Hyun Joo Lee, Navyasree Sriramoju

CIS 5550-01

Netflix Data Analysis

December 15th, 2023

California State University, Los Angeles

Submitted to

Dr. Taimoor A Khan

Index

Netflix Data Analysis

Introduction

In our data analysis endeavor, we investigate a dataset that captures the breadth of Netflix's collection, spanning movies and TV series. As a leading entity in the entertainment realm, Netflix commands an audience of more than 200 million subscribers globally. This dataset serves as a gateway to understanding the viewing habits, content evolution, and regional content strategies, delivering vital knowledge that could benefit Netflix and its counterparts in the industry.

Our investigation has been a deep dive into Netflix's diverse content portfolio, employing various visual tools to illustrate the platform's content dynamics. We have used visual analytics to dissect the composition of Netflix’s offerings, track content production across different regions, monitor the pattern of releases over time, and categorize the content by viewer ratings.

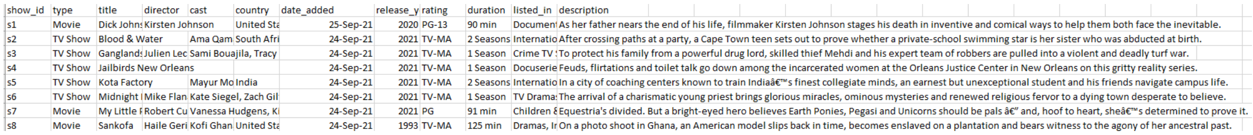
For instance, our pie charts reveal a dominant preference for films over series. At the same time, our trend lines provide a historical perspective on how content releases have fluctuated, especially in the face of worldwide disruptions such as the pandemic. Additionally, our maps highlight the hotspots of content creation that supply Netflix's extensive library. Moreover, the bar charts we've created illuminate the range of content maturity levels on Netflix, emphasizing material suited for mature viewers. However, a considerable portion caters to younger audiences as well.

These visualizations lay the foundation for our comprehensive analysis, offering numerical evidence to support our qualitative interpretations. Through this lens, we seek to devise strategies that could refine Netflix's approach to content curation, optimize its promotional activities, and amplify audience engagement. The insights gained from these visuals are crucial for a deeper understanding of Netflix's content strategy and identifying potential innovation avenues in an industry constantly in flux.

Data Collection

Dataset Link: https://www.kaggle.com/datasets/shivamb/netflix-shows

We collected this Dataset from Kagle. The Netflix dataset provides a detailed overview of the streaming platform's vast collection of movies and TV shows. It is structured in CSV format and includes a variety of fields. We selected this Dataset as it has diverse Fields like Show id, type(either TV Show or Movie), title, director, Cast members of the movie, Country, date_added, release_year, rating of the movie, duration, listen in what format, and Description through which we can delve into Netflix's content trends and strategy evolution offers a prime opportunity for insightful analysis. Analyzing this data can uncover trends, preferences, and patterns pivotal in understanding content performance, aiding future content creation.



show_id	type	title	director	cast	country	date_added	release_y	rating	duration	listed_in	description
s1	Movie	Dick Johns	Kirsten Johnson		United Sta	25-Sep-21	2020	PG-13	90 min	Document	As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable.
s2	TV Show	Blood & Water		Ama Qam	South Afri	24-Sep-21	2021	TV-MA	2 Seasons	Internatio	After crossing paths at a party, a Cape Town teen sets out to prove whether a private-school swimming star is her sister who was abducted at birth.
s3	TV Show	Ganglands	Julien Lec	Sami Bouajila, Tracy		24-Sep-21	2021	TV-MA	1 Season	Crime TV	To protect his family from a powerful drug lord, skilled thief Mehdi and his expert team of robbers are pulled into a violent and deadly turf war.
s4	TV Show	Jailbirds New Orleans				24-Sep-21	2021	TV-MA	1 Season	Docuserie	Feuds, flirtations and toilet talk go down among the incarcerated women at the Orleans Justice Center in New Orleans on this gritty reality series.
s5	TV Show	Kota Factory		Mayur Mo	India	24-Sep-21	2021	TV-MA	2 Seasons	Internatio	In a city of coaching centers known to train Indiaâ€™s finest collegiate minds, an earnest but unexceptional student and his friends navigate campus life.
s6	TV Show	Midnight	Mike Flan	Kate Siegel, Zach Gil		24-Sep-21	2021	TV-MA	1 Season	TV Drama	The arrival of a charismatic young priest brings glorious miracles, ominous mysteries and renewed religious fervor to a dying town desperate to believe.
s7	Movie	My Little	Robert Cu	Vanessa Hudgens, Ki		24-Sep-21	2021	PG	91 min	Children &	Equestria's divided. But a bright-eyed hero believes Earth Ponies, Pegasi and Unicorns should be pals â€” and, hoof to heart, sheâ€™s determined to prove it.
s8	Movie	Sankofa	Haile Geri	Kofi Ghan	United Sta	24-Sep-21	1993	TV-MA	125 min	Dramas, Ir	On a photo shoot in Ghana, an American model slips back in time, becomes enslaved on a plantation and bears witness to the agony of her ancestral past.

Sample Screenshot of Dataset

Field Name	Data Description	Example Value
show_id	Unique identifier for each title	s1

type	The format of the title (Movie/TV Show)	Movie
title	Title of the movie or TV show	Dick Johnson Is Dead
director	Name(s) of the director(s)	Kirsten Johnson
cast	List of actors and actresses involved	“Ama Qamata, Khosi Ngema,...”
country	Country or countries where the title was produced	United States
date_added	Date when the title was added to Netflix	25-Sep-21
release_year	The year when the title was originally released	2020
rating	The age rating of the title	PG-13
duration	Duration of the title (in minutes or number of seasons)	90 min
listed_in	Genres or categories the title is listed under	Documentaries

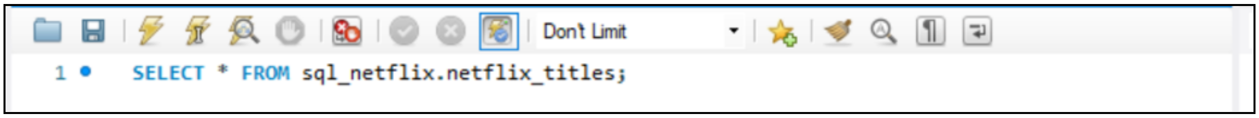
description	A brief description of the title	“As her father nears the…”

Data Cleaning

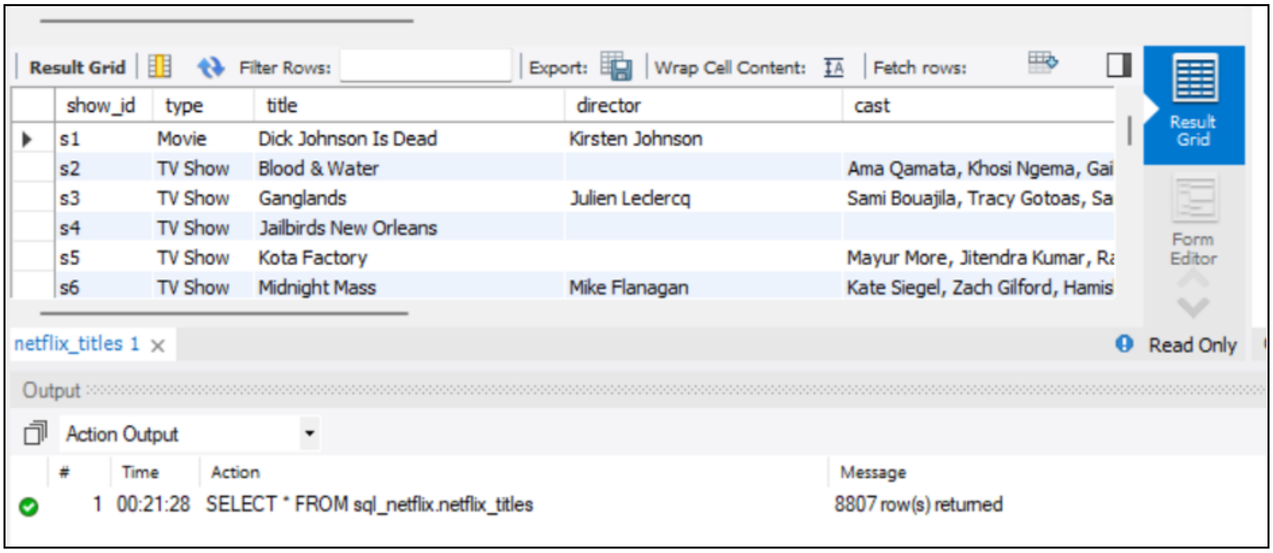
In the process of enhancing the Netflix titles dataset, comprehensive data cleaning was conducted using SQL. Null values were effectively handled by employing strategies such as imputation or removal, ensuring data completeness and accuracy. Duplicates were identified and eliminated, streamlining the dataset for analysis. Missing rows were populated where necessary, mitigating gaps and bolstering the dataset's comprehensiveness. To optimize the dataset's relevance, irrelevant columns were dropped, refining the dataset to contain only pertinent information. Additionally, column splitting was performed, organizing data into more granular segments for improved analysis and insights. These meticulous data cleaning procedures were instrumental in preparing a refined, robust dataset conducive to in-depth analysis and meaningful conclusions for the project.

View the dataset:

SELECT * FROM sql_netflix.netflix_titles;



```
SELECT * FROM sql_netflix.netflix_titles;
```

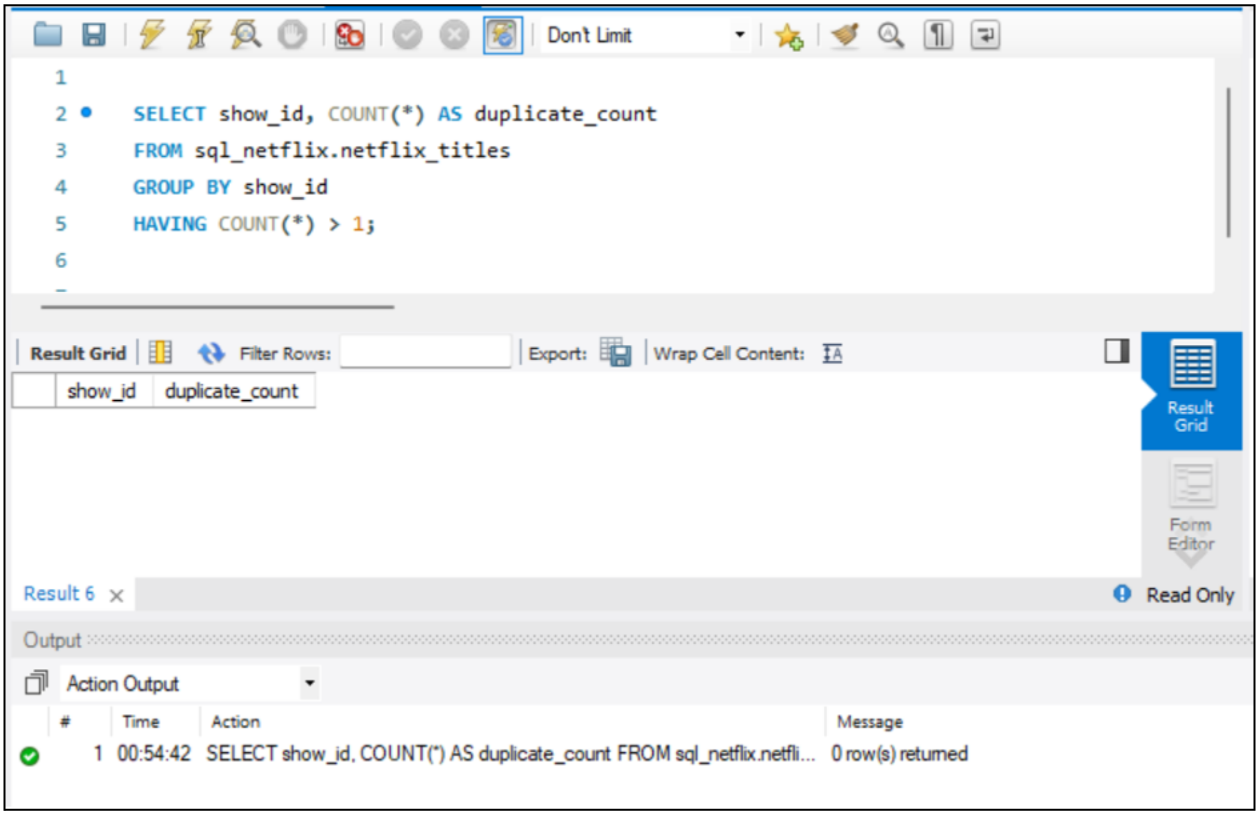


show_id	type	title	director	cast
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	
s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gai
s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Sa
s4	TV Show	Jailbirds New Orleans		
s5	TV Show	Kota Factory		Mayur More, Jitendra Kumar, Ra
s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamis

```
#   Time      Action                                    Message
1   00:21:28  SELECT * FROM sql_netflix.netflix_titles  8807 row(s) returned
```

Check for Duplicates:

SELECT show_id, COUNT(*) AS duplicate_count
FROM sql_netflix.netflix_titles
GROUP BY show_id
HAVING COUNT(*) > 1;



```sql
SELECT show_id, COUNT(*) AS duplicate_count
FROM sql_netflix.netflix_titles
GROUP BY show_id
HAVING COUNT(*) > 1;
```

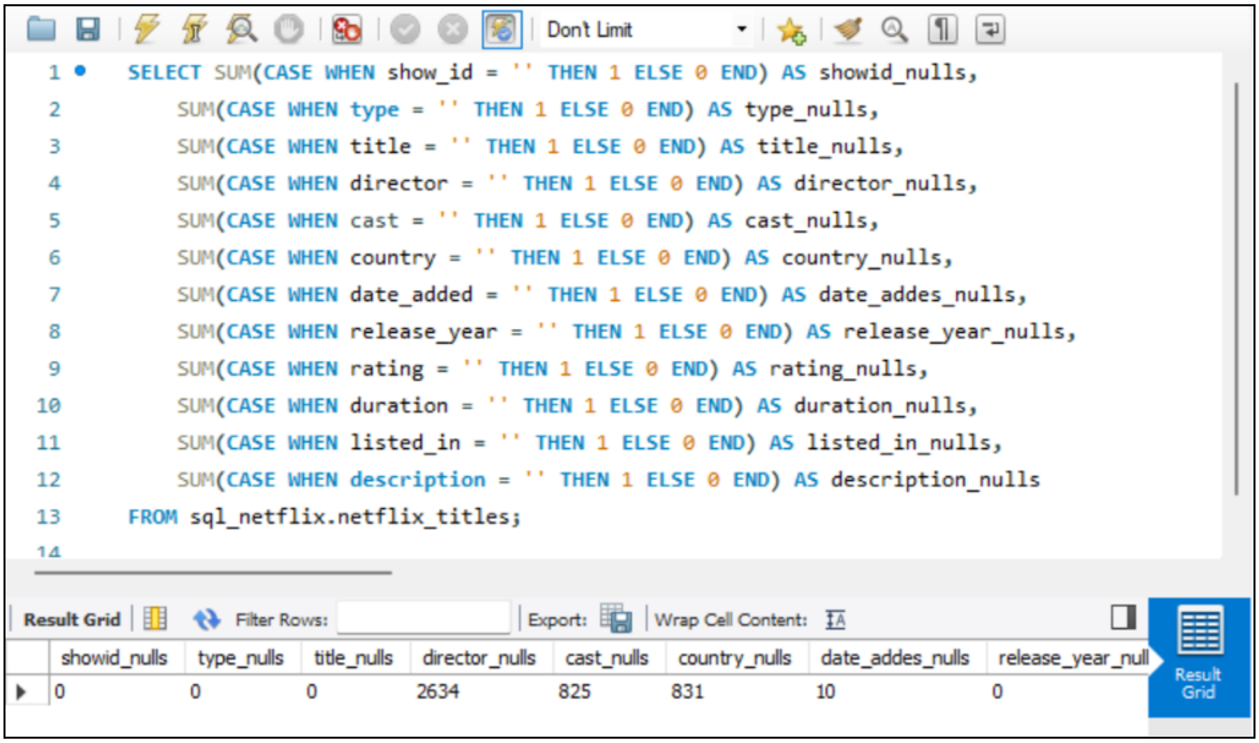
show_id	duplicate_count

```
#   Time      Action                                                      Message
1   00:54:42  SELECT show_id, COUNT(*) AS duplicate_count FROM sql_netflix.netfli...  0 row(s) returned
```

No duplicates found

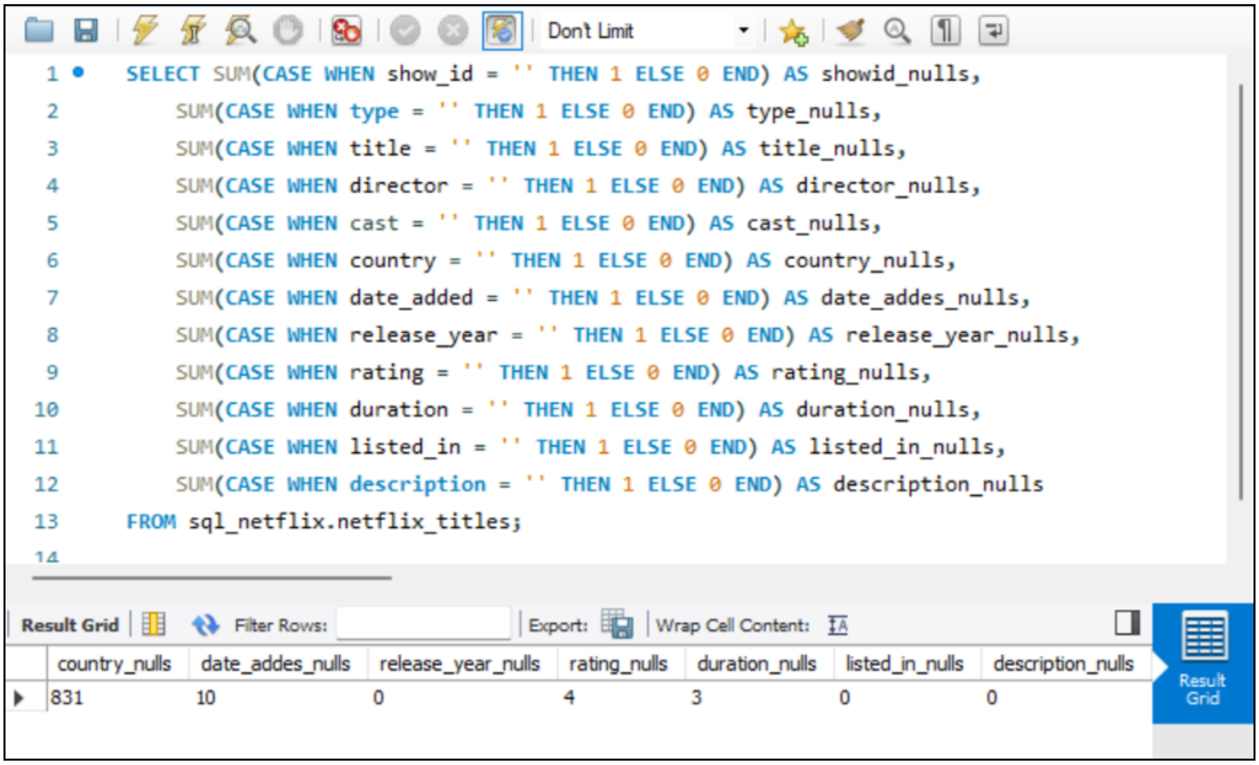
Check for Null Values:

```
SELECT SUM(CASE WHEN show_id = " THEN 1 ELSE 0 END) AS showid_nulls,
   SUM(CASE WHEN type = " THEN 1 ELSE 0 END) AS type_nulls,
   SUM(CASE WHEN title = " THEN 1 ELSE 0 END) AS title_nulls,
   SUM(CASE WHEN director = " THEN 1 ELSE 0 END) AS director_nulls,
   SUM(CASE WHEN cast = " THEN 1 ELSE 0 END) AS cast_nulls,
   SUM(CASE WHEN country = " THEN 1 ELSE 0 END) AS country_nulls,
   SUM(CASE WHEN date_added = " THEN 1 ELSE 0 END) AS date_addes_nulls,
   SUM(CASE WHEN release_year = " THEN 1 ELSE 0 END) AS release_year_nulls,
   SUM(CASE WHEN rating = " THEN 1 ELSE 0 END) AS rating_nulls,
   SUM(CASE WHEN duration = " THEN 1 ELSE 0 END) AS duration_nulls,
   SUM(CASE WHEN listed_in = " THEN 1 ELSE 0 END) AS listed_in_nulls,
   SUM(CASE WHEN description = " THEN 1 ELSE 0 END) AS description_nulls
FROM sql_netflix.netflix_titles;
```



```
SELECT SUM(CASE WHEN show_id = '' THEN 1 ELSE 0 END) AS showid_nulls,
    SUM(CASE WHEN type = '' THEN 1 ELSE 0 END) AS type_nulls,
    SUM(CASE WHEN title = '' THEN 1 ELSE 0 END) AS title_nulls,
    SUM(CASE WHEN director = '' THEN 1 ELSE 0 END) AS director_nulls,
    SUM(CASE WHEN cast = '' THEN 1 ELSE 0 END) AS cast_nulls,
    SUM(CASE WHEN country = '' THEN 1 ELSE 0 END) AS country_nulls,
    SUM(CASE WHEN date_added = '' THEN 1 ELSE 0 END) AS date_addes_nulls,
    SUM(CASE WHEN release_year = '' THEN 1 ELSE 0 END) AS release_year_nulls,
    SUM(CASE WHEN rating = '' THEN 1 ELSE 0 END) AS rating_nulls,
    SUM(CASE WHEN duration = '' THEN 1 ELSE 0 END) AS duration_nulls,
    SUM(CASE WHEN listed_in = '' THEN 1 ELSE 0 END) AS listed_in_nulls,
    SUM(CASE WHEN description = '' THEN 1 ELSE 0 END) AS description_nulls
FROM sql_netflix.netflix_titles;
```

showid_nulls	type_nulls	title_nulls	director_nulls	cast_nulls	country_nulls	date_addes_nulls	release_year_null
0	0	0	2634	825	831	10	0



```sql
SELECT SUM(CASE WHEN show_id = '' THEN 1 ELSE 0 END) AS showid_nulls,
    SUM(CASE WHEN type = '' THEN 1 ELSE 0 END) AS type_nulls,
    SUM(CASE WHEN title = '' THEN 1 ELSE 0 END) AS title_nulls,
    SUM(CASE WHEN director = '' THEN 1 ELSE 0 END) AS director_nulls,
    SUM(CASE WHEN cast = '' THEN 1 ELSE 0 END) AS cast_nulls,
    SUM(CASE WHEN country = '' THEN 1 ELSE 0 END) AS country_nulls,
    SUM(CASE WHEN date_added = '' THEN 1 ELSE 0 END) AS date_addes_nulls,
    SUM(CASE WHEN release_year = '' THEN 1 ELSE 0 END) AS release_year_nulls,
    SUM(CASE WHEN rating = '' THEN 1 ELSE 0 END) AS rating_nulls,
    SUM(CASE WHEN duration = '' THEN 1 ELSE 0 END) AS duration_nulls,
    SUM(CASE WHEN listed_in = '' THEN 1 ELSE 0 END) AS listed_in_nulls,
    SUM(CASE WHEN description = '' THEN 1 ELSE 0 END) AS description_nulls
FROM sql_netflix.netflix_titles;
```

country_nulls	date_addes_nulls	release_year_nulls	rating_nulls	duration_nulls	listed_in_nulls	description_nulls
831	10	0	4	3	0	0

This SQL query uses conditional aggregation to count empty or null values in all columns within the table. By employing `SUM` and `CASE` statements, it evaluates each column to identify empty cells. If a column contains an empty string (''), it assigns a value of 1; otherwise, it assigns 0 for each row. The `SUM` function then aggregates these values, providing a count of missing data for each attribute in the Netflix titles dataset.

We can see that there are NULLS.

director_nulls = 2634

cast_nulls = 825

country_nulls = 831

date_added_nulls = 10

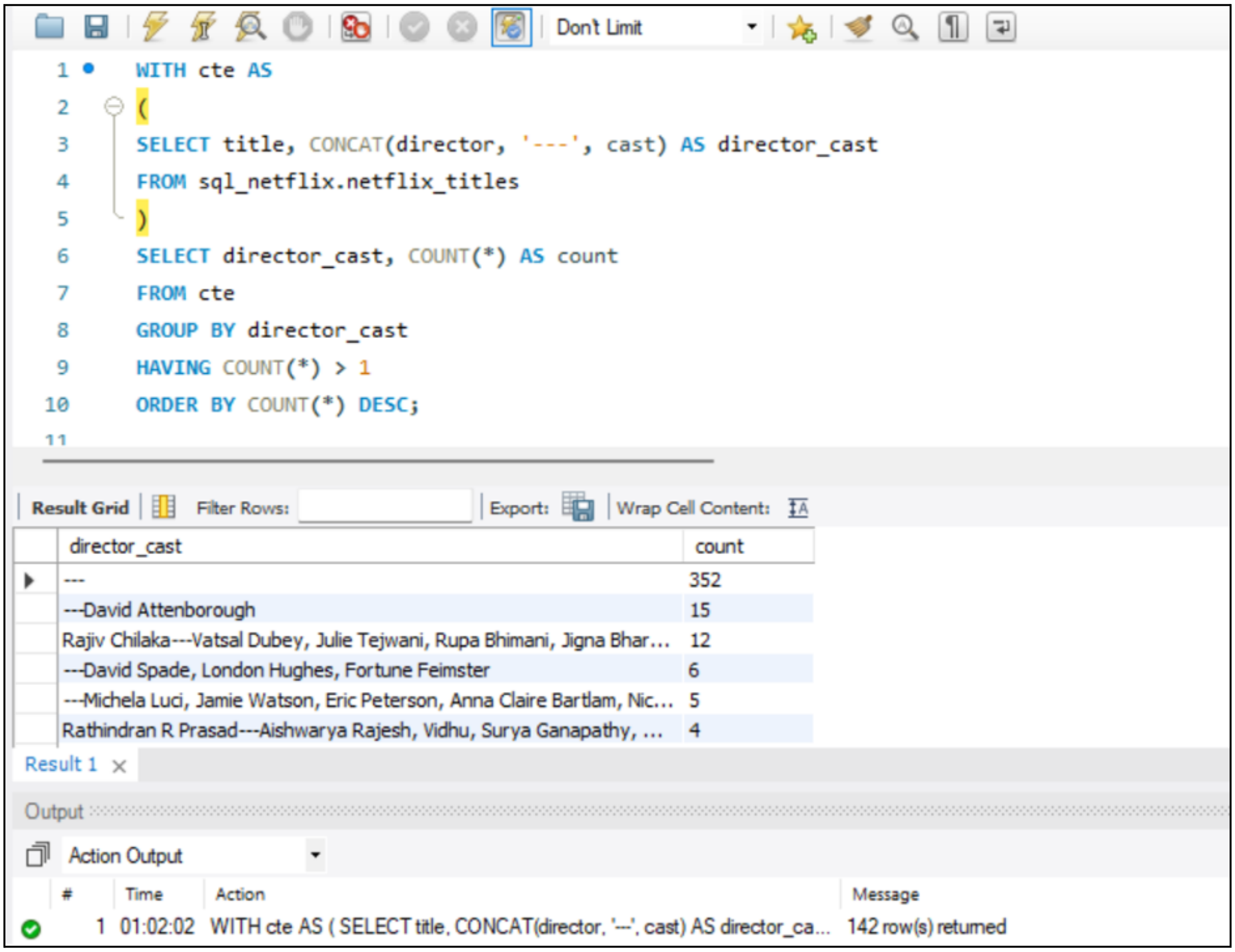
rating_nulls = 4

duration_nulls = 3

Populate the Rows:

We've discovered that about 30% of the director column contains null value. Instead of removing these entries, we rather have another column to populate it. We are utilizing the information from the 'cast' column to populate the missing 'director' data.
We created a field named 'director_cast' and grouped the results and counted how often each appears to identify the most frequent collaborations. With the output from previous Query, we updated the Director column by comparing Cast members

```
WITH cte AS
(
SELECT title, CONCAT(director, ' --- ', cast) AS director_cast
FROM sql_netflix.netflix_titles
)
SELECT director_cast, COUNT(*) AS count
FROM cte
GROUP BY director_cast
HAVING COUNT(*) > 1
ORDER BY COUNT(*) DESC;
```

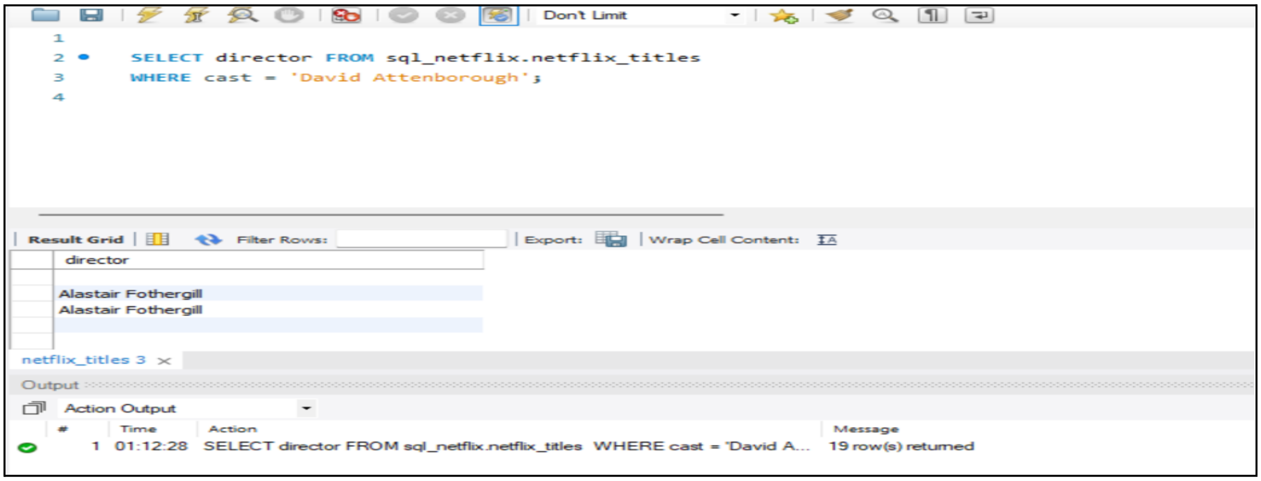


```sql
WITH cte AS
(
SELECT title, CONCAT(director, '---', cast) AS director_cast
FROM sql_netflix.netflix_titles
)
SELECT director_cast, COUNT(*) AS count
FROM cte
GROUP BY director_cast
HAVING COUNT(*) > 1
ORDER BY COUNT(*) DESC;
```

director_cast	count
---	352
---David Attenborough	15
Rajiv Chilaka---Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhar...	12
---David Spade, London Hughes, Fortune Feimster	6
---Michela Luci, Jamie Watson, Eric Peterson, Anna Claire Bartlam, Nic...	5
Rathindran R Prasad---Aishwarya Rajesh, Vidhu, Surya Ganapathy, ...	4

```
#   Time      Action                                                                   Message
1   01:02:02  WITH cte AS ( SELECT title, CONCAT(director, '---', cast) AS director_ca...  142 row(s) returned
```

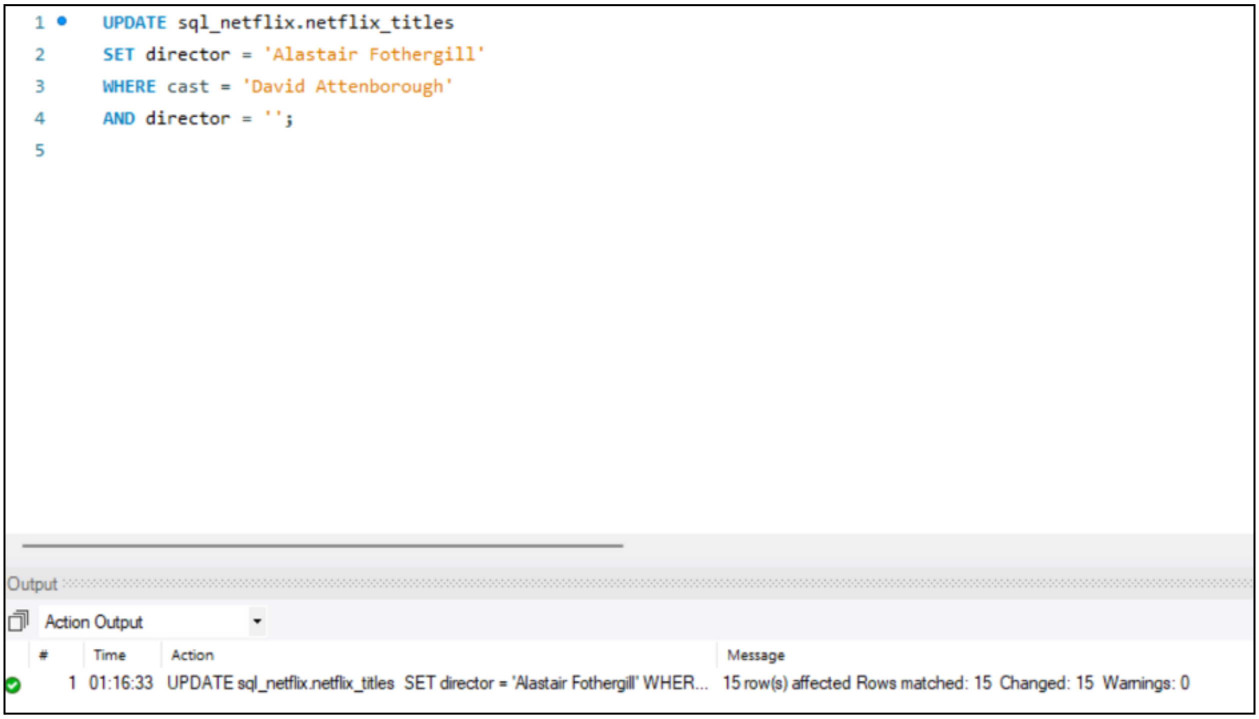
SELECT director FROM sql_netflix.netflix_titles
WHERE cast = 'David Attenborough';



```sql
SELECT director FROM sql_netflix.netflix_titles
WHERE cast = 'David Attenborough';
```

director
Alastair Fothergill
Alastair Fothergill

```
#   Time      Action                                                                Message
1   01:12:28  SELECT director FROM sql_netflix.netflix_titles  WHERE cast = 'David A...  19 row(s) returned
```



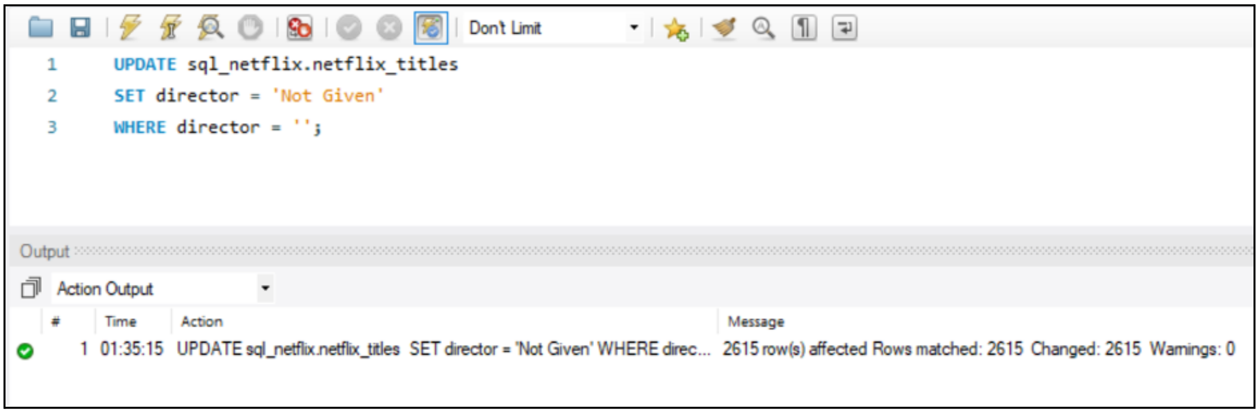
```
UPDATE sql_netflix.netflix_titles
SET director = 'Alastair Fothergill'
WHERE cast = 'David Attenborough'
AND director = '';
```

```
#   Time       Action                                                                 Message
1   01:16:33   UPDATE sql_netflix.netflix_titles  SET director = 'Alastair Fothergill' WHER...   15 row(s) affected Rows matched: 15  Changed: 15  Warnings: 0
```

Repeat this step to populate the rest of the director nulls and Populate the rest of the NULL in director as "Not Given".

For those records where a director could not be determined, we took a different approach. Instead of leaving the field empty, we updated the 'director' column to 'Not Given'. This ensures that our dataset remains clean and we avoid any misleading analysis due to blank fields.

UPDATE sql_netflix.netflix_titles
SET director = 'Not Given'
WHERE director = '';



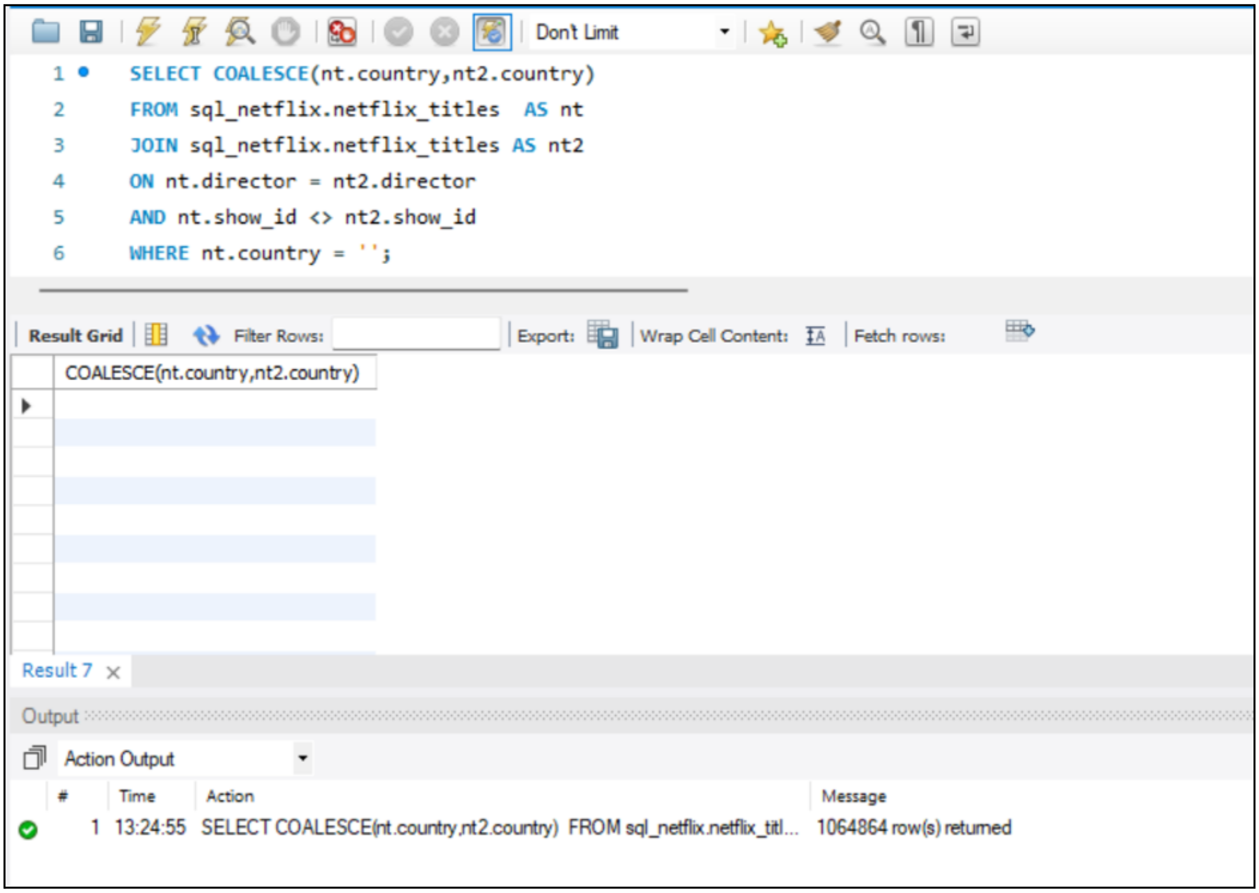
```
UPDATE sql_netflix.netflix_titles
SET director = 'Not Given'
WHERE director = '';
```

```
#   Time       Action                                                                 Message
1   01:35:15   UPDATE sql_netflix.netflix_titles  SET director = 'Not Given' WHERE direc...   2615 row(s) affected Rows matched: 2615  Changed: 2615  Warnings: 0
```

Populate the country using the director column

```
SELECT COALESCE(nt.country,nt2.country)
FROM sql_netflix.netflix_titles AS nt
JOIN sql_netflix.netflix_titles AS nt2
ON nt.director = nt2.director
AND nt.show_id <> nt2.show_id
WHERE nt.country = ' ';
```

SQL query selects and retrieves the country values associated with Netflix titles from the 'sql_netflix.netflix_titles' table based on a self-join operation using the 'director' column. It uses the COALESCE function to handle NULL values between two instances of the same table (aliased as nt and nt2). The condition 'nt.show_id <> nt2.show_id' ensures that the join occurs between different Netflix titles directed by the same individual.



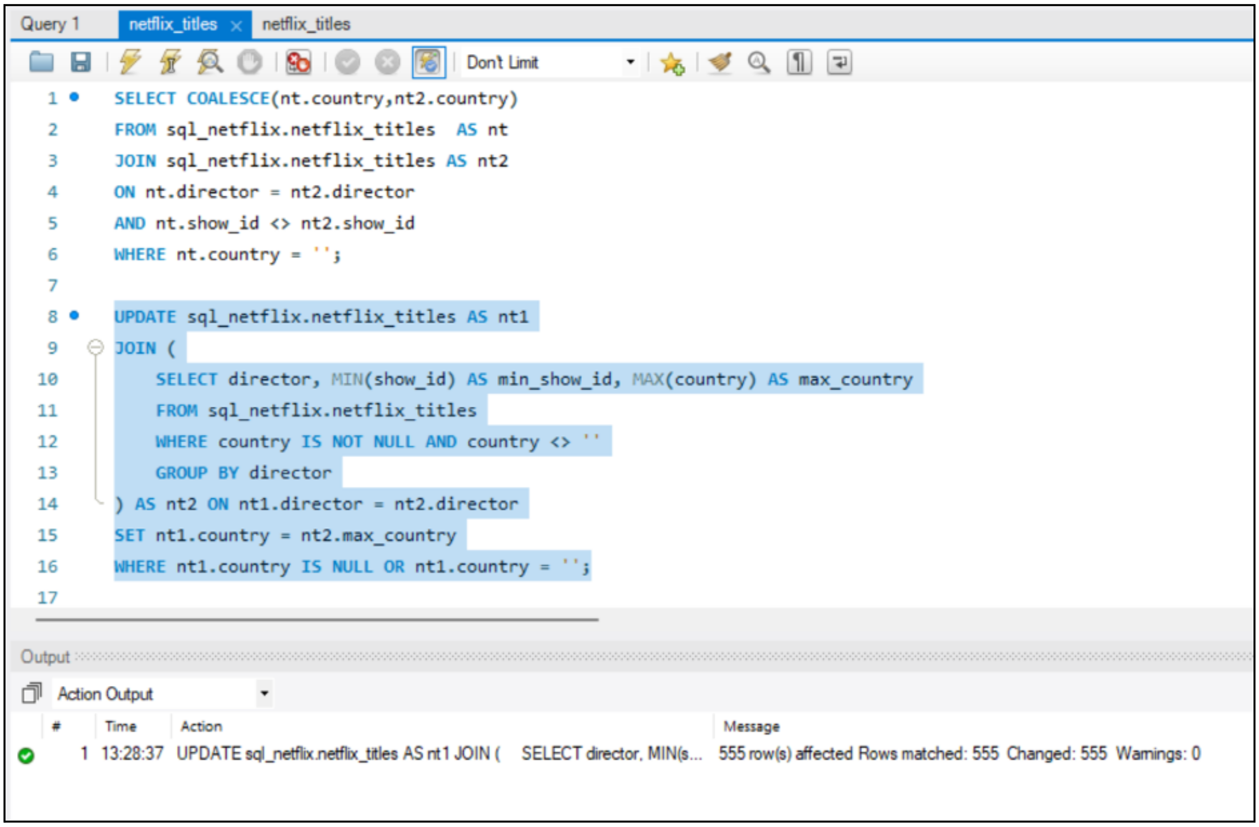
```
SELECT COALESCE(nt.country,nt2.country)
FROM sql_netflix.netflix_titles  AS nt
JOIN sql_netflix.netflix_titles AS nt2
ON nt.director = nt2.director
AND nt.show_id <> nt2.show_id
WHERE nt.country = '';
```

```
COALESCE(nt.country,nt2.country)
```

```
1  13:24:55  SELECT COALESCE(nt.country,nt2.country) FROM sql_netflix.netflix_titl...  1064864 row(s) returned
```

Just like the director column, we will not delete the nulls in country. Since the country column is related to director and movie, we are going to populate the country column with the director column

```
UPDATE sql_netflix.netflix_titles AS nt1
JOIN (
   SELECT director, MIN(show_id) AS min_show_id, MAX(country) AS max_country
   FROM sql_netflix.netflix_titles
   WHERE country IS NOT NULL AND country <> "
   GROUP BY director
) AS nt2 ON nt1.director = nt2.director
SET nt1.country = nt2.max_country
WHERE nt1.country IS NULL OR nt1.country = ' ';
```



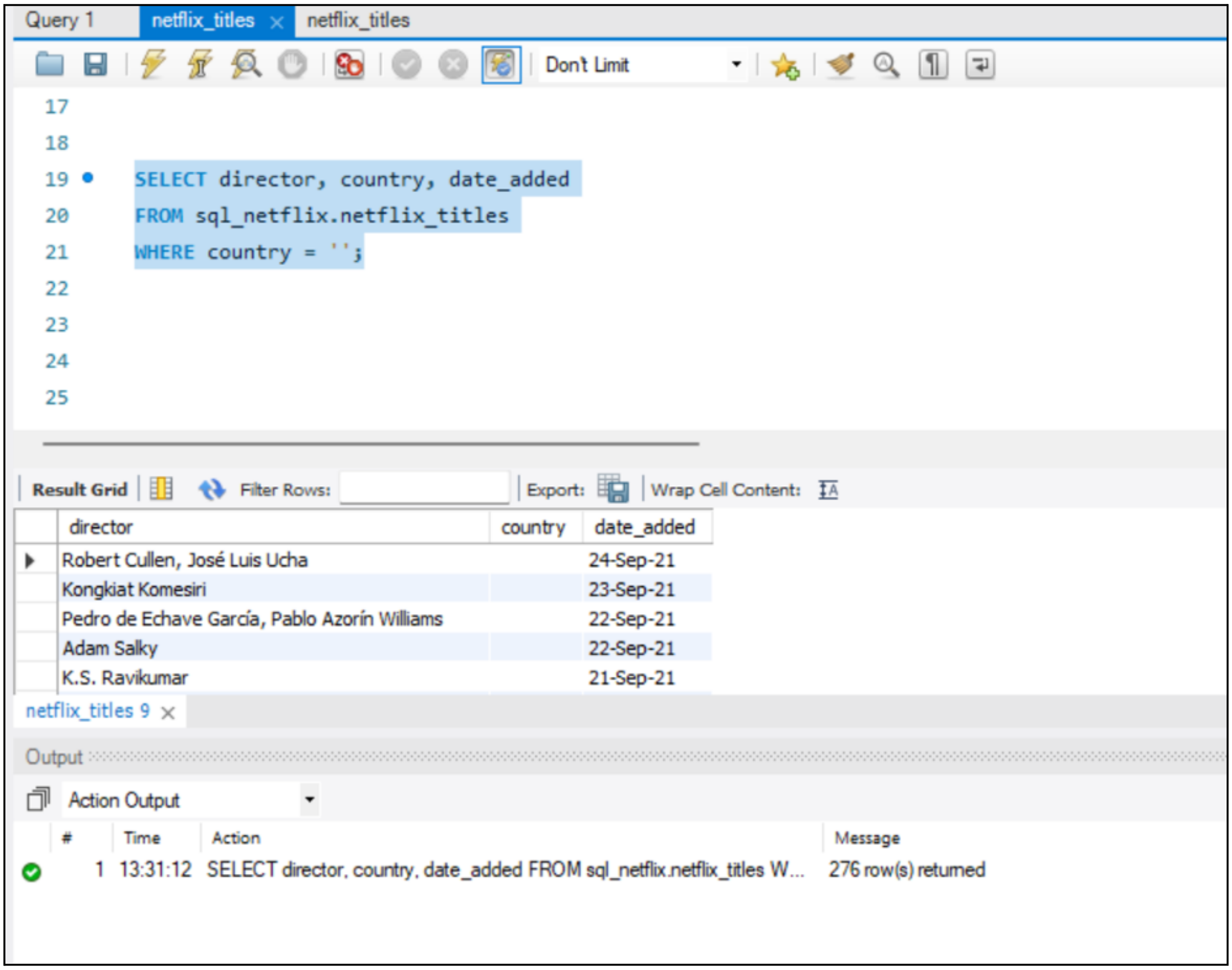
```
SELECT COALESCE(nt.country,nt2.country)
FROM sql_netflix.netflix_titles  AS nt
JOIN sql_netflix.netflix_titles AS nt2
ON nt.director = nt2.director
AND nt.show_id <> nt2.show_id
WHERE nt.country = '';

UPDATE sql_netflix.netflix_titles AS nt1
JOIN (
    SELECT director, MIN(show_id) AS min_show_id, MAX(country) AS max_country
    FROM sql_netflix.netflix_titles
    WHERE country IS NOT NULL AND country <> ''
    GROUP BY director
) AS nt2 ON nt1.director = nt2.director
SET nt1.country = nt2.max_country
WHERE nt1.country IS NULL OR nt1.country = '';
```

```
   #  Time      Action                                                          Message
   1  13:28:37  UPDATE sql_netflix.netflix_titles AS nt1 JOIN (   SELECT director, MIN(s...  555 row(s) affected Rows matched: 555  Changed: 555  Warnings: 0
```

To confirm if there are still directors linked to country that refuse to update

```
SELECT director, country, date_added
FROM sql_netflix.netflix_titles
WHERE country = ' ';
```



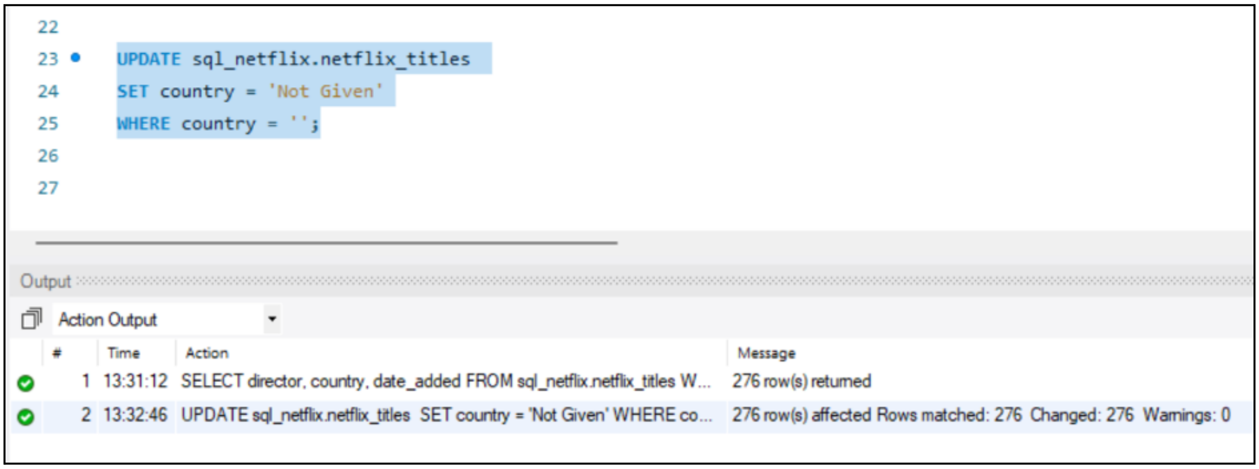
```
SELECT director, country, date_added
FROM sql_netflix.netflix_titles
WHERE country = '';
```

director	country	date_added
Robert Cullen, José Luis Ucha		24-Sep-21
Kongkiat Komesiri		23-Sep-21
Pedro de Echave García, Pablo Azorín Williams		22-Sep-21
Adam Salky		22-Sep-21
K.S. Ravikumar		21-Sep-21

#	Time	Action	Message
1	13:31:12	SELECT director, country, date_added FROM sql_netflix.netflix_titles W...	276 row(s) returned

Populate the rest of the NULL in director as "Not Given"

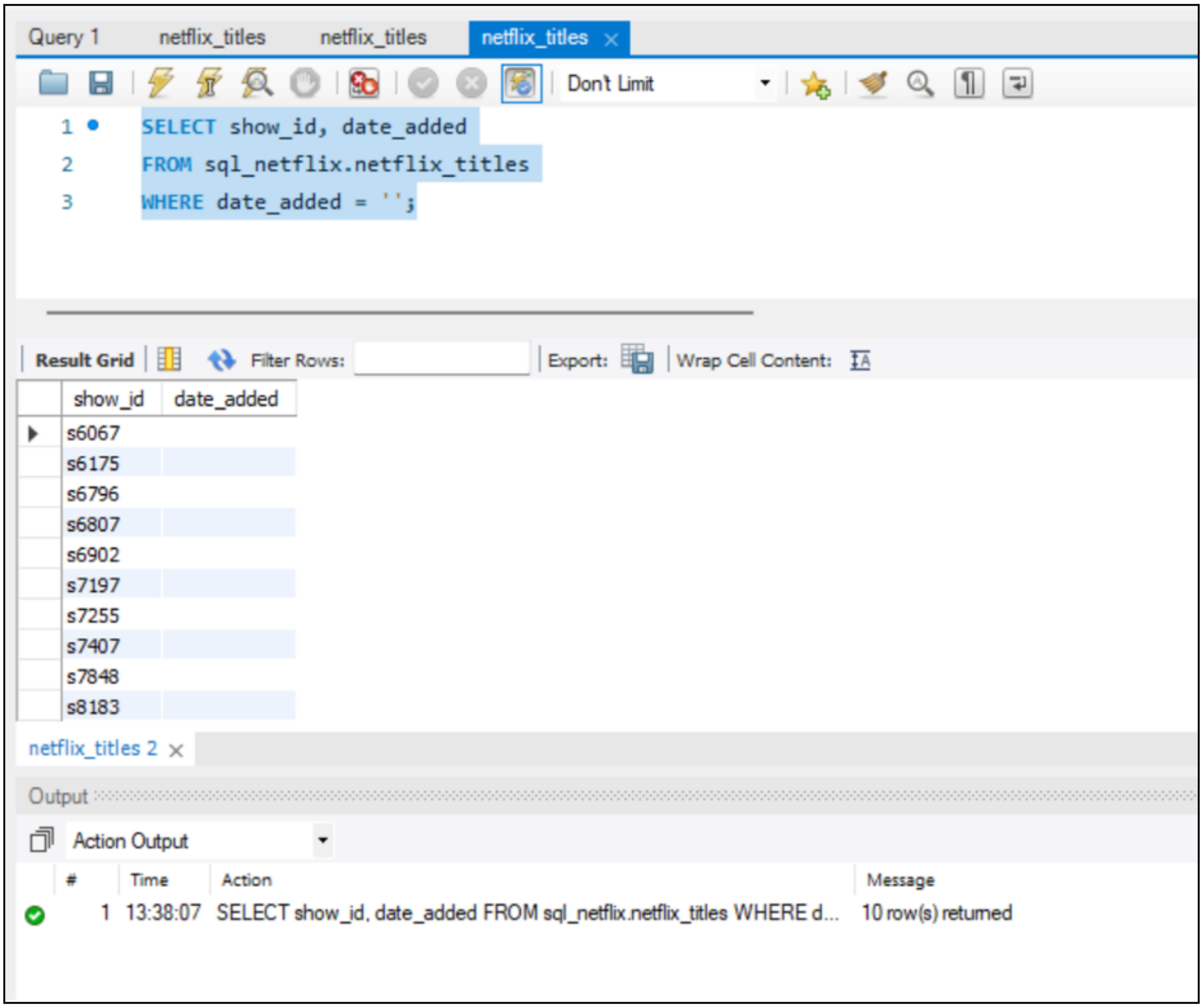
UPDATE sql_netflix.netflix_titles
SET country = 'Not Given'
WHERE country = ' ';



```
UPDATE sql_netflix.netflix_titles
SET country = 'Not Given'
WHERE country = '';
```

	#	Time	Action	Message
✓	1	13:31:12	SELECT director, country, date_added FROM sql_netflix.netflix_titles W...	276 row(s) returned
✓	2	13:32:46	UPDATE sql_netflix.netflix_titles SET country = 'Not Given' WHERE co...	276 row(s) affected Rows matched: 276 Changed: 276 Warnings: 0

Show date_added nulls:
SELECT show_id, date_added
FROM sql_netflix.netflix_titles
WHERE date_added = ";



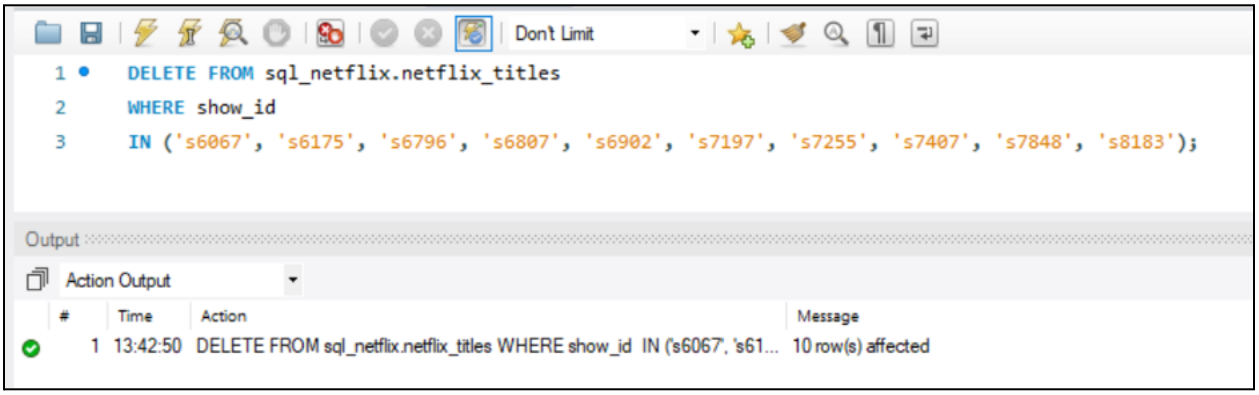
```
SELECT show_id, date_added
FROM sql_netflix.netflix_titles
WHERE date_added = '';
```

show_id	date_added
s6067	
s6175	
s6796	
s6807	
s6902	
s7197	
s7255	
s7407	
s7848	
s8183	

#	Time	Action	Message
1	13:38:07	SELECT show_id, date_added FROM sql_netflix.netflix_titles WHERE d...	10 row(s) returned

Delete Nulls:

DELETE FROM sql_netflix.netflix_titles
WHERE show_id
IN ('s6067', 's6175', 's6796', 's6807', 's6902', 's7197', 's7255', 's7407', 's7848', 's8183');

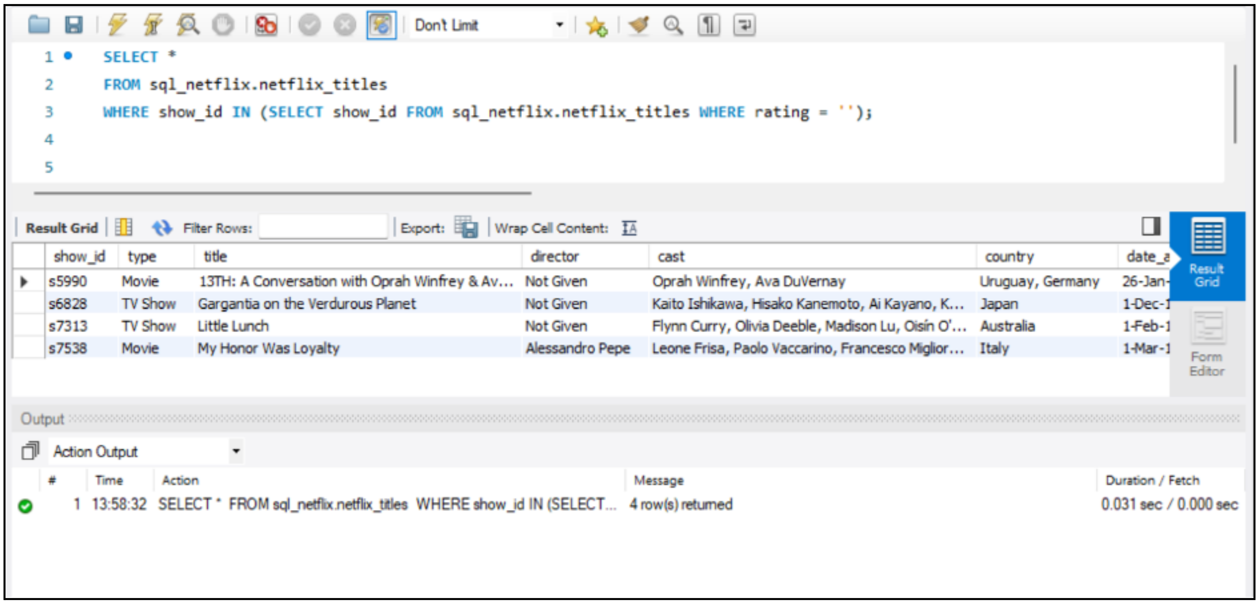


```
DELETE FROM sql_netflix.netflix_titles
WHERE show_id
IN ('s6067', 's6175', 's6796', 's6807', 's6902', 's7197', 's7255', 's7407', 's7848', 's8183');
```

	#	Time	Action	Message
✓	1	13:42:50	DELETE FROM sql_netflix.netflix_titles WHERE show_id IN ('s6067', 's61...	10 row(s) affected

Show rating NULLS:

SELECT *
FROM sql_netflix.netflix_titles
WHERE show_id IN (SELECT show_id FROM sql_netflix.netflix_titles WHERE rating = '');



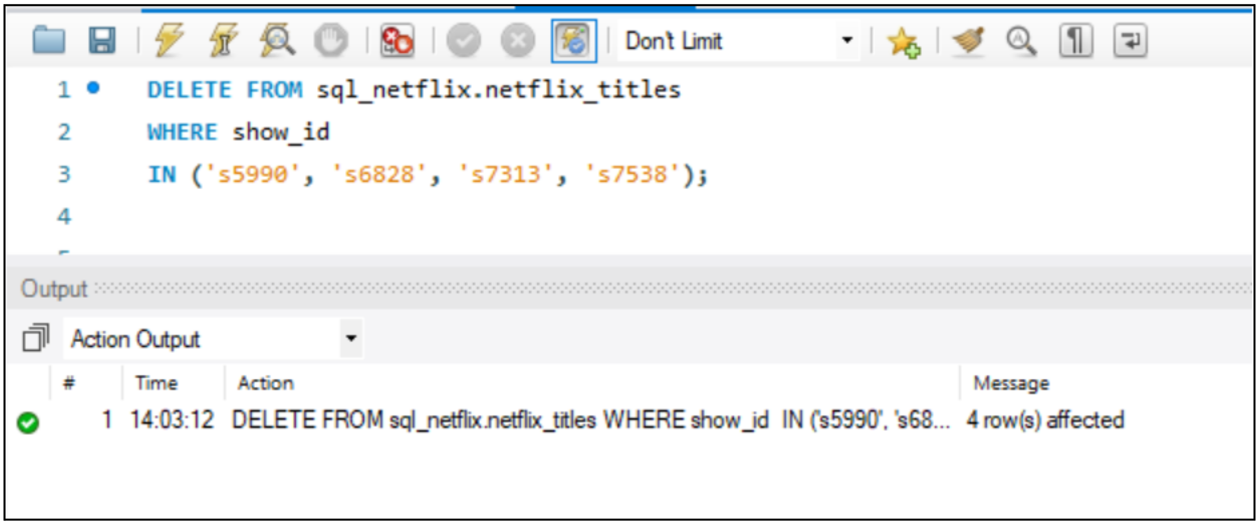
```
SELECT *
FROM sql_netflix.netflix_titles
WHERE show_id IN (SELECT show_id FROM sql_netflix.netflix_titles WHERE rating = '');
```

show_id	type	title	director	cast	country	date_a
s5990	Movie	13TH: A Conversation with Oprah Winfrey & Av...	Not Given	Oprah Winfrey, Ava DuVernay	Uruguay, Germany	26-Jan-
s6828	TV Show	Gargantia on the Verdurous Planet	Not Given	Kaito Ishikawa, Hisako Kanemoto, Ai Kayano, K...	Japan	1-Dec-1
s7313	TV Show	Little Lunch	Not Given	Flynn Curry, Olivia Deeble, Madison Lu, Oisín O'...	Australia	1-Feb-1
s7538	Movie	My Honor Was Loyalty	Alessandro Pepe	Leone Frisa, Paolo Vaccarino, Francesco Miglior...	Italy	1-Mar-1

	#	Time	Action	Message	Duration / Fetch
✓	1	13:58:32	SELECT * FROM sql_netflix.netflix_titles WHERE show_id IN (SELECT...	4 row(s) returned	0.031 sec / 0.000 sec

Delete the Null Values of Rating:

DELETE FROM sql_netflix.netflix_titles
WHERE show_id
IN ('s5990', 's6828', 's7313', 's7538');



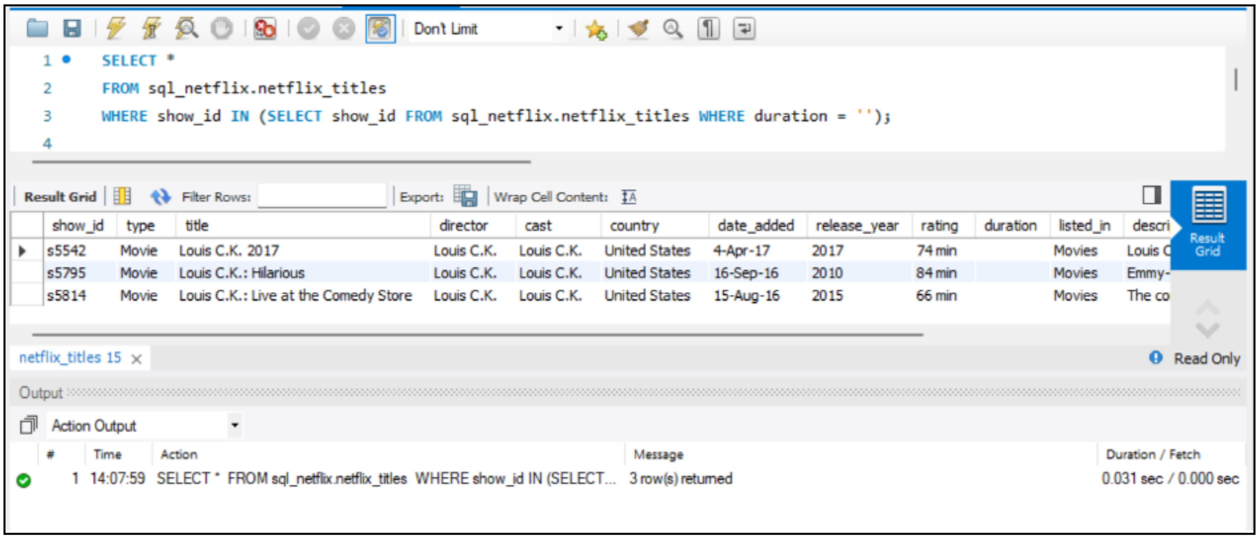
```
DELETE FROM sql_netflix.netflix_titles
WHERE show_id
IN ('s5990', 's6828', 's7313', 's7538');
```

Action Output

```
#  Time      Action                                                            Message
1  14:03:12  DELETE FROM sql_netflix.netflix_titles WHERE show_id  IN ('s5990', 's68...  4 row(s) affected
```

View the Duration nulls:

SELECT *
FROM sql_netflix.netflix_titles
WHERE show_id IN (SELECT show_id FROM sql_netflix.netflix_titles WHERE duration = '');



```
SELECT *
FROM sql_netflix.netflix_titles
WHERE show_id IN (SELECT show_id FROM sql_netflix.netflix_titles WHERE duration = '');
```

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descri
s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	4-Apr-17	2017	74 min		Movies	Louis C
s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	16-Sep-16	2010	84 min		Movies	Emmy-
s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	15-Aug-16	2015	66 min		Movies	The co

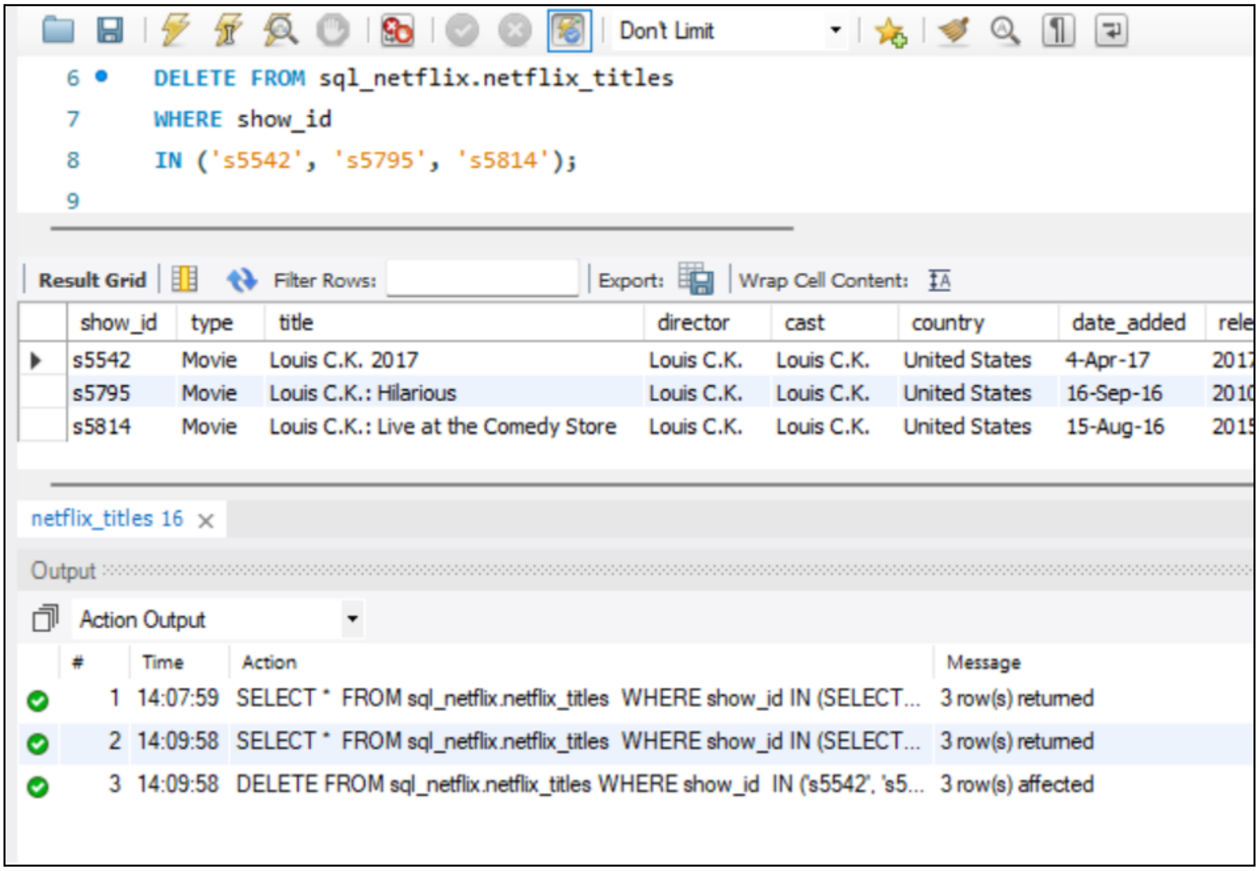
netflix_titles 15

Action Output

```
#  Time      Action                                                           Message              Duration / Fetch
1  14:07:59  SELECT *  FROM sql_netflix.netflix_titles  WHERE show_id IN (SELECT...  3 row(s) returned    0.031 sec / 0.000 sec
```

Delete Duration Nulls:

DELETE FROM sql_netflix.netflix_titles
WHERE show_id
IN ('s5542', 's5795', 's5814');

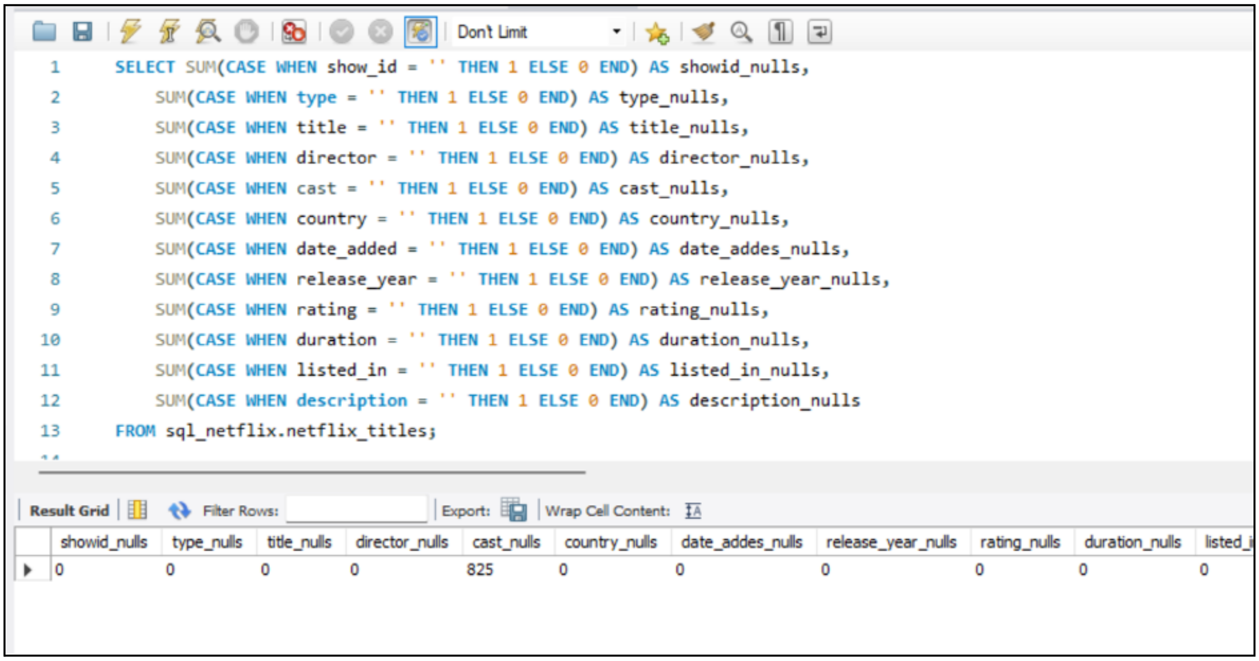


```sql
DELETE FROM sql_netflix.netflix_titles
WHERE show_id
IN ('s5542', 's5795', 's5814');
```

show_id	type	title	director	cast	country	date_added	rele
s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	4-Apr-17	2017
s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	16-Sep-16	2010
s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	15-Aug-16	2015

#	Time	Action	Message
1	14:07:59	SELECT * FROM sql_netflix.netflix_titles WHERE show_id IN (SELECT...	3 row(s) returned
2	14:09:58	SELECT * FROM sql_netflix.netflix_titles WHERE show_id IN (SELECT...	3 row(s) returned
3	14:09:58	DELETE FROM sql_netflix.netflix_titles WHERE show_id IN ('s5542', 's5...	3 row(s) affected

Now check for the nulls in all Columns again



```sql
SELECT SUM(CASE WHEN show_id = '' THEN 1 ELSE 0 END) AS showid_nulls,
    SUM(CASE WHEN type = '' THEN 1 ELSE 0 END) AS type_nulls,
    SUM(CASE WHEN title = '' THEN 1 ELSE 0 END) AS title_nulls,
    SUM(CASE WHEN director = '' THEN 1 ELSE 0 END) AS director_nulls,
    SUM(CASE WHEN cast = '' THEN 1 ELSE 0 END) AS cast_nulls,
    SUM(CASE WHEN country = '' THEN 1 ELSE 0 END) AS country_nulls,
    SUM(CASE WHEN date_added = '' THEN 1 ELSE 0 END) AS date_addes_nulls,
    SUM(CASE WHEN release_year = '' THEN 1 ELSE 0 END) AS release_year_nulls,
    SUM(CASE WHEN rating = '' THEN 1 ELSE 0 END) AS rating_nulls,
    SUM(CASE WHEN duration = '' THEN 1 ELSE 0 END) AS duration_nulls,
    SUM(CASE WHEN listed_in = '' THEN 1 ELSE 0 END) AS listed_in_nulls,
    SUM(CASE WHEN description = '' THEN 1 ELSE 0 END) AS description_nulls
FROM sql_netflix.netflix_titles;
```

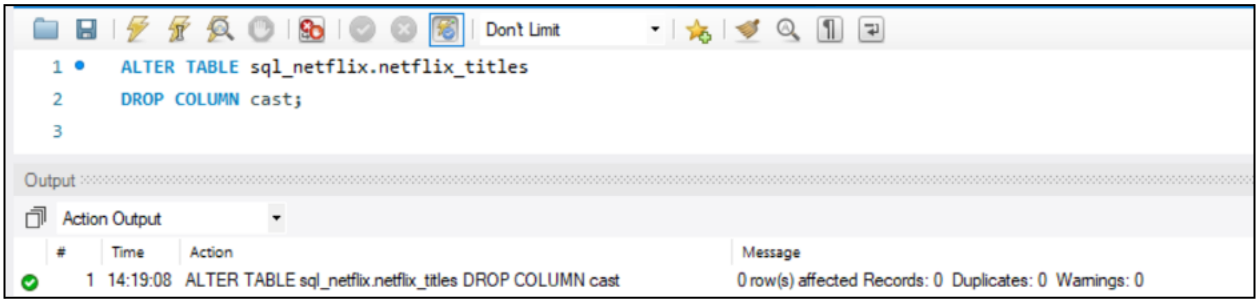
showid_nulls	type_nulls	title_nulls	director_nulls	cast_nulls	country_nulls	date_addes_nulls	release_year_nulls	rating_nulls	duration_nulls	listed_i
0	0	0	0	825	0	0	0	0	0	0

After Clearing and populating the Nulls in all cells, Only Cast Column is left with 825 Nulls which is too many and it has many irregularities as well.

Delete Unneeded Columns:

We've decided to remove the 'cast' column from our dataset. The reasoning is that the field contains multiple names, creating a level of complexity that isn't necessary for our analysis. By dropping this column, we simplify our data and ensure that our dataset is streamlined. We executed a simple query using DROP COLUMN.

ALTER TABLE sql_netflix.netflix_titles
DROP COLUMN cast;



```sql
ALTER TABLE sql_netflix.netflix_titles
DROP COLUMN cast;
```

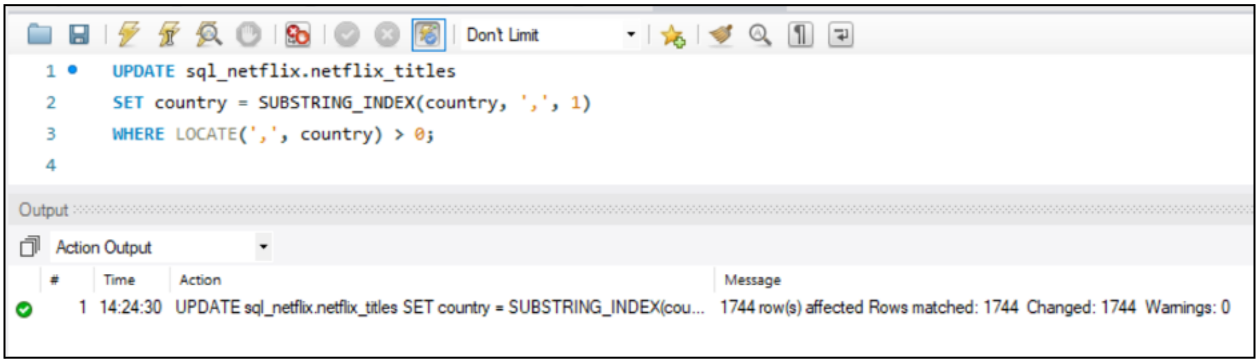
#	Time	Action	Message
1	14:19:08	ALTER TABLE sql_netflix.netflix_titles DROP COLUMN cast	0 row(s) affected Records: 0 Duplicates: 0 Warnings: 0

Split Columns:

We found out that some of the rows in the country column have multiple countries. For visualization, only one country is needed per row to create map visualization clean and easy. Therefore, we're going to split the country column and retain the first country by the left which would be the original country of the movie.

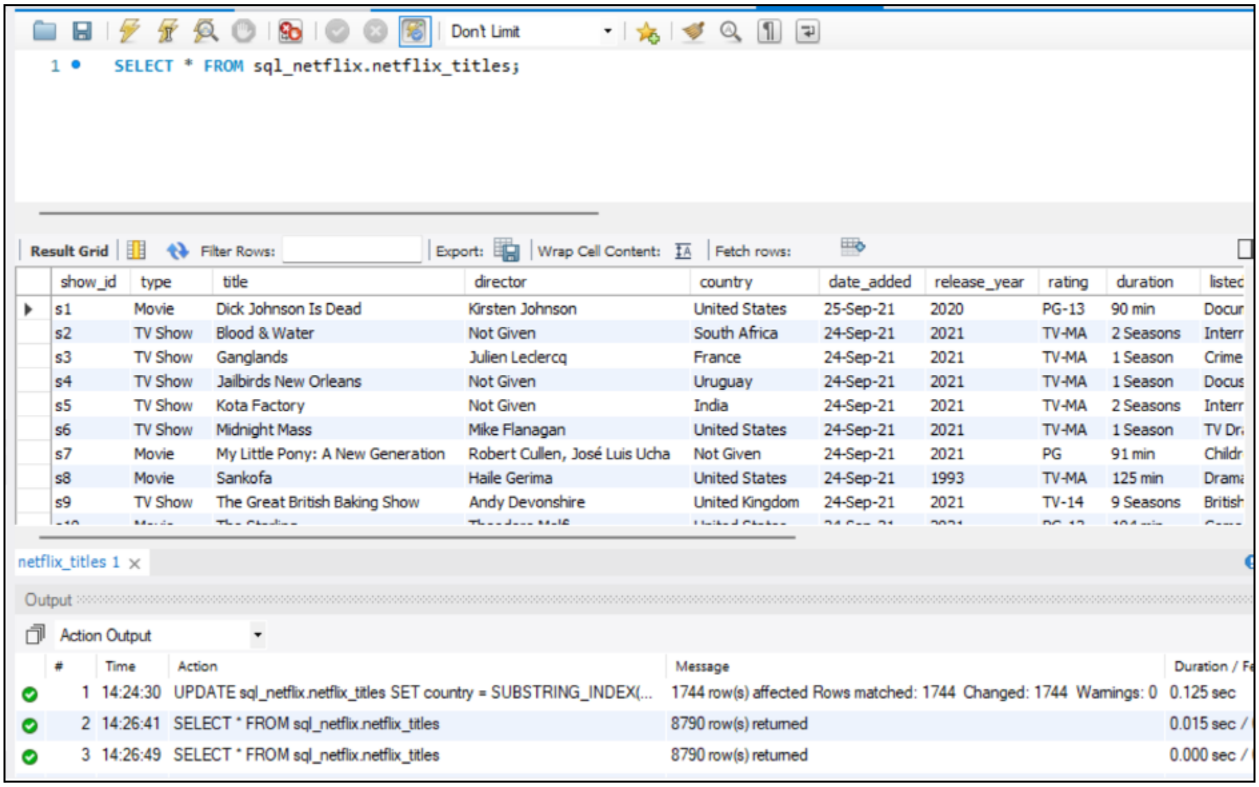
We used the function 'SUBSTRING_INDEX' and it helps us extract the country before any comma, which we assume is the main country of origin for the title.

```
UPDATE sql_netflix.netflix_titles
SET country = SUBSTRING_INDEX(country, ',', 1)
WHERE LOCATE(',', country) > 0;
```



```
UPDATE sql_netflix.netflix_titles
SET country = SUBSTRING_INDEX(country, ',', 1)
WHERE LOCATE(',', country) > 0;
```

```
#   Time      Action                                                     Message
1   14:24:30  UPDATE sql_netflix.netflix_titles SET country = SUBSTRING_INDEX(cou...   1744 row(s) affected Rows matched: 1744  Changed: 1744  Warnings: 0
```



```sql
SELECT * FROM sql_netflix.netflix_titles;
```

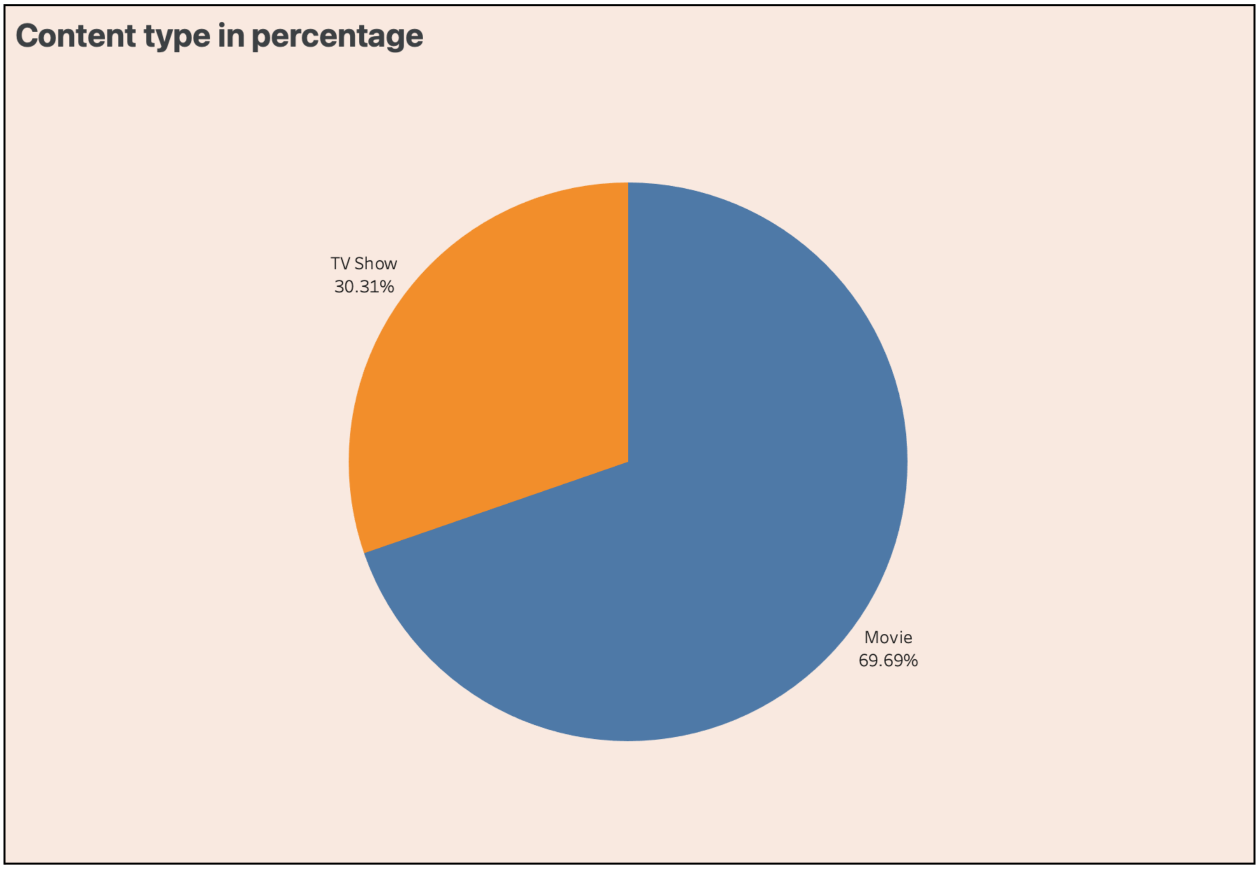
show_id	type	title	director	country	date_added	release_year	rating	duration	listec
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	25-Sep-21	2020	PG-13	90 min	Docur
s2	TV Show	Blood & Water	Not Given	South Africa	24-Sep-21	2021	TV-MA	2 Seasons	Interr
s3	TV Show	Ganglands	Julien Leclercq	France	24-Sep-21	2021	TV-MA	1 Season	Crime
s4	TV Show	Jailbirds New Orleans	Not Given	Uruguay	24-Sep-21	2021	TV-MA	1 Season	Docus
s5	TV Show	Kota Factory	Not Given	India	24-Sep-21	2021	TV-MA	2 Seasons	Interr
s6	TV Show	Midnight Mass	Mike Flanagan	United States	24-Sep-21	2021	TV-MA	1 Season	TV Dr.
s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Not Given	24-Sep-21	2021	PG	91 min	Childr
s8	Movie	Sankofa	Haile Gerima	United States	24-Sep-21	1993	TV-MA	125 min	Dram.
s9	TV Show	The Great British Baking Show	Andy Devonshire	United Kingdom	24-Sep-21	2021	TV-14	9 Seasons	British

	#	Time	Action	Message	Duration / Fe
✓	1	14:24:30	UPDATE sql_netflix.netflix_titles SET country = SUBSTRING_INDEX(...	1744 row(s) affected Rows matched: 1744 Changed: 1744 Warnings: 0	0.125 sec
✓	2	14:26:41	SELECT * FROM sql_netflix.netflix_titles	8790 row(s) returned	0.015 sec /
✓	3	14:26:49	SELECT * FROM sql_netflix.netflix_titles	8790 row(s) returned	0.000 sec /

Now we can see that there is only one country in the Country column and 8790 Rows are returned after the Data Cleaning from 8807 Rows

Exploratory Data Analysis

1. Distribution of Content Types on Netflix

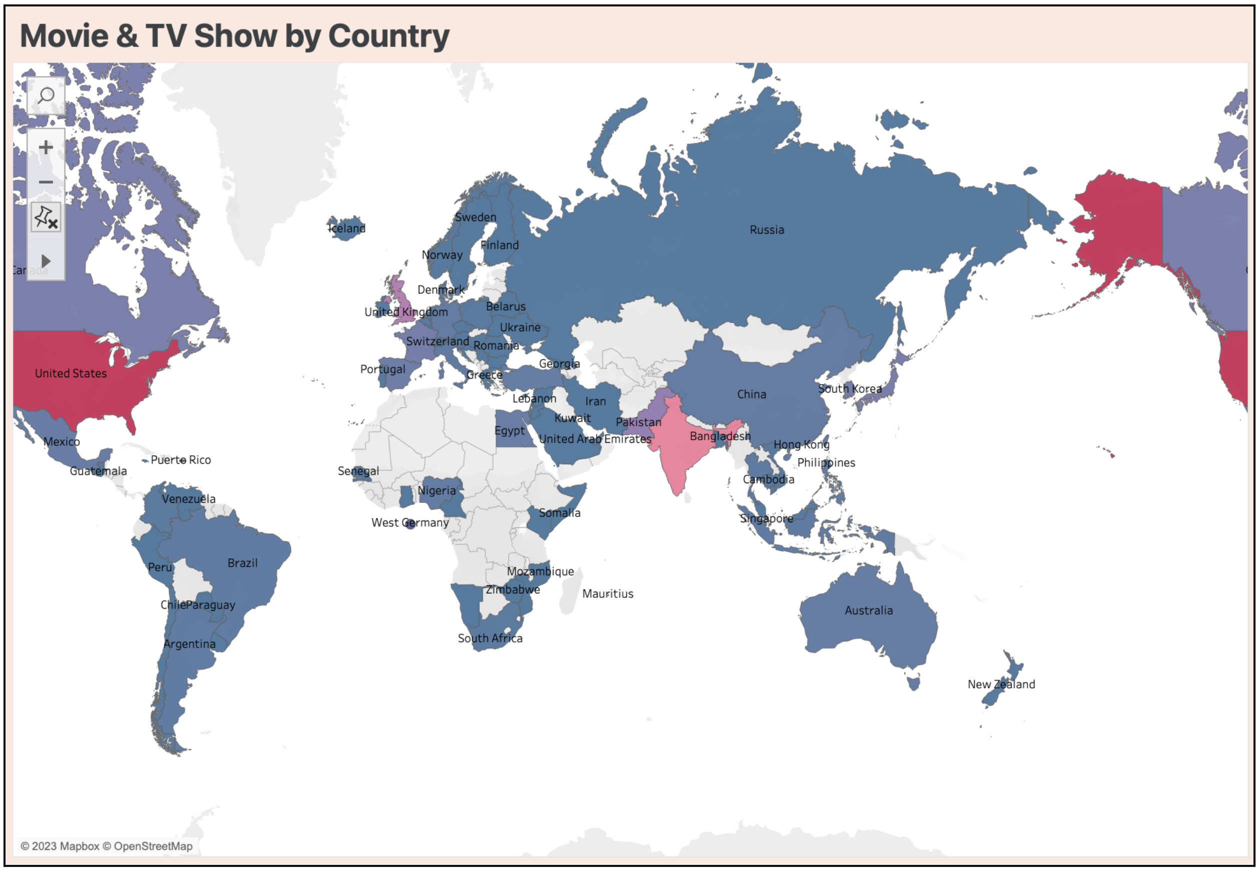


The pie chart distinctly segregates Netflix's offerings into movies and TV series, visually representing the platform's content inventory. The chart indicates that movies are predominant, with 69.69% of the content, suggesting they are the bulk of Netflix's catalog. In contrast, while less numerous, with 30.31% of the content, TV series still form a significant portion of the content mix. This distribution suggests a tilt in Netflix's content strategy towards films, which could align with viewer preferences or reflect the company's content acquisition policies.

The more significant number of movies in the dataset could reflect the different production and licensing dynamics that distinguish movies from TV series. Movies are often standalone works and are thus more plentiful. In contrast, TV series encompass multiple episodes and seasons, signifying a heavier commitment regarding production resources and space on the platform.

The pie chart explains Netflix's position as a critical movie distributor while effectively mapping out the proportions of content types while maintaining a diverse array of TV series. It lays the groundwork for a deeper analysis of how content preferences among subscribers might influence the platform's content acquisition and production strategies.

2. Geographic Distribution of Netflix Content Production

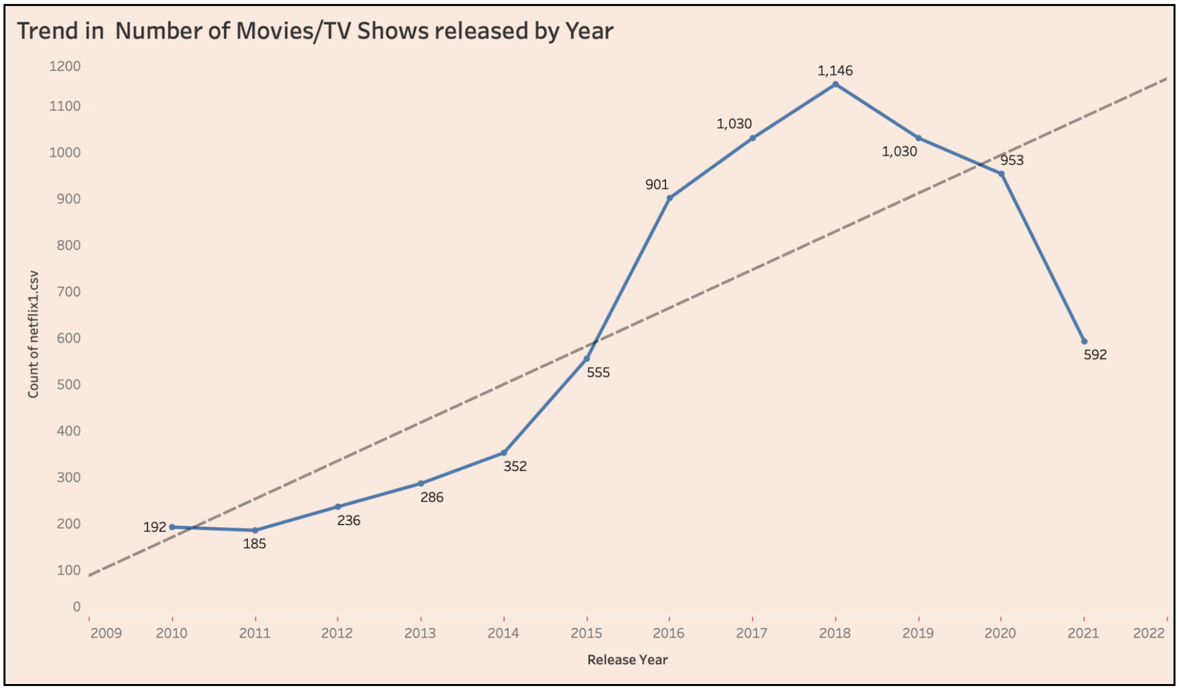

Movie & TV Show by Country

The map visualization, categorized by the intensity of color, indicates the popularity of movies and TV shows by country. The red color signifies the regions where content is most popular or watched, while blue denotes areas where content is less popular.

The red areas, likely representing countries such as the United States, indicate a high popularity of media content. This could suggest a strong domestic production of movies and TV shows that are also globally recognized. Given the United States' established film and television industry, it's plausible that American content is highly consumed both locally and internationally.

The distribution of color intensities provides an overview of the platform's content strategy effectiveness. High-intensity red regions are likely well-served by the platform's current offerings, while lighter blue regions may need different content strategies to engage viewers.

3. Annual Trends in Netflix Content Release



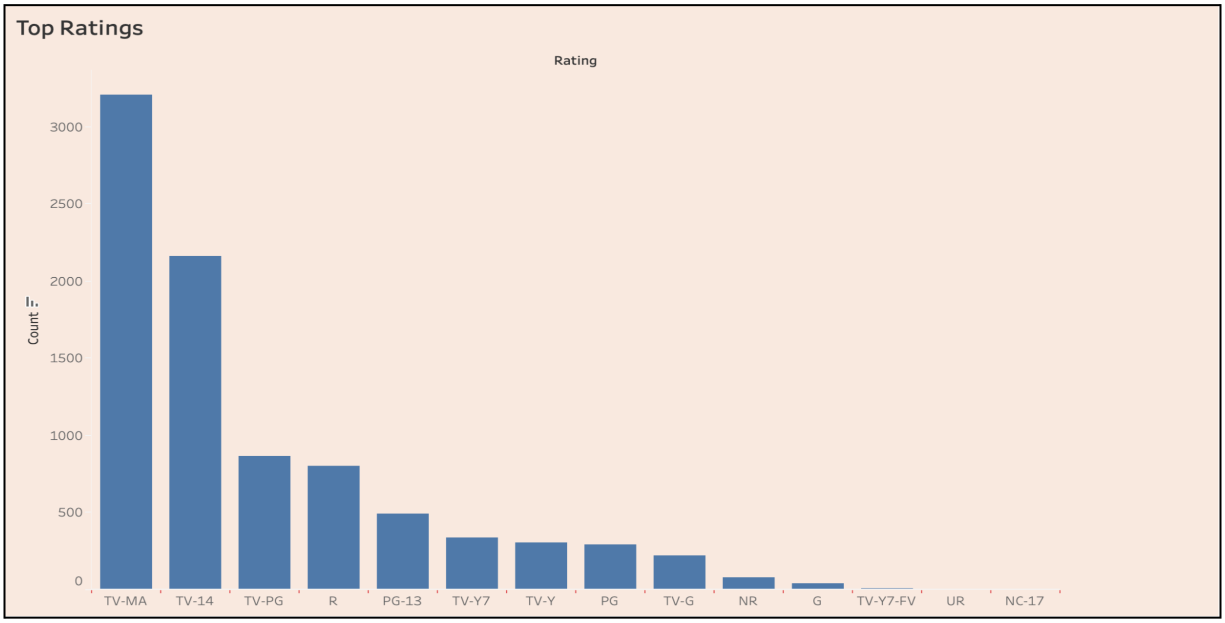
The line chart traces the evolution in the number of movies and TV series released by Netflix from 2009 through 2022. A noticeable upward trend in content releases leading up to 2020 reflects Netflix's vigorous growth and its increased emphasis on content production and acquisition. Starting from just 71 titles in 2009, there has been consistent year-over-year growth.

A significant leap in content release was evident around 2015, when the number of releases more than doubled from the previous year, jumping from 352 to 555, marking Netflix's transition towards a dual role as both distributor and originator of original content.

The zenith of releases is observed in 2018, boasting 1146 titles, before experiencing a slight retraction in 2019. The most pronounced downturn is witnessed in 2021, with a release count dwindling to 592 titles, likely a repercussion of the global disruptions caused by the COVID-19 pandemic that impacted the entertainment industry's production and release schedules across the globe.

This chart serves as a numeric reflection of the shifts in Netflix's content strategy over the years. It lays the groundwork for analyzing the impact of external factors, such as the pandemic, on content release patterns. It prompts further inquiries into the effects of these trends on the platform's subscriber base expansion and its competitive stance in the streaming market.

4. Distribution of Netflix Content by Rating



The bar chart presents an overview of the Netflix content spectrum categorized by viewer ratings, ranging from TV-MA to NC-17. These ratings delineate the suitability of content for different age groups, indicating the level of maturity of the programming.

The most prominent bar corresponds to TV-MA-rated content, suggesting that a significant portion of Netflix's catalog is crafted for an adult audience, with mature themes possibly unsuitable for children. Following this, the substantial presence of TV-14 and PG-13-rated content shows that Netflix caters to teenagers and pre-teens, with a considerable selection appropriate for viewers aged 14 and above.

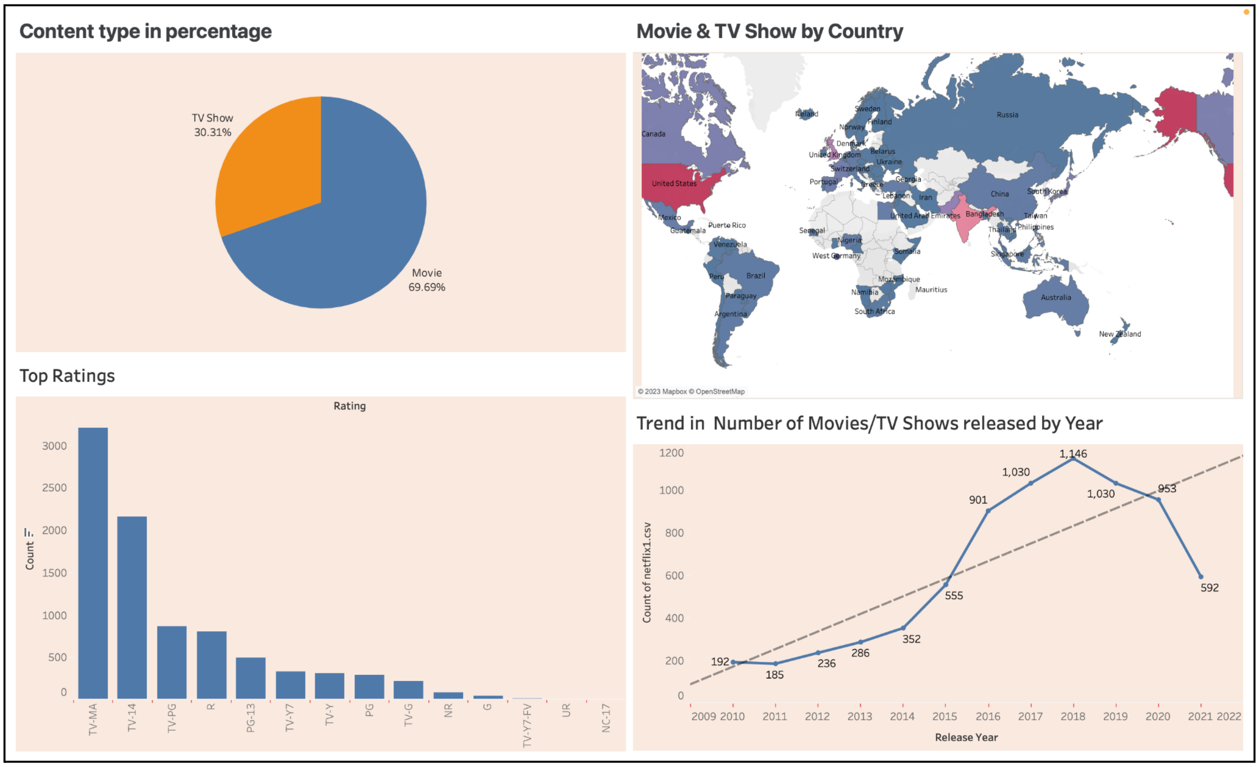
As the chart progresses towards ratings like TV-PG and PG, there's a noticeable downtrend in the count, pointing to a smaller content inventory for young families. The scarcity of G-rated content (suitable for all ages) may imply a more limited selection for family viewing.

Titles with ratings of TV-Y7, TV-Y, and TV-G, deemed appropriate for younger children, including toddlers, represent even fewer entries. The bar chart further shows that NR (Not

Rated), UR (Unrated), and NC-17 rated content, which are less common, occupy the smallest share, with NC-17 catering exclusively to adult viewers.

The chart reflects a diverse content library spanning various maturity levels, with a noticeable tilt toward adult programming. This may mirror Netflix's strategic targeting of a broader adult demographic, potentially aligning with the preferences of most of its subscriber base. The distribution also offers insights into the strategic direction of content creation and acquisition, indicating areas where Netflix may expand or adapt to meet the changing demands of its audience.

Dashboard:



The interactive Tableau Dashboard integrates a pie chart showcasing Netflix's content distribution between movies and TV series. Upon selecting different countries on the map graph, the pie chart dynamically adjusts to illustrate the proportion of each content type specific to the chosen country. This visualization underscores that movies constitute a substantial 69.69% of Netflix's content, outweighing TV series, which account for 30.31%. This clear depiction hints at Netflix's strategy favoring movies while maintaining a significant collection of TV series. It also implies potential factors influencing content acquisition strategies and viewer preferences, suggesting a deliberate tilt towards films.

The synergy among the pie chart, line chart, bar chart, and map graph within the Tableau Dashboard facilitates an interconnected view of Netflix's global content distribution, evident in various visualizations. When a specific country is selected on the map, the remaining charts dynamically adapt to display corresponding insights. For instance, the map graph showcases the geographic popularity of Netflix content, while the other charts alter to represent content types, annual release trends, and viewer ratings specific to the chosen country. This real-time synchronization empowers a comprehensive analysis of Netflix's content strategy on a country-by-country basis, offering valuable insights into content preferences and strategic considerations in different regions.

These integrated visualizations in the Tableau Dashboard provide an immersive and interconnected exploration of Netflix's content landscape. As users interact with the map graph to select different countries, the accompanying pie, line, and bar charts seamlessly adjust, unraveling nuanced insights about content distribution, release trends, and audience preferences tailored to specific geographic regions. This synchronized visualization approach fosters a deeper understanding of Netflix's content strategy, enabling informed decision-making and strategic planning within the platform's global content ecosystem.

Methodology

The dataset, sourced from Kaggle, arrived in CSV format, which we converted into JSON format for ease of manipulation. Then imported the dataset into MySQL Workbench Editor, the data exhibited irregularities, particularly within the 'Director' column. To tackle these inconsistencies, a comprehensive data cleaning process was executed using SQL. Various SQL functions such as JOIN, GROUP BY, and HAVING were leveraged to identify and rectify anomalies. Null values were systematically addressed—rows were populated and unnecessary columns were removed, Columns were split for improved data analysis efficiency.

Following the meticulous data cleansing in MySQL Workbench, the refined dataset was seamlessly extracted as csv into Tableau for comprehensive visualization and analysis. To derive meaningful insights, diverse chart types were employed, including Pie Charts for categorical distributions, Maps for geographical trends, Line Graphs for temporal patterns, and Bar Graphs for comparative analysis. These visualizations served as a powerful tool to unravel patterns, trends, and correlations within the data, enabling stakeholders to grasp the information intuitively. The utilization of Tableau's features facilitated the creation of visually appealing and informative representations, empowering stakeholders to gain actionable insights from the cleaned dataset, thereby aiding in informed decision-making and strategic planning.

Key Findings

Netflix's content composition is predominantly movie-centric, with films making up nearly 70% of its offerings. This substantial movie library showcases a strategic emphasis on a wide range of standalone films. On the other hand, television series, which account for about 30% of the content, represent a notable commitment from Netflix, given the extensive resources required for episodic content development.

Regarding geographic consumption, the United States is a significant content consumer, indicating that Netflix's strategies are particularly effective in this region. However, the map also points to the other areas, marked in blue, where Netflix could benefit from creating more tailored content strategies to increase its popularity.

The trend in Netflix's content release has seen a general upward trajectory, with a peak in 2018. But, in 2021, there was an unexpected decline in titles released, a likely consequence of the global COVID-19 pandemic affecting production and release schedules.

As for the ratings distribution within Netflix's catalog, TV-MA-rated programs predominate, catering to the adult demographic. This focus not only meets the preferences of a mature audience but also solidifies Netflix's position as a leading distributor of content that appeals to adult viewers.

References

- *James, K. (2022, December 12). The rise of Netflix: A Data Analysis. Medium. https://medium.datadriveninvestor.com/the-rise-of-netflix-a-data-analysis-9cbd3e00d736*

- *SQL data cleaning techniques for accurate analytics*. Airbyte. (n.d.). https://airbyte.com/data-engineering-resources/sql-data-cleaning

- Singh, G. (2021, October 14). *A step by step guide for Data Visualization using tableau*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/step-by-step-guide-data-visualization-tableau/