yipitDATA

Data 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 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. The purpose of this assignment is to gauge your ability to understand and analyze a new dataset and express your findings clearly and concisely.

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 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 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.

yipitDATA

Goal

Your goal is to use the attached Excel file 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 Specialist role. You will be graded on accuracy, communication, and logic.

This assignment should take approximately 1-2 hours to complete. Please reach out to Irene Ear (iear@yipitdata.com) with any questions!

Questions

Note - Be sure to exclude the week of the outage when answering questions 1, 2, and 3

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?
3. 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.
4. What are the risks of ignoring the data from the week of May 22nd when calculating the metrics from the previous questions?
    - Please limit your response to 150 words or less.
5. While we've indicated that the 'weekly_hours_viewed' data for the week of May 22nd cannot be used in our estimates, we may want to fill in this missing information for other analyses. As a Data Specialist, what methodology would you propose to estimate the 'weekly_hours_viewed' for this missing week?
    - Please limit your response to 150 words or less.

Part 2

6. Do you think the information presented in this report is positive or negative for NFLX as a company? Why?
    - Please limit your response to 300 words or less.
7. List 3 questions our clients may have after reading the report.
8. Please clarify the level of proficiency in any coding skills that you have. (Note: coding skills are not a prerequisite for the position. If you do not possess any coding skills, please write "NA")

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, 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.

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