Assignment 1

STA 141A

In this assignment you'll examine the College Scorecard data set, which powers https://collegescorecard.ed.gov/. Use the `college_scorecard.rds` data file available on Canvas. The RDS file is a subset of the complete College Scorecard data set, with some years removed and the feature names changed to the official "developer-friendly" names.

Documentation about the data set is also available on Canvas. It is a good idea to read the "About" section on page 2 of the documentation to learn more about the data set.

1. What is the purpose of this data set? Who created it? What are the sources for the data?
2. How many rows are there? What do rows represent in this data set?
3. How many columns are there? What do columns represent?
4. What range of years does the data set span? How many colleges are recorded for each year?
5. What are the 5 states with the most colleges? How many colleges do they have? What are the states with the fewest colleges? Make a hypothesis about why some states have a lot of colleges. Can you confirm your hypothesis (possibly using outside sources)?
6. For public schools in the 2014 academic year, create a scatter plot of average net price versus median student earnings after 10 years (`earn_10_yrs_after_entry.median`). Comment on any patterns you see, interpreting what they mean for college students.
7. Create the plot from the previous question for private for-profit schools. How do the two plots compare? Whether you see similarities or differences, discuss what your results imply about public schools and private for-profit schools.
8. Continuing from the previous two questions, what can you say about private non-profit schools? Use evidence to support your claims.
9. Create a bar plot that shows the number of schools recorded for each year. Use a separate color for each year and a separate group of bars for each kind of ownership. Are there any trends? Comment on what you see.
10. The ID for UC Davis is 110644. Create a line plot (with points marked) that shows the admission rates for UC Davis for each year. Do the admission rates change much from year to year?
11. Discuss the R data types and statistical data types of the features. Does each R data type map to just one statistical data type? Give examples.

12. List 3 questions you think can be answered with this data set. For each question, explain why the question is compelling (Who would benefit from knowing the answer, and why?), which variables you would use to answer the question, and how these variables help you answer the question. You do not need to write any code for this problem.

    Due **October 9 at 10:30am**. Submit a digital copy of your report (PDF) and code (R script) to Canvas. Additionally, please submit a printed copy in lecture. Maximum 10 pages (excluding code).

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