DSC640 Weeks 1 & 2 Exercise

September 8, 2024

- Sauda Haywood
- Weeks 1 & 2 Exercise

0.1 Summary

My Audience: Movie Creators For this assignment, my chosen audience is movie creators. These individuals have a good understanding of movie performance and are focused on developing new shows and films that align with audience preferences. I believe they are interested in knowing what viewers are paying the most attention to, as well as insights from viewership data, including trends, genre performance, and regional audience preferences. Having this data will help them identify opportunities for new movie development based on the insights.

Call to Action I will recommend that movie creators develop new movies and TV shows that align with the high-performing genres identified in the analysis. Specifically, focus on producing more movies similar to top-ranked movies in regions where they have performed exceptionally well. Additionally, consider filling gaps in underrepresented regions by creating culturally relevant content, which will help increase Netflix's global presence.

Medium and Story Approach I will be using PowerPoint as the medium. All graphs will be generated using Jupyter Notebook. Since movie creators are familiar with data in their industry, my approach will incorporate technical language and detailed performance metrics. I will present insights based on hours viewed, genre popularity, and regional preferences, highlighting which types of movies and shows are performing best. This will help creators make informed decisions about future movie and show development.

Ethical Considerations The data used for this assignment is sourced directly from Netflix and is legally obtained. There are no significant legal or regulatory risks in using this data, as it is publicly available and ethically acquired.

0.2 Data Exploration and Visualization

[7]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

[2]:

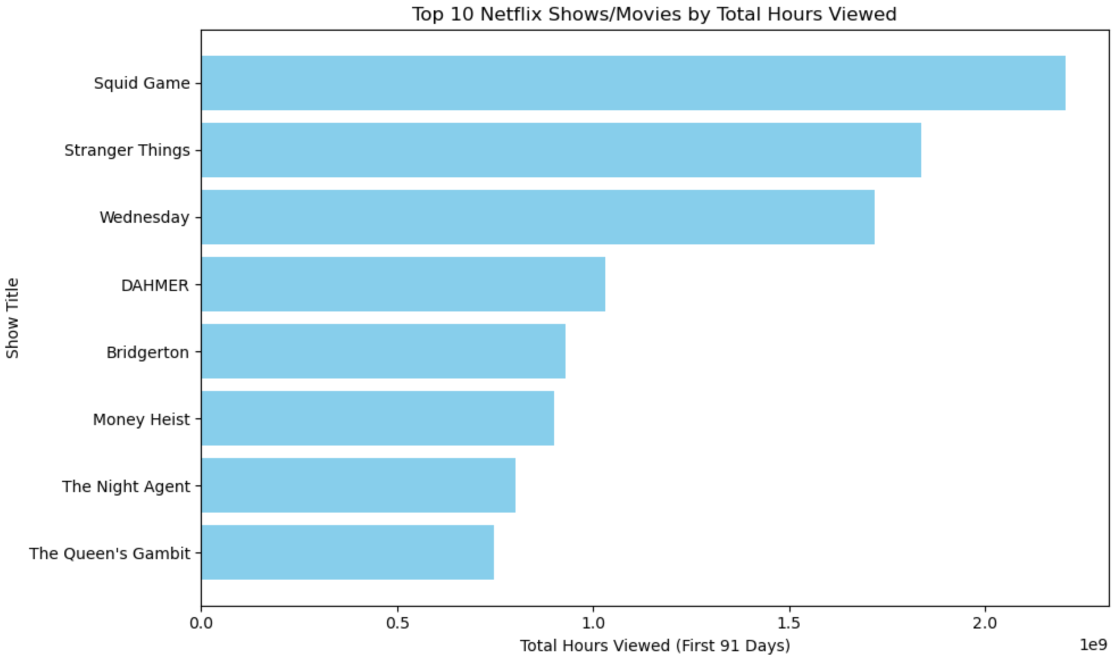
```
# Load the 1ST dataset
most_popular_df = pd.read_excel('most-popular-netflix.xlsx')

# Preview the data
print(most_popular_df.head())
```

```
          category  rank              show_title season_title  \
0  Films (English)     1              Red Notice          NaN
1  Films (English)     2           Don't Look Up          NaN
2  Films (English)     3        The Adam Project          NaN
3  Films (English)     4                Bird Box          NaN
4  Films (English)     5  Leave the World Behind          NaN

   hours_viewed_first_91_days  runtime  views_first_91_days
0                   454200000   1.9667            230900000
1                   408600000   2.3833            171400000
2                   281000000   1.7833            157600000
3                   325300000   2.0667            157400000
4                   339300000   2.3667            143400000
```

- This dataset contains shows and movies that have gained popularity based on hours viewed in the first 91 days. It highlights the top-ranked content globally.

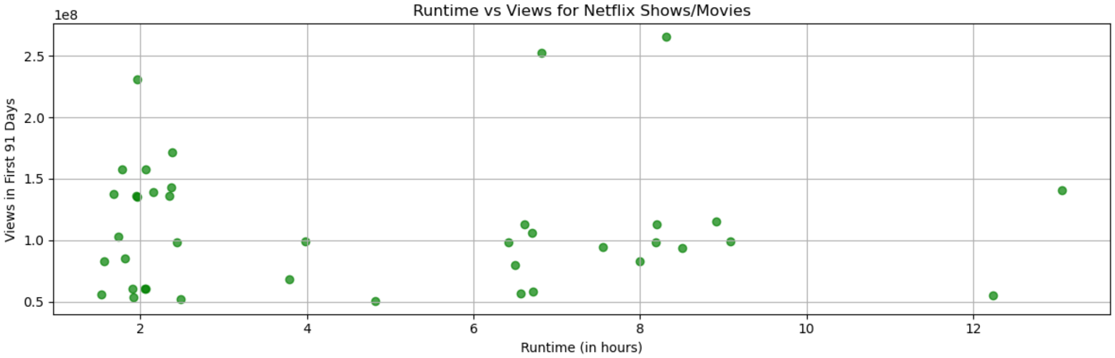
[4]:

```
# Sort the dataset
plt.figure(figsize=(10, 6))
top_10_sorted = most_popular_df.sort_values(by='hours_viewed_first_91_days',␣
 ↪ascending=False).head(10)

# Create a horizontal bar chart
plt.barh(top_10_sorted['show_title'],␣
 ↪top_10_sorted['hours_viewed_first_91_days'], color='skyblue')

# Set the labels and title
plt.xlabel('Total Hours Viewed (First 91 Days)')
plt.ylabel('Show Title')
plt.title('Top 10 Netflix Shows/Movies by Total Hours Viewed')

# Invert the y-axis to have the highest value at the top
plt.gca().invert_yaxis()
plt.tight_layout()

# Display the chart
plt.show()
```



- Most of the top-performing shows are thriller, fantasy , adventure-oriented, or have strong comedic elements. This trend suggests that these genres are the most appealing to a broad audience across regions.

[3]:
```
# Create a scatter plot
plt.figure(figsize=(12, 4))
plt.scatter(most_popular_df['runtime'], most_popular_df['views_first_91_days'],␣
 ↪color='green', alpha=0.7)

# Set the labels and title
plt.xlabel('Runtime (in hours)')
plt.ylabel('Views in First 91 Days')
plt.title('Runtime vs Views for Netflix Shows/Movies')

# Add a grid for better readability
plt.grid(True)

# Adjust layout for better fitting
plt.tight_layout()

# Display the chart
plt.show()
```



- The scatter plot shows a wide spread of points, indicating that there is no clear or strong correlation between runtime and the number of views. Movies and shows with shorter runtimes and longer runtimes alike have both high and low views, suggesting that runtime alone may not be a decisive factor in determining viewership success.

Recommendation: * Rather than limiting content based on runtime, movies and show creators can ensure that the content quality, storylines, and genres align with viewer preferences. * Experiment with Different Lengths as it seems like audiences seem to enjoy both types, so offering variety in runtime can help attract a wider range of viewers with different time commitments and content preferences.

[9]:
```python
# Load the 2nd datasets
df_all_weeks_global = pd.read_excel('all-weeks-global-netflix.xlsx')

# Preview the data
print(df_all_weeks_global.head())
```

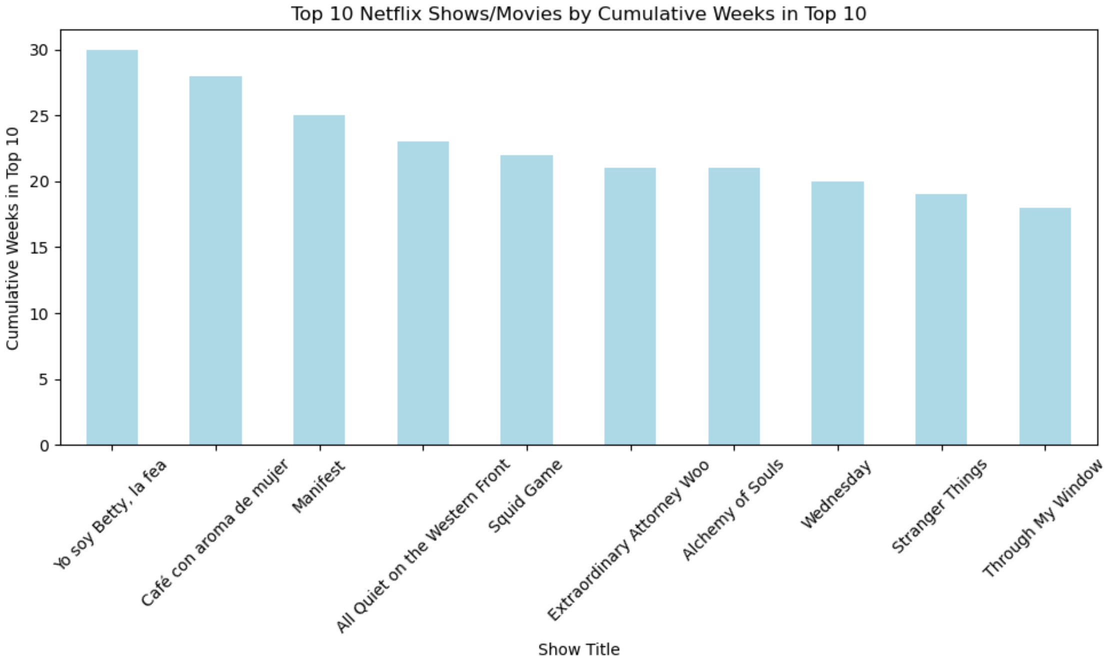
```
         week          category  weekly_rank                     show_title  \
0  2024-04-14  Films (English)            1              What Jennifer Did
1  2024-04-14  Films (English)            2  Woody Woodpecker Goes to Camp
2  2024-04-14  Films (English)            3                          Scoop
3  2024-04-14  Films (English)            4                          Glass
4  2024-04-14  Films (English)            5                   Megan Leavey

  season_title  weekly_hours_viewed  runtime  weekly_views  \
0          NaN             26100000   1.4500    18000000.0
1          NaN             19600000   1.6667    11800000.0
2          NaN             14600000   1.7167     8500000.0
3          NaN             11000000   2.1500     5100000.0
4          NaN              9700000   1.9333     5000000.0

   cumulative_weeks_in_top_10  is_staggered_launch episode_launch_details
0                           1                False                    NaN
1                           1                False                    NaN
2                           2                False                    NaN
```

```
3                             2                False                        NaN
4                             1                False                        NaN
```

- This dataset captures weekly rankings for shows and movies globally, showing the number of hours watched, views, runtime, and how long the a movie or a show has stayed in the Top 10

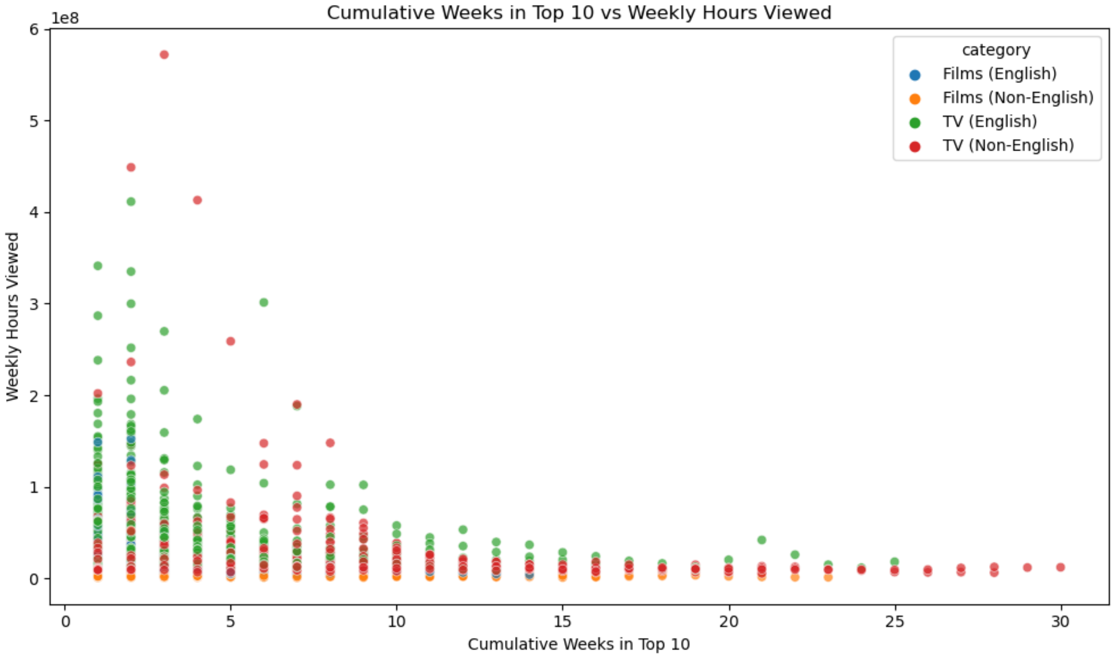
[13]:

```python
# Rank the movies by cumulative weeks in top 10
ranked_movies = df_all_weeks_global.groupby('show_title')['cumulative_weeks_in_top_10'].max().sort_values(ascending=False)

# Prepare the data for the top 10 shows
top_10_ranked_movies = ranked_movies.head(10)

# Plot the data
plt.figure(figsize=(10, 6))
top_10_ranked_movies.plot(kind='bar', color='lightblue')

# Set the labels and title
plt.xlabel('Show Title')
plt.ylabel('Cumulative Weeks in Top 10')
plt.title('Top 10 Netflix Shows/Movies by Cumulative Weeks in Top 10')

# Rotate x-axis labels for readability
plt.xticks(rotation=45)

# Display the plot
plt.tight_layout()
plt.show()
```



Key Insights: * Shows like Yo soy Betty, la fea, Café con aroma de mujer, and Manifest have spent significantly more time in the top 10 compared to others, indicating sustained popularity. * Diverse Genres: The top-ranked shows span different genres and languages, showing that both international content and a variety of genres can perform well over time on Netflix.

[14]:
```
# Visualization: Scatter plot of cumulative weeks in top 10 vs weekly hours␣
↪viewed
plt.figure(figsize=(10, 6))
sns.scatterplot(x='cumulative_weeks_in_top_10', y='weekly_hours_viewed',␣
↪hue='category', data=df_all_weeks_global, alpha=0.7)

# Set the labels and title
plt.xlabel('Cumulative Weeks in Top 10')
plt.ylabel('Weekly Hours Viewed')
plt.title('Cumulative Weeks in Top 10 vs Weekly Hours Viewed')

# Display the chart
plt.tight_layout()
plt.show()
```



- The scatter plot shows that shows/movies with high weekly hours viewed are distributed across various numbers of cumulative weeks in the top 10. Different categories, such as Films (English), Films (Non-English), TV (English), and TV (Non-English), are spread throughout the plot. Non-English TV shows seem to maintain high weekly hours viewed.

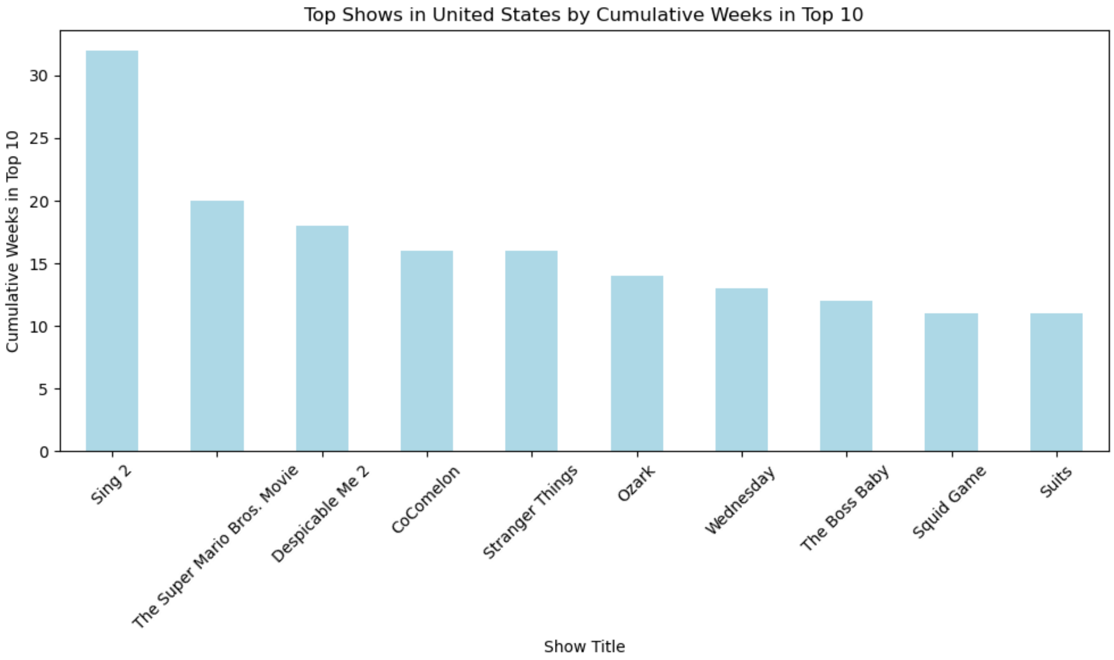
```
[15]: # Load the 3rd datasets
df_all_weeks_countries = pd.read_excel('all-weeks-countries-netflix.xlsx')
# Preview the data
print(df_all_weeks_countries.head())
```

```
  country_name country_iso2        week category  weekly_rank  \
0    Argentina           AR  2024-04-14    Films            1
1    Argentina           AR  2024-04-14    Films            2
2    Argentina           AR  2024-04-14    Films            3
3    Argentina           AR  2024-04-14    Films            4
4    Argentina           AR  2024-04-14    Films            5

                      show_title season_title  cumulative_weeks_in_top_10
0                  The Tearsmith          NaN                           2
1                         Stolen          NaN                           1
2                  Love, Divided          NaN                           1
3   Woody Woodpecker Goes to Camp          NaN                           1
4                  Rest In Peace          NaN                           3
```

- This dataset breaks down the weekly rankings by country, showing which shows are most popular in various regions, along with cumulative weeks spent in the top 10.

[17]:

```
# Filter the data for United States
country_data = df_all_weeks_countries[df_all_weeks_countries['country_name'] ==␣
 ↪'United States']

# Group by show title and calculate the maximum cumulative weeks in top 10 for␣
 ↪each show
top_shows_by_country = country_data.
 ↪groupby('show_title')['cumulative_weeks_in_top_10'].max().
 ↪sort_values(ascending=False).head(10)

# Plot the top shows in Argentina by cumulative weeks in top 10
plt.figure(figsize=(10, 6))
top_shows_by_country.plot(kind='bar', color='lightblue')

# Set the labels and title
plt.xlabel('Show Title')
plt.ylabel('Cumulative Weeks in Top 10')
plt.title('Top Shows in United States by Cumulative Weeks in Top 10')

# Rotate x-axis labels for readability
plt.xticks(rotation=45)

# Display the plot
plt.tight_layout()
plt.show()
```

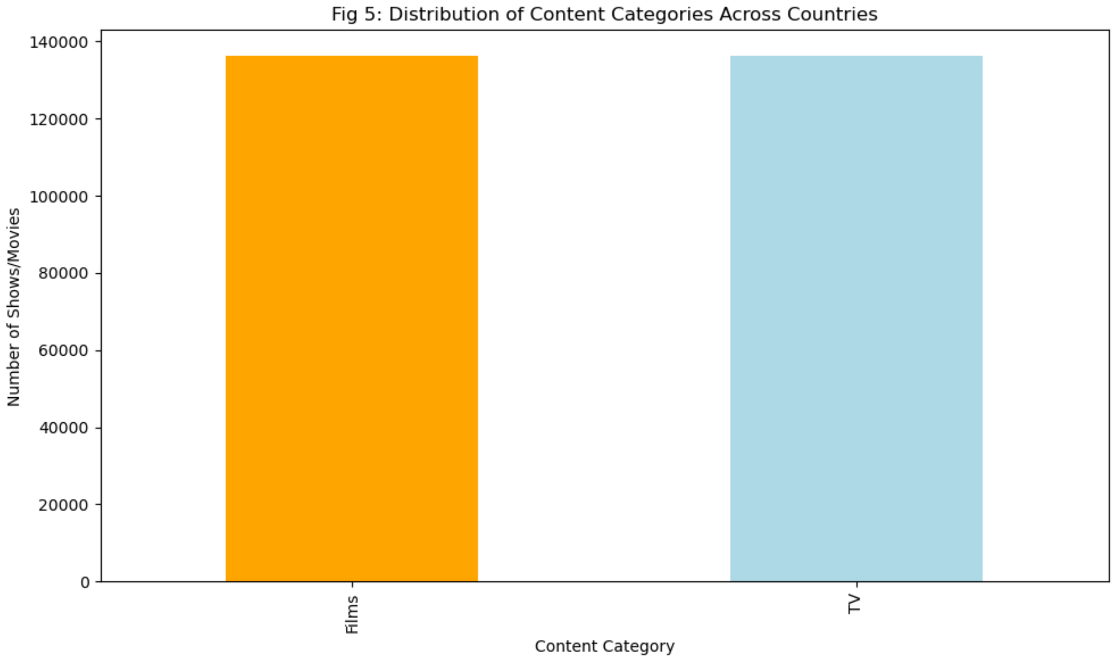


Key Insights: * Sing 2 has spent the most time in the top 10, with over 30 cumulative weeks * Popular Family-Friendly Titles: Titles like The Super Mario Bros. Movie, Despicable Me 2, and CocoMelon—which are family-oriented—have stayed in the top 10 for a significant number of weeks. This suggests a strong demand for family content. * Diverse Content: Alongside family shows, there are darker, more mature series like Stranger Things, Ozark, and Squid Game, reflecting the wide range of genres that perform well in the U.S. market.

Recommendations: * Increase Focus on Family Content: Given the strong performance of family-oriented titles, I would recommend focusing more on family-friendly content, particularly animated movies or shows. * Maintain Diverse Offerings: It's essential to maintain a mix of genres. While family content performs well, darker shows like Stranger Things and Ozark also consistently performed well, indicating the need for diversity in movies and shows offerings.

[24]:

```
# Load the dataset
df_all_weeks_countries = pd.read_excel('all-weeks-countries-netflix.xlsx',␣
 ↪sheet_name='Top 10')

# Group by category and calculate the total count of entries per category
category_distribution = df_all_weeks_countries['category'].value_counts()

colors = ['lightblue' if category != 'Films' else 'orange' for category in␣
 ↪category_distribution.index]

# Plot the distribution of content categories
plt.figure(figsize=(10, 6))
category_distribution.plot(kind='bar', color=colors)

# Set the labels and title
plt.xlabel('Content Category')
plt.ylabel('Number of Shows/Movies')
plt.title('Fig 5: Distribution of Content Categories Across Countries')

# Display the plot
plt.tight_layout()
plt.show()
```



Key Insights: * Balanced Representation: There's a relatively even split between Films and TV Shows, this indicates that audiences are consuming a wide range of both content types.