



Presentation By :Group 3

# *PREDICTING MOVIE PERFORMANCE USING IMDB & BOX OFFICE DATA*



DVF-PT06

# Project Overview

This project analyzes movie metadata, audience ratings, and box office performance to understand what factors contribute to a movie's financial and critical success.

Using data from the IMDb title.basics dataset, IMDb ratings, and box office gross revenue data, we explored relationships between genres, runtime, release year, and audience ratings with worldwide gross revenue.

We applied exploratory data analysis to identify key drivers of performance. The results show that genre, audience rating, and gross have measurable relationships with box office success.

We recommend that the business use these insights to guide content investment decisions, optimize release strategies, and prioritize projects with higher predicted return potential.

# Business Problem

The film industry faces high financial risk when deciding which movies to fund, distribute, or promote. Each project requires significant investment, yet only a small number achieve strong box office and audience success. The goal of this project is to identify movie characteristics that drive higher revenue and ratings and use data-driven insights to support better decision-making.

## Key Pain Points

- Uncertainty in predicting box office performance before release
- Difficulty comparing movie projects across genres and time periods
- Limited clarity on factors influencing audience reception and revenue

# Business Problem

## Data Analysis Focus

- Relationship between movie features (genre, runtime, release year) and audience ratings
- Link between IMDb ratings and box office revenue
- Identification of consistently high-performing genres
- Use of metadata and ratings to build a baseline revenue prediction model

## Business Value

These insights support:

- Smarter project selection and investment decisions
- More effective marketing budget allocation
- Improved revenue forecasting and financial planning

# Data Understanding

## Data sources

**IMDb Title Basics:** Movie metadata including title, release year, runtime, and genres

**IMDb Ratings Dataset:** Audience ratings and number of votes per movie

**Box Office Gross Dataset:** Domestic, foreign, and worldwide revenue figures

These datasets were merged and cleaned titles to support analysis of performance drivers.

## Data Preparations

- Filled in the missing gross revenue (Foreign) records
- Filled in the missing runtime minutes with median values
- Removed the outliers on runtime minutes (51,420 ) and the year (2115) using boxplots.

# Data Understanding

## Data sources

- What the Data Represent
- Each row represents a released movie
- The sample includes films with both audience ratings and box office data
- Variables capture movie characteristics, audience reception, and financial performance

### **Target Variable**

- Worldwide Gross Revenue
- Used as a proxy for financial success

# Data Understanding

## Data sources

### Key Variables & Properties

**Numerical:** Runtime (minutes), Release Year, Average Rating, Number of Votes, Worldwide Gross

**Categorical:** Genres and Title

Revenue is highly skewed, requiring transformation for analysis.

## Key Insights

- Higher IMDb ratings correlate with higher revenue
- Genre significantly influences box office performance
- Runtime shows a weak but noticeable effect

## Business Impact

- Supports movie project selection
- Improves marketing budget allocation
- Enables early-stage revenue forecasting

# Recommendations

- Focus on high-performing genres
- Use ratings as early performance signals
- Incorporate budget, cast popularity, and advanced models

Presentation By :Group 3

# THANK YOU!

**Group Members : Adelaide Achieng, Bramwel Mwangi, Denis Mwilaria, Patrick Kathurima**