A Decade of Movies: Exploring Trends and Patterns in IMDB Data from 2006-2016

|  |
| --- |
| **Introduction** |

Cinema refers to the art of creating motion pictures, also known as movies or films, using a combination of visuals, sound, and narrative to tell a story or convey an idea. It is a form of mass communication and entertainment that has become an important part of modern culture, offering a variety of benefits to society. The socio-economic benefits of film industry are -

1. Provides employment opportunities and stimulates economic growth.

2. Promotes cultural exchange and diversity.

3. Generates revenue for local economies through tourism.

4. Encourages education and awareness-raising efforts.

5. Stimulates innovation and creativity.

6. Promotes technological advancements.

7. Boosts the development of related industries such as fashion and music.

8. Increases international trade and business opportunities.

9. Enhances a country's global cultural image.

10. Provides a platform for social commentary and positive social change.

Overall, cinema has become an important part of our society, offering numerous socio-economic benefits. Through film, we can enjoy entertainment, promote cultural exchange and diversity, stimulate economic growth and development, promote tourism, and raise awareness of important social issues.

|  |
| --- |
| **About The Dataset** |

For our project, we decided to pull data describing the world’s most popular **IMDb** movies from 2006 to 2016. IMDb stands for the Internet Movie Database. It is an online database that contains information about movies, TV shows, and video games. IMDb provides details such as cast and crew information, release dates, ratings, fan and critical reviews, trivia, and other related information about each title.

The website was launched in 1990 and is now owned by Amazon. It is one of the most popular websites for movie enthusiasts, industry professionals, and casual viewers alike. Users can create accounts, rate movies, and contribute their own reviews and trivia. The site also features a list of the top-rated movies of all time, as voted by the site's users, as well as industry news and updates. Overall, IMDb serves as a comprehensive resource for anyone interested in movies and TV shows. Although this data is coming from IMDb as the root source, it is hosted on [Kaggle](https://www.kaggle.com/), and the specific dataset can be found [here](https://www.kaggle.com/PromptCloudHQ/imdb-data?fbclid=IwAR1RReRdXCLWdFUwMOIsj-vvW8fP6tCUvjipIrT957lCctyASQDc-H5CgD8).

https://www.kaggle.com/datasets/PromptCloudHQ/imdb-data?fbclid=IwAR1RReRdXCLWdFUwMOIsj-vvW8fP6tCUvjipIrT957lCctyASQDc-H5CgD8

The dataset we have on IMDb movies comprises information on the top 1,000 most well-liked movies on IMDb during the 2006-2016 period. This particular dataset is made up of 1,000 rows, or instances, and 11 columns, or attributes. The variables in this data set are a combination of quantitative and qualitative attributes, with a total of 5 quantitative attributes and 6 qualitative attributes..

The movie dataset comprises of various attributes, including both quantitative and qualitative features. The former includes the movie's **Title, Genre, Description, Director** and **Actor**, while the latter includes **Year** of release, **Runtime (minutes), Rating, Votes, Revenue(millions),** and **Metascore**.

The Title attribute refers to the name of the movie, while Genre is a classification based on the comma-separated list of genres the film belongs to. Description summarizes the plot in a single sentence, and Director refers to the person who directed the film. Actors lists the main stars of the movie.

The Year attribute indicates the year of release, expressed as an integer. The Runtime (Minutes) attribute specifies the duration of the movie in minutes. Rating is a user-rated score, ranging from 0 to 10. The Votes attribute represents the number of votes the movie has received. Revenue (Millions) is the total box office earnings of the film, given in millions. Finally, Metascore represents an aggregate average of critic scores, ranging from 0 to 100, where higher scores reflect positive reviews.

Collectively, these attributes provide a comprehensive summary of each movie, allowing for easy comparison and analysis of different films based on various criteria.

|  |
| --- |
| **Understanding the information in our data** |

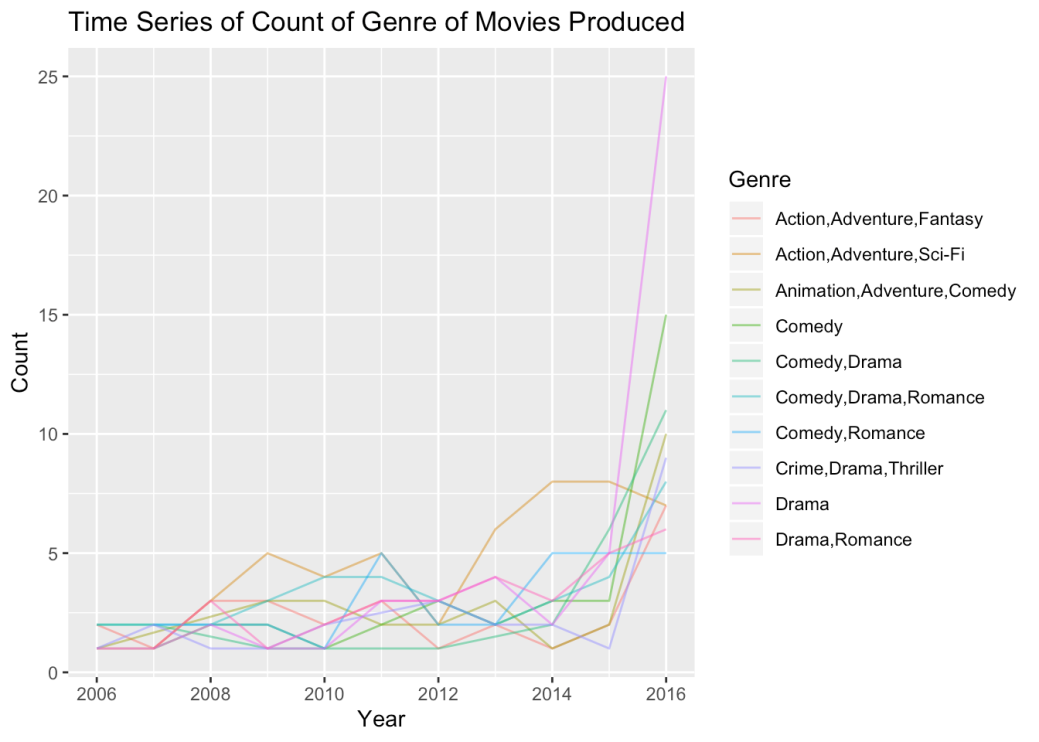
To effectively analyze and plot our data in order to answer our research questions, it's important to have a good understanding of the data we're working with. In particular, we'll be focusing on attributes such as genre, movie performance, user and critic ratings, title, and descriptions, so it's crucial that these columns are in a format we can work with. One challenge we encountered was that our dataset contains over 200 subsets of genres, with many overlapping each other. To simplify the analysis, we decided to focus on the top 10 genres with the most frequency of movies represented in the dataset. We also checked the data types of each attribute, and found that some were of character type (such as Title, Genre, Description, Director, and Actor) while others were of double type (Year, Runtime (Minutes), Rating, Votes, Revenue (Millions), and Metascore). As we will be using the user and critic ratings and movie fiscal performance attributes extensively, it is appropriate that they are all the same data type and quantitative, which they are, so no further cleaning is needed. With a good understanding of our data, we can now dive into our research questions.

**Can a movie's genre influence its long-term popularity and financial success?**

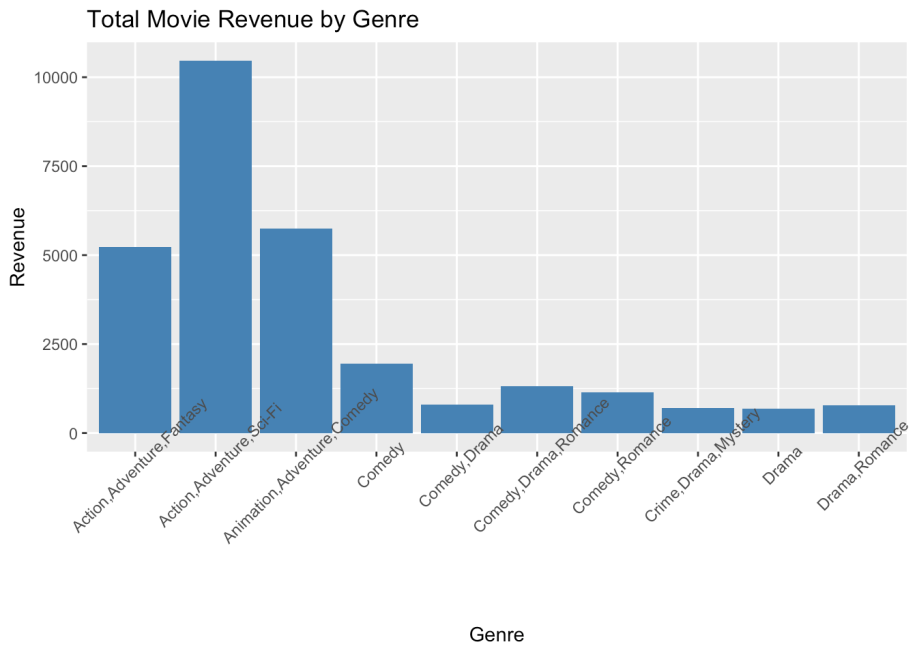
Our initial area of interest is exploring the correlation between a movie's genre and its level of popularity. We aim to investigate if there are any specific genres that frequently result in underperforming movies, and which genres are preferred by the audience. By examining this, we can provide valuable insights for critics to anticipate the success of a particular movie and offer guidance to producers and writers regarding genres that are either neglected in the industry or typically not well-received by the general public.

|  |
| --- |
| **Plots** |

To determine the popularity of a genre, we can analyze how many movies of a certain subset of genres are produced each year. This is because as the demand for a certain type of movie increases, more movies of the same genre are made. Since our dataset contains 207 different subsets of genres, we will focus on the top 10 combinations of genres. These 10 combinations cover the majority of movies produced and are not overly specialized.



We observed a noticeable rise in the production of movies across most genres, with the exception of Romantic Comedies and Action, Adventure, Sci-Fi movies, which experienced a decline. Among all genres, Drama movies had the highest increase in production. Having studied the relationship between genre and movie production, we will now investigate how genre affects a movie's revenue (in millions).

  
Based on our analysis, the Action, Adventure, Sci-Fi genre seems to generate the highest revenue, followed closely by Action, Adventure, Fantasy, and Animation, Adventure, Comedy. Adventure seems to be a common element in the top-grossing genres, which is not surprising given the popularity of blockbuster franchises like the Marvel Cinematic Universe. On the other hand, Drama, Romance, Crime, and Comedy mixed with Drama are among the lowest grossing genres. While this does not necessarily mean that these genres are not enjoyed by moviegoers, it suggests that they may not be as well-suited for the big screen and could be better suited for streaming services like Netflix or Hulu. For those in the film industry looking to maximize box-office revenue, focusing on Action, Adventure, Sci-Fi and Animation, Adventure, Comedy genres could be a profitable strategy, and collaborations with companies like Marvel or Disney could be beneficial.

|  |
| --- |
| **Statistical Tests and Analysis** |

In order to further validate our findings from the plots, we will utilize a two-sample t-test to compare the mean performance of adventure movies to that of movies not in the adventure genre. To do this, we will divide our dataset into two distinct groups.

What is two sample t-test?

A two-sample t-test is a statistical test used to determine if two independent samples have different means. It is also known as an independent samples t-test. The test assumes that the data in both samples are normally distributed, and have equal variances. The t-test calculates a t-statistic, which measures the difference between the means of the two samples relative to the variation in the data. The larger the t-statistic, the more evidence there is that the means of the two samples are different. The test also calculates a p-value, which represents the probability of observing a difference as large as the one observed if the null hypothesis (that the means of the two samples are equal) is true. If the p-value is less than a pre-determined significance level (usually 0.05), the null hypothesis is rejected, and it is concluded that the means of the two samples are significantly different.

Once we have separated our data into the two groups, we can begin conducting the two-sample t-test to determine if there is a statistically significant difference between the mean performance of adventure movies and non-adventure movies. This will help us confirm or refute our initial hypothesis regarding the impact of genre on a movie's performance.

|  |
| --- |
| **Code** |

|  |
| --- |
| **Book3 <- read\_excel("C:/Users/AYAN/OneDrive/Desktop/project/Book3.xlsx")**  **x<-Book3$`adventure revenue`**  **y<-Book3$`non adventure`**  **res<-t.test(x, y, var.equal=TRUE)**  **res** |

|  |
| --- |
| **output** |

|  |
| --- |
| **Two Sample t-test**  **data: x and y**  **t = 5.7874, df = 282, p-value = 1.9e-08**  **alternative hypothesis: true difference in means is not equal to 0**  **95 percent confidence interval:**  **56.29973 114.33717**  **sample estimates:**  **mean of x mean of y**  **162.15542 76.83697** |

The results of our two-sample t-test indicate a statistically significant difference between the mean revenues of adventure movies and those of non-adventure movies. This is supported by a very small p-value (p =1.9e-08), which is lower than our predetermined significance level of 0.05. Therefore, we reject the null hypothesis that the means of revenue for adventure movies and non-adventure movies are equal.

|  |
| --- |
| **Learnings** |

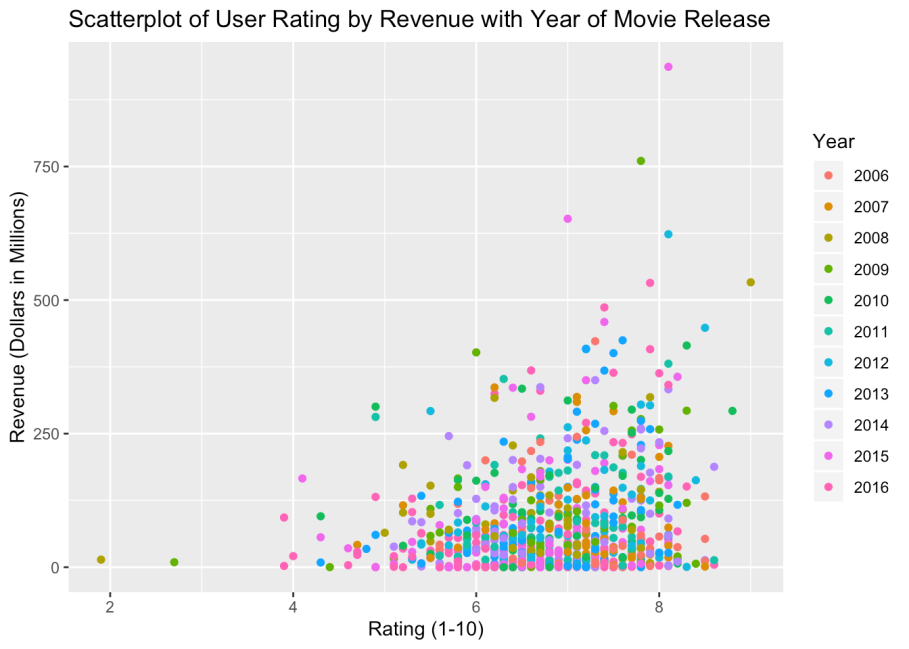
Based on our findings, it can be inferred that adventure movies have a higher gross revenue than all other genres, as both our statistical analysis and graph indicate. It is noteworthy that the production of adventure movies has increased recently, as revealed by our initial graph. Although the causal relationship between these two factors cannot be ascertained, it is generally recognized that an increase in demand for a product, which is reflected in higher revenue, results in a corresponding increase in supply (in this case, the number of adventure movies produced)

**Do well-rated movies, based on critics and users (general public0 reviews earn better revenue?**

Our next area of interest is understanding the correlation between movie ratings and their financial success. When deciding which movie to watch, we often rely on ratings to make a decision. However, the question is whether high ratings necessarily translate to high revenue. Some movies, such as the Twilight Saga, have received poor ratings but still made millions of dollars, prompting us to investigate whether there is a relationship between ratings and revenue.

We will explore whether revenue increases as the public rates the movie higher and what general trends we observe when comparing the two variables.

|  |
| --- |
| **Plots** |



Based on our analysis, there appears to be a positive correlation between higher movie ratings and box office revenue. The dataset suggests that movies with higher ratings tend to earn more revenue. However, this relationship is not entirely clear from the initial chart we created.

To further investigate this relationship, we created additional charts that show the average and median revenues of each rating category, both in whole digits and their original decimal format. We also color-coded the bar graphs based on revenue to make it easier to identify the best-performing movies within each rating category.

These additional charts allow for a clearer visualization of the relationship between ratings and revenue. While it is not a perfectly linear correlation, there is a general trend towards higher revenue for movies with higher ratings.



From this, where we truncated the decimal digits, we can see that the average revenue and the median revenue do actual do increase with the increase of the rating. The graph is left-skewed, with lower ratings indeed producing lower revenues.

These charts made the data much more clear in interpretation than the scatterplot, but they were both valid ways of viewing the data.

|  |
| --- |
| **Statistical tests and analysis** |

Visualizing trends through graphs is useful, but it is also important to verify them with statistical analysis. To investigate the correlation between ratings and revenue, we will perform linear regression.

What is linear regression?

Linear regression is a statistical method that models the relationship between a dependent variable (also known as the response or outcome variable) and one or more independent variables (also known as predictor or explanatory variables) as a linear equation. The aim of linear regression is to find the best-fitting line through the data points in a way that minimizes the distance between the actual data points and the predicted values on the line. Linear regression is commonly used to identify the strength and direction of the relationship between variables, as well as to predict the value of the dependent variable based on the values of the independent variables.

|  |
| --- |
| **Code** |

|  |
| --- |
| **data = read\_xlsx("C:/Users/AYAN/OneDrive/Desktop/project/book1.xlsx")**  **data**  **head(data)**  **y<-data$`Revenue (Millions)`**  **x<-data$Rating**  **v<-lm(formula = y ~x, data = data)**  **print(summary(v))** |

|  |
| --- |
| **Output** |

**Call:**

**lm(formula = y ~ x, data = data)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-126.94 -68.25 -26.35 31.05 818.83**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) -91.60 27.62 -3.316 0.000951 \*\*\***

**x 25.85 4.02 6.431 2.13e-10 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 102.1 on 836 degrees of freedom**

**Multiple R-squared: 0.04713, Adjusted R-squared: 0.046**

**F-statistic: 41.35 on 1 and 836 DF, p-value: 2.134e-10**

Based on our statistical analysis using linear regression, we found that both the p-values of the coefficients and the f-statistic are less than our chosen significance level of 0.05. This leads us to reject the null hypothesis, which assumes that there is no significant relationship between the rating and the revenue (in millions). Instead, we can conclude that there is a relationship between these two variables.

An alternative approach to explore the relationship between ratings and revenue is to test the averages of ratings against each other. As ratings can be considered as an ordinal categorical grouping, we can use the ANOVA test to analyze variance and determine if there is a significant difference in mean revenues between the different rating categories. This would provide us with another perspective to examine the relationship between ratings and revenue.

What is ANOVA?

ANOVA stands for Analysis of Variance, and it is a statistical method used to compare means between two or more groups. It assesses whether the means of different groups are statistically significantly different from each other. The assumption of ANOVA is that the data is normally distributed and that the variances across the groups are equal. Another assumption is that the observations within each group are independent.

Assumptions of ANOVA

The three assumptions of ANOVA are:

1. Independence of observations: This means that the observations are independent of each other. In other words, the values of one observation do not affect the values of any other observation.

2. Homogeneity of variance: This means that the variances of the populations from which the samples are drawn are equal.

3. Normality: This means that the data are normally distributed within each group.

|  |
| --- |
| **Code** |

|  |
| --- |
| **data = read\_xlsx("C:/Users/AYAN/OneDrive/Desktop/project/book1.xlsx")**  **data**  **head(data)**  **y<-data$`Revenue (Millions)`**  **x<-data$Rating**  **v<-lm(formula = y ~x, data = data)**  **print(summary(v))**  **anova(v)** |

|  |
| --- |
| **Output** |

**Df Sum Sq Mean Sq F value Pr(>F)**

**as.factor(Rating) 49 998397 20375 1.971 0.000123 \*\*\***

**Residuals 788 8145391 10337**

**---**

**Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1**

Based on the results of the ANOVA test, we can conclude that the null hypothesis of equal average revenues across all rating groups is rejected. The p-value is less than 0.05, indicating that there is a statistically significant difference in revenues between at least one of the rating groups. Therefore, we can infer that there exists a relationship between the rating of a movie and its revenue(million).

|  |
| --- |
| **Learnings** |

The correlation between user ratings and movie revenue seems logical, given the wealth of available data at our fingertips. With review aggregation websites like Rotten Tomatoes, it's easy to get a sense of what others are thinking about a particular film. Similarly, social media platforms such as Facebook and Instagram can provide insight into how our friends are responding to new releases. These sources of information heavily influence our decision-making when it comes to going to the movies, and it's likely that this is why there is a correlation between user ratings and revenue. Essentially, positive user ratings drive us to the theaters, while negative ones may deter us. By relying on user ratings, we can make more informed choices about which movies to spend our money on.

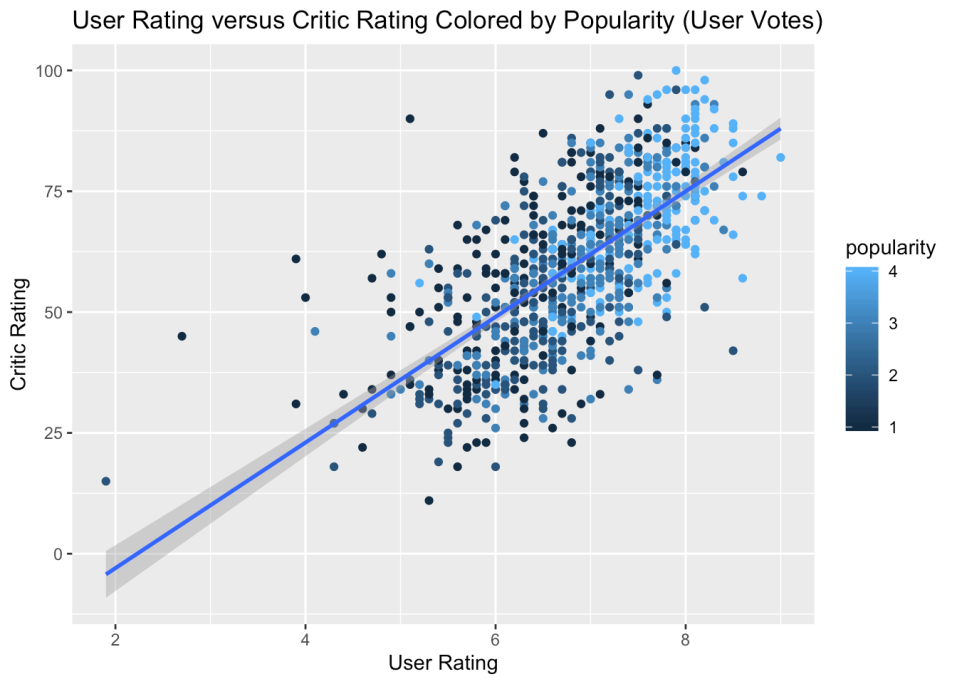
**Is there a consensus between critics and the general public in their movie ratings and what makes a good movie?**

In order to better understand the relationship between user ratings and critic ratings, we will examine their correlation. Film criticism is a profession where film critics analyze and evaluate movies based on various elements, such as themes, motifs, acting, and plot. In contrast, user ratings are typically based on personal feelings and do not always delve into the intricate details of a film. However, this doesn't necessarily mean that user reviews are unable to predict a good movie, or that movie critics are always correct in their reviews and the response of the general public to a movie's success. We will explore these questions further in the following plots.

|  |
| --- |
| **Plot** |

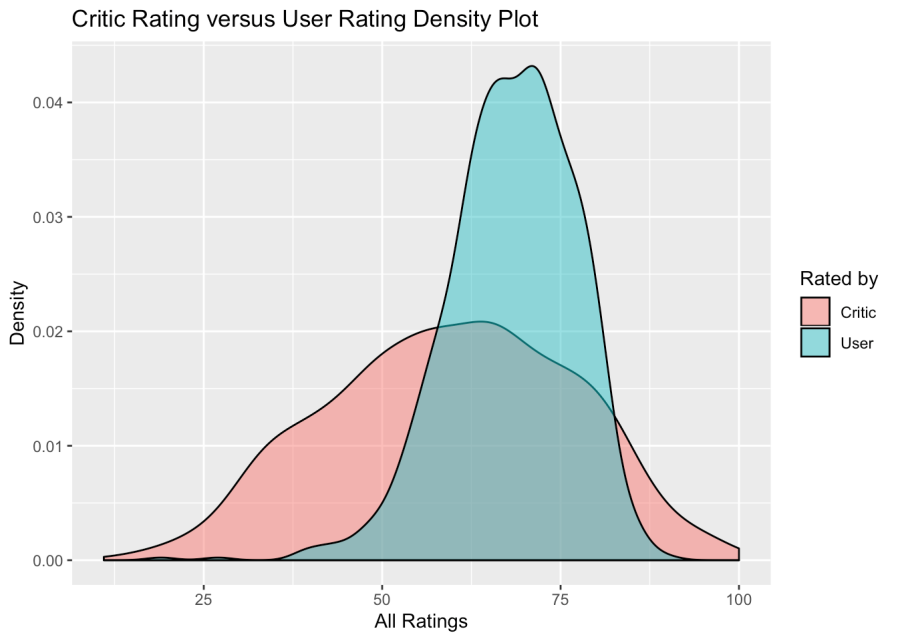
Before conducting any statistical analysis, it is important to understand the variables we will be working with. In this case, we will be analyzing the relationship between critic ratings (Metascore) and user ratings (Rating) of movies. The Metascore ranges from 0 to 100 while the user Rating ranges from 0 to 10. However, to compare the two variables, we need to standardize their ranges.

Furthermore, to better visualize the popularity of movies based on user ratings, we can define user popularity as the number of votes a movie received from users. To create a more meaningful display, we will group movies into five tiers based on their popularity, using quantiles. Additionally, to address the issue of overplotting, we will employ a smoothing method, such as a straight line, to better visualize patterns in the data.



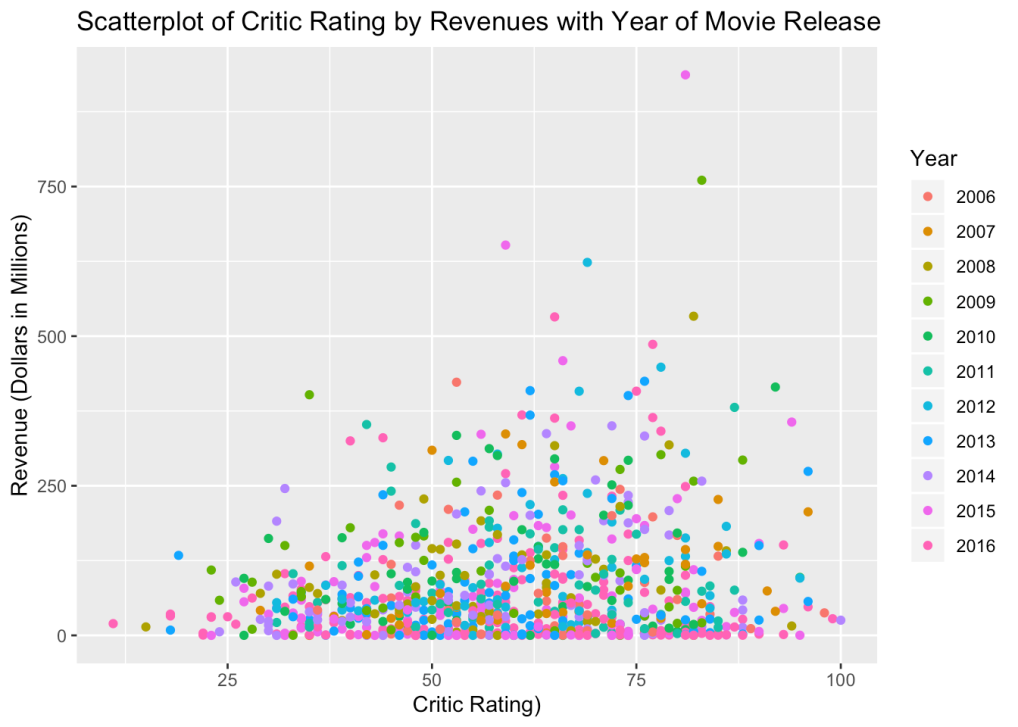
The graph shows a general agreement between the ratings of users and critics in terms of identifying good and bad movies, with only a few exceptions. Furthermore, it seems that movies with higher ratings tend to receive more ratings from users.

To better compare the two rating systems, we need to convert the user ratings to the same 0 to 100 scale as the Metascores.



The distribution of user ratings is relatively more uniform, while critic ratings tend to have more extreme values. Critic ratings show a wider range of scores with a flatter curve, whereas user ratings are more concentrated around the middle. The modes of the two distributions are different, with critics having a mode in the 60s and users in the 70s. Overall, both distributions are fairly normal.

To further investigate the correlation between movie ratings and revenue, we created a scatter plot of critic ratings (Metascore) against movie revenue by year. The color-coded points allowed us to visualize which years critics accurately predicted movie success, and which years they were off the mark. This information can provide insights into the performance of movie critics over time and their ability to predict the success of movies.



The scatter plot comparing critic ratings and revenue by year reveals that a high Metascore does not always translate into high revenue. Interestingly, some movies with poor Metascores generated higher revenue than those with excellent ratings. The plot also shows that the most successful movies were released in recent years, particularly in 2016. However, as we move towards the late 2010s, the relationship between Metascore and revenue became more accurate, indicating that critics were able to better predict the public's opinion of a movie's success during those years. This plot provides valuable insights into the performance of movie critics over time and their ability to gauge movie success.

|  |
| --- |
| Statistical Tests and Analysis |

To assess whether the distribution of critic ratings and user ratings are similar, we conducted a two-sample Kolmogorov-Smirnov test. This statistical test was chosen due to its ability to compare two samples in a non-parametric manner and its sensitivity to differences in both location and shape of the empirical cumulative distribution functions of the two groups.

What is Kolmogorov-Smirnov test?

The Kolmogorov-Smirnov test is a statistical test used to determine whether two sets of data are from the same distribution or if they differ significantly. It is a non-parametric test, meaning that it does not rely on any assumptions about the underlying distribution of the data. The test works by comparing the cumulative distribution functions (CDFs) of the two datasets and calculating the maximum difference between them. The test produces a p-value that represents the probability of obtaining such a difference if the two datasets are actually from the same distribution. If the p-value is small enough (typically less than 0.05), the null hypothesis (i.e., the two datasets are from the same distribution) is rejected in favor of the alternative hypothesis (i.e., the two datasets are from different distributions).

|  |
| --- |
| **Code** |

|  |
| --- |
| **y <- read\_excel("C:/Users/AYAN/OneDrive/Desktop/project/Book1.xlsx")**  **y**  **x<-y$Rating**  **z<-y$Metascore**  **ks.test(x,z)** |

|  |
| --- |
| **Output** |

|  |
| --- |
| **Two-sample Kolmogorov-Smirnov test**  **data: x and z**  **D = 1, p-value < 2.2e-16**  **alternative hypothesis: two-sided** |

The results of the Kolmogorov-Smirnov test indicate that the distributions of critic ratings and user ratings are significantly different. The p-value obtained (p-value < 2.2e-16) is much smaller than our chosen level of significance (0.05), which led us to reject the null hypothesis that the two distributions are the same. This implies that there are significant differences in the distribution of critic ratings and user ratings, which cannot be attributed to chance alone.

Next, we use a t-test to determine if the average critic rating is different from the average user rating.

|  |
| --- |
| **Code** |

|  |
| --- |
| **y<- read\_excel("C:/Users/AYAN/OneDrive/Desktop/project/Book.xlsx")**  **x<-y$`Metascore`**  **z<-y$`Rating`**  **res<-t.test(x, z, var.equal=TRUE)**  **res** |

|  |
| --- |
| **Output** |

|  |
| --- |
| **Two Sample t-test**  **data: movies$Metascore and userratings**  **t = -12.993, df = 1255.7, p-value < 2.2e-16**  **alternative hypothesis: true difference in means is not equal to 0**  **95 percent confidence interval:**  **-9.861773 -7.274265**  **sample estimates:**  **mean of x mean of y**  **59.57518 68.14320** |

Based on the results of our test, we can conclude that there is a significant difference between the average values of critic ratings and user ratings. The 95% confidence interval, which ranges from -9.495726 to -6.998189, is entirely negative, indicating that the true difference in averages is not zero. Furthermore, since the interval does not contain 0, it suggests that critics generally rate movies lower than users do.

|  |
| --- |
| **Learnings** |

Although the scatterplot of critic and user ratings suggests a positive trend, indicating that the groups rate movies similarly, it does not present the complete picture. Upon examining the density plot, it becomes apparent that the distributions of critic and user ratings are dissimilar. This is confirmed by the Kolmogorov-Smirnov test, which also indicates a significant difference in how critics and users rate movies. Furthermore, the scatterplot comparing critic ratings and revenue demonstrates that critic ratings do not always align with public opinion, as several poorly rated movies still managed to generate substantial revenue. This reinforces the conclusions drawn from the density plot. The revenue a movie generates serves as an indirect user rating system, as people would not continue to pay money to watch a movie that did not receive positive feedback.

|  |
| --- |
| **Conclusion** |

After conducting our research, we have gained valuable insights into the most popular IMDb movies from 2010 to 2016. The data has revealed that adventure and action movies have consistently generated the highest revenue, whereas lighter genres such as comedy, romance, and drama tend to perform better on DVD or streaming platforms. This finding aligns with the general notion that adventure and action movies are best enjoyed in theaters, providing a more immersive and larger-than-life experience for the audience.

Our analysis of the relationship between user and critic ratings has also shed light on the importance of public opinion in determining a movie's success. The positive correlation between user ratings and revenue highlights the fact that a movie's popularity ultimately drives its financial success. Critic ratings, on the other hand, do not necessarily align with public opinion, as we observed instances where poorly-reviewed movies outperformed highly-rated ones in terms of revenue.

Furthermore, our density plot and statistical tests confirmed that there is a significant difference in how critics and users rate movies. This suggests that the role of critics in shaping public opinion and driving movie revenue may need to be reevaluated. Ultimately, our findings emphasize the importance of considering public opinion when producing and marketing movies, as the general public's approval plays a critical role in a movie's financial success.