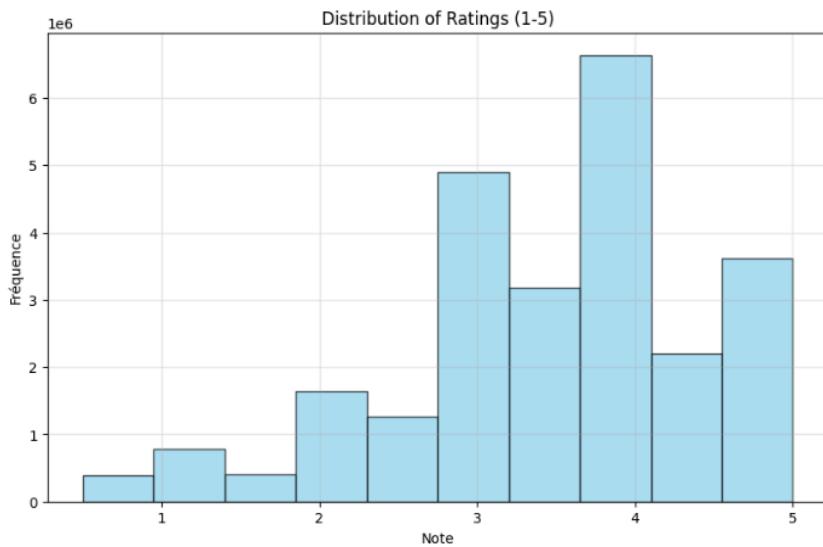


# MovieLens - ANALYTICS

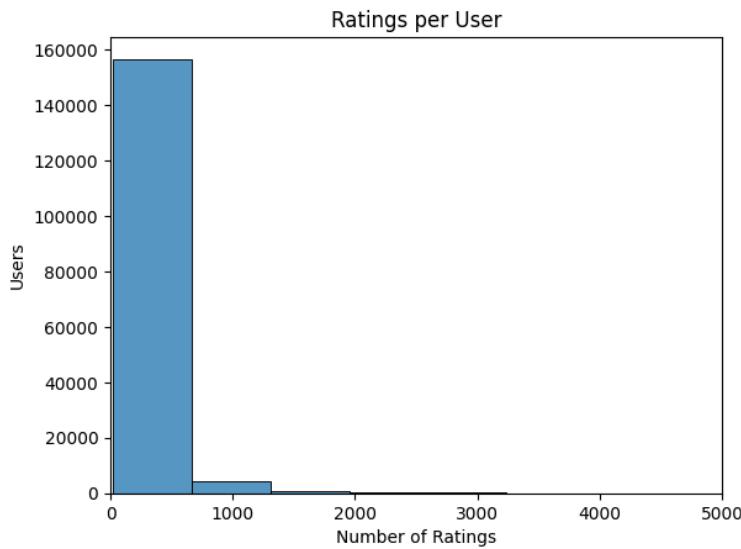
## 1. Distribution of rating

- a. **Ratings are skewed toward higher values** – Most ratings fall between **3 and 5**, especially around **4**.
- b. The model/product being evaluated is **well-received overall**, as most ratings are concentrated between **3 and 5**, especially around **4**.
- c. The **low number of poor ratings (below 2)** indicates **high user satisfaction and reliability**.
- d. The presence of some **mid-range ratings (~3)** suggests **scope for improvement** despite generally positive feedback.



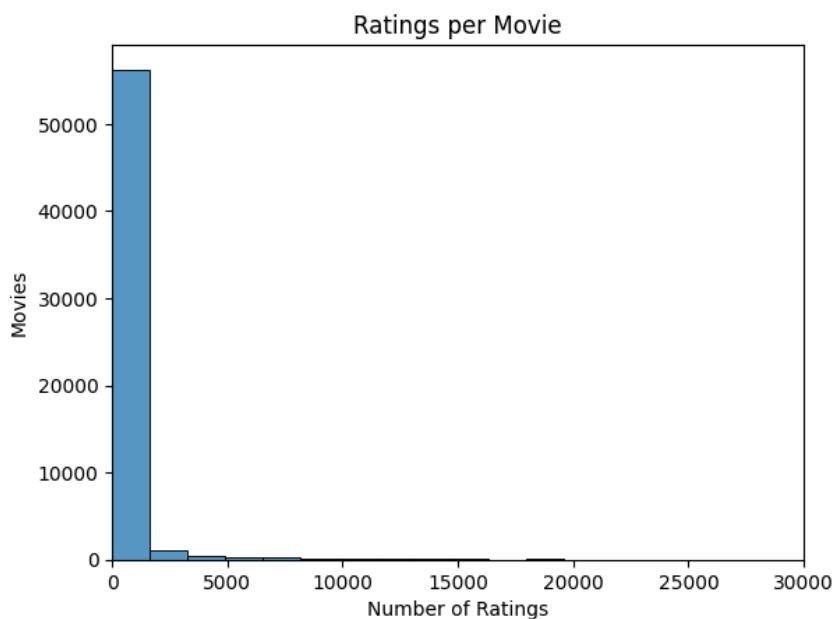
## 2. Number of Ratings per User

- a. **Most users are light users**, giving only a small number of ratings, as shown by the heavy concentration near the lower end.
- b. There are **very few highly active users** with a large number of ratings, forming a long right tail (outliers).
- c. The data is **highly right-skewed**, indicating an imbalanced contribution where a small group of users generates a large portion of the ratings.

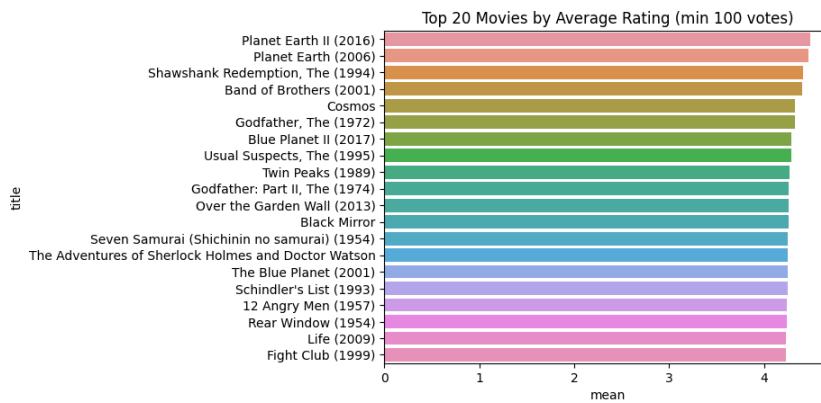


### 3. Number of Ratings per Movie

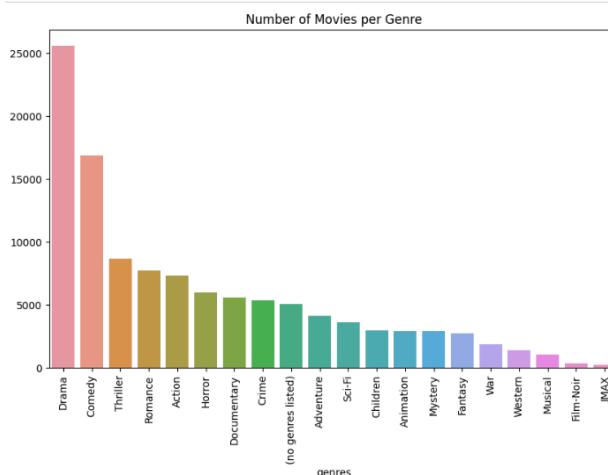
- Most movies receive very few ratings**, indicating that a large portion of the catalog is sparsely rated.
- A small number of popular movies receive extremely high numbers of ratings**, creating a long right-tail distribution.
- The distribution is **highly skewed**, which can lead to popularity bias in recommendation systems.



4. Top 20 Movies by Average Rating (min 100 ratings)
- All top-ranked movies have **very high average ratings (above ~4.2)**, indicating exceptionally strong audience approval.
  - The list is dominated by **critically acclaimed classics, documentaries, and high-quality series**, showing that quality-driven content consistently receives higher ratings.
  - Since a **minimum of 100 votes** was enforced, these high ratings are **statistically reliable** and not due to small-sample bias.

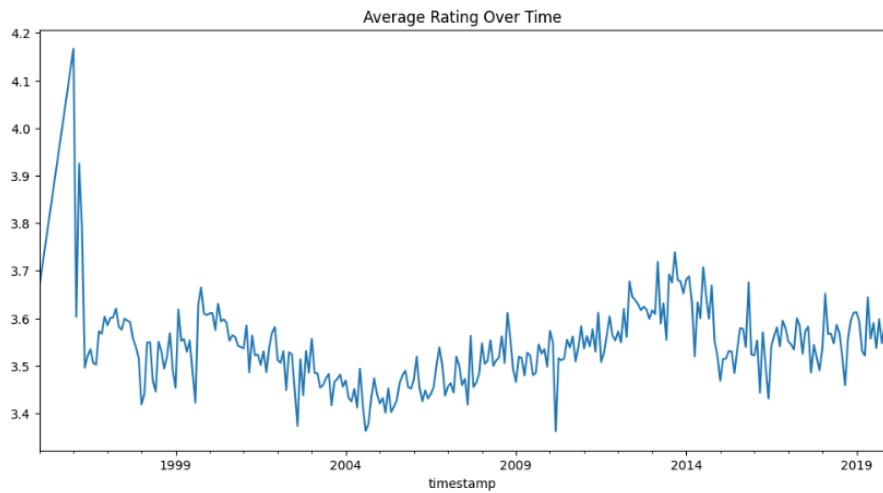


5. Number of Movies per Genre
- Drama and Comedy dominate the dataset**, having the highest number of movies compared to all other genres.
  - Genres like Thriller, Romance, and Action** also have strong representation, indicating viewer preference for mainstream entertainment.
  - IMAX, Film-Noir, and Musical** have very few movies, showing that these are **niche or less-produced genres** in the dataset.



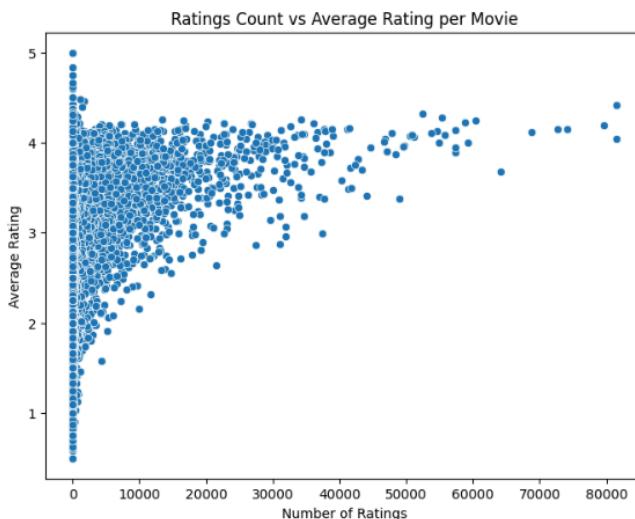
## 6. Average Rating Over Rating

- a. The **average rating remains fairly stable** over the years, mostly fluctuating between **3.4 and 3.7**, indicating consistent user rating behavior.
- b. There is a **slight upward trend around 2010–2014**, suggesting a period of improved user satisfaction or higher-quality content.
- c. Short-term **spikes and dips reflect temporary variations**, but no extreme long-term decline or growth is observed.



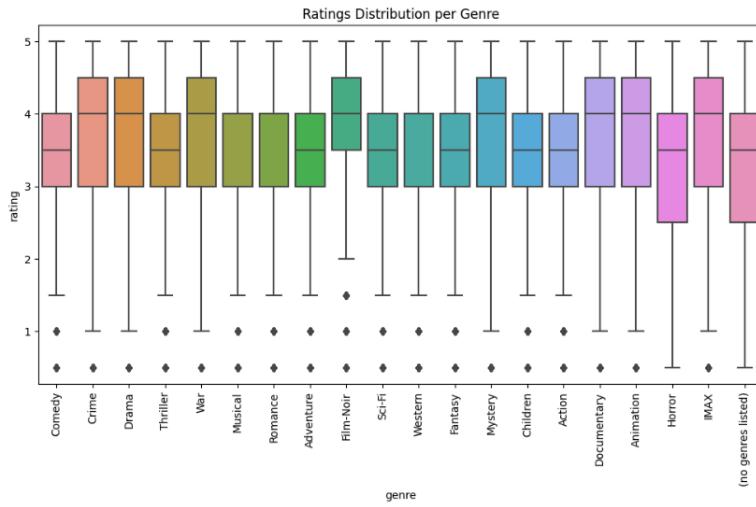
## 7. Rating Count vs Average Rating per Movie

- a. Movies with **few ratings show a wide spread of average ratings**, indicating **low statistical reliability** for less-rated movies.
- b. As the **number of ratings increases, the average rating stabilizes**, mostly between **3.5 and 4.2**, showing more reliable consensus.
- c. **Highly rated movies usually have a large number of ratings**, suggesting that popular movies tend to maintain consistently good ratings.



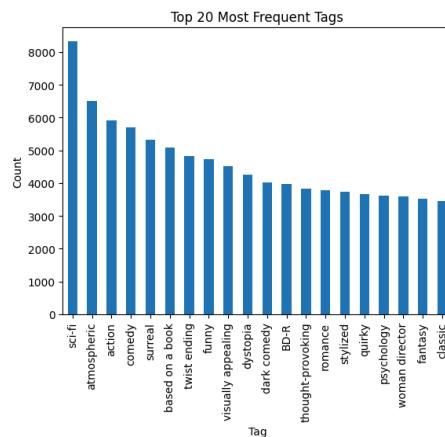
## 8. Rating Distribution per Genre

- a. **Most genres have median ratings between 3.5 and 4.0**, indicating generally positive audience reception across genres.
- b. **Documentary, Animation, and IMAX** show slightly **higher median ratings**, suggesting these genres are perceived as higher quality.
- c. **Horror and Children** exhibit **wider variability and more low-rating outliers**, indicating mixed audience responses.



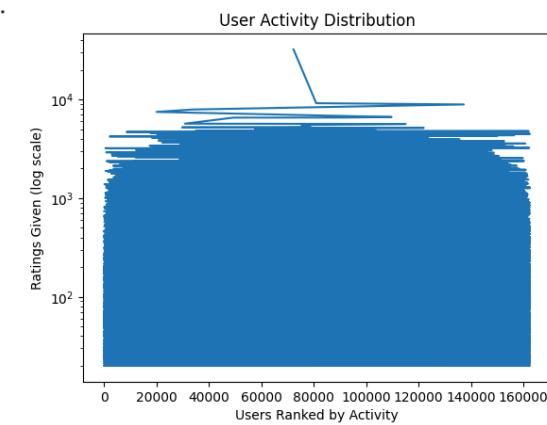
## 9. Top 20 Most Frequent Tags

- a. **Sci-fi, atmospheric, and action** are the most common tags, meaning users frequently engage with visually immersive, fast-paced genres.
- b. Many tags describe **tone and style** (e.g., surreal, dark comedy, thought-provoking) rather than just genre, showing users care about mood and storytelling style.
- c. Tags like "based on a book", "quirky", and "psychology" indicate that audiences also value unique narrative elements and intellectual depth beyond mainstream genres.



## 10. User Activity Distribution

- The distribution shows that most users provide only a small number of ratings, while only a few users are extremely active and contribute thousands of ratings.
- This pattern reflects a typical long-tail behavior, where a minority of “power users” generate a large portion of the data used for collaborative filtering.
- Because many users have very limited interactions, the system must handle cold-start scenarios and rely on content-based or hybrid strategies to recommend items effectively.



## 11. Early vs Late Rating Drift

- The plot suggests that user ratings fluctuate over time, meaning users may rate similar movies differently depending on when they watched them — this indicates **rating drift** in behavior.
- Since the drift ranges both positive and negative, users do not maintain consistent strictness or leniency, which implies **personal rating style evolves** with experience, mood, or exposure to more movies.
- This variation highlights the importance of **time-aware or session-aware recommendation models**, rather than treating all historical ratings equally.

