STAT-S 670: EXPLORATORY DATA ANALYSIS

IMDB MOVIES

Team Members
Netaji Sai Pavan Neerukonda
Kaushik Parvathaneni
Venkata Sai Pavan Anvesh Tamidala

Statement of goals

The objective of our project is to conduct exploratory data analysis (EDA) on the top 1000 movies on IMDb and gain insights.

The research questions that can be answered through this dataset are:

- What are the most popular genres among the top-rated movies ?
- Who are the most popular directors and actors among the top-rated movies?
- How does the gross revenue of the movies vary by decade?
- What is the relationship between the IMDb rating, Metascore, number of votes and the gross revenue of the movies?
- How did the runtime of the movies vary over the years?
- In what decade what duration films are more popular?

These questions are significant for the subsequent reasons:

- Recognizing what audiences are interested in viewing can be determined by exploring the most popular genres, directors, and actors. Filmmakers, studios, and streaming services can use this data to decide which type of fifilms to invest in.
- This also helps audience to understand how well a movie has been received by other viewers and whether it is worth watching. This information can be particularly useful for audiences who are trying to decide between multiple movies to watch.
- We can gain a better understanding of how a film is doing by examining the relationship between the IMDb rating, Metascore, number of votes, and gross revenue. This can help studios and investors estimate how much money to invest into a film and what kind of profits to anticipate.
- An understanding of how movie runtimes and certificates have transformed over time could demonstrate to us significant data about the motion picture business. When making decisions about their upcoming projects, studios and filmmakers could utilize this data to be beneficial.
- Audiences can manage their expectations and plan for their movie-watching experiences by being conscious of how movie runtimes and certifications have evolved over time. For instance, if viewers are aware that films are often getting longer, they may need to make plans to watch them for longer lengths of time or alter their schedules accordingly.

Figuring out the answers to these questions can help the production houses and inverstors invest in and produce movies more strategically and can also give substantial data about audience preferences and market dynamics.

Data Description

We have taken the dataset for our project from Kaggle.
https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

The columns in this dataset are:

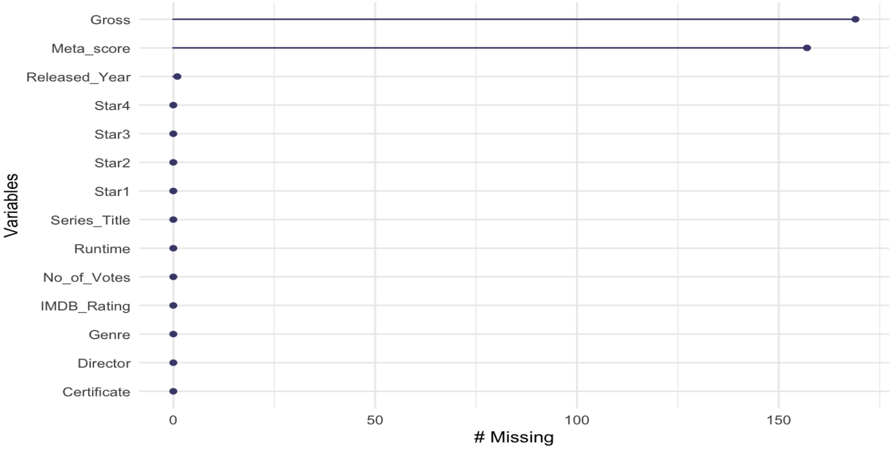
- Poster_Link - Link of the poster used on IMDb
- Series_Title - Name of the movie
- Released_Year - Year the movie released
- Certificate - Certificate earned by that movie
- Runtime - Total runtime of the movie
- Genre - Genre of the movie
- IMDB_Rating - Rating of the movie on IMDb
- Overview - Mini story/ summary of the movie
- Meta_score - Score earned by the movie
- Director - Name of the Director
- Star1,Star2,Star3,Star4 - Name of the Stars
- No_of_votes - Total number of votes
- Gross - Money earned by that movie

Genre: This column can be used to identify the most common genres among the top-rated movies and TV shows.
Director and Star1,Star2,Star3,Star4: These columns can be used to identify the most popular directors and actors among the top-rated movies and TV shows.
Released_Year: This column can be used to track changes in movie and TV show production over time.
Runtime: This column can be used to analyze changes in the length of movies and TV shows over time, and to identify trends in the preferred length of movies and TV shows.
IMDB_Rating, No_of_Votes, Gross and Metascore: These columns can be used to analyze the relationship between the quality of movies and TV shows (as measured by rating and Metascore) and their popularity (as measured by number of votes and revenue).
Certificate: This column can be used to identify trends in movie and TV show certification over time, and to analyze the relationship between certification and other variables such as rating and revenue.

Quantitative variables are variables that can be measured or counted numerically, and they can be further classified as discrete or continuous, whereas, ***Qualitative variables*** are variables that cannot be measured numerically, and they can be further classified as nominal or ordinal.
Here are the quantitative and qualitative variables in the dataset:
Quantitative variables - *Released_Year, Runtime, IMDB_Rating, Metascore, No_of_Votes, Gross*
Qualitative variables - *Series_Title, Certificate, Genre, Director, Star1, Star2, Star3, Star4*

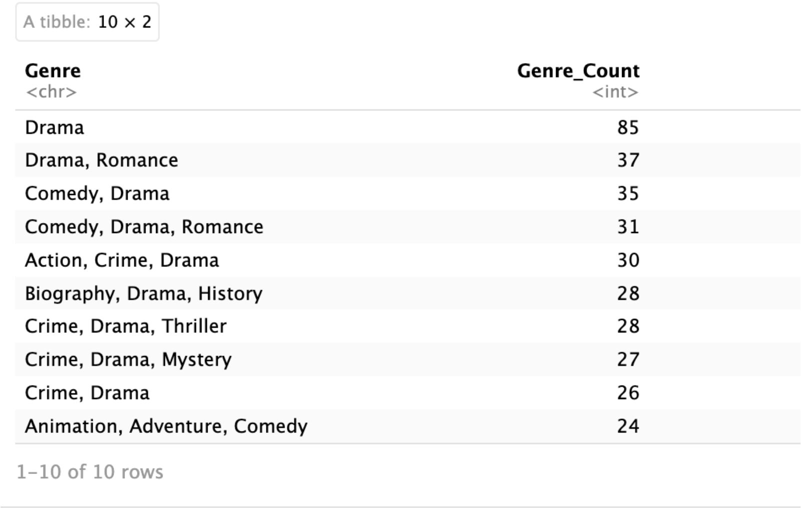
Data Preprocessing:

For data preprocessing, we have removed the attributes ***Poster_Link*** and ***Overview*** as these are discrete and qualitative attributes that might not be helpful for data analysis. Also, there are few null values in the columns ***Gross*** and ***Meta_Score*** so we have removed the rows related with these null values for further analysis.
We also converted the ***Runtime*** and ***Released_Year*** columns to numeric data types, and removed commas and dollar signs from the ***Gross*** column before converting it to integer format.



Answering our Questions

What are the most popular genres among the top-rated movies ?



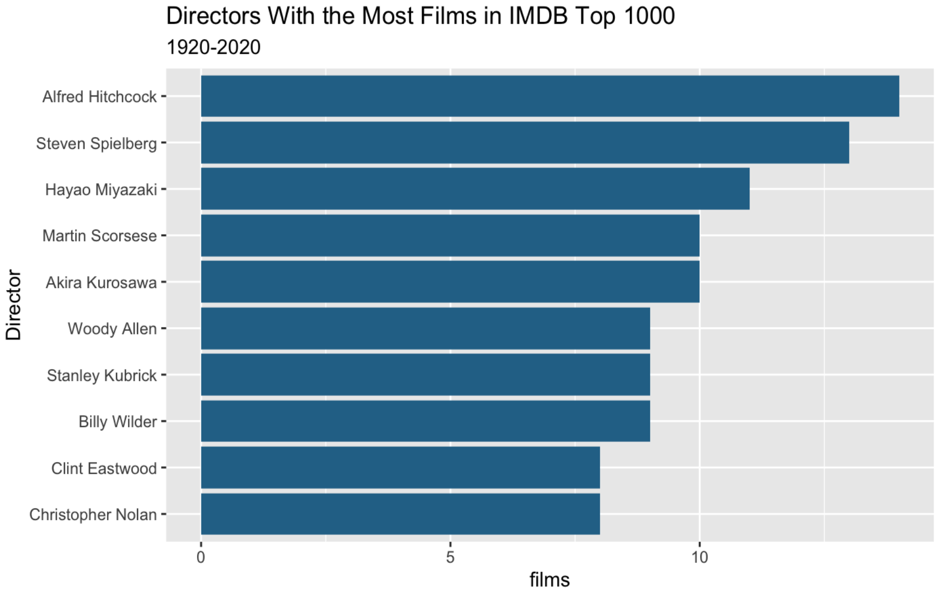
A tibble: 10 × 2

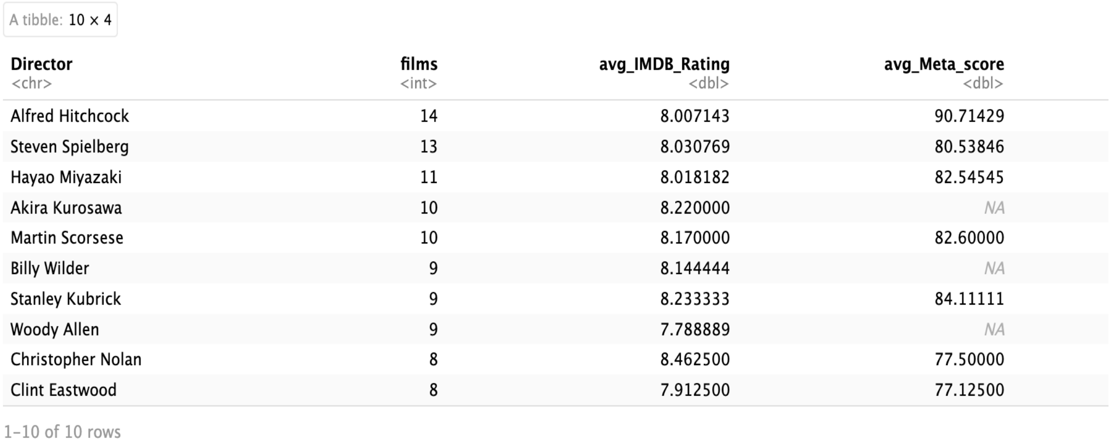
Genre <chr>	Genre_Count <int>
Drama	85
Drama, Romance	37
Comedy, Drama	35
Comedy, Drama, Romance	31
Action, Crime, Drama	30
Biography, Drama, History	28
Crime, Drama, Thriller	28
Crime, Drama, Mystery	27
Crime, Drama	26
Animation, Adventure, Comedy	24

1–10 of 10 rows

The data clearly indicates that Drama is the preferred genre among movie-goers, followed closely by Drama and Romance. This provides valuable insights for film studios, allowing them to produce emotionally-driven narratives that resonate deeply with audiences. Focusing on creating high-quality drama films can lead to increased ticket sales and profitability. Importantly, this popularity of the Drama genre is enduring and not a fleeting trend, providing studios with a strategic advantage in the competitive world of film-making. By leveraging these insights, studios can make informed decisions about content creation and distribution, ultimately leading to greater success.

Who are the most popular directors among the top-rated movies?


Directors With the Most Films in IMDB Top 1000
1920-2020


A tibble: 10 × 4

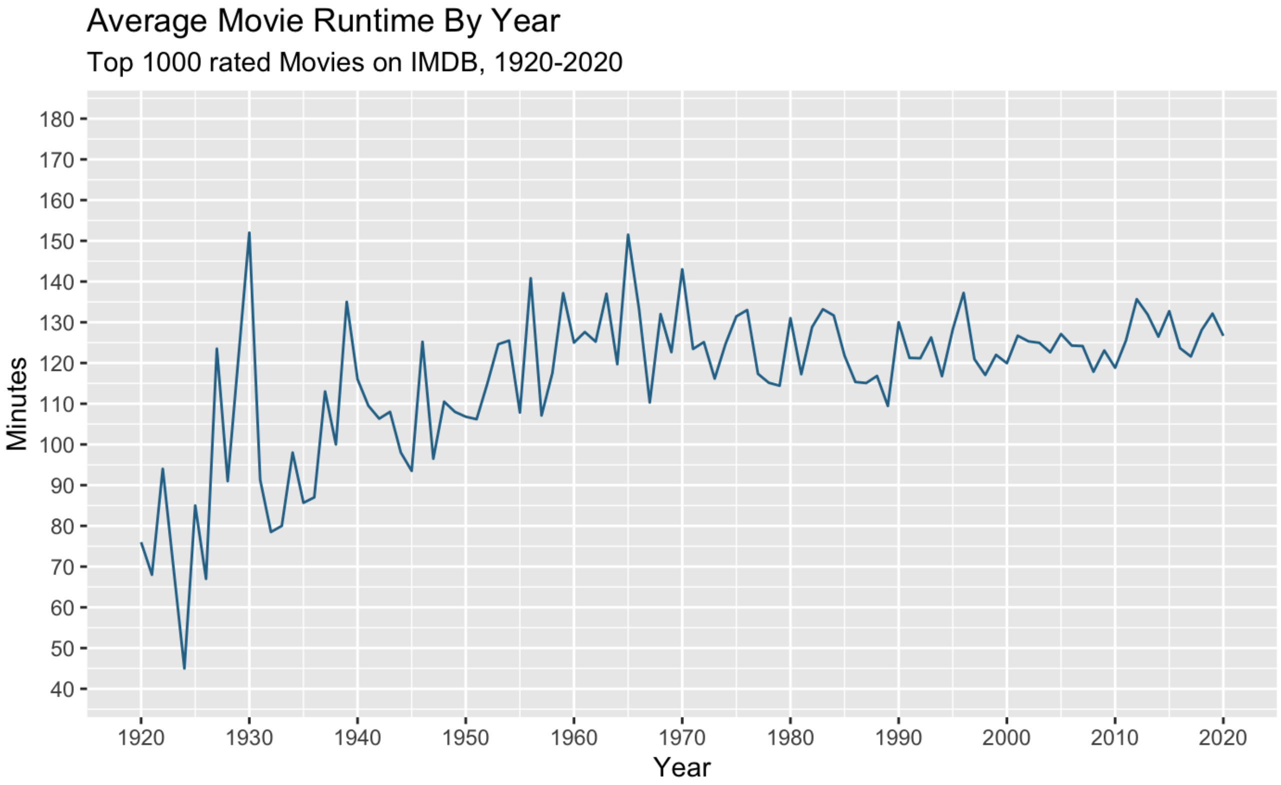
Director	films	avg_IMDB_Rating	avg_Meta_score
<chr>	<int>	<dbl>	<dbl>
Alfred Hitchcock	14	8.007143	90.71429
Steven Spielberg	13	8.030769	80.53846
Hayao Miyazaki	11	8.018182	82.54545
Akira Kurosawa	10	8.220000	NA
Martin Scorsese	10	8.170000	82.60000
Billy Wilder	9	8.144444	NA
Stanley Kubrick	9	8.233333	84.11111
Woody Allen	9	7.788889	NA
Christopher Nolan	8	8.462500	77.50000
Clint Eastwood	8	7.912500	77.12500

1–10 of 10 rows

For each director, we calculated the average IMDb rating and meta score of the films they directed and also the total number of films directed by them. The resulting table is sorted in descending order based on the number of films directed and includes the top ten directors.

The resulting table provides information about the top ten directors with the most films in the dataset, along with their average IMDb rating and meta score from which we can interpret that **Alfred Hitchcock** has the most number of films in the top 1000 IMDb movies and **Christopher Nolan** has the highest avg_IMDB_Rating which makes him the top rated director.

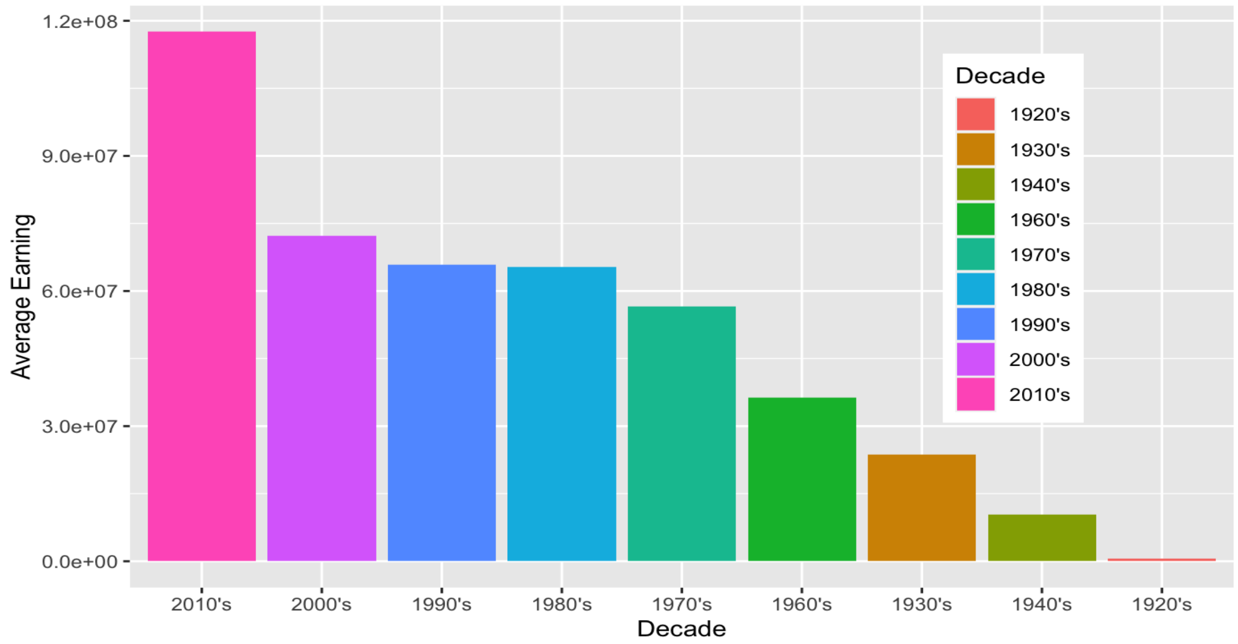
How did the runtime of the movies vary over the years?



We created a line plot using ggplot that shows the trend of average movie runtime over the years from 1920 to 2020. The x-axis represents the years and the y-axis represents the average movie runtime in minutes. The plot shows that the average movie runtime has varied over time, with a significant increase in the runtime in the early 1930s, followed by a steady decline in the 1940s and 1950s. There is also a significant increase in the runtime in the early 1970s, followed by a steady decline in the 1980s and 1990s The plot also shows that the average movie runtime has been relatively stable since the early 2000s.

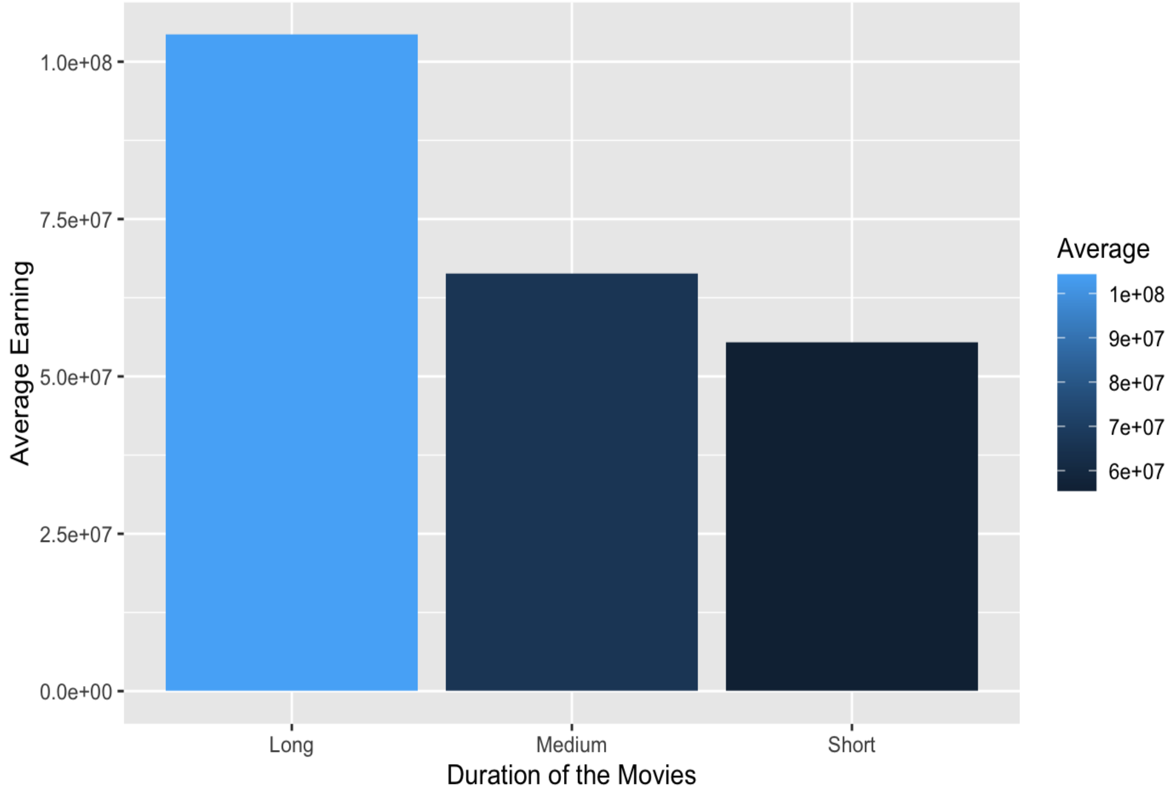
Are the earnings continuously increasing with respect to decade or there is a downfall in earnings? (How does the gross revenue of the movies vary by decade)

From the below plot we can clearly see that there is always an increase in the collections but there are exceptions here. In 1940s the earnings decreased compared to 1930s this is due to World War happened at that time. And afterwards earnings have decreased in 2020s due to the pandemic COVID-19.

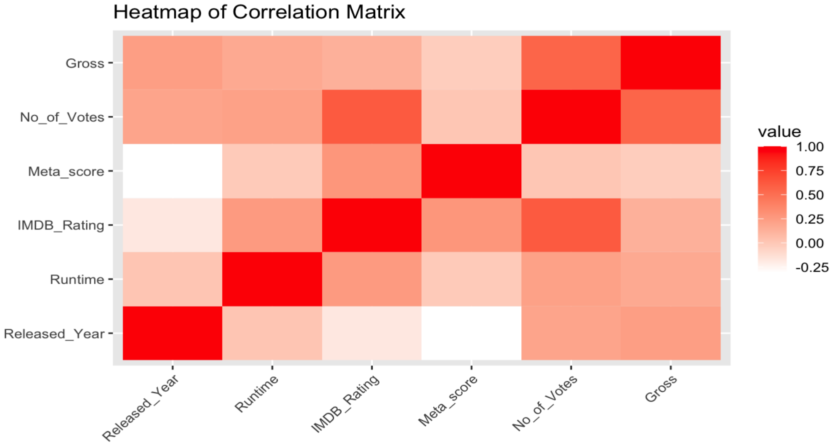


How does the runtime of the movies affect the earnings of the movie?

From the plot we can observe that if the runtime of the movies is directly proportional to the earnings of the movie. But this is not always the same as the years passes people tend to watch movies with the medium runtime rather than long or short runtime.



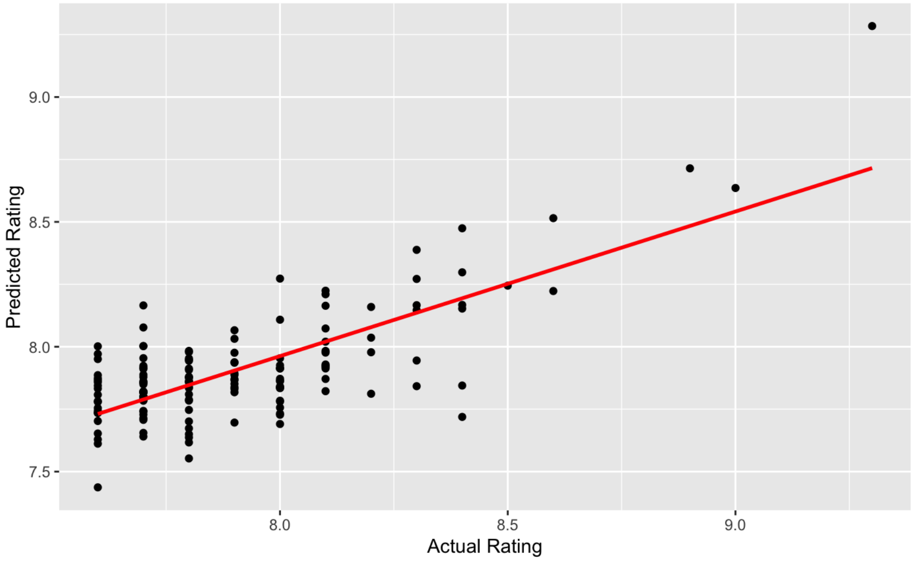
What is the relationship between the IMDb rating, Metascore, number of votes and the gross revenue of the movies?



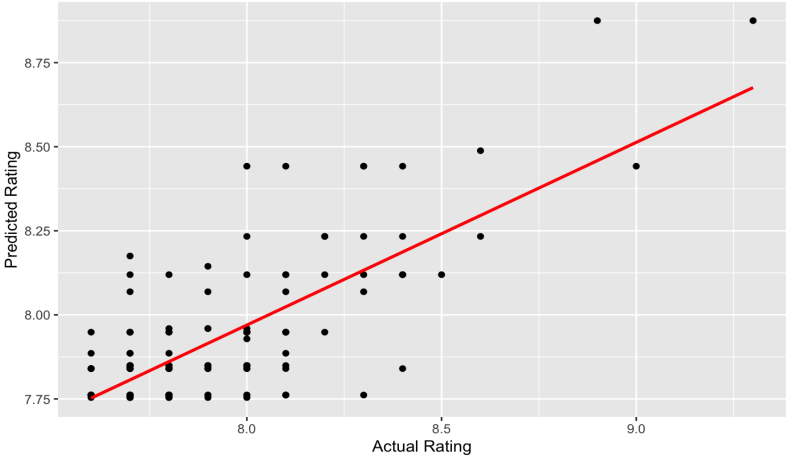
The color scale goes from *white* (no correlation) to *red* (strong positive correlation). The heatmap's findings demonstrate a strong positive correlation between the number of votes and the gross, demonstrating that movies with high collections i.e. more gross frequently receive more votes. There is also a positive correlation between the movie's rating and the total number of votes, demonstrating that highly rated films frequently receive more votes. The meta score, which is determined by expert critics, and the movie's rating also have a positive correlation. This implies that movies are more likely to receive higher ratings if their meta scores are greater.

Regression Analysis on Movie Ratings:

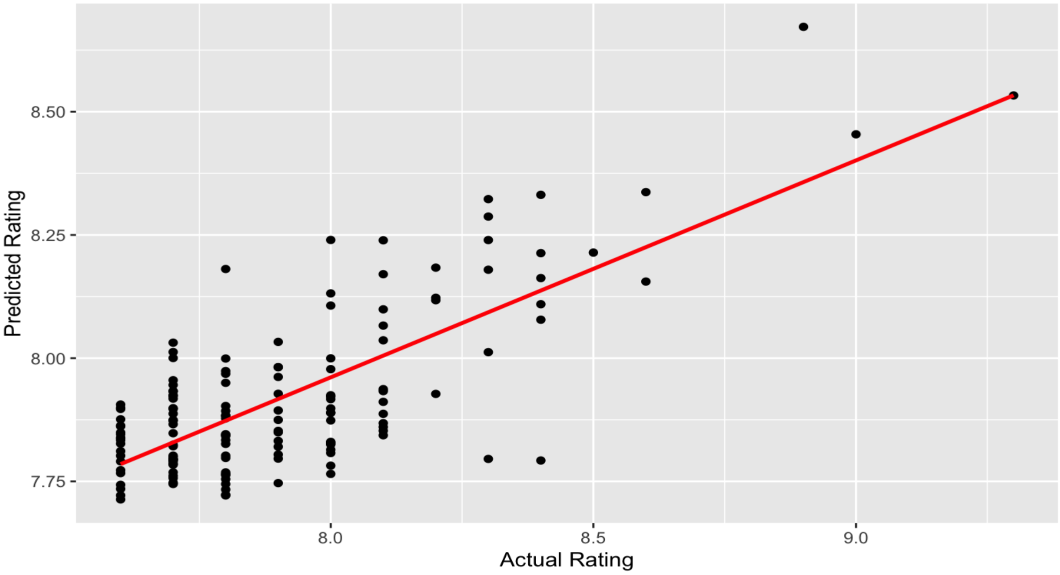
Linear Regression Model: RMSE: 0.1926021



Decision Tree Regression model: RMSE: 0.1928382



Random Forest Regression model: RMSE: 0.1974437

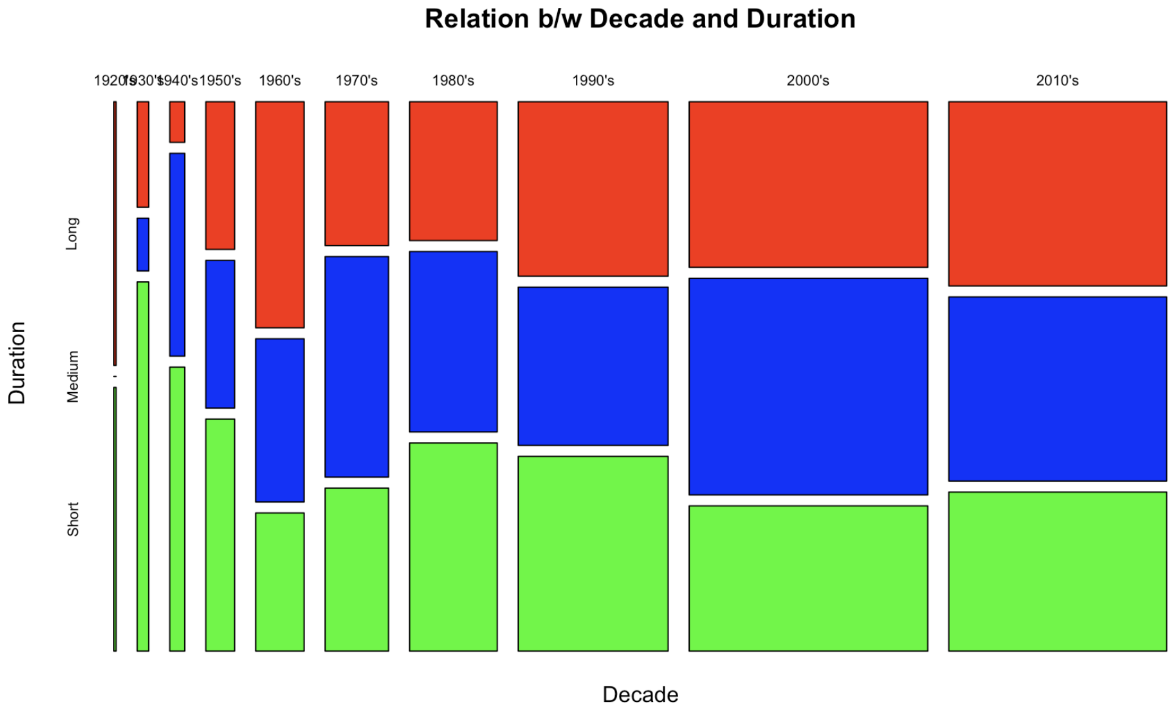


We have fitted three different models for our data. The models were Decision Tree, Random Forest and Linear Regression. We observed that Decisio tree and Linear Regression had similar performance with linear regression having a slightly lower RMSE values. Selecting the best model is defined on various factorsbut here we are considering the RMSE value so the best value here is Linear Regression model as the data points align nearly with the line and also RMSE value proves it.

A low RMSE value indicates that the model is able to predict the target variable with high accuracy, while a high RMSE value indicates that the model has poor predictive performance.Therefore, when comparing models, a model with a lower RMSE value is generally considered to be a better model than one with a higher RMSE value. However, it's important to keep in mind that RMSE is just one metric for evaluating model performance, and it's always a good idea to consider multiple evaluation metrics when selecting the best model for your specific problem.

What duration of movies are loved by people in respective decade?

The use of mosaic plots has allowed us to gain insights into audience preferences for movie length across different decades. Interestingly, the data shows a shift in preference towards medium-length movies as time progresses, with a decrease in popularity for both short and long-length movies. This finding has important implications for the movie industry in terms of content creation and distribution, as studios may need to adapt to changing audience preferences in order to remain successful.



Limitations:

The factors included in the dataset are not enough to gain complete insight on the movies so we can add other external factors like production budget, marketing spend, and demographic information on the audience. Furthermore, potential biases in the dataset, such as sampling and measurement bias, need to be taken into account to ensure the analysis is accurate and representative. For example, the dataset may only include movies from certain countries or time periods, which can limit its representativeness. Also our graphs may not be accurate as variables like ratings may be subject to measurement bias as they rely on individual user ratings that may not be fully representative of broader opinions. Addressing these limitations in the analysis may involve incorporating additional data sources or refining the methodology to account for potential biases. By doing so, a more comprehensive understanding of the movie industry's trends and patterns can be gained, leading to better decision-making and strategic planning for the project.

Findings:

- Drama, Action, and Comedy are the most popular genres among the top-rated movies.
- Christopher Nolan is the most popular director among the top-rated movies, while Tom Hanks is the most popular actor.
- The IMDb rating and the gross revenue of the movies have been increasing over the years, with a slight dip in 2020 due to the COVID-19 pandemic.
- There is a weak positive correlation between the IMDb rating, meta score, and the gross revenue of the movies.
- The runtime of the movies varies significantly by genre, with drama movies having the longest runtime.

Future work:

- Conducting statistical analyses to test the significance of the relationships observed in the EDA.
- Building a predictive model to predict the IMDb rating and gross revenue of the movies based on their characteristics.
- Analyzing the impact of various factors such as the release date, marketing budget, and critical acclaim on the success of the movies.
- Collecting additional data on user reviews, awards, and nominations to gain more insights into the success of the movies.

Conclusion:

- This project can be used to examine a variety of data aspects in the context of the IMDb movies dataset, including the distribution of movie ratings, the relationship between movie budget and income, the appeal of different genres, and the skills of different actors and directors.
- We can gain a deeper understanding of the data and identify any potential issues, such as missing data, outliers, or conflicts.
- This project also helps with the selection of relevant statistical methods and models for additional research.
- This project provide insightful information that can aid in decision-making and improve the accuracy and dependability of the results.