# **🎬 IMDB Movie Data Analysis Project Report**

## 

## **📘 1. Introduction**

The purpose of this project was to perform an end-to-end analysis of a large movie dataset (IMDB movies) to uncover insights related to **genres, durations, languages, directors, budgets, and box office performance**.  
 The project involved comprehensive **data cleaning**, **exploratory data analysis (EDA)**, **correlation analysis**, and creation of a **Summary Dashboard** to visualize and interpret key insights.

## 

## **🧹 2. Data Cleaning and Preparation**

### **2.1 Initial Dataset Overview**

* **Initial Rows:** 5,000+
* **Initial Columns:** Included movie metadata such as Title, Genre, Duration, Language, Director, Budget, Gross, IMDB Rating, etc.
* **Issues Identified:**
  + Missing values in several numeric and categorical fields
  + Duplicate entries
  + Presence of outliers (unrealistic or extreme values in Budget & Gross)

### 

### 

### **2.2 Cleaning Operations Performed**

1. **Removed Duplicates:**
   * Identified and deleted duplicate movie entries using conditional formatting and unique filtering.
   * Ensured only unique movie titles remained.
2. **Handled Missing Values:**
   * For numeric columns (e.g., Budget, Gross, IMDB Rating), replaced missing values with median or appropriate estimates.
   * For categorical fields (e.g., Genre, Language), used mode or removed incomplete rows if minimal.
3. **Removed Outliers:**
   * Applied statistical methods (IQR/Z-score) to identify outliers in **Budget** and **Gross**.
   * Outlier rows were removed for clean statistical correlation.
4. **Result After Cleaning:**
   * **Final Row Count:** **4,936 rows**
   * **Dataset became consistent and analysis-ready.**

### 

### 

### **2.3 Data Organization**

After cleaning, several dedicated sheets were created for structured analysis:

| **Sheet Name** | **Purpose** |
| --- | --- |
| **Analysis** | **Initial cleaning overview (missing values, duplicates count)** |
| **Data Cleaning** | **Final cleaned dataset** |
| **Important Columns** | **Selected relevant fields for focused analysis** |
| **Important Columns Backup** | **Safety copy of important data** |
| **Genre Analysis** | **Genre-wise performance metrics** |
| **Duration Analysis** | **Distribution & correlation of movie duration** |
| **Language Analysis** | **Language-wise IMDB score patterns** |
| **Director Analysis** | **Performance comparison of top directors** |
| **Budget Analysis** | **Profitability, correlation between Budget & Gross** |
| **Cleaned for Scatter** | **Dataset used to create the Budget vs Gross scatter plot** |
| **Summary Dashboard** | **Consolidated summary and interpretation of results** |

### **3.1 Genre Analysis**

* **Total genres analyzed: 20**
* **Most common genre: Drama (2,594 movies)**
* **Highest average IMDB rating: War (Avg 7.36)**
* **Insights: Drama dominates production count, while War and Biography genres have the highest critical acclaim.**

### **3.2 Duration Analysis**

* **Average Duration: 108.15 minutes**
* **Median Duration: 104 minutes**
* **Range: 7–330 minutes**
* **Correlation (Duration vs IMDB Rating): Weak (between -0.3 and 0.3)**
* **Insights: Most movies fall between 90–120 minutes, the ideal duration for audience engagement.**

| **Language** | **Mean IMDB** | **Count** |
| --- | --- | --- |
| **English** | **6.37** | **4607** |
| **Japanese** | **7.42** | **17** |
| **German** | **7.34** | **19** |
| **French** | **7.02** | **72** |

* **Most Common Language: English**
* **Highest Average IMDB Rating: Japanese**
* **Insight: While English dominates in quantity, Japanese and European languages often deliver higher-quality ratings.**

### **3.4 Director Analysis**

* **Top Directors by Movie Count: Steven Spielberg (26), Woody Allen (22), Martin Scorsese (20)**
* **Top Director by Average IMDB Rating: Martin Scorsese (7.66)**
* **Most Consistent: Woody Allen (lowest standard deviation = 0.53)**
* **Insight: Spielberg leads in volume, Scorsese in quality, and Allen in consistency.**

### **3.5 Budget Analysis**

* **Top 3 Most Profitable Movies:**
  1. ***Avatar* – Profit Margin: $523,505,847 – Ratio: 3.21**
  2. ***Jurassic World* – Profit Margin: $502,177,271 – Ratio: 4.35**
  3. ***Titanic* – Profit Margin: $458,672,302 – Ratio: 3.29**
* **Budget vs Gross Correlation: 0.4519 → Moderate Positive**
* **R² Value: ≈ 0.20**
* **Insight: Higher budgets often yield better gross, but the relationship isn’t strong — creativity and audience appeal matter more.**

### 

### 

### **3.6 Cleaned for Scatter**

* **Created a cleaned numeric subset (“Budget\_Filtered”) for regression visualization.**
* **Scatter plot shows a moderately increasing trend between Budget and Gross.**

**📈 4. Summary Dashboard**

| **Metric** | **Value** |
| --- | --- |
| **Most Common Genre** | **Drama** |
| **Highest Rated Genre** | **War** |
| **Average Duration** | **108.15 mins** |
| **Most Common Language** | **English** |
| **Top Director (Avg IMDB)** | **Mitchell Altieri — 8.70** |
| **Budget ↔ Gross Corr.** | **0.1119 (Weak Positive)** |

**Correlation Interpretation: Weak positive relationship — higher budget *slightly* improves box office, but not always.**

### **4.2 Insights**

* **🎭 Drama dominates the dataset, confirming it as a universally preferred storytelling genre.**
* **🎬 War, Biography, and Documentary genres achieve the best IMDB ratings.**
* **🌍 English movies lead in volume, but Japanese and German movies show higher viewer satisfaction.**
* **🕒 Movies between 90–120 minutes tend to perform better.**
* **💰 Spending more doesn’t guarantee success — story quality and audience connection remain key drivers.**

### 

### **4.3 Recommendations**

* **Producers: Focus on strong scripts and balanced budgets.**
* **Studios: Prioritize mid-budget films with proven genre-audience alignment.**
* **Streaming Platforms: Feature non-English high-rated titles for international diversity.**
* **Filmmakers: Blend creativity and pacing within ~2-hour duration range.**
* **Analysts: Explore future studies connecting actor popularity, release year, and audience reviews.**

## **🧾 5. Conclusion**

**This project successfully cleaned, structured, and analyzed a movie dataset to derive actionable insights.  
 From data wrangling to dashboard visualization, each step enhanced the dataset’s integrity and analytical clarity.  
 The final Summary Dashboard provides a one-glance understanding of the dataset’s key trends in genre, language, duration, direction, and profitability.**

## **🧩 6. Tools & Techniques Used**

* **Platform: Google Sheets**
* **Techniques: Data Cleaning, Filtering, IQR Outlier Removal, Correlation Analysis, Dashboard Creation**
* **Visuals: Scatter Plot, Summary Table, Correlation Charts**
* **Functions Used:**
  + **COUNTIF, UNIQUE, FILTER, CORREL, INDEX-MATCH, IF, ROUND, AVERAGE, STDEV, SORT, MAX, MIN**

| **Metric** | **Value** |
| --- | --- |
| **Total Rows (after cleaning)** | **4,936** |
| **Duplicates Removed** | **64** |
| **Missing Values Treated** | **Several (varied by column)** |
| **Outliers Removed** | **Extreme budget/gross values** |
| **Final Sheets Created** | **9** |