# Capstone Project: Analyzing Movies and TV Shows

Justin Kim, Anya Debelynska, Brandon Lequang, Jimmy Chu, Alexander Okonkwo
CS 4650: Big Data Analysis and Cloud Computing
Professor David Johannsen
December 12, 2023

## 1. Dataset

The dataset we chose is the “HBO and HBO Max Content Dataset” and contains a few thousand movies and TV shows (The, 2023). The dataset includes columns for the title, type (movie or TV), year of release, age rating, IMDB and Rotten scores, genres, and platforms the movie or TV show is available. This dataset is up-to-date up to the year 2020 and goes back nearly a century to 1915.

We also chose to use the “IMDB Movies Dataset” for a brief analysis regarding the IMDB scores, gross profit, and runtime (Shankhdhar, 2021). It contains movies from IMDB up to the year 2020 and includes other columns such as starring actors and number IMDB votes.

## 2. Purpose of Analysis

Our goal was to examine the potential factors influencing IMDb and Rotten scores. We focused our data analysis on the different years, genres, platforms, runtimes, and gross profits to determine their correlation with the scores. By using descriptions and visualizations, this report aims to provide valuable insights into creating a highly-rated movie or TV show.

After some preliminary analysis of the data we decided to answer the following questions:

- What effects does age rating, available platforms, genres, the year of release, runtime, or gross profit have on IMDb or Rotten scores, if any?
- What are the most popular genres by IMDb or Rotten score?
- What are the most popular genres by number of occurrences?
- How did the audience’s sentiment for certain genres change over time?
- How do the IMDb or Rotten scores differ?

## 3. Approach and Methodology

### *3.1. Reading Datasets*

For the HBO and HBO MAX Content Dataset, there were two comma-separated values files for different platforms (HBO and HBO MAX). We read these files into two separate data frames called `hbo_content_df` and `hbo_max_content_df` using the pandas library. For the IMDB Movies Dataset, there was only one file, so that was read into a single data frame.

*3.2. Preliminary Data Analysis*

We first performed preliminary data analysis to describe the statistics of the dataset. Using the `describe` function for the data frames of the dataset, we determined multiple statistics like the mean and standard deviation of each numeric column. We also printed the head of the data frame. By using these techniques, we determined that some columns, like genre and platform, are represented with boolean variables. Other columns, like type and age rating, were represented with string values. Since not all columns were numeric, we would have to convert them to an appropriate value which we will describe later in the report.

Next, we determined what columns had null or invalid values. The following are the number of null values in columns that had null values:

| HBO Content Dataframe | |
|---|---|
| Column Name | Number of Nulls |
| type | 1157 |
| rating | 297 |
| imdb_score | 60 |
| rotten_score | 588 |
| imdb_bucket | 60 |

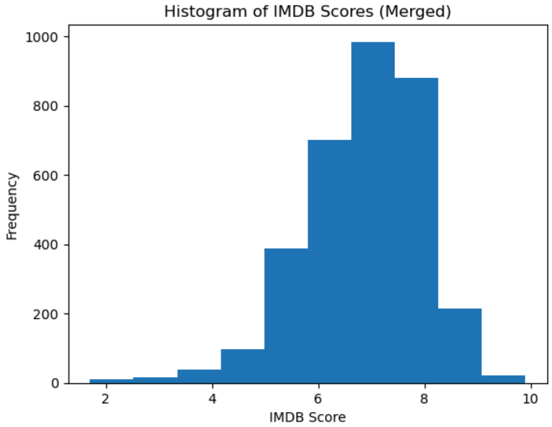
| HBO MAX Content Dataframe | |
|---|---|
| Column Name | Number of Nulls |
| type | 1712 |
| rating | 550 |
| imdb_score | 62 |
| rotten_score | 722 |
| imdb_bucket | 62 |

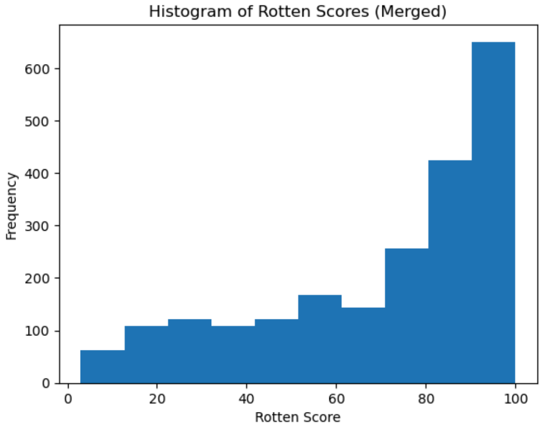
From the head of the data frames and the list of null values, we determined which columns should be dropped, modified, or left alone. For the type, we determine that a null value represents a movie and the string “TV” represents a TV show. We determined that this should be modified to a numeric boolean value. For the rating, we determined that some movies and TV shows did not have an age rating. In this case, we decided to leave it unchanged until the regression steps we performed. For the IMDb and Rotten scores, there was a huge difference in the number of null values. In further analysis, such as regression and data visualizations, we dropped these null values. For the IMDb bucket column, we decided to drop the column entirely since it directly correlates to the IMDb score.

Next, we checked for duplicate values in the dataset. We did not find any duplicate values, so no changes were made in this step.

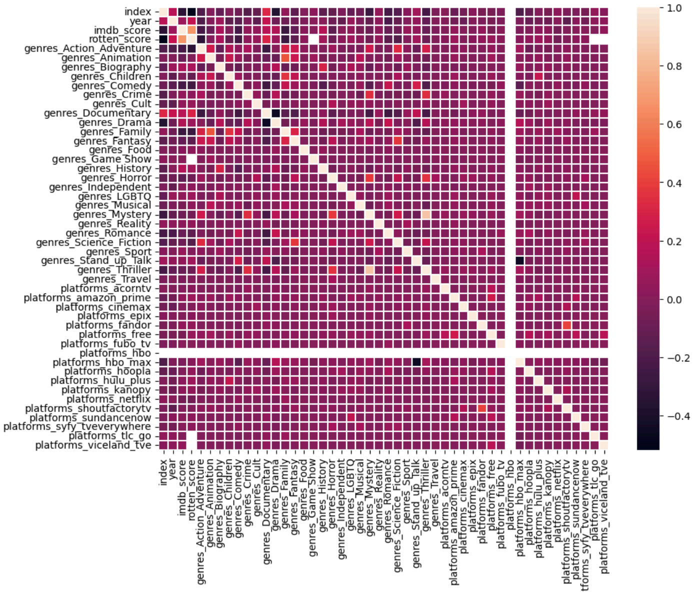
Since we wanted to analyze both HBO and HBO MAX datasets, we decided to merge both datasets based on an outer join of both datasets. After merging these datasets, we determined the difference of columns between the two datasets by comparing the columns. During this step, we determined that most platforms and two genres do not exist in one of the datasets.

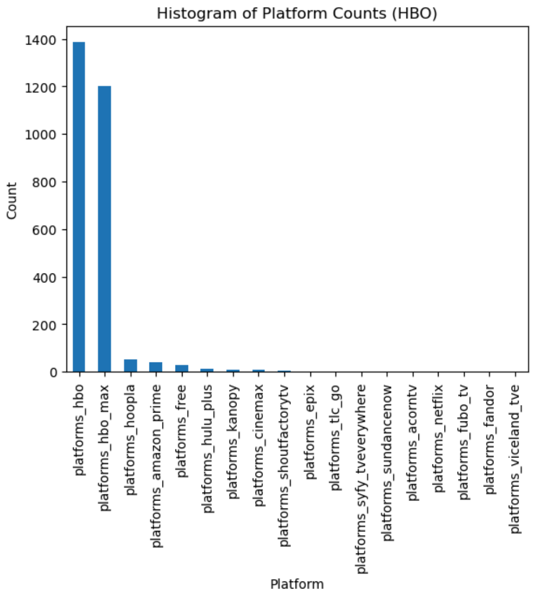
Next, we plotted histograms for the IMDb and Rotten scores. From these plots, we determined that IMDb scores below 5 can be considered as outliers, and that Rotten scores are left skewed. We will use these insights for the data cleaning of the datasets.

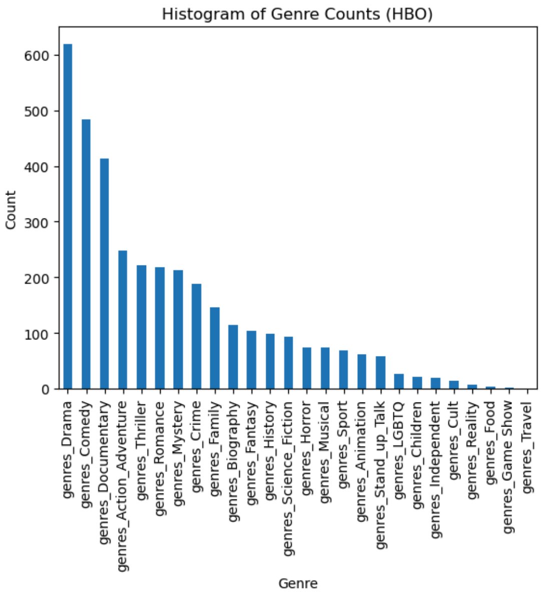




We also create heat maps, platform histograms, and genre histograms. For the HBO content dataset, these were the corresponding heat map, and platform and genre histograms:

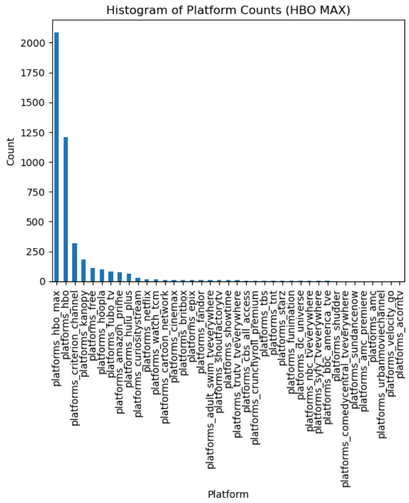






In the context of the HBO content dataset, this dataset is best for analyzing the genre instead of the platform since the platform histogram is heavily skewed compared to the genre. From the above heat map, at a glance, we saw no clear correlation between any columns to the IMDb or Rotten scores. However, we decided to go forward with the analysis to see if we can find a relationship regardless of correlation.

We made heat maps and histograms for the other two datasets. In the context of the HBO MAX content dataset, we determined that it is slightly better for analyzing platforms since it is slightly less skewed than the HBO content dataset. In addition, this dataset had a genre histogram similar to the HBO content dataset, so it may be appropriate to analyze the relationship between scores and genre.



For the merged dataset, we determined that the corresponding histograms were similar to the ones in the HBO MAX content dataset. For these reasons, we determined that it was appropriate to analyze the relationship between genre, platform, and scores with this dataset.

### *3.3. Data Cleaning*

To clean the data sets, we dropped certain columns that were not important to the analysis, and dropped rows with invalid values or outliers. The following steps were done for all three datasets:

1. Drop genre columns with less than *n* occurrences where *n* is 50 for HBO and HBO MAX content and 100 for merged content.
2. Drop certain platform columns (all except `platform_hbo` and `platform_hbo_max` for HBO content, otherwise any platform with less than 100 occurrences).
3. Drop `title`, `decade`, `imdb_bucket`, and `index` columns.
4. After dropping genre and platform columns, drop any rows that do not have a genre or a platform.

During the linear regression steps, we did further data cleaning to make the results more accurate. Since the Rotten scores were somewhat skewed, we decided to flatten this curve by dropping scores greater than 80. In addition, we applied similar transformations to IMDb scores by dropping rows with scores below 5.

### *3.4. Feature Engineering*

For all three datasets, we create boolean variables for age rating. This was done by creating separate columns in each data frame for the unique values in the rating column. This step created the following boolean columns:

- `rating_G/TV-G`
- `rating_PG-13/TV-14`
- `rating_PG/TV-PG`
- `rating_R/TV-MA`

Finally, we transformed the type column to a numeric value. To do this, we map “TV” to the value one and null values to the value 0. 3.5. Data Visualization To visualize the data, we used a combination of histograms, line graphs, and bar graphs. We used these graphs to represent the trend of genres, platforms, IMDb and Rotten scores, and age ratings. Some graphs also involve averages over a five-year period or an exponential moving average every five years.

### *3.6. Linear Regression*

To perform linear regression, we used the `sklearn` library. We aim to analyze the relationship between scores and genre, platform, and age rating. We also performed linear regression for all combinations of analyzing genre, platform, and age rating. For example, we analyzed genre only by dropping the platform and age rating columns. In another example, we analyzed genre and platform by dropping age rating. For the linear regression step, we used a training set size of 80% and a testing set size of 20%. To test the performance of the linear regression models, we evaluated using mean square error and the R^2 value.

To try to improve the models further, we used `StandardScaler` to scale the training and testing datasets. In addition, to better evaluate the performance of the models, the test results will include both the scale and unscaled mean squared error and R^2 value.

### *3.7. Logistic Regression*

To perform logistic regression, we utilized the ‘sklearn’ library. We aim to analyze how features like genre, platform, and age rating influence the IMDB score of various movies and TV shows on HBO and HBO MAX. Our approach to performing logistic regression involves categorizing IMDB scores into two distinct classes. In the case of this method, we use ‘0’ and ‘1’ to classify the scores, which, for logistic regression, suits binary outcomes.

We messed around with different combinations of features to see how they affected IMDB scores. For example, we checked whether genre alone significantly impacted scores, whether features such as the platform that hosts the movie or show, or the age rating. The performance of our logistic regression model was evaluated using various metrics appropriate for our classification needs, such as accuracy score, precision, recall, F1-score and confusion matrix. These metrics allow us to determine our model’s effectiveness in predicting the significance of movies and TV shows to be successful by its IMDB score classification.

## 4. Results and Visualizations

### *4.1. Linear Regression*

For the first linear regression model, the model was trained with a target of the IMDb score and with training columns year, genres, platforms, age rating, and type. After training and testing, it resulted in the following test metrics:

```
Mean Squared Error: 0.573352753353398
R^2 Score: 0.27299727598535584
Min prediction: 5.5046710471713105
Max prediction: 7.961482369559478
Mean Squared Error Scaled: 0.7039673864164372
R^2 Score Scaled: 0.27299727598535584
```

Looking at the results, we can see that the model was not very accurate. Even though the model was able to predict the score with an average error of around 0.8 (average error = MSE ^ 0.5), the range of predictions was limited. This range was from around 6 to 8. When compared to the range of IMDb scores, which is 0 to 10, it seemed like the model was trying to guess the most common values. This would mean that the combination of genres, platforms, and age rating did not significantly contribute to predicting the score.

We created many more models with different combinations of the columns for genres, platforms, and age ratings which resulted in testing six additional models. These models, however, had the same or worse performance when compared to the first model. Due to these reasons, we concluded that no combination of genres, platforms, or age rating would effectively describe the IMDb scores column.

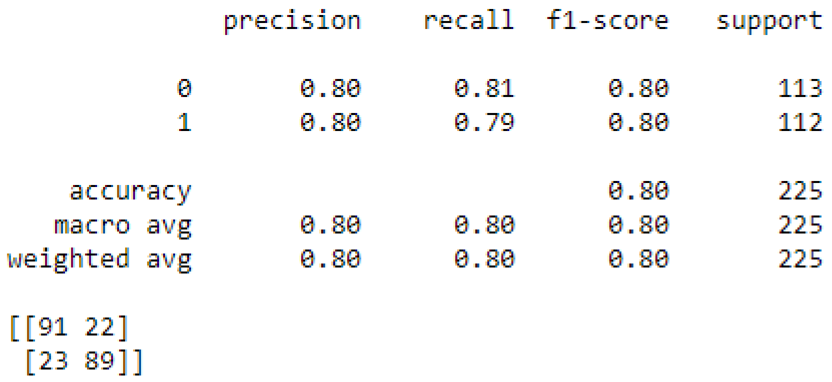
Next, we tried building the same models, but with Rotten scores as the target column. We performed the same steps as above and created a total of seven models. The following is the performance for a model with year, genres, platforms, age rating, and type as the training dataset columns:

```
Mean Squared Error: 377.4730183602868
R^2 Score: 0.13979433794109097
Min prediction: 33.34310805228177
Max prediction: 80.535142859449
Mean Squared Error Scaled: 0.8780077986879442
R^2 Score Scaled: 0.1397943379410912
```

We also built the six additional models which all had the same or worse performance as the above model. From the results of this model, we can see that there is an average error of around 20 points (average error = MSE ^ 0.5). Since the range of Rotten scores is 0 to 100, this error is very significant. We can conclude that neither IMDb or Rotten scores can be predicted accurately using all of the above columns alone.

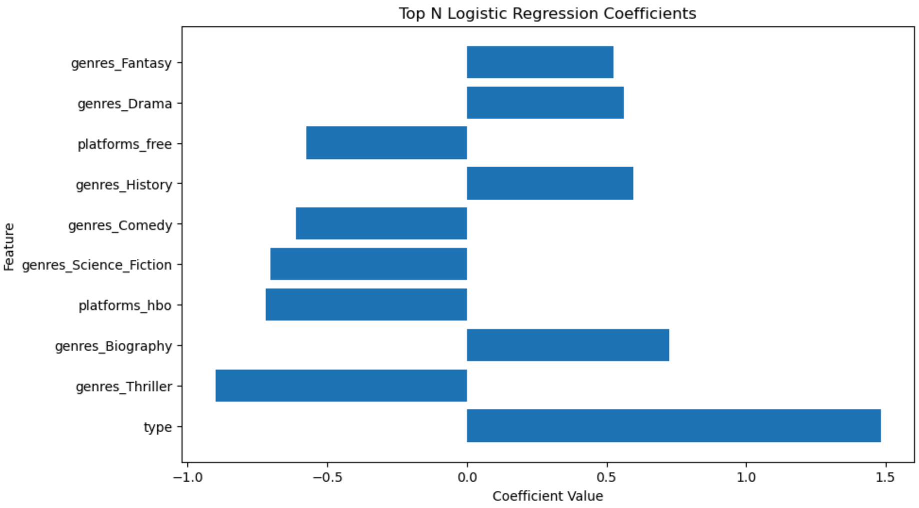
*4.2. Logistic Regression*

For our logistic regression model, we performed a traditional logistic regression analysis, as well as a feature importance plot and a coefficient plot, to further see the significance of features concerning the change in IMDB scores. Touching on our logistic regression analysis model first, we focused on understanding how different features such as genres, platforms, and age ratings affect the classification of IMDB scores into two categories: zero or one. We trained our model using the following features and evaluated its performance in our dataset. The model produced the following values:

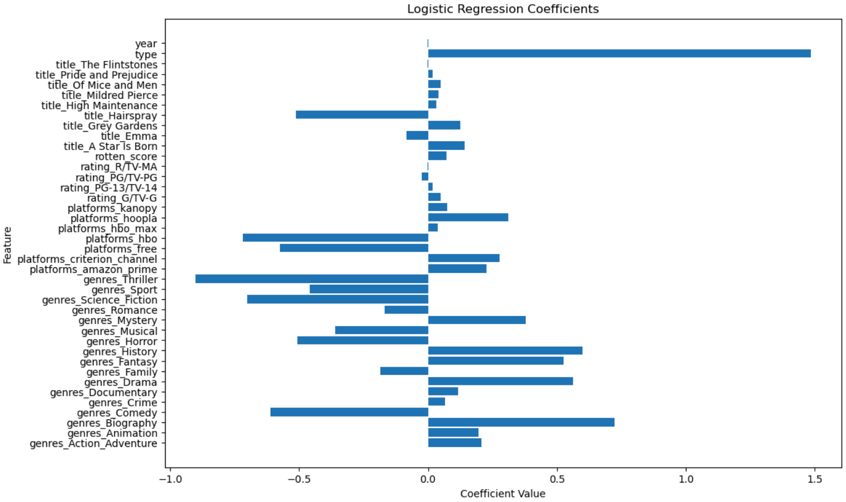


```
              precision    recall  f1-score   support

           0       0.80      0.81      0.80       113
           1       0.80      0.79      0.80       112

    accuracy                           0.80       225
   macro avg       0.80      0.80      0.80       225
weighted avg       0.80      0.80      0.80       225

[[91 22]
 [23 89]]
```

From our results, we observe that the model achieved a balanced performance in terms of our precision, recall, and f1-score across both of our categories. This indicates that our model can effectively distinguish between high and low IMDB scores. Our accuracy of 80% suggests that our features are predictive of IMDB score classification. Our model’s confusion matrix, showing 91 true negatives and 89 true positives, indicates that our model is pretty effective overall in classifying both classes. However, our indication of 22 false positives and 23 false negatives shows that our model can be improved upon to increase classification accuracy. While our results overall suggest that our features do have a notable impact on IMDB score classification, the presence of misclassifications indicates that there may be other factors that do have some level of significance that affect our model. Because of this, we perform a feature importance model and a coefficient model to see if we can reveal any influential factors.



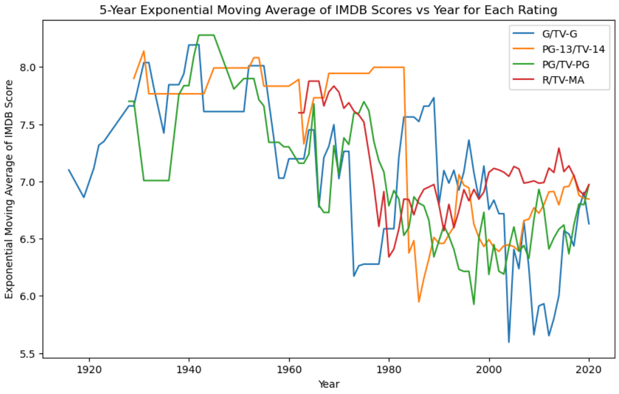
Our feature importance model shows the magnitude and influence certain features may have in coefficient values.  Positive coefficients indicate that a feature will have a higher significance of being hindered from getting a higher IMDB score, while negative coefficients indicate the opposite. Features such as genres_Biography, type, genres_History, generes_Drama, and generes_Fantasy have a high positive coefficient, which indicates that the probability of having a higher IMDB score based on its category is higher. On the other hand, genres_Thriller and generes_Science_Fiction have a negative coefficient, indicating that a movie or TV show categorized with the following features is less likely to have a higher IMDB score. This analysis model lets us look at how categories affect IMDB from a different perspective.

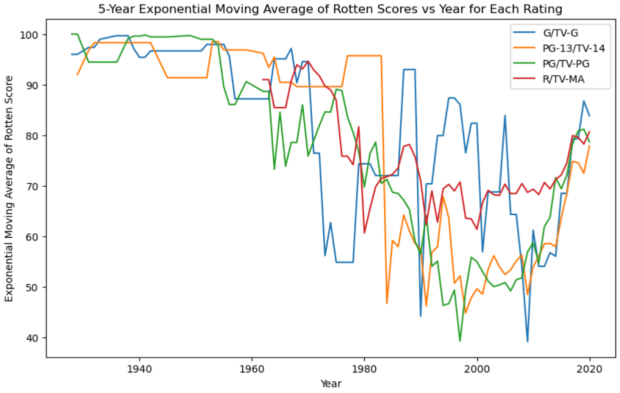


Our coefficient plot shows the relative importance of the various features in predicting IMDB scores. For example, positive coefficients such as ‘generes_Action_Adventure’ and ‘platforms_hbo’ indicate that these features are generally associated with high

IMDB scores. On the other hand, negative coefficients for ‘type’ indicate that the feature is unlikely to be associated with a high IMDB score. This model shows that various features hold a significantly higher value that is likely to be associated with a high score. Overall, the model presents information that is indicative of the idea that certain genres and platforms have an influence on determining IMDB scores.

### *4.3. Audience Age Rating Sentiment*





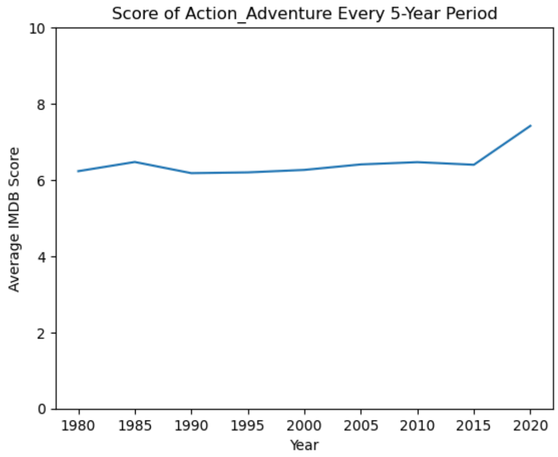
Above are EMA (Exponential Moving Average) plots of IMDB and Rotten Scores vs Year for each age rating. Looking at the plots, one can’t conclude much. The scores for each rating seem to fluctuate roughly with each other, with a rough and slight downwards trend over time across all ratings. Thus, there are no significant conclusions to be drawn from these plots.

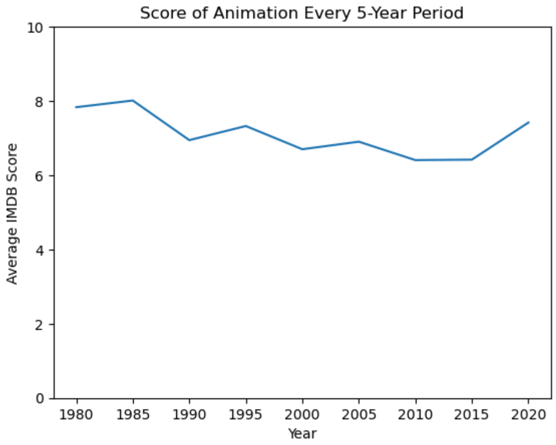
Additionally, some of the data seems to be conflicting. For example, around the year 2020, the R ratings for IMDB vs Rotten appear to be inconsistent. Due to these points, we can conclude that there seems to be no meaningful relationship between age rating and the scores.

*4.4. Audience Genre Sentiment*

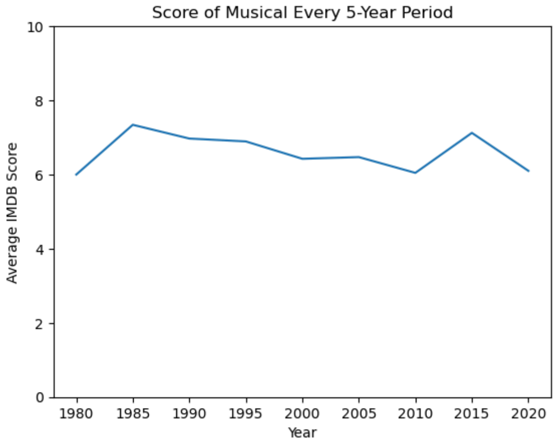
We also created plots for each genre for the average IMDb score versus time. The aim for this part of the analysis is to gauge how audience's preferences have changed in the past few decades (after 1980).

There were a couple genres that had a significant upwards trend in the score in the last two decades. These genres included: Action Adventure and Animation.

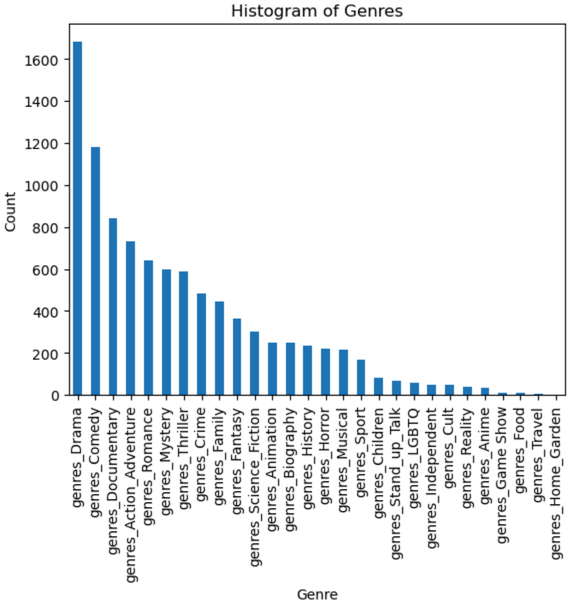




This up trend is significant since it may indicate an audience's interest in these particular genres. One genre, however, had a noticeable downward trend which included: Musicals.

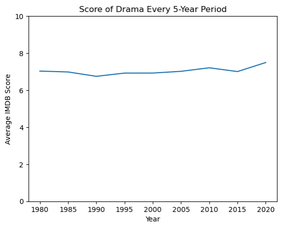


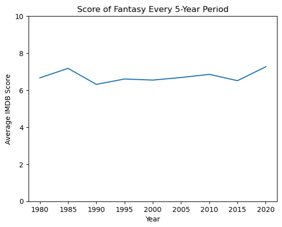
Although this may seem like audiences are disinterested with musicals, this may be due to a skewed dataset in terms of the number of Musical genres. We can see this skew in the following histogram on the number of genres in the dataset:

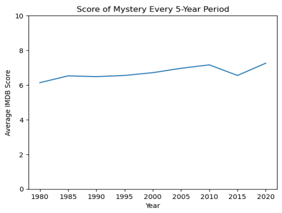


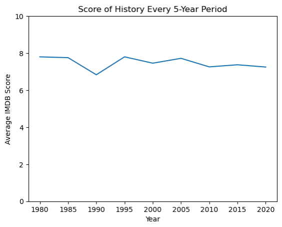
The Musical genre is near the middle of the histogram meaning this genre represents a small fraction of the dataset. Since data for the Musical genre is limited, it may not be possible to accurately determine the audience's sentiment for this genre. In addition, a similar statement could be made for the Animation genre, since it has a similar number of occurrences as the Musical genre.

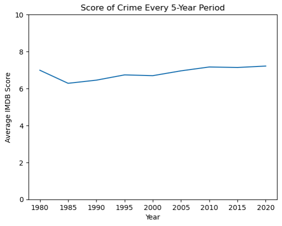
Other than the Action Adventure genre, every other genre seems to have a flat curve or is steadily rising or decreasing. The genres that seem to rise slightly over the past 40 years are Mystery, Crime, Thriller, and Science Fiction. All other genres seem to have a flat or decreasing curve. This may indicate that the audience is interested in Mystery, Crime, Thriller, and Science Fiction genres.

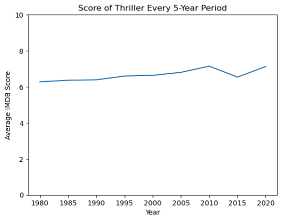


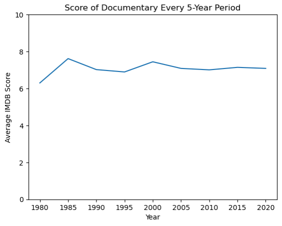


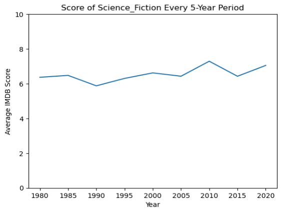


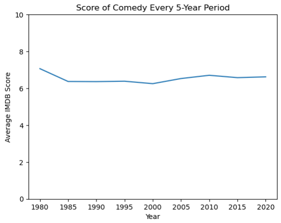


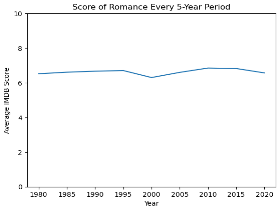


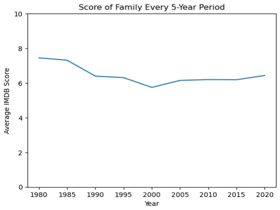


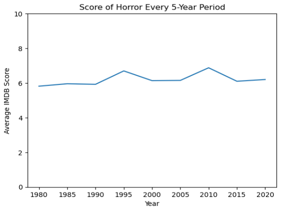


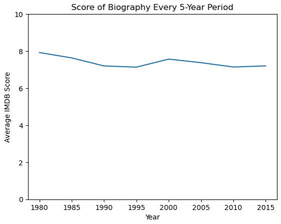


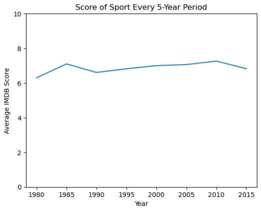






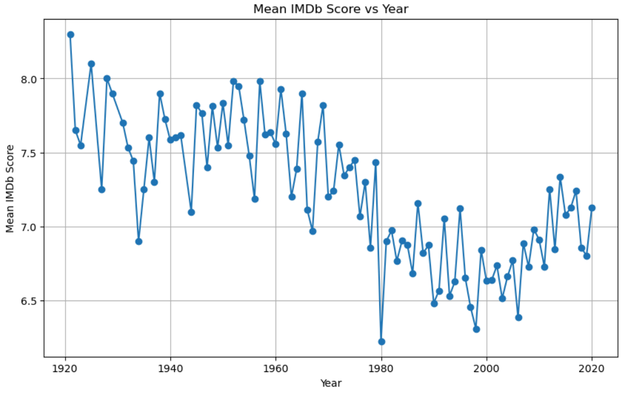




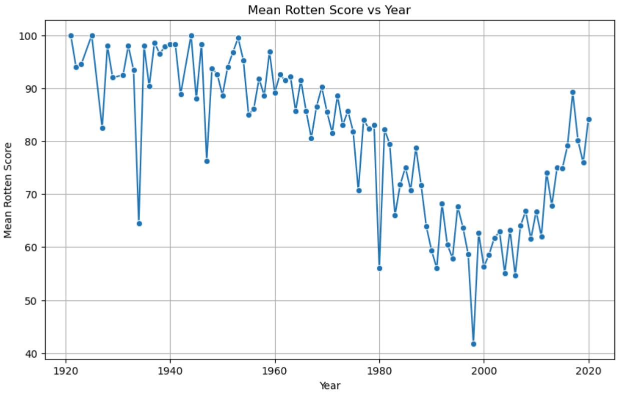


### *4.5. IMDb and Rotten Scores, and Years Comparison*

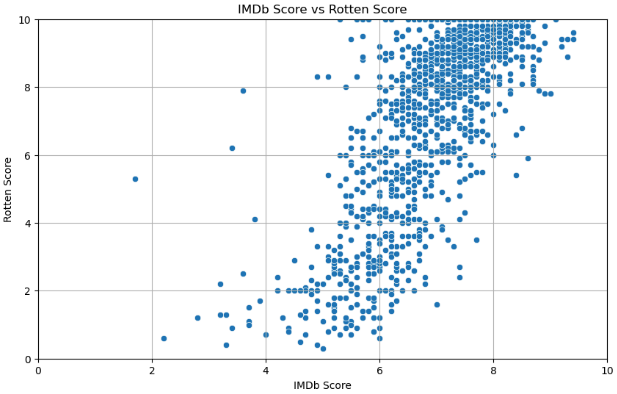
We compared IMDb and Rotten scores in relation to release years to see if there is a correlation between film ratings and the years those movies and TV shows were released. Additionally, IMDb Scores were compared with Rotten scores to see how much they differ.



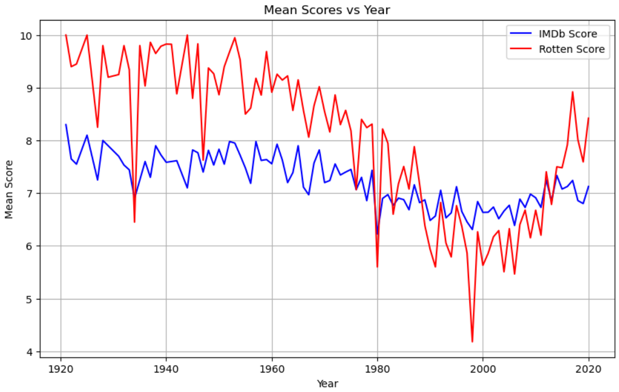
The above graph represents the relation between mean IMDb Score (y-axis) vs Year (x-axis). As we can see, the mean IMDb Score stayed relatively low starting in the 1980s, with the lowest mean rating drop in 1980 at almost 6 out of 10 mean IMDb score for a year.



We did the same examination with Rotten Scores. On the graph above, y-axis represents the mean Rotten Score, and the x-axis represents the Year. We can see that graphs slightly differ from the previous one, and Rotten Scores have had higher scores in recent years. Yet, it started to slowly drop after the 1960s with it rising around the 2000s, after it reached the lowest mean point of almost 40 in 1998.The highest recent point was at a score of about 90 out of 100 in 2017.



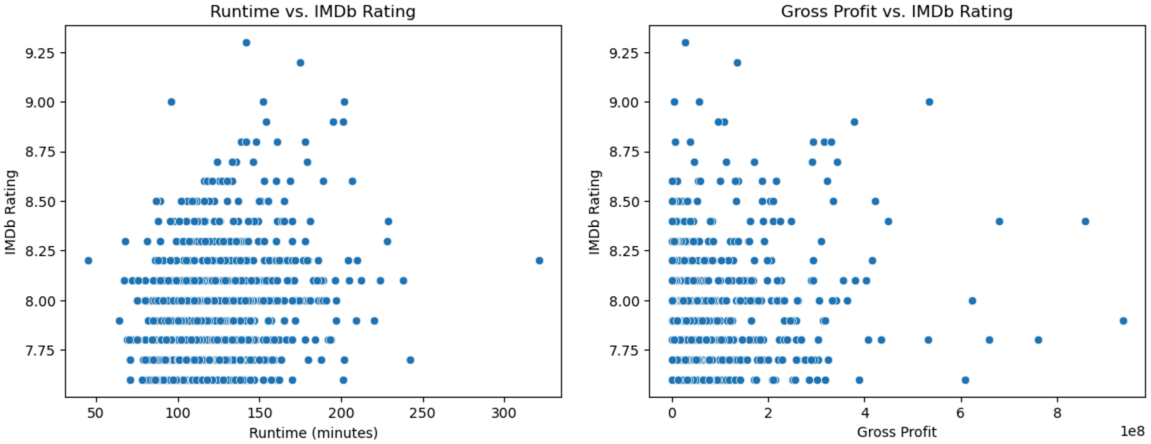
In this graph we compared IMDb Scores (x-axis) and Rotten Scores (y-axis) to see how much they differ. To make this comparison, we had to divide Rotten Scores by 10 as they are on a scale of 100, while IMDb Scores are on a scale of 10. We can see that there are a lot of Rotten scores under the rating of two while only one IMDb score with such a low rating. Yet, this scattered graph of IMDb Score vs Rotten Score shows that Rotten Scores are averaging a higher rating overall than IMDb Scores as there are a lot of points at the top second to the right box, and not as many points pass 8 on the IMDb score axis.



This graph has IMDb Score as a blue line and Rotten Score as a red line, both mean values are on the y-axis, and it's compared to the Year on x-axis. We can observe that the mean Rotten Score is higher compared to the IMDb Score vs Year graph with a slight exception around the year of 2000. But focusing on the overall rating vs release year, this graph shows that the overall rating of movies was higher in the previous years, especially before 1980s with both mean IMDb and Rotten Scores dropping between 1960s and 2000s. After which scores start to slightly increase, but still they don't go above the highest scores before, which were there before the 1960s.

From this analysis we may assume that old-movies had better ratings and respectively were more popular. Possible reason may be because things such as movies and TV shows were new to everyone while later people might have become used to it, and now there is a much bigger variety of movies and shows releasing each day compared to only a few in the 1920s.

*4.6. Additional Analysis*



In our pursuit of discerning trends within our dataset, an augmentation involving the inclusion of data from the Top 1000 Movies by IMDB Rating was implemented. The subsequent examination of the relationship between the runtime of films and their corresponding IMDB ratings failed to unveil any conspicuous trends. Although a concentration of films with the highest IMDB ratings (8.75 or above) was noted within the runtime range of 100 to 200 minutes, the prevalence of movies within this temporal span precludes any definitive assertion that a runtime falling within the 100 to 200-minute range invariably results in higher IMDB ratings.

Subsequently, an analysis juxtaposing Gross Profit against IMDB Rating was conducted to elucidate potential trends. The resultant plot depicted a preponderance of data points in the lower left quadrant, thereby underscoring the absence of a discernible trend implying a direct correlation between higher or lower gross profits and elevated IMDB ratings. However, a noteworthy observation emerged, revealing that movies with the highest profit margins predominantly inhabit the IMDB rating range of 7.75 to 8.5.

In summary, after looking at combined data from different sources, we couldn't prove that there's a direct connection between how long a movie is or how much money it makes and its IMDB rating. This highlights how tricky it is to figure out what exactly influences how people rate movies, showing that there are many factors at play in determining a movie's success.

## 5. Conclusion

### *5.1. What Worked Well And Did Not*

The analysis of the datasets using simple scatter and line plots worked well for our analysis. It allowed us to find any relationships between the genre, age rating, score, and time in years like the trend of audience sentiment for genre and age rating. Another aspect of our analysis that worked well is using histograms to describe the dataset in the preliminary analysis. By using histograms, it allowed us to highlight any outliers or issues with the dataset for the genre and platforms columns. Histograms also allowed us to identify any skewness in the shape of the histogram. Some examples are histograms for the scores, and that the Rotten scores histogram was skewed to the left. Another aspect of the analysis that worked well was using an exponential moving average (EMA) for the score versus time per age rating line graph. Without EMA, it would be difficult to interpret the graph since, initially, it was extremely jagged and unstable.

However, there were some aspects that did not work well for this project. One aspect that did not work well is using linear regression for the analysis. We initially thought linear regression would be able to accurately predict the scores, but that was not the case. It was extremely inaccurate and would not work well. After this analysis, we think linear regression did not work because there may not be a linear relationship in this dataset. Another aspect that did not work well for us is using the original dataset. The first dataset we analyzed was very limited on the types of data that can be analyzed. Therefore, we decided we had to find another dataset with more columns like gross profit and runtime.

### *5.2. Summary of Results*

By using linear regression and data visualizations, we can conclude that no combination of genre, platform, platform, gross profit, or age rating can predict either IMDb or Rotten score. However, we gave some insight into the audience's sentiment for certain age ratings and genres. There was increasing positive sentiment for the Mystery, Crime, Thriller, and Science Fiction genres, and especially for the Action Adventure genre which increased in popularity in the last decade. We also propose that since scores were higher in the past, it may imply that audiences were more interested due to the novel nature of movies and TV shows during that era.  In addition, we had some success with predicting high and low scores using logistic regression. This may indicate that linear regression is not suited for this dataset, but logistic regression can provide insights into whether a movie or TV show will fail or not.

# References

The Devastator. (2023). *HBO and HBO Max Content Dataset*. Kaggle. https://www.kaggle.com/datasets/thedevastator/hbo-and-hbo-max-content-dataset

Shankhdhar, Harshit. (2023). *IMDB Movies Dataset*. Kaggle. https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows