

Movie and Show Trend Analysis with Netflix

Netflix has become one of the most popular streaming platforms worldwide, with millions of users using it every day, offering a wide variety of movies and shows. Understanding users' behaviours and movie trends on the platform plays a crucial role in improving the strategy to engage more users. This project aims to analyze the patterns and trends in movies and shows over time while looking into the pattern of genres over the season, the effects of release month on ratings and predict the popularity score based on runtime, release year, and average rating.

For the dataset, it was a cleaned and combined dataset from the non-commercial IMDb raw data and another dataset online about Netflix. Although Netflix's dataset provides information about date added and release year, it does not have viewer ratings or a metric to measure the popularity score. To understand more about the user's engagement, IMDb's dataset was combined to get the average ratings for each movie or show and the number of people who voted on this platform. IMDb, or Internet Movie Database, is a trusted source with millions of user ratings and reviews from around the world and a diverse library for movie and show information. It provides standardized ratings and popularity metrics, which allow us to calculate the popularity score and analyze the popularity of Netflix content. Some movies and shows are listed under multiple genres. Different platforms may provide slightly different genre labels, so I combined the genre information for each title to create a consistent dataset before analyzing the trend. The genres information in the Netflix dataset was also represented differently from that in IMDb, with plural forms, while the other is in singletons. This issue was handled with the inflect library in Python with a helper function for some special cases, such as "Romantic" to "Romance". In addition, there was some inconsistent whitespace in the string that needs to be removed so we can separate the genres string into an array of individual genre strings. Another cleaning step for the Netflix dataset was to remove the words "TV", "Shows", and "Movies". In the original dataset, all TV shows included an extra label "TV Show", which was hard to standardize and combine with the IMDb dataset. It is also not an actual genre, so I removed this extra string to retain the meaningful information for the genre while maintaining the consistency with the other dataset.

After cleaning and combining the two datasets, I have enough information to calculate the popularity score to give a quantitative measure of how much attention a movie or show actually received. The dataset provides the average rating for each title with a votes number, but these two values can not reflect the true popularity of each title. Some of them can receive very few votes, but each of them is high, which makes their average ratings high. This can create an imbalance and may affect the interpretation of the popularity trend. The popularity trend was calculated based on how IMDb calculates its popularity trend with a Bayesian estimation. This estimation can adjust the weight of votes and balance the extremely high or low ratings to give a fair measure across all titles. The formula for the popularity score is:

$$\frac{v}{v + m} * R + \frac{m}{v + m} * C$$

Where v is the given number of votes for the title, R is the given average ratings, m is the median of all votes, so we know how many votes a title needs before trusting the rating fully, and C is the average ratings of all titles.

I also added a season column based on the month added to Netflix for each title. This categorizes each movie or show into Spring, Summer, Fall, and Winter to analyze the seasonal patterns in the distribution. To give a better format, I exploded the dataset to give one genre for each movie or show. This means a title can have multiple rows in the dataset, but each of them only includes a single genre. At the end, I decided to write the final dataset to an CSV file so I can import and use Pandas to do the analysis with visualization to interpret the result.

I first examined whether some genres appear more often than others in some seasons using the Chi-square test. To investigate the movie and show trends over time, I first focused on “Do some genres appear more often than others in some seasons?” to understand if the audience’s interest changes. The season for each title was created based on the date Netflix added it, so we still need more data to understand this pattern, but Netflix usually adds content based on users’ interests. They can be influenced by other factors such as holidays, school breaks or their mood, so this test can give us more insight into whether each genre likely has the same number of titles in any season. I created a crosstab between the genres and season to give a numeric table frequency for each pair of them, and used the Chi-square test to give a p-value. The test gives a p-value of 0.0003756030542558157, which is significantly lower than 0.05, so we reject the null hypothesis, which is that the genre added to Netflix does not depend on the season.

Additionally, I used the stacked bar plot in *Figure 1* below to have more insight into the distribution of each genre across all seasons. We can see that the Drama genre has the most amount of titles along with Comedy, International, Romance, and Adventure. With the visualization below, it is more certain that the genre distribution changes across the season. Most of the title was released during the winter. The production team may have wanted to utilize the free time during the holidays and tried to get more viewers at that time.

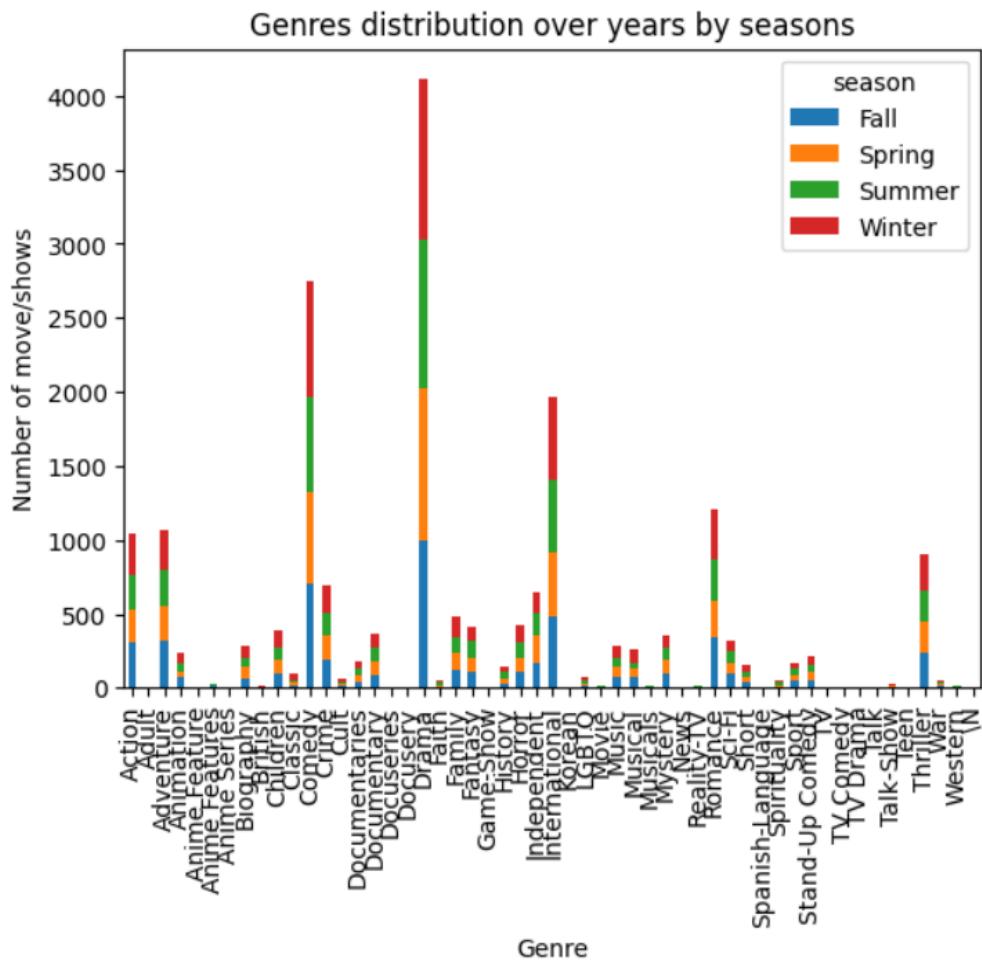


Figure 1

Moreover, examined the relationship between the movie's release year and its popularity score in order to determine whether more recent films tend to receive higher levels of audience engagement with a linear regression model. This model can be used to evaluate the relationship between these two variables: release year and popularity score. As a result, the correlation coefficient was approximately -0.21, which was close to 0 and negative. This indicates that the release year does not have a meaningful linear relationship with a movie's popularity. Although the regression line slopes in *Figure 2* slightly downward, which indicates a minor decline in popularity for more recent movies, the correlation coefficient shows that this relationship is very weak. With this result, we can interpret that the release year does not influence popularity scores, or the more recent films cannot be determined to have higher popularity scores with release years.

Relationship between year released and popularity score on Netflix

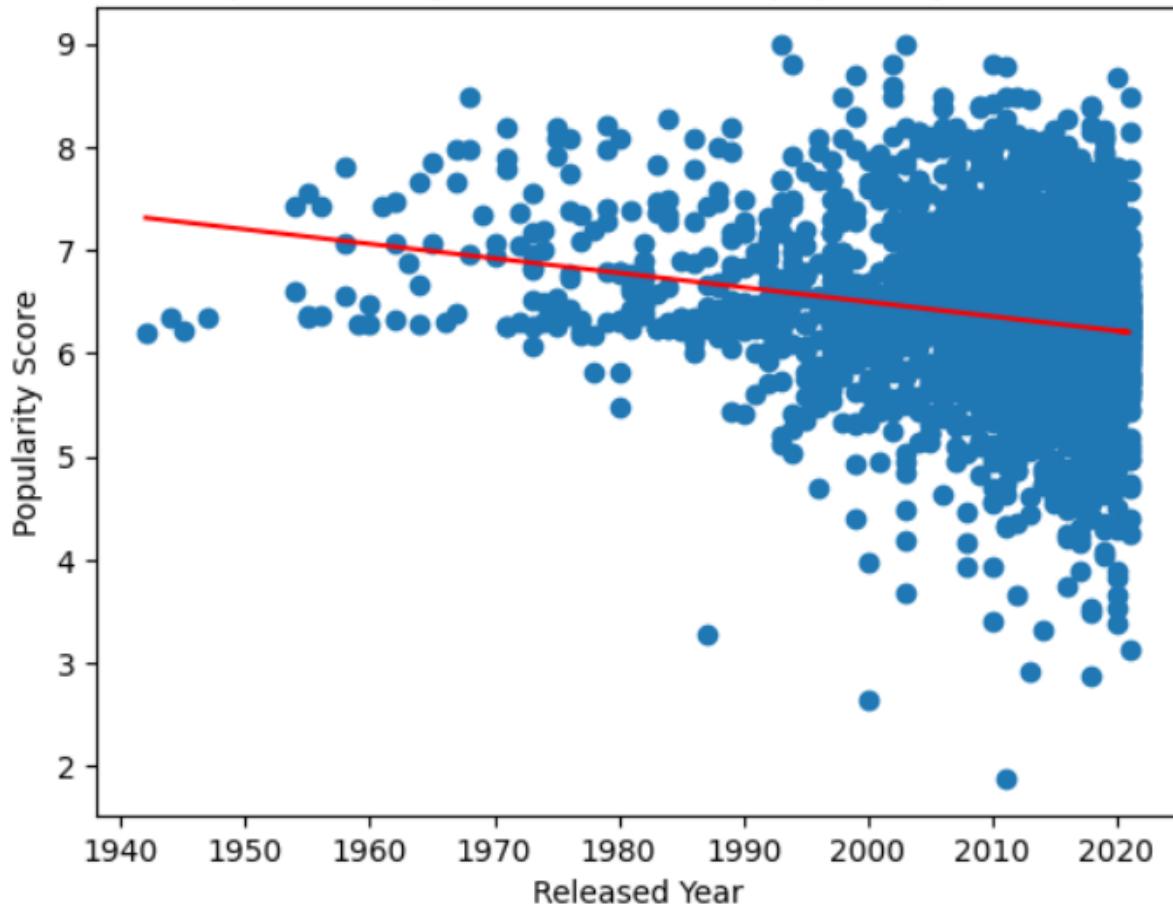


Figure 2

Additionally, Random Forest Regressor was used with 100 trees, each has a max depth of 5 to split and a minimum number of samples that a leaf node must have of 10 to predict the popularity score given the running time in minutes, number of votes and release year. The model achieved a score of 35% accuracy on the training set and 32% on the validation set. With this model, I also got the feature coefficient of around 0.23, 0.70 and 0.06 corresponding to running time, number of votes release year. This reveals that the number of votes influences the prediction mostly, with 70% in total. Runtime contributes approximately 23% to the model's prediction, while release year accounts for only about 6%. This aligns with the earlier linear regression analysis, which demonstrates that release year has minimal influence on popularity scores. However since the model has a limited capacity with only 35% and 32%

accuracy on training and validation, respectively, further additional factors and analysis, such as t-tests or ANOVA, need to be tested to understand what contributes to the movie or show's popularity the most.

To conclude, the Drama genre has the highest amount, with most of the movies/shows released during the winter. The timing might be close to the holiday season, so the production team potentially can take advantage of viewers having more free time. However, the Chi-square test indicates that the addition of a genre is not dependent on the season. My initial hypothesis for popularity was that more recently released films would be more popular due to advances in technology and changes in viewer behaviour. Surprisingly, both the linear regression and random forest indicate that the release year has little effect on popularity scores. With the limited time, the prediction model with a random forest regressor using runtime, number of votes, and release year has a limited capacity with a pretty low accuracy. In future iterations, there is still room for improvement by adjusting hyperparameters and incorporating additional features to better capture the factors that contribute to the popularity.

One of the main challenges in this project was dealing with inconsistencies in the raw data, including missing values, duplicate entries, with mismatched data types in the genres column. I encountered several challenges while obtaining and preparing the dataset. Merging three large datasets took a long time with only pandas DataFrame, so I decided to migrate this step to PySpark, which resulted in an estimated 70% faster running time. Besides, performing the necessary cleaning required more than expected due to a large number of inconsistencies in genres' format, which required additional standardization across the dataset.