# Predicting Top Movie Ratings

Group B
Aidan Mantelmacher, Jiarui Chen, Simon Soendergaard, Sizhe Wang

# Agenda

1. Background
2. Predictive analytics problem and design
3. Data information
4. Data cleaning
5. Models/Model Selection
   a. Results
6. Insights
7. Takeaways

# Background

- Since 1995 on average more than 1 billion movie theater tickets has been sold in North America alone each year
- According to Statista, 83% of Americans 18 years and older have a subscription video service in 2022
    - Approximately 220 million Americans

THE SHAWSHANK REDEMPTION The Godfather

IRON MAN TITANIC


# Predictive Analytics Problem

- Writers, directors, and actors can heavily influence the quality of a movie
- Is there a simpler formula to produce a well-liked movie?
- Question: Can we use film features, budget, and sales to predict consumer ratings?
    - Model predicts film ratings on a scale of 1-10 and compares to average user-generated ratings from online

# Description of Data

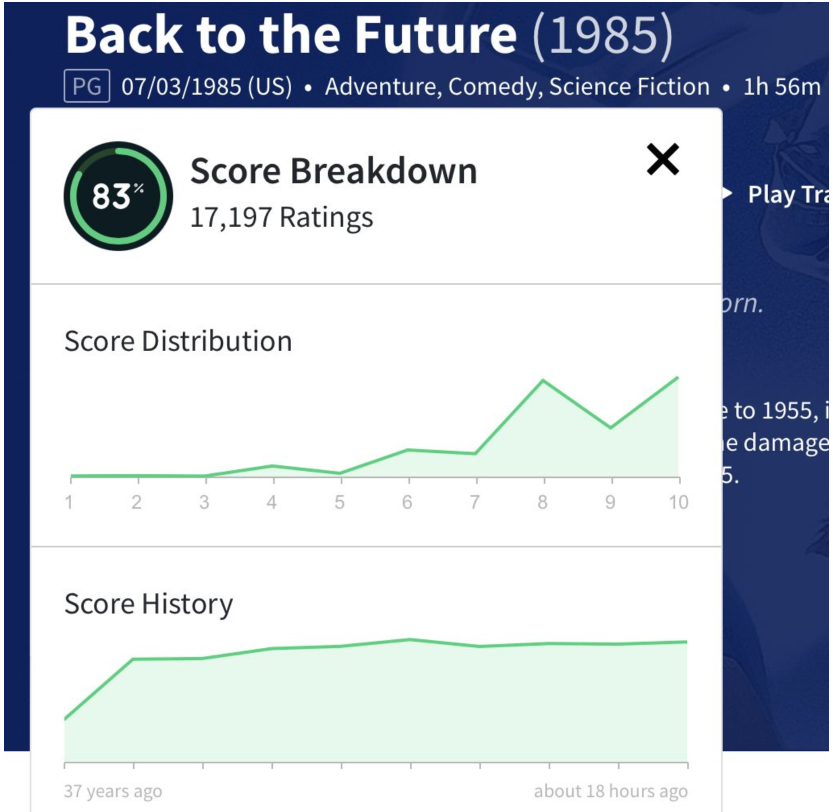
# Data Information

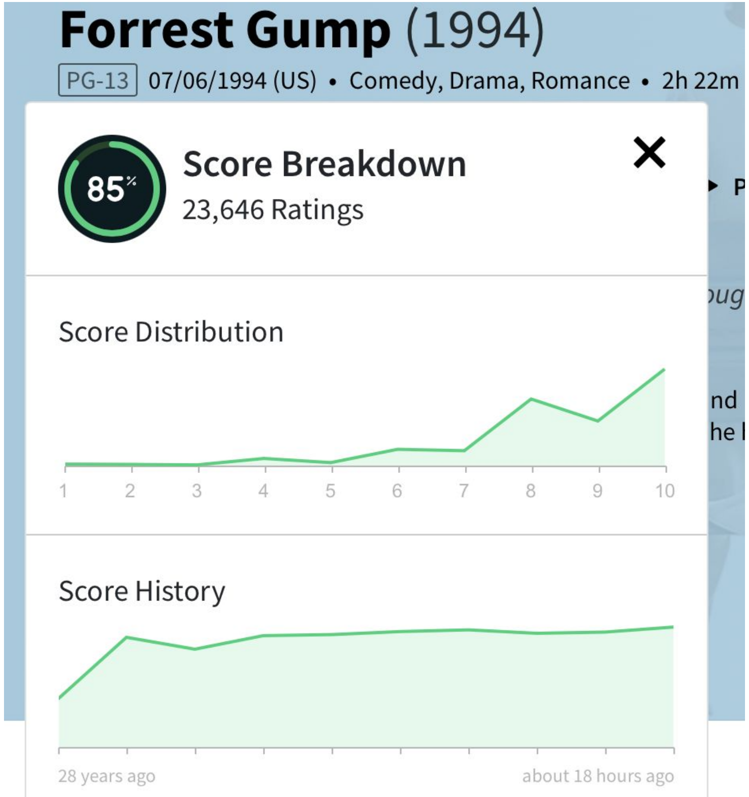
Movies Metadata (from TMDB (Movies Database))

- 737339 observations, 20 columns
- Title, Genre, Popularity, Release Date, Budget, Runtime, Votes, **Vote Average**

Highest Hollywood Grossing Movies

- 918 observations, 11 columns
- Title, Distributor, World Sales (in $), License (Appropriateness Rating)




Example scores from TMDB

# Data Cleaning/Preprocessing

- TMDB data has more than just films (TV shows, shorts, games) → Remove duplicate names and keep higher rated films
- Clean movie titles in both datasets to merge on them
- Remove unneeded attributes
- Remove observations with null, blank, or zero values
- Convert categorical values to dummy variables
  - 4 appropriateness ratings (G, PG, PG-13, R)
  - 32 distributors → 7 dummies (i.e. Warner Bros, 20th Century Fox, Disney, Universal)
    - 7 dummies chosen all had 10+ films in the dataset
  - 7 genres (Action, Adventure, Family, Scary, Sci-Fi, Comedy, Drama)
    - Most films had multiple genres, kept first/most fitting one

# Final Dataset

- 341 observations (films) with title, 24 prediction variables, 1 outcome variable
- Dummy Variables: G + PG + R + PG_13 + Warner_Bros + Universal + Disney + TwentiethC_Fox + Sony + Paramount + Lionsgate + Action + Adventure + Family + Scary + Sci_Fi + Comedy + Drama
- Example: {<u>Title</u>: 'A Quiet Place', <u>World_Sales</u>: 340,952,971, <u>Popularity</u>: 73.248, <u>Budget</u>: 17,000,000, <u>Runtime</u>: 91, **vote_average**: 7.397, <u>vote_count</u>: 12,137, <u>Release_Month</u>: 4, <u>PG_13</u>: 1, <u>Paramount</u>: 1, <u>Scary</u>: 1}
- Split 60/40 into train and validation set
  - Avoid overfitting
  - Provided better results than 70/30 or more

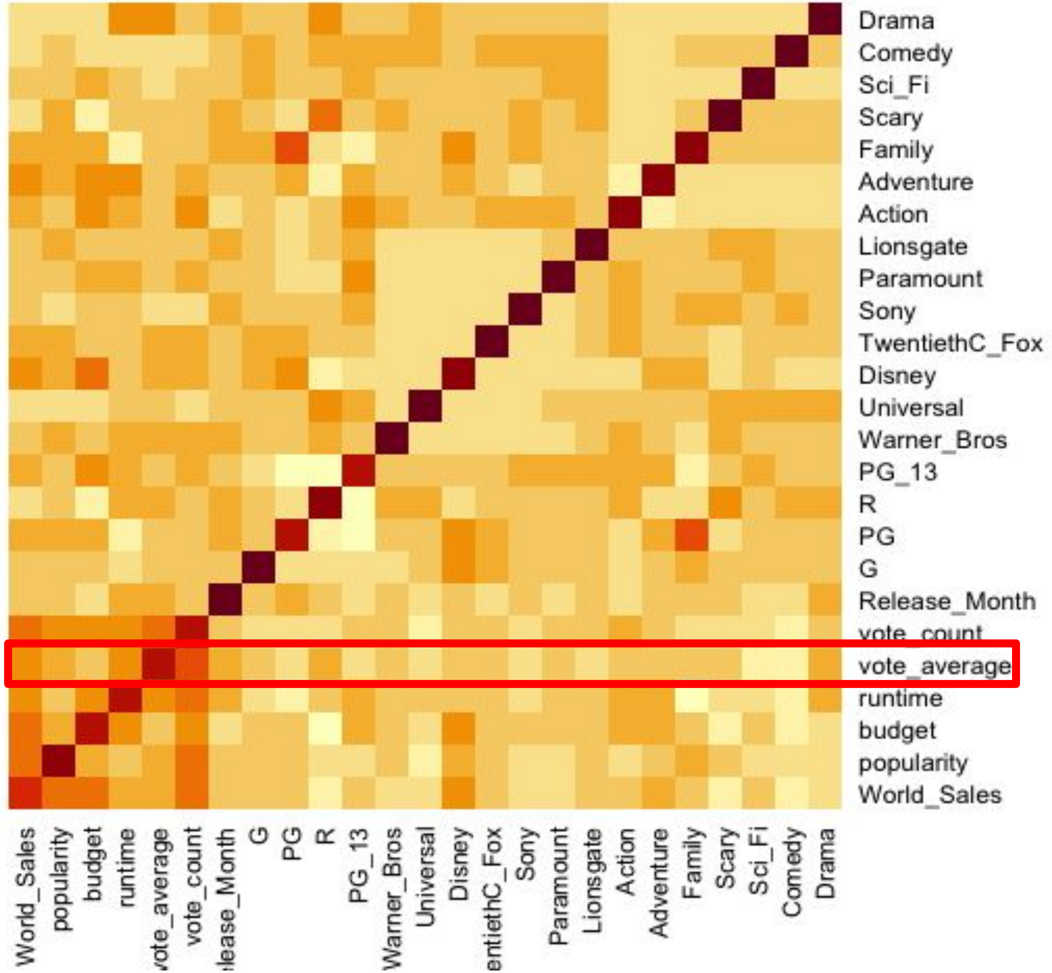
| Title | World_Sales | popularity | budget | runtime | vote_average | vote_count | Release_Month |
|---|---|---|---|---|---|---|---|
| Schindler's List | 322161245 | 44.655 | 22000000 | 195 | 8.565 | 13456 | 12 |
| The Dark Knight | 1005973645 | 70.953 | 185000000 | 152 | 8.503 | 28547 | 7 |
| Pulp Fiction | 213928762 | 74.145 | 8000000 | 154 | 8.492 | 24075 | 9 |
| Forrest Gump | 678226133 | 66.762 | 55000000 | 142 | 8.481 | 23593 | 6 |
| The Lord of the Rings: The Return of the King | 1146030912 | 95.023 | 94000000 | 201 | 8.478 | 20542 | 12 |
| Spider-Man: Into the Spider-Verse | 375540831 | 88.916 | 90000000 | 117 | 8.416 | 11926 | 12 |
| The Lord of the Rings: The Fellowship of the Ring | 897690072 | 115.983 | 93000000 | 179 | 8.388 | 21796 | 12 |
| Interstellar | 701729206 | 212.500 | 165000000 | 169 | 8.382 | 29785 | 11 |
| The Lord of the Rings: The Two Towers | 947495095 | 103.997 | 79000000 | 179 | 8.371 | 18942 | 12 |
| Inception | 836836967 | 84.678 | 160000000 | 148 | 8.359 | 32609 | 7 |
| Se7en | 327333559 | 65.788 | 33000000 | 127 | 8.358 | 17876 | 9 |
| Avengers: Endgame | 2797501328 | 236.025 | 356000000 | 181 | 8.278 | 22117 | 4 |
| Green Book | 321752656 | 58.716 | 23000000 | 130 | 8.244 | 9525 | 11 |

Example of dataset without the dummy variables

# Correlation Matrix

Heatmap showing the best fitting variables:

- Vote_count, World_Sales, runtime and the Drama categories are the best fit for vote_average
- Majority of the other variables seems to be more insignificant to the vote_average



# Modeling

# Model Selection Methodology

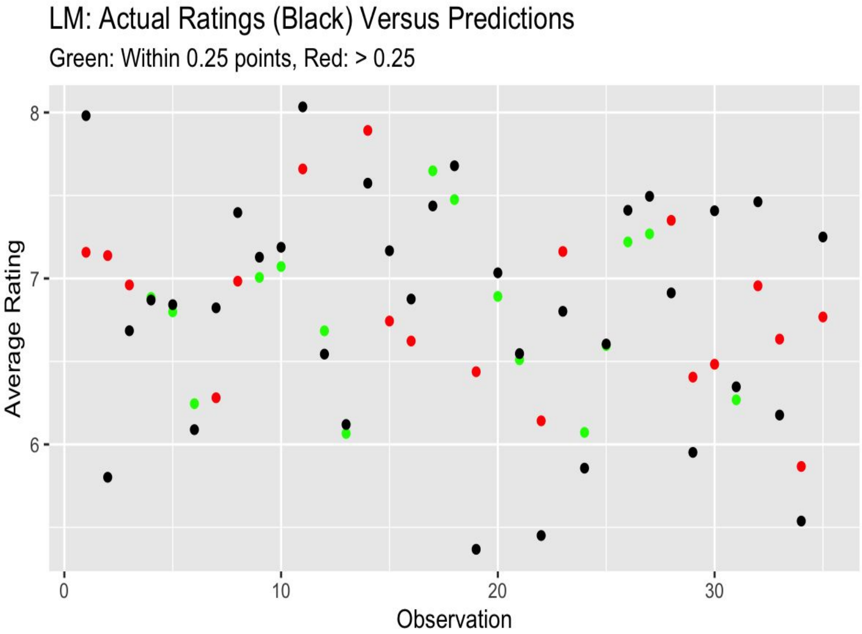
- Predicting numerical output (vote_average)
- Useful methods we have learned:
  - Multiple Linear Regression
  - Stepwise Regression
  - Neural Network
  - Logistic Regression (classification)
  - KNN (mainly for classification)
    - Did not learn KNN for regression
    - Need substantial amount of observations for # of predictors

# Multiple Linear Regression

- Performed on all variables
  - AIC: -273.6
- R-Squared: 0.6629, Adj: 0.62
- RMSE for training: 0.4546
- RMSE for validation: 0.5149

| Predicted | Actual | Residual |
|---|---|---|
| 7.158 | 7.981 | 0.823209 |
| 7.138 | 5.802 | -1.336052 |
| 6.961 | 6.685 | -0.275845 |
| 6.885 | 6.870 | -0.015130 |
| 6.799 | 6.842 | 0.042981 |

| | Estimate |
|---|---|
| (Intercept) | 5.2956982650851 |
| World_Sales | 0.0000000000977 |
| popularity | -0.0008147317836 |
| budget | -0.0000000024266 |
| runtime | 0.0098361416119 |
| vote_count | 0.0000747729598 |
| Release_Month | 0.0183008431321 |
| G | 0.3547375637829 |
| PG | 0.2267556144616 |
| R | 0.1156925604161 |
| PG_13 | NA |
| Warner_Bros | -0.1037503692582 |
| Universal | 0.1303504519254 |
| Disney | 0.1936399987978 |
| TwentiethC_Fox | 0.0520128187181 |
| Sony | 0.0584073214714 |
| Paramount | 0.0923006330407 |
| Lionsgate | 0.0707137583241 |
| Action | -0.2710753679224 |
| Adventure | -0.3016284469747 |
| Family | 0.0845626959650 |
| Scary | -0.2124533896252 |
| Sci_Fi | -0.6900816288388 |
| Comedy | -0.3584056515873 |
| Drama | 0.1423592436294 |



# Stepwise Regression

- Similar methodology for multiple linear regression, but eliminating variables
- Utilize “both” (forward and backward) method of stepwise regression
- Formula with lowest AIC (-289.2):
  - `vote_average ~ budget + runtime + vote_count + G + PG + Warner_Bros + Disney + Action + Adventure + Scary + Sci_Fi + Comedy`
- Adjusted R-Squared: 0.63
- Lower AIC should lead to better model
- RMSE on validation data is: 0.5176

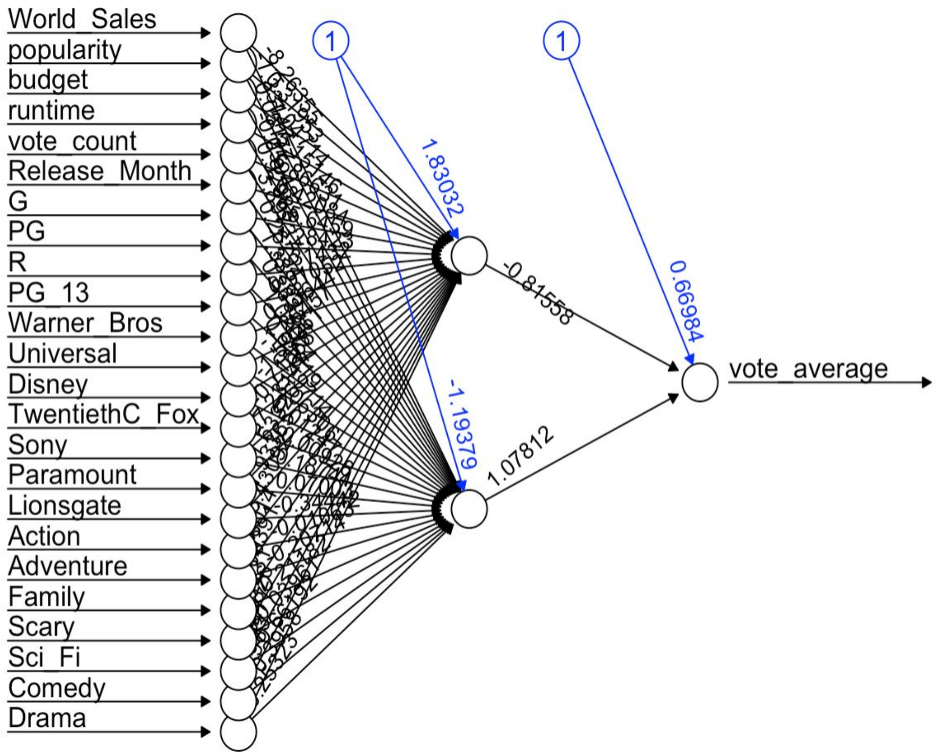
```
Coefficients:
                   Estimate
(Intercept)  5.478784035099
budget      -0.000000002796
runtime      0.011529478986
vote_count   0.000072744498
G            0.338343179027
PG           0.193305154627
Warner_Bros -0.154169778160
Disney       0.168328715375
Action      -0.398892692880
Adventure   -0.454128181193
Scary       -0.313452253886
Sci_Fi      -0.831061180435
Comedy      -0.489991178188
```

# Regression Discussion

- The reduced stepwise regression fits the model slightly better according to the adjusted R-squared metric
- Original regression has a lower RMSE on the validation dataset
- 24 versus 12 predictor variables
    - AIC score of -273.6 versus -289.2
    - Simplified model with very similar performance
    - In general, more predictors lead to overfitting
- We asked ourselves, is the data actually linear?

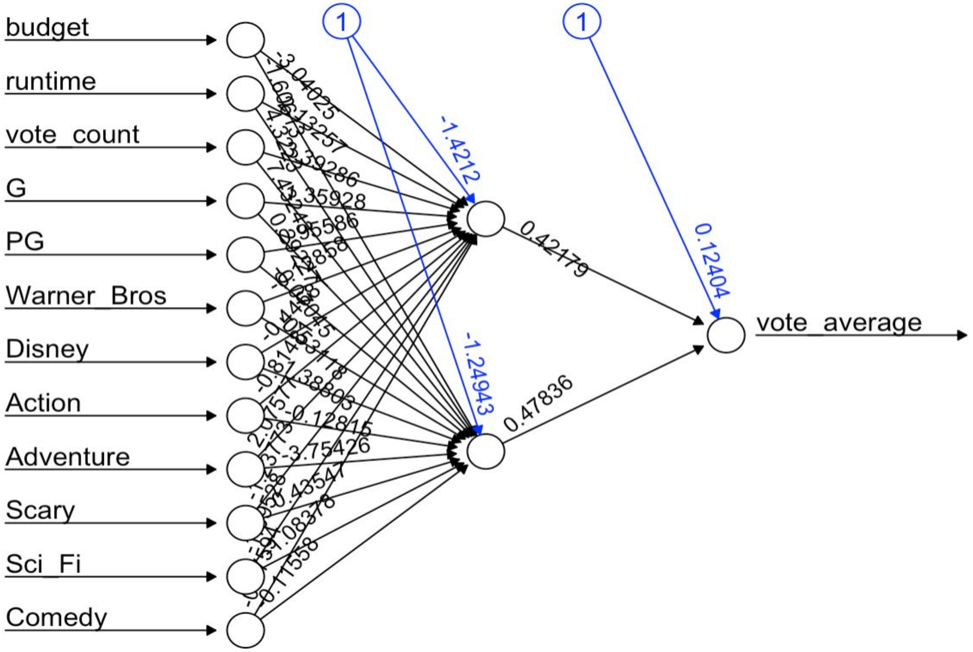
# Neural Network

- Normalize (scale) data from 0-1 using preprocess
- Attempted many combinations of nodes and layers
    - Trial and error method
    - Analyze on RMSE
- 2 nodes, 1 hidden layer yielded best results
- 2 nodes limited RMSE on validation set: 0.5328



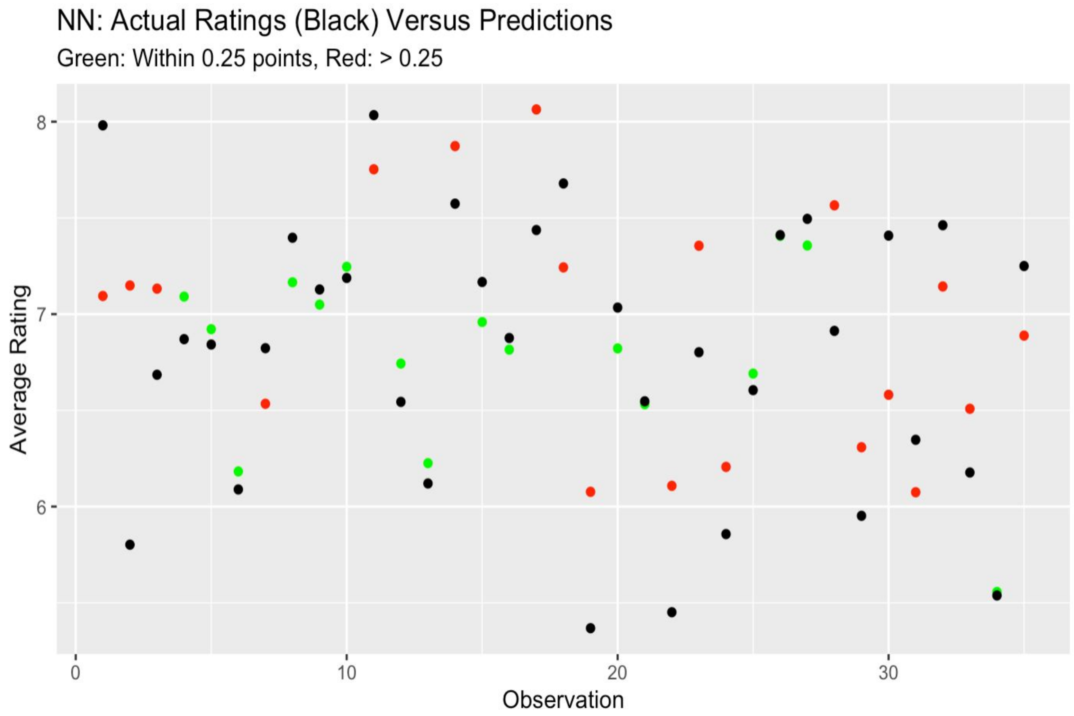
# Second Neural Network

- Fit model on equation from stepwise regression
    - 12 variables
        - 9 dummies
- 2 nodes, 1 hidden layer
- RMSE on validation: <u>0.5021</u>



# Visuals for Second Neural Network

| Predicted | Actual | Residual |
|---|---|---|
| 7.094 | 7.981 | 0.88887 |
| 7.149 | 5.802 | -1.35218 |
| 7.132 | 6.685 | -0.45045 |
| 7.091 | 6.870 | -0.21856 |
| 6.922 | 6.842 | -0.05376 |



# Insights

- All of our models have an RMSE around 0.50.
  - Example: A film with an actual rating of 7.5 is predicted to be 7.0 or 8.0
- This is not amazing, but it still shows that the variables can be used to predict the vote average
- Stepwise regression: Positive coefficients for the models are runtime, vote count, G, PG, and Disney
  - Similar to what we saw from the heatmap

# Takeaways

- A few variables seem to have a significant impact on the vote average ⟶ Errors
- Difficult to choose which model works best
- Neural networks randomness and no proven methods for choosing nodes/layers leads to uncertainty
    - Model, weights, and RMSE change each time it's run
- Multiple linear regression is a proven and simple method
- Stepwise regression does the same, but with lesser variables on a consistent basis

| Model | RMSE |
|---|---|
| MLR | 0.5149 |
| Stepwise | 0.5176 |
| 1st NN | 0.5328 |
| 2nd NN | 0.5021 |

Questions?

# Sources

https://www.kaggle.com/datasets/sanjeetsinghnaik/top-1000-highest-grossing-movies?select=Highest+Holywood+Grossing+Movies.csv

https://www.kaggle.com/datasets/akshaypawar7/millions-of-movies

https://www.the-numbers.com/market/

https://worldpopulationreview.com/countries/united-states-population

https://www.statista.com/statistics/318778/subscription-based-video-streaming-services-usage-usa/