Assignment 2

STA 141A

Due **October 16 at 10:30am**.

In this assignment, you'll continue to analyze the U.S. Department of Education's College Scorecard data set from Assignment 1. The data set and documentation are available on Canvas.

1. Are there any features with no missing values? Which features have the most missing values? Explore the missing values and report any patterns you find.
2. Explore student populations for the universities. Are there any schools with unusual populations? What is the relationship between undergraduate and graduate populations? Are there exceptions to the relationship?
3. Explore the program percentages for the universities. What programs are the most popular? What programs are the least popular? Are there any program percentages that show patterns different from the others?
4. How does tuition vary across different states? Is there a relationship between the number of universities in a state and tuition? Do these characteristics differ for in-state tuition and out-of-state tuition?
5. Which colleges have the most diverse demographics? Make sure to explain how you measured "diversity" for this problem, in addition to discussing your conclusions.
6. Answer 2 of the questions you invented for Assignment 1, Problem 12. Use statistical summaries (including graphics) as evidence to support your conclusions. Also make sure to clearly state each question before your answer!
7. Reflect on the questions you answered in Problem 6. Did they lead to interesting conclusions? Why or why not? Did they raise new questions? Is it the question that makes a result interesting, the data, or both? Explain.
8. List and answer 2 follow-up questions raised by any of the work you did for this assignment. Along with each question and answer, make sure to explain what raised the question for you.

Assemble your answers into a report. Please do not include any raw R output. Instead, present your results as neatly formatted tables or graphics, and write something about each one. You must **cite your sources**. Your report should be **no more than 10 pages** including graphics, but excluding code and citations. Think carefully about what information is important to include.

When you are finished, submit:

- A digital copy on Canvas. The digital copy must contain your report (1 PDF) and your code (1 or more R scripts).
- A printed copy in lecture. The printed copy must contain your report and your code (in an appendix). Please print double-sided to save trees.

Your submission will be graded according to the STA 141A grading standards, which are available on Canvas.