Problem Set #1 - The College Dataset

Econ 224

Due Date: Sunday, September 2nd by 11:59pm

Instructions

Submit your code and solutions to the following questions as an RMarkdown document. In particular, upload both the `.Rmd` file you used to generate your report and the resulting `.html` output to *canvas* by the due date listed above. If you need extra help with R Markdown, we suggest that you complete Chapters 1 and 2 of the course "Reporting with R Markdown" on *Datacamp*. Late work will only be accepted in exceptional circumstances, so you are better off submitting an incomplete problem set rather than nothing at all.

You may discuss this problem set with your classmates, provided that you adhere to the *empty hands* policy: after any such discussion, all parties must leave the room *empty-handed* i.e. without code files or written notes or any kind. In other words, the final code and write-up that you produce must be entirely your own work. If you discuss the problem set with any other students be sure to list their names at the top of your problem set. You are likewise welcome to consult printed or internet resources provided that you cite them. Violations of this policy constitute cheating and will be reported to the Office of Student Conduct.

Questions

Begin by installing the R package `ISLR` using the command `install.packages('ISLR')`. Then load both the `ISLR` and `tidyverse` packages. The dataset we'll be working with is `College` from the `ISLR` package. Before proceeding, we'll convert it to a tibble as follows:

```
college <- as.tibble(College, rownames = 'name')
```

The preceding command adds a column called `name` to the tibble that contains the rownames of the original `College` dataset. (We'll need these to know which college is which!) Note the differences between `College` with a lowercase "c" versus `college` with an uppercase "c" – the latter is the tibble we'll use in this assignment. Answer each of the following using appropriate `dplyr` and `ggplot2`.

1. Read the help file for the `College` dataset using the command `?College`. What is the source of this dataset? How many variables and observations are there, and what information does each variable contain?
2. Create a column called `acceptRate` that contains the number of acceptances divided by the total number of applications, and store it in the `college` tibble.
3. What are the five most selective institutions in the dataset, based on their acceptance rates? What are the five most selective public institutions in the dataset?
4. Judging by acceptance rates, which are more selective on average: private or public institutions?
5. A college's matriculation rate is defined as the number of students who enroll as a fraction of those who are accepted. Create a corresponding column called `matricRate` and store it in the `college` tibble.
6. Make a scatterplot of acceptance rates versus matriculation rates, using different colors to indicate public and private institutions. Be sure to add an appropriate title to your plot.
7. Calculate the sample correlation between acceptance and matriculation rates separately for public and private institutions. Briefly discuss your results.
8. Pose a question that interests you and answer it by computing summary statistics and making appropriate plots based on the `college` dataset. This question will be graded both on the interestingness of your question and the quality of your answer to it. (Suggestion: a question such as "what is the average graduation rate?" is too simple to be very interesting.)