Lab 1 Practice

Risa Gelles-Watnick

2025-09-05

Part 1: Set up your work environment

1. Create a folder for this class.
2. Create a project for this class in that folder.
3. Put this RMarkdown file into that folder.

Part 2: Equations practice

Type the equation for the mean in LaTeX code so that it prints on a new line.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Figure 1: Equation for Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Part 3: Knitting

1. Try to knit this document as is.
2. Now create a new `R` chunk and edit the chunk options so that it is included in the knitted PDF output. Try knitting again to make sure it worked.

```
x <- "hello"
```

Part 4: Write a function

Write a function that tells you whether an input is between 6 and 36. (hint - the `ifelse()` command might be helpful here – look it up in the console to see how it works)

```
between <- function(number) {
  ifelse(number > 6 & number < 36, TRUE, FALSE)
}
```

Uncomment these lines to test your function

```
between(4)
```

```
## [1] FALSE
```

```
between(30)
```

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

Part 5: Practicing tidyverse commands

1. Load in the `tidyverse` package.

```
pacman::p_load(tidyverse)
```

2. Load in a dataset of LGBTQ movies. This data is from TidyTuesdays, which is a cool source to poke around if you're looking for interesting (and already cleaned – huge plus) datasets!

```
# loading in data
lgbtq_movies <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data
```

```
## Rows: 7165 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): title, original_title, original_language, overview, genre_ids
## dbl  (4): id, popularity, vote_average, vote_count
## lgl  (2): adult, video
## date (1): release_date
##
## 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.
```

3. How many rows and columns are in this dataset?

```
dim(lgbtq_movies)
```

```
## [1] 7165   12
```

4. Create a dataset of the top 10 most popular (`popularity`) Spanish language (`es`) LGBTQ movies.

```
pop_es <- lgbtq_movies %>%
  filter(original_language == "es") %>%
  arrange(popularity %>% desc()) %>%
  slice(1:10)
```

5. Create a variable that contains the average rating for a movie only if it has more than 100 votes (the ratings for each movie can be found in the `vote_average` column). Otherwise this variable should be `NA`.

```
lgbtq_movies_rate <- lgbtq_movies %>%
  mutate(
    vote_average_100 = ifelse(vote_count > 100, vote_average, NA)
  )
```

Part 6: Summary statistics

1. Which movies have higher ratings on average – those released in the 20th century or those released in the 21st century?

```
# movies released in the 20th century
lgbtq_movies %>%
```

```
  filter(release_date >= "1900-01-01" &
           release_date < "2000-01-01") %>%
  pull(vote_average) %>% # this extracts the column you want as a vector rather than a data frame colum
  mean(., na.rm = TRUE)
```

```
## [1] 3.986006
```

```
# movies released in the 21st century
lgbtq_movies %>%
  filter(release_date >= "2000-01-01") %>%
  pull(vote_average) %>%
  mean(., na.rm = TRUE)
```

```
## [1] 3.263054
```

20th century movies have higher ratings on average.

2. Create a dataset of summary statistics by language that contains:

   (i) The earliest release date of an LGBTQ movie in that language

   (ii) The latest release date of an LGBTQ movie in that language

   (iii) The average rating of LGBTQ movies in that language

```
# summary statistics table
sum_stats <- lgbtq_movies %>%
  group_by(original_language) %>%
  summarize(
    earliest = min(release_date, na.rm = TRUE),
    latest = max(release_date, na.rm = TRUE),
    avg_rating = mean(vote_average, na.rm = TRUE)
  )
```

```
## Warning: There were 2 warnings in `summarize()`.
## The first warning was:
## i In argument: `earliest = min(release_date, na.rm = TRUE)`.
## i In group 2: `original_language = "am"`.
## Caused by warning in `min.default()`:
## ! no non-missing arguments to min; returning Inf
## i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

What language was the earliest LGBTQ film made in?

English (according to this dataset)