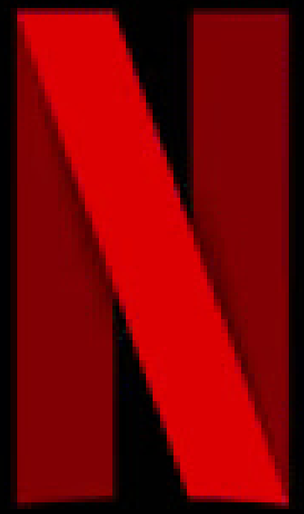
GROUP85 Home TV Shows Movies New & Popular My List KIDS

UNVEILING NETFLIX SUCCESS: ANALYZING MOVIE FACTORS

Presented by:

- Vincent Nguyen 1168263
- Issac Tse 834150
- Christopher Kenneth Yapharis 1360684
- Koquiun Li Lin (Franklin) 1319881



INTRODUCTION OF RESEARCH QUESTION

Research Question

"Which attributes serve as predictors of a movie's success on Netflix?"

Data Sources

Data is gathered from "titles" and "credits" csv files, offering insights into movie genres, runtime, production countries, and key personnel involved in filmmaking.

METHOD, TECHNIQUES, AND TOOLS

Machine Learning Models

- Supervised model: Linear regression
- Unsupervised model: K-means clustering

Data Pre-processing and Wrangling

- Data merging
- Natural language processing:
    - Tokenization
    - Stemming
- Data imputation
- Text processing: TF-IDF
- Linear regression

Analytical Techniques

- Pearson Correlation
- K-means Clustering

Visualization

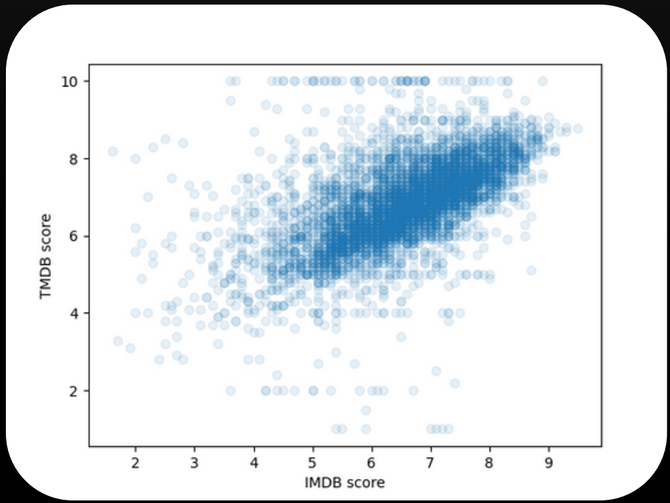
- Bar plots
- Box plots
- Scatter plots
- Tables

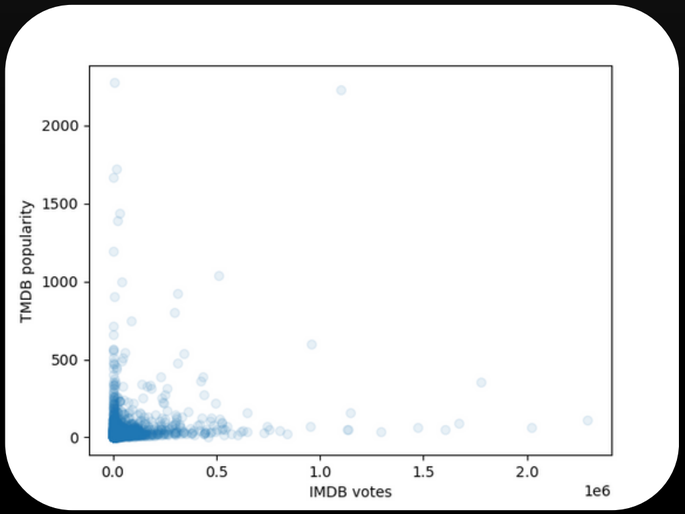
Libraries

- Pandas, Numpy, Matplotlib, Seaborn and Scikit-Learn

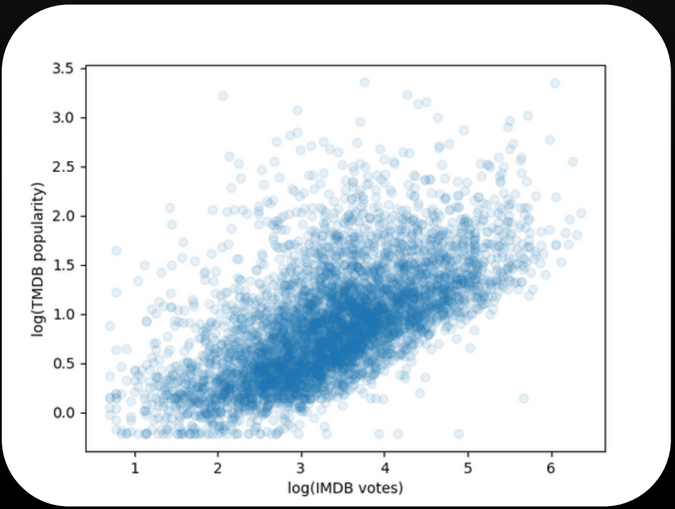
MACHINE LEARNING MODELS

LINEAR REGRESSION



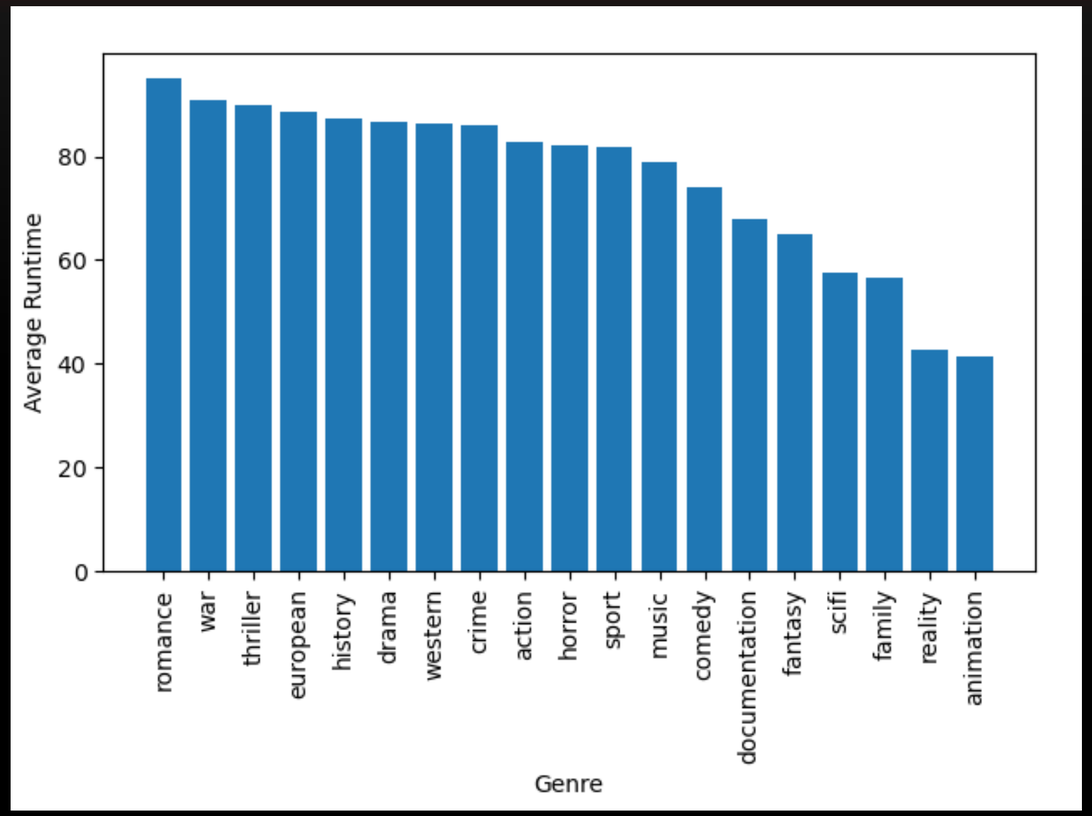


Untransformed data

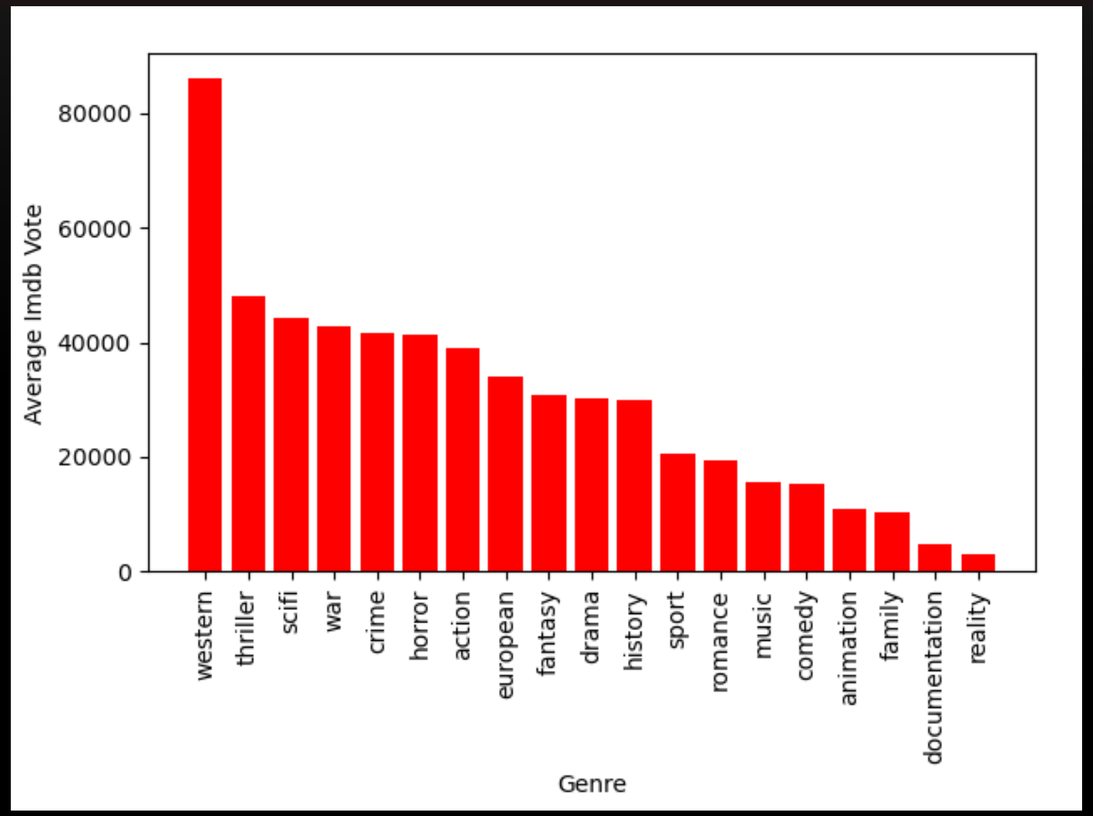


Log transformed data

K-MEANS CLUSTERING



Bar plot of average runtime of all genres



Bar plot of average imdb_votes of all genres

LIST OF FINDINGS AND IN-DEPTH INTERPRETATION

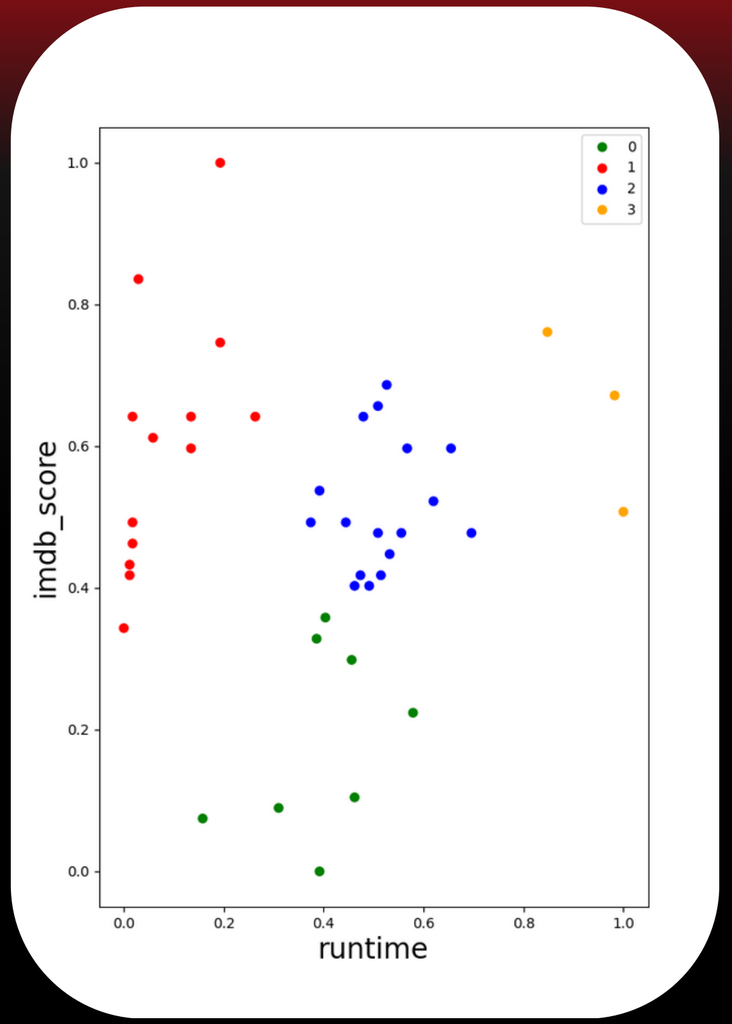
Runtime and IMDB Scores

Findings

- Low negative Pearson correlation except for **Western Genres**
- Short-runtime movies tend to have higher IMDB scores compared to long-runtime ones
- Medium runtime cluster performs worse than other clusters

Cluster	runtime	imdb_score
Red	0.083221	0.605052
Blue	0.517716	0.514486
Green	0.393275	0.184701
Orange	0.943470	0.646766

Centroid of Western Genre



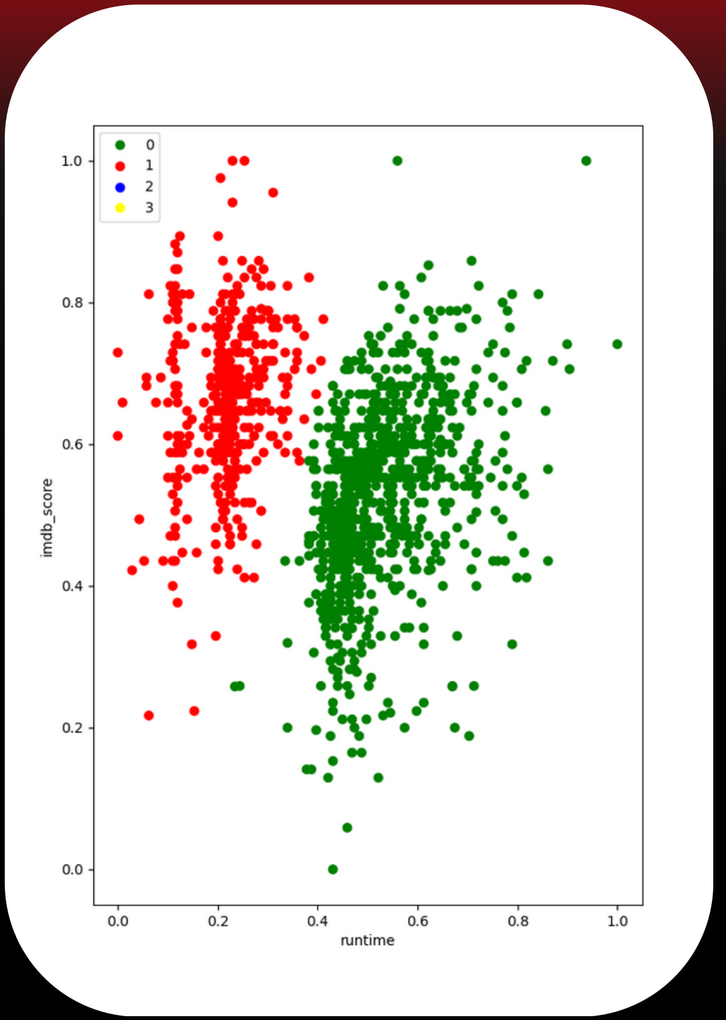
Clustering of runtime vs imdb_score of Western movies with k = 4

LIST OF FINDINGS AND IN-DEPTH INTERPRETATION

Runtime and IMDB Scores

Interpretation

- Longer movies tend to decrease viewer retention.
- The audience prefers more concise storytelling.
- Filmmakers might avoid having movies with medium runtime.
- Runtime is not a definitive predictor, but consistent findings across methods offer potential contextual use for filmmakers



Clustering of runtime vs imdb_score of Thriller movies with k = 2

LIST OF FINDINGS AND IN-DEPTH INTERPRETATION

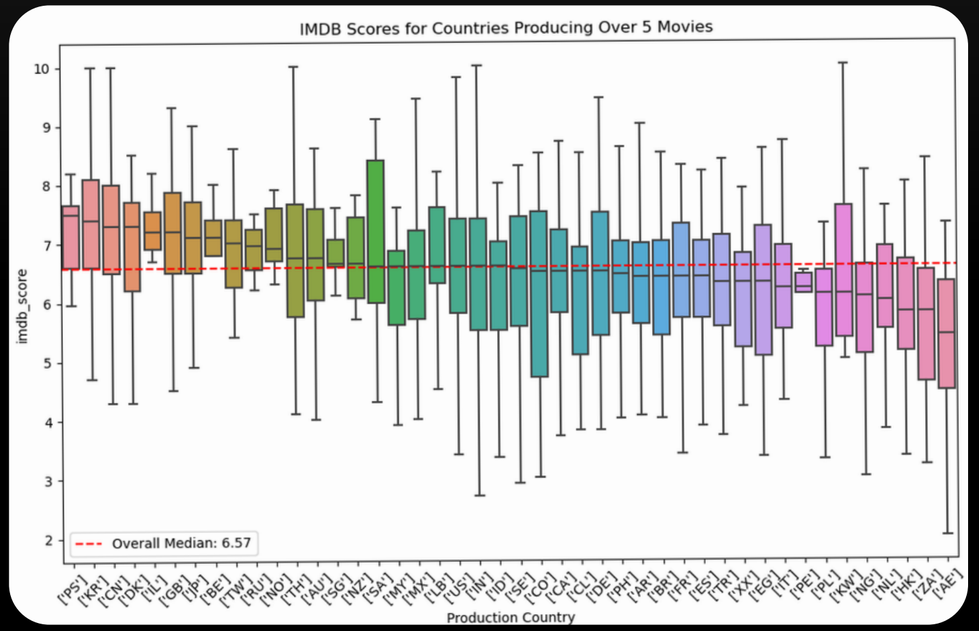
Production Countries and Success

Findings

- Rich countries benefit from international collaboration
- Niche countries can thrive independently
- Positive correlation with the industry size and movie's success

Interpretation

- International collaboration benefits rich countries, highlighting the importance of diversifying talent and resources
- Niche countries can succeed with unique, independent productions
- A country's film industry size impacts movie variety, emphasizing resources and capabilities



Box plot of counties producing more than 5 movies

LIST OF FINDINGS AND IN-DEPTH INTERPRETATION

Actors and Directors

Findings

- Director's Impact: A strong correlation exists between a movie's IMDB score and the average success of its directors, emphasizing the vital role of a successful director in movie/TV show success
- Actor's Influence: There is a moderate correlation between a movie's IMDB score and the average success of its actors, highlighting the importance of good actors, albeit diluted compared to directors, likely due to the larger number of actors per show

Interpretation

- Director's Vital Role: Strong correlation underscores the director's vital role in movie/TV show success.
- Actor's Importance: Moderate correlation highlights the importance of good actors, albeit less pronounced than directors.

LIMITATIONS AND IMPROVEMENT OPPORTUNITIES

Limitations

Cluster Interpretation Challenges

- VAT analysis yielded different interpretations of cluster numbers, potentially causing confusion and inconsistent representation of results.

Assumption on IMDb Metrics

- The assumption that IMDB metrics are a better fit for measuring movie success than TMDB metrics was made based on external information not verified by the provided data. We need external variables to get a better prediction

Interaction between variables not sufficiently investigated

- Only the interaction between genre and runtime was investigated. There are much more interactions between other variables.

LIMITATIONS AND IMPROVEMENT OPPORTUNITIES

Improvement Opportunities

Broaden Predictor Consideration

- Explore a wider range of predictors to obtain a more comprehensive understanding of movie success factors.

Enhanced Cluster Validation

- Employ advanced cluster validation techniques to provide more consistent and reliable cluster results.
- Investigate the use of alternative clustering algorithms that may better suit the data and research objectives.

Data Verification

- Verify external information used to support assumptions, such as IMDB metrics being better suited for movie success measurement.

Thank you for joining us on this journey to uncover the secrets of the success of Netflix's movies