**Data Support Specialist**

**Netflix Pre-Screen Exercise**

*Please be sure to read the instructions carefully.*

YipitData is the on-demand data team for the largest institutional investors in the world. We identify, screen, license, clean, and analyze alternative datasets to help investors answer their key questions with actionable insights.

You will be analyzing web-scraped NFLX viewership data in Part 1 and analyzing insight of a NFLX chart in Part 2.

**Background**

Netflix, Inc. is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a library of films and television series through licensing deals and in-house production.

Each week, Netflix [publishes](https://top10.netflix.com/films) its top 10 titles, along with how many hours users spent watching each one. These are split up into four categories: Films (English), Films (Non-English), TV (English), and TV (Non-English). Investors have a number of key questions about Netflix that this data can help address. For example: Is Netflix producing and licensing engaging content? Are Netflix's content investments in new genres or geographies generating significant viewership? How is viewership trending over time, and what implications does this have for Netflix's subscriber numbers?

To answer these questions, we scrape the Netflix top 10 website every week. We run our systems on a set schedule, assuming that Netflix will continue to publish the list each week. We then scrape [IMDb](https://www.imdb.com/) to get information including a movie or show’s running time and ratings. Once we have this data, we clean and analyze it to provide insights to our clients.

The system we use to collect data from the Netflix website experienced an issue during the week of May 22nd, 2022, specifically affecting the 'weekly\_hours\_viewed' column. All other columns were accurately collected. Due to this, we cannot use this week in our estimates, as the 'weekly\_hours\_viewed' column is critical for tracking viewership metrics. You can assume that this is the only week with incomplete viewership data.

**Goal**

**Your goal is to use the attached Excel file and chart to answer the below questions.** These questions are representative of the types of questions you will have the opportunity to answer in the Data Support Specialist role. You will be graded on accuracy/attention to detail, communication, and logic.

This assignment should take approximately 30 minutes to complete. Please reach out to Irene Ear ([iear@yipitdata.com](mailto:iear@yipitdata.com)) with any questions!

Please insert your answers into the **app**. Please perform your calculations entirely in Excel/Google Sheets, Python, SQL, or R. You do not need to learn any new programming languages or packages to complete this assignment. You should be able to complete the entire analysis in Excel/Google Sheets if that is your preference.

**Questions**

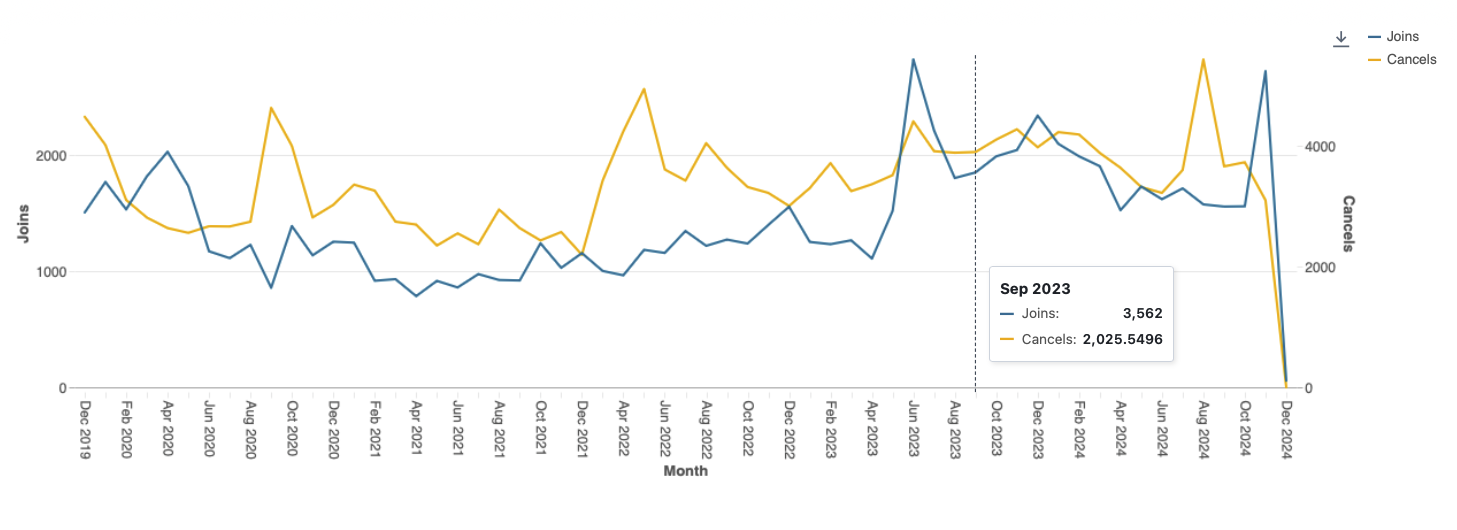
***Note - Be sure to exclude the week of the outage when answering***

**Part 1**

1. In the ‘Films (English)’ category, which film has the most appearances in our data set (NFLX Top 10 tab of the Sheet)? What were its average weekly hours viewed?
2. In the 'Films (English)' category, which film has the lowest IMDb rating? What were its average weekly hours viewed?

**Part 2**

1. The chart below (also attached in the resources) shows the amount of users who joined or cancelled from a subsegment of Netflix users. **Do you notice any potential errors on the chart? Considering any potential errors, is this chart a positive or negative reflection for Netflix, why?**
   * Please limit your response to 250 words or less.
2. Please record a 1-3 min video summarizing your key takeaways of the information presented in this chart. Please attach the file to your submission.

**Part 2.1 Chart**

**Dataset Description and Data Dictionary**

The attached Excel workbook contains the 2 tabs below.

**NFLX Top 10 -** *list of weekly Netflix rankings for each category*

* Columns:
  + date\_added - The date YipitData scraped the information from the Netflix website
  + week - The week of the ranking. The date represents the last day of the week
  + category - Category classification for each title. The available categories are: Films (English), Films (Non-English), TV (English), and TV (Non-English)
  + show\_title - The title of the show or film
  + season\_title - The season of the show (if applicable)
  + weekly\_rank - The rank of the title for the associated week, split by category (each category has a top 10)
  + cumulative\_weeks\_in\_top\_10 - The number of total weeks (not necessarily consecutive) a title has spent in the top 10
  + weekly\_hours\_viewed - The total number of hours Netflix users spent watching the title in the week

**IMDb Ratings** *- IMDb rating (if applicable) for each title*

* Columns:
  + title - The title of the show or film
  + rating - The rating of the show or film scraped from IMDb. This value can range from 1 - 10
    - If a title’s rating is unavailable on IMDb, we input it as 0 and exclude it from any ratings analyses
* Notes:
  + This list contains each title in the NFLX Top 10 sheet and additional titles that are not in the NFLX Top 10 sheet. You are expected to match the titles across the datasets and join in the rating to answer the questions.
  + You do not need to worry about ratings over time. You can assume the ratings remain constant.