# Analysis of Factors Affecting Movies and TV Shows Scores and Popularity

COMP20008 23S2 Assignment 2

**Group Members:**

- Xinxiu Dai 1313108 xinxiud@student.unimelb.edu.au
- Muhan Chu 1346862 muhanc@student.unimelb.edu.au
- Jiayuan Li 1404463 jiayuan6@student.unimelb.edu.au
- Qiurong Li 1383954 qiurong@student.unimelb.edu.au

# Table of Contents

# 1. Executive Summary

The objective of this project is to analyze the factors influencing scores and popularity of movies and shows. This project can assist streaming platforms in more effectively identifying movies or shows that are better received by audiences and enable viewers to discover highly rated movies and shows that align with their preferences, which can be further used for Netflix management team to see what visual productions are worth being imported more easily.

After a comprehensive data exploration analysis and data preprocessing, a linear regression model on imdb scores and a decision tree classification model on tmdb popularity are employed to achieve the analysis and prediction based on given dataset. The aims are to identify the characteristics that are associated with higher score in movies and shows and which turn out to be popular among the viewers. The linear regression model does not fit well with the limited given data but it is found that tmdb scores and imdb votes affects the imdb scores most as they show relatively higher correlation than other features. As for the decision tree model, it has successfully identified which visual production is popular with accuracy as high as 72%.

However, modelling and analysis can be improved if more information is given. We can explore ways to further deepen and broaden our understanding and analysis of the factors influencing the ratings of film and television productions, such as if the movie or TV show is award winning. More advanced machine learning algorithms will also help.

# 2. Introduction

To reveal what affects the scores and popularity of movies and TV shows, a linear regression model and decision tree model are used after a comprehensive exploration of the data and data preprocessing. We have also found some interesting insights when exploring the data, which will be expanded later in section 4 Data Exploration and Analysis.

After model evaluation, various visualization techniques such as histograms, bar charts, heat maps, scatter plots and pie charts are used to help understand the data distribution, feature importance and model performance.

# 3. Methodology

## 3.1 Dataset Understanding

### 3.1.1 Data Source

Two experimental datasets, Titles.csv and Credit.csv, are available for the analysis. They are mainly derived from two major online databases, i.e., Internet Movie Database (IMDb) and The Movie Database (TMDb), both of which are regarded as authoritative platforms for providing information on film and television productions.

### 3.1.2 Data Overview

The title.csv contains all the basic information for each movie and show. Before any data preprocessing, there are 5850 rows and 15 fields. Main fields including:

- title: name of the movie and show
- type: which category the visual media production falls under. There are only two types: MOVIE and SHOW.
- description: a concise and informative summary for what the movie or show is about.
- release_year: when the visual work is released.
- age_certification: categorize movies and shows based on the content into different age groups.
- runtime: a continuous field of how long the production is in minutes.
- genres: a list of categories of what the movie or show is about.
- production_countries: a list of which countries are involved in movies or TV shows production.
- tmdb_score: a continuous filed ranges from 1-10 points rated by viewers and sourced from tmdb
- imdb_score: similar to tmdb_score but sourced from imdb
- imdb_votes: the total number of votes on IMDb for the production
- tmdb_popularity: the index reflects the popularity of the work on TMDb, it is a dimensionless index and usually contains decimals.

Credits.csv file is also provided and contains 77,801 rows and 5 columns of data. Each row of data is a production in which the director or actor has acted, and each column is part of the information of the film or television production, such as the role played by the actor and the id of the production.

However, it's believed that directors and actors are not decisive factors for a visual production to be highly rated or popular. For the sake of our analysis, only title.csv is used for further exploration and modelling.

## 3.2 Data Pre-processing

For a more in-depth data analysis, we cleansed and improved the existing dataset mainly by removing and filling missing values, data scaling and encoding. To address the varying conditions within each feature, we applied the following data pre-processing techniques to manipulate the data and prepare the data for modelling.

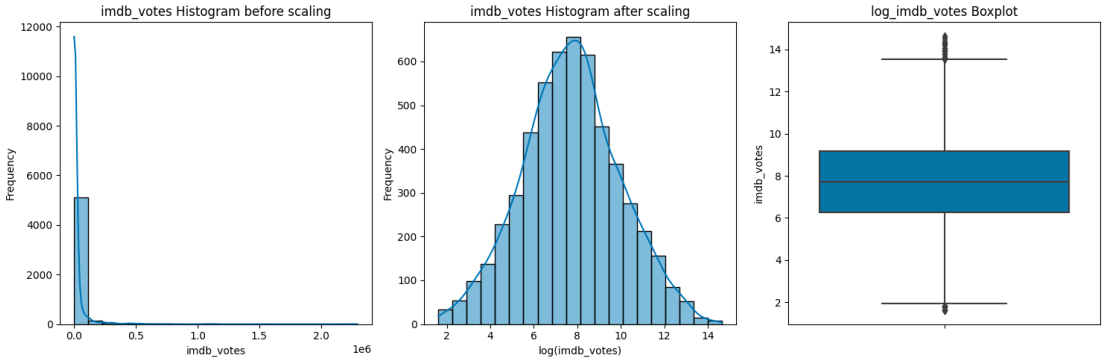
### 3.2.1 Handling Missing Values

The data describes most of the basic information about each visual production. However, the data is not complete. We found missing values mainly in below fields:

- Only one missing title and 18 missing descriptions are found in the data. We removed those missing values as they are very small volumes.
- As for age certification, 2619 rows are found missing, occupying almost 45% of the whole dataset, which is almost impossible for us to fill in the missing values according to the available information. Also, age certification is irrelevant to our research topic because it is commonly accepted that age certification is mainly used for age groups classification and cannot determine whether a movie or TV show is good or bad. Therefore, age certification is not used for this project.
- 229 empty lists were found in the field of production countries. It is filled with “US” because “US” produces most of the visual works.
- 59 empty lists are found in genres and the filed is filled in with “drama” as most of the visual works are drama.

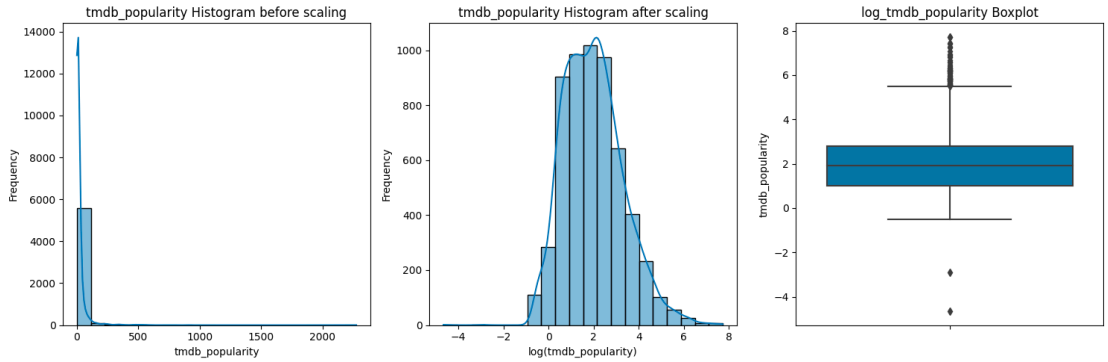
### 3.2.2 Adjusting Data Ranges via Scaling

Both imdb_votes and tmdb_popularity represent how popular the movie or TV show is among the viewers, but they have different data units. We used log scaling technique to process both fields to make sure they are compatible and can be used for modelling.




*Picture 1. Comparison of imbd_votes before and after scaling*

It can be seen from Picture 1, the histogram shows a nice normal distribution after scaling. Therefore, mean value of the scaled votes is used to fill in the missing value.




*Picture 2. Comparison of tmdb_popularity before and after scaling*

Picture 2 shows that tmdb_popularity is a bit right skewed after scaling and a few outliers are detected from the boxplot. Therefore, outliers were removed, and missing values are filled with median. Because median is more robust to outliers and provides better estimate of central tendency than mean value for skewed data.

### 3.2.3 Data Transformation Via Encoding

For classification modelling, below categorical features are encoded into numerical format:

- Type: movie is 0 and show is 1
- Genres and production countries are transformed from text into list first, then one hot encoding is applied to both.

Additionally, outliers in runtime were also removed. Overall, through the application of these techniques, the dataset has been rid of redundant data, streamlined, and ready for modelling and analysis.

## 3.3 Modelling using Machine Learning Methods

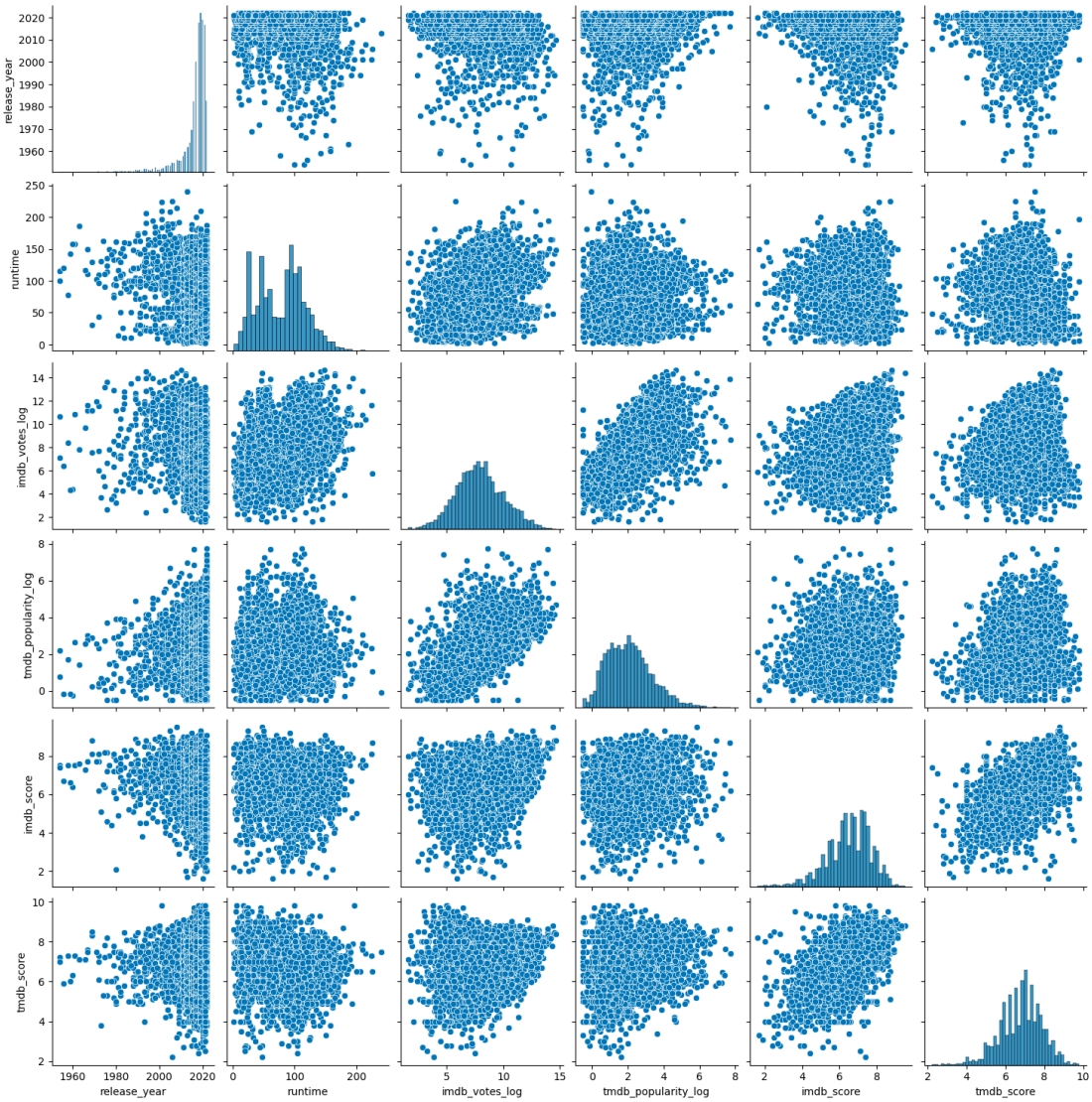
Given our objective to predict certain outcomes based on available features using supervised methods, the implementation of regression and classification models is imperative. Standardization is applied to numerical features to facilitate feature comparisons in both modelling process. Both models split data into 80% training set and 20% test set to evaluate the performance and generalization.

### 3.3.1 Linear Regression

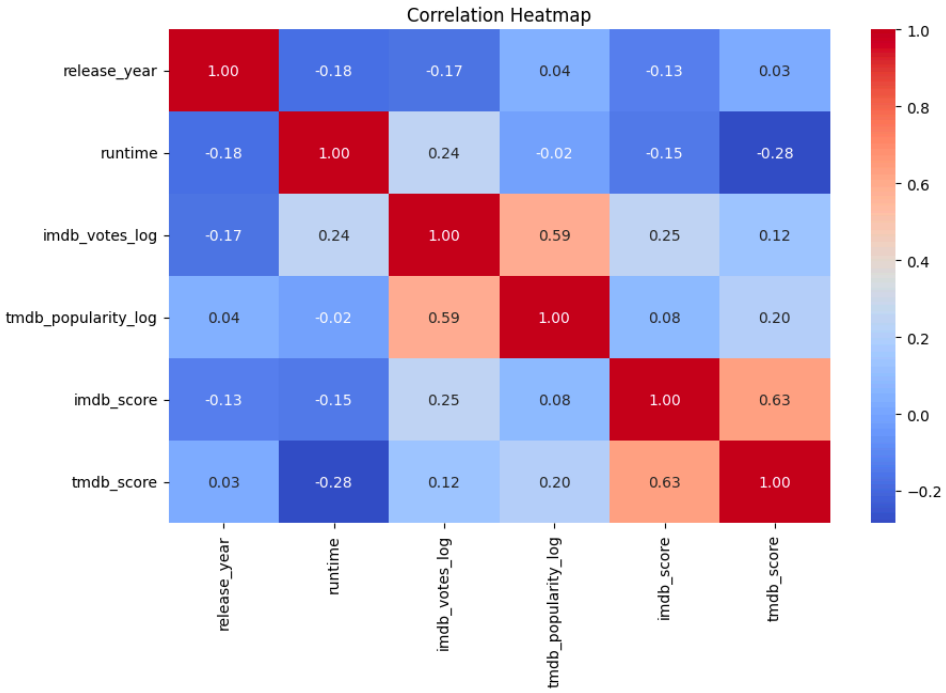
A linear regression model is employed to predict the imdb score of a movie or TV show.

1) Correlation Analysis and Feature Selection

A correlation analysis was conducted first to find which features are highly correlated to the label.




*Picture 3. Pairplot of Numerical Fields*




*Picture 4. Correlation Heatmap*

As can be seen from Picture 3 and 4, tmdb score and imdb score are highly correlated with 0.63 in the heatmap and scatter plot between the two also shows a linear tendency. This means that there is a potential linear relationship between the two. Followed by imdb votes with correlation 0.25, imdb votes may be a good feature to fit the model.

For the other features, they do not seem to have much impact on the label. This may indicate that there is no significant relationship between these features and the ratings of film and television productions, or that these relationships are difficult to capture using current data and correlation analysis.

Therefore, a linear regression model was trained using the processed features, including imdb votes and tmdb_score, for the prediction of imdb score.

2) Performance Metrics and Evaluation Method

- The performance metrics used to appraise the linear regression model are R-squared and RMSE. Higher R-squared and lower RMSE mean that the model fits the data better.
- Additionally, 5-fold cross validation is also applied to see the generalizability of the model.

Both are presented and further explained in Section 5 Results.

### 3.3.2 Decision Tree

A decision tree model was trained using processed features to predict whether a movie or TV show is popular or not. Tmdb popularity is categories as low and high as per its values and used as the label for the model.

1) Feature Selection

One-hot encoding is performed on the categories of genres and production countries which top 10 countries are used as selected features due to the sparsity of production countries.

Encoded type, scaled imdb votes and runtime are also selected as relevant features.

2) Performance Metrics and Evaluation Method

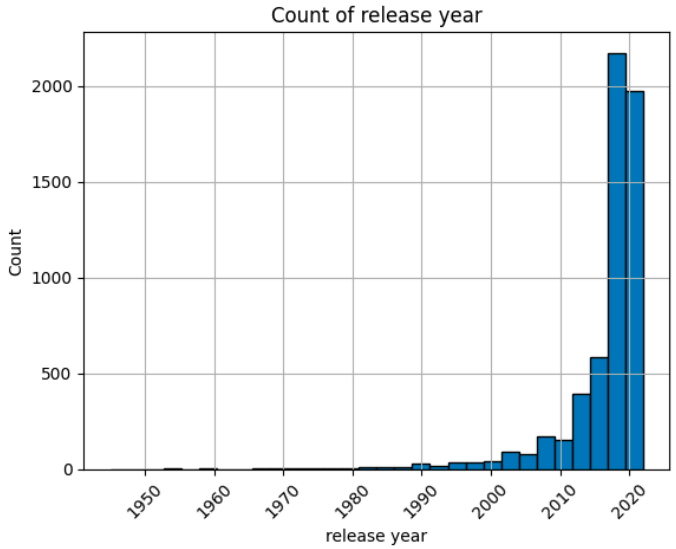
- Accuracy and f1 scores are mainly used to appraise the prediction of our model.
- Confusion matrix also gives more insights about the classification results.
- Performance of 0R model is used as threshold as this model makes predictions solely based on the most frequent label in the dataset without considering any features.
- Moreover, 5-fold cross validation is applied to see the generalizability of the model and whether the model overfits the data.

All the results are analysed in detail in Section 5 Results.

# 4. Data Exploration and Analysis

Some interesting insights in release year, movies and TV shows, genres and production countries are found through exploring the data.

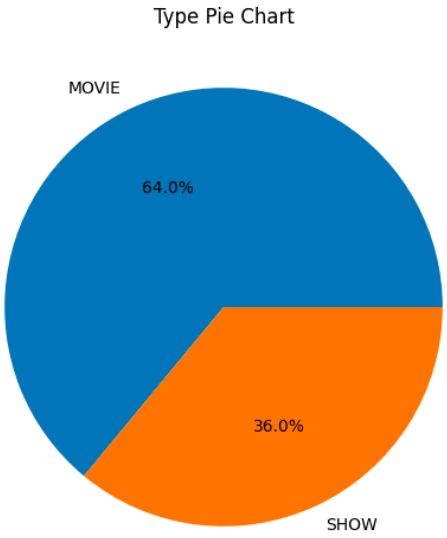
## 4.1 Release year




*Picture 5. Count of Release Year*

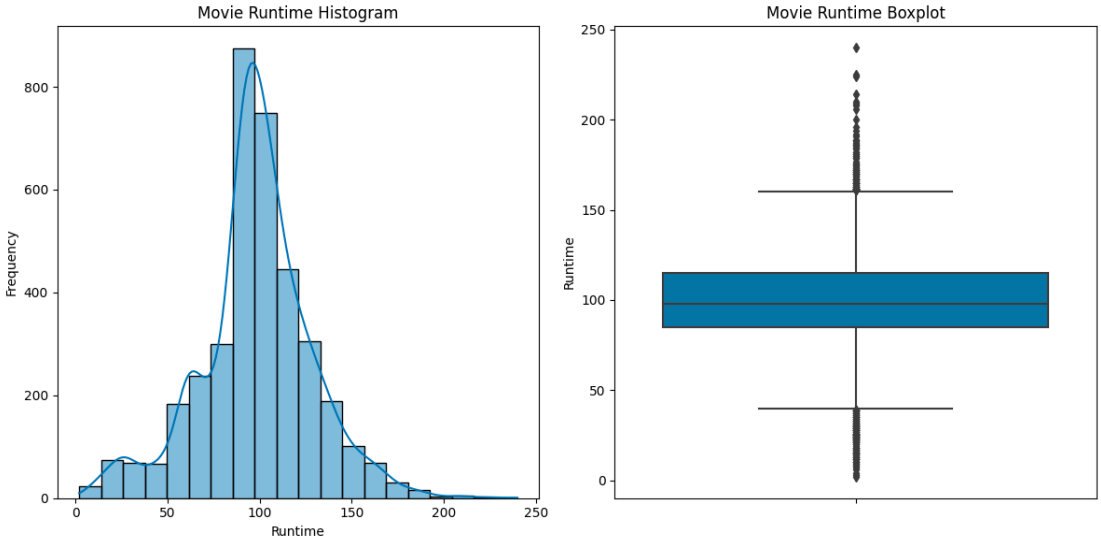
From Picture 5, we observe that the release years of film and television works are mainly concentrated in 2016 and later. To be more specific, approximately 96% of the production were released during the period from 2000 and 2022, indicating Netflix's films and television library are relatively new and keep updating.

## 4.2 Movies vs TV shows

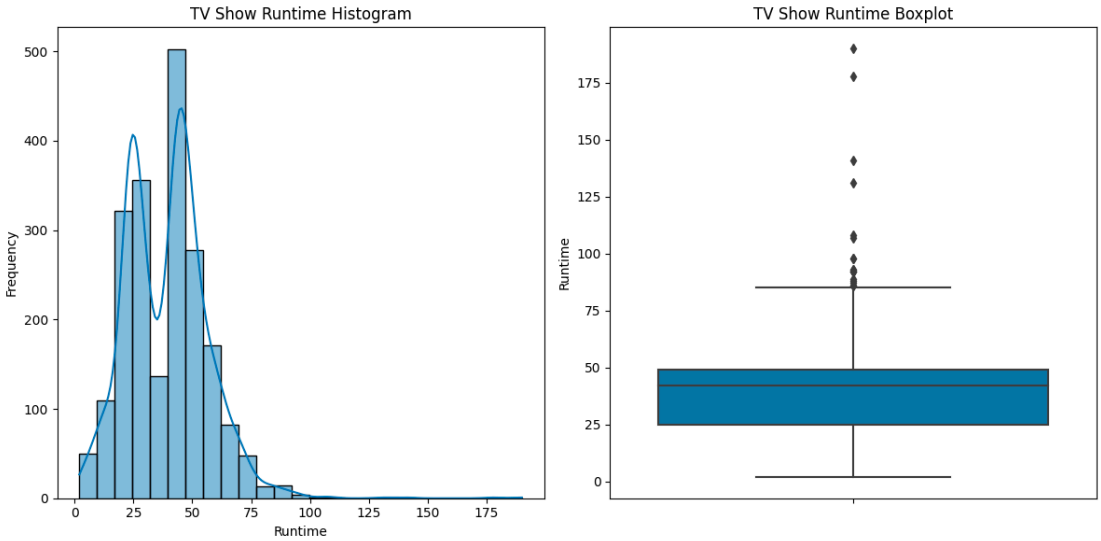



*Picture 6. Movie vs TV Show*

Above pie chart shows most of the data is movie, accounting for 64% (3744 out of 5850), while TV shows accounts for 36% (2106 out of 5850). It is also discovered that runtime differentiates between movie and TV shows.



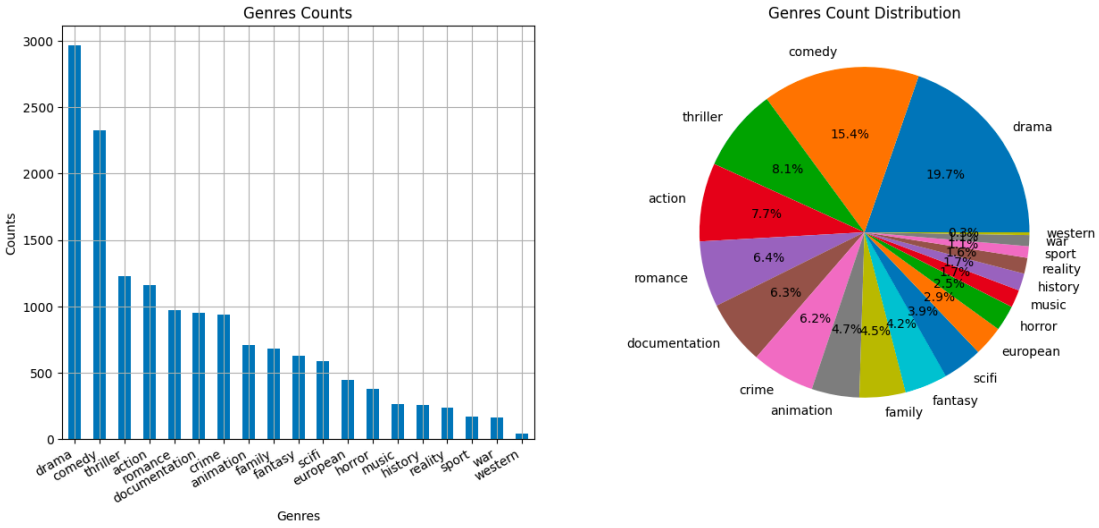

*Picture 7. Movie Runtime*




*Picture 8. TV Show Runtime*

The distributions for movies and tv shows are quite different as is widely known that runtime for TV shows is usually shorter than movies. From above boxplots, we found that the median run time for movies around 100 minutes, and for show is around 40 minutes. And TV shows runtime focus on around 25 minutes mostly for comedy and 45 minutes mostly for drama.

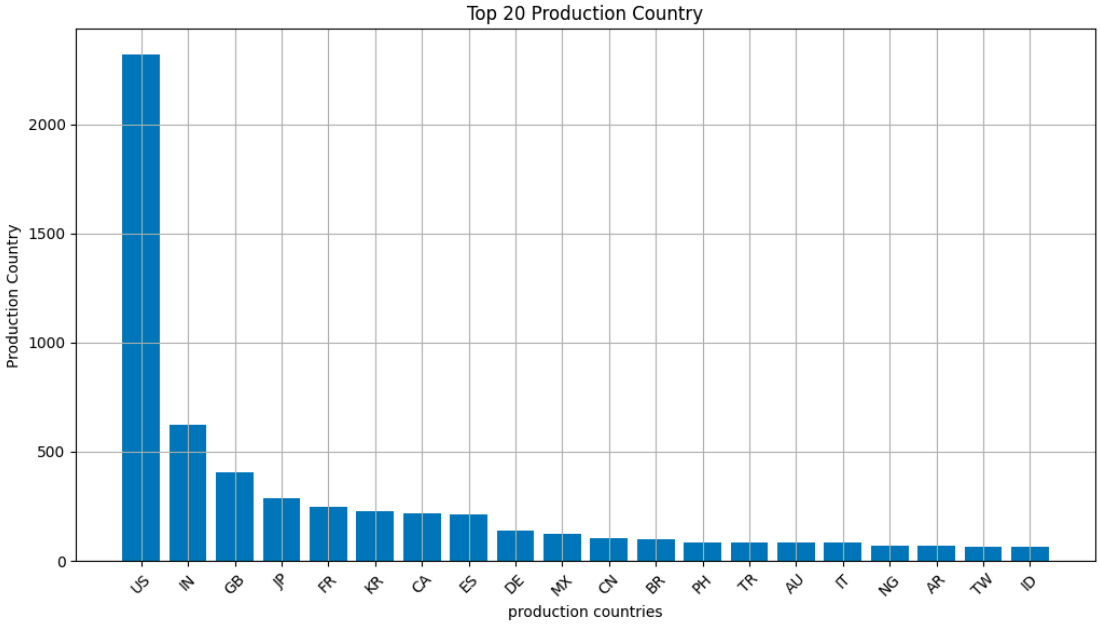
## 4.3 Genres




*Picture 9. Genres*

There are 19 different genres in the data set. And about 3000 visual productions are drama, accounting for 19.7%, following by comedy which accounts for 15.4%.

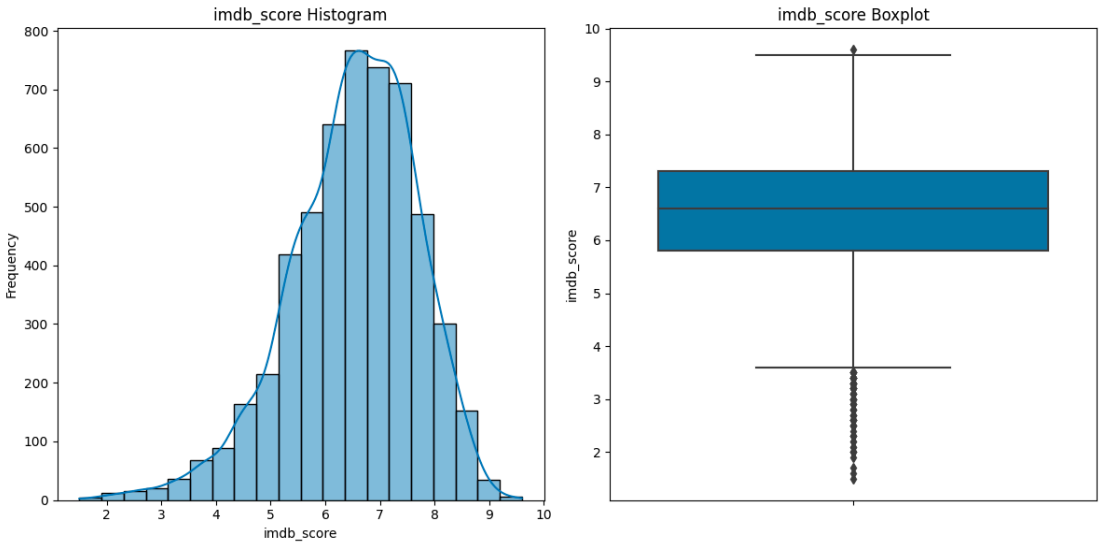
## 4.4 Production Countries



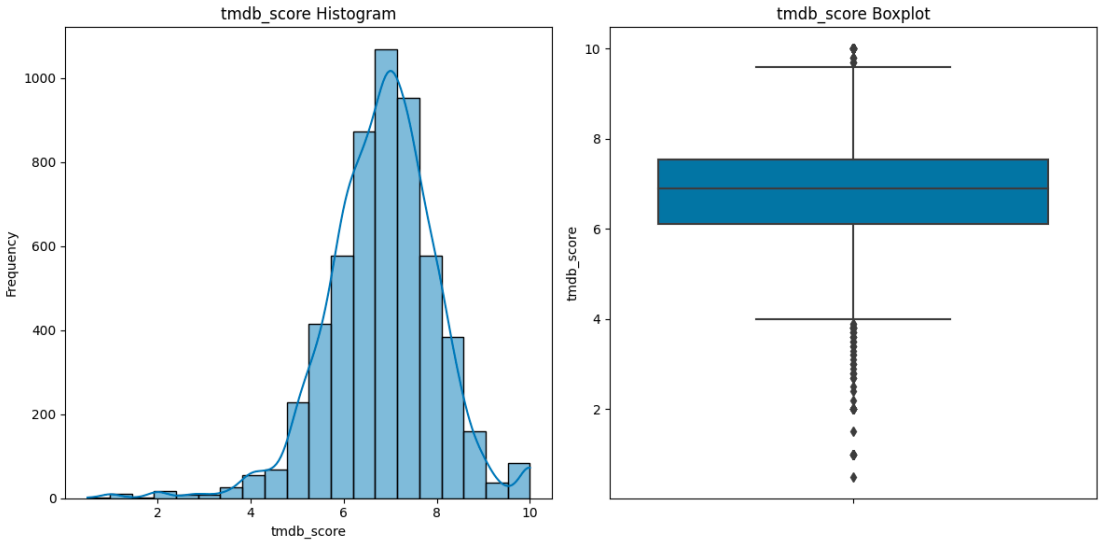

*Picture10. Top 20 Production Countries*

The dominance of films produced in the United States (US) is clearly observed in the graph, with its proportion in the overall dataset significantly higher than that of other countries or regions.

## 4.5 imdb score vs tmdb score




*Picture11. Distribution of imdb score*




*Picture12. Distribution of tmdb score*

These two graphs are used to describe the imdb_score and tmdb_score. From the histogram, we observe that both the scores exhibits a little left-skewed distribution characteristic, ranging from 6 to 8. The box plots tells us some of the points are outliers which we have removed during data pre-processing process.

# 5. Results

## 5.1 Linear Regression

### 5.1.1 Coefficients

According to the correlation analysis, only tmdb score and imdb votes are used for linear regression on imdb score. As for linear regression, the greater the coefficient the bigger influence on the label. We found that tmdb_score affects the imdb_score most with the highest coefficient 0.67 while the coefficient for imdb votes is only 0.09.

### 5.1.2 R-squared and RMSE

Our R-squared scores are relatively low, with an average of 0.39 from 5-fold cross validation, indicating that we are unable to explain a substantial portion of the variance in the label. Furthermore, our RMSE scores are relatively high, with an average of 0.88 from 5-fold cross validation, signifying that the predictive accuracy of our model is limited. Both indicate that linear regression may not be a viable choice used for the given data set.

## 5.2 Decision Tree

### 5.2.1 Feature Importance

As per the output of feature importance for the model, it is shown that imdb votes, run time, Indian production, drama genre and animation are the top 5 significant features on whether the movie or show is popular. Especially with imdb votes, the importance is as high as 0.47, following by runtime with 0.15.

### 5.2.2 Performance Metrics

```
Decision Tree Classification Report:
              precision    recall  f1-score   support

        high       0.72      0.70      0.71       535
         low       0.72      0.74      0.73       550

    accuracy                           0.72      1085
   macro avg       0.72      0.72      0.72      1085
weighted avg       0.72      0.72      0.72      1085

Cross Validation Accuracy:
[0.79447005 0.7437788  0.74261993 0.70848708 0.67712177]
Cross Validation Mean Accuracy: 0.73
```

*Picture13. Performance Metrics for Decision Tree*

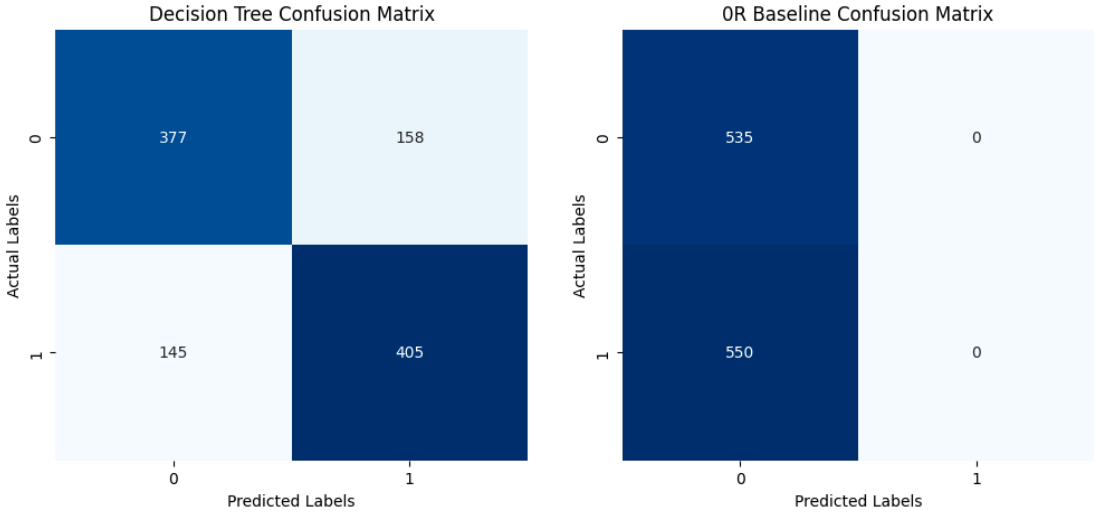
0R Baseline Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| high | 0.49 | 1.00 | 0.66 | 535 |
| low | 0.00 | 0.00 | 0.00 | 550 |
| accuracy | | | 0.49 | 1085 |
| macro avg | 0.25 | 0.50 | 0.33 | 1085 |
| weighted avg | 0.24 | 0.49 | 0.33 | 1085 |

*Picture14. Performance Metrics for 0R Model*

Above two pictures show the performance metrics for decision tree model and 0R model. Accuracy of the Decision Tree model is 0.72 which is higher than 0.49 as threshold from 0R model. Moreover, f1 score of decision tree is 0.71 for high popularity and 0.73 for low popularity, which are also higher than 0R model.

Mean accuracy from 5-fold cross validation is 0.73 which also means our model is not overfitting and generalizes well.




*Picture15. Confusion Matrix for Decision Tree and 0R Model*

From above CONFUSION MATRIX, we can conclude that Decision Tree is more suitable for our Data set than 0Rmodel and our model has successfully predicted whether a visual production is popular.

# 6. Findings and Interpretation

In the linear regression model, the coefficient of imdb votes is only 0.09, which means that popularity does not necessarily guarantee the quality of a movie or show. Film and TV production creation teams understanding the factors that influence scores can help them produce content that matches audience preferences, while Netflix can optimise their content acquisition and recommendation strategies to help film and TV productions achieve higher ratings. So, mapping content to audience preferences may affect audience engagement, platform subscriptions, and word-of-mouth.

Additionally, even though US is the biggest production country, decision tree model shows Indian visual productions are more popular than other countries. India has larger population which implies more audience. Also, most of Indian movies and shows are English speaking which suits most of Netflix viewers and have more dramatic story plots with singing and dancing to attract the audience.

# 7. Limitation and Improvement Opportunities

There are limitations and improvement opportunities to better process the data and model for this research topic.

The data set has selection bias, for example, most of the production countries are US, which may affect the accuracy and reliability of the analysis results. Another issue with data set is that not all the text fields in the given data are English, there are other languages such as Chinese, Turkish and etc. We did limited processing on text fields, such as title and description, which could have been used to explore more insights.

Additionally, the linear regression could have achieved better prediction results if more information is given. When a new movie or TV show is released and there will not be much information about it on the internet, which means our model is not capable of predicting the score for new productions. It is believed that regression model can be improved with better information such as award-winning or explore more advanced regression methods that can more effectively capture the complex relationships and interactions between features.

Finally, by making in-depth fine-tuning of the model parameters, we can further improve the models performance.

# 8. Conclusions

Ensuring proper data preprocessing is essential as inadequate data cleaning can significantly impact model performance. Every dataset has an optimal model that suits it, making the selection of an appropriate model crucial. Linear regression may not be the most suitable choice for predicting IMDb scores. Classification models excel at forecasting the popularity of movies or TV shows.