Homework 9 Graded

Sheridan Grant

Must be uploaded to Canvas under "Homework 9 Graded" by
Monday, June 1 at 11:59pm

Instructions

Format your .RMD file using the template on the course website. **Submit the .RMD file, the knitted .html output, and any other files (covid and 115th Congress Senate members data) or folders needed as a single .zip file.**

The grader will be compiling your .RMD file and making sure it knits. Any libraries/packages needed should be near the top of the .RMD file, so the grader can make sure they're installed. **If your code does not knit and there is no immediate fix, the grader will grade your HTML for a [-10pts] penalty.**

Any time I ask you to demonstrate something, show something, generate something, etc., you must provide the code and/or text commentary that does so.

Finally, we will be giving [5pts] for code style and cleanliness. For any function you define, include a comment on the line above the function saying what the function expects as input and what it outputs. If you do this and the rest of your code is reasonably neat then this is an easy [5pts].

1 Time is of the Essence

In this question, you will time how long (in seconds) the computer takes to run code using the `Sys.time` function. Make sure you understand how it works. Make sure you have `tidyverse` and `data.table` installed. Each part of this question requires that you time just a single line of code for each package.

(a) How long do `tidyverse`'s `readr` and `data.table` take to read the coronavirus data? Which is faster, and by how much? You may use the data you originally downloaded in mid-April, and don't use any optional arguments to the functions that read the data (i.e. column formatting). [2pts]

(b) How long do `tidyverse`'s `dplyr` and `data.table` take to sort the coronavirus data by date? Which is faster, and by how much? [2pts]

(c) How long do tidyverse's dplyr and data.table take to spread the coronavirus data into the format where Confirmed and Deaths are their own columns? Which is faster, and by how much? The line of code you write for each package should assign a data.frame/data.table with 6 columns—Date, Country_Region, Province_State, Admin2, Confirmed, and Deaths—to a variable. [6pts]

2 Trolling the Trolls

I have put a small subset of the Quora "Troll Question" data on the website. Remember, for the target variable, 1 means Troll, 0 means Legit Question.

(a) What percent of the questions start with "Why," "Where," "Who," "When," "What," "How," or something else? Capitalization doesn't matter. [2pts]

(b) What percent of the questions include a single period as the end of a sentence? Are these questions more or less likely to be a Troll question than a randomly selected question? [3pts]

(c) In class Wednesday, we explored two concepts—discussion of vaccines and use of multiple question marks in a row—that led to a higher likelihood of being a Troll. Use those two concepts (just look for discussion of vaccines, not autism) as well as the one from the previous part to build a logistic regression model with target as the outcome and those 3 concepts as (binary) covariates. Train it on the first $90,000$ rows of the data (no shuffling) and use a prediction rule to test its accuracy on the rest of the data. Does your prediction rule do better in terms of pure accuracy than just guessing "not troll" every time? Explain when accuracy may be a bad prediction metric for a statistical model. [5pts]

3 Project Prep

Write one of your project scientific questions of interest that can be answered with a statistical model we've used in class, write some code to usefully display summary tables and/or graphs of the variables involved, write the code that answers the scientific question, and write a paragraph explaining your results as you would in the final report. Don't submit your data file—we'll just look at the compiled HTML to grade this part. Make sure we can understand what's going on without having access to the data! [10pts]