# Netflix Analysis

### 1. Identify the English TV show with the most appearances in the top 10 list (you can treat each row in the data as a separate appearance). What were the average weekly viewed hours for that show across all appearances?

| Title | Average Weekly Hours Viewed |
|---|---|
| **YOU** | 43193333.33 |

### 2. For the "Films (Non-English)" category, identify the film with lowest IMDb rating. What were the average weekly hours viewed for that film?

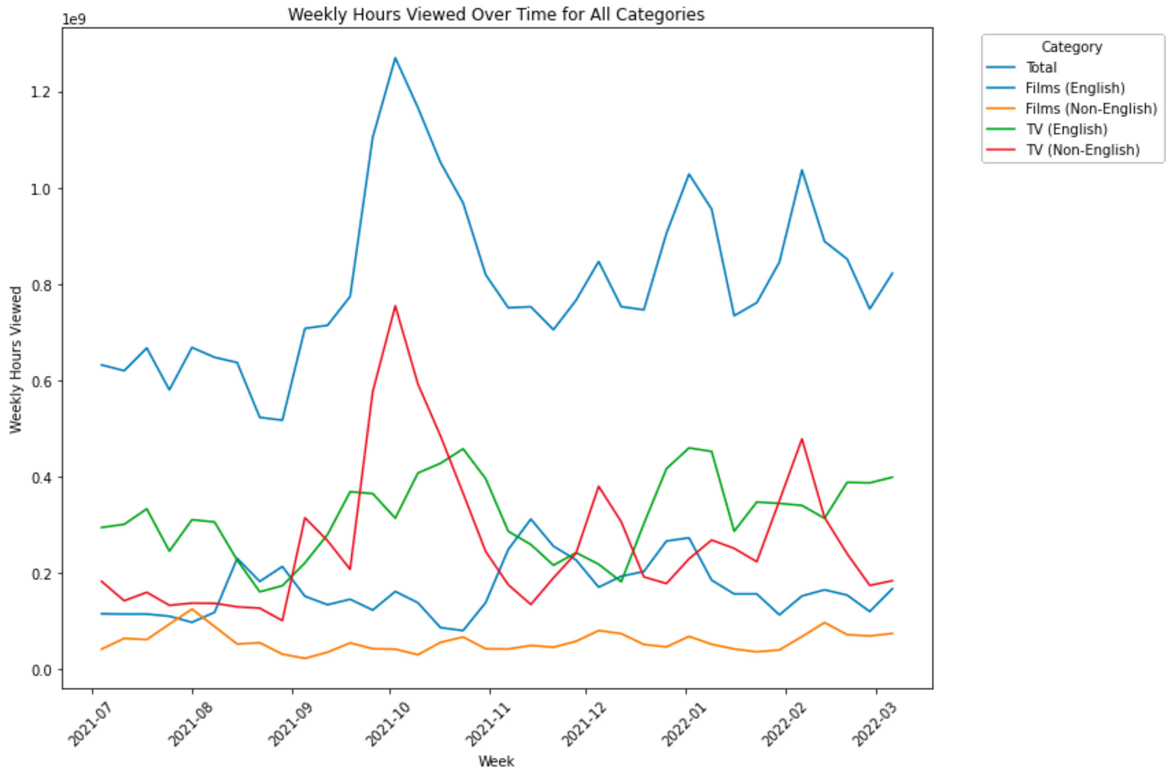
| Title | Average Weekly Hours Viewed |
|---|---|
| **Nobody Sleeps in the Woods Tonight 2** | 4610000.00 |

### 3. Identify the film in the “Films (English)” category with the most cumulative weeks in the top 10. How could you approximate how many users watched this film? What assumptions would you make? What risks are there to your approach?

"Films(English)" with the most cumulative weeks is **Red Notice**. We can estimate an approximate number of users by dividing the sum of weekly viewing hours by the runtime.
Assumptions are as below:
1. Each user completes watching the entire movie.
2. Each user watches the movie only once and does not repeat playbacks.
3. The 'weekly_hours_viewed' data is unaffected by actions like fast-forwarding, rewinding, or changing playing speed, it records the net playback time of the film.

Potential risks could be:
1. Ignoring users who only watched part of the film, this will cause the estimated result lower than the actual result.
2. Not counting multiple viewings by the same user/account, this makes the estimated result higher than the reality.
3. Not considering the impact of different playback behaviors.
4. Some of the users may share a screen to watch the film, we cannot accurately identify.

**4. Plot weekly hours viewed over time (as an aggregate and for each of the four categories), and describe *__the 2 most noticeable__* trends you see and explain the underlying drivers.**



The two most noticeable trends observed are as follows:

1.Both English and non-English TV shows have consistently higher total weekly viewing hours compared to movies for most of the time.
Reasons:
1) TV shows typically have longer total durations compared to movies.
2) People tend to allocate fragmented time for TV shows rather than movies.
3) Some users may prefer going to cinemas to watch movies, diverting the viewing time spent for movies on Netflix.

2. Non-English movies consistently exhibit the least amount of viewing time, coupled with the least variability. While non-English TV shows displays high volatility with significant raises and declines.
Reasons:
1) User appeal: Non-English movies might have a relatively weaker appeal to broader audiences due to cultural differences, understanding barriers and quality compared to TV shows.
2) Breakout hits: It is harder to have a hit Non-English movie than TV show, which attracts millions of users.

**5. Another key investor question is how many US subscribers Netflix has each quarter. Name *one type of alternative dataset* you could use to answer this question. How would this data source help you estimate Netflix's US subscribers?**

We could obtain a dataset, which records how many viewers (in total) watched each TV program and their geographical distribution.

We can use the previous method (by quarter) to roughly estimate how many people worldwide on Netflix have watched a specific TV program (for example, Program A).

Initially, we assume that the proportion of viewers watching the program is equal. That is, the proportion of all viewers watching Program A is consistent with the proportion of viewers watching it on Netflix. Consequently, by knowing the viewership of Program A on Netflix, we can estimate the total number of subscribers on Netflix.

Then, assuming that there is no difference in the geographical distribution, we can calculate the number of subscribers in the United States, based on the proportion of US viewers.

To minimize prediction errors, we can select multiple popular programs and perform a weighted average calculation on the results obtained.

**6. List three reasons why the web-scraping methodology may be inaccurate.**

1. For the Top10 dataset, we do not have data like the complete rating of a program, what is the replay rate, these could affect our estimate result of the number of users.
2. For IMDb Rating dataset, some of the ratings could be out of date, some programs' rating could be quite different according to different era, which we may need to consider.
3. For Runtime dataset, as we lack the respective number of episodes of TV shows, we cannot obtain the total duration of the series. When evaluating the quality or attractiveness of a piece of content based on user viewing time, comparing movies and TV shows with different total durations may require a modification of the evaluation criteria.