STA130H1S – Winter 2021

Week 1 Problem Set

Yixing Xu

Instructions

How do I hand in these problems for the January 14th?

Your complete Rmd file that you create for these practice problems AND the resulting pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/206597/assignments/512542) by 11:59 a.m. ET, on Thursday, January 14th. Late problem sets are NOT accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focussed on writing. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

[Question 1]

Below is a 'math square puzzle'. The value for each row and column is shown after the equals signs, but the operations are missing. The possible operations are: addition, subtraction, multiplication and division.

For example, if a row said 2 blank 7 = 14, we could figure out that blank has to be multiplication (*).

a) Write out the equations below and assign them to the appropriate names. The first rows has been completed as an example.

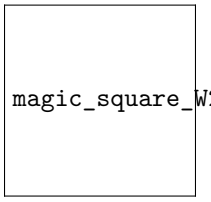

Figure 1: A math square puzzle

```r
# Row 1
r1 <- 4 * 3

# Row 2 (r2)
r2 <- 2+8

# Column 1 (c1)
c1 <- 4/2

# Column 2 (c2)
c2 <- 3-8
```

b) Now, let's check you are right using what we've learned about logicals. Create a vector called `my_answers` in the order: rows 1 and 2 and then columns 1 and 2. Use the names you assigned, i.e. 'r1' and 'c2', not the numbers.

```r
my_answers <- c(r1, r2, c1, c2)
```

c) Create a numeric vector called `square_answers` with the values from after the equal signs from the math square above, i.e. 12, 10, 2, and -5, in that order.

```r
square_answers <- c(12, 10, 2, -5)
```

d) Assign the name `check` to the result of this code: `square_answers == my_answers`.

```r
check <- square_answers == my_answers

# after saving the result of `square_answers == my_answers` to check, you can type "check" (no quotes)
check
```

```
## [1] TRUE TRUE TRUE TRUE
```

```r
# alternatively, write check by itself on a new line after writing the code and run the whole chunk aga
```

e) Which of the following best describes what `check` is?

(i) A single value counting how many correct rows and columns you calculated.
(ii) A numeric vector of the differences between the math square answers and your answers (should be all 0s if you got them all right).
(iii) A character vector of 'TRUE' and 'FALSE', 'TRUE' for each answer that matches and 'FALSE' for any that don't.
(iv) A logical vector of TRUE and FALSE, TRUE for each answer that matches and FALSE for any that don't.
(v) A single logical value TRUE or FALSE, TRUE if all the values match, FALSE if any of the values don't match.

Answer: iv

f) Use the `sum()` function (and the fact of coercion) to get a single numeric value for the number of rows and columns you got right. If you got everything right your result should be 4.

```
corrects <- sum(c(check))
print(corrects)
```

```
## [1] 4
```

[Question 2]

For this question we will work with data about the TV show Avatar: The Last Airbender. See the data description video on Quercus for more information and attributions:

a) The name of the data set is `avatar.csv`. Load the data and save it under the name "avatar".

```
library(tidyverse)
avatar <- read_csv(file = "avatar.csv")

# Tip: don't forget to put quote marks around the name of the dataset inside the function
```

b) We have learned two functions this week that let us quickly get an idea of our data. Apply both of them to the avatar data.

```
glimpse(avatar)
```

```
## Rows: 9,992
## Columns: 10
## $ book            <chr> "Water", "Water", "Water", "Water", "Water", "Water...
## $ book_num        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ chapter         <chr> "The Boy in the Iceberg", "The Boy in the Iceberg",...
## $ chapter_num     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ character       <chr> "Katara", "Sokka", "Katara", "Sokka", "Katara", "Ka...
## $ full_text       <chr> "Water. Earth. Fire. Air. My grandmother used to te...
## $ character_words <chr> "Water. Earth. Fire. Air. My grandmother used to te...
## $ mention_appa    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA...
## $ director        <chr> "Dave Filoni", "Dave Filoni", "Dave Filoni", "Dave ...
## $ imdb_rating     <dbl> 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8...
```

```
head(avatar)
```

```
## # A tibble: 6 x 10
##   book  book_num chapter chapter_num character full_text character_words
##   <chr>    <dbl> <chr>         <dbl> <chr>     <chr>     <chr>
## 1 Water        1 The Bo~           1 Katara    Water. E~ Water. Earth. ~
## 2 Water        1 The Bo~           1 Sokka     It's not~ It's not getti~
## 3 Water        1 The Bo~           1 Katara    [Happily~ Sokka, look!
## 4 Water        1 The Bo~           1 Sokka     [Close-u~ Sshh! Katara, ~
## 5 Water        1 The Bo~           1 Katara    [Struggl~ But, Sokka! I ~
## 6 Water        1 The Bo~           1 Katara    [Exclaim~ Hey!
## # ... with 3 more variables: mention_appa <lgl>, director <chr>,
## #   imdb_rating <dbl>
```

c) Based on your answer to b) answer the following:

How many observations does the `avatar` data frame include? 9992

How many variables are measured for each observation? 10

How many rows and columns does the `avatar` data frame have? 9992 rows and 10 columns

[Question 3]

In this question, you will consider another example of survivor bias. In 1987, a study published in the Journal of the American Veternary Medicine Association reported that cats that survived falls from higher floors in high-rise buildings suffered fewer injuries than cats who fell from lower floors (e.g. more than 6 stories vs less than 6 stories). While this finding seems counterintuitive, the authors suggested that this might be due to the cats relaxing and re-positioning themselves for a relatively safer landing after they reached maximum speed during their fall. The data for this study was collected from cats who suffered falls and were brought to veternary clinics.

(a) Is this sample representative of all cats who suffer falls?

No, because the sample only includes data collected from cats who suffered falls and were brought to veternary clinics. This means that cats who suffered falls and instantly died or cats who suffered falls and were simply not brought to veternary clinics are excluded from this study.

(b) Do you expect that the average number of injuries for cats suffering falls calculated from this sample will be close to the true value?

No, because this sample is not representative of all cats who suffer falls, as the data does not include the cats who suffered falls and were not brought to veternary clinics. Also, cats who may have instantly died from the falls may not have been brought to the veternary clinics and the recorded injuries from those cats are not accounted for in the average number of injuries for cats suffering falls from this sample.

Part 2

Writing prompt:

Identify one of the course learning objectives (from the syllabus) that you are most excited about. Paraphrase the objective (i.e. describe in your own words). You are not allowed to directly quote the objective you have identified. After you have paraphrased the objective, describe why this objective is especially interesting to you.

Some guidlines/general reminders

- Remember to start with a small introduction that explains what STA130 is, and what you will be discussing in your writing prompt.
- Try to not send more than 20 minutes on the prompt.
- Aim for more than 200 words but less than 400 words.
- Use full sentences.
- While grammar is not the main focus of the assessment, it is important that you communicate in a clear and professional manner. I.e., no slang or emojis should appear.
- Be specific. A good principle when preparing a writing sample is to assume that your audience is not aware of the topic matter (or in this case has not read the syllabus). Therefore, you need to properly communicate what the objective is by paraphrasing it, putting it in your own words. You should not use quotations.

Ever since acquiring a passion for math since my competitive math days in middle school, I have always known I would want to do something heavily related to math as a career in the future. One of the possible routes for me was Data Science/Analysis, since math plays a major role. Because of this, I decided to take STA130, which is an Introductory Statistics/Data Analysis course. One of my main goals – which is also one of its learning objectives – is to be able to mathematically execute statistical analyses, especially with the aid of R. The reason why this knowledge and skill is so important to me is because I wish to be able to combine coding and mathematics, and data analysis using R and Python and more coding languages is what I want to learn. Coding is more of a means to an end to me, where the end goal is to be able to make a mathematical/statistical deduction; coding is also like an art form to me. It's a tool I can use to paint and create many masterpieces. That is why it is important to me to be familiar with R as a coding language, and it is also important to me that I can combine my passion for both coding and mathematics, like data science does.