NETFLIX TOP 10 DATA ANALYSIS

Introduction:
Netflix Top 10 is a list top 10 movies or TV shows of the week, listed for each category by Netflix. This is valuable data as it gives insights on user preferences, trend, and revenue direction. The goal of this analysis is to conduct exploratory data analysis, and answer a few questions to get valuable insights.

Data Overview:
There are 3 datasets utilized in this analysis : NFLX Top 10, IMDb Ratings, and Runtime. The dataset ranges from 7/4/2021 to 3/6/2022 (35 weeks).

1. *NFLX Top 10*
   This is our primary dataset containing list of weekly Netflix rankings for each category.
   - category - Category classification for each title. The 4 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
   - week - The week of the ranking.

2. *IMDb Ratings*
   IMDb rating for each title.This list contains each title in the NFLX Top 10 sheet and additional titles that are not in the NFLX Top 10 too.
   - title - The title of the show or film
   - rating - The rating of the show or film. This value can range from 1 - 10.

3. *Runtime*
   running time, in minutes,for each title. This list contains each title in the NFLX Top 10 sheet and additional titles that are not in the NFLX Top 10 too.
   - title - The title of the show or film
   - runtime - The running time, in minutes for each title

(Working can be found in this Google Sheet : ☒ Netflix Top 10 Data Analysis.xlsx)

Data Cleaning and Integration:

Cleaning
In the IMDb Ratings and Runtime tables, some titles have multiple values of ratings or runtime values. As there is uncertainty about what the true values could be, to capture the best information - I averaged the multiple values, using Pivot Tables. This way, data is not erroneous or ambiguous. These averages can be found in the sheets : 'Data Cleaning : IMDb Ratings Average' and 'Data Cleaning : Runtime Average'.

Integration:
Combined the three datasets(in the sheets : 'NFLX Top 10' , Data Cleaning : IMDb Ratings Average' and 'Data Cleaning : Runtime Average') into one sheet '*Data Manipulation : Combined Data*' by mapping titles and utilizing VLookup.
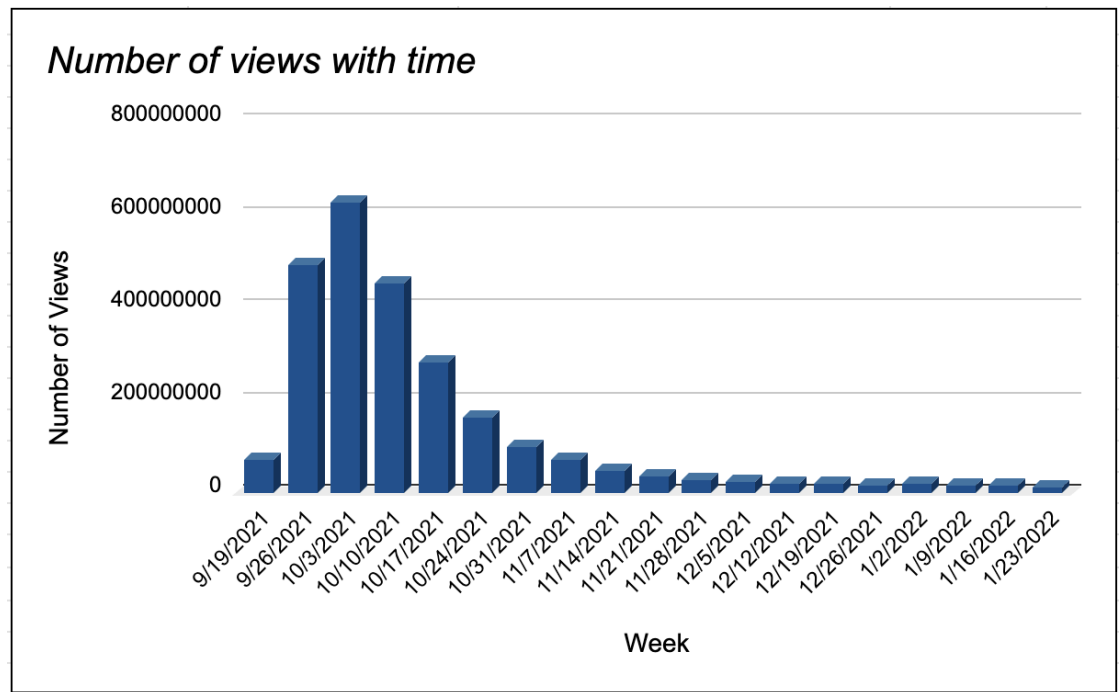
Analysis and Results:

1. What are the total number of ***unique*** titles? In each category as well.

    A. There are a total of 496 unique titles that appeared in the top 10 for the 35 weeks.

category	Unique Titles
Films (English)	173
Films (Non-English)	147
TV (English)	90
TV (Non-English)	87
Grand Total	**496**

2. Which TV show or film has the **highest cumulative weeks** in the top 10 across all categories?

    A. TV Show ***Squid Game*** stayed the longest number of weeks continuously for about **19** weeks in the Top 10.

3. What TV show and what film have the **highest average weekly hours** viewed?

    A. The TV show with the highest average weekly hours viewed is ***Inventing Anna***, with over 120 million average hours and the Film with the highest average weekly hours viewed is ***Don't Look Up*** with over 48 million hours.

4. For the TV show with the highest cumulative weeks, what is the trend of number of views like?

   A. The TV show with highest cumulative weeks in Squid Game and it's trend of number of views is as show in the graph below. It is like a bell curve, with the highest number of views happening in the third week, and again gradually decreasing.

Number of views with time

5. Is there a correlation between IMDb ratings and total number of views for titles in the NFLX top 10?

   Yes, there is a correlation between IMDb ratings and the total number of views for titles in the NFLX top 10. The calculated linear regression is as below :
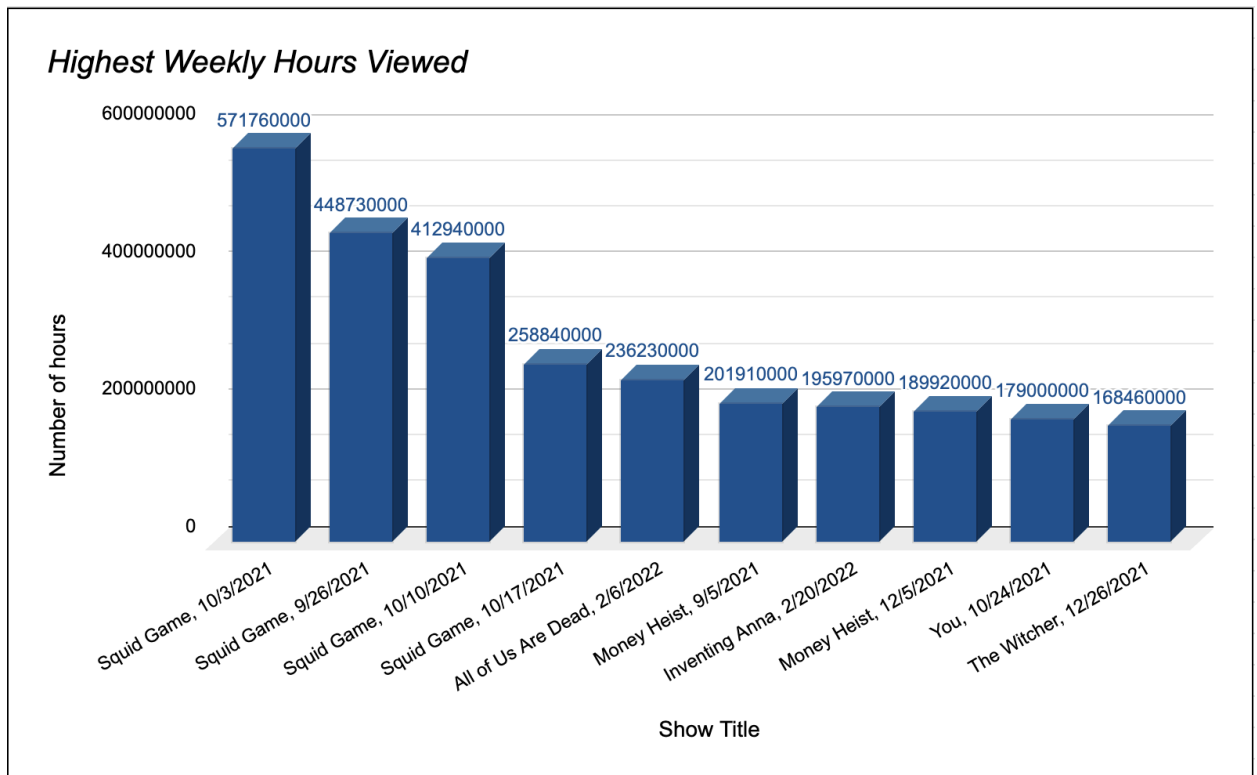
   *Y = 34670153.77 * X + -165986793.3*

   This equation suggests a positive relationship, indicating that as IMDb ratings increase, the total number of views tends to increase, with each unit increase in IMDb rating associated with an estimated increase of 34,670,153.77 weekly hours viewed.

6. What is the **average weekly hours** viewed for each category (Films (English), Films (Non-English), TV (English), TV (Non-English))?

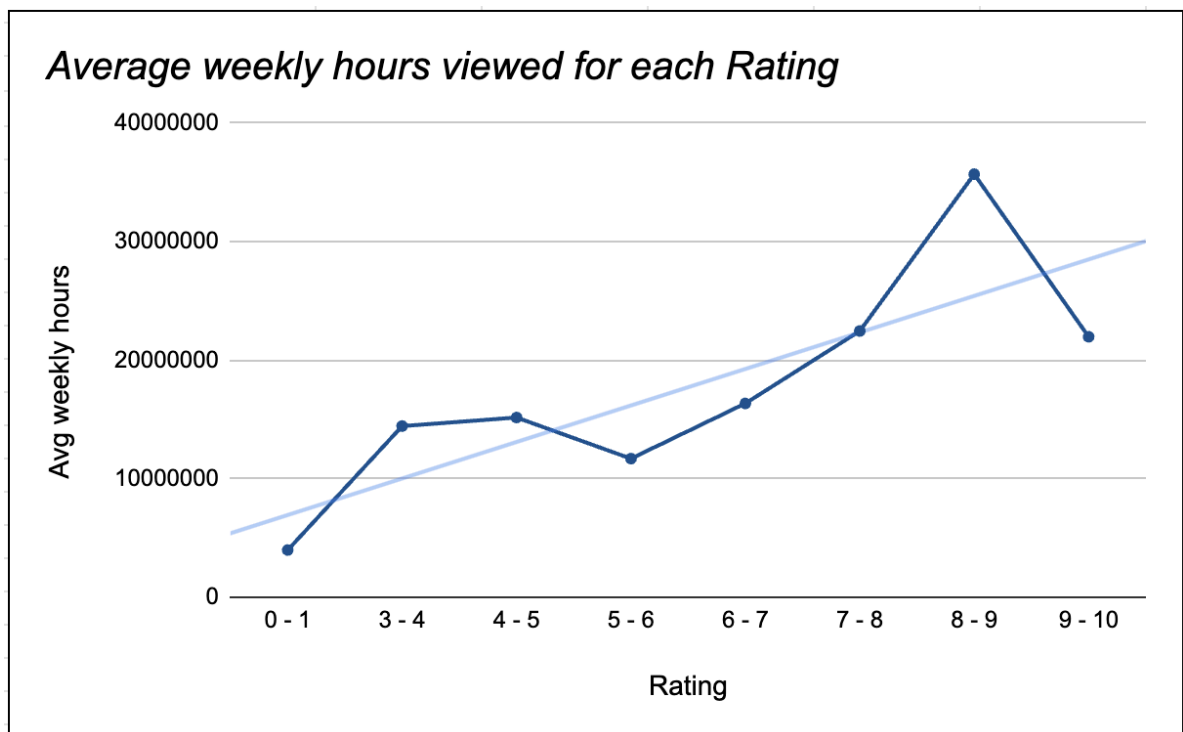| category	Average of weekly_hours_viewed
Films (English)	16534111.11
Films (Non-English)	5693777.778
TV (English)	31692083.33
TV (Non-English)	26521611.11
Grand Total	**20110395.83**

7. Which TV show or film had the **highest weekly hours** viewed overall and in which week?

   A. **_Squid Game_** Season 1 has the highest weekly hours viewed with over 570 million hours in week 10/3/2021.

Highest Weekly Hours Viewed

Number of hours

Show Title	Value
Squid Game, 10/3/2021	571760000
Squid Game, 9/26/2021	448730000
Squid Game, 10/10/2021	412940000
Squid Game, 10/17/2021	258840000
All of Us Are Dead, 2/6/2022	236230000
Money Heist, 9/5/2021	201910000
Inventing Anna, 2/20/2022	195970000
Money Heist, 12/5/2021	189920000
You, 10/24/2021	179000000
The Witcher, 12/26/2021	168460000

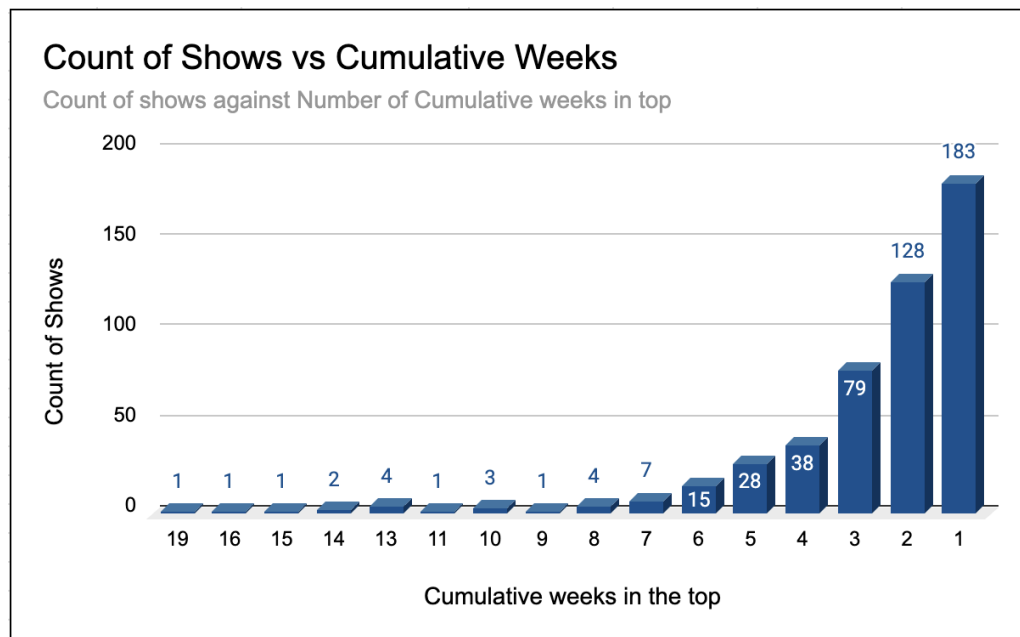The top 10 are all shows, with an obvious reason that shows have higher running time.

8. Identify the English TV show with the **most appearances** in the top 10 list .

    A. The English TV show which appeared the most number of times in the top 10 is *You*, with a total of 18 appearances in the top 10 in 35 weeks.

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

    A. The Film *Nobody Sleeps in the Woods Tonight 2* has the lowest IMBb rating of 3.9 in Films (Non-English) category with over 4.6 million

10. What is the trend of average weekly hours viewed for titles against different IMDb ratings like?

    A. Overall the average hours viewed is increasing as the IMDb rating of the show increases, with an exception of 5-6 and 9-10.

Average weekly hours viewed for each Rating

(Line chart. X-axis: Rating, with categories 0 - 1, 3 - 4, 4 - 5, 5 - 6, 6 - 7, 7 - 8, 8 - 9, 9 - 10. Y-axis: Avg weekly hours, from 0 to 40000000. The data line rises from ~4000000 at 0-1, up to ~14000000 at 3-4, ~15000000 at 4-5, dips to ~11500000 at 5-6, rises to ~16000000 at 6-7, ~22500000 at 7-8, peaks at ~35500000 at 8-9, and drops to ~22000000 at 9-10. A trend line rises steadily from ~5000000 to ~30000000.)

There are very few movies in the NFLX Top 10 with ratings in the range of 9 - 10.

11. Are there any **outliers** in terms of weeks in the top 10 compared to their IMDb ratings?

    A. There are very few shows that stayed consistently for more than 10 weeks in the NFLX top 10.

Count of Shows vs Cumulative Weeks
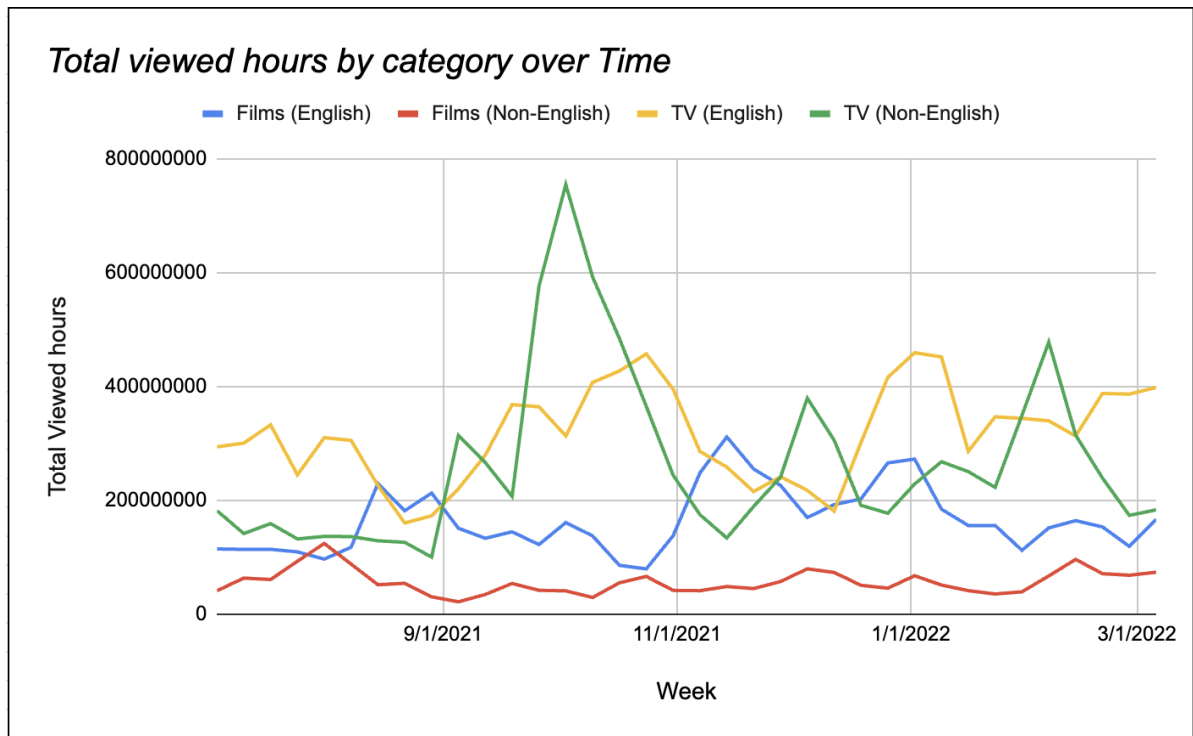Count of shows against Number of Cumulative weeks in top

Of these few, there are **4 shows** with ratings in range of 6-7, which can be considered pretty low. They are Blood Red Sky,Red Notice,Carinha de Anjo and Carrossel.

12. Which TV show and which film has the highest number of views overall?

    A. **Squid Game** and **Red Notice** have the highest number of views with over 2 billion views and 228 millions views respectively.

13. If you plot weekly hours viewed over time (as an aggregate and for each of the four categories), what trends do you notice?

    A. The graph is follows :

Total viewed hours by category over Time

Legend: — Films (English) — Films (Non-English) — TV (English) — TV (Non-English)

Y-axis: Total Viewed hours (0 to 800000000)

X-axis: Week (9/1/2021, 11/1/2021, 1/1/2022, 3/1/2022)

The sum of the weekly hours shows a graph with a high fluctuation in the TV (Non-English) category, before dwelling deep I also observed the average hours graph for the 4 categories in case there could be a big variation in the runtime. However, both the sum and the average of the weekly hours are similar. On exploring further, the trends and observations are as follows :

- The high value of TV(Non-English) in October '21 is due to the TV show titled 'Squid Game Season' 1 which has the highest hours viewed throughout all categories, all-time in this data.
- The next highest peak for the same category is in Feb '22, this time due to 'All of Us Are Dead', which is next to only 'Squid Game' in the highest hours viewed.
- The TV(English) category has the highest in October'21 as well due to the TV show titled 'You' which, as we observed in previous questions, appeared most in the top 10.

We can observe that the trend in these categories pertains to the new content that is being released and has less to do with the holiday season, or any such common seasonal activity

(Working can be found in this Google Sheet : ⊠ Netflix Top 10 Data Analysis.xlsx)