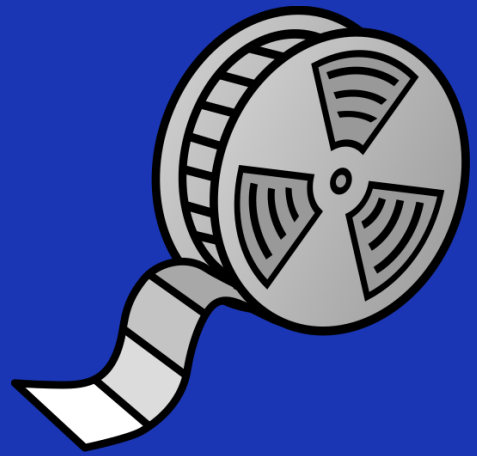
What Makes a Movie Successful? Insights from the Top Movies, Ratings, and More

By: amani and brian



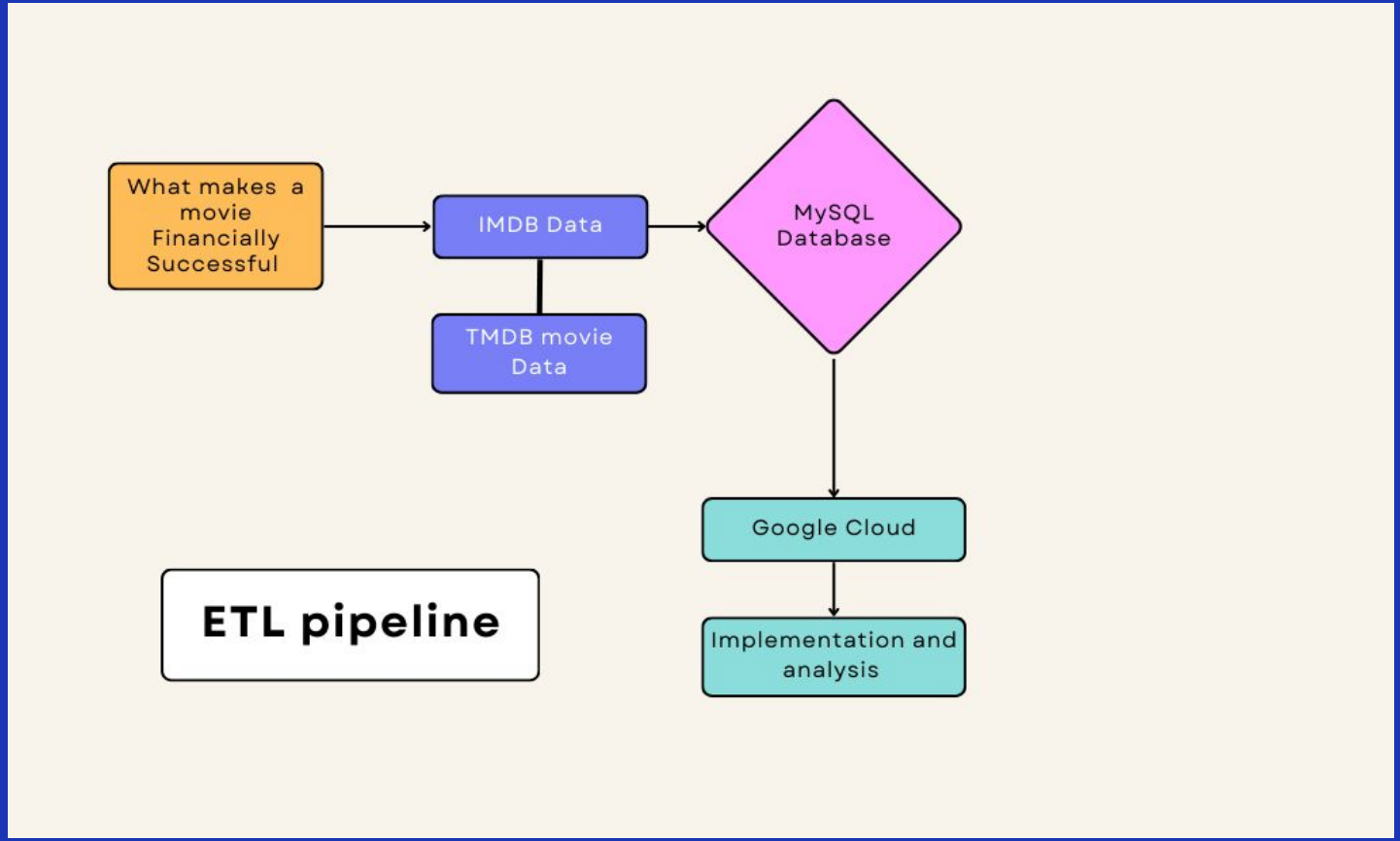
Our Objective and Data Sources

- Predict movie success by analyzing gross revenue, votes, runtime, and ratings
- Overview of IMDB and TMDB data
  - **Comprehensive Data:** Revenue, popularity, genres, runtime, and release dates are all crucial predictors of movie success.
  - **Reliable Sources:** Both are widely used, authoritative platforms with accurate, up-to-date information.
  - **Data Diversity:** TMDB focuses on financial performance, while IMDb provides user ratings, reviews, and metadata.
  - **Global Reach:** Includes data for a wide range of movies, catering to different genres, audiences, and markets.
  - **Actionable Insights:** The combination allows for multifaceted analysis, blending financial and audience engagement metrics.

Preparing the Data: Challenges and Solutions

- **Data Quality Issues:**
  - Over 3 million missing values requiring imputation or removal.
  - Irrelevant and out-of-context entries in key columns.
- **Dataset Selection:**
  - Difficulty finding reliable, well-documented movie datasets.
  - Many datasets contained fabricated or incomplete data.
- **Formatting Challenges:**
  - Inconsistent column structures and schema mismatches.
  - Need for extensive cleaning and restructuring to suit analysis.
- **Scale of Data:**
  - Large dataset size necessitated efficient processing methods like batch loading and parallel processing.

ETL Pipeline



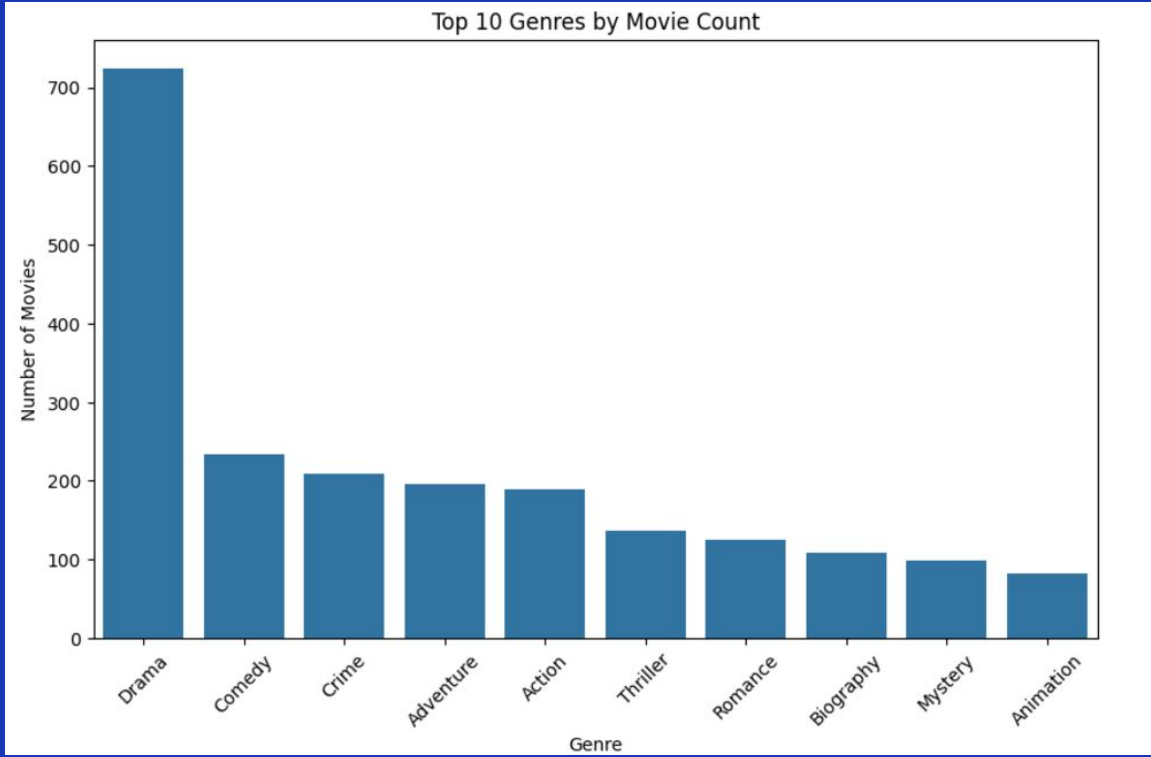
So, How Do You Gauge Success?

Highest Grossing movie on first list: Star Wars: Episode VII - The Force Awakens: $936,662,225, IMDB Rating of 7.9, meta score of 80

However, the highest SCORING (based on IMDB) movie was The Shawshank Redemption with an IMDB Rating of 9.3, meta score of 80 and a Gross Revenue of $28,341,469

Some might argue that The Shawshank Redemption is more successful than Star Wars: Episode VII… Others might say the opposite…

What Drives Movie Success? Early Insight



What Drives Movie Success? Early Insight

Correlation Between Genre and IMDB Rating

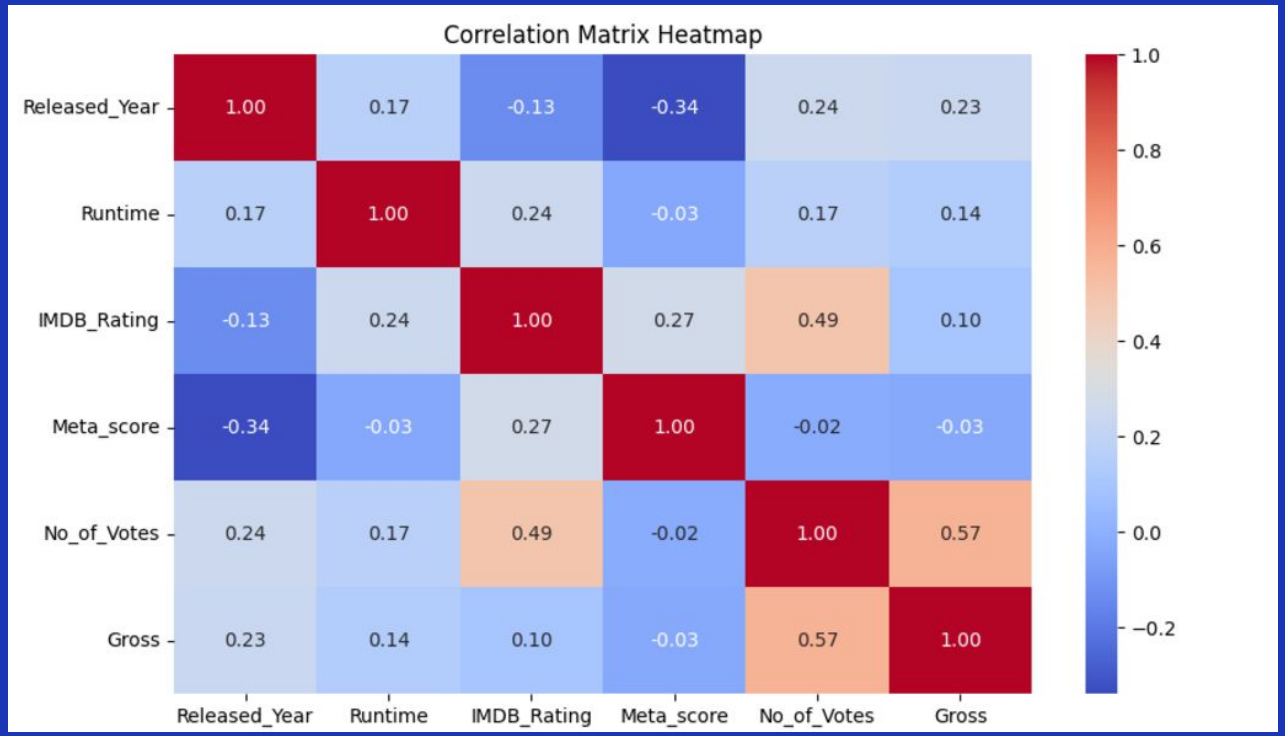
ANOVA Test Result: F-statistic = 1.0209, p-value = 0.4327
No significant difference in IMDB ratings across genres.

Chi-Square Test Result: Chi2 = 0.0000, p-value = 1.0000
No significant association between Genre and IMDB Rating bins.

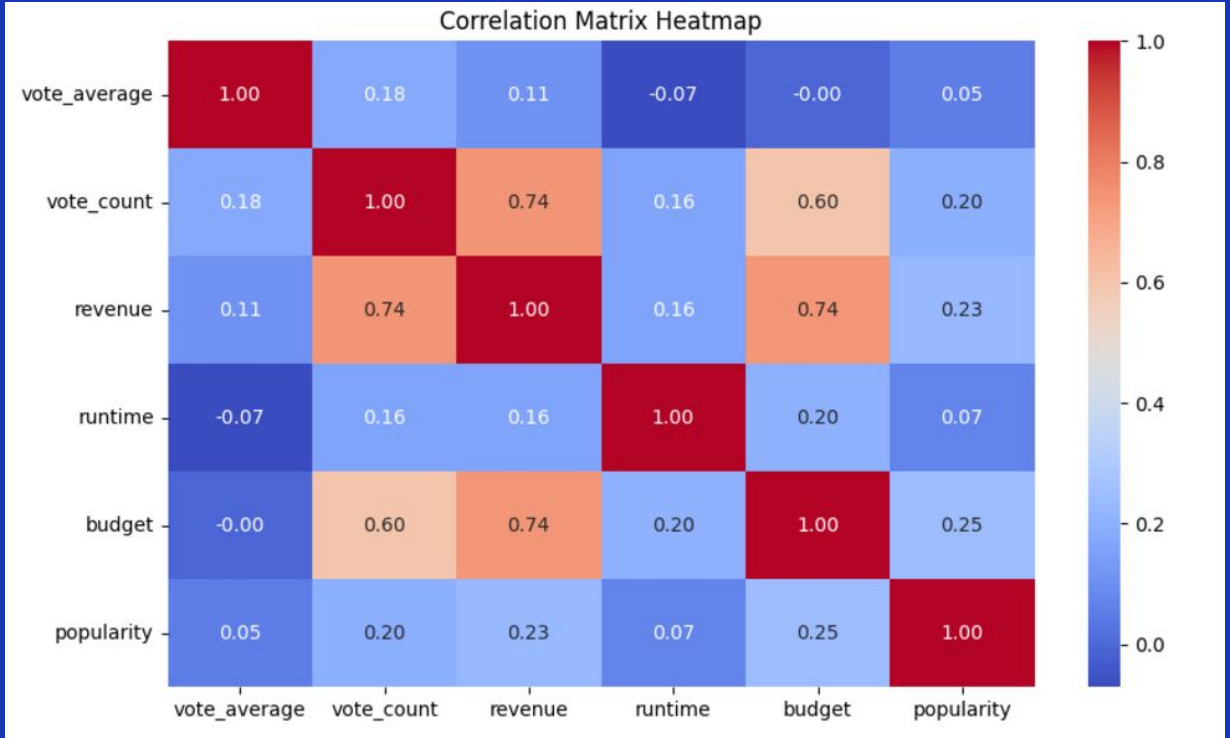
HOWEVER,

Correlation between Gross Revenue and IMDB Rating: 0.10 (p-value: 0.006)

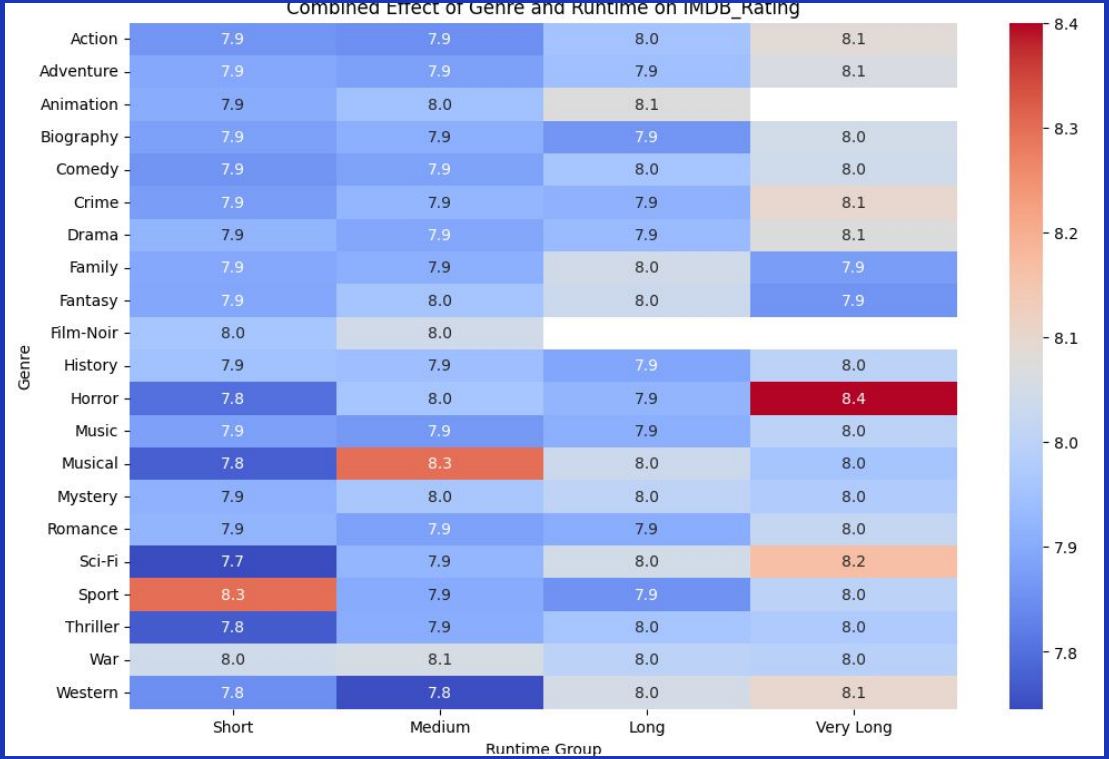
What Drives Movie Success? Early Insight



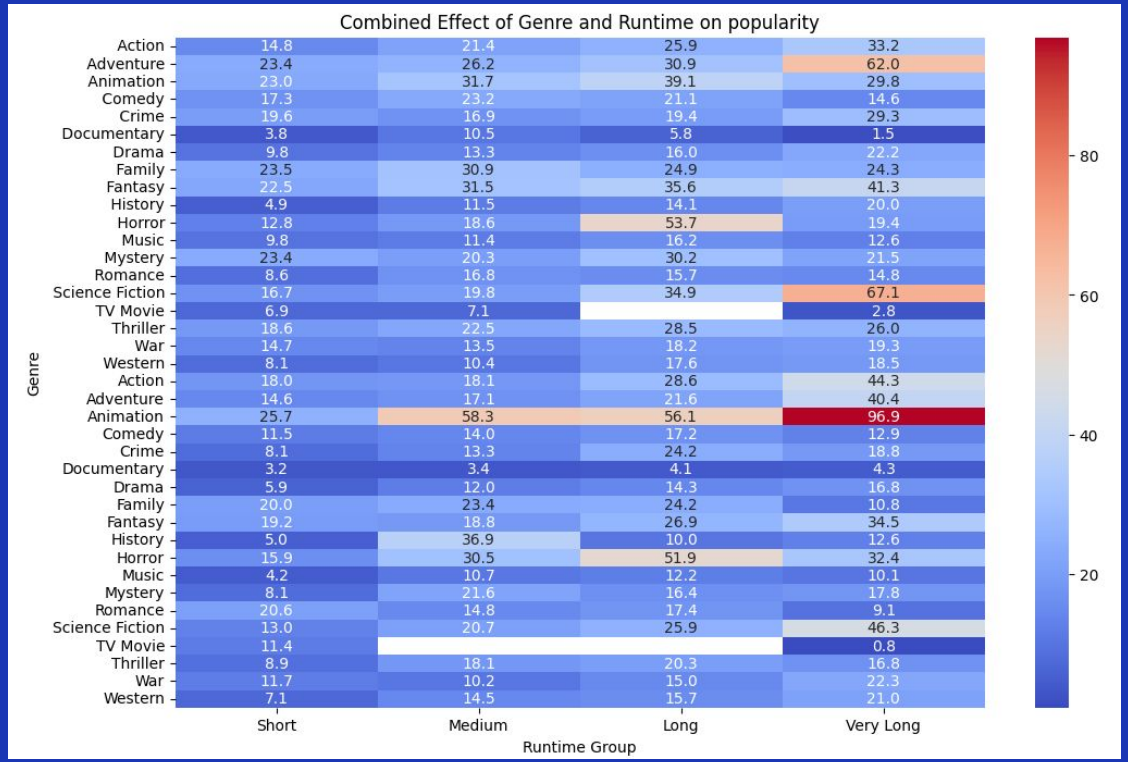
What Drives Movie Success? Early Insight



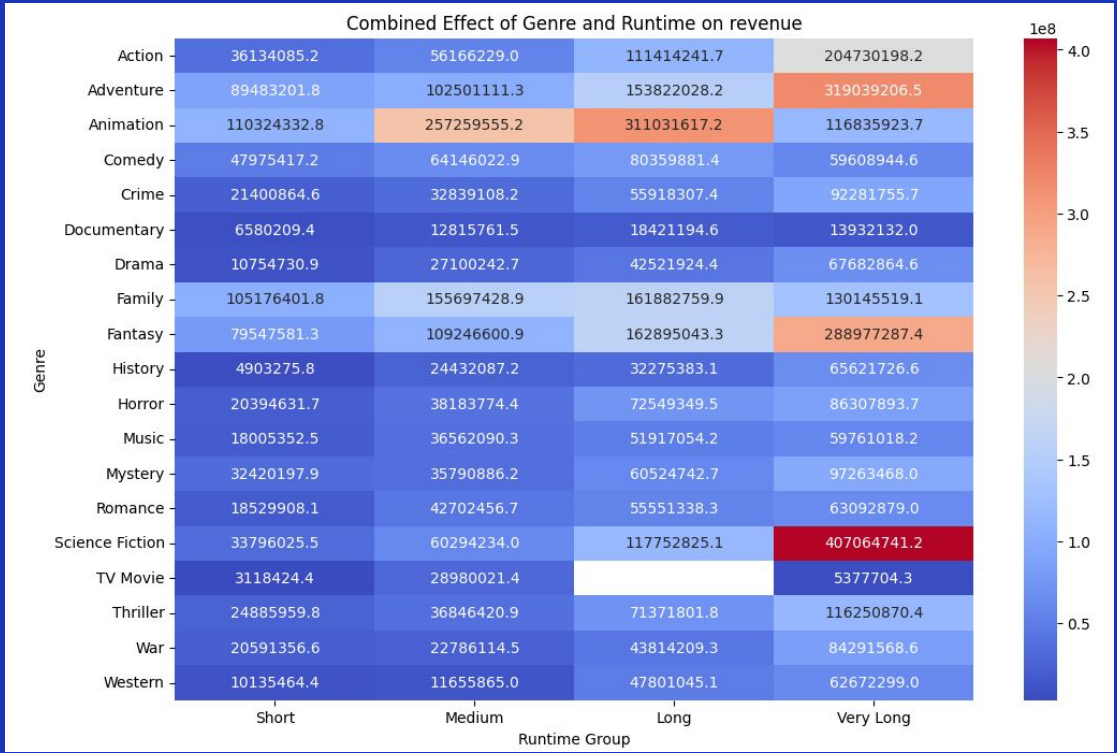
What Drives Movie Success? Early Insight



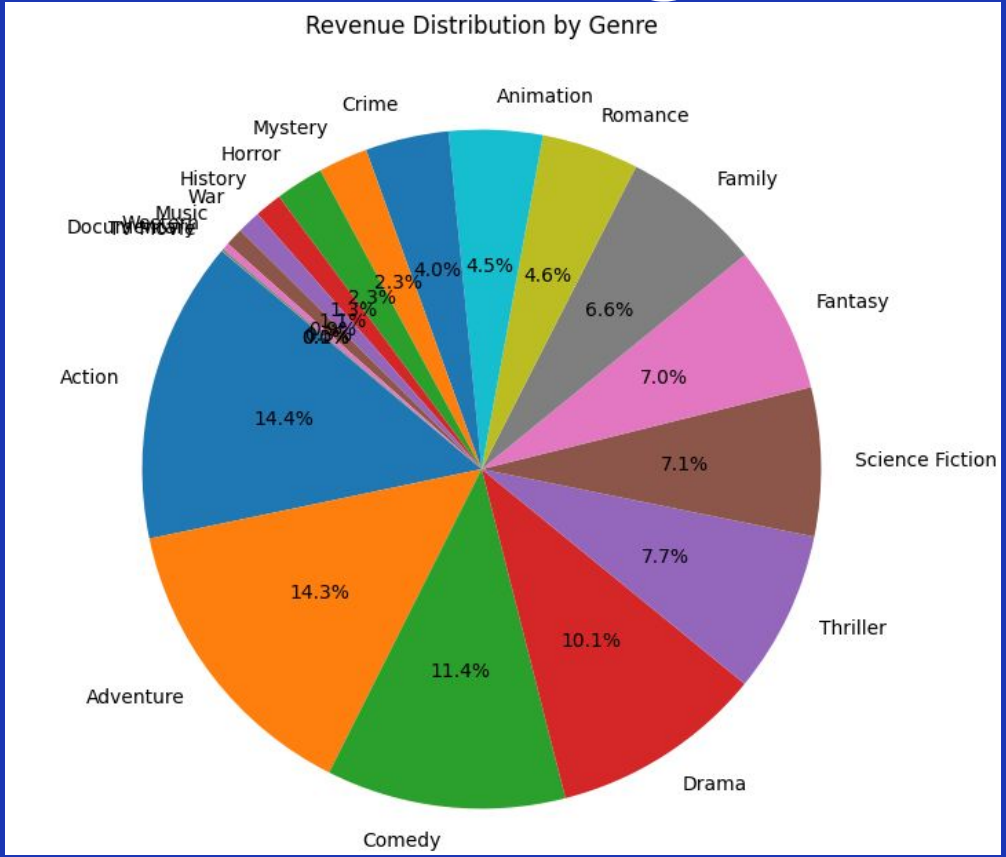
What Drives Movie Success? Early Insight



What Drives Movie Success? Early Insight



What Drives Movie Success? Early Insight



Making a “Successful” Movie

Movie 1, based off of 1st dataset (highest average IMDB Score):

Director Frank Darabont
Genre Animation, Drama, War
Star1 Tim Robbins
Star2 Morgan Freeman
Star3 Bob Gunton
Star4 William Sadler
Runtime 175 min

Movie 2, based off of 2nd dataset (highest average revenue):

budget $237000000
genres Adventure
production_companies Dune Entertainment, Lightstorm Entertainment, 20th Century Fox, Ingenious Media
production_countries Japan, Spain, United Kingdom, United States of America
spoken_languages English, Japanese, Xhosa
original_language Chinese
runtime 181 min

Challenges, Lessons and Future Directions

Challenges Faced:

- Large datasets with 3M+ missing values and inconsistent formats.
- Complex ETL setup with reproducibility issues.
- Difficulty integrating cloud storage SDKs.

Skills Gained:

- Data cleaning, ETL optimization, cloud integration.
- Improved teamwork and communication.

Major Takeaways:

- Clean data is essential for reliable analysis.
- Modular pipelines and visualization enhance impact.
- Collaboration drives project success.

Thank you for your time!

Questions?