Jackson Schmidt

Tuesday, December 17th

Data Wrangling

Project Report

Trends in IMDb Top 250 Movies

1. Introduction

The film industry is a dynamic and ever-evolving field, with new trends emerging each year. Understanding these trends can provide valuable insights for filmmakers, critics, and audiences alike. This project aims to explore the trends within the IMDb Top 250 movies, focusing on factors such as ratings, genres, directors, and release years.

Initially, my idea with this project was to analyze movies released in 2024 using data from Rotten Tomatoes and a Kaggle dataset. However, due to challenges in scraping data from Rotten Tomatoes, I have to shift my focus to the IMDb Top 250 list. Although this dataset forced me to steer away from my original ideas with this project, it still allowed for a comprehensive analysis about the top 250 rated movies.

By examining this dataset, I aimed to uncover patterns and correlations that can shed light on what makes a movie successful and critically acclaimed. This analysis will help to highlight the characteristics of top-rated movies and directors.

2. Data

This project uses data from the IMDb Top 250 Movies dataset available on Kaggle. https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset

2.1 **IMDb Top 250 Movies**

The dataset contains information about the top 250 movies as rated by IMDb users. The data includes movie titles, release years, genres, directors, and IMDb ratings.

2.2 **Data Collection Challenges**

Initially, the plan was to scrape data from Rotten Tomatoes to complement the Kaggle dataset. However, the dynamic nature of the Rotten Tomatoes website made it difficult to scrape data using Selenium and Beautiful Soup. After multiple attempts, I decided that it would be best to switch to the IMDb Top 250 dataset, which was successfully scraped using Selenium.

2.3 **Data Merging Issues**

While attempting to merge the IMDb data with other datasets, it was found that there were not enough movies for a comprehensive analysis. This led to the decision to find a more suitable dataset on Kaggle that could be effectively joined with the IMDb data. The second dataset I used contained the top 250 IMDb movies from 2021. The list had changed slightly since 2021, resulting in only 240 usable movies, but this number was still much higher than the original dataset making for a more comprehensive analysis.

3. Analysis

In my analysis, I aimed to uncover correlations between various attributes using scatterplots. Specifically, I explored the relationship between IMDb ratings and worldwide box office sales, clustered by different certificates. This helped identify patterns in how movie ratings and certifications impact box office performance.

Additionally, I examined the influence of directors by identifying those with the most movies rated 9 or higher, as well as those with the most movies rated 8 or higher. This provided insights into which directors consistently produce highly rated films.

Furthermore, I analyzed the relationship between movie budgets and worldwide box office sales to understand how production investments correlate with financial success. This comprehensive approach allowed me to identify key factors that contribute to a movie's success both critically and commercially. With these questions in mind, this allowed me to create a comprehensive analysis using the following charts below.

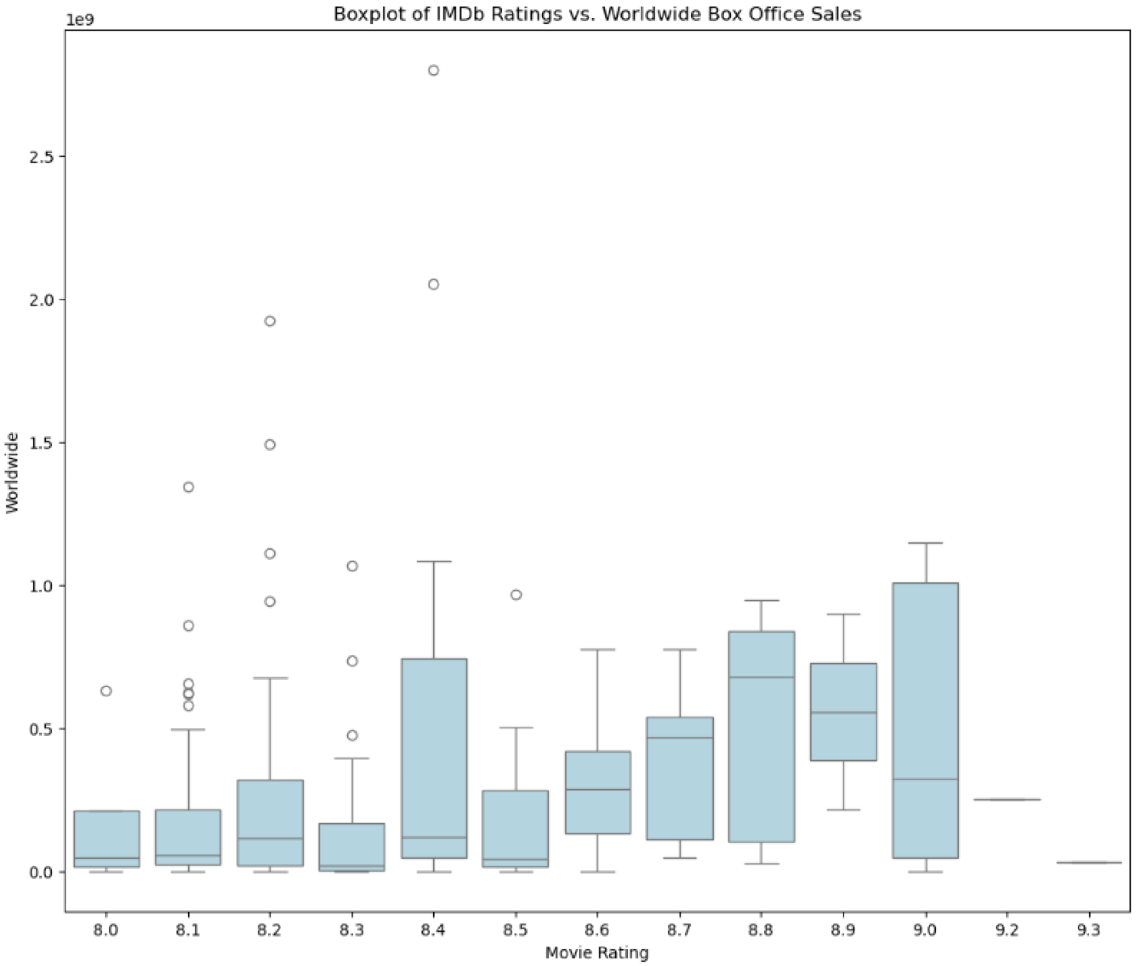


Figure 1: Boxplot of Ratings vs. Box Office

The boxplot illustrates a positive correlation between IMDb ratings and worldwide box office sales. As ratings increase, the central tendency and upper range of sales generally shift upwards, suggesting that higher-rated movies tend to generate more revenue. However, the spread of the boxes and presence of outliers indicate significant variability within each rating category. This suggests that while good reviews may increase the likelihood of box office success, other factors play a crucial role in determining a film's financial performance.

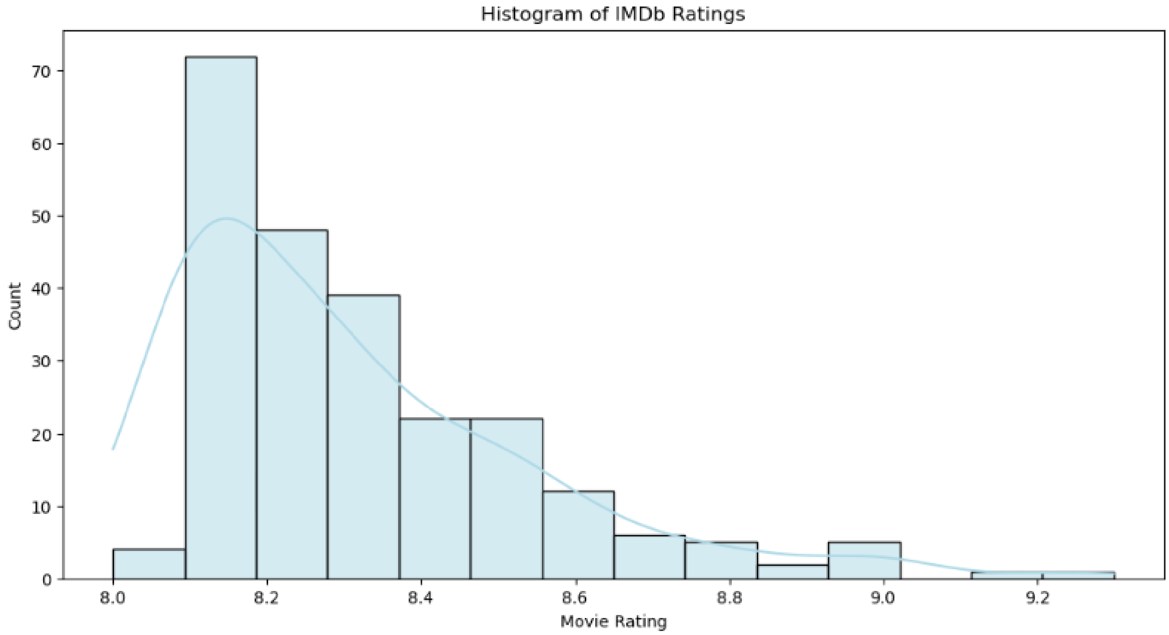


Figure 2: Histogram of IMDb Ratings

This histogram shows a right-skewed distribution, but with a significant difference. If we were looking at a large dataset with movies consisting of multiple ratings, this histogram would look a little different. However, the right-skewness suggests that even among the top 250, there's a concentration of movies with ratings in the higher-average range with fewer movies reaching the peak of critical acclaim. This implies that while these are top-rated movies, achieving a truly exceptional rating is still a rare feat.

In essence, the right-skewed distribution within the top 250 suggests that while these movies are highly regarded, there's a subtle yet significant differentiation in critical reception even among the elite.

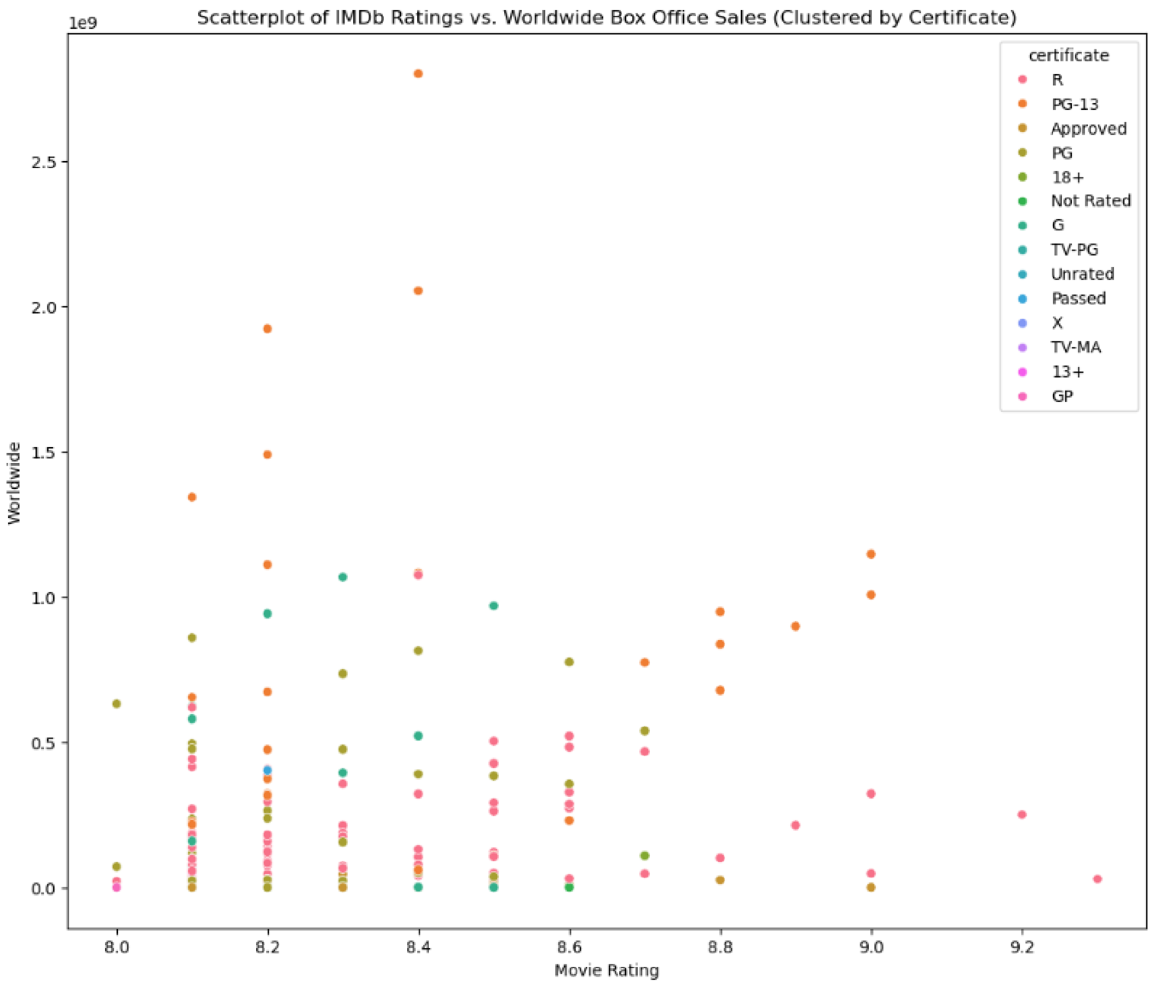


Figure 3: Scatterplot of Ratings vs. Box Office

The scatterplot illustrates the relationship between IMDb ratings and worldwide box office sales, with data points clustered by film certificate. A general positive trend is observed, suggesting higher-rated movies tend to generate more revenue. However, the spread of data points indicates significant variability within each rating category. For instance, some movies with high ratings have relatively low sales, while others with lower ratings achieve substantial box office success.

Furthermore, the clustering by certificate reveals potential influences of age restrictions on box office performance. Certain certificates seem to be associated with a particular range of sales figures. For example, movies with a "PG" certificate appear to have a wider range of box office performance compared to movies with an "R" certificate. This suggests that age restrictions might play a role in shaping a movie's audience and, consequently, its box office potential. I also noticed that two of the highest rated movies are “R” certificate movies, suggesting that these “R” certified films, have a higher chance of receiving a higher IMDb rating.

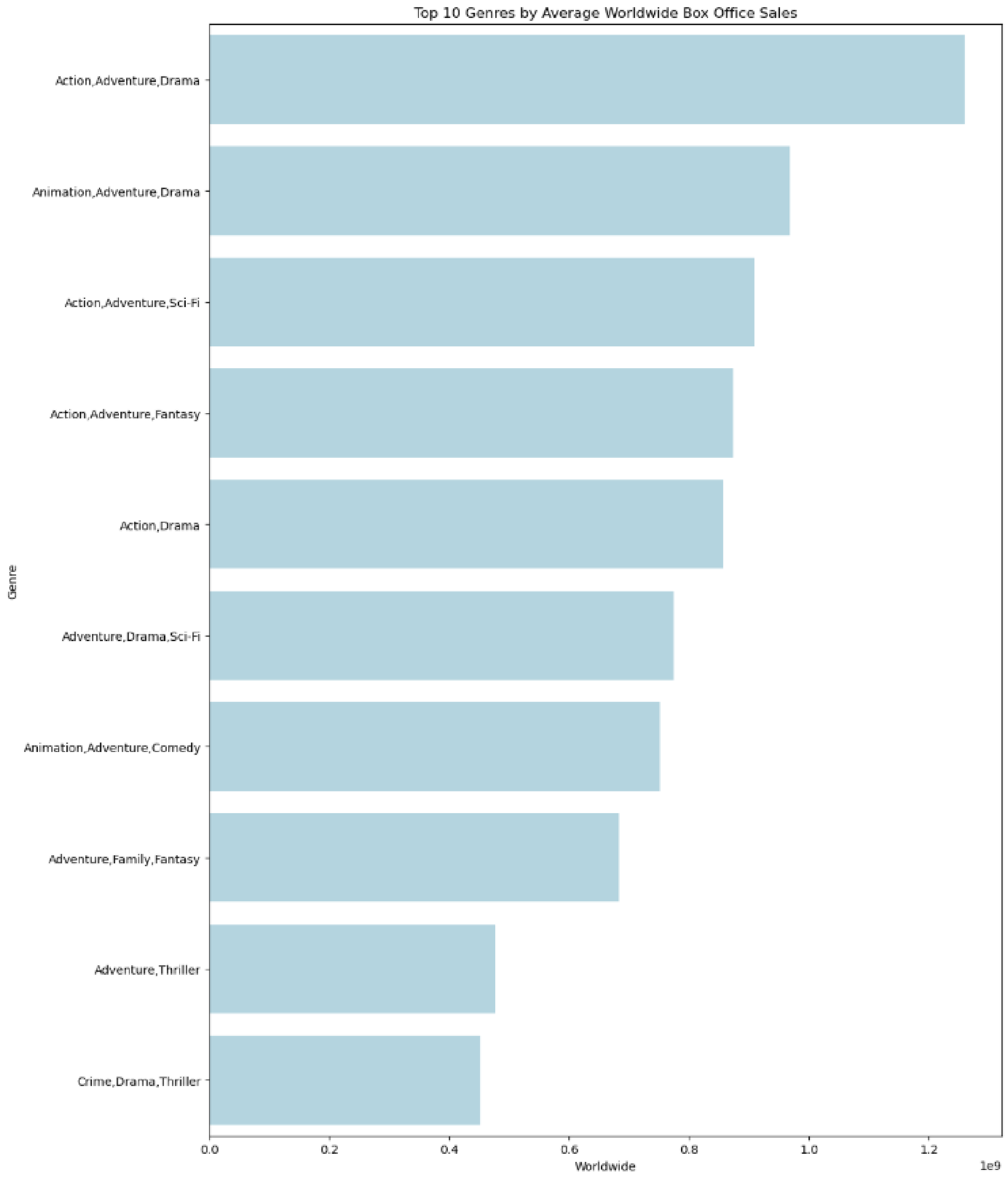


Figure 4: Horizontal Bar Chart of Top Genres

This bar chart presents an overview of the top 10 genres by average worldwide box office sales. "Action, Adventure, and Drama" emerges as the top-grossing genre, showcasing its strong commercial appeal. The varying lengths of the bars underscore the significant disparity in average box office revenue across different genres. This suggests that audience preferences and market trends heavily influence the financial success of movies within specific genres.

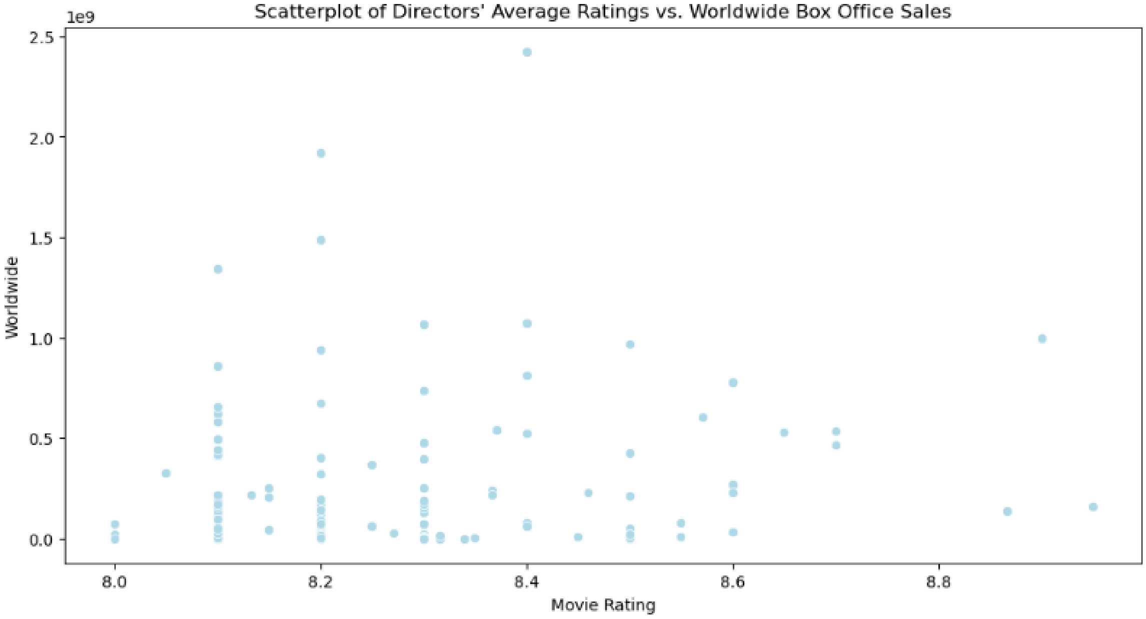
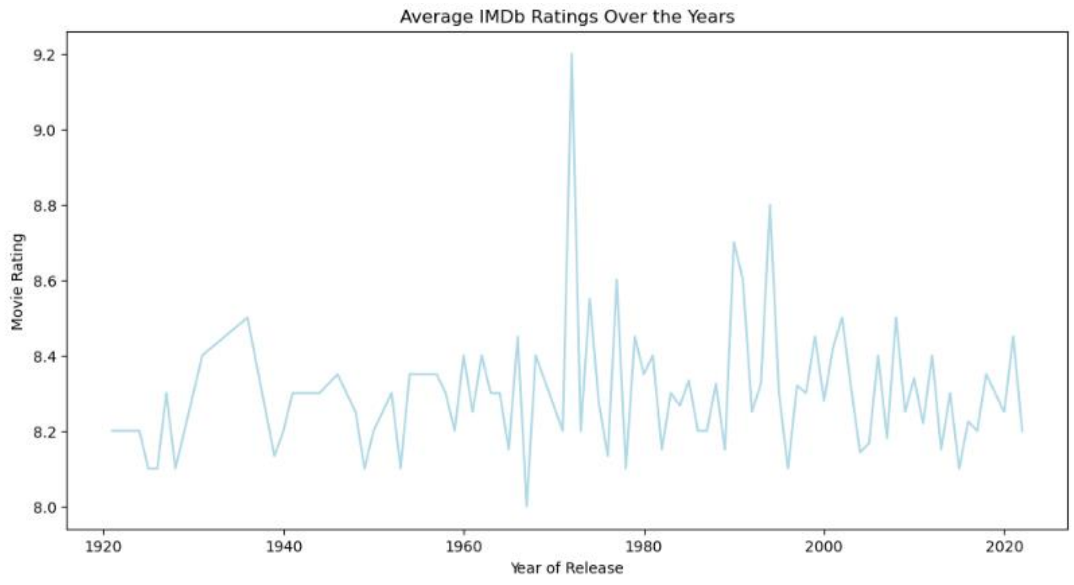


Figure 5: Scatterplot of Movie Rating vs. Box Office

The scatterplot explores the relationship between directors' average ratings and worldwide box office sales. While there is a slight positive trend visible, suggesting that directors with higher average ratings may tend to have higher-grossing movies, the relationship is not strong. A wide scatter of data points indicates that there is some variability when it comes to measuring these two attributes. Many directors with high average ratings have decent box office performance, and on the other hand, some directors with lower average ratings have achieved blockbuster success.

This suggests that a director's average rating is not the sole determinant of a movie's box office performance. Other factors, such as genre, star power, marketing campaigns, and overall industry trends, likely play a more significant role in determining a movie's commercial success.



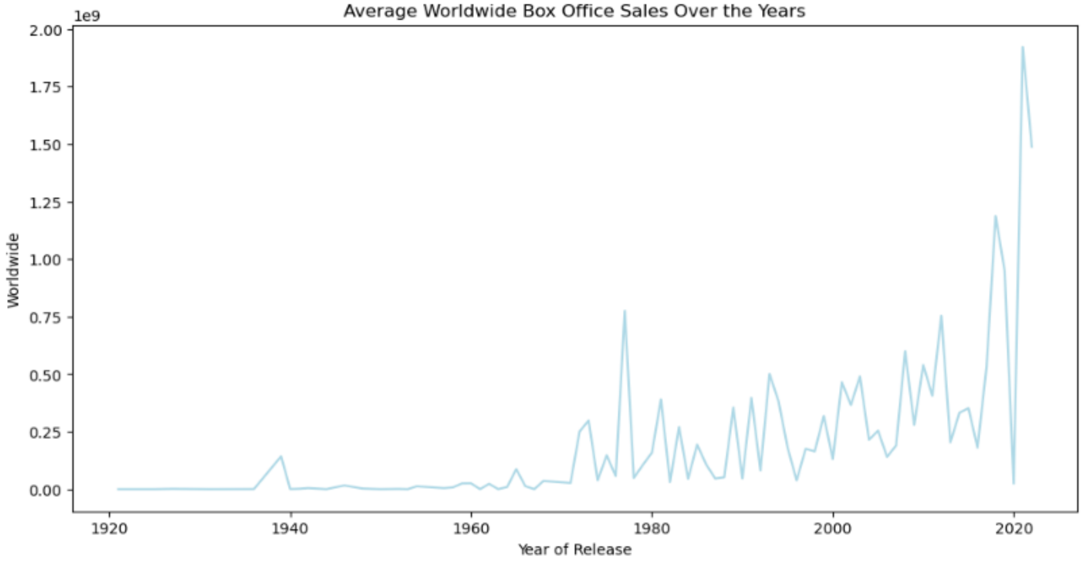


Figure 6: Line Charts of Box Office and IMDB Ratings Over Time

The line graphs above both use "Year of Release" on the x-axis, but the y-axis alternates between Box Office sales and IMDb ratings. When analyzing these two charts, it is interesting to note that there does not seem to be a direct correlation between worldwide box office revenue and IMDb ratings. There appears to be some correlation before 1980, but after that, the relationship weakens significantly. This could suggest that, over time, IMDb reviews and box office performance are not always aligned. One possible explanation is that audiences may enjoy movies more than critics do, which would increase box office revenue, while not increasing IMDb rating. Additionally, audiences might not be as informed about what makes a movie critically good compared to critics, which could also be a contributing factor.

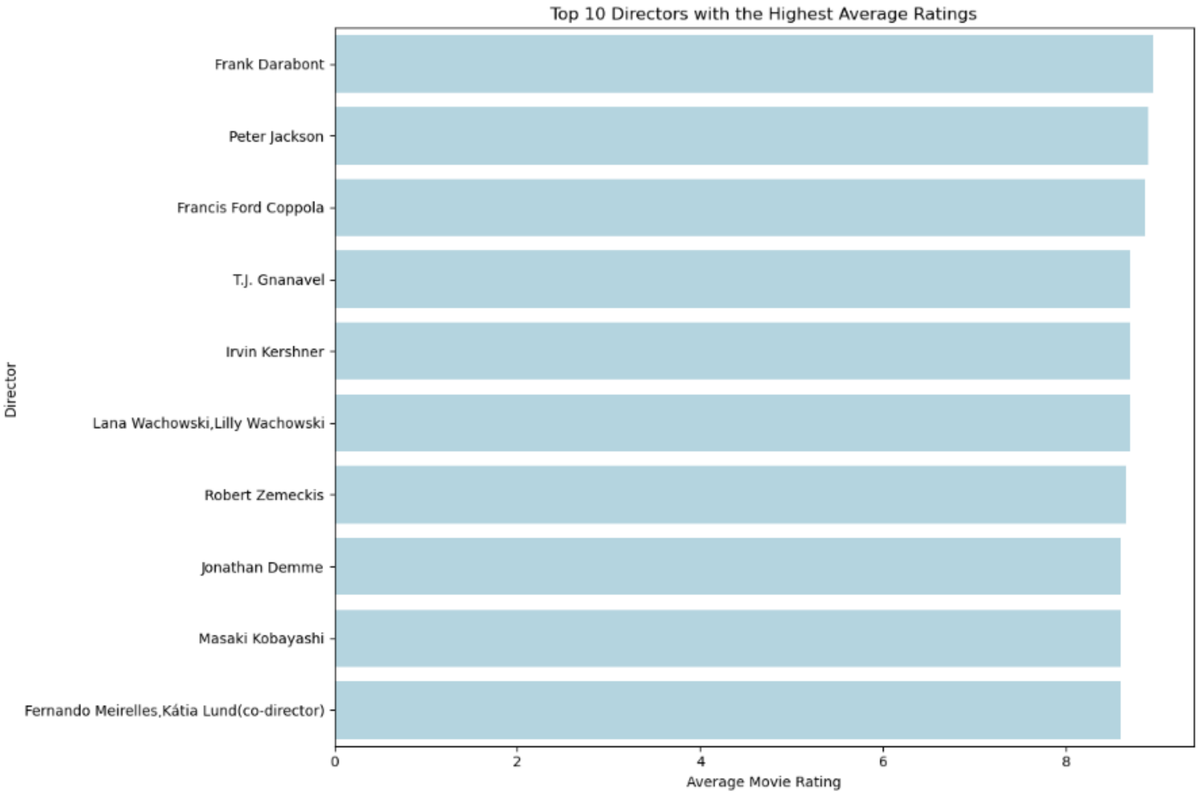


Figure 7: Top 10 Directors with the Highest Ratings

This chart presents a ranking of directors based on the average ratings of their movies. Each bar represents a director, and the length of the bar corresponds to their average movie rating. The chart reveals that Frank Darabont leads the list with the highest average rating, followed by Peter Jackson and Francis Ford Coppola. Notably, the top 10 directors demonstrate a relatively high range of average ratings, suggesting a high level of critical acclaim among this group.

This chart shows how well-regarded films by these famous directors are. It highlights their consistent ability to produce high-quality, acclaimed movies, confirming their status as top figures in the film industry.

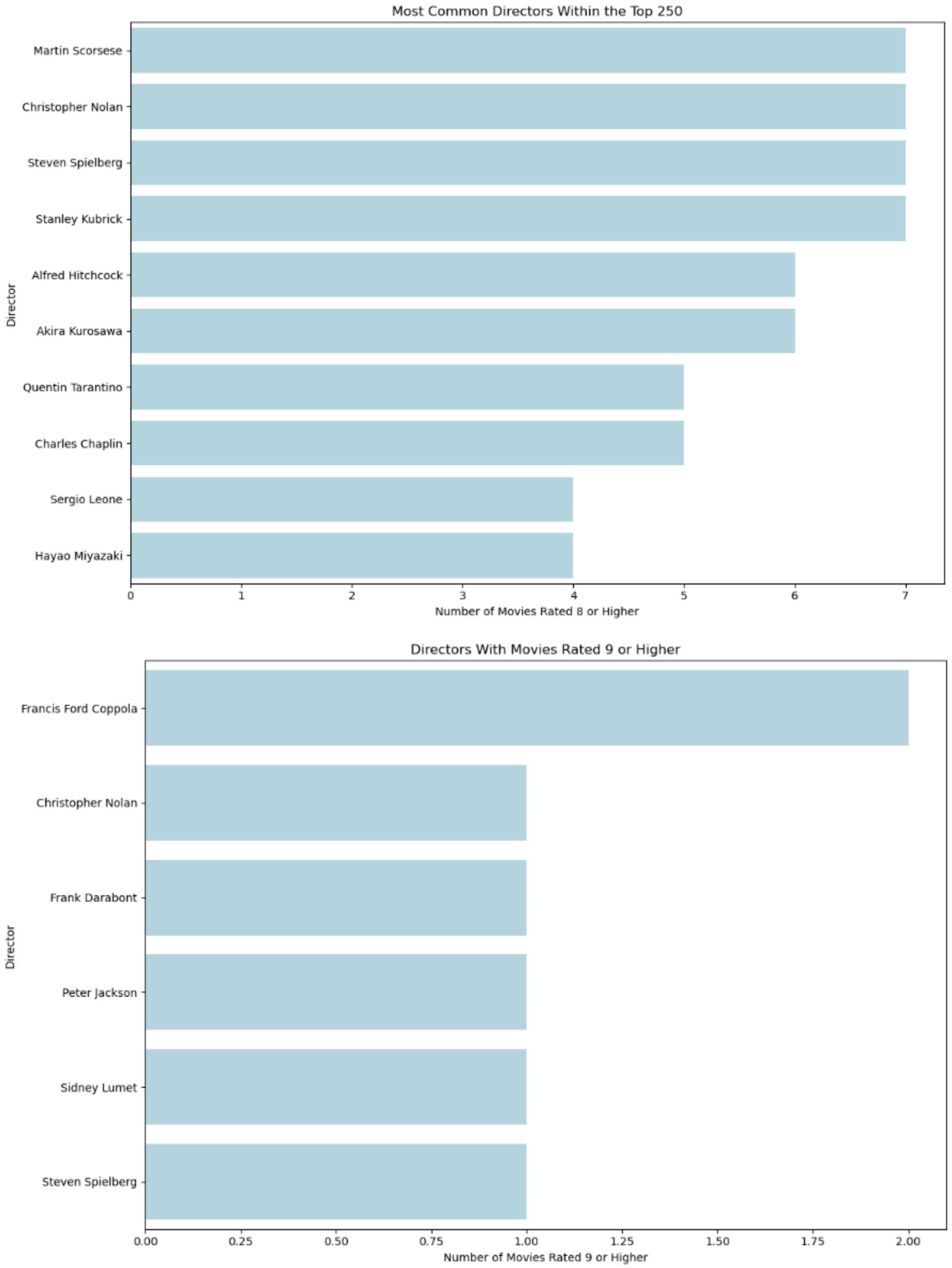


Figure 8: Horizontal Bar Charts with Directors with High Rated Movies

When creating these two charts, my initial analysis focused on directors with movies rated 9 or higher, aiming to identify those who consistently deliver exceptional, critically acclaimed films. This approach provided valuable initial insights into the directors who excel at achieving the highest ratings amongst critics. However, by expanding the analysis to include directors with movies rated 8 or higher, it allowed for me to gain a more comprehensive understanding of the directors who consistently produce highly rated films

Including this broader range of highly rated films likely revealed a wider pool of directors who consistently deliver high-quality films that resonate with critics. This is evident in the change in results between the two bar charts. For example, Martin Scorsese leads the pack when it comes to movies rated eight or higher but is absent from the bar chart of directors with movies rated nine or higher. Analyzing this relationship identifies that even amongst the best of directors, there are only a select few who consistently direct almost perfect films.

Analyzing directors with movies rated 9 or higher focuses on the very best films. Including movies rated 8 or higher covers more directors who consistently make good films. Comparing these two groups helps us understand different levels of critical success and identify top directors at each level.

4. Conclusion

In this project, I analyzed several aspects influencing movie box office success: director impact, genre trends, and the relationship between IMDb ratings and box office sales.

After Completing this project, I found the following results:

1. **Is there a strong correlation between a movie's review score and its box office sales?**

There is a positive correlation between IMDb ratings and box office sales, but the relationship is not linear. Higher-rated movies tend to generate more revenue, but significant variability exists within each rating category.

2. **How do genre trends impact box office success?**

Certain genres, such as "Action, Adventure, and Drama," consistently demonstrate higher average box office sales, suggesting that audience preferences and market trends significantly influence the commercial success of movies within specific genres.

3. **What is the impact of director influence on box office success?**

Directors with consistently high average ratings, such as Frank Darabont and Peter Jackson, tend to deliver critically acclaimed films. However, a director's average rating is not the sole determinant of box office success. Other factors, such as genre, star power, and marketing, significantly influence commercial performance.

Project Limitation's:

This project has a few limitations, like possible sampling bias in the dataset and the subjective nature of genre classifications.

In the future, it would be interesting to look at how different marketing strategies affect box office performance, analyze the link between production costs and box office returns, and see how genre trends have evolved over time.

This analysis lays the groundwork for understanding what makes films successful. By digging deeper into these relationships and considering more variables, we can get a better grasp of the complex dynamics in the film industry.