Netflix

TV Shows & Movies Analysis

By Anna Meudec

About the Dataset

Netflix is a highly popular streaming service that allows users worldwide to access a wide variety of movies, TV shows and Netflix's own original content. The dataset used in this analysis was downloaded originally from Kaggle and contains content added to Netflix from 2008 to 2021. This dataset was cleaned through PostgreSQL during my analysis, and visualisations were created through Tableau. The original dataset, cleaned dataset, a complete version of my cleaning code and my tableau visualisations can all be accessed in the “Files Employed” section at the end of this analysis.

The overall goal of this analysis is to practise my data cleaning skills in PostgreSQL and try out Tableau to create visualisations. With this, my analysis entails a step-by-step walkthrough of how I created my database, cleaned the dataset and produced visualisations in Tableau.

Note: Most explanations can be found as **comments** within my code.

Creating the Database

Firstly, I began by creating my database in PostgreSQL.

```
--Creating my database
CREATE DATABASE netflix_titles;

--Creating my table
CREATE TABLE netflixtbl
(
show_id VARCHAR(5) PRIMARY KEY,
net_type VARCHAR(7),
net_title VARCHAR(105),
net_director TEXT,
net_cast TEXT,
net_country TEXT,
date_added DATE,
release_year INT,
net_rating VARCHAR(8),
net_duration VARCHAR(10),
listed_in VARCHAR(80),
net_description TEXT
);

--Importing my data
Copy netflixtbl (show_id, net_type, net_title, net_director, net_cast, net_country, date_added,
release_year, net_rating, net_duration, listed_in, net_description)
FROM 'C:\Users\Public\netflix_titles.csv'
DELIMITER ','
csv Header;

--Viewing our dataset
SELECT *
FROM netflixtbl
```

Cleaning the Dataset

Moving on to cleaning the dataset, I will focus on treating NULLS, duplicates, missing rows, unnecessary columns and splitting columns.

```
 --Checking for duplicates in the show_id column. Since it's our primary key & there should be no duplicates.
SELECT show_id, COUNT(*)
FROM netflixtbl
GROUP BY show_id
ORDER BY show_id DESC;

--Outcome: No duplicates in show_id column

--Checking for NULL values across all columns
SELECT COUNT(*) FILTER (WHERE show_id IS NULL) AS showid_nulls,
      COUNT(*) FILTER (WHERE net_type IS NULL) AS type_nulls,
      COUNT(*) FILTER (WHERE net_title IS NULL) AS title_nulls,
      COUNT(*) FILTER (WHERE net_director IS NULL) AS director_nulls,
      COUNT(*) FILTER (WHERE net_cast IS NULL) AS cast_nulls,
      COUNT(*) FILTER (WHERE net_country IS NULL) AS country_nulls,
      COUNT(*) FILTER (WHERE date_added IS NULL) AS date_added_nulls,
      COUNT(*) FILTER (WHERE release_year IS NULL) AS release_year_nulls,
      COUNT(*) FILTER (WHERE net_rating IS NULL) AS rating_nulls,
      COUNT(*) FILTER (WHERE net_duration IS NULL) AS duration_nulls,
      COUNT(*) FILTER (WHERE listed_in IS NULL) AS listed_in_nulls,
      COUNT(*) FILTER (WHERE net_description IS NULL) AS description_nulls
FROM netflixtbl;
```

After checking, we can see that NULLS do exist.

director_nulls = 2634
movie_cast_nulls = 825
country_nulls = 831
date_added_nulls = 10
rating_nulls = 4
duration_nulls = 3

director_nulls is over 30% of the entire column therefore, I won't remove them. I will instead populate it with another column.

```
--Checking if directors are likely to work with certain cast members
WITH cte AS
(
SELECT net_title, CONCAT(net_director, '---', net_cast) AS director_cast
FROM netflixtbl
)
```

```
SELECT director_cast, COUNT(*) AS count
FROM cte
GROUP BY director_cast
HAVING COUNT(*) > 1
ORDER BY COUNT(*) DESC;

--Having done this, we can now populate NULL rows in directors with movie_cast
UPDATE netflixtbl
SET net_director = 'Alastair Fothergill'
WHERE net_cast = 'David Attenborough'
AND net_director IS NULL ;

--Repeating this step to populate all director_nulls
--Populating remaining NULL in director as "Not Provided"

UPDATE netflixtbl
SET net_director = 'Not Provided'
WHERE net_director IS NULL;
```

Similar to the director column, I won't delete the NULLS in the country column but rather populate them with the director column.

```
--Populate the country using the director column
SELECT COALESCE(nt.net_country,nt2.net_country)
FROM netflixtbl  AS nt
JOIN netflixtbl AS nt2
ON nt.net_director = nt2.net_director
AND nt.show_id <> nt2.show_id
WHERE nt.net_country IS NULL;
UPDATE netflixtbl
SET net_country = nt2.net_country
FROM netflixtbl AS nt2
WHERE netflixtbl.net_director = nt2.net_director and netflixtbl.show_id <> nt2.show_id
AND netflixtbl.net_country IS NULL;

--Checking to see if any directors refuse to update
SELECT net_director, net_country, date_added
FROM netflixtbl
WHERE net_country IS NULL;

--Outcome: There are still NULLS that have to be populated

--I will populate the remaining NULLs as "Not Provided"
UPDATE netflixtbl
SET net_country = 'Not Provided'
WHERE net_country IS NULL;
```

Taking a look at date_added NULLS. We only have 10, so deleting them will likely not affect our visualisations or analyses.

```
--Showing the date_added NULLS
SELECT show_id, date_added
FROM netflixtbl
WHERE date_added IS NULL;

--Deleting the date_added NULLS
DELETE FROM netflixtbl
WHERE show_id
IN ('6797', 's6067', 's6175', 's6807', 's6902', 's7255', 's7197', 's7407', 's7848', 's8183');
```

Again, we only have 4 nulls in net_rating. Therefore, we will just delete them.

```
--Showing the net_rating NULLS

SELECT show_id, net_rating
FROM netflixtbl
WHERE net_rating IS NULL;

--Deleting the net_rating NULLS
DELETE FROM netflixtbl
WHERE show_id
IN (SELECT show_id FROM netflixtbl WHERE net_rating IS NULL)
RETURNING *;
```

Lastly, we only have 3 nulls in net_duration. Therefore, we will just delete them.

```
--Showing the net_duration NULLS

SELECT show_id, net_duration
FROM netflixtbl
WHERE net_duration IS NULL;

--Deleting the net_duration NULLS
DELETE FROM netflixtbl
WHERE show_id
IN (SELECT show_id FROM netflixtbl WHERE net_duration IS NULL)
RETURNING *;
```

Lastly, just to make sure there are no more NULLs in our columns, we will re-run our previous NULL-checking query.

```
--Checking to make sure there are no more NULLS in our columns

SELECT COUNT(*) FILTER (WHERE show_id IS NULL) AS showid_nulls,
	COUNT(*) FILTER (WHERE net_type IS NULL) AS type_nulls,
	COUNT(*) FILTER (WHERE net_title IS NULL) AS title_nulls,
	COUNT(*) FILTER (WHERE net_director IS NULL) AS director_nulls,
	COUNT(*) FILTER (WHERE net_country IS NULL) AS country_nulls,
	COUNT(*) FILTER (WHERE date_added IS NULL) AS date_added_nulls,
	COUNT(*) FILTER (WHERE release_year IS NULL) AS release_year_nulls,
	COUNT(*) FILTER (WHERE net_rating IS NULL) AS rating_nulls,
	COUNT(*) FILTER (WHERE net_duration IS NULL) AS duration_nulls,
	COUNT(*) FILTER (WHERE listed_in IS NULL) AS listed_in_nulls
FROM netflixtbl;
```

Now, I am going to drop the net_cast and net_description columns as I won't be using them for my visualisations or analysis.

```
ALTER TABLE netflixtbl
DROP COLUMN net_cast,
DROP COLUMN net_description;
```

The net_country column contains multiple countries per row. For visualisation purposes, I will just take the first country as the original country of where the movie was produced.

```
SELECT *,
	SPLIT_PART(net_country,',',1) AS country,
	SPLIT_PART(net_country,',',2),
	SPLIT_PART(net_country,',',4),
	SPLIT_PART(net_country,',',5),
	SPLIT_PART(net_country,',',6),
	SPLIT_PART(net_country,',',7),
	SPLIT_PART(net_country,',',8),
	SPLIT_PART(net_country,',',9),
	SPLIT_PART(net_country,',',10)
FROM netflixtbl;

--Updating the netflixtbl
ALTER TABLE netflixtbl
ADD country1 varchar(500);
UPDATE netflixtbl
SET country1 = SPLIT_PART(net_country, ',', 1);
-- This creates a new column called "country1" and inserts just the 1st country.
```

Now, I am going to delete the net_country column as it is no longer useful to us.

```
ALTER TABLE netflixtbl
DROP COLUMN net_country;

--Checking to verify the column has been dropped
SELECT *
FROM netflixtbl;
```

Lastly, I am just renaming the "country1" column to “net_country” and copying my new cleaned data to a CSV file.

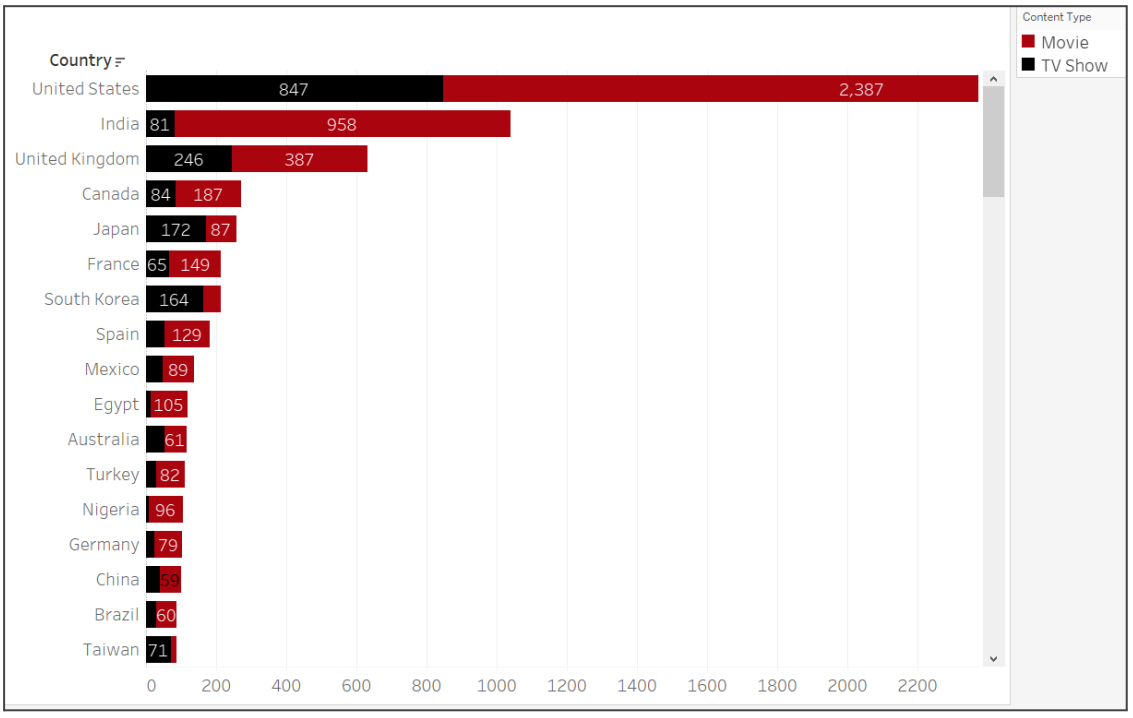
```
ALTER TABLE netflixtbl
RENAME COLUMN country1 TO net_country;


COPY (SELECT * FROM netflixtbl) TO 'C:\Users\Public\netflix_titles_cleaned.csv' WITH
CSV HEADER;
```

Creating Visualisations in Tableau

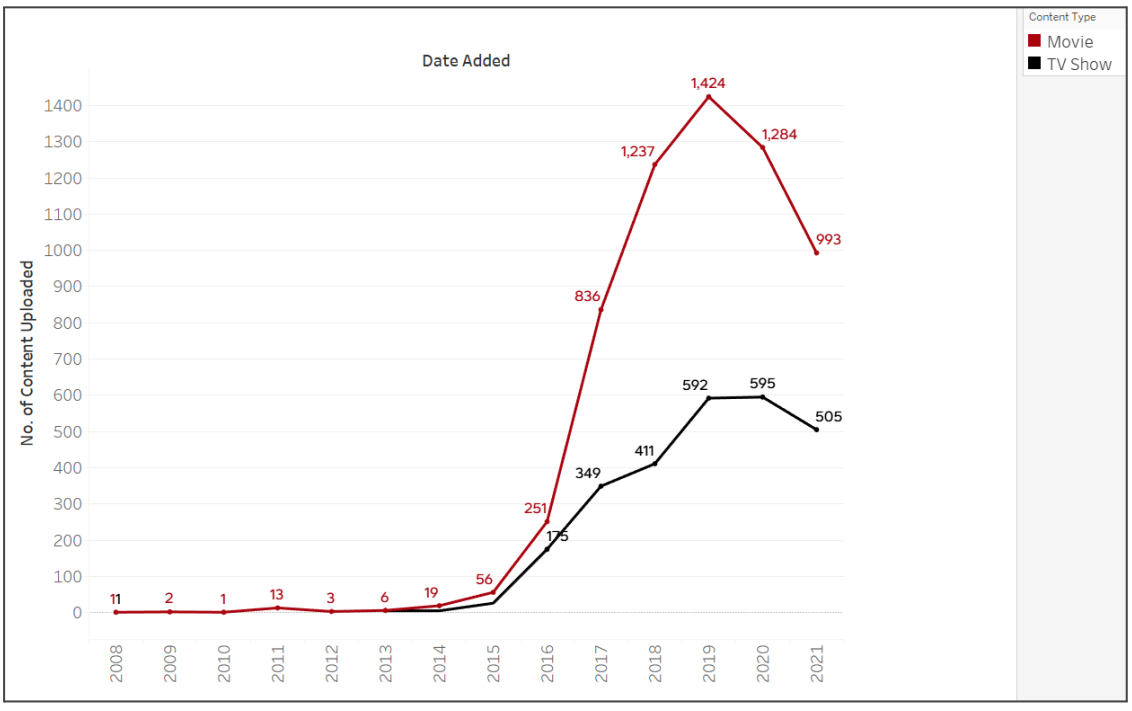
With my dataset cleaned, I can now begin creating some interesting visualisations in Tableau. In the section “Files Employed”, you can click to see my complete Tableau Dashboard and interact with it on Tableau Public. Additionally, you can download my file from there to view all individual sheets and how they were created.

Sheet 1: Number of Content Uploaded by Country (2008 - 2021)



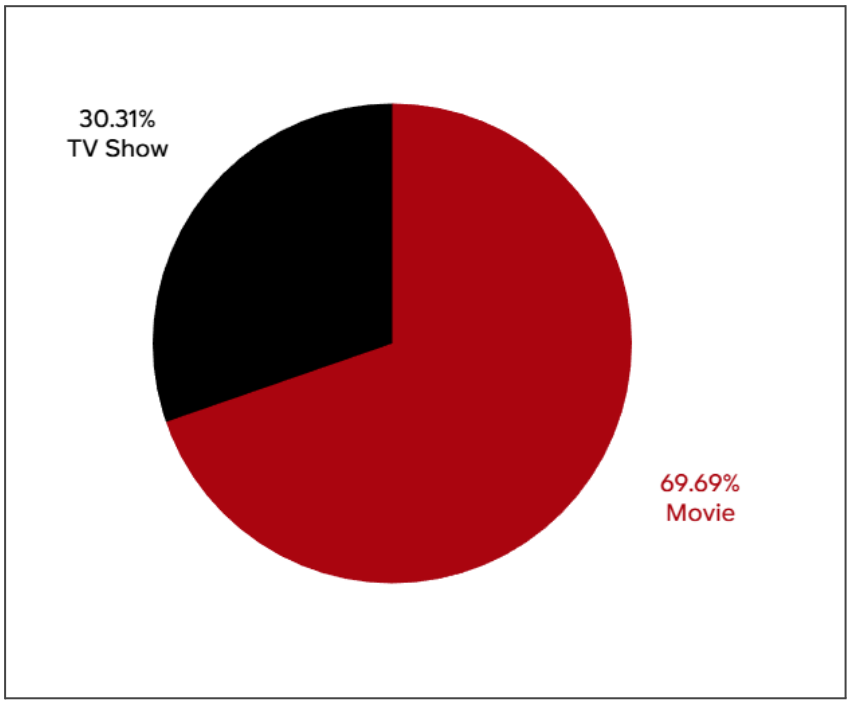
This bar chart visualisation shows an alternative view of the “TV Shows & Movies by Country” visualisation found on sheet 4. However, this bar chart allows us to view the breakdown of both content types (TV shows and movies) and clearly see the total count of each type per country.

Sheet 2: Content Type Uploaded between 2008 - 2021



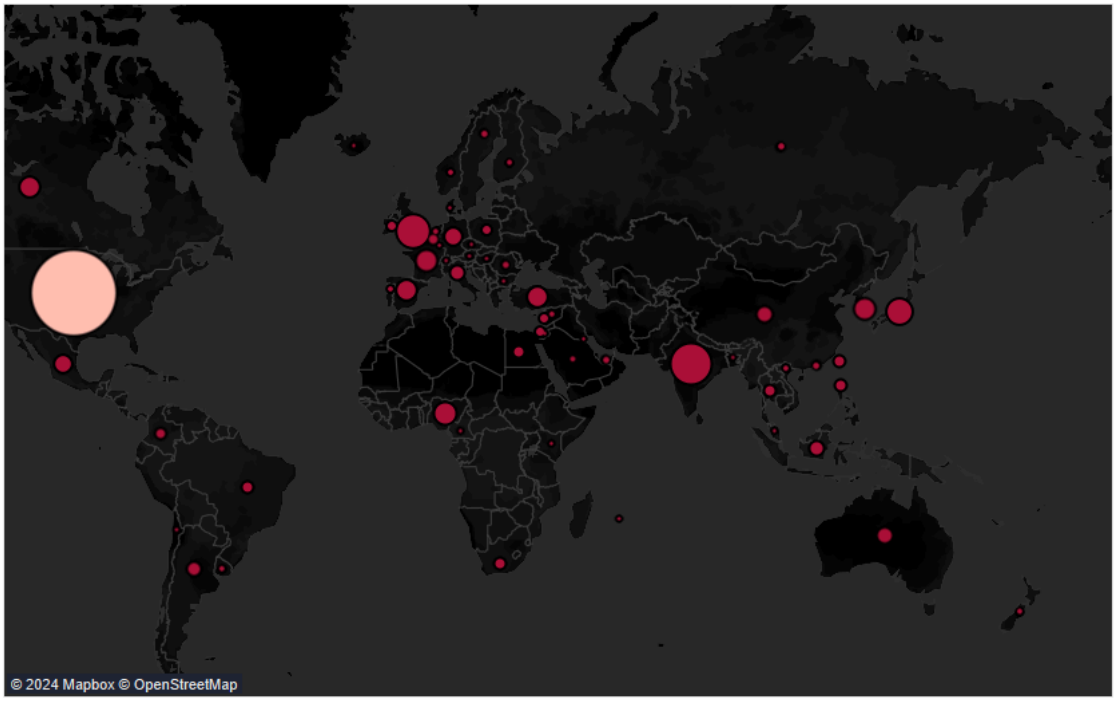
Above is a line chart comparing the number of TV shows and movies uploaded over the years. Upon viewing, we can see that movies were initially the primary type of content available on Netflix for the first 5 years. In 2016, we see a sharp increase in the number of TV shows added, from just 26 available to stream in 2015 to 175 in 2016. Within the tooltip on the Tableau Dashboard, you can navigate through the visualisation and better view all counts for the TV shows and movies.

Sheet 3: Percentage of Content Type



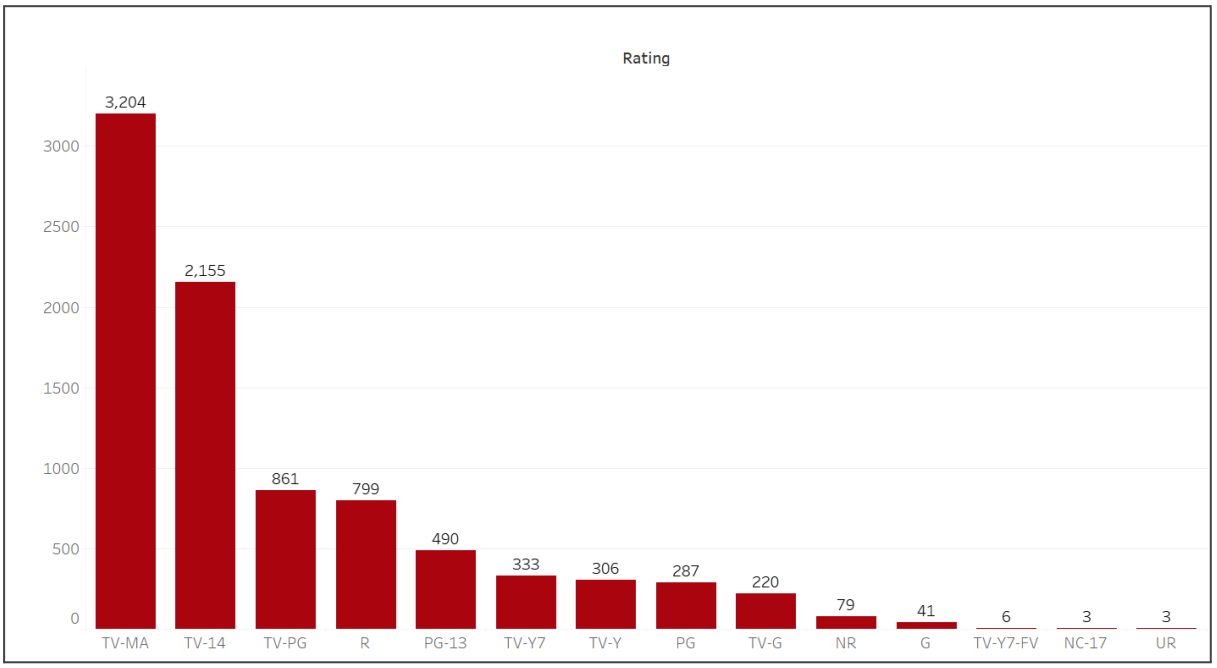
This chart represents the dataset's 2 content types (TV shows and movies). As we can see from the pie chart, movies make up the majority of Netflix's content, at 69.7%. Additional details can be found in the interactive tooltip on my dashboard, revealing the count of both TV shows and movies.

Sheet 4: TV Shows & Movies by Country



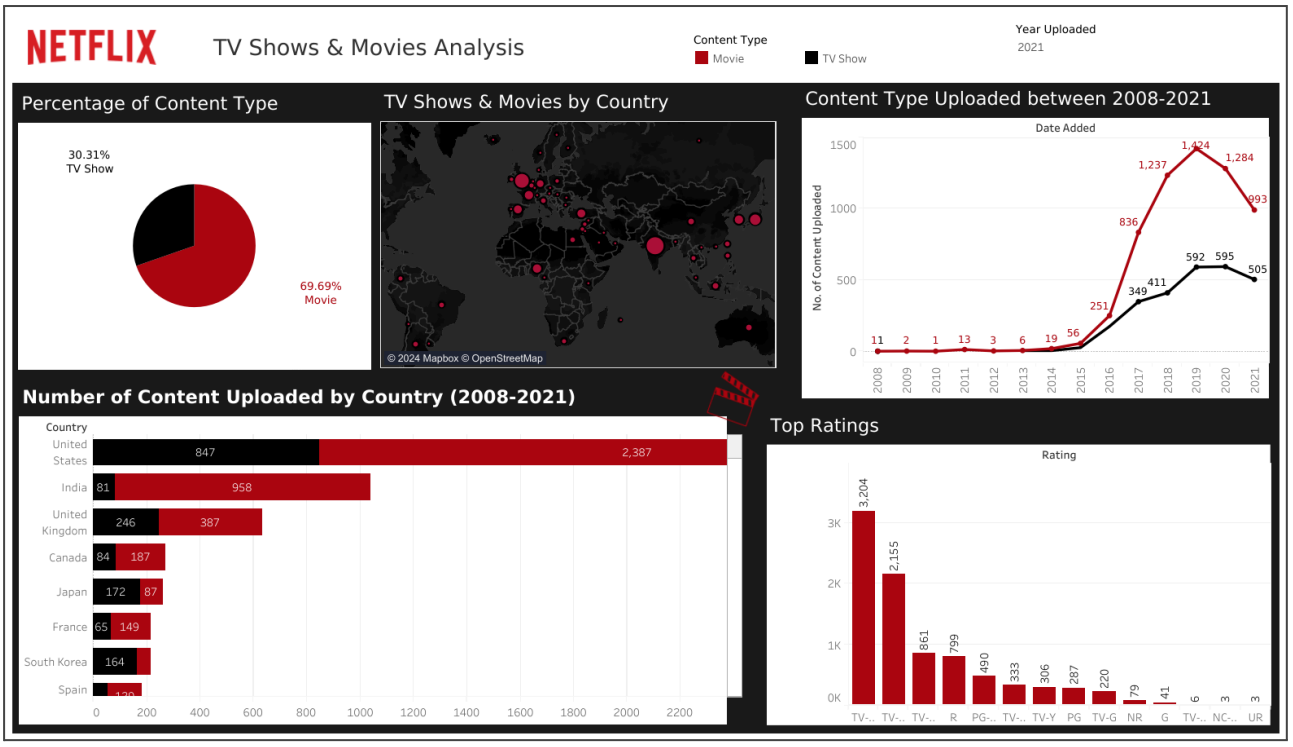
Here, we have a map visualisation showcasing the total amount of TV shows and movies per country in 2021. The total amount is depicted by the colour and size of the circular bubbles. From the visualisation, the United States of America has the most TV shows and movies, followed by India and then the United Kingdom. Upon viewing my interactive dashboard, you will be able to filter by year (2008 - 2021) and see the total count of TV shows and movies for that country for that specific year.

Sheet 5: Top Ratings



In this visualisation, we have the top-rating types on Netflix. TV shows and movies on Netflix have ratings that recognise the recommended age/audience type for viewing content. From the bar chart, we can see that “TV-MA”, standing for mature audiences, has the most content on Netflix, followed by content recommended for ages 14+ and more kid-friendly content (rated PG) coming in third.

Dashboard:



Here is my complete Tableau Dashboard that combines all my separate visualisations to create an interesting overview of Netflix's available content all over the world. Again, the purpose of this analysis was to practise my SQL data cleaning skills and try out Tableau Dashboard to make a variety of visualisations. Please see the section below, whereby all the files used for this analysis can be accessed, as well as my interactive dashboard hosted on Tableau Public.

Files Employed

Original CSV File: **netflix_titles.csv (Google Drive)**

Cleaned CSV File: **netflix_titles_cleaned.csv (Google Drive)**

Data Cleaning SQL Code: **netflix_titles_code.sql (Google Drive)**

Link to Tableau Dashboard: **Netflix TV Shows & Movies Analysis (Tableau Public)**