Project: Final Report
What are the factors affecting IMBD ratings?
Group 5

1. Statement of the research problem or questions being addressed

Movies are considered a global cultural phenomenon due to advancements in technology, the accessibility of streaming platforms, and the globalization of the film industry. Also, movies are portrayed as powerful platforms for creative expression, social criticism, and narratives that transcend geographic boundaries. Furthermore, the film industry's impact extends beyond mere entertainment, with significant contributions to the global economy, evident by statistics from the Motion Picture Association indicating substantial employment, compensation, and support for numerous enterprises.

To enhance the overall success of the film industry, it is important to identify factors that influence movie ratings because they are considered crucial indicators of a film's quality, impacting its success or failure in the highly competitive and dynamic industry. These ratings shape audience perceptions, drive box office performance, and influence critical acclaim. Understanding these factors becomes essential for industry professionals to make informed decisions and optimize production, marketing, and distribution strategies. The research paper aims to delve into three key areas that potentially influence movie ratings: movie duration, film critics and reviews, and genre popularity.

This study is meant to answer the question “What are the factors affecting IMBD ratings?” The fundamental research questions which are of this report can be sorted as follows:

1. How does movie duration impact movie ratings? The research will explore whether longer or shorter movie durations tend to receive higher ratings and how temporal aspects affect audience perceptions and preferences.
2. To what extent do film critics and reviews influence movie ratings? The study will examine the influence of film critics on movie ratings and whether certain words have a greater impact on movie ratings.
3. Which movie genres tend to receive higher ratings from audiences? The research will identify and analyze the movie genres that typically receive higher ratings, including the number of genres, combination of genres, etc.

2. Briefly discuss the data and justify its suitability

The dataset used for this research is "IMDB Top 250 Movies." It contains details of the top 250 movies as per their IMDB ratings. This dataset comprises various features related to movie titles, imbd_votes, imbd_ ratings, duration, genre, cast, director, writer, and user reviews.

This dataset is suitable for our research questions because of the following reasons. First of all, this dataset is perfectly relevant to our research questions as the dataset contains information on movie ratings, duration, and genre, which directly align with the research questions. It allows for a comprehensive exploration of the factors influencing movie ratings. Besides, this dataset contains a large number of sample sizes, which includes information on the top 250 movies, providing a substantial sample size for analysis. This large sample size enhances the statistical significance and generalizability of the findings. What's more, this dataset includes a diverse variety of features such as cast, director, and writer information, which could potentially be used to control for confounding factors and gain deeper insights into the research question though we didn't have the chance to consider these factors into our questions. Also, the dataset is based on real-world movie ratings and reviews from the IMDB platform, making the findings applicable and relevant to the film industry and its stakeholders. The data is reliable as it comes directly from a reputable and widely recognized source like IMDB, a well-established movie database known for its comprehensive movie information and ratings, adding to the dataset's reliability. We also did data preparation to appropriate data manipulation and preparation techniques, including handling missing values and sentiment analysis on user reviews, ensuring its reliability and usability for the research. Considering these factors, the "IMDB Top 250 Movies" dataset is well-suited for addressing the research questions related to the relationship between movie ratings and film duration, critic reviews, and genre popularity, which can provide valuable insights for industry professionals and researchers in the film industry.

3. **"Influence of movie duration on movie rating"**

After careful consideration of the research questions and feedback received on the project proposal, our team decided to conduct an in-depth analysis of the relationship between movie duration and its ratings. We were convinced that exploring this specific correlation, rather than spreading our resources thin over a broad range of topics, would yield a more concrete understanding and actionable insights.

Linear regression and correlation analysis were chosen as the primary analytical techniques due to their ability to measure the strength and direction of the relationship between two variables effectively. In this context, we wanted to answer the research question: "Is there a significant correlation between a movie's duration and its IMDb rating?"

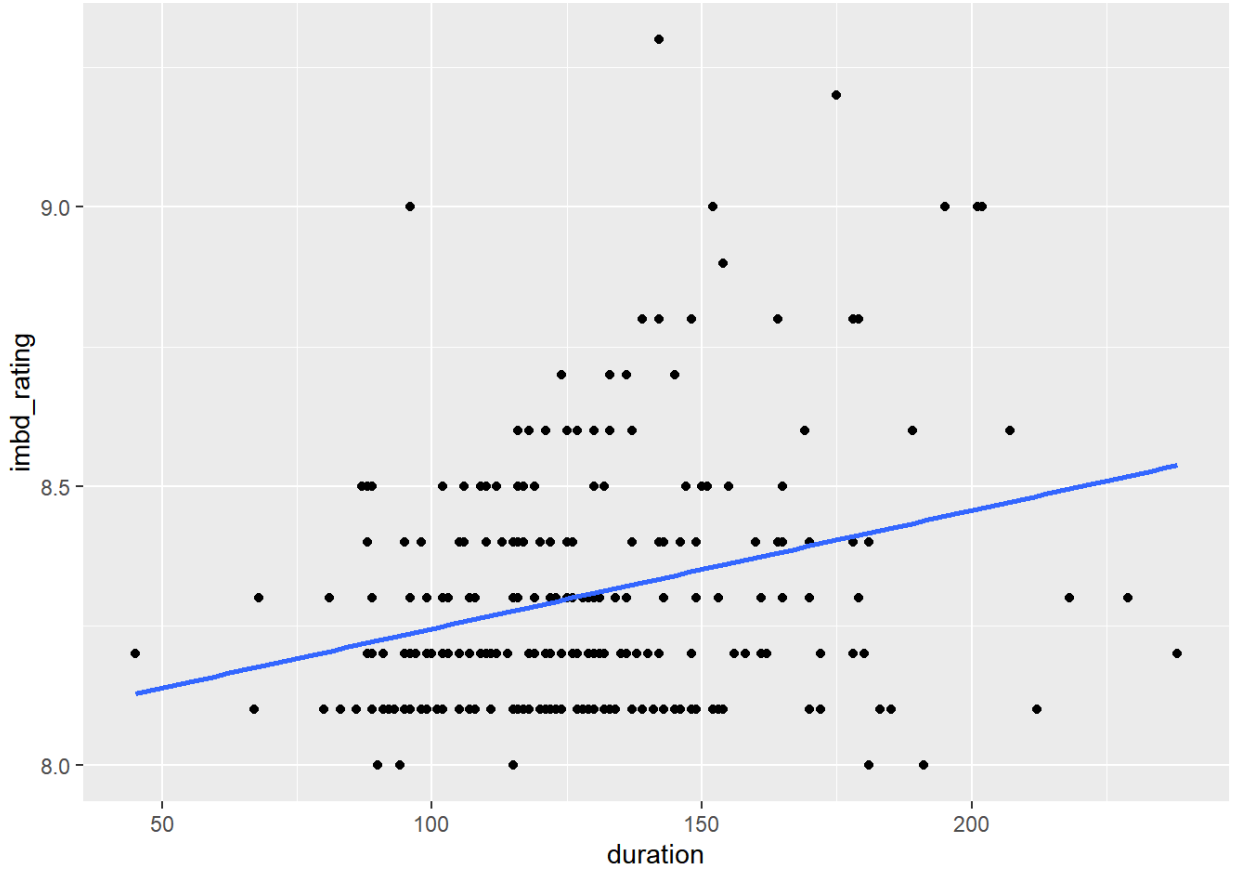
   a. **Reasons behind the choice of analytical techniques. Explain why the technique is suitable for the questions being addressed**

The initial step in our analysis involved the creation of a scatterplot with movie duration plotted against the corresponding IMDb ratings. We then fitted a linear regression model to the plotted data points to observe any potential correlation visually. The purpose of this was to explore the nature of the relationship between movie duration and rating at a glance.

We then conducted a Pearson correlation test to quantify the degree of this relationship. This step was essential as it provided a numerical measure to understand the strength and direction of the relationship between the two variables.

b. Discuss the results from the analysis run. Wherever possible use charts to convey results succinctly

The results of the correlation test showed a correlation coefficient of 0.2794597. While this value indicated a positive correlation between movie duration and ratings, the strength of this relationship was relatively weak. This finding suggested that as movie duration increased, the movie rating also tended to increase, but the effect was not very strong.



1 - Linear regression: movie duration vs. imdb rating

Further scrutiny of the linear regression model revealed an estimated coefficient for movie duration of 0.0021199. This suggested that, assuming all other factors remain constant, an increase in movie duration by one unit (minute) would lead to an expected increase in the movie rating by 0.0021199 units. However, it is important to note that the R-squared value from our model was 0.0781, meaning that only 7.81% of the variability in movie ratings could be explained by movie duration. This result indicated that a considerable proportion of the variability in movie ratings was due to factors not included in our model.

c. Discuss the conclusions from the analysis and offer recommendations in a form that is simple and understandable to decision makers. Support conclusions with relevant charts

Given the insights from the initial findings, the chosen analytical techniques appeared suitable for the research question at hand. However, recognizing the limitations of our model, future iterations of our analysis could include other influential variables such as genre, cast, and direction to capture a broader perspective of factors influencing movie ratings.

Within the context of the film industry, our analysis revealed that while movie duration appeared to have some impact on the ratings, it was far from being the sole or even the primary determinant. Therefore, while crafting strategies to improve movie ratings, industry professionals should not overly focus on movie duration but also consider other factors.
Our analysis highlights the importance of considering a variety of factors when attempting to improve movie ratings. While movie duration does have a part to play, it contributes to less than 8% of the variability in ratings. Therefore, industry professionals might see greater improvements in ratings by focusing on aspects such as storyline quality, casting, direction, and genre.

For future improvements, we recommend using multivariable models to provide a more comprehensive understanding of the factors influencing movie ratings. By including a wider range of variables, decision-makers in the film industry will have a more holistic tool to aid in their strategic planning, ultimately leading to the creation of better-rated movies.

Question 2: Impact of Film Critics and Reviews on Movie Ratings: Regression Tree with Emotion and TF-IDF Features

Impact of Film Critics and Reviews on Movie Ratings: Regression Tree with Emotion and TF-IDF Features
The purpose of this analysis is to gain insight into the impact of film critics and reviews on movie ratings through the use of two different regression tree models with emotion features and TF-IDF features. The dataset consists of IMDb ratings, film critics' scores, movie reviews, and either emotion or TF-IDF features.

Regression Tree with Emotion Features:
The first regression tree model explored the relationship between emotion features and IMDb ratings. Our analysis considered the following aspects:

- Reason for using regression tree with emotion features
    - It can capture Nonlinear Relationships, this is due to the fact that regression tree models excel in capturing non-linear relationships between emotion features and IMDb ratings. Emotions are known to have intricate effects on movie ratings, and the regression tree's hierarchical structure can identify significant combinations of emotions that influence ratings.
    - It also features importance and interpretability, while The regression tree provides an interpretable model with feature importance rankings. The decider can easily understand how specific emotions impact movie ratings, helping them make informed marketing and content creation decisions.
- Results from regression Tree with emotion features

    - We performed a linear regression analysis to understand the relationship between the movie ratings and the emotional features. Here are the results of the regression model: The intercept of the model is 8.464, indicating that if all predictors that contain emotion features are zero, then the expected value of movie ratings for imbd_rating would be 8.464. The result, shows the coefficients for all emotion predictors are very close to zero, suggesting that these emotion features have a relatively small influence on movie ratings. The p-values for all predictors are greater than 0.05, indicating that the influence of these emotional features on movie ratings is not statistically significant. The R-squared value of the model is 0.04592, and the adjusted R-squared value is -0.01847. The low R-squared value indicates that the model explains only a small portion of the variability in movie ratings, hinting that the emotion features alone are not sufficient to predict movie ratings accurately. The F-statistic of the model is 0.7131, with a corresponding p-value of 0.7248. This suggests that, as a whole, the emotional features in the model are not statistically significantly related to movie ratings.
- Conclusion
    - Based on the results of the linear regression analysis, the emotional features extracted from movie reviews do not have a major impact on movie ratings. The model's low R-squared value indicates that factors other than emotions play a significant role in determining movie ratings. Therefore, relying only on emotional features from movie reviews may not be an effective way to predict movie ratings accurately.
- Recommendation:
    - To improve the accuracy of movie rating predictions, it is essential to consider other relevant factors such as movie genre, director, actors, and plot complexity. A more comprehensive model that incorporates these additional features may lead to a better prediction of movie ratings.

Regression Tree with TF-IDF Features:
The second regression tree model looks at the impact of TF-IDF features, representing the importance of keywords in movie reviews, on IMDb ratings. Our analysis highlighted the following:

- Reason for using regression tree with emotion features
    - Through analysis of the textual data, the regression tree models are well organized for analyzing non-linear relationships between TF-IDF features and IMDB ratings. This allows the identification of important keywords that influence movie ratings.
    - Feature importance and visualization, the regression tree offers a clear ranking of the most important TF-IDF terms, which allows the decision-makers to focus on keywords that significantly impact movie ratings. The visualization aid in presenting the results effectively.
- Results from regression Tree with emotion features

    - From the tree algorithm called CART (Classification and Regression Trees), we used its result to build a regression tree model to predict movie ratings based on TF-IDF features. The results of the decision tree model are: the RMSE of the prediction is 0.2699699, which points out the average prediction error of the model. The smaller the RMSE for the predictive model the better it is. The coefficient of determination R square of the model is 0.03317102. The value represents the percentage of variance in movie ratings that the model is shown. The closer the R-squared value is to 1, the stronger the explanatory power of the model.
- Conclusion:
    - The decision tree model using TF-IDF features has a moderate predictive performance with a relatively higher RMSE and a low R-squared value. This indicates that more TF-IDF features are required to accurately predict movie ratings.
- Recommendation:
    - We recommend enhancing the prediction accuracy and including additional relevant features, such as movie genre, director, and review sentiment, along with TF-IDF features. A more comprehensive model that considers multiple factors is likely to improve the ability to predict movie ratings effectively.

Overall Conclusion:
In this analysis, we explored two different approaches to analyze the impact of film critics and reviews on movie ratings. The linear regression model used emotion features to indicate that emotion alone has a minimal impact on movie ratings, while the decision tree model used TF-IDF features and demonstrated average predictive performance. However, both models have limitations in accurately predicting movie ratings, indicating the importance of considering other relevant factors to improve prediction accuracy.

Final Recommendation:
For deciders involved in assessing the predicting movie ratings, it is essential to adopt an extensive approach that contains multiple factors, such as director, movie genre, actors, and review sentiment. Then apply the textual analysis techniques beyond TF-IDF. Through the combination of these elements, it can drive more accurate and reliable movie rating predictions that can be achieved. Lastly, it will prove valuable insights for the movie industry.

Question 3: "Which movie genres tend to receive higher ratings from audience"

Reasons behind the choice of analytical techniques. Explain why the technique is suitable for the questions being addressed

The genre of film is a key factor to understand what resonates with audiences, and this is critical for filmmakers, investors and platforms alike. We explores the relationship between film genres and IMDb ratings through statistical analysis of linear regression models, with the aim of revealing which genres has stronger relationship with higher ratings.

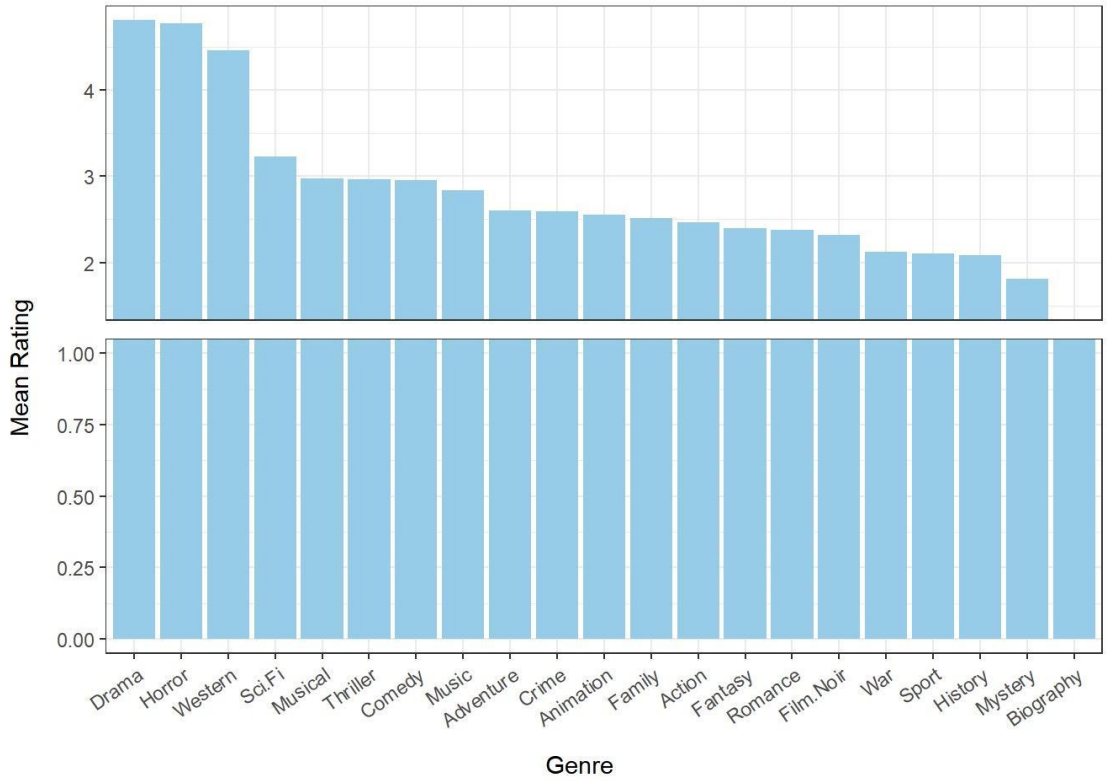
Selecting linear regression analysis to predict IMDb ratings based on film genres is for its ability to model a linear relationship between dependent(rating) and independent variables(genre). Our dataset comprises various genres, the regression model was built, encapsulating 21 different genres as predictors for IMDb ratings. The choice of linear regression was also informed by its interpretability and flexibility, allowing a more straightforward understanding of how each genre impacts ratings, and it is preferred when there is concern about overfitting, especially if the dataset is not very large. Thus, this model will be more preperable for broader business understanding.

d. Discuss the results from the analysis run. Wherever possible use charts to convey results succinctly

Significant findings

The summary of the linear regression analysis revealed important insights into the relationship between film genres and IMDb ratings; Because there was a significant influence of genres, as almost all genres were found to have a statistically significant impact on the ratings.

Genres like Drama, Horror, and Western emerged as strong predictors of higher ratings. This may be attributed to the emotional resonance and storytelling elements commonly found in these genres. A bar chart showing the mean ratings for each genre clearly illustrate the popularity of different genres. The x-axis would represent the genres, and the y-axis would represent the mean ratings. The genres is ordered by mean ratings, and we can see that Drama is the highest, followed by genres like Horror and Western.



In terms of the model's goodness of fit, the R-squared value of 0.9545 demonstrated that 95.45% of the variability in IMDb ratings could be explained by the genres, indicating a robust fit. This high value highlight the important role of genre selection in shaping audience perception and preferences.

Combination of Genres

We have also examined the genre combinations that receive the highest ratings, and we visualize the top 5 genre combinations with the highest average ratings using a bar chart. The x-axis represents the genre combinations (e.g., "Adventure, Western"), and the y-axis represents the average rating. The top five combinations include blends of Adventure, Western, Action, Drama, Sci.Fi, Crime, Family, and Fantasy. This suggests that certain combination of genres are particularly appealing to audiences.

```
## # A tibble: 5 x 3
##   GenreCombination         AverageRating NumberOfVotes
##   <chr>                            <dbl>         <int>
## 1 Adventure, Western                 8.8        769389
## 2 Action, Adventure, Drama           8.7       8118258
## 3 Adventure, Drama, Sci.Fi           8.6       1870060
## 4 Crime, Drama, Fantasy              8.6       1317747
## 5 Drama, Family, Fantasy             8.6        467723
```

Number of Genres

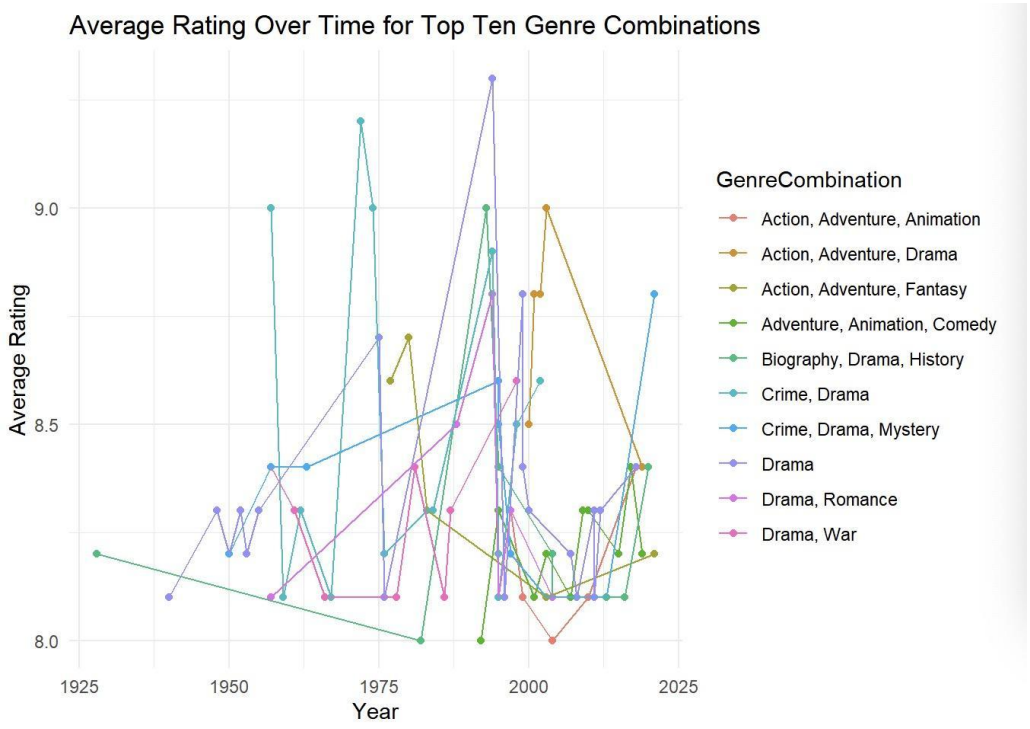
The number of genres in a movie correlates with the rating. However, it looks like there isn't a significant difference in average ratings between movies with different numbers of genres. From the line chart, the genre count on the x-axis and average rating on the y-axis illustrate the relationship between the number of genres in a movie and its rating. This plot shows a relatively flat line, indicating that the number of genres doesn't significantly affect the rating. The number of genres in a film does not have a strong influence on its popularity.

```
## # A tibble: 3 x 2
##   GenreCount AverageRating
##        <int>         <dbl>
## 1          1          8.32
## 2          2          8.32
## 3          3          8.30
```

Time Series

Finally we used Time Series Analysis showing the average rating over time for the top ten genre combinations. The x-axis would represent the years, and the y-axis would represent the

average rating. Different lines would represent different genre combinations. This plot would reveal trends and changes in popularity for different genre combinations over time.



e. Discuss the conclusions from the analysis and offer recommendations in a form that is simple and understandable to decision makers. Support conclusions with relevant charts

The analysis reveals some important insights that can inform decisions in the film industry. Drama appear as the most popular genre, receiving the highest average ratings among all genres, which suggests a strong audience affinity for dramatic content. Simultaneously, certain genre combinations such as "Adventure, Western" and "Action, Adventure, Drama" have proven to be particularly appealing to audiences, which highlights the creative potential of blending genres in unique ways. Interestingly, the analysis shows that the number of genres within a film doesn't significantly influence its rating, so the filmmakers have flexibility in combining genres without affecting appeal.

Furthermore, by examining trends over time through a line plot of the top ten genre combinations, the study uncovers shifts in genre popularity across years. This historical perspective allows for the identification of potential future trends in audience preferences.

Based on these conclusions, the recommendations for decision makers is investing in Drama films or incorporating dramatic elements. Exploring successful genre combinations offers a creative pathway to success. Flexibility in number of genres means creators can blend elements as they see fit. Filmmakers could also trends over time and get insights for future content acquisition, creation, and marketing strategies. These insights would informed decision-making in the film industry.