WebScrape Insights: A consistency analysis report of movie ratings Based on Metacritic

Project Name & Team Members

Project Name: WebScrape Insights
Team Member: Chuyi Wang
USC Email: chuyi@usc.edu
USC ID: 5458365139

Short Description

This project focuses on the Metacritic movie platform and systematically collects movie rating data through web scraping techniques to analyze the consistency between professional critic scores and general user scores. The primary objective is to examine the degree of agreement between expert-based evaluation systems and crowd-sourced audience evaluations, and to further investigate whether movie characteristics such as genre, production company scale, and release era influence rating discrepancies. Through data cleaning, statistical analysis, and multiple visualization methods, this study establishes a complete data analysis pipeline and reveals structural differences between the two rating systems as well as their potential underlying causes.

Data

Data Sources

All data used in this project were collected from publicly available movie pages on the Metacritic website. The data were obtained through a self-developed Python web scraper without relying on any official APIs. The scraper was implemented using Requests and BeautifulSoup to parse both movie listing pages and individual movie detail pages. The collected information includes movie titles, release years, genres, production companies, as well as critic and user rating scores and review statistics. To reduce the risk of being flagged as abnormal traffic, the scraping process incorporated random request delays, User-Agent spoofing, and a resume-from-checkpoint mechanism, ensuring stable and responsible data collection.

During project implementation, a large-scale feasibility and stability test was conducted by setting the scraping target to 200 movies. The results show that complete data were successfully retrieved for 197 movies, while only 3 movies could not be fully parsed due to abnormal page structures or missing key fields, resulting in an overall success rate of 98.5%. No access restrictions or bans were triggered during the entire process, indicating that the adopted anti-scraping strategies were effective under the current access frequency.

Number of Data Samples

The large-scale scraping stage primarily served to validate the robustness and scalability of the data collection system. For the final statistical analysis, a curated dataset was selected based on data quality and reproducibility requirements. The final dataset consists of 60 movies with fully cleaned, consistent fields and no missing values. These movies span multiple release eras, cover 23 genres, and involve 133 production companies, providing sufficient diversity to support rating consistency analysis and multi-dimensional pattern exploration while ensuring stable and repeatable analysis workflows.

Data Cleaning, Analysis & Visualization

Data Cleaning

After data collection, a systematic data cleaning and preprocessing process was applied. First, all fields were checked for completeness, and a small number of movies with missing fields, such as writers or review components, were identified. For these cases, the scraper attempted to re-fetch the corresponding pages to supplement missing information. Duplicate records were then removed, and all data were converted into a unified structured tabular format.
After cleaning, the final dataset contained 60 observations and 30 variables, with no remaining missing values, corresponding to an overall missing rate of 0%. This result was further validated using a missing data heatmap. Since Metacritic uses a 0–100 scale for critic scores and a 0–10 scale for user scores, critic scores were normalized to a 0–10 scale to ensure comparability between the two rating systems. Data types were standardized, and basic outlier checks were performed to prepare the dataset for analysis.

Data Analysis

The analysis began with descriptive statistics of critic and user ratings. The results indicate that critic scores are highly concentrated, with an average score of 9.86, a

minimum of 9.7, and a maximum of 10.0, reflecting strong internal consistency within the critic rating system. In contrast, user scores exhibit a wider distribution, with an average score of 7.68, a minimum of 5.0, and a maximum of 9.0. This difference suggests that critics and general audiences apply different evaluation standards and rating scales.

To quantitatively assess rating consistency, Pearson correlation analysis was conducted between critic and user scores. The resulting Pearson correlation coefficient was −0.098, indicating an extremely weak negative linear relationship. In addition, score difference analysis revealed that critic scores are systematically higher than user scores, with an average difference of 2.17 points.

Data Visualization

To visually illustrate the relationship between critic and user ratings, a series of visualizations was constructed around two core dimensions: score correlation and score difference. Scatter plots comparing normalized critic scores and user scores show that most data points deviate substantially from the ideal diagonal line of perfect agreement, and the fitted trend line has a low slope, further confirming the lack of strong linear consistency.

Histograms of score differences, defined as "critic score minus user score," reveal that the distribution is clearly shifted toward positive values, with an average difference of approximately 2 points. This indicates that rating disagreement is directional rather than random, with critics consistently assigning higher scores.

Grouped analyses further illustrate structural patterns in rating disagreement. Temporal trend plots show that critic scores remain relatively stable across different eras, while user scores tend to decline in more recent periods. Genre-based comparisons indicate larger discrepancies for science fiction, documentary, and history films, while comedies and musicals exhibit relatively smaller gaps. Outlier analyses highlight individual movies with particularly large discrepancies, demonstrating that disagreement is also pronounced at the case level. Finally, correlation heatmaps show strong internal correlations within critic-related variables and within user-related variables, but weak correlations across the two systems, suggesting that critic and user rating mechanisms operate largely independently.

Hypothesis / Premise & Conclusions

The core premise of this project is that critic ratings and user ratings should exhibit a certain degree of consistency when evaluating movie quality. To test this premise, rating consistency was explicitly operationalized using two complementary measures. The first measure is correlation analysis, including both Pearson correlation (to assess linear relationships) and Spearman rank correlation (to assess monotonic relationships), which captures the extent to which the two rating systems move together. The second measure is score difference analysis, defined as the mean and

dispersion of the difference between normalized critic scores and user scores, which captures systematic deviations between the two systems.
In this study, rating systems were considered highly consistent only if the correlation coefficient exceeded 0.7 and the average score difference was less than 1 point. The empirical results show that the correlation between critic and user scores is approximately 0.15 and that the average score difference is about 2.07 points, which clearly fails to meet the consistency criteria. Therefore, the findings do not support the hypothesis that critic and user ratings are highly consistent. Further analyses indicate that this inconsistency persists across genres, production backgrounds, and release eras, suggesting that professional critics and general audiences rely on fundamentally different evaluation dimensions and value frameworks when assessing films.

Changes from Original Proposal

Compared to the original project proposal, this study expanded both the scale of data collection testing and the scope of analysis. While the proposal primarily focused on rating consistency, the final implementation additionally validated scraper performance at scale and incorporated outlier analysis, internal consistency analysis, and grouped analyses by genre, production company, and release era. These extensions allowed for a more comprehensive and nuanced interpretation of rating discrepancies.

Future Work

With additional time and resources, future work could further expand the dataset to include more movies, thereby improving the robustness of statistical conclusions. Incorporating sentiment analysis of textual reviews would provide content-level explanations for rating differences. In addition, comparing Metacritic ratings with those from other platforms such as IMDb or Rotten Tomatoes could offer broader insights into how different rating systems shape audience and critic evaluations.