What is contained within a sheet? make your own notes about the content.

1. There is one sheet.
2. There are 8809 movies/tv shows listed rows.
3. There are 12 attributes of data on each movie/tv show.

show_id: Unique identifier for each title.

type: Indicates whether the title is a movie or a TV show.

title: Name of the title.

director: Name of the director (may be missing for some titles).

cast: Names of the main cast members (may also be missing).

country: Country of origin.

date_added: When the title was added to Netflix.

release_year: Year the title was originally released.

rating: The maturity rating (e.g., PG-13, TV-MA).

duration: Duration in minutes for movies or number of seasons for TV shows.

Unnamed columns (16–25): These seem to be unnecessary or contain null values and can likely be ignored or removed during analysis.

What would you like to discover from the data set?

1. When was the earliest recorded TV Show/Movies in the Netflix?
2. Summarize the data: number of movies vs. TV shows, distribution of release years
3. Top directors, actors, and countries contributing to the platform.
4. Is there a trend in content release on Netflix based on genre or country?
5. What is the distribution of content based on the age ratings across different countries?

The context of your questions, the industry the dataset is from, typical data for that industry

The dataset likely comes from the streaming media industry, which has been rapidly expanding with major players such as Netflix, Hulu, Amazon Prime, and Disney+. A key trend in this industry is the globalization of content, where platforms are increasingly offering international content to cater to diverse audiences. Another important trend is the shift from licensed content to original productions, as companies seek to retain viewers with exclusive shows and films.

Typical Data in the Streaming Industry: Common attributes in datasets from this industry often include:

- **Title**, **Genre**, and **Type** (Movie or TV Show)
- **Release Date**
- **Country of Production**
- **Director and Cast**
- **Duration** or **Seasons**
- **Language** and **Subtitle Options**
- **Viewer Ratings** and **Popularity Metrics** (though this might not always be available publicly)

The attributes in the IMDb dataset:

title.akas.tsv.gz

- titleId (string) - a tconst, an alphanumeric unique identifier of the title
- ordering (integer) – a number to uniquely identify rows for a given titleId
- title (string) – the localized title
- region (string) - the region for this version of the title
- language (string) - the language of the title
- types (array) - Enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay". New values may be added in the future without warning
- attributes (array) - Additional terms to describe this alternative title, not enumerated
- isOriginalTitle (boolean) – 0: not original title; 1: original title

Write in your report the research you undertake and the reasons for choosing your questions and area of interest.

The reasons for choosing these questions might stem from the desire to understand content strategies, the global impact of streaming services, and how platforms like Netflix respond to audience preferences globally.

Identify the sheet(s) and the columns you may use to answer your questions.

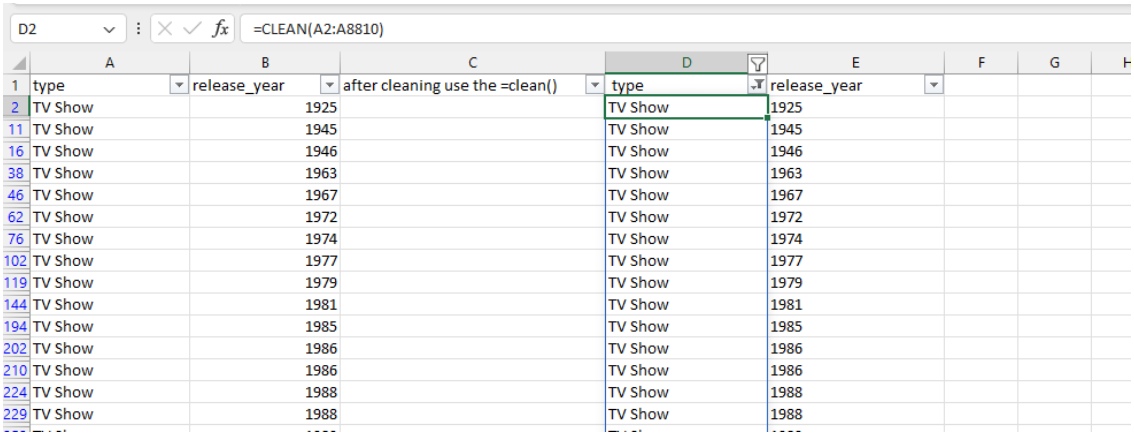
1.When was the earliest recorded TV Show/Movies in the Netflix?

Columns to Use

- `type`: To distinguish between TV shows and movies.
- `release_year`: To determine the earliest release year.

Steps

- clean the data first use the clean function



	A	B	C	D	E	F	G
1	type	release_year	after cleaning use the =clean()	type	release_year		
2	TV Show	1925		TV Show	1925		
11	TV Show	1945		TV Show	1945		
16	TV Show	1946		TV Show	1946		
38	TV Show	1963		TV Show	1963		
46	TV Show	1967		TV Show	1967		
62	TV Show	1972		TV Show	1972		
76	TV Show	1974		TV Show	1974		
102	TV Show	1977		TV Show	1977		
119	TV Show	1979		TV Show	1979		
144	TV Show	1981		TV Show	1981		
194	TV Show	1985		TV Show	1985		
202	TV Show	1986		TV Show	1986		
210	TV Show	1986		TV Show	1986		
224	TV Show	1988		TV Show	1988		
229	TV Show	1988		TV Show	1988		

- select the result

The earliest recorded Movie was published in 1942.

	type	release_year
3	Movie	1942
4	Movie	1942
5	Movie	1943
6	Movie	1943
7	Movie	1943
8	Movie	1944
9	Movie	1944

The earliest recorded TV Show was published in 1942.

	type	release_year
2	TV Show	1925
11	TV Show	1945
16	TV Show	1946
38	TV Show	1963
46	TV Show	1967

2.Summarize the data: number of movies vs. TV shows, distribution of release years

We can use the same table in the previous question.

Columns to Use:

- `type`: To categorize between "Movie" and "TV Show".
- `release_year`: To create a distribution of titles by year.

Steps:

Use this function `=COUNTIF(D2:D8810,"TV Show")` . There are 2677 TV Shows in the netfilx.

Use this function `=COUNTIF(D2:D8810,"Movie")` . There are 6132 movies in the netfilx.

The amount of movies is 2.29 times than TV shows in the netfilx.

3.Top directors, actors, and countries contributing to the platform.

Columns to Use:

- `director`: For top directors.
- `cast`: For top actors.
- `country`: For top contributing countries.

Steps:

4.Is there a trend in content release on Netflix based on genre or country?

Columns to Use:

- `release_year`: This shows the year the content was released. It is critical for understanding trends over time.
- `listed_in`: This column contains information about the genre(s) of each title, which will help analyze trends in genre distribution.
- `country`: This shows the country of origin for the content and will be useful to analyze trends in content releases across different regions.

5.What is the distribution of content based on the age ratings across different countries?

Columns to Use:

`rating`, `country`