Rating Patterns in IMDb Top 250 Films:
A Descriptive Analysis Across Time, Genre, and Runtime

Part I : Introduction

Team members

Ester Hu (jhu05525@usc.edu)
Qi Shen (qshen840@usc.edu)

Research Question

What patterns can be observed in audience ratings of IMDb Top 250 films across time, genre, and runtime, and how do these film attributes relate to higher or lower IMDb ratings?

We focus on three closely related sub-questions:

1. Genre & Rating: Do certain film genres consistently receive higher average IMDb ratings than others within the Top 250 list?
2. Runtime & Rating: Is there a relationship between a film's runtime and its IMDb rating? In other words, do longer films tend to be rated more highly, or is rating largely independent of film length?
3. Temporal Patterns: How do IMDb ratings vary across release years and decades? Are films from certain eras systematically rated higher than those from others?

Short Description

Rather than treating movie ratings as purely subjective opinions, this project approaches IMDb ratings as aggregated signals of audience evaluation and examines how they relate to observable film attributes. Using data from the IMDb Top 250 films, we conduct a structured exploratory analysis to identify rating patterns across genre, runtime, and release period.
The IMDb Top 250 list provides a meaningful dataset: all films included are already highly rated, allowing us to focus on variation within excellence rather than broad success versus failure. By analyzing cleaned, movie-level and genre-level datasets, we aim to uncover whether certain genres dominate the upper end of ratings, whether longer films are systematically rewarded with higher scores, and how historical context (captured through release year and decade) shapes rating distributions. Overall, this project emphasizes descriptive data analysis and visualization to support comparative reasoning, helping us move from individual movie opinions toward broader insights about rating patterns in highly acclaimed films.

Part II : Data Collection

Approach & Obstacles

In order to analyze the patterns in IMDb Top250 films through time, genre and runtime, it's necessary to collect at least these three types of information for all 250 films. Although this information is available on the webpage, it is scattered across different locations. Some details, such as film genres, require clicking on links to view. This undoubtedly makes manually reviewing and collecting the data highly impractical. The optimal approach is to utilize scraping to gather information for each of the 250 films and consolidate all data into a uniformly formatted spreadsheet for convenient subsequent analysis.

Having clarified this fundamental objective, in the beginning we attempted to scrape the IMDb Top 250 page by directly parsing the HTML structure with BeautifulSoup. This method worked well at first, but we quickly discovered that only 25 movie entries were being loaded in the static HTML. The rest of the movies were rendered dynamically by IMDb's front-end JavaScript framework, which means they do not appear as normal HTML tags when the page is fetched through Python. Because of this, a traditional HTML parsing approach was not able to capture the full Top 250 list.

To solve this limitation, we examined the page more closely and luckily discovered that IMDb embeds a complete JSON object inside a <script type= "application/ld+json"> tag. This JSON block contains structured metadata for all 250 movies, including each item's title, IMDb ID, ranking order, genre, rating, and a link to its dedicated IMDb page. Once we identified this script tag, we extracted the JSON text and loaded it with Python's JSON library. This allowed us to retrieve clean and complete metadata without relying on dynamically rendered HTML.

While the JSON script provided most of the information needed for analysis, it did not include two important factors: the release year and the runtime. To obtain these missing variables, we used the URL already included in each JSON entry and performed a second round of scraping. We looped through all 250 URLs and sent individual HTTP requests to each movie's detailed page. After retrieving the HTML for each film, we parsed it with BeautifulSoup and searched for the specific tags corresponding to the release year and duration. Because IMDb's structure is not uniform across films, we used flexible tag-matching logic instead of relying on fixed positions.

By combining JSON-based extraction for the main list with page-level scraping for additional attributes, we were able to construct a complete dataset with all variables required for further analysis. This two-stage collection process ensured both completeness and accuracy, overcoming limitations of dynamic HTML rendering while still allowing access to detailed metadata for each movie.

Changes & Challenges

Compared to our original project proposal, the most significant change in our analysis plan involved the treatment of popularity-related metrics, particularly vote counts. In the initial proposal, we intended to examine how IMDb ratings relate not only to film attributes such as genre and runtime, but also to popularity indicators, including the number of user votes. However, during data collection and preparation, we found that the scraped IMDb Top 250 dataset did not consistently include vote count information. As a result, incorporating popularity-based analysis would have required either additional data sources or assumptions that could not be reliably validated within the scope of this project. Therefore, we refocus the project on rating-based patterns alone. This shift led us to reformulate the research question to emphasize comparative analysis across time, genre, and runtime within a curated set of highly rated films. By narrowing the scope in this way, we were able to ensure that all analyses were based on consistent and fully observed variables, strengthening the internal validity of the project.

Another challenge encountered during data preparation involved data completeness and formatting. Although the original dataset contained 250 films from the IMDb Top 250 list, one film was excluded during the cleaning process because its runtime information could not be reliably parsed into a numeric format. This resulted in a final movie-level dataset of 249 observations. All subsequent analyses and visualizations were conducted using this consistent subset, and the exclusion was documented to maintain transparency.

Overall, these adjustments led to a more focused and coherent analysis pipeline. By prioritizing data integrity and analytical clarity over expanding the number of variables, the revised project design allowed us to produce interpretable results that align closely with the available data and the goals of descriptive data analysis.

Part III : Data Analysis

Analysis techniques

We conducted a structured descriptive and group-based analysis on the IMDb Top 250 dataset after cleaning and feature engineering. Our workflow combined:

a. Data cleaning & type standardization: we validated numeric fields (e.g., year, rating) and converted runtime strings (e.g., “2h 22m”) into a numeric variable runtime_min for quantitative comparison. We also created a derived time bucket, decade, by mapping each film’s release year to its corresponding decade (e.g., 1994 to 1990s).

b. Aggregation (groupby) and summary statistics: we used decade- and genre-level aggregations to compute: Representation counts by decade (how many films per decade appear in Top 250); Average rating by decade; Average rating by genre (with genre-level data produced by exploding multi-genre labels so that each genre entry is analyzed consistently).

c. Association analysis (non-causal): we examined the relationship between runtime and rating within each decade using correlation summaries. This was treated as an exploratory measure of association rather than evidence of causality.

Findings

Recent decades are more heavily represented in the IMDb Top 250. Films from the 2000s and 2010s account for the largest share of the Top 250, reflecting the composition of the ranking rather than an inherent superiority of recent films. This pattern is likely influenced by ongoing voting dynamics and larger contemporary audiences on IMDb.

Average ratings across decades show limited variation within the Top 250.Although small differences in decade-level average ratings exist, they are modest and sensitive to sample size. Overall, films across different eras cluster within a similarly high rating range, reinforcing that the Top 250 represents an elite group rather than a decade-based hierarchy.

Genre and runtime reveal descriptive patterns rather than deterministic effects.Certain genres appear more frequently and with slightly higher average ratings in the Top 250, while the relationship between runtime and rating is weak and inconsistent across decades. These results suggest that genre and runtime function as contextual attributes within highly rated films, not as direct drivers of rating outcomes.

Data Visualization (figures shown in appendix)

Figure 1: Average Rating Trend by Decade

This bar chart visualizes how the average IMDb rating of Top 250 movies changes across decades. The x-axis represents each decade, and the y-axis shows the average rating of movies in the list released in that decade. Each bar corresponds to one decade, and its height reflects the average rating for that period. I also narrow the y-axis range (around 8–8.5) to make small differences easier to see and add a simple trend line to represent general trends. This figure provides an overview of how overall film quality varies over time.

Figure 2: Number of Top Movies by Decade

This bar chart shows how many Top 250 films were released in each decade. The x-axis lists different decades, and the y-axis indicates the count of movies. Each bar represents one decade, and its height reflects how many films from that decade belong to the Top 250. This figure reveals the difference in time periods that contributed the movies to the list.

Figure 3: Average Rating by Genre

This bar chart compares the average IMDb rating across different movie genres. The x-axis lists genres, and the y-axis shows the average rating of all Top 250 films belonging to that category. Bars are sorted from highest to lowest rating to highlight which genres tend to receive stronger audience approval. I also narrow the y-axis range (around 8–8.5) to make small differences more visible. The chart summarizes differences in quality on genre-level.

Figure 4: Runtime–Rating Correlation by Decade

This bar chart presents the correlation between movie runtime and IMDb rating across decades. The x-axis shows each decade, and the y-axis shows the correlation coefficient ranging from –1 to 1. I add a horizontal line at 0 to help distinguish positive from negative relationships. Each bar indicates whether longer films in that decade tended to receive slightly higher, lower, or unrelated ratings. This figure does not claim that "longer movies are better," but instead illustrates how the relationship between runtime and rating varies across decades.

Observations and Conclusion

First, we observe that the distribution of films across decades is uneven. Certain decades, particularly the 2000s and 2010s, are more heavily represented in the Top 250 than earlier periods. This suggests that Top 250 membership reflects not only rating levels but also broader contextual factors such as audience exposure, contemporary viewing habits, and the dynamic nature of IMDb rankings. At the same time, when examining average ratings by decade, differences are relatively small. This indicates that the Top 250 functions as a high-rating tier in which films from different eras tend to cluster within a similar rating range, rather than showing strong decade-based separation.

Second, at the genre level, we observe clearer differentiation. Some genres appear more frequently in the Top 250 and also exhibit slightly higher average ratings within this curated set. This pattern suggests that certain genres are more likely to be consistently recognized as "top-rated," possibly due to long-standing narrative traditions, critical reception, or audience expectations associated with those genres. However, because many films belong to multiple genres, these findings should be interpreted as descriptive patterns of representation and association rather than as evidence that genre alone determines a film's rating.

Third, regarding runtime, our observations indicate that the relationship between movie length and rating is neither strong nor uniform. The direction and magnitude of the runtime–rating association vary across decades, and no single trend persists over time. This suggests that longer runtimes do not systematically correspond to higher ratings within the Top 250. Instead, runtime appears to interact with historical context and filmmaking conventions, reinforcing the idea that film length should be understood as a contextual attribute rather than a direct indicator of audience evaluation.

Impact

The impact of this analysis is that it provides a clearer and more structured understanding of how IMDb ratings behave within the Top 250 list. By examining ratings across decades, genres, and runtime, the project shows that while Top 250 films are unevenly distributed across time, their average ratings remain relatively stable, indicating that the list represents a consistently high-rating group rather than a decade-driven hierarchy. The genre-level results further clarify which types of films are more frequently represented among highly rated titles, helping distinguish patterns of representation from assumptions about overall film quality. In addition, the weak and inconsistent relationship between runtime and rating demonstrates that commonly assumed indicators, such as film length, do not reliably explain rating differences within this elite set. Overall, this project contributes to a more informed and critical understanding of how ratings and rankings function in contemporary film culture.

Part IV : Future Work

Given more time, several extensions could meaningfully strengthen this project. First, incorporating additional data metrics—such as online view counts and box office performance—would allow a deeper level of analysis: how popularity measures relate to ratings. Meanwhile, given our current dataset focuses purely on attributes available from the Top 250 list, we could access more data from distinctive sources to make it possible to compare these highly rated films to the broader universe of movies and assess whether the patterns observed here are unique to the list or generalizable.

Another improvement direction would be applying more advanced techniques to data analysis and visualization processes. For example, if we collect data from more dimensions, we can introduce more diverse chart representations, such as bubble plots to show relationships between three factors. Additionally, some factors such as runtime, may be grouped into categorical bins (short, medium, long films) rather than being treated independently. It may help us to build a more comprehensive understanding of the factors associated with high ratings.

Appendix.

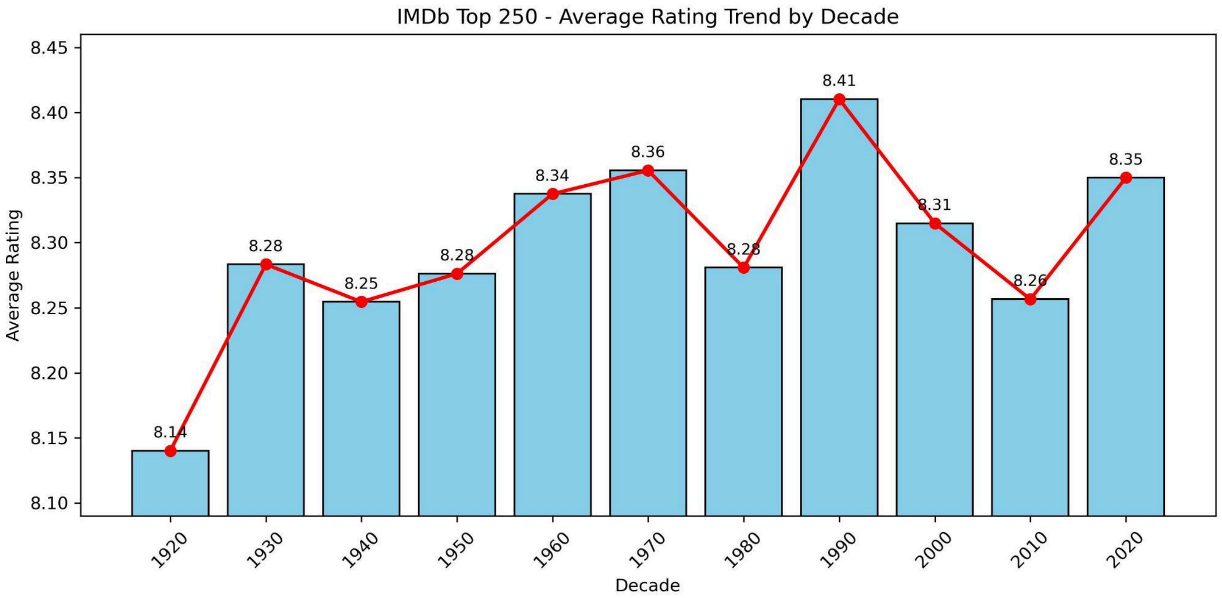

Figure.1

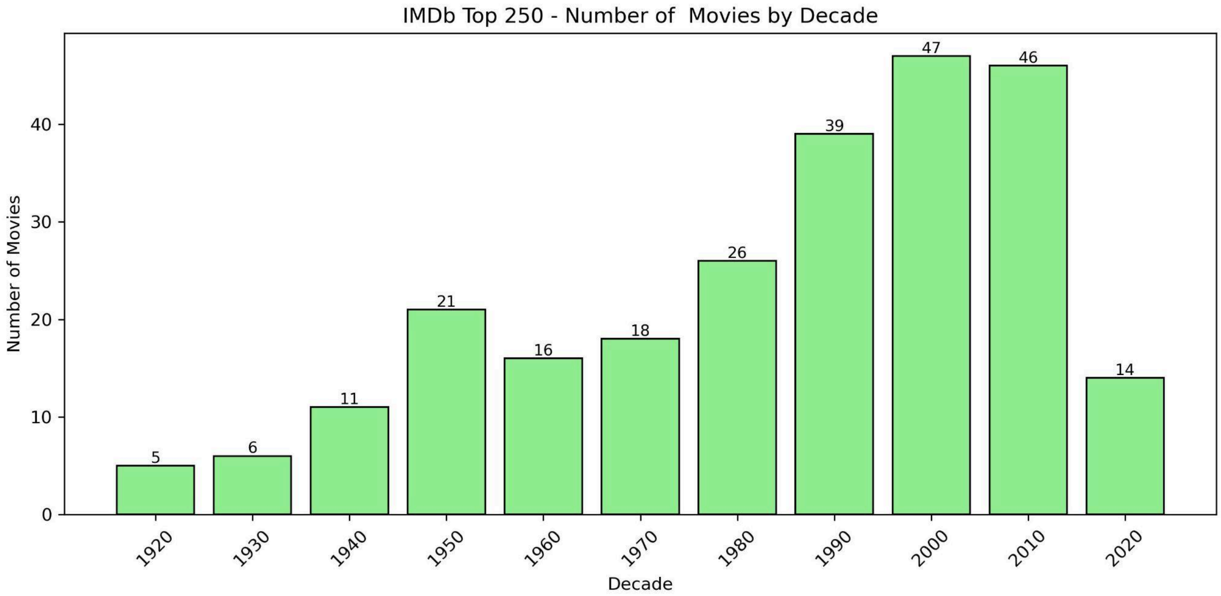

Figure.2

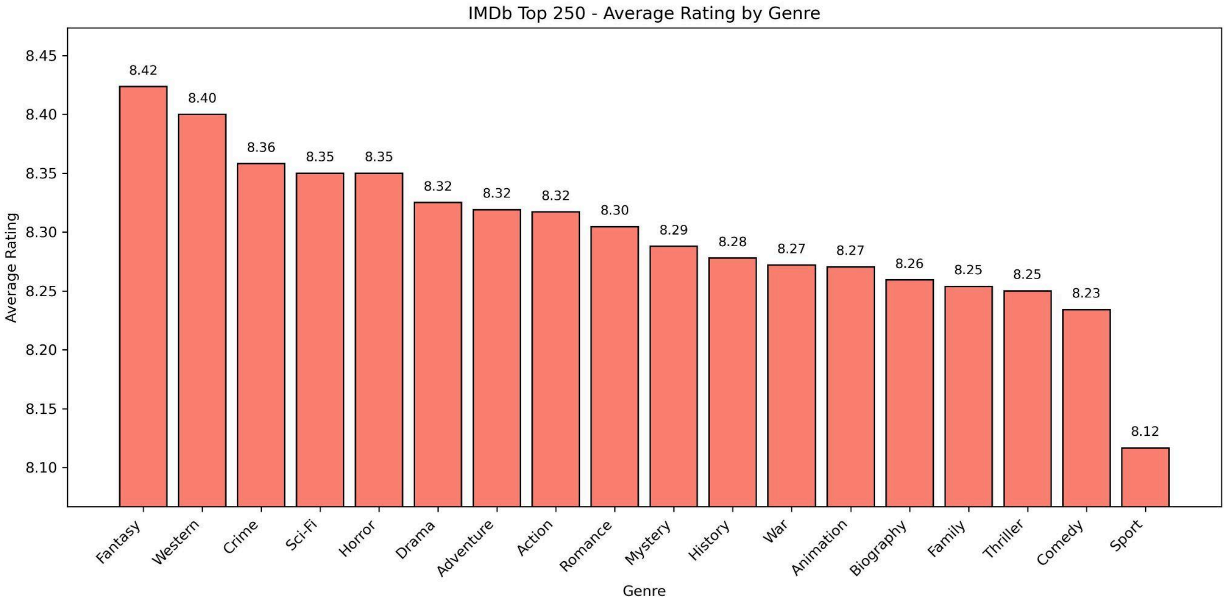

Figure. 3

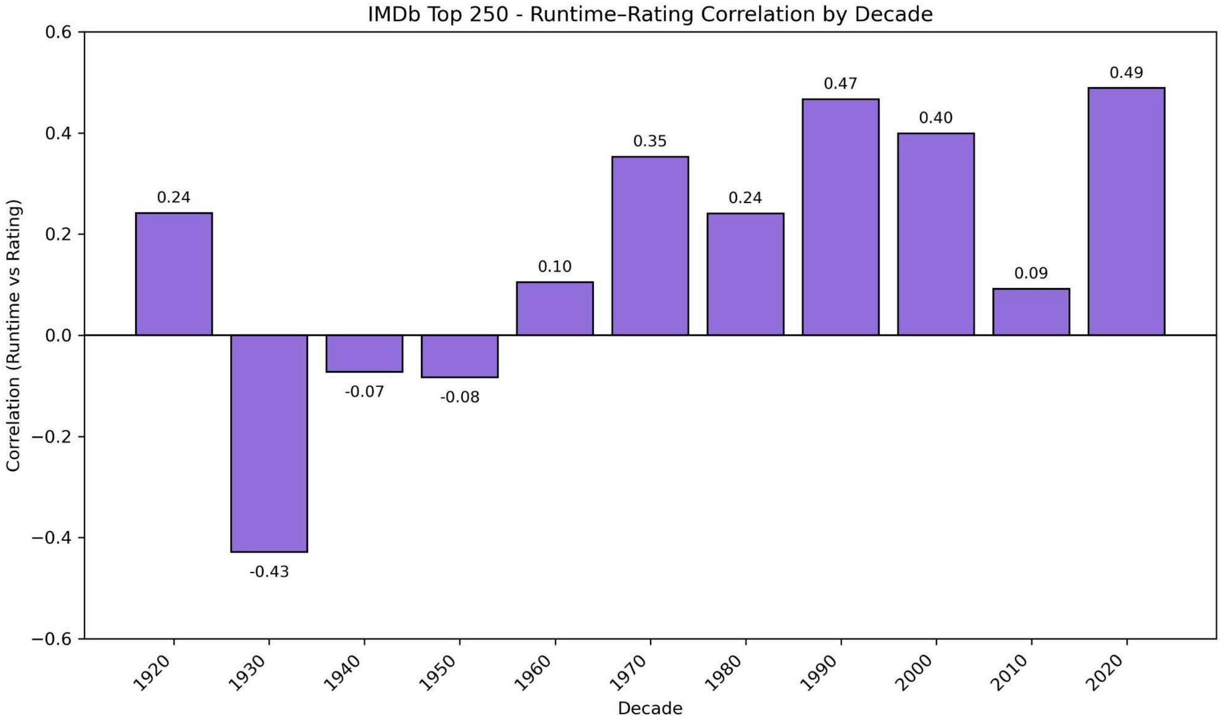

Figure.4