

Netflix & Disney+ Analysis

Prepared by:

Ahmed Hosny

Abdulfattah Gamal

Supervised by:

Eng. Marwan Mokhtar

Course:

Microsoft Power BI

Group code / Tech Company / Gov:

MNF1_DAT2_M1d / NHC / Menoufia

Content:

- Introduction
- Data Collection and Understanding
- Business questions
- Data Modeling
- Database Creation
- Data Preprocessing
- Analysis business questions
- Dashboard
- Insights

Skills:

- Business background
- Programming (SQL)
- Visualization & Analysis (Power BI)

Tools:

- Power Query (Excel & Power BI) for preprocessing the data
- PostgreSQL for SQL query
- Power BI for building dashboards

Introduction:

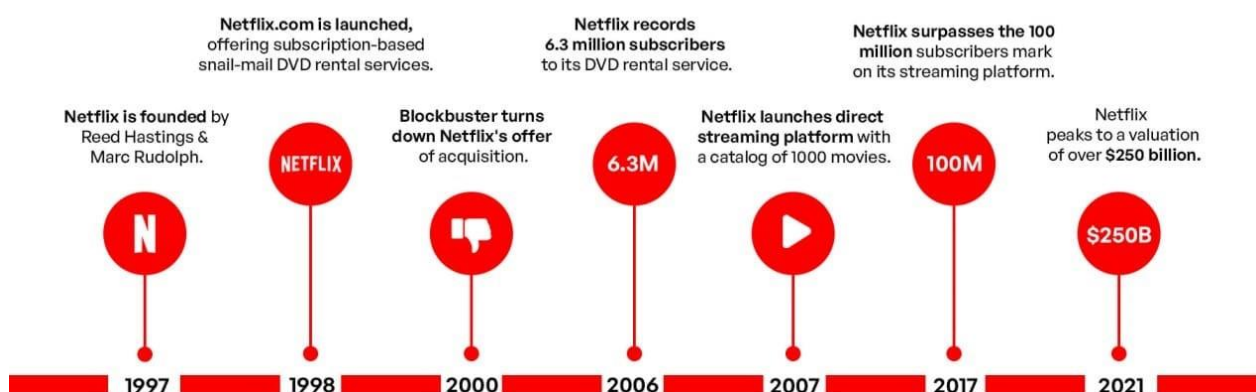
Netflix

Netflix is a streaming service that offers a wide variety of award-winning TV Shows, Movies, Anime, Documentaries and more – on thousands of internet-connected devices.

You can watch as much as you want whenever you want without a single ad – all for one low monthly price. There's always something new to discover and new TV Shows and Movies are added every week!

History of Netflix

Netflix is an American subscription video on-demand over-the-top streaming service. The service primarily distributes original and acquired films and television shows from various genres and it is available internationally in multiple languages. Launched on January 16, 2007, nearly a decade after Netflix Inc. began its pioneering DVD-by-mail movie rental service, Netflix is the most-subscribed video on demand streaming media service with over 277.7 million paid memberships in more than 190 countries as of July 2024.



Disney+

Disney+ is a streaming service that offers a vast collection of movies, TV shows, documentaries, and more from Disney, Pixar, Marvel, Star Wars, National Geographic, and beyond – all accessible on numerous internet-connected devices. You can enjoy unlimited ad-free viewing for a single monthly fee. With new content added regularly, there's always something fresh to watch!

History

Disney+ is an American subscription video-on-demand streaming service. It primarily features original and acquired content from various genres, available internationally in multiple languages. Launched on November 12, 2019, Disney+ quickly became a major player in the streaming industry. As of March 2024, it boasts over 153.8 million paid subscribers across more than 150 markets

Data Collection and Understanding

In this stage, we focus on two main steps:

1. Collecting the Data: We sourced the data from the Kaggle website.
2. Understanding the Data: We analyzed the datasets from Netflix and Disney+, examining their columns and defining our responsibilities.

This process sets the foundation for our Exploratory Data Analysis (EDA).

1.collecting the data from the Kaggle website:

get the raw data from the links:

Netflix: [Netflix Movies and TV Shows \(kaggle.com\)](#)

Netflix subscription: [Netflix subscription fee in different countries \(kaggle.com\)](#)

Disney+: [Disney+ Movies and TV Shows \(kaggle.com\)](#)

2.Exploratory Data Analysis (EDA):

Netflix & Disney+ Dataset - Column Descriptions

1. **show_id:**
A unique identifier assigned to each title (movie or TV show) on Netflix.
2. **type:**
Indicates whether the title is a **Movie** or a **TV Show**.
3. **title:**
The official name of the title (movie or TV show) on Netflix.
4. **director:**
The name(s) of the director(s) responsible for the movie or TV show. If there are multiple directors, they are listed separately by commas.
5. **cast:**
The list of actors and actresses who appear in the movie or TV show. Multiple cast members are separated by commas.
6. **country:**
The country or countries where the movie or TV show was produced. If there are multiple countries, they are separated by commas.
7. **date_added:**
The date when the title was added to the Netflix library.
8. **release_year:**
The year the movie or TV show was originally released.
9. **rating:**
The maturity rating of the title (e.g., **TV-MA**, **PG-13**), indicating the appropriate audience based on age restrictions.

10. **duration:**

The length of the movie in minutes or the number of seasons for TV shows (e.g., **90 min** or **2 Seasons**).

11. **listed_in:**

The genres or categories the title belongs to (e.g., **Documentaries, Comedies**).
Multiple genres are separated by commas.

12. **description:**

A short synopsis or summary of the title, describing its plot or content.

Content Ratings Column:

1. **NR (Not Rated)**: Indicates that the show or movie has not been rated by the Motion Picture Association of America (MPAA) or any other rating system.
2. **TV-Y**: Suitable for all children, usually content appropriate for very young audiences.
3. **TV-Y7**: Suitable for children aged 7 and older. Content may contain mild fantasy violence.
4. **TV-Y7-FV**: Similar to TV-Y7 but includes fantasy violence.
5. **TV-G**: General audiences, suitable for all age groups.
6. **TV-14**: Parents strongly cautioned. Some material may be inappropriate for children under 14.
7. **TV-MA**: Mature audiences may contain content unsuitable for children under 17.
8. **G**: General audiences, suitable for all age groups.
9. **NC-17**: Adults only. Content is not suitable for viewers under 17.
10. **PG**: Parental guidance suggested. Some material may not be suitable for children.
11. **TV-PG**: Parental guidance suggested. Content may be unsuitable for younger children.
12. **PG-13**: Parents strongly cautioned. Some material may be inappropriate for children under 13.
13. **R**: Restricted. This rating indicates that viewers under 17 require accompanying parent or adult guardian due to strong content such as violence, language, or sexual situations.
14. **UR (Unrated)**: the content has not been officially rated by a recognized rating board like the MPAA (Motion Picture Association of America). This often happens with independent or less mainstream content

Business Questions:

The following business questions are designed to extract valuable insights from the Netflix and Disney+ datasets. These questions aim to uncover trends, patterns, and key metrics that can inform strategic decisions and enhance our understanding of the content available on these platforms.

1. Find the Top Countries with the Most Content on Netflix.
2. Find the Most Common Rating for Movies and TV Shows.
3. Identify the Longest Movies.
4. Find Content Added in the Last 5 Years.
5. Find Each Year and the Average Number of Content Released in Egypt on Netflix.
6. Find the Top 10 Actors Who Have Appeared in the Highest Number of Movies.
7. Number of Actors for Each Country as Percentage of Total Actors on Netflix.
8. Number of Directors for Each Country as Percentage of Total Directors on Netflix.
9. Growth Content for Top 5 Countries with Content on Netflix (Since 2009).
10. Show the Difference in the Number of TV Shows Added Year-over-Year for Each Country.
11. Top Casts for Distinct Genres.
12. Distinct Genres for Each Country (Genre Diversity).
13. Number of Age Groups for Each Country.
14. The Top 5 Most Common Genres Across All Shows and Their Total Count.
15. Find the Next Show's Title (Lead) and Previous Show's Title (Lag) for Each Show Based on the Release Year.

16. Top 5 Longest Movies by Duration and Their Directors.
17. Percentage of TV Shows vs Movies Added in the Last 4 Years (2018-2021).
18. Top 3 Countries with the Most Content in 2020 and 2021.
19. Find the Top 5 Directors with the Most Shows Listed.
20. Calculate the Average Duration of Movies by Rating.
21. Count the Number of Content Items in Each Genre.
22. Find the Most Common Rating for Movies and TV Shows (Top 5 for Each Type).
23. Running Total of Movies Added Year-Over-Year by Rating.
24. Number of Content Items for Each Release Year.
25. Number of Content Items Added to Netflix Each Year.

Data Modeling:

In this stage, we focused on improving the database design to enhance data integrity and efficiency. Here are the steps we took:

1. Director column has multiple values in one row
2. Cast column has many values in one row
3. Country column has multiple values in one row
4. Listed_in column has multiple values in one row

Problem Statement

The current database design involves storing multiple values in a single column, which leads to data duplication and inefficient storage. Specifically, the Listed_in column contains multiple genres separated by commas. This design complicates data analysis and retrieval.

Proposed Solution

To address these issues, the following steps are proposed:

1. Normalization of Data:

- **Current Approach:** Each genre is stored as a comma-separated value within the Listed_in column.
- **Issues:** This approach increases the number of columns and leads to data duplication, resulting in inefficient storage and complex queries.

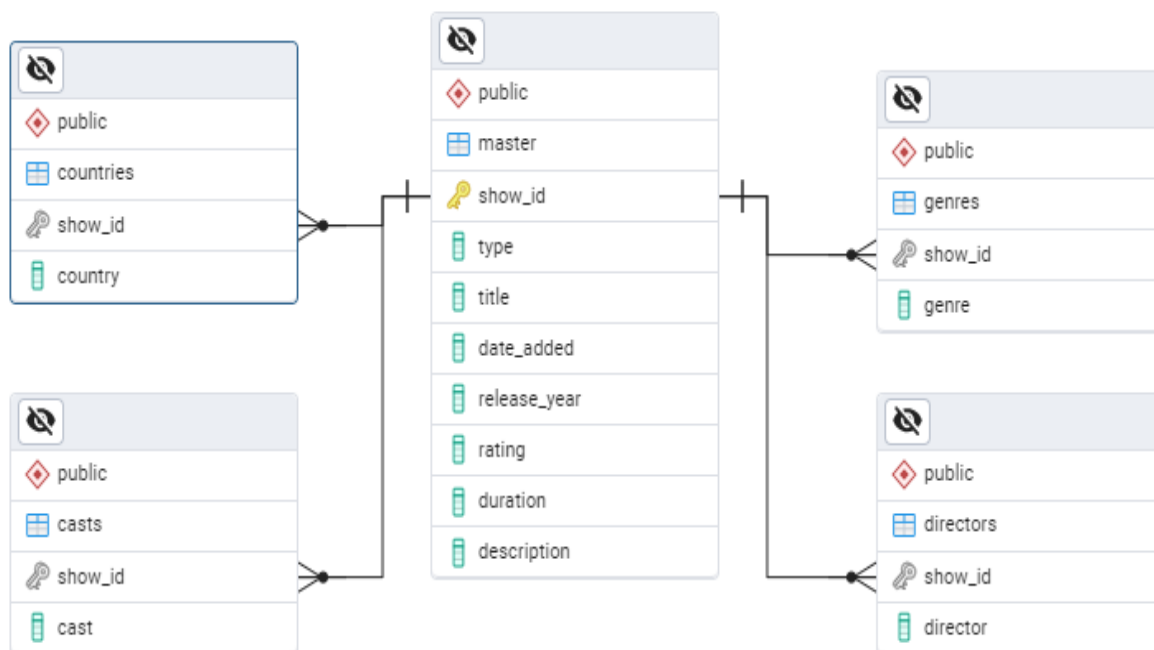
2. Using UNNEST Function:

- **Current Approach:** The UNNEST(STRING_TO_ARRAY(column, ',')) AS country function is used to split the comma-separated values into individual rows.
- **Issues:** This method requires repeated execution and results in data duplication.

3. Redesigning the Entity-Relationship Diagram (ERD):

- **New Approach:** Normalize the Listed_in column by creating a separate table for genres. This table will use show_id and a new genre_id as a composite key.
- **Benefits:** This design improves data integrity, reduces duplication, and enhances query performance.

ERD:



Database Creation:

-- Table: public.master

-- DROP TABLE IF EXISTS public.master;

CREATE TABLE IF NOT EXISTS public.master

(

show_id character varying(7) COLLATE pg_catalog."default" NOT NULL,

type character varying(7) COLLATE pg_catalog."default",

title character varying(110) COLLATE pg_catalog."default",

date_added character varying(10) COLLATE pg_catalog."default",

release_year character varying(5) COLLATE pg_catalog."default",

rating character varying(20) COLLATE pg_catalog."default",

duration character varying(15) COLLATE pg_catalog."default",

description character varying(300) COLLATE pg_catalog."default",

CONSTRAINT master_pkey PRIMARY KEY (show_id)

)

TABLESPACE pg_default;

ALTER TABLE IF EXISTS public.master

OWNER to postgres;

-- Table: public.casts

-- DROP TABLE IF EXISTS public.casts;

CREATE TABLE IF NOT EXISTS public.casts

(
 show_id character varying(6) COLLATE pg_catalog."default",
 actor character varying(100) COLLATE pg_catalog."default"
)

TABLESPACE pg_default;

ALTER TABLE IF EXISTS public.casts

 OWNER to postgres;

-- Table: public.countries

-- DROP TABLE IF EXISTS public.countries;

CREATE TABLE IF NOT EXISTS public.countries

(
 show_id character varying(6) COLLATE pg_catalog."default",

```
country character varying(25) COLLATE pg_catalog."default"  
)
```

```
TABLESPACE pg_default;
```

```
ALTER TABLE IF EXISTS public.countries  
OWNER to postgres;
```

```
-- Table: public.directors
```

```
-- DROP TABLE IF EXISTS public.directors;
```

```
CREATE TABLE IF NOT EXISTS public.directors
```

```
(  
    show_id character varying(6) COLLATE pg_catalog."default",  
    director character varying(35) COLLATE pg_catalog."default"  
)
```

```
TABLESPACE pg_default;
```


ALTER TABLE IF EXISTS public.directors

OWNER to postgres;

-- Table: public.genres

-- DROP TABLE IF EXISTS public.genres;

CREATE TABLE IF NOT EXISTS public.genres

(
 show_id character varying(6) COLLATE pg_catalog."default",
 genre character varying(30) COLLATE pg_catalog."default"
)

TABLESPACE pg_default;

ALTER TABLE IF EXISTS public.genres

OWNER to postgres;

Data Preprocessing:

In this stage, put significant effort into preparing the data to ensure it was in the best possible form. This involved multiple steps and careful attention to detail.

1. For the country table we undertook several steps to improve data quality and usability:

1. **Splitting Multi-Valued Rows:** split columns containing multiple values to eliminate multiplicity.
2. **AI Grouping:** We used AI to group countries based on continents and cultures.
3. **Trimming Values:** We removed any leading or trailing spaces (Trim()) to clean the data.
4. **Rearranging Columns:** Columns were rearranged to enhance visual appeal and accessibility.
5. **Creating New Classifications:** Using conditional statements, we created new columns with classifications provided by AI.
6. **Merging Tables:** We merged the country table with the subscriptions table to determine the subscription type for each country (basic, standard, or premium).



1. Sample from the table before any preprocessing

= Csv.Document(File.Contents("G:\Tasks DEPI\Netflix\Final\Netflix\	
Column1	Column2
Valid 100%	Valid 100%
Error 0%	Error 0%
Empty 0%	Empty 0%
772 distinct, 612 unique	95 distinct, 37 unique
1 show_id	country
2 s1	United States
3 s2	South Africa
4 s5	India
5 s8	United States
6 s8	Ghana
7 s8	Burkina Faso
8 s8	United Kingdom
9 s8	Germany
10 s8	Ethiopia
11 s9	United Kingdom
12 s10	United States
13 s13	Germany
14 s13	Czech Republic
15 s16	United States
16 s18	Mexico
17 s22	Turkey
18 s25	India
19 s26	Australia
20 s28	United States
21 s29	United States
22 s30	United States
23 s30	India
24 s30	France

2. The table after using first rows as headers

show_id		country	
<p>Valid 100%</p> <p>Error 0%</p> <p>Empty 0%</p> <p>772 distinct, 612 unique</p>		<p>Valid 100%</p> <p>Error 0%</p> <p>Empty 0%</p> <p>94 distinct, 36 unique</p>	
1	s1	United States	
2	s2	South Africa	
3	s5	India	
4	s8	United States	
5	s8	Ghana	
6	s8	Burkina Faso	
7	s8	United Kingdom	
8	s8	Germany	
9	s8	Ethiopia	
10	s9	United Kingdom	
11	s10	United States	
12	s13	Germany	
13	s13	Czech Republic	
14	s16	United States	
15	s18	Mexico	
16	s22	Turkey	
17	s25	India	
18	s26	Australia	
19	s28	United States	
20	s29	United States	
21	s30	United States	
22	s30	India	
23	s30	France	
24	s33	United Kingdom	

3. The table after trim

AB _C show_id		AB _C country	
<div> <div>Valid 100%</div> <div>Error 0%</div> <div>Empty 0%</div> </div> <div>772 distinct, 612 unique</div>		<div> <div>Valid 100%</div> <div>Error 0%</div> <div>Empty 0%</div> </div> <div>62 distinct, 18 unique</div>	
1	s1	United States	
2	s2	South Africa	
3	s5	India	
4	s8	United States	
5	s8	Ghana	
6	s8	Burkina Faso	
7	s8	United Kingdom	
8	s8	Germany	
9	s8	Ethiopia	
10	s9	United Kingdom	
11	s10	United States	
12	s13	Germany	
13	s13	Czech Republic	
14	s16	United States	
15	s18	Mexico	
16	s22	Turkey	
17	s25	India	
18	s26	Australia	
19	s28	United States	
20	s29	United States	
21	s30	United States	
22	s30	India	
23	s30	France	
24	s33	United Kingdom	

4. After adding continent column

show_id	Country	Continent
Valid 100%	Valid 100%	Valid 100%
Error 0%	Error 0%	Error 0%
Empty 0%	Empty 0%	Empty 0%
772 distinct, 612 unique	62 distinct, 18 unique	
s1	United States	North America
s2	South Africa	Africa
s5	India	Asia
s8	United States	North America
s8	Ghana	Africa
s8	Burkina Faso	Africa
s8	United Kingdom	Europe
s8	Germany	Europe
s8	Ethiopia	Africa
s9	United Kingdom	Europe
s10	United States	North America
s13	Germany	Europe
s13	Czech Republic	Europe
s16	United States	North America
s18	Mexico	North America
s22	Turkey	Europe
s25	India	Asia
s26	Australia	Oceania
s28	United States	North America
s29	United States	North America
s30	United States	North America
s30	India	Asia
s30	France	Europe
s33	United Kingdom	Europe

5. After adding culture column

AB _C show_id	AB _C Country	ABC 123 Continent	AB _C Culture
<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>772 distinct, 612 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>62 distinct, 18 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% 	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>9 distinct, 0 unique</p>
s1	United States	North America	North American / American
s2	South Africa	Africa	Sub-Saharan African
s5	India	Asia	Asian
s8	United States	North America	North American / American
s8	Ghana	Africa	Sub-Saharan African
s8	Burkina Faso	Africa	Sub-Saharan African
s8	United Kingdom	Europe	European
s8	Germany	Europe	European
s8	Ethiopia	Africa	Sub-Saharan African
s9	United Kingdom	Europe	European
s10	United States	North America	North American / American
s13	Germany	Europe	European
s13	Czech Republic	Europe	European
s16	United States	North America	North American / American
s18	Mexico	North America	North American / American
s22	Turkey	Europe	European
s25	India	Asia	Asian
s26	Australia	Oceania	Oceanian
s28	United States	North America	North American / American
s29	United States	North America	North American / American
s30	United States	North America	North American / American
s30	India	Asia	Asian
s30	France	Europe	European
s33	United Kingdom	Europe	European

6. After merging with subscription fees

show_id	Country	Continent	Culture	Monthly subscription cost - Basic (\$)	Monthly subscription cost - Standard (\$)
Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 93% Error 0% Empty 7%	Valid 93% Error 0% Empty 7%
772 distinct, 612 unique	62 distinct, 18 unique		9 distinct, 0 unique	32 distinct, 4 unique	35 distinct, 5 unique
s1	United States	North America	North American / American	8.99	13.99
s8	United States	North America	North American / American	8.99	13.99
s10	United States	North America	North American / American	8.99	13.99
s16	United States	North America	North American / American	8.99	13.99
s28	United States	North America	North American / American	8.99	13.99
s29	United States	North America	North American / American	8.99	13.99
s30	United States	North America	North American / American	8.99	13.99
s39	United States	North America	North American / American	8.99	13.99
s41	United States	North America	North American / American	8.99	13.99
s42	United States	North America	North American / American	8.99	13.99
s43	United States	North America	North American / American	8.99	13.99
s44	United States	North America	North American / American	8.99	13.99
s45	United States	North America	North American / American	8.99	13.99
s47	United States	North America	North American / American	8.99	13.99
s49	United States	North America	North American / American	8.99	13.99
s56	United States	North America	North American / American	8.99	13.99
s68	United States	North America	North American / American	8.99	13.99
s82	United States	North America	North American / American	8.99	13.99
s83	United States	North America	North American / American	8.99	13.99
s91	United States	North America	North American / American	8.99	13.99
s2	South Africa	Africa	Sub-Saharan African	6.26	10.00
s47	South Africa	Africa	Sub-Saharan African	6.26	10.00
s26	Australia	Oceania	Oceanian	7.84	12.10

7. After getting rid of missing values in subscription columns

show_id	Country	Continent	Culture	Monthly subscription cost - Basic (\$)	Monthly subscription cost - Standard (\$)
Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
772 distinct, 612 unique	62 distinct, 18 unique		9 distinct, 0 unique		
s1	United States	North America	North American / American	8.99	13.99
s8	United States	North America	North American / American	8.99	13.99
s10	United States	North America	North American / American	8.99	13.99
s16	United States	North America	North American / American	8.99	13.99
s28	United States	North America	North American / American	8.99	13.99
s29	United States	North America	North American / American	8.99	13.99
s30	United States	North America	North American / American	8.99	13.99
s39	United States	North America	North American / American	8.99	13.99
s41	United States	North America	North American / American	8.99	13.99
s42	United States	North America	North American / American	8.99	13.99
s43	United States	North America	North American / American	8.99	13.99
s44	United States	North America	North American / American	8.99	13.99
s45	United States	North America	North American / American	8.99	13.99
s47	United States	North America	North American / American	8.99	13.99
s49	United States	North America	North American / American	8.99	13.99
s56	United States	North America	North American / American	8.99	13.99
s68	United States	North America	North American / American	8.99	13.99
s82	United States	North America	North American / American	8.99	13.99
s83	United States	North America	North American / American	8.99	13.99
s91	United States	North America	North American / American	8.99	13.99
s2	South Africa	Africa	Sub-Saharan African	6.26	10.00
s47	South Africa	Africa	Sub-Saharan African	6.26	10.00
s26	Australia	Oceania	Oceanian	7.84	12.10

2. For the genres table, we undertook several steps to improve data quality and usability:

1. **Splitting Multi-Valued Rows:** split columns containing multiple values to eliminate multiplicity.
2. **AI Grouping:** We used AI to group categories based on genres.
3. **Trimming Values:** We removed any leading or trailing spaces (trim) to clean the data.
4. **Rearranging Columns:** Columns were rearranged to enhance visual appeal and accessibility.
5. **Creating New Classifications:** Using conditional statements, we created new columns with classifications provided by AI.



(1) A sample from table before any preprocessing

	Column1	Column2
	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>459 distinct, 105 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>43 distinct, 3 unique</p>
1	show_id	genre
2	s1	Documentaries
3	s2	International TV Shows
4	s2	TV Dramas
5	s2	TV Mysteries
6	s3	Crime TV Shows
7	s3	International TV Shows
8	s3	TV Action & Adventure
9	s4	Docuseries
10	s4	Reality TV
11	s5	International TV Shows
12	s5	Romantic TV Shows
13	s5	TV Comedies
14	s6	TV Dramas
15	s6	TV Horror
16	s6	TV Mysteries
17	s7	Children & Family Movies
18	s8	Dramas
19	s8	Independent Movies
20	s8	International Movies
21	s9	British TV Shows
22	s9	Reality TV
23	s10	Comedies
24	s10	Dramas



(2) After adding the genre segmentation column

	ABC C show_id	ABC C genre	ABC 123 Genre segmentation
	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>458 distinct, 104 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>42 distinct, 2 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0%
1	s1	Documentaries	Documentary
2	s2	International TV Shows	International shows
3	s2	TV Dramas	Drama
4	s2	TV Mysteries	Crime
5	s3	Crime TV Shows	Crime
6	s3	International TV Shows	International shows
7	s3	TV Action & Adventure	Action&Adventure
8	s4	Docuseries	Documentary
9	s4	Reality TV	Comedy
10	s5	International TV Shows	International shows
11	s5	Romantic TV Shows	Romantic
12	s5	TV Comedies	Comedy
13	s6	TV Dramas	Drama
14	s6	TV Horror	Horror
15	s6	TV Mysteries	Crime
16	s7	Children & Family Movies	Family
17	s8	Dramas	Drama
18	s8	Independent Movies	Independent
19	s8	International Movies	International shows
20	s9	British TV Shows	International shows
21	s9	Reality TV	Comedy
22	s10	Comedies	Comedy
23	s10	Dramas	Drama
24	s11	Crime TV Shows	Crime

(3) After trimming the columns

AB _C show_id	AB _C genre	AB _C Genre segmentation
<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>458 distinct, 104 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>42 distinct, 2 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>19 distinct, 0 unique</p>
s1	Documentaries	Documentary
s2	International TV Shows	International shows
s2	TV Dramas	Drama
s2	TV Mysteries	Crime
s3	Crime TV Shows	Crime
s3	International TV Shows	International shows
s3	TV Action & Adventure	Action&Adventure
s4	Docuseries	Documentary
s4	Reality TV	Comedy
s5	International TV Shows	International shows
s5	Romantic TV Shows	Romantic
s5	TV Comedies	Comedy
s6	TV Dramas	Drama
s6	TV Horror	Horror
s6	TV Mysteries	Crime
s7	Children & Family Movies	Family
s8	Dramas	Drama
s8	Independent Movies	Independent
s8	International Movies	International shows
s9	British TV Shows	International shows
s9	Reality TV	Comedy
s10	Comedies	Comedy
s10	Dramas	Drama
s11	Crime TV Shows	Crime

3. For the Cast table, we undertook several steps to improve data quality and usability:

1. **Splitting Multi-Valued Rows:** split columns containing multiple values to eliminate multiplicity.
2. **Trimming Values:** We removed any leading or trailing spaces (trim) to clean the data.
3. **Rearranging Columns:** Columns were rearranged to enhance visual appeal and accessibility.


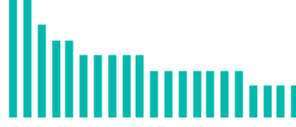
(1) A sample from the data before any preprocessing

Column1	Column2
<p>Valid 100%</p> <p>Error 0%</p> <p>Empty 0%</p> <p>116 distinct, 6 unique</p>	<p>Valid 100%</p> <p>Error 0%</p> <p>Empty 0%</p> <p>903 distinct, 845 unique</p>
show_id	Actor
s2	Ama Qamata
s2	Khosi Ngema
s2	Gail Mabalane
s2	Thabang Molaba
s2	Dillon Windvogel
s2	Natasha Thahane
s2	Arno Greeff
s2	Xolile Tshabalala
s2	Getmore Sithole
s2	Cindy Mahlangu
s2	Ryle De Morny
s2	Greteli Fincham
s2	Sello Maake Ka-Ncube
s2	Odwa Gwanya
s2	Mekaila Mathys
s2	Sandi Schultz
s2	Duane Williams
s2	Shamilla Miller
s2	Patrick Mofokeng
s3	Sami Bouajila
s3	Tracy Gotoas
s3	Samuel Jouy
s3	Nabiha Akkari

(2) After using first rows as headers

AB _C show_id	AB _C Actor
<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>115 distinct, 5 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0% <p>903 distinct, 845 unique</p>
s2	Ama Qamata
s2	Khosi Ngema
s2	Gail Mabalane
s2	Thabang Molaba
s2	Dillon Windvogel
s2	Natasha Thahane
s2	Arno Greeff
s2	Xolile Tshabalala
s2	Getmore Sithole
s2	Cindy Mahlangu
s2	Ryle De Morny
s2	Greteli Fincham
s2	Sello Maaake Ka-Ncube
s2	Odwa Gwanya
s2	Mekaila Mathys
s2	Sandi Schultz
s2	Duane Williams
s2	Shamilla Miller
s2	Patrick Mofokeng
s3	Sami Bouajila
s3	Tracy Gotoas
s3	Samuel Jouy
s3	Nabiha Akkari
s3	Sofia Lesaffre

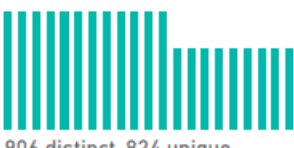
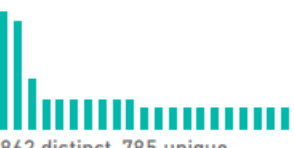
(3) After trimming the columns

AB _C show_id	AB _C Actor
<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0%  <p>115 distinct, 5 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0%  <p>903 distinct, 845 unique</p>
s2	Ama Qamata
s2	Khosi Ngema
s2	Gail Mabalane
s2	Thabang Molaba
s2	Dillon Windvogel
s2	Natasha Thahane
s2	Arno Greeff
s2	Xolile Tshabalala
s2	Getmore Sithole
s2	Cindy Mahlangu
s2	Ryle De Morny
s2	Greteli Fincham
s2	Sello Maake Ka-Ncube
s2	Odwa Gwanya
s2	Mekaila Mathys
s2	Sandi Schultz
s2	Duane Williams
s2	Shamilla Miller
s2	Patrick Mofokeng
s3	Sami Bouajila
s3	Tracy Gotoas
s3	Samuel Jouy
s3	Nabiha Akkari
s3	Sofia Lesaffre

4. For the Directors table, we undertook several steps to improve data quality and usability:

1. **Splitting Multi-Valued Rows:** split columns containing multiple values to eliminate multiplicity.
2. **Trimming Values:** We removed any leading or trailing spaces (trim) to clean the data.
3. **Rearranging Columns:** Columns were rearranged to enhance visual appeal and accessibility.

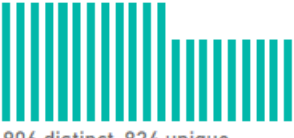

(1) Before any preprocessing

	Column1	Column2
	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
	 906 distinct, 824 unique	 862 distinct, 785 unique
1	show_id	Director
2	s1	Kirsten Johnson
3	s3	Julien Leclercq
4	s6	Mike Flanagan
5	s7	Robert Cullen
6	s7	José Luis Ucha
7	s8	Haile Gerima
8	s9	Andy Devonshire
9	s10	Theodore Melfi
10	s12	Kongkiat Komesiri
11	s13	Christian Schwochow
12	s14	Bruno Garotti
13	s17	Pedro de Echave Garc�a
14	s17	Pablo Azor�n Williams
15	s19	Adam Salky
16	s21	Olivier Megaton
17	s23	K.S. Ravikumar
18	s24	Alex Woo
19	s24	Stanley Moore
20	s25	S. Shankar
21	s27	Rajiv Menon
22	s28	Dennis Dugan
23	s29	Scott Stewart
24	s30	Robert Luketic

(2) After using first rows as headers

show_id		Director	
<p>Valid 100%</p> <p>Error 0%</p> <p>Empty 0%</p> <p>906 distinct, 824 unique</p>		<p>Valid 100%</p> <p>Error 0%</p> <p>Empty 0%</p> <p>862 distinct, 785 unique</p>	
1	s1	Kirsten Johnson	
2	s3	Julien Leclercq	
3	s6	Mike Flanagan	
4	s7	Robert Cullen	
5	s7	Jos�� Luis Ucha	
6	s8	Haile Gerima	
7	s9	Andy Devonshire	
8	s10	Theodore Melfi	
9	s12	Kongkiat Komesiri	
10	s13	Christian Schwochow	
11	s14	Bruno Garotti	
12	s17	Pedro de Echave Garc��a	
13	s17	Pablo Azor��n Williams	
14	s19	Adam Salky	
15	s21	Olivier Megaton	
16	s23	K.S. Ravikumar	
17	s24	Alex Woo	
18	s24	Stanley Moore	
19	s25	S. Shankar	
20	s27	Rajiv Menon	
21	s28	Dennis Dugan	
22	s29	Scott Stewart	
23	s30	Robert Luketic	
24	s31	Ashwiny Iyer Tiwari	

(3) After trimming the columns

	AB _C show_id	AB _C Director
	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0%  <p>906 distinct, 824 unique</p>	<ul style="list-style-type: none"> Valid 100% Error 0% Empty 0%  <p>862 distinct, 785 unique</p>
1	s1	Kirsten Johnson
2	s3	Julien Leclercq
3	s6	Mike Flanagan
4	s7	Robert Cullen
5	s7	Jos�� Luis Ucha
6	s8	Haile Gerima
7	s9	Andy Devonshire
8	s10	Theodore Melfi
9	s12	Kongkiat Komesiri
10	s13	Christian Schwochow
11	s14	Bruno Garotti
12	s17	Pedro de Echave Garc��a
13	s17	Pablo Azor��n Williams
14	s19	Adam Salky
15	s21	Olivier Megaton
16	s23	K.S. Ravikumar
17	s24	Alex Woo
18	s24	Stanley Moore
19	s25	S. Shankar
20	s27	Rajiv Menon
21	s28	Dennis Dugan
22	s29	Scott Stewart
23	s30	Robert Luketic
24	s31	Ashwiny Iyer Tiwari

5. For the Master table, we undertook several steps to improve data quality and usability:

1. **Segmenting Age Ratings:** segmented the age ratings into “Age Groups” based on common target age groups shared among each rating.
2. **Segmenting Movies by Duration:** We categorized movies into three groups based on their durations:
 - Under 60 minutes
 - 60-90 minutes
 - Over 90 minutes
3. **Segmenting TV Shows by Duration:** We categorized TV shows into three groups based on their seasons:
 - Under 5 seasons
 - 5-10 seasons
 - Over 10 seasons

A ^B _C rating	A ^B _C Age group	1 ² ₃ duration value	A ^B _C duration unit	A ^B _C Duration segmentation
<div> <div>Valid</div> <div>Error</div> <div>Empty</div> </div> <div>100%</div> <div>0%</div> <div>0%</div> 	<div> <div>Valid</div> <div>Error</div> <div>Empty</div> </div> <div>100%</div> <div>0%</div> <div>0%</div> 	<div> <div>Valid</div> <div>Error</div> <div>Empty</div> </div> <div>100%</div> <div>0%</div> <div>0%</div> 	<div> <div>Valid</div> <div>Error</div> <div>Empty</div> </div> <div>100%</div> <div>0%</div> <div>0%</div> 	<div> <div>Valid</div> <div>Error</div> <div>Empty</div> </div> <div>100%</div> <div>0%</div> <div>0%</div> 
PG-13	Parents strongly cautioned or children under 13		90 min	60-90 min
TV-MA	Mature audiences		2 Seasons	Under 5 seasons
TV-MA	Mature audiences		1 Season	Under 5 seasons
TV-MA	Mature audiences		1 Season	Under 5 seasons
TV-MA	Mature audiences		2 Seasons	Under 5 seasons
TV-MA	Mature audiences		1 Season	Under 5 seasons
PG	Parental guidance suggested for children		91 min	Over 90 min
TV-MA	Mature audiences		125 min	Over 90 min
TV-14	Parents strongly cautioned or children under 14		9 Seasons	5-10 seasons
PG-13	Parents strongly cautioned or children under 13		104 min	Over 90 min
TV-MA	Mature audiences		1 Season	Under 5 seasons
TV-MA	Mature audiences		1 Season	Under 5 seasons
TV-MA	Mature audiences		127 min	Over 90 min
TV-PG	Parental guidance suggested for younger		91 min	Over 90 min
TV-MA	Mature audiences		1 Season	Under 5 seasons
TV-MA	Mature audiences		4 Seasons	Under 5 seasons
TV-MA	Mature audiences		67 min	60-90 min
TV-MA	Mature audiences		2 Seasons	Under 5 seasons
TV-14	Parents strongly cautioned or children under 14		94 min	Over 90 min
TV-MA	Mature audiences		1 Season	Under 5 seasons
TV-14	Parents strongly cautioned or children under 14		1 Season	Under 5 seasons
TV-14	Parents strongly cautioned or children under 14		5 Seasons	5-10 seasons
TV-PG	Parental guidance suggested for younger		161 min	Over 90 min

6. Some records are shifted

A ^B _C rating	A ^B _C duration	A ^B _C listed_in
2 distinct, 2 unique	2 distinct, 2 unique	1 distinct, 0 unique
74 min		Movies
84 min		Movies
66 min		Movies
	37 min	Movies
	1 Season	Anime Series, International T...
	1 Season	Kids' TV, TV Comedies
	115 min	Dramas

7. Handling Missing Values: identified missing values in the Directors (2634), Casts (825), and Countries (831) columns.

8. Error Correction: deleted 2 rows that contained errors during recording to ensure data accuracy and integrity

A ^B _C show_id	A ^B _C type	A ^B _C title	A ^B _C director
2 distinct, 2 unique	2 distinct, 2 unique	1 distinct, 0 unique	2 distinct, 2 unique
and probably will."			
Flying Fortress"	William Wyler		United States

date_added	release_year	A ^B _C rating	A ^B _C duration	A ^B _C listed_in	A ^B _C descrip
		2 distinct, 2 unique	2 distinct, 2 unique	1 distinct, 0 unique	1 distinct, 1
null	null				
Error	Error	Classic Movies, Documentaries	This documentary centers on ...		

These steps ensured the data was in the best possible form for analysis.

Analysis Business questions:

-- 1. Find the Top Countries with the Most Content on Netflix

```

SELECT *
FROM (
    SELECT
        country AS country,
        COUNT(m.*) AS country_content,
        SUM(COUNT(m.*)) OVER() AS total_content,
        DENSE_RANK() OVER(ORDER BY COUNT(m.show_id) DESC) AS r
    FROM
        master m
    INNER JOIN countries c ON m.show_id = c.show_id
    GROUP BY 1
    ORDER BY 2 DESC
) AS tab
WHERE r <= 5;

```

	country character varying (25)	country_content bigint	total_content numeric	r bigint
1	United States	3686	9993	1
2	India	1045	9993	2
3	United Kingdom	804	9993	3
4	Canada	445	9993	4
5	France	391	9993	5

-- 2. Find the Most Common Rating for Movies and TV Shows

```
SELECT type, rating
```

```
FROM (
```

```
  SELECT
```

```
    type,
```

```
    rating,
```

```
    COUNT(*) AS rating_count,
```

```
    RANK() OVER(PARTITION BY type ORDER BY COUNT(*) DESC) AS rank
```

```
  FROM master
```

```
  GROUP BY type, rating
```

```
) AS tab
```

```
WHERE rank <= 3;
```

	type character varying (7) 🔒	rating character varying (20) 🔒
1	Movie	TV-MA
2	Movie	TV-14
3	Movie	R
4	TV Show	TV-MA
5	TV Show	TV-14
6	TV Show	TV-PG

-- 3. Identify the Longest Movie

SELECT

type,

title,

duration

FROM master

WHERE type = 'Movie'

ORDER BY CAST(SPLIT_PART(duration, ',', 1) AS INT) DESC;

	type character varying (7) 🔒	title character varying (110) 🔒	duration character varying (15) 🔒
5	Movie	Lock Your Girls In	233 min
6	Movie	Raya and Sakina	230 min
7	Movie	Once Upon a Time in America	229 min
8	Movie	Sangam	228 min
9	Movie	Lagaan	224 min
10	Movie	Jodhaa Akbar	214 min
11	Movie	Seven Souls in the Skull Castle: Season Moon Kagen	212 min
12	Movie	The Irishman	209 min
13	Movie	Kabhi Khushi Kabhie Gham	209 min
14	Movie	No Direction Home: Bob Dylan	208 min
Total rows: 1000 of 6131 Query complete 00:00:00.174 Ln 31, Col 1			

-- 4. Find Content Added in the Last 5 Years

SELECT *

FROM master

WHERE TO_DATE(date_added, 'MM/DD/YYYY') >= CURRENT_DATE - INTERVAL '5 years';

show_id [PK] character varying (7)	type character varying (7)	title character varying (110)	date_added character varying (10)	release_year character varying (5)	rating character varying (20)	duration character varying (10)
s1	Movie	Dick Johnson Is Dead	9/25/2021	2020	PG-13	90 min
s2	TV Show	Blood & Water	9/24/2021	2021	TV-MA	2 Seasons
s3	TV Show	Ganglands	9/24/2021	2021	TV-MA	1 Season
s4	TV Show	Jailbirds New Orleans	9/24/2021	2021	TV-MA	1 Season
s5	TV Show	Kota Factory	9/24/2021	2021	TV-MA	2 Seasons
s6	TV Show	Midnight Mass	9/24/2021	2021	TV-MA	1 Season
s7	Movie	My Little Pony: A Ne...	9/24/2021	2021	PG	91 min
s8	Movie	Sankofa	9/24/2021	1993	TV-MA	125 min

1 rows: 1000 of 3920 Query complete 00:00:00.183 Ln 38, Col 1

-- 5. Find each year and the average numbers of content release in Egypt on Netflix

WITH total_egypt_shows AS (

SELECT COUNT(show_id) AS total_shows

FROM countries

WHERE country = 'Egypt'

)

SELECT

m.release_year,

COUNT(m.show_id) AS total_release,

ROUND(COUNT(m.show_id)::numeric / total_shows * 100, 2) AS avg_release

FROM master m

INNER JOIN countries c ON m.show_id = c.show_id

CROSS JOIN total_egypt_shows

WHERE c.country = 'Egypt'

GROUP BY m.release_year, total_shows

ORDER BY avg_release DESC

LIMIT 5;

	release_year character varying (5)	total_release bigint	avg_release numeric
1	2018	15	12.82
2	2017	14	11.97
3	2012	9	7.69
4	2011	8	6.84
5	2016	7	5.98

-- 6. Find the Top 10 Actors Who Have Appeared in the Highest Number of Movies

```
SELECT
    actor,
    COUNT(*) AS movie_count
FROM casts
GROUP BY actor
ORDER BY movie_count DESC
LIMIT 10;
```

	actor character varying (100)	count bigint
1	Anupam Kher	43
2	Shah Rukh Khan	35
3	Julie Tejewani	33
4	Naseeruddin Shah	32
5	Takahiro Sakurai	32
6	Rupa Bhimani	31
7	Om Puri	30
8	Akshay Kumar	30
9	Yuki Kaii	29
Total rows: 1000 of 36388 Query complete		

-- 7. Number of Actors for Each Country as Percentage of Total Actors on Netflix

```
SELECT *,
        ROUND((actor_No::numeric / TotalActors::numeric) * 100, 2) AS percentage
FROM (
    SELECT
        country,
        COUNT(DISTINCT actor) AS actor_No,
        SUM(COUNT(DISTINCT actor)) OVER() AS TotalActors
    FROM casts a
    INNER JOIN countries c ON a.show_id = c.show_id
    GROUP BY 1
    ORDER BY 2 DESC
) AS tab;
```

	country character varying (25)	actor_no bigint	totalactors numeric	percentage% numeric
1	United States	15059	50752	29.67
2	India	3888	50752	7.66
3	United Kingdom	3856	50752	7.60
4	Canada	2734	50752	5.39
5	France	2696	50752	5.31
6	Japan	1791	50752	3.53
7	Germany	1560	50752	3.07
8	South Korea	1401	50752	2.76
9	Spain	1319	50752	2.60
Total rows: 115 of 115 Query complete 00:00:00.735 Ln 79, Col 1				

-- 8. Number of Directors for Each Country as Percentage of Total Directors on Netflix

```
SELECT *,
        ROUND((director_No::numeric / TotalDirector::numeric) * 100, 2) AS percentage
FROM (
    SELECT
        country,
        COUNT(DISTINCT director) AS director_No,
        SUM(COUNT(DISTINCT director)) OVER() AS TotalDirector
    FROM directors d
    INNER JOIN countries c ON d.show_id = c.show_id
    GROUP BY 1
    ORDER BY 2 DESC
) AS tab;
```

	country character varying (25)	director_no bigint	totaldirector numeric	percentage% numeric
1	United States	2243	6480	34.61
2	India	718	6480	11.08
3	United Kingdom	511	6480	7.89
4	France	321	6480	4.95
5	Canada	308	6480	4.75
6	Germany	183	6480	2.82
7	Spain	172	6480	2.65
8	Japan	120	6480	1.85
9	China	118	6480	1.82
Total rows: 117 of 117 Query complete 00:00:00.183 Ln 86, Col 1				

-- 9. Growth Content for Top 5 Countries with Content on Netflix (Since 2009)

WITH country_release AS (

SELECT

c.country,

m.release_year::INTEGER,

COUNT(m.show_id) AS current_release

FROM

master m

INNER JOIN countries c ON m.show_id = c.show_id

WHERE

m.release_year::INTEGER > 2009

AND c.country IN ('United States', 'India', 'United Kingdom', 'Canada', 'France')

GROUP BY

c.country, m.release_year

),

release_with_lag AS (

SELECT

country,

release_year,

current_release,

LAG(current_release, 1, 0) OVER (PARTITION BY country ORDER BY release_year) AS
previous_release

FROM country_release

)

SELECT

country,

release_year,

current_release,

previous_release,

ROUND((current_release - previous_release) * 100.0 / NULLIF(previous_release, 0), 2)
AS growth_rate

FROM release_with_lag

WHERE previous_release <> 0

ORDER BY country DESC, release_year;

	country character varying (25)	release_year integer	current_release bigint	previous_release bigint	growth_rate numeric
1	United States	2011	85	85	0.00
2	United States	2012	101	85	18.82
3	United States	2013	111	101	9.90
4	United States	2014	137	111	23.42
5	United States	2015	210	137	53.28
6	United States	2016	348	210	65.71
7	United States	2017	441	348	26.72
8	United States	2018	457	441	3.63
9	United States	2019	427	457	-6.56
Total rows: 55 of 55 Query complete 00:00:00.121 Ln 130, Col 32					

-- 10. Show the Difference in the Number of TV Shows Added Year-over-Year for Each Country

```
SELECT
    country,
    release_year,
    COUNT(*) AS tv_show_count,
    COALESCE(COUNT(*) - LAG(COUNT(*)) OVER (PARTITION BY country ORDER BY
release_year), 0) AS year_over_year_diff
FROM master m
INNER JOIN countries c ON m.show_id = c.show_id
WHERE type = 'TV Show'
GROUP BY country, release_year;
```

	country character varying (25) 🔒	release_year character varying (5) 🔒	tv_show_count bigint 🔒	year_over_year_diff bigint 🔒
1	Argentina	2006	1	0
2	Argentina	2015	2	1
3	Argentina	2016	1	-1
4	Argentina	2017	2	1
5	Argentina	2018	2	0
6	Argentina	2019	5	3
7	Argentina	2020	6	1
8	Argentina	2021	1	-5
9	Australia	1986	1	0
Total rows: 456 of 456 Query complete 00:00:00.108 Ln 137, Col 1				

-- 11. Top Casts for Distinct Genres

```
SELECT
    actor,
    COUNT(DISTINCT g.genre) AS total_genres
FROM casts c
INNER JOIN genres g ON c.show_id = g.show_id
GROUP BY actor
ORDER BY total_genres DESC;
```

	actor character varying (100)	total_genres bigint
1	Ron Perlman	17
2	Gary Cole	16
3	Kiernan Shipka	16
4	Glenn Close	15
5	Nicholas Hoult	14
6	John Leguizamo	14
7	Guy Pearce	14
8	Carla Gugino	14
9	Jay Baruchel	14
Total rows: 1000 of 36388 Query complete 00		

-- 12. Distinct Genres for Each Country (Genre Diversity)

```
SELECT
    country,
    COUNT(DISTINCT g.genre) AS distinct_genres,
    (SELECT COUNT(DISTINCT genre) FROM genres) AS total_genres
FROM countries c
INNER JOIN genres g ON c.show_id = g.show_id
GROUP BY country
ORDER BY distinct_genres DESC;
```

	country character varying (25) 🔒	count bigint 🔒	all_genres bigint 🔒
1	United States	42	42
2	United Kingdom	37	42
3	Canada	37	42
4	Australia	36	42
5	France	35	42
6	India	35	42
7	Japan	35	42
8	Spain	32	42
9	Germany	32	42
Total rows: 122 of 122 Query complete 00:00:00.204			



-- 13. Number of Age Groups for Each Country

```
SELECT
    country,
    COUNT(DISTINCT rating) AS distinct_agegroup
    ,(SELECT COUNT(DISTINCT rating) FROM master) AS total_rating
FROM
    countries c
    INNER JOIN master m ON c.show_id = m.show_id
GROUP BY
    country
ORDER BY
    distinct_agegroup DESC;
```

	country character varying (25) 🔒	distinct_agegroup bigint 🔒	total_rating bigint 🔒
1	United States	15	15
2	Canada	13	15
3	France	13	15
4	United Kingdom	12	15
5	Spain	12	15
6	Italy	12	15
Total rows: 122 of 122 Query complete 00:00:00.120 Ln 15			

-- 14. Top 5 Most Common Genres Across All Shows and Their Total Count

```
SELECT  
    rating,  
    COUNT(*) AS genre_count  
FROM master  
WHERE type = 'Movie'  
GROUP BY rating  
ORDER BY genre_count DESC  
LIMIT 5;
```

	rating character varying (20) 	genre_count bigint 
1	TV-MA	2062
2	TV-14	1427
3	R	797
4	TV-PG	540
5	PG-13	490

-- 15. Find the Next Show's Title (Lead) and Previous Show's Title (Lag) for Each Show Based on the Release Year

SELECT

title,

release_year,

LAG(title) OVER(ORDER BY release_year) AS previous_show,

LEAD(title) OVER(ORDER BY release_year) AS next_show

FROM master;

	title character varying (110)	release_year character varying (5)	previous_show character varying	next_show character varying
1	Pioneers: First Women Filmmakers*	1925	[null]	Prelude to War
2	Prelude to War	1942	Pioneers: First Women Filmmakers*	The Battle of Midway
3	The Battle of Midway	1942	Prelude to War	WWII: Report from the Aleutians
4	WWII: Report from the Aleutians	1943	The Battle of Midway	Why We Fight: The Battle of Russia
5	Why We Fight: The Battle of Russia	1943	WWII: Report from the Aleutians	Undercover: How to Operate Behind Enemy Lines
6	Undercover: How to Operate Behind Enemy Lines	1943	Why We Fight: The Battle of Russia	Tunisian Victory
7	Tunisian Victory	1944	Undercover: How to Operate Behin...	The Memphis Belle: A Story of a
8	The Memphis Belle: A Story of a	1944	Tunisian Victory	The Negro Soldier

-- 16. Top 5 Longest Movies by Duration and Their Directors

SELECT

title,

director,

duration

FROM master m

INNER JOIN directors d ON m.show_id = d.show_id

WHERE type = 'Movie' AND duration IS NOT NULL

ORDER BY CAST(SPLIT_PART(duration, ',', 1) AS INT) DESC

LIMIT 5;

	title character varying (110) 🔒	director character varying (35) 🔒	duration character varying (15) 🔒
1	The School of Mischief	Houssam El-Din Mustafa	253 min
2	No Longer kids	Samir Al Asfory	237 min
3	Lock Your Girls In	Fouad El-Mohandes	233 min
4	Raya and Sakina	Hussein Kamal	230 min
5	Once Upon a Time in America	Sergio Leone	229 min

-- 17. Percentage of TV Shows vs Movies Added in the Last 4 Years (2018-2021)

SELECT



type,

ROUND(COUNT(*) * 100.0 / SUM(COUNT(*) OVER(), 2) AS percentage

FROM master

WHERE EXTRACT(YEAR FROM TO_DATE(date_added, 'MM/DD/YYYY')) BETWEEN 2018
AND 2021

GROUP BY type;

	type character varying (7) 	percentage numeric 
1	Movie	70.12
2	TV Show	29.88

-- 18. Top 3 Countries with the Most Content in 2020 and 2021

WITH recent_shows AS (

SELECT

country,

COUNT(m.*) AS total_shows

FROM master m

INNER JOIN countries c ON m.show_id = c.show_id

WHERE release_year::NUMERIC IN (2020, 2021)

AND country IS NOT NULL

GROUP BY country

)

SELECT country, total_shows

FROM recent_shows

ORDER BY total_shows DESC

LIMIT 3;

	country character varying (25) 🔒	total_shows bigint 🔒
1	United States	557
2	India	112
3	United Kingdom	104

-- 19. Find the Top 5 Directors with the Most Shows Listed

```
SELECT *
FROM (
    SELECT
        director,
        COUNT(m.*) AS total_shows,
        DENSE_RANK() OVER (ORDER BY COUNT(m.*) DESC) AS rank
    FROM master m
    INNER JOIN directors d ON m.show_id = d.show_id
    WHERE director IS NOT NULL
    GROUP BY director
) AS tab
WHERE rank <= 5;
```

	director character varying (35) 🔒	total_shows bigint 🔒	rank bigint 🔒
1	Rajiv Chilaka	22	1
2	Jan Suter	21	2
3	RaÃ¶l Campos	19	3
4	Marcus Raboy	16	4
5	Suhas Kadav	16	4
6	Jay Karas	15	5

-- 20. Calculate the Average Duration of Movies by Rating

SELECT

rating,

ROUND(AVG(CAST(SPLIT_PART(duration, ',', 1) AS INT)), 2) AS avg_duration

FROM master

WHERE type = 'Movie' rating != 'Not detected'

GROUP BY rating;

	rating character varying (20) 🔒	avg_duration numeric 🔒
1	G	90.27
2	NC-17	125.00
3	NR	94.53
4	PG	98.28
5	PG-13	108.33
6	R	106.72
7	TV-14	110.29
8	TV-G	79.67
9	TV-MA	95.89

-- 21. Count the Number of Content Items in Each Genre

SELECT

genre,

COUNT(*) AS total_content

FROM genres



GROUP BY genre

ORDER BY total_content DESC;

	genre character varying (30)	total_content bigint
1	International Movies	2752
2	Dramas	2427
3	Comedies	1674
4	International TV Shows	1351
5	Documentaries	869
6	Action & Adventure	859
7	TV Dramas	763
8	Independent Movies	756
9	Children & Family Movies	641
Total rows: 42 of 42 Query complete 00:00:00.217		

-- 22. Find the Most Common Rating for Movies and TV Shows (Top 5 for Each Type)

```
SELECT
    type,
    rating
FROM (
    SELECT
        type,
        rating,
        COUNT(*) AS rating_count,
        RANK() OVER(PARTITION BY type ORDER BY COUNT(*) DESC) AS rank
    FROM master
    GROUP BY type, rating
) AS tab
WHERE rank <= 5;
```

	type character varying (7) 	rating character varying (20) 
1	Movie	TV-MA
2	Movie	TV-14
3	Movie	R
4	Movie	TV-PG
5	Movie	PG-13
6	TV Show	TV-MA
7	TV Show	TV-14
8	TV Show	TV-PG
9	TV Show	TV-Y7
Total rows: 10 of 10 Query complete 00:00:00.15		

-- 23. Running Total of Movies Added Year-Over-Year by Rating

```
SELECT
    rating,
    release_year,
    COUNT(*) AS yearly_movie_count,
    SUM(COUNT(*)) OVER (PARTITION BY rating ORDER BY release_year) AS running_total
FROM master
WHERE type = 'Movie'
GROUP BY rating, release_year;
```

	rating character varying (20) 🔒	release_year character varying (5) 🔒	yearly_movie_count bigint 🔒	running_total numeric 🔒
1	G	1956	1	1
2	G	1958	1	2
3	G	1964	1	3
4	G	1968	1	4
5	G	1969	1	5
6	G	1971	2	7
7	G	1973	1	8
8	G	1977	1	9
9	G	1986	1	10
Total rows: 408 of 408 Query complete 00:00:00.188 Ln 254, Col 31				

-- 24. Number of Content Items for Each Release Year

```
SELECT *,
    ROUND(content_ReleaseYear::numeric * 100 / total_content::numeric, 2) || '%' AS
    Year_percentage
FROM (
    SELECT
        release_year,
        COUNT(show_id) AS content_ReleaseYear,
        SUM(COUNT(show_id)) OVER() AS total_content
    FROM master
    GROUP BY release_year
    ORDER BY release_year DESC
) AS tab;
```

	release_year character varying (5)	content_releaseyear bigint	total_content numeric	year_percentage text
1	2021	592	8807	6.72%
2	2020	953	8807	10.82%
3	2019	1030	8807	11.70%
4	2018	1147	8807	13.02%
5	2017	1032	8807	11.72%
6	2016	902	8807	10.24%
7	2015	560	8807	6.36%
8	2014	352	8807	4.00%
9	2013	288	8807	3.27%
Total rows: 74 of 74 Query complete 00:00:00.229 Ln 257, Col 1				

-- 25. Number of Content Items Added into Netflix Each Year

```
SELECT *,
        ROUND(content_Added::numeric * 100 / total_content::numeric, 2) || '%' AS
Year_percentage
FROM (
    SELECT
        EXTRACT(YEAR FROM date_added::DATE) AS Netflix_year,
        COUNT(show_id) AS content_Added,
        SUM(COUNT(show_id)) OVER() AS total_content
    FROM master
    WHERE date_added IS NOT NULL
    GROUP BY 1
    ORDER BY 1 DESC
) AS tab;
```

	netflix_year numeric	content_added bigint	total_content numeric	year_percentage text
1	2021	1498	8797	17.03%
2	2020	1879	8797	21.36%
3	2019	2016	8797	22.92%
4	2018	1649	8797	18.75%
5	2017	1188	8797	13.50%
6	2016	429	8797	4.88%
7	2015	82	8797	0.93%
8	2014	24	8797	0.27%
9	2013	11	8797	0.13%
Total rows: 14 of 14 Query complete 00:00:00.123 Ln 266, Col 1				

Dashboard:



Insights & Recommendations

Netflix Content Insights (1945-2021)

1. Total Content Released:

- Netflix has released 3,735 Movies and 1,955 TV Shows from 1945 to 2021 (a span of 66 years). The number of movies is nearly double that of TV shows, highlighting Netflix's strong focus on movie content.

2. Top Countries by Content Production:

- United States, India, and Japan are the top three countries producing the most titles, leading the global content contribution on Netflix.

3. Release Trends in the Last Decade:

- In the last decade, the peak release for Movies was in 2019, with 540 movie titles, while TV Shows reached their highest release in 2020, with 306 titles.
- Content production showed a steady increase over the decade, peaking in 2019 for movies and 2020 for TV shows. However, content releases experienced a sharp decline starting in 2021.

4. Age Ratings Distribution:

- Netflix content is classified under 11 different age certifications. The majority of titles are rated TV-MA, followed by R-rated content.
- This suggests that the primary audience on Netflix is mature viewers, with most content being suitable for audiences aged 17 and older.

5. Content Duration Analysis:

- The largest portion of movie titles, 1,823 titles, fall within the 90-120 minutes duration range, making this the most common movie length.
- Titles with durations between 30-60 minutes come second, suggesting a demand for short-form content, particularly among TV Shows and mini-series.

6. Actor Participation:

- The year 2019 saw the highest number of actors participating in titles, with 8,892 actors featured across Netflix's content catalog that year.

7. Director Participation:

- Similarly, 2019 also saw the highest participation of directors, with 623 directors involved in content production.
- Raúl Campos stands out as the most prolific director, having directed 21 titles on Netflix, making him a leading figure in content direction on the platform.

Recommendations

- **Expand Regional Content:** Netflix should continue producing and acquiring content from top-performing regions like U.S., India, the U.K., and emerging markets such as South Korea and Spain.
- **Decline in 2021:** The sharp drop in content releases in 2021 could be attributed to external factors, such as the global pandemic. Netflix can focus on rebounding content production in 2022 and beyond by re-engaging directors and actors from past peak years.
- **Diversify Age Ratings:** Increase the availability of family-friendly and youth-oriented content to attract a younger demographic and family audiences.
- **Leverage Popular Genres and Talent:** Focus on promoting and expanding dominant genres (e.g., **Comedy, Drama, International**) and high-performing actors to enhance user engagement.
- ***Tailor Recommendations by Region:** Use insights into genre preferences by country to deliver more personalized content recommendations and promotions.
- **Focus on Short-form Content:** Catering to audiences looking for quick content options (e.g., under 60 minutes) can capture the attention of viewers with limited time.

- **Data-Driven Production Decisions:** Utilize historical performance data to guide decisions on future content creation and acquisition strategies, ensuring high engagement and viewer satisfaction.
- **Culture regions**