Film Data Analysis for Microsoft

Flatiron School Data Science Phase 1 Project

Student Name: Rafael V Rabinovich (mailto:rafvrab@gmail.com)

Student Pace: Flex pace

Instructors: Morgan Jones, Mark Barbour

Blog URL: https://medium.com/@rafvrab (https://medium.com/@rafvrab)

Business Understanding

Microsoft's venture into the film-making industry has prompted a comprehensive analysis of provided datasets to deliver actionable recommendations. Commissioned by Microsoft, our task is to delve into the complexities of the movie industry. Specifically, our goal is to conduct data analysis aimed at uncovering the key factors driving successful box office performance. These insights will serve as a compass, guiding strategic decisions for Microsoft's upcoming movie studio.

The primary stakeholders vested in this analysis are the Board of Directors at Microsoft. Our findings will play a pivotal role in shaping their decision-making processes, aiding in the identification of lucrative film genres, potential directors, and critical success factors for maximizing movie performance.

Beyond the scope of analysis, this project holds immense significance by offering actionable insights that empower Microsoft to curate a portfolio of high-potential movies.

Data Understanding

Data Sources Overview

The project utilizes the following data files:

- **Box Office Mojo**:
  - File: `bom.movie_gross.csv.gz`
  - Size: 53,544 bytes

- **IMDB**:
    - File: `im.db.zip`
    - Size: 67,149,708 bytes
    - Detailed tables:
        - `df_mb` (movie_basics): 146,144 rows, 6 columns
        - `df_dir` (directors): 291,174 rows, 2 columns
        - `df_kf` (known_for): 1,638,260 rows, 2 columns
        - `df_akas` (movie_akas): 331,703 rows, 8 columns
        - `df_ratings` (movie_ratings): 73,856 rows, 3 columns
        - `df_persons` (persons): 606,648 rows, 5 columns
        - `df_principals` (principals): 1,028,186 rows, 6 columns
        - `df_writers` (writers): 255,873 rows, 2 columns
- **Rotten Tomatoes**:
    - Movie information (file: `rt.movie_info.tsv.gz`)
        - Size: 498,202 bytes
        - Shape: 156 rows, 12 columns (`id`, `synopsis`, `rating`, `genre`, `director`, `writer`, `theater_date`, `dvd_date`, `currency`, `box_office`, `runtime`, `studio`)
    - Reviews (file: `rt.reviews.tsv.gz`)
        - Size: 3,402,194 bytes
        - Shape: 54,432 rows, 8 columns (`id`, `review`, `rating`, `fresh`, `critic`, `top_critic`, `publisher`, `date`)
- **The Movie DB**:
    - File: `tmdb.movies.csv.gz`
    - Size: 827,840 bytes
    - Shape: 26,517 rows, 10 columns (`Unnamed: 0`, `genre_ids`, `id`, `original_language`, `original_title`, `popularity`, `release_date`, `title`, `vote_average`, `vote_count`)
- **The Numbers**:
    - File: `tn.movie_budgets.csv.gz`
    - Size: 153,218 bytes
    - Shape: 5,782 rows, 6 columns (`id`, `release_date`, `movie`, `production_budget`, `domestic_gross`, `worldwide_gross`)

Data Exploration Process

The data exploration process involved opening and exploring the datasets using Python libraries for initial insights. The conclusions derived from this exploration are summarized above in the tables presented in this section.

The project utilizes the following data files:

Data Preparation

Data Preparation and Cleaning

This section outlines the steps involved in preparing and cleaning the dataset for analysis. The dataset consists of various movie-related tables obtained from different sources.

Retrieval and Unzipping

The provided notebook includes instructions and code to retrieve the raw data and unzip the files containing movie-related information from various sources. Multiple datasets were acquired, such as Box Office Mojo, Rotten Tomatoes movie info and reviews, The Movie DB, and The Numbers, among others.

Exploring SQL Database Tables

Utilizing SQL database files, the notebook connected to the IMDb database to retrieve and explore its tables. Detailed information regarding the database tables' contents and their corresponding Pandas DataFrames is provided.

Statistical Analysis

Various statistical analyses were performed on different DataFrames derived from the dataset, showcasing descriptive statistics for attributes such as movie ratings, runtime, and other relevant movie-related data.

Data Shape and Structure

The README presents tabular forms displaying the shape and structure of each table, including the number of rows and columns they contain. Detailed descriptions of column titles and their respective tables are also provided for clarity.

Frame Mergers

The process of merging different DataFrames to combine relevant information into a single consolidated DataFrame is described step-by-step. This merging process aimed to link directors, actors, movie titles, genres, ratings, and other critical information into a unified dataset for analysis.

Data Cleaning

Following the merging process, steps were taken to clean the dataset, including the removal of duplicates and unnecessary columns. The cleaned dataset focuses on retaining essential information required for analysis while eliminating redundant or irrelevant data points.

Merger of Other Data Frames

The following code snippets demonstrate the process of merging different DataFrames and performing relevant data cleaning steps:

Merging 'movie_basics' with 'movie_ratings'

The code snippet shows how the DataFrames 'df_mb' and 'df_ratings' are merged using the `pd.merge()` function on the 'movie_id' column with a "left" merger.

Checking for Duplicates and Cleanliness

The code checks for duplicates in the resulting merged DataFrame 'df_im_mgd'. It also verifies if there are any null values present in the dataset.

DataFrame Shape and Display

The code snippet provides insights into the shape of the DataFrame 'df_im_mgd' and displays the first few rows of the merged DataFrame to visualize the data.

Merging 'df_mg', 'db_movies', and 'db_movie_budgets'

Demonstrates the merging process of 'df_mg', 'db_movies', and 'db_movie_budgets' DataFrames. It explains the merging methodology and drops unnecessary columns from the resulting DataFrame 'unified_df'.

Final Merging Process

Displays the final merging process by merging 'unified_df' with 'db_movie_budgets' based on specified columns. It drops redundant columns and displays the resulting 'final_df' DataFrame.

Additional Merging with 'filtered_merged_df'

Describes the additional merging process with 'filtered_merged_df' based on the 'primary_title' and 'title' columns, displaying the 'merged_final_df' DataFrame.

Data Preparation for `merged_final_df`

The data preparation phase involved cleaning the dataset and organizing it for analysis. Columns were converted to numeric values by removing commas and dollar signs from the 'worldwide_gross' and 'production_budget' columns. Additionally, the desired column order was defined to structure the dataset for further analysis.

Exploratory Data Analysis

Return on Investment (ROI) Calculation

The ROI and net profit for each movie were computed using the provided formulas: Net Profit = Gross Revenue - Budget, and ROI = (Net Profit / Budget) * 100. These metrics were calculated to gauge the profitability of movies in the dataset.

Data Visualization

Top 25 Film Genres by ROI

A bar plot was generated to showcase the top 25 film genres based on their mean Return on Investment (ROI). This visualization helps identify genres that tend to yield higher returns, aiding in decision-making for potential movie productions.

Various Metrics Analysis

Multiple bar plots were created to analyze different metrics such as profit, ROI, average rating, number of votes, and runtime minutes across the dataset. These visualizations offer insights into the distribution and trends of these metrics, assisting in understanding the dataset's characteristics.

Production Budget Analysis and Categorization

To strategize production budget recommendations, the dataset was categorized into three brackets: lower, middle, and higher budget segments. These categories were derived based on quartiles of the 'production_budget'. A bar plot visualizes the median production budgets for each category, providing a clear comparison of budget ranges.

Genre Analysis

The analysis delved into finding the best Return on Investment (ROI) per film across different genres within each budget bracket. The top three genres with the best ROI were identified for lower, middle, and higher budget films. This insight is valuable for understanding the most profitable genres within specific budget constraints.

Seasonal Analysis

Understanding the influence of release months on ROI, the analysis grouped films by month and year, calculating the average ROI. This was visualized through a line plot showcasing the average ROI over time, enabling insights into ROI fluctuations across different months and years.

Further exploring monthly ROI trends within budget categories, the analysis highlighted the top three performing months in terms of average monthly ROI for lower, middle, and higher budget films. This insight assists in understanding the seasonality impact on ROI based on budget categories.

Staff Analysis

Exploring staff roles and their impact on the Return on Investment (ROI), the analysis focused on different professions within the film industry. Initially, the dataset was filtered to include professionals with specific roles and relevant birth and death years.

The unique values in the 'primary_profession' column were extracted, showcasing a wide array of professions ranging from directors, producers, and writers to actors, editors, and various other specialties.

To provide a clearer understanding, the individual professions were separated, resulting in a condensed list that presents specific roles such as director, producer, writer, actor, and others.

Visualizing the average ROI per profession in a bar plot offered insights into the varying impacts of different roles within the film industry. The top 10 professions with the highest average ROI were identified, shedding light on the key roles that tend to generate higher returns.

Further analysis delved into identifying the top three individuals for each top-performing profession within different budget categories (Lower, Middle, Higher Budget). This involved filtering alive individuals with the highest ROI for specific professions within each budget category, highlighting notable contributors in different roles.

The findings present key professionals across different budget categories, emphasizing their significance in achieving higher ROIs within the film industry.

Recommendations

The recommendations section synthesizes key insights and offers strategic guidance for film production across three distinct budget brackets: Lower, Middle, and Higher Budget.

Lower Budget Recommendations

Production Budget Range: 1 Million to 15 Million US$

Genre Recommendations

1. Best Recommendation: Comedy, Romance, Sport
2. Second Recommendation: Drama, Fantasy
3. Third Recommendation: Horror

Seasonal Recommendations

1. Best Recommendation: February
2. Second Recommendation: August
3. Third Recommendation: May

Staff Recommendations

1. Director, Actor, Camera Department: Levan Gabriadze
2. Miscellaneous, Production Manager, Producer: Jamie Buckner
3. Producer, Writer, Cinematographer: Tom Boyle

Middle Budget Recommendations

Production Budget Range: 15 Million to 50 Million US$

Genre Recommendations

1. Best Recommendation: Horror, Mystery, Thriller
2. Second Recommendation: Action, Sci-Fi, Thriller
3. Third Recommendation: Comedy, Fantasy

Seasonal Recommendations

1. Best Recommendation: July
2. Second Recommendation: November
3. Third Recommendation: January

Staff Recommendations

1. Director, Producer, Actress: Sam Taylor-Johnson
2. Writer, Music Department, Producer: Seth MacFarlane
3. Actor, Producer, Animation Department: Conrad Vernon

Higher Budget Recommendations

Production Budget Range: Above 50 Million US$

Genre Recommendations

1. Best Recommendation: Biography, Drama, Music
2. Second Recommendation: Action, Biography, Drama
3. Third Recommendation: Adventure, Drama, Sport

Seasonal Recommendations

1. Best Recommendation: April
2. Second Recommendation: June
3. Third Recommendation: July

Staff Recommendations

1. Director, Animation Department, Visual Effects: Kyle Balda
2. Animation Department, Director, Writer: Chris Buck
3. Writer, Miscellaneous, Producer: Jennifer Lee, Jared Bush

Conclusions

The analysis presents actionable insights for optimizing film production strategies across various budget brackets. These insights, ranging from genre preferences and release timings to key staff roles, offer valuable recommendations tailored to different budget categories. Stakeholders can utilize this information as a starting point for informed investment and production decisions.

Limitations

- The analysis is limited to the provided dataset and may not encompass all factors impacting future performance.
- Data limitations, up until 2018, might overlook recent trends within the past six years.

- The analysis provides a high-level overview and might benefit from a more detailed breakdown for enhanced accuracy.
- Financial data hasn't been adjusted for inflation rates, which could influence the final outcomes.

Next Steps

Future steps involve a deeper exploration of the US and Foreign film markets to understand their influence on the industry. Accessing data from the last six years will complement existing datasets for a comprehensive analysis of recent trends. Further exploration of the original datasets may offer insights lost during the merging process, enhancing the completeness and relevance of the research.

Structure Map

GitHub Repository: https://github.com/rafvrab/MovieAnalysis/tree/main (https://github.com/rafvrab/MovieAnalysis/tree/main)

┌ .gitignore
├ Film_Analysis.pdf
├ README.md
└ analysis.ipynb

Data Files Repository: https://github.com/learn-co-curriculum/dsc-phase-1-project-v2-4/tree/master/zippedData (https://github.com/learn-co-curriculum/dsc-phase-1-project-v2-4/tree/master/zippedData)

┌ bom.movie_gross.csv.gz
├ im.db.zip
├ rt.movie_info.tsv.gz
├ rt.reviews.tsv.gz
├ tmdb.movies.csv.gz
└ tn.movie_budgets.csv.gz