Letterboxd Movie Intelligence Report

A Data-Driven Deep Dive into Popularity, Quality, and Strategy

By: Aditya Dutta

Tools Used: Python (Pandas, Seaborn, XGBoost, NLTK), Power BI

Contents

1. Dataset Overview
2. Business Problems & Key Questions
3. Technical Approach & Challenges
4. Dashboard, Audience Insights, Genre Engagement and Sentiments
5. Strategic Insights (10 Questions Answered)
6. Summary & Recommendations

Dataset Description

We used the **Letterboxd Movie Classification Dataset** (Kaggle), containing over 9,000+ films with metadata like:

- Film Title, Description, Runtime
- Average Rating, Total Watches, Likes, Fans
- Genres, Studios, Director
- Original Language
- Breakdown of 1★, 3★, 5★ ratings

Data Cleaning & Preparation

We enriched the dataset by:

- Cleaning and parsing genres/languages
- Creating popularity and quality segments
- Running sentiment analysis on descriptions
- Building clustering, classification & regression models
- Developing a recommender system

Business Problems & Key Questions

Letterboxd, as a social film platform, needs to optimize:

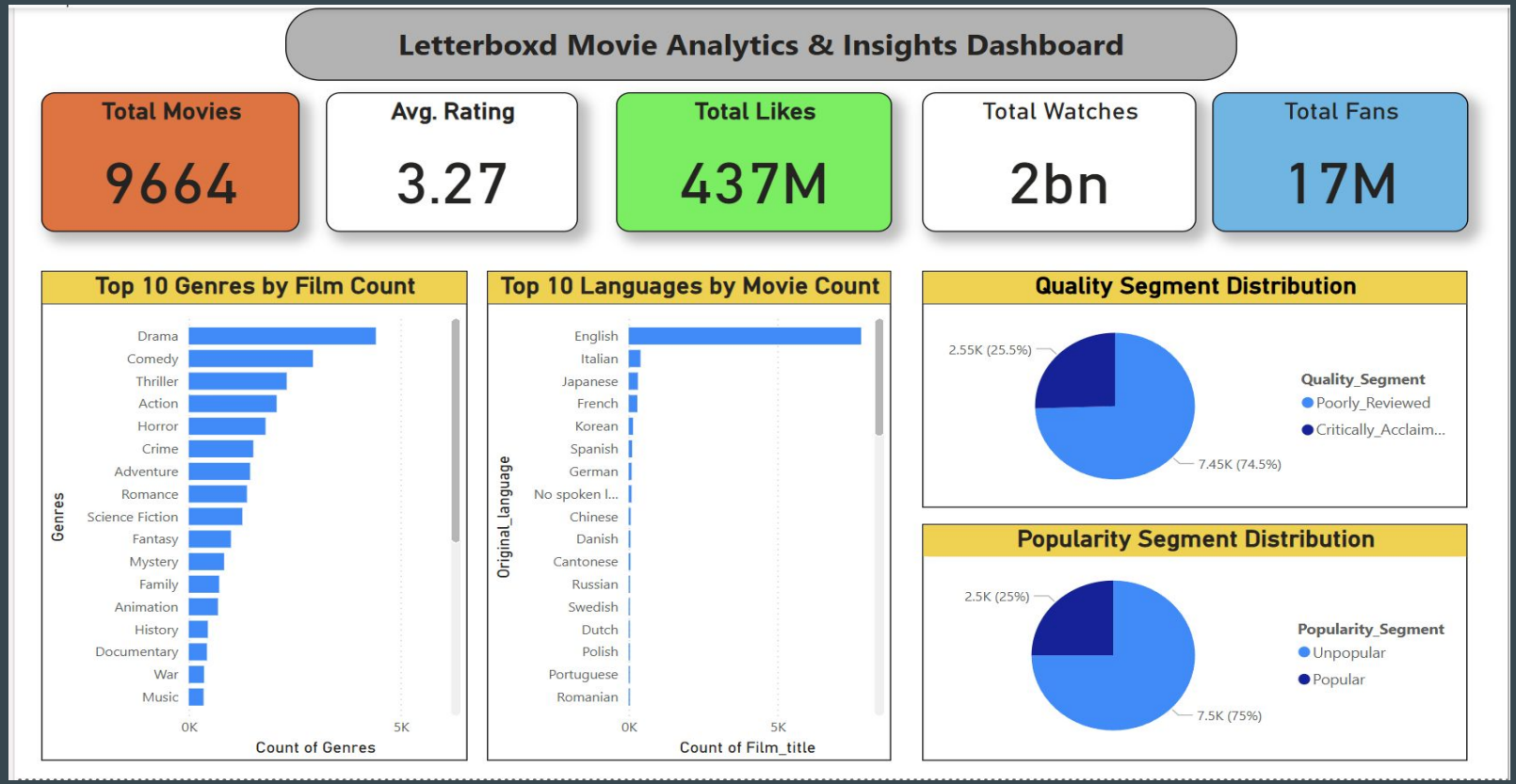
- **User Engagement**: What makes a movie popular?
- **Quality Metrics**: Can we predict what critics will love?
- **Genre Strategy**: Are we ignoring valuable niche films?
- **Recommendation UX**: Can we improve personalized content?
- **Investment Strategy**: What metadata signals success?

Technical Approach

We tackled this through:

- **EDA**: Language, genre, studio trends
- **Visualizations**: Distribution, bar charts, scatterplots, boxplots
- **Modeling**:
  - Regression (XGBoost): Predict average rating
  - Classification (XGBoost): Predict quality band
  - Clustering (KMeans): Segment films
  - NLP: TF-IDF for recommendations + Sentiment analysis
- **Power BI Dashboard**: 4 Pages — KPIs, Engagement, ML, Clustering

Overview (Power BI screenshots)



Audience Insights

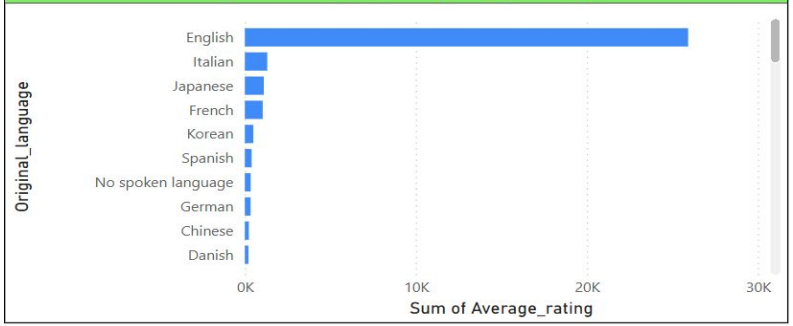
Audience Reception & Quality Segmentation

Top Avg Rating

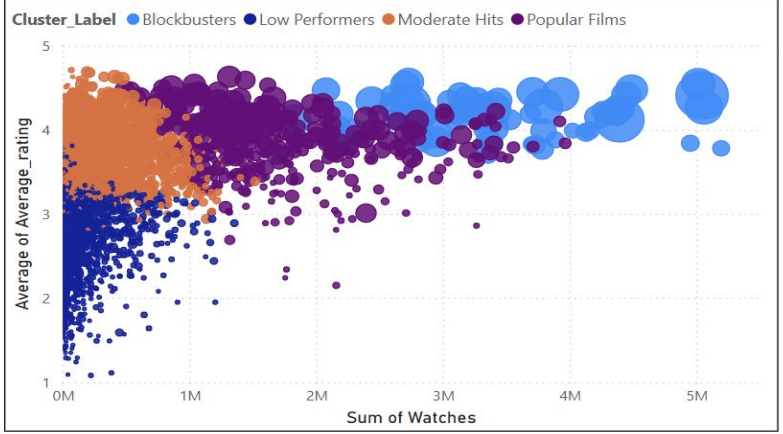
3.63

Cluster_Label	Avg_Rating
Blockbusters	4.16
Popular Films	3.80
Moderate Hits	3.66
Low Performers	2.91

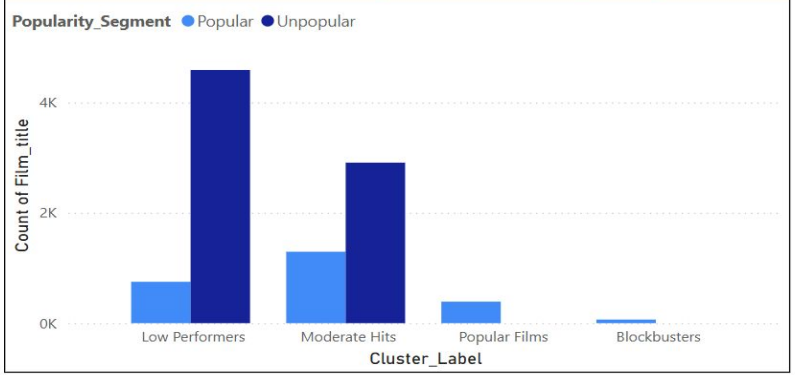
Top 10 Languages by Avg Rating



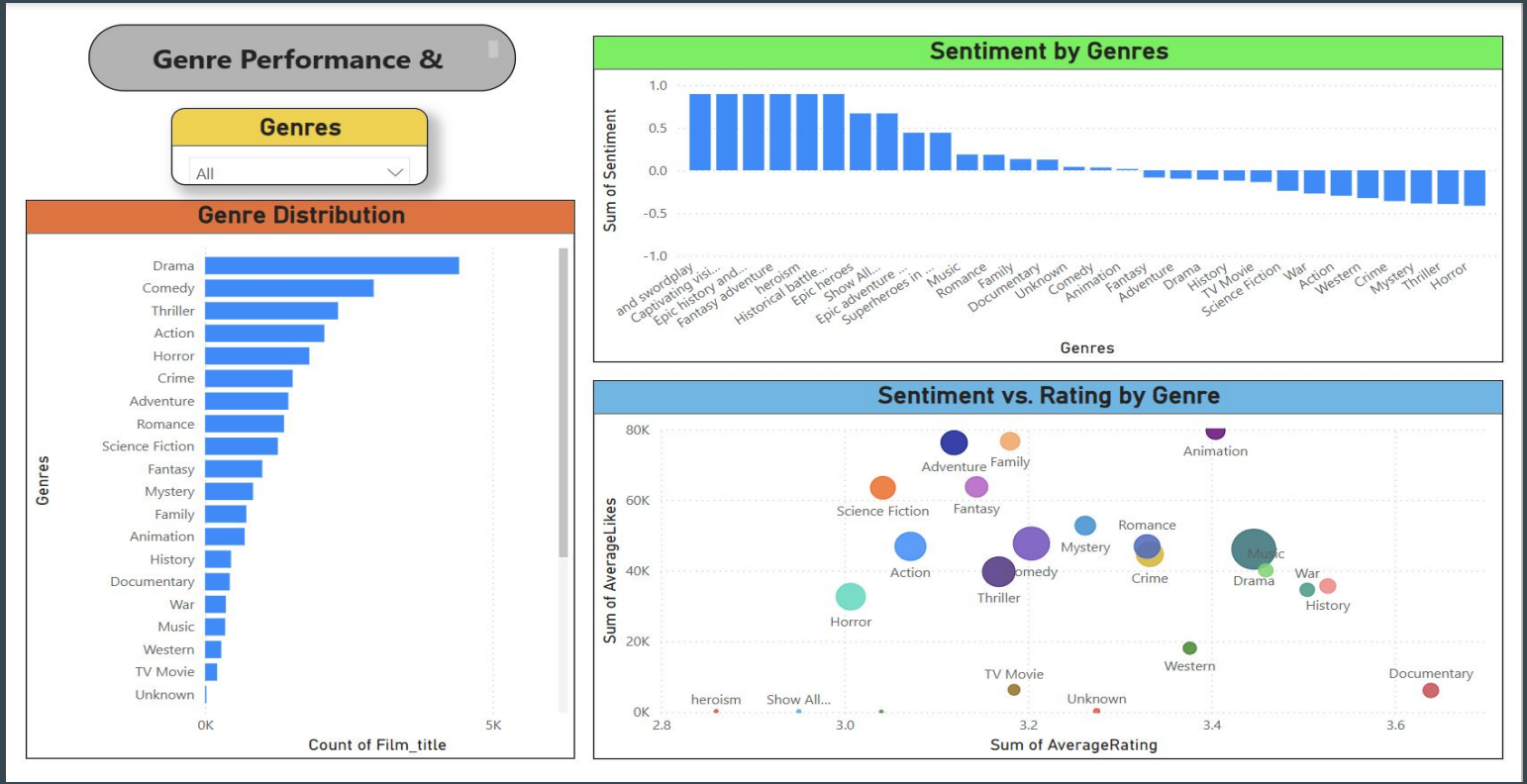
Watches vs Rating (Fan Size & Clusters)



Movie Distribution by Cluster and Popularity



Genre Engagements & Sentiment Based on Descriptions

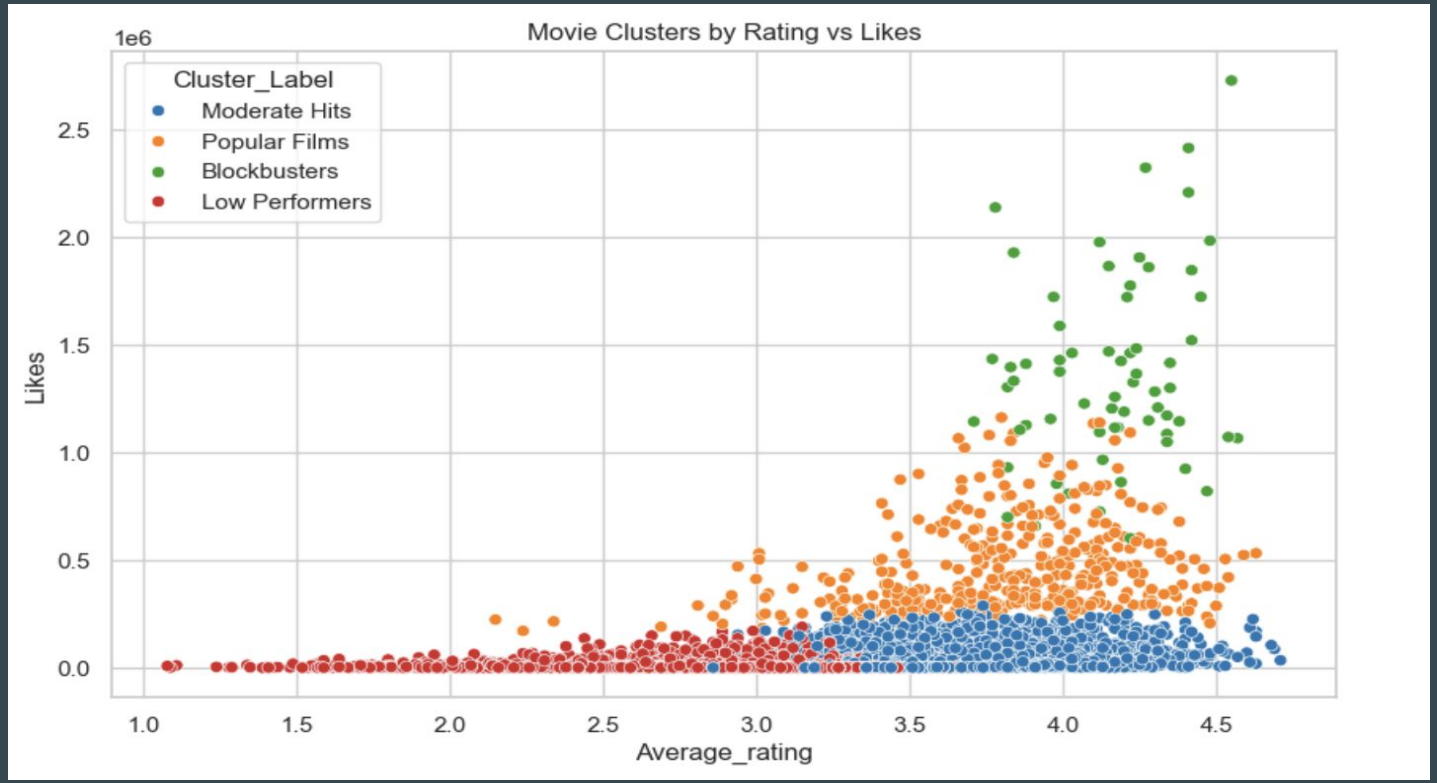


Q1: What traits define a blockbuster film?

Insight: Blockbusters tend to have: Higher runtime, Large fan counts and likes, Above-average watches

Cluster_Label	Average_rating	Watches	Likes	Fans	Runtime
Blockbusters	4.158462	3.418297e+06	1.377551e+06	87230.769231	130.923077
Low Performers	2.910075	5.471439e+04	7.208985e+03	78.335581	91.661269
Moderate Hits	3.660343	1.303489e+05	3.150708e+04	961.385969	115.663100
Popular Films	3.803872	1.642736e+06	4.533944e+05	16396.300000	121.407692

Movie Clusters by Ratings vs Likes



Key Findings:

Blockbusters consistently fall into a distinct cluster characterized by:

- **High Likes** and **Fan counts**
- **Above-average Runtime**
- Strong performance across **engagement metrics** (Watches, List Appearances)

These traits sharply differentiate them from low-performing or niche films.

Business Use Case

Use these insights to:

- **Forecast early-stage films** with blockbuster potential by checking for alignment on these traits
- **Prioritize distribution or homepage promotion** for titles matching this cluster

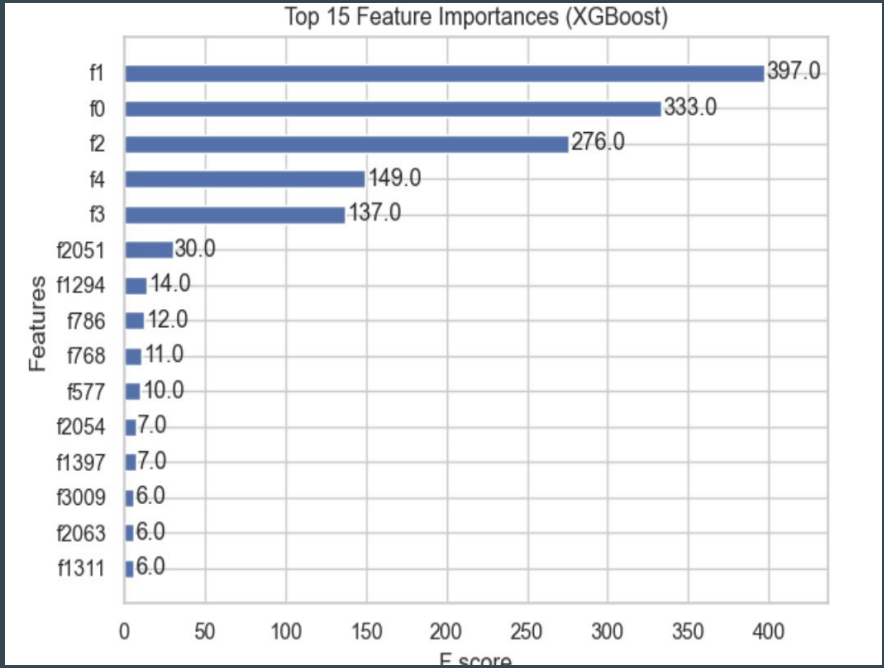
Strategic Recommendation

Create a **“Trending Blockbuster Picks”** carousel for the homepage featuring films from this cluster — it will likely boost both watch time and user engagement.

Q2: How Well Does Metadata Predict Film Success Pre-Release?

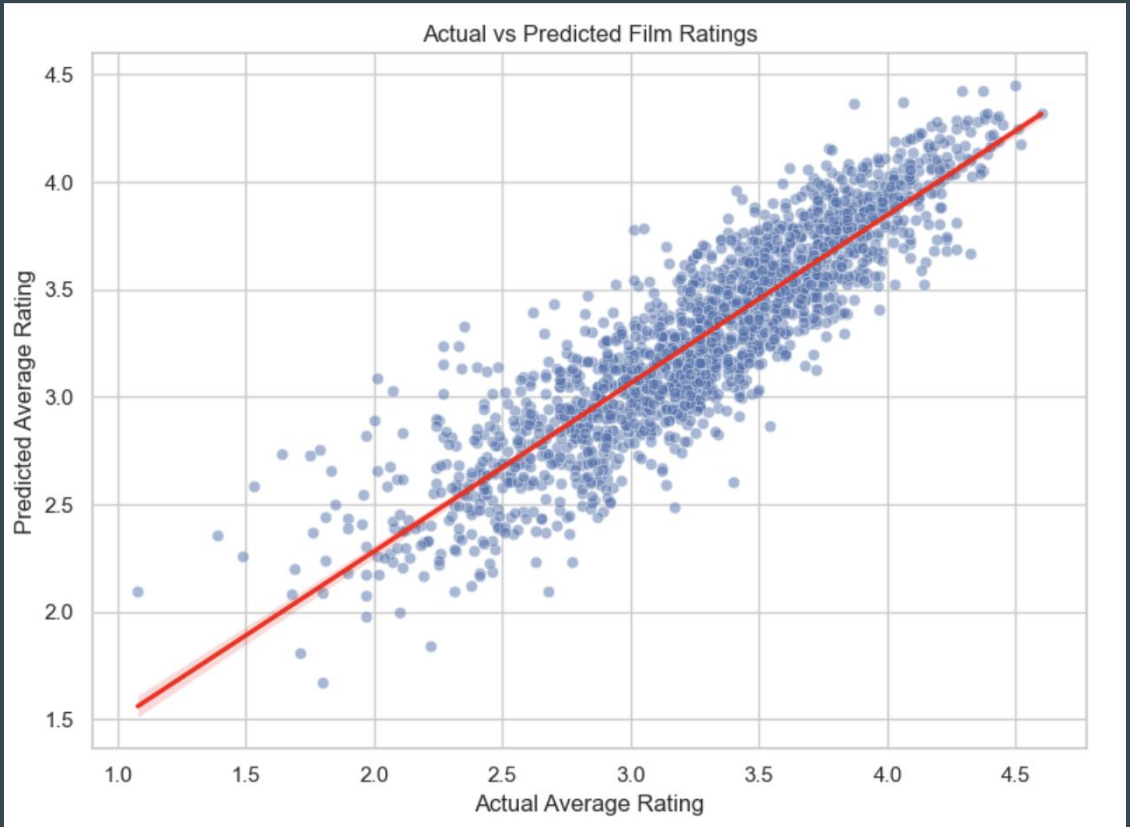
Source: XGBoost Regression Model on Metadata

(Features: Genre, Director, Studio, Runtime, Language, Watches, Likes, Fans, List Appearances)



	Actual_Average_Rating	Predicted_Average_Rating
0	3.42	3.066190
1	4.35	4.138627
2	3.73	3.809510
3	3.81	4.028441
4	3.69	3.645100
5	4.09	3.713751
6	3.35	3.490617
7	2.51	2.631384
8	2.98	2.952615
9	3.54	3.638750

Actual VS Predicted Film Ratings Scatterplot



Key Findings :

The regression model achieved:

- **R^2 Score ≈ 0.80** → 80% of the variance in average ratings can be explained by metadata
- **Top Predictive Features**:
  - Likes
  - Fans
  - Specific Genres
  - Director & Studio encoded with high importance

Business Use Case

- **Greenlighting & Budget Allocation**: Before release, use metadata to predict reception and confidently allocate promotional budget

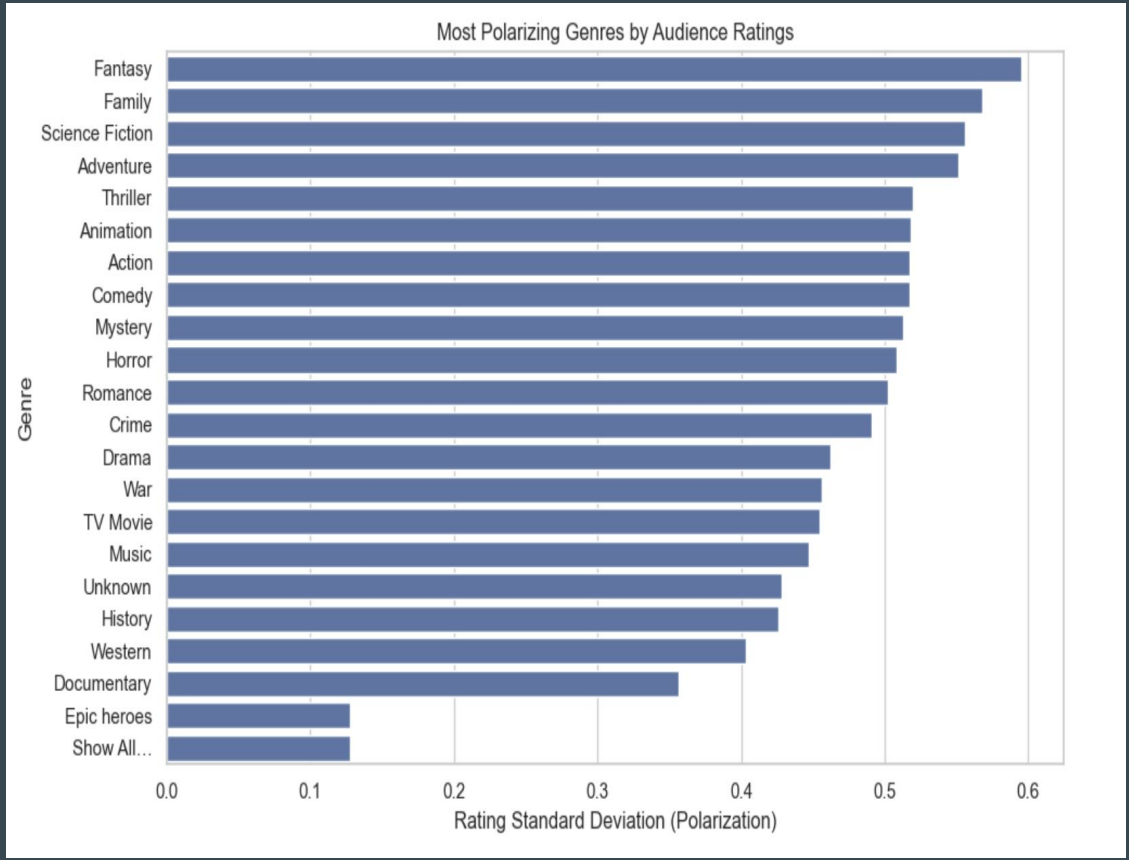
- **Early Quality Flags**: Detect likely underperformers for creative intervention or repositioning

Strategic Recommendation

Integrate this model into the **content acquisition pipeline** — enable studio partners to test scripts or metadata-based pitches against this tool.

Q3: Which Genres Are Most Polarizing?

Source: Standard Deviation of Ratings by Genre
(Computed from Average_rating distribution within each genre)



Standard Dev of Ratings by Genre

Key Findings

- **Genres with Highest Rating Variance**:
    - **Fantasy**, **Family**, and **Adventure** genres exhibited the **widest spread of audience ratings**.
    - These genres often attract **niche or passionate audiences**, leading to either high praise or strong criticism.
- Conversely, genres like **Drama** and **Documentary** had tighter rating distributions, indicating more **consistent audience perception**.

Business Use Case

- **Marketing Strategy**:
  - For polarizing genres, **segment audience targeting** more precisely (e.g., tailor trailers and social media to superfans).

- **UX Personalization**:
  - Recommend such content to users with matching historical sentiment patterns or genre affinity.

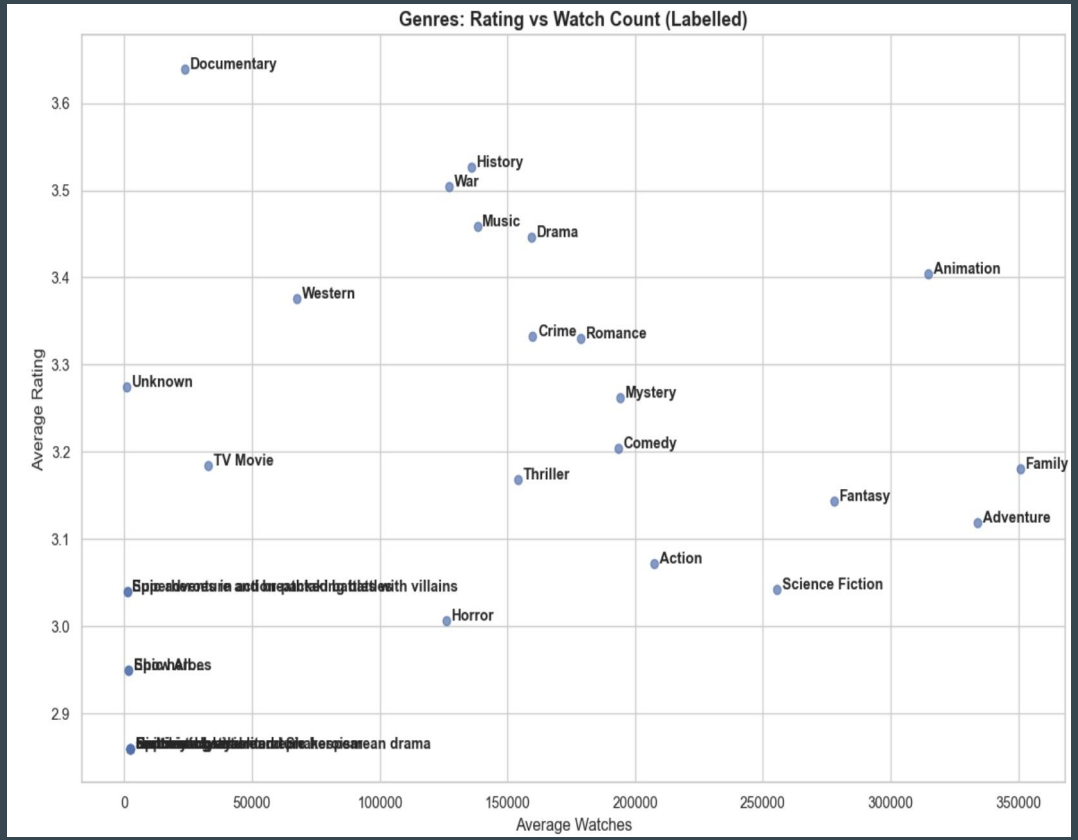
- **Partnerships & Funding Decisions**:
  - Exercise **creative caution or refinement** in polarizing genres unless a proven talent (director/studio) is involved.

Strategic Recommendation

Use genre-level variance to tag titles as "broad-appeal" vs "niche-risk" and adjust promotional effort, platform placement, and recommendation aggressiveness accordingly.

Q4: What Genres Are Undervalued but High-Quality?

- Aggregated Average_rating and Watches per genre.
- Filtered out genres with extremely low movie counts to reduce noise.
- Created a **scatter plot** of *Average Rating vs Average Watches*, labeling the genres.



Rating vs Watch Count Plot

Key Findings :

From the scatterplot:

- **Documentary, History, Music, War**:
  - Have **above average ratings** (3.4–3.6).
  - But relatively **low watch counts**.
  - These genres are **critically appreciated** but under-viewed.

- **Animation, Adventure, Family**:
  High watch count **and** strong average ratings → Already mass-market successes.

- **Fantasy, Action, Sci-Fi, Thriller**:
  - Popular but average-to-low rating genres — wide appeal but **polarizing quality**.

- **Horror**:
  - High visibility but **lowest average rating** — potential overproduction in this segment.

Business Use Case

This analysis surfaces **“Hidden Gem Genres”** — content areas with **high quality but low reach**.

These genres:

- Can be featured in **editorial promotions** (“Critically Acclaimed, Overlooked Films”).
- Offer **curation value** to cinephiles.
- Can guide **platform recommendation tweaks** to introduce users to high-quality niche content.

Strategic Recommendation

- **Elevate underappreciated genres** like *Documentary*, *War*, and *History* through:
  - Homepage features
  - Recommendation weighting
  - Thematic lists or campaigns (e.g., “Truth in Film” week)

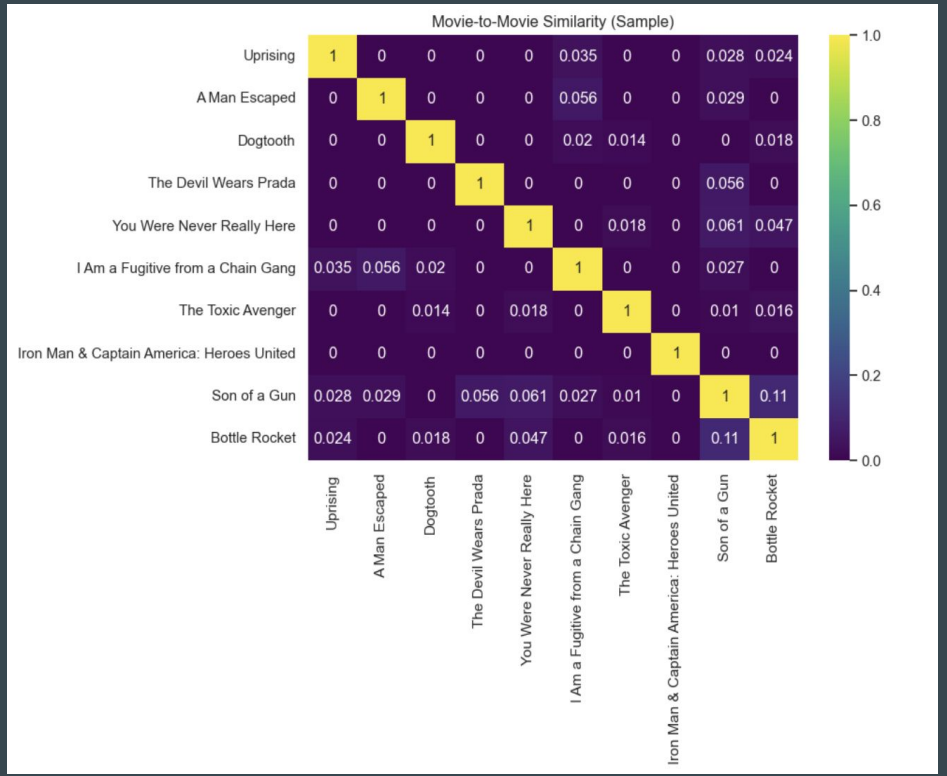
- **Partner with studios** known for high-quality genre films (e.g., indie documentary producers).

- **User Retention**: Offering well-reviewed but less-seen films increases the value for returning users tired of mainstream content.

Q5: Can we personalize recommendations by genre/language?

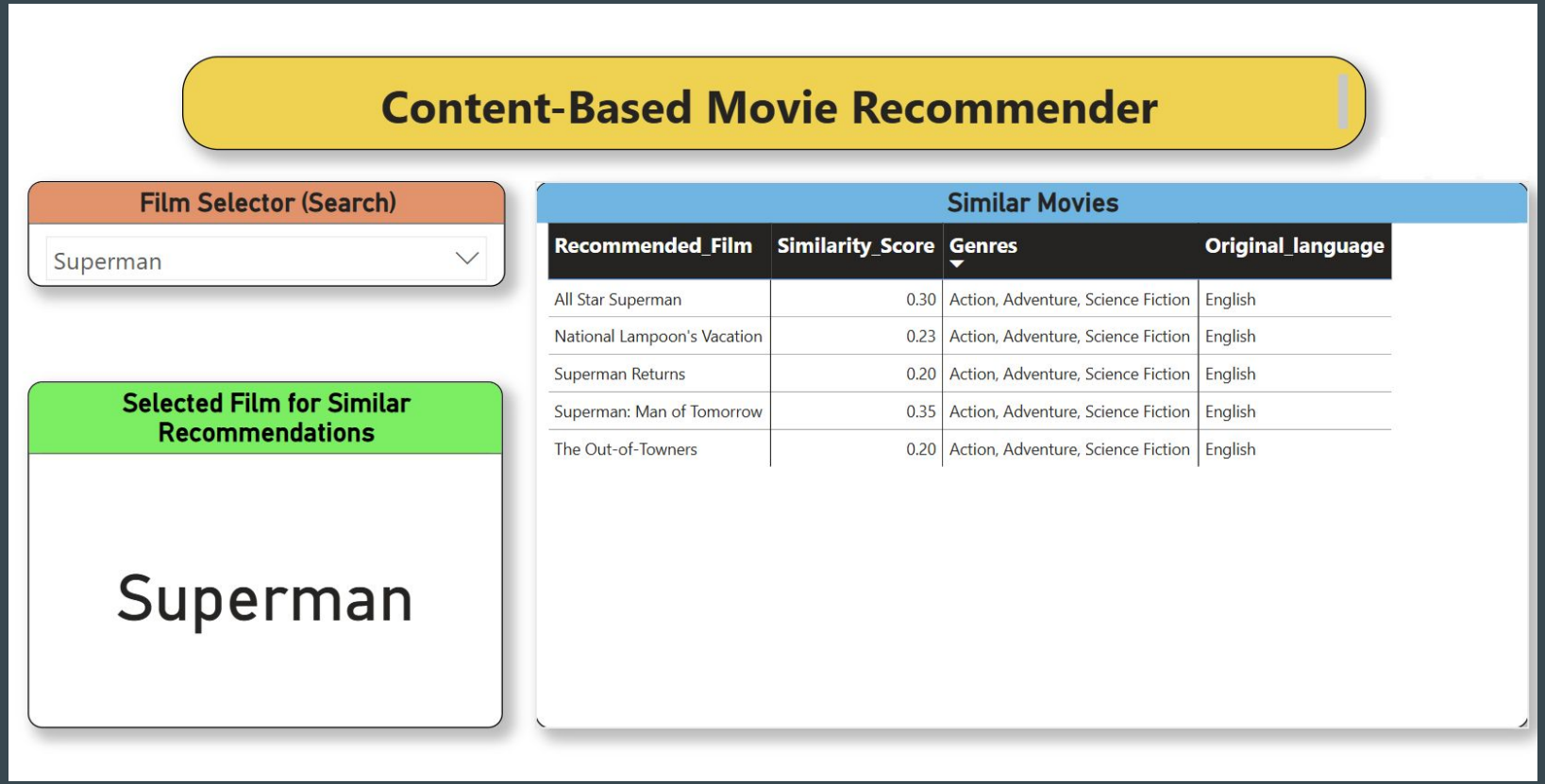
Yes — we built a **content-based recommender system** using **TF-IDF** on movie descriptions + **cosine similarity** to suggest similar films. These recommendations can be filtered or prioritized by **genre, language**, or **quality segment**.

In the similarity matrix, each cell shows how similar two movies are based on their descriptions. Bright yellow = highly similar. These scores drive personalized suggestions.



Similarity Matrix

Working Movie Recommendation System



Key Findings

- Movies with similar plots have strong cosine similarity — great for follow-up suggestions.
- Combined with metadata (e.g. language, genre), recommendations can be made more **relevant** and **localized**.
- Sentiment and popularity tags (e.g., “Blockbusters”, “Critically Acclaimed”) add additional filters.

Business Use Case

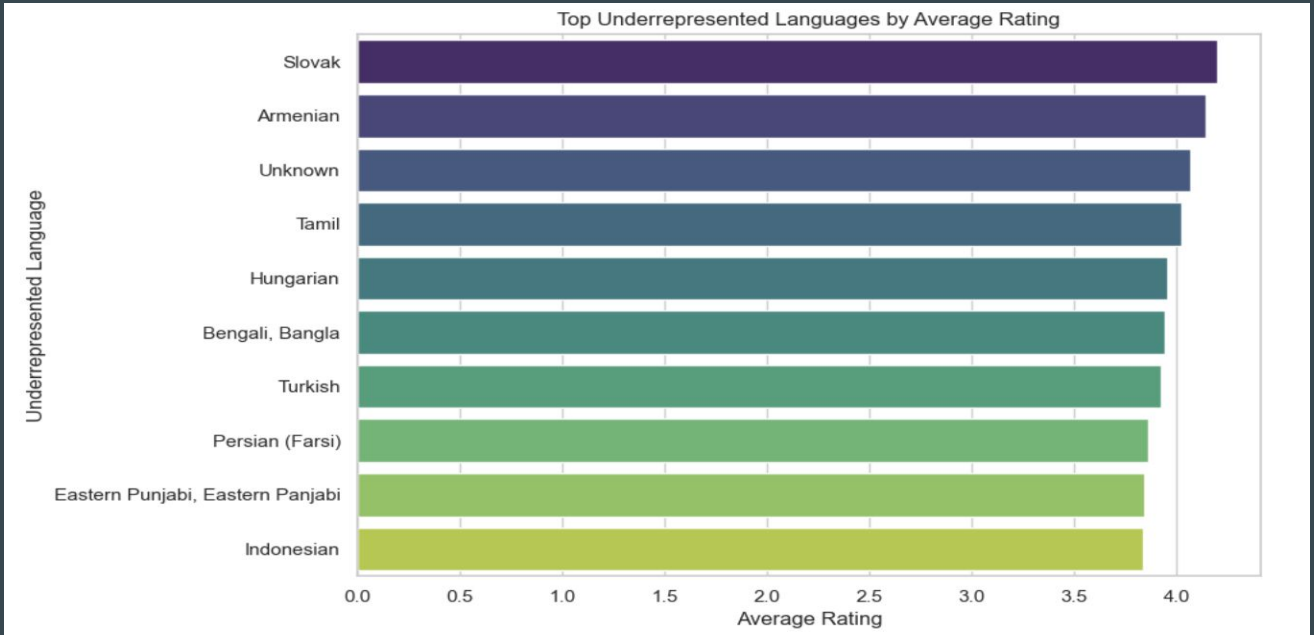
- Enables **personalized discovery**: users get similar films in genres/languages they prefer.
- Boosts **watch time** and **engagement retention** by offering smart suggestions immediately after viewing.
- Enhances **user satisfaction**, reducing bounce rates from aimless browsing.

Strategic Recommendation

- Integrate this recommender into Letterboxd's UX — show "Because you liked X..." with filters.
- Allow users to refine recommendations by **genre** (e.g., "Sci-Fi like Interstellar") or **language** (e.g., "Spanish-language thrillers").
- Promote lesser-known but similar titles — especially **hidden gems** with high sentiment.

Q6: Which Underrepresented Languages Show High Average Ratings?

Analysis of films grouped by **Original Language**, filtered to show **languages with low watch count** but **high average ratings**.



Key Findings

- **Top underrepresented languages** with consistently high ratings include:
  - **Slovak**, **Armenian**, **Tamil**, **Hungarian**, and **Bengali**.

- These languages receive **less exposure** (fewer total views) yet outperform mainstream languages in **average rating**.

- Indicates **strong content quality** in regional or niche cinema that is currently **underleveraged**.

Strategic Use Case

- **Promotional Strategy**: Spotlight films from these languages on curated lists or homepage banners.

- **Global Expansion**: Identify markets (e.g. Armenia, Slovakia) for platform growth or partnerships with regional studios.

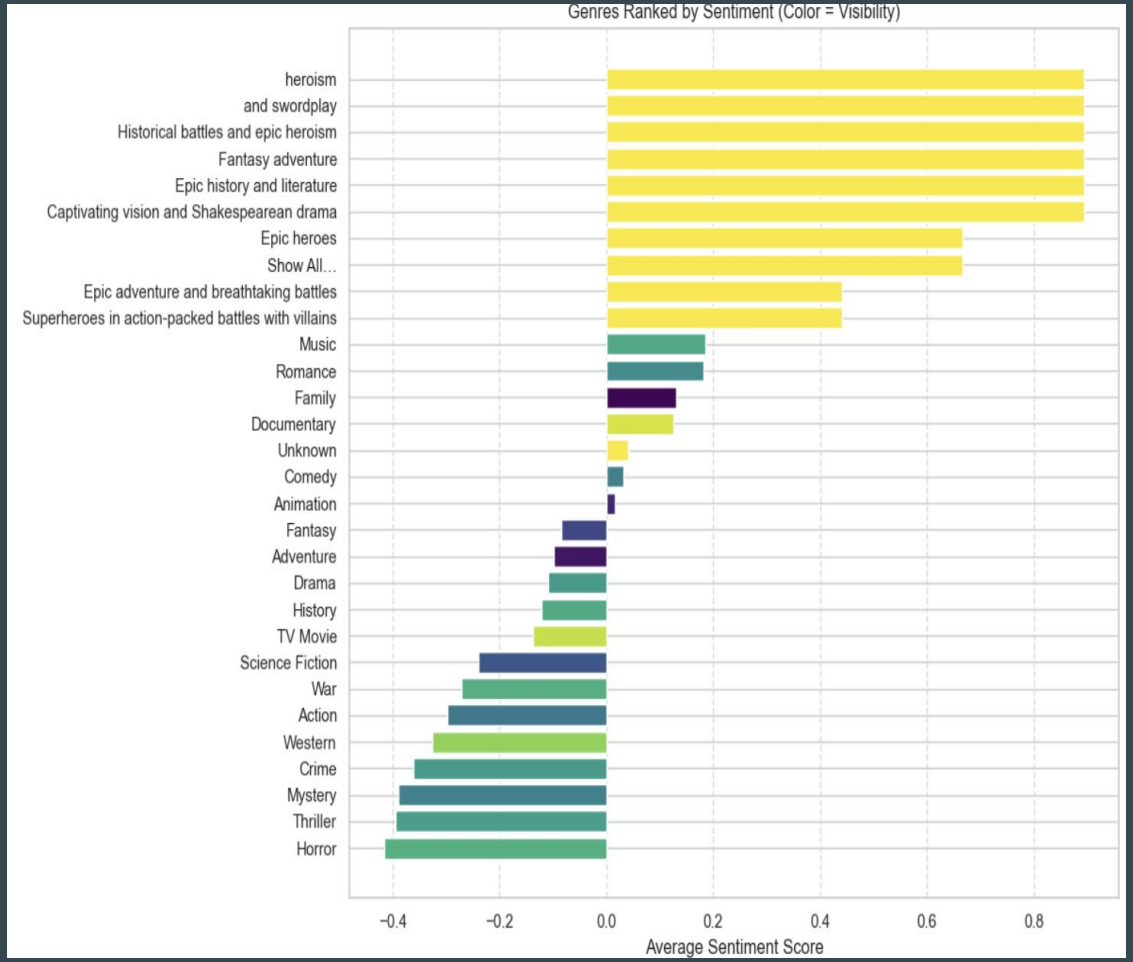
- **Recommendation Engine**: Boost these high-quality, low-visibility films in personalized recs to diversify user experience.

Business Impact

- Helps **retain cinephile users** seeking niche or international cinema.
- Opens up **cross-cultural content promotion** pathways and **long-tail engagement** opportunities.

Q7: Which genres combine high sentiment AND low visibility?

Some genres resonate deeply with audiences (high sentiment) but receive relatively low attention or viewership. Identifying these can guide promotion and investment strategies.



Genres Ranked by Sentiment

The horizontal bar chart above shows genres ranked by **average sentiment**, with **bar color** representing **visibility** (based on average watches).

- **Yellow bars**: High watch count (high visibility)
- **Darker bars**: Low watch count (low visibility)

Key Findings

- Genres like **"Documentary"** and **"Family"** receive strong audience sentiment yet appear less watched.
- High-sentiment thematic clusters like **"Epic history"**, **"Fantasy adventure"**, and **"Heroism"** also show varying degrees of low-to-moderate visibility.

Business Use Case

These genres are emotionally impactful yet underwatched — **perfect candidates for strategic promotion**. By surfacing these films to more users (e.g., featured carousels or email campaigns), platforms like **Letterboxd** can boost user engagement and discoverability.

Strategic Recommendation

- Develop curated playlists around **"Hidden Emotional Gems"**.
- Use **sentiment + visibility filters** in recommender systems.
- Invest in content partnerships for overlooked yet high-impact genres like **Documentary** and **Historical Fantasy**.

Summary & Recommendations

Key Wins

- Built 3 ML models (Classification, Regression, Clustering)
- Created actionable audience and content segmentation
- Identified promotional & partnership opportunities
- Delivered Power BI dashboard with 4 insight-packed pages

Suggested Next Steps for Letterboxd

1. Use cluster segments to launch targeted recommendation emails
2. Boost promotion of high-sentiment but low-view genres
3. Use predicted quality scores for unreleased film validation
4. Localize experience by top-performing underrepresented languages
5. Create a *"Director/Studio Spotlight"* feature for engagement