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 to calculate the exact decimal number of seconds a line of code took to run (don't just print the start time and end time). Make sure you have `tidyverse` and `data.table` installed. Each part of this question requires that you time one or two lines of code for each package.

(a) How long do `tidyverse`'s `readr` and `data.table` take to read the coronavirus data? Which is faster? You may use the data you originally downloaded in mid-April. Reading the data involves making sure the columns are the correct data types, so for `data.table` you'll need to set the Date type and for `readr` use the code template below. [2pts]

```
read_csv("../data/coronavirus.csv",
         col_types = cols(Admin2 = col_character(),
                          Date = col_date('%m/%d/%Y')))
```

(b) How long do `tidyverse`'s `dplyr` and `data.table` take to sort the coronavirus data by date? Which is faster? Make sure the sorted data are stored in the original variable name in your line of code. [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? 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—and half the number of rows of the original data frame 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. Please put the answer, a vector of 7 numerics that add up to 1 corresponding to the order in the previous sentence, in a vector called `qTypes`. [2pts]

(b) What percent of the questions include a single period, that does not begin the question, followed by exactly one space? (These questions will either include sentences, or words such as "Mr." and "Mrs.," but probably more of the former.) Are these questions more or less likely to be a Troll question than a randomly selected question? Use the regular expression `'^.*\\.{1} {1}'`. [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 of the form $f_{cutoff}(x) = \mathbf{1}[x > cutoff]$ 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 (either one we've used in class or not), 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! **If you are working with a partner, please each submit something different. This should be considered a "first draft" of your final project's answer to these questions, especially for partner pairs as you will have to be more thorough and perhaps ask more difficult questions to meet the word limit.** [10pts]