STA130 Rstudio Homework

Problem Set 1

Josh Speagle & Scott Schwartz

Instructions

Complete the exercises in this `.Rmd` file and submit your `.Rmd` and knitted `.pdf` output through Quercus by 11:59 pm E.T. on Sunday, January 22.

Part 1: R Coding Practice

Question 1

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

- The data is stored in the file `avatar.csv` in the same directory as this file (HW1).

  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)`.

(a) Load the data set from the file `avatar.csv` using `read_csv` and save it as an object named "avatar".

```
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 learned about two functions in class 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. (See HW0 for some additional examples of using the "pipe".)

```
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,~
```

```
avatar %>% head()
```

```
## # A tibble: 6 x 10
##   book  book_num chapter chapt~1 chara~2 full_~3 chara~4 menti~5 direc~6 imdb_~7
##   <chr>    <dbl> <chr>     <dbl> <chr>   <chr>   <chr>   <lgl>   <chr>     <dbl>
## 1 Water        1 The Bo~       1 Katara  Water.~ Water.~ FALSE   Dave F~     8.1
## 2 Water        1 The Bo~       1 Sokka   It's n~ It's n~ FALSE   Dave F~     8.1
## 3 Water        1 The Bo~       1 Katara  [Happi~ Sokka,~ FALSE   Dave F~     8.1
## 4 Water        1 The Bo~       1 Sokka   [Close~ Sshh! ~ FALSE   Dave F~     8.1
## 5 Water        1 The Bo~       1 Katara  [Strug~ But, S~ FALSE   Dave F~     8.1
## 6 Water        1 The Bo~       1 Katara  [Excla~ Hey!    FALSE   Dave F~     8.1
## # ... with abbreviated variable names 1: chapter_num, 2: character,
## #   3: full_text, 4: character_words, 5: mention_appa, 6: director,
## #   7: imdb_rating
```

(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.

```
avatar
```

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

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

It's head() that output a 'tibble'

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

The head() function.

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

The glimpse() function.

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

9,992 observations.

- How many variables are measured for each observation?

10 variables.

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

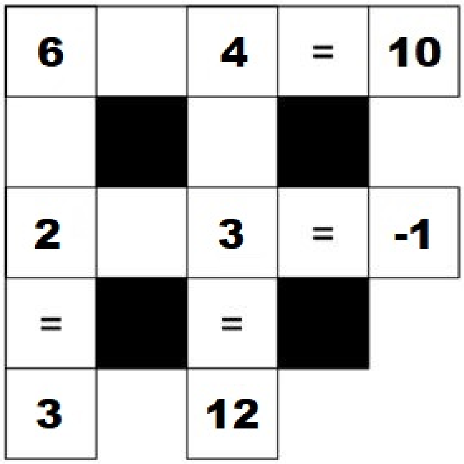
9,992 rows and 10 columns.

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

Available from glimpse() function.

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 (`*`) operation.



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.

```
# Row 1 (r1)
r1 <- 6 + 4

# Row 2 (r2)
r2 <- 2 - 3

# Column 1 (c1)
c1 <- 6 / 2

# Column 2 (c2)
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?

It consumes much less time when using 'c()' and 'all()' in a big dataset than just writing everything out with logical '==' and '&' operators. When using '==' and '&' operators when need to write the '==' and '&' to check every single variables. However, when using 'c()' and 'all()', we just need to combine every values into a vector.

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
```

1. When using 'all(correctness)' code above, it is checking whether each single value in the c() is match and equal to other values in another c(), returning Ture if matches and returning False if not matches.
2. When using 'sum(correctness)' code above, the aim is also checing whether each single value in the c() is match and equal to other values in another c(). However, the difference is, 'sum(correctness)' turn each True into number 1 and each False into number 0. This is the reason why it shows 4 in this circumstances.

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 this section.

Question 1

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

I am interesting in the area of data science. Firstly, taking STA130 make me get a chance, according to the university website, to go into this scientific area of study at the end of first year. Besides, I'm also excited about carrying out a variety of statistical analyses in R and interpreting the results of the analysis which I've never been able to touch before. Pursuing this course will enable me to accumulate one of the most essential skills in statistical analysis.

(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 career aspiration has always been and will always be to become an excellent researcher, working at the vanguard of data science and computer science. This thought came to me when I was in the last year of high school. I was in three internships at that year and I thought I don't love the environment in a company. And I really love the surroundings in university, most of the people are engaged and pure in scientific research.

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?

From my prospective, No it wouldn't. Because the data is only from my friends which really relates to personal preferences. Moreover, 10 songs is a really small sampling compared to the whole song production industry. The time is also a problem, it doesn't control the listening to be the same to compare.

(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?

No it wouldn't. As mentioned before, the data is only from my friends which really relates to personal preferences. Maybe they just chose some very famous songs in the past and some special-interest songs after Jan 1, 2000, which is an unfair compare.

(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?

Actually, we are currently teenagers and most of our friends are teenagers. We are unlikely to be born by our parents before the year 2000, which means that the songs they chose are possibly the most prevalent songs in the past and they've heard about it. After listening to them, they might think they are good and list those songs in the 10 songs produced prior to Dec 31, 1999. Because they are not in that time, they are not familar with other songs. However, the songs after 2000 are much better matching their own taste. Therefore, the 10 songs produced prior to Dec 31, 1999 that our friend selected might be expected to have a higher number of listens

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 the first Tutorial class.