Data Analyst

Project 1

Prepared by:

Bisan Abdelrahim

- **Full analysis**

1. Dataset Specification
Dataset Analyzed: TMDb 5000 Movies Dataset
- Source: The Movie Database (TMDb)
- Original size: 4,803 movies with 20 variables
- Cleaned dataset: 3,229 movies with 31 variables
- Time period: 1916-2016 (100 years of cinema)
- Key variables: Budget, Revenue, Genres, Ratings, Runtime, Release Dates

2. Research Questions Posed
Primary Questions:
1. What factors influence movie profitability?
   - How do budget, genre, and runtime affect profit?
   - Is there a sweet spot for budget that maximizes ROI?

2. How have movie trends changed over time?
   - Are budgets and revenues increasing over the years?
   - Have certain genres become more or less popular?

3. What makes a movie successful?
   - What's the relationship between ratings and financial success?
   - Do longer movies make more money?

Secondary Questions:
4. Which genres are most profitable and popular?
5. How do different budget levels perform?
6. Do higher-rated movies always make more money?

3. Investigation Methodology
Analysis Techniques:
- Descriptive statistics for financial and rating variables
- Data visualization (15+ plot types: histograms, scatter plots, heatmaps, bar charts)
- Correlation analysis between key variables
- Categorical analysis by genres, budget ranges, rating levels
- Time series analysis of decade trends
- Statistical grouping and comparative analysis

Tools Used:
- Python with Pandas for data manipulation
- NumPy for numerical computations
- Matplotlib & Seaborn for visualizations
- Custom helper functions for analysis and formatting

4. Data Wrangling Documentation

Data Cleaning Steps:

1. Removed invalid records:
   - Filtered out 1,574 movies with budget = $0
   - Removed movies with revenue = $0
   - Final dataset: 3,229 movies (67% of original)

2. Missing value handling:
   - Runtime: Filled with median (110 minutes)
   - Converted release_date to datetime format

3. Feature Engineering:
   - Created derived variables: Profit, ROI, Revenue per minute
   - Added categorical variables: Budget/Revenue/Rating/Runtime categories
   - Extracted release year and decade
   - Parsed JSON genre data into usable format

5. Summary Statistics & Key Results

Financial overview:

- Average Budget: $40.7 million
- Average Revenue: $121.2 million
- Average Profit: $80.6 million
- Average ROI: 295,382% (median: 130%)
- Average Rating: 6.3/10
- Average Runtime: 110.7 minutes

Key Findings:

Budget & Profitability:

- Strong budget-revenue correlation (0.705)
- ROI sweet spot: Low-budget movies (<$5M) achieve 1,917,837% average ROI
- Most profitable: Avatar ($2.55B profit), Gone with the Wind (9,904% ROI)

Genre Performance:

- Top revenue genre: Animation ($279M average)
- Best overall: Animation, Adventure, Fantasy, Family
- Most prolific: Comedy (1,110 movies)

Time Trends:

- Budget inflation: $6M (1960s) → $51M (2010s)
- Production peaked in 2000s-2010s
- Runtimes gradually increasing

Quality vs Success:

- Excellent movies (8+ rating): $232M average revenue
- Poor movies (<5 rating): $44M average revenue
- Rating-revenue correlation: 0.188 (moderate)
- Longer movies perform better financially

Strategic Insights:

- Portfolio approach: Mix high-budget blockbusters + low-budget high-ROI films
- Animation and Adventure genres consistently successful
- Quality matters but marketing/popularity crucial (vote count correlation: 0.756)

6. Include These Visualizations

From your notebook, include these key plots:

1. Distribution plots (budget, revenue, profit, ratings)
2. Budget vs Revenue scatter plot
3. Genre performance bar charts
4. Time trends line charts
5. Correlation heatmap
6. Any other compelling visualizations from your analysis