Business Intelligence
Mini Project

Made by:
Med Malek Kaouach
Alya Barcous

Project name: **Netflix Data Analysis**

I/ Introduction:

1/ What is Netflix?

- Netflix is a platform that provides their users a way to stream movies without the need to renting movies from the store and some of the newly movies are available there.
- Netflix is considered to be one of the best in providing online streaming as they contain a vast amount of films and television series, including those that are produced themselves.
- Netflix's subscribers kept on increasing from 167 millions subscribers on 2019 to 182 millions subscribers by the first quarter of 2020.
- As a result of Covid-19, most prefer to avoid public places such as cinema which was the main reason of the increased amount of subscribers for Netflix.
- Directors and movie's producers are attracted to secure an agreement with Netflix to showcase their movies or television series on the platform.

2/ Project aims:

The aim of this project is to extract, clean, analyze and visualize the dataset by studying the pattern of the distribution of films and Tv shows on Netflix that are able to provide better insights for researchers on what content is available in different countries and which type of movie or Tv shows is popular and most produced.
Beside that, we want to prepare a network analysis of the most interesting directors and find if Netflix has focused more on TV shows than movies in recent years.

3/ Project objective:

The main objective of this project is to:

- Analyze and gain insight into the kind of content available in the Netflix Database based on type, age, duration, geographic location.
- Find how many shows are added annually based on their types.
- Find most popular directors

4/ About the data set:

The dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as cast, directors, ratings, release year, duration, etc.

5/ Problem statement:

- Due to rough competition in the show's production market, directors had to challenge themselves in making a movie genre that is suitable for everyone.
- Most directors are unsure on what type of movies that can attract a large audience.
- Most directors are unsure of what age category they should focus on to gain views.

II/ Roadmap:

1/ Data Gathering:

- Raw data collected from Kaggle:
    1) https://www.kaggle.com/datasets/shivamb/netflix-shows **netflix_titles.csv**
    2) Txt Document to convert the rating to age category: **rate_netflix.txt**
    3) Txt Document to convert the duration for movies to intervals (example if the movie duration is “50min” it will be transformed to “less than 60min”): **duration.txt**

netflix_titles.csv:

- The dataset contains 8809 rows of records and 12 columns of attributes.
- The dataset contains two types of attributes which are string and integer.
- Their dataset refers to the movie’s id, type, title along with the director of the show, the cast, country available, date added, release year, rating, genre and their description.
- The oldest show available in the dataset is from the year 1925 and latest is from 2021.

rate_netflix.txt:

rate_netflix - Bloc-notes

Fichier Edition Format Affichage Aide

```
Rating;Age_category
PG-13;Teens
TV-MA;Adults
PG;Older Kids
TV-14;Teens
TV-PG;Older Kids
TV-Y;Kids
TV-Y7;Older Kids
R;Adults
TV-G;Kids
G;Kids
NC-17;Adults
NR;Adults
TV-Y7-FV;Older Kids
UR;Adults
```

duration.txt:

duration - Bloc-notes

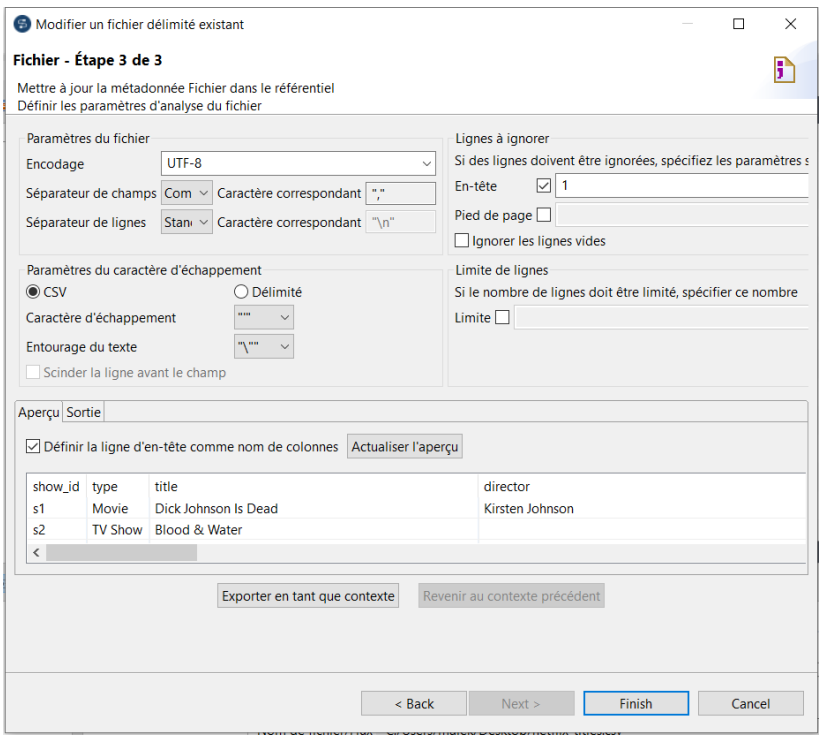
Fichier Edition Format Affichage Aide

```
Less than 60min;Between 60 and 90min;Between 90 and 120 min;More than 120min
<60min;[60min..90min];[90min..120min];>120min
```

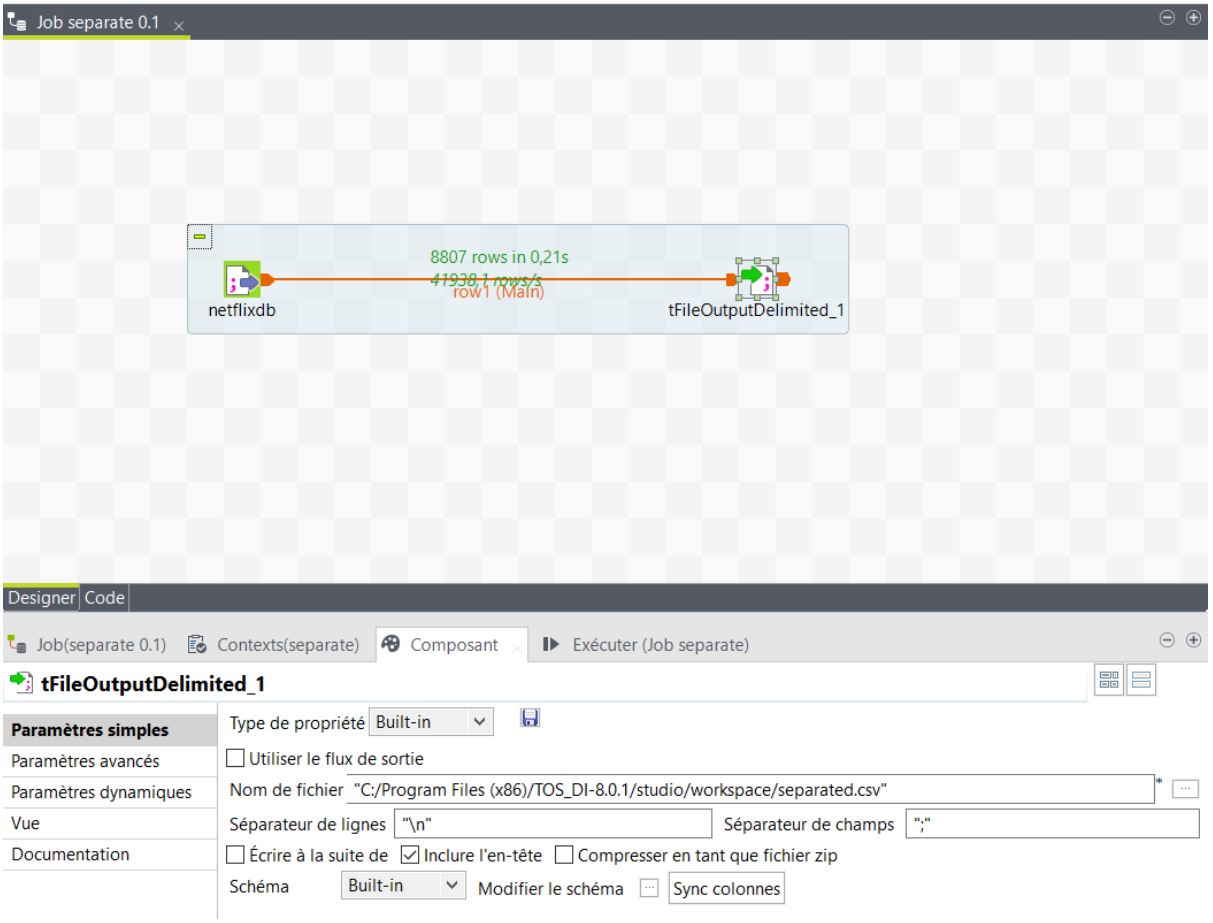
NB: you will find all the resources in a file called resources.

2/ Data cleaning and correcting faulty values

1) The first step consists of importing the raw data to talend.
   NB: data are separated by comma “,” .However, there’s a particular condition: we shouldn’t separate data that are between “ “ that contains inside them a comma “,”. So we should use “ as an escape character as shown down below.



To test the first step we wreated a new job called **separate**.



The result is an output file called **separated.csv**.

2) After splitting the database into columns, we would check if there's some issues in our database.
We created a copy from the **separated.csv** called **test.csv**
and used the UNIQUE function in excel to test the values of rating column.
The result was:

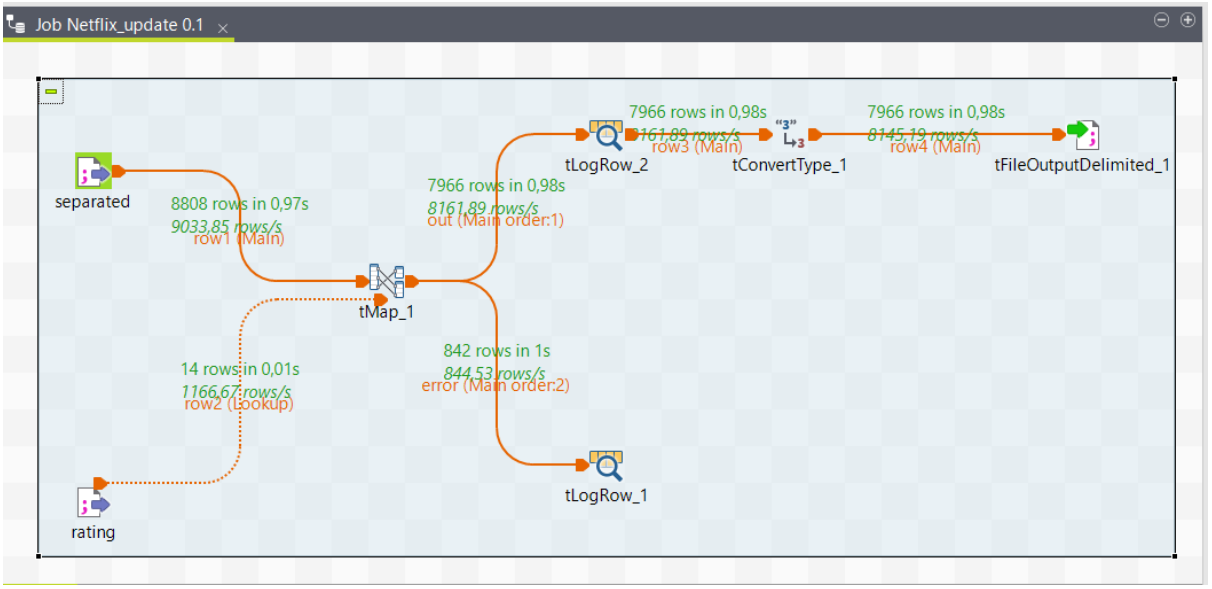
Unique rating
PG-13
TV-MA
PG
TV-14
TV-PG
TV-Y
TV-Y7
R
TV-G
G
NC-17
74 min
84 min
66 min
NR
TV-Y7-FV
UR
Classic Movies, Docun

In the rating column we found non appropriate values: 74min/ 84min/ 66min/ "classic movies, documentaries" which is a value of the type of the movie.

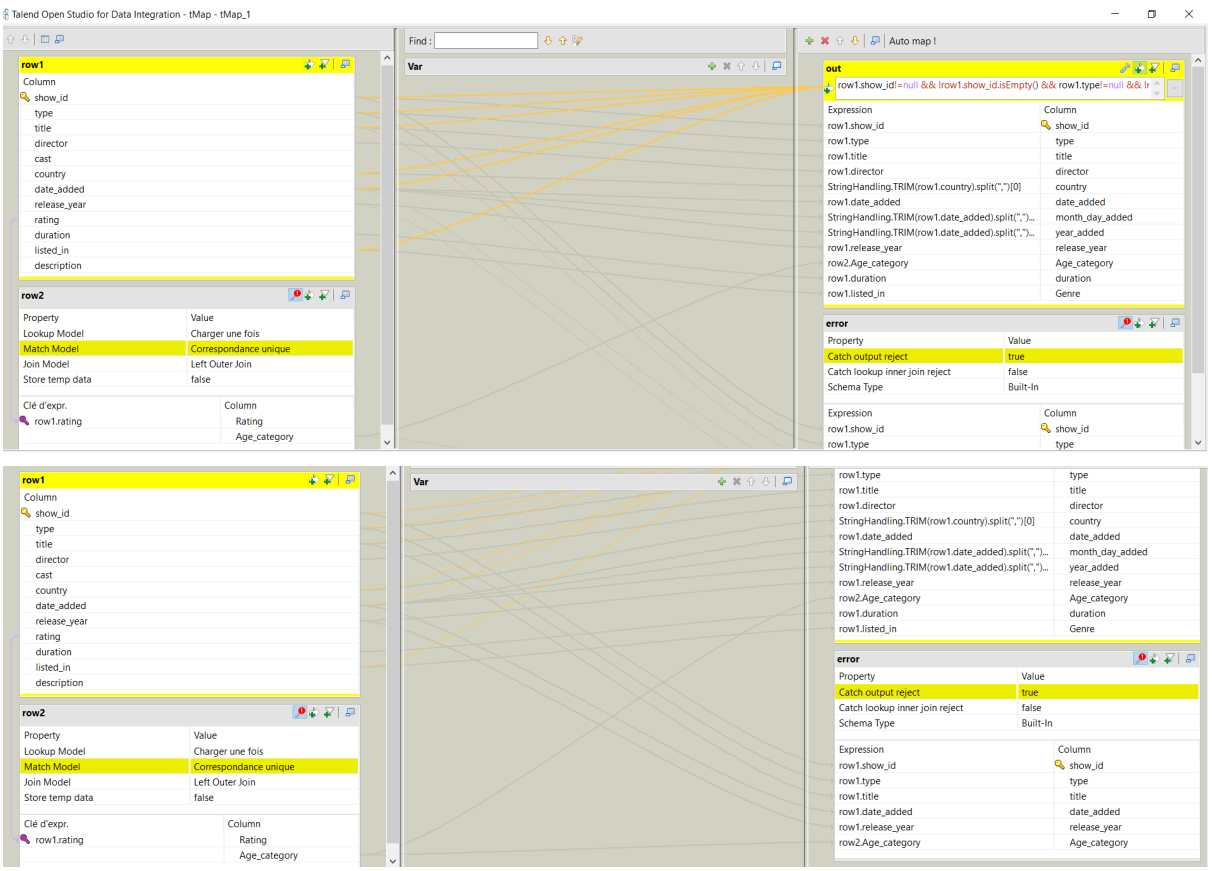
⇒ the solution is to: use the search method in excel (cntrl+F) to search for the shows corresponding to these weird values. Inspecting columns with '74min' , '66min' , '84min' rating manually with excel.
Remark: It appears that these weird ratings actually belong in the duration column.
So we fix it manually using excel.
NB: we fix the data in **separated.csv**

NR(nonrelated) and UR(unrelated) are the same
⇒ We should use just one of these values.

3) As a next step, we should change all rating of shows to other format which is Age_category.
We will create a new job called **Netflix_update** where we will use tMap to join the **separated.csv** with **rate_netflix.txt**.



Further details in tMap:



In tMap we put some condition for the output **out:**
row1.show_id!=null && !row1.show_id.isEmpty() && row1.type!=null && !row1.type.isEmpty() && row1.title!=null && !row1.title.isEmpty() && row1.date_added!=null && !row1.date_added.isEmpty() && row1.country!=null && !row1.country.isEmpty() && row1.listed_in!=null && !row1.listed_in.isEmpty()
⇒ show_id, type, title, date_added, country, listed_in shouldn't contain null or empty fileds.
Any error of these conditions is added to tLogRow.

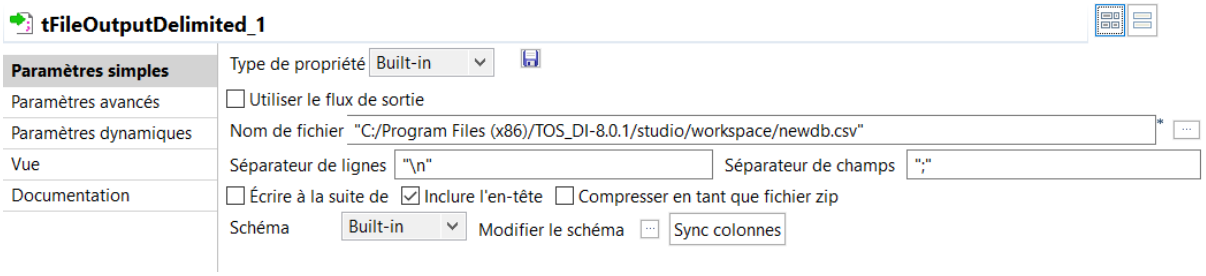
We also changed the name of listed_in column to genre to make it more understandable.

For the country column: we choose to simplify the data in it and choose only the first country in every list available.
we used: StringHandling.TRIM(row1.country).split(",")[0]
to delete every space or tab in column and separate data between comma "," and choose the first value in every list.
To make the split function work we should ensure that every country value shouldn't be null or empty.

We should simplify the date_added column because it's composed to day, month and year (the format is for example "September 25, 2021").
we will just keep the year and transform the date_added type from string to date in format of "yyyy".
We separated the column date_added into:
month_day_added: StringHandling.TRIM(row1.date_added).split(",")[0]
year_added: StringHandling.TRIM(row1.date_added).split(",")[1]

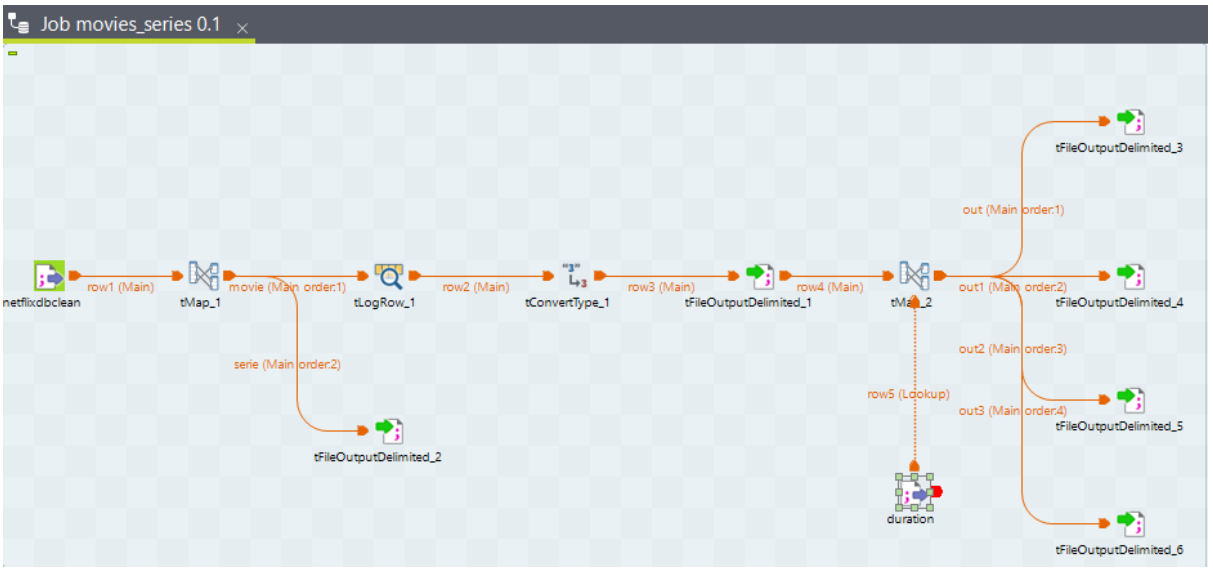
The output **out** is sent to tLogRow and then we added tConvertType component to change the year_added type from string to date type in format of "yyyy".
We also eliminated date_added and month_day in tConvertType by eliminating them from the output.

Then the final output of the job **Netflix_update** is sent to file called **newdb.csv** in workspace using the tFileOutputDelimited component.

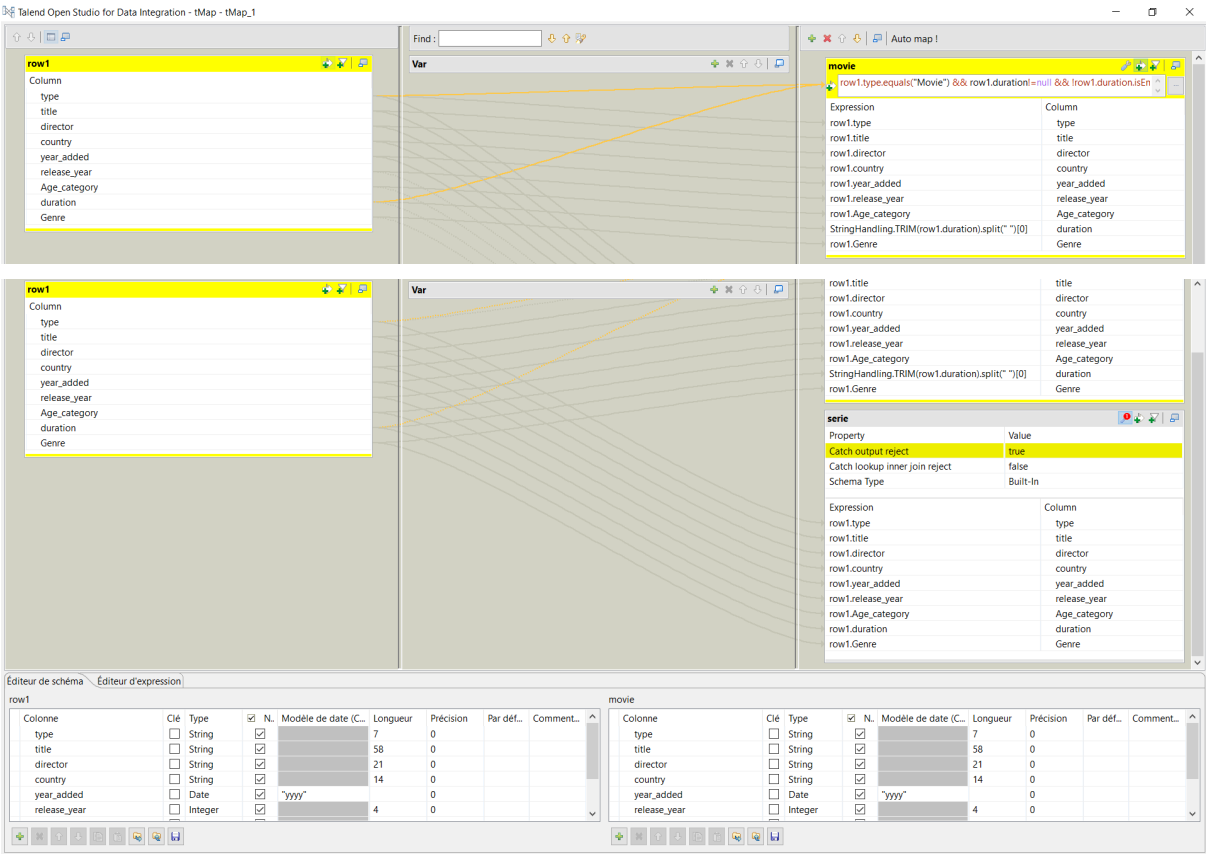


4) This step consists of grouping movies together and tv shows together and putting them in the same csv file.
Beside that, we will transform the duration data of every movie from minutes to intervals based on **duration.txt** file.
for example: 55min would be transformed to <60min
⇒the solution is to transform the duration for movies from string type to integer type in order to compare them and then transform them again to string type.

We will create a job called **movies_series**



The first step in the job is to separate movies and series to different csv files.
We will use for that tMap component.



For the movie output we made a condition:
the type field should be “Movie” and the duration shouldn’t be null or empty.
row1.type.equals("Movie") && row1.duration!=null && !row1.duration.isEmpty()

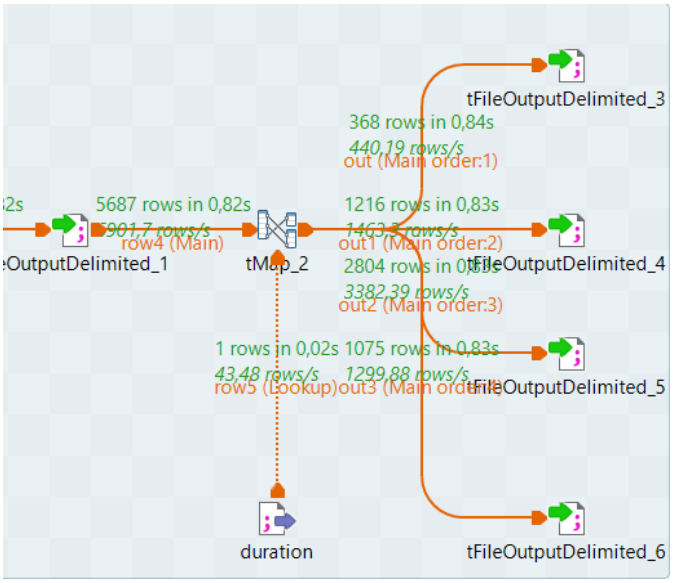
In the movie table output, we will transform the format of the duration.
<u>for example:</u> “74 min” is transformed to “74”
by using the following expression: StringHandling.TRIM(row1.duration).split(" ")[0]

in order to convert it later into string and compare it in future steps.

The output of the tv shows is sent to file called **series.csv** in the workspace.

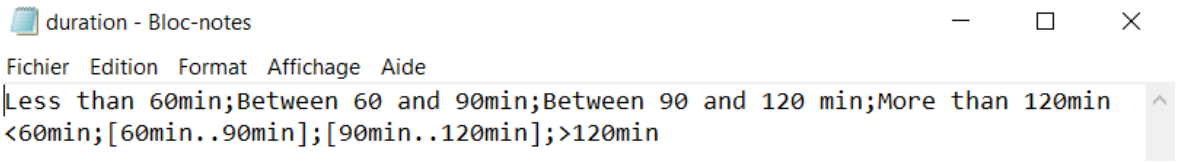
However, the movie output is sent to tLogRow component where it will pass by tConvertType in order to change the type of the duration from **string** to **integer**.

And then it will pass by tFileOutputDelimited where the result will be sent to file called **movies.csv** in workspace to test if the job is working correctly.

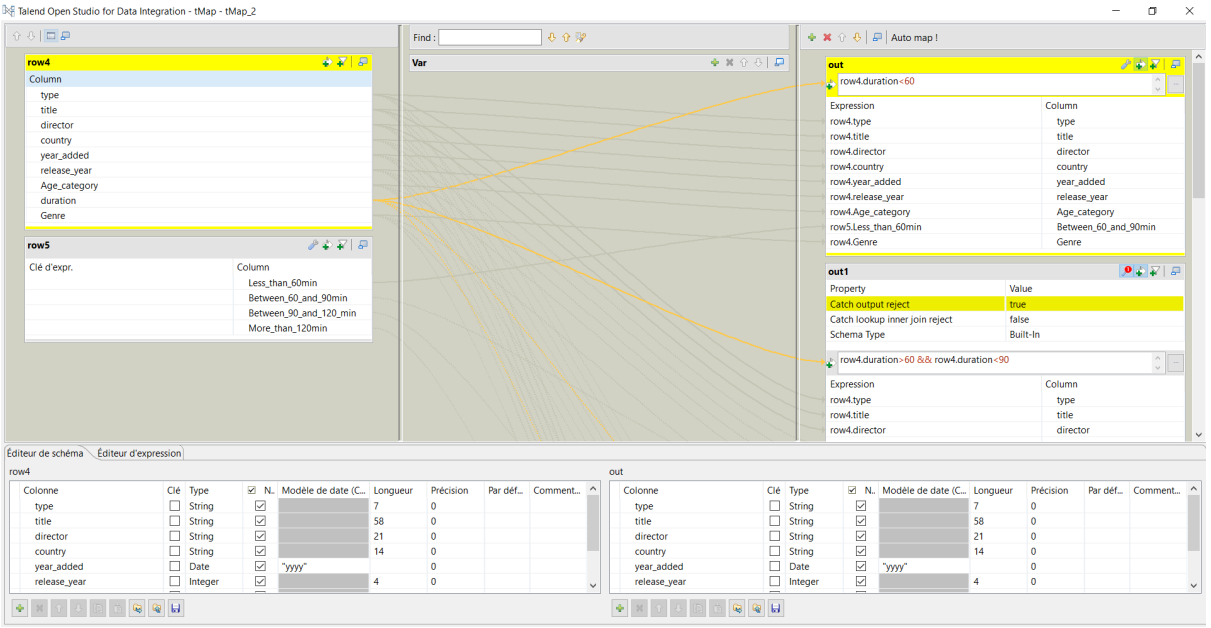


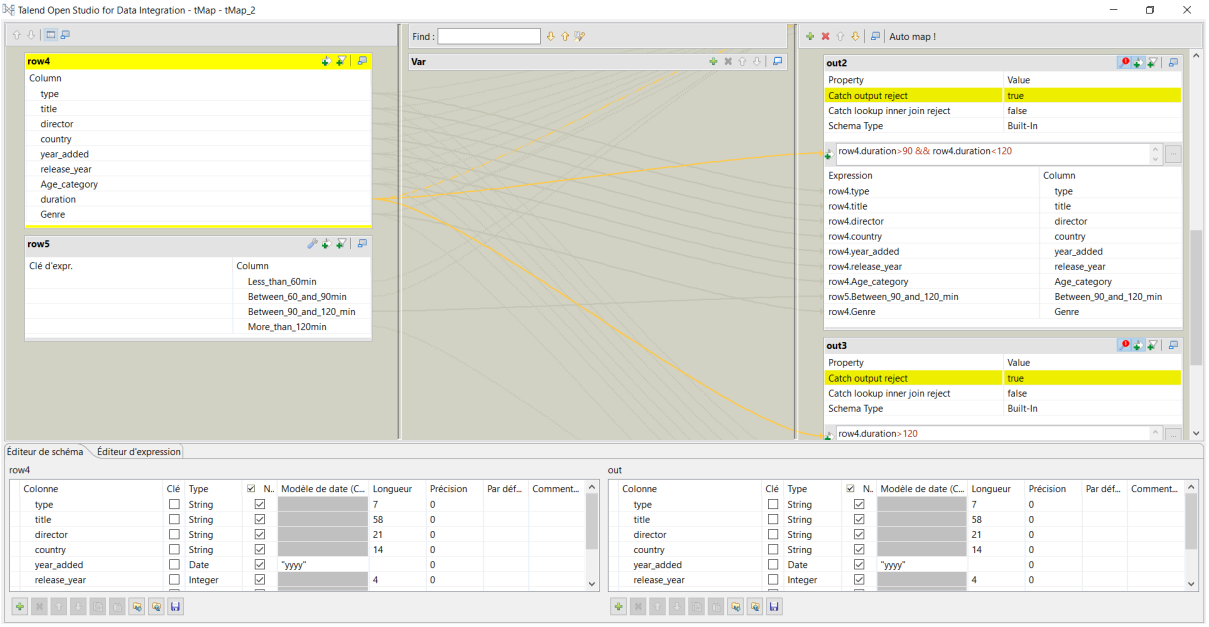
As a next step, we will filter the **movie.csv** depending on their duration.
we will divide the file into 4 files where:
file1: contain movies with duration less than 60mins
file2: contain movies with duration between 60 and 90mins
file3: contain movies with duration between 90 and 120mins
file4: contain movies with duration more than 120mins.

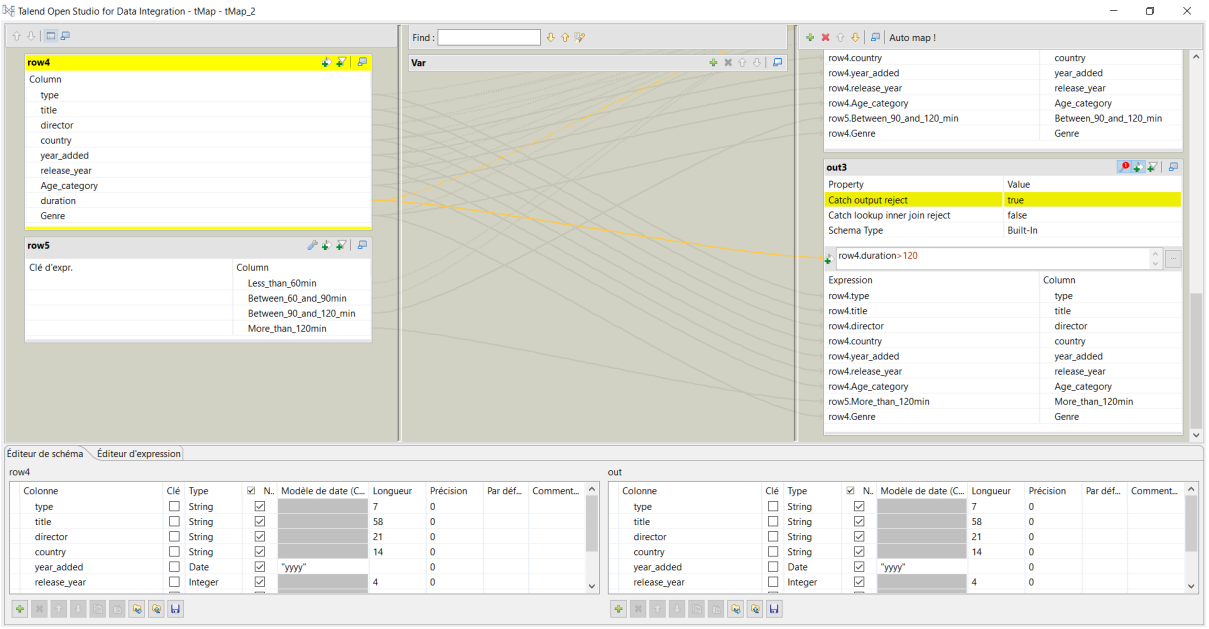
Meanwhile, the duration data type will be transformed depending on the **duration.txt** using tMap.



tMap component details:







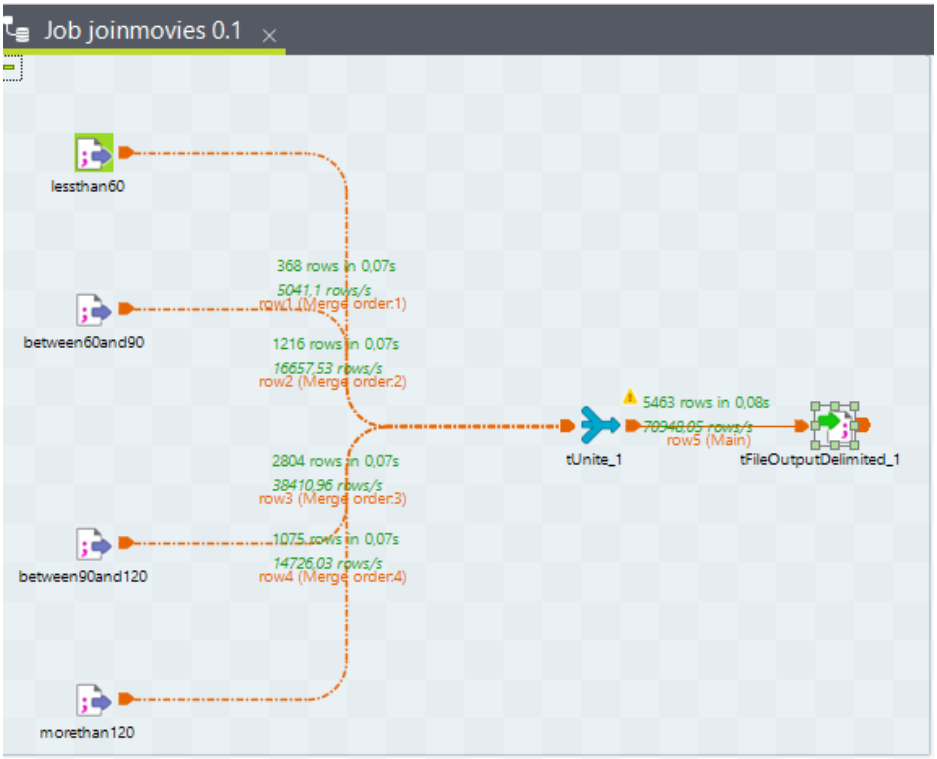
for **out** we put the following condition: row4.duration<60
for **out1** we put the following condition: row4.duration>60 && row4.duration<90
for **out2** we put the following condition: row4.duration>90 && row4.duration<120
for **out3** we put the following condition: row4.duration>120

the 4 files will be sent in a document in the workspace called **duration**:

between_60_and_90	20/01/2023 20:31	Fichier CSV Micros...	139 Ko
between_90_and_120	20/01/2023 20:31	Fichier CSV Micros...	321 Ko
lessthan60	20/01/2023 20:31	Fichier CSV Micros...	41 Ko
morethan120	20/01/2023 20:31	Fichier CSV Micros...	115 Ko

Now we have our movies separated based on their durations.
As a next step we should collect them into one file containing all movies.

We will create a new job called **joinmovies**.

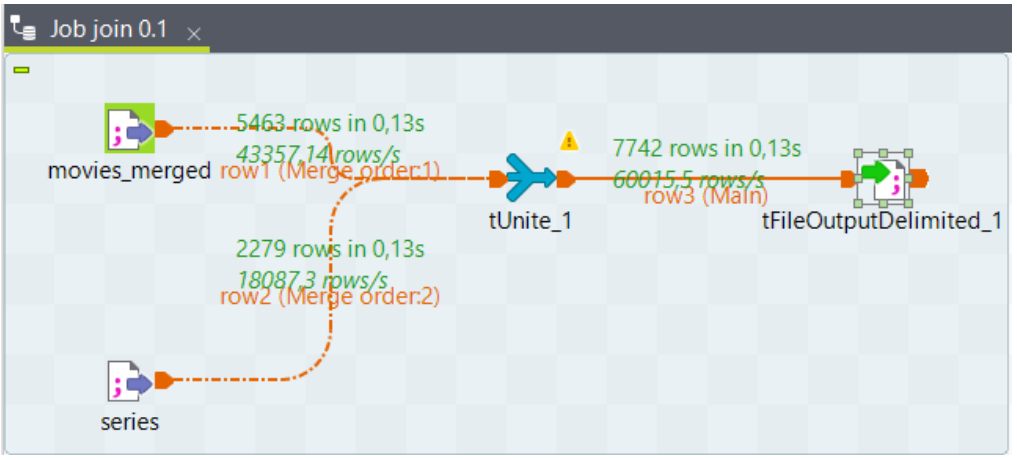


We will use tUnite component to merge all files into one file.
The main condition of tUnite component is that all files should have the same structure and data types in order to merge them together.

The result of tUnite component will be sent to tFileOutputDelimited component where the result will be sent to **movies_merged.csv** in workspace.

5) In this final step we are required to merge the **movies_merged.csv** and **series.csv** into one csv file.

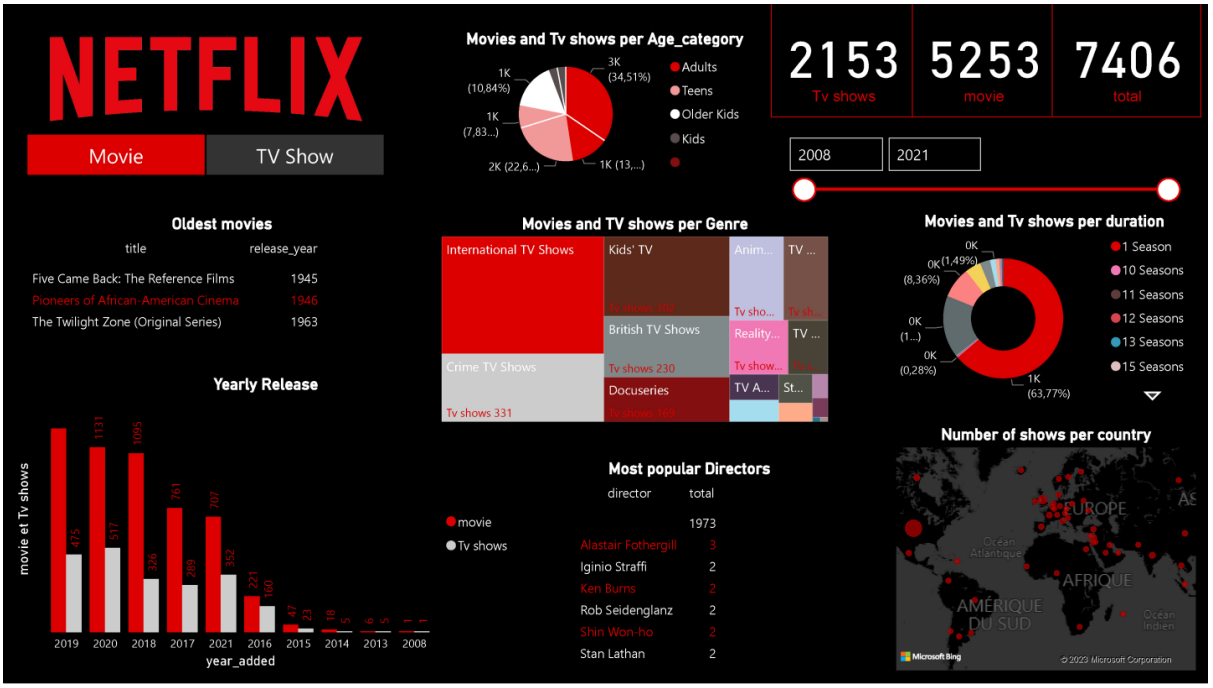
we will create a job called **join**.



We will use tUnite componenet to merge the files of movies_merged and series into one file.
The result of tUnite component will be sent to tFileOutputDelimited component where the result will be sent to **final_netflix_db.csv** in workspace.

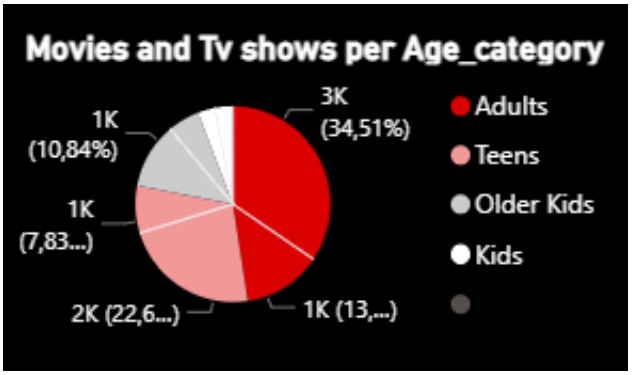
⇒**Final version of the cleaned database:** final_netflix_db.csv

3/ Data analysis:

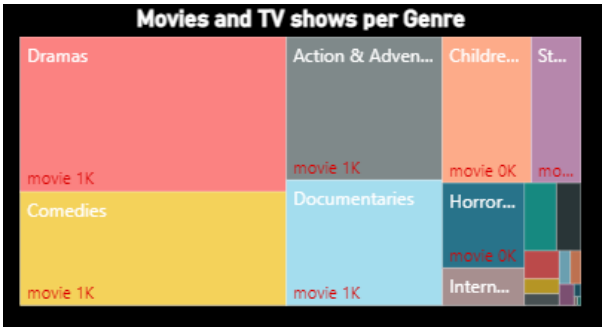


1) What type of content is available in Netflix database?

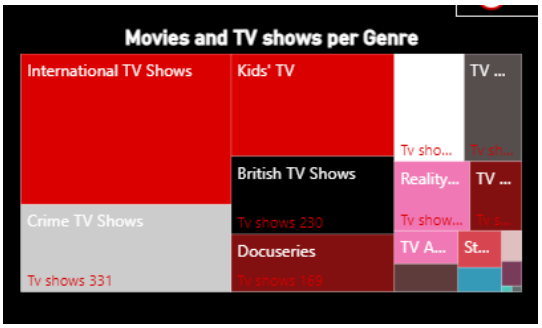
What is the ratio of shows vs movies on netflix?
We have in total 7406 shows: 2153 of them are TV shows and 5253 are movies.
29% of shows are TV shows and 71% are movies.
⇒ movies type of shows are dominating the Netflix database compared to TV shows.
What amount of content was added each year?
During the latest 3 years:
2021: 707 movies and 352 TV shows were added. ⇒ 33.2% of TV shows
2020: 1131 movies and 517 TV shows were added. ⇒31,3% of TV shows
2019: 1247 movies and 475 TV shows were added. ⇒27,5% of TV shows
2018: 1095 movies and 326 TV shows were added. ⇒22.9% of TV shows
2017: 761 movies and 289 TV shows were added. ⇒27.5% of TV shows
⇒ the ratio of TV shows began to increase from 2019.
⇒Since 2019, Netflix began to focus more on TV shows more than movies.
What are some of the oldest Tv shows and movies on Netflix?
The oldest movie is: *Five came Back: The reference Films*



The chart shows that the majority of Netflix's movies and Tv shows offerings are geared towards adults and teenagers. This could suggest that these age groups are the primary target demographics for the streaming service. It could also indicate that there is a larger market for content that appeals to adults and teens, and therefore Netflix is focusing on producing and acquiring content that will appeal to these audiences.

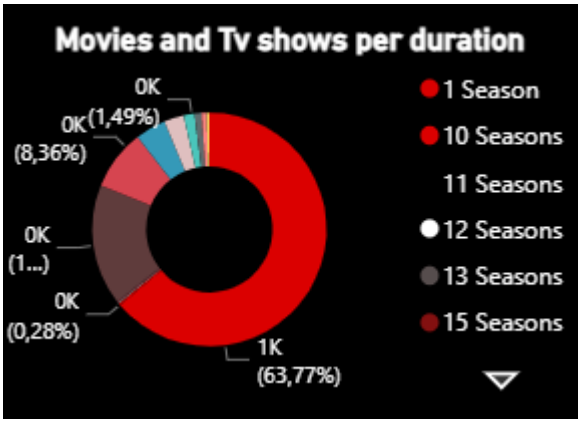


The majority of Netflix's movie offerings are in the genres of dramas, comedies, action and adventure, and documentaries. This could suggest that these are the genres that are most popular among Netflix's audience and therefore Netflix is focusing on producing and acquiring content that falls within these genres. It could also indicate that these genres are the most profitable for the streaming service, and therefore they are investing more resources in creating and acquiring content in these genres.
It may also mean that these genres are the most versatile, and can appeal to a broad audience, in contrast to genres like horror, for example, that may only appeal to a specific niche.

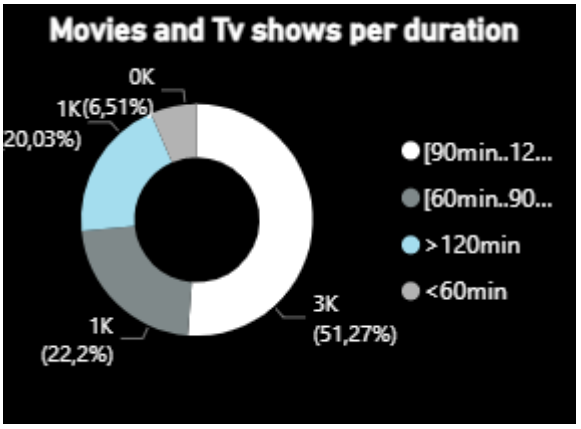


It seems that the majority of Netflix's TV show offerings are in the genres of international TV shows and crime TV shows, followed by kids TV shows. This could suggest that these are the genres that are most popular among Netflix's audience for TV shows, and therefore the company is focusing on producing and acquiring content that falls within these genres. It could also indicate that these genres are the most profitable for the streaming service and

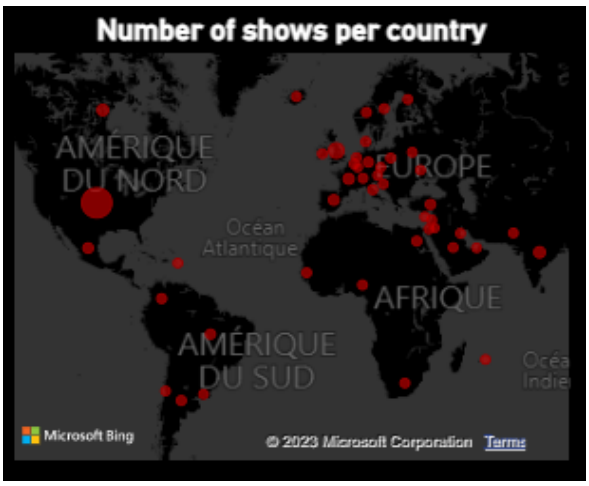
therefore they are investing more resources in creating and acquiring content in these genres.
It may also indicate that Netflix is trying to diversify its offer by incorporating international shows and breaking into the crime genre. Additionally, the presence of kids TV shows could mean that Netflix is trying to appeal to families and provide content for children as well



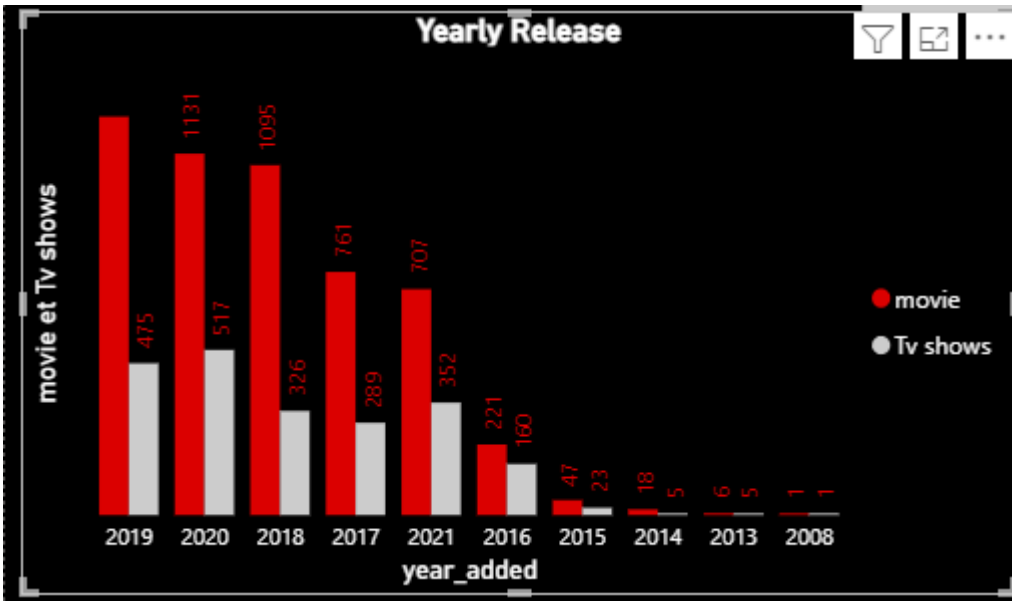
This chart suggests that Netflix is releasing a large number of TV shows that only have one season. This could indicate that Netflix is prioritizing the release of new, original content over continuing existing shows for multiple seasons. It could also mean that Netflix is canceling shows after one season if they do not perform well or do not align with the company's programming strategy. Another possibility is that Netflix is releasing shorter series, known as "limited series" which are pre-planned to have a set number of episodes and a defined story arc.



This chart suggests that Netflix is releasing a large number of movies that fall within the 90-120 minute range. This could indicate that Netflix is focusing on producing movies that are of a certain length, as they may have found that this length is popular among viewers. It could also suggest that Netflix is targeting a specific audience that prefers movies of this length. Another possibility is that they have found that this is the optimal length to keep the audience engaged without making the movie too long or dragging. Additionally, this length is common for many films, regardless of the platform they are being released on.



This map suggests that the majority of the films and TV shows on Netflix are set or produced in the United States, with Europe following closely behind but with a lower percentage. This could indicate that Netflix is primarily focused on producing and acquiring content from these regions. It could also reflect the fact that the United States and Europe have well-established film and television industries, making it easier for Netflix to access and acquire content from these regions. Additionally, this could also be due to the fact that most of the audience is based in these regions making it more profitable for Netflix to produce content that aligns with the audience's preferences.



This chart suggests that most of the movies and TV shows on Netflix were released between 2016 and 2019, with a significant spike in releases in 2019. This could indicate that Netflix was particularly focused on increasing the amount of content it released during this period. The spike in 2019 specifically could be due to the company's strategy to increase the amount of original content it produces and releases. It could also reflect the company's growing budget and resources to produce more content. Additionally, this could also be due to the increasing competition in the streaming industry and the need to keep up with the audience's demand.