Analysing the Trends in Netflix and Predicting the Next Block-Buster

HWAYUN, JOH* and SOHYEON, PARK*, Ewha Womans University

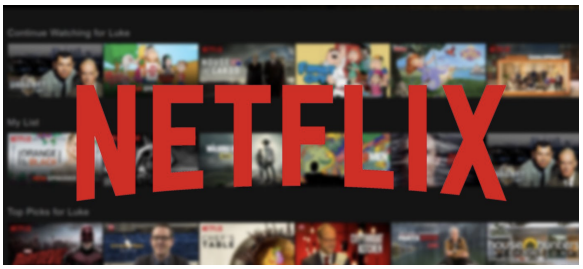

Fig. 1. Netflix, the most popular online streaming platform worldwide

CCS Concepts: • **Data analytics**;

Additional Key Words and Phrases: Data Analysis, Netflix Data, Digital Content, Trend Analysis

ACM Reference Format:
Hwayun, Joh and Sohyeon, Park. 2018. Analysing the Trends in Netflix and Predicting the Next Block-Buster. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Human society faced the most disastrous pandemic in 2019, Coronavirus, and even got named Covid-19. Due to the pandemic, many countries had to go through quarantine which made people to be forced to stay in their house for months. This caused them to use various online streaming platforms instead of going to the movies, and the most popular online streaming platform known worldwide is Netflix. BBC News reported that due to Covid-19 lockdown, Netflix received 16 million new subscribers [2]. However, as time passes by, the original subscribers of Netflix are turning their backs to other online platforms and it is now time for Netflix to make a new business decision in order to bring back its past popularity [3]. To do so, we perform data analysis based on the contents included in the Netflix's movie and TV list and suggest a business guideline to give insight for Netflix in taking the next step, such as what type of content they should include.

There were many prior projects that made an attempt to perform data analysis based on the TV and Movies data in Netflix [4, 6]. Thiago *et al.* analyzes the number of contents in TV shows and movies each, and also explores the top 5 oldest and newest movies and TV shows [6]. However, focusing only on the date of when each content were included in the Netflix list does not give

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Netflix the proper insight in which business decision the company should make in order to bring back the former popularity. Moreover, Saurav *et al.* analyzes the data based on countries, presenting business analysts the insight in where the content should be directed in [4]. The project also presents the most appeared cast in Movies and TV shows in a global scale, which would help Netflix in deciding what content they should include in their repository. However, it lacks of analysis in most presented cast and director based on country. It is necessary to perform the analysis since Netflix uses different repositories depending on the country. Such analysis would make it easier for Netflix to decide in which content to include depending on the country.

The Netflix dataset we used in the analysis consists of meta details about the movies and tv shows such as the title, director, and cast of the shows/movies. It also includes detailed information of each content such as the release year, the rating, duration etc. More information about the data can be found in Kaggle (https://www.kaggle.com/shivamb/netflix-shows). Through the analysis, such as understanding what content is available in different countries, identifying similar content by matching text-based features, performing analysis of Actors/Directors of the movies and TV shows, we would find interesting insights.

2 SOURCE DATA AND DEVELOPMENT TOOLS AND ENVIRONMENT

The data we used for analysis was stored in Kaggle and was last updated in April, 2021 [1]. It consisted of in total 7787 unique values of TV shows and movies that were published between 1925 and 2021 and are currently included in the Netflix list. There were in total 12 attributes: *show_id, type, title, director, actor, country, date_added, release_year, rating, duration, listed_in, description*, where *show_id* attribute acted as the primary key that distinguished each value from the others. The original Kaggle data consisted of different types of data type depending on the attributes, however as we uploaded the data to Google Dataflow Cloud, we updated the data type for all attributes as string type for convenience. There were in total 2389 null values in the data which we had to eliminate based on the analysis we performed on the data.

We used Google Bigquery to create the data warehouse, since it provided a free trial for 90 days. We previously tried to use IBM Db2 Database version 11.5, but unfortunately we constantly faced an error while installing the program. We used Google Cloud to store the Netflix data and used Structured Query Language (SQL) to interact with the relational databases. Moreover, to visualize the results of the queries, Google Data Studio was used.

3 DESIGN OF DATA WAREHOUSE

The result of our relational database as an ER diagram is shown as in Figure 2. The *Netflix_origin* table is the primary parent table to all other tables and consists of all the attributes that would be used as a foreign key in other child tables. All the first child tables reference the primary key of the table as their primary and foreign key. The child table for *Netflix_origin* table are as follows: *TV Show*, *Movie*, *All Date*, and *Director and Actor*. The *TV Show* table consists of only 'TV Show' type related records. Similarly, *Movie* table consists of only 'Movie' type related records. Lastly, *All Date* table contains date related attributes on both types of content, such as release data and added date.

We created in total 6 views to simplify the selection process when counting the number of records for certain values, where two of them were extrapolated from the parent table and the other four from the child table.

4 SCHEMA IMPLEMENTATION

We used bitmap indexing to count the records based on ratings in both types of content, in order to easily count without fully scanning the whole table. Since the process of using bitmap index is

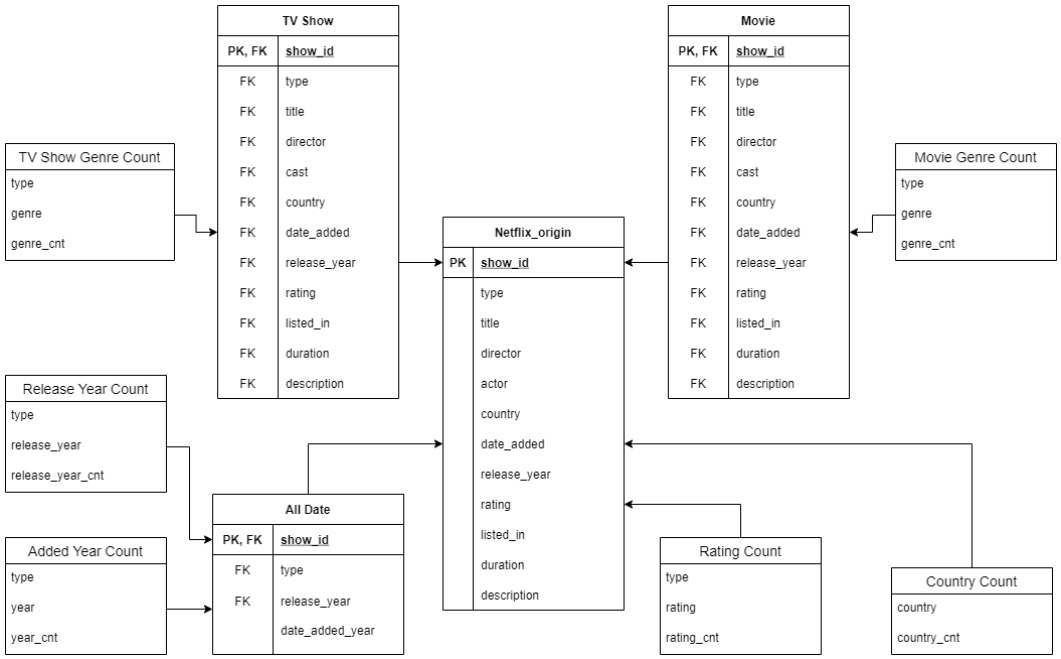

Fig. 2. ER Diagram of the data warehouse.

the same for all types of rating indexes, we only show some of the examples in each TV show and movie type content (N=10).

Figure 3 is the bitmapped index we created based on the rating system presented in Netflix [5]. The rating system is divided into 3 types of rating scale: kid (TV-Y, TV-Y7, G, TV-G, PG, TV-PG), teen (PG-13, TV-14), and adult (R, TV-MA, NC-17). According to Julia *et.al* [7], it reveals that the age range with the most users in Netflix is in their 20s and 30s. Therefore, we decided to count the records in the adult rating (*e.g.,* R, TV-MA, NC-17, etc.) using bitmap index. The index value changed depending on the rating value. We created two rating indexes (*i.e.,* R, TV-MA) and conducted a boolean test. As shown in Figure 3, for the first boolean test, we used an *OR* operation to combine the counting of both rating indexes. Then, we performed an *AND* operation with both content type indexes to provide the final result index.

5 QUERY ANALYSIS

5.1 Movie vs. TV Show

To understand what type of content is most created and included in the Netflix list and to distinguish what type of content should be included to earn popularity, we analyzed the ratio of Movie and TV show contents. To do so, we used selection query for two types of content each, and counted the number of records in each type as below:

```
SELECT * FROM Netflix_origin WHERE type="TV Show";
SELECT * FROM Netflix_origin WHERE type="Movie";
```

Program Type	Program Rating	R Index		TV-MA Index		R / TV-MA Index		TV Show Index		Final Index
TV Show	TV-14	0		0		0		1		0
Movie	R	1		0		1		0		0
Movie	R	1		0		1		0		0
Movie	TV-MA	1		1		1		0		0
TV Show	TV-MA	1	OR	1	=	1	AND	1	=	**1**
Movie	TV-MA	1		1		1		0		0
Movie	TV-14	0		0		0		0		0
TV Show	TV-MA	1		1		1		1		**1**
Movie	TV-14	0		0		0		0		0
Movie	TV-PG	0		0		0		0		0

(a) TV show contents with rating R and TV-MA.

Program Type	Program Rating	R Index		TV-MA Index		R / TV-MA Index		Movie Index		Final Index
TV Show	TV-14	0		0		0		0		0
Movie	R	1		0		1		1		**1**
Movie	R	1		0		1		1		**1**
Movie	TV-MA	1		1		1		1		**1**
TV Show	TV-MA	1	OR	1	=	1	AND	0	=	0
Movie	TV-MA	1		1		1		1		**1**
Movie	TV-14	0		0		0		1		0
TV Show	TV-MA	1		1		1		0		0
Movie	TV-14	0		0		0		1		0
Movie	TV-PG	0		0		0		1		0

(b) Movie contents with rating R and TV-MA

Fig. 3. Table on the far left shows the program types that Netflix provide and the rating of each program. The rows colored in green indicates the result after performing the bitmap indexing. (N=10).

5.2 Content Added Year

As mentioned before, Netflix has recently been experiencing a decline in the number of subscribers. To understand the reason, we analyzed the number of contents included in the Netflix list in an annual manner since the year Netflix had been founded. We first created a new table, *All Date*, by splitting the front (month) and last part (year) of the string data stored in the 'date_added' attribute. In the *All Date* table, we performed a selection query while eliminating the null data, and counted the number of records in each year. Both of the queries are as below:

```
SELECT show_id, type, release_year, RIGHT(date_added, 4) AS date_added_year,
FROM Netflix_origin;

SELECT type, date_added_year AS year, COUNT(date_added_year) AS year_cnt
FROM all_date
WHERE date_added_year IS NOT NULL
GROUP BY type, date_added_year;
```

5.3 Content Released Year

We first analyzed the trend in the number of contents being released in both types of content, TV show and movie. This will allow us to identify what type of content Netflix would have to include in order to increase its declining popularity. We first changed the data type of the 'release_year' values to an integer type, then counted the number of contents in both TV Show and Movie type

using selection query on *All Date* table and counted the number of records in each 'release_year' value as below:

```
SELECT type, CAST(release_year AS INT64) AS release_year,
COUNT(release_year) AS release_year_cnt
FROM all_date
GROUP BY release_year
ORDER BY release_year DESC;
```

5.4 Rating

We analyzed the ratings of both types of contents to understand if Netflix is widely available for all age ranges, considering the fact that Netflix may be attracting only a certain type of age range. We used a selection query on *Netflix_origin* table and counted the number of records in each rating and content type value, while eliminating the null values as below:

```
SELECT type, rating, count(rating) AS rating_cnt
FROM Netflix_origin
WHERE rating IS NOT NULL
GROUP BY type, rating
```

5.5 Number of Contents in Each Country

To identify where Netflix is mostly receiving its popularity, we analyzed the origin of the contents included in the Netflix list. We used selection query on *Netflix_origin* table and counted the number of contents under each country and sorted from highest to lowest as below:

```
SELECT country, COUNT(country) AS county_cnt
FROM Netflix_origin
GROUP BY country
ORDER BY country_cnt DESC;
```

5.6 Top 20 Movie and TV Show Genre

To provide Netflix the insight in what type of genre they should consider including for both TV show and movie type content, we analyzed the top 20 genres of both types of contents included in the Netflix list. We performed the same selection query on *TV Show* and *Movie* table each, and split the string value of the 'listed_in' attribute by using comma as a denominator. Then we counted the number of records in each genre and sorted the result data in a descending order. Below is the query used in both types of content, where only the name of the table differed:

```
SELECT type, listed_in AS genre, count(listed_in) AS genre_cnt
FROM movie, UNNEST(SPLIT(listed_in, “, ")) AS listed_in
GROUP BY type, genre
ORDER BY genre DESC;
```

6 RESULTS AND SIMPLE PERFORMANCE ANALYSIS

6.1 Query Execution Results

Through performing data analysis on Netflix TV show and movie data, we noticed that there are higher ratio of movie contents compared to TV show contents as in Figure 7. However, while

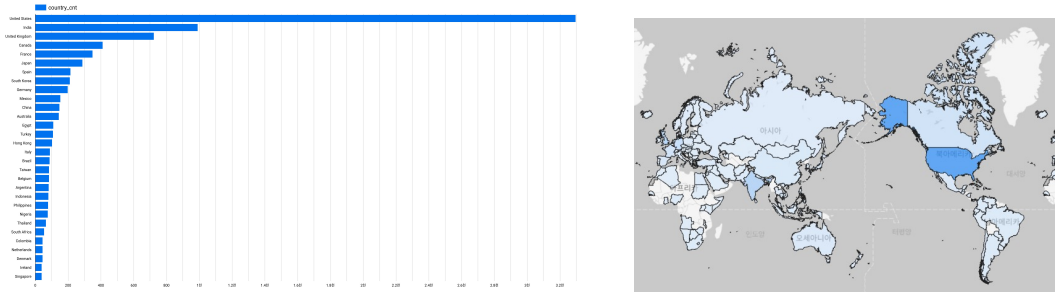

Fig. 4. **Left**: Top 30 countries that has the most contents, **Right**: The number of contents each country possesses visualized in a world map.

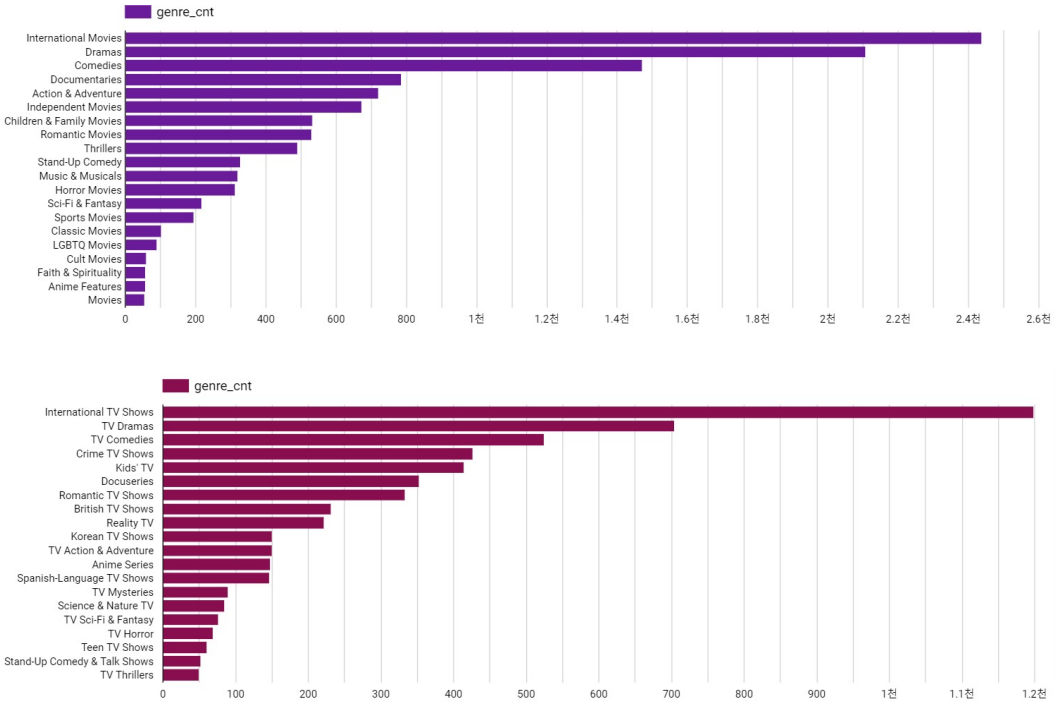

Fig. 5. **Top**: Top 20 movie genre, **Bottom**: Top 20 TV Show genre.

Figure 8 reveal that the number of contents for both types of content that have been included in the Netflix list have declined, **??** shows that the release of TV shows has been growing consistently over the past few years, while the number of movie releases declined. This shows that Netflix has been lacking focus in including the released TV shows, which might have caused the subscribers to move to other platforms.

Moreover, Netflix lacks multiple contents in a diverse range of types and our analysis findings would enable us to provide insight into what types of content Netflix should include in their content repository. For instance, as in Figure 4, Netflix contains contents that are mostly created in the

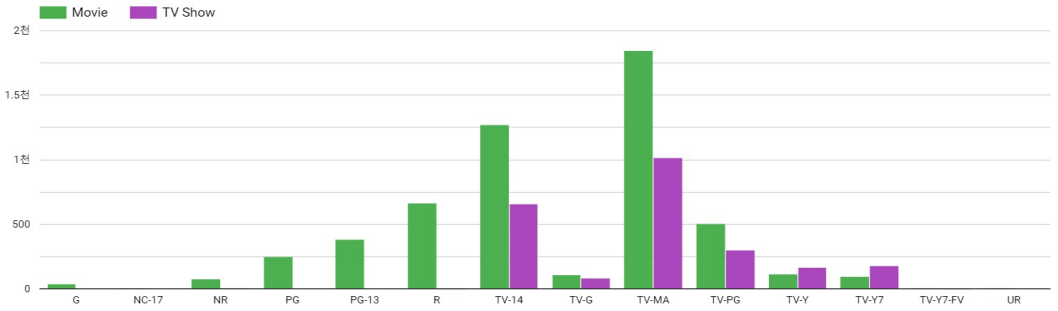

Fig. 6. The distribution of the ratings in both movie and TV show type contents.

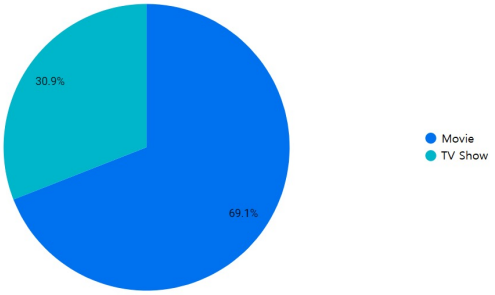

Fig. 7. Ratio in the number of contents between TV shows and movies.

United States. Even considering the fact that the United States has the biggest fandom in media worldwide, the result still shows that the contents in the Netflix list is quite biased. Also, analyzing the genres that have the most contents in both genres identifies that it also has a bias. Figure 5 reveals that the difference in the number of contents between the top 3 and the other genres in both TV show and movie is large. Last but not least, for the analysis of the number of contents for both content type in each rating provides us a result as in Figure 6. This clearly shows that there is a lack of a wide variety of content with different ratings, which may limit the choice of view for people in a certain age range. As such, the results of the analysis we performed above provides Netflix the insight that it would be better to increase the diversity in their contents to attract more subscribers.

6.2 Simple Performance Analysis

We first analyzed the performances for creating the tables: *TV Show*, *Movie*, and *All Date*. Each table took 2.8, 4, and 4.3 seconds to create and used 1.87, 4.26, and 6.2MB of data storage. The average time it took for the query executions was 1.48 seconds and the average storage use was 19.701KB. One of the query executions, counting the records of the release year of both content types, raised the average storage use since we used string split during the selection query.

7 CONCLUSION

Netflix received an exponential growth in the number of subscribers since Covid-19 occurred, due to various countries ordering quarantine. However, recent statistics revealed that Netflix is experiencing loss in their number of users to other online platforms. For Netflix to regain its

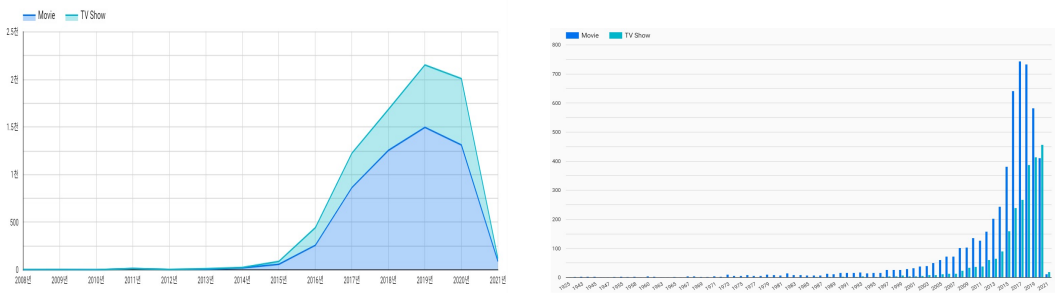

Fig. 8. **Left**: The trend in the number of contents added in Netflix every year in both TV show and movie. **Right**: The trend in the number of contents released every year in both movies and TV shows.

original popularity, it should make new and insightful business decisions through the data analysis we performed in this paper. We analyzed the overall Netflix's movie and TV show content data and reveal problems Netflix need to resolve while also providing advice to their future content creations. Through the analysis, we identified that Netflix lacks diversity in their contents, which might have enticed only certain types of people. Netflix should eliminate the existing bias in their overall contents to attract a wide range of users.

REFERENCES

[1] Shivam Bansal. [n.d.]. *Netflix Movies and TV Shows.* Retrieved May 20, 2021 from https://www.kaggle.com/shivamb/netflix-shows

[2] BBCNews. [n.d.]. *Netflix gets 16 million new sign-ups thanks to lockdown.* Retrieved May 14, 2021 from https://www.bbc.com/news/business-52376022

[3] Yasin Ebrahim. [n.d.]. *Netflix Slumps as Dwindling Subscriber Growth Points to Scary Things Ahead.* Retrieved May 14, 2021 from https://www.investing.com/news/stock-market-news/netflix-earnings-revenue-beat-in-q1-2480011

[4] Saurav Joshi. [n.d.]. *Netflix EDA and Data Visualization Plotly.* Retrieved May 14, 2021 from https://www.kaggle.com/sauravjoshi23/netflix-eda-and-data-visualization-plotly

[5] MediaSmarts. [n.d.]. *Understanding the rating systems for movies.* Retrieved May 31, 2021 from https://mediasmarts.ca/tipsheet/understanding-rating-systems

[6] Thiago Panini. [n.d.]. *Insights from Netflix: The show must go on!* Retrieved May 14, 2021 from https://www.kaggle.com/thiagopanini/insights-from-netflix-the-show-must-go-on

[7] Julia Stoll. [n.d.]. *Share of Netflix users in the United Kingdom (UK) in 1st quarter 2017 and 1st quarter 2018, by age group.* Retrieved May 31, 2021 from https://www.statista.com/statistics/963432/netflix-users-by-age-united-kingdom-uk/