TMDB Movie Dataset Analysis

Midterm Project Presentation

By – Hinal Pujara and Sravani Yedoti

1. Introductio n

Analysis of The Movie Database (TMDB) dataset

Comprehensive examination of movie industry trends and patterns

Focus on financial performance, audience reception, and production characteristics

2. Problem Definition

Research Questions

What factors contribute to a movie's financial success?

How do production budgets influence revenue and ROI?

What relationships exist between audience ratings and financial performance?

How have movie characteristics evolved over time?

Business Value

Guide investment decisions in film production

Optimize budget allocation

Understand audience preferences

Identify profitable market segments

3. Dataset Overview

Source

TMDB Movie Dataset

- 342,267 unique movies after cleaning
- 20 features including financial, temporal, and categorical data
- Time span: Multiple decades of film data

Key Features

- Financial: Budget, Revenue, ROI
- Audience Reception: Vote Average, Vote Count
- Production: Runtime, Genres, Production Companies
- Temporal: Release Dates
- Descriptive: Title, Overview, Keywords

4. Exploratory Data Analysis

Removed unnecessary columns

Handled missing values

Standardized dates

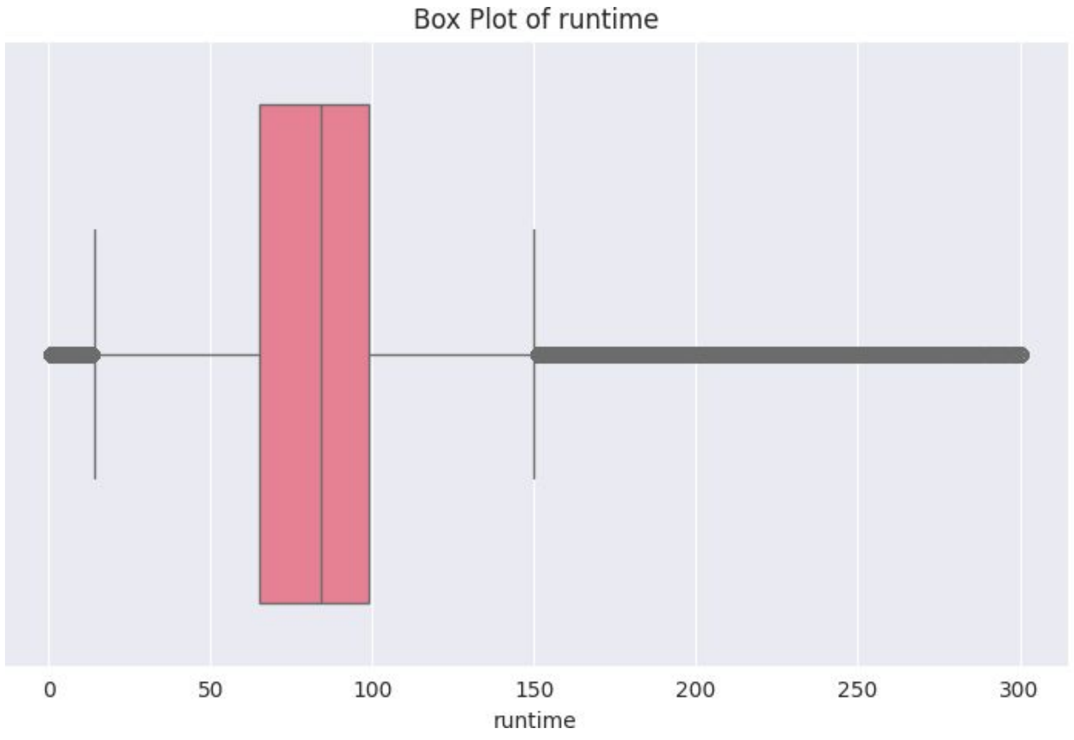
Created derived features (ROI, release year/month)

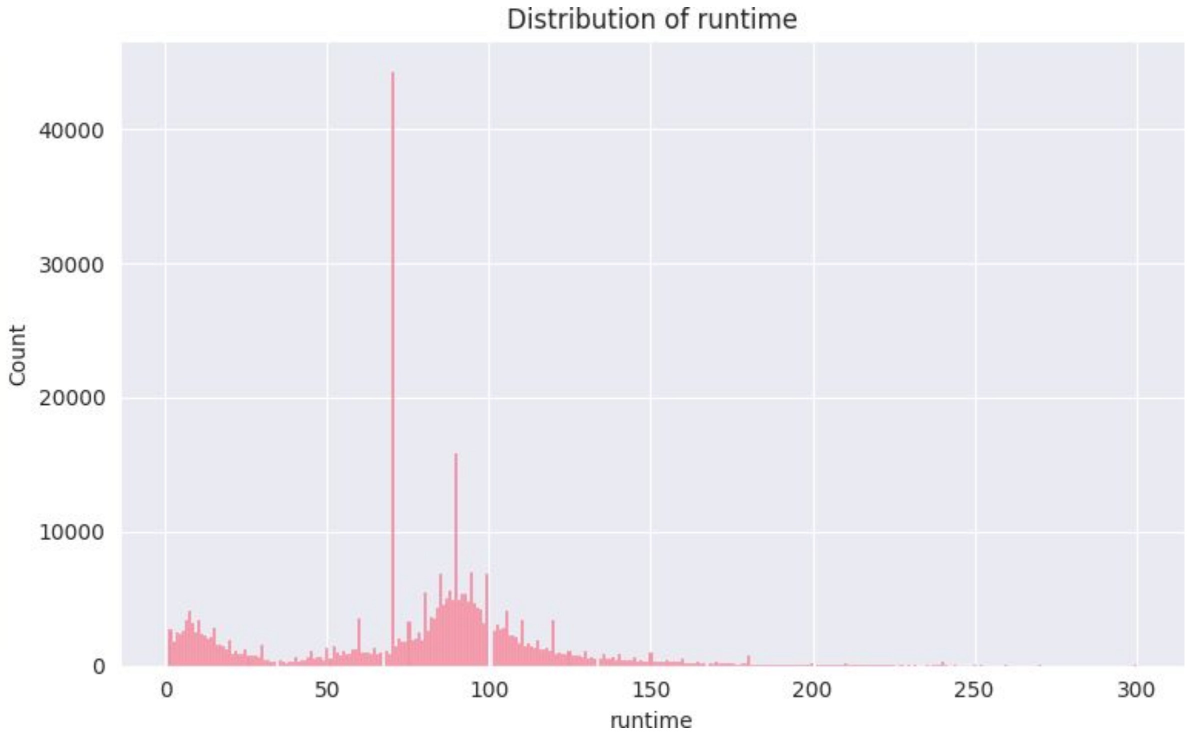
Removed duplicates and outliers

4.1 Data Cleaning Process

4.2 Key Visualizations

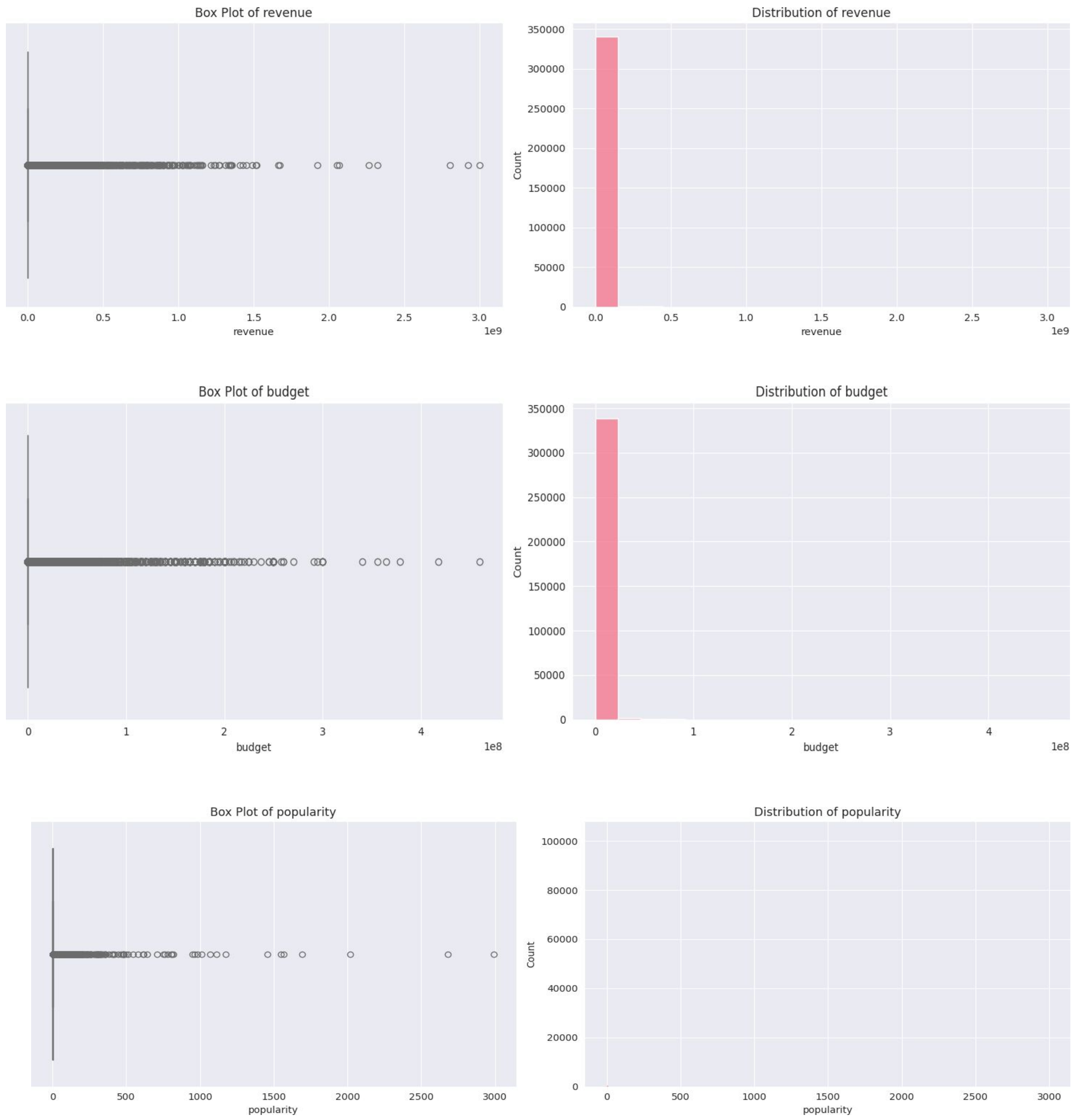
1. Distribution Analysis
    1. Runtime: Most movies between 64-100minutes
    2. Vote averages: Normal distribution
    3. Revenue/Budget: Right-skewed distribution

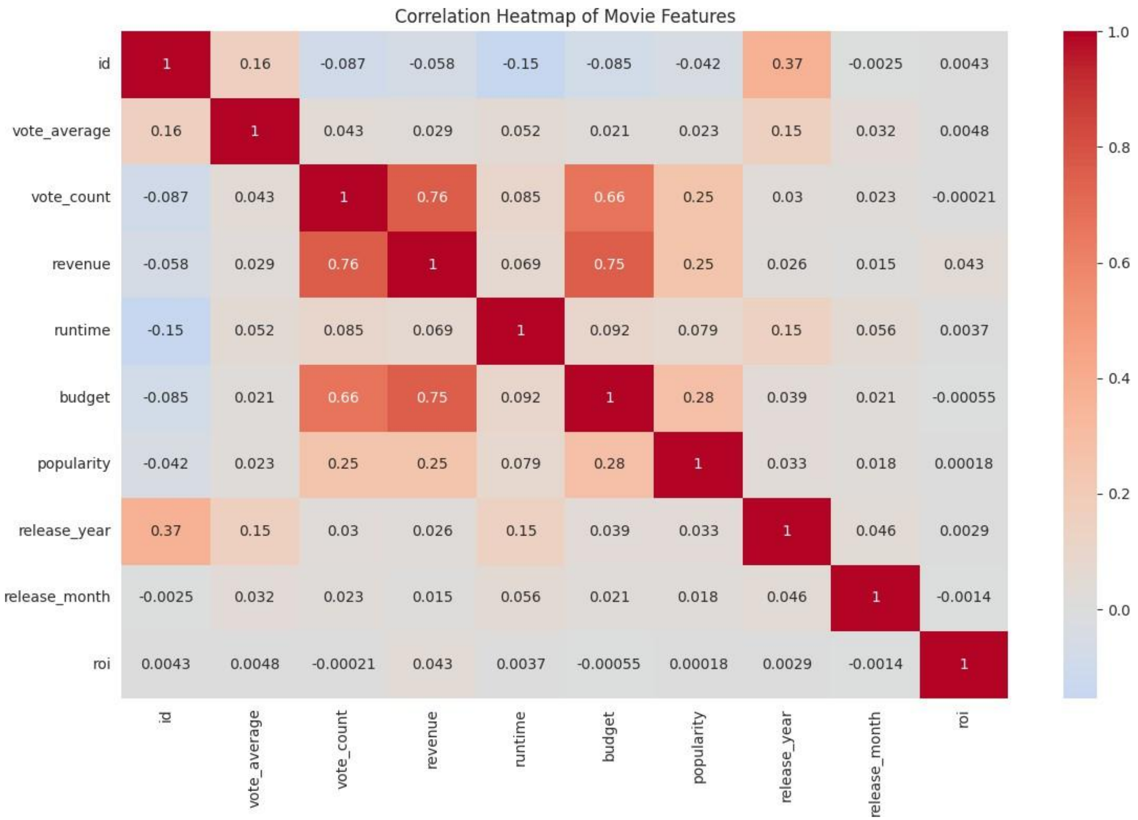




Outlier Analysis

- Box plots reveal extreme values in:
  - Revenue distribution
  - Budget allocation
  - Popularity metrics





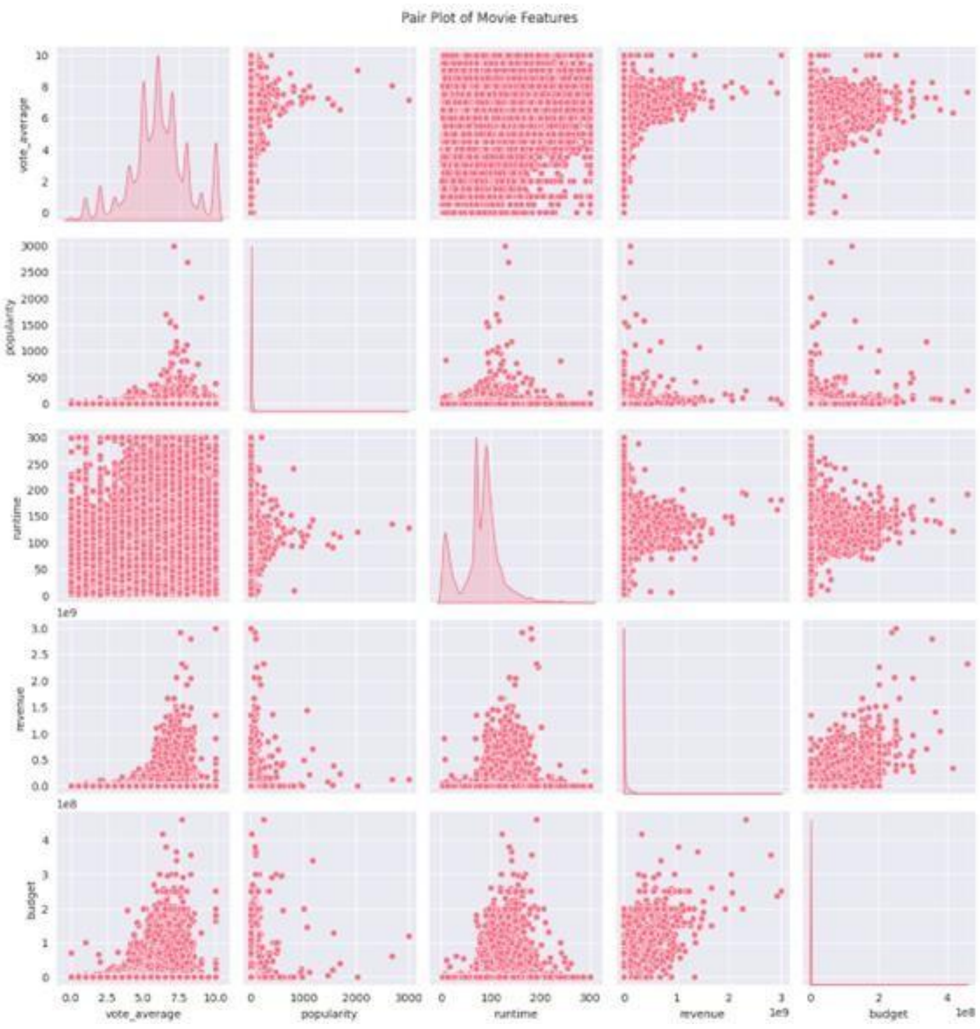
Correlation Analysis

- Strong positive correlation (0.75) between budget and revenue
- Moderate correlation between vote

count and financial metrics

- Weak correlation between runtime and other metrics

- Strong budget-revenue correlation suggests higher investments often lead to higher returns
- However, outliers indicate some low-budget films can achieve significant success
- Movie ratings follow a normal distribution
- High ratings don't guarantee high popularity or revenue
- Suggests quality and commercial success aren't always aligned
- Runtime shows little correlation with success metrics
- Suggests focusing on content quality over length
- Most successful films fall within standard runtime ranges
- Industry shows a "blockbuster" pattern where few movies capture most revenue
- High budget films tend to generate higher revenue but with varying ROI



5. Results and Insights

1. **Financial Patterns**

- Mean revenue: $2.11M
- High variance in ROI
- Budget strongly predicts revenue

5.2 Production Characteristics

- Optimal runtime range identified
- Genre influence on success
- Studio performance variations

- **5.3 Audience Reception**
- Vote patterns
- Popularity metrics
- Rating correlations
- **6. Technical Implementation**
- Python-based analysis
- Libraries: Pandas, NumPy, Seaborn, Matplotlib
- Custom functions and classes
- Exception handling
- Data transformation techniques

7. Conclusions

- **Key Findings:**

1. Strong budget-revenue relationship
2. Audience ratings impact financial success
3. Genre diversity affects market performance
4. Production quality indicators

- **Recommendations:**

5. Budget optimization strategies
6. Genre selection guidance
7. Production planning insights
8. Marketing focus areas

8. Future Work

- Deep learning predictions
- Sentiment analysis of reviews
- Real-time data integration
- Market segment analysis

10. References

- https://www.kaggle.com/datasets/juzershakir/tmdb-movies-dataset

Thank you for your attention!