BIBDA, DMST, Aueb

Athens, 2024

# MOVIE DATASET ANALYSIS

Members:

Pula Sergi, 8200141

Malandrakis Georgios, 8200094

# DATASET & ORIGIN

In this section we will talk about :

- The origin of our main dataset
- How it changed based on other datasets
- The final structure it got

# OUR FIRST DATASET

Our first dataset was exported by kaggle
**‘TMDB_movie_dataset_v11.csv’**

977.779 rows - 19 columns

ASANICZKA · UPDATED 24 MINUTES AGO

148 | New Notebook | Download (190 MB)

## Full TMDB Movies Dataset 2024 (985K Movies)

Complete dataset containing movie data from TMDb. Updated Daily

Data Card | Code (7) | Discussion (3) | Suggestions (0)

### About Dataset

The TMDb (The Movie Database) is a comprehensive movie database that provides information about movies, including details like titles, ratings, release dates, revenue, genres, and much more.

**Usability**
10.00

**License**
ODC Attribution License (ODC-...

# MOST INTERESTING COLUMNS

Production Companies

Production Countries

Spoken Languages

Genres

imdb_id

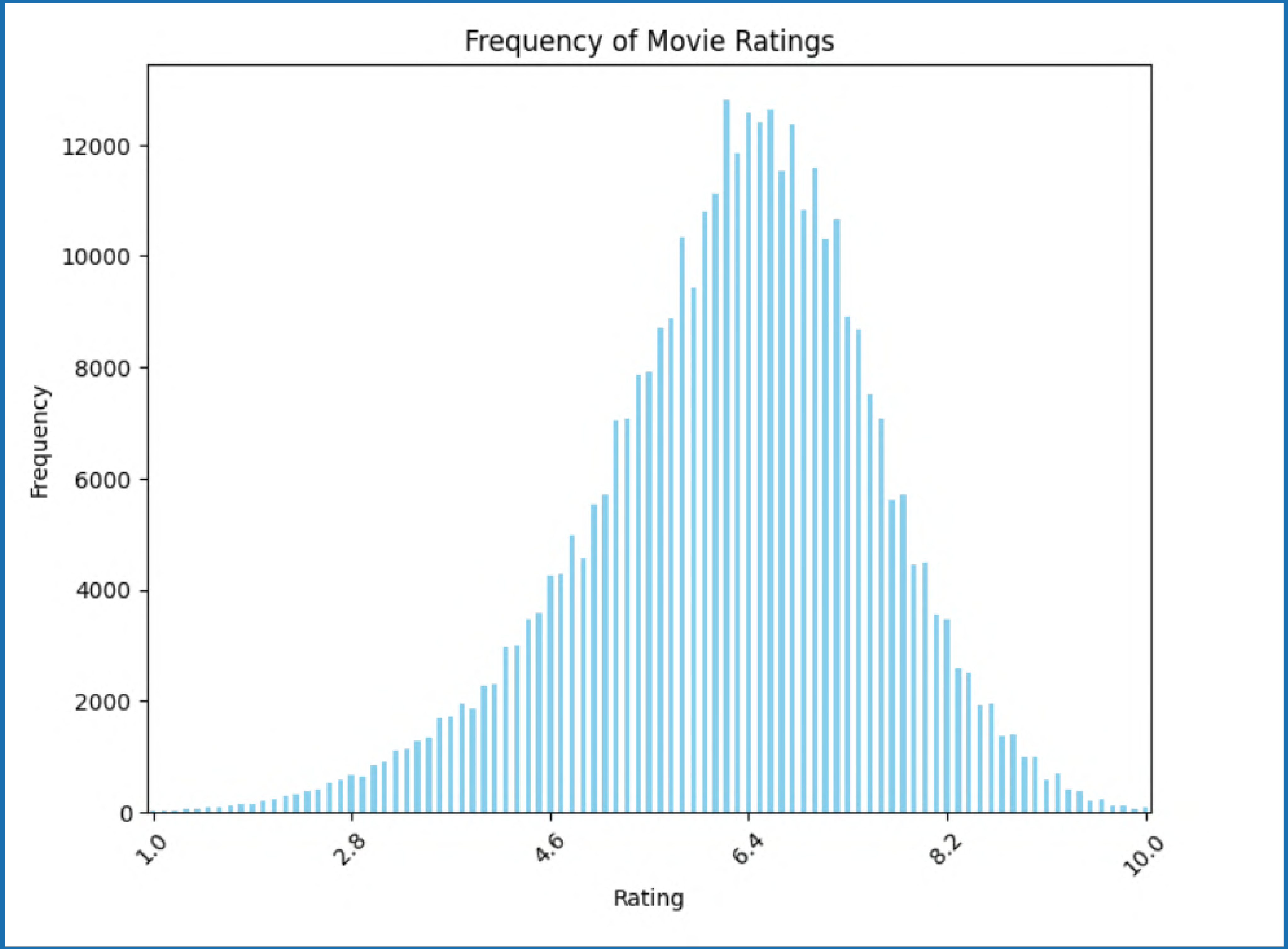
Revenue

Runtime

Budget

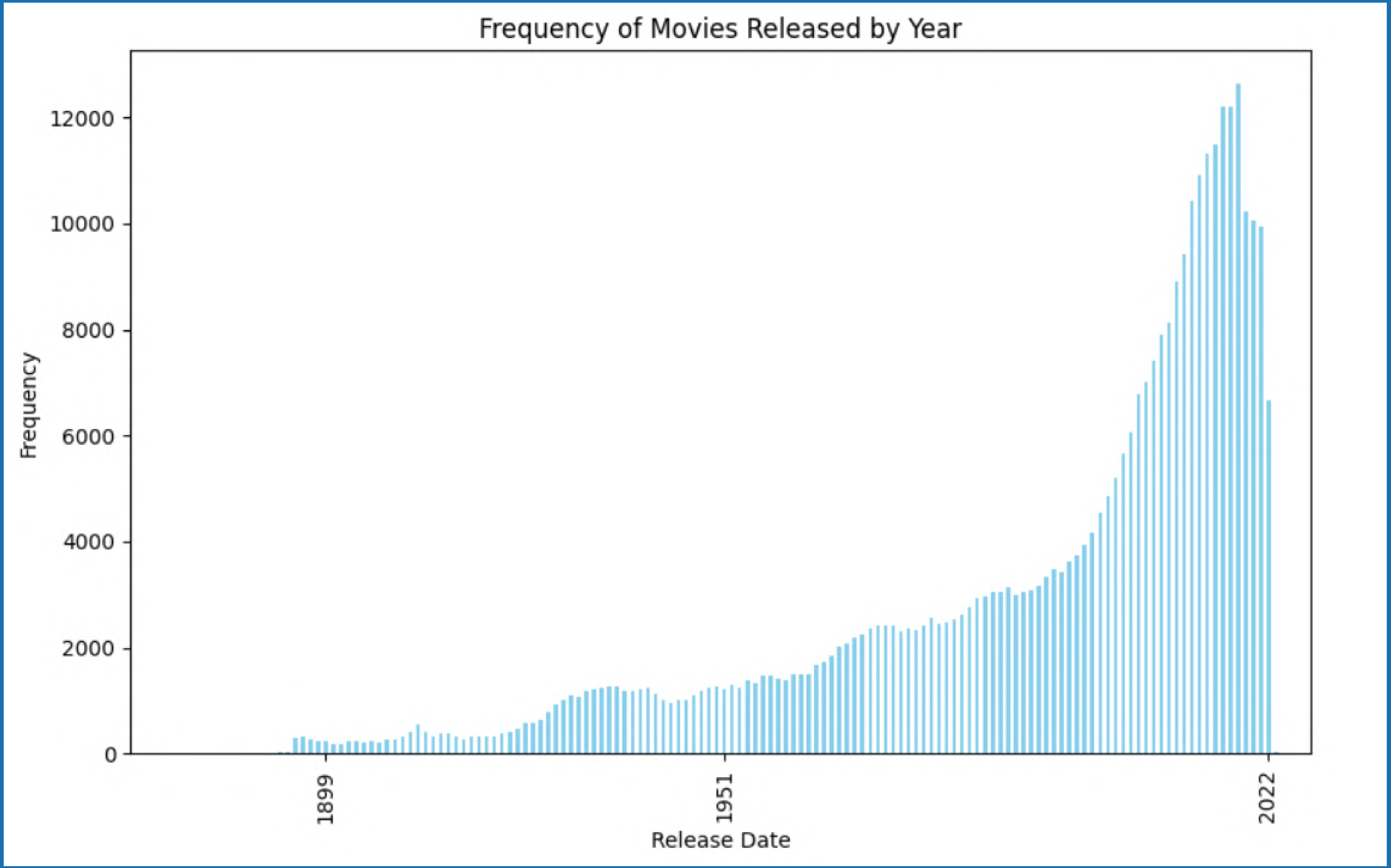
# MOST INTERESTING COLUMNS

Average Rating



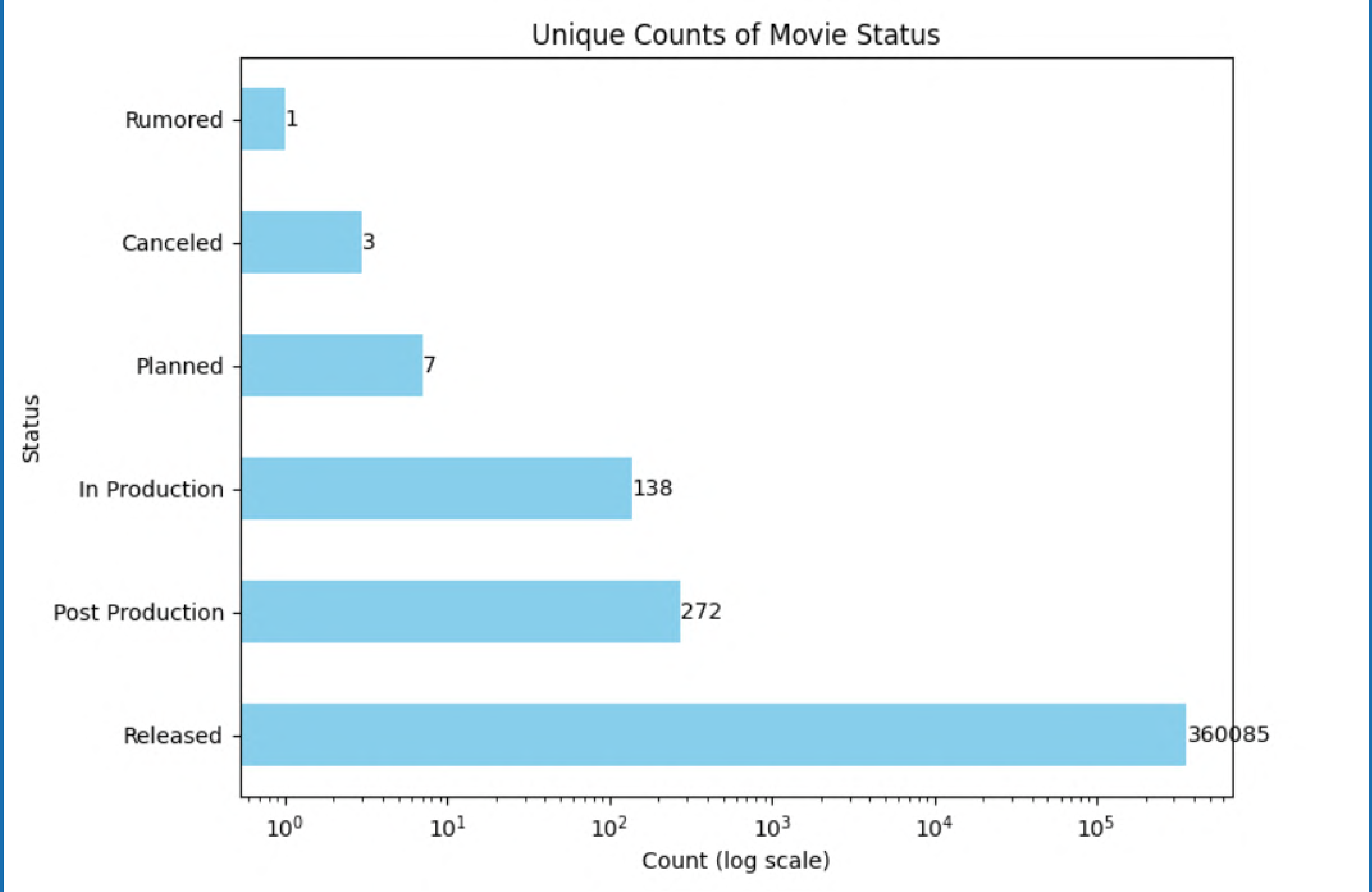
# MOST INTERESTING COLUMNS

Release Date



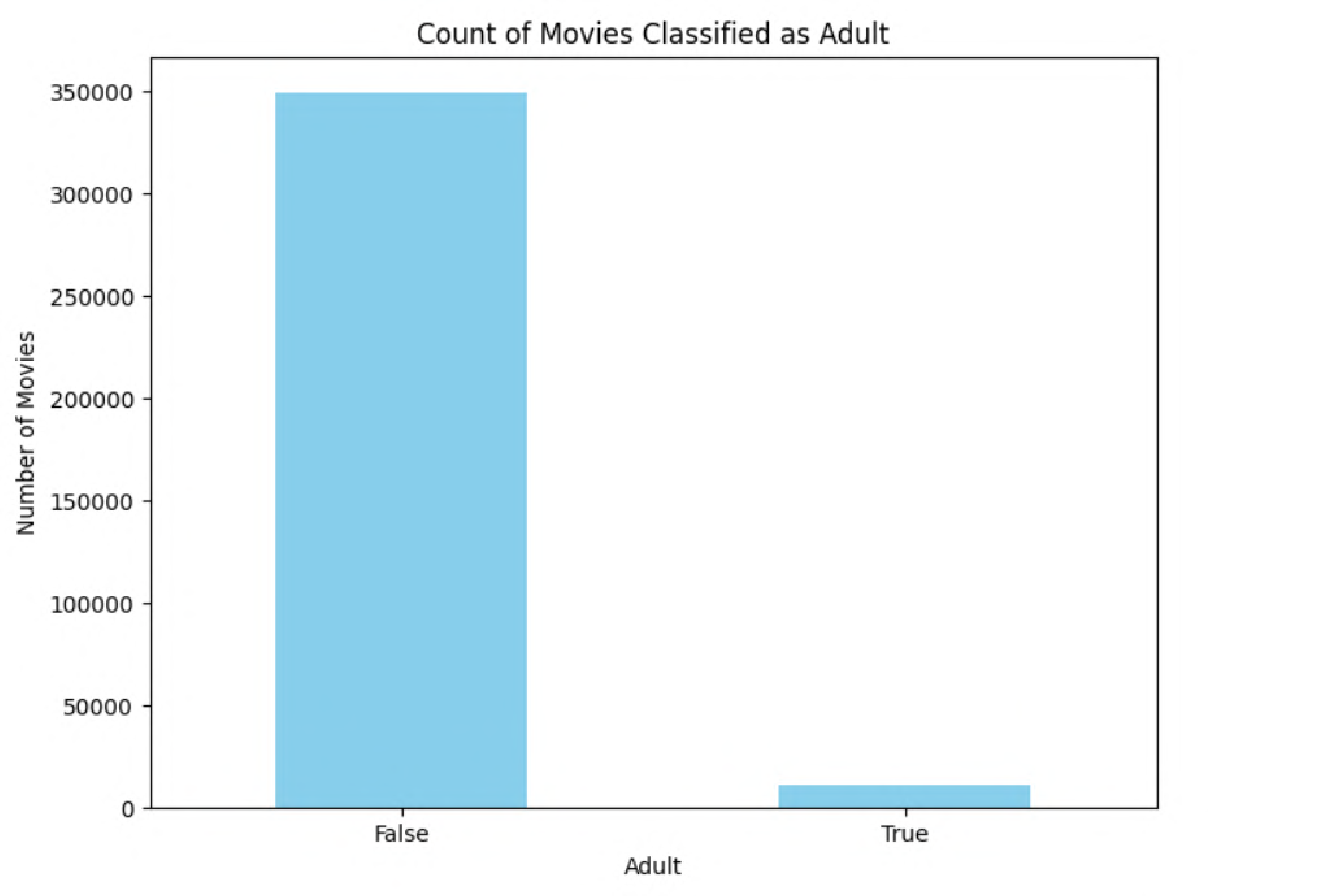
# MOST INTERESTING COLUMNS

Status

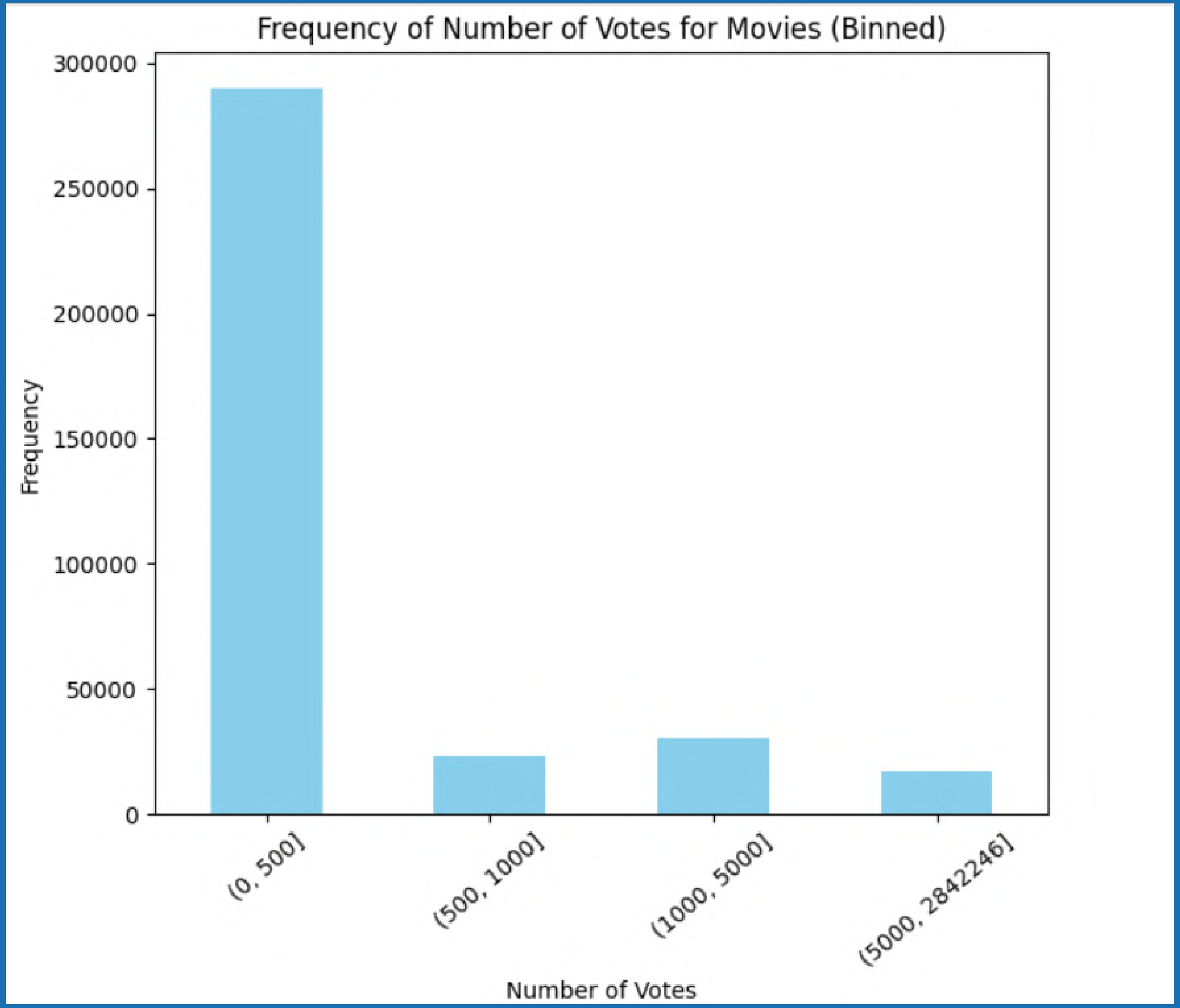


# MOST INTERESTING COLUMNS

## Adult




## Number of Votes



# Merging with our 2nd dataset from IMDB

IMDb typically has a larger user base, resulting in a greater number of votes and a more comprehensive rating system.

Using datasets from IMDb Developer we acquire our second dataset:

| | tconst | averageRating | numVotes |
|---|---|---|---|
| **0** | tt0000001 | 5.7 | 2014 |
| **1** | tt0000002 | 5.7 | 270 |
| **2** | tt0000003 | 6.5 | 1937 |
| **3** | tt0000004 | 5.5 | 178 |
| **4** | tt0000005 | 6.2 | 2712 |
| **...** | ... | ... | ... |
| **1391426** | tt9916730 | 7.6 | 11 |
| **1391427** | tt9916766 | 7.1 | 23 |
| **1391428** | tt9916778 | 7.2 | 36 |
| **1391429** | tt9916840 | 8.8 | 6 |
| **1391430** | tt9916880 | 8.2 | 6 |

1391431 rows × 3 columns

# Merging with our 2nd dataset from IMDB

By merging them on the IMDb ID, we narrow down our movies to only those associated with IMDb, while adding two important columns.

409.387 rows

# Merging with 6 more datasets!

Before starting our analysis, we decided to incorporate six additional datasets to include a column named '**Platform**', which will indicate whether a movie belongs to one of the following platforms:

# Merging with 6 more datasets!

We merged our datasets with our main dataset based on the IMDb ID:

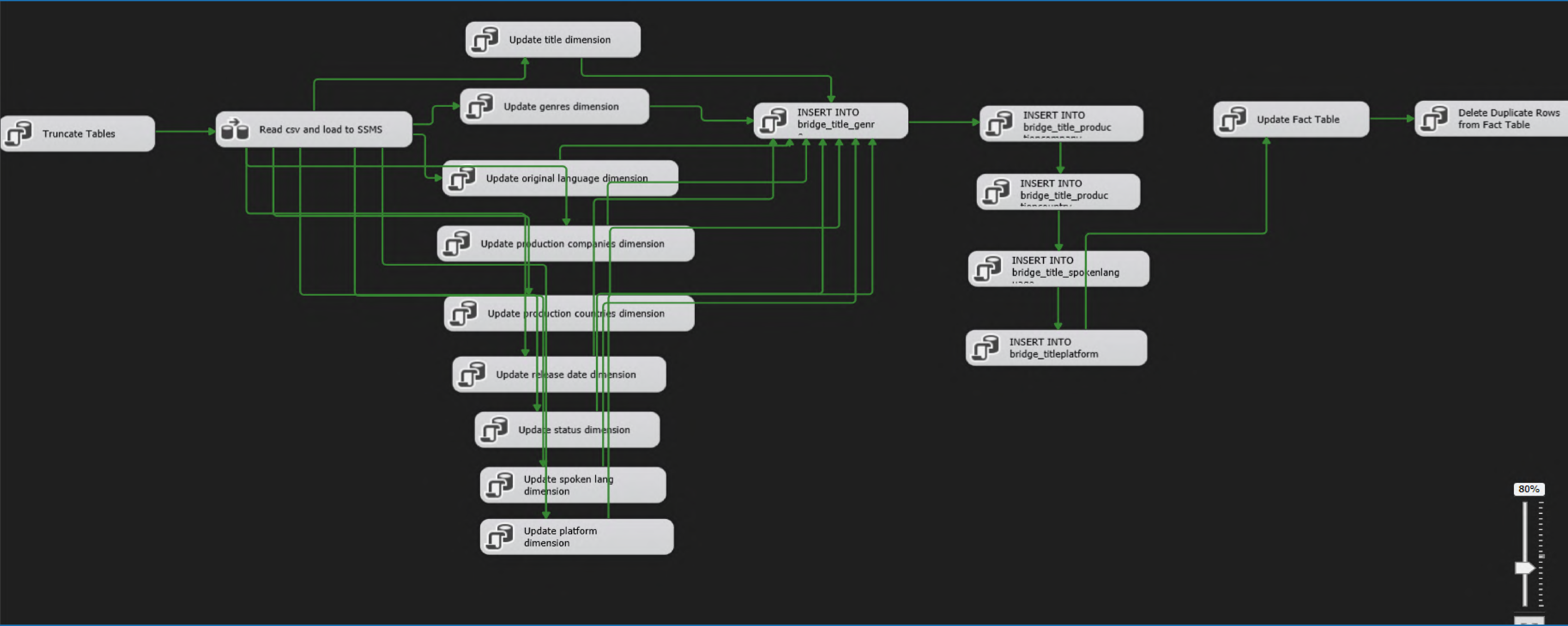
| title | Netflix | Amazon | Disney | Apple | Paramount | HBO |
|---|---|---|---|---|---|---|
| Inception | False | False | False | False | False | False |
| Interstellar | False | False | False | False | True | True |
| The Dark Knight | True | False | False | False | False | False |
| Avatar | False | False | True | False | False | False |

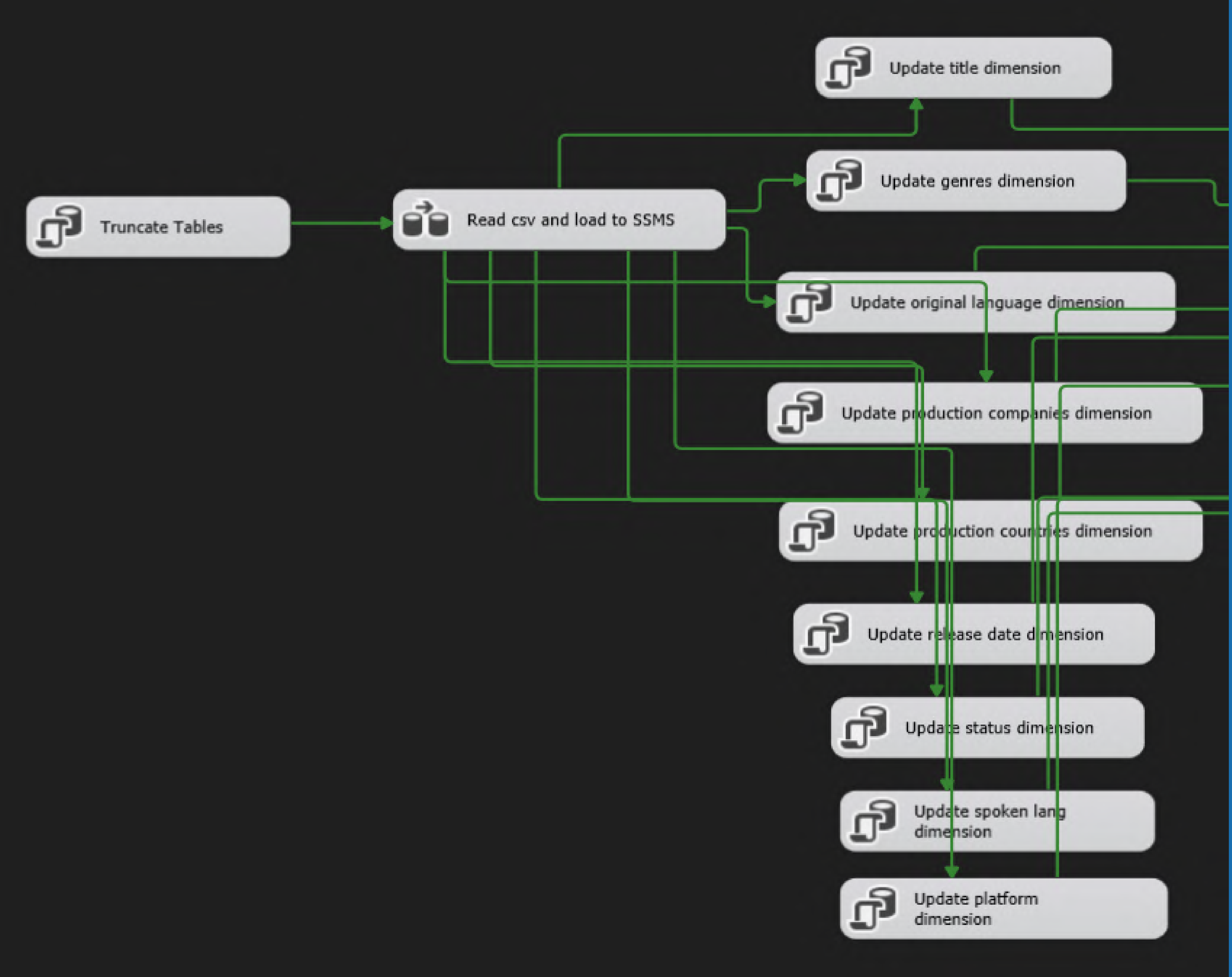
# ETL PROCESS

In this section we will talk about :

- The ETL process we followed
- SQL code depicting the process

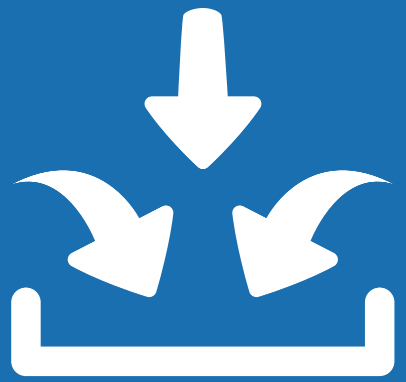
# Overal Process

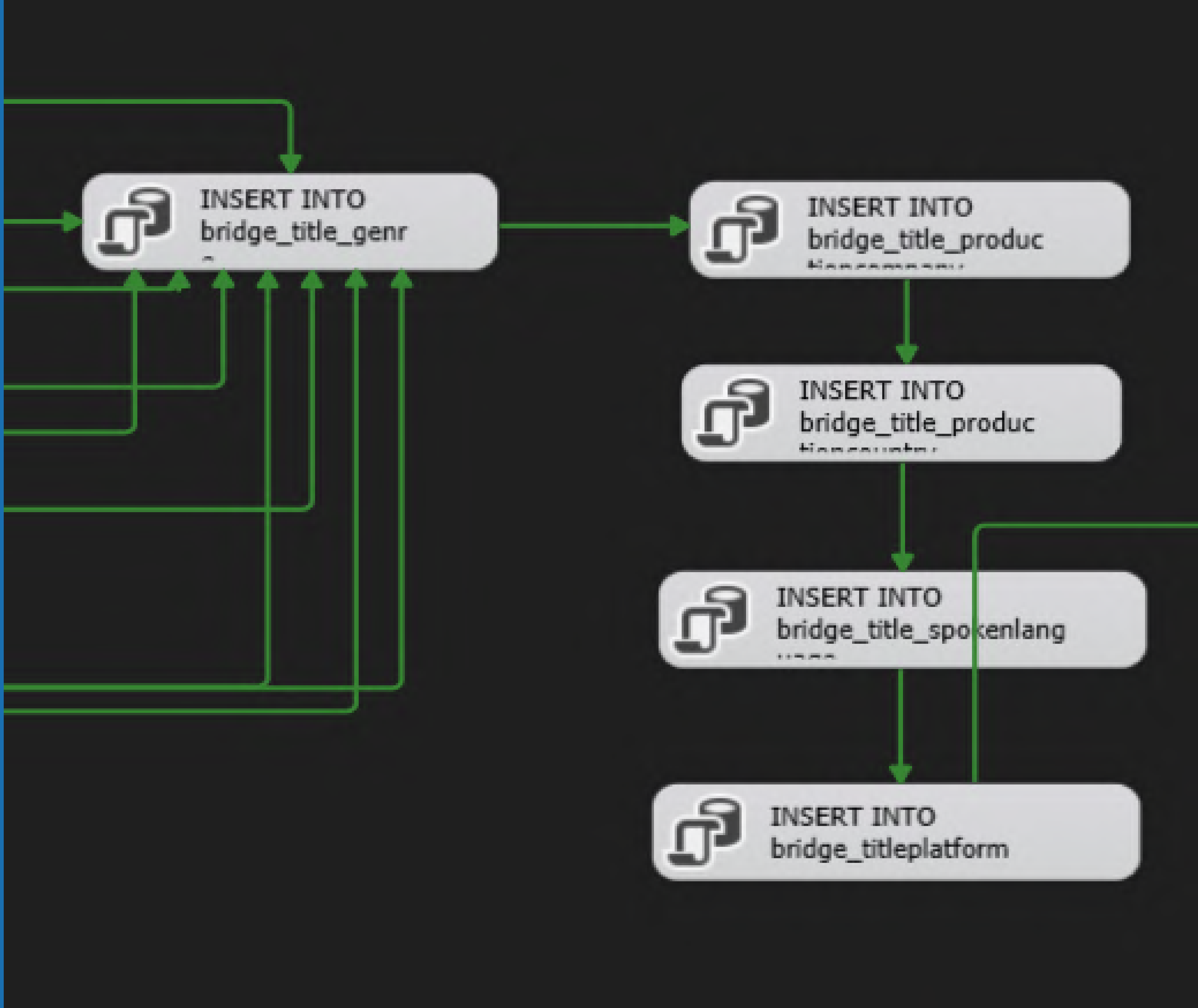




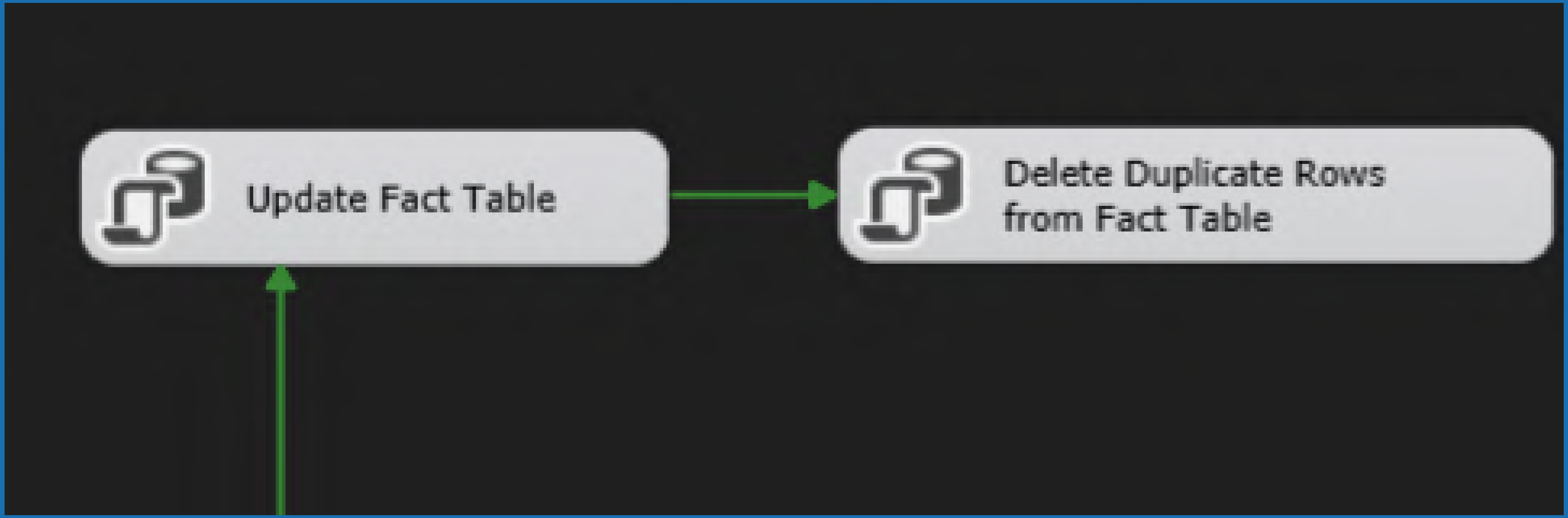

- Deleting all previous entries from tables.
- Read CSV files and load into SSMS.
- Update all dimension tables.

- Insert all unique combinations into bridge tables (also known as factless fact tables).

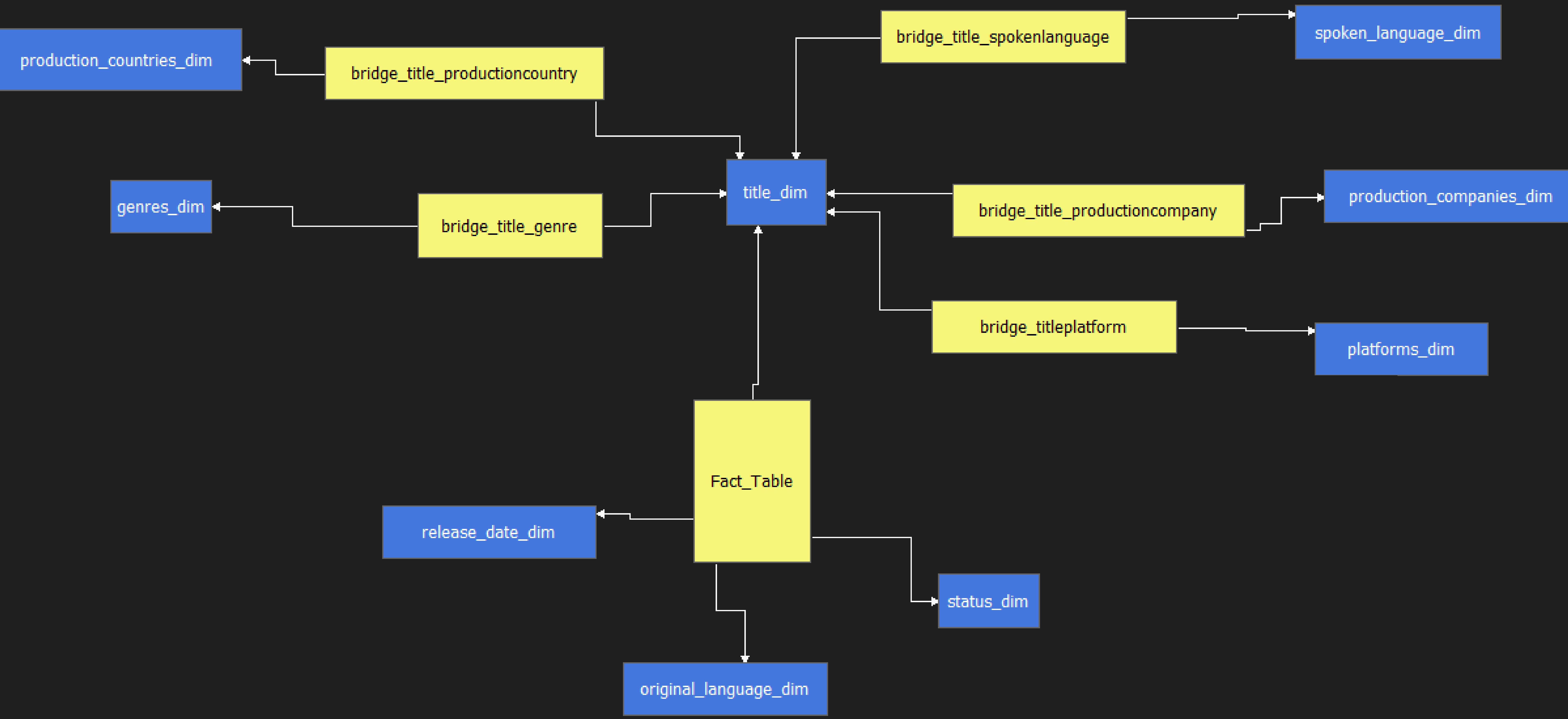


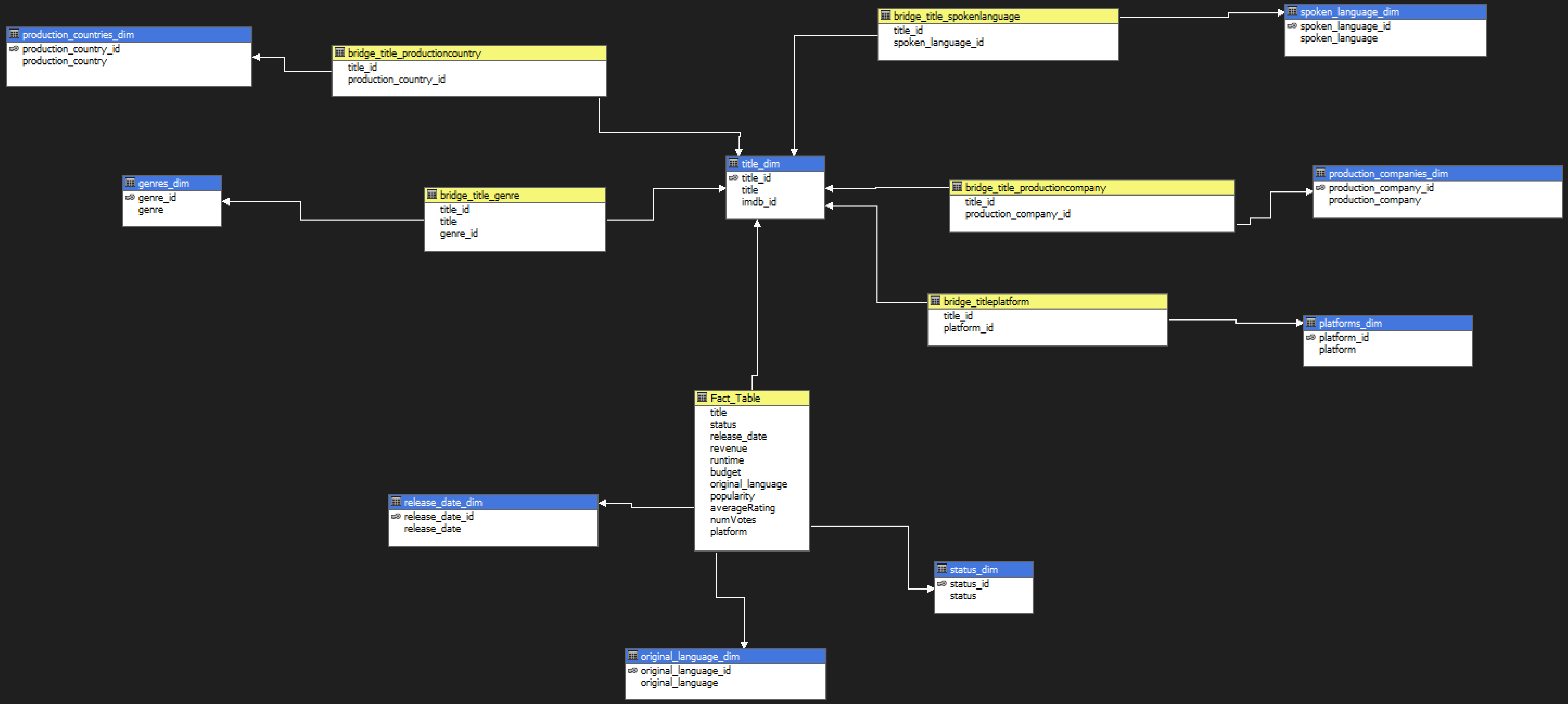


- Update the fact table by populating it with data.
- Delete duplicate rows from the fact table.

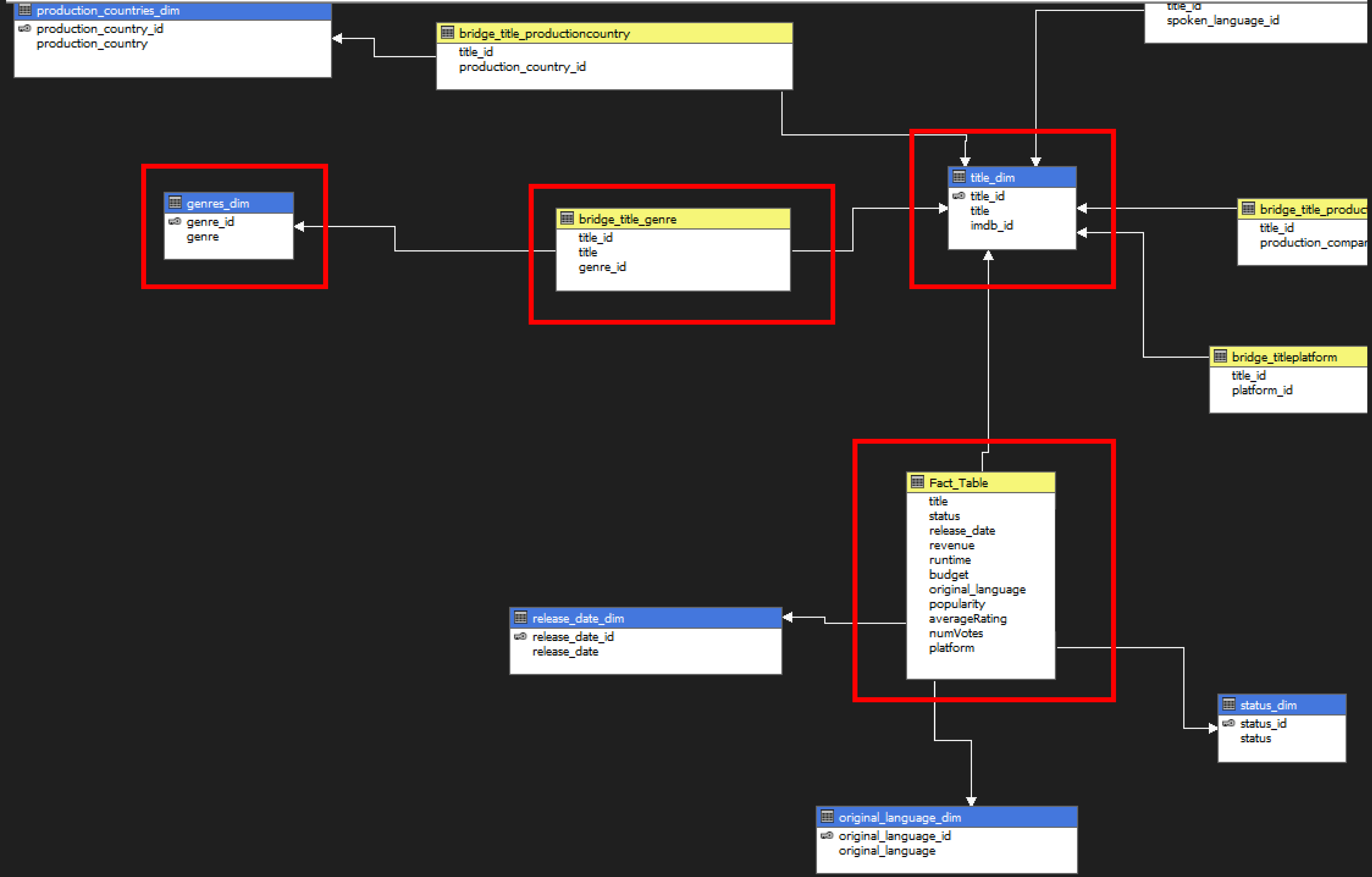


# STAR SCHEMA / CUBE


production_countries_dim
bridge_title_productioncountry
bridge_title_spokenlanguage
spoken_language_dim
genres_dim
bridge_title_genre
title_dim
bridge_title_productioncompany
production_companies_dim
bridge_titleplatform
platforms_dim
Fact_Table
release_date_dim
status_dim
original_language_dim



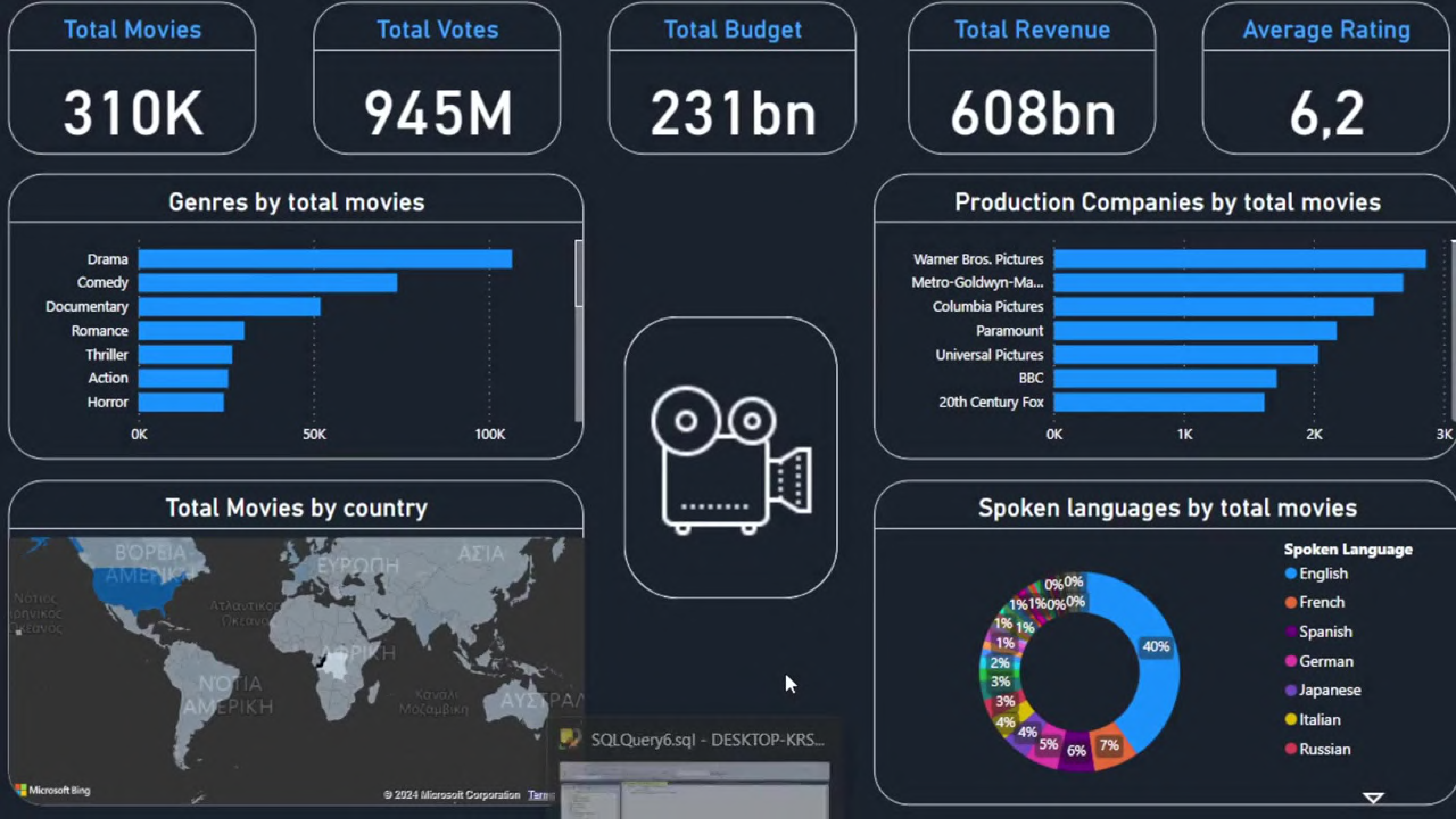
production_countries_dim
production_country_id
production_country
bridge_title_productioncountry
title_id
production_country_id
bridge_title_spokenlanguage
title_id
spoken_language_id
spoken_language_dim
spoken_language_id
spoken_language
genres_dim
genre_id
genre
bridge_title_genre
title_id
title
genre_id
title_dim
title_id
title
imdb_id
bridge_title_productioncompany
title_id
production_company_id
production_companies_dim
production_company_id
production_company
bridge_titleplatform
title_id
platform_id
platforms_dim
platform_id
platform
Fact_Table
title
status
release_date
revenue
runtime
budget
original_language
popularity
averageRating
numVotes
platform
release_date_dim
release_date_id
release_date
status_dim
status_id
status
original_language_dim
original_language_id
original_language


production_countries_dim
production_country_id
production_country
bridge_title_productioncountry
title_id
production_country_id
spoken_language_id
genres_dim
genre_id
genre
bridge_title_genre
title_id
title
genre_id
title_dim
title_id
title
imdb_id
bridge_title_produc
title_id
production_compa
bridge_titleplatform
title_id
platform_id
Fact_Table
title
status
release_date
revenue
runtime
budget
original_language
popularity
averageRating
numVotes
platform
release_date_dim
release_date_id
release_date
status_dim
status_id
status
original_language_dim
original_language_id
original_language

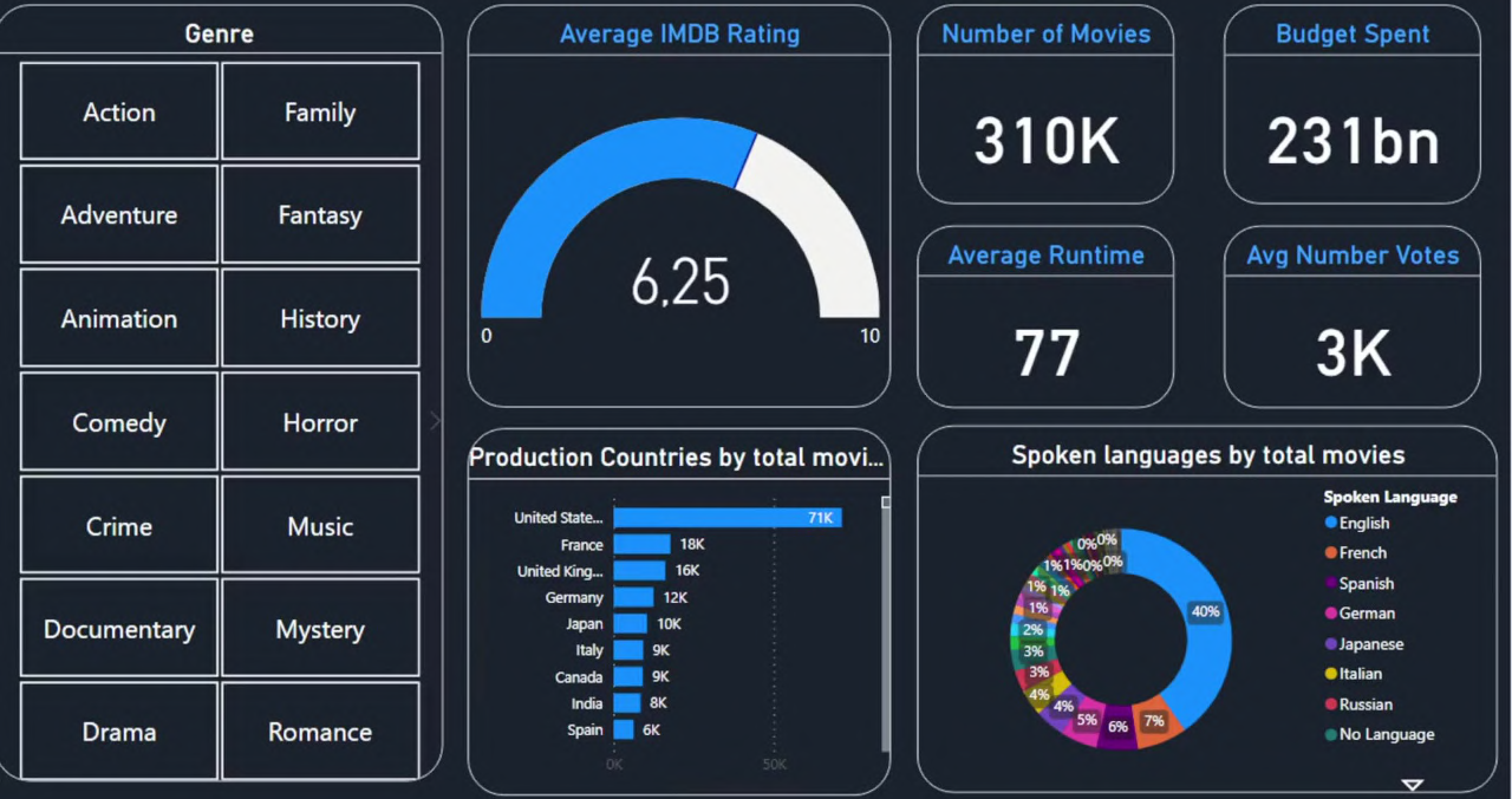
# VISUALIZATIONS

**In this section we will show:**

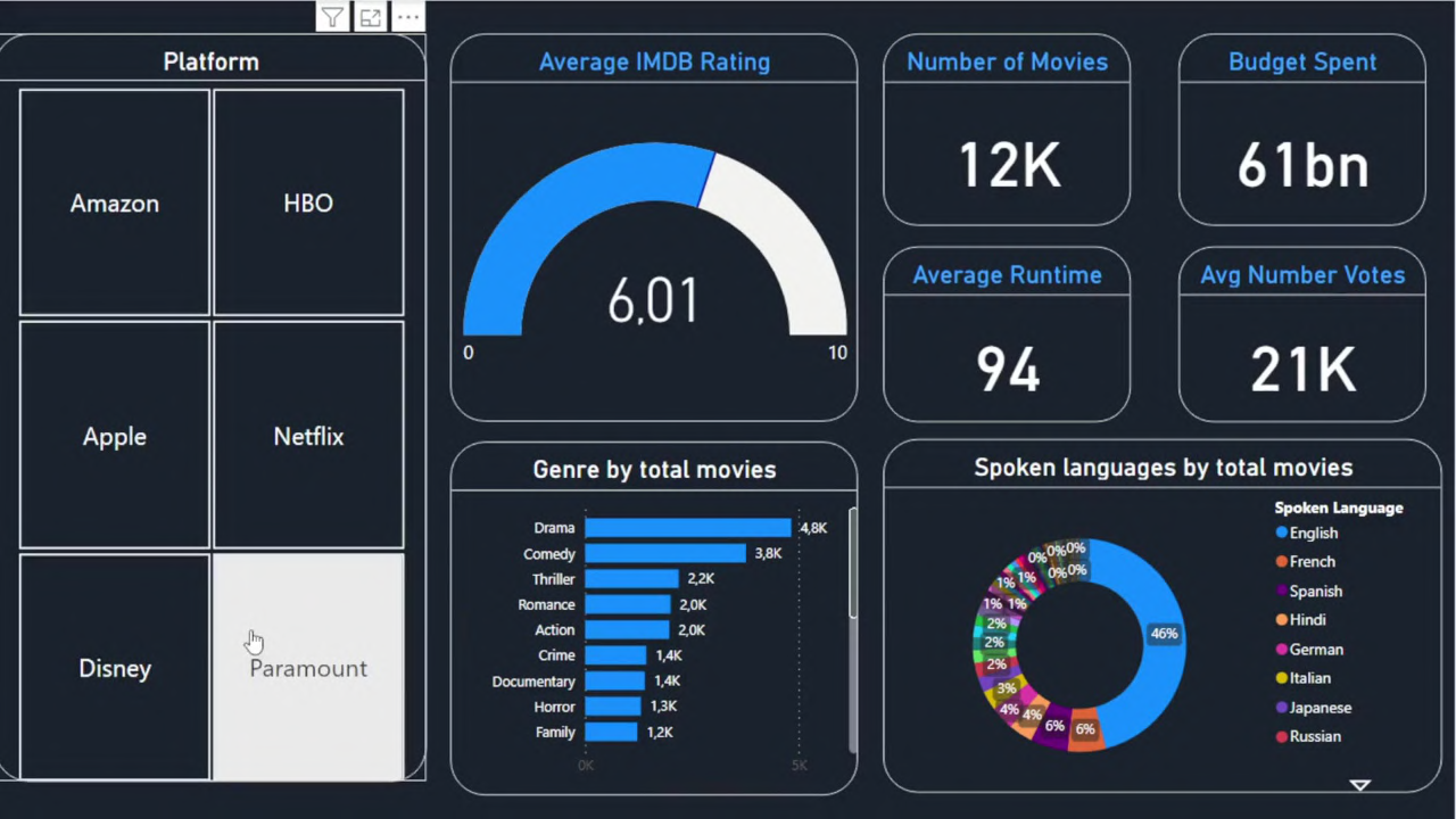
- **All Power bi visualizations**



Total Movies
310K
Total Votes
945M
Total Budget
231bn
Total Revenue
608bn
Average Rating
6,2
Genres by total movies
Drama
Comedy
Documentary
Romance
Thriller
Action
Horror
0K
50K
100K
Production Companies by total movies
Warner Bros. Pictures
Metro-Goldwyn-Ma...
Columbia Pictures
Paramount
Universal Pictures
BBC
20th Century Fox
0K
1K
2K
3K
Total Movies by country
ΒΟΡΕΙΑ ΑΜΕΡΙΚΗ
ΕΥΡΩΠΗ
ΑΣΙΑ
ΝΟΤΙΑ ΑΜΕΡΙΚΗ
Microsoft Bing
© 2024 Microsoft Corporation
Spoken languages by total movies
Spoken Language
English
French
Spanish
German
Japanese
Italian
Russian
40%
7%
6%
5%
4%
4%
3%
3%
2%
1%
1%
1%
1%
1%
0%
0%
0%
0%
SQLQuery6.sql - DESKTOP-KRS...



Genre
Action
Family
Adventure
Fantasy
Animation
History
Comedy
Horror
Crime
Music
Documentary
Mystery
Drama
Romance
Average IMDB Rating
6,25
0
10
Number of Movies
310K
Budget Spent
231bn
Average Runtime
77
Avg Number Votes
3K
Production Countries by total movi...
United State... 71K
France 18K
United King... 16K
Germany 12K
Japan 10K
Italy 9K
Canada 9K
India 8K
Spain 6K
0K
50K
Spoken languages by total movies
40%
7%
6%
5%
4%
4%
3%
3%
2%
1%
1%
1%
1%
1%
0%
0%
0%
0%
Spoken Language
English
French
Spanish
German
Japanese
Italian
Russian
No Language


Platform
Amazon
HBO
Apple
Netflix
Disney
Paramount
Average IMDB Rating
6,01
0
10
Number of Movies
12K
Budget Spent
61bn
Average Runtime
94
Avg Number Votes
21K
Genre by total movies
Drama 4,8K
Comedy 3,8K
Thriller 2,2K
Romance 2,0K
Action 2,0K
Crime 1,4K
Documentary 1,4K
Horror 1,3K
Family 1,2K
0K
5K
Spoken languages by total movies
Spoken Language
English
French
Spanish
Hindi
German
Italian
Japanese
Russian
46%
6%
6%
4%
4%
3%
2%
2%
2%
1%
1%
1%
1%
0%
0%
0%
0%
0%

## Platform

| | |
|---|---|
| Amazon | HBO |
| Apple | Netflix |
| Disney | Paramount |

## Genre

| | | |
|---|---|---|
| Action | Comedy | Drama |
| Adventure | Crime | Family |
| Animation | Documentary | Fantasy |

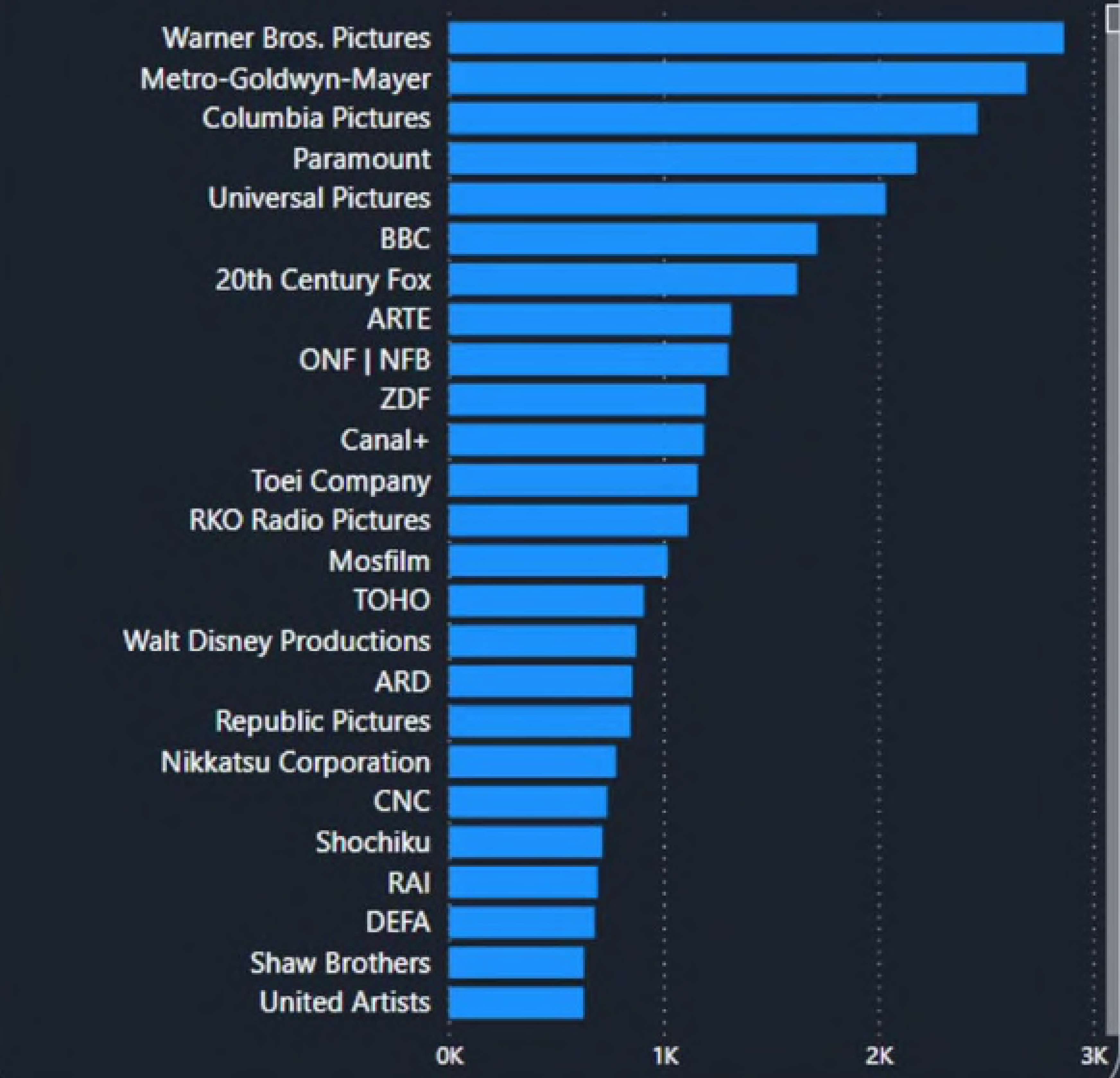
### Top Movies based on IMDB Rating

| Title | Average Rating |
|---|---|
| Avatar: The Last Airbender | 9,30 |
| The Shawshank Redemption | 9,30 |
| The Godfather | 9,20 |
| 12 Angry Men | 9,00 |
| ARCANE | 9,00 |
| Schindler's List | 9,00 |
| The Dark Knight | 9,00 |
| The Godfather Part II | 9,00 |
| The Lord of the Rings: The Return of the King | 9,00 |
| Pulp Fiction | 8,90 |
| The Lord of the Rings: The Fellowship of the Ring | 8,90 |
| Blackadder Goes Forth | 8,80 |
| Bojack Horseman | 8,80 |
| Fight Club | 8,80 |
| Forrest Gump | 8,80 |
| Inception | 8,80 |
| The Lord of the Rings: The Two Towers | 8,80 |
| Attack on Titan The Final Chapters: Special 2 | 8,70 |

## Genre

| | | |
|---|---|---|
| Action | Drama | Mystery |
| Adventure | Family | Romance |
| Animation | Fantasy | Science Fiction |
| Comedy | History | Thriller |
| Crime | Horror | TV Movie |
| Documentary | Music | War |

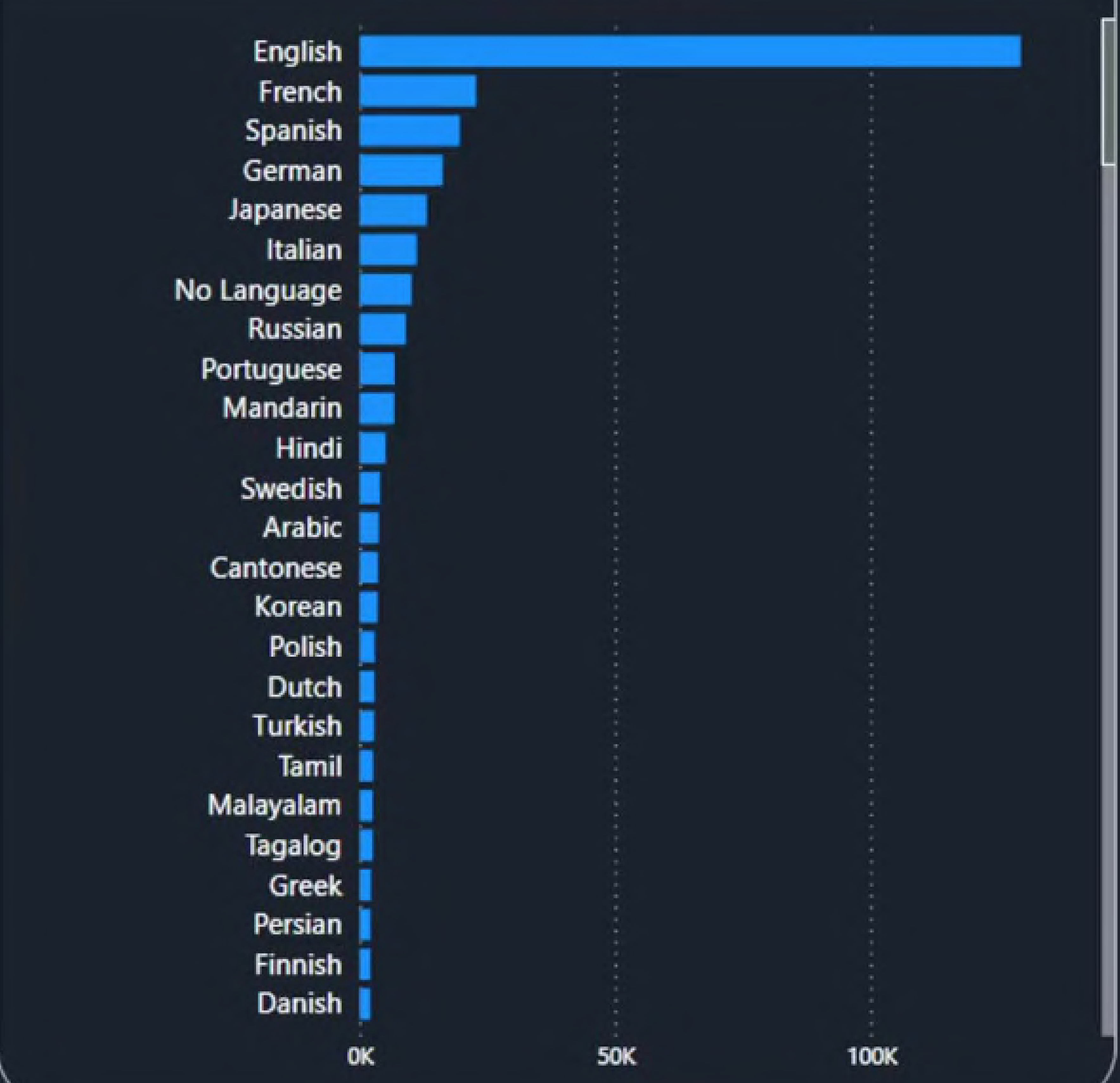
## Production Companies by total movies



## Genre

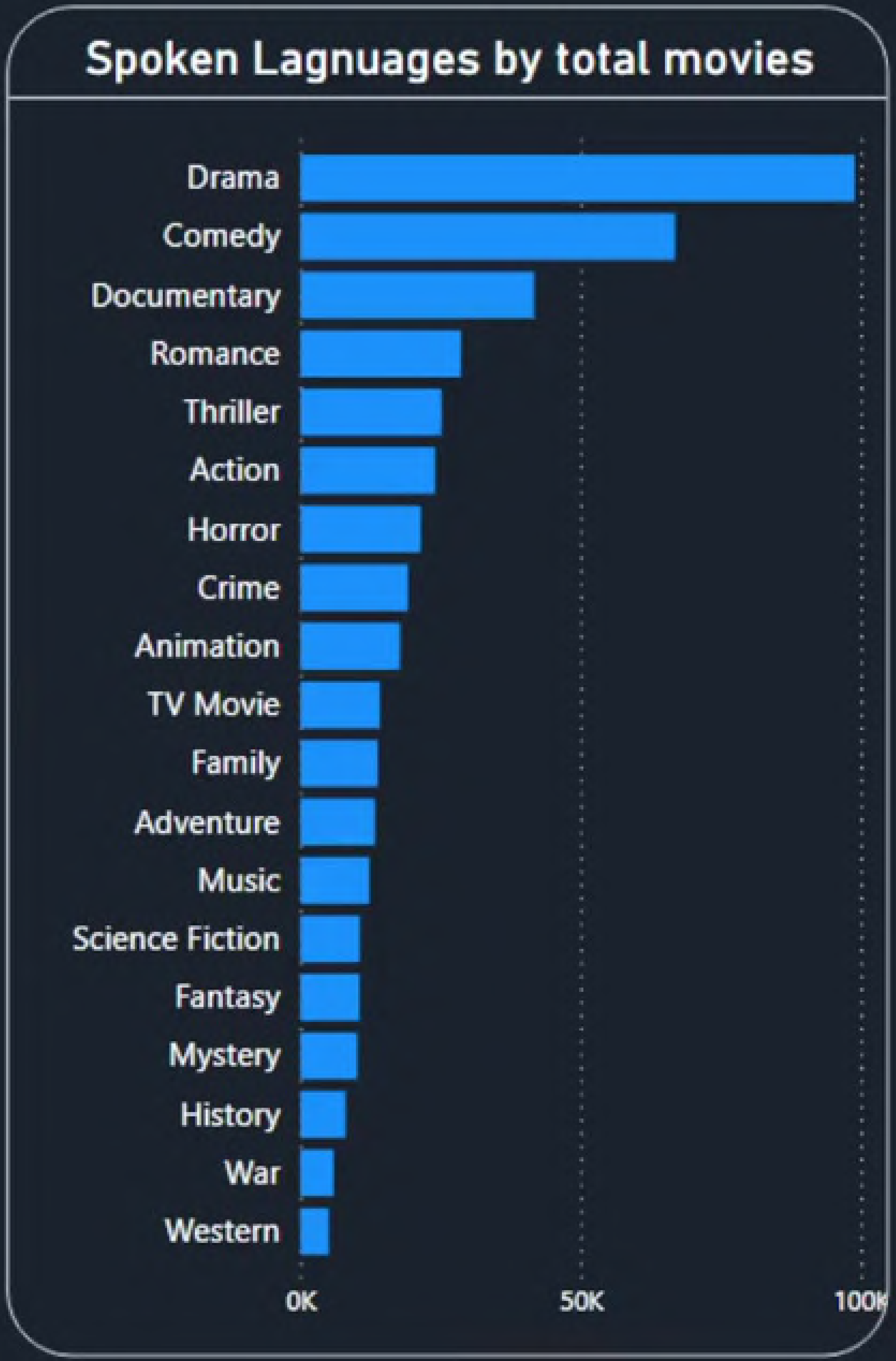
| | | |
|---|---|---|
| Action | Drama | Mystery |
| Adventure | Family | Romance |
| Animation | Fantasy | Science Fiction |
| Comedy | History | Thriller |
| Crime | Horror | TV Movie |
| Documentary | Music | War |

## Spoken Languages

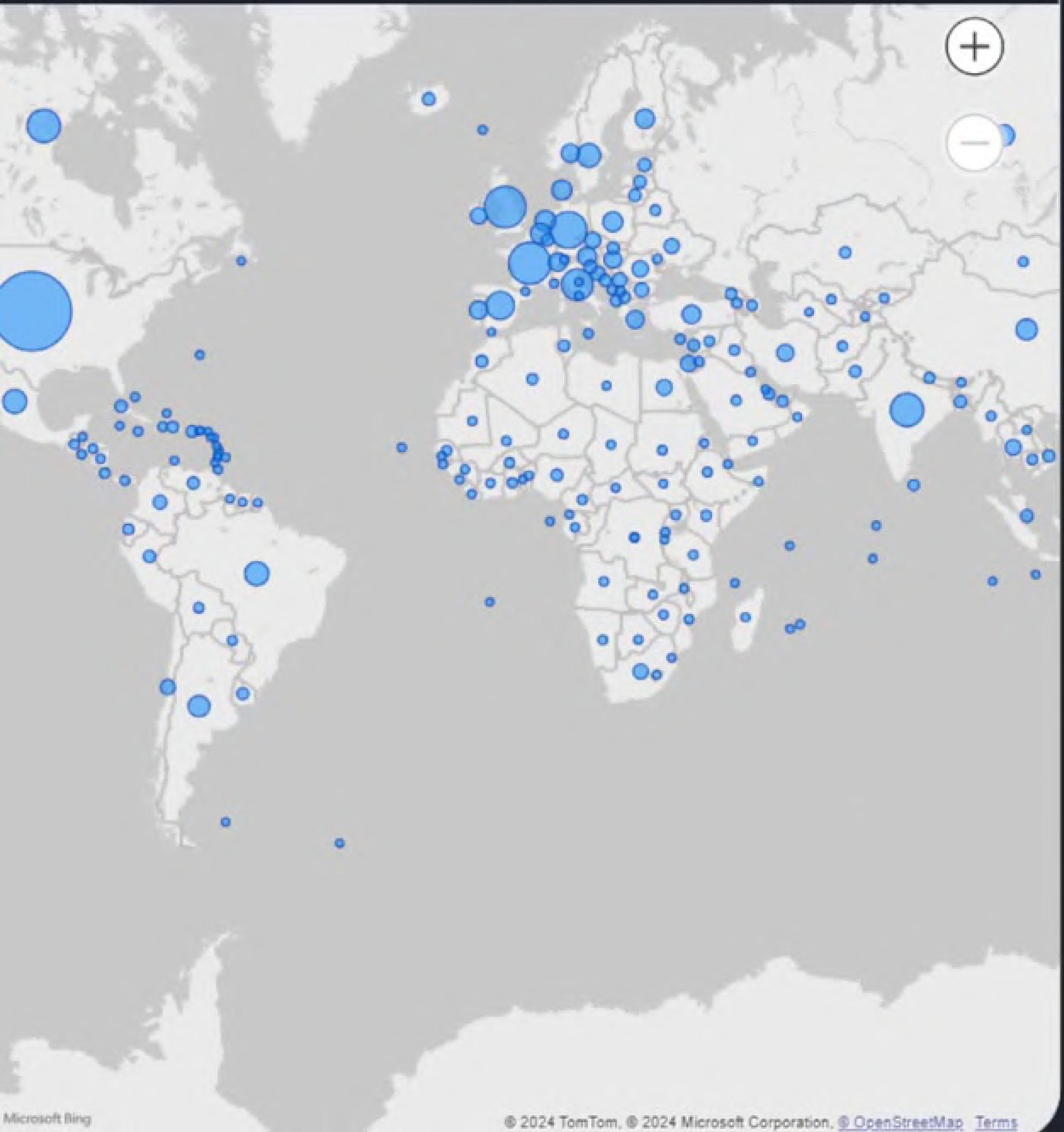



Production Countries by total movies
United States of ...
France
United Kingdom
Germany
Japan
India
Canada
Italy
Spain
Brazil
Mexico
Soviet Union
Sweden
Hong Kong
Australia
Argentina
Russia
China
Netherlands
South Korea
Belgium
Denmark
Poland
Philippines
Finland
0K
50K
100K

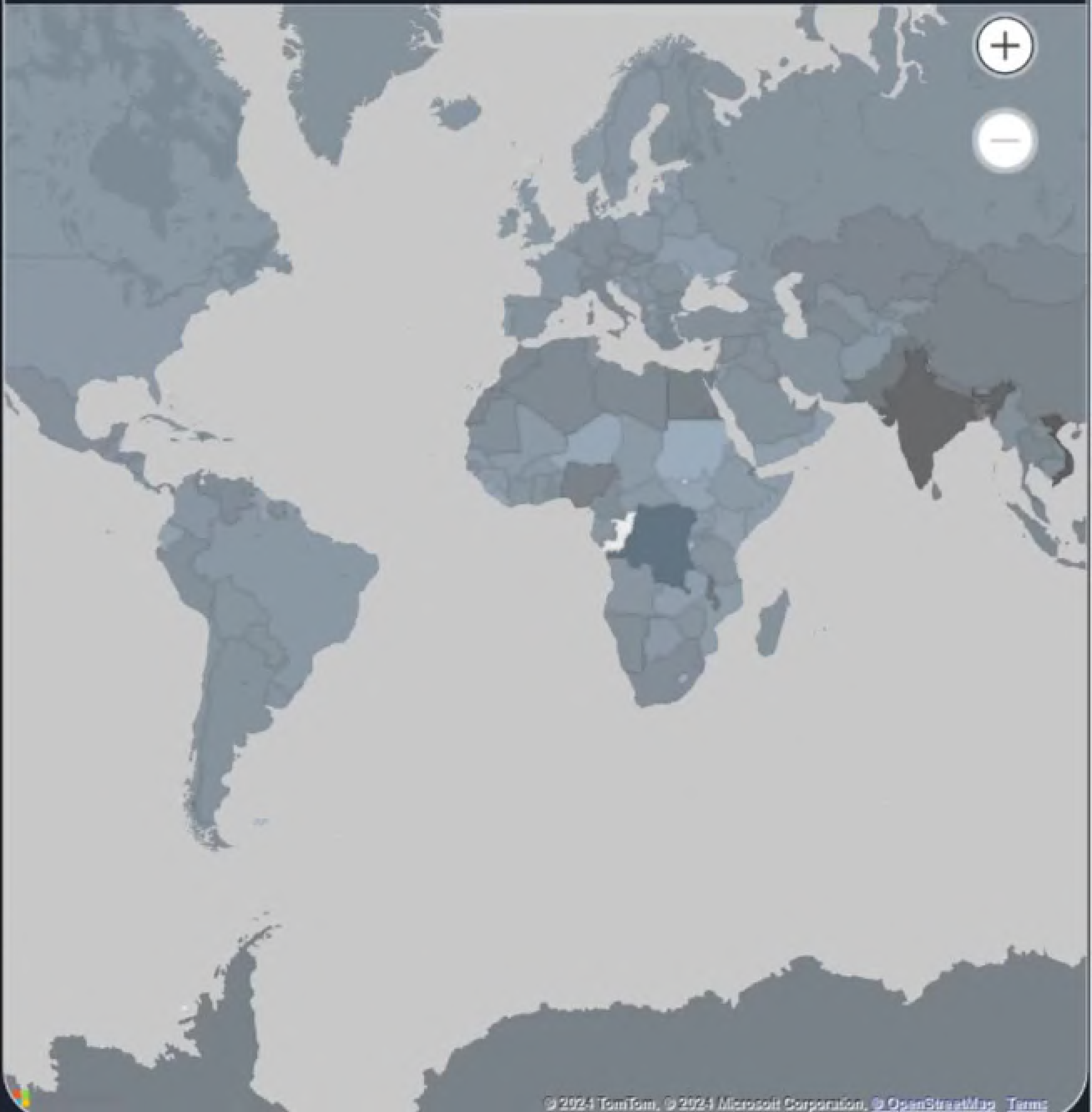

Spoken Lagnuages by total movies
English
French
Spanish
German
Japanese
Italian
No Language
Russian
Portuguese
Mandarin
Hindi
Swedish
Arabic
Cantonese
Korean
Polish
Dutch
Turkish
Tamil
Malayalam
Tagalog
Greek
Persian
Finnish
Danish
0,0M
0,1M


Spoken Lagnuages by total movies
Drama
Comedy
Documentary
Romance
Thriller
Action
Horror
Crime
Animation
TV Movie
Family
Adventure
Music
Science Fiction
Fantasy
Mystery
History
War
Western
0K
50K

## Number of Movies by Production Country




## Average Runtime by Production Country



# DATA MINING MODELS

In this section we will talk about :

- Classification model for rating forecasting
- Association rules

# Classification model for rating forecasting

Our goal is to create a model that can predict the rating of a movie based on various metrics.

**We will categorize the movies into low rating (<=4), medium rating (<=7), and high rating (>7).**

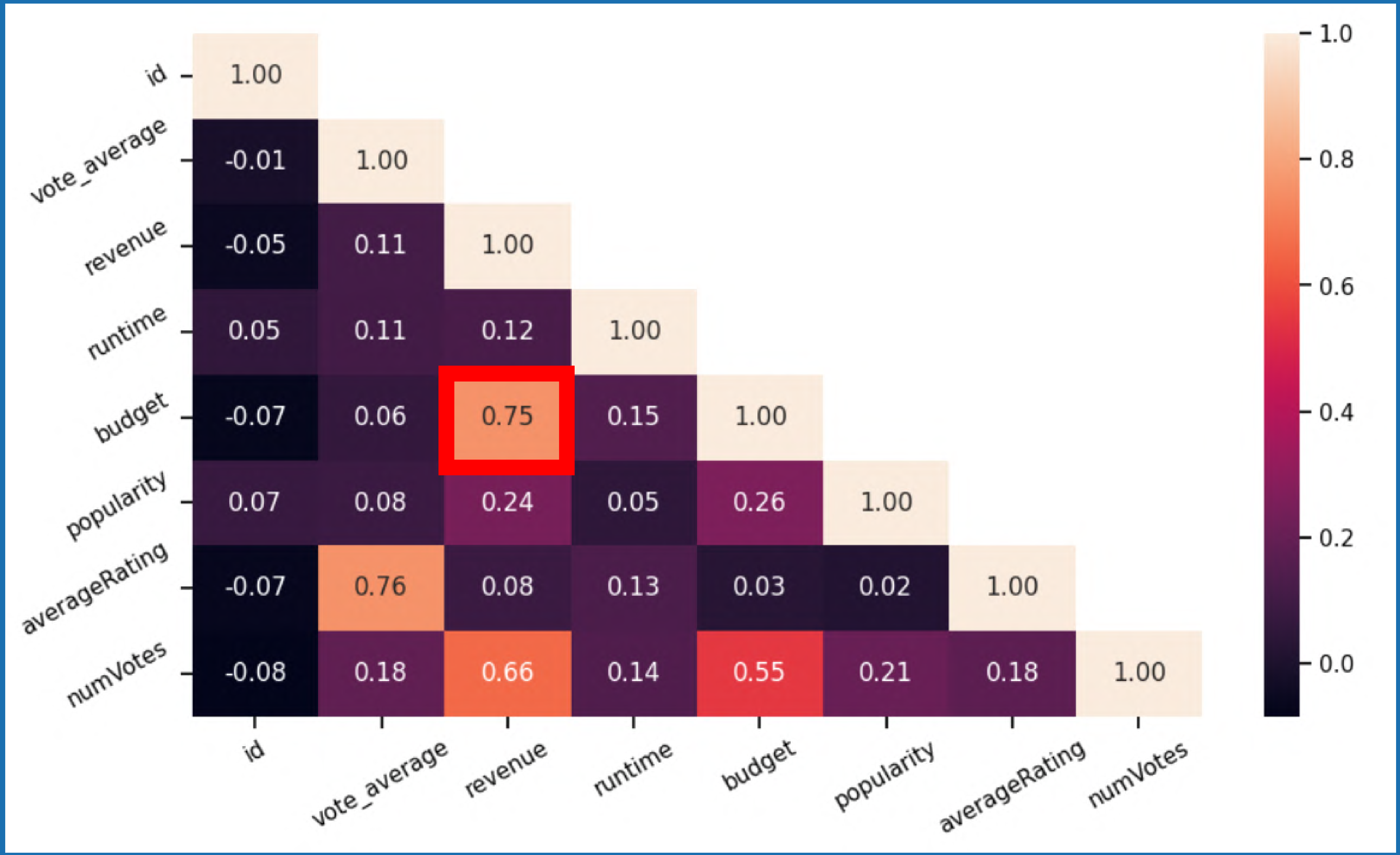
```
def categorize_rating(rating):
    if rating <= 4:
        return 0
    elif rating <= 7:
        return 1
    else:
        return 2

# Add a new column 'rating_category' to the DataFrame
df['rating_category'] = df['averageRating'].apply(categorize_rating)
```

# Classification model for rating forecasting

In order to create a trustworthy model, we aim to avoid high correlation among metrics, as our model's effectiveness could become overly dependent on them.

Therefore, based on the graph, we decide to combine revenue and budget into a single metric named 'Profitability'.



# Classification model for rating forecasting

For our model, we will include all the metrics shown in the graph, as well as additional metrics such as:

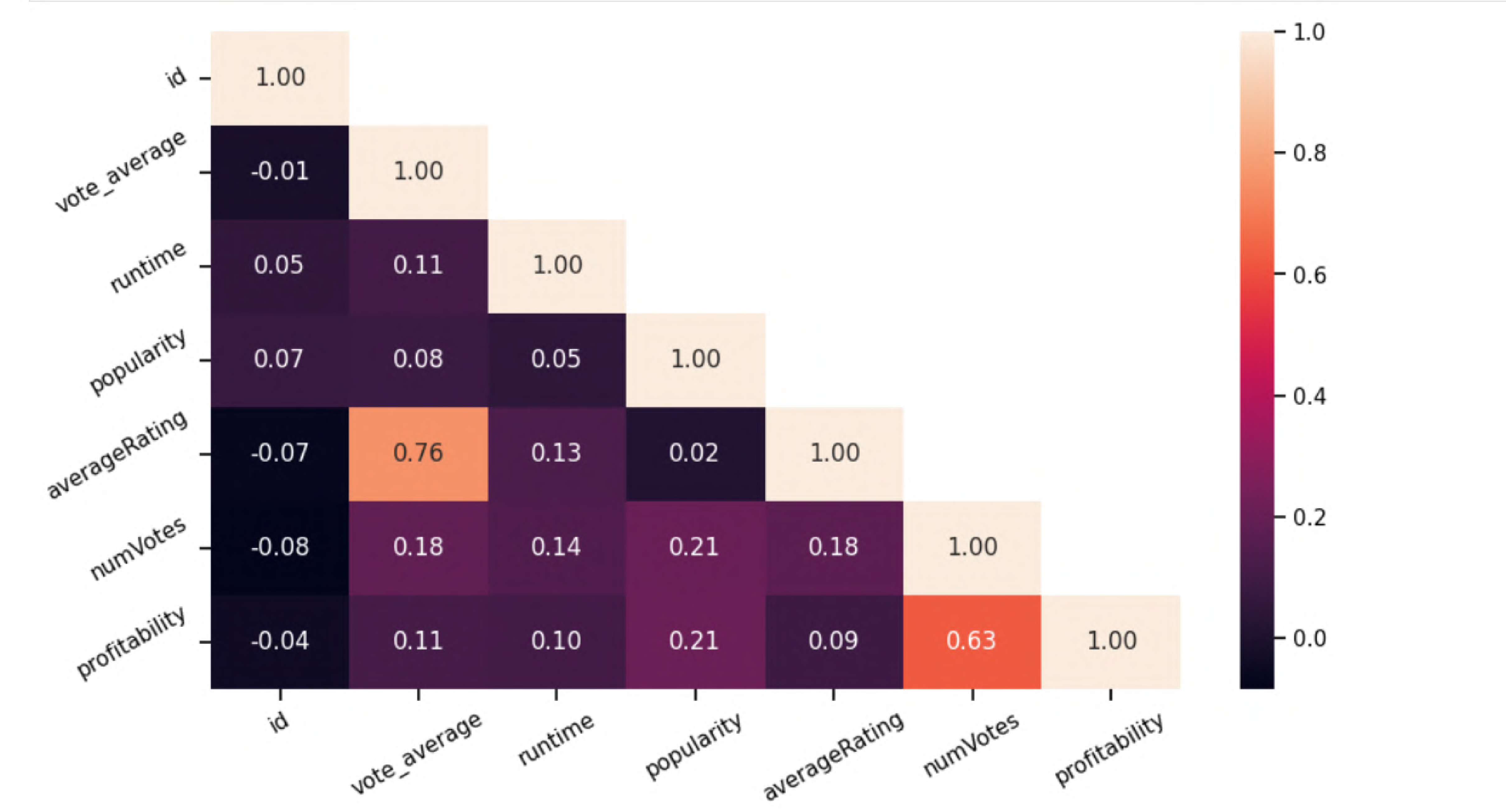
Platform

Genre

Production Countries

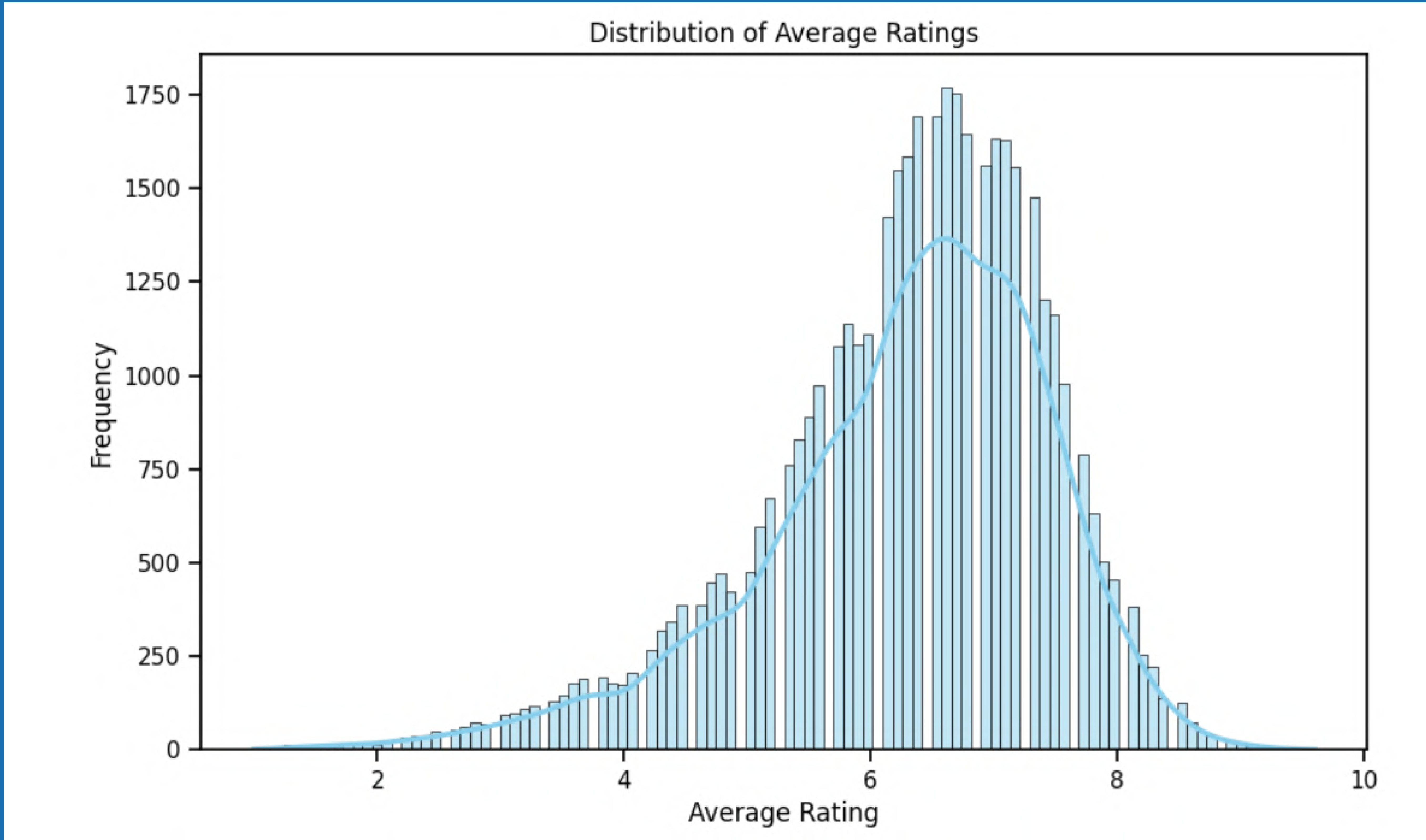
Spoken Languages

Years Since Release



# Classification model for rating forecasting

We can also examine the rating distribution among our IMDb movies. We observe that the majority of them fall <u>within the medium to large</u> rating classes



# Classification model for rating forecasting

Our model has an **accuracy** of nearly **88%**, making it a reliable choice for classifying the future rating of a movie based on the discussed metrics.

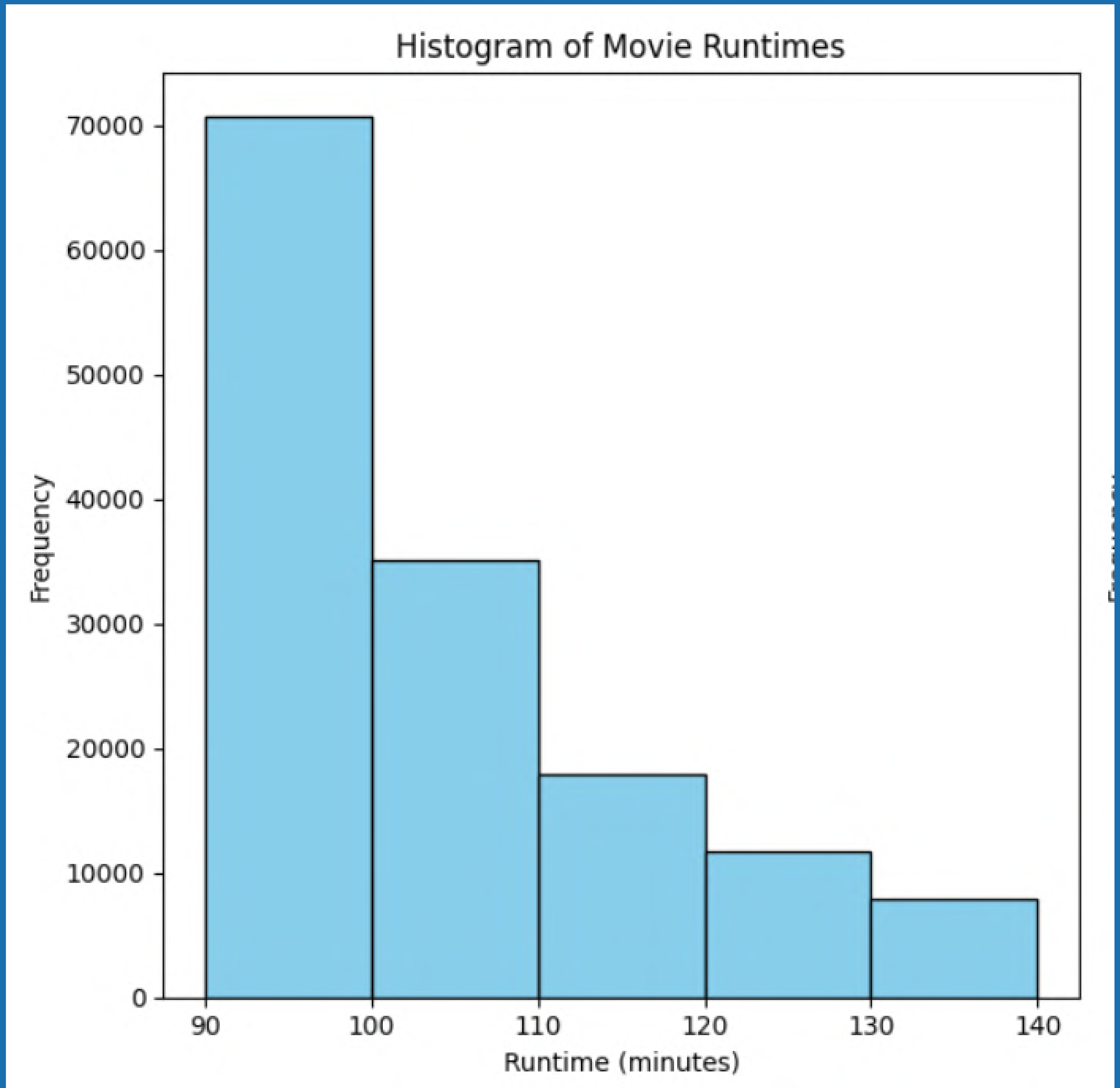
```
[93]: from sklearn import metrics
      from xgboost import XGBClassifier
      xgb = XGBClassifier()
      xgb.fit(X_train, np.ravel(y_train,order='C'))
      xgbprd = xgb.predict(X_test)
      cnf_matrix = metrics.confusion_matrix(y_test, xgbprd)
      print(cnf_matrix)
      print("Accuracy:",metrics.accuracy_score(y_test, xgbprd))
```

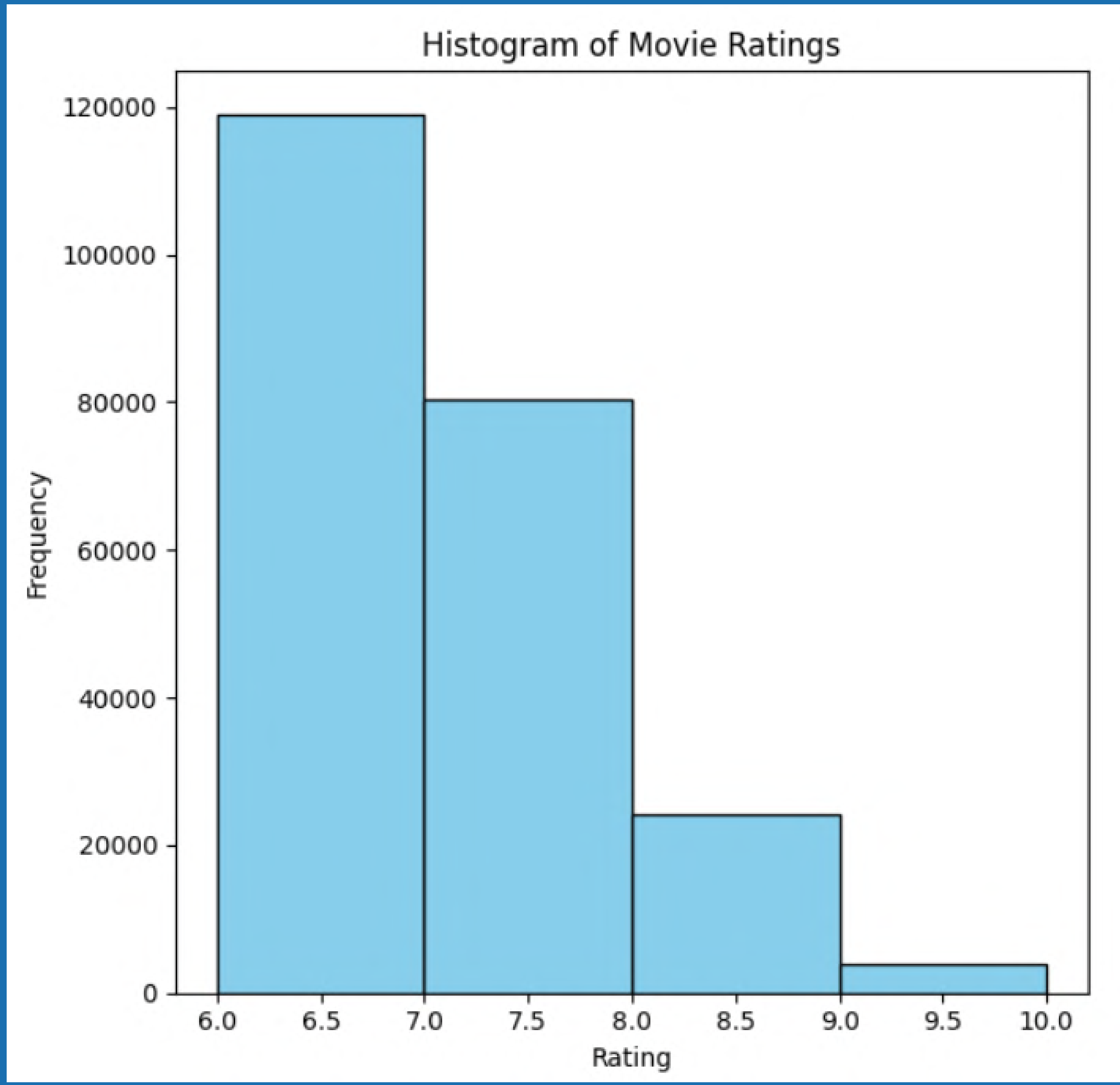
```
[[ 249  173    1]
 [  75 5396  323]
 [   1  488 1909]]
Accuracy: 0.8768427161926872
```

# Association Rules

This model will be based on the following features:

**Genre**
**Platform**
**Runtime**
**Rating**



# Association Rules

We have grouped runtime and rating into classes to make them easier to use in our Association Rules.

| runtime_low | runtime_medium | runtime_high | averageRating_low | averageRating_medium | averageRating_high |
|---|---|---|---|---|---|
| False | False | True | False | False | True |
| False | False | True | False | False | True |
| False | False | True | False | False | True |
| False | False | True | False | False | True |
| False | False | True | False | False | True |

# Association Rules

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| **46** | (Comedy) | (Amazon) | 0.355256 | 0.479980 | $\boxed{0.134435}$ | 0.378417 | 0.788403 |
| **48** | (Amazon) | (Drama) | 0.479980 | 0.441472 | 0.237178 | 0.494143 | $\boxed{1.119307}$ |
| **50** | (Thriller) | (Amazon) | 0.206679 | 0.479980 | 0.127109 | 0.615004 | $\boxed{1.281313}$ |

**Amazon - Genres**

- If a movie is a **Comedy**, it is **less likely** to be hosted on Amazon.
- If it is **Drama**, it is 1.12 times **more likely** to be on Amazon.
- If it is **Thriller**, it is 1.28 times **more likely** to be on Amazon. We are 61% confident in our finding.

# Association Rules

## Amazon - Average Rating

- We observe a tendency for Amazon to host low-rating movies; if a movie is on Amazon, it is 1.17 times more likely to be low-rating.
- If a movie has a high rating, it is 0.85 times more likely to be on Amazon.

[113]:

[113]:

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| **24** | (Amazon) | (averageRating_low) | 0.479980 | 0.349463 | 0.196115 | 0.408591 | 1.169195 |
| **38** | (Amazon) | (averageRating_high) | 0.479980 | 0.323053 | 0.132220 | 0.275470 | 0.852709 |

# Association Rules

**Drama - Runtime**

If the runtime of a movie is high, it is 1.40 times more likely to be a Drama movie.

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 22 | (runtime_high) | (Drama) | 0.322201 | 0.441472 | 0.19833 | 0.615547 | 1.394306 |

# Association Rules

**Comedy - Drama - Romance**

- Comedy movies are 0.68 times less likely to also be Drama at the same time.
- Comedy movies are 1.43 times more likely to be Romance movies.

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| **0** | (Comedy) | (Drama) | 0.330965 | 0.468937 | 0.106422 | 0.321549 | $\boxed{0.685697}$ |
| **3** | (Comedy) | (Romance) | 0.330965 | 0.178344 | 0.084413 | 0.255051 | $\boxed{1.430102}$ |
| **4** | (Romance) | (Drama) | 0.178344 | 0.468937 | 0.112249 | 0.629393 | 1.342170 |

BIBDA, DMST, Aueb

Athens, 2024

# THANK YOU