Aris Chalini

Streaming Services Project

Executive Summary

Netflix and Spotify datasets were analyzed from the lens of seeking to understand user preference for content. In addition to the production of some summary statistics, it is shown that two useful datasets can be produced in order to better understand the distribution of user preferences:

a) An analysis of Netflix content genre-tags and a comparison of their distributions among all Netflix content and among Netflix content that has made it to the top 10 list. b) The creation of a metric which would aid in the comparison of user-preference between the top 200 lists of different Spotify regions.

With these two techniques we can better understand the content-preferences favored by a population and also calculate the similarity between the content-preferences of any two sub-populations.

Some other interesting findings include:

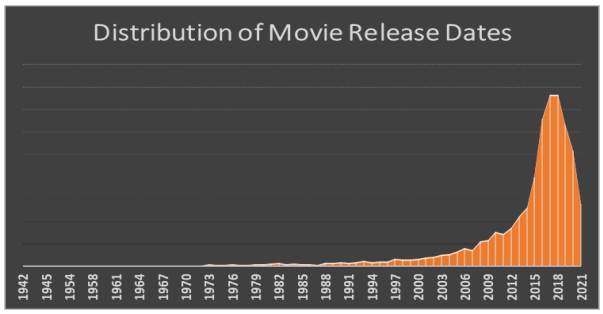
- Foreign content makes up more than half of all the content on netflix.
- Netflix tends to add more content at the beginning of each month.
- The most often appearing genre tags in Netflix Top 10 are: International TV Shows, TV Dramas, International Movies, Faith and Spirituality, Comedies, and Action & Adventure. ● Ed Sheeran has had remarkable success both internationally and domestically when we look at total streams and features in the Spotify Top 200.

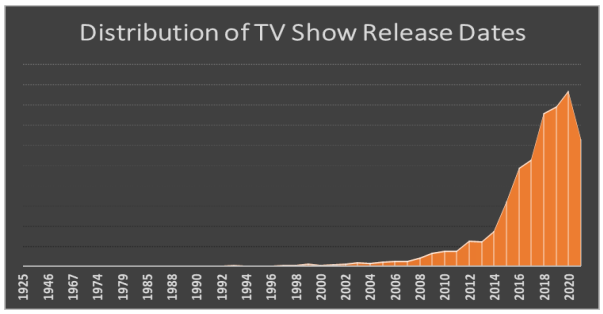
Introduction

Any company that procures content for streaming services should seek to understand the distribution of its user preferences. When this is achieved, a company is better informed to purchase the best content for its users or to market content to the optimal user-populations. In this brief report, I will be conducting an analysis of three tables in the Streaming Services Analysis Dataset: List of Netflix Shows, Top Trending and Watched Netflix Shows, and Top Daily Songs for Spotify. For both these streaming services (Netflix and Spotify), I will provide a general summary of key findings and then present a recommendation for how to understand their respective user-preferences.

Netflix General Findings

Is there a difference between Movie and TV Show release data distribution?
Although it appears minor, TV-Shows tend to have more recent release dates.





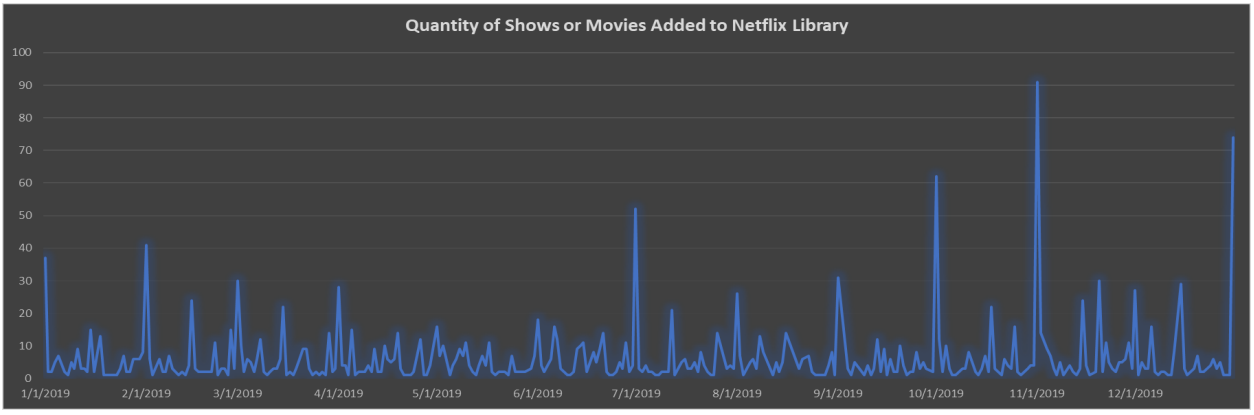
What proportion of Netflix content is US-produced?
Foreign content makes up a greater proportion of the content on Netflix. Only about 35% of content is purely produced in the United States.

Content Origin	Quantity of Content	Percent of Total Content
Foreign	3838	48.12%
US	2818	35.33%
Foreign Mixed	927	11.62%
US Mixed	393	4.93%

What type of content more often makes it into the Netflix Top 10?
TV shows represent **72%** and Movies represent **28%** of content in the Netflix Top 10.

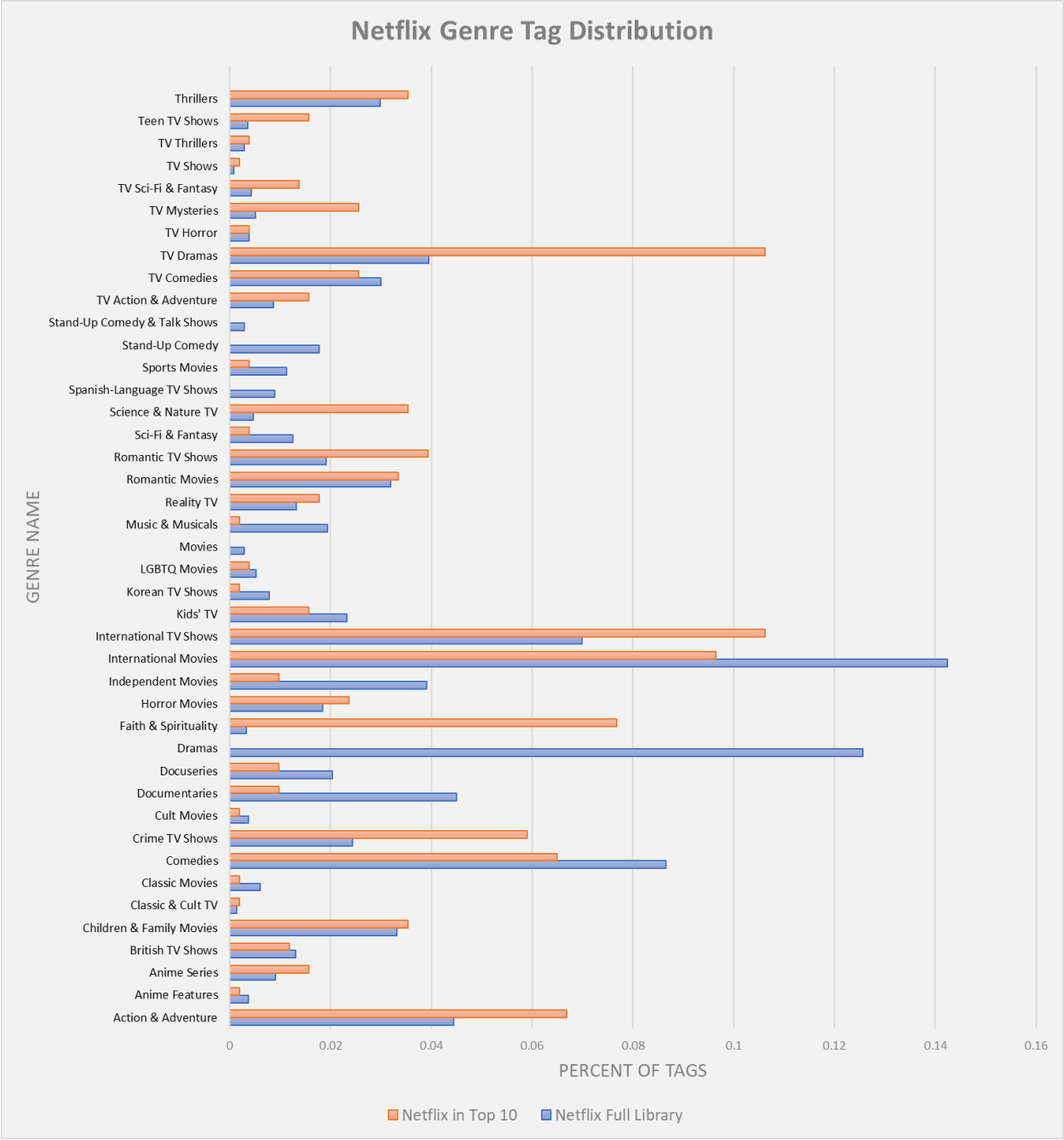
Is English the dominant language for movie content? What about tv-show content?
Among movie content, **73%** is in english. Among tv-show content, **42%** is in english.

When does Netflix tend to add the most shows or movies to their library? Are there any trends?
There does not appear to be any significant seasonal trends when Netflix adds content. However, when we look at years in isolation, there seems to be a day-based trend of adding more content near the beginning of the month. Below is an example of this trend over the course of the year 2019. Notice the peaks on or near the first for each month.



Netflix Genres Distribution

The goal here is to paint a picture of the genre composition of the content that makes it into the top 10. Thus my methodology will involve the bucketing of genre labels listed in the “listed_in” column of the data. We will compare this distribution of genres with that of Netflix’s whole library.



It must be noted that in this case the netflix top10 data is more up to date than the netflix full library data. Thus, not all records in the top 10 will be used in this analysis.

From this comparison we can easily find genres for which the full Netflix library genre proportion greatly differs from the top 10 genre proportion. This could indicate opportunities to add specific content to the full Netflix library in order to more accurately cater to user preference.

We can also use the weekly_hours_viewed column in the top 10 dataset in order to add some weights to the distribution. However since this information is not available for Netflix's whole library, a comparison with this weighted distribution will not be possible. Instead, let's just see which genres represent the majority of viewing hours among movies and shows that make it to the Netflix top 10. Below are the top 20 tags in the Netflix top 10 dataset ranked by proportion of total genre-tag-hours.

TV Thrillers	28.66%
Action & Adventure	10.15%
British TV Shows	9.40%
TV Dramas	7.71%
TV Sci-Fi & Fantasy	6.59%
TV Action & Adventure	4.39%
Comedies	4.10%
TV Comedies	3.39%
TV Mysteries	3.25%
Faith & Spirituality	3.16%
Science & Nature TV	2.18%
Anime Series	1.85%
Horror Movies	1.63%
TV Horror	1.62%
International TV Shows	1.48%
Teen TV Shows	1.30%
Crime TV Shows	1.20%
LGBTQ Movies	1.11%
International Movies	0.92%

Here we find that TV Thrillers are responsible for a disproportionate amount of hours viewed among shows that have made it to the Netflix top 10. Although this second analysis provides additional information about user preferences, it should be noted that lengthier content will naturally have an advantage in this form of ranking.

Spotify General Findings

What are the most streamed songs of all time in the United States?

1	HUMBLE.	Kendrick Lamar	339677217
2	XO TOUR Llif3	Lil Uzi Vert	316206696
3	Congratulations	Post Malone	285451131
4	Shape of You	Ed Sheeran	282319891
5	Mask Off	Future	241828211

What are the most streamed songs of all time internationally?

1	Shape of You	Ed Sheeran	2993988783
2	Despacito - Remix	Luis Fonsi	1829621841
3	Despacito (Featuring Daddy Yankee)	Luis Fonsi	1460802540
4	Something Just Like This	The Chainsmokers	1386258295
5	HUMBLE.	Kendrick Lamar	1311243745

Which artists appear in the Spotify top 200 the most in the United States?

1	Drake	3210
2	Kendrick Lamar	1879
3	Lil Uzi Vert	1860
4	Post Malone	1827
5	Ed Sheeran	1592
6	Migos	1585
7	The Weeknd	1569
8	Future	1524

Which artists appear in the Spotify top 200 the most internationally?

1	Ed Sheeran	127064
2	The Chainsmokers	68321
3	Drake	49054
4	Calvin Harris	38460
5	Imagine Dragons	37992
6	J Balvin	37320
7	Martin Garrix	35043
8	Ozuna	34409

Notable Observations:

- “Humble” and “Shape of You” are featured in both the top 5 streamed songs of the United States and overall Internationally.
- “Drake” and “Ed Sheeran” are among the top 3 artists that make the top 200 list the most often both in the United States and internationally.

Spotify Region Similarity Metric

I believe that it would be interesting to learn how the user-preferences of other Spotify regions compare to user-preferences in the United States. How could we quantify similarity? Here I propose a metric that uses two important measures: (α) the proportion of songs that are shared by the two top 200 lists being compared and (β) the SSE between the total days on their respective top 200 list for each song.

$$Song\ Preference\ Similarity\ Metric\ = \beta\ \times\ (1\ -\ \alpha)$$

Under this mathematical definition, a smaller similarity metric indicates more similarity between the song preferences of the two regions being compared. A larger similarity metric indicates less similarity between the song preference of the two regions being compared.

Country Being Compared to US	Percent Shared Songs on Top 200	SSE of Days in Top 200 for Shared Songs	Proposed Similarity Metric
Canada	0.383	1994519	1231307
Mexico	0.141	3319470	2850719
Japan	0.165	4166652	3477991

For this example data I chose three regions to compare to: Canada, Mexico, and Japan. Canada borders the United States and speaks the same language. Mexico borders the United States and does not speak the same language. Japan does not border the United States and does not speak the same language. When my similarity metric is applied to these three countries (comparing them to the United States top 200 composition) we find that out of the three, Canada is ranked the most similar and Japan the least.

This type of metric generation could be quickened and automated with the use of Python. A metric like this one could be used in order to understand which regions have similar tastes in music. Values from this similarity metric could inform Spotify marketing campaigns or this data could be sold for sociological studies. For example, continuous monitoring of this metric could help researchers who seek to understand how the cultural influence between any two countries has changed over time.

Recommendations for Next Steps

This project demonstrates that it is possible to achieve meaningful representations and metrics that help us understand user-content-preference for streaming services. For the Netflix-genre-distribution, next steps would include (1) stratifying results in order to understand user-preference for sub-categories of content and (2) using machine learning techniques in order to calculate weights for a weighted distribution model that could predict genre composition of content that makes it to the Netflix top 10 list. For the Spotify similarity metric, the next steps would be to (1) calculate the metric for all regional comparisons possible and then (2) find a way to connect more users to each other who belong to regions of high similarity.