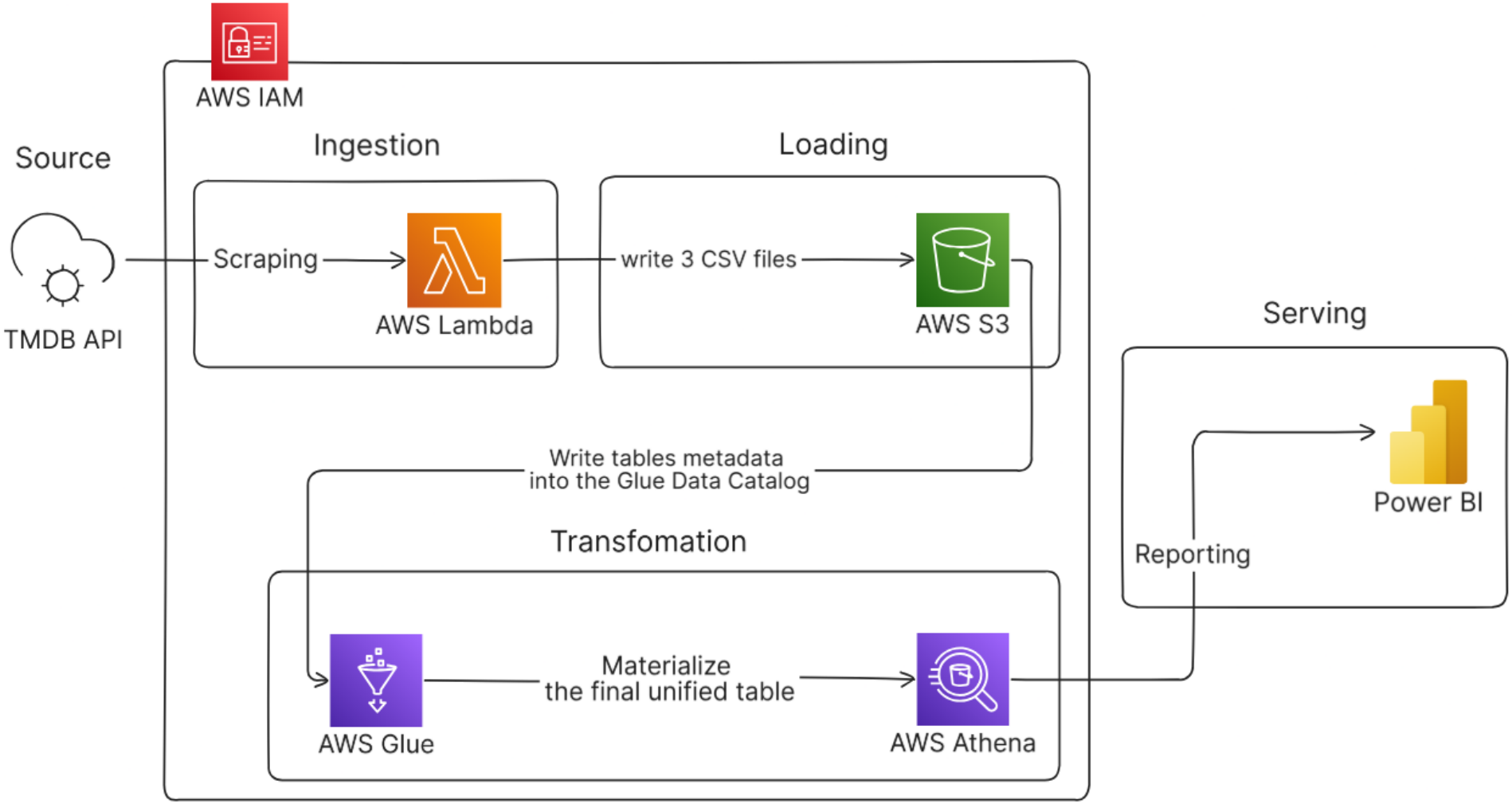

AWS IAM
Source
Ingestion
Loading
TMDB API
Scraping
AWS Lambda
write 3 CSV files
AWS S3
Serving
Power BI
Reporting
Write tables metadata
into the Glue Data Catalog
Transfomation
AWS Glue
Materialize
the final unified table
AWS Athena

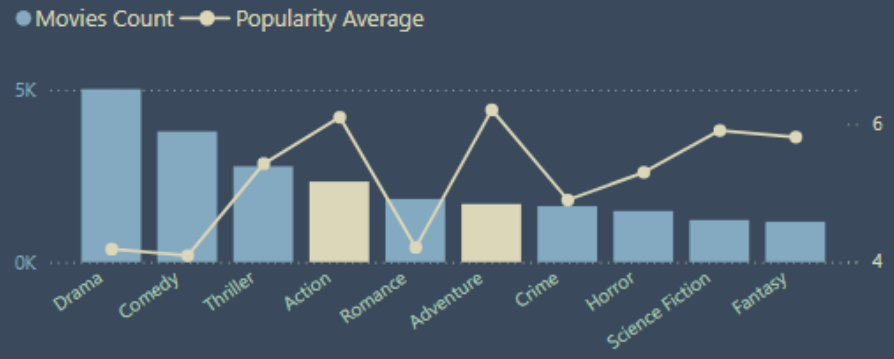
# US Box office (1902:2024) Votes & Popularity Analysis

**English Movies**
Vote Average
6.5

**Non-English Movies**
Vote Average
6.1

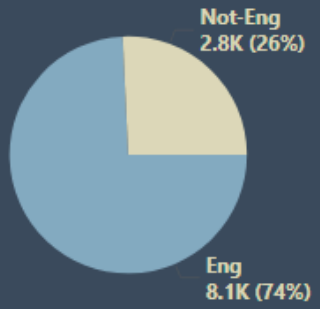
YEAR
1902 — 2024

## Top 10 Genres by movies count vs its popularity Average



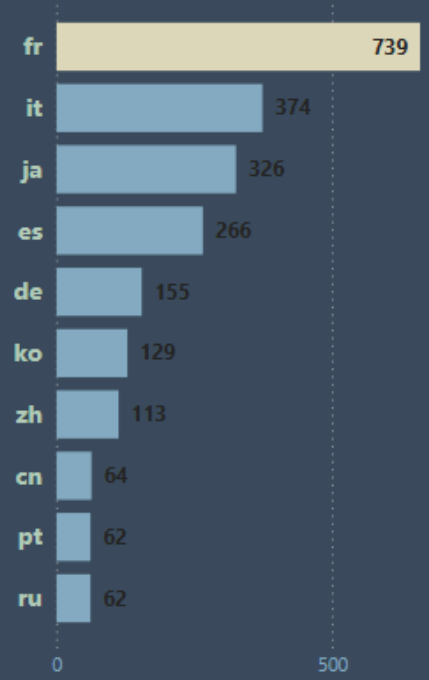

## Eng Movies VS Non-English

(By movies count)



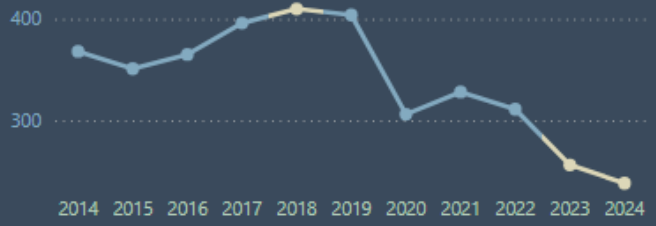

## Top 10 Foreign Languages

(By movies count)



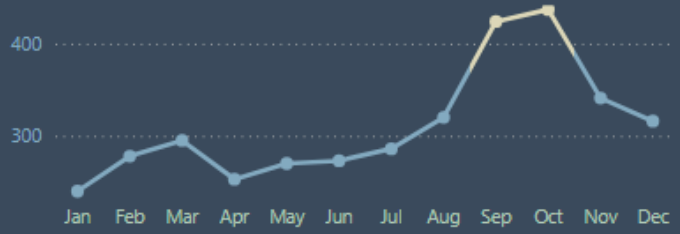

## Movies Count (Over Years)

Last 10 years




## Movies Count (Over Months)

Last 10 years



# Hollywood in the Cloud

## An End-to-End Data Pipeline for Box Office Analysis

(22 nov 2025)
-: Presenter :-
Maher Mahmoud Maher

# ? Project Overview & Objectives

- **The Goal :**

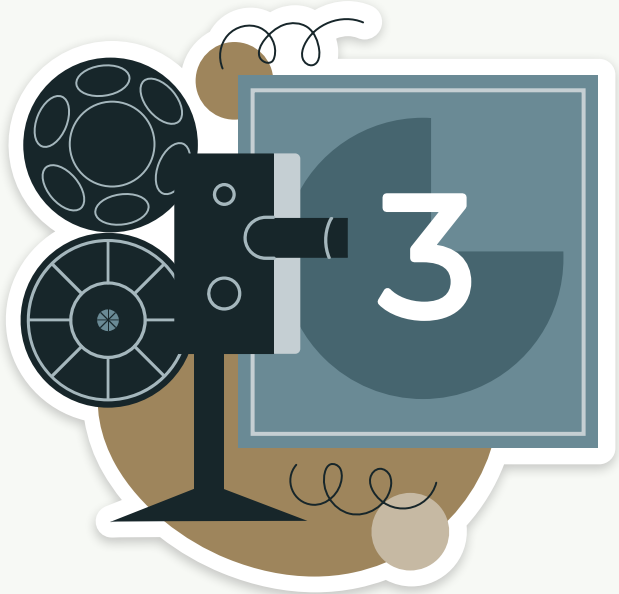
  To build a fully automated, serverless data pipeline to ingest, process, and analyze over 100 years of US Box Office data (1902-2024).

- **Key Questions :**

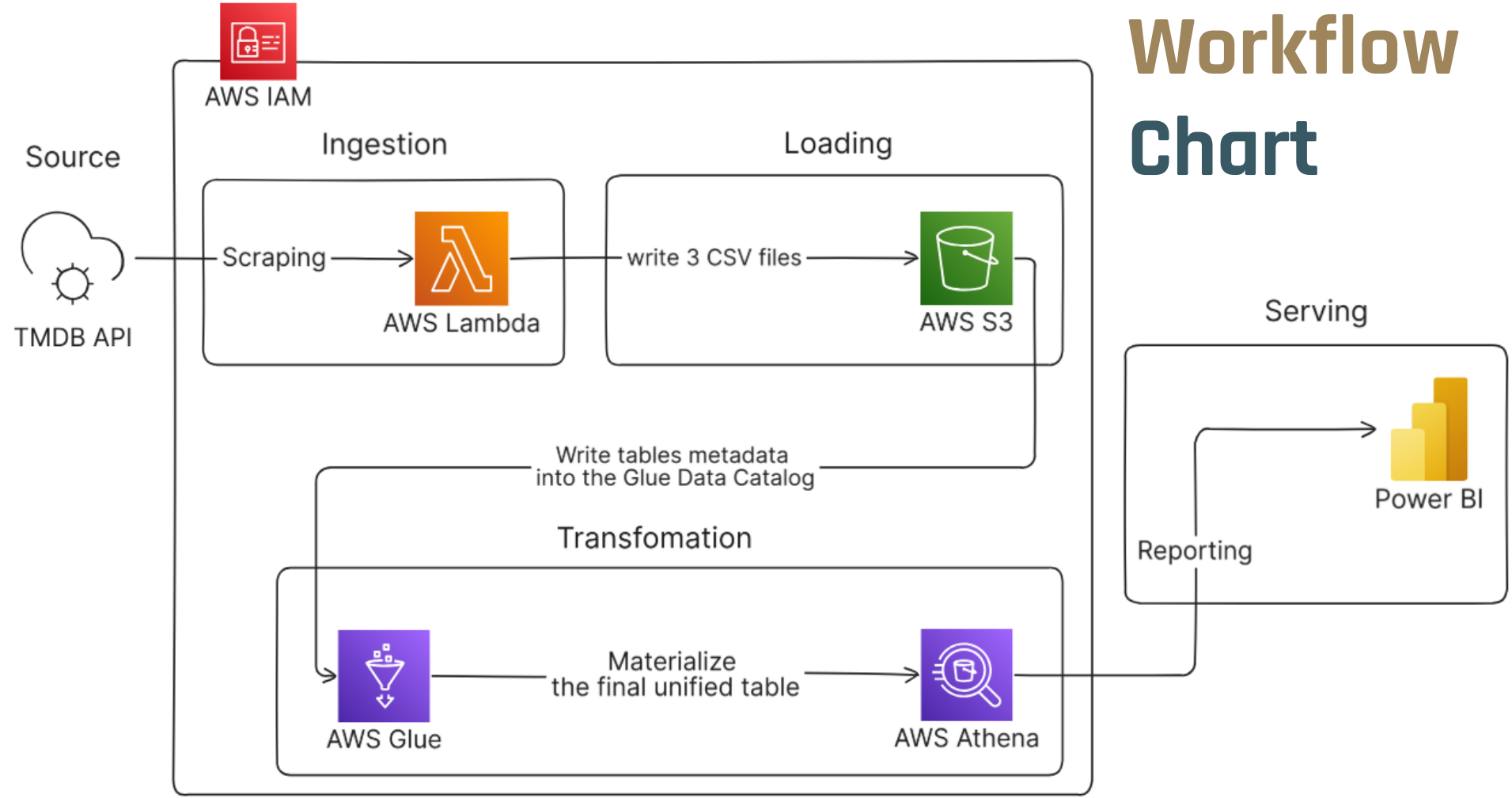
  What are the long-term trends in movie production?
  Which genres are most common vs. most popular?
  How do English and Non-English films compare?
  What are the most popular foreign languages in film?

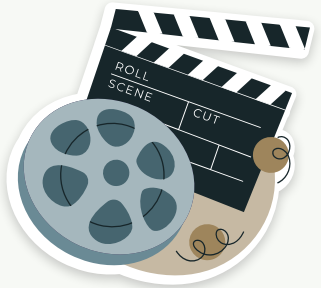
- **Target Audience & Stakeholders :**

  1- Production Studios – green-lighting new projects
  2- Film Distributors – picking peak release windows



# Workflow Chart

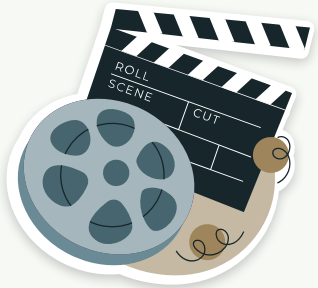





# Proposed Tech Stack

**An End-to-End Cloud Data Pipeline**

| | | |
|---|---|---|
| **Security** | AWS **IAM** | Creating the roles with the needed permissions |
| **Ingest** | AWS **Lambda** | Fetches data from API. |
| **Load** | AWS **S3** | Raw CSVs stored in S3 Data Lake. |
| **Transform** | AWS **Glue** | Crawls metadata and cleans data. |
| **Serve** | AWS **Athena** | Queries the optimized table. |
| **Visualize** | MS **Power BI** | Consumes data via Athena. |

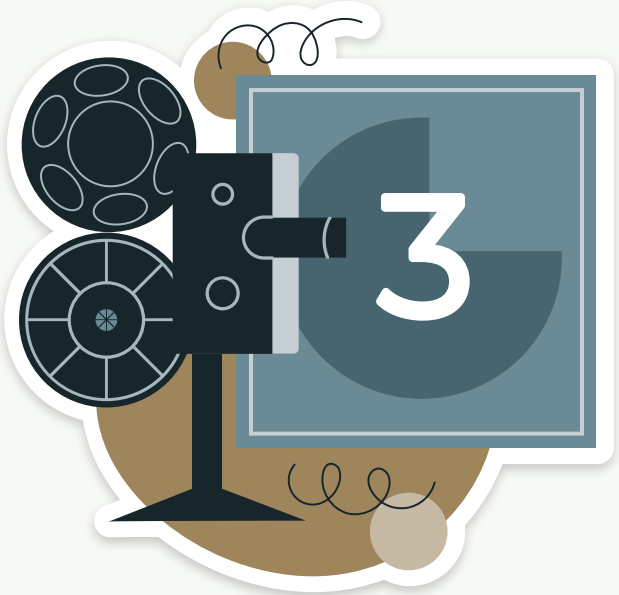



# Some of faced Challenges

| | |
|---|---|
| AWS **IAM** | Selecting the needed permissions for each role |
| AWS **Lambda** | Timed out the code from 3sec to 15min<br>Adding the Environment Variables |
| AWS **S3** | Splitting each lambda csv output into separated locations for the crawlers sake (mandatory) |
| AWS **Glue** | Configuring the glue crawlers (one per each csv) for parsing the schema |
| AWS **Athena** | Creating the unified Table that will be stored in s3 |
| MS **Power BI** | Using the ODBC service to connect with Athena |

# Data Storytelling

## (The Dashboard)



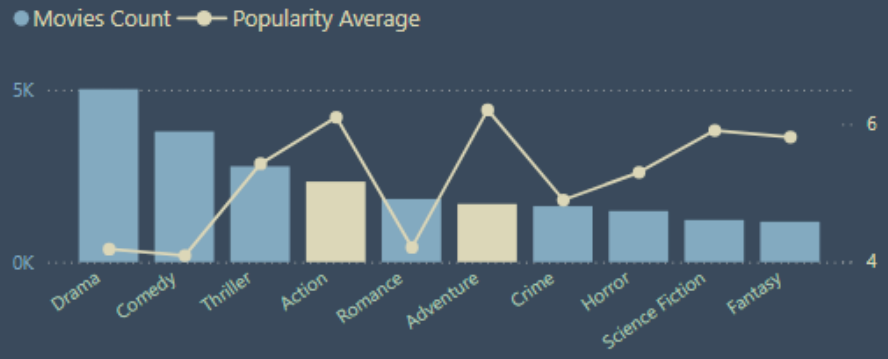
# US Box office (1902:2024) Votes & Popularity Analysis

**English Movies**
Vote Average
6.5

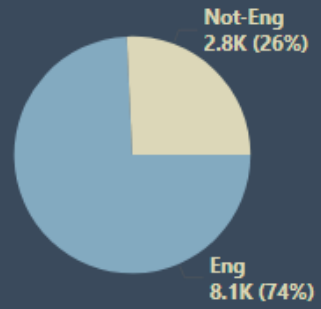
**Non-English Movies**
Vote Average
6.1

YEAR
1902 — 2024

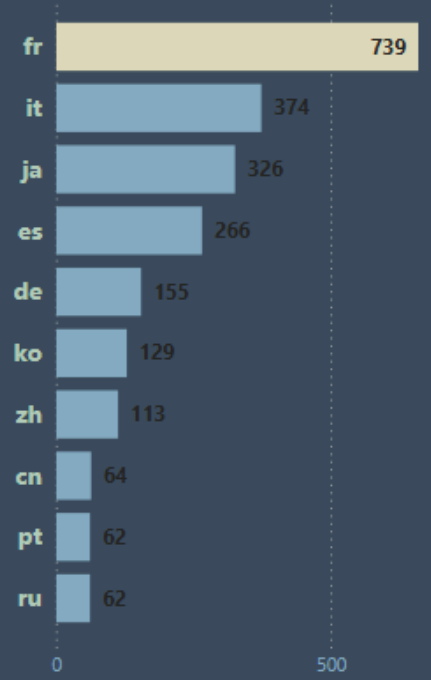
## Top 10 Genres by movies count vs its popularity Average



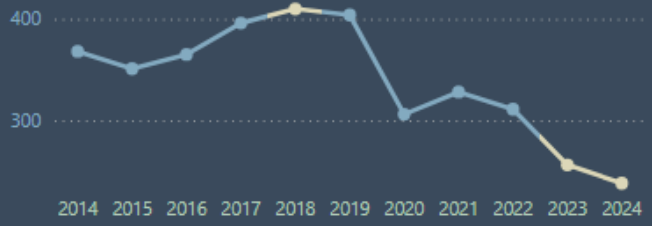

## Eng Movies VS Non-English
(By movies count)



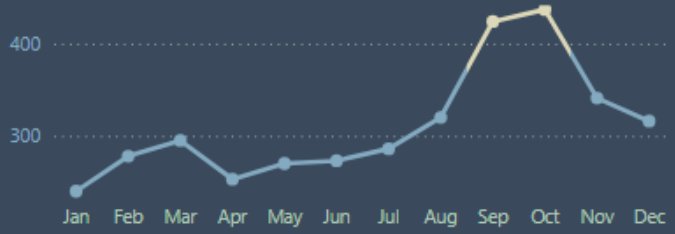

## Top 10 Foreign Languages
(By movies count)




## Movies Count (Over Years)
Last 10 years




## Movies Count (Over Months)
Last 10 years



# Some of the Key insights

- **Quality vs. Quantity :**

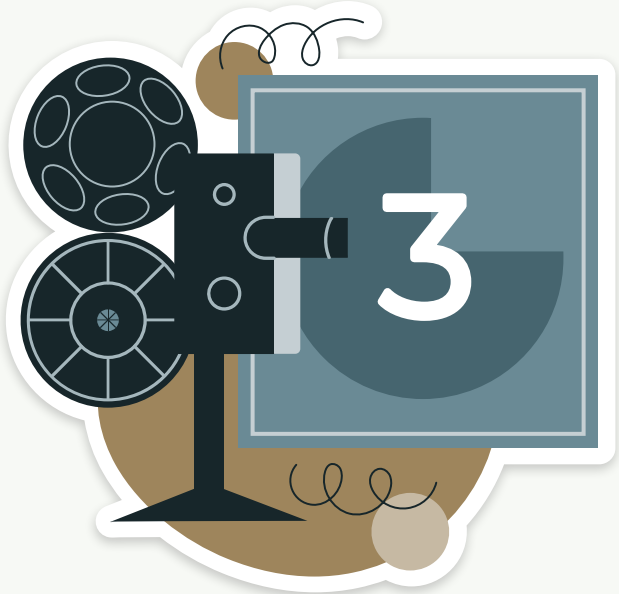
  Drama and Comedy have the highest volume, but Adventure and Action hold the highest popularity ratings.

- **Seasonality :**

  The industry clearly targets Q4 (Oct/Nov) for peak releases, likely aiming for holiday audiences and award seasons.

- **Language Trends :**

  French is the dominant foreign language in the US market (739 films), nearly double that of Italian or Japanese.



# THANK YOU

## Any questions?