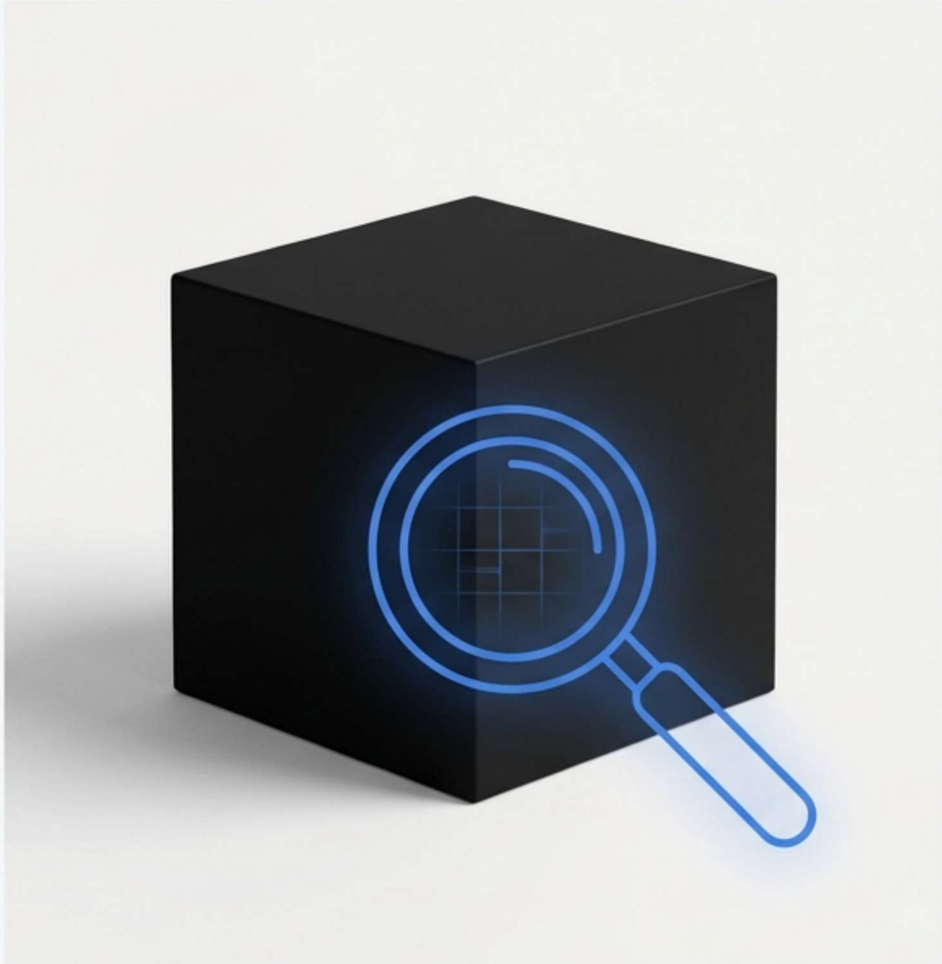
Decoding the Netflix Library 🎬

What 10,000 movies tell us about streaming trends and viewer tastes.

Exploratory Data Analysis • Python • Pandas • Data Visualization



The Netflix Library is a Black Box of Content

Netflix hosts thousands of titles, but simply browsing the catalog doesn't reveal the strategic big picture.

The Core Challenge: How can we move from unstructured data to business intelligence? We need to answer critical questions:

- Which genres truly dominate the platform?
- How has the volume of movie production changed over time?
- Does the platform's content strategy prioritize quality or quantity?

Our Approach: Auditing the World's Biggest Digital Bookstore

We treated the Netflix dataset like a massive bookstore in need of an audit. Our exploratory data analysis followed a clear, three-step process to create an inventory of its most valuable assets.

1

Organize the Shelves (Data Cleaning)

We fixed inconsistent data and handled missing values, just like dusting off old books and repairing covers.

2

Map the Sections (Genre & Release Analysis)

We categorized every movie by genre and release year to see which 'aisles' were largest and which were brand new.

3

Check the Borrower Logs (Vote Analysis)

We analyzed viewer ratings to see which titles are popular favorites versus those left gathering dust.

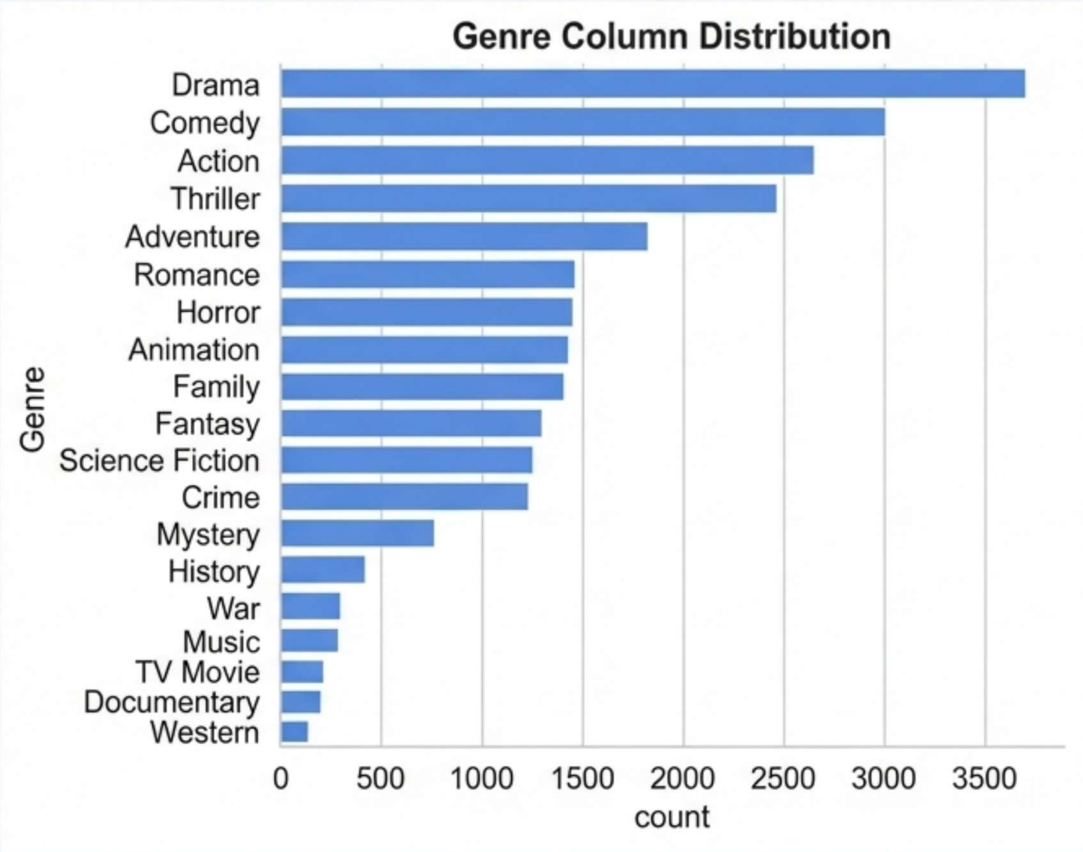
Finding 1: Drama and Comedy Form the Bedrock of the Netflix Catalog

Key Insight

An analysis of nearly 10,000 movies reveals that Drama is the single most frequent genre, followed closely by Comedy and Action. **This indicates a content strategy focused on mass-appeal categories.**

Supporting Points

- Niche genres like Westerns, Documentaries, and TV Movies represent a very small fraction of the library.
- This heavy investment in popular genres is likely a core tactic for broad subscriber retention.



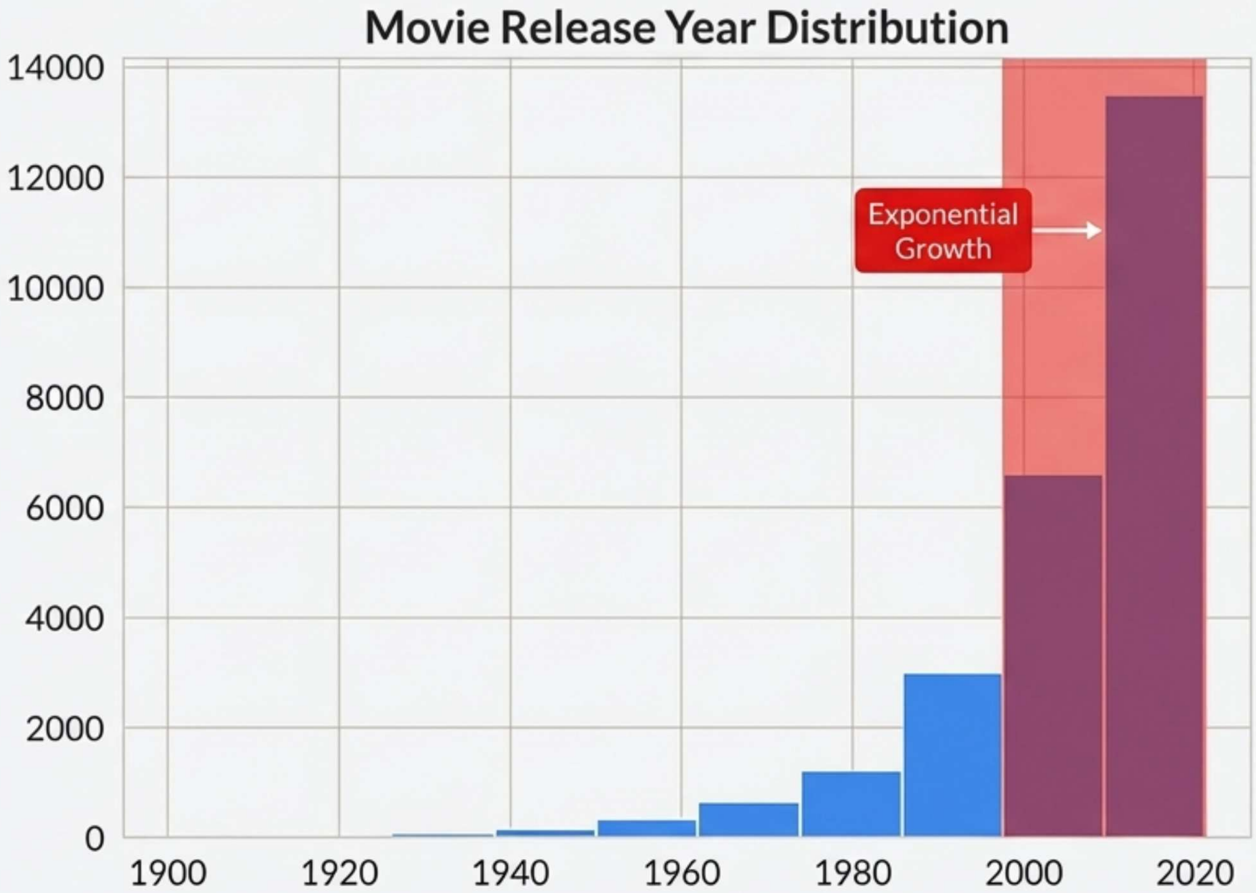
Finding 2: The Digital Era Triggered an Explosion in Movie Production

Key Insight

While movie releases were steady throughout the 20th century, production volume surged exponentially after the year 2000.

The 'Why'

This trend directly correlates with the rise of digital filmmaking and the streaming era, which dramatically lowered the barriers to producing and distributing content globally. The Netflix library is overwhelmingly modern.

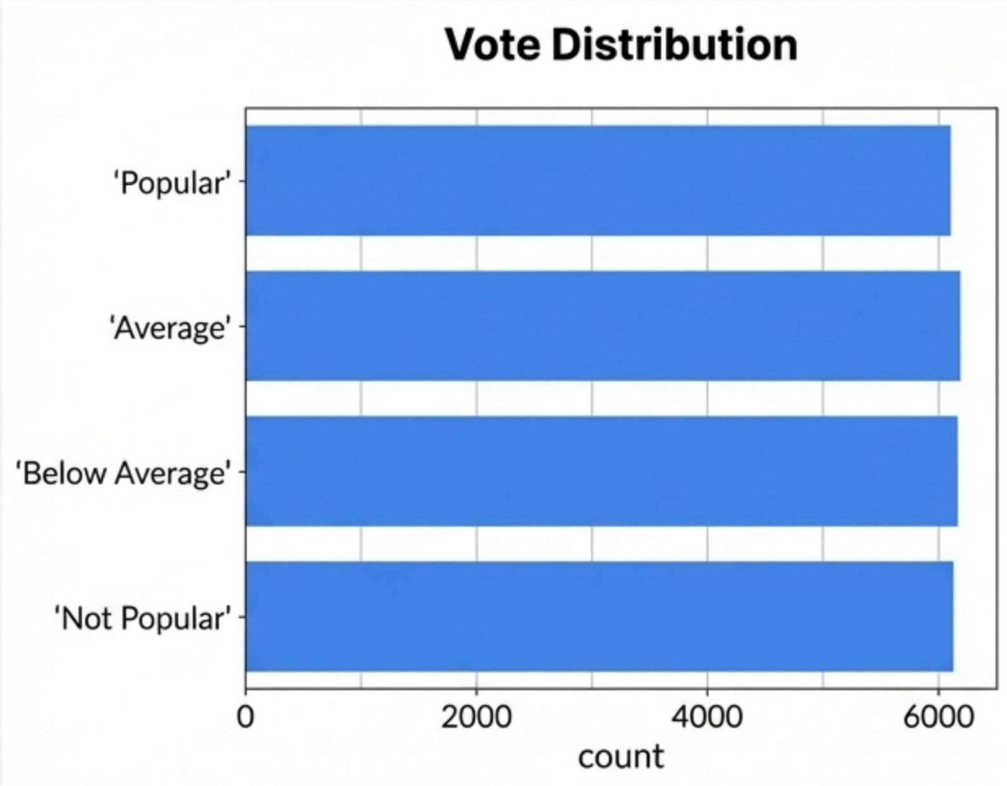


Finding 3: The Library is Curated to Favor "Popular" and "Average" Content

Methodology Note

To analyze viewer satisfaction, we engineered a new feature. We converted raw numerical scores into four distinct categories—"Popular", "Average", "Below Average", and "Not Popular"—by segmenting the dataset into equal-sized quartiles.

Key Insight
The distribution across these categories is even, which demonstrates that highly "Popular" content has significant representation. There is a clear effort to ensure a consistent supply of well-regarded titles.



NotebookLM

A Closer Look at ‘Vote Categories’ Reveals a Strategic Quality Filter

The categories aren’t just labels; they correspond to specific viewer rating brackets. The data shows a clear filtering mechanism at the low end of the spectrum.

Popular (Score >= 7.0)
Well-received titles with strong audience appeal.

Average (Score 5.0–7.0)
The bulk of the library, representing reliable content.

Below Average (Score 3.0–5.0)
A smaller, but present, cluster of movies.

Not Popular (Score < 3.0)
Minimal representation. Netflix actively avoids acquiring or retaining extremely low-rated content.

The Anatomy of the Analysis: Tools and Techniques

This project was built using a standard Python data science stack to perform data cleaning, feature engineering, and visualization.

Tech Stack

- Python
- Pandas
- NumPy
- Seaborn
- Matplotlib
- Jupyter Notebook

Key Technical Steps

1. Data Cleaning

```
[ ]: df.dropna(inplace=True)
```

Handled nulls and standardized column formats for accuracy.

```
df['Genre'].str.split(',')
df.explode('Genre')
```

2. Feature Engineering

Created the "Vote Category" from continuous scores and "exploded" the complex "Genre" column, splitting comma-separated lists into individual rows for precise counting.

```
categorize_col(df, 'Vote_A
```

The Big Picture: A Strategy of Modern, Mass-Appeal, Quality-Controlled Content

When combined, our findings paint a clear picture of the Netflix movie acquisition and retention strategy.

Focus on Broad Appeal

The catalog is heavily weighted towards Drama, Comedy, and Action—genres that attract and retain the widest possible audience.

Embrace the Now

The library overwhelmingly favors modern-era movies (post-2000), aligning with the production boom of the digital age.

Enforce Quality Control

The rating distribution demonstrates a clear floor on quality, suggesting a selective acquisition process that filters out poorly performing titles.

From Data to Decisions: How These Insights Drive Business Value

Understanding the composition of the content library has direct, actionable implications for key business teams.

For Content Acquisition Teams

The analysis identifies potential gaps in genre coverage (e.g., Westerns, Documentaries) and validates the success of the current mass-appeal focus.

For Recommendation Engine Teams

Insights on genre dominance and popularity can be used to refine algorithms that power the 'Because you watched...' and 'Top 10' features, boosting viewer engagement.

For Marketing & Strategy Teams

This data provides a quantitative understanding of audience preferences, informing campaigns and strategic planning for future content investments.