**Netflix’s Project**

This project is an analysis of Netflix's vast library of movies and TV shows, as well as the company's pricing strategies across different countries. As the leading global streaming platform, Netflix offers a wide range of content, including original productions and licensed titles. The project delves into Netflix’s content database, examining key information such as show titles, cast, genre, and the release year. Additionally, it explores Netflix's subscription pricing in various markets, comparing the number of available shows and movies with the costs associated with different subscription tiers. The study aims to understand the correlation between Netflix’s content offerings, pricing strategies, and the platform's success in different regions.

**Objective:**

**Objective:**

The objective of this dashboard and report is to analyze Netflix's content performance across various categories and regions. Key goals include identifying which category receives the highest ratings, determining the number of TV shows and movies available in each country, and examining the budget and international sales of top-grossing movies. This analysis will provide actionable insights into content preferences, market trends, and financial success across different regions, aiding Netflix in optimizing its content strategy and resource allocation.

**"General insights on the dataset of the project"**

**Netflix Titles (Table):**

A list of movies and TV shows available for streaming on Netflix, including original content and licensed titles. This column typically includes the title, release year, genre, and sometimes a brief synopsis or other relevant details about the content as following:

1. **show\_id**: A unique identifier for each show (either TV show or movie).
2. **type**: Indicates whether the entry is a "Movie" or a "TV Show".
3. **title**: The name of the movie or TV show.
4. **director**: The director(s) of the movie or TV show.
5. **cast**: The main actors or cast members in the show or movie.
6. **country**: The country where the movie or TV show was produced.
7. **date\_added**: The date when the show or movie was added to Netflix.
8. **release year**: The year when the show or movie was originally released.
9. **rating**: The age rating of the movie or TV show (e.g., TV-MA, PG-13, etc.).
10. **duration**: The number of seasons for TV shows or the runtime for movies (usually displayed as "1 Season" for TV shows or "90 min" for movies).
11. **listed in**: The genres or categories the show or movie belongs to (e.g., Drama, Comedy, Docuseries).
12. **description**: A summary or description of the movie or TV show.

**Netflix Price in Different Countries (Table):**

The subscription cost for accessing Netflix services in various countries. This column reflects the pricing tiers available (e.g., Basic, Standard, Premium) and may vary based on local currency and regional pricing strategies, offering insights into how Netflix adapts its pricing for different markets as following:

1. **Country**: The name of the country where Netflix is available.
2. **Total Library Size**: The total number of titles (TV shows and movies) available in that country.
3. **No. of TV Shows**: The total number of TV shows available in the Netflix library for that country.
4. **No. of Movies**: The total number of movies available in the Netflix library for that country.
5. **Cost Per Month - Basic ($)**: The monthly subscription cost for the Basic plan in that country, in U.S. dollars.
6. **Cost Per Month - Standard ($)**: The monthly subscription cost for the Standard plan in that country, in U.S. dollars.
7. **Cost Per Month - Premium ($)**: The monthly subscription cost for the Premium plan in that country, in U.S. dollars.

**Hollywood's Highest-Grossing Movies (Table):**

A list of films produced by Hollywood that have achieved the highest total box office revenue, both domestically and internationally. This category typically includes titles that have made significant financial impacts at the box office, showcasing trends in audience preferences, production budgets, and marketing strategies as following:

1. **Title**: The name of the movie.
2. **Movie Info**: A summary or synopsis of the film's storyline, outlining the main plot points, characters, and key themes. This provides an overview of the film's narrative, helping viewers understand the central conflict and the journey of the characters.
3. **Year**: The year in which the movie was released.
4. **Distributor**: The company responsible for distributing the film to theaters and other platforms.
5. **Budget (in $)**: The total financial investment made in producing the movie, often covering production costs, marketing, and other expenses.
6. **Domestic Opening (in $)**: The revenue generated by the movie in its opening weekend in the domestic market (typically the U.S. and Canada).
7. **Domestic Sales (in $)**: The total revenue generated by the movie in the domestic market throughout its theatrical run.
8. **International Sales (in $)**: The revenue earned by the movie from international markets outside the domestic territory.
9. **Worldwide Sales (in $)**: The total revenue generated by the movie from both domestic and international markets combined.
10. **Release Date**: The specific date on which the movie was released in theaters.
11. **Genre**: The category or categories that describe the movie's style or themes (e.g., Action, Comedy, Drama).

**The age rating contains:**

**Film Ratings:**

* **G (General Audience)**: Suitable for all ages; contains no content that would offend parents if viewed by children.
* **PG (Parental Guidance)**: Contains some material that may not be suitable for children; parents are advised to provide supervision.
* **PG-13 (Parents Strongly Cautioned)**: Some material may be inappropriate for children under 13; parents are strongly cautioned to provide guidance.
* **R (Restricted)**: Contains adult material; parents are strongly advised to provide proper supervision for children under 17.
* **NC-17 (No One 17 and Under Admitted)**: Contains explicit content; not suitable for viewers under 17.
* **NR (Not Rated)**: The content has not been submitted for a rating.

**Television Ratings:**

* **TV-Y (All Children)**: Suitable for all children; contains no inappropriate content.
* **TV-Y7 (Directed to Older Children)**: Suitable for children age 7 and older; may contain mild themes or humor.
* **TV-Y7-FV (Directed to Older Children - Fantasy Violence)**: Similar to TV-Y7 but may include mild fantasy violence.
* **TV-G (General Audience)**: Suitable for all ages; contains no inappropriate content.
* **TV-PG (Parental Guidance Suggested)**: May contain some material unsuitable for young children; parental guidance is suggested.
* **TV-14 (Parents Strongly Cautioned)**: May not be suitable for children under 14; parents are strongly cautioned to provide guidance.
* **TV-MA (Mature Audiences)**: Contains content specifically designed for mature audiences; may not be suitable for viewers under 17.

**"Optimizing Netflix Dataset for Analysis: A Comprehensive Data Cleaning and Transformation Approach"**

**"Optimizing Netflix Dataset for Analysis: A Comprehensive Data Cleaning and Transformation Approach"**

**Data Cleaning Process:**

* **Handling Missing Values:**1- In the first table (Netflix Titles): some columns were missing names, so I assigned appropriate names to them. There were also some empty rows, but since they contained valid data, I decided to keep them for future use Instead of deleting columns with missing values, I identified that the "Director" column contains some empty fields that could still provide useful insights when visualized. To retain these records, I replaced the missing values with the placeholder text "Director Missing" to ensure data consistency while allowing for analysis of such cases.  
    
  2- In the second table (Highest Hollywood Grossing Movies): I adjusted the data types for certain columns. For example, the Budget column was originally in text format, so I converted it to a decimal type. As a result of this conversion, some errors appeared, which I addressed by removing the problematic rows
* **Converting Data Types**:

The "Released Date" column was initially stored as text, which could hinder accurate date-based analysis. I converted the data type from text to date format, enabling proper time-based visualizations and calculations.

* **Standardizing Duration**:

Upon reviewing the "Duration" column, I found that it contained different formats (e.g., minutes for movies and seasons for TV shows). To standardize this for further analysis, I created a conditional column that separates the values based on keywords like "min" (for movies) and "season" (for TV shows). The new column captures the numeric part of the duration, converting it to a whole number for easier aggregation and calculation.

**Adding Tables & columns:**

* **(Casting for Each show & Category for Each show):**

I developed two tables to facilitate my analysis. The first table provides a detailed of the casting for each show, derived from the Netflix dataset. Given that the cast names were separated by commas, I employed a split function to separate them into individual rows. In the second table, I applied a similar methodology to categorize each show, ensuring that all categories are clearly represented.

* **(Countries for Each Show ):**

I added a table for Countries for Each Show by creating a duplicate of the Netflix\_Title table. In this new table, I split the Country column so that whenever a comma is found, it separates the countries into individual rows. This allows for a detailed analysis of the countries associated with each show.

* **(Movies That Netflix Added):**

I established a third table that presents the titles of each movie added by Netflix. This table was constructed by referencing both the Netflix\_Title dataset and the Highest Hollywood Grossing Movies dataset. I conducted a merge operation between the two tables to align the title names, which enabled me to retrieve the corresponding Show\_ID for each title identified in the Highest Hollywood Grossing Movies dataset.

* **Adding Columns: (Season):**

I incorporated a column titled "Season" by utilizing a measure that classifies titles according to their addition dates. This measure evaluates whether the month of the [date\_added] falls within specified ranges: if it identifies December (12), January (1), or February (2), it designates the season as "Winter." Similarly, it assesses March (3), April (4), and May (5) to classify the season as "Spring"; June (6), July (7), and August (8) as "Summer"; and September (9), October (10), and November (11) as "Fall." This approach effectively categorizes each title into one of the four seasons based on its addition date.

* **Added top movies:**

I created a new column that checks if the **Title** in the **Netflix\_Title** table matches the **Title** in the **Highest Hollywood Grossing Movies** table. If there is a match, the column is flagged as "Yes"; if not, it is marked as "No." This helps identify which Netflix titles are part of the highest-grossing Hollywood movies.

**Transformation Data in SQL:**

* **Data Preparation and Table Creation in SQL Server:**

To organize and clean the data, I created separate tables to eliminate commas and separate values in columns such as **Genres**, **Countries**, and **Directors**. This process involved structuring each table individually to ensure a clear representation of the data.

For example, I created a table named **Netflixcasting** using the following SQL code:

A screen shot of a computer

Description automatically generated

This table uses a composite primary key (Show\_Id, Cast) to uniquely identify each entry. I repeated a similar table creation process for other attributes like **Genres**, **Countries**, and **Directors**, adjusting the table names and column names accordingly.

* **Separation Process:**

To populate these tables, I applied SQL code to split columns such as **Cast**, **Genres**, **Director**, and **Country** based on commas. For instance, I inserted data into the **Netflixcasting** table by extracting the Show\_Id from the **Netflix\_Title** table and splitting the Cast column values, as shown below:

A screen shot of a computer program

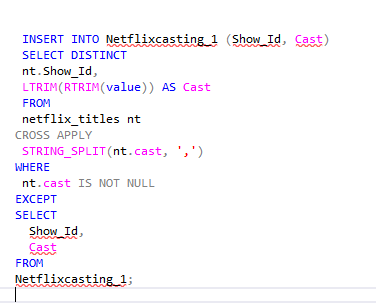
Description automatically generated

This query uses the STRING\_SPLIT function to separate each value within the Cast column and creates individual rows for each cast member associated with a show.

I applied the same method to other tables like **Genres**, **Director**, and **Country** to ensure each attribute is stored in separate rows, facilitating detailed analysis.

* **Handling Duplicates:**

In scenarios where primary key duplicates existed, I removed them using the following SQL code:



This query ensures that only unique combinations of Show\_Id and Cast are added to the new table, effectively eliminating duplicates. I implemented the same logic for other tables that required separation of values, such as **Genres**, **Directors**, and **Countries**.

By using this approach, I achieved a structured and clean dataset, making it easier to perform further analysis on individual attributes.

**Measures & Grouping Used**

**Measures Used**

* **Measures:**
* **Total Content**: This measure counts the total number of unique Show IDs within the library.
* **Total Movies**: This measure counts the number of unique movie titles present in the library.
* **Total TV Shows**: This measure counts the number of unique TV show titles in the library.
* **Percentage of Movies**: This measure calculates the proportion of total content that comprises movies, expressed as a percentage of the overall library.
* **Percentage of TV Shows**: This measure calculates the proportion of total content that consists of TV shows, expressed as a percentage of the overall library.
* **Total Sales**: This measure aggregates the sum of all values in the Worldwide Sales column.
* **Profit from Total Sales**: This measure calculates the total profit by subtracting total costs from total sales, providing insights into overall financial performance.
* **Profit Margin**: This measure calculates profit by dividing the profit from total sales by total sales, expressed as a percentage.
* **Average Budget**: This measure computes the average of the budget column.
* **Average Worldwide Sales**: This measure calculates the average of the worldwide sales column.
* **Grouping:**
* Country by continents : I grouped each country into its respective continent. For example, countries like France, Germany, and Portugal were categorized under the **Europe** continent. This grouping allows for regional analysis and easier aggregation of data at the continent level.
* **Subscription Packages (Basic)**: I categorized the subscription costs into three tiers—**Low Tier**, **Mid Tier**, and **High Tier**—based on their pricing, allowing for a clearer comparison of package affordability and market segmentation.
* **Subscription Packages (Standard)**: I categorized the subscription costs into three tiers—**Low Tier**, **Mid Tier**, and **High Tier**—based on their pricing, allowing for a clearer comparison of package affordability and market segmentation.
* **Subscription Packages (Premium)**: I categorized the subscription costs into three tiers—**Low Tier**, **Mid Tier**, and **High Tier**—based on their pricing, allowing for a clearer comparison of package affordability and market segmentation.

1.**What percentage of Netflix content is classified as TV shows versus movies?**

**Business Questions:**

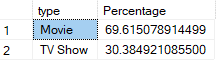
SQL Code:

SELECT type,

COUNT(\*) \* 100.0 / (SELECT COUNT(\*) FROM netflix\_titles) AS Percentage

FROM netflix\_titles

GROUP BY type;



* Insights :  
  This query calculates the percentage of content classified as "TV Shows" vs "Movies".

2.**How many TV shows are added to Netflix in each season?**

SQL Code :

WITH Season\_Categorized AS (

SELECT show\_id, Title,

CASE

WHEN MONTH(date\_added) IN (12, 1, 2) THEN 'Winter'

WHEN MONTH(date\_added) IN (3, 4, 5) THEN 'Spring'

WHEN MONTH(date\_added) IN (6, 7, 8) THEN 'Summer'

WHEN MONTH(date\_added) IN (9, 10, 11) THEN 'Fall'

END AS Season

FROM netflix\_titles

WHERE type = 'TV Show' and date\_added is not null),

Season\_Totals AS (

SELECT Season, COUNT(show\_id) AS TotalShows

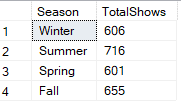
FROM Season\_Categorized

GROUP BY Season)

SELECT Season,TotalShows

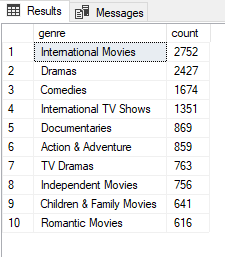
FROM Season\_Totals

ORDER BY Season desc



* **Insights:**
* **Summer Peak**: The data shows that the highest number of TV shows (716) were added during Summer, suggesting that Netflix tends to release more content during this period. This could be due to the increased leisure time that people have during vacations and school breaks, which encourages higher consumption of entertainment.
* **Fall Consistency**: With **655 shows** added in **Fall**, it’s the second-highest season, possibly due to the build-up for holiday season entertainment (like preparing for Thanksgiving, Christmas, etc.). Netflix might release popular content in Fall to engage viewers ahead of the winter holidays.
* **Winter Releases**: **Winter** has slightly fewer releases (**606 shows**) compared to Fall. This might indicate that while Netflix releases significant content during the holiday season (especially December), it may rely more on content from other months in Fall and Winter combined for viewer retention.
* **Spring is the Lowest**: **Spring** has the lowest number of new shows added (**601**). This might indicate a lighter release schedule after the busy Winter and before the high-content Summer period.

3.**Which genres are the most common on Netflix?**



SQL Code :

SELECT TOP 10 genre, COUNT(\*) AS count

FROM NetflixGenres

GROUP BY genre

ORDER BY count DESC

* **Insights :**
* **Content Strategy Alignment:** By analyzing the most common genres, Netflix can align its content acquisition and production strategy with audience demand. For example, investing more in popular genres (like drama or thriller) could increase viewer satisfaction and engagement.
* **Targeted Marketing Campaigns:** Understanding which genres are most common allows Netflix to create targeted marketing campaigns that highlight these genres, catering to viewers' preferences and increasing viewership.
* **Viewer Retention:** Common genres may indicate high viewer demand, so Netflix can curate personalized recommendations based on a user's history in these genres, boosting retention and increasing content consumption.
* **Genre Saturation Management:** If certain genres dominate the platform, Netflix can balance its offerings by introducing more niche or underrepresented genres to diversify its catalog, attracting a broader audience base.
* **Localized Content Strategy:** The popularity of specific genres might vary by region. By identifying the most common genres globally, Netflix can adapt its strategy to introduce more localized versions of these genres in different markets, enhancing its international appeal.

4. **Which countries have the highest number of Netflix subscriptions, and how does this correlate with the total library size?**

SQL Code :

SELECT

country,Total\_Library\_Size

FROM [netflix price in different countries]

GROUP BY country , Total\_Library\_Size

ORDER BY Total\_Library\_Size desc;

A screenshot of a data

Description automatically generated

* **Insights :**
* **Content Demand and Localization:** Countries with a large number of subscriptions may demand more localized content. Netflix can focus on expanding the library of region-specific shows and movies in these countries to further grow its subscriber base.
* **Library Size Impact on Subscription Growth:** A larger library size may correlate with higher subscriptions. In countries where the library size is smaller, Netflix could expand its content offerings to boost subscriptions and improve user retention.
* **Strategic Investment in Content:** Insights into countries with both high subscription rates and a large library could guide Netflix’s content investment strategy, prioritizing regions with higher content consumption and demand.
* **Market Penetration Opportunities:** If high subscription countries have relatively smaller libraries, Netflix can seize the opportunity by increasing the variety of content to stimulate further growth and retain users.
* **Subscription Patterns and Engagement:** A diverse and large content library can keep subscribers engaged longer. By analyzing this correlation, Netflix can optimize its content strategy for different regions to maximize viewer satisfaction and minimize churn rates.

5. **Which countries have the highest number of TV shows and movies available on Netflix?**

SQL Code :

SELECT country, count(title) as count

FROM netflix\_titles

where country is not null

GROUP BY country

ORDER BY count desc;

A screenshot of a computer

Description automatically generated

* **Insights:**
* **Top-Producing Countries**: The results will show which countries contribute the most content to Netflix. Countries like the **United States**, **India**, or **United Kingdom** are likely to be at the top due to their strong film and TV industries.
* **Global Content Strategy**: This query provides insight into **Netflix’s global content distribution**. A high count for a particular country could suggest that Netflix is investing more in local content production or acquiring more shows and movies from that country.
* **Regional Trends**: You might observe that certain regions or countries produce more niche content. For example, **South Korea** might rank high due to the popularity of **K-dramas**, while countries like **Japan** may have a significant number of **anime** titles.

6. **How has the number of titles (movies/TV shows) added per year changed over time?**

SQL Code :

SELECT year(date\_added) as date\_added, type, COUNT(\*) AS total\_titles

FROM netflix\_titles

GROUP BY year(date\_added), type

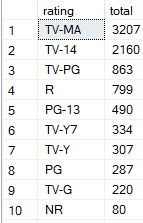
ORDER BY year(date\_added)desc ;

A screenshot of a computer

Description automatically generated

* **Insights** **:** Analyze content growth over the years and Track content growth for movies and TV shows separately over time.

7. **What are the most common content ratings on Netflix?**



SELECT rating, COUNT(\*) AS total

FROM netflix\_titles

Where rating is not null

GROUP BY rating

ORDER BY total DESC;

* **Insights**: Determine how much content is suitable for different age groups (e.g., TV-MA,PG-13).

8.**Which directors on Netflix have the longest and most consistent careers in terms of content releases?**

select distinct Do.director,COUNT(\*) OVER (partition by Do.director) AS total\_content,

first\_value(title) OVER (partition by Do.director order by nt.date\_added) as first\_film,

first\_value(title) OVER (partition by Do.director order by nt.date\_added desc) as last\_film,

first\_value(nt.date\_added) OVER (partition by Do.director order by nt.date\_added) as date\_of\_first\_film,

first\_value(nt.date\_added) OVER (partition by Do.director order by nt.date\_added desc) as date\_of\_last\_film,

DATEDIFF(day, first\_value(nt.date\_added) OVER (partition by Do.director order by nt.date\_added),

first\_value(nt.date\_added) OVER (partition by Do.director order by nt.date\_added desc)) as date\_difference\_in\_days

from netflix\_titles Nt inner join Netflix\_directors DO on nt.show\_id=Do.Show\_Id

where do.director is not null

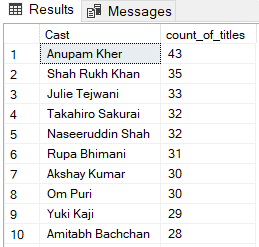
order by total\_content desc;

A screenshot of a computer

Description automatically generated

* insights:
* This helps a business or content acquisition team assess:
* Which directors have been consistently producing content over time. The potential for future collaborations based on a director's history of releases.
* How active a director has been on Netflix, potentially identifying those who might benefit from new opportunities.

9. **How can Netflix leverage insights from cast members with the highest number of appearances to improve content recommendations and increase viewer engagement?**



select ca.Cast , COUNT (\*) as count\_of\_titles

from Netflixcasting ca inner join netflix\_titles ti

on ca.Show\_Id = ti.show\_id

group by ca.Cast

order by count\_of\_titles desc

* Insights :
* **Targeted Content Promotion:** Cast members with the highest number of titles are likely to have a strong fan base. Netflix can feature these actors in its marketing campaigns or highlight their movies and shows in personalized recommendations to boost engagement.
* **Content Acquisition and Investment:** If a specific actor or set of actors consistently performs well across multiple titles, Netflix can focus on acquiring or producing more content featuring them. This can ensure a consistent draw for subscribers and increase retention.
* **Recommendation Algorithm Improvement:** Data on cast members who frequently appear in popular titles can be used to refine Netflix’s recommendation engine. For example, viewers who have watched multiple shows featuring a specific actor can receive suggestions for other content featuring the same cast.
* **Partnership Opportunities:** Knowing which cast members are highly featured and popular on Netflix could open doors for exclusive partnerships, collaborations, or original content deals with these actors, further cementing Netflix's competitive edge in the streaming market.

10.**Which Netflix subscription plan (Basic, Standard, Premium) is the most prevalent across continents for each tier, and what is the average cost for each plan across these continents**

**Insight from the Question:**

1. **Identifying Plan Prevalence Across Continents:**
   * By determining which Netflix subscription plan (Basic, Standard, Premium) is the most prevalent across different continents for each tier, we can understand regional preferences. This helps in assessing if certain regions favor more affordable plans (like Basic) or if they prefer plans with higher features (like Standard or Premium).
2. **Analyzing Average Cost for Each Plan:**
   * Calculating the average cost for each plan across continents provides insights into how Netflix's pricing strategy varies by region. For instance, if a continent prefers the Premium plan, is it because the cost difference between Premium and Standard is smaller in that region compared to others?