**Data Support Specialist**

**Netflix 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.

This assignment builds upon your first pre-screen assignment and is split into two parts. In Part 1, you will be analyzing web-scraped NFLX viewership data. In Part 2, you will read a brief report on our Streaming Video product (which covers a variety of services, including NFLX) and explain your interpretation of our findings in the live technical review. The purpose of this assignment is to gauge your ability to understand and analyze a new dataset and express your findings clearly and concisely.

**Reminder**

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, chart, and PDF 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, communication, and logic.

This assignment should take approximately 45 minutes to complete.

Please insert your answers into the **attached template**, then save your file as a **PDF** before submitting it via the link shared in our previous email (where we asked you when you want to start the exercise). 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. When submitting your work, please submit both the template and your working Excel/Google Sheets or coding file.

**Questions**

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

**Part 1**

1. In the 'Films (Non-English)' category, which film has spent the most weeks in the top 10? To estimate the number of users who watched this film, what assumptions would you make and what risks are involved?
   * Please limit your response to 150 words or less.
2. What are the risks of ignoring the data from the week of May 22nd when calculating the metrics from the previous questions? (first question above and prior pre-screen questions)
   * Please limit your response to 150 words or less.

**Part 2**

1. Please review the attached report and be prepared to answer live questions regarding its insights into NFLX's performance as a company.

**Dataset Description and Data Dictionary**

The attached Excel workbook contains the 3 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.

**Runtime** *- running time, in minutes, (if applicable) for each title*

* Columns:
  + title - The title of the show or film
  + runtime - The running time, in minutes, (if applicable) for each title
    - If a title’s runtime is unavailable on IMDb, we input it as 0 and exclude it from any runtime 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 runtime to answer the questions.
  + Some television shows do not have an accurate (complete) runtime, but you will not need to perform any analyses on these titles