Box Office Gold

# Data-Driven Insights for a Profitable Movie Studio Launch

**CRoss Industry Standard Process for Data Mining**

### ***Step 1: Business Understanding***

The goal is to understand the domain of the problem and the project requirements from a business perspective.

***Overview/Background***

The company, Box Office Gold, is looking to enter the film industry by creating a new movie studio. It, however, lacks expertise in creating movies and needs data-driven insights. The film industry is competitive with high financial stakes and risks. Box office performance is a key metric for measuring movie success

***Challenges***

There is a limited understanding of the types of films that are currently successful at the box office, coupled with no prior experience in film production or movie studio operations. Entering the industry involves high investment costs and significant financial risks, especially in a competitive market dominated by well-established major studios and production companies. Additionally, the challenge is compounded by constantly evolving consumer preferences and viewing habits.

***Proposed Solution***

* Analyze the collected data to understand it
  + Gather and organize movie data from trusted industry sources like IMDb, Box Office Mojo, Rotten Tomatoes, The Movie DB, and The Numbers.
* Clean the data
  + Clean and standardize the datasets to ensure consistency and accuracy before analysis.
  + Combine the different data sources into a single, well-structured dataset for deeper insights.
* Exploratory Data Analysis: Perform exploratory data analysis on the structured data to gauge the box office performance
* Predictive Analysis: Identify patterns and trends that correlate with box office success

***Brief Conclusion***

The analysis will provide actionable insights to guide the new movie studio in selecting film types with the highest potential for box office success, thereby reducing risk and increasing the likelihood of profitability.

Launching a successful movie studio requires more than creativity—it demands strategic, data-informed decisions. This project leverages box office analytics to uncover what drives film profitability, helping the company make confident choices about genre, budget, and release strategy. With clear insights and recommendations, the company will be well-positioned to enter the competitive film industry with a strong foundation for success.

***Problem Statement***

Your company now sees all the big companies creating original video content and they want to get in on the fun. They have decided to create a new movie studio, but they don’t know anything about creating movies. You are charged with exploring what types of films are currently doing the best at the box office. You must then translate those findings into actionable insights that the head of your company's new movie studio can use to help decide what type of films to create.

***Objectives***

1. To analyze movie data to identify patterns and trends related to box office performance
2. To identify the highest grossing films.
3. To determine key factors such as genre, budget and runtime that influence revenue
4. To present the findings visually through dashboard and non-technical presentation and give recommendations to the business stakeholders.

### ***Step 2: Data Understanding***

The goal is to collect/gather/mine and explore the data to identify quality issues and gain initial insights.

***Data Sources***

1. **IMDB Database (im.db.zip)**
   * movie\_basics table:
     + movie\_id: Unique identifier for each movie
     + primary\_title: Title of the movie
     + start\_year: Year the movie was released
     + runtime\_minutes: Length of the movie in minutes
     + genres: Categories the movie belongs to
   * movie\_ratings table:
     + movie\_id: Unique identifier for each movie
     + averagerating: Average user rating (likely on a scale of 1-10)
     + numvotes: Number of votes contributing to the rating
2. **Box Office Mojo Data (bom.movie\_gross.csv.gz)**
   * title: Name of the movie
   * studio: Production studio
   * domestic\_gross: Box office revenue within the domestic market (likely USD)
   * foreign\_gross: Box office revenue in international markets (likely USD)
   * year: Release year
3. **Additional Potential Data Sources**
   * Rotten Tomatoes (critic and audience scores)
   * TheMovieDB (additional metadata)
   * The Numbers (detailed financial information)

***Data Quality Issues to Check For:***

* **Uniformity**: Consistent naming conventions across datasets (movie titles, studios)
* **Null values**: Missing data in key fields like gross revenue or ratings
* **Outliers**: Unusually high or low box office performances that may skew analysis
* **Missing values**: Incomplete records for certain movies or years
* **Data currency**: How recent is the data and does it reflect current market trends
* **Duplicates**: Same movies appearing multiple times with different spellings or formats
* **Data types**: Ensuring numerical data is formatted correctly for analysis

### ***Step 3: Data Preparation***

The goal for this step is to clean, write, and transform the data to gain deeper insights and information as well as having it in a format that is appropriate for the model.

***Data Selection and Integration***

* Join IMDB tables using the relevant primary and foreign keys
* Using the box office data from BOM
* Use additional data sources as needed based on relevance to analysis

***Data Cleaning Steps***

* Handle missing values in critical fields (especially financial data)
* Remove duplicates arising from different versions of the same film
* Standardize movie titles across datasets
* Convert currency values to a consistent format
* Convert the dates to the correct format
* Normalize studio names to account for subsidiaries and acquisitions

***Feature Engineering***

* Calculate total\_gross by combining domestic\_gross and foreign\_gross

***Data Format Transformations***

* Convert date strings to datetime objects
* Ensure all financial figures are floats
* Convert categorical variables to appropriate formats for analysis
* Normalize numerical features where appropriate

***Distribution Analysis***

* Plot distribution of box office grosses to identify skewness
* Check for imbalanced representation of genres and other key categories
* Analyze the temporal distribution of releases and performance

***Analytical Approaches***

* Descriptive statistics for box office performance by genre, rating, studio, etc.
* Trend analysis of film performance over time
* Correlation analysis between film characteristics and financial performance
* Segmentation analysis to identify patterns across film categories

***Visualization Techniques***

* Bar charts comparing average revenue by genre
* Scatter plots showing relationship between ratings and box office performance
* Time series plots of box office trends
* Heat maps of correlation between different variables
* Box plots for revenue distribution across different categories
* Regression plots to show correlation between two different variables

***Key Metrics***

* Domestic and foreign box office gross
* Total gross revenue (where budget data is available)
* Audience and critic rating
* Genre performance consistency

### ***Step 4: Evaluation***

The goal is to assess the results against business objectives and determine the next steps.

***Results Assessment***

* Evaluate findings against initial business objectives
* Verify the statistical significance of observed patterns
* Consider the market context and external factors
* Assess the limitations of the analysis and potential biases
* Identify the most reliable and actionable insights

***Business Value Assessment***

* Quantify the potential financial impact of recommendations
* Estimate risk factors for different film types
* Evaluate alignment with company capabilities and resources
* Consider the competitive landscape for recommended film categories

***Key Findings Documentation***

* Document the three most significant patterns discovered
* Highlight unexpected or counter-intuitive findings
* Specify constraints and considerations for each recommendation
* Identify knowledge gaps requiring further research

### ***Step 5: Deployment***

The goal is to organize and present the knowledge gained in a way that the business can use.

***Deliverables***

1. A Non-technical presentation featuring three key visualizations and business recommendations
2. Jupyter Notebook with complete technical analysis and documentation
3. GitHub repository with organized project files and documentation

***Implementation Plan***

* Provide clear, actionable guidance for film type selection
* Outline considerations for production budgeting based on findings
* Suggest timing strategies for film releases
* Recommend approaches for measuring success against predictions

***Future Work***

* Identify opportunities for ongoing analysis as new films are released
* Suggest additional data sources that could strengthen future analyses
* Recommend potential areas for more specialized analysis (e.g., marketing effectiveness, specific genre deep-dives)
* Outline process for monitoring and updating recommendations as market conditions change

***Important Links***

***Trello:*** <https://trello.com/b/ILhuizXU/group-4-phase-2-project>

***Git Hub:*** <https://github.com/KarimzKarimz/Group4-Phase2-Project>