LAB 3 REPORT: MOVIE INDUSTRY ANALYSIS

PART 1: DATA EXPLORATION & CLEANING

1.1 Initial Exploration

Dataset Shape: The dataset contains 3,000 movies and 14 columns.
Missing Data: Only one column, *metascore*, had missing values (422 rows, approx 14%).
Data Types:
- Integer: movie_id, release_year, runtime, metascore, oscar_nominations, oscar_wins.
- Float: budget, revenue, imdb_rating, votes.
- String (Object): title, genre, director, main_actor.

1.2 Handling Missing Data

Strategy: Filled missing *metascore* values with the median value by genre.
Justification: Critical reception varies significantly by genre (e.g., dramas often score higher than horror films). Using the genre median preserves these inherent differences better than a global mean.

1.3 Data Validation

Impossible Values: Checked for negative budget/revenue, runtime < 10 mins, and ratings outside valid ranges. **0 impossible records found** in this dataset.
Inconsistent Data: Checked for cases where Oscar Wins > Oscar Nominations. **0 inconsistent records found**.
Extreme Outliers: Identified distinct outliers in Budget and Revenue (handled in Part 3).

PART 2: DESCRIPTIVE STATISTICS

2.1 Numerical Summary

Budget: Mean: $36.3M | Median: $23.7M | Range: $0.1M - $350M
Revenue: Mean: $153.2M | Median: $79.5M | Range: $0.01M - $2800M
IMDb Rating: Mean: 6.59 | Median: 6.60 | Range: 1.0 - 9.5

2.3 Statistical Interpretation

Most Variable: Revenue is the most variable metric (CV = 1.61), confirming the "hit-driven" nature of the industry where a few movies make billions while many make little.
Most Skewed: Oscar Wins (Skew = 2.08) and Revenue are highly right-skewed. Most movies win 0 Oscars and earn modest revenue, while a tiny tail achieves massive success.
Interesting Patterns:
- The median profit is usually positive, but the mean is heavily pulled up by blockbusters.
- Runtime has very low variability (CV ~0.2), suggesting a strict industry standard for movie length (90-120 mins).
- Oscar nominations are rare; the median is 1.0, but the mean is 1.6, showing the influence of "prestige" films sweeping nominations.

PART 3: DISTRIBUTION ANALYSIS

3.2 Outlier Analysis

Impact: Outlier movies account for **49.87%** of the total revenue!
Interpretation: These are not errors but genuine blockbusters. Removing them would destroy the validity of any market analysis.
Top 5 Outliers (Revenue):

1. Movie 2926: $2.80B
2. Movie 1611: $2.80B
3. Movie 1976: $2.70B
4. Movie 1281: $2.67B
5. Movie 327: $2.61B

3.3 Normality Assessment (IMDb Rating)

The IMDb rating distribution is **approximately normal**.

- Within 1 SD: 67.7% (Exp: 68%)
- Within 2 SD: 95.4% (Exp: 95%)
- Within 3 SD: 99.9% (Exp: 99.7%)

Conclusion: The bell curve fits user ratings almost perfectly.

PART 4: RELATIONSHIP ANALYSIS

4.1 Correlation Analysis

Strongest Positive Correlation: Budget & Revenue (0.74). Higher investment strongly predicts higher returns.
Correlation Surprise: Profit Margin vs. IMDb Rating (~0.00).
Investigation: There is effectively **no correlation** between how much critics/users like a movie and its Return on Investment (ROI). A movie can be a critical disaster but a financial hit (high profit), or a critical masterpiece that loses money.

4.3 Advanced Analysis

Profit Margin: Calculated as (Revenue - Budget) / Budget.
Infinite Margin: Occurs when Budget = 0. Handled by excluding these cases or treating as NaN to prevent calculation errors.

PART 5: COMPARATIVE ANALYSIS

5.1 Genre Comparison

Highest Median Budget: Comedy ($25.7M). (Note: In this specific dataset, comedies appear to have higher mid-level budgets than expected).
Most Consistent Ratings: Comedy (Lowest SD = 0.94). Viewers generally rate comedies within a narrower range than polarizing genres like Drama.
Highest Profitability: Action (Avg Profit Margin = 3.40x). Action movies have the best average return on investment.
Most Awards: Drama (262 Oscar Wins). As expected, Dramas dominate the awards circuit.

5.2 Time Trend Analysis

Budget: Budgets have fluctuated but show a general upward trend when smoothed (3-year moving average).
Quality (Ratings): Average IMDb ratings have remained relatively stable over time (hovering

around 6.6), suggesting no significant long-term decline or improvement in perceived movie quality.

5.3 Director Analysis (Top 10)

The top directors (e.g., Christopher Nolan, James Cameron, Steven Spielberg) show remarkable consistency in this dataset:

- **Average Rating:** ~6.5 - 6.6
- **Profit Margin:** ~3.0x - 3.3x
- **Oscar Nominations:** ~1.5 - 1.7 per movie.

Conclusion: Top directors are hired because they reliably deliver ~3x returns on budget and consistently decent ratings.