Data Wrangling - Project Report

January 2024

1 Project Group Members

1. Rafaëlla van Nee, rne330
2. Hari Joshithaa Aghilah Senthilprathiban, han980
3. Mara-Iuliana Dragomir, mdr317
4. Viktoriya Li, vli210

2 Research Question

Movies have been a main source of entertainment for long periods of time. With improvements in technology, the production of different types of movies have arisen. Thus, we decided to analyse the prevalence and popularity of different movie genres over time in the film industry. To do this, we have the following main research question:

"How have the prevalence and popularity of different movie genres evolved over time in the film industry?"

To answer our main research question, we have the following subquestions:

1. What does the distribution of movie genres look like in general, and how did it change over time? Which genres were more prevalent in the movie industry in each decade?
2. How have audience and critic ratings for specific movie genres (such as action, drama, comedy) evolved over the decades? Are certain genres experiencing an increase or decrease in popularity?
3. Is there a correlation between movie genres and the revenue generated, and if so, which genres tend to exhibit a stronger correlation with higher revenues?
4. Is there a correlation between a movie's genre and its likelihood of being part of a franchise?

3 Data Sources

In this paper, we use the "Movies Dataset" available on Kaggle at: `https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=movies_metadata.csv`. The entire dataset consists of several CSV files such as movies_metadata.csv, ratings.csv, links.csv, credits.csv, and keywords.csv. To answer our research question(s) we only use the movies_metadata.csv and ratings.csv files. The last access date for the dataset is 19th January 2023.

4 Data Wrangling Methods

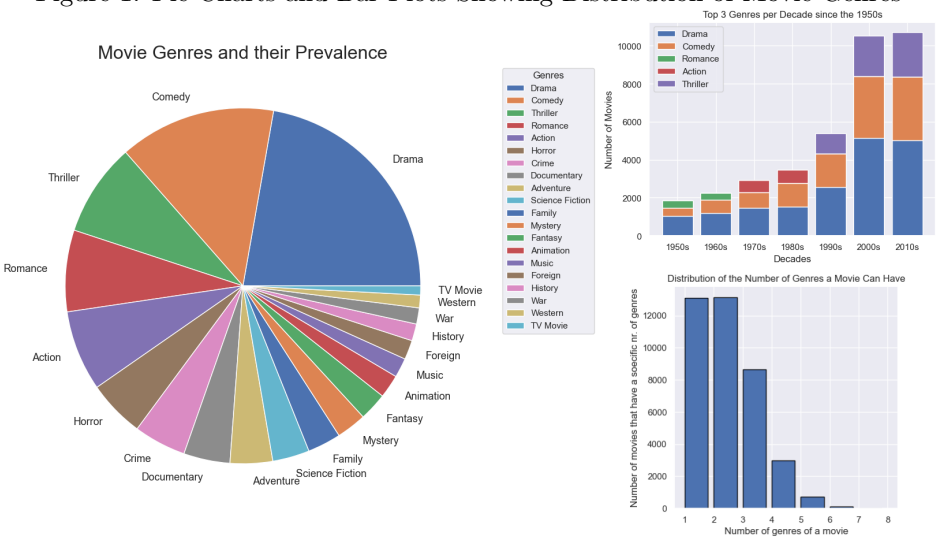
The movies_metadata.csv file, which was used for all subquestions, has 45466 rows and 24 columns, where entries are of diverse data types: text, numbers, boolean, JSON-format and URLs. It is mostly a clean dataset, but there were some missing values and duplicates that needed to be removed; regarding missing values, there were some acceptable ones, as a few columns were optional, but the rest of 1418 rows were removed; and duplicates were considered by the movie's ID, and they were 30. Afterwards, we needed to extract relevant data for our work. We did some operations first as they were relevant for all questions, namely: extracting the names of the genres of each movie from the JSON-formatted entries, and filtering

the movies to only include released movies. Then, we had to create separate copies of the dataset to perform different extractions necessary for each subquestion. Each subquestion selected a different subset of columns from the dataset, but some were required to extract specific data from some columns, such as the decade from the 'release_date' column, necessary for the first and second questions, and a boolean column called franchise, based on 'belongs_to_collection', which was used in the last subquestion. Moreover, the third subquestion merged the main dataset with another file, 'ratings.csv': it has 26024289 rows and 4 columns. Each column presents data in varied formats, including numeric identifiers for users and movies, numeric values for ratings, and timestamps indicating when each rating was recorded. The dataset is clean and does not have any missing values and duplicate rows. The movieId column from the ratings.csv was aligned with the id column in the movies_metadata.csv, ensuring that each rating was correctly associated with the corresponding movie. The resulting merged dataset, merged_df, became the foundation for further subquestion in our analysis.

4.1 RQ1: Distribution of Genres in the Dataset

Now, after performing data wrangling operations, we can analyse the outcomes in order to answer our research questions. First of all, we are interested in the distribution of genres over movies in the dataset, which is essential for answering our research question. We chose 3 methods to measure this: the number of movies made with each genre; a timeline of top 3 most common genres by number of movies in each decade; and the number of genres that each movie can have. Each measurement gives us unique insight into the movie industry, therefore we used them to generate 3 plots for data visualization: a pie chart, a stacked bar plot and a histogram, which can be seen below. The pie chart shows us how common each genre is in the movie dataset. We can notice the dataset has a diverse palette of genres, though there are a few visible peaks: the two most common genres are Drama and Comedy, followed by Thriller, Romance and Action, then most other genres have a smaller proportion, but are relatively close to each other, and the least used genre is TV Movie. The stacked bar chart demonstrates how the movie industry has evolved over the years, since the 1950s until the 2010s. It is notable that in each decade from this range, Drama and Comedy are always the two most featured movie genres, in this order, while the third most common genre changes approximately every 20 years. And the histogram on the right illustrates the distribution of the number of genres a movie can have. As one can expect, there are significantly more movies with just one or two genres, while more complex genre combinations are less common.

Figure 1: Pie Charts and Bar Plots Showing Distribution of Movie Genres

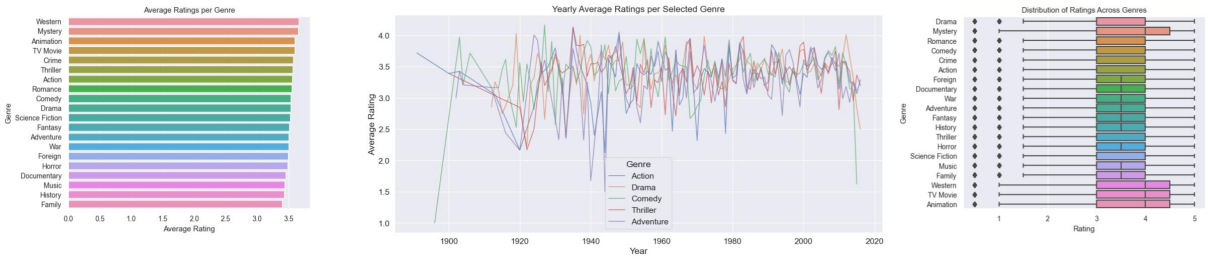


4.2 RQ2: Association between Genres and Rating

We obtain a greater understanding of audience preferences and trends in movie genres by using both the "movies_metadata.csv" and "rating.csv" datasets in our research. Rating.csv contains ratings and timestamps. Through the process of merging, user ratings and movie metadata were linked together using an

identifier.

Figure 2: Bar Charts, Box Plots and Line Graphs Showing Relation between Ratings and Genres

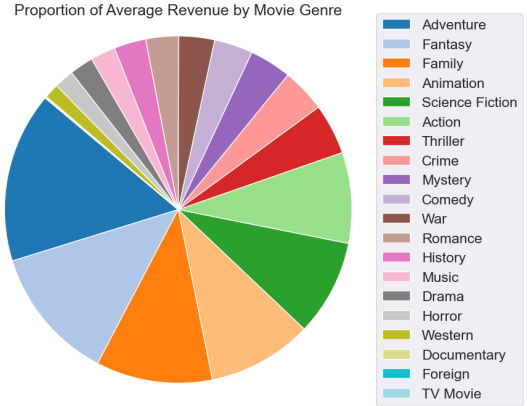


The line chart illustrates the yearly average ratings for various movie genres over an extended period. It provides a visual representation of how audience perceptions of different genres have evolved over the years. The chart includes a multitude of genres, resulting in a complex and detailed visualization. It is observed that there is significant fluctuation in ratings across genres and over time, indicating changing audience tastes and the impact of individual movies. The average ratings for each genre are shown in a bar chart, which highlights an interesting trend: no specific movie genre significantly outperforms or underperforms the others on average. This finding suggests a degree of consistency in viewer evaluations, suggesting that while individual tastes in genre may differ, people generally regard diverse genres of films to be of similar quality and enjoyment. The similarity in average scores suggests that each genre has devoted supporters and admirers. Whether it's the humor and lightheartedness of comedies, the tension and excitement of thrillers, or the thought-provoking quality of documentaries, each genre appeals to different interests and preferences while maintaining a level of overall pleasure. The boxplot shown illustrates the way ratings are distributed throughout different movie genres and gives an overview of how ratings are distributed within each genre. Each genre in the graphic is depicted by a horizontal box that displays the ratings' interquartile range and a line that indicates the median rating inside the box. An interesting aspect of this distribution is that some genres have a more dense distribution with ratings centred around the median, whilst other genres show a wide range of ratings, including many outliers. This may indicate different degrees of agreement among viewers on the level of films in each category.

4.3 RQ3: Relationship between Revenue Generated and Genres

Understanding the correlation between movie genres and their respective revenues is crucial for the film industry. By focusing on the correlation between movie genres and their respective revenues, we gain insight into the commercial aspects of genre popularity. As seen in the pie chart genres such as Adventure, Fantasy, and Family emerged at the top, suggesting a higher revenue generation capacity. On the other end of the spectrum, genres like Documentary, Foreign, and TV Movie appeared to be less profitable, indicating lower average revenues.

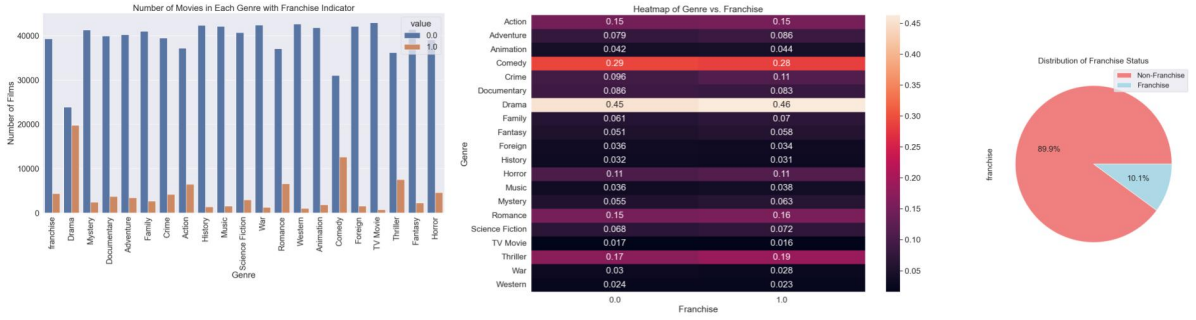
Figure 3: Pie Chart Analysing Genres and Revenue



4.4 RQ4: Correlation Between Franchise Movies and Genres

Analysing the relation between the genres and whether the movie is part of a franchise can also indicate if certain genres are more popular and prevalent than others. However, there is a class imbalance in the dataset, i.e., there are more movies which are not part of a franchise (90%) than the movies which are part of a franchise. This observation is crucial in our analysis as there are very few movies which are part of a franchise in the dataset.

Figure 4: Pie Chart, Bar Plot and Heatmap Analysing Genres and Franchise



The bar chart shows the genres and the number of movies for each genre which are part of a franchise or not. As seen, for all genres, there are more movies that are not part of a franchise than movies that are part of a franchise. However, the genre with the least difference in movies which are part of a franchise and not is Drama. Interestingly, Drama movies are more prominent in movies which are part of a franchise and least prominent in movies that are not part of a franchise. The heatmap shows the correlation between the franchise and the genres for the movies in the dataset. As seen, Comedy and Drama and non-franchise movies have a high positive correlation whereas all other genres have a no correlation with non-franchise movies. Within franchise movies, Comedy and Drama again movies have a high positive relation whereas the remaining genres have no correlation, indicating that they are not highly part of franchise movies.

5 Conclusion

In conclusion, our findings show that most common genres in general are Drama and Comedy. In terms of ratings, Western, Animation and Mystery movies are rated higher than other genres. While Adventure and Fantasy movies perform well in terms of revenue, the genre Drama is more common in both franchise and non-franchise movies. This study successfully demonstrated the changing nature of movie genre popularity over time. Through rigorous data wrangling and analysis, we've identified notable trends in genre prevalence, audience and critic ratings, revenue correlations, and the relationship between movie genres and franchise likelihood.

Our analysis has been thorough, but it is not without limitations. Future research could benefit from incorporating more diverse datasets, such as international cinema, to provide a more complete picture of global movie genre trends. Furthermore, investigating the impact of digital streaming platforms on genre popularity may provide additional insights into the rapidly changing landscape of the film industry. In conclusion, the insights gained from this project not only serve as a strategic guide for current industry professionals but also pave the way for further exploratory studies in the realm of film analytics. The film industry's stakeholders can use these findings as a strategic guide to help them make well-informed decisions about upcoming projects.