Hieu Tran & Ava Oberle
CS 181

Mid-Semester Project Proposal

Central Question: The average duration and ratio between movies and TV shows of popular streaming services.

Data Sets from Kaggle:

- Each of these datasets cover one streaming service so we can get input from each one:
    - Disney+ (1451 rows) https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows
    - Netflix (8808 rows) https://www.kaggle.com/datasets/shivamb/netflix-shows
    - Amazon Prime (9669 rows) https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows
    - Hulu (3073 rows) https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows
- Each dataset contains a list of movies and TV shows of each platform, updated 2021, with 12 columns ("show_id", "type", "title", "director", "cast", "country", "date_added", "release_year", "rating", "duration", "listed_in", "description"). For the purpose of this project, we will only take in 4 columns: "show_id", "type" (movie or TV show), "title", and "duration" (in minutes or in seasons).

Outline:

1. **Parse the datasets into DataFrames**

- pandas.read_csv() to read and parse the dataset

2. **Operate on the DataFrames**

- pandas.DataFrame.groupby() to split each dataset based on the “type” column (“movie” or “TV Show”)

- pandas.DataFrame.replace() to get rid of the " min" or " seasons" in the "duration" column
- int() to convert the duration to integers

3. Get the relevant data

- pandas.DataFrame.shape[0] to get the number of rows (count of each type) in each
- Divide the count of each type to get the ratio of movies and TV shows in each
- pandas.DataFrame.describe() to get numerical statistics of the duration (mean, min, max, percentiles,...) for comparison between each provider
- sum() to sum the number of movies and TV shows in all services so that we can average the duration of each and get the average duration of a movie/TV show as well as the overall ratio of movies and TV shows across all platforms