Summary Analysis on Netflix

By Yun Xiao, Jake Kahrs, Andrew Xiao, Priya Ramakrishnan, Ni Zhan, Binglei Hong

Introduction:

One of the most universal ways the modern person consumes entertainment is through streaming services. Streaming services provide an easy and convenient medium to allow consumers to choose from a variety of titles of both movies and tv shows. During the peak of the Covid-19 pandemic and the resulting quarantine, many turned to streaming services to keep themselves entertained. This led to a boom in the streaming service industry, with nearly every platform seeing an increase in subscribers. Due to the ever growing popularity of streaming services, our group took an interest in what sort of viewing habits consumers have while streaming and how companies can leverage this information to make more popular content.

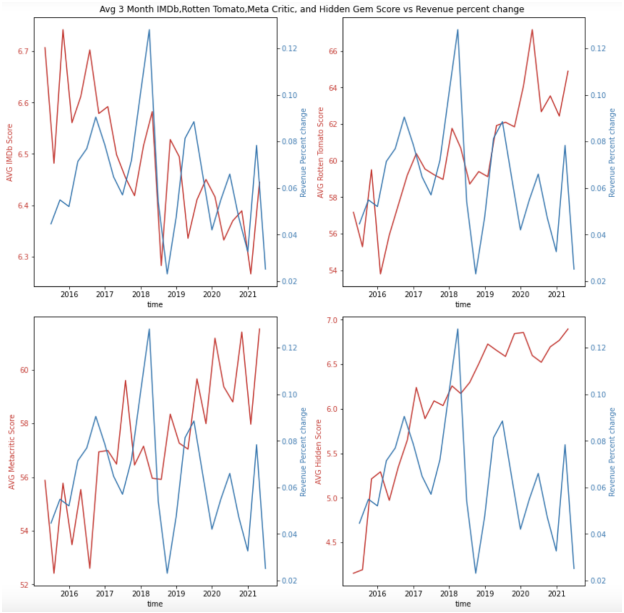
Netflix was founded in 1997 and came from humble beginnings. Long before it was the streaming giant it is today, Netflix originally was a mail order movie rental service, where consumers would order a movie online and then have it shipped to their homes where after a few days they would have to return it. In 2007 the company would undergo a radical change when they started their streaming services. This allowed customers to choose a movie in their own home and then instantaneously be able to watch it. This revolutionized the entertainment industry, changing the way in which consumers were able to watch both movies and tv shows.

With Netflix being the most successful streaming service, it is only natural that they would have the most data on consumer viewing habits. This is in part due to Netflix's size, meaning that the scale of the data collected would serve as an adequately representative sample of the total population. Netflix is also known for keeping meticulous records of consumer viewing habits, which they use to recommend new content. Lastly, Netflix has made a sizeable amount of this data publicly available. The data set we found comes from Kaggle and has data for over 15,000 different movies with 26 different variables per movie with information like writer/director, runtime, languages, and various statistics from different review platforms. The goal of our

group's project was to find any interesting consumer viewing habits so that Netflix could be advised on what to do next.

Analysis:

The first question we tackled was looking at the different ratings and seeing which one had the closest relationship with revenue or subscriber change. In our data set we had ratings from IMDb, Rotten Tomatoes, Meta critic, and their own hidden gem score. To evaluate this we looked at the change in revenue by quarter with their corresponding average rating.



Looking at this figure we can see that imdb score has most resemblance to changes in revenue. After analyzing this figure we decided to look exclusively at IMDb scores. After we determined the best Index to use, we looked at the top 5 rated releases when the revenue percent change was highest and when it was lowest. The highest revenue gain was in the first quarter of 2018 and the lowest revenue gain was in the third quarter of 2018.

For the highest revenue gain, of the top shows, three were foreign and two were American. While the highest IMDb score was 9.1, the average was 6.5.

	Title	IMDb Score	Language
8924	Reply 1988	9.1	Korean
8753	Hey Duggee	8.9	English
8758	A Touch of Green	8.5	Mandarin
8947	Queer Eye	8.5	English
9029	Mob Psycho 100	8.5	Japanese

For the lowest revenue gain four contained English and one was in a foreign language. While the highest IMDb score was 9.2, the average was 6.3.

	Title	IMDb Score	Language
7176	Conspiracy	9.2	English
7170	Im Sorry	9.2	NaN
7230	Car Masters: Rust to Riches	8.8	NaN
8297	Mr. Sunshine	8.8	Korean, Japanese, English
12653	Age of Rebellion	8.7	Mandarin

After determining IMDb as the optimal index for revenue and doing a preliminary search of how these scores relate, we delve further into how language and IMDb scores are related to connect different language films to revenue. We want to use this information to advise Netflix on what sort of new content they should produce and increase their market share in other countries. To accomplish this we compared the average IMDb scores of movies made in languages besides English. After choosing 10 languages, the movies were split into two segments, those shown exclusively in one language and those shown in multiple languages but containing the selected languages. For example, if a movie was only uploaded in Spanish, it would go into the first aka individual segment. If a movie was uploaded in both Spanish and Chinese, it would go into the second aka combination segment and be counted in both of those languages' scores. The same steps were repeated again for series' available on Netflix.

The count of movies that fall into both categories for movies in the languages selected are as follows:

Individual:

```
Spanish = 267
Japanese = 768
Korean = 296
Chinese = 19
Hindi = 294
Russian = 25
French = 196
```

Combos:

```
Spanish =  911
Japanese =  1080
Korean =  451
Chinese =  90
Hindi =  495
Russian =  300
French =  896
Swedish =  139
Arabic =  256
Turkish =  149
```

```
Swedish = 64
Mandarin = 92
Arabic = 89
Turkish = 112
```

The count of series that fall into both categories for movies in the languages selected are as follows:

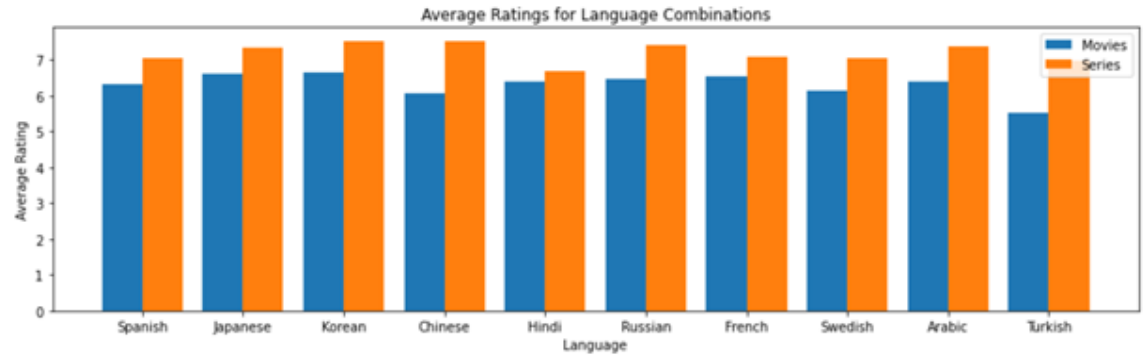
Individually:

```
Spanish = 115
Japanese = 445
Korean = 245
Chinese = 36
Hindi = 35
Russian = 7
French = 56
Swedish = 20
Mandarin = 65
Arabic = 13
Turkish = 19
```

Combos:

```
Spanish =  231
Japanese =  587
Korean =  284
Chinese =  47
Hindi =  44
Russian =  35
French =  160
Swedish =  39
Arabic =  44
Turkish =  27
```

Japanese has by far the largest portion of Netflix's selection when not including English. This is due to anime, which is very popular globally, leading to Netflix increasing the amount of available anime titles to stream. When comparing the IMDb scores for these selected languages for each category, you following charts are created:



The first clear pattern in the data is that universally, series receive a higher average rating than movies. Therefore, it would be advisable for Netflix to focus its attention on creating and purchasing new series to stream. The three languages with the highest average scores for movies are Japanese, Korean, and French. The results are largely the same for the series', with the highest three being Japanese, Korean, and Chinese. Thus it would make the most sense for Netflix to continue to produce or purchase content in these languages in order to have the highest chance of increasing their market share.

For the third question, we extracted keywords from the reviews and gave Netflix some suggestions on the popular elements of the series, so as to better cater to the preferences of the audience.

Based on the average score and the number of series in each genre, we selected the six most popular genres. Then, for each genre, we scraped 1000 series reviews from IMDB according to the scores of the series. After that, we performed text preprocessing techniques on series reviews on each genre. First, we removed stop words and all irrelevant characters, such as numbers, punctuation, white spaces and symbols. Also, we converted all characters into lowercase. Next, we used the python function word_tokenize to split each review into individual words based on certain delimiter. After that, we did two text normalization techniques, stemming and lemmatization, to transform reviews into canonical forms. Lemmatization performed better than stemming.

Then, we converted cleaned data into a numerical format where each word is represented by a matrix. We calculated the TF-IDF score for bigrams, which gave us a better outcome without having to process more reviews.

Comedy

name	tfidf
watch	0.364040
character	0.356799
series	0.279778
drama	0.264637
good	0.236330

Drama

name	tfidf
design	0.330955
steal	0.326811
character bit	0.310825
season great	0.310233
watch hook	0.300168

Family

name	tfidf
watch reboot	0.396871
character plot	0.267439
episode series	0.226115
joon ki	0.214419
great work	0.212860

Fantasy

Sci-Fi

Thriller

name	tfidf
cgi	0.324532
start	0.319153
watch series	0.310785
second half	0.277914
funny watch	0.276121

name	tfidf
character funny	0.353267
saw	0.324429
funny character	0.300397
watch main	0.293788
son	0.261946

name	tfidf
scenery	0.408305
watching	0.308326
foster	0.281759
stay	0.274767
chance	0.271970

Another approach we used to do text mining was topic modeling. Two main methods are LDA and NMF. They're slightly different. LDA generates topics based on word frequency and works better with longer texts. We performed both two methods on a fantasy series. Since most of our reviews are not that long, the NMF method performed better.

Fantasy series with LDA:

```
Fantasy:
The top 5 words for topic # 0
['read book', 'right choice', 'good series', 'kid entertain', 'dance watch', 'old son', 'line good', 'little girl',
'kid learn', 'review year', 'good review', 'daughter watch', 'story line', 'life lesson', 'original series', 'lesson
bad', 'beat bug', 'old daughter', 'old obsess', 'year old']


Fantasy:
The top 5 words for topic # 1
['female lead', 'great actor', 'waste time', 'character good', 'character series', 'type anime', 'character act', 'ac
t story', 'good evil', 'episode watch', 'character development', 'original series', 'voice act', 'happy end', 'charac
ter great', 'second season', 'story line', 'story good', 'time travel', 'main character']


Fantasy:
The top 5 words for topic # 2
['rating review', 'watch rating', 'year watch', 'romantic comedy', 'sad end', 'story wonderful', 'twist good', 'plot
sense', 'teach valuable', 'valuable lesson', 'episode thank', 'anime character', 'bad act', 'great plot', 'story char
acter', 'great story', 'th episode', 'anime plot', 'twist end', 'plot twist']


Fantasy:
The top 5 words for topic # 3
['good way', 'comedy action', 'series second', 'second half', 'episode lot', 'storyline interesting', 'watch actres
s', 'real life', 'watch tell', 'beautiful love', 'love story', 'amaze worth', 'shin min', 'overall good', 'watch dram
a', 'start watch', 'deng lun', 'binge watch', 'watch series', 'worth watch']


Fantasy:
The top 5 words for topic # 4
['st episode', 'episode story', 'rumiko takahashi', 'big fan', 'waste time', 'episode episode', 'movie watch', 'korea
n drama', 'animation good', 'watch watch', 'episode character', 'series plot', 'finish watch', 'episode lot', 'enjoy
watch', 'start watch', 'episode series', 'episode day', 'stop watch', 'watch episode']
```

Then, we chose the NMF method with TfidfVectorizer to do further data analysis. We ignored terms that appeared in less than 2 and more than 95% of the documents. After that, we extracted bigrams and generated five topics for each genre with ten most frequent words. However, the results didn't bring any sufficient insight compared to using TF-IDF score.

Comedy	
1	'great series', 'watch year', 'old love', 'episode year', 'wrong kind', 'beat bug', 'old son', 'old year', 'old daughter', 'year old'
2	'st cheesy', 'inwas wrong', 'chinese drama', 'great cast', 'lot korean', 'drama start', 'great act', 'drama hilarious', 'watch korean', 'korean drama'

3	'good animation', 'character good', 'glitch tech', 'fun watch', 'good evil', 'video game', 'episode watch', 'original series', 'watch episode', 'main character'
4	'bad language', 'bit confusing', 'good story', 'series good', 'thing good', 'female lead', 'great story', 'line great', 'good thing', 'story line'
5	'real life', 'start end', 'rimba racer', 'start watch', 'male lead', 'deng lun', 'female lead', 'chinese drama', 'watch drama', 'worth watch'
Drama	
1	'descendant sun', 'series watch', 'fight scene', 'perfect tv', 'wow wow', 'act young', 'social issue', 'great act', 'watch tv', 'tv series'
2	'watch drama', 'second season', 'love story', 'female lead', 'episode series', 'main character', 'character development', 'worth watch', 'drama watch', 'watch episode'
3	'lead actor', 'rise phoenix', 'great story', 'brilliant story', 'story sense', 'range emotion', 'drama series', 'year old', 'chinese drama', 'story line'
4	'prince historian', 'drama lot', 'love korean', 'writer director', 'series end', 'drama netflix', 'drama series', 'south korean', 'watch korean', 'korean drama'
5	'korean culture', 'character wonderful', 'love story', 'watch lot', 'watch plot', 'series quality', 'plot twist', 'good series', 'series watch', 'korean series'
Family	
1	'daughter watch', 'nursery rhyme', 'watch year', 'old obsess', 'watch episode', 'old daughter', 'old year', 'old love', 'old son', 'year old'
2	'face light', 'old watch', 'rainbow ruby', 'kid watch', 'new episode', 'sit watch', 'kid love', 'love month', 'old love', 'month old'
3	'watch series', 'good good', 'bad guy', 'seuss book', 'main character', 'seuss adaptation', 'voice act', 'egg ham', 'green egg', 'dr seuss'
4	'favorite episode', 'character adorable', 'finish season', 'old son', 'love love', 'theme song', 'old baby', 'main character', 'little boy', 'trash truck'
5	'great fun', 'son love', 'documentary style', 'movie documentary', 'music sound', 'cut funny', 'constant cut', 'wait season', 'funny sound', 'sound effect'
Fantasy	
1	'read book', 'right choice', 'good series', 'kid entertain', 'dance watch', 'old son', 'line good', 'little girl', 'kid learn', 'review year', 'good review', 'daughter watch', 'story line', 'life lesson', 'original series', 'lesson bad', 'beat bug', 'old daughter', 'old obsess', 'year old'
2	'female lead', 'great actor', 'waste time', 'character good', 'character series', 'type anime', 'character act', 'act story', 'good evil', 'episode watch', 'character development', 'original series', 'voice act', 'happy end', 'character great', 'second season', 'story line', 'story good', 'time travel', 'main character'
3	'rating review', 'watch rating', 'year watch', 'romantic comedy', 'sad end', 'story wonderful', 'twist good', 'plot sense', 'teach valuable', 'valuable lesson', 'episode thank', 'anime character', 'bad act', 'great plot', 'story character', 'great story', 'th episode', 'anime plot', 'twist end', 'plot twist'
4	'good way', 'comedy action', 'series second', 'second half', 'episode lot', 'storyline interesting', 'watch actress', 'real life', 'watch tell', 'beautiful love', 'love story', 'amaze worth', 'shin min', 'overall good', 'watch drama', 'start watch', 'deng lun', 'binge watch', 'watch

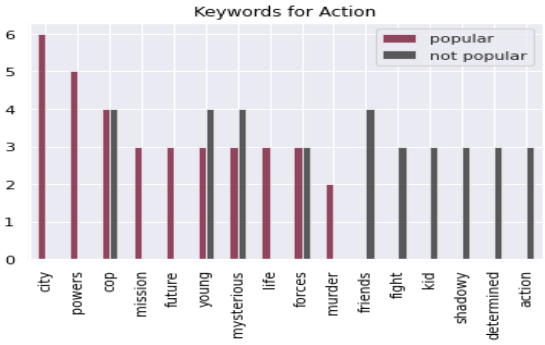
	series', 'worth watch'
5	'st episode', 'episode story', 'rumiko takahashi', 'big fan', 'waste time', 'episode episode', 'movie watch', 'korean drama', 'animation good', 'watch watch', 'episode character', 'series plot', 'finish watch', 'episode lot', 'enjoy watch', 'start watch', 'episode series', 'episode day', 'stop watch', 'watch episode'
Sci-Fi	
1	'story bad', 'storyline sofiane', 'good thing', 'animation good', 'series main', 'character personality', 'story place', 'series great', 'character annoy', 'main character'
2	'animation bit', 'old watch', 'mad kid', 'good action', 'old daughter', 'watch episode', 'son year', 'sci fi', 'old son', 'year old'
3	'comic relief', 'series time', 'quality tv', 'mighty morphin', 'ranger fan', 'new cast', 'morphin power', 'hope season', 'ranger series', 'power ranger'
4	'time watch', 'anime story', 'binge watch', 'plot twist', 'waste time', 'season great', 'finish day', 'watch second', 'worth watch', 'second season'
5	'tv series', 'voice actor', 'great story', 'good story', 'suit gundam', 'watch season', 'long time', 'gundam series', 'mobile suit', 'story line'
Thriller	
1	'development main', 'plot sense', 'video game', 'character development', 'easy understand', 'series main', 'interesting character', 'great anime', 'watch episode', 'main character'
2	'long time', 'good review', 'series time', 'mighty morphin', 'new cast', 'ranger fan', 'morphin power', 'hope season', 'ranger series', 'power ranger'
3	'time good', 'series plot', 'watch tv', 'binge watch', 'second season', 'series character', 'character development', 'real life', 'good tv', 'tv series'
4	'serial killer', 'episode actor', 'bad edit', 'watch episode', 'bad act', 'story line', 'act waste', 'complete waste', 'plot character', 'waste time'
5	'wait season', 'bad review', 'story line', 'watch series', 'watch season', 'good story', 'series good', 'worth watch', 'act good', 'good act'

As for the fourth question, we aimed at analyzing keywords in summaries of most popular and least popular content on Netflix. By looking at the choices of those words, we want to provide suggestions about choosing targeted keywords.

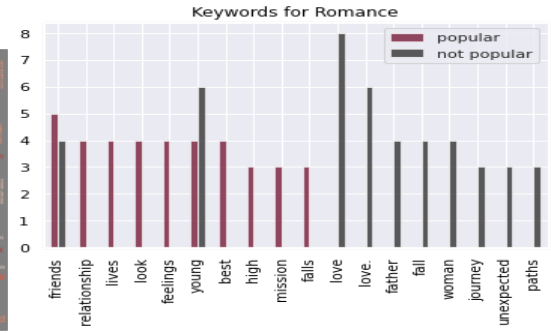
Firstly, we chose “IMDb votes” as the measurement of popularity. Because it is also related to those in other OTT platforms and cinemas, from three datasets, we used the index and map function to get those only available on Netflix. After dropping null values, we split the dataset into different groups by Genre.Then from 28 genres, we selected those including the most content: thriller, action, romance, drama, comedy and documentary. Based on IMDb votes, we chose the top 10% most popular and the

bottom 10% least popular content for every genre. After dropping stopwords, we built wordcloud charts for popular movies and shows and created bar charts of the top 10 keywords for both popular and not popular content. And the following charts are the most interesting of the genres we analyzed. Apparently the word selections in summaries for popular and not popular content are different. We have similar findings for other genres.

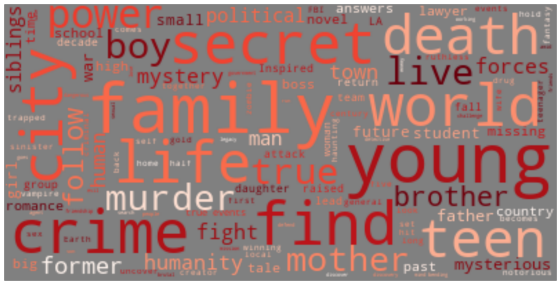
Action:

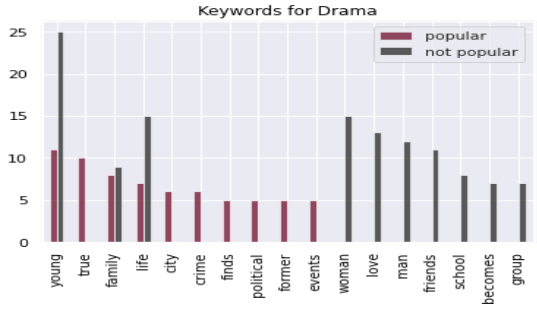


Romance:



Drama:





Conclusion:

IMDb is the most influential rating site in terms of Netflix's revenue. The most content is in Japanese. Korean is the most popular by average score for both movies and series.

Text analysis is powerful and managers should use it to improve user experience.

Keyword selections for popular and unpopular content are different. To attract audiences, the company should choose more targeted keywords when writing summaries for their popular content.

Bibliography:

Kariuki, P. (2021, October 22). *How and when did netflix start? A brief history of the company*. MUO. Retrieved November 20, 2021, from https://www.makeuseof.com/how-when-netflix-start-brief-company-history/#:~:text=2007%3A%20Netflix%20begins%20streaming%20content,month%20physical%20DVD%20subscription%20tier

Stoll, J. (2021, November 8). *Global number of SVOD subscribers by service 2026*. Statista. Retrieved November 20, 2021, from https://www.statista.com/statistics/1052770/global-svod-subscriber-count-by-platform/#:~:text=As%20of%20September%202020%2C%20the,market%2C%20with%20117%20million%20users

Datasets:

Netflix Data:
https://www.kaggle.com/ashishgup/netflix-rotten-tomatoes-metacritic-imdb

TV shows on Netflix, Prime Video, Hulu and Disney+:
https://www.kaggle.com/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney

Movies on Netflix, Prime Video, Hulu and Disney+:

https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney