Ontario Tech Talent

Canopy's curated movies

Final Assignment

Mahalakshmi Surya

6-9-2023

Table of Contents

Introduction

Canopy is a boutique streaming service that plans to cater to the viewers of French-language movies. Their current business goals, as a streaming service provider, are to:

- offer curated selections of the best-rated French-language movies from the existing content.
- make French-language movies available to various age groups.
- identify the least tapped genres to provide the filmmakers with data to make original content for Canopy.

Our Goal

In this project, we are going to analyze movie datasets to gather insights that adhere to the strategic objectives to expand Canopy streaming services.

The key analytical details that are answered as a result of the analysis include:

1. Find the top 20 movies in each genre that is available in the French language.
2. Analyze the genre distribution in French-language movies.
3. Analyze French movies by streaming platform.
4. Make curated collection of movies by -
    a. Rating
    b. Year of release
    c. Rating of the directors

Data Description

We are going to use two datasets in csv file format for the analysis.

Dataset-1: Movies Info

Dataset-2: Movies Streaming Info

Why Python for our data analysis?

Python is one of the best programming languages available for data analysis. It offers a variety of libraries such as NumPy, Pandas, Matplotlib etc. specifically designed for analysis and visualizations purposes. These libraries simplify the complexity of data analysis tasks and help to derive valuable insights and trends.

As the data source used for the analysis is of csv type, Python provides a simple coding syntax for loading and reading csv type data. The Pandas library helps to convert the data into dataframe for easy readability and data manipulation.

With visualization libraries we can easily convert the manipulated data into charts and plots.

Data Processing Pipeline

Import Python Library

We are going to use the following python libraries,

NumPy: Numerical Python contains functions that can be used for all kinds of numerical operations in the data analysis process using Python.

Pandas: Pandas provide fast, flexible, and expressive data structures designed to make working with 'relational' or 'labelled' data both easy and intuitive.

Matplotlib: Matplotlib is used to create data visualizations elements such as charts, graphs and plots.

Seaborn: Seaborn provides higher level API for creating statistical visualizations.

Load datasets from csv files

The csv files are downloaded and saved into a folder called dataset. The *read_csv()* function helps to import the data from csv files to Pandas dataframe.

Combine the datasets

The *merge()* function helps to combine the two datasets into one. This helps in combining the categories that are common to both dataframe. From our analysis, *Title* of the movie is the common column connecting both datasets. The combined dataset will be henceforth called as *movies dataset.*

Handling the missing data

Handling the missing data is an important part of data analysis and manipulation. In order to clean data, we must find the total number of rows & columns in the dataset, columns names, non-null count and data type. This can be achieved using *info()* method of Python.

In our movies dataset, the total number of rows is 13423 and the non-Null count against each column is found using *info()* method.

- The column Rotten Tomatoes has a minimum non-null count of 4121 rows and is not critical for the analysis. Hence, it can be removed from the dataframe with the help of drop() method.
- The NaN values in the object data type columns such as Age, Language, Directors, Genres, Country are filled with the value 'Not Specified' using the fillana() method.
- The NaN values in the integer data type columns such as IMDb, Runtime are filled with '-1' using the fillana() method.

Statistical Summary

The describe() method shows the statistical summary of numerical columns present in the dataframe. With the parameter *include='all'* provides the statistical summary of all the columns in the dataframe. The output shows the total count of movies, with minimum & maximum year of release, top language & genre etc.

Finding the top 20 movies in each genre that is available in the French-language

Let's analyze the dataset to find top 20 movies in French language. Firstly, the data set is filtered to French language movies by setting a filter on language column with the value French. Also, the data set has other language movies that has been dubbed in French. We are going to use the French dubbed movies as well in our analysis. To categorize movies, we apply the below logic,

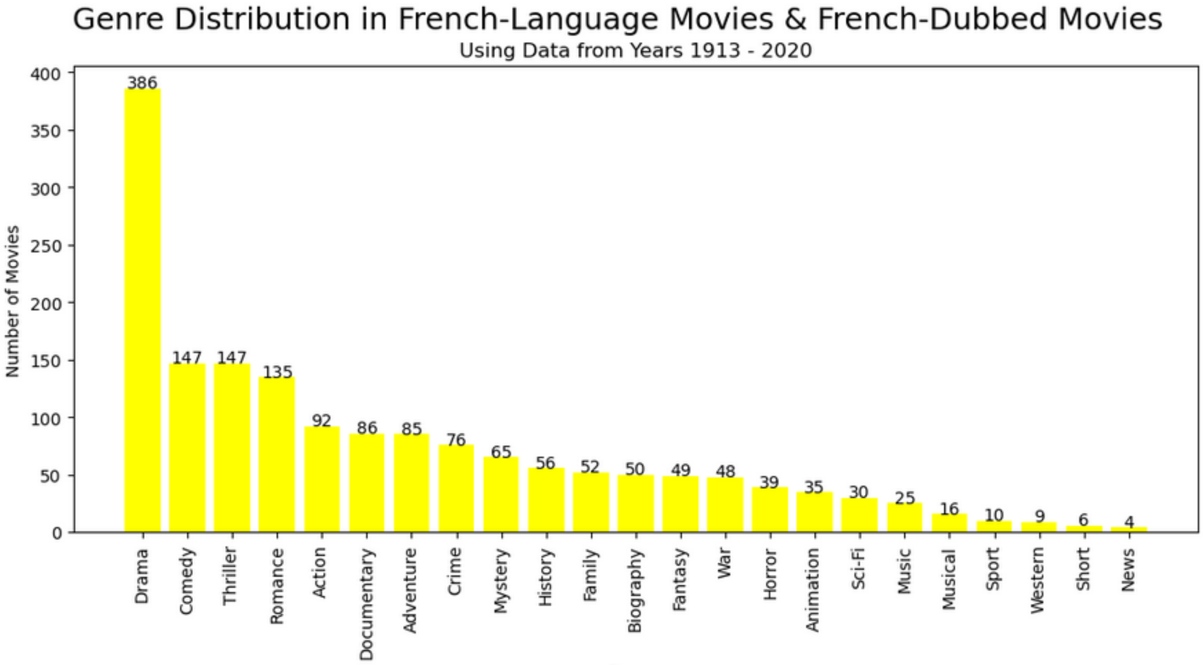
1. If the language of the movie is only French, it is categorized as a French Movie
2. If the language of the movie has multiple languages and starts with French, it is categorized as a French Movie
3. If the language of the movie has multiple languages and French is not the first language and part of the language list, it is categorized as a French Dubbed
4. Else, it is categorized as Not French

After categorizing the movies, we are going to categorize the genre for each movie. Once the individual genre for each movie is found, we are going to merge the rest of the information from the movies dataset with it. This allows us to add columns like Directors, Age, Year, Streaming platforms etc. for further analysis. To find the top 20 movies, we are going to use group the data with Genre and use nlargest() method.

This analysis will help Canopy to stream the top-rated French-language movies and French-dubbed movies for each genre.

Analyzing the genre distribution in French-language movies

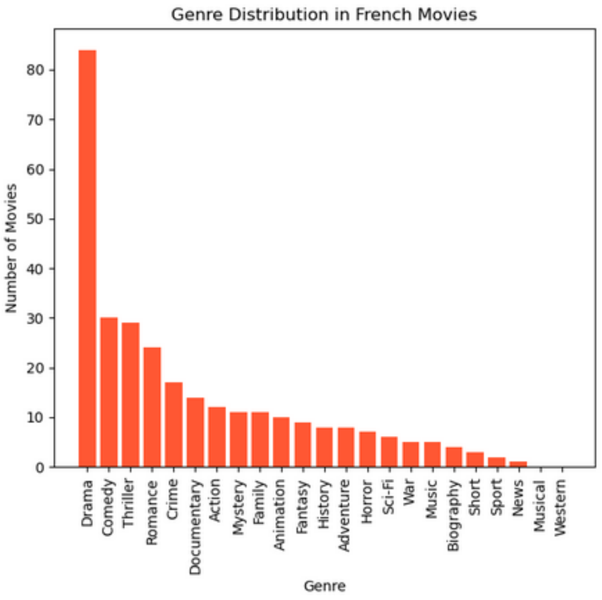
In order to find the genre distribution, lets group the dataframe created for French movies with genre and aggregate using count() method to find the total number of movies against each genre.



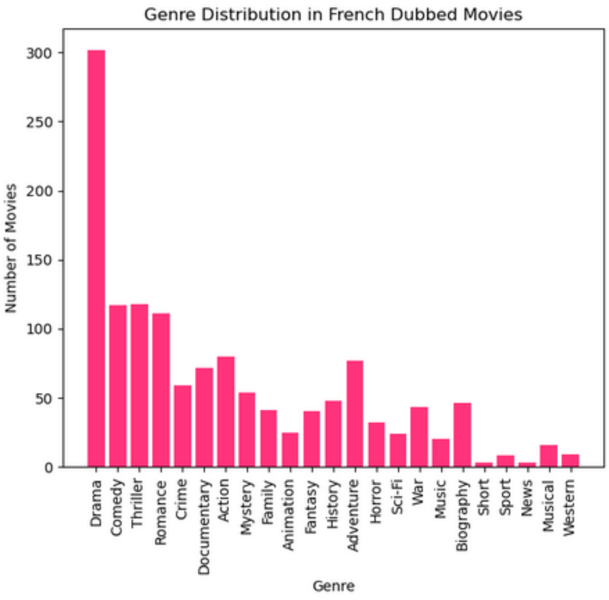
Identifying the least tapped genres in French movies

In order to find the lease tapped genres made in French movies, we are going to compare the number of movies made in French language with the one that is dubbed as French movie.

When looking at the below analysis, the least tapped genres are Western, Musical, News and Sport. French filmmakers can make original content in these genres.

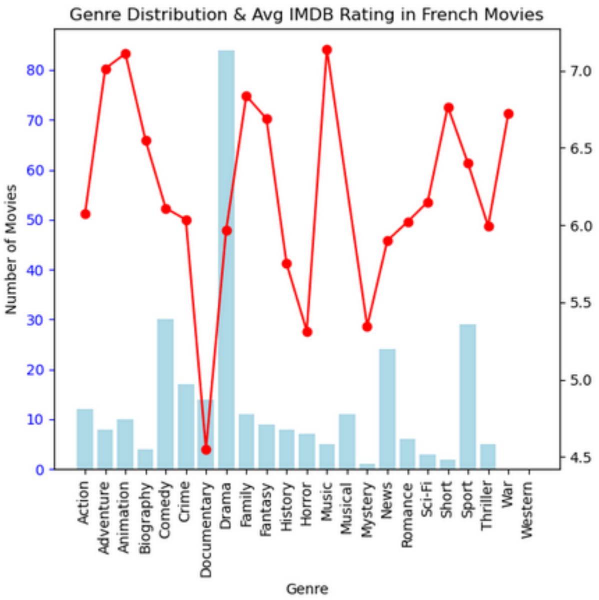


Genre Distribution in French Movies

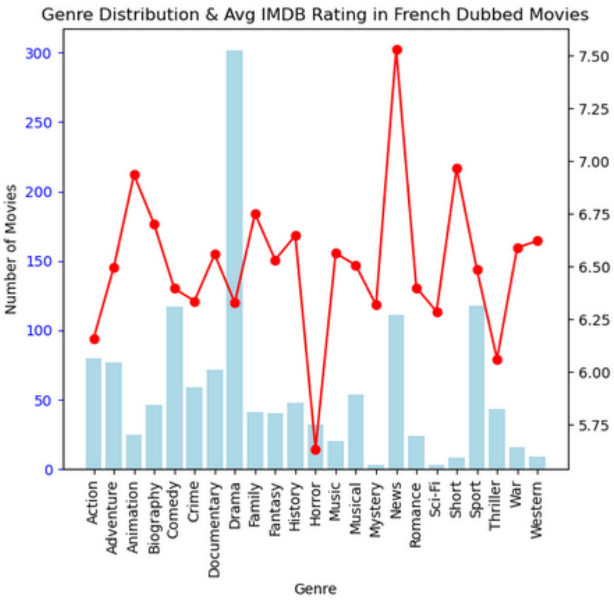


Genre Distribution in French Dubbed Movies

The below analysis is based on average rating in each genre of French language and French dubbed movie.



Genre Distribution & Avg IMDB Rating in French Movies

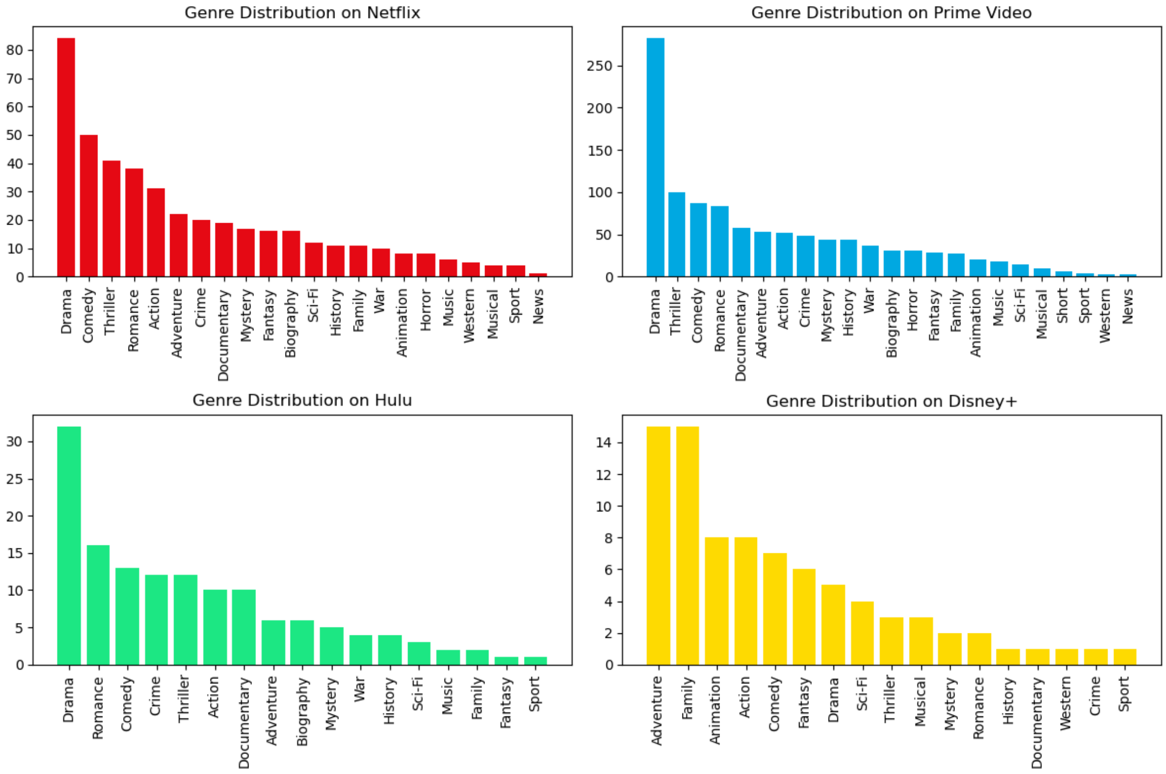


Genre Distribution & Avg IMDB Rating in French Dubbed Movies

News is the genre in the French-dubbed movies with the highest rating which French film makers can make original content. The least tapped genres in French-language movies are News, Short, Animations, Family, Biograph, History and so on.

French Movies by Streaming Platform

In order to find the genres streamed across various platforms, we are going to group the dataset with platforms and the number of movies.



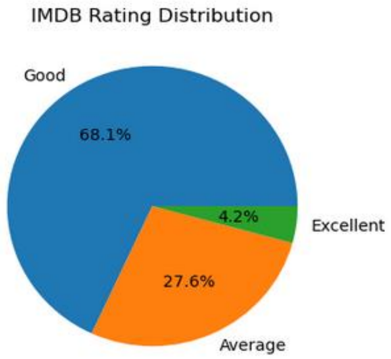
According to the above analysis, the least popular of all the platforms are Sport, Musical, Western, War, Sci-Fi and so on. These genres are underrepresented and could be a potential opportunity for Canopy to offer movies in these genres.

Making curated collections of movies

Canopy can offer curated collections of movies by rating, year of release, directors, actors, age group, genres etc. Below are three collections that canopy might be interested in.

a. Top rated movies

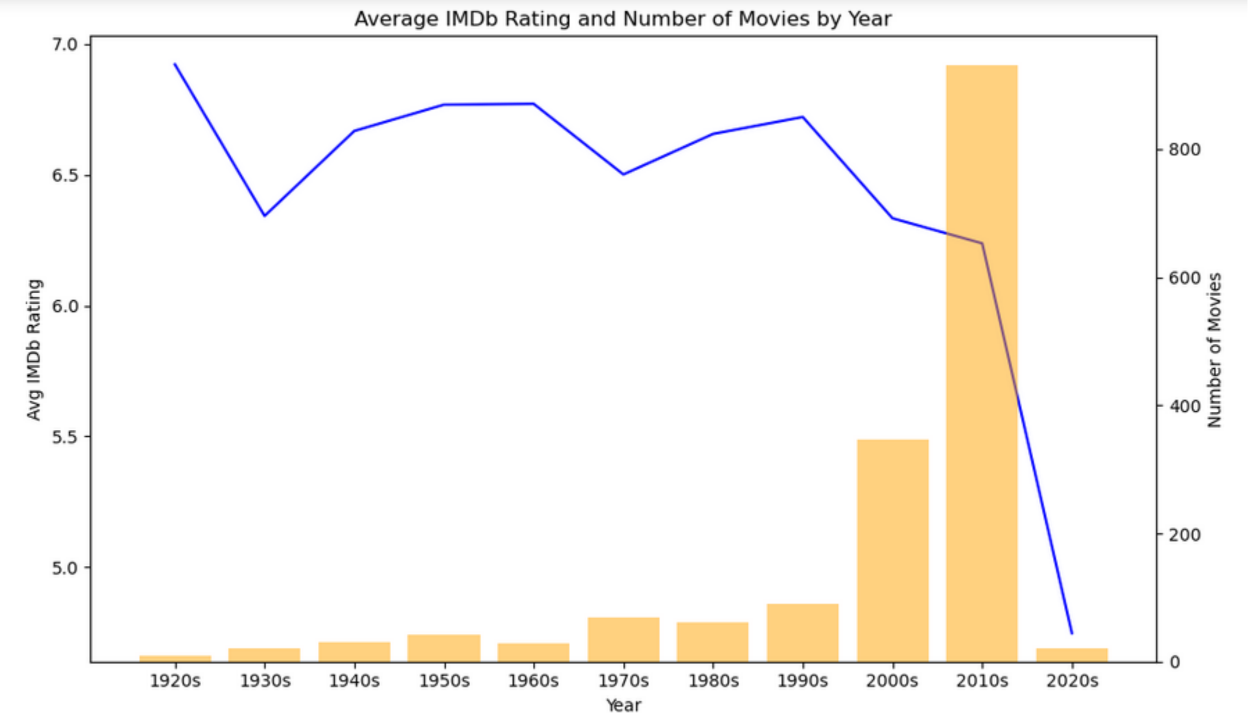
In this analysis, we are going to categorize IMDb rating as Excellent, Good and Average and save them in a new column as IMDB Category.



Canopy can create a collection of top-rated movies that falls under the Excellent category as ‘People’s choice’ and the Good category as ‘Popular content’.

b. Year of release

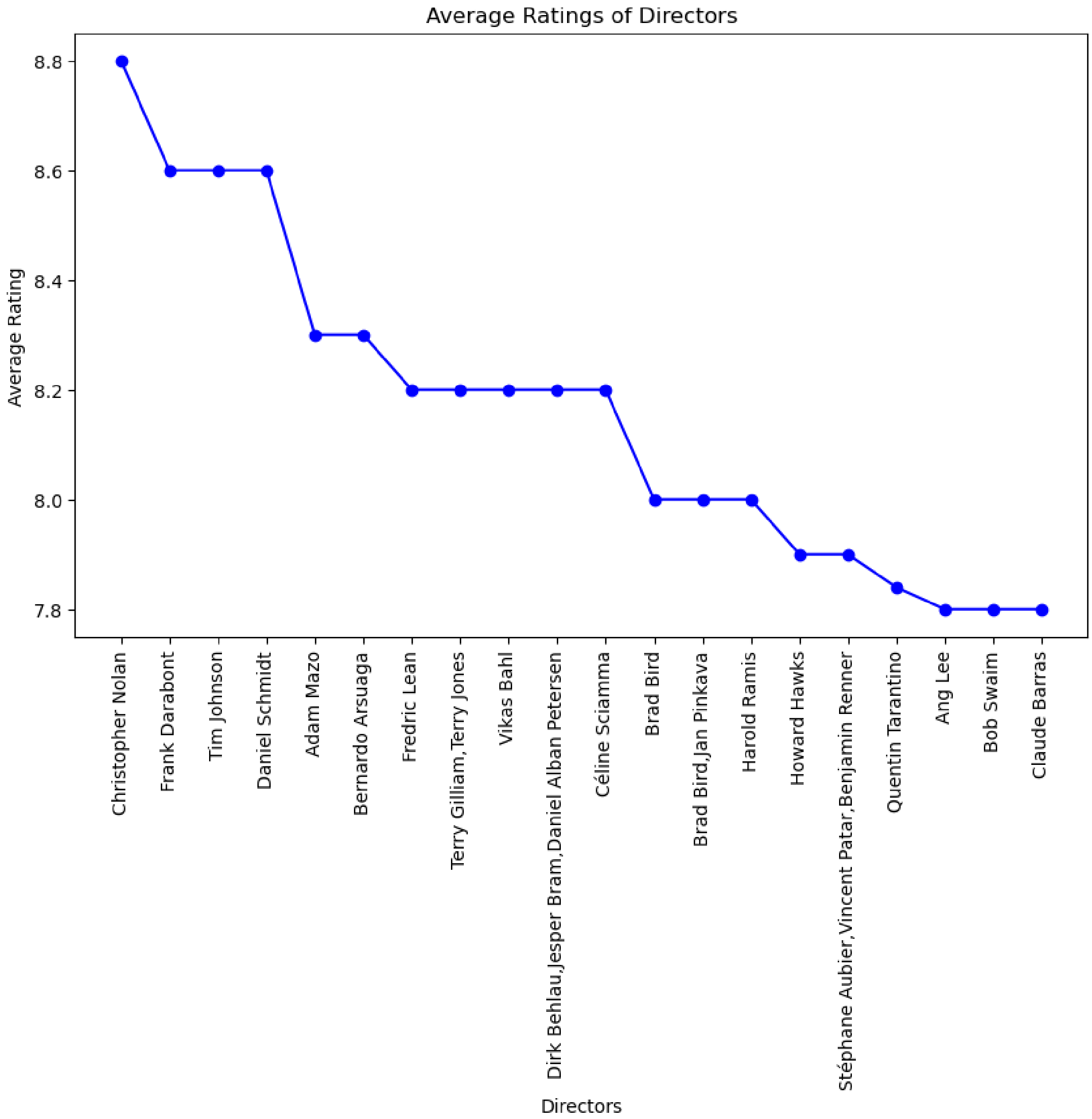
In the next collection, we are going to group French movies based on year of release. Collecting movies by each decade can create groups like 1920s, 1980’s, 2000’s etc. Along with that, the average IMDb rating for each decade is calculated to see the rating trend over years.



This helps Canopy to create collections like Nostalgic movies, Cult Classics etc. It can be catered according to the age of the viewers as their childhood movies, all-time favorite movies etc.

c. Top rated directors

Directors are one of the top reasons to watch a movie. It makes it easier for the viewers to choose top directors movies from a director list. We are going to find the top-rated French movie directors with the number of movies and average rating. In the image below, Christopher Nolan, Frank Darabont and Tim Johnson are the top-rated directors with an average rating of 8.


Average Ratings of Directors
Average Rating
8.8
8.6
8.4
8.2
8.0
7.8
Christopher Nolan
Frank Darabont
Tim Johnson
Daniel Schmidt
Adam Mazo
Bernardo Arsuaga
Fredric Lean
Terry Gilliam,Terry Jones
Vikas Bahl
Dirk Behlau,Jesper Bram,Daniel Alban Petersen
Céline Sciamma
Brad Bird
Brad Bird,Jan Pinkava
Harold Ramis
Howard Hawks
Stéphane Aubier,Vincent Patar,Benjamin Renner
Quentin Tarantino
Ang Lee
Bob Swaim
Claude Barras
Directors

Conclusion

Based on the analysis, French movies (original and dubbed) made in Drama, Comedy, Thriller, and Romance are the most popular. The least tapped genres are News, Short, Western, and Sport. So, Film makers can make original content in these genres. Among all the streaming platforms, the least popular genres are Sport, Musical, Western, and War. These genres are potential opportunities for Canopy, and it could bring a newer set of audiences. Canopy must make a curated collection of movies with ratings, age groups, year of release, top directors etc. as highlighted in our analysis.

Word count summary

Total number of words = 1472

Counted from page 2 (Introduction) to page 14 (Conclusion)

Cover page, table of contents, and word count summary pages have been excluded from the word count.