ANALYSIS OF THE IMDB TOP 1000 MOVIES & TV SHOWS

BY LEE-ANNE VAN DER MERWE

INTRODUCTION

In this report, a deep dive into the insights discovered from the IMDB Top 1000 Movies and TV Shows data (Kaggle, 2025) will be reviewed. The data collected were both cleaned and feature engineered before applying statistical methods and the usage of visualization libraries to uncover the important insights. Some of the questions answered include:

- Who is the highest grossing director on average?
- The average IMDB ratings, number of votes and gross of movies and TV shows?
- What are the top ten movie and TV show genres?
- What is the effect of thecertificate rating on IMDB ratings?

TRENDING HIGHLIGHTS

- Number of votes could predict movie and TV show gross
- Most Popular Genre is Dramas
- Director Athony Russo grosses above 550 million dollars on average
- Top Actor among top IMDB rated movies and TV shows is Elijah Wood

Data Collection

The data used for the insights into the IMDB Top 1000 Movies and TV Shows were imported from Kaggle using the code in figure 1. The Kagglehub library was utilized throughout the data collection, preparation and visualization phases.

Figure 1: Python Code for Data Collection

```
import kagglehub
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
### Importing data into dataframe from Kaggle
path= r"c:\Users\Dell\.cache\kagglehub\datasets\harshitshankhdhar\imdb-dataset-of-top-1000-movies-and-tv-shows\versions\1\imdb_top_1000.csv"
print( "\n Path to dataset files:" , path)

df= pd.read_csv(path)
```

Data Preparation

Next, the data collected were then cleaned by deleting any null values and checking for duplicate values which were zero, to prevent skewed statistical results in the next phase. The code used for data preparation can be seen in figure 2.

Figure 2: Python Code for Data Cleaning

```
### Data Cleaning ###
#
# Percentage of missing values and removal
total_rows= len(df)
missing=df.isnull().sum().sum()
missing_percentage= (missing/total_rows)*100
print("\nPercentage of Missing values:", missing_percentage.round(2), "%")
#
# Row with missing data
print(df.isnull().sum())
dfA=df.dropna().copy(deep=True)
#
# Check and remove duplicate rows
print("\nDuplicates:", dfA.duplicated().sum())
# No duplicates found to remove
```

Feature Engineering

Feature engineering was then utilized to convert runtime from an object datatype to a numerical type, extract the decades from the release years and to create a new column called, “Lead_Actors”. The code utilized to perform the feature engineering can be seen in figure 3.

Figure 3: Python Code Utilized to Perform Feature Engineering

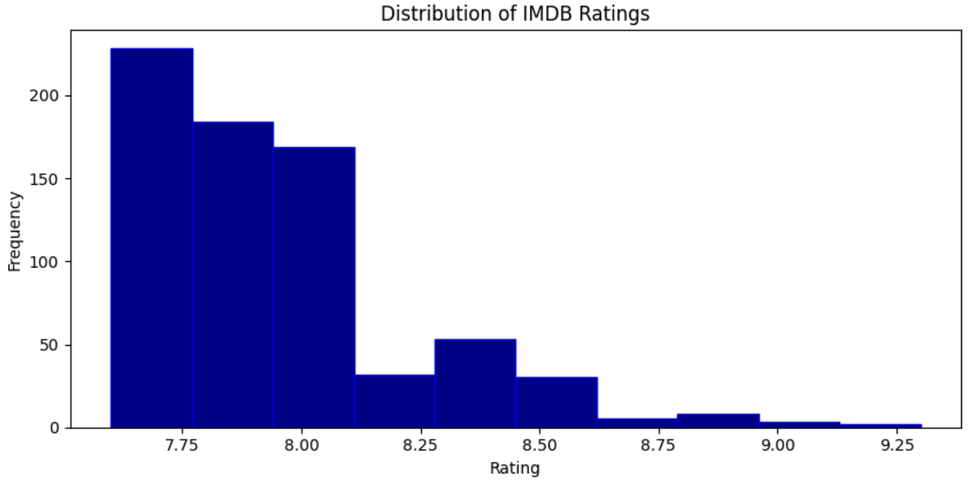
```
### Feature Engineering ###
#
#Converting Runtime dtype:object to dtype:int
#
dfA.loc[:,'Runtime'] = dfA.loc[:,'Runtime'].str.extract(r'(\d+)').astype(int)
#
#Extracting the Decade from the release year and making a new column named "Decade"
pd.set_option('future.no_silent_downcasting', True)
dfA.loc[:,'Released_Year'] = pd.to_numeric(dfA.loc[:,'Released_Year'], errors='coerce')
dfA.loc[:,"Released_Year"]= dfA.loc[:,"Released_Year"].fillna(0)
dfA.loc[:,"Released_Year"]= dfA.loc[:,"Released_Year"].infer_objects(copy=False).astype(int)
dfA.loc[:,'Decade'] = (dfA.loc[:,'Released_Year'] // 10 * 10).astype('Int64').astype(str) + 's'
#
#Combining Star1,2,3 and 4 into a single column-Lead_Actors
dfA.loc[:,"Lead_Actors"]= dfA.loc[:,"Star1"]+(", ")+dfA.loc[:,"Star2"]+(", ")+dfA.loc[:,"Star3"]+(", ")+dfA.loc[:,"Star4"]
#
```

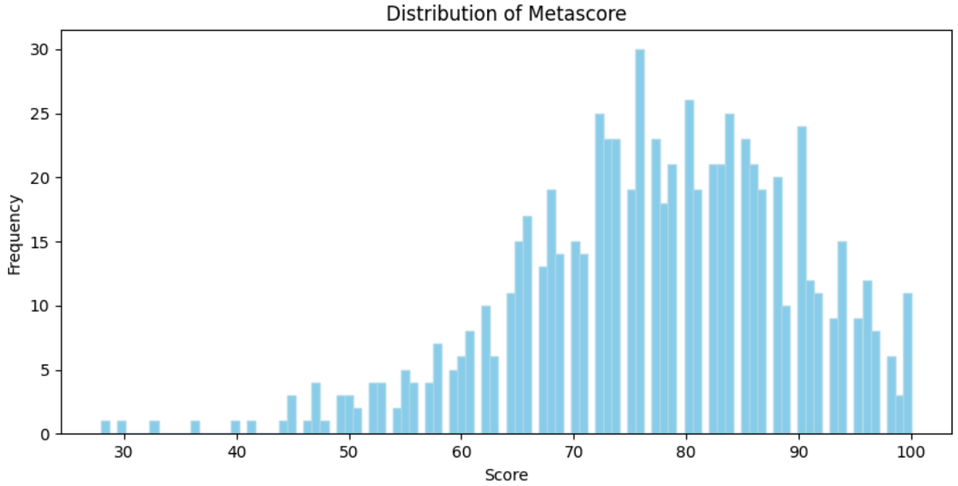
STATISTICS

Statistics			
	Gross	No of Votes	IMDB Rating
Average	78,513,585.24	356,134.82	7.94
Median	34,850,145.50	236,602.50	7.9
Standard Deviation	114,977,950.33	353,901.13	0.29
Outliers	60	N/A	N/A

- **The average gross for movies and TV shows were above 78 million dollars with a standard deviation of just under 115 million dollars. The big discrepancy between the average and standard deviation could be caused by the 60 outliers found in the gross, causing a skewed distribution and high variance among results.**
- **On average movies and TV shows had 356 thousand votes with a standard deviation of 353 thousand votes**
- **The average IMDB rating for a movie or TV show were 7.94 with a standard deviation of 0.29. The low standard deviation indicates low variance between ratings, therefore, implying high accuracy.**

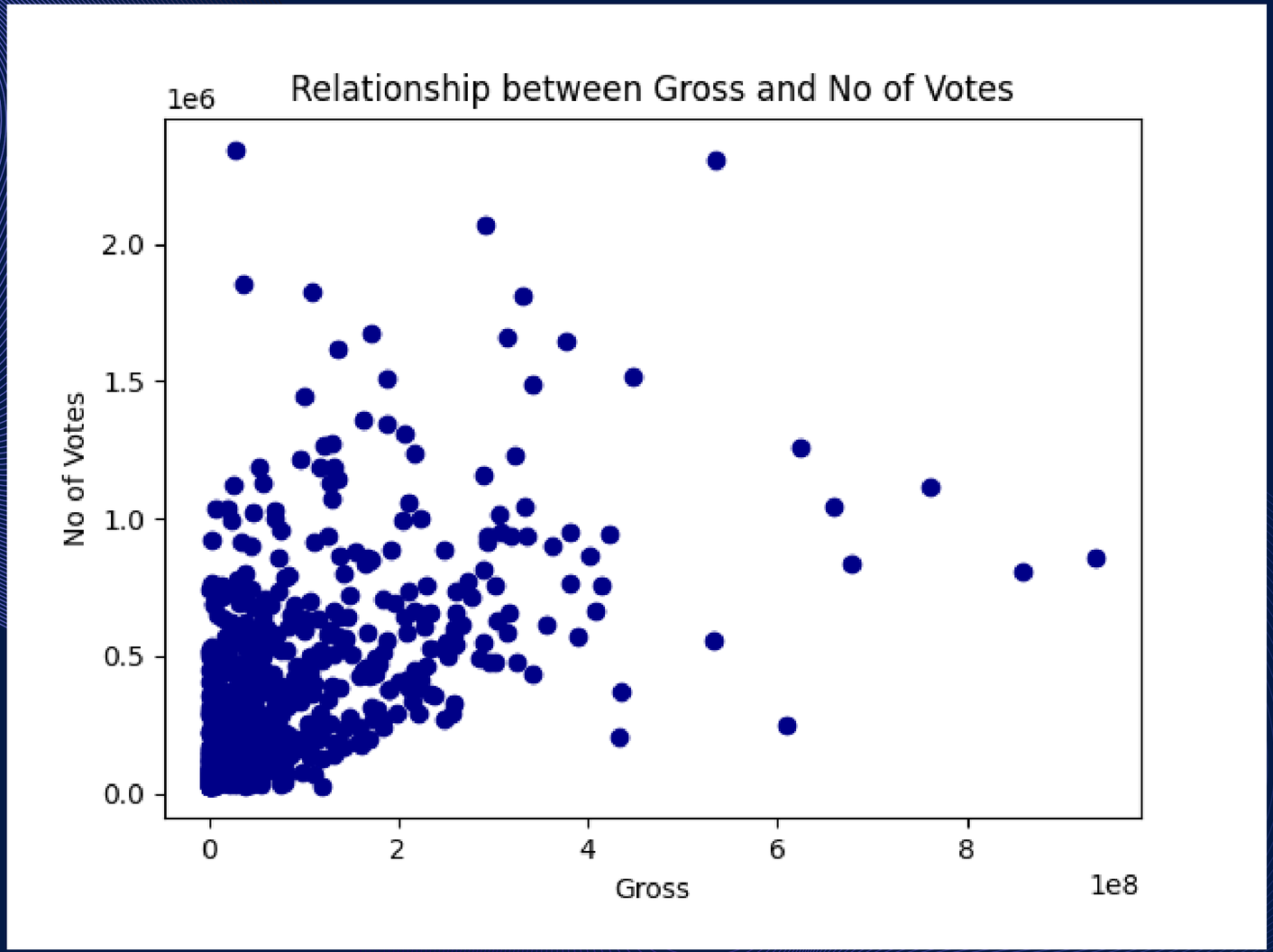
THE DIFFERENCE IN IMDB RATINGS VS METASCORE





The IMDB ratings lean more towards the left (right-skewed), with higher frequencies of ratings of 7.75 to 8 and have an average rating of 7.9. Meanwhile, the Metascores have scores leaning right (left-skewed) towards 70 to 80 , with an average score of 77. When comparing IMDB ratings to Metascores, IMDB ratings are slightly higher and have a lower standard deviation compared to Metascores, therefore, being a better option when using viewer ratings to distinguish the success of a movie or TV show.

HIGHER NUMBER OF VOTES PREDICT HIGHER GROSSING MOVIES



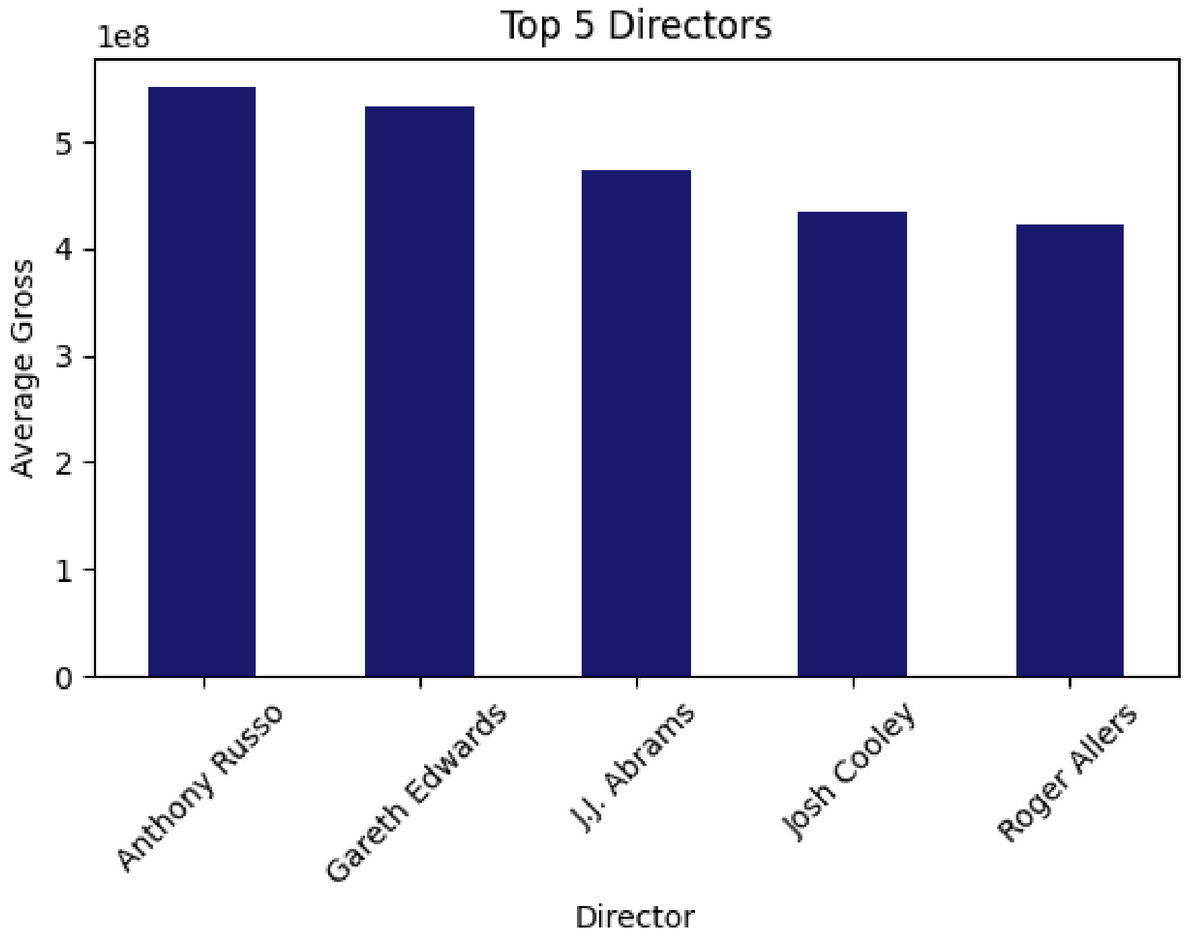
Slightly Strong Correlation (0.545)

- Number of votes could be used to predict gross but can not guarantee it
- Movies with votes in the range of zero to 1 million had a gross up to 4 billion, therefore, movies with higher votes tend to be higher grossing.

TOP 5 DIRECTORS

NO. 1 DIRECTOR

Athony Russo



- Athony Russo was the number one director with an averge gross of over 550 million dollars. This was followed by Gareth Edwards, J.J. Abrams, Josh Cooley and Roger Allers.
- On average, these five directors made films with the highest gross.

THE DIRECTOR CHOSEN MAY INCREASE MOVIE SALES, THEREFORE, INCREASING OVERALL GROSS

Top 5 Directors

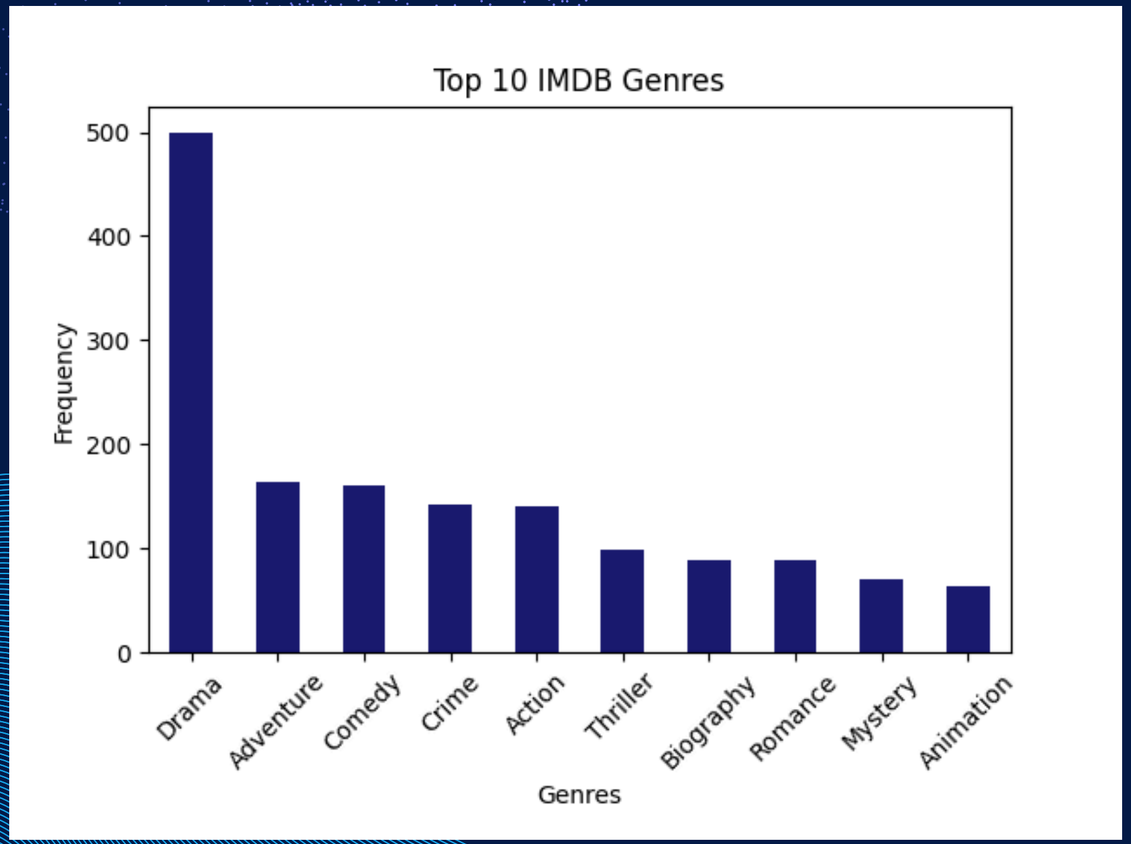
The top five directors were found utilizing the code in figure 4 and from this code, the director with the highest grossing movies on average was found. The first line of code groups the data by directors and then calculates the average gross per director. Next, the top five directors are found by sorting the data in descending order and selecting the first five rows. Lastly, the top five directors and their average gross for their films were plotted utilizing Matplotlib.

Figure 4: Python Code Utilized to find Top 5 Directors and the Director with Average Highest Grossing Movies

```
avg_gross_by_director = dfA.groupby('Director')['Gross'].mean()
top_5_directors = avg_gross_by_director.sort_values(ascending=False).head(5)
print("Top 5 Directors by Highest Average Gross:\n", top_5_directors)

#Plot Top 5 Directors
top_5_directors.plot(kind='bar', color='midnightblue')
plt.xlabel('Director')
plt.ylabel('Average Gross')
plt.xticks(rotation=45)
plt.tight_layout
plt.title('Top 5 Directors')
plt.show()
#
```

TOP 10 GENRES



The type of genre is important to consider when marketing or investing in the entertainment industry. Hence, these were the top 10 IMDB genres according to the frequency at which these genres were found in the top 1000 movies and TV shows.

WHICH ACTOR APPEARED MOST IN THE TOP MOVIES?

ELIJAH WOOD

In the IMDB top 1000 movies and TV shows, Elijah Wood appeared in at least three movies with IMDB ratings higher than 8.5. Hence, he is the actor to have appeared the most in the top rated IMDB movies and shows.

Investment Suggestion:

- **Invest in movies and shows with at least one of the top 5 genres.**
- **Investing in a movie or show with more than one top 10 genre could be financially more successful.**
- **Top actors also increase IMDB ratings**

Top 10 Movies and Tv Show Genres

The top 10 movie and TV show genres were found utilizing the code in figure 5. The cleaned up the genre column into separate genres before ordering and plotting the results on a bar chart. The python library utilized for this was Matplotlib.

Figure 5: Python Code Utilized to find Top 10 Genres

```
#Find out the number of different genres
#
split_genres = dfA["Genre"].dropna().str.split(',')
flattened_genres = split_genres.explode()
cleaned_genres = flattened_genres.str.strip()
genre_counts = cleaned_genres.value_counts()
top_10_genres = genre_counts.head(10)

print("Top 10 Genres:\n", top_10_genres)


# Plot bar chart of top_10_genres
top_10_genres.plot(kind='bar', color='midnightblue')
plt.xlabel('Genres')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.tight_layout
plt.title('Top 10 IMDB Genres')
plt.show()
#
```

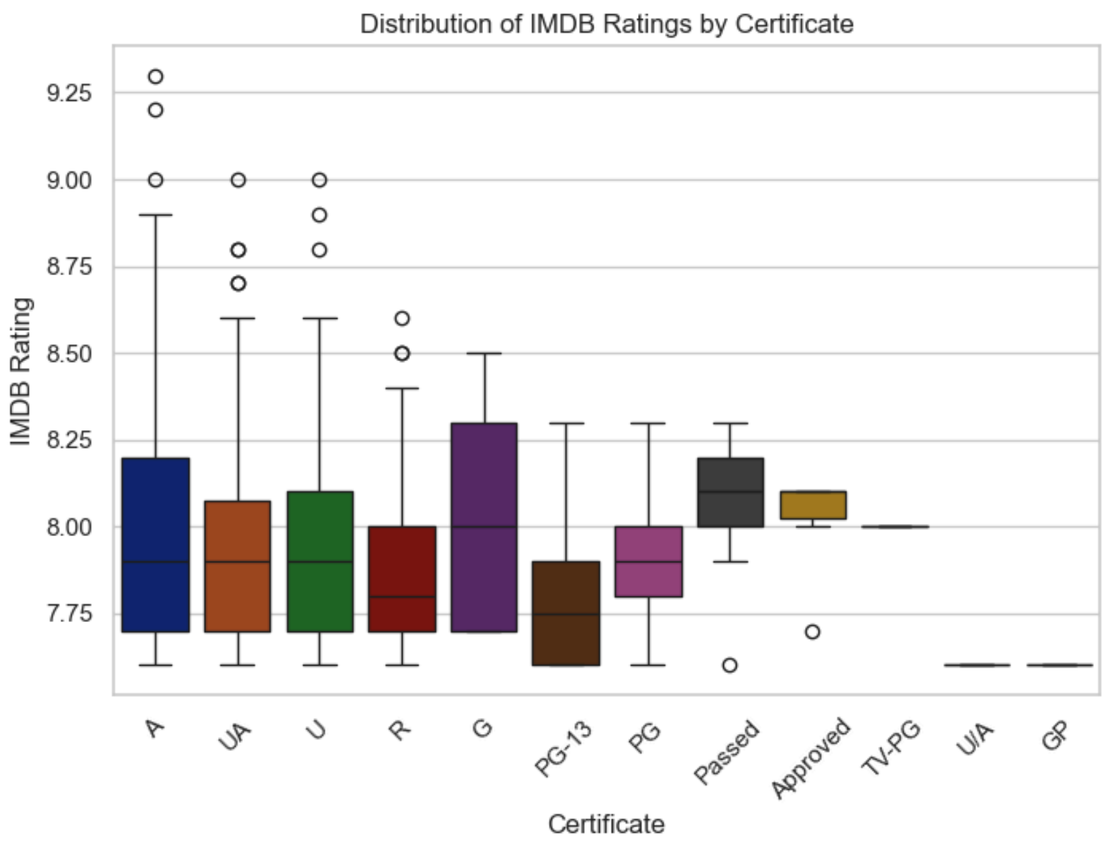
Top Actor

The actor most frequently starring in movies and TV shows IMDB rated above 8.5 was found utilizing the code in figure 6. Firstly, the IMDB ratings were filtered by ratings above 8.5 before searching for the most frequently starring actor among those movies and TV shows. Finally, those actors names were cleaned before counting and ordering in descending order of most frequently starring.

Figure 6: Python Code Utilized to find the Top Actor

```
# Actor which appears most frequently
#
top_rated = dfA[dfA['IMDB_Rating'] > 8.5]
top_rated['Star1'] = top_rated['Star1'].astype(str).str.split(',')
all_actors = top_rated['Star1'].explode().str.strip()
actor_counts = all_actors.value_counts()
most_frequent_actor = actor_counts.head(1)
print("Top Actor:", most_frequent_actor)
#
```

THE IMPACT OF CERTIFICATES ON IMDB RATING?



Certificate: A, UA, U

- Highest IMDB rating of 8.9, 8.6 and 8.6
- Three Outliers
- Median IMDB ratings of 7.9 for all certificates

Certificate: R, G

- Highest IMDB ratings of 8.4, 8.5
- The certificates rated R have two outliers
- Median IMDB ratings of 7.8 and 8

Certificate: PG-13, PG

- Highest IMDB ratings of 8.3 for both
- Median IMDB ratings of 7.75 and 7.9

Certificate: Passed

- Highest IMDB ratings of 8.3
- Median IMDB ratings of 8.15

Highlights:

- **Certificates rated A, AU or U, on average have higher top IMDB ratings**
- **Certificates rated A, UA, U or R have more outliers among IMDB rated films or TV shows, which could pull up IMDB ratings slightly**
- **PG-13 films and TV shows had the lowest median IMDB ratings of 7.75**

Recommendation:

- **Aim for PG certificates instead of PG-13 due to the higher median IMDB rating of 7.9 associated with them**
- **Overall, A rated certificates give a higher chance of top IMDB ratings which is associated with higher gross**

CONCLUSION

Overall, the dataset provided some useful insights for entering the movie and TV show production industry. Movies and TV shows in the drama genre were the most produced followed by adventure. While the highest average grossing director was Athony Russo whom produced on average more than 550 million dollars; the top IMDB rated movies and TV shows frequently had the actor Elijah Wood starring in them. Also, the correlation between number of votes and a movie or TV shows gross were high enough to use as a prediction tool for future. In conclusion, the use of these insights would benefit any company considering to fund the production of a movie or TV show.

APPENDIX

Data Source:

- Kaggle- https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

Code:

- Python 3.13.5

Coding libraries:

- Kagglehub - version 0.3.12
- Pandas - version 2.3
- Seaborn - version 0.13.2
- Matplotlib.pyplot - version 3.10.3