

Exercise 3

For this exercise, you will not be working with a single dataset, but will instead practice new skills using both your NLSY data from last seminar, as well as online data concerning movies.

1. Revisit your NLSY97 dataset from last week

- 1 Create an indicator for sex using a vectorized conditional statement.
- 2 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")
```

2. Load the IMDB Top 250 Movies

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()
```

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()
```

Steps 2c-d

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

```
top250_eng <- top250_eng.pre %>%
  select("Rank...Title", "IMDb.Rating") %>%
  rename(Title = Rank...Title,
         Rating = IMDb.Rating) # Keep only Title and Year Columns
```

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))")
```

2e-2g

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

```
top250_eng$Year <- str_extract(top250_eng$Title,
                              "(?<=[:punct:])[[:digit:]]+")
```

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)).+")
```

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

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


Important Tables and Figures

View Top 250 dataframe

```
head(top250_eng)
```

Invisible table!

Elite R Programmer

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

