Winter 2025-2026 DS 687 Capstone Progress Report

Signal Extraction from IMDb Ratings and Metadata

Jan McConnell
Advisor: Dr. Brian Maeng
MS in Data Science
School of Technology & Computing (STC)
City University of Seattle (CityU)
janmcconnell@cityuniversity.edu, maengjooyol@cityu.edu

Abstract

The Internet Movie Database (IMDb) provides large, public relational datasets widely used in applied analytics. IMDb ratings and vote counts serve as compressed signals of audience engagement, aggregating individual evaluations into numeric measures. This project examines how production attributes such as genre, release period, and cast and crew roles relate to observed engagement patterns. A relational dataset is constructed using SQL, followed by exploratory and statistical analysis in Python and R. Results indicate a positive but variable relationship between ratings and vote volume, with substantial dispersion across genres and release periods. These findings suggest that ratings reflect evaluative intensity while vote counts capture participation scale, and that both are influenced by structural factors beyond perceived quality. The study highlights limitations in treating IMDb metrics as direct indicators of preference and supports their use as context-dependent signals in analytical and decision-support settings.

Keywords: Audience engagement, IMDb ratings, production metadata, statistical analysis, decision support

1. INTRODUCTION

Public film datasets such as IMDb are widely used in applied analytics because they combine large-scale audience feedback with structured production metadata. Millions of titles are represented through production metadata and aggregated audience ratings, creating a dataset that appears simple but reveals complexity during analysis.

IMDb (Internet Movie Database) ratings and vote counts provide a compressed signal of audience engagement by reducing individual user evaluations to numeric scores and participation volume. Although these measures are often treated as direct indicators of audience preference or quality, they reflect only part of how audiences engage with films and are subject to structural and behavioral limitations.

This project examines how production attributes such as genre, release period, and cast and crew roles relate to observable audience engagement as reflected in IMDb ratings and vote counts. The focus is on interpretation rather than prediction or recommendation.

IMDb is well suited to this analysis because titles, people, roles, and ratings are stored across multiple tables. As a result, schema design and data integration directly shape what can be analyzed, making data modeling a central part of the work rather than a preliminary step. This framing motivates an exploratory, relational approach focused on interpretation of engagement metrics.

1.1 Problem Statement

IMDb's public datasets provide structured information on film titles, production attributes, and aggregated audience ratings. In practice,

IMDb ratings and vote counts are often treated as direct indicators of audience preference or film quality, even though they compress complex audience behavior into simplified numeric signals.

The problem is that the relationship between ratings-based metrics and production characteristics is frequently assumed rather than examined. As a result, it is unclear how reliably these metrics support analytical or decision-oriented conclusions when used in isolation. Therefore, a clearer evidence-based understanding is needed, and where those signals are limited.

1.2 Motivation

Aggregated ratings are widely used in applied analytics because they provide a compact and easily communicated summary of audience response. However, ratings are not a context-free measurement. Prior work shows that ratings can shift based on contextual handling and how rating information is structured, which affects how meaningfully values can be compared across items or groups (Baltrunas & Ricci, 2009).

This matters because rating metrics are frequently reused for comparison, trend analysis, and decision support without explicit validation of what else may be influencing the signal. Research has shown that incorporating complementary information, such as reviews, can reveal dimensions of evaluation that ratings alone do not capture, suggesting that numeric scores by themselves may be incomplete or misleading indicators of audience response (McAuley & Leskovec, 2013; Leino & Räihä, 2007). For IMDb-style analysis, this motivates examining ratings and vote counts in context rather than treating them as standalone measures.

1.3 Approach

The analysis focuses on examining how observable production attributes align with aggregated measures of audience engagement. Guided by prior work that treats ratings as composite and context-dependent signals, the analysis focuses on examining how observable production attributes align with aggregated measures of audience engagement.

The first step involves constructing a relational analytical dataset from IMDb's normalized tables using SQL. This includes joining title, ratings, and personnel data; filtering records to relevant subsets; and engineering features that represent production attributes such as genre, release period, and participation of cast and crew roles. Particular attention is paid to schema design and join logic, as these choices directly shape which relationships can be meaningfully examined.

Exploratory analysis is then conducted using Python to examine distributions, correlations, and group-level patterns in ratings and vote counts across production attributes. Visualization and summary statistics are used to surface variation, concentration effects, and potential biases in engagement metrics. Statistical analysis in R supports validation of observed patterns and helps distinguish systematic relationships from noise.

Throughout the analysis, results are interpreted in light of documented limitations of ratings-based data, including selection bias and data sparsity in observed feedback (Steck, 2013; Wang et al., 2021). Because IMDb ratings and vote counts arise from voluntary participation rather than controlled sampling, engagement metrics reflect behavioral and platform effects in addition to audience evaluation. Rather than producing predictions or rankings, the approach prioritizes analytical transparency and contextual interpretation to support responsible use of IMDb engagement metrics in decision-oriented analytics.

1.4 Conclusions

This project clarifies how IMDb ratings and vote counts relate to production attributes such as genre, release period, and cast and crew participation, and to highlight where these engagement metrics are informative and where they are limited. By examining ratings-based signals within their relational and production context, the analysis moves beyond treating aggregated scores as direct proxies for audience preference or quality.

The anticipated outcome is a more grounded understanding of how audience engagement metrics behave across different production characteristics and data structures. From a practical perspective, the project demonstrates how careful relational modeling and exploratory analysis can support responsible interpretation of public datasets for decision support, particularly in settings where existing metrics are reused without full awareness of their assumptions and constraints.

1.5 Value to the Student

I am interested in this topic because it reflects the kind of data problems I want to work on after graduation: large, imperfect datasets where the challenge is not just modeling, but understanding what the data actually represents and how it can

be used responsibly. IMDb's datasets are complex, relational, and messy in ways that mirror real organizational data. Working with this data requires careful schema design, thoughtful joins, and deliberate feature construction before any meaningful analysis can take place, which directly strengthens my applied data analysis skills.

This project also aligns with my interest in decision support rather than predictive modeling for its own sake. In many professional settings, analysts are asked to interpret existing metrics, explain patterns, and communicate limitations, not simply produce predictions or rankings. By focusing on ratings as signals of audience engagement and examining where those signals are informative and where they fall short, this work helps me practice analytical judgment and communication. These are skills I expect to rely on in future roles where data is used to inform decisions rather than automate them.

1.6 Alignment to Program Outcomes

This project draws on core MS in Data Science competencies through applied work in relational data modeling, exploratory analysis, and analytical interpretation. It requires integrating data across multiple tables using SQL, constructing interpretable features, and analyzing relationships between ratings, vote counts, and production attributes using Python and R. Emphasis is placed on understanding data provenance, structure, and limitations rather than treating the data solely as input for modeling.

The analysis also reflects critical and ethical thinking by explicitly examining what ratings-based metrics can and cannot represent and by avoiding over-interpretation of compressed signals. Quantitative literacy is demonstrated through evaluation of patterns in the data to support interpretation rather than prediction. Communication skills are developed by framing findings for decision support, with clear explanation of patterns, uncertainty, and limitations using appropriate analytical narratives and visualizations.

2. BACKGROUND

Public film datasets such as IMDb represent large-scale observational systems in which audience ratings emerge from voluntary participation rather than controlled experimental design. In online rating environments, observed scores are shaped not only by individual evaluation but also by participation patterns and platform dynamics. Prior research in recommender systems shows that rating datasets are subject to selection bias, as users choose which items to rate and which to ignore (Steck, 2013; Wang et al., 2021). As a result, aggregated ratings reflect participation patterns in addition to underlying preferences.

Social influence further complicates interpretation of rating signals. Users may be affected by visible aggregate scores or prevailing opinions when forming their own evaluations. Conformity modeling research demonstrates that public rating environments can influence subsequent user behavior, meaning that ratings may reflect interactive social processes rather than independent judgments (Liu et al., 2016). Consequently, average ratings should be interpreted as outcomes of collective dynamics rather than direct measures of intrinsic quality.

In addition to behavioral factors, item attributes are often systematically associated with rating behavior. Research shows that user demographics and rating tendencies can be inferred from item-rating interactions, indicating structured relationships between item characteristics and audience response (Weinsberg et al., 2012). Joint modeling of ratings and review information also reveals that rating dimensions contain interpretable latent structure linked to item features (McAuley & Leskovec, 2013; Wu et al., 2016). These findings suggest that rating systems encode structured signals that can be analyzed in relation to observable item attributes.

Within this project, audience engagement refers to observable interactions reflected in aggregated ratings and vote counts. Production attributes denote descriptive characteristics associated with film creation and release, including genre classifications, temporal context, and creative roles. Ratings-based signals are therefore treated as indirect indicators shaped by participation bias, social influence, and structured item-attribute relationships. Interpretation in this study refers to statistical evaluation of these observable relationships within the constraints of large-scale observational data.

These perspectives frame IMDb ratings not as direct measures of audience preference but as observable signals produced through interaction between users, items, and platform structure.

3. RELATED WORK

Research in recommender systems has increasingly moved beyond treating user ratings

as isolated numerical values. Prior work shows that ratings often reflect multiple underlying factors, including item characteristics, contextual influences, and audience interpretation, which are not fully captured by a single score. As a result, researchers have explored ways to integrate auxiliary information, particularly textual reviews, to better understand and model audience engagement.

This related work synthesizes research that treats ratings as composite signals informed by review text, latent factors, and contextual structure. It focuses on approaches that jointly model ratings and reviews, examine multidimensional evaluation frameworks, and balance predictive performance with interpretability. Rather than cataloging individual methods, the review highlights how these strands of research collectively inform modeling and evaluation choices for analyzing audience response. While these approaches inform how ratings-based engagement signals can be interpreted, this project does not implement review-aware or latent factor models and instead uses this literature to guide exploratory analysis and interpretation.

Prior studies demonstrate that ratings alone provide an incomplete representation of audience response and that review text contains latent signals that help explain observed rating behavior. Aspect-aware and review-integrated models address this limitation by decomposing ratings into underlying components associated with item characteristics or user concerns, enabling more nuanced reasoning about audience engagement (Cheng et al., 2018).

Building on this idea, several approaches jointly model ratings and review text to improve both predictive performance and interpretability. Unified and hybrid models align latent rating dimensions with textual topics, allowing review content to inform how ratings are generated and interpreted (Ling et al., 2014; McAuley & Leskovec, 2013). These methods are particularly valuable in sparse or cold-start settings, where textual information can compensate for limited rating data.

More recent work extends review-aware modeling by incorporating multiple engagement signals and explicitly distinguishing between different forms of user interaction. Multi-factor collaborative approaches demonstrate that clicks, interactions, and ratings represent distinct but related audience signals that can be jointly modeled to improve recommendation quality (Liu et al., 2025). Similarly, co-clustering approaches such as Poisson Additive Co-Clustering (PACO) provide interpretable structures that link review topics to numeric ratings, offering explanations grounded in observed textual patterns and demonstrated on IMDb data (Wu et al., 2016).

Beyond modeling techniques, evaluation-oriented research emphasizes the importance of analyzing recommender systems across multiple dimensions rather than relying solely on aggregate accuracy metrics. Multidimensional frameworks that integrate recommender outputs with OLAP-style analysis enable structured exploration of performance across items, users, and contextual conditions, supporting more meaningful interpretation of engagement signals (Krohn-Grimberghe et al., 2010). Complementary work highlights limitations of rating prediction as a sole objective, noting the effects of selection bias, sparsity, and polarized rating distributions on observed outcomes (Steck, 2013; Wang et al., 2021).

3.1 Review Conclusions

The reviewed literature consistently supports treating ratings and reviews as complementary signals rather than independent sources of information. Aspect-aware, hybrid, and review-integrated models demonstrate that incorporating textual data improves predictive performance while enabling more interpretable representations of audience engagement (Cheng et al., 2018; Ling et al., 2014; Wu et al., 2016). At the same time, these studies reveal important trade-offs between model complexity and interpretability, reinforcing the need to align methodological choices with analytical objectives.

Evaluation-focused perspectives further show that relying solely on rating prediction accuracy can obscure important patterns related to bias, sparsity, and polarization in real-world datasets (Steck, 2013; Wang et al., 2021). Multidimensional evaluation frameworks provide a structured way to analyze recommender performance across multiple factors and better connect modeling results to practical interpretation (Krohn-Grimberghe et al., 2010). Together, this body of work establishes a strong foundation for analyzing audience engagement by jointly modeling ratings, review text, and contextual structure while remaining attentive to interpretability and evaluation limitations.

4. APPROACH

4.1 Overview

This project adopts a structured analytical

approach designed to extract interpretable engagement signals from large-scale relational IMDb data. Rather than treating IMDb ratings as direct measures of quality, the analysis treats ratings and vote counts as observable proxies for audience engagement embedded within a broader production context.

The methodology integrates three components:

1. Relational data modeling and integration
2. Structured feature engineering based on production metadata
3. Exploratory and statistical analysis for signal interpretation

Unlike recommendation-system or predictive modeling approaches, this project emphasizes interpretation and decision-support insight rather than prediction or classification performance. The objective is not to predict ratings but to examine how production attributes relate to observable engagement metrics.

4.2 User Requirements
The primary users of this analytical framework are decision-makers in content strategy, media analytics, and production planning. The approach is designed to meet the following requirements:

- Integrate multiple IMDb tables into a coherent analytical dataset
- Preserve relational integrity across titles, people, roles, genres, and ratings
- Enable structured comparison across production characteristics
- Provide interpretable outputs suitable for reporting and presentation
- Maintain methodological transparency and reproducibility

4.3 Design
IMDb data is distributed across multiple normalized tables, including title basics, ratings, crew, and name information. The first design step constructs a relational analytical dataset using structured SQL queries. Key joins are performed on unique title and person identifiers to preserve referential integrity and maintain the relational structure of the source data.

Filtering criteria are applied to:

- Limit analysis to feature films to ensure comparability across title type
- Remove incomplete records, including titles with missing release year, runtime, genre, or rating information
- Exclude titles with fewer than 5,000 votes to reduce instability in engagement measures
- Ensure consistent temporal metadata by retaining titles with valid release year information

Titles with very low vote counts can produce unstable engagement measures because average ratings derived from minimal participation are sensitive to small numbers of votes. To improve comparability across titles, the analytical dataset is restricted to films with at least 5,000 recorded votes. This threshold removes low-engagement titles while preserving a large dataset suitable for statistical analysis.

Even within this filtered dataset, vote counts remain strongly right-skewed, as illustrated in Figure 2. Some titles accumulate substantially more engagement than others. To address this, vote counts are log-transformed using the natural logarithm (ln(1 + numVotes)) to stabilize variance and support linear modeling assumptions prior to regression analysis.

The distribution of raw and log-transformed vote counts is examined in Section 6.1.

4.4 Feature Engineering
To move beyond raw metadata, production attributes are transformed into analytical features. These include:

- Genre categories
- Release period bins
- Vote count thresholds
- Role-based aggregation, such as numbers of credited actors or directors

Derived features enable structured comparison across titles and production characteristics. This step bridges raw relational data and interpretable engagement analysis.

4.5 Analytical Framework
The analytical component evaluates engagement signals through structured statistical examination rather than predictive modeling. The objective is to determine whether observable variation in ratings and vote counts across production attributes reflects systematic patterns rather than random fluctuation.

The analysis includes:

- Descriptive statistics of ratings and vote counts, including mean, median, variance, and distributional shape
- Group-level comparisons across genres,

release periods, and role-based attributes
- Correlation analysis between engineered production features and engagement metrics
- Regression modeling using categorical predictors to estimate the magnitude and direction of differences in engagement relative to reference categories, where coefficient values represent the change in log vote counts associated with each category
- Visual analysis through distribution plots and coefficient plots to support interpretability

The emphasis is on identifying consistent directional patterns and assessing whether production characteristics are associated with statistically distinguishable differences in engagement metrics.

4.6 Comparison with Existing Approaches
Existing research using IMDb data frequently focuses on:

- Recommendation algorithms
- Sentiment analysis of reviews
- Box office prediction models

These approaches primarily emphasize prediction and classification performance. In contrast, this project focuses on interpreting engagement signals within a relational production context.

Table 1 summarizes the differences between common IMDb analytical approaches and the interpretive framework adopted in this study. Predictive modeling prioritizes forecasting performance, while sentiment analysis focuses on extracting opinion signals from review text. The approach used in this project instead emphasizes transparent interpretation of engagement signals derived from relational metadata.

Approach Type	Primary Goal	Strength	Limitation
Predictive Modeling	Forecast ratings or revenue	High predictive power	Limited interpretability
Sentiment Analysis	Extract review sentiment	Text-level insight	Detached from production metadata
Proposed Approach	Interpret engagement signals in relational context	Transparent and interpretable	Does not optimize predictive accuracy

Table 1: Approach Comparisons

This comparison highlights the methodological trade-off guiding the project design: prioritizing analytical transparency and interpretability rather than maximizing predictive accuracy.

4.7 Implementation and Technologies
The implementation environment includes:

- SQLite for relational database management
- SQL for data extraction and transformation
- Python for data processing and statistical analysis
- Pandas and NumPy for structured computation
- Matplotlib for visualization
- VS Code for development and version control

The primary regression specification was independently replicated in R to confirm coefficient stability and model fit consistency across analytical environments.

All analysis steps are documented and reproducible. The workflow separates data loading, transformation, feature engineering, and analysis to maintain clarity and auditability.

4.8 Ethical and Methodological Considerations
Although IMDb data is publicly available, the analysis explicitly avoids overstating ratings as measures of objective quality. The study recognizes sampling bias inherent in voluntary rating systems and clearly distinguishes correlation from causation when interpreting statistical relationships.

These safeguards help ensure that conclusions remain analytically grounded and defensible.

4.9 Interpretation Criteria
Interpretation for this project is grounded in regression-based evaluation of engagement differences across production categories.

1. Coefficient Interpretation:
   Categorical regression coefficients estimate differences in engagement relative to an omitted reference category (Drama genre). Coefficients represent expected differences in log vote counts or rating compared to the baseline group.
2. Statistical Significance:
   Statistical significance is assessed using 95% confidence intervals (CI). Effects are considered statistically distinguishable when confidence intervals do not cross zero.
3. Model-Level Evaluation:
   Overall model fit is assessed using R^2 and

the F-test. R^2 provides the proportion of variance in engagement explained by observable production attributes. The overall F-test evaluates whether the model explains significantly more variance than a null model.

4. Substantive Interpretation:
   Even when statistically significant, coefficients are interpreted cautiously. Small R^2 values indicate that production attributes explain only part of engagement variation, reinforcing that ratings reflect multiple structural and behavioral influences.

Metric	Value
Observations	318,870
R^2	0.169
Adjusted R^2	0.169
F-statistic	3,096
p-value (F-test)	< 0.001

Table 2: Regression Model Summary

As shown in Table 2, the regression model explains a modest portion of variation in engagement outcomes. The model R^2 is 0.169, indicating that observable production attributes account for 16.9% of the variance in log vote counts. The overall F-test is statistically significant ($F = 3,096$, $p < 0.001$), indicating that the predictors jointly explain more variance than a null model.

These criteria ensure that conclusions reflect structured relational patterns rather than overinterpretation of noisy signals.

5. DATA COLLECTION

5.1 Data Source

This study uses the publicly available IMDb non-commercial datasets distributed through IMDb's developer portal. The data are provided as tab-separated value (TSV) files and include normalized tables describing titles, ratings, crew, and related metadata.

The primary IMDb tables used in this study include the following:

- name.basics.tsv.gz, which provides reference information for personnel appearing in IMDb titles
- title.basics.tsv.gz, which contains core title metadata such as type, release year, runtime, and genres
- title.crew.tsv.gz, which records director and writer relationships for each title
- title.principals.tsv.gz, which lists principal cast and production roles associated with each title
- title.ratings.tsv.gz, which contains IMDb average rating and vote count metrics

These datasets follow a relational structure in which each film or television title is identified by a unique alphanumeric identifier (tconst). Personnel are similarly identified by a unique identifier (nconst). These identifiers enable joins across multiple tables, linking title-level attributes such as genre, release year, and runtime with audience engagement metrics such as average rating and vote count, as well as production roles including directors, writers, and principal cast members.

The datasets are distributed in UTF-8 encoded TSV format with explicit markers for missing values. Because the files represent normalized relational tables rather than a single analytical dataset, multiple tables were integrated through key-based joins to construct a consolidated dataset suitable for statistical analysis.

No human subjects were involved in this study. All data were secondary, publicly available, and aggregated; therefore Institutional Review Board approval was not required.

The datasets were downloaded and stored locally for reproducible analysis in February 2026.

5.2 Data Storage and Database Construction

The raw TSV files were imported into a local SQLite database (imdb_data.db) to preserve relational structure and enable structured SQL querying. Each IMDb table was loaded as a separate relational table.

Primary keys and identifiers such as tconst (title identifier) and nconst (name identifier) were used to maintain relational integrity across tables. The normalized schema was preserved during import.

A filtered analytical table, filtered_movies, was constructed using SQL joins between title_basics and title_ratings. This table contains one row per title and includes:

- tconst
- primaryTitle
- startYear
- genres
- averageRating
- numVotes

The filtered_movies table contains one row per

film. The dataset spans films released from the early twentieth century through recent years, reflecting the historical coverage of the IMDb database. IMDb ratings are measured on a 1-10 scale, and vote counts range from thousands to several million observations per title.

5.3 Filtering and Cleaning Criteria

To ensure analytical stability and comparability, several filters were applied:

- Only feature films were retained (excluding episodes, shorts, and other non-feature titles).
- Titles without available audience ratings were excluded.
- Titles with missing production metadata (startYear, runtimeMinutes, or genres) were removed during preprocessing in Python.

IMDb title and rating tables were processed in SQLite to construct a dataset of feature films with available rating information. Because rating information is stored separately from title metadata, the dataset was created by joining the title_basics and title_ratings tables on the shared identifier tconst. The resulting records were exported to Python for additional preprocessing, including logarithmic transformation of vote counts and the construction of genre indicator variables. The final analytical dataset used for regression modeling contains 318,870 film records.

5.4 Feature Construction

Several derived features were constructed for analysis:

- Log-transformed vote count (log_votes) using log(1 + numVotes) to reduce right skew.
- Genre indicator variables were constructed using the first listed genre in the comma-separated genres field. IMDb allows up to three genre labels per title and lists them alphabetically rather than hierarchically. Because the dataset does not designate a primary genre, the first listed genre was used as a consistent encoding rule for regression analysis.
- Release year retained as a continuous control variable to account for temporal growth in participation.

5.5. Data Limitations

The IMDb public dataset provides aggregated engagement metrics only. Voter-level information, including demographic or geographic characteristics, is not available. As a result, engagement signals reflect collective participation without the ability to model individual-level variation.

Additionally, production country, marketing spend, distribution scale, and platform exposure variables are not included in the dataset. These omitted factors likely contribute to unexplained variance in engagement metrics and are acknowledged as structural limitations of the data source.

6. DATA ANALYSIS

6.1 Distributional Characteristics of Key Variables

Before estimating regression models, the distributional properties of key variables were examined. As shown in Figure 1, IMDb average ratings are bounded between 1 and 10 and exhibit an approximately symmetric distribution centered near the mid-range of the scale.

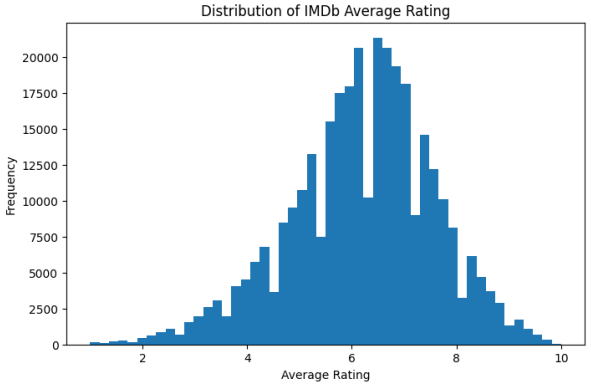

Figure 1: Distribution of IMDb average ratings across all titles in the analytic sample.

As shown in Figure 2, raw vote counts exhibit strong right skew, with a small number of titles receiving extremely high engagement relative to the majority.

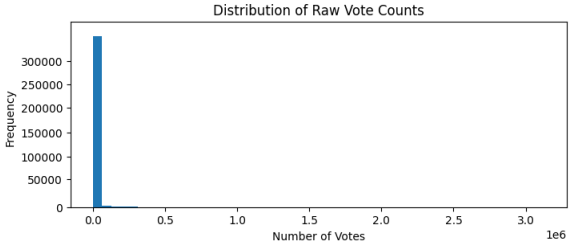

Figure 2: Distribution of raw IMDb vote counts, illustrating pronounced right skew.

The transformation reduces the influence of extreme values and improves adherence to linear modeling assumptions. Using log(1 + numVotes) also allows coefficients to be interpreted approximately as proportional differences in audience engagement.

Figure 3 shows that vote counts were log-transformed prior to regression analysis to stabilize variance and support linear modeling assumptions. This transformation allows coefficients to be interpreted as approximate proportional differences in engagement.

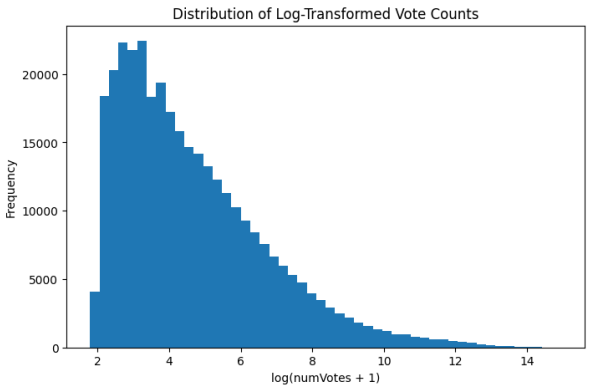

Figure 3: Distribution of log-transformed vote counts used as the regression outcome variable.

6.2 Genre Encoding and Structural Metadata

To avoid perfect multicollinearity among genre indicator variables, one genre, Drama, the most prevalent genre, was omitted from the regression specification. This does not remove Drama films from the analysis; rather, it establishes a reference category within the linear model. Coefficients for other genres therefore represent the marginal association between the presence of that genre and engagement relative to Drama, holding other included genres constant.

In addition to individual genre indicators, a derived variable, genre_count, was considered. This variable captures the number of genre labels assigned to a title. Because IMDb permits up to three genre labels per title, genre_count ranges from one to three and reflects platform classification structure rather than a substantive production attribute. IMDb does not designate a primary genre for multi-genre titles; consequently, all genre indicators are treated symmetrically in the regression specification.

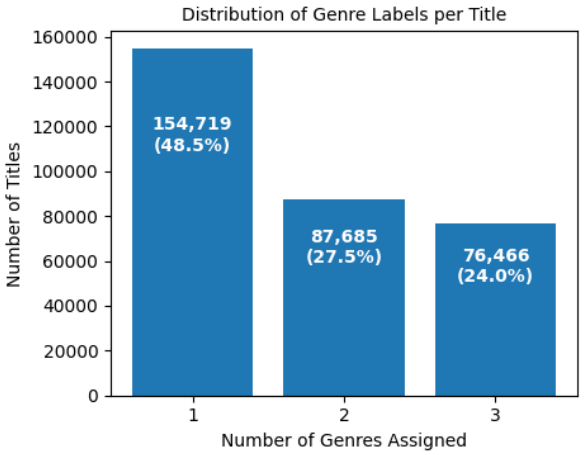

Figure 4: Frequency distribution of the number of genre labels assigned per title (genre_count).

As shown in Figure 4, nearly half of titles receive a single genre label, while the remainder are classified under two or three genres. Multi-genre labeling is therefore common within the dataset. Because broader categorization may influence discoverability or audience reach, this structural feature may introduce systematic variation in observed engagement.

6.3 Model Specification and Estimation

To evaluate whether multi-genre classification alters observed genre associations, two specifications were estimated. The first model included genre_count, capturing potential structural effects associated with multi-genre labeling. However, because genre_count is mechanically derived from the same genre indicators included in the regression, its inclusion alters the interpretation of individual genre coefficients. When genre_count is included, coefficients represent the marginal association of a specific genre conditional on a title having the same number of genre labels. When it is excluded, coefficients reflect the overall association between genre presence and engagement, including any structural effect related to multi-genre classification.

For interpretability and transparency, both specifications are presented and compared.

6.4 Replication Validation Across Analytical Environments

To assess the stability and reproducibility of the engagement model, the primary regression specifications were independently replicated in R using the modeling-ready dataset exported from Python. Both environments implemented identical ordinary least squares specifications,

identical genre indicator encoding, the same omitted reference category (Drama), and the same log-transformed engagement outcome variable.

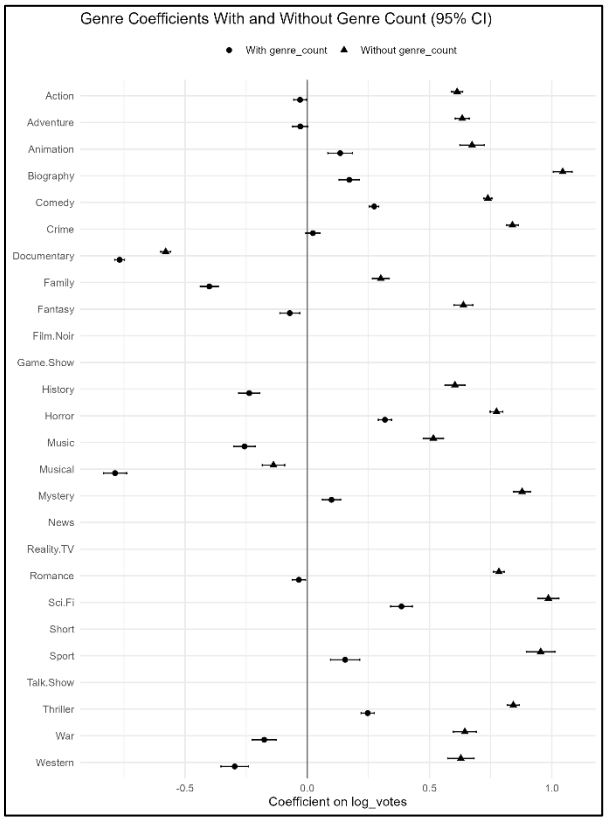

Figure 5: Cross-environment comparison of genre coefficient estimates from Python and R implementations (OLS, 95% confidence intervals).

Figure 5 shows the near-perfect alignment of coefficient estimates across environments, confirming structural equivalence of the design matrix, predictor encoding, and estimation procedure.

The replication procedure followed three validation criteria:

1. Structural equivalence of the model matrix, including identical predictor sets and reference category omission.
2. Consistency of coefficient signs and relative magnitudes across environments.
3. Consistency of model fit statistics and confidence interval estimates.

After correcting the R model specification, the estimates aligned with the Python results. Coefficient signs were identical across all genres, relative ordering of magnitude was preserved, and confidence intervals overlapped precisely. Model fit statistics were also consistent across environments, indicating that the underlying regression specification is stable and not dependent on software-specific implementation.

The replication confirms that the estimated genre associations are not artifacts of a particular analytical environment but reflect structural properties of the dataset and specified model. This cross-platform validation strengthens internal credibility and reduces the likelihood that observed effects arise from coding errors or software-specific defaults.

6.5 Interpretation of Genre Effects Across Specifications

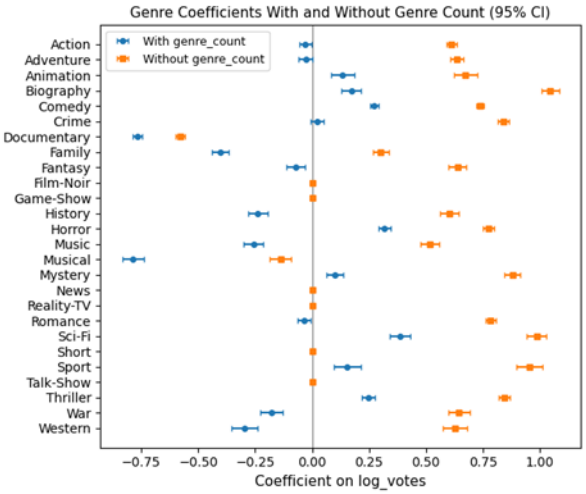

Figure 6: Genre coefficient estimates from models with and without genre_count (OLS, 95% confidence intervals).

Figure 6 compares estimated genre coefficients from the two specifications. Coefficient signs remain consistent across models, indicating that the direction of genre associations is not driven solely by multi-genre classification structure.

Coefficient magnitudes are reduced when genre_count is included, reflecting partial capture of structural variation associated with multi-genre labeling. R^2 increases from 0.145 to 0.169, indicating that classification structure explains a modest additional share of engagement variation. This change does not materially alter the relative ordering or direction of genre effects.

High-engagement genres, including Action, Adventure, Sci-Fi, Thriller, and Biography, exhibit consistently positive associations with log-transformed vote counts, whereas Documentary, Musical, and Family remain negatively associated

relative to the reference category. Although metadata structure influences coefficient magnitudes, genre identity remains the primary organizing dimension of observable engagement.

6.6 Release Period and Engagement

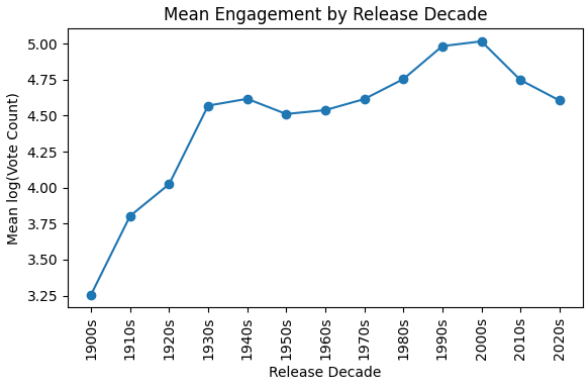

Figure 7: Mean log-transformed vote count by release decade.

As shown in Figure 7, mean engagement rises from the early 1900s through the 1990s and 2000s, reflecting expansion in audience participation and platform adoption rather than inherent differences in quality. Because IMDb voting behavior is shaped by historical and technological context, engagement metrics are not time-neutral. The modest decline in the 2010s and 2020s reflects recency effects, as newer titles have had less time to accumulate votes. These findings indicate that release period constitutes an important contextual dimension when interpreting engagement-based metrics.

6.7 Cast and Crew Participation

To examine whether engagement varies with production scale, the number of credited principal cast and crew members per title was calculated using the normalized title.principals table. This count reflects production scale and relational breadth rather than specific celebrity influence.

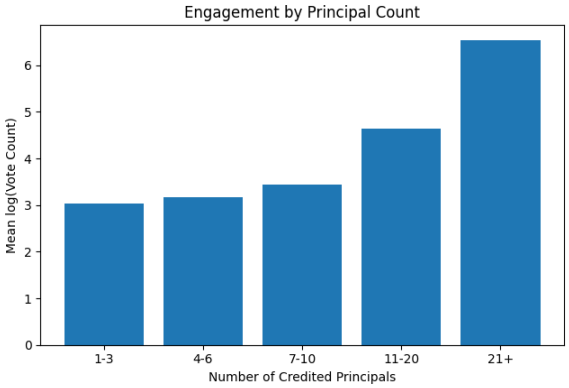

Figure 8: Engagement by Principal Count.

Figure 8 shows that engagement increases steadily across principal-count bins. Titles with 1-3 credited principals exhibit mean log vote counts near 3.0, while titles with 21 or more credited principals exhibit mean log vote counts exceed 6.5. This monotonic pattern indicates that engagement aligns strongly with production scale. Larger productions, which involve more credited personnel, accumulate substantially more votes on average.

This relationship reflects association rather than causation. Principal count likely correlates with production budget, distribution reach, and audience exposure. Nonetheless, the results reinforce the interpretation that IMDb vote counts function as structured engagement signals shaped not only by genre and historical context but also by production characteristics embedded in relational metadata.

6.8 Relationship Between Rating and Engagement

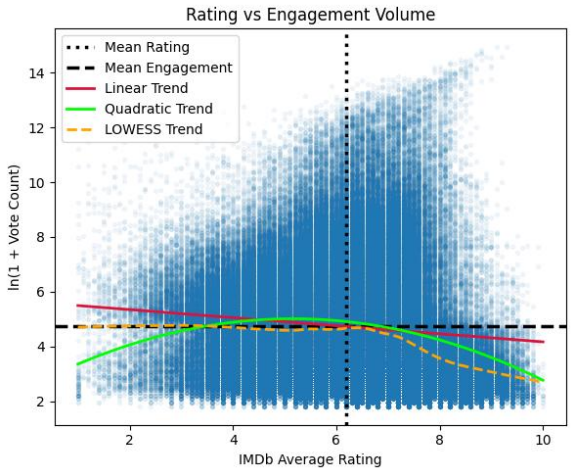

Figure 9: Relationship between IMDb average rating and log-transformed vote count.

Figure 9 shows the relationship between IMDb average rating and log-transformed vote counts. The figure includes reference lines for the mean rating and mean engagement level, along with linear, quadratic, and non-parametric LOWESS trend lines to summarize the overall pattern.

The distribution shows substantial dispersion across the rating spectrum. Titles with moderate ratings exhibit wide variation in engagement volume, and highly rated titles do not uniformly correspond to the largest vote counts. The linear trend indicates only a weak overall association between rating level and engagement volume.

The quadratic trend reveals a general non-linear structure, with engagement increasing into the mid-range of ratings and declining at higher rating levels. However, the LOWESS smoother provides a more flexible, data-driven view of the relationship. It shows that the decline in engagement beyond approximately rating 7 is more pronounced than suggested by the quadratic fit. This indicates that the relationship is not symmetric and that engagement does not continue to increase at the highest rating levels.

This pattern likely reflects participation dynamics rather than evaluative quality. Titles with mid-range ratings tend to attract broader audiences and higher participation, while highly rated titles may reflect more selective or niche engagement. In addition, newer or less widely distributed titles may achieve high ratings with smaller voting populations, contributing to lower observed engagement at the upper end of the rating scale.

Rating level and engagement volume therefore represent related but distinct signals. Average rating reflects evaluative intensity among voters, whereas vote count captures participation scale. The dispersion across the rating spectrum, together with the non-linear pattern identified by both parametric and non-parametric trends, indicates that engagement is influenced by factors beyond rating level alone. These findings support modeling engagement volume directly and reinforce the distinction between quality perception and participation magnitude.

7. FINDINGS

7.1 Genre Identity and Engagement Structure

Across both regression specifications, genre identity exhibits stable associations with engagement volume. Coefficient signs remain consistent regardless of whether multi-genre classification structure is controlled. Genres such as Action, Adventure, Sci-Fi, Thriller, and Biography are positively associated with engagement volume relative to the reference category (Drama), while Documentary, Musical, and Family are negatively associated.

Including genre_count increases model fit modestly, indicating that multi-genre labeling explains additional variation in engagement. However, the direction and relative ordering of genre effects remain largely unchanged. This pattern suggests that while classification structure influences magnitude, genre identity itself remains a primary organizing dimension of observed engagement differences.

These results indicate that engagement variation across genres is not an artifact of multi-label tagging alone but reflects stable structural differences within platform voting behavior.

7.2 Engagement as a Function of Production and Historical Context

Engagement volume varies systematically across release periods and production scale. Descriptive analysis shows that mean log-transformed vote counts increase steadily across decades through the late twentieth century, reflecting growth in audience participation and platform adoption. More recent decades display lower mean engagement, consistent with platform recency effects rather than differences in rating level.

Similarly, engagement increases monotonically with the number of credited principal cast and crew members. Titles involving larger credited teams accumulate substantially more votes on average. This pattern indicates that engagement volume is associated with production scale and exposure, reinforcing the interpretation of vote count as a participation signal rather than a purely evaluative metric.

7.3 Distinction Between Rating Level and Participation Volume

Average rating and vote count are associated but exhibit substantial dispersion across titles. Highly rated titles do not uniformly correspond to the largest vote counts, and moderately rated titles display wide variation in engagement volume.

The relationship is not strictly linear. Engagement tends to increase into the mid-range of ratings and then decline at higher rating levels, indicating that participation peaks among broadly rated titles rather than the highest-rated ones.

This distinction supports treating rating level and participation volume as separate but related signals. Average rating reflects evaluative intensity among voters, whereas vote count captures participation scale. Engagement volume therefore reflects structural and contextual influences beyond perceived quality alone.

7.4 Robustness and Replicability

Independent replication of the regression models in R produced coefficient estimates and fit statistics consistent with the original Python implementation. Cross-environment validation confirmed that the design matrix and predictor encoding produced equivalent estimates.

This replication strengthens confidence that the

findings reflect properties of the dataset and model specification rather than software-specific estimation behavior.

Taken together, these findings indicate that engagement metrics on IMDb reflect a combination of genre identity, production scale, historical participation patterns, and non-linear relationships between rating level and participation rather than a single underlying measure of perceived quality.

8. CONCLUSIONS

This study examined whether observable audience engagement signals can be extracted from large-scale IMDb rating and metadata structures. Rather than treating ratings as direct measures of quality, the analysis evaluated how engagement metrics relate to production characteristics and historical participation patterns.

The findings show that engagement on IMDb reflects multiple structural influences. Genre identity exhibits consistent associations with engagement volume, indicating stable differences in voting behavior across content categories. Production scale, represented by the number of credited contributors, is also associated with higher participation levels. In addition, participation patterns vary across historical periods, reflecting broader changes in platform adoption and audience behavior.

The results further demonstrate that rating level and vote count capture distinct dimensions of audience response. While ratings measure evaluative intensity among participants, vote counts reflect participation scale and exposure. Treating these signals separately provides a clearer interpretation of audience engagement dynamics. The relationship between rating and engagement is not strictly linear, with participation peaking at mid-range ratings rather than at the highest-rated titles.

Overall, the study demonstrates that structured relational metadata can reveal interpretable engagement patterns through transparent statistical analysis, even without complex predictive models. These findings highlight the value of transparent analytical frameworks for examining large media datasets while acknowledging the limitations inherent in aggregated participation data.

9. FUTURE WORK

9.1 Temporal Controls in Multivariate Models

The release-decade analysis demonstrates that engagement is not time-neutral. A natural extension would be to incorporate release period directly into the regression framework, either through decade indicators or continuous year controls. Including temporal controls would allow estimation of genre effects net of historical participation growth and platform adoption effects, strengthening interpretability of structural associations.

9.2 Role-Specific Personnel Effects

The present analysis operationalizes production scale using total principal count. Future work could disaggregate personnel roles by category, such as actor, director, writer, or producer, to examine whether specific role types exhibit distinct associations with engagement volume. This would allow more granular interpretation of how production structure relates to audience participation.

9.3 Interaction Effects Between Genre and Time

Genre popularity may evolve across historical periods. Extending the model to include interaction terms between genre indicators and release period could clarify whether genre engagement patterns are stable or historically contingent. This would deepen understanding of how structural metadata interacts with temporal context.

9.4 Distributional Modeling Beyond OLS

Although log transformation improves adherence to linear modeling assumptions, engagement volume remains highly skewed. Future work could evaluate alternative modeling approaches, such as generalized linear models for count data or quantile regression, to assess whether structural associations differ across engagement levels.

9.5 Incorporation of Review Text or External Signals

Prior literature suggests that numeric ratings capture only part of audience response. Integrating textual review data or external performance indicators could provide a richer representation of engagement and enable comparison between quantitative and qualitative signals.

10. LIMITATIONS

Several limitations should be considered when interpreting the results of this study. The analysis

relies on aggregated engagement metrics and observable metadata from the IMDb public dataset. As a result, the models capture structural associations rather than causal relationships between production characteristics and audience participation.

The regression framework also incorporates a limited set of explanatory variables. Important factors such as marketing exposure, distribution scale, production budget, and platform promotion are not available in the dataset and therefore cannot be included in the model. These unobserved influences likely contribute to the unexplained variance in engagement outcomes.

In addition, IMDb participation reflects voluntary user behavior rather than a representative sample of the broader viewing population. Voting patterns may therefore reflect platform-specific participation dynamics, community preferences, and historical adoption patterns rather than universal audience responses.

These limitations do not invalidate the observed structural patterns but highlight that engagement metrics should be interpreted as indicators of platform participation rather than direct measures of content quality or audience preference.

11. REFERENCES

Reference 1

Baltrunas, L., & Ricci, F. (2009). Context-based splitting of item ratings in collaborative filtering. *Proceedings of the Third ACM Conference on Recommender Systems*, 245-248. https://doi.org/10.1145/1639714.1639759

Summary
The research shows that specific attributes (like release seasons or demographics) can separate item ratings into homogeneous groups to improve prediction accuracy by analyzing a pre-filtering technique called item splitting, dividing item ratings based on contextual conditions.

Value
This study provides a methodology for the project to analyze how production attributes act as contexts that split or define audience engagement signals.

Reference 2

Cheng, Z., Ding, Y., Zhu, L., & Kankanhalli, M. (2018). Aspect-aware latent factor model: Rating prediction with ratings and reviews. *Proceedings of the 2018 World Wide Web Conference*, 639-648. https://doi.org/10.1145/3178876.3186145

Summary
This paper is a useful example of how rating data can be treated as more than a single summary measure. The work illustrates an alternative way of organizing rating information that makes it easier to reason about how different item characteristics may influence audience responses.

Value
This paper supports my approach of examining ratings as composite signals that can be analyzed to better understand audience response, instead of relying on a single numeric score.

Reference 3

IMDb. (n.d.). *IMDb non-commercial datasets*. https://datasets.imdbws.com/

Summary
IMDb provides publicly available tab-separated datasets containing structured information on titles, ratings, genres, personnel, and related metadata. The datasets are normalized across multiple relational tables and are updated regularly for non-commercial research use.

Value
This source provides the primary data used in the project. It establishes the structure, scope, and limitations of the IMDb engagement metrics analyzed in this study, including aggregated ratings and vote counts without voter-level detail.

Reference 4

Krohn-Grimberghe, A., Nanopoulos, A., & Schmidt-Thieme, L. (2010). Integrating OLAP and recommender systems: An evaluation perspective. *Proceedings of the ACM 13th International Workshop on Data Warehousing and OLAP*, 85-92. https://doi.org/10.1145/1871940.1871959

Summary
This paper proposes a multidimensional framework that integrates relational data warehouses and Online Analytical Processing (OLAP) with recommender systems. It details a three-tier architecture, Source, Warehouse, and OLAP, that allows analysts to perform ad-hoc queries, such as drill-downs and slice/dice operations, to evaluate performance against business objectives.

Value
This reference informs my SQL-based relational workflow by demonstrating how OLAP-style slicing and drill-down operations can be applied to recommender evaluation, supporting structured analysis of engagement signals across multiple dimensions.

Reference 5

Ling, G., Lyu, M. R., & King, I. (2014). Ratings meet reviews: A combined approach to recommend. *Proceedings of the 8th ACM Conference on Recommender Systems*, 105-112. https://doi.org/10.1145/2645710.2645728

Summary
The authors propose a unified model that aligns latent topic spaces with rating dimensions to alleviate the cold-start problem. The model is able to learn interpretable topics that link items to prior knowledge, such as famous directors or specific production roles.

Value
This reference adds value by supporting the feature engineering of cast and crew roles and examining their relationship to observable audience ratings.

Reference 6

Liu, J., Li, T., Yu, M., Yang, S., Tang, Z., & Yang, Z. (2025). A multi-factor collaborative prediction for review-based recommendation. *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. https://doi.org/10.1145/3705328.3748062

Summary
This research examines the complex relationship between different feedback factors, specifically mining the correlation between click behaviors and rating behaviors. It proposes a multi-factor collaborative prediction method that extracts click factors and rating factors from user reviews to improve recommendation accuracy.

Value
This paper informs the project's framing of audience engagement signals by illustrating how different forms of user feedback represent distinct dimensions of engagement. While IMDb does not provide clickstream data, the distinction between participation volume (vote counts) and evaluation intensity (average ratings) supports treating these measures as related but non-equivalent signals.

Reference 7

Liu, Y., Cao, X., & Yu, Y. (2016). Are you influenced by others when rating? Improve rating prediction by conformity modeling. *Proceedings of the 10th ACM Conference on Recommender Systems*, 269-272. https://doi.org/10.1145/2959100.2959141

Summary
The authors model user conformity on online rating sites, analyzing how public opinions, highlighted via average ratings and distributions, implicitly contribute to ongoing user behavior.

Value
This supports background and motivation regarding how observed ratings signals emerge from social interaction rather than just item quality.

Reference 8

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. *Proceedings of the 7th ACM Conference on Recommender Systems*, 165-172. https://doi.org/10.1145/2507157.2507163

Summary
This study combines latent rating dimensions with latent review topics learned through models like LDA (Latent Dirichlet Allocation; A topic modeling algorithm that discovers abstract, underlying "topics" in a collection of documents by treating each document as a mix of topics and each topic as a distribution of words) to provide textual labels for rating dimensions. The model improves rating prediction and facilitates automated genre discovery by harnessing information present in the review text of even a single review.

Value
It provides a methodological precedent for using review text to surface interpretable dimensions that help explain observed audience ratings.

Reference 9

Steck, H. (2013). Evaluation of recommendations: Rating-prediction and ranking. *Proceedings of the 7th ACM Conference on Recommender Systems*, 213-220. https://doi.org/10.1145/2507157.2507160

Summary
Steck examines the differences between rating-

prediction accuracy and ranking accuracy. The research identifies that predicting observed ratings only solves a small part of the real-world problem due to selection bias in the data combined with data sparsity.

Value
This reference is critical for explaining the limitations of engagement signals and identifying where observed IMDb ratings may be biased or limited.

Reference 10

Wang, X., Zhang, R., Sun, Y., & Qi, J. (2021). Combating selection biases in recommender systems with a few unbiased ratings. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 427-435. https://doi.org/10.1145/3437963.3441799

Summary
This research identifies that recommendation datasets are prone to selection biases due to user self-selection and system item selection. It proposes a novel objective using a small set of unbiased data to improve the training of systems on biased data.

Value
It helps the project explain limitations and trade-offs when working with real-world audience signals that are not collected in a controlled environment.

Reference 11

Weinsberg, U., Bhagat, S., Ioannidis, S., & Taft, N. (2012). BlurMe: Inferring and obfuscating user gender based on ratings. *Proceedings of the Sixth ACM Conference on Recommender Systems*, 195-202. https://doi.org/10.1145/2365952.2365989

Summary
The authors demonstrate that a user's demographic features, such as gender, can be inferred with high accuracy based solely on the items they have rated. The paper identifies specific movies for which the rating given is strongly correlated with categorical metadata.

Value
It establishes that item attributes can be strongly correlated with rating behavior, supporting my analysis of how observable title attributes align with audience engagement patterns, while also highlighting ethical and inferential limits.

Reference 12

Wu, C.-Y., Beutel, A., Ahmed, A., & Smola, A. J. (2016). Explaining reviews and ratings with PACO: Poisson additive co-clustering. *Proceedings of the 25th International Conference Companion on World Wide Web*, 127-128. https://doi.org/10.1145/2872518.2889400

Summary
The authors introduce the PACO model to jointly analyze ratings and review texts through additive co-clustering. Utilizing IMDb data, the research identifies clusters of items and associated audience engagement topics (e.g., "plot," "acting," "worst") to provide interpretable explanations for numeric ratings.

Value
This paper directly matches the IMDb data type and supports the project's goal of identifying patterns and relationships between production attributes and audience engagement signals.

Reference 13

Zhao, X., Zhu, Z., Zhang, Y., & Caverlee, J. (2020). Improving the estimation of tail ratings in recommender systems with multi-latent representations. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 762-770. https://doi.org/10.1145/3336191.3371810

Summary
This paper addresses large estimation errors in tail ratings (ratings far from the mean) caused by the uni-modal assumptions of popular models. By proposing multi-latent representations, the authors significantly improve predictions for the most polarized items.

Value
This adds value by helping to interpret items with polarized rating distributions, common in IMDb audience measures.