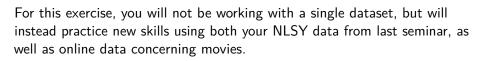
Exercise 3

Exercise 3 1 / 11



Exercise 3 2 / 11

1. Revisit your NLSY97 dataset from last week

- Oreate an indicator for sex using a vectorized conditional statement.
- Recode the schooltype variable into text values, corresponding to:
 - "Public" if the value is 1
 - "Private, religious" if the value is 2
 - "Private, non-religious" if the value is 3
 - "Other" if the value is 4.

nlsy97 <-import("nlsy97.rds")</pre>

Exercise 3 3 / 11

2. Load the IMDB Top 250 Movies

Exercise 3 4 / 11

2a: Scrape the data from the "Top 250 Movies as rated by IMDb users

From https://www.imdb.com/chart/top:

```
# IMDB Top 250 Movies
top250_basic <-
  read_html("https://www.imdb.com/chart/top/") %>%
  html_table() %>% as.data.frame()
```

Exercise 3 5 / 11

2b. Notice that IMDB scrapes the data in Swedish by default.

To get the data in English, use html_session() in place of read_html(), adding the option:

```
add_headers("Accept-Language"="en-US, en;q=0.5")
```

You may need to load the httr package to use add_headers().

```
top250_eng.pre <-
   html_session("https://www.imdb.com/chart/top/",
   add_headers("Accept-Language"="en-US, en;q=0.5")) %>%
   html_table %>% as.data.frame()
```

Exercise 3 6 / 11

Steps 2c-d

2c: Keep only the columns "Rank... Title" and "IMDb.Rating", suitably renaming them.

2d. Create a ranking variable by extracting the values that appear before the dot in the title column.

```
top250_eng$Ranking <- top250_eng$Title %>%
str_extract("[0-9]+(?=(.\n))")
```

Exercise 3 7 / 11

2e-2g

2e. Create a year variable, by extracting the numbers inside a parenthesis from the title column.

2f. Redefine the title variable by extracting the string information that appear after the dot in the title column.

```
top250_eng$Title <- top250_eng$Title %>%
str_extract("(?<=(.\n)).+")</pre>
```

2g. Trim the white space on both sides of the title.

```
top250_eng$Title %<>% str_trim(side = "both")
```

Exercise 3 8 / 11

Important Tables and Figures

Exercise 3 9 / 11

View Top 250 dataframe

head(top250_eng)

Invisible table!

Exercise 3 10 / 11

Elite R Programmer

Keep working and one day soon, this can be you!



Exercise 3 11 / 11