Project Report

Title: Analysis of Netflix Movies and TV Shows

Course: DS8003 - Management of Big Data and Big Data Tools

Instructor:

Dr. Roy Kucukates

Group Number:

10

Group Members:

Ishrat Jaben Bushra (501338510)

Sahil Arora(501098147)

Sam Ensafi(501015750)

Table of Contents:

1. Problem Definition:

This section defines five key problems regarding Netflix's content strategy. For each problem, the struggle faced by Netflix is identified, the real world impact or outcomes is explained, and the specific analytical direction of the project is outlined.

Problem 1: Research shows that Netflix's catalog growth has disproportionately favoured movies over TV series, indicating the platform's format balance is skewed in practice (Vu, 2022). Due to this strategy, many major TV shows were not preserved because Netflix allocated more budget and focus to movies. This resulted in harm, with 13 percent of cancellations in Q1 2022 coming from long-term subscribers who left after losing the content they valued (Roth, 2022). This project will analyze the extent of this imbalance by examining the ratio between movies and TV series and determining how significantly the catalog is weighted toward one format.

Problem 2: Research indicates that Netflix's catalog has increasingly centered around a limited set of trend-driven genres, showing a real-life issue of genre concentration and reduced diversity in the platform's offerings (Zahmadi, 2025). A survey of more than 1000 former subscribers found that 42 percent cancelled Netflix due to a lack of interesting content, demonstrating the harm caused when viewers cannot find variety beyond dominant genres (Barnes, 2019). This project will analyze the distribution of content across genres and measure the degree of concentration to determine how strongly the catalog has shifted toward a small number of dominant genres over time.

Problem 3: Research shows that Netflix's libraries across countries are strongly influenced by the U.S.-origin content and show less contribution from many other regions, indicating a real-world imbalance in country-of-origin representation (Lotz et al., 2022). This imbalance also results in significant revenue dependency, since about 44 percent of Netflix's total revenue comes from North America alone, showing over-reliance on a single region (Business of Apps, 2025). This project will examine the extent of this dominance by analyzing the proportion of content from different countries and determining how heavily the catalog relies on a small group of source nations.

Problem 4: Research indicates that a large majority of Netflix's original content is adult-rated, showing a real-life imbalance in age-group targeting (Parents Television & Media Council, 2023). Due to this strategy, in many regions such as Latin America, Disney+ is much more often chosen because kids' content is the main subscription motivator for its users, while for Netflix users, children's programming is not among the main reasons for subscribing (BB Media, 2024). This project will analyze the distribution of maturity ratings to

determine how strongly the catalog is skewed toward adult audiences and to measure how limited the availability of children's and teen content currently is.

Problem 5: Research shows that streaming movies have been getting longer while series lengths are shrinking and platforms are experimenting with shorter runs, indicating that Netflix faces a real-life shift in content duration strategy (Omdia Report, 2025). This shift has contributed to viewer drop-off, shown by Netflix movie completion rates of below 35 percent within 90 days of release (Fletcher, 2023). Furthermore, TV shows with short season lengths can lead to immediate subscription cancellations because viewers finish them quickly and feel less reason to keep their subscription (NewscastStudio, 2025). This project will analyze how content duration has changed over time by examining trends in movie length and the number of seasons in TV series to identify the scale and direction of this shift.

2. Data Description:

2.1. Attributes description

Netflix is one of the most popular streaming platforms worldwide, with over 200 million subscribers and more than 8,000 films and television series available as of mid-2021. The dataset used in this analysis lists every movie and TV show offered on Netflix, including details about directors, cast, country of production, ratings, release year, duration, and other relevant metadata. It comprises 8,807 rows with no exact duplicates or duplicate show IDs, ensuring data consistency. Missing values are mainly found in the director (29.91%), country (9.44%), and cast (9.37%) columns, while all other attributes have minimal or no missing data. Overall, the dataset provides a clear and comprehensive view of Netflix's global catalog, capturing key descriptive attributes such as title, genre, duration, and rating. The columns that will be used to inform our analysis are as follows:

1. show_id (string, nullable = true) – A unique identifier assigned to every movie or TV show, contains 0 missing values (0.0%).
2. type (string, nullable = true) – Indicates whether the entry is a Movie or TV Show, contains 0 missing values (0.0%).
3. title (string, nullable = true) – The title of the movie or TV show, contains 0 missing values (0.0%).
4. country (string, nullable = true) – The country where the movie or TV show was produced, contains 831 missing values (9.44%).
5. date_added (string, nullable = true) – The date the title was added to Netflix, contains 10 missing values (0.11%).

6. rating (string, nullable = true) – The official Netflix TV rating of the title, such as PG, R, TV-MA, contains 4 missing values (0.05%).
7. duration (string, nullable = true) – Represents either total duration in minutes (for movies) or number of seasons (for TV shows), contains 3 missing values (0.03%).
8. listed_in (string, nullable = true) – The genre or category under which the title is classified, contains 0 missing values (0.0%).

2.2. Data Preparation and Statistics of the data (use the tools learnt in this course to generate the data statistics):

PySpark is used to clean and summarize the Netflix dataset. "duration" was standardized and missing values imputed,missing production countries were filled with "Unknown" and its numeric value extracted (minutes for Movies, seasons for TV). Then computed descriptive statistics for the numeric fields and exported the results as CSV for reporting.

Table 1: Descriptive statistics of numeric features

For each numeric feature (movie_duration_min, tv_season, release_year) the table reports count, mean, standard deviation (sd), min, Q1, median, Q3, max, IQR (= Q3–Q1), and bounds (lower_bound = Q1–1.5·IQR, upper_bound = Q3+1.5·IQR).

1. Typical values are given by the median (robust to outliers).
2. Spread, variability is indicated by IQR and sd.
3. Bounds help flag potential outliers beyond normal variation.

Here, movies cluster around 98 minutes (IQR ≈ 87–114),TV content is typically 1 season (IQR 1–2), and titles are concentrated in the 2010s (median 2017), consistent with Netflix's catalog growth.

feature	count	mean	sd	min	q1	median	q3	max	iqr	lower_bound	upper_bound
movie_duration_min	6128	99.58	28.29	3	87	98	114	312	27	46.5	154.5
tv_season	2676	1.76	1.58	1	1	1	2	17	1	–0.5	3.5
release_year	8807	2014.18	8.82	1925	2013	2017	2019	2021	6	2004	2028

Table 1: Descriptive statistics of numeric features

Table 2: Rating distribution (Top 10)

This table summarizes the maturity mix of Netflix titles, displaying the ten most common content ratings and their respective shares of the catalog. The distribution is adult leaning TV-MA being the single largest bucket (36.7%), followed by TV-14 (24.7%). Together, they account for 61% of all titles, indicating a strong emphasis on mature and older-teen audiences. Mid-tier guidance categories such as TV-PG (9.9%) and theatrical R (9.1%) form a secondary cluster, while family, younger-audience ratings PG-13 (5.6%), TV-Y7 (3.8%), TV-Y (3.5%), PG (3.3%), and TV-G (2.5%) are comparatively smaller. NR (0.9%) is rare, suggesting most titles carry a standardized rating. Higher percentages indicate a larger share of the catalog under that rating. Comparing adjacent rows quickly reveals the platform's audience targeting, such as the dominance of TV-MA and TV-14 over TV-PG and TV-G, which highlights a catalog skew toward mature series and films.

rating	count	percentage
TV-MA	3207	36.7
TV-14	2160	24.7
TV-PG	863	9.9
R	799	9.1
PG-13	490	5.6
TV-Y7	334	3.8
TV-Y	307	3.5
PG	287	3.3
TV-G	220	2.5
NR	80	0.9

Table 2: Rating distribution (Top 10)

Table 3: Genre distribution (Top 20)

This table shows the most common genres in the catalog and their share of titles. The mix is international- and drama-heavy: International Movies (16.3%) and Dramas (14.4%) lead, with Comedies (9.9%) and International TV Shows (8.0%) next. Together, the top four account for 48.6% of all titles. Mid-tier categories include Documentaries and Action & Adventure, each 5.1%, TV Dramas and Independent Movies (each 4.5%). The rest lsuch as, Children & Family Movies (3.8%), Romantic Movies (3.6%), TV Comedies and Thrillers 3.4% each, and

niche segments such as Docuseries (2.3%), Music & Musicals (2.2%), Horror Movies (2.1%), Stand-Up Comedy (2.0%), and Reality TV (1.5%). Higher percentages indicate genres that dominate Netflix's listings. While many smaller niches contribute depth, the concentration at the top points to a strategy focused on international programming and story-driven drama and comedy.

genre	count	percentage
International Movies	2752	16.3
Dramas	2427	14.4
Comedies	1674	9.9
International TV Shows	1351	8
Documentaries	869	5.1
Action & Adventure	859	5.1
TV Dramas	763	4.5
Independent Movies	756	4.5
Children & Family Movies	641	3.8
Romantic Movies	616	3.6
TV Comedies	581	3.4
Thrillers	577	3.4
Crime TV Shows	470	2.8
Kids' TV	451	2.7
Docuseries	395	2.3
Music & Musicals	375	2.2
Romantic TV Shows	370	2.2
Horror Movies	357	2.1

Stand-Up Comedy	343	2
Reality TV	255	1.5

Table 3: Genre distribution (Top 20)

"datastat.py" py file contains the necessary PySpark code for data statistics

3. Work Distribution:

3.1 Ishrat Jaben Bushra: Used PySpark for questions 1, 2,5, data description, solution description, and presentation development, problem definition.

3.2 Sahil Arora: Used Hive for questions 1,4, future work, problem definition, and presentation development.

3.3 Sam Ensafi: Used PySpark for question 3, used Elastic and Kibana for question 1 to 5 visualizations, problem definition, and presentation development

4. Solution Description

4.1. HDFS (Hadoop Distributed File System)

4.1.1. How was it used

HDFS worked at the Data Storage Level and was used to store the dataset, providing a robust and distributed storage solution that ensured data availability and reliability. It served as the backbone of this project, as Hive and Spark rely heavily on it.

4.1.2. Why it was chosen for that particular problem space

As it is used to store our netflix_titles.csv dataset, it may not be useful for answering any question individually. Still, it is essential to access all the tools that we worked with, as we need the netflix_titles.csv file for our entire project.

4.1.3. Code snippets and explaining the logic behind the code snippets

The file netflix_titles.csv was added to HDFS using the **hadoop fs -put** command. Then, it was used in later tasks. Then, in the hadoop **mkdir -p** command was used to save format CSV results in a directory, **hadoop fs -get** command was used to copy the results as a CSV format from HDFS to the local, which we later used in Elastic Search and Kibana for visualization.

4.2. Apache Spark

4.2.1. How was it used

Apache Spark worked at the Data Processing Level for data cleaning, providing data statistics, and then it was used to address a part of Q1, and Q2, Q3, Q5 analysis. The workflow of how Spark is used is as follows:

Q1 analysis: (How does Netflix balance short- and long-form content,does its catalog lean more toward TV shows or films?)

We answered “How does Netflix balance short- and long-form content” with pyspark.

Q2 analysis:(Which genres are most prevalent in Netflix’s catalog, and how have their shares changed over time?)

Q3 analysis:(Which countries contribute the most titles to Netflix’s catalog?)

Q4 analysis:(How is Netflix’s content distributed across different maturity ratings?)

Q5 analysis:(Are Netflix movies getting longer over time by genre and year? Are Netflix TV series trending toward multi-season runs or limited series?)

4.2.2. Why it was chosen for that particular problem space

Spark can handle large-scale data processing and analysis efficiently. Although the dataset is fairly small and manageable, in case the scalability of the project is increased, Spark can still handle it well. Therefore, Spark is the best choice for fast processing of big data.

4.2.3. Code snippets and explaining the logic behind the code snippets

Q1 analysis: (How does Netflix balance short- and long-form content,does its catalog lean more toward TV shows or films?)

1. Ingestion & Schema: Explicit schema **StructType** with String , Integer fields is defined to prevent type drift and unstable auto-inference on multiline text.CSVfile read configured with **.option("multiLine","true")**, proper quote and escape handling, and **PERMISSIVE** mode so commas or newlines inside descriptions do not break during parsing.

2. Data Cleaning: nz(col) trims whitespace and converts empty strings to **NULL**, to avoid “empty but not null” issues in counts and joins. Normalization is applied to all the string columns so downstream aggregations treat blanks repeatedly .**release_year** is cast to an integer to enable durable numeric operations.

3. Quality checks: Total rows are computed using **df.count()**. Exact duplicates computed as **df.count() - df.dropDuplicates().count()**.

4. Scope Filter: Records restricted to type Movie, TV Show(**df_q1**). Duration parsing logic assumes minutes for movies and seasons for TV shows.

5. Duration Imputation: Frequency table is built with **groupBy(type, duration).count()** on non-null durations.Per type maximum frequency is identified and joined back to obtain **mode_duration**, imputing with the most common duration string preserves unit consistency within each type.**mode_duration** is left-joined by **type**, null **duration** values are filled with the per type mode. Imputation impact is documented as: missing before, missing after, and total filled.

6. Numeric Duration Extraction:duration_num created from the cleaned **duration**.“min” extracts digits as **minutes**, “Season” extracts digits as **seasons**, otherwise **NULL**.

7.Summary Statistics: Mean by type computed as **avg(duration_num)** is rounded and collected for Movies vs TV.Median by type computed with **stat.approxQuantile("duration_num", [0.5], 0.01)** on Movies and TV separately. Medians reduce the influence of outliers, such as very long films or multi-season packages.

#Note: Q1 PY file is named as "**q1_test_project.py"**

Q2 analysis:(Which genres are most prevalent in Netflix's catalog, and how have their shares changed over time?)

1. Ingestion & Schema: Explicit schema StructType with String, Integer fields to prevent type drift on multiline text. CSV file read uses **.option("multiLine","true")**, proper quoting and escaping, and **PERMISSIVE** mode so commas, newlines in descriptions don't break during parsing.

2. Data Cleaning: nz(col) trims whitespace and transforms empty strings to NULL to avoid "empty but not null" issues. Applied to all string columns. **release_year** cast to integer.

3. Quality checks: duplicate **show_id** is assessed.

4. Date parsing (year of addition): date_added is parsed against multiple patterns,year_added and month_added are derived. Rows with unparseable or NULL **date_added** are explicitly dropped for time-series calculations to keep yearly trends consistent.

5. Genre normalization & weighting: listed_in is split and exploded to one row per (title × genre). Equal-share weighting **1/k** for titles with **k** genres, so each title contributes a total of 1 across its tags.

6.Yearly genre prevalence & share: For each **year_added, genre ,**weighted counts summed as **approx_title_count**. Yearly denominators **titles_in_year** from distinct **show_id. Share = approx_title_count / titles_in_year.** Rank within each year keeps the Top 10.

#Note: Q2 and Q5 are done in the same PY file, and it is named as "**q2_and_q5_test_project.py"**

Q3 analysis :(Which countries contribute the most titles to Netflix's catalog?)

1. Ingestion & Schema: Used explicit schema **StructType** with **String**, **Integer** fields to prevent type drift on multiline text. CSV file read is configured with **.option("multiLine","true")**, proper quote escape, and **PERMISSIVE** mode so commas, newlines in descriptions don't break during parsing.

2. Data cleaning: All string columns are trimmed, empty strings are replaced with **NULL** by **clean_str(...)** to avoid "empty but not null" issues during aggregations.

3. Country null handling: country replaced with"**Unknown**" wherever the field is **NULL**, so country coverage is exhaustive, and rows are retained rather than dropped.

4. Scope filter: Records are limited to type Movie, TV Show, to keep counts aligned with the project's movie or TV show focus.

5. Multi-country normalization: country is split on commas and **exploded** to one row per **(title × country)**, so

co-productions are counted for each listed country.

6. Overall country contribution: Grouped by country and counted as **total_titles**, then it's sorted in descending order to identify the largest contributing countries in the catalog.

7. Country × type breakdown: Grouped by country and type) and counted as **total_titles** to show each country's distribution across Movies vs TV Shows.

8. Row drops: No rows are dropped for a missing country. Only the **type** filter is applied, the parser runs in **PERMISSIVE** mode to keep malformed lines where possible.

#Note: Q3 PY file is named as "**q3.py"**

Q5 analysis :(Duration trends: movies by minutes; TV shows by seasons)

1. Duration imputation: Per-type mode of the duration string used to fill nulls. The units are: minutes for Movies, seasons for TV.

2.Numeric duration extraction: duration_num by regex on **duration_filled**. "min" is minutes,"Season(s)" is season count, otherwise NULL.

3.Movie duration trends (by year × genre): Calculations used: titles = countDistinct(show_id),mean_minutes = avg**(duration_num) (rounded),median_minutes = percentile_approx(duration_num, 0.5)**,Medians emphasized.

4. TV duration trends (by year × genre): Calculations used: **titles = countDistinct(show_id),mean_seasons, median_seasons via percentile_approx**,Medians indicate shift toward limited series 1 season vs multi-season runs from 3 to 4 .

5. Interpretation: Medians tell the typical viewer experience, flag and filter bins with very small titles, remember trends reflect catalog additions supply timing rather than content production year.For duration, we don't 1/k-weight,each tagged genre inherits the title's full **duration_num**.

#Note: Q2 and Q5 are done in the same PY file, and it is named as "**q2_and_q5_test_project.py"**

4.3. Hive

4.3.1. How was it used

Hive was used at the Data Querying & Analysis Level. After the dataset was cleaned and stored in HDFS, Hive was used to run analytical SQL queries on top of structured tables. Specifically, Hive helped in answering Question 1 and Question 4 efficiently by running aggregation queries, filtering logic, groupings, and data transformations directly on the dataset.

It also allowed us to create managed external tables, load the dataset from HDFS, and convert CSV data into a queryable tabular format. Hive was used to extract patterns and relationships such as movie vs TV show proportions (Q1) and maturity ratings trends and insights (Q4).

4.3.2. Why it was chosen for that particular problem space

Hive was ideal for this problem space because:

Reason	Explanation
SQL-like interface	Hive lets us query big data using simple SQL syntax (SELECT, GROUP BY, WHERE). No complex coding required like PySpark.
Works directly with HDFS	Since our data was already in HDFS, Hive could directly read from it without additional data movement.
Fast aggregations on large data	For questions like “how many movies vs TV shows?” or “rating distribution?”, Hive performs optimized aggregation via MapReduce execution.
Schema-on-read approach	We could define table structure while reading CSV files, no need to modify raw data.
Easy output export	Results generated in Hive were exported to CSV and later used in Elastic & Kibana for visualization.

Therefore, Hive was the most suitable tool for structured analysis & query-based insights, especially for Q1 & Q4, where we needed quick answers through SQL-like logic.

4.3.3. Code snippets and explaining the logic behind the code snippets

Question 1 queries:

“CREATE TABLE netflix (show_id STRING, type STRING, title STRING, country STRING,date_added STRING, rating STRING, duration STRING, listed_in STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE

TBLPROPERTIES ("skip.header.line.count"="1");

LOAD DATA INPATH '/project/netflix_titles.csv' INTO TABLE netflix;”

Purpose:

1. Creates the Hive table and loads data directly from HDFS.
2. **skip.header.line.count** is used to ignore the first row containing column names.

“SELECT type, COUNT(*) AS total_titles

FROM netflix

GROUP BY type;”

Purpose:

1. This query helps answer Q1 by calculating how many **Movies vs TV Shows** are present.
2. Used **GROUP BY** aggregation to find the overall content proportion.

Question 4 queries:

"SELECT country, rating, COUNT(*) AS total FROM netflix

WHERE country IS NOT NULL GROUP BY country, rating

ORDER BY total DESC;"

Purpose:

1. Shows which countries contribute to rated content the most.

4.4. Elastic Search And Kibana

4.4.1. How was it used

ElasticSearch was used to index the cleaned output CSV files generated from Spark and Hive, so the data could be searched, filtered, and aggregated quickly. After uploading each CSV into the Index Management section, custom field mappings were applied to ensure the fields were read correctly. Data views were then created in Kibana for each index. Using these indexed datasets, Kibana was used to build visual dashboards for all five questions. Different chart types such as bar, pie, and line graphs were created by selecting fields for horizontal axis, vertical axis, and breakdown, and applying functions like Count and Average to show trends and comparisons clearly.

4.4.2. Why it was chosen for that particular problem space

ElasticSearch delivers high performance for fast searching and aggregating data, and Kibana provides an intuitive interface for building visualizations without requiring programming. They were also selected because they integrate smoothly within the same ecosystem. This allows an efficient connection between indexed data and interactive dashboards.

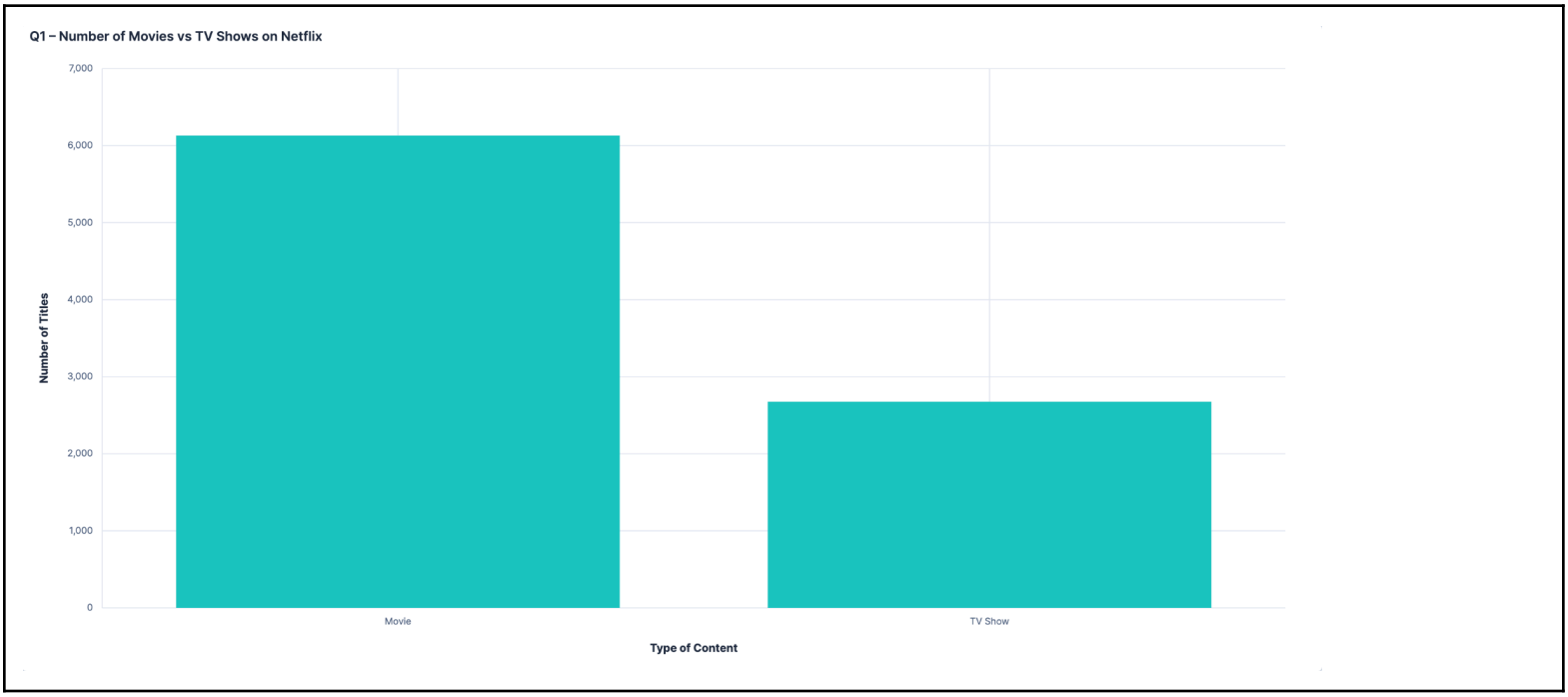
4.4.3. Code snippets and explaining the logic behind the code snippets

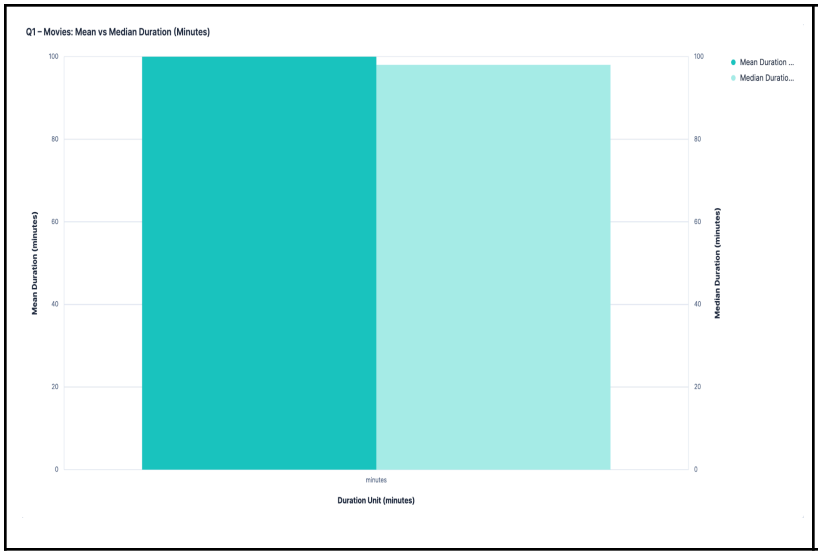
The processed CSV files (type counts, genre trends, country contribution, maturity distribution, duration trends, etc.) were uploaded into ElasticSearch as separate indices. While uploading, mappings were assigned to define field types such as text, keyword, or numeric. After indexing, a Data View was created in Kibana, and charts were built by selecting the indexed fields and applying functions such as Count for total titles or Average for duration trends. Additional options like Rank direction or limit on number of values to display were set to control chart clarity.

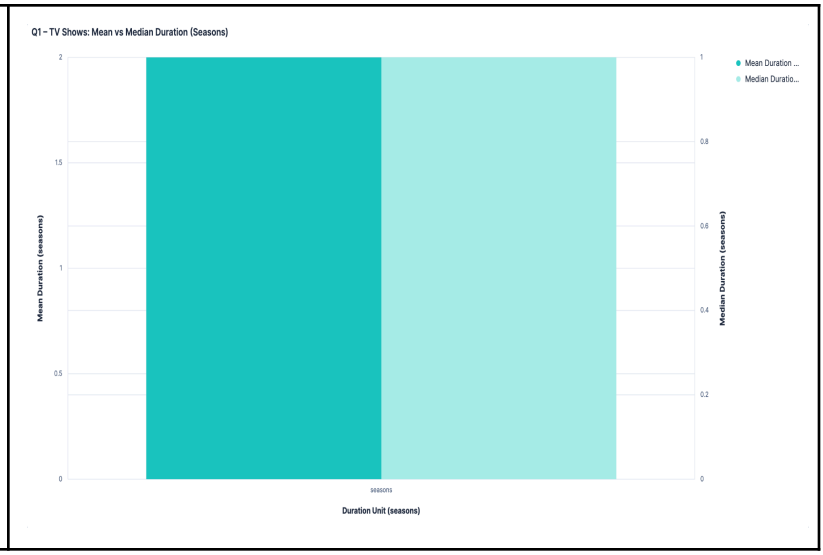
5. Insights gathered from data

5.1 Q1 insight:

The analysis shows a clear imbalance in Netflix's catalog, with around 6100 movies compared to only 2700 TV shows, meaning movies outnumber series by more than double. Duration trends confirm this imbalance, since the average movie is about 100 minutes long while half of all TV shows end after only one season, indicating a strong focus on short form content rather than long running series. This suggests that Netflix emphasizes content that is quick to produce and consume instead of multi season shows that build ongoing viewer commitment. Since longer shows are a major reason users maintain subscriptions, this imbalance could lead to reduced loyalty and increased cancellations. Increasing investment in multi season TV series and expanding the volume of long running shows could strengthen viewer retention and improve long term engagement.

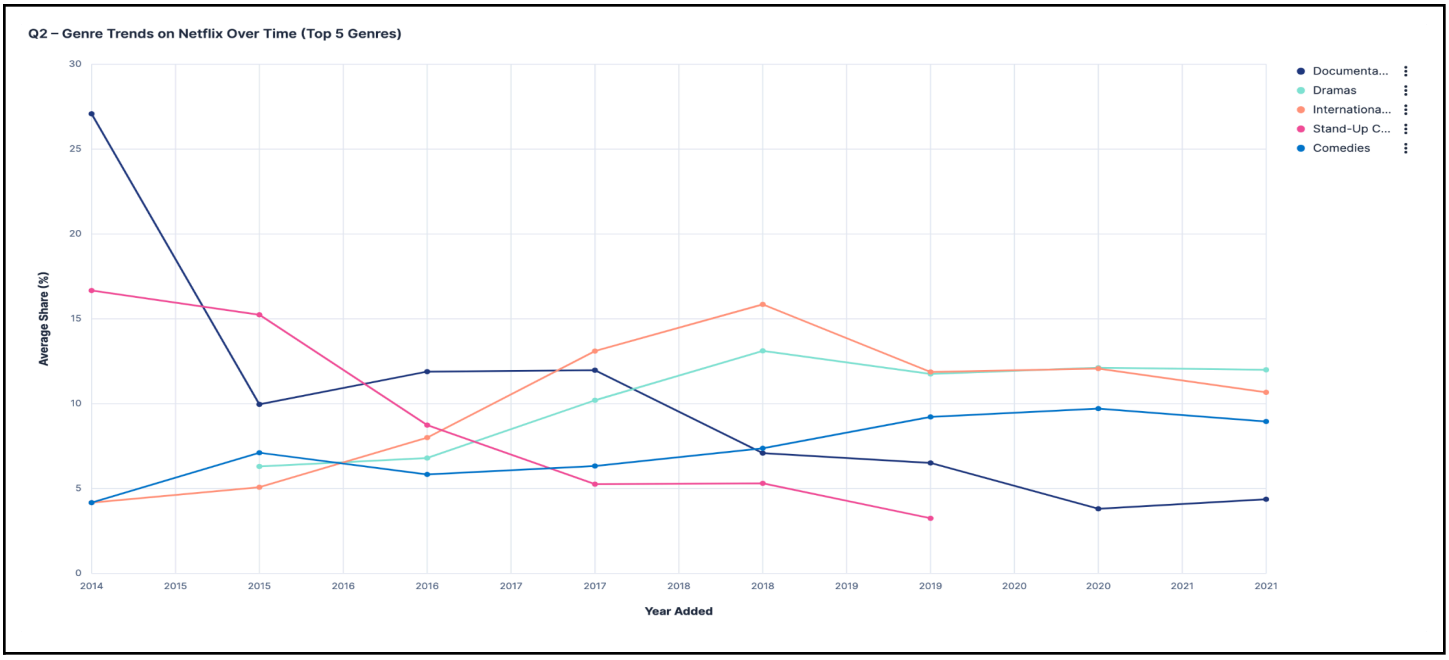


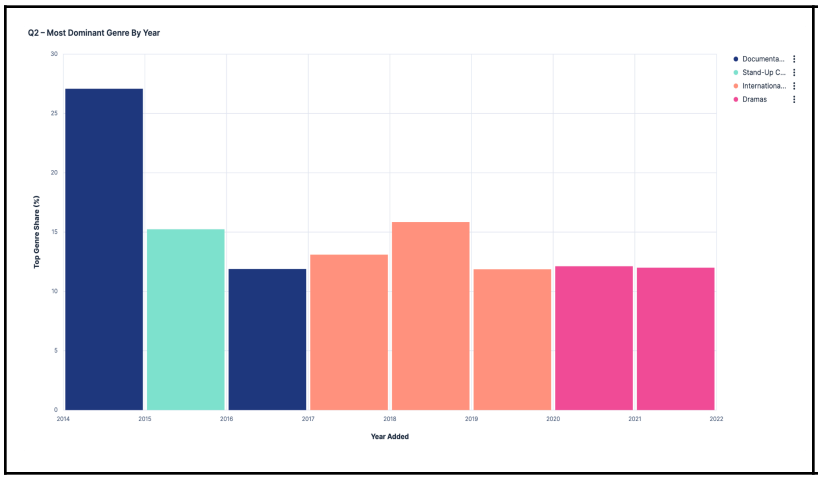


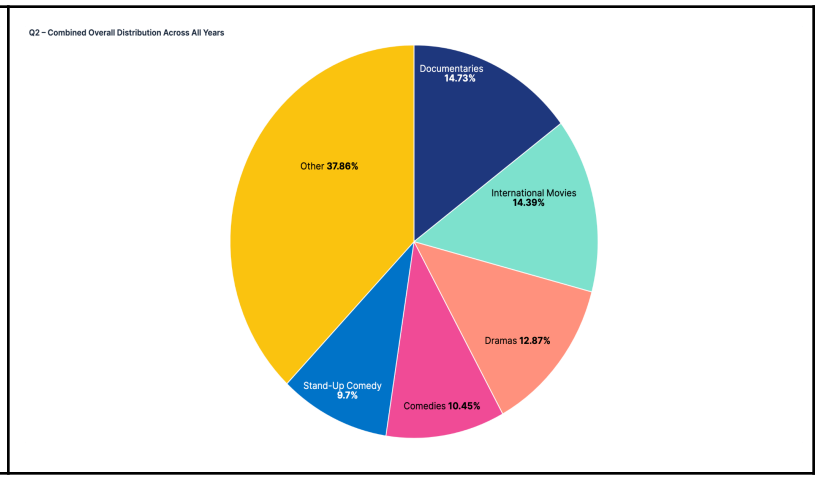


5.2 Q2 insight:

The genre analysis shows that Netflix's catalog has shifted toward a small group of dominant genres while reducing diversity over time. Documentaries and Stand Up Comedy were leading categories early but declined sharply, while Dramas and International content grew and became the most dominant genres from 2018 onward. The overall distribution confirms that a few genres take most of the catalog share, and all remaining non dominant genres together make up only about 38 percent, suggesting Netflix is focusing on trend driven content instead of maintaining balance. The analysis also shows that strong genres were reduced in favor of short term trends, leading to noticeable drops in their share. Although expanding international content supports global growth, reducing proven genres risks viewer dissatisfaction. A more balanced strategy that brings back strong categories and mixes trend based content with consistent long term options could help maintain wider audience interest and improve retention.

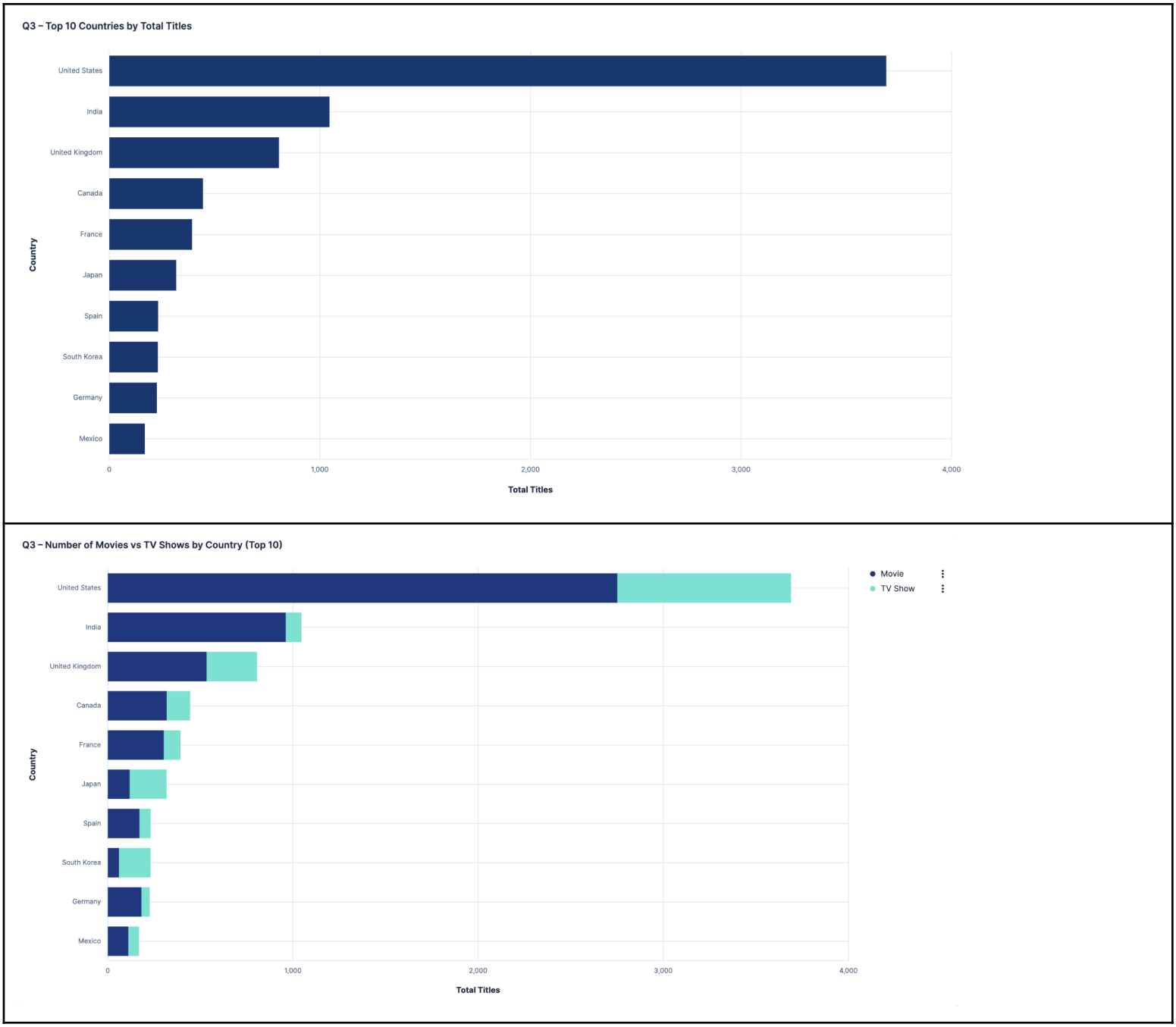






5.3 Q3 insight:

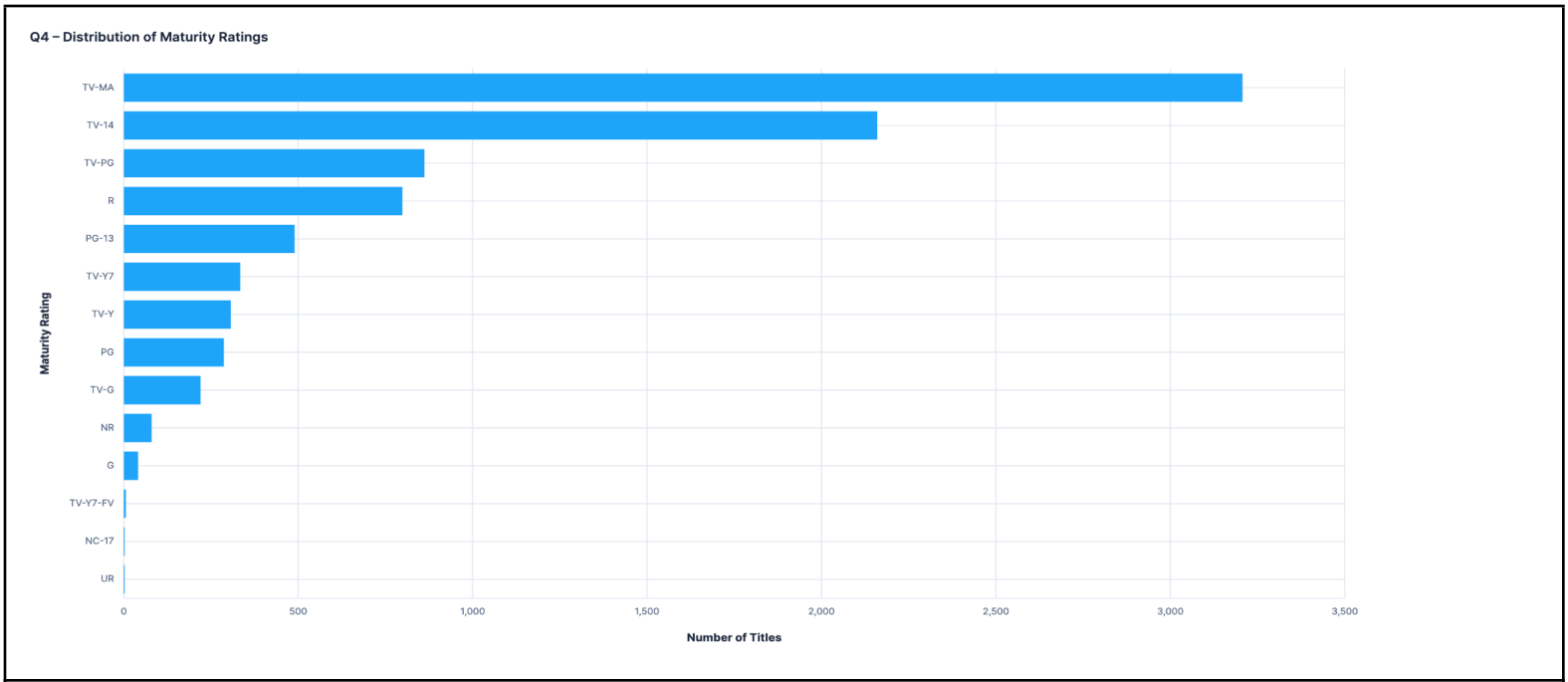
The analysis of country contributions shows that Netflix's catalog is heavily dominated by the United States, which has around 3800 titles compared to India with about 1050 and the United Kingdom with about 800, while all other countries drop far below these levels. The breakdown of movies and TV shows by country confirms the same pattern, with the United States leading in both categories by a large gap and most countries offering many more movies than TV shows. This indicates strong over reliance on U.S. produced content and limited international diversity, which may reduce appeal for global audiences who prefer more local options. Increasing investment in international productions and partnerships with regional creators could help balance the catalog, expand cultural variety, and improve global viewer satisfaction.

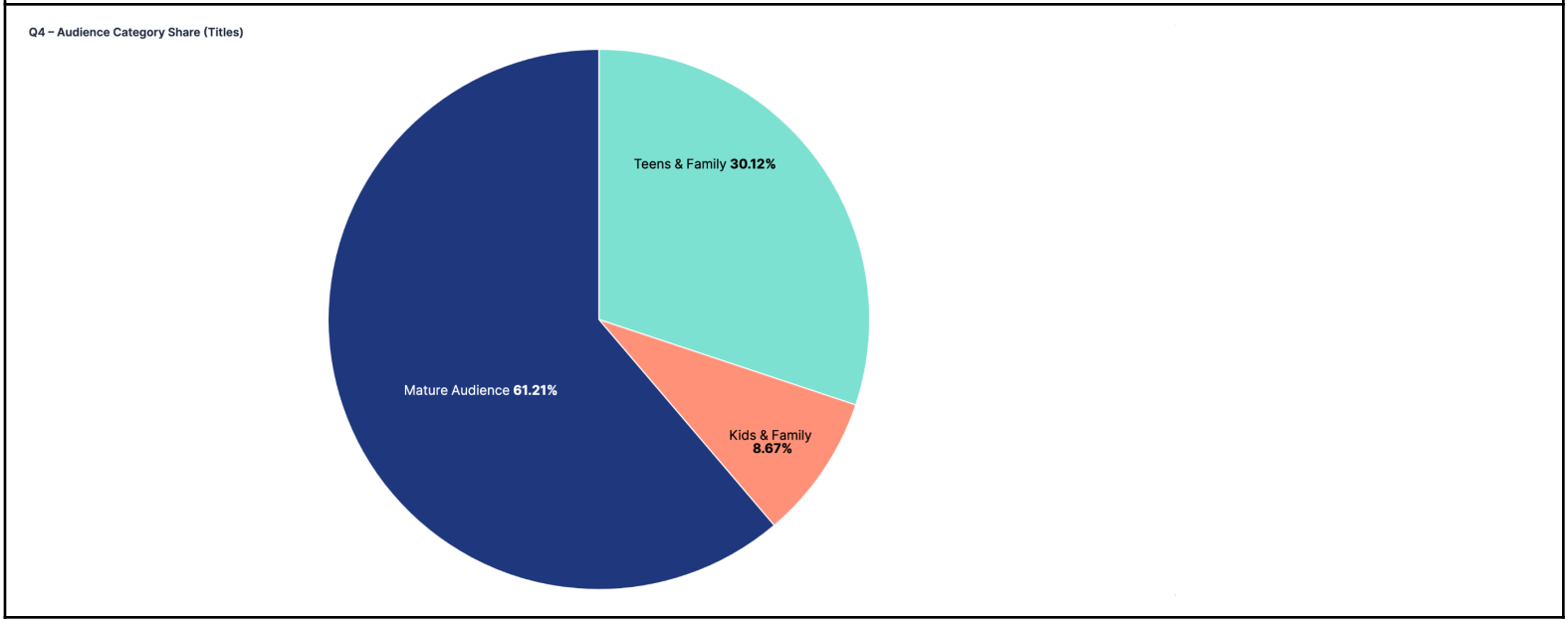


5.4 Q4 insight:

The analysis of maturity ratings shows that Netflix's catalog is strongly dominated by adult focused content, with TV MA having the highest count at more than 3200 titles and TV 14 also very high, while children and family friendly ratings such as TV Y, TV Y7, TV G, and PG each have only a small number of titles. The audience share chart confirms that more than 60 percent of all content targets adults, while less than 10 percent is designed for young children. This significant imbalance indicates that Netflix prioritizes mature audiences and may be overlooking family viewers, who often make subscription decisions. To maintain competitiveness against platforms that specialize in children's programming, Netflix may benefit from expanding investment in

high quality family and children content to achieve a more balanced maturity distribution and better serve a wider audience.

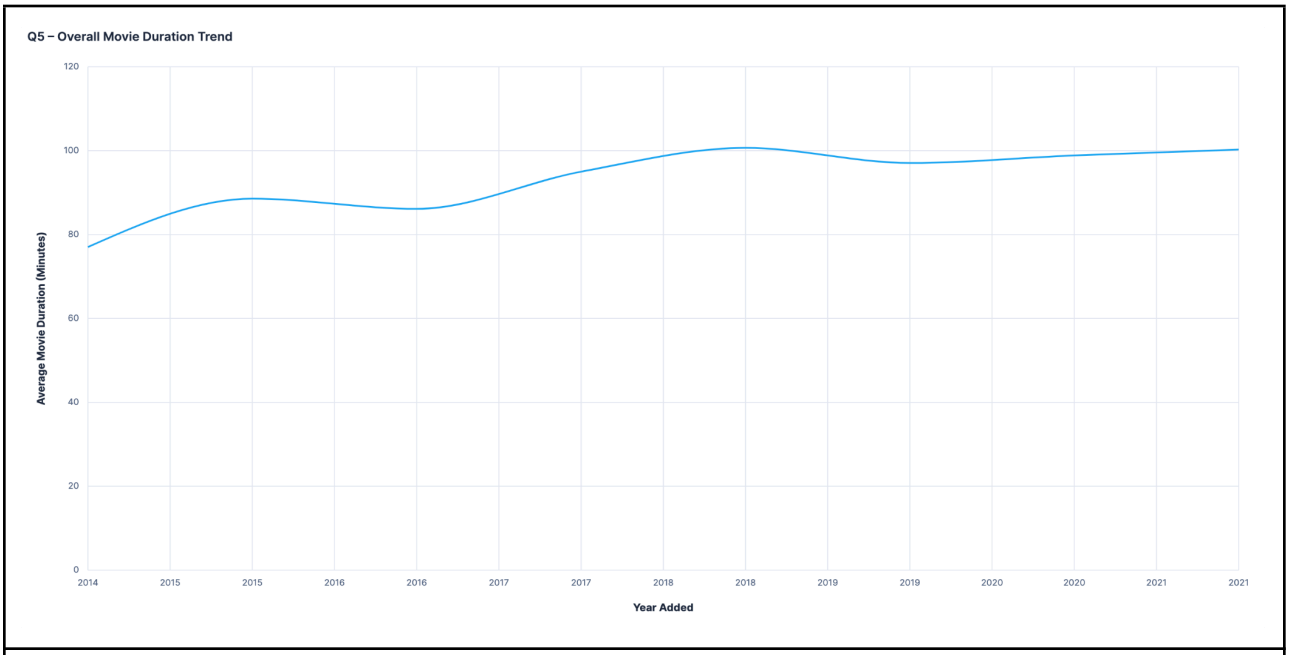


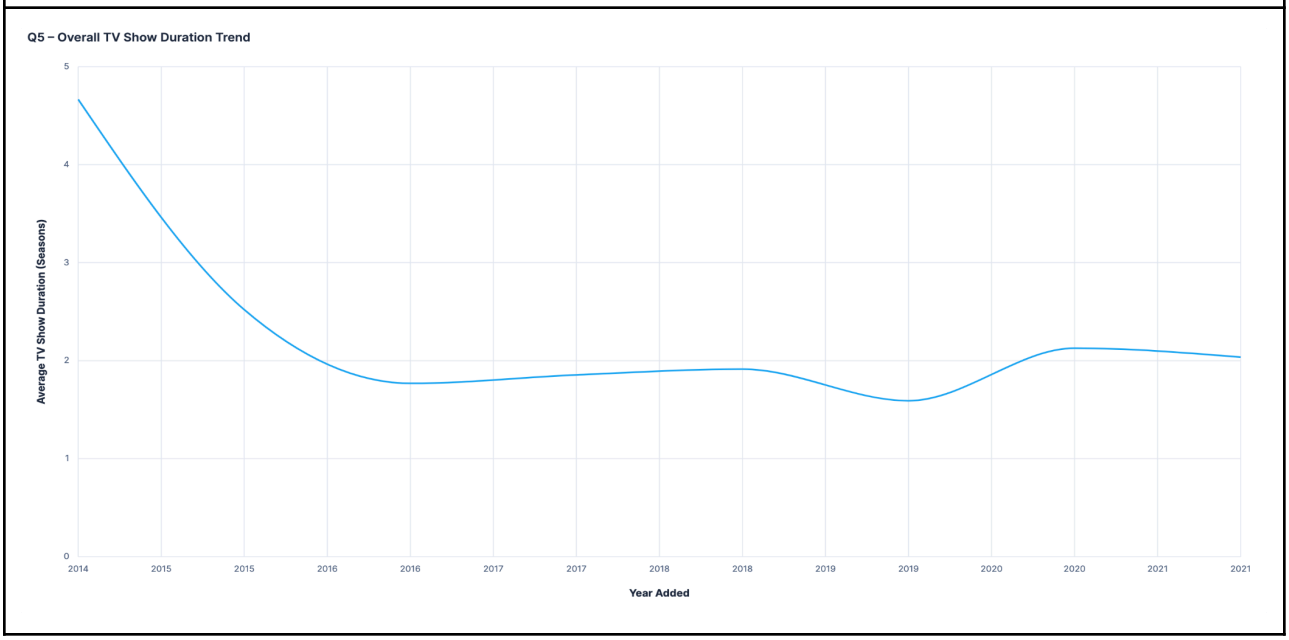


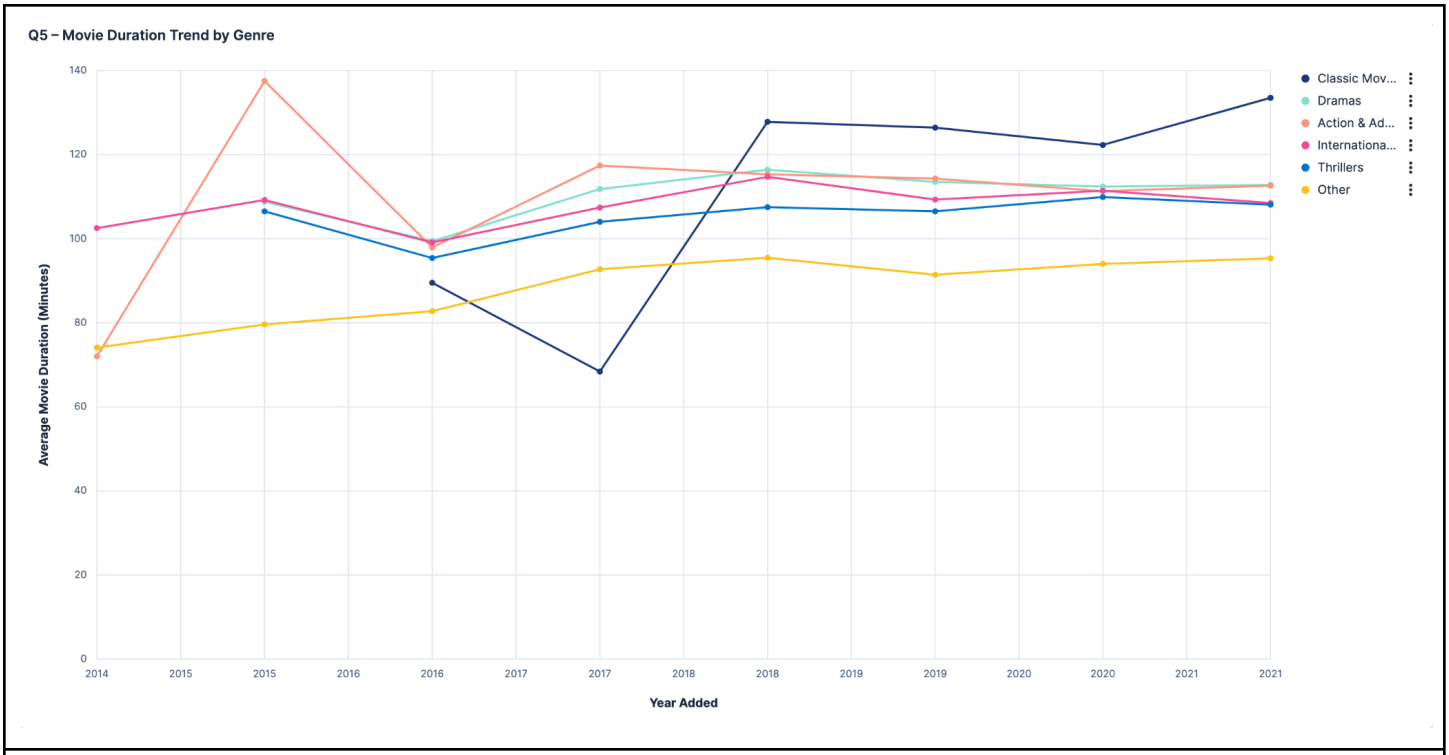
5.5 Q5 insight:

The duration analysis shows a clear shift in Netflix's content structure, with average movie length increasing from about 77 minutes in 2014 to around 100 minutes by 2021, and many major genres, such as Action and Adventure, Dramas, and International Films, now reaching more than 110 minutes. At the same time, TV shows have become much shorter, with the average number of seasons dropping from about 4.6 in 2014 to about 2 by 2021, and this decline appears across many genres, including TV Comedies and TV Mysteries. These trends indicate that Netflix is moving toward longer movies and shorter series, which may reduce viewer engagement since long movies require more commitment and short series do not build long term attachment.

Balancing movie lengths and increasing investment in multi season shows could help strengthen audience loyalty and improve subscriber retention.







5.6 Overall summary of insights:

Overall, the insights across all five questions reveal a consistent pattern in Netflix's content strategy, showing strong imbalance across multiple dimensions. The catalog is dominated by movies rather than TV shows, genres are concentrated around a few trend driven categories, the majority of titles originate from the United States, most content targets mature audiences, and movie durations are increasing while TV series are becoming shorter. Together, these findings suggest Netflix is prioritizing fast production and trend based

content instead of maintaining long term balance, diversity, and viewer retention stability. These strategic choices may risk subscriber loss, reduced satisfaction, and weaker long term engagement, especially among international viewers, families, and fans of multi season series. Strengthening catalog balance, expanding international and family focused options, and reinvesting in multi season shows could help Netflix improve loyalty and sustain competitive advantage.

6. Future Work

1. Use Hive partitions and buckets to accelerate queries on large datasets.
2. Implement Spark Streaming to simulate a real-time pipeline that processes new Netflix titles.
3. Implement Elastic ingestion pipelines with processors to automatically clean and transform data before indexing.
4. Automate this pipeline using Airflow so the entire process,ingestion, cleaning, statistics, and visualization,runs on a schedule

7. References

Barnes, J. (2019, November 19). *Netflix users are cancelling due to price increases & lack of content, survey shows*. *Cord Cutters News*.
https://cordcuttersnews.com/netflix-users-are-cancelling-due-to-price-increases-lack-of-content-survey-shows

BB Media. (2024, December 6). *The role of kids' content in subscription choices: BB Media analysis*. *Broadcast Pro ME*.
https://www.broadcastprome.com/news/the-role-of-kids-content-in-subscription-choices-bb-media-analysis

Bansal, S. (n.d.). *Netflix movies and TV shows* [Dataset]. Kaggle. https://www.kaggle.com/datasets/shivamb/netflix-shows

Business of Apps. (2025). *Netflix revenue and usage statistics (2025)*.
https://www.businessofapps.com/data/netflix-statistics

Fletcher, B. (2023, December 19). *What's behind Netflix's engagement report data dump? StreamTV Insider*.
https://www.streamtvinsider.com/video/whats-behind-netflixs-engagement-report-data-dump

Lotz, A. D., Eklund, O., & Soroka, S. (2022). Netflix, library analysis, and globalization: Rethinking mass media flows. *Journal of Communication, 72*(4), 511 to 521. https://doi.org/10.1093/joc/jqac020

NewscastStudio. (2025, June 27). *Streaming platforms abandon binge model for hybrid release strategies*.
https://www.newscaststudio.com/2025/06/27/streaming-platforms-abandon-binge-model-for-hybrid-release-strategies

Omdia. (2025, April 4). *Key insights from Netflix's 2H24 "What We Watched" report: Shorter series and seasonal swings*.
https://omdia.tech.informa.com/om129546/key-insights-from-netflixs-2h24-what-we-watched-shorter-series-and-seasonal-swings

Parents Television & Media Council. (2023). *Families need not subscribe: OTT report 2023*.
https://www.parentstv.org/resources/Families-Need-Not-Subscribe-2023-OTT-Report-Final.pdf

Roth, E. (2022, May 18). *Survey shows Netflix is losing more long-term subscribers*. *The Verge*.
https://www.theverge.com/2022/5/18/23125424/netflix-losing-long-term-subscribers-streaming

Vu, L. (2022). *Netflix data analysis: Content preferences on Netflix—Movies vs TV shows*. *Medium*. https://medium.com/@linhvu.nt/data-analysis-and-recommendations-on-netflix-content-28707163553

Zahmadi, I. (2025, May). *Exploring Netflix content trends by genre, rating, and format: What the catalog says*. *Medium*. https://medium.com/@zahmadi_76657/exploring-netflix-content-trends-by-genre-rating-and-format-what-the-catalog-says-b66353868ca0