in part b. also have the largest amount of `grad_pop`?
8. For schools that are *for-profit* in `ownership` and issue Bachelor's degrees as their `primary_degree`, do the following:

   **Part a.** Visualize `revenue_per_student` and `spending_per_student` and describe the relationship. What issues may arise when fitting a linear regression model?

   **Part b.** Create a new variable called `total_net_income`. Think carefully about how this variable would be calculated. Visualize the top 5 earning schools.

9. Now, examine the relationship between `avg_sat` and `admission` for all schools.

   **Part a.** Use an appropriate plot to visualize the relationship. Split the data into **two groups** based on their combination of `avg_sat` and `admission`. Justify your answer. *Hint: How does the variance of `admission` depend on values of `avg_sat`?*. Define this variable as `group`.

   **Part b.** Using code to justify your answers, comment on how the following continuous variables change depending on `group`:

   (a) `med_10yr_salary`

   (b) The percentage of `race_white` and `race_asian` combined

   (c) The percentage of graduate students enrolled at a university

   **Part c.** Using code to justify your answers, comment on whether the categorical variables are dependent or independent of `group`:

   (a) `open_admission`

   (b) `main_campus`

   (c) `ownership`

   (d) Whether the university has more than 1 `branch` or not

10. Examine the relationship between `avg_10yr_salary` using `avg_family_inc` for all schools.

    **Part a.** Use an appropriate plot for these two variables. Fit a linear regression model that predicts `avg_10yr_salary` using `avg_family_inc`. Add this line to the plot you used. Investigate the groups of points that may be affecting the regression line.

    **Part b.** Describe a categorical variable that would improve the fit of the regression line based on your investigation in part a. What would the levels of this variable be?

Assemble your answers into a report. Please do not include any raw R output. Instead, present your results as neatly formatted[3] tables or graphics, and write something about each one. You must **cite your sources**. Your report should be **no more than 8 pages** including graphics, but excluding code and citations. The page limit is deliberately low so that you will think carefully about what information is important to include.

### What To Submit

Email a digital copy to `spring18stat141a@gmail.com`. The digital copy must contain your report (as a PDF) and your code (as one or more R scripts).

Additionally, submit a printed copy to the box in the statistics department office[4]. The printed copy must contain your report and your code (in an appendix). Please print double-sided to save trees. It is your responsibility to make sure the graphics are legible in the printed copy!

### Data Documentation

The dataset contains the following features:

| | |
|---|---|
| `unit_id` | unique campus ID number |
| `ope_id` | unique college ID number |
| `main_campus` | whether this the main campus |
| `branches` | number of campuses for this college |
| `open_admissions` | whether this college has open admissions |

[1] `https://collegescorecard.ed.gov/data/`

[2] These features can but do not necessarily have to be present in the dataset!

[3] See the graphics checklist on Canvas.

[4] 4th floor of Mathematical Sciences Building

| | |
|---|---|
| `name` | name |
| `city` | city |
| `state` | state |
| `zip` | zip code |
| `online_only` | whether college is online-only |
| `primary_degree` | most common degree awarded |
| `highest_degree` | highest degree awarded |
| `ownership` | ownership (public, nonprofit, or for profit) |
| `avg_sat` | mean SAT score of students |
| `undergrad_pop` | undergraduate population |
| `grad_pop` | graduate student population |
| `cost` | estimated total cost without financial aid |
| `net_cost` | estimated total cost with financial aid |
| `tuition` | in-state tuition cost |
| `tuition_nonresident` | out-of-state tuition cost |
| `revenue_per_student` | amount college earns per student |
| `spend_per_student` | amount college spends per student |
| `avg_faculty_salary` | mean faculty salary |
| `ft_faculty` | % of full-time faculty |
| `admission` | % of applicants admitted |
| `retention` | % of students that stay more than 1 year |
| `completion` | % of students that graduate within 6 years |
| `fed_loan` | % of students that take out federal loans |
| `pell_grant` | % of students that receive Pell grants |
| `avg_family_inc` | mean family income of students |
| `med_family_inc` | median family income of students |
| `avg_10yr_salary` | mean salary of students 10 years after starting college |
| `sd_10yr_salary` | standard deviation of salary of students 10 years after starting college |
| `med_10yr_salary` | median salary of students 10 years after starting college |
| `med_debt` | median debt of students at graduation |
| `med_debt_withdraw` | median debt of students at withdrawal |
| `default_3yr_rate` | % of students that default on loans after 3 years |
| `repay_5yr_rate_withdraw` | % of withdrawn students that have partially or completely repaid loans after 5 years |
| `repay_5yr_rate` | % of graduated students that have partially or completely repaid loans after 5 years |
| `avg_entry_age` | mean student age at entry |
| `veteran` | % of students that are veterans |
| `first_gen` | % of first-generation college students |
| `male` | % of male students |
| `female` | % of female students |
| `race_white` | % of white students |
| `race_black` | % of black students |
| `race_hispanic` | % of Hispanic students |
| `race_asian` | % of Asian students |
| `race_native` | % of Native American students |
| `race_pacific` | % of Pacific Islander students |
| `race_other` | % of students of mixed/unspecified race |

For more detailed information, see the original documentation provided by the Department of Education: <`https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf`>.

The `clean_college_scorecard.R` file in the `extras/` directory on Canvas shows how feature names in this dataset correspond to the original.

### Relevant Functions

`getwd()`, `setwd()`, `readRDS()`, `names()`, `colnames()`, `rownames()`, `nrow()`, `ncol()`, `dim()`, `length()`, `str()`, `summary()`, `table()`, `prop.table()`, `mean()`, `median()`, `sd()`, `quantile()`, `fivenum()`, `cor()`, `max()`, `min()`, `plot()`, `boxplot()`, `density()`, `hist()`, `dotchart()`, `matplot()`, `legend()`, `smoothScatter()`, `par()`, `which.max()`, `which.min()`, `order()`, `sort()`, `is.na()`, `typeof()`, `class()`, `sapply()`