AMERICAN UNIVERSITY OF BEIRUT

SULIMAN S. OLAYAN SCHOOL OF BUSINESS

MSBA 305 – Data Processing Framework

Professor: Ahmad El-Hajj

Project Report:

Recommending a new content strategy for Netflix

Rita Abdel Nour

Simon Feghali

Lea Kanaan

Christelle Khoury

Wadih Nassif

Date: 22nd of April, 2025

Table of Contents

LIST OF FIGURES

I. Introduction & Dataset Description

A. Introduction

In today's saturated entertainment market, streaming platforms like Netflix are in a constant race to capture audience attention while optimizing profitability. As content investment decisions become increasingly data-driven, Netflix's production content team is seeking guidance on what types of movies to prioritize—balancing viewer ratings with financial return. In this project, we position ourselves as data consultants for Netflix, tasked with helping the content production team make smarter decisions to increase their movies' ratings and their overall profitability.

To achieve this, we work with a large-scale dataset from The Movie Database (TMDB), originally containing over 1 million movie entries. The raw dataset includes variables such as title, genre, production companies, audience ratings, budget, revenue, release dates, adult flag, popularity score, and more. After cleaning and preprocessing, we curated a high-quality, analysis-ready subset of roughly 12000 movies that contain complete and reliable financial and rating information. This cleaned dataset was stored both locally and backed up on Google cloud for safe collaboration and version control.

Our primary objective is to help Netflix's production team identify what types of content lead to higher audience ratings and profitability. Specifically, we aim to answer questions such as: *Which genres consistently deliver high returns? How do production companies influence movie success? Does producing adult content affect profitability or ratings? What is the ideal mix of budget, release timing, and genre?*

To accomplish this, we implemented a full data processing pipeline, following principles of modern data engineering. Our approach combines SQL-based querying and Exploratory Data Analysis (EDA) using tools like pandas, matplotlib, and seaborn in a Jupyter Notebook environment. While the methodology section will dive into the detailed steps taken—from acquisition and storage to cleaning and analysis—at a high level, our process included:

Ingesting the TMDB dataset from Kaggle; Cleaning and transforming the data by filtering missing values, formatting dates, parsing strings, creating derived features like ROI and profit, and engineering new variables; Exploring the cleaned dataset visually and statistically to uncover insights; And finally, querying the data to extract answers to business questions using SQLite.
The outcome of this project will be a set of strategic, evidence-based recommendations for Netflix's content production team, helping them focus investment on the right types of movies to boost both viewer satisfaction and profitability.

B. Dataset Description

The dataset, sourced from The Movie Database (TMDB) and available on Kaggle, provides a comprehensive collection of movie information. It includes essential details like movie titles, IDs, release dates, status, and runtimes, alongside key performance metrics such as budget, revenue, user vote averages, vote counts, and popularity scores. Furthermore, it contains rich descriptive attributes like genres, keywords, overviews, taglines, and information about the production companies, production countries, and spoken languages, making it a valuable resource for analyzing movie trends, success factors, and building recommendation systems.

II. Approach & Methodology

In this project, we implemented a comprehensive data processing pipeline to analyze the TMDB dataset, which contains over 1 million movie entries. The goal was to assist Netflix in making data-driven content decisions by identifying what types of content lead to higher ratings and profitability.

A. Data Acquisition & Ingestion

The dataset was sourced from Kaggle's TMDB dataset, containing structured data about movies, including information such as title, genre, production companies, ratings, budget, revenue, and release dates. Using Python's pandas library, the dataset was ingested into the project, allowing easy manipulation and analysis. We filtered the dataset to focus on movies with complete financial and rating information, ensuring the data was analysis-ready.

B. Data Storage & Integration

Initially, the dataset was stored locally in CSV format within the Jupyter Notebook environment. This allowed for easy manipulation and quick processing during the analysis phase. For future scalability, we planned integration with a SQLlite database, as its relational database structure is well-suited for efficient querying and handling large datasets.

C. Data Cleaning & Transformation

This phase was essential for ensuring that the data quality was optimal for analysis. Key steps included:

- **Drop duplicate, redundant rows**
- **Drop redundant columns** that are irrelevant to our analysis such as 'homepage','tagline', 'overview', and 'original_title'.
- **Handling Missing Values:** Rows with missing critical fields such as title and release_date were removed, while missing runtime values were imputed with the median.
- **Data Type Conversion:** Dates were parsed using pandas' pd.to_datetime function, and numeric fields such as budget and revenue were properly formatted.

- **Outlier Removal:** Any unrealistic values (e.g., revenue equal to 0) were filtered out to maintain analysis integrity, by imputing them with NaN. These values were then filtered out into a new df_financial dataframe for the analysis of the movies' financial performance, when calculating profit/ROI but are kept for popularity/vote analysis. Rows with movies having future releases - release_date beyond a defined cutoff year - were dropped.
- **Feature Engineering:**
  Release Month and Year were extracted from the release date for better time-based analysis.
  Using the df_financial dataframe, new features were created to assist with analysis, such as:
  - **ROI (Return on Investment)** calculated as revenue divided by budget.
  - **Profit** calculated as revenue minus budget.
  - **Production Company Count** to capture diversity in movie production.

  Comma-separated strings columns were parsed into a list of strings, stripping whitespace from each item, and handling NaN or empty strings. Then, primary categorical information (genre, production company, country) were extracted from these columns into new ones 'main_genre', 'main_company', etc. For rows where the list is empty, values for new columns were imputed with 'Unknown'.

D. Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns in the data. Various visualizations were generated using libraries such as matplotlib and seaborn to explore relationships between movie genres, budget, revenue, and ratings. Key visualizations included:

- **Heatmaps** to explore correlations between different variables (e.g., budget and revenue). Heatmaps were used to show the correlation between genres and production companies, highlighting strong genre-company relationships and the performance of specific studios.
- **Scatterplots** to assess the profitability of movies of different genres and the impact of budget on revenue, enabling clear identification of profitable genres and trends. These visual representations were directly tied to the results from SQL queries, reinforcing the business insights gathered.
- **Correlation Matrices** to identify significant relationships between financial success and other variables like audience ratings and popularity.

E. Querying

- SQL queries played a crucial role in extracting actionable insights from the cleaned and transformed dataset. Using SQLite, we performed specific queries to answer Netflix's business questions. These queries were designed to filter and analyze large datasets efficiently:
- **Profitability by Budget and runtime range:** Using SQL, we generated queries to group movies by Budget and runtime range, calculating the average profit and

rating. This helped identify which Budget and runtime ranges were most aligned with financial success.

- **Profitability by Genre and Production Company:** Using SQL, we generated queries to group movies by genre and production company, calculating the average revenue, budget, and profitability. This helped identify which genres were most aligned with financial success and which production companies consistently delivered high returns.
- **Top Movies by Rating and Revenue:** Queries were also crafted to select movies with the highest average viewer ratings or revenue, sorted by genre or production company, to offer strategic insights into what works well for Netflix in terms of content types and partnerships.
- These insights were critical for guiding Netflix's content strategy, specifically in terms of optimizing the mix of genres, production companies, and financial investments.

F. Storage & Archival Strategy

- During the course of the project, the dataset and resulting outputs were stored both locally and on Github to ensure version control and easy access for collaboration among team members.
- **Local Storage:** The dataset was initially processed in CSV format within the Jupyter Notebook. Intermediate steps, such as cleaned and transformed data, were saved as new CSV files.
- **Google Cloud:** For secure cloud-based storage, all finalized datasets and SQL database were uploaded to Google Cloud in a bucket. This not only ensured that files were safely stored. In addition to that some queries were done on BigQuery where we got the same results locally and on cloud.

III. Results & Analysis

To achieve our objective, successfully recommending best movie practices for Netflix’s production team to attain higher ratings and profitability, two methods will be utilized on our large dataset to extract actionable insights — EDA (using matplotlib and seaborn libraries) and Queries (via SQLite).

A. From EDA

Figures Figure 1: Correlation Matrix (EDA) and Figure 2: Scatterplots for key metrics (EDA) both highlight that while greater production budgets are in general correlated with greater revenues and profits, this relationship is not a guarantee. Actually, a lot of high-budget movies underperform, which reveals the risk of diminishing returns at scale. Although revenue is still a good indicator of profit, there are weak or erratic correlations between financial success and other variables like popularity, vote average, and runtime.

Actionable Insights: These findings reveal that Netflix needs to do more than just invest in pricey or highly acclaimed content. Rather than relying solely on prestige appeal or social buzz, budgets should be strategically matched with audience demand, genre-specific benchmarks, and expected return on investment (ROI).

The Adventure, Family, Fantasy, and Science Fiction genres are consistently very profitable and popular, as seen in Figures Figure 3: Most profitable genres (median) and Figure 4: Most popular genres (median), suggesting a strategic sweet spot for content investment. These genres frequently draw large, international audiences, particularly younger ones, and are successful without exclusively depending on substantial production budgets.

Actionable Insights: The presence of this overlap between financial return and audience engagement strongly suggests that Netflix should focus on franchise-capable, visually immersive, and family-friendly content. Prioritizing these genres can aid the platform in balancing profitability with sustained viewer appeal, which in return supports subscription growth and retention efforts in a more than ever competitive streaming landscape.

The heatmap in Figure Figure 5: Heatmap (Genre vs Company) illustrates how the strength of the production studio has a significant impact on revenue performance in addition to being genre-driven. Walt Disney Pictures leads with the highest median revenues in categories like adventure, romance, and thrillers, demonstrating its unparalleled skills in global distribution, franchise building, and audience trust. Strong genre-studio alignment is also shown by other big studios, such as Universal and Paramount, especially in Family and Adventure films. On the other hand, Metro-Goldwyn-Mayer and other smaller studios' poor success emphasizes the value of scale and brand leverage.

Actionable Insights: For Netflix, this highlights the necessity to not solely invest in high-performing genres but to copy the production levels of the best studios. This includes poaching top creative and production talent, co-producing with high-performing studios, or acquiring IP tied to proven genre success. By taking such actions, Netflix would be able to increase the original content's commercial viability while keeping control over its quality and attractiveness to viewers.

B. From Queries

Figures Figure 6: Most profitable genres (Avg) and Figure 7: Most profitable movies per language (Avg) reveal that genres such as Adventure, Science Fiction, Animation, Fantasy, and Family consistently bring in the highest average profits, which reflects high audience demand for visually immersive, globally relatable storylines. On the other hand, high-volume categories like Comedy and Thriller deliver lower returns, suggesting a necessity for more selectivity in these highly saturated spaces. In addition, while some barely known languages generate exemplary profits, their small sample sizes make them unreliable indicators. More ironclad opportunities lie in East Asian languages like Cantonese, Mandarin, and Japanese, which show sustained profitability and audience reach.

Actionable Insights: For Netflix, this calls for a dual strategy: focus on the high-performing genres with global reach and scalability, and attempt scaling up localized production in Asia’s highest performing language markets, where not just regional but international streaming value can be captured.

Figures Figure 8: Profit & ratings per budget range and Figure 9: Performance by runtime illustrate that budget size and runtime are closely linked to profitability and good ratings. Movies with budgets over $200M and runtimes above 150 minutes bring in the highest returns and audience ratings, but even movies in the $50M-$150M budget range and 130-149 minute runtime perform really well, which suggests that there is a sweet spot where scale aligns with efficiency. However, shorter and lower-budget films tend to drive modest financial outcomes and lower ratings.

Actionable Insights: For Netflix, this calls for a strategy of selectively creating and investing in mid-to-high-budget, long duration content. Of course, within proven genres where rich storytelling and production quality can generate both commercial success and buzz.

Figure Figure 10: Performance by production country's analysis reveals that the United States, China, and Japan are the top three countries in terms of average movie profitability, with Japan leading the pack at $123M per film. Notably, Asian markets dominate the top ranks, with Hong Kong, Taiwan, and India all achieving competitive profits and strong average ratings. The U.S. continues to be a volume leader, but its profitability per film lags behind key Asian counterparts.

Actionable Insights: This presents Netflix with a significant chance to increase its investments in or co-produce content from high-performing Asian markets, particularly China and Japan, where viewer engagement and financial returns are strong. In these nations, customized regional approaches may have significant financial and cultural benefits.

The query in Figure Figure 11: Performance by production company finds production companies that have scored highly on audience ratings during the previous ten years. Walt Disney Animation Studios (7.40), Marvel Studios (7.37), and Pixar (7.52) top the list, all of which have consistently achieved remarkable average ratings for a number of their films. These studios are known for their strong narratives and devoted fan bases, but they also worked in genres like family entertainment, animation, and superhero franchises.

Actionable Insights: This highlights the importance of consistent quality over time for Netflix and implies that collaborating with or imitating these businesses' creative standards, whether through co-production agreements, IP licensing, or talent acquisition could boost the platform's reputation and audience satisfaction.

According to the analysis in Figure Figure 12: Average profit of R-rated and PG films, over the past ten years, adult-rated films (such as R-rated or 18+) have underperformed financially, losing an average of $5.6 million per film, while non-adult (general audience) content has made a healthy $36.5 million profit. Even though adult-rated films might have more depth or critical appeal, their lower financial performance is probably caused by their smaller audience, less flexible distribution, and advertising restrictions.

Actionable Insights: This implies that, although some mature content may foster brand diversity, Netflix's core content strategy should give priority to widely accessible, family-friendly, or PG/PG-13 productions because of their much greater potential for scale and return on investment.

Combining genres can greatly increase profitability, as shown by Figure Figure 13: Most profitable genre combinations's analysis, which shows that combinations like Action, Adventure, and Science Fiction can produce average profits of over $500M. Other high-yield combinations include Animation-Comedy-Family and Family-Adventure-Drama, which support previous findings that genre fusion appeals to a wide audience and generates financial success, especially when it involves high-concept, visually driven, or emotionally resonant formats. Notably, Adventure, Fantasy, and Action are frequently found in the top combinations, demonstrating their adaptability and synergy with other genres.

Actional Insights: This means that in order to optimize audience reach and return on investment, Netflix should not only focus on high-performing standalone genres but also purposefully create cross-genre content, particularly in family, sci-fi, and fantasy contexts.

IV. Final Recommendation for Netflix

A. Recommendation with justifications

- Invest in adventure, sci-fi, fantasy, animation, and family films because these genres are known for their high box office receipts and enduring appeal. These appeal to younger, international audiences and provide scalability without always requiring large financial outlays.
- Produce content that spans and combines genres, especially Action, Adventure, Fantasy, and Family mixes, as these genres have been shown to greatly increase average profitability. This approach encourages both audience diversity and creative innovation.
- Find the ideal balance between runtime and budget. Productions with budgets between $50 million and $150 million and runtimes between 130 and 150 minutes should receive the majority of resources because they provide the best balance between cost, story-telling complexity, ratings, and profitability.
- Scale up regional content in Asia. Give top priority to co-productions or acquisitions in languages like Cantonese, Mandarin, and Japanese, which have demonstrated both profitability and potential for international streaming, as well as high-performing Asian markets like China, Hong Kong, and Japan.
- Emulate and collaborate with top studios. Take inspiration from Pixar, Marvel, and Disney Animation Studios, or collaborate with them to match their track record of consistently satisfying audiences by poaching top creative talent, licensing established intellectual property, or co-creating original content with comparable high standards.
- Don't rely too much on adult-rated/R-rated material. Films aimed at adults (18+) do not do well financially. Given its wider marketability and greater potential for profit, Netflix should concentrate its primary investments on family-friendly or PG/PG-13 content.

B. Potential Risks

- Outlier inflation: Averages may be disproportionately skewed by some extremely successful films; further variance analysis could support budget and genre choices.
- Cultural/contextual variation: Localized success doesn't always scale, and success in some markets (like Asia) might not always translate internationally.
- Genre fatigue risk: Without constant innovation, over-saturation of popular genres could result in diminishing marginal engagement.
- Sample size sensitivity: It is important to exercise caution when extrapolating from small groups because some high-profit combinations or languages have low volume.

C. Next Steps

- Run profitability clustering models: Determine profitability segments across genres, runtimes, and regions using k-means or decision trees to guide production targeting.
- Pilot regional co-productions: Start a series of mid-budget joint productions in China, South Korea, and Japan, then track the impact on global engagement and retention.
- Develop internal benchmarks: Using past performance data, establish ROI benchmarks unique to Netflix for genre, budget, and runtime combinations.
- Build a genre fusion content strategy: Establish a framework that supports tried-and-true genre pairings, particularly in franchise-capable verticals like Fantasy Adventure or Sci-Fi Family.
- Conduct regression analysis for interpretability: To measure the marginal effects of runtime, production budget, and other characteristics on revenue and profit, use linear regression models. For instance, based on the market and genre, "Every additional $1M in budget yields $X in revenue, holding all other predictors constant."
- Analyze talent influence: Use performance analytics for actors and directors to identify the people who consistently increase viewership, ratings, and profits. Make use of this to inform casting choices and possible exclusive collaborations.

V. References & Links

- Github link: https://github.com/simonfeghali/TMDB_FINANCIAL/tree/955cf06c754fce602a1bc6d91be13742da1dfa91
- Kaggle link for the dataset: https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies/code
- Queries done on cloud: https://console.cloud.google.com/bigquery?ws=!1m7!1m6!12m5!1m3!1smsba305project-457316!2sus-central1!3sae9dcfcf-376b-4fd1-abb0-df0c0aea89d4!2e1

VI. Appendix A – Graphs & Tables

Figure 1: Correlation Matrix (EDA)

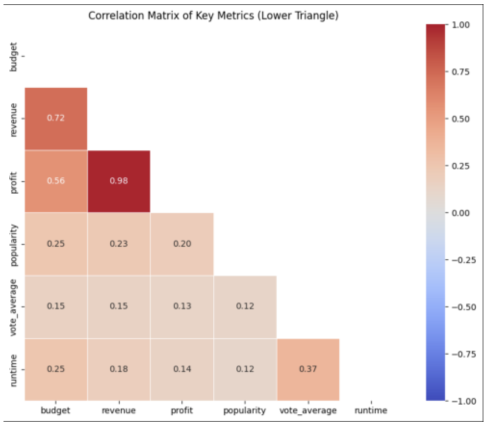


Figure 2: Scatterplots for key metrics (EDA)

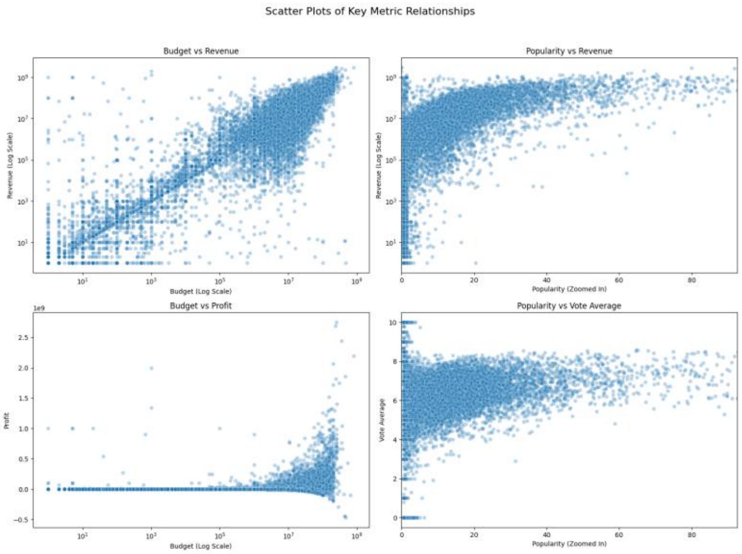


Figure 3: Most profitable genres (median)

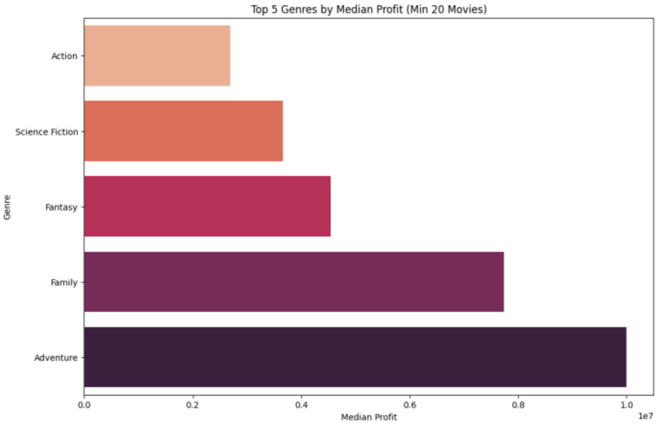


Figure 4: Most popular genres (median)

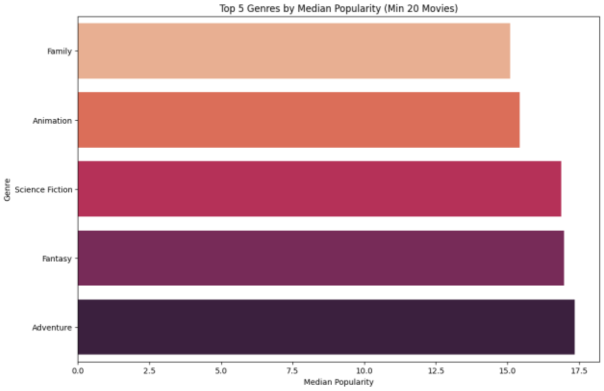
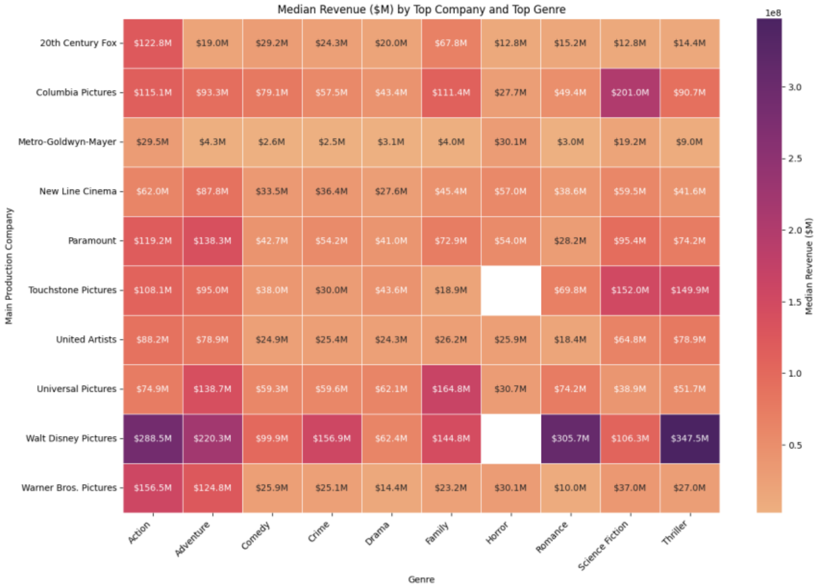
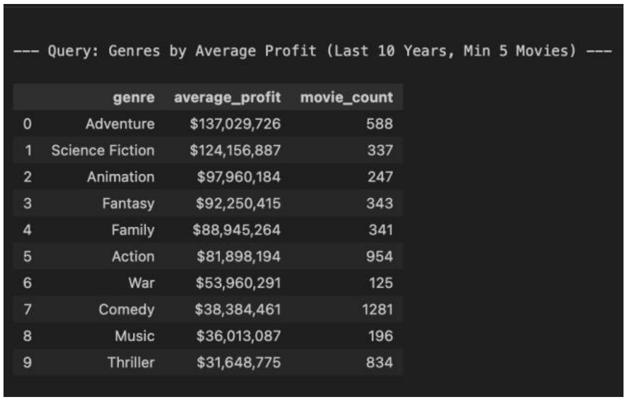


Figure 5: Heatmap (Genre vs Company)



	Action	Adventure	Comedy	Crime	Drama	Family	Horror	Romance	Science Fiction	Thriller
20th Century Fox	$122.8M	$19.0M	$29.2M	$24.3M	$20.0M	$67.8M	$12.8M	$15.2M	$12.8M	$14.4M
Columbia Pictures	$115.1M	$93.3M	$79.1M	$57.5M	$43.4M	$111.4M	$27.7M	$49.4M	$201.0M	$90.7M
Metro-Goldwyn-Mayer	$29.5M	$4.3M	$2.6M	$2.5M	$3.1M	$4.0M	$30.1M	$3.0M	$19.2M	$9.0M
New Line Cinema	$62.0M	$87.8M	$33.5M	$36.4M	$27.6M	$45.4M	$57.0M	$38.6M	$59.5M	$41.6M
Paramount	$119.2M	$138.3M	$42.7M	$54.2M	$41.0M	$72.9M	$54.0M	$28.2M	$95.4M	$74.2M
Touchstone Pictures	$108.1M	$95.0M	$38.0M	$30.0M	$43.6M	$18.9M		$69.8M	$152.0M	$149.9M
United Artists	$88.2M	$78.9M	$24.9M	$25.4M	$24.3M	$26.2M	$25.9M	$18.4M	$64.8M	$78.9M
Universal Pictures	$74.9M	$138.7M	$59.3M	$59.6M	$62.1M	$164.8M	$30.7M	$74.2M	$38.9M	$51.7M
Walt Disney Pictures	$288.5M	$220.3M	$99.9M	$156.9M	$62.4M	$144.8M		$305.7M	$106.3M	$347.5M
Warner Bros. Pictures	$156.5M	$124.8M	$25.9M	$25.1M	$14.4M	$23.2M	$30.1M	$10.0M	$37.0M	$27.0M

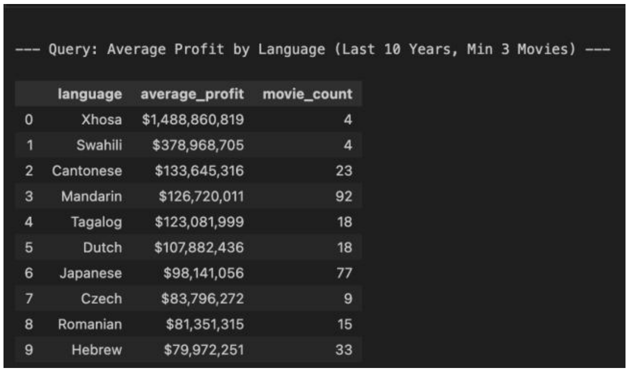
Figure 6: Most profitable genres (Avg)



```
--- Query: Genres by Average Profit (Last 10 Years, Min 5 Movies) ---

             genre  average_profit  movie_count
0        Adventure    $137,029,726          588
1  Science Fiction    $124,156,887          337
2        Animation     $97,960,184          247
3          Fantasy     $92,250,415          343
4           Family     $88,945,264          341
5           Action     $81,898,194          954
6              War     $53,960,291          125
7           Comedy     $38,384,461         1281
8            Music     $36,013,087          196
9         Thriller     $31,648,775          834
```

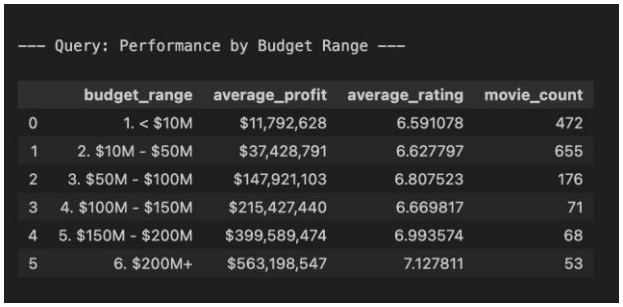
Figure 7: Most profitable movies per language (Avg)



```
--- Query: Average Profit by Language (Last 10 Years, Min 3 Movies) ---

    language  average_profit  movie_count
0      Xhosa  $1,488,860,819            4
1    Swahili    $378,968,705            4
2  Cantonese    $133,645,316           23
3   Mandarin    $126,720,011           92
4    Tagalog    $123,081,999           18
5      Dutch    $107,882,436           18
6   Japanese     $98,141,056           77
7      Czech     $83,796,272            9
8   Romanian     $81,351,315           15
9     Hebrew     $79,972,251           33
```

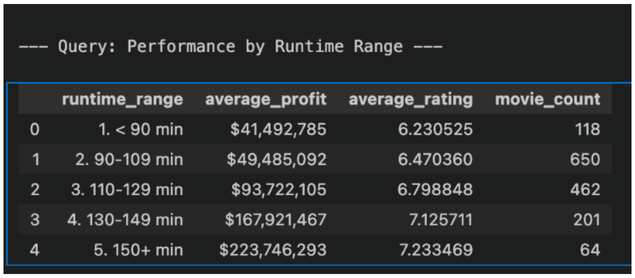
Figure 8: Profit & ratings per budget range



```
--- Query: Performance by Budget Range ---
```

	budget_range	average_profit	average_rating	movie_count
0	1. < $10M	$11,792,628	6.591078	472
1	2. $10M - $50M	$37,428,791	6.627797	655
2	3. $50M - $100M	$147,921,103	6.807523	176
3	4. $100M - $150M	$215,427,440	6.669817	71
4	5. $150M - $200M	$399,589,474	6.993574	68
5	6. $200M+	$563,198,547	7.127811	53

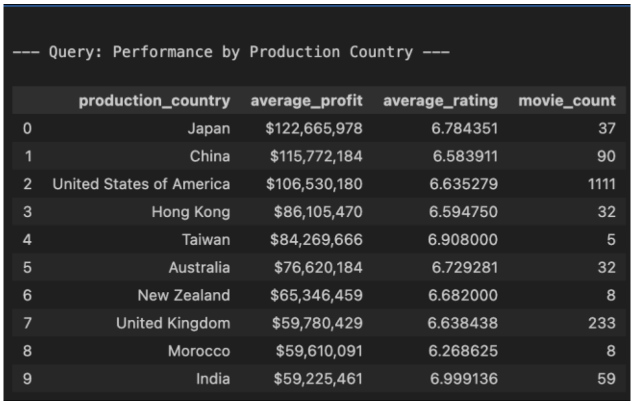
Figure 9: Performance by runtime



```
--- Query: Performance by Runtime Range ---
```

	runtime_range	average_profit	average_rating	movie_count
0	1. < 90 min	$41,492,785	6.230525	118
1	2. 90-109 min	$49,485,092	6.470360	650
2	3. 110-129 min	$93,722,105	6.798848	462
3	4. 130-149 min	$167,921,467	7.125711	201
4	5. 150+ min	$223,746,293	7.233469	64

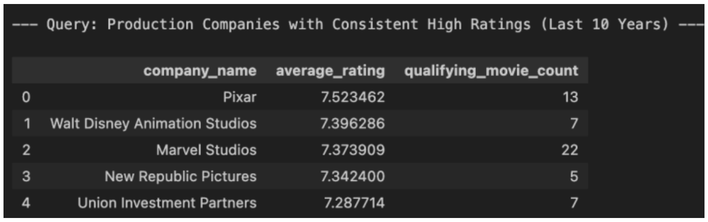
Figure 10: Performance by production country



```
--- Query: Performance by Production Country ---
```

	production_country	average_profit	average_rating	movie_count
0	Japan	$122,665,978	6.784351	37
1	China	$115,772,184	6.583911	90
2	United States of America	$106,530,180	6.635279	1111
3	Hong Kong	$86,105,470	6.594750	32
4	Taiwan	$84,269,666	6.908000	5
5	Australia	$76,620,184	6.729281	32
6	New Zealand	$65,346,459	6.682000	8
7	United Kingdom	$59,780,429	6.638438	233
8	Morocco	$59,610,091	6.268625	8
9	India	$59,225,461	6.999136	59

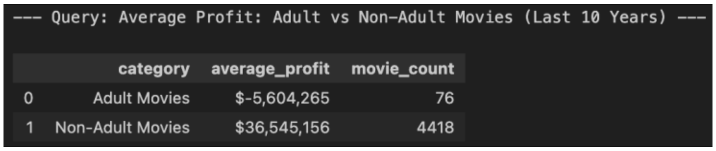
Figure 11: Performance by production company



```
--- Query: Production Companies with Consistent High Ratings (Last 10 Years) ---
```

	company_name	average_rating	qualifying_movie_count
0	Pixar	7.523462	13
1	Walt Disney Animation Studios	7.396286	7
2	Marvel Studios	7.373909	22
3	New Republic Pictures	7.342400	5
4	Union Investment Partners	7.287714	7

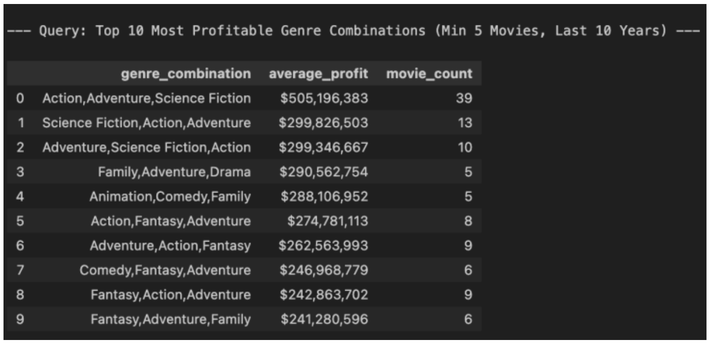
Figure 12: Average profit of R-rated and PG films



```
--- Query: Average Profit: Adult vs Non-Adult Movies (Last 10 Years) ---

           category  average_profit  movie_count
0      Adult Movies     $-5,604,265           76
1  Non-Adult Movies     $36,545,156         4418
```

Figure 13: Most profitable genre combinations



```
--- Query: Top 10 Most Profitable Genre Combinations (Min 5 Movies, Last 10 Years) ---

                 genre_combination  average_profit  movie_count
0  Action,Adventure,Science Fiction    $505,196,383           39
1  Science Fiction,Action,Adventure    $299,826,503           13
2  Adventure,Science Fiction,Action    $299,346,667           10
3           Family,Adventure,Drama    $290,562,754            5
4          Animation,Comedy,Family    $288,106,952            5
5         Action,Fantasy,Adventure    $274,781,113            8
6         Adventure,Action,Fantasy    $262,563,993            9
7         Comedy,Fantasy,Adventure    $246,968,779            6
8         Fantasy,Action,Adventure    $242,863,702            9
9          Fantasy,Adventure,Family    $241,280,596            6
```