NETFLIX DATA CLEANING

NETFLIX DATA
Author: Prayag Das
Website: https://prayagpds.wixsite.com/my-site-1
Contact: prayag.pds@gmail.com

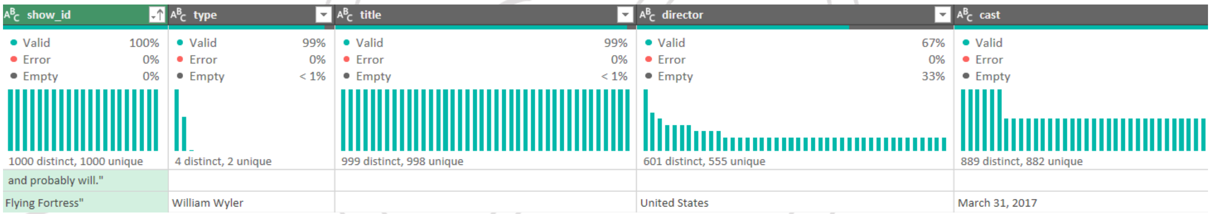
CONTENT-

INTRODUCTION-

Netflix is a popular streaming platform. It contains many categories of movies and TV shows. Prior to its launch in 2007 as streaming platform, it was a DVD based rental company which launched in 1996. It is the most popular on-demand streaming platform with 300 million+ paid membership in more than 190 countries. As a result, there is a lot of diversity in content. This data is of the year 2021. Cleaned using Excel, POWER QUERY and DAX.

DATA CLEANING-

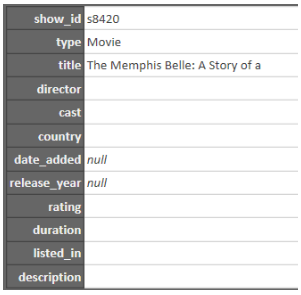
- Loaded the netflix_titles dataset.
- In Power Query, it is easier to notice the inconsistencies.



show_id	type	title	director	cast
● Valid 100% ● Error 0% ● Empty 0%	● Valid 99% ● Error 0% ● Empty < 1%	● Valid 99% ● Error 0% ● Empty < 1%	● Valid 67% ● Error 0% ● Empty 33%	● Valid ● Error ● Empty
1000 distinct, 1000 unique	4 distinct, 2 unique	999 distinct, 998 unique	601 distinct, 555 unique	889 distinct, 882 unique
and probably will."				
Flying Fortress"	William Wyler		United States	March 31, 2017

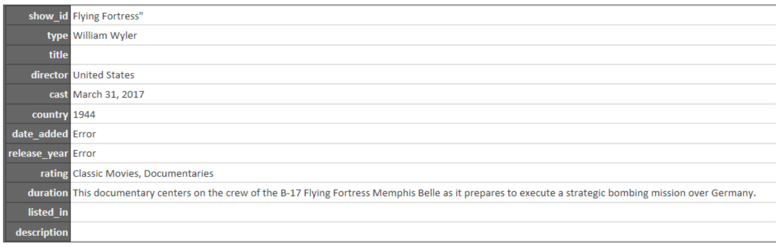
The first column had 2 errors, and subsequent columns also had either missing values or error values, so removed these 2 rows.

By exploring other columns found an entire row having inconsistent value.



show_id	s8420
type	Movie
title	The Memphis Belle: A Story of a
director	
cast	
country	
date_added	null
release_year	null
rating	
duration	
listed_in	
description	

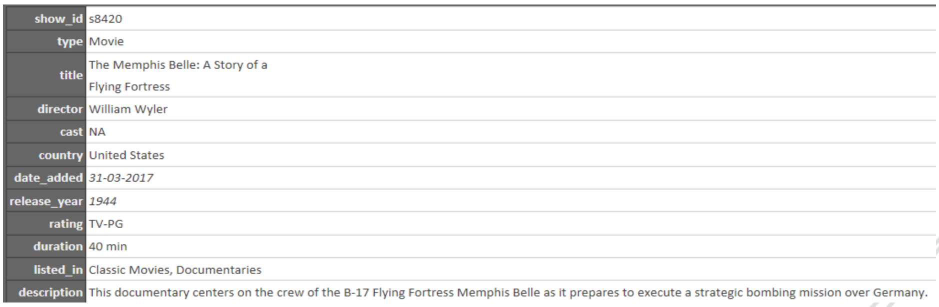
The show_id was available and by doing some more digging, found out that the previous 2 removed rows had some similar data.



show_id	Flying Fortress"
type	William Wyler
title	
director	United States
cast	March 31, 2017
country	1944
date_added	Error
release_year	Error
rating	Classic Movies, Documentaries
duration	This documentary centers on the crew of the B-17 Flying Fortress Memphis Belle as it prepares to execute a strategic bombing mission over Germany.
listed_in	
description	

Used excel to manually replace the values.

By using the available data and some google, Final result-



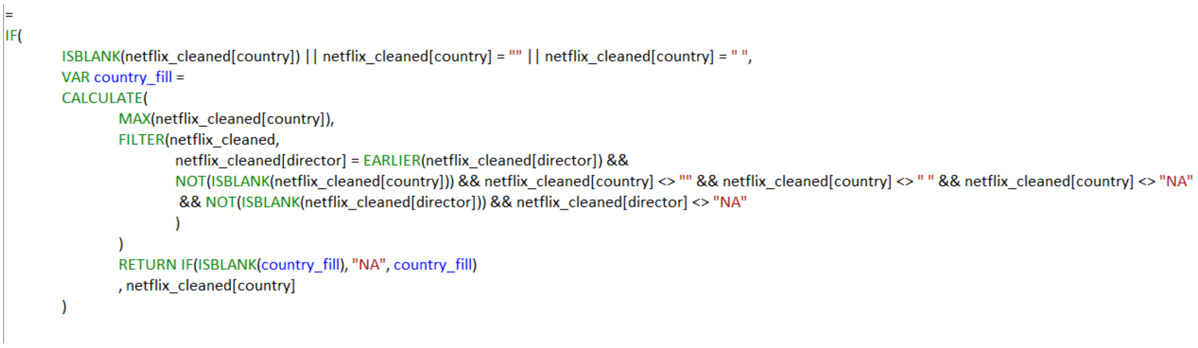
show_id	s8420
type	Movie
title	The Memphis Belle: A Story of a Flying Fortress
director	William Wyler
cast	NA
country	United States
date_added	31-03-2017
release_year	1944
rating	TV-PG
duration	40 min
listed_in	Classic Movies, Documentaries
description	This documentary centers on the crew of the B-17 Flying Fortress Memphis Belle as it prepares to execute a strategic bombing mission over Germany.

Explored the remaining columns and rows, and filled the null and error values.

- The nulls in director, cast and country were too many, so replaced those with “NA” for better consistency.
- In the country column, several entries were missing. However, many of these rows had the director field completed, and numerous films and shows shared directors whose country information was available elsewhere. By leveraging DAX within the data model, I calculated a new column that inferred missing country values based on their association with the director.
  Even after trimming the country column, it had some blank spaces, took that into consideration.

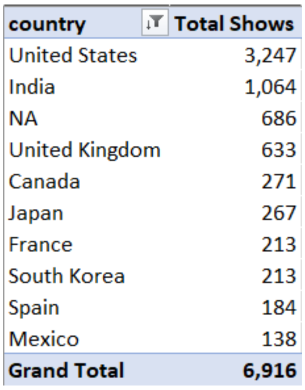
```
=IF(

        ISBLANK(netflix_cleaned[country]) || netflix_cleaned[country] = "" ||
netflix_cleaned[country] = " ",

        VAR country_fill =

        CALCULATE(

                MAX(netflix_cleaned[country]),

                FILTER(netflix_cleaned,

                        netflix_cleaned[director] = EARLIER(netflix_cleaned[director]) &&

                        NOT(ISBLANK(netflix_cleaned[country])) && netflix_cleaned[country] <> "" &&
netflix_cleaned[country] <> " " && netflix_cleaned[country] <> "NA"

                         && NOT(ISBLANK(netflix_cleaned[director])) && netflix_cleaned[director] <>
"NA"

                        )

                )

                RETURN IF(ISBLANK(country_fill), "NA", country_fill)

                , netflix_cleaned[country]

        )
```

Formula in DAX-



```
=
IF(
    ISBLANK(netflix_cleaned[country]) || netflix_cleaned[country] = "" || netflix_cleaned[country] = " ",
    VAR country_fill =
    CALCULATE(
        MAX(netflix_cleaned[country]),
        FILTER(netflix_cleaned,
            netflix_cleaned[director] = EARLIER(netflix_cleaned[director]) &&
            NOT(ISBLANK(netflix_cleaned[country])) && netflix_cleaned[country] <> "" && netflix_cleaned[country] <> " " && netflix_cleaned[country] <> "NA"
             && NOT(ISBLANK(netflix_cleaned[director])) && netflix_cleaned[director] <> "NA"
            )
        )
        RETURN IF(ISBLANK(country_fill), "NA", country_fill)
        , netflix_cleaned[country]
    )
```

- Here are the top 10 countries that have their filmography on Netflix:



country	Total Shows
United States	3,247
India	1,064
NA	686
United Kingdom	633
Canada	271
Japan	267
France	213
South Korea	213
Spain	184
Mexico	138
Grand Total	**6,916**

- The data is cleaned for further analysis.

CONCLUSION-

- The 2021 Netflix titles dataset was thoroughly cleaned and prepared for analysis using Excel, Power Query, and DAX.
- Critical errors and inconsistencies were resolved by removing problematic rows and manually correcting key entries.
- Null values in important columns such as director, cast, and country were standardized by replacing them with “NA” for consistency.
- DAX was utilized to intelligently infer and fill missing country values based on shared director information, enhancing data completeness.
- The resulting dataset is accurate, consistent, and reliable, providing a strong foundation for in depth analysis of Netflix’s global content trends.