Popular Video Content Recommendation System for Over-The-Top (OTT) Streaming Services

CHAN Kar Chun
20729353
The Hong Kong University of Science and Technology
paddy.chan@connect.ust.hk

LAKHANI Harsh Sunil
20910249
The Hong Kong University of Science and Technology
hslakhani@connect.ust.hk

TSANG Kai Ho
20905476
The Hong Kong University of Science and Technology
khtsangak@connect.ust.hk

WONG Yan Ho
20605624
The Hong Kong University of Science and Technology
yhwongay@connect.ust.hk

ABSTRACT

This report encapsulates a comprehensive study into the content offering dynamics of major Over-The-Top (OTT) platforms, specifically Netflix, Amazon Prime, HBO Max, and Disney+. Our analysis, derived from datasets spanning 18 attributes including content type, run-time, genre, age-certification, and actor-director information, along with a key element, the 'popularity score', reveals a distinct pattern in the content strategy of these platforms. Initial exploratory data analysis (EDA) indicates Amazon Prime as the leader in terms of volume. Paradoxically, Netflix, despite having lesser content, enjoys the highest popularity, prompting an investigation into the quality versus quantity paradigm. We hypothesize that Netflix's success lies in the quality of content, as measured by the 'popularity score', which acts as a proxy for content quality. To substantiate this, we developed a content recommendation tool that deeply analyzes each attribute and its impact on content popularity. This tool serves as an invaluable resource for content producers and OTT platforms to strategize their content creation and acquisition, leading to higher user retention and attraction.

Keywords

OTT platform; Data Visualisation; Visualisation System Design; Netflix; Streamlit; Tableau; Visualisation Dashboard;

1 Introduction

The rapid rise of OTT platforms has revolutionized the entertainment industry, leading to a seismic shift in content consumption patterns. This report delves into a comparative analysis of four leading streaming platforms - Netflix, Amazon Prime, HBO Max, and Disney+, with a focus on understanding the relationship between content offerings and user popularity.

The datasets under study comprise 18 distinct attributes for each platform. These attributes range from basic information such as content type (movie or show), run-time, genre, age-certification, to more specific details like actor-director information and popularity scores. Our preliminary findings through EDA unveiled an intriguing contradiction - while Amazon Prime leads in terms of sheer volume of content, it is Netflix that triumphs in popularity.

Given this conundrum, we propose a hypothesis - Netflix's popularity could be attributed to the quality of its content. The 'popularity score' in our dataset provides a quantifiable measure for content quality, making it an instrumental factor in our analysis.

To validate this hypothesis and provide actionable insights, we have developed a content recommendation tool that scrutinizes each attribute and its influence on the popularity of the content. Our ultimate aim is to assist content producers and OTT platforms in creating and procuring content that resonates with viewers, aiding in user attraction and retention. The insights derived from this study not only unravel the content strategies of these OTT platforms but also lay the groundwork for data-driven decision-making in the rapidly evolving digital entertainment industry.

1.1 Related Work

1.1.1 Motion picture attributes analysis

Motion Picture data visualization is a growing field of research that aims to explore the trends and patterns in user-generated movie ratings. One study [1] analyzed data from the Internet Movie Database (IMDb) and The Numbers, an online movie industry information service, to examine the popularity of different movie genres over time. The authors used various visualization techniques, including Microsoft Excel, Many Eyes, and Google Fusion Tables, to represent the data and identify possible influences on movie genre popularity. The study found evidence that cycles of current events and movie profitability are linked to the production of movies within certain genres. This research sheds light on the factors that influence movie genre popularity and demonstrates the importance of data visualization in understanding complex patterns in movie ratings.

In another related work [2], researchers explored the use of publicly available movie datasets to extract knowledge and visualize it for the purpose of designing accurate movie recommender systems. The authors argue that having a good understanding of the data and user behaviour is essential to building better recommender systems, and that machine learning and neural networks alone are not sufficient. The study aimed to extract insights from movie datasets and visualize them in a way that would provide valuable insights for designing accurate movie recommender systems. This research demonstrates the importance of data visualization in understanding user behaviour and preferences, which can help to improve the accuracy and efficiency of recommender systems for movies and other industries.

1.1.2 Collaborative network analysis

Another related work [3] that can be discussed is a study that analyzed the collaboration networks of actors in the movie industry. The study was inspired by the growing interest in social networks and aimed to model and analyzed the network formed by music artists all around the world. The authors compared their analytic results with generic online friendship models and discovered the most influential nodes in the network using centrality measures. The study also highlighted the importance of music producers in terms of meritocracy versus topological positioning, and discussed the differentiation between collaboration networks using a network fidelity approach. This research sheds light on the importance of collaboration in the movie industry and demonstrates how social network analysis can be used to understand the structure and dynamics of collaboration networks in the entertainment

2 Datasets

To explore the best video content for the OTT platforms primarily involves analyzing the datasets cataloguing all the content available on the platforms. Four datasets used in this project are obtained from Kaggle [4][5][6][7] and summarized in Table 1.

Table 1: Dataset description

OTT Platform	# of Features	# of Videos	% TV Shows	% Movies
Amazon Prime Video	18	9871	14%	86%
Disney+	18	1535	27%	73%
HBO Max	18	3294	23%	77%
Netflix	18	5850	36%	64%

Each dataset contains 15 columns that provide comprehensive information about each available content on the OTT platforms, as shown in Table 2.

Table 2: Data columns

Attributes	Description	Attributes	Description
id	Video ID on JustWatch	description	Description of the video
title	Name of the video	seasons	# of seasons
type	TV show or movie	release_year	Year of release of the video
age_certfication	Age certification for a video	genres	Genres of a video
runtime	Length of a movie or an episode of a TV show	production_countries	Countries produced the video
name	Name of the actor/actress	character	Character name in the
role	Actor or director	imdb_id	IMDb's ID
imdb_votes	IMDb votes for a video	tmdb_score	TMDb score of a video
imdb_score	IMDb score of a video	tmdb_popularity	TMDb popularity

3 Design Requirements

3.1 Design Process

The development of a data visualization system encompasses various essential components that determine the tasks to be executed. These components involve a thorough analysis of the specific purpose of the analytical goal and a comprehensive understanding of the data characteristics, including different types of data attribute. Once the task and data abstraction are comprehended, we can proceed to identify the appropriate visual techniques and formulate the design accordingly. This includes determining the type of dashboard for effective communications, selecting the processing algorithm to model and represent the analytical results, and ensuring a smooth user experience and flow within the data visualization system. Figure 1 illustrates the design tasks and the logical process involved in the development of a data visualization system.

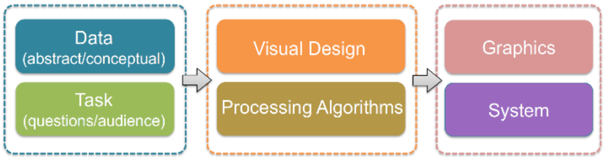


Figure 1: Process of developing a data visualization system

3.2 Task and Data Abstraction

We develop an analysis framework for task analysis, which aims to extract relevant and crucial information from the data visualization system. The framework comprises three levels: low level (Query), middle level (Search), and high level (Analyze). Each level focuses on different aspects of the data analysis process.

3.2.1 Low level (Query): Characteristic of the item

The low level task aims to provide users with an understanding of the statistical information regarding content quantity across the four OTT streaming platforms. This includes:

- Distribution of the number of released videos
- Number of released videos over the past 120 production years
- Popularity of released videos over the past 120 production years

3.2.2 Middle level (Search): Relationship between attributes and popularity

The middle level task involves exploring the relationship between each attribute and content popularity. Through data exploratory analysis as shown in Figure 3, it suggests that there is no significant statistical correlation between individual attributes and content popularity. This indicates that an individual attribute is insufficient to make a video popular. Different combinations of the attributes can result in different levels of sensation in the society. Users can self-explore and identify potential relationships between attributes and popularity, enabling them to form hypotheses and delve into historical data to uncover factors influencing content popularity.

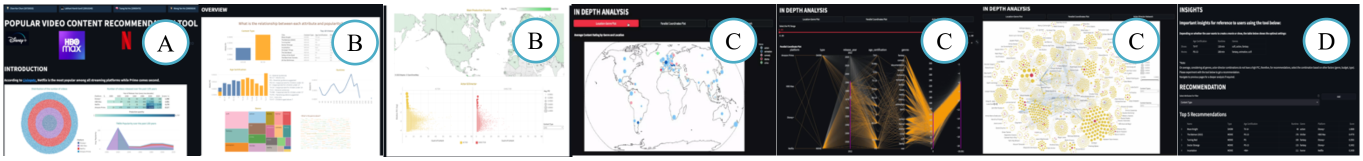


Figure 2: Overview of visualization system: A) Introduction; B) Overview; C) In-depth Analysis; D) Recommendation

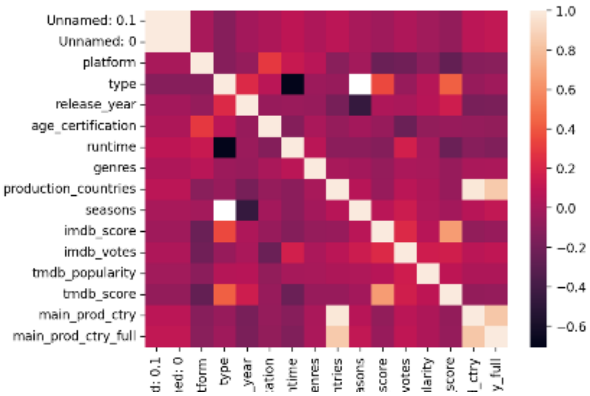


Figure 3: Correlation between all columns

3.2.3 High level (Analyze): Consumption and production of information

The high level task focuses on consuming and producing information by conducting combined analyses of attribute sets. Rather than examining the relationship between attributes and popularity independently, users are provided with the opportunity to gain insights from various dimensions. The following high level tasks are defined to produce analytical information:

- Production Country & Genre vs. Popularity
- Actor & Director vs. Popularity
- All attributes vs. Popularity

4 Visualization System Design

To address the tasks outlined in Section 4.2, our visualisation system is designed with the structure shown in Figure 4.

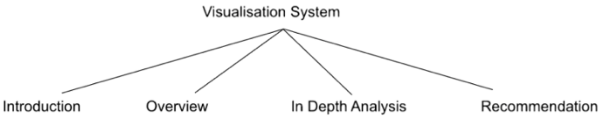


Figure 4: Visualization system structure

We adopt a top-down approach to design the user experience and flow of the website. The "Introduction" page begins by presenting statistics on the quantity of content to users, raising users' curiosity and interest in understanding the specific reasons behind content popularity. It then leads them to other pages, namely "Overview", "In-Depth Analysis", and "Recommendation", which provide users with visualizations designed for answering their curiosity and for their exploration. Overview of these pages are provided in Figure 4.

4.1 Introduction Page

We develop the "Introduction" page specifically for the low level task.

4.1.1 Distribution of the number of videos

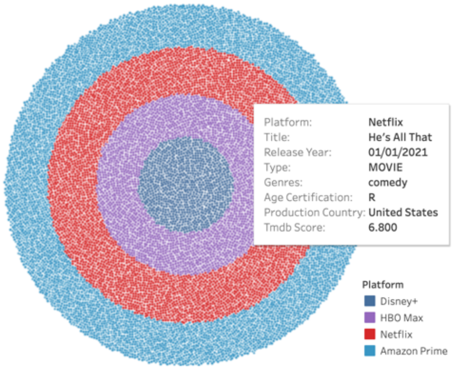

Platform: Netflix
Title: He's All That
Release Year: 01/01/2021
Type: MOVIE
Genres: comedy
Age Certification: R
Production Country: United States
Tmdb Score: 6.800

Figure 5: Distribution of the number of videos on each platform

<u>Purpose</u> : To visualize the size of content distribution among the four OTT platforms

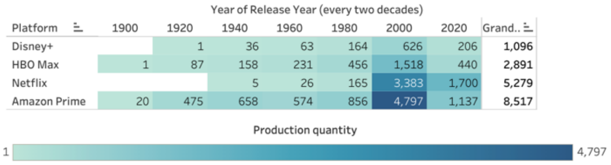
<u>Data type</u> :
- Content video: Numerical
- OTT platform: Categorical

<u>Chart type</u> : Bubble Graph

Figure 5 illustrates the size of distribution of each OTT platform by representing each content video as a data point. The graph uses four distinct colours to encode four categories and area to encode the number of videos available on the platform, allowing for easy comparison of distribution sizes among the platforms.

While both bubble graphs and pie charts utilize area to represent the size of data, the characteristics of the bubble graph can be leveraged to enhance its effectiveness in illustrating data size. By arranging the categories in the bubble graph from the largest to the smallest, with the largest category placed in the outermost position and the smallest in the innermost position, it creates a vivid illustration of size comparisons.

4.1.2 Number of released videos over the past 120 years



	Year of Release Year (every two decades)							
Platform	1900	1920	1940	1960	1980	2000	2020	Grand..
Disney+		1	36	63	164	626	206	1,096
HBO Max	1	87	158	231	456	1,518	440	2,891
Netflix			5	26	165	3,383	1,700	5,279
Amazon Prime	20	475	658	574	856	4,797	1,137	8,517

Production quantity
1
4,797

Figure 6: Number of videos released over the past 120 years

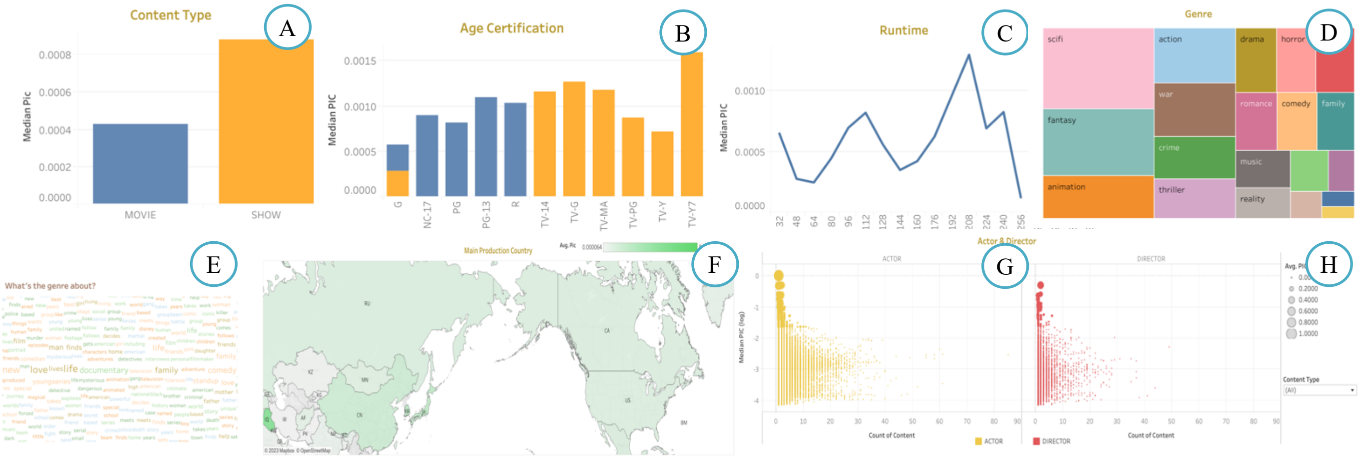


Figure 7: Visualizations in the Overview page: A) Content type vs. Popularity; B) Age certification vs. Popularity; C) Runtime vs. Popularity; D) Genre vs. Popularity; E) Word cloud by genres; F) Production country vs. Popularity; G) Actor vs. Popularity; H) Director vs. Popularity

<u>Purpose</u> : To visualize the change of the number of video available on each OTT platform over time

<u>Data type</u> :
- Content video: Numerical
- OTT platform: Categorical
- Production year: Ordinal

<u>Chart type</u> : Table

Figure 6 presents the number of videos released over the past 120 production years on each platform in a summary table format. To avoid excessive wide table, the number is aggregated every two decades. This table format eliminates the complexity of 3D charts and provides a straightforward representation of the data, facilitating clear comparisons between datasets with three dimensions of values.

In terms of colour coding, a sequential teal colour scheme with uniform saturation is employed. This single hue variation represents the level of quantity, where the colour changes in response to the variation in the number of contents. This colour coding aids in visually understanding the varying quantities across the different production years.

4.1.3 Popularity over the past 120 years

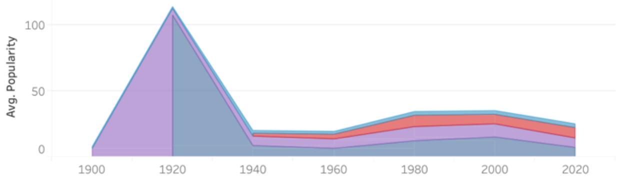


Figure 8: TMDb popularity over the past 120 years

<u>Purpose</u> : To visualize the average popularity of videos available on each OTT platform over time

<u>Data type</u> :
- Popularity: Continuous
- OTT platform: Categorical
- Production year: Ordinal

<u>Chart type</u> : Stacked band plot

Figure 8 is designed to facilitate the comparison of the average popularity of content offered by the four OTT platforms. In contrast to a line graph that uses the direction of lines to indicate upward and downward trends, the stacked band plot uses the size of the bands to represent the level of popularity.

While the use of size as an encoder may not be as effective when two platforms have similar levels of popularity, it still provides a general trend of popularity across all platforms. By visually comparing the sizes of the stacked bands, users can gain insights into the relative popularity of the platforms and observe the overall popularity trends.

4.2 Overview Page

We develop the "Overview" page specifically for the middle level task. This page incorporates all significant attributes into a single interactive system to demonstrate their influence on popularity. Users can experiment with different subcategories of attributes such as show (*content type*), TV-G (*age certification*), or horror (*genre*) to observe their impact on popularity.

To enhance the user experience and avoid overwhelming human memory, we design the page with a juxtaposed layout format instead of superimposition. This layout allows users to easily compare and analyze the visualizations without the need for constant scrolling.

While all attributes on this page are designed to interact with each other, the main objective is to provide a quick understanding of which subcategories of an attribute positively or negatively affect the popularity of a video. To avoid inconvenience in experimenting with different attribute combinations, we undertake more complex visualizations in Section 4.3.

The "Overview" page incorporates a range of visualizations as illustrated in Figure 7. Unlike Section 4.1, given the number of visualizations, their design rationales are summarized in Table 3 to provide an overview of the key considerations and principles behind the design choices made for each visualization.

Table 3: Overview of key considerations and principles for each visualization in the Overview page

Item	Data Type	Chart Type	Design Rationale
Content Type (A)	Nominal	Bar chart	Bar chart enables easy magnitude comparison between categorical items using length.
Age certification (B)	Ordinal	Bar chart	Bar chart enables easy magnitude comparison between categorical items using length.
Runtime (C)	Interval	Line graph	Line graph enables easy pattern, trend and variation identification among continuous (interval) items using direction.
Genre (D)	Nominal	Tree map	Tree map enables easy magnitude comparison between multiple categorical items using area. Compared to a bar chart, a tree map provides a clear visual representation that enables users to discern the varying magnitudes among the 20 genres without overcrowding the plot.
Word Cloud (E)	Nominal	Word cloud	Word cloud vividly displays the most frequent words appearing in the description of a specific genre using size.
Production Country (F)	Geographical	Map	Map provides a visual representation of geographic regions and allows users to easily identify data associated with specific locations. Together with a sequential colour hue, users can discern areas of higher or lower popularity based on the colour intensity.
Actor (G) & Director (H)	Nominal	Scatter plot	Scatter plot enables two-dimensional comparisons and effectively displays the relationship between the involvement frequency of an actor/director and their popularity. The size of the points enhances the emphasis on popularity.

4.3 In-Depth Analysis Page

We develop the "In-depth Analysis" page specifically for the high level task. We shift our focus to a more comprehensive analysis of content popularity by considering multiple factors that contribute to it.

4.3.1 Production country & genre vs. popularity

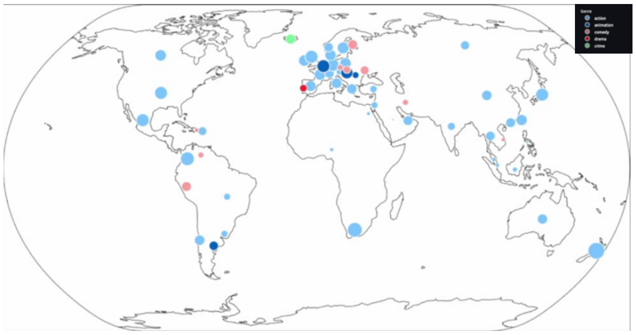


Figure 9: Map comparing popularity with production country & genre

Purpose : To investigate which genre gained the most popularity for each production country

Chart type : Map

Figure 9 represents the most popular genres produced in different countries, utilizing three channels for effective representations:

1) Colour: The colour of the circles on the map corresponds to different genres.
2) Size: The size of the circles indicates the magnitude of popularity. Larger circles represent higher popularity.
3) Position: The geometry of circles on the map indicates the production counties.

By combining these three channels in a single map visualization, users can gain valuable insights into the factors contributing to content popularity, specifically considering the production country and genre. Compared to a tabular visualization, the map provides several advantages. It eliminates the need for duplicated data rows, which can be confusing and require extra effort from users to extract key information. Additionally, the map visualization offers a clustering effect that aids in discovering insights. The visual representation of locations on a map allows users to discern patterns and relationships, such as identifying countries that excel in producing specific genres of content.

4.3.2 Actor & director vs. popularity

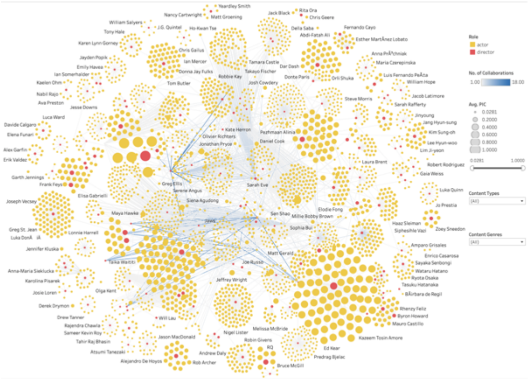


Figure 10: Actor & director's popularity network

<u>Purpose</u> : To investigate which actors and directors were involved in the most popular videos, considering the number of collaborations

<u>Chart type</u> : Network diagram

As illustrated in Figure 10, actors and directors are represented as nodes, and their collaborations are depicted as edges connecting them. The colour intensity of the edges is used to represent the number of collaborations between individuals. The size of the nodes correspond to the overall popularity of the videos they have been involved in.

The network chart surpasses other design options in effectively visualizing network relationships and their strength. By utilizing multiple channels and marks to represent information, it enhances the visual experience and enables users to grasp the relationship structure and strength simultaneously. In addition, it overcomes the limitations of textual formats and 3D charts. Textual formats lack the capability to capture nuanced difference in collaboration, while 3D charts can sometimes mislead with their depth perception.

4.3.3 All attributes vs. popularity

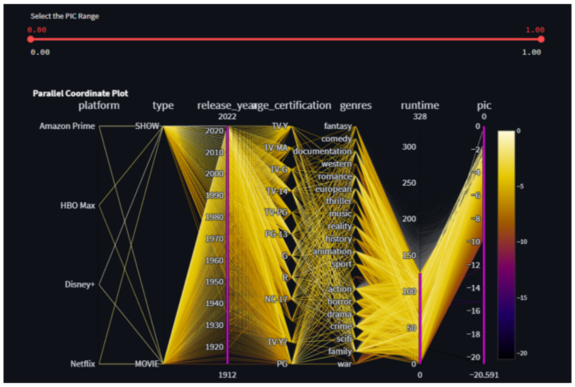


Figure 11: Parallel coordinates comparing all attributes with popularity

<u>Purpose</u> : To investigate which combinations of attributes of a content gained most popularity

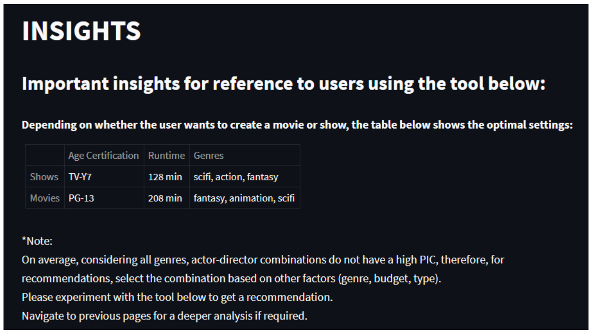
<u>Chart type</u> : Parallel coordinates plot

In Figure 11, the parallel coordinates chart visually presents each attribute as parallel vertical lines connected by edges. The colour intensity of these edges corresponds to the magnitude of popularity.

This chart offers a compact 2D representation that accommodates multiple attributes, enabling users to observe the overall pattern and understand how popularity changes along the lines connecting the attribute axes. We design the parallel coordinates chart with a focus on flexibility. For instance, users can interact with it with a slider that filters the range of popularity, and observe the characteristics of the content in that popularity bucket. Then, users can interact with the plot freely to explore any insights, like filtering out only "drama" in the genre's axis, moving the axis around to get a better view, etc.

4.4 Recommendation Page

The last page of the visualization system consists of two items.



INSIGHTS

Important insights for reference to users using the tool below:

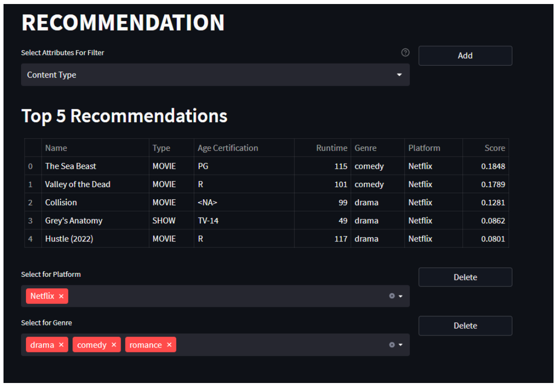
Depending on whether the user wants to create a movie or show, the table below shows the optimal settings:

	Age Certification	Runtime	Genres
Shows	TV-Y7	128 min	scifi, action, fantasy
Movies	PG-13	208 min	fantasy, animation, scifi

*Note:
On average, considering all genres, actor-director combinations do not have a high PIC, therefore, for recommendations, select the combination based on other factors (genre, budget, type).
Please experiment with the tool below to get a recommendation.
Navigate to previous pages for a deeper analysis if required.

Figure 12: Insights provided in the visualization system

First, as shown in Figure 12, a recap of our biggest insights which answers our core question in the first page – what combinations of attributes will provide the best content popularity. There is a final combination for both TV shows and movies.



RECOMMENDATION

Select Attributes For Filter
Content Type

Top 5 Recommendations

	Name	Type	Age Certification	Runtime	Genre	Platform	Score
0	The Sea Beast	MOVIE	PG	115	comedy	Netflix	0.1848
1	Valley of the Dead	MOVIE	R	101	comedy	Netflix	0.1789
2	Collision	MOVIE	<NA>	99	drama	Netflix	0.1281
3	Grey's Anatomy	SHOW	TV-14	49	drama	Netflix	0.0862
4	Hustle (2022)	MOVIE	R	117	drama	Netflix	0.0801

Select for Platform
Netflix

Select for Genre
drama, comedy, romance

Figure 13: Recommendation tool

The second item is a simple recommendation tool that provides users with content recommendations as illustrated in Figure 13. The page works by filtering out contents with different attributes, such as genre, production countries, and runtime, to suggest titles that are most likely to appeal to the user.

It is a simple yet effective way to help users discover new and exciting content that they may have otherwise missed. The page can be used in tandem with all the previous visualizations on the website. For instance, if a user knows there's a country with good movies in the Overview section, they can easily find title recommendations from that country on the recommendation page and filter them by genre or rating.

Overall, while not being a visualization in itself, the recommendation page is an indispensable tool to complete the user's experience with the exploration.

5 Insights & Findings

As mentioned in the earlier sections, the dashboard has the following pages - introduction, overview, in-depth analysis and recommendation. In this section, we break-down the pages to understand the insights from each plot within each page thereby justifying the usefulness of the plots used.

5.1 Introduction Page

Figure 5 indicates that Amazon Prime has the highest number of videos available for streaming, followed by Netflix, HBO Max, and Disney+. Figure 6 shows a steady increase in the number of videos released from 1900 to 1980, followed by a sudden increase in 2000 and a decline in 2020. Figure 8 finds that the average popularity was high back in 1920, possibly due to the fewer content available at that time. The popularity followed a similar trend as the number of videos released, indicating that the popularity of videos is influenced by the volume of content available.

Overall, this page provides insights into the trends and patterns of video releases and popularity over the past 120 years, highlighting the impact of the number of videos available on their popularity.

5.2 Overview Page

The "Overview" page, as shown in Figure 7, with its multitude of informative visualizations, serves as a launchpad to deeper explorations into the factors that drive content popularity. This segment presents essential introductory data, priming the user for more sophisticated plots on subsequent pages. Specifically, it opens the door to discern how each attribute of the dataset influences a content's popularity measure, encouraging user curiosity and facilitating the extraction of actionable insights.

5.2.1 Content Type

Through a straightforward bar graph, we contrast the popularity median scores of movies and TV shows. Employing the median as our measure of central tendency reduces the impact of outliers, thereby ensuring a more accurate comparison. As the linear scale didn't yield clear insights, we shifted to a logarithmic scale. The graph vividly reveals that, in general, TV shows garner higher popularity scores than movies.

5.2.2 Age Certification

Our next point of analysis, age certification, is visualized using another simple bar graph. The graph compares the distinct world-standard age certifications against the popularity score. Here we must understand that each of these certifications are official world standards, therefore, even though some of them may make sense to combine intuitively, we must consider them individually to draw a fair analysis. The analysis demonstrates that TV shows with a PG-13 rating command higher popularity scores. Meanwhile, other age certifications show a negligible variance in their popularity scores, highlighting the distinct appeal of PG-13 rated content.

5.2.3 Runtime

We employ a line graph to illustrate the relationship between runtime and popularity score, treating movies and TV shows separately. Both graphs show a pattern of increasing popularity scores with runtime, peaking at an optimal point before dropping off sharply. For movies, the optimal runtime is about 208 minutes, while for TV shows, it's roughly 128 minutes. This indicates a user preference for content that is adequately long but not excessively so.

5.2.4 Genre

A tree map serves to elucidate the influence of genre on the popularity measure of content. The box size corresponds to the average popularity score of the genre, with distinct colours representing different genres. An accompanying word cloud highlights the most common words in content descriptions for each genre. This graph is interactive too where users can change the type of content - movie or tv show to analyze them separately and users can also click on a specific genre to make the word cloud on the side display the most common words used for that specific genre.

From the graph, we observe that the genres sci fi, action and fantasy are the most popular genres for TV shows compared to the rest. We see something similar for movies as well, where the top genres are fantasy, animation and sci-fi. From this analysis, we deduce that regardless of content type, fantasy and sci fi genres are the most popular.

5.2.5 Geographical Plot

Users are next presented with a geographical plot, revealing the countries producing the most popular content, including details on content type and genre. Interesting observations include the most popular content overall being a TV show from Iraq and the most popular Sci-Fi movie hailing from Sweden. The interactive nature of the plot allows users to derive their unique conclusions based on their requirements and interests.

5.2.6 Actor & director

The final element of the overview page is a scatter plot analyzing the impact of directors and actors on content popularity. This interactive plot highlights collaborations between actors and directors across different projects. It reveals that experienced actors and directors tend to have a lower average rating, while newcomers or those with fewer projects might show potential with higher ratings. This insight is especially useful for content producers deciding on the ideal team for their next project.

5.3 In Depth Analysis Page

The "In-Depth Analysis" page, as shown in Figure 9, Figure 10, and Figure 11, is a powerhouse of rich insights, providing a deep dive into the attributes that influence content popularity. Beginning with a geographical plot in the first section, users explore the relationship between production country, genre, and their collective impact on popularity, thereby uncovering the geographical intricacies that shape audience preferences.

The users are then presented with a parallel coordinates plot in the second section and finally an actor-director network graph in the third section. The former unravels the intricate dynamics of diverse attribute combinations and their effect on popularity, while the latter reveals the complex web of collaborations among actors and directors, enriched with additional attributes like genre and popularity. Together, these tools deliver a multifaceted understanding of the factors driving content success described in greater detail below.

5.3.1 Production country & genre vs. popularity

The analysis reveals that certain genres and production countries are associated with higher levels of popularity. In general, action movies dominate the market, while Iceland is particularly strong in producing crime content. Western countries such as the US, Canada, and New Zealand produce many successful action movies, while Belgium is known for its strong animation content.

5.3.2 Actor & director vs. popularity

The analysis shows that most popular and successful actors and directors do not collaborate very often, indicating that they may be one-hit wonders. Most actor-director clusters are localized, meaning that collaborations tend to occur within specific subnetworks.

By clicking on specific actors and directors within the network graph, we can see the other clusters of actors and directors they have worked with. This allows us to identify potential collaborations based on shared interests and similarities. For example, clicking on actor Bob Odenkirk reveals that he has collaborated with director Vince Gilligan on multiple occasions. Clicking on Gilligan, in turn, reveals a cluster of actors such as Aaron Paul and RJ Mitte who have also collaborated with him. In fact they collaborated in the popular TV show ‘Breaking Bad’.

These insights could be useful for content creators and production companies seeking to create successful movies and shows. By identifying frequent collaborators and potential collaborations based on shared interests and similarities, they may be able to create content that resonates with audiences and has a higher likelihood of success.

Additionally, the analysis highlights the importance of collaboration in the movie and show industry. Successful collaborations between actors and directors can lead to the creation of engaging and memorable content that resonates with audiences. Production companies and content creators may benefit from fostering collaborations between actors and directors, particularly those with shared interests and similarities.

5.3.3 All attributes vs. popularity

The analysis reveals that good content is generally associated with movies produced after 1990, longer runtimes (over 120 minutes), and specific genres such as documentary, romance, animation, and action.

The finding that good content is generally associated with movies produced after 1990 could be attributed to advancements in technology and the availability of better resources for filmmakers during this time period. It is also possible that changes in audience preferences and consumption habits have influenced the types of movies and shows that are successful in recent years. Since Streaming services are only popular in recent years, it is reasonable to assume not many audiences would watch older content intentionally.

The finding that longer runtimes are generally associated with good content is interesting. This could be due to the fact that longer runtimes allow for more complex and engaging storylines, character development, and world-building.

The analysis also reveals that specific genres such as documentary, romance, animation, and action are generally associated with good content. These genres may have specific characteristics that make them more appealing to audiences, such as the ability to convey compelling and emotional stories (romance and documentary) or to transport viewers to imaginative and fantastical worlds (animation and action).

On the other hand, the analysis also shows that certain genres such as war, history, comedy, western, and music have less good content associated with them. This could be due to a variety of factors, such as a lack of innovation and creativity within these genres or changes in audience preferences.

6 Method

6.1 Python

Python is a cross-functional language with many support and offers an extensive range of libraries for different analytical tasks as per use cases. We make use of Python to perform data preprocessing tasks including data cleansing, data transformation and data integration.

6.2 Streamlit

Streamlit is a Python library that allows users to build interactive dashboards with real-time data processing capabilities. Our aim with this project was to build a tool that enables users to get a deeper analysis into different content attributes and get a closer look into how popular content is created. Therefore, Streamlit fits our needs perfectly because it isn’t code-intensive such as other tools including JavaScript or .Net, it allows for integration with other visualization tools such as Tableau as well as other Python visualization libraries such as Plotly Express and Seaborn, and finally because it’s open source and allows for a simple process to deploy the application.

6.3 Tableau

Tableau is a popular open source visualization tool that provides many inherent features that help to create interactive single-page dashboards. We make use of Tableau to build the individual pages in our overall system as described in earlier sections. Using Streamlit, we were able to combine all of the Tableau single-page dashboards into a multi-page web application tool which makes this combination highly efficient and effective.

7 Conclusion

This comprehensive study has brought to light significant findings including an intriguing paradox - while Amazon Prime leads in terms of content volume, it is Netflix that prevails in user popularity, pointing to a potential correlation between content quality and user preference.

The project's centrepiece, a content recommendation tool, has been created using state-of-the-art tools like Streamlit and Tableau, resulting in a multi-page web application. This tool has proven to be instrumental in providing deep insights into the attributes that determine the popularity of content. By facilitating a detailed analysis of each attribute, content producers and OTT platforms can now gain a deeper understanding of what makes content popular among viewers.

The use of diverse visualization plots significantly enhanced our comprehension of the data, allowing us to draw many useful insights. It brought to light the potential hidden patterns and correlations between the attributes and the popularity of the content. This visualization-based approach has also highlighted the potency of data-driven decision-making in the digital entertainment industry.

Finally, our recommendation tool does more than just provide an analysis of attributes; it also offers a pathway for content creators and OTT platforms to understand what type of content to create for increased popularity. This functionality of the tool, acting both as an analysis instrument and a recommendation system, can serve as a catalyst in the decision-making process for content creation and acquisition. It is expected to drive OTT platforms towards higher quality content, thereby increasing their ability to attract and retain users.

In essence, this study underscores the importance of quality over quantity in content strategy, and the potential of data-driven decision-making in the entertainment industry. We anticipate that the insights gained from this project will stimulate further research and development in the field of content analytics and recommendation systems, thereby contributing to the ongoing evolution of the OTT platform ecosystem.

REFERENCES

[1] Dauenhauer, J., Hockett, J., Mammarelli, J., & Yarem, M. (2014). Information analysis of movie genres.

[2] Zarif, N., Chhabra, D., & Polakova, L. What can we learn from Movie Ratings?.

[3] A. Topirceanu, G. Barina and M. Udrescu, "MuSeNet: Collaboration in the Music Artists Industry," 2014 European Network Intelligence Conference, Wroclaw, Poland, 2014, pp. 89-94, doi: 10.1109/ENIC.2014.10.

[4] HBO Max TV Shows and Movies: https://www.kaggle.com/datasets/victorsoeiro/hbo-max-tv-shows-and-movies

[5] Amazon Prime TV Shows and Movies: https://www.kaggle.com/datasets/victorsoeiro/amazon-prime-tv-shows-and-movies?select=titles.csv

[6] Disney TV Shows and Movies: https://www.kaggle.com/datasets/victorsoeiro/disney-tv-shows-and-movies?select=titles.csv

[7] Netflix TV Shows and Movies:https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies?select=titles.csv