# STAT 413/613 Homework on Web Data: APIs and Scraping

Richard Ressler

2020-09-29

## Instructions

- Write your solutions **in this starter file**.
  - Modify the "author" field in the YAML header.
- Commit R Markdown and HTML files (no PDF files). **Push both .Rmd and HTML files to GitHub**.
  - Make sure you have knitted to HTML for your final submission.
- **Only include necessary code and data** to answer the questions.
- Most of the functions you use should be from the tidyverse. **Too much base R** will result in point deductions.
- Submit a response on Canvas that your assignment is complete on GitHub
- Feel free to use Pull requests and or email (attach your .Rmd) to ask me any questions.

### Learning Outcomes:

- Collect and tidy data from web sites using APIs or web scraping techniques.
- Apply concepts and methods from STAT 412/612.

## Scoring Rubric

| Question.Part: | Points | Topic |
|---|---|---|
| 1.API Connect | 1.00 | Collect Data |
| 1.API Tidy | 3.00 | Tidy Data |
| 1.Plot | 2.00 | Appropriate Plot with proper labels |
| 1.Plot | 1.00 | Appropriate Question and Interpretation of the plot |
| 2.Scrape HTML | 1.00 | Scrape HTML Data with no extra tags |
| 2.Convert Nodes to Text | 1.00 | Convert Nodes to Text |
| 2.Clean | 3.00 | Clean using stringr functions |
| 2.Missing Elements | 1.00 | Correct two elements with the fewest entries |
| 2.Tidy | 2.00 | Tidy the data frame using dplyr functions |
| 2.Length vs gross-a | 2.00 | Appropriate Plot |
| 2.Length vs gross-b | 1.00 | Proper Interpretation |
| 2.Stars versus Metacritic-a | 2.00 | Appropriate Plot |
| 2.Stars versus Metacritic-b | 1.00 | Proper Interpretation |
| 2.Gross vs Rating -a | 2.00 | Appropriate Plot |
| 2.Gross vs Rating -b | 1.00 | Correct Highest median gross receipts |
| 2.Gross vs Rating -c | 1.00 | Correct two R-Rated movies in the top 10 of Gross Receipts |
| 2.Analysis of Variance of Gross versus Rating | 1.00 | Correct results and interpretation |

| Question.Part: | Points | Topic |
|---|---|---|
| 3.Extra Credit: Podcast Thoughts | 1.00 | Cogent Responses |
| Total | 25 | plus 2 possible extra credit |

# 1 Using APIs

- Pick a website of your choice and use an API to download a data set. Convert elements of interest into a tibble and create a graph to answer a question of interest.
- State the question and interpret the plot

# 2 IMDB List of Oscar Winners

IMDB has a list of the Oscar Best Picture Winners.

Scrape the following elements, convert the data into a tibble, tidy it, and clean it to answer the questions below: - Number - Title - Year - MPAA Rating - Length in minutes - Genre - Star Rating - Metascore Rating - Gross Receipts

Convert the data into a tibble, tidy it, and clean it to answer the following questions:

1. Which two elements are missing the most from the movies?
2. Create a plot of the length of a film and its gross, color coded by rating.

- Does MPAA rating matter?

3. Create a plot with a single Ordinary Least Squares smoothing line with no standard errors showing for predicting stars rating based on metacritic scores.

- Is there a meaningful relationship in terms of the $p$-value and adjusted R-Squared?

4. Use an appropriate plot to compare the gross receipts by MPAA rating.

- Which MPAA rating has the highest median gross receipts?
- Which are the R-rated movies in the overall top 10 of gross receipts?
- Extra Credit (1 pts): Use one-way analysis of variance to assess the level of evidence for whether all ratings have the same mean gross receipts. Provide your interpretation of the results.

# 3 Extra Credit 1 Pts

- Listen to the AI Today podcast on Machine Learning Ops and provide your thoughts on the following questions:

1. Does knowing about Git and GitHub help you in understanding the podcast?
2. How do you think the ideas of ML OPs will affect your future data science projects?

You may also want to check out this article on Towards Data Science