Data Visualisation: Project Development Group 10

Rebekka Schade
up202203042@up.pt

Deborah Dore
up202202823@up.pt

Mariana Barbosa
up201508455@up.pt

I. Introduction

In the present work, we conducted an exploratory analysis of the IMDB data set and aimed to answer the following research question: what makes a series successful? In order to retrieve from the available data some insights on how to achieve success, we first needed to reach a definition of success. In other words, we need to figure out how to measure success. After filtering and cleaning the data sets (Section II), we conducted some preliminary analysis that could give us some insights on what features we should use to consider a series as successful (Section III).

We decided to only include series from 2010 onward in our analysis because the rise of internet connectivity and the increasing adoption of streaming platforms such as Netflix and HBO changed the entertainment landscape in the last decade. Given this paradigm shift, we cannot expect that the definition for success in the series industry with the prospective audiences of Millenials and Generation X would be the same as for their predecessors. After defining our success criteria and identifying the most successful series based on those criteria, we investigated this successful group in dimensions such as *genre*, *number of seasons*, *runtime minutes*, *actors* and *directors*, looking for differences (compared to the general series data set from the same period) and patterns that could provide us with insights into what the ingredients are for a series to reach success (Section IV).

II. Dataset

For each data set, we performed cleaning and filtering. In every data set for the columns `tconst` and `nconst` we checked that every entry consists of “tt” or “nm” followed by 7 or 8 digits and then transformed it to integers corresponding to the digits. Missing values denoted as “\\N” or empty character were transformed to *NA*.

The data sets were then combined and only the columns crucial to our research were kept. This information included the following: the series' identifier, type, most well-known title, start year, run time, genres, average rating, the number of votes, directors, actors, the number of translations and the number of seasons. For most series, the average run time is the average duration of one duration in minutes. However, for some series this variable contained the total amount of run time. We left these series in the data set and in the graphs we decided to cut the axis for run time at 100 minutes. After selecting the columns, we dropped all the rows which had missing values in ratings, number of votes, number of translations or the run time.

While working, we also took a close look at each season by creating a unique data set with the attributes listed below for each season of each series: a unique identifier for each season, created by us, the series ID to which it belongs, the season's number, the average number of minutes each episode, the genres, the directors and and actors involved, the number of votes, the average rating, the minimum and highest ratings, and the number of episodes.

The results are two data sets that have been filtered and cleaned, ready for the following stage of analysis.

III. How to measure success

Looking at Figure 1, the most expressive result is the positive correlation between number of translations and number of votes. Apparently, series that are translated in more languages are also rated by more people. From this plot, we may also conclude that there are more translations of the most recent series since there is a positive correlation between the number of translations and the start year of each series.

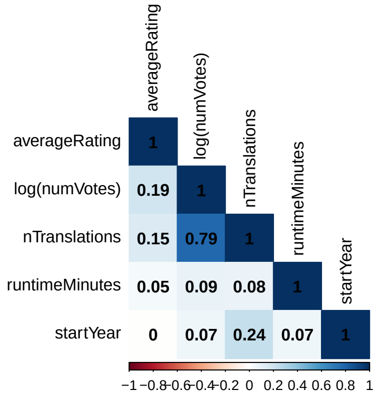
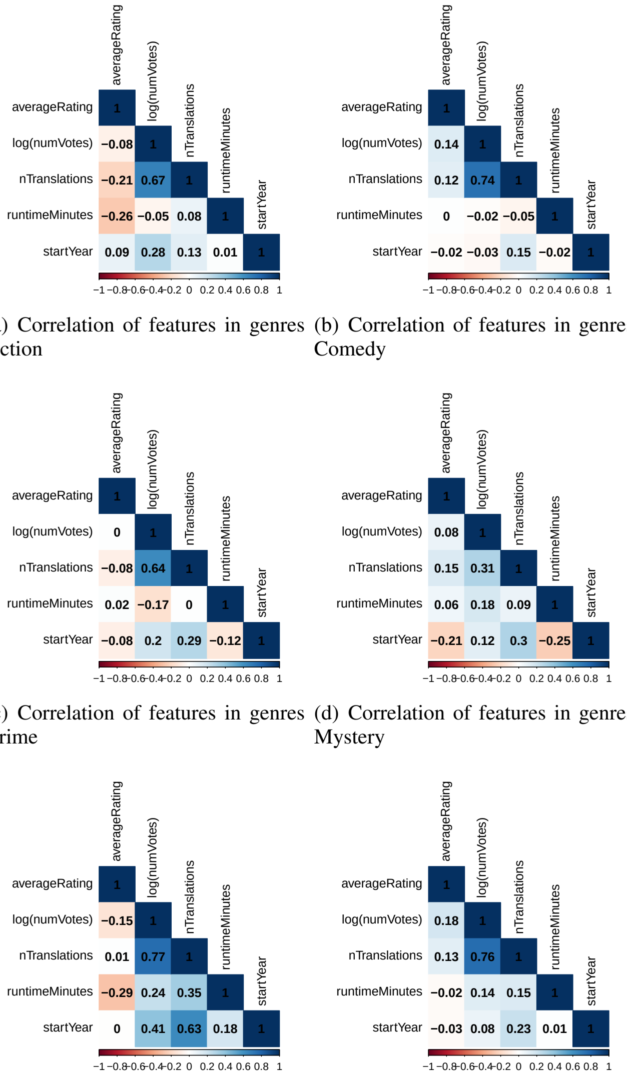


Fig. 1. Correlation between all the genres

The positive correlation between the number of translations and the number of votes can be found in all genres, with Sci-Fi being the strongest, followed by Drama, Comedy, Crime, and Mystery coming in last. The positive correlation between number of translations and start year is especially high for Sci-Fi, followed by Mystery and Crime (Figure 2(d) and Figure 2(c)), meaning that more recent series are translated more often. Also in the Sci-Fi genre is where we can find the strongest positive correlation between number of votes and start year, followed by Action and Crime. This is an interesting finding since we had expected that older series which have been published for a longer time also gave more opportunities to the viewers to vote for them. Because of the other two mentioned high correlations, we assume that the start year and the number of votes are only correlated through the number of translations, not directly.



(a) Correlation of features in genres Action

(b) Correlation of features in genres Comedy

(c) Correlation of features in genres Crime

(d) Correlation of features in genres Mystery

(e) Correlation of features in genres Sci-Fi

(f) Correlation of features in genres Drama

Fig. 2. Correlation plots

Looking into the genre Mystery in Figure 2(d), it's interesting to see that for this genre, the start year has a negative correlation with both the average rating and the run time. This might indicate that more recent Mystery series have shorter episodes but are also rated worse by the viewers. Interestingly, for Sci-Fi and Action series 'less seems better', since there is a negative correlation between the average rating and the run time: shorter Sci-Fi and Action series tend to be better rated than longer ones. Another interesting thing is that, for instance, in the genre Action, there is a negative correlation between average rating and the number of translations. Given this result, in case our definition for success would be popularity, in the quantitative sense, we should only use number of votes and number of translations as measures of it. But if in the short-term a series that reached lots of viewers can be very profitable, in the long-run only the remarkable ones will be remembered. In this sense, we believe success implies also 'quality' on top of 'popularity', and so we decided to include for measures of 'success' also the 'average rating'.

A. *Success Factors*

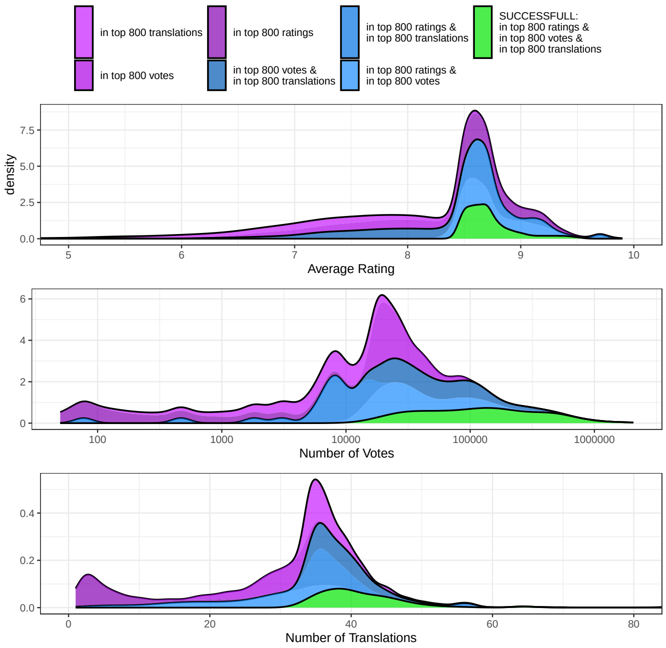


Fig. 3. Average rating, number of votes and number of translations density plot

To identify the most successful series, we first created three subgroups for the series that are among the 800 best rated series, the 800 series with most votes and with most translations. Then we intersected these subgroups, which yielded a sample of exactly 100 series, which are hereinafter defined as "successfull".

Figure 3 shows the distribution of these subgroups in relation to the three variables. The blue part depicts the density of the series which are among the 800 best in only one of the variables, the purple part contains the series which are among the 800 best in two variables and the green part consists of the series which are included in the intersection of all three variables, hence, the successful series. As the figure shows, this combination of features results that our group of successful series is constituted by 100 series that have more than 10000 votes, more than 30 translations and more than 8.5 points in the average rating.

IV. How to get success

A. Number of Seasons

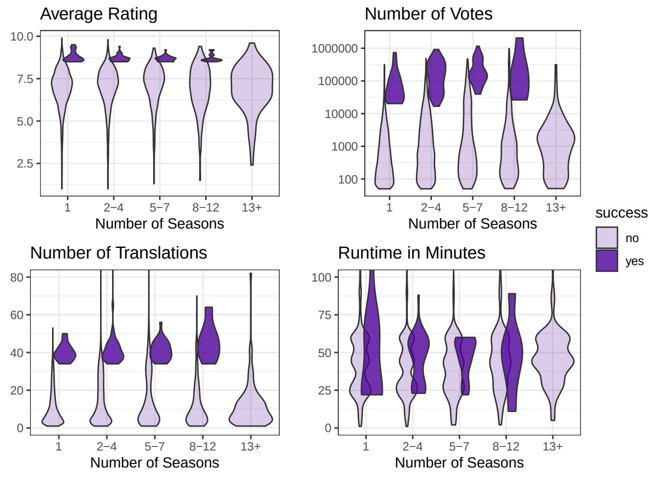


Fig. 4. Violin plot of the number of seasons

Now that we defined our criteria for success, in order to find the 'recipe' that leads to success, with need to explore the characteristics of the series matching our criteria and compare them with the rest.
As for the number of seasons, we can see that the 100 series that match our criteria for success have less than 13 seasons. Hence, if the goal is to produce a successful series, the number of seasons should not extent 12 seasons. We can see that for our successful group, there are no major variations in the average ratings between number of seasons, which tells us that the average rating of a successful series is independent from the number of seasons it has in total. When looking at the series which do not meet our criteria for success, the series with only one season have a clear mode at around 7.4, but the distribution widens in the categories with more seasons in total. Also, series with more seasons have a slightly better average rating which can be explained by the fact that series who are not rated well will not be continued for long.

Furthermore, we can see an increase in the number of votes in the series that have 5 to 7 seasons and 8 to 12, where the former group has the highest minimum. In fact, the latter reaches the highest number of votes with more than one million. These results let us conclude that series with more seasons and thereby more total run time tend to get more votes. As for the group of series which are not considered successful by our definition, the mode for the number of votes is a lot higher in the series with more than 12 seasons. This finding contributes to our conclusion that the series which are continued for so many seasons - probably because they are somehow successful, although it might not fit into our definition - in general receive more votes.

For the number of translations, there is a similar pattern to the number of votes where the number of translations rises if a successful series has more seasons and series with more than 12 seasons in general are translated to more languages.

The distribution of the run time has a more bumpy shape, meaning that there are certain run times which occur very often. These are around 25 minutes, 40 minutes and 60 minutes. We cannot see any mentionable differences between the categories of number of seasons. The rather high run times among the successful series with only one season is probably due to errors in the data where the run time for these short series is not the one per episode but the total run time.

B. Zoom in of five series

Let's examine more closely five of the feature distributions of series which we manually chose in Figure 5.

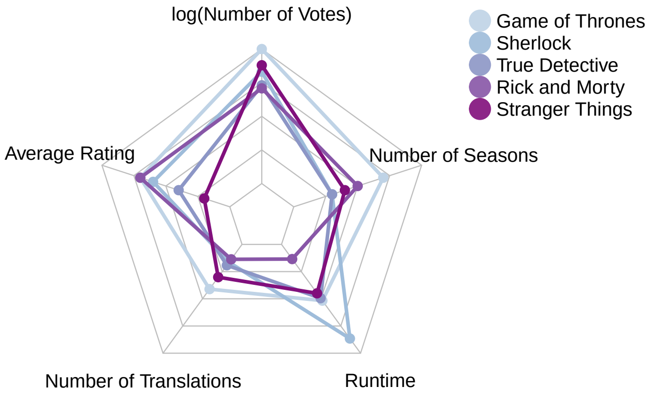


Fig. 5. Radar plot

Four of the five series have shorter runtime minutes than the others in Figure 5. The series with higher runtime minutes within the 5 analyzed is *Sherlock*.

Game of Thrones is the series that is mostly translated in the Figure 5, in terms of quantity, followed by *Stranger Things*, while the others have a similar number of translations.
In terms of average rating, *Game of Thrones* and *Rick and Morty* have the highest rating, followed by *Sherlock*, *True Dective* and *Stranger Things*. *Game of Thrones* has the most votes, which could indicate that it was the most voted series and thus people like it a lot, but *Game of Thrones* also has the most seasons compared to the others, which could also suggest that people had more time to vote for it because they had more seasons to watch.

C. Distribution of Genres

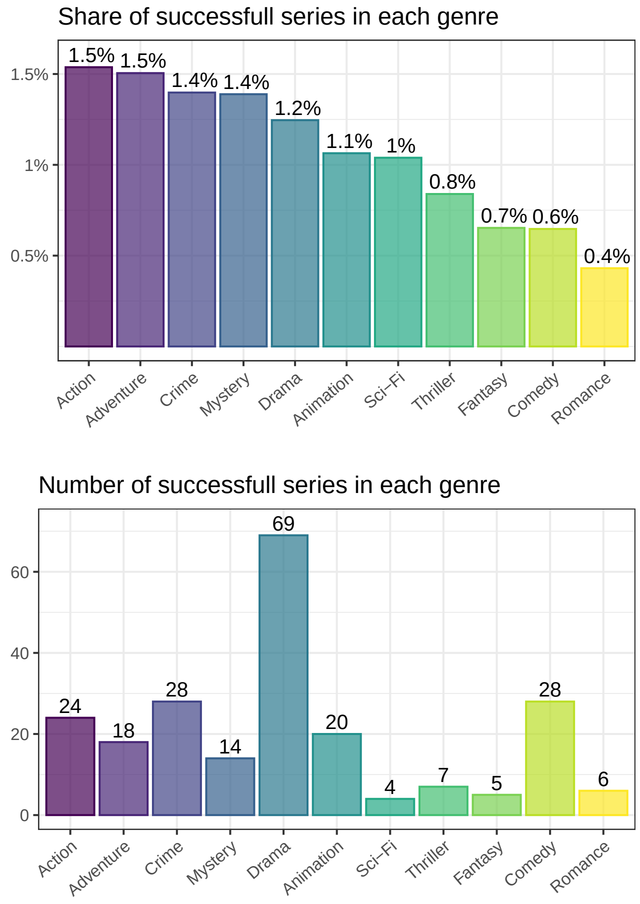


Fig. 6. Distribution of Genres

Later in our studies, we turned our attention to the differences between the genres. In order to select the most relevant genres for further analysis, we decided to explore the distribution of genres in the most successful series. As can be seen in Figure 6, Drama is the most frequent genre, followed by Comedy, Crime and Action. But if we compare this distribution with the rest of the data set and look at the share of successful series in each genre, we can see that Action, Adventure, Crime and Mystery have high percentages. This means that among all series of these genres, the share of successful ones is higher. We decided to include the genres Action, Crime, Mystery, Drama and Comedy in further analysis. The reason that Adventure is not included is the high concurrence between Adventure and Action and therefore a similar behaviour, which was revealed by using an interactive dashboard.

D. Analysis of selected Genres

Concerning the analysis of the selected genres, our primary focus, as illustrated in Figure 7, was on the differences between the *average rating*, *number of votes*, *number of translations* and *runtime in minutes*.

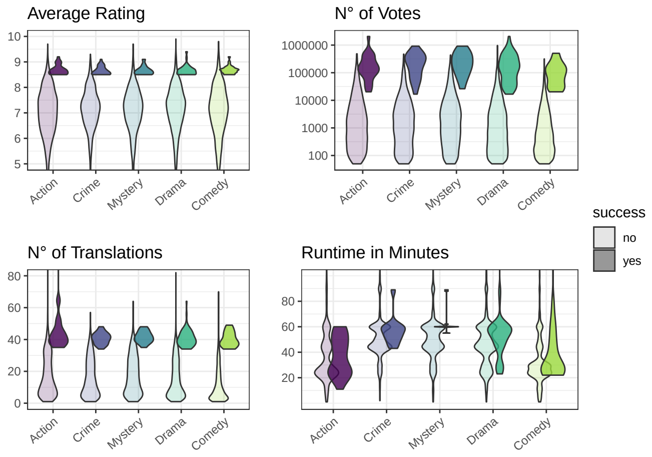


Fig. 7. Violin Plot of Genres

We displayed the distribution of each genre separated by its success (successful or not) for each feature.
According to our findings, the distribution of the average ratings for the unsuccessful series ranges between 5 and 10 with 7 being the mean value, whereas the distribution of the average rating for the successful series is concentrated between 8.5 and 9.5. This is indeed attributable to our definition of success, which only considers successful series with ratings above 8.5. Every genre considered follows this pattern.
A successful series also has over 10,000 votes. While the number of votes for unsuccessful series varies between 50 and 1 million, it decreases after 10,000, indicating that there are more successful series than unsuccessful series after 10,000 votes. Different genres tend to have a similar distribution. The distribution of the number of votes for genres for successful series is similar, with *Action* and *Drama* being the ones that surpass 1 million votes.
In terms of the number of translations, the situation is similar: unsuccessful series are translated less than successful ones. Typically unsuccessful shows, for example, are translated from 0 to 40 different languages. Popular shows, on the other hand, are translated 35 times or more. Some unsuccessful series have been translated more than 40 times, but this number is very low. Again, we may discover samples that have been translated 80 times and more for the genre *Action*.
Finally, there are some disparities in the distribution of successful and unsuccessful series by run time. This is the first feature in which there is no discernible difference between successful and unsuccessful series distribution. Aside from the *Mystery* genre, the distribution of each genre of successful and unsuccessful series is relatively similar. The only noteworthy variation is in the *Mystery* genre, which, having a few successful series, has a lower distribution, with successful series having a run time of 60 minutes per episode.

E. Development of Seasons

We expected to find patterns when looking into the development of seasons from the first to the last. Figure 8 shows, for different genres, the development of the average ratings per season, the number of votes per season and the run time. The colored, non-black lines represent series which meet our criteria for success. For this plot, the sample of series to be presented is reduced to those which have individual ratings for their episodes.

From Figure 8 we can clearly see that successful series (highlighted in color) have an average rating and number of votes per episode that are higher than the others and remain more or less stable through the seasons. Compared to figure 4, among the higher ratings, there is a lot more overlapping between successful seasons and others. Apparently, the general rating of a series considered successful is better than the ratings for each of their episodes. What the graphs shows as well is that seasons tend to decrease in ratings, but not many series increase their ratings from season to season.
When looking at the second row depicting the development of the number of votes, there clearly is a decreasing trend, both for the successful series and for the others. Each new season tends to have less votes than the previous season. However, in the genres Action and Drama, there are a few series which do increase regarding the votes.
The run time shows less mobility between seasons, most series stay consistent with the duration of their episodes. The genre Mystery has most variability in the run time of successful series. Most successful series have a run time of 45 minutes or more. In fact, the share of successful series is much higher among run times over 50 minutes. Only in the genres Action, Drama and Comedy, there are successful series with about 20 or 25 minutes run time. It seems advisable to produce episodes of 45 minutes run time or more and leave the run time consistent. Only if the genre is Mystery, a slight change in run time could add that extra spice which viewers seem to like.

F. *Collaboration of Actors and Directors*

A remaining variable of interest for our purpose are the actors and directors who are involved in successful series, especially their collaborations with each other. Figure 9 and figure 10 are networks which is supposed to analyze the collaboration patterns between actors and directors, respectively. In the network graphs, one node depicts one person and an edge between two persons exists if these people collaborated in at least one successful series. The population is limited to actors and directors in successful series where figure 9 include only actors and figure 10 only directors. The nodes are colored by one of the most frequent genres of the series which a person has participated in. Since there can only be one color per node, but the actors and directors are involved in series with more than one genre, the coloring does not give the entire information available regarding genres. We therefore suggest to conduct the analysis separately for each genre, for instance in an interactive dashboard.

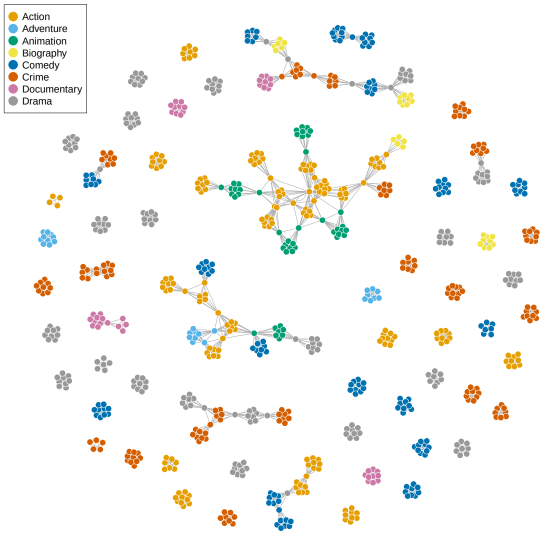


Fig. 9. Collaborations between actors

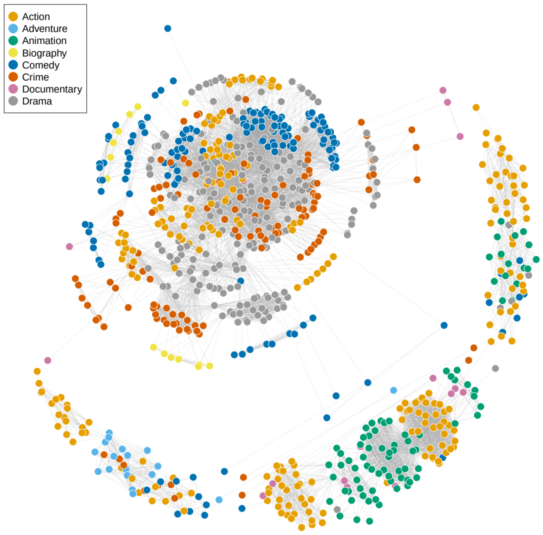


Fig. 10. Collaborations between directors

Figure 9 shows that the actors of successful series do not collaborate a lot in between series. Most series have their set of actors who don't have major roles in other successful series. There are four bigger nets where the actors of more than four series series are collaborating, but in most cases this only applies to one actor who has a role in two series. The biggest net shows a closer collaboration between series of the genres Action and Animation and further analysis have shown that most of the series in this net are animated. In conclusion we can derive that in order to make a series successful, it is recommended not to focus on having actors from other successful series in the cast unless it is an animated (Action)

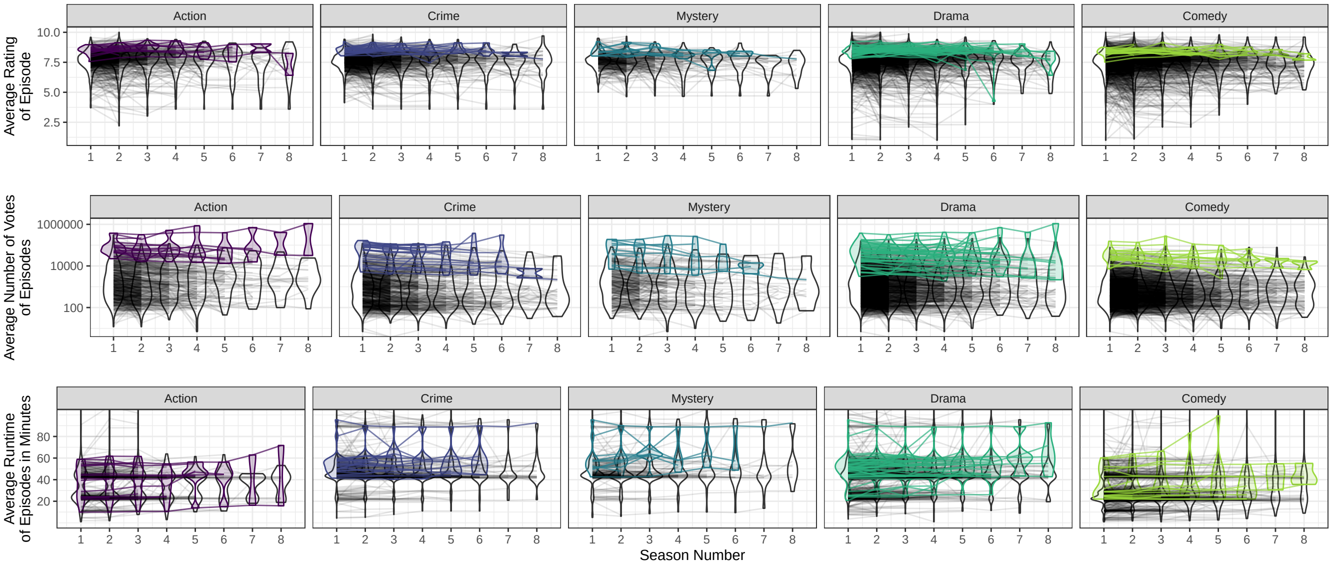


Fig. 8. Development of Seasons - colored lines depict successful series

series.

When it comes to directors, figure 10 presents an entirely different image. There are no clear patterns as we saw in the previous network. Instead, the network consists of a lot more edges, which tells us that the directors of successful series are collaborating a lot. While focusing on the coloring by the genre, there is also no obvious pattern. The analysis of each genre separately in an interactive dashboard revealed that most of the directors involved in the biggest hive are directing series of the genre Drama while the directors of animated series are all outside of the big hive.

G. Most successful Directors

Since we concluded from figure 10 that the directors of successful series are highly collaborative, we were interested in who these directors are who seem to have a positive influence on the success of a series. Figure 11 takes a closer look at the directors who participated in more than 3 successful series.

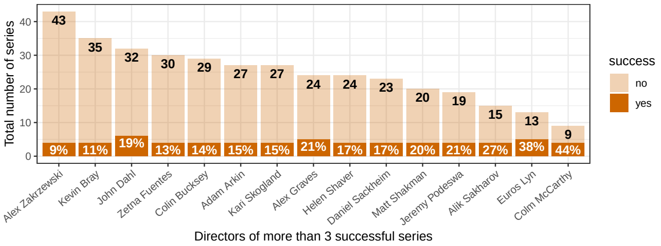


Fig. 11. Directors of more than 3 successful series

A remarkable director for instance is John Dahl who scores high with 6 successful series in his career. His share of 19% of successful series is also quite impressive. There are other directors like Colm McCarthy with a higher share of 44%. He ”only” did 4 successful series, but since he only did 9 series in total, at least according to out data set, his share is extremely high. In conclusion, any of the directors mentioned in the graph would be a good investment for producing a new series to raise the chance of it being successful.

V. Conclusion

From all these results we can build some guidelines which will increase the chance of producing a successful series. As a genre, the producer should consider Action, Adventure, Crime or Mystery. Series of the genres Fantasy, Comedy and Romance have a high risk of not being successful according to our definition. The run time of the serious heavily depends on the genre. In general, successful series tend to have a higher run time. For Mystery, a run time of about 60 minutes, with tiny changes in between, has been proven to be the right choice. Since the ratings stay stable throughout the seasons, good ratings after the first season can predict that the following seasons will have high ratings as well. For the number of votes, however, it is normal that they will decrease throughout the seasons. Independently of how well a series goes, the number of seasons should not extent 12. After all, figure 5 hints to conclude that even within successful series, each series can have a very individual pattern of the average rating, the number of votes, the number of translations, the run time and the number of seasons. When searching for a cast, there is no need to recruit actors from other successful series. A new cast does not seem to impede the success of a series. Nevertheless, an animated series might benefit from a cast which was already involved in other successful series. When choosing directors, it is more important to consider directors from other successful series. All the directors from figure 11 will increase the chance of having success.