STA130H1S – Fall 2022

Problem Set 1

() and STA130 Professors

Instructions

Complete the exercises in this `.Rmd` file and submit your `.Rmd` and knitted `.pdf` output through Quercus on September 15 by 5:00 p.m. ET.

Part 1: R Coding Practice

Question 1

For this question we will work with data about the old TV show Avatar: The Last Airbender.

- The data is stored in the file `avatar.csv` in the same directory as this file.

  This data was posted on github by user averyrobbins1 and subsequently featured on Tidy Tuesday. For more information see the above links; or, install the package with `devtools::install_github("averyrobbins1/appa")` and then type `help(appa)`.

```
# Write your answer below
# Don't forget to put quote marks around the data set name in the function

### (a) Load the data set from the file `avatar.csv` using `read_csv` (not `read.csv`) and save it as a
library(tidyverse)

   avatar<- read_csv("avatar.csv")
```

Hints to help fix common "gotchas"

- `Error in read_csv(avatar.csv) : could not find function "read_csv"`
  - *Have you loaded the appropriate libraries? I.e., `library(tidyverse)`?*
- `Error in standardise_path(file) : object 'avatar.csv' not found`
  - *Do you have quotes around the file name?*
- `Error: 'avatar.csv' does not exist in current working directory (...).`
  - *Are you running code as `<ctrl-shift-end>` (PC) or `<cmd-shift-enter>` (Mac)?*

(b) We have seen two functions that let us quickly get an idea of our data: `glimpse()` and `head()`. Using `%>%` "pipe" the avatar object you created into each of these functions.

```r
# Write your answer below

avatar %>% glimpse()
```

```
## 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, 1, ~
## $ chapter         <chr> "The Boy in the Iceberg", "The Boy in the Iceberg", "T~
## $ chapter_num     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ character       <chr> "Katara", "Sokka", "Katara", "Sokka", "Katara", "Katar~
## $ full_text       <chr> "Water. Earth. Fire. Air. My grandmother used to tell ~
## $ character_words <chr> "Water. Earth. Fire. Air. My grandmother used to tell ~
## $ mention_appa    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE~
## $ director        <chr> "Dave Filoni", "Dave Filoni", "Dave Filoni", "Dave Fil~
## $ imdb_rating     <dbl> 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8.1, 8.1,~
```

```r
# Write your answer below

avatar %>% head(

)
```

```
## # 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 Boy in the~           1 Katara    Water. E~ Water. Earth. ~
## 2 Water        1 The Boy in the~           1 Sokka     It's not~ It's not getti~
## 3 Water        1 The Boy in the~           1 Katara    [Happily~ Sokka, look!
## 4 Water        1 The Boy in the~           1 Sokka     [Close-u~ Sshh! Katara, ~
## 5 Water        1 The Boy in the~           1 Katara    [Struggl~ But, Sokka! I ~
## 6 Water        1 The Boy in the~           1 Katara    [Exclaim~ Hey!
## # ... with 3 more variables: mention_appa <lgl>, director <chr>,
## #   imdb_rating <dbl>
```

(c) Run the two code chunks below using (PC) or (MAC) or the "play" button, and then compare their output to the output of the `glimpse()` and `head()` functions above.

```r
avatar
```

```r
avatar %>% head(12) # <- try another number instead of 3... maybe 12?
```

- Is the `glimpse()` output or the `head()` output a `tibble`?

head()

- Which function allows you to look at the first `n` rows of a data set?

head()

- Which function lists data set columns vertically rather than horizontally so you can immediately see them all?

glimpse()

- 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?

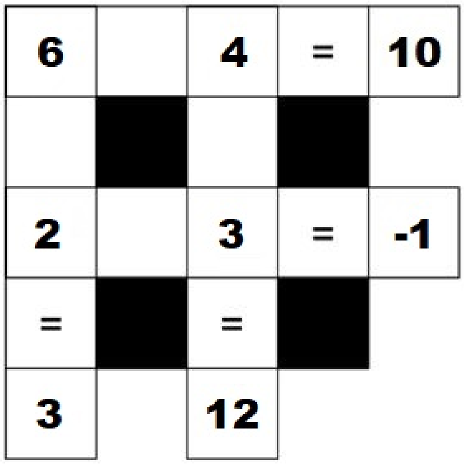
9992 rows and 10 columns

- Is the information for the three previous questions available from the `glimpse()` function or the `head()` function?

glimpse()

Question 2

Below is a 'math square puzzle'. The value for each row and column is shown after the equals signs, but the operations (`+`, `-`, `*`, `/`) producing the resuts are missing. For example, a row with "2 blank 7 = 14" is missing a multiplication.



6		4	=	10
2		3	=	-1
=		=		
3		12		

Figure 1: A math square puzzle

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

```
# Addition
r1 <- 6 + 4


# Subtraction
r2 <- 2 - 3

# Division
c1 <- 6/2

# Multiplication

c2 <- 4*3
```

(b) Now, let’s check each of your answers individually with the logical `==` operation.

```
r1 == 10
```

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

```
r2 == -1
```

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

```
c1 == 3
```

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

```
c2 == 12
```

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

(c) Now, let’s check each of your answers at the same time with logical `&` operations.

```
(r1 == 10) & (r2 == -1) & (c1 == 3) & (c2 == 12)
```

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

```

my_answers <- c(r1,r2,c1,c2)
square_answers <- c(10,-1,3,12)
my_answers == square_answers
```

Consider the code above relative to the code below using the `c()` “concatentation” function which “combines” objects into a *vector*.

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

```

correctness <- my_answers == square_answers
all(correctness)
```

Consider the code above relative to the code below using the `all()` function which checks if every element of a vector is `TRUE`.

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

(d) What is the benefit of using the `c()` and `all()` functions compared to just writing everything out with logical `==` and `&` operators?

**The number of operations in the instance of the logical == and & operators for suppose a bigger puzzle (ex.1000) would occur sequentially. It would be slower because the time taken would be (Time for one) multiplied by the number of operatioins. The vertorized computer operations would perform each one in parralel reducing the time. Using the `c()` and ‘all() would preform the operations simulaneouly so, it would be faster and less tedious * *

Hints

- Right now we just have `r1`, `r2`, `c1` and `c2`. But what if we had a bigger math square that went all the way up to, say, `r100` and `c100`?
- “Vectorized” computer operations do a series of individual observations in parallel, rather than sequentially. So just like writing out things sequentially takes a long time, doing operations sequentially with a computer also takes more time than just computing them in parallel.

(e) What is the difference between the code below and the `all(correctness)` code above?

```
sum(correctness)
```

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

The code below returns an integer value of the number of correct answers as opposed to the code above which provides a boolean value

Part 2: TUT communication/writing exercises *Primer Questions*

You are expected to be efficient with your time in this section, and should ***spend no more than 30 minutes* on Part 2**.

Question 1

(a) How is it that you have come to take STA130?

My reasoning for taking this course is because regardless of career goals the need to prove one's point in the workplace or at home can only be furthered with an indept knowledge of statistical analysis and impressive looking data-representation

(b) Do you currently have a sense of the kind of career (e.g., industry, company, type of work) you think you might want to pursue? Please describe your current thinking on this aspect of your university experience.

My goal is to become the chief economist for the World Bank.It's been my dream since middleschool and I intend to try and reach this goal.

Question 2

Suppose you ask your friends to name 10 songs produced prior to Dec 31, 1999 and 10 songs produced after Jan 1, 2000. Then, suppose you check the song statistics on Spotify.

(a) If the total number of times the older songs have been listened to is greater than the newer songs, would this confirm that music from earlier periods is better than music now?

No because that data is complelely based on my preferences. If an accurate understanding was neccessary it would require me to pick to songs that depict each time to the greatest extent

(b) If the average number of times (per user and per year) the older songs have been listened to is greater than the newer songs, would this confirm that music from earlier periods is better than music now?

Yes. The data seems to be of a bigger sample so it will be more accurate and make it more reliable so songs earlier would be better. It may depend on the integrity of the data but if that is correct, songs from the past would be better

(c) Could there be a systematic reason that the 10 songs produced prior to Dec 31, 1999 that your friend selected might be expected to have a higher number of listens?

It's probably because the songs prior to 1999 were just the only songs listened to be popular enough to still be listened to and the rest just died off earlier that didnt ever come on spotify so the older songs may be the best of there time while there is still time until the songs of the 2000s die off becuase they are relatively newer

Hints

- Are the 10 songs produced prior to Dec 31, 1999 that your friend selected fairly representative of all songs produced prior to Dec 31, 1999?
- This question is addressing ***survivorship bias***, which will be considered further in your upcoming tutorial class.

Additional Recommended Study Material

Did you finish quickly? Do you still have an unused course study time allocation? Would you like to cover this material a little bit more?

More Introductory R Material

The following materials all approach learning R slightly differently, but they each have quite good perspectives.

- R for Data Science Chapter 4: Workflow basics
- Rstudio Primer: Programming Basics
- The Department of Statistical Science Toolkit

Markdown

Markdown supports efficiency and productivity, and it's important for our class for the course project (and beyond).

- Markdown Tutorial
- RStudio Markdown
- RStudio Markdown Cheatsheet