Report IMDB data analysis

Team 1

20-03-2025

Motivation

The renewal of a TV series for a second season is a multifaceted decision shaped by several factors. Traditionally, networks and streaming platforms have relied on audience viewership, critical perception, and financial considerations to determine whether a show continues. However, with the expansion of streaming services and globalization of content, additional factors may now influence these decisions. The goal of this research is to explore which factors influence the likelihood of renewal of a series for a second season. While this is an exploratory analysis, there are two factors that are expected to have a big impact on the renewal chances: genre and average ratings. The genre of a TV show is expected to have a big impact on renewal likelihood, as certain genres are more popular than others and attract loyal audiences. For instance, crime dramas and reality TV have historically shown strong audience retention, making them more likely to be renewed (What's Behind a Show Renewal, n.d.-b). On the other hand, niche or experimental genres with smaller followings may struggle to secure additional seasons, regardless of critical acclaim. Recognizing which genres have a higher likelihood of renewal can help producers and platforms optimize their content strategies. Moreover, average ratings are frequently considered a key factor in renewal decisions. High ratings generally indicate strong audience engagement, suggesting that a show is well-received and likely to perform well in future seasons (Analyzing TV Ratings Systems - Examining the Influence of Viewership Data on Show Renewals | Common Good Ventures, 2001). However, the correlation between ratings and renewal is complex, as other factors – such as production costs, competition, and strategic objectives of the platform – can also influence the final decision. Based on these factors, the question arises to which factors in general influence the likelihood of series renewal, and in particular to what extent genre popularity and average ratings influence this likelihood. The findings of this study can contribute to both academic and industry discussions by shedding light on the key factors that influence TV series renewal decisions. By identifying patterns in these decisions, this research can assist content creators, streaming platforms, and production companies in making more strategic and data-driven choices regarding future productions. Understanding these trends can also help media executives allocate resources more effectively and develop content that aligns with audience preferences. Furthermore, the automated and reproducible workflow used in this study ensures that the research process remains transparent and accessible. This not only enhances the reliability of the findings but also makes the study a valuable resource for other students, researchers, and the broader scientific community. Future studies can build upon this framework to explore additional factors influencing the renewal of TV series, contributing to a deeper understanding of decision-making in the entertainment industry.

Data

The data for this study was obtained from IMDb via https://developer.imdb.com/non-commercial-datasets/. This website contains seven different datasets with information from IMDb. For this study, the following four datasets were used: 'title.akas', 'title.basics', 'title.episode' and 'title.ratings'.
None of the information that is being collected contains sensitive information about people for example, as we only collect information about the series themselves, like release year or genre.

Methodology

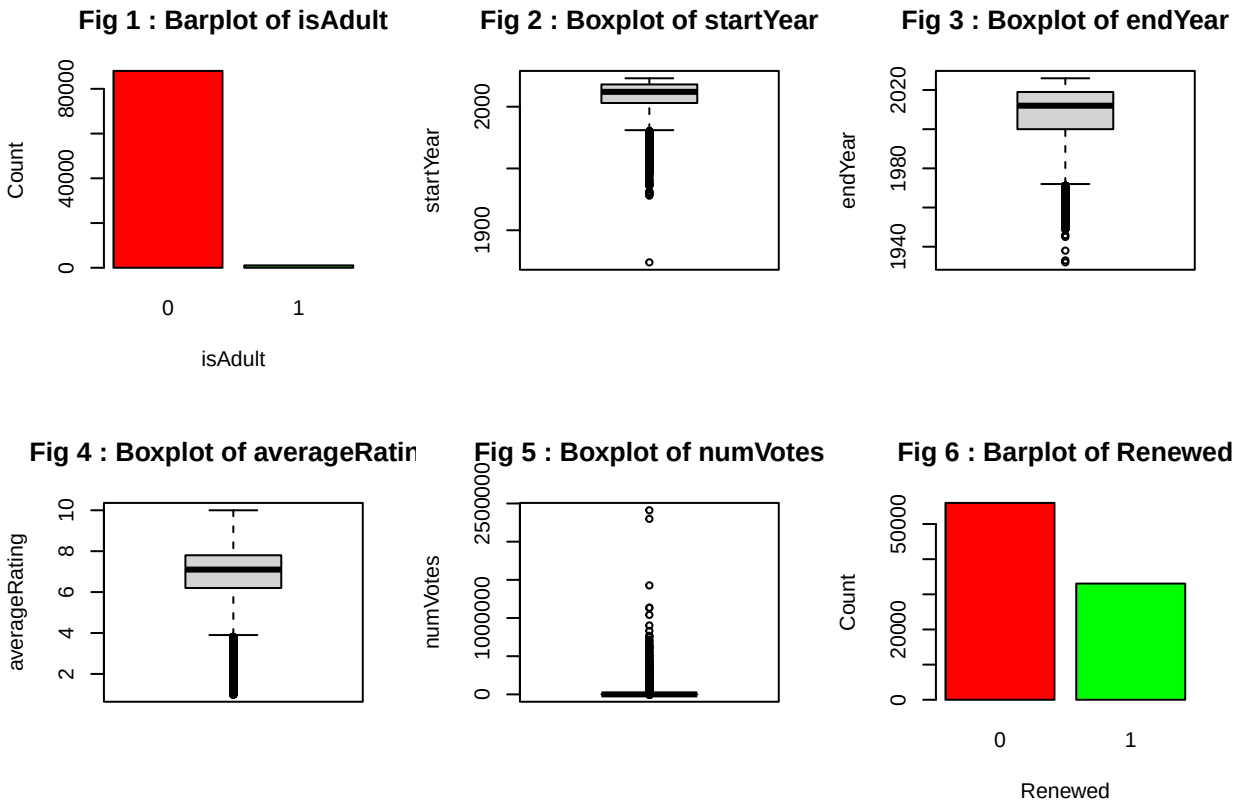
To explore what factors influence the likelihood of series renewal, we will conduct a logistic regression analysis. The renewal status will be the dependent variable, while isAdult, startYear, endYear, Genre1_encoded, Genre2_encoded, Genre3_encoded, averageRating and numVotes are the independent variables. These are either numerical or dummy encoded. According to Lee and Wang (2003), logistic regression is a useful method for analyzing binary variables, which is our dependent variable, because it models and predicts the probability of a specific outcome. This method is useful as it can handle both continuous and categorical predictors, making it versatile for various types of data.

New dataset construction

Before performing logistic regression, the data collected from IMDb needs to be cleaned. The first step is merging four different data sets into a single one using the unique series IDs present in all of them. This new, complete data set is then enhanced by adding a "renewed" variable, which indicates whether a series was renewed for a second season. Additionally, the "genre" variable is split into three separate variables, as each series originally had multiple genres, but now each variable represents a single genre. Next, unnecessary variables are removed from the data set, along with those that contain only NA values. Data points with missing values for both the average rating and the number of votes are also discarded, as these are key indicators for the regression analysis.

Below you can see the first rows of the final data set we used to answer our research question.

```
## # A tibble: 6 x 17
##   parentTconst primaryTitle       originalTitle isAdult startYear endYear Genre1
##   <chr>        <chr>              <chr>           <dbl>     <dbl>   <dbl> <chr>
## 1 tt0035803    The German Weekly~ Die Deutsche~       0      1940    1945 Docum~
## 2 tt0039120    Americana          Americana           0      1947    1949 Family
## 3 tt0039123    Kraft Theatre      Kraft Televi~       0      1947    1958 Drama
## 4 tt0039125    Public Prosecutor  Public Prose~       0      1947    1951 Crime
## 5 tt0040021    Actor's Studio     Actor's Stud~       0      1948    1950 Drama
## 6 tt0040028    Talent Scouts      Talent Scouts       0      1948    1958 Comedy
## # i 10 more variables: Genre2 <chr>, Genre3 <chr>, title <chr>, ...11 <dbl>,
## #   averageRating <dbl>, numVotes <dbl>, Renewed <dbl>,
## #   Genre1_encoded_encoded <dbl>, Genre2_encoded_encoded <dbl>,
## #   Genre3_encoded_encoded <dbl>
```



Interpretation of the plots The analysis of the TV-series data reveals several interesting patterns: isAdult: The bar plot clearly shows that the number of TV series with an adult rating is significantly smaller than those without. This suggests that adult-themed shows are relatively rare compared to general audience shows.

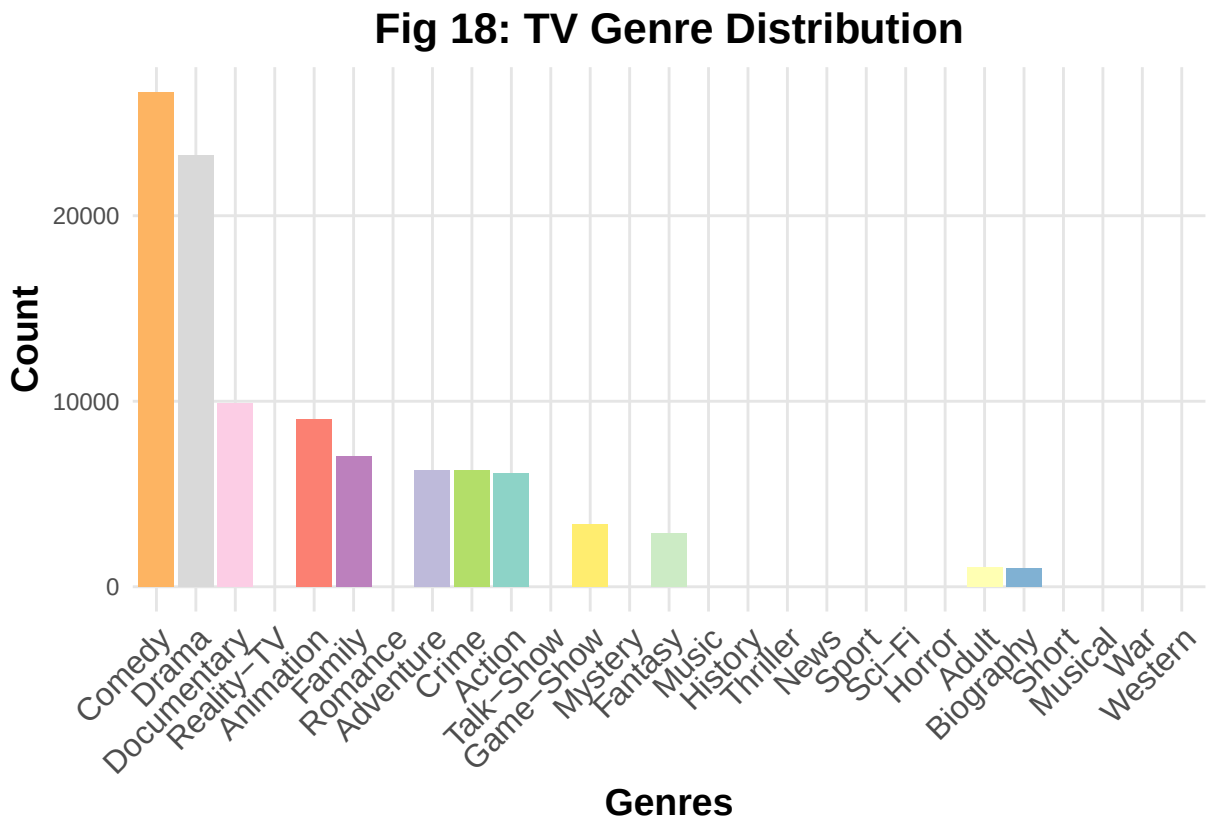
Start Year: The data indicates that significantly more TV series began airing in the early to late 2000s than in the late 1900s, reflecting a negative skew. This is likely because IMDb, which has only been around since 1990, wasn't widely used by showrunners earlier on. As IMDb gained popularity, submitting shows to the platform became a bigger priority.

End Year: Similar to the start year, the end year also shows a higher frequency of shows concluding in the 2000s, which follows the same reasoning as the start year trend.

Average Rating: The distribution of average ratings is mostly normal but slightly negatively skewed. This suggests that, on average, viewers tend to rate shows positively, as an average rating of 5 would indicate a neutral sentiment.

Number of Votes: The number of votes is positively skewed, indicating that most TV shows receive a relatively low number of votes, while a small number of shows attract a large number of votes. This suggests that popularity is not evenly distributed, with a few shows receiving most of the attention.

Next, we turn our attention to the genre of TV shows. Since a show can have multiple genre tags, we can't simply plot a single variable. Instead, we use the pivot_longer function to count the percentage of occurrences of each genre across all shows. The percentages in the table below represent how often each genre is tagged in total, rather than how frequently a genre appears for individual TV shows.



As shown in the graph, the most common genres are Comedy and Drama, which are clearly dominant. On the other hand, genres like Western, War, and Musical are much less popular, appearing far less frequently across the data set. This highlights the genre preferences and trends within TV shows.

Data Analysis

To explore which variables influence series renewal a logistic regression was performed. First the assumptions of the model are tested. To assess whether the independence assumption holds, a Durbin-Watson test was conducted to detect autocorrelation in the regression residuals. For multicollinearity concerns, a Variance Inflation Factor (VIF) test was performed. The results indicate that startYear and endYear have VIF values greater than 10, suggesting severe multicollinearity. This implies that these variables are highly correlated, and it may have been beneficial to combine them into a single feature. Meanwhile, all other predictors have VIF values below 5, meaning that while some correlation exists between independent variables, it does not reach a level that would severely distort the regression model's estimates.

The logistic regression results provide valuable insights into the factors that influence the likelihood of a TV series being renewed. The intercept, with an exponentiated value approaching 1, suggests that when all predictors are at their baseline, the probability of renewal is nearly 100%. However, the individual predictors significantly modify this probability, indicating that various factors contribute to renewal decisions. Among categorical variables, isAdult has an odds ratio of 0.969, meaning that adult-rated series have about a 3.1% lower chance of renewal compared to non-adult series. However, this effect is not statistically significant, suggesting that a show being classified as adult content does not meaningfully influence renewal outcomes. Similarly, averageRating, with an odds ratio of 0.995, shows no significant impact on the likelihood of renewal, indicating that a show's audience rating alone is not a strong predictor of its continuation. For numerical predictors, startYear and endYear play significant roles. StartYear has an odds ratio of 0.632, meaning that each additional year in the start date reduces the probability of renewal by 36.8%, suggesting

that older shows are more likely to be canceled. Conversely, endYear has an odds ratio of 1.593, meaning that with each additional year a show remains active, its likelihood of renewal increases by 59.3%, reinforcing that newer or ongoing series have better chances of being renewed. Genre also affects renewal probability. The encoded genre variables—Genre1_encoded (0.978), Genre2_encoded (0.986), and Genre3_encoded (0.975)—all have odds ratios below 1, indicating that certain genres are less likely to be renewed. Among them, Genre1_encoded and Genre3_encoded exhibit stronger negative effects, reducing renewal likelihood by 2.2% and 2.5%, respectively. This suggests that some genres face higher cancellation risks, possibly due to lower audience demand or shifting network priorities. The numVotes variable, with an odds ratio of 1.000034, shows that an increase in audience engagement slightly improves the chances of renewal. While this effect is minimal, it indicates that shows with more votes—likely reflecting larger or more dedicated audiences—are marginally more likely to be renewed. Overall, the analysis highlights that start year, end year, and genre are the most influential factors in determining whether a TV series gets renewed. Older shows face a higher risk of cancellation, while ongoing or newer series have better renewal prospects. Additionally, certain genres appear to have a disadvantage in renewal decisions, while audience engagement through votes has a minor yet positive impact.

Sources

Lee, E. T., & Wang, J. W. (2003). Statistical methods for survival data analysis. In Wiley series in probability and statistics. https://doi.org/10.1002/0471458546 What's behind a show renewal. (n.d.). Parrot Analytics. https://www.parrotanalytics.com/insights/whats-behind-a-show-renewal/