



IMDB Movie Success: An Analysis of

Movie Release Factors and Box Revenue

by Four-mula

Background

- **Subject:**
 - IMDb Movies
- **Motivation:**
 - What drives a movie's box office success
 - How to better predict movie's box office performance
- **Research Question:**
 - What production and release factors have the greatest impact on a movie's total revenue?



IMDb Top Box Office

1. Snow White
Weekend Gross: \$42M
Total Gross: \$45M
Weeks Released: 1
★ 1.6 (149K) Rate
2. Black Bag
Weekend Gross: \$4.3M
Total Gross: \$15M
Weeks Released: 2
★ 7.3 (11K) Rate
3. Captain America: Brave New World
Weekend Gross: \$4M
Total Gross: \$192M
Weeks Released: 6
★ 5.9 (82K) Rate
4. Novocaine
Weekend Gross: \$3.7M
Total Gross: \$16M
Weeks Released: 2
★ 6.8 (9.4K) Rate
5. Mickey 17
Weekend Gross: \$3.7M
Total Gross: \$40M
Weeks Released: 3
★ 7.0 (59K) Rate
6. The Alto Knights
Weekend Gross: \$3.2M
Total Gross: \$3.5M
Weeks Released: 1

IMDb Top 250 Movies

1. The Shawshank Redemption
1994 2h 22m R
★ 9.3 (3M) Rate
2. The Godfather
1972 2h 55m R
★ 9.2 (2.1M) Rate
3. The Dark Knight
2008 2h 32m PG-13
★ 9.0 (3M) Rate
4. The Godfather Part II
1974 3h 22m R
★ 9.0 (1.4M) Rate
5. 12 Angry Men
1957 1h 36m Approved
★ 9.0 (918K) Rate
6. The Lord of the Rings: The Return of the King
2003 3h 21m PG-13
★ 9.0 (2.1M) Rate
7. Schindler's List
1993 3h 15m R
★ 9.0 (1.5M) Rate
8. Pulp Fiction

Dataset

- **Origin**

- Kaggle
- Collected by Anand Shaw from the IMDb website

- **Columns**

- 903263 unique values
- 21 columns on characteristics of a specific movie (basic information, ratings, monetary values)

`title`: Movie name

`vote_average`: Average rating from 0 to 10

`vote_count`: The total number of votes

`status`: The current state of the movie

`release_date`: The date the movie was officially released

`revenue`: The total earnings (usually in USD)

`runtime`: The duration of the movie in minutes

`adult`: Whether classified as adult content

`budget`: Total cost of production (usually in USD)

`imdb_id`: The unique identifier for the movie on IMDb

`original_language`: Original language produced

`popularity`: Popularity of the movie based on views or ratings

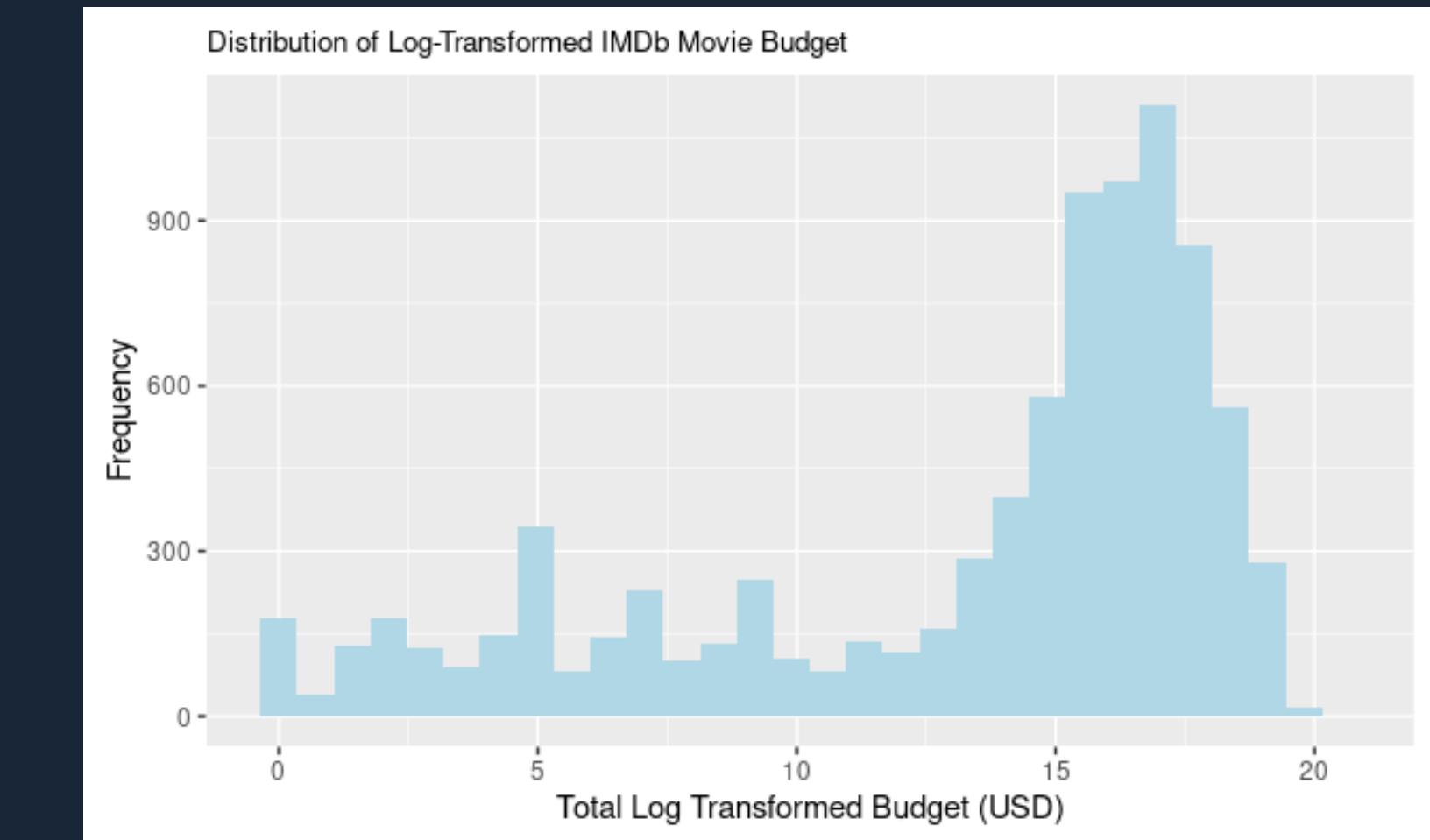
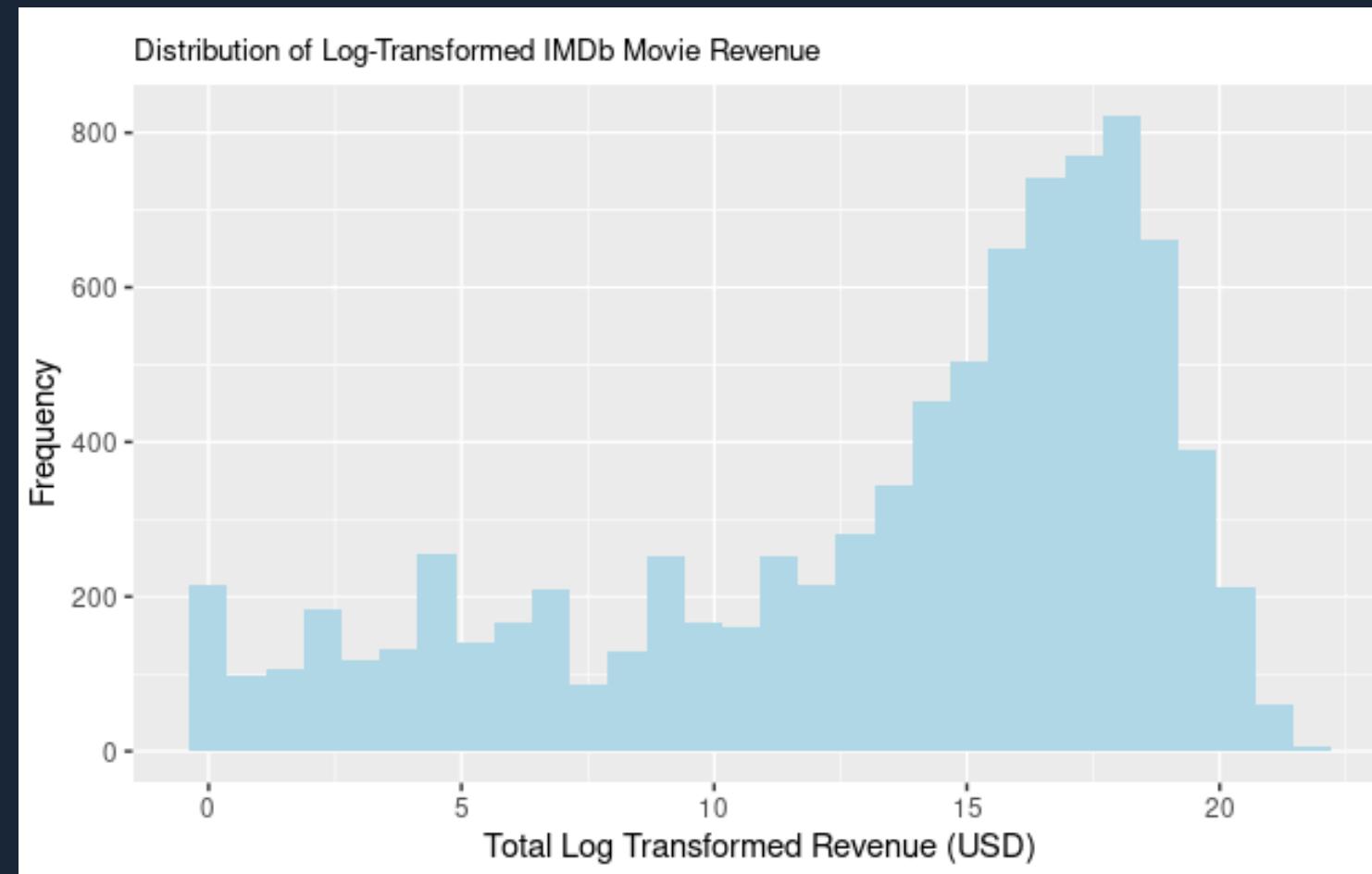
`genres`: The categories or genres the movie belongs to

`production_countries`: Production countries

`spoken_languages`: The languages spoken in the movie



Univariate EDA

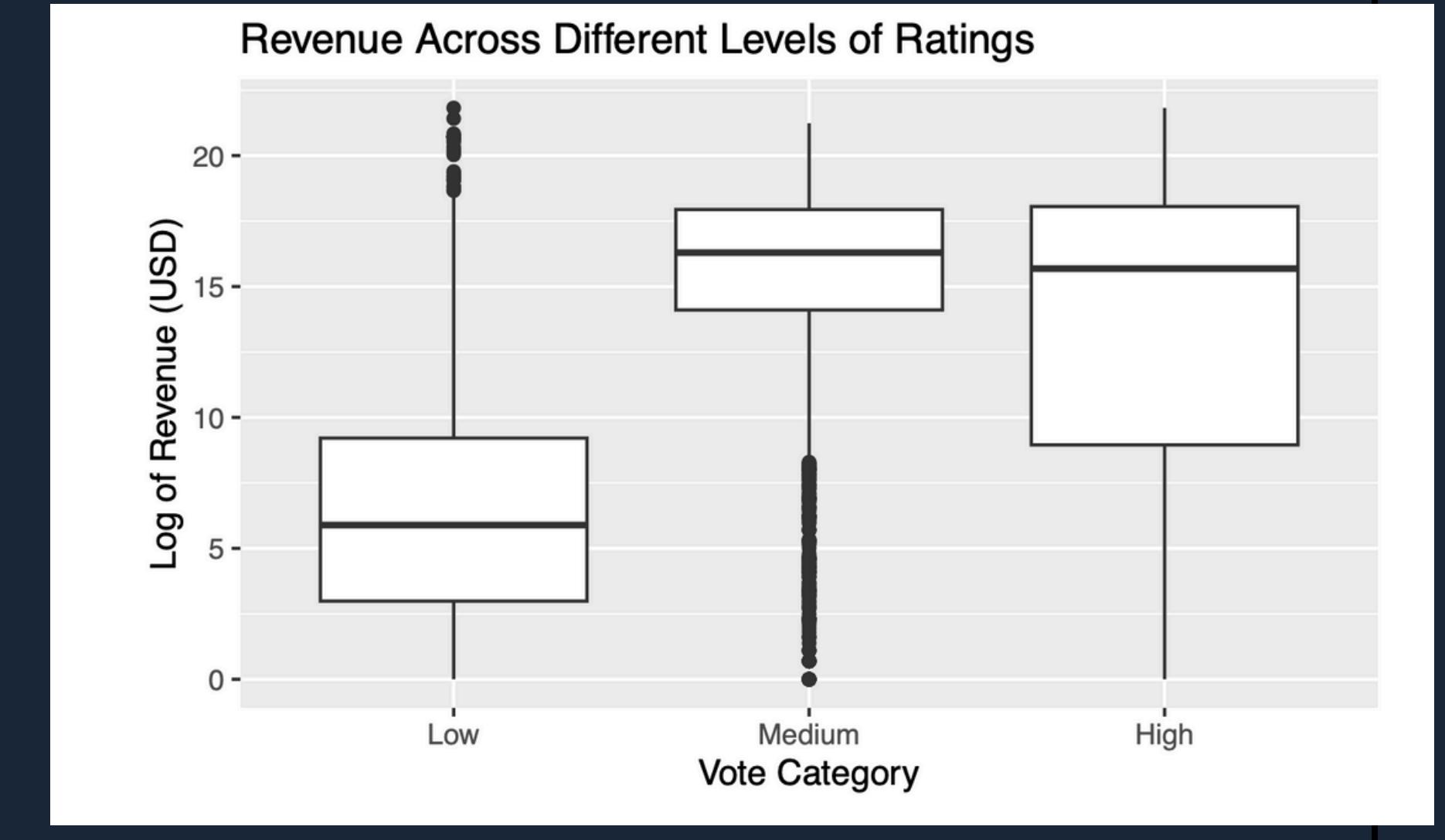
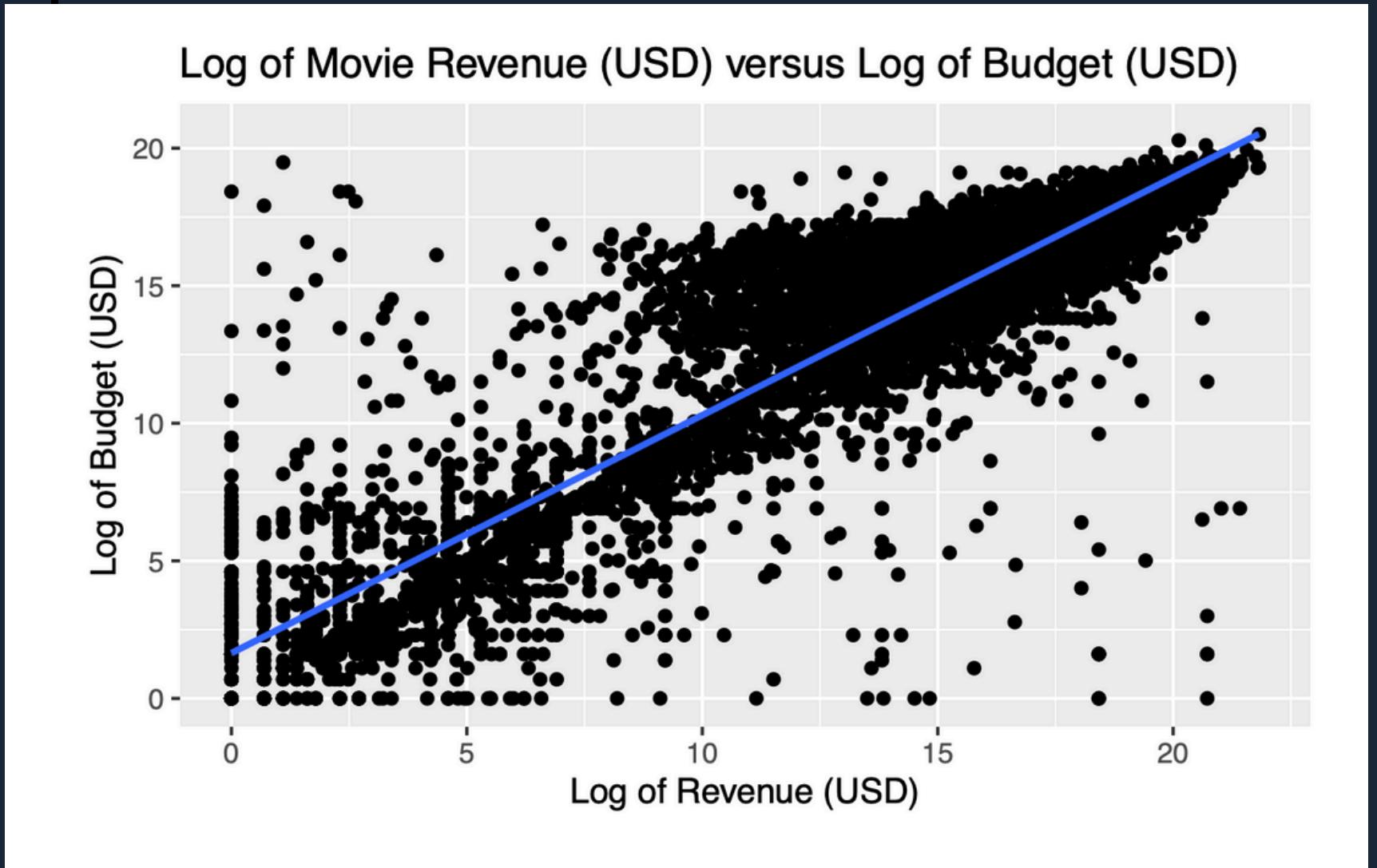


- Left-skewed, unimodal
- Center of approximately \$1,800,000 USD
- Spread of approximately \$18,600,000 USD

- Left-skewed, unimodal
- Center of approximately \$5,000,000 USD
- Spread of approximately \$25,000,000 USD



Bivariate EDA



Log of revenue versus log of budget

- Positive, linear correlation
- Many potential outliers
- High concentration of movies with high budgets and high revenue

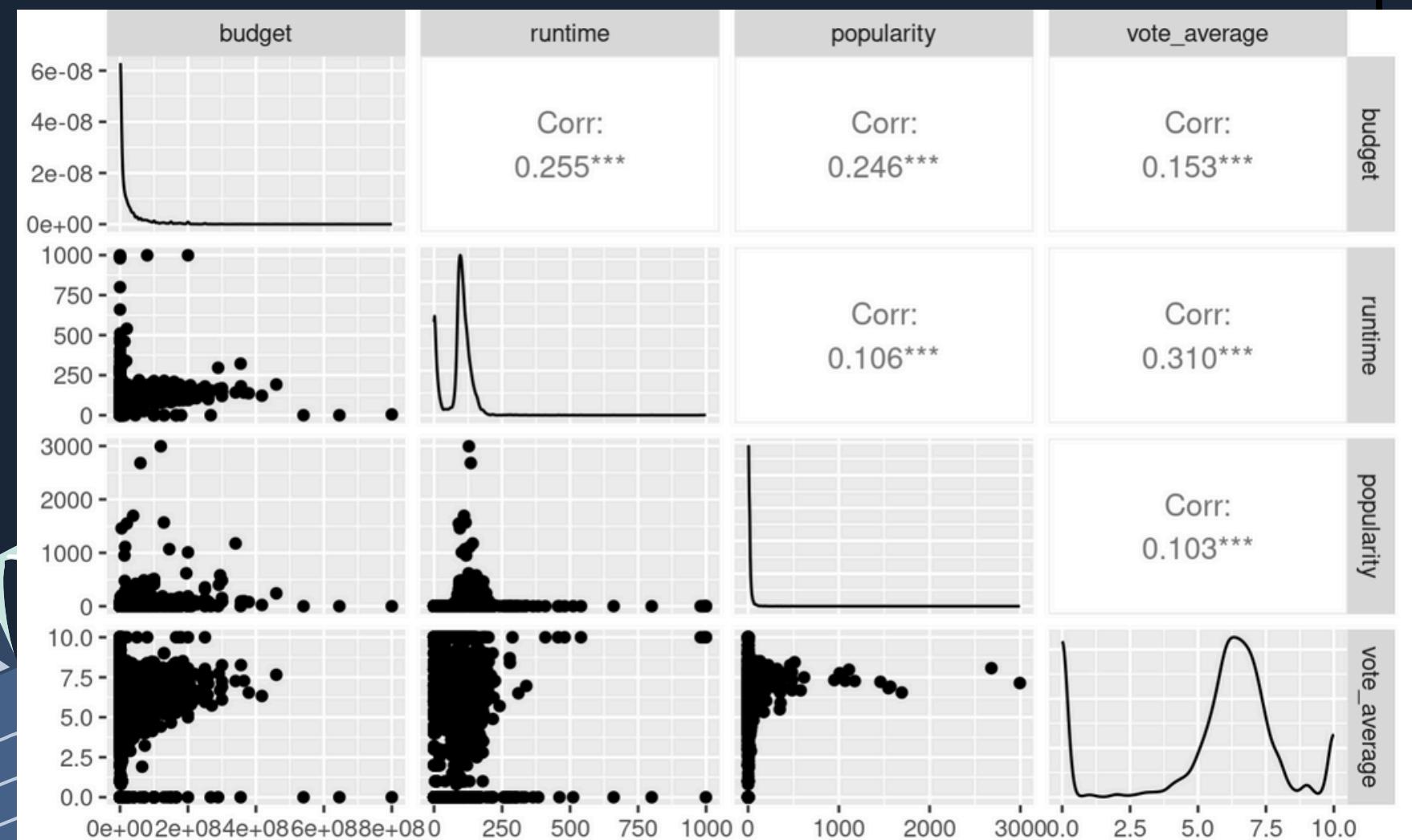
Vote category versus log of Revenue

- “Low”: movies with average rating 0-3
- “Medium” : movies with average rating 3-7
- “High” : movies with average rating 8-10
- Median revenue is highest for “medium” rated movies

Modeling Strategies/Results

- Potential interaction effects (from conceptual evidence)
 - budget & popularity - not significant
 - vote average & vote count - significant
 - action movies & runtime - significant

- Modeling strategies:
 - logistic model
 - select predictors
 - log transformations



Next Steps

- Further analyze collinearity with categorical predictors
- Interpret the model and test the research question/prediction (higher budgets, higher average ratings)
- Reflect on limitations and applications from our findings

