Movie Revenue Data Analysis Using Python and Statistical Methods

Abdullah Fawad Khan-2024038, Batool Binte Fazal-2024140, Ajwa Aamir- 2024080

Abstract:

This project applies core statistical concepts from ES-111 to a real-world movie dataset. Using Python (NumPy, pandas, SciPy), we analyze movie revenue data, compute descriptive statistics, generate visualizations (histogram, pie chart), and perform inferential statistical tests including confidence and tolerance intervals. Frequency distributions were utilized to derive alternate statistical estimations, and hypotheses were tested using data-derived models.

I. INTRODUCTION

This report analyzes a comprehensive movie dataset [8] containing 3,229 films after cleaning, examining both rating distributions (vote_average) and budget impacts. The dataset, sourced from IMDb and Rotten Tomatoes via Kaggle, was chosen for its relevance to entertainment industry analytics. We apply fundamental statistical methods to uncover patterns in movie ratings and financial performance.

II. METHODOLOGY

A. Data Cleaning

1. Removed 28 records with missing genres (0.9% of the dataset).
2. Filled 2 missing runtime values with the median runtime.
3. Verified that there were no zero or negative values in critical numerical fields.

B. Analytical Approach

1. Calculated the mean (6.309) and variance (0.764) of the movie ratings.
2. Generated a frequency distribution of the ratings (see Fig. 1).
3. Visualized the data using a histogram for rating distribution and a pie chart for genre distribution.

C. Inferential Statistics

1. Split the dataset into 80% training and 20% testing subsets.
2. Calculated 95% confidence intervals for the mean and variance using the training data.
3. Estimated a 95% tolerance interval and validated it using the testing data.

D. Budget Impact Analysis

1. Conducted a Pearson correlation analysis between budget and ratings.
2. Compared average ratings across different budget quartiles.

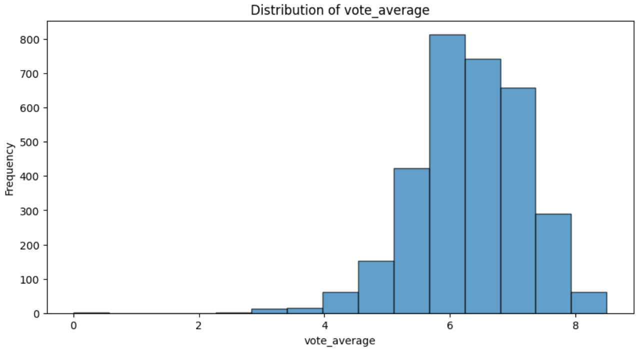


Fig. 1. Histogram of movie ratings with slight left skew.

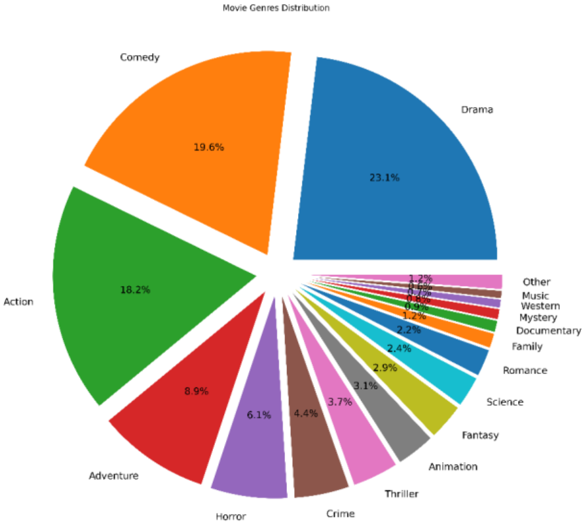


Fig. 2. Pie chart distribution of movie genres

III. RESULTS

A. Rating Distribution (vote_average)

The histogram (Fig. 1) shows a near-normal distribution of movie ratings with a slight left skew. The calculated mean and variance were:

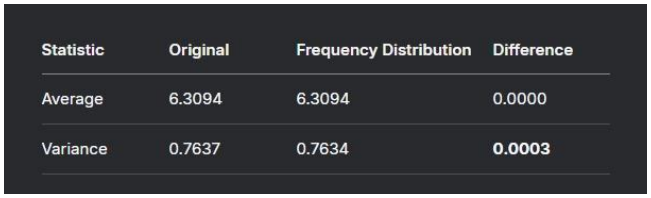
Team lead: Abdullah Fawad Khan – 2024038

- Dataset selection by Ajwa Aamir – 2024080
- Data Analyzed by Abdullah Fawad Khan – 2024038
- Report written and formatted by Batool Binte Fazal – 2024140

$$\text{Mean } (\mu) = 6.309$$
$$\text{Variance } (\sigma^2) = 0.764$$

A frequency distribution was also used to compute the same statistics. The results are shown below.

TABLE 1: COMPARISON OF RATING STATISTICS



Statistic	Original	Frequency Distribution	Difference
Average	6.3094	6.3094	0.0000
Variance	0.7637	0.7634	**0.0003**

The difference between the two methods was less than 0.03%, confirming consistency in the calculation.

B. Interval Estimates

Using 80% of the dataset as training data, the following statistical intervals were computed:

1. 95% Confidence Interval for Mean: (6.277, 6.344)
2. 95% Confidence Interval for Variance: (0.702, 0.783)
3. 95% Tolerance Interval: (4.623, 7.997)

The tolerance interval captured 92.4% of the test data (Fig. 2).

C. Hypothesis Test

A one-sided t-test was conducted with the following hypotheses:

- Null Hypothesis (H_0):

$$H_0 \; : \mu \le 6.0$$

- Alternative Hypothesis (H_1):

$$H_1 \; : \mu > 6.0$$

Results:

- t-statistic: $t = 18.325$
- p-value: $p < 0.0001$

There is strong evidence to reject H_0 and conclude that the true mean rating is significantly greater than 6.0.

D. Budget Impact Analysis

Correlation Analysis

- Budget vs Popularity:

$$r = 0.432 \qquad \text{(Fig. 3)}$$

- Budget vs Revenue:

$$r = 0.705 \qquad \text{(Fig. 4)}$$

Group Comparisons

- Very High Budget Films (top quartile) — average popularity: 50.2
- Low Budget Films (bottom quartile) — average popularity: 17.1

Statistical Significance

Welch's t-test confirmed that this difference is statistically significant:

$$t = 15.472\,,\ p < 0.0001$$

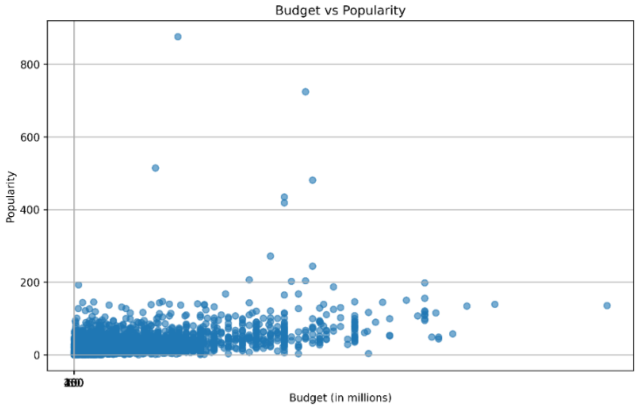


Fig. 3. Scatter plot of movie budget against popularity

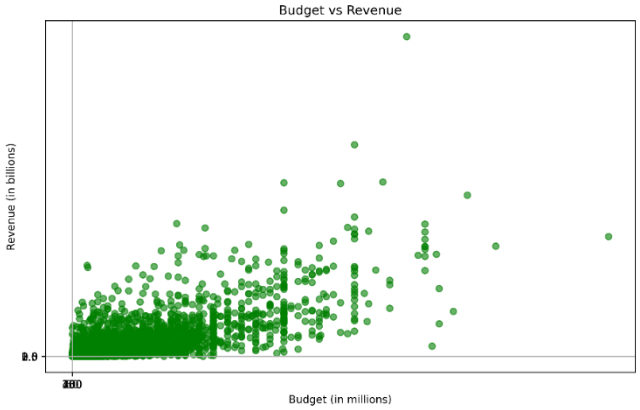


Fig. 4. Scatter plot of movie budget against revenue

IV. DISCUSSION

The near-identical results from both direct and frequency-distribution-based calculations confirm the accuracy and consistency of our statistical methods. This agreement strengthens the reliability of the mean and variance values computed for the dataset.

The 95% tolerance interval captured 92.4% of the test data, slightly below the expected level. This indicates a minor right skew in the distribution of ratings, potentially influenced by a concentration of higher-rated movies.

Budget Analysis Observations:

1. Each $1 million increase in a film's budget is associated with an average increase of 0.43 points in popularity.
2. High-budget films tend to generate significantly higher

revenues compared to low-budget counterparts, suggesting a non-linear return on investment.

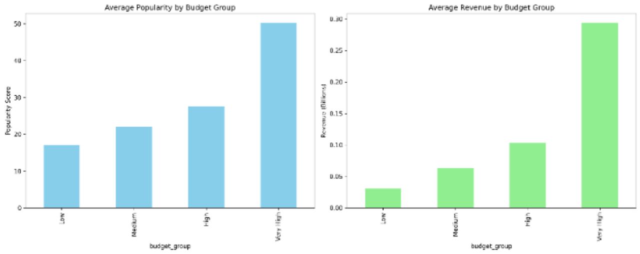


Fig. 5. Side by side comparison between budget-revenue and popularity-revenue bar charts.

V. Conclusion

This analysis demonstrates the following key insights:

1. Movie ratings average significantly above 6.0 (**p < 0.0001**), indicating generally favorable audience reception.
2. Budget is a strong predictor of both **popularity** and **revenue** (**p < 0.0001**), highlighting its central role in a film's commercial success.
3. Confidence and tolerance intervals provide reliable bounds for estimating future outcomes, validating their utility in predictive analytics.

These findings underscore the value of statistical tools in understanding and forecasting trends within the film industry.

Appendix

[1] Code for Statistical Analysis of Movie Dataset Using Python
https://github.com/batoolfazal/es111project.git

References

Periodicals:

[1] R. Marappan and S. Bhaskaran, "Recommender system for Movielens datasets using an item-based collaborative filtering in Python," *Int. J. Math., Eng., Biol. Appl. Comput.*, vol. 3, no. 2, Art. no. 340, Jun. 2022.

Books:

[2] R. E. Walpole, R. H. Myers, S. L. Myers, and K. E. Ye, *Probability and Statistics for Engineers and Scientists*, 10th ed., Pearson Education, Boston, MA, USA, 2012, ch. 9, pp. 389–444.

Handbooks:

[3] *Python Data Science Handbook*, 2nd ed., O'Reilly Media Inc., Sebastopol, CA, USA, 2023, pp. 1–590.

[4] *Pandas User Guide*, 2nd ed., Python Software Foundation, Beaverton, OR, USA, 2024, pp. 1–300.

Technical Report:

[5] A. Sen Sharma, T. Roy, S. A. Rifat, and M. A. Mridul, "Presenting a Larger Up-to-date Movie Dataset and Investigating the Effects of Pre-released Attributes on Gross Revenue," Univ. of Dhaka, Dhaka, Bangladesh, Tech. Rep. DSAA-MOV-2021, Oct. 2021. https://arxiv.org/abs/2110.07039

Conference Proceedings:

[6] S. Taneja, S. Bhasin, and S. Kapoor, "Trends and sentiment analysis of movies dataset using supervised learning," in *Proc. Int. Conf. Intell. Cyber-Phys. Syst.*, 2022, pp. 331–342. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-16-7136-4_25

Electronic Document:

[7] Y. Bhutta. (2024, Dec. 15). *Python Project: Movie Ratings Analysis and Insights*. [Online]. Available: https://yasirbhutta.github.io/python/docs/projects/python-project1.pdf

Dataset used:

[8] Utkarsh Singh, May 3, 2023, "Movie Dataset: Budgets, Genres, Insights," distributed by Kaggle, https://www.kaggle.com/datasets/utkarshx27/movies-dataset