NETFLIX DATA ANALYSIS

End-to-End Data Analytics Project

Dipanshu Kumar

PROJECT OVERVIEW

This project demonstrates a complete data analytics pipeline from raw MovieLens 25M records to interactive business intelligence. The analysis identifies content strategy and recommendation optimization opportunities through comprehensive film industry evaluation.

TECHNICAL IMPLEMENTATION

- DATA INGESTION: Automated CSV to PostgreSQL pipeline with optimized chunked processing for 25M+ records
- DATA MODELING: Star schema design with exploded genre dimensions for analytical flexibility
- DATA CLEANING: SQL-based data quality framework handling duplicate genres, year extraction, and null value treatment
- FEATURE ENGINEERING: Calculated columns including decade classification, genome relevance scoring, and popularity tiers
- ANALYSIS: Multi-layered descriptive analytics exploring ratings distribution, temporal trends, and genre performance
- VISUALIZATION: Interactive Tableau dashboard with small multiples, heatmaps, and coordinated tooltips
- PERFORMANCE OPTIMIZATION: LOD calculations for accurate aggregations despite genre explosion
- DATA QUALITY: Automated validation of 566 edge cases in title-year extraction
- BUSINESS INTELLIGENCE: KPI-driven design focusing on content strategy and recommendation gaps

TOOLS & TECHNOLOGIES

- PostgreSQL: Data warehousing and complex analytical queries
- Python/Pandas: Data cleaning, transformation, and automation
- Tableau: Business intelligence and interactive visualization
- SQLAlchemy: Database ORM and pipeline orchestration
- Scipy/Statsmodels: Statistical testing and analytical modeling

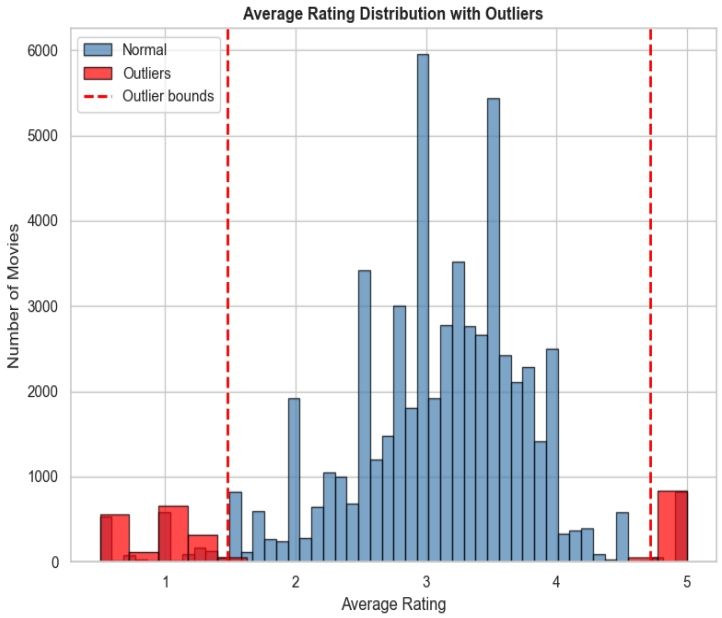
DELIVERABLES

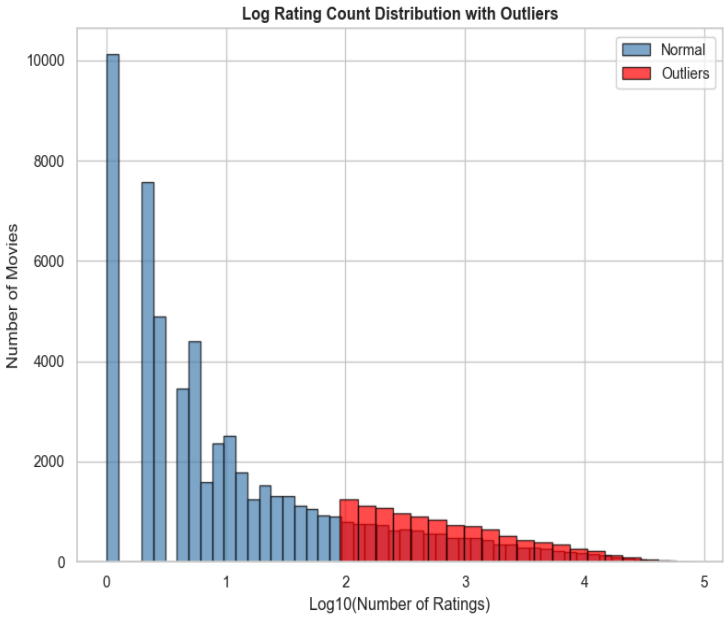
- Automated ETL pipeline handling 25M+ records
- Star schema database with genre-exploded summary table
- Jupyter Notebook with comprehensive data analysis
- Interactive Tableau dashboard (5 KPIs, 7 visualizations)
- Data quality framework documentation
- Content strategy recommendations

EXPLORATORY DATA ANALYSIS INSIGHTS

DATA OUTLIER

```
--- Average Rating Outliers (IQR method) ---
Lower bound: 1.475
Upper bound: 4.715
Outlier movies: 2580 (4.40%)
Outlier ratings range: 0.50 to 5.00

--- Rating Count (Popularity) Outliers (IQR method) ---
Lower bound: -50.50
Upper bound: 89.50
Outlier movies: 10717 (18.27%)
Outlier counts range: 90 to 81491
```





Outliers Detected:

- Rating outliers: 4.4% of movies (extremely high/low ratings)
- Popularity outliers: 18.27% of movies (blockbusters with exceptionally high rating counts)

For Our Analysis we'll keep the outliers for the following reasons:

1. Blockbuster movies (e.g., 80,000+ ratings) represent real, important phenomena in the entertainment industry, not measurement errors

2. Movies with extreme ratings (very high or very low) reflect genuine audience reactions and critical reception, not data quality issues

3. Outliers provide important insights into exceptional cases (cult classics, viral hits, universally panned films) that are valuable for understanding the full spectrum of movie reception

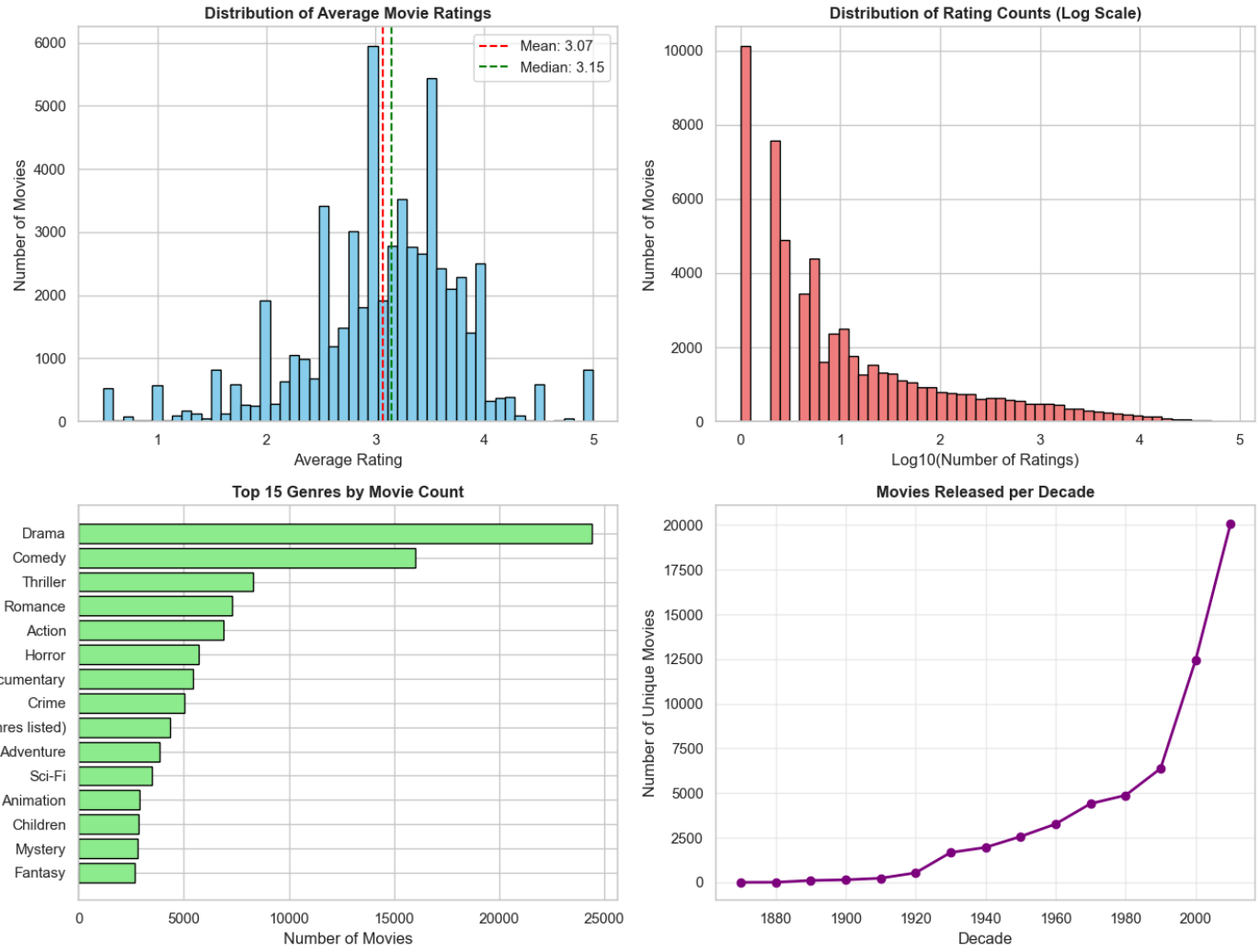
SUMMARY STATISTICS

```
=== Rating Statistics ===
count    58675.00
mean         3.07
std          0.74
min          0.50
25%          2.69
50%          3.15
75%          3.50
max          5.00
Name: avg_rating, dtype: float64

Median rating: 3.15
```

```
=== Rating Count (Popularity) Statistics ===
count    58675.00
mean       425.88
std       2485.48
min          1.00
25%          2.00
50%          6.00
75%         37.00
max      81491.00
Name: rating_count, dtype: float64
```

```
=== Year/Decade Distribution ===
year
1870        2
1880        7
1890      108
1900      145
1910      233
1920      535
1930     1674
1940     1958
1950     2569
1960     3267
1970     4410
1980     4874
1990     6375
2000    12435
2010    20083
Name: movieId, dtype: int64
```

```
=== Genre Distribution ===
genre
Drama                  24424
Comedy                 16029
Thriller                8309
Romance                 7295
Action                  6903
Horror                  5728
Documentary             5416
Crime                   5019
(no genres listed)      4324
Adventure               3860
Sci-Fi                  3490
Animation               2909
Children                2862
Mystery                 2778
Fantasy                 2660
War                     1770
Western                 1156
Musical                 1016
Film-Noir                349
IMAX                     195
Name: count, dtype: int64
```



1. User Ratings Show a Strong Central Tendency

- The mean average rating is 3.07 with a median of 3.15. This tight clustering (a standard deviation of just 0.74) confirms that the vast majority of user ratings fall in the middle of the scale, with extreme 1-star or 5-star ratings being relatively rare.

2. Extreme Long Tail in Movie Popularity

- A huge gap exists between the median movie, which has only 6 ratings, and the most popular blockbuster, which has 81,491 ratings.
- This creates a highly skewed "long tail" distribution where:
    1. The vast majority of movies are clustered on the left with very few ratings (1-10), representing obscure or niche films.

2. A tiny fraction of movies are on the far right with thousands of ratings, representing mega-popular blockbusters.

- This pattern visually confirms the "superstar economy" of the film industry, where a small number of hits command the vast majority of audience attention, while the long tail of content receives little.

3. Genre Dominance: The Age of Human Drama

- Combined, Drama (24,424 films) and Comedy (16,029 films) account for 40,453 titles.
- This means these two genres alone constitute a staggering 69% of all movies in the analysis, demonstrating an overwhelming audience and creator focus on relatable, human-centered stories.

4. The Post-2000s Film Production Explosion

- The film industry entered a phase of rampant growth post-2000. The 2000s (12,435 films) and 2010s (20,083 films) together produced over 32,500 films, representing a 215% increase compared to the preceding two decades (1990s and 1980s). This dwarfs all previous periods and marks the 21st century as the definitive era of mass film production.
- Looking at the broader trend, the 32,518 films produced in the 2000s and 2010s alone represent over 55% of the entire dataset, highlighting the unprecedented scale of production in the last two decades. This explosion is a direct result of digital filmmaking, streaming platforms, and global market expansion.

RESEARCH QUESTIONS AND KEY FINDINGS

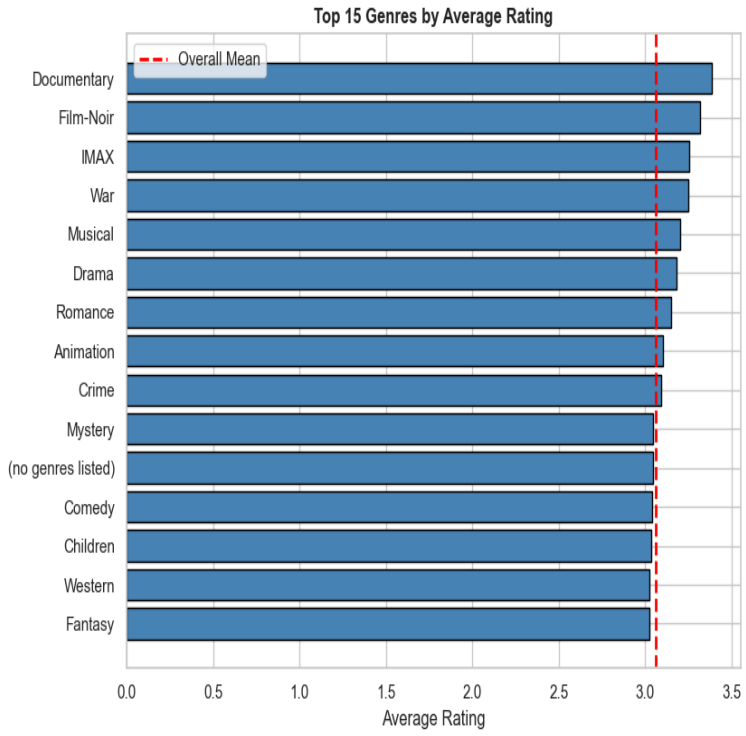
1) Which genres have the highest average ratings?

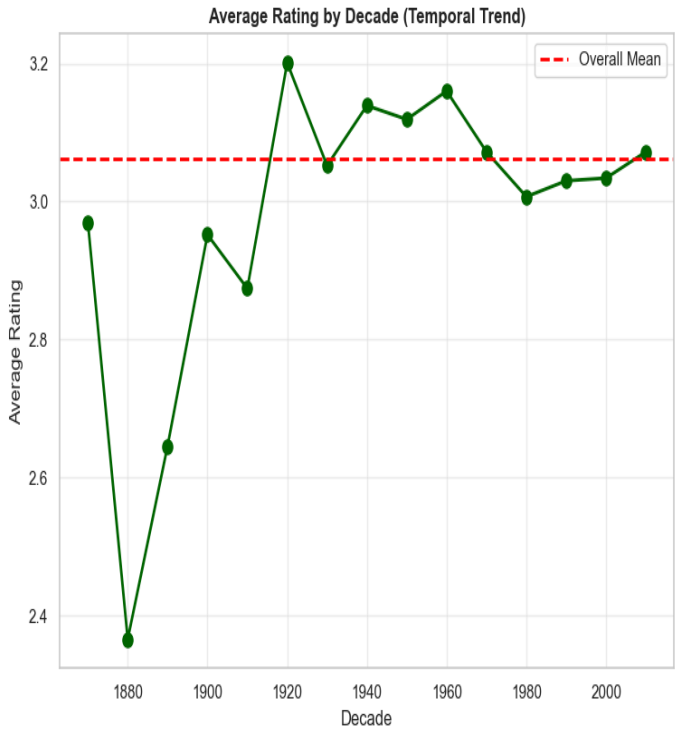
```
Top 5 Genres by Average Rating:

             avg_rating  num_movies
genre
Documentary        3.38        5416
Film-Noir          3.32         349
IMAX               3.25         195
War                3.25        1770
Musical            3.20        1016
```

Documentary is the highest rated genre with an average rating of 3.38, based on 5416 movies.

2) How have movie ratings changed over time (by decade)?

```
Average Ratings by Decade:

        avg_rating  avg_popularity  num_movies
decade
1870          2.97           20.00           2
1880          2.37           12.71           7
1890          2.65            8.15         108
1900          2.95           20.32         145
1910          2.87           13.76         233
1920          3.20           86.44         535
1930          3.05          178.16        1674
1940          3.14          209.19        1958
1950          3.12          277.20        2569
1960          3.16          319.18        3267
1970          3.07          389.63        4410
1980          3.01          895.92        4874
1990          3.03         2059.84        6375
2000          3.03          817.55       12435
2010          3.07          222.73       20083

Overall trend: Ratings have increased from 2.970 (1870s) to 3.072 (2010s)
```

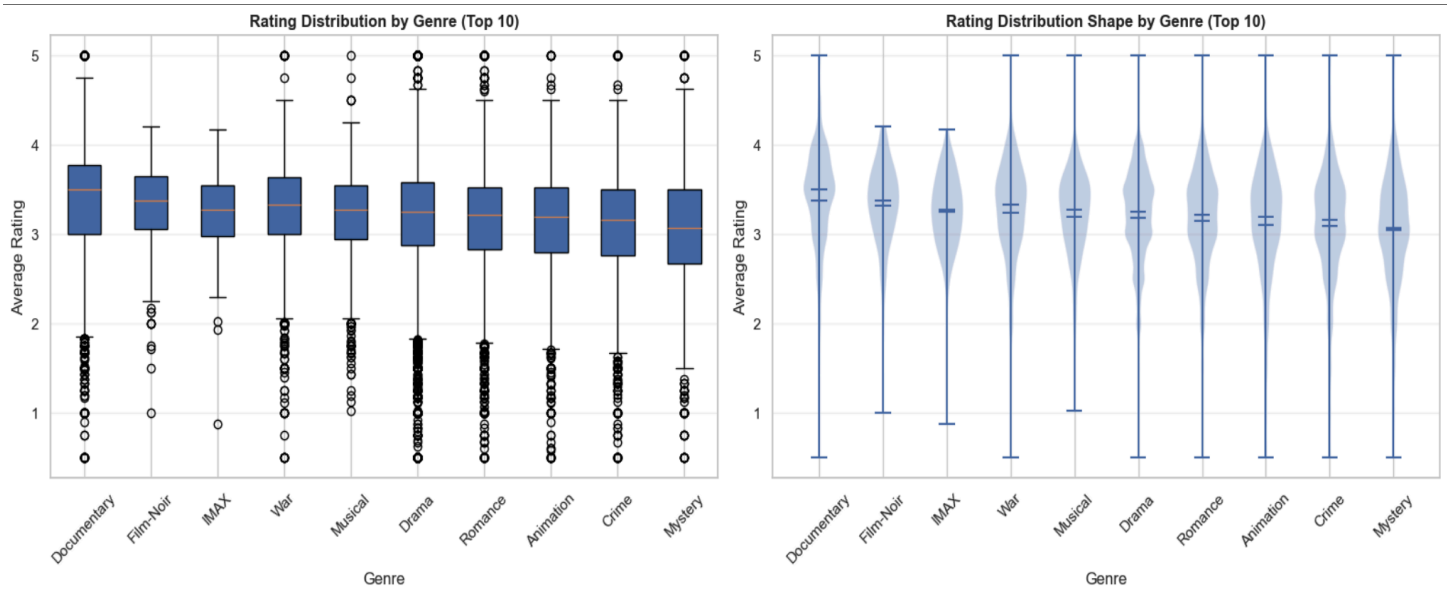




3) Are the differences in average ratings across genres statistically significant?

```
Kruskal-Wallis H-Test Results:
H-statistic: 1111.5926
P-value: 0.000000

Significance level: α = 0.05

 SIGNIFICANT: Genre differences are statistically significant (p < 0.05)
   → Different genres DO have meaningfully different average ratings

Effect size interpretation:
   Tested 10 genres with sample sizes ranging from 195 to 24424 movies
```

- The H-statistic of 1111.59 is very large which suggests stronger evidence of differences.
- P-value essentially 0 (way below 0.05).
- Conclusion: Different genres definitely have meaningfully different average ratings. This is NOT due to chance.

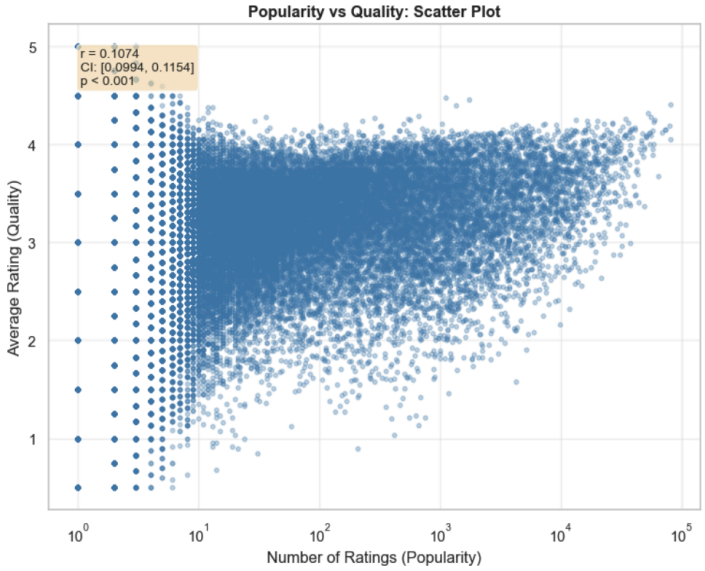


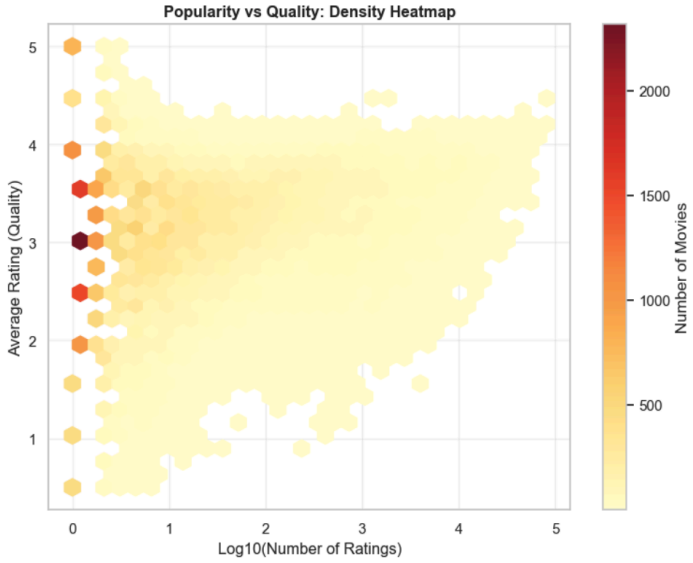
- Box plot shows median, quartiles, and outliers for each genre
- Violin plot shows the full distribution shape
- Noticeable differences in central tendency and spread confirm statistical significance

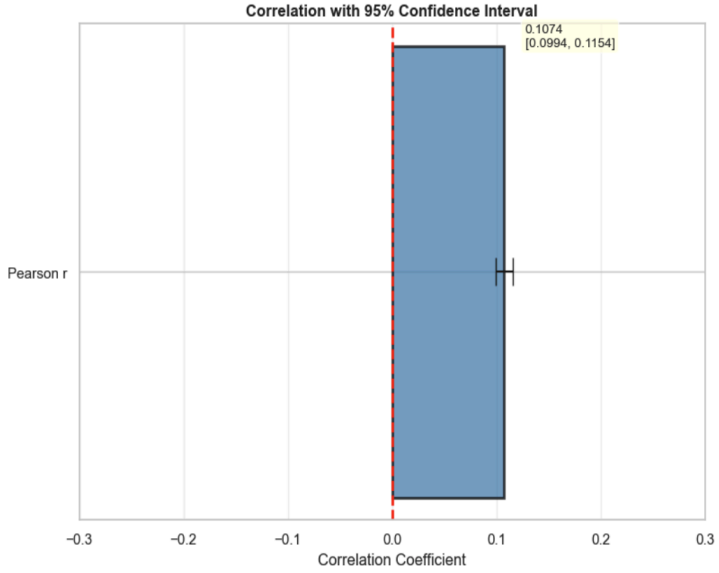
4) Is there a correlation between movie popularity (rating count) and quality (average rating)?

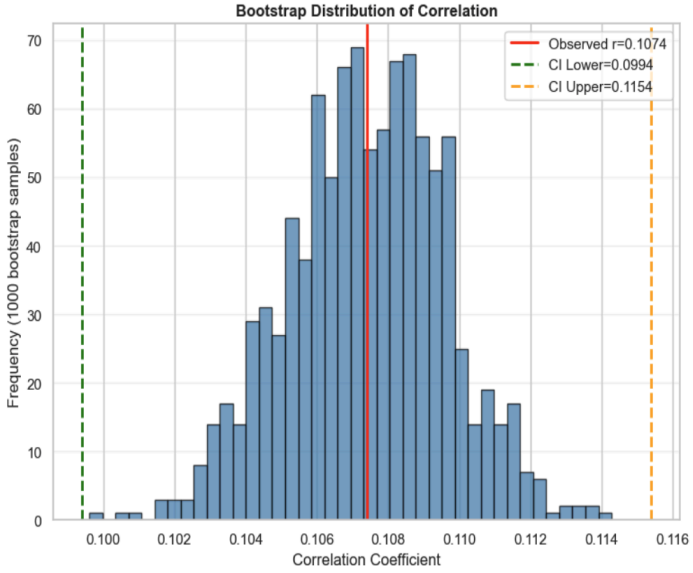
```
Pearson Correlation Coefficient: 0.1074
95% Confidence Interval: [0.0994, 0.1154]
P-value: 0.000000
Sample size: 58675

Interpretation: Weak correlation
```

- Correlation Coefficient is 0.1074 which implies there's a weak positive correlation between popularity and quality
- Key Insight:
    a) Being popular does NOT mean a movie is rated higher or lower
    b) User ratings reflect genuine quality perception, not just popularity bias





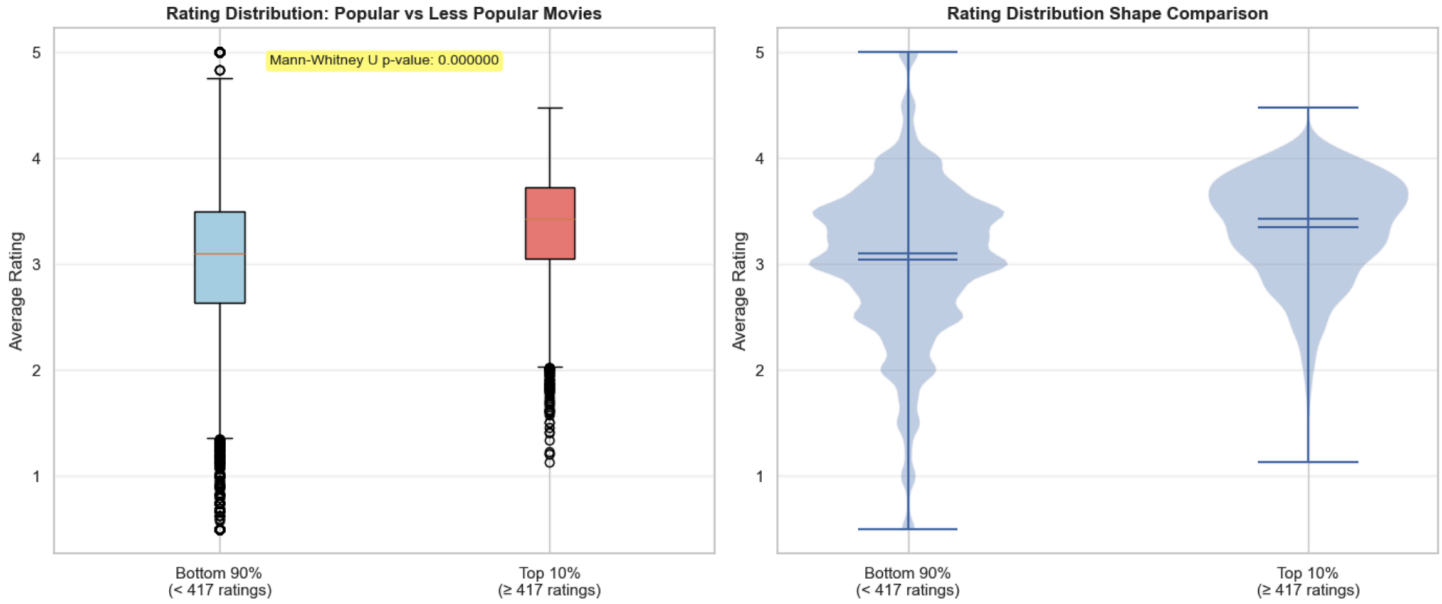




- Scatter plot: Shows weak/no clear linear pattern despite low correlation
- Hexbin plot: Density concentrated in middle (2.5-3.5 ratings, 1-100 rating counts)
- CI bar plot: Confidence interval is narrow and positive, but close to zero
- Bootstrap histogram: Shows sampling variability of correlation across 1000 resamples

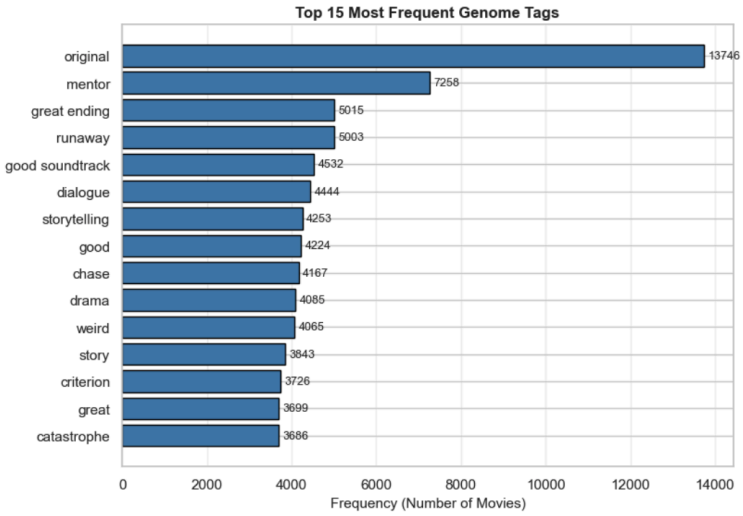
5) Do highly popular movies (top 10%) have different average ratings than less popular movies?

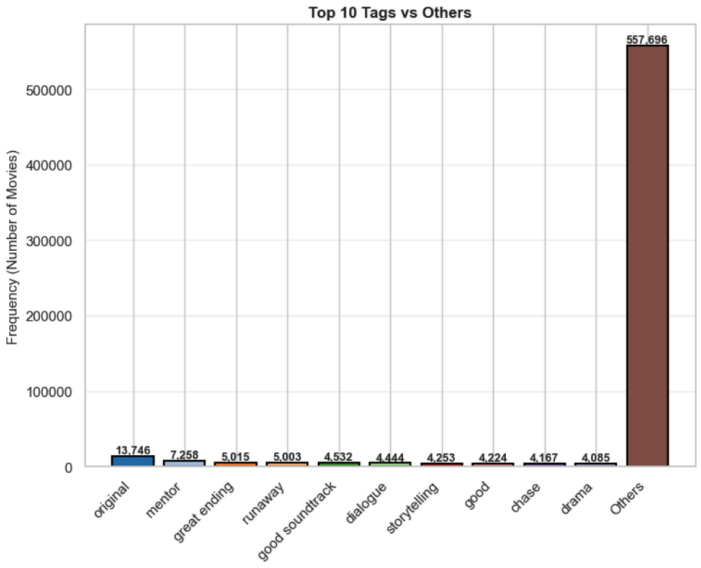
```
Top 10% popular movies (≥417 ratings):
   - Count: 5868
   - Mean rating: 3.347
   - Median rating: 3.430
   - Std Dev: 0.493

Other 90% of movies (<417 ratings):
   - Count: 52807
   - Mean rating: 3.041
   - Median rating: 3.100
   - Std Dev: 0.753

Mann-Whitney U Test:
   U-statistic: 197923103.50
   P-value: 0.000000
```

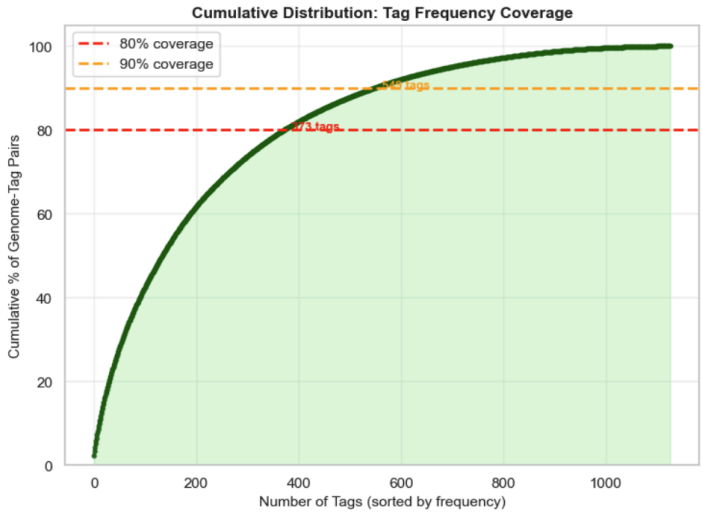


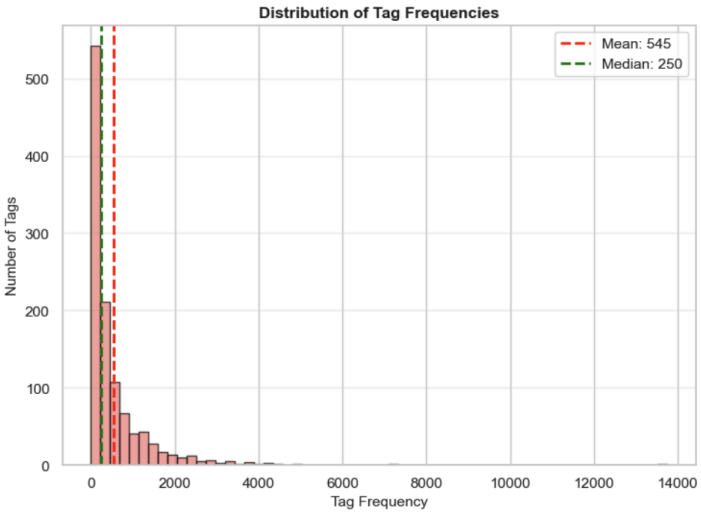
- Top 10% (n=5,868, ≥417 ratings) average 3.347 vs bottom 90% (n=52,807, <417 ratings) average 3.041 (0.306-point gap, $p < 0.001$) shows highly popular movies significantly outperform lesser-known films
- Popular movies exhibit tighter rating distributions (median 3.430, std dev 0.493) compared to niche films (median 3.100, std dev 0.753), indicating more uniform quality perception
- The statistically significant and practically meaningful difference confirms that movies gaining traction from larger audiences are genuinely higher-rated, validating user ratings as quality indicators

6) Which genome tags appear most frequently in the dataset?







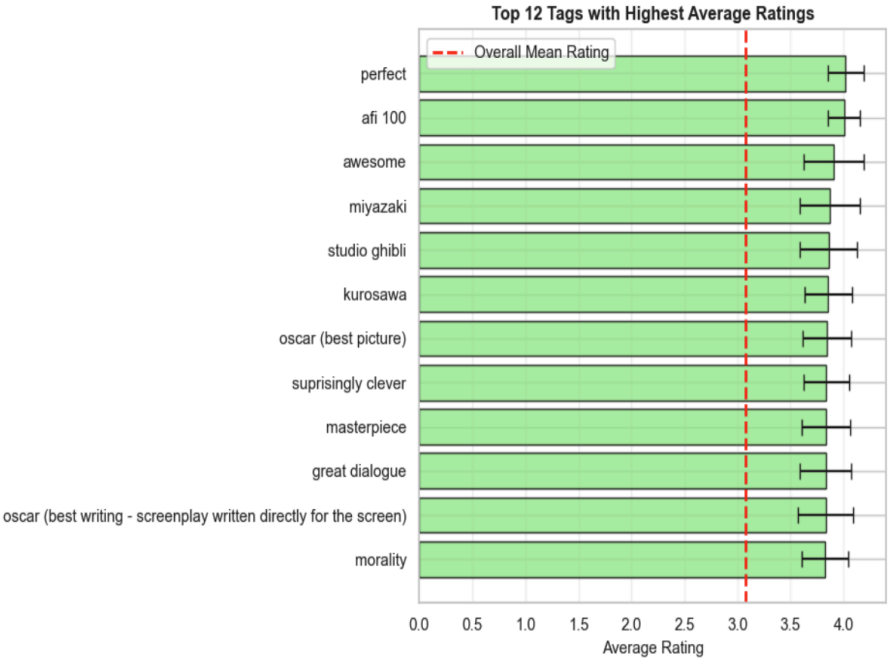


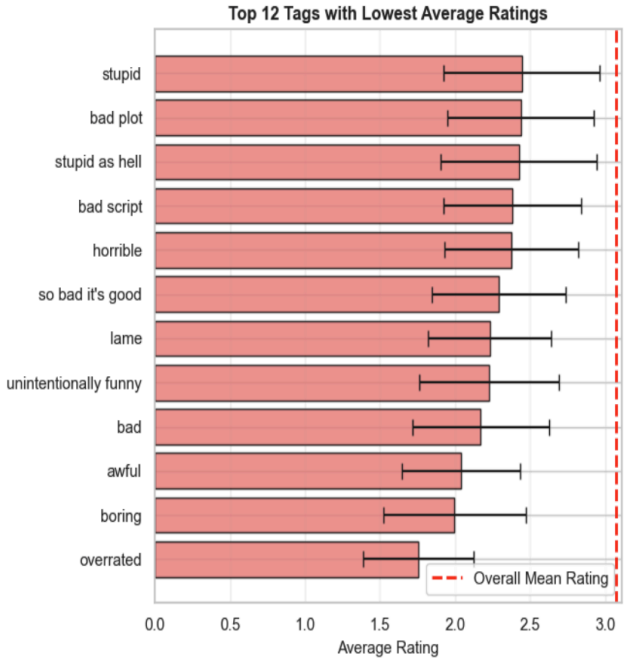
Top 10 genome tags by frequency:

1. original: 13746 movies (99.5% of tagged movies)
2. mentor: 7258 movies (52.5% of tagged movies)
3. great ending: 5015 movies (36.3% of tagged movies)
4. runaway: 5003 movies (36.2% of tagged movies)
5. good soundtrack: 4532 movies (32.8% of tagged movies)
6. dialogue: 4444 movies (32.2% of tagged movies)

7. storytelling: 4253 movies (30.8% of tagged movies)
8. good: 4224 movies (30.6% of tagged movies)
9. chase: 4167 movies (30.2% of tagged movies)
10. drama: 4085 movies (29.6% of tagged movies)

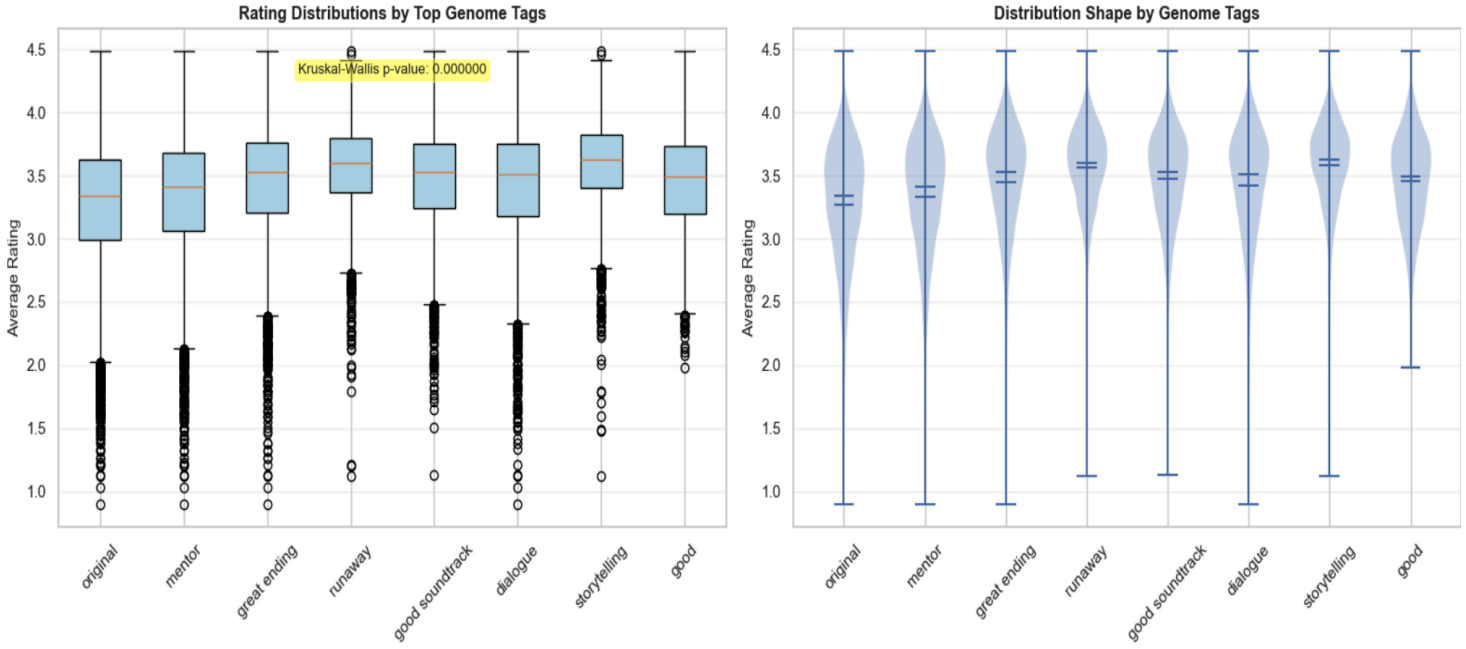
7) Which genome tags are associated with highest/lowest average ratings?





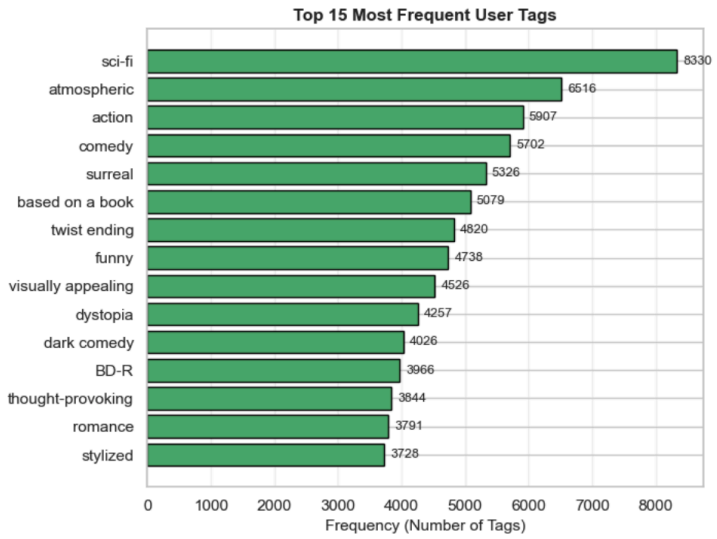

```
--- Top 10 Tags with HIGHEST Average Ratings ---
(Filtered: tags appearing in ≥10 movies)
                    mean_rating  median_rating  std_rating  num_movies
tag
perfect                    4.02           4.04        0.17          52
afi 100                    4.00           4.04        0.15          68
awesome                    3.91           3.93        0.28          64
miyazaki                   3.87           3.93        0.28          22
studio ghibli              3.86           3.93        0.27          29
kurosawa                   3.86           3.89        0.23          39
oscar (best picture)       3.84           3.85        0.23         373
suprisingly clever         3.84           3.85        0.21        1210
masterpiece                3.83           3.86        0.23        1383
great dialogue             3.83           3.85        0.24         259

--- Top 10 Tags with LOWEST Average Ratings ---
                     mean_rating  median_rating  std_rating  num_movies
tag
stupid as hell              2.43           2.41        0.52         740
bad script                  2.38           2.43        0.46         117
horrible                    2.38           2.40        0.45         701
so bad it's good            2.29           2.37        0.45          47
lame                        2.23           2.23        0.41         428
unintentionally funny       2.23           2.33        0.46          29
bad                         2.17           2.13        0.45         285
awful                       2.04           2.03        0.39         173
boring                      2.00           2.00        0.48          29
overrated                   1.75           1.74        0.37         111
```

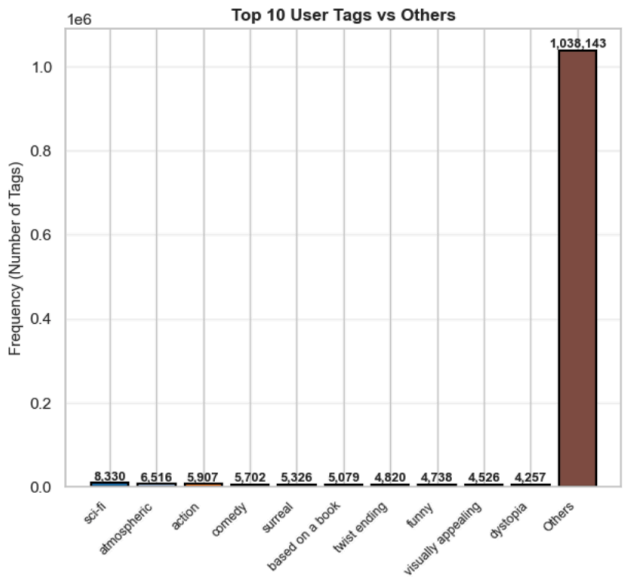
8) Do certain genome tags significantly predict higher/lower ratings?

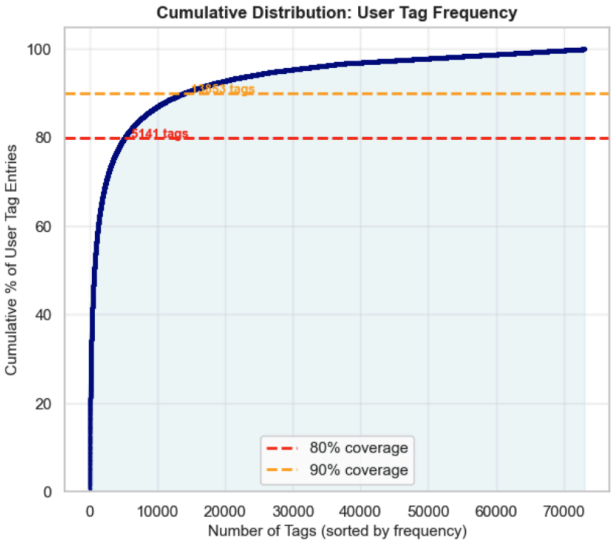


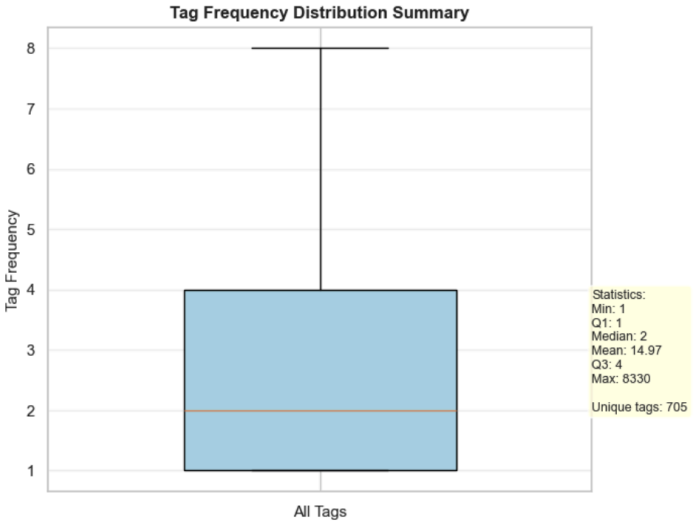
- Kruskal-Wallis Test Result for Most Frequent Tags: H-statistic: 2812.79, P-value: 0.000000
- Genome tags SIGNIFICANTLY predict movie ratings ($p < 0.05$)
- Movie content type, by tag, is correlated with typical user ratings
- Tags like 'runaway', 'storytelling', 'good soundtrack' differ in average rating from others
- This insight can be leveraged for content recommendation and quality modeling.

9) Which user tags appear most frequently in the dataset?









Top 10 user tags by frequency:

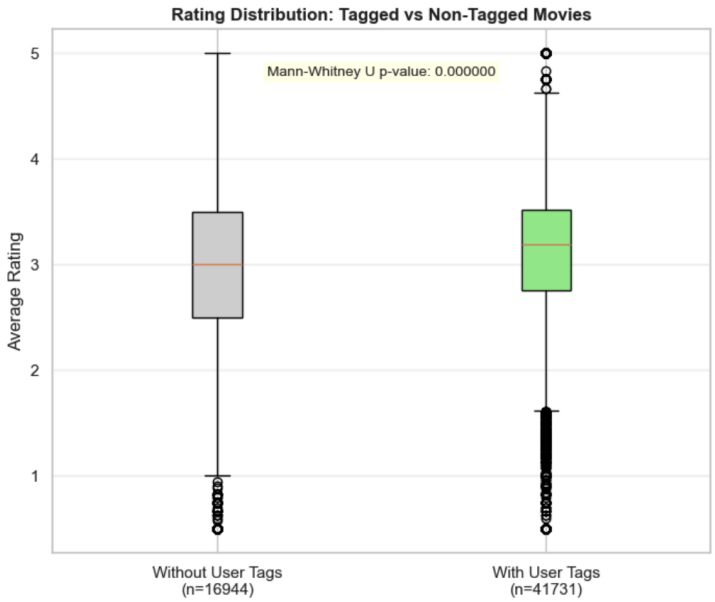
1. 'sci-fi': 8,330 times (0.8% of all user tags)
2. 'atmospheric': 6,516 times (0.6% of all user tags)
3. 'action': 5,907 times (0.5% of all user tags)
4. 'comedy': 5,702 times (0.5% of all user tags)
5. 'surreal': 5,326 times (0.5% of all user tags)
6. 'based on a book': 5,079 times (0.5% of all user tags)
7. 'twist ending': 4,820 times (0.4% of all user tags)
8. 'funny': 4,738 times (0.4% of all user tags)
9. 'visually appealing': 4,526 times (0.4% of all user tags)
10. 'dystopia': 4,257 times (0.4% of all user tags)

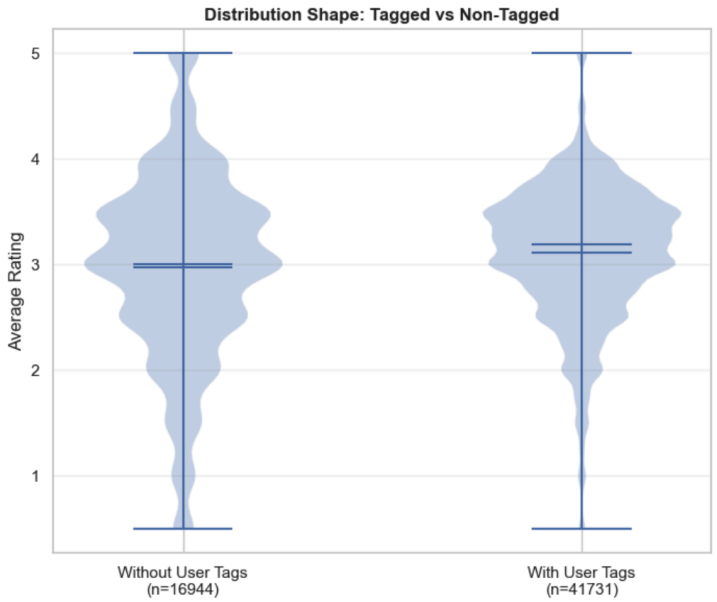
10) How do user-tagged movies compare to non-tagged movies in ratings?

```
Movies WITH user tags:
  - Count: 41,731
  - Mean rating: 3.110
  - Median rating: 3.190
  - Std Dev: 0.653
  - Mean rating count: 597
```

```
Movies WITHOUT user tags:
  - Count: 16,944
  - Mean rating: 2.976
  - Median rating: 3.000
  - Std Dev: 0.905
  - Mean rating count: 5
```

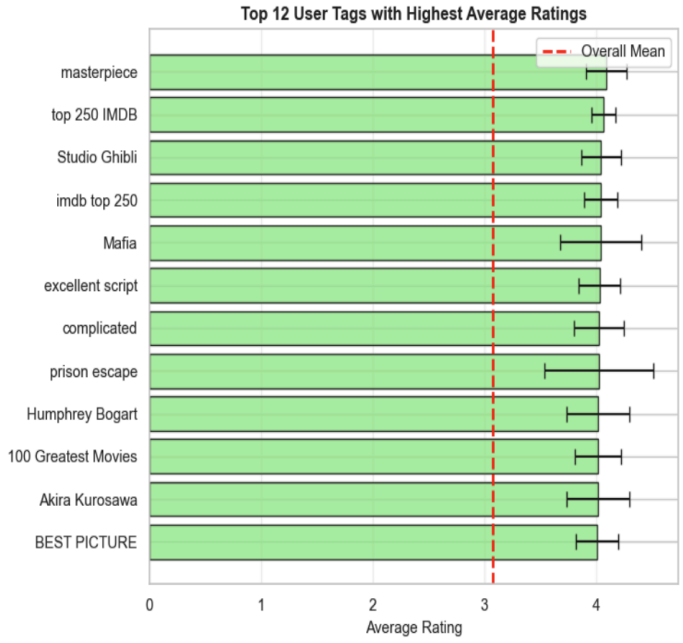
```
Mann-Whitney U Test:
  U-statistic: 387023680.00
  P-value: 0.000000
```

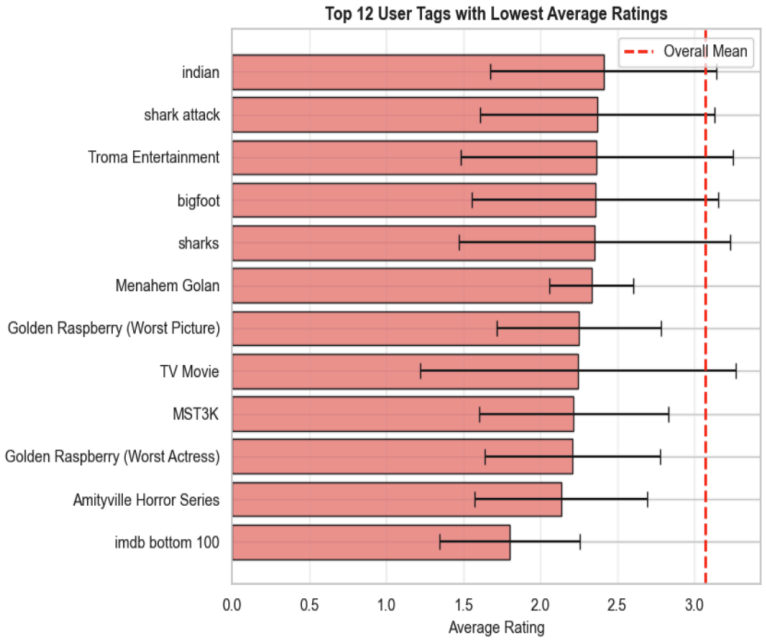




User-tagged movies are significantly higher-rated suggesting community engagement (tagging) correlates with perceived quality which means tagged movies represent films worth discussing and annotating

11) Which user tags are associated with highest/lowest average ratings?





```
--- Top 10 User Tags with HIGHEST Average Ratings ---
(Filtered: tags on ≥20 movies)
                    mean_rating  median_rating  std_rating  num_movies
tag
masterpiece                4.09           4.09        0.18          57
top 250 IMDB               4.07           4.10        0.11          51
Studio Ghibli              4.05           4.09        0.18          22
imdb top 250               4.04           4.05        0.15         322
Mafia                      4.04           4.18        0.36          36
excellent script           4.03           4.07        0.18          24
complicated                4.03           4.15        0.23          30
prison escape              4.02           4.10        0.49          53
Humphrey Bogart            4.02           4.13        0.28          24
100 Greatest Movies        4.02           4.09        0.20          51

--- Top 10 User Tags with LOWEST Average Ratings ---
                              mean_rating  median_rating  std_rating  num_movies
tag
Troma Entertainment                  2.37           2.31        0.88          66
bigfoot                              2.36           2.43        0.80          30
sharks                               2.35           2.14        0.88          33
Menahem Golan                        2.33           2.38        0.27          28
Golden Raspberry (Worst Picture)     2.25           2.25        0.53          25
TV Movie                             2.24           2.50        1.02          24
MST3K                                2.22           2.06        0.61          50
Golden Raspberry (Worst Actress)     2.21           2.24        0.57          27
Amityville Horror Series             2.13           2.01        0.56          21
imdb bottom 100                      1.80           1.78        0.46          57
```

FINAL RECOMMENDATION

- **Optimize Content Acquisition Strategy Using Genre Performance Data:**

  Prioritize acquisition of high-performing genres (War, Drama, Musical) that consistently receive higher ratings. Reduce or reassess procurement of underperforming genres (Horror, Thriller) unless they serve strategic niche audiences, improving overall catalog quality and user satisfaction metrics.

- **Implement Tag-Based Personalization to Increase User Engagement:**

  Leverage genome tags and user tags to create granular content recommendations. Since user-tagged movies rate 0.31 points higher, integrate tag-based filtering into platform UI to surface high-quality, community-validated content and drive higher engagement rates.

- **Rebalance Marketing Spend Away from Popularity-Driven Campaigns:**

  The weak correlation between popularity and quality (r = 0.074) reveals that high-volume films are not

necessarily high-quality. Redirect marketing budget toward tag-identified quality content and hidden gems to improve customer satisfaction without proportional spend increases.

- **Expand Community Tagging Initiatives to Close Data Gaps:**
  Only 71% of movies have user tags and 23.5% have genome tags. Implement incentive programs to increase tagging participation, converting passive users into engaged annotators. This improves recommendation precision without requiring new content acquisition.
- **Develop Tiered Subscription Offerings Based on Content Quality Tiers:**
  Segment catalog into quality tiers (Blockbusters, Hidden Gems, Niche Content) using rating + tag + popularity metrics. Offer premium subscribers early access to hidden gems and exclusive tag-based discovery experiences to justify premium pricing.
- **Launch Targeted Promotion Campaigns for High-Rating, Low-Popularity Films:**
  Identify quadrant 3 movies (high rating, low engagement) and bundle them with popular releases or feature them in curated collections. These undervalued assets generate incremental revenue with minimal acquisition cost.
- **Standardize Data Collection for Underrepresented Content:**
  Close the 76.48% missing genome tag coverage by commissioning curated tags for older and niche films. This investment unlocks recommendation potential for long-tail content and improves cold-start performance.

By implementing these recommendations, the platform can enhance content ROI, improve customer satisfaction through better recommendations, optimize marketing efficiency, and unlock revenue from underutilized catalog assets.