Module 2 synchronous class

Prof. Caetano

2025-01-14

The first part of this file has the code for our synchronous class, while the second part includes code from the videos. While editing this file in the source pane you can use the stacked lines button at the top right of the source pane to view a table of contents for this document.

Synchrounous class

Quick recap of RMarkdown basics

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter* or *Cmd+Shift+Return*.

Set up library chunk

```
x <- 2

age = c(1,2,3,10)
```

The mean age is 4 years old.

Sometimes, when you load a package, R prints some messages to tell us what it just did. If you don't want the messages above to appear in my final document, you can put 'message=FALSE' to the top part of the chunk.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I* or *Cmd+Option+I*.

When you save the notebook, an PDF file containing the code and output will be saved alongside it (click the *Knit* button or press *Ctrl+Shift+K* or *Cmd+Shift+K* to Knit the PDF file).

Avatar data

Let's start by loading our data.

Our data only appears in the Environment pane if we SAVE it as an object in R, using the assignment operator.

Let's view the data the point and click way. We will often talk about the rows as the observations and the columns as the variables.

Let's look at the data in more 'code-y' ways.

Pipes

Let's do something kind of silly. What do you expect to get as the result of this code? (Note: Keyboard shortcut for pipes: *Ctrl+Shift+M* or *Cmd+Shift+M*)

```
# head(avatar_data) %>% glimpse()
```

Notice the number of rows.

Please post any questions you have about R, RStudio or JupyterHub in Piazza. You are encouraged to answer your classmates' questions for quickest response times, and the teaching team will review all answers to make sure they are complete and add to them if necessary! If you have a question, it is very likely someone else has the same question too!

Tidyverse

In this block of code try to reduce your data frame to only contain lines stated by Katara.

Now try to reduce your data frame to only contain the variables of `mention_appa` and `director`.

Exercise: Take 3 minutes to write some code that will calculate the number of lines that Katara and Aang each say, and of those lines which proportion mention Appa.

Hypothesis

Let's try to run a hypothesis test to see if Aang or Katara mentions Appa more. (Note, the results may vary based off the employed test).

Null and Alternative Hypothesis:

Test Stat

$$Z^* = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

(https://online.stat.psu.edu/stat415/lesson/9/9.4)

Go to pollev.com/sta to try this out!

```
#avatar_data %>% select(mention_appa, character) %>% head(4)

#avatar_data %>% filter(character=="Aang") %>% select(mention_appa)

#avatar_data %>% select(mention_appa) %>% group_by(character) %>% head()

#avatar_data %>% group_by(character) %>% select(mention_appa)
```

Visualizations

Create a histogram of `imdb_rating` in base R:

Create a histogram of `imdb_rating` in using ggplot:

More info to play with here: https://ggplot2.tidyverse.org/reference/geom_histogram.html

Quantitative vs. Qualitative Variables

What is an example of a quantitative variable in the data?

What is an example of a qualitative/categorical variable in the data?

Which visualizations are appropriate for either quantitative vs qualitative variables?

Quantitative:

Qualitative/categorical:

Other Useful Functions

Summary

Use the `summary()` function to learn more about the data

Missing-ness

You can use the function `is.na()` to assess if a value is missing, and `!` means NOT and can use `filter()` in conjunction with the other two functions to remove missing values in the data frame.

Let's create a new data that removes the observations with missing imbd ratings.

Simulation

There are some functions within R that allow you to simulate data. Some useful functions are `set.seed()`, `sample()`, `sample_n()`, `rnorm()`, 'runif()", etc. If time permits we can simulate some data.

Video code

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

R Basics (Part 1)

Using the console as a calculator

```
2 + 2
```

```
[1] 4
```

```
314 - 15
```

```
[1] 299
```

```
77 * 88
```

```
[1] 6776
```

```
14/2
```

```
[1] 7
```

```r
2^4
```

```
[1] 16
```

```r
(2+4)*3.5
```

```
[1] 21
```

```r
# note: space don't matter 2+2 is the same as 2 + 2
```

Saving objects in R

```r
x <- 2+2
my_name <- "Prof. Caetano"
```

Vectors

```r
my_vector <- c(1, 1, 2, 3, 5, 8, 13)
is.numeric(my_vector)
```

```
[1] TRUE
```

```r
is.character(my_vector)
```

```
[1] FALSE
```

Comments in R

```r
# I don't want the computer to read this comment about how I am afraid computers will take ov
my_vector <- c(1, 1, 2, 3, 5, 8, 13)
my_vector
```

```
[1]  1  1  2  3  5  8 13
```

Meet the data

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

```
Rows: 9992 Columns: 10
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (6): book, chapter, character, full_text, character_words, director
dbl (3): book_num, chapter_num, imdb_rating
lgl (1): mention_appa

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Note that the output below is R being helpful and telling us how it has
# interpretted each column of our csv file. Red text isn't always an error!
```

R Basics (Part 2)

The trouble with Tibbles

This is just the code shown in the video, for completeness. We don't need to run it again so I have set eval (whether or not the chunk should be evaluated) to FALSE.

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

glimpse() and head()

```
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, 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,~
```