**Box Office Blueprint: Data-Driven Insights for Microsoft’s Movie Studio Launch**

## Box Office Intel for Microsoft’s Studio Launch – Tittle

[GitHub Link](https://github.com/Hafsa-mohamed-Ad3n/Box-Office-Blueprint-Data-Driven-Insights-for-Microsoft-s-Movie-Studio-Launch) -

## Group Members:

1. Lynn – Group Leader & Presentations
2. Hafsa M. Aden - Statistics
3. Liz Ogutu - Visualization
4. Ryan – Git Hub
5. Rose – SQL

**Main Goal:**

Demonstrate our skills in:

1. Data Communication
2. Creating Jupyter Notebooks
3. Data Manipulation and Analysis using Pandas and Tableau

What we must Submit:

* Three PDFs: We’ll need to upload the following to Canvas:

1. Presentation (presentation.pdf)

* Create slides summarizing the project and export it as PDF.
* Rename to `presentation.pdf`
* Add it to our GitHub repo and push.
* Start with a clear introduction to the project.
* End with a distinct, persuasive conclusion that reinforces the project’s value.

1. Jupyter Notebook (notebook.pdf)

* Run all cells so outputs are visible.
* Save or print the notebook as a PDF.
* Alternatively, use an online converter or `nbconvert` for a clean PDF.

1. GitHub Repository (github.pdf)

* Go to your GitHub repo page in Chrome.
* Use "Print to PDF" in the browser to save the webpage.
* Name it something like `github.pdf`.
* Your GitHub URL - Paste the link to our GitHub repository in the comment box when submitting on Canvas.

**Grading Criteria (Rubric):**

What “Exceeds Objective” Means In each of the criteria

* Data Communication – We create and explain good visualizations that clearly answer business questions
* Jupyter Notebooks – We use Markdown + comments well to make your notebook easy to read and well-organized
* Pandas & Tableau – We use both tools properly to manipulate data and answer questions efficiently and clearly

## Project Overview

State the project goals clearly — what problem are we solving?

Describe the data used — where it came from, and why it matters.

Explain the methods — briefly describe our approach (e.g., analysis, modeling) in non-technical terms.

Present the results — show what we discovered or achieved.

explain why it matters to the business (saves time, reduces costs, improves accuracy, drives decisions).

In conclusion Summarize the Project’s Impact - Recap how the project helps solve the problem. Re-emphasize the value your solution brings to the stakeholders.

This project was assigned as Phase 2 Moringa School’s Data Science Program. We were presented with a business scenario in which Microsoft is launching a new movie studio and wants to understand what types of films tend to succeed at the box office.

As a team of data scientists, we were required to:

* Explore public movie datasets (Rotten Tomatoes, IMDb, TMDB, and Box Office Mojo),
* Analyzing factors that influence box office performance,
* Define success primarily in terms of gross revenue and profit
* And deliver a data-driven presentation with actionable insights to inform strategic decision-making by a hypothetical CEO regarding the types of films Microsoft should consider producing.

## Project Objectives

1. Identify key characteristics that influence a film’s box office success.
2. Use public movie data to analyze trends in film profitability and revenue.
3. Determine which genres, release months, runtimes, ratings, and budgets are linked to higher profits or gross earnings.
4. Translate data findings into strategic recommendations for Microsoft’s entry into film production.
5. Communicate results using clear visualizations and storytelling tailored for executive decision-making.

## Analysis

**Profitability & Revenue Drivers –**

Which genres of movies are the most profitable?

* Genre Analysis
* Which genres tend to generate the highest gross and profit?
* Are some genres more consistently profitable than others?
* Budget vs. Revenue
* What is the relationship between a film’s budget and its gross revenue/profit?
* Is there a point of diminishing returns (e.g., high budget but low profit)?
* Studio Performance
* Which studios produce the most profitable films?
* Do certain studios specialize in specific high-performing genres?
* Release Timing
* What are the best months/seasons to release a movie for maximum profitability?
* Are summer or holiday releases more financially successful?

**2. Film Attributes & Box Office Success**

* Runtime
* Is there a correlation between a movie’s runtime and its performance at the box office?
* Ratings
* How do IMDB ratings relate to financial success?
* Do critically acclaimed films perform better?
* Country of Release
* Do films released in certain regions or countries (from imdb.title.akas) perform better?

**3. Talent Influence (People Involved)**

What effect does Director, Actor and Producer choice have on movie grossing?

* Director & Actor Analysis
* Are there directors or actors who are consistently associated with successful films?
* Does having a well-known director or cast significantly influence box office success?
* Crew Composition
* Is there any pattern in crew composition (e.g., number of writers, producer involvement) that correlates with success?

**4. Temporal Trends**

* Trends Over Time
* Are certain types of movies becoming more profitable over the years?
* How have budget and gross changed over time?

**Suggested Visualizations**

1. Bar charts of average profit by genre/studio
2. Scatter plots of budget vs gross
3. Box plots of profit by release month
4. Line graphs for trends over time
5. Heatmaps showing combinations (e.g., genre + month)

**Further Areas of Exploration**

1. Deeper Insights into Film Budgeting and Finances

While our analysis uses available budget and gross data as proxies for success, film production budgets are often simplified in public datasets. They may not fully reflect marketing, distribution, or post-production costs, which can significantly affect profitability. A more granular understanding of film financing would allow for more precise calculations of profitability and risk. Future work could involve sourcing or scraping richer financial data to gain this perspective.

2. The Rise of Streaming Platforms

With the growing dominance of streaming services like Netflix, Amazon Prime, and Disney+, it’s increasingly important to consider how these platforms may disrupt traditional box office models. Future analysis could explore how streaming availability correlates with reduced box office gross, or whether certain genres are shifting to become more streaming-friendly. This could open up insights into potential hybrid release strategies or licensing decisions for Microsoft’s new studio.

3. Expanding the Definition of “Success”

In this project, we have primarily used gross revenue and profit as our success metrics. However, a more comprehensive analysis might include other key performance indicators such as:

* Return on Investment (ROI): Profit relative to budget
* Audience ratings and critical reception (e.g. IMDb, Rotten Tomatoes)
* Streaming viewership data (if available)
* Award nominations/wins as indicators of prestige or long-term brand value

Marketing and Social Media Influence

Marketing budgets and online engagement (such as social media buzz, trailer views, or sentiment analysis from platforms like Twitter or YouTube) are increasingly correlated with a film's success. Future work could incorporate external marketing or engagement metrics to understand the role of pre-release hype in driving box office performance

## Git & GitHub

* Create A repository – Share link with group
* Add collaborators
* Clone the Repository to Your Computer
* Before starting work, always pull the latest version:
* git pull origin main
* ⚠️ Always git pull before you start working
* ✅ Always git add . → git commit -m "what you did" → git push when done
* Make changes (e.g. Jupyter notebooks, data cleaning scripts, etc.)
* Add and commit your work locally:
* git add.
* git commit -m "Brief message about what you changed"
* Push changes to GitHub:
* git push origin main

**IMPORTANT TIPS FOR COLLABORATION**

* Always pull before you push
* Use clear commit messages
* Avoid working on the same notebook simultaneously
* Create a shared document for task tracking (e.g., Google Docs or Notion)
* Use branches if multiple people want to work on the same file at once

**RECOMMENDED STRUCTURE**

Phase 2 Project

* Data
* notebooks # Jupyter notebooks per person
* 01\_eda.ipynb
* 02\_modeling.ipynb
* visuals # Charts, images, presentation graphics
* README.md # Project overview and instructions
* Project title
* Team members
* Project description
* Objectives
* Data sources
* Insights / Key findings
* .gitignore # Files to ignore (e.g., large datasets)

Incorporating a multi-metric framework would allow Microsoft to tailor its production strategy not only for profitability but also for audience impact, brand equity, and longevity.

## Deliverables

- `notebook.ipynb`: Jupyter Notebook with analysis

- `presentation.pdf`: Slide deck summarizing our work

- `github.pdf`: PDF snapshot of this GitHub repo

Cloning The repository

Tableau Dashboard

## Data sets

| **File Name** | **What It Is** | **How to Open in Python** |
| --- | --- | --- |
| **bom.movie\_gross.csv.gz** | Compressed CSV (Box Office Mojo data) | Use pd.read\_csv("bom.movie\_gross.csv.gz") |
| **im.db.zip** | Zipped SQLite Database (IMDB data) | First unzip it → then use sqlite3.connect() |
| **rt.movie\_info.tsv.gz** | Compressed TSV (Rotten Tomatoes movie info) | Use pd.read\_csv("rt.movie\_info.tsv.gz", sep="\t") |
| **rt.reviews.tsv.gz** | Compressed TSV (Rotten Tomatoes reviews) | Use pd.read\_csv("rt.reviews.tsv.gz", sep="\t") |
| **tmdb.movies.csv.gz** | Compressed CSV (TheMovieDB data) | Use pd.read\_csv("tmdb.movies.csv.gz") |
| **tn.movie\_budgets.csv.gz** | Compressed CSV (The Numbers data) | Use pd.read\_csv("tn.movie\_budgets.csv.gz") |

### The numbers

https://chatgpt.com/c/68418e0a-bca8-8010-a481-10f6d0bdb037

**🔍 Step-by-Step Analysis Plan**

**✅ Step 1: Clean the Data**

You need to convert the dollar columns (production\_budget, domestic\_gross, worldwide\_gross) into numeric types for analysis:

# Remove $ and commas, convert to int

cols\_to\_clean = ['production\_budget', 'domestic\_gross', 'worldwide\_gross']

for col in cols\_to\_clean:

tn[col] = tn[col].replace('[\$,]', '', regex=True).astype(int)

# Convert release\_date to datetime

tn['release\_date'] = pd.to\_datetime(tn['release\_date'])

**✅ Step 2: Create Key Metrics**

Add a **profit** and **ROI (Return on Investment)** column:

tn['profit'] = tn['worldwide\_gross'] - tn['production\_budget']

tn['roi'] = tn['profit'] / tn['production\_budget']

**✅ Step 3: Analyze Profitability by Budget Range**

This helps address:

📌 "Determine optimal budget ranges"

import matplotlib.pyplot as plt

import seaborn as sns

# Create budget bins

bins = [0, 50e6, 100e6, 150e6, 200e6, 300e6, 500e6]

labels = ['<50M', '50-100M', '100-150M', '150-200M', '200-300M', '300M+']

tn['budget\_range'] = pd.cut(tn['production\_budget'], bins=bins, labels=labels)

# Average ROI by budget range

roi\_by\_budget = tn.groupby('budget\_range')['roi'].mean().sort\_values(ascending=False)

roi\_by\_budget.plot(kind='bar', title='Average ROI by Budget Range', ylabel='ROI')

plt.tight\_layout()

plt.show()

**✅ Step 4: Analyze Seasonality of Release**

This helps answer:

📌 "Strategize release timings"

# Extract release month

tn['release\_month'] = tn['release\_date'].dt.month\_name()

# Average worldwide gross by month

monthly\_avg = tn.groupby('release\_month')['worldwide\_gross'].mean().sort\_values(ascending=False)

monthly\_avg.plot(kind='bar', title='Avg Worldwide Gross by Release Month')

plt.ylabel('Average Worldwide Gross ($)')

plt.tight\_layout()

plt.show()

**✅ Step 5: Join with Genre and Rating Data (Optional but Valuable)**

If you have access to IMDB or TMDb data with genres and ratings:

* Merge with this tn dataset on movie name and release year.
* Then analyze profitability or ROI by **genre** and **rating**:

# Example

roi\_by\_genre = merged\_df.groupby('genre')['roi'].mean().sort\_values(ascending=False)

roi\_by\_genre.plot(kind='bar')

**✅ Step 6: Build Recommendation Summary**

Use your findings to make recommendations like:

* Best **budget ranges** for strong ROI
* Best **months** for release
* Most **profitable genres** (if genre data added)
* Average expected **ROI or profit** based on spending levels

**✨ Bonus Ideas**

* Identify **outliers**: Movies that lost a lot of money despite big budgets (bad investments).
* Use regression to **predict profitability** based on budget, release month, and genre.
* Use clustering to group movies with similar financial profiles.

Let me know if you'd like help merging the genre/rating data or running a specific analysis.