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. New datasets are being created every day, and our clients need to incorporate them to remain competitive.

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.

Goal

Your goal is to use the attached Excel file 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 Michael (mwhite@yipitdata.com) with any questions!

yipit DATA

Questions

- Identify the TV show (English) 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?
- For the "Films (Non-English)" category, identify the film with lowest IMDb rating. What were the average weekly hours viewed for that film?
- 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 show? What assumptions would you make? What risks are there to your approach?
    - Please limit your response to 150 words or less.
- If you plot weekly hours viewed over time (as an aggregate and for each of the four categories), what trends do you notice?
    - Please limit your response to 150 words or less.
- Another key investor question is how many US subscribers Netflix has each quarter. Name one type of dataset you could use to answer this question. How would this data source help you estimate Netflix's US subscribers?
    - Please limit your response to 150 words or less.
- List three reasons why our web scraping methodology may be inaccurate.
- What is your undergraduate or graduate GPA, both overall and for your major?
    - If you do not have easy access to this, that is acceptable, but please note that in your response to this question!
- What are the scores for your undergraduate standardized tests (eg. SAT) and graduate standardized tests (eg. GRE)? Please break out the scores into individual sections (eg. Math, Reading, etc.).
    - If you did not take an undergraduate or graduate standardized test (or do not have easy access to your results), that is acceptable, but please note that in your response to this question!
- 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 submit your work as a PDF according to this template, which is based on the criteria above. When submitting your work, please submit **both** the template and your working Excel or coding file.

- https://docs.google.com/document/d/13Z3yFJyV_jOZmZqfykHPiVetzTInA4F_jJqCqZmkDYM/edit ← Direct link to the template

yipitDATA

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:
  - category - Category classification for each title. The available categories are: Films (English), Films (Non-English), TV (English), and TV (Non-English)
  - 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
  - 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)
  - show_title - The title of the show or film
  - 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.

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.
- 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
- 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