Movie Success Trend analysis

Aminata Bangura
Ehren Dietrick
Huong Giang Ho

Introduction

This project aim to use data mining to predict movie success and inform entertainment industry strategy:

- The global box office is a multi-billion dollar market where a majority of films fail to recoup their investments
- Conventional multi-factor predictive models are complex, relying on variables like budget, marketing spend, and genre, which can be difficult to quantify or obtain early in a film's lifecycle

Related work

Studies have shown that data mining methods can reveal patterns of movie success factors. For example, budget can predict revenue, while genres can significantly impact performance. The methodologies used in these studies include the use of regression analysis and machine learning, providing a framework for our analysis [4], [5], [6].

[4]
J. Lee, “Exploratory Data Analysis With Movies,” *Medium*, Sep. 06, 2020. https://medium.com/data-science/exploratory-data-analysis-with-movies-3f32a4c3f2f3 (accessed Nov. 24, 2025).

[5]
H. Yang, Y. Pei, and Z. Wang, “Movie Data Analysis and a Recommendation Model,” *Proceedings of the 2022 International Conference on Bigdata Blockchain and Economy Management (ICBBEM 2022)*, pp. 739–751, Dec. 2022, doi: https://doi.org/10.2991/978-94-6463-030-5_74.

[6]
L. Kharb, D. Chahal, and Vagisha, “Forecasting Movie Rating Through Data Analytics,” *Data Science and Analytics*, pp. 249–257, 2020, doi: https://doi.org/10.1007/978-981-15-5830-6_21.

Objective

01	02	03
Investigate if a single feature from each dataset has significant standalone predictive power for success	Evaluate the consistency of this rating-performance correlation across multiple, distinct movie datasets.	Assess the viability of minimal-feature predictive systems to offer early, accessible insights for filmmakers, studios, and analysts.

Data overview

Three datasets were selected from Kaggle for diversity in scope and movie type:

- Movies Box Office Dataset (2000–2024): Contains revenue, genre, rating, and other metadata.
- IMDB Top 250 Movies Dataset: Includes only highly rated films curated by audience scores.
- Popular Movies Dataset: Contains trending films with mixed commercial and critical performance.

Each dataset varies in rating distribution, genre representation, and movie success metrics

=> we conclude that it will be the most appropriate to pick different variables to define success from each dataset

STEP 1

Handling missing rating and revenue values by either imputing or removing incomplete rows

STEP 2

Converting categorical variables to numeric representation where required for correlation checks

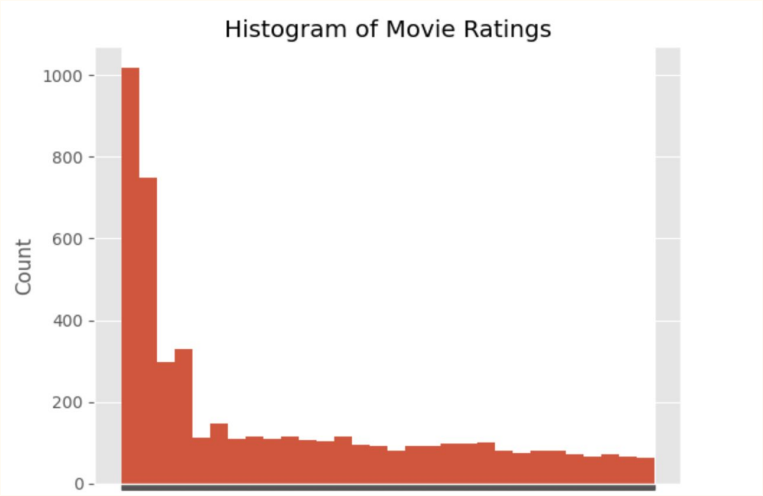
STEP 3

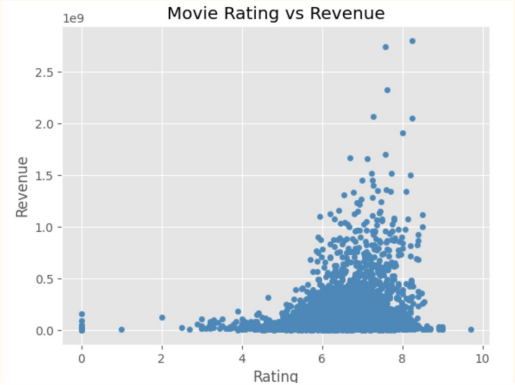
Make sure rating is normalized when comparing across time to reduce inflation bias

STEP 4

Defining a "success" value for regression evaluation

Data pre-processing

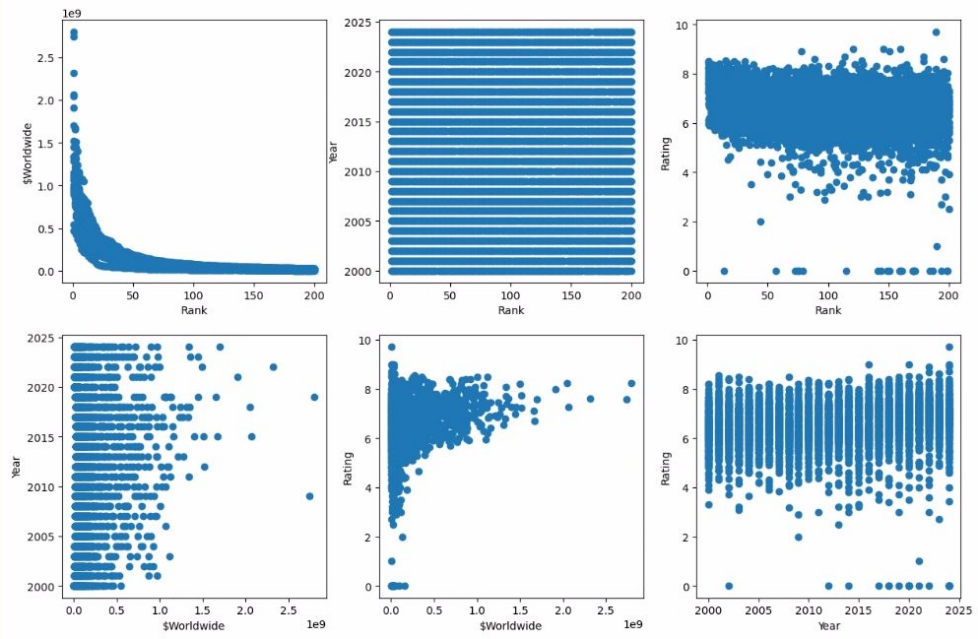




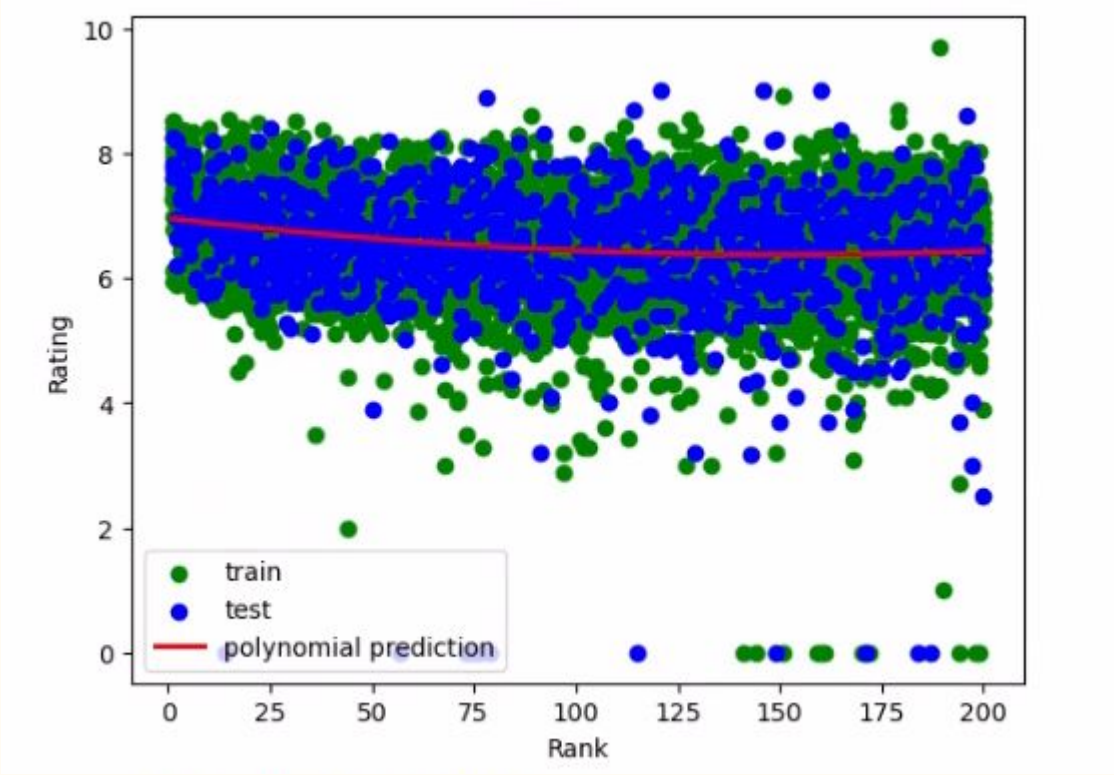
Dataset 1

(Movies Box Office (2000-2024))

EDA



Results


10
8
6
4
2
0
Rating
0
25
50
75
100
125
150
175
200
Rank
train
test
polynomial prediction

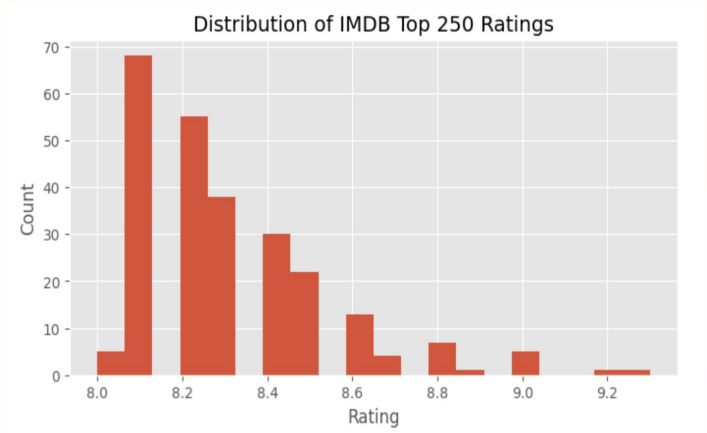
Evaluation

```
MSE: 1.1629072316342282
RMSE: 1.0783817652548786
R^2: 0.01891969763708068
Rating = -0.007719590143510946*Rank + 2.5449826943293823e-05*Rank^2 + 6.957751165526493
```

The polynomial regression performed on Rank → Rating provides further confirmation of weak correlation.
The fitted function is:
Rating = -0.0077(Rank) + 2.5449e-5(Rank²) + 6.9577
The metrics demonstrate that Rank is a very poor predictor of Rating:

- NSE: 1.1629 (values > 1 indicate the model performs worse than the mean)
- RMSE: 1.878 (large error relative to rating scale)
- R²: 0.018 (only 1.8% of variance in rating is explained)

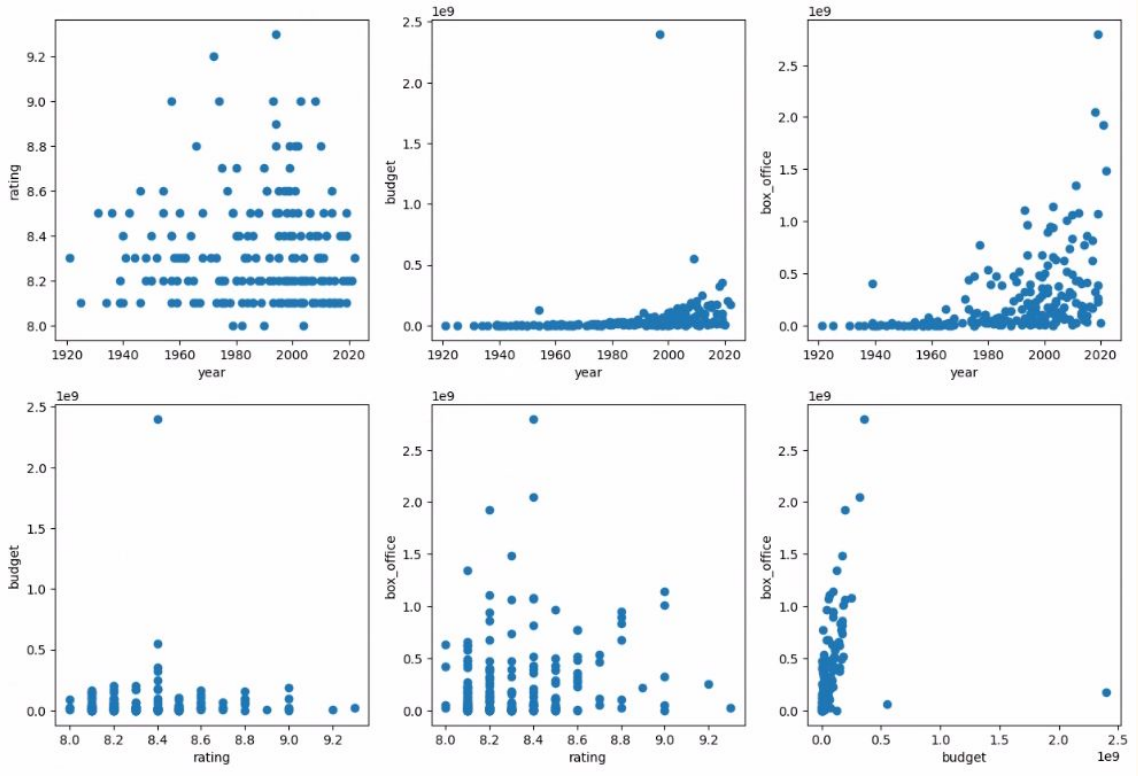
The regression line in the plot appears nearly flat with a barely noticeable downward bend, showing that increases in rank offer almost no predictive power for rating.



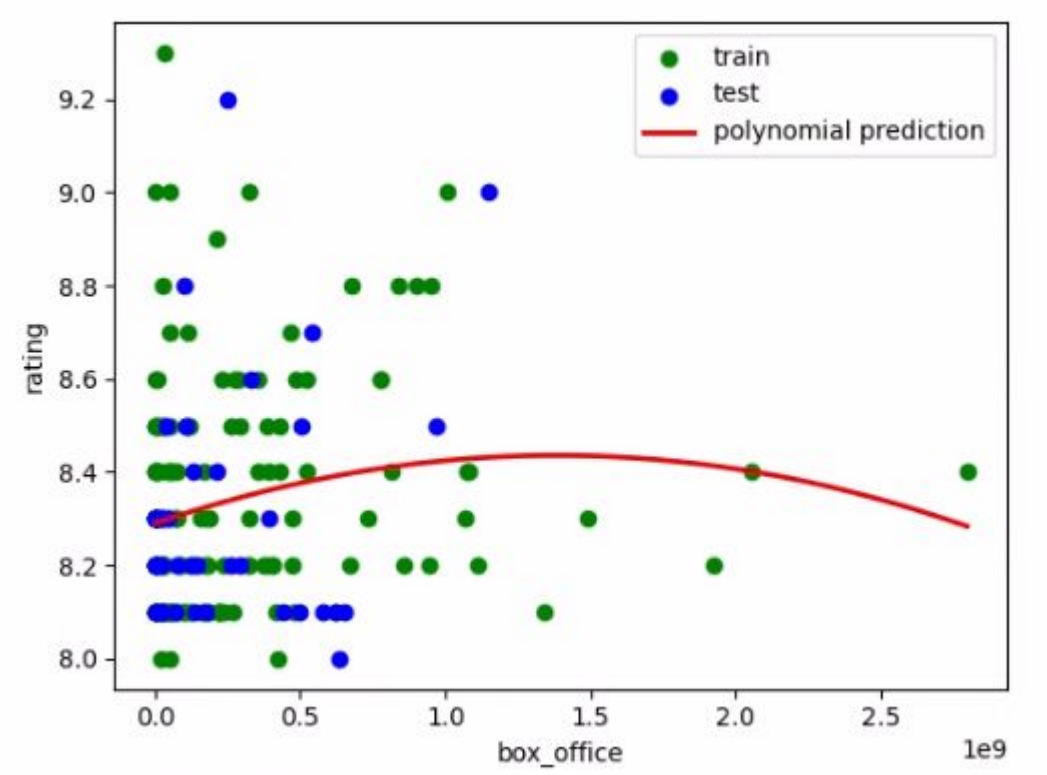
Dataset 2

(IMDB top movies)

EDA



Results


train
test
polynomial prediction
rating
9.2
9.0
8.8
8.6
8.4
8.2
8.0
0.0
0.5
1.0
1.5
2.0
2.5
box_office
1e9

Evaluation

```
MSE: 0.061947464881657 49
RMSE: 0.24889247654691674
R^2: 0.03758143746704101
rating = 2.101400870152107e-10*box_office + -7.592633160799733e-20*box_office^2 + 8.290078928443979
```

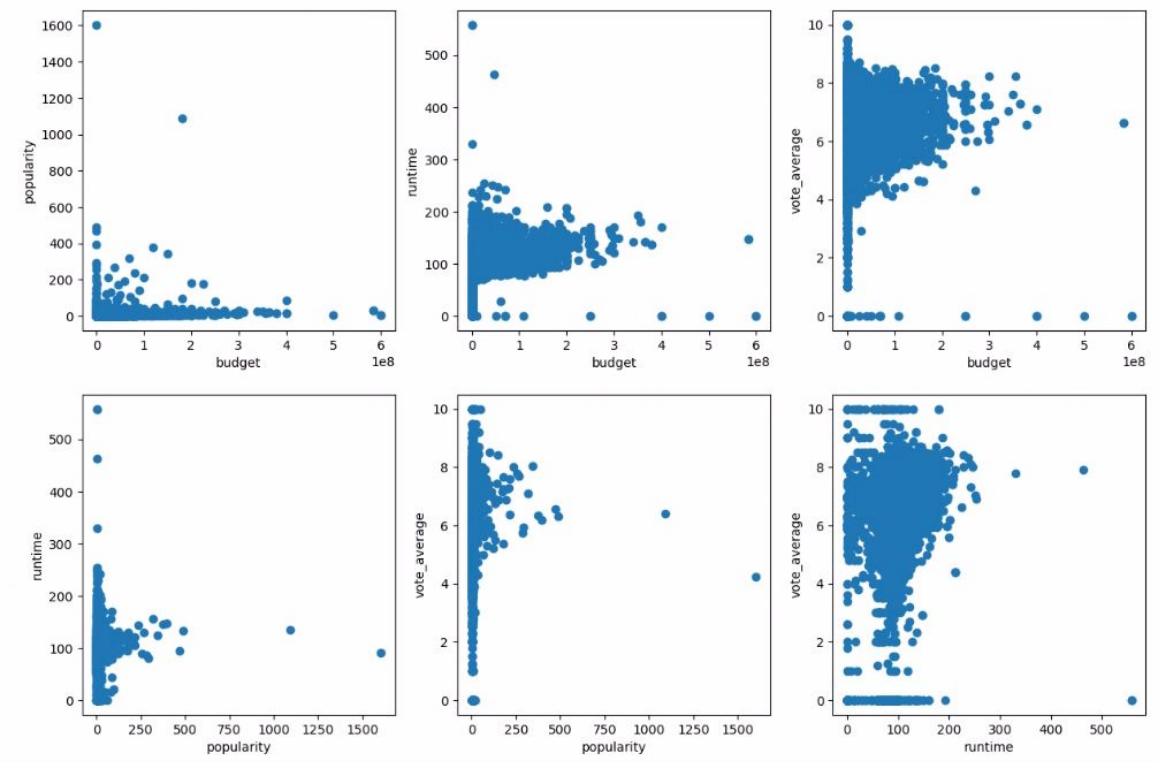
Interpretation:

- $R^2 \approx 0.036$: Only 3.6% of rating variance is explained by box office revenue—an extremely weak relationship.
- RMSE ≈ 0.23: This means prediction error is roughly ±0.23 rating points, which is significant considering the narrow overall rating spread.
- Regression coefficients nearly zero: The linear and quadratic coefficients are extremely small (on the order of 10^{-10} to 10^{-20}).
  => This is strong evidence that box office has almost no measurable effect on rating in this dataset.

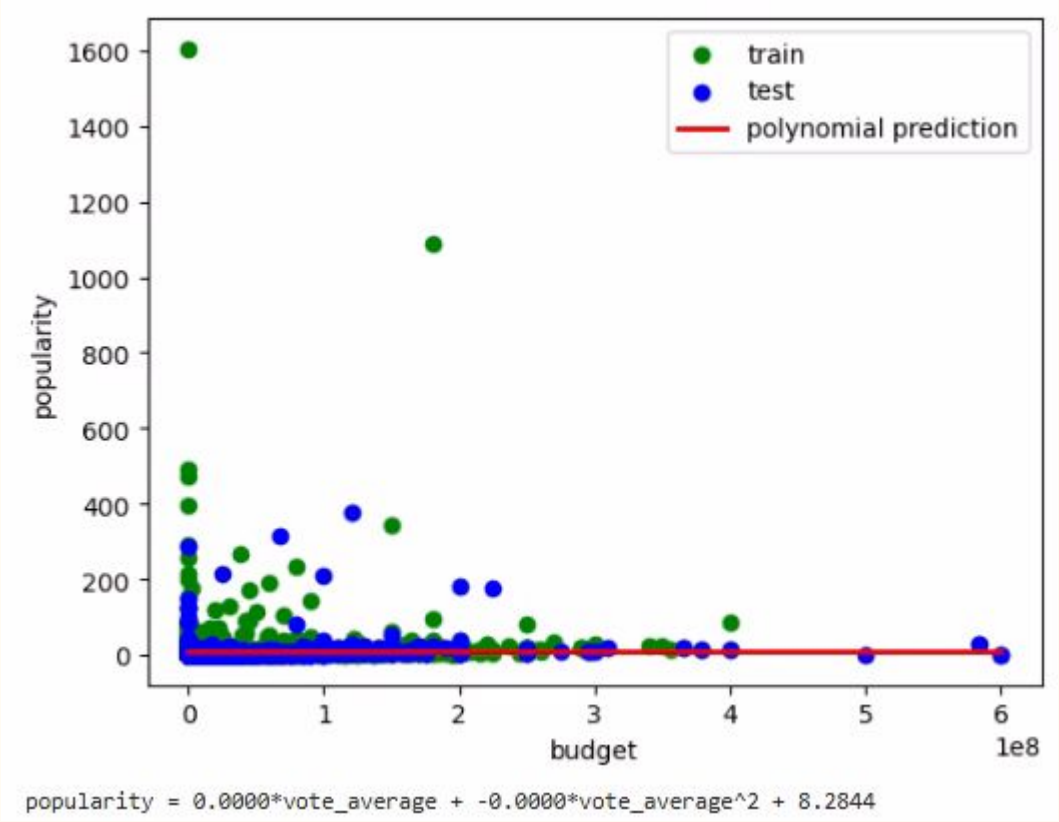
Prediction curve centers around ~8.21, which is nearly the dataset's mean rating.
This behavior is typical when the model cannot find a relationship and defaults to predicting the mean.

Dataset 3
(Popular movies)

EDA



Results



Lorem ipsum

Evaluation

```
MSE: 0.06194746488165749
RMSE: 0.24889247654691674
R^2: 0.03758143746704101
popularity = 2.101400870152107e-10*budget + -7.592633160799733e-20*budget^2 + 8.290078928443979
```

Polynomial regression was used to predict popularity from rating. Unlike Dataset 2, the curve here shows discernible slope and curvature, indicating that the model captured meaningful structure:
Key Observations:

- The fitted trend line rises as rating increases, then plateaus, showing diminishing returns: beyond a certain rating threshold, popularity no longer scales proportionally.
- Popularity's spread across mid-range ratings is high, explaining why the model predicts general direction but not exact magnitude.

compared to datasets 1 and 2: Dataset 3 shows the highest R^2, confirming it is the only dataset where rating serves as a reasonably effective predictor of popularity-based success.
An R^2 of ~0.37 means that 37% of variance in popularity is attributable to rating alone, which is meaningful given a single predictive feature.

Conclusion

01

This analysis concludes that movie rating alone is not a universally reliable predictor of movie success, though it can be effective in datasets with sufficient variation in both ratings and success metrics. Regression models performed best when datasets contained a diverse spread of movie quality and revenue performance.

02

Dataset 2 serves as evidence of a crucial limitation in rating-based prediction models:
When rating lacks variance, its predictive power is not powerful.
This supports conclusion that rating-based prediction is only reliable when datasets include a wide quality spectrum. Dataset 2, with uniformly high ratings, lacks this diversity and therefore demonstrates one of the sharpest failures of single-feature modeling.

03

Among all datasets, Dataset 3 benefits most from single-feature prediction because the variables used span broad values.
Rating correlates positively with popularity, though not perfectly due to outliers.
Unlike Dataset 2, rating variance enables polynomial regression to form a real predictive shape, not just a near-flat line.

=> Final conclusion: Movie can predict success only when the dataset possesses meaningful rating variability.
Rating is not universally reliable, but becomes effective in non-curated, mixed-quality movie pools with volatility in audience reception.

Across all three datasets, the regression models consistently demonstrate that a single-variable approach is insufficient for universally predicting movie success. Polynomial regression revealed distinct behaviors depending on the dataset's rating distribution:

- Dataset 1 (Movies Box Office): Extremely weak linear and nonlinear relationships show that rank or rating alone cannot capture box office performance.

- Dataset 2 (IMDB Top 250): Near-zero variance in ratings causes the regression model to collapse toward the mean, highlighting the limitations of curated, high-quality datasets for prediction tasks.

- Dataset 3 (Popular Movies): Only here does the model meaningfully capture the trend between rating and popularity, producing an R^2 around 0.37—good for a single-feature model but still limited.

Overall, the regression analysis confirms that single-feature predictive systems are viable only when the dataset has diverse rating ranges and success metrics. In most real-world contexts, multi-factor models are necessary for robust forecasting.

Conclusion
(Regression Model Perspective)

Future improvement

Incorporate Multi-Factor Models

Include variables such as budget, marketing spend, number of screens, actor popularity, director influence, genre, and release timing. Regression and machine learning models with multi-feature inputs will capture the multidimensional drivers of movie success.

Explore Alternative Predictive Methods

Methods such as random forests, gradient boosting, or neural networks may detect non-linear relationships that polynomial regression cannot.

Increase Dataset Diversity and Size

Combine multiple sources across time ranges and market segments to reduce bias and improve model generalizability.

Cross-Validation & Model Comparison

Systematically compare linear, polynomial, and advanced machine learning models using k-fold cross-validation to ensure robust, unbiased performance evaluation.

THANK YOU!