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

```r
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.

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

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

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.

```
glimpse(avatar_data)
```

```
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,~
```

```
head(avatar_data)
```

```
# 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!
# i 3 more variables: mention_appa <lgl>, director <chr>, imdb_rating <dbl>
```

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*)

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

```
Rows: 6
Columns: 10
$ book            <chr> "Water", "Water", "Water", "Water", "Water", "Water"
$ book_num        <dbl> 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
$ 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
$ 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
```

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.

```
Katara_data <- avatar_data %>%
  filter(character == "Katara")
```

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

```
avatar_data %>%
  select(mention_appa, director)
```

```
# A tibble: 9,992 x 2
   mention_appa director
   <lgl>        <chr>
 1 FALSE        Dave Filoni
 2 FALSE        Dave Filoni
 3 FALSE        Dave Filoni
 4 FALSE        Dave Filoni
 5 FALSE        Dave Filoni
 6 FALSE        Dave Filoni
 7 FALSE        Dave Filoni
 8 FALSE        Dave Filoni
 9 FALSE        Dave Filoni
10 FALSE        Dave Filoni
# i 9,982 more rows
```

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.

```
# A tibble: 2 x 3
  character num_lines prop_Appa
  <chr>         <int>     <dbl>
1 Aang           1796    0.0551
2 Katara         1437    0.0188
```

Table 1: Statistics about lines said by characters Aang and Katara

character	Number of lines	proportion of lines mentioning Appa
Aang	1796	0.0551225
Katara	1437	0.0187891

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). 2

Null and Alternative Hypothesis:

$$H_0 : p_A = p_K$$
$$H_A : p_A \neq p_K$$

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)

```r
phat1 = 0.0551
phat2 = 0.0188
n1= 1796
n2= 1437

Z_star = (phat1-phat2)/sqrt(phat1*(1-phat1)/n1 + phat2*(1-phat2)/n2)
Z_star
```

```
[1] 5.61287
```

```r
pvalue = 2*(1-pnorm(Z_star))
```

The p-value is extremely small, thus there is evidence against the claim that Aang and Katara mention Appa at the same frequency. In fast, there is evidence to suggest that Aang mentions Appa more.

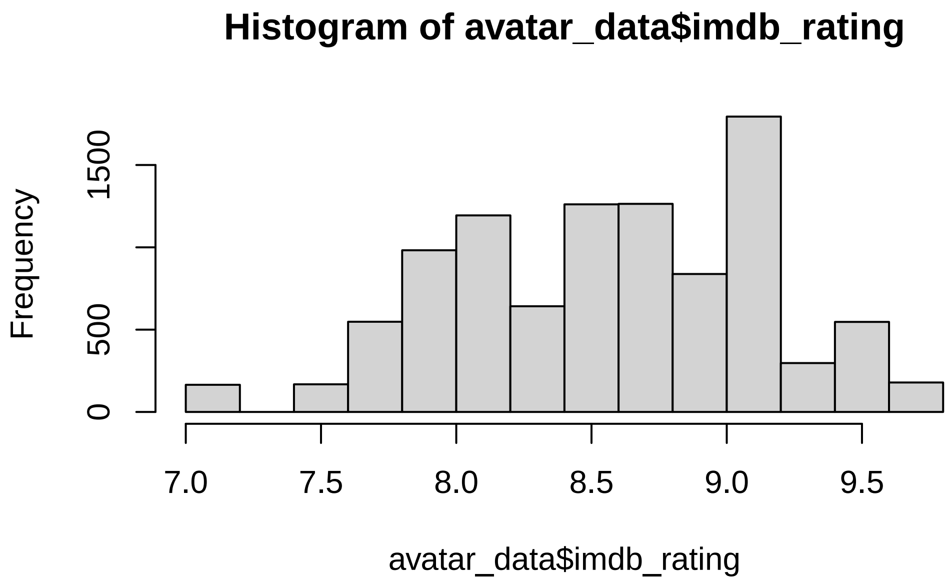
Go to pollev.com/sta to try this out!

```r
#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:

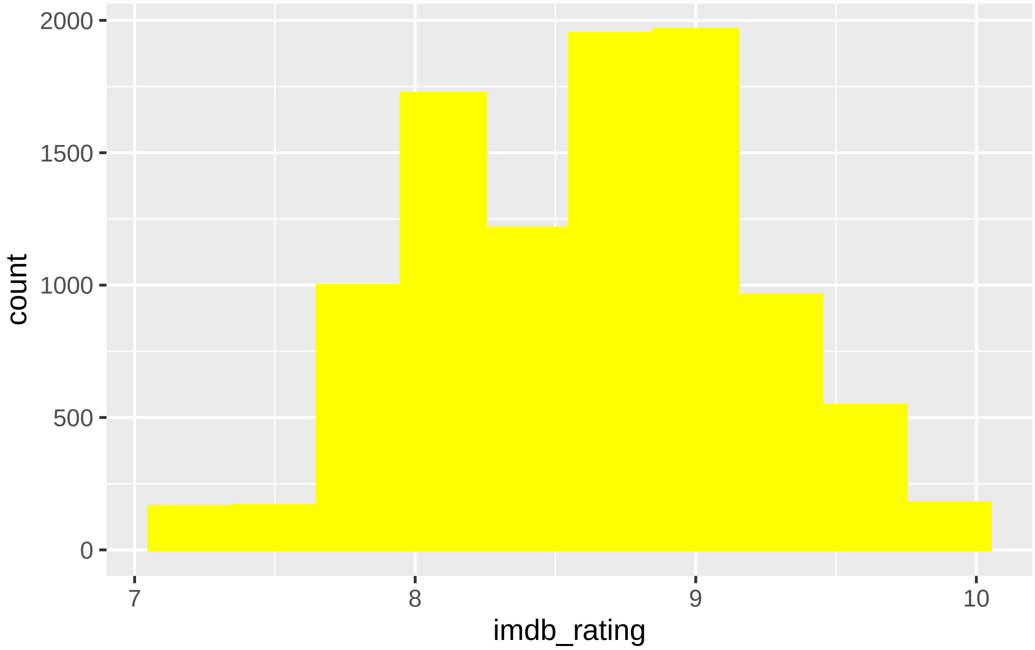
```
hist(avatar_data$imdb_rating) # hist(avatar_data[, column])
```



Create a histogram of `imdb_rating` in using ggplot:

```
ggplot(data = avatar_data, aes(x=imdb_rating)) +
  geom_histogram(colour="yellow", fill="yellow",
                 bins=10)
```

```
Warning: Removed 113 rows containing non-finite outside the scale range
(`stat_bin()`).
```



```
ggplot(data = avatar_data, aes(x=imdb_rating)) +
  geom_histogram(colour="maroon", fill="gray", bins=10) +
  labs(title = "Figure 1: Histogram of IMDb rating for
       lines of \nAvatar the last Airbender TV show",
       x="IMDb rating", y="Number of Lines") +
  theme_light()
```

```
Warning: Removed 113 rows containing non-finite outside the scale range
(`stat_bin()`).
```

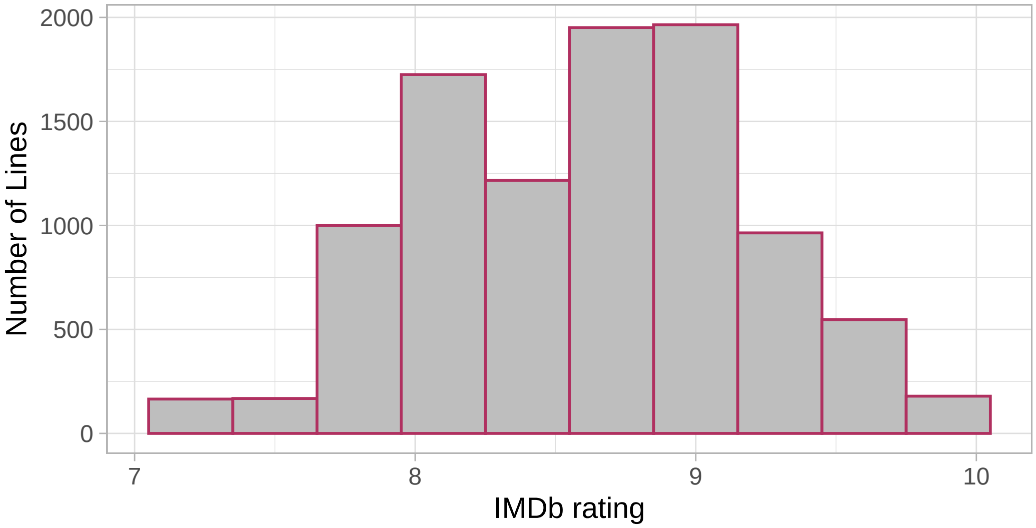


Figure 1: Histogram of IMDb rating for lines of
Avatar the last Airbender TV show

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? Chapter Number, imdbrating

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

Which visualizations are appropriate for either quantitative vs qualitative variables?

Quantitative: histogram, dotplot, scatterplot (needs two quantitative variables)

Qualitative/categorical: barplot, piechart (not preferred)

Other Useful Functions

Summary

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

```
summary(avatar_data)
```

```
     book              book_num       chapter           chapter_num
 Length:9992        Min.   :1.000   Length:9992        Min.   : 1.00
 Class :character   1st Qu.:1.000   Class :character   1st Qu.: 6.00
 Mode  :character   Median :2.000   Mode  :character   Median :11.00
                    Mean   :1.981                      Mean   :10.53
                    3rd Qu.:3.000                      3rd Qu.:15.00
                    Max.   :3.000                      Max.   :21.00
  character          full_text         character_words    mention_appa
 Length:9992        Length:9992        Length:9992        Mode :logical
 Class :character   Class :character   Class :character   FALSE:9813
 Mode  :character   Mode  :character   Mode  :character   TRUE :179



   director          imdb_rating
 Length:9992        Min.   :7.100
 Class :character   1st Qu.:8.200
 Mode  :character   Median :8.600
                    Mean   :8.616
                    3rd Qu.:9.100
                    Max.   :9.800
                    NA's   :113
```

Go to pollev.com/sta to test your knowledge

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.

```r
avatar_noNAs <- avatar_data %>% filter(!is.na(imdb_rating))
```

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

```r
library(tidyverse)
```

R Basics (Part 1)

Using the console as a calculator

```r
2 + 2
```

```
[1] 4
```

```r
314 - 15
```

```
[1] 299
```

```r
77 * 88
```

```
[1] 6776
```

```r
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

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

Vectors

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

```
[1] TRUE
```

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

```
[1] FALSE
```

Comments in 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,~
```