What makes a Drama TV Series successful

Inês Marques
Department of Computer Sciences
Faculty of Sciences of the University of Porto
Rua do Campo Alegre s/N, 4169-007
Porto, Portugal
up201604601@fc.up.pt

João Cunha
Department of Computer Sciences
Faculty of Sciences of the University of Porto
Rua do Campo Alegre s/N, 4169-007
Porto, Portugal
up202103227@fc.up.pt

Maria Bessa
Department of Computer Sciences
Faculty of Sciences of the University of Porto
Rua do Campo Alegre s/N, 4169-007
Porto, Portugal
up201704601@fc.up.pt

***Abstract* – IMDb, the Internet Movie Database, is an online database that contains information and statistics about movies and TV shows as well as cast and crew members. The data we were given for this project was related to IMDb and the goal was to understand what makes a series successful, i.e., what makes it popular. We decided to reduce the data to information relating the top 100 voted drama series. We created visualizations with Python's libraries matplotlib, seaborn and plotly to try and find patterns that would lead to a successful drama series. Results showed that there aren't specific patterns strong enough to justify the success of a drama series.**

Keywords—drama series, success, IMDb

I. Introduction

The measurement of TV series' success comes down to one thing: popularity. The goal of the film and TV industry is to make money, meaning they will take quantity over quality. Now more than ever, with the rise of streaming services, such as Netflix and HBO, the only thing that takes to keep a TV series alive is the number of viewers. If a TV series manages to lure new viewers and, in the case of streaming services, new subscribers, then it is more likely that it will keep being renewed. We are not saying that a TV series' rating is completely discarded when it comes to measuring its success, but if a TV studio must make the decision of renewing a series with a high rating but a low viewership rating or a series with a low rating but a high viewership rating, the studio will always choose the latter because popularity is what makes more money.

The goal of this project was to try to understand what makes a TV series successful, i.e., what it takes for a TV series to have a high number of viewers.

II. Dataset

A. Original Dataset

The data we were given data was divided into seven distinct datasets: *names.basics, title.akas, title.basics, title.crew, title.episode, title.principals* and *title.ratings.*

The *names.basics* dataset contained 11972690 rows with information for names, and each row had significant information about each name, such as the name for which the person is most credited for, birth year and death year (when applicable), their primary professions and the titles for which they're most known for. There were no duplicate or negative values, however, there were missing values: *primaryName* had 1, *birthYear* had 11409886, *deathYear* had 11765986, *primaryProfession* had 2577463 and *knownForTitles* had 2121080. Some entries had wrong birth/death information, leading to the existence of 8 people with negative lifespans.

The *title.akas* dataset had 33362201 entries and 8 features: *title, ordering, title, region, language, types, attributes* and *isOriginalTitle*. No duplicated values were found; however, the following variables had missing values: *title* (5), *region* (1863867), *language* (6279784), *types* (28058248), *attributes* (33114486) and *isOriginalTitle* (2187). This dataset mainly contained information about the variation of titles in different countries.

The *title.basics* dataset contained 9267897 entries and 9 columns that described basic information about titles: *tconst* (ID of title), *titleType*, *primaryTitle*, *originalTitle*, *isAdult* (boolean), *startYear*, *endYear*, *runtimeMinutes* and *genres*. There were no duplicate or negative values. However, we detected some missing values: *primaryTitle* and *originalTitle* had 11, *isAdult* had 1, *startYear* had 1237218, *endYear* had 9171150, *runtimeMinutes* had 6787240 and *genres* had 427925.

The *title.crew* dataset had the directors and writers information for all the titles in IMDb. It had 9267897 names. No duplicated values were found; however, the variables *directors* and *writers* had 3966150 and 4506648 missing values, respectively.

The *title.episode* dataset consisted of information related to TV series. It had 6991827 entries and 4 columns that described basic related information about TV series, consisting of the following variables: *tconst* (ID of title), *parentTconst*, *seasonNumber* and *episodeNumber*. No duplicated values were found, however the variables *seasonNumber* and *episodeNumber* both had 1468784 missing values.

The *title.principals* dataset was the largest one with 52370246 entries, containing information about the main cast and crew member, such as the category of the job being executed, the specific job title and the characters played (in this case, only for actors). No duplicated values were found but the variables *job* and *characters* had 43778297 and 26796202 missing values, respectively.

The title.ratings dataset had 1263181 rows, and each row corresponded to a single movie or TV Series with its average rating (*averageRating*), from 1 to 10, and the

IMDb’s number of votes (*numVotes*) the film or series has received. There were no duplicate values, no missing values, and no negative values.

B. *Transformations of the Dataset*

We carried out several transformations on the different datasets.

The first one we performed was dividing the *title.basics* dataset into two, one in which the *titleType* was only tvSeries (*serie_basics*), and other where the *titleType was* only tvEpisodes (*episode_basics*). We did this because we were only going to work with TV series, so the rest of the data was not of our interest.

Next, we merged the *series_basics* dataset we obtained previously with *title.ratings*, creating the *series_ratings* dataset so that we could obtain the *averageRating* and *numVotes* features for each TV series. With this new dataset (*series_ratings*), only the variable *startYear* had missing values, having 31 of them. We decided to delete them, since they were few comparatively with the number of rows we had in this dataset. After that, we filtered the *series_ratings* dataset so that we only had TV shows whose *startYear* was greater or equal to 1990 and the *numVotes* was greater or equal to 20000. Afterwards, we were left with only 875 series. Of those 875 series, we had 211 missing values for *endYear* and 17 for *runtimeMinutes*. Since *endYear* wouldn’t affect our future analysis, we did nothing to those missing values; however, we deleted the missing values of *runtimeMinutes* for the same reason we deleted the *startYear* missing values, previously. Still regarding the *series_ratings* dataset, we trimmed it so that it only contained series that had the word "Drama" on the *genre* variable and only kept the top 100 ones that met that criteria. We also ordered those 100 drama series by number of votes. In the end of all these transformations, *series_ratings* had only one column with missing values, the *endYear* column, having 22 of them.

After this, we merged the *title.akas* dataset with the *series_ratings* dataset, creating the *akas_ratings* dataset, but in order to do it, we had to rename the *titleID* column of *title.akas* to *tconst*, since they represented the same thing, and so that we would be able to join these two datasets. The *akas_ratings* dataset only contained our top 100 voted drama series and had 1004 missing values for *endYear*, 113 for *region*, 3033 for *language*, 438 for *types* and 4432 for *attributes*. The only change we performed here was to transform the variable *runtimeMinutes* into an integer because it was classified as a string. We did nothing regarding the missing values.

Another transformation was merging *series_ratings* with *title.crew* to obtain a new dataset (*crew_ratings*). We only had one column with missing values, *endYear*, with the same number of missing values as in *series.ratings*, since the *title.crew* dataset didn't have any missing values.

We did also merge *series_ratings* with *title.principals* (*principal_ratings*), and then merged it with *name.basics*. This *principal_ratings* dataset had 996 rows and the following missing values: 216 for *endYear*, 96 for *birthYear* and 981 for *deathYear*.

Lastly, we merged *title.episode* with *title.ratings* to obtain a new dataset called *episodes.rating*, which was merged with our previously obtained dataset *series_ratings*, ending with a dataset called *tvseriesepisode_ratings*, that contained all the information related to all episodes of the tv shows. It had 9253 entries and 1966 missing values in *endYear*.

C. *Final Dataset*

Our final datasets had only information regarding the top 100 voted drama series.

III. Preliminary Analysis

In this part of the report, we will show some of the data transformations we made, what we were expecting before performing them and why we made those transformations.

The first step of our preliminary analysis was to confirm the gradual rise of the total number of TV series made each year (Fig. 1) and the total number of voters for the TV series made at the corresponding year (Fig. 2).

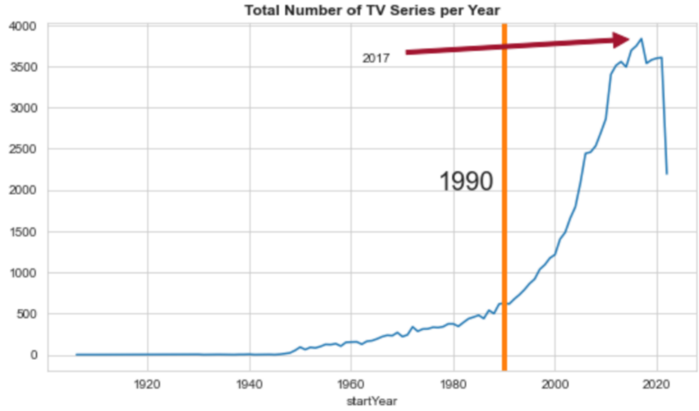

Fig. 1. Total Number of TV Series per Year

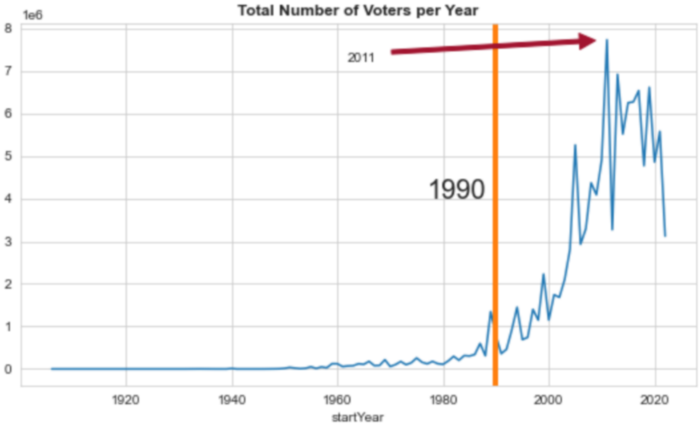

Fig. 2. Total Number of Voters per Year

These plots kind of confirm our initial assumptions though we weren’t expecting a sudden drop after each respective peak. Is this because IMDb is losing its popularity? This could be an interesting research point.

We decided to only keep the TV series made after 1990 because the total number of TV Series and the total number of votes start to steadily increase after 1990. Also, the 90’s witnessed a gradual shift in status and quality of TV, with shows like Twin Peaks, The X-Files, Buffy the Vampire Slayer and The Sopranos.

After only keeping the TV series after 1990, the dataset had about 73000 series. The series with the lowest number of votes had 113 votes. Series with such a low number of votes are not popular. So, to only deal with the most popular series, we decided to keep the ones with more than 20000 votes. We decided on this number by going to the IMDb page and sort the top 250 TV series by number of votes. The 249th series had 24000 votes and the 250th had 12000 votes, so we chose a number in between.

After eliminating some missing values (very few), we had a dataset with 858 TV series. Almost all the series had multiple genres, so, out of those series, we counted the number of TV series per unique genre (Fig. 3).

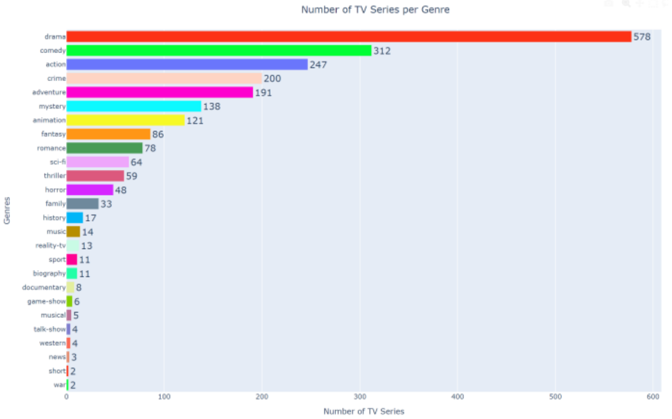

Fig. 3. Number of TV Series per unique genre

As expected, the top spots were taken by drama and comedy. Documentaries having a lower number than reality-tv shows is not unexpected even though documentaries tend to have a much higher rating than reality-tv shows, which just confirms that the measurement of success is mainly calculated by how many people watch the TV series (number of votes), and not its rating.

Fig. 4 shows that drama series don’t have the highest mean average rating, despite being the genre with the highest number of TV series with more than 20000 votes, showing, once again, why we should prioritize the number of votes over rating.

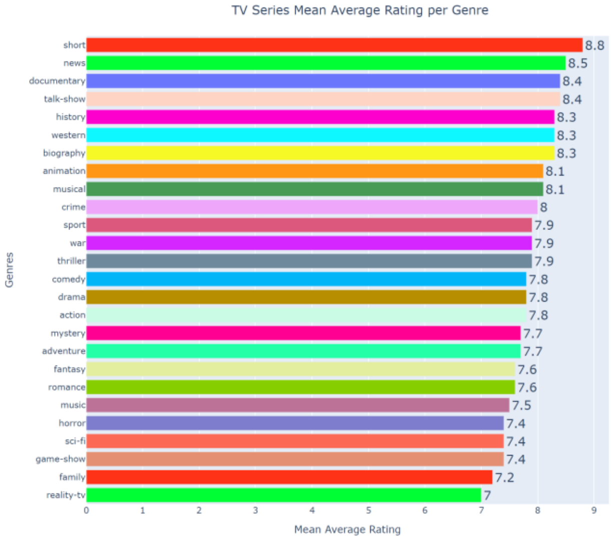

Fig. 4. TV Series Mean Average Rating per Genre

At the end of this preliminary analysis, we decided to try and figure out what makes our top 100 voted drama TV series successful.

IV. Analysis – Top 100 Voted Drama Series

Our first step was analyzing the distribution of the top drama TV series’ rating (Fig. 5). Even though we have been saying that the most successful series are the ones with the highest number of votes and not the highest rating, we were expecting that our top 100 voted drama series would have a high rating.

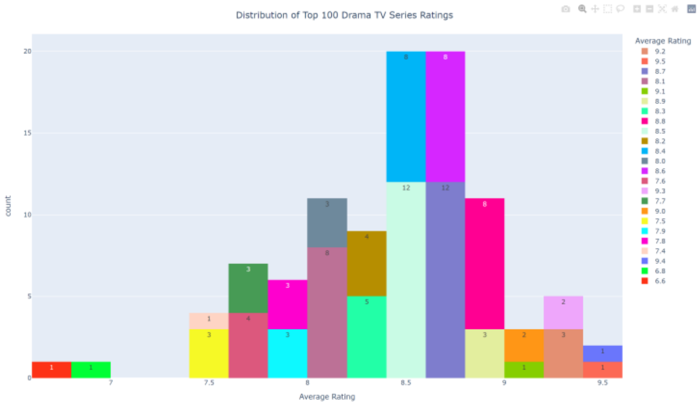

Fig. 5. Distribution of Top 100 Drama TV Series Ratings

As expected, most of the top 100 voted drama series have a very high average rating. Two of the series have an average rating below 7. A 6.8 (Glee) and a 6.6 (Riverdale) are not bad ratings, but they are also not great. Still, both series have over 140000 votes, showing, once again, that a series doesn’t have to be great to be popular/successful.

We found the top 10 rated drama series (Fig. 6) and the top 10 voted drama series (Fig. 7). Only three of them were in both top 10’s (Breaking Bad, Game of Thrones and Sherlock). If the quality of a series would directly translate into the number of voters of a series, these top 10’s would be the same. Nevertheless, the top 10 voted drama series have a very high rating. So, from this point on, since we were only dealing with series with a very high number of votes, we also started using the quality of a TV Series (average rating) as a measurement of a series’ success.

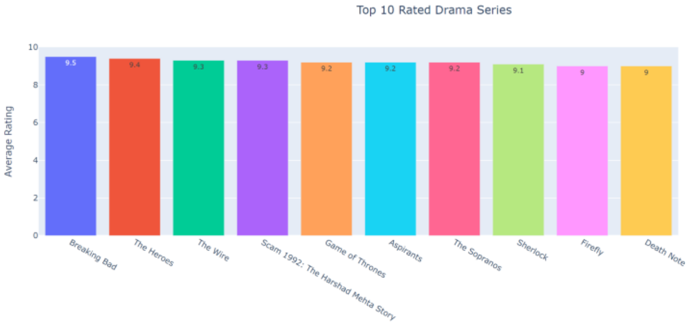

Fig. 6. Top 10 Rated Drama Series

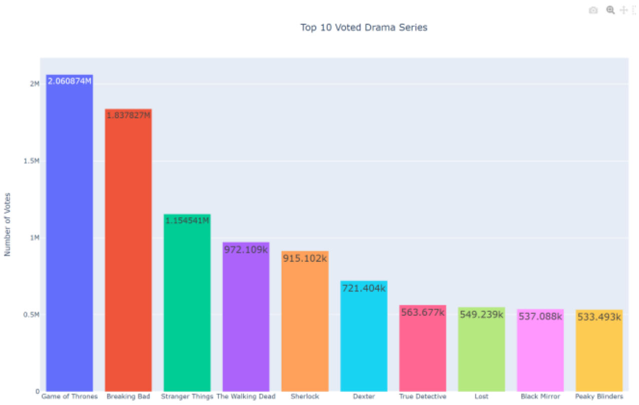

Fig. 7. Top 10 Voted Drama Series

At this point, we started looking for specific patterns that lead to a drama series having a high rating and a high number of votes.
To see if drama TV series tend to have a higher average rating when its episodes' runtime is longer, we made the plot below (Fig. 8):

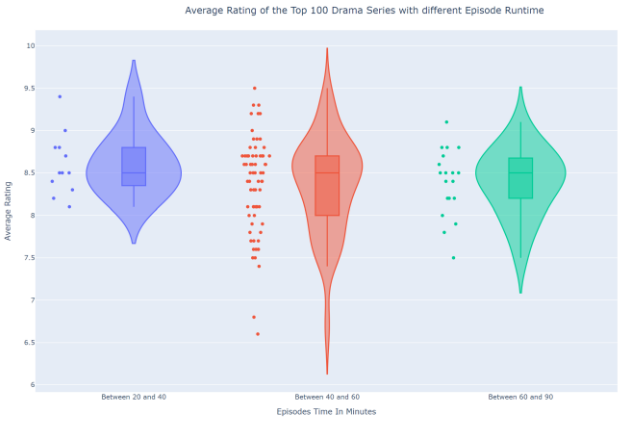

Fig. 8. Average Rating of the Top 100 Drama Series with different Episode Runtime

There is almost no difference between the average rating of drama series with runtimes between 20 and 40, 40 and 60, or 60 and 90, meaning the runtime of a drama series does not influence its average rating. On the other hand, it seems to influence, if only slightly, the number of votes of a drama series (Fig. 9).

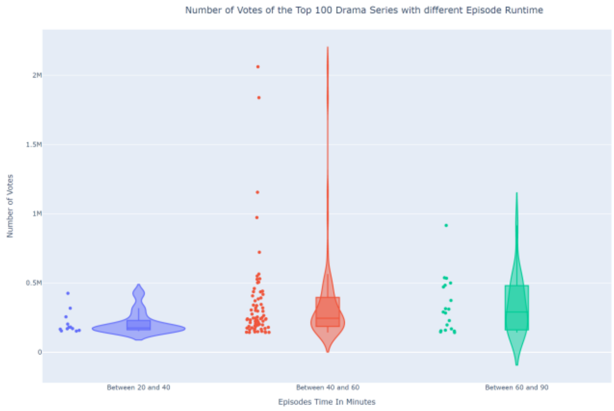

Fig. 9. Number of Votes of the Top 100 Drama Series with different Episode Runtime

The number of votes seems to increase when the drama series have longer episodes, which can be explained by the fact that, when a series is popular (which is the case of the 100 series we're analyzing), people want to see more minutes of it. But we can't conclude that series with longer episodes will automatically be more successful.

Then, to see how many different drama series are showed in different countries (the higher the number, more different drama series are showed), we made the plot below (Fig 10.):

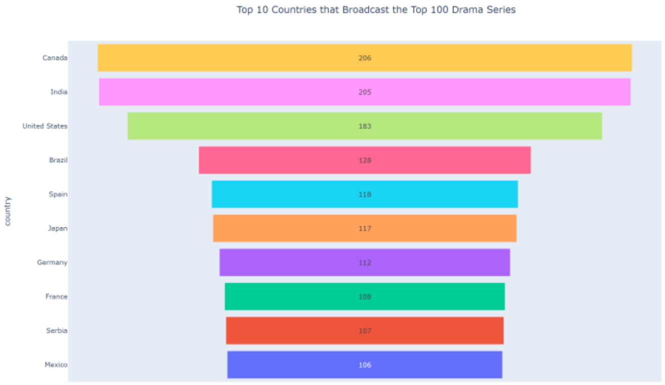

Fig. 10. Top 10 Countries that Broadcast the Top 100 Drama Series

This could mean that if a drama series is popular, then maybe it will be broadcasted in one of these countries.

Next, we found that the vast majority (more than 90%) of our top 100 drama series had more than 5 writers and more than 3 directors. This comes as no surprise since series have several episodes with different types of scenes, hence, several different types of writers and directors. We tried finding a pattern regarding the number of writers and directors of our top 10 voted drama series, but there wasn't one (Fig.11). The result was similar for the number of writers and directors of our top 10 rated drama series (Fig. 12).

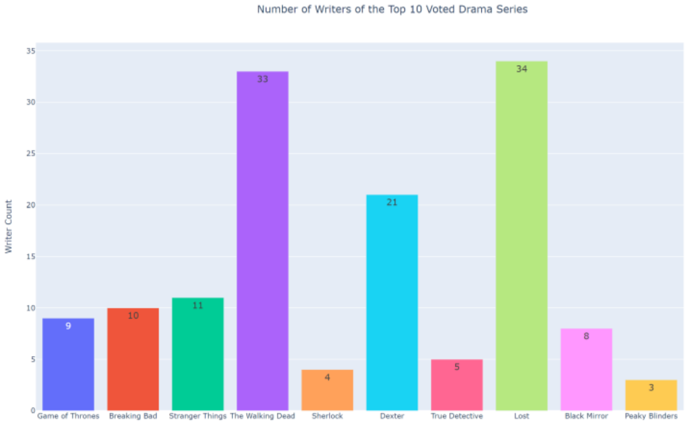

Fig. 11. Number of Writers of the Top 10 Voted Drama Series

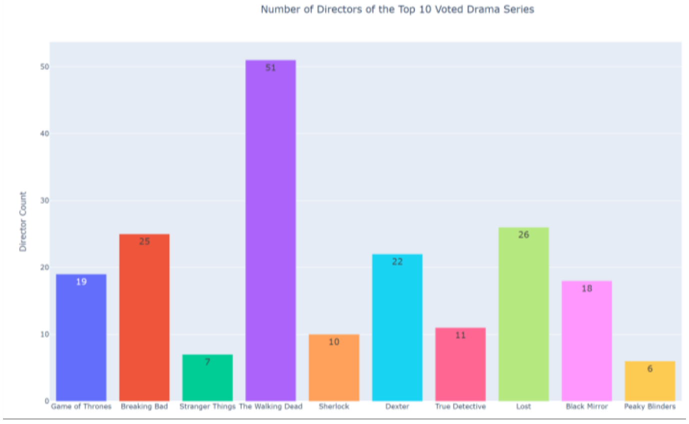

Fig. 12. Number of Directors of the Top 10 Votes Drama Series

The next step was trying to find relations between actors, writers, and directors to see if certain collaborations were a step towards having a successful drama series. For that, we chose the actors, writers and directors that were in more than one of our top 100 drama series. 5 directors, 5 writers and 29 actors were in 2 series. 1 writer and 2 actors (Aaron Paul and Gillian Anderson) were in 3 series. By using a neural network that showed the actors, writers, and directors with the most direct drama TV series' connections (Fig. 13), we were able to see that the only people who worked together more than once on a top drama series were Bob Odenkirk (actor), Jonathan Banks (actor) and Vince Gilligan (writer), who worked on Breaking Bad and Better Call Saul (highlighted by a red circle).

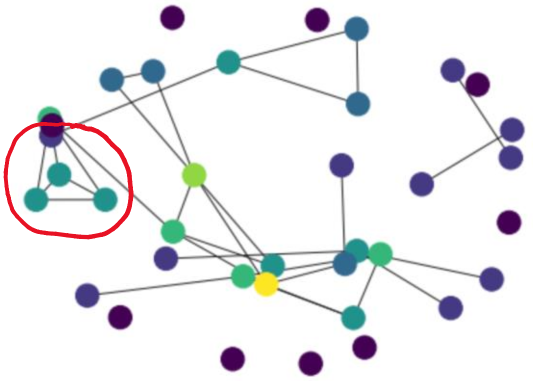

Fig. 13. Actors, writers, and directors with the most direct drama TV Series' connections

Did we find a successful collaboration? We can't know for sure, but our guess is no. Better Call Saul is a spinoff of Breaking Bad, meaning both series are set in the same universe and both actors play the same characters in both series. What we can conclude from this is that the selection of the cast and crew in a drama series is a non-determining factor in the success of a series. More well-known actors and directors might draw a lot of viewers, but it won't be enough to make a series successful.

Next, we tried finding out which of our top 100 drama series were overrated or underrated by comparing their rating with the mean rating of their episodes. We considered a series overrated if its rating was higher than its mean episode rating, i.e., if they were above the green line (Fig. 14)

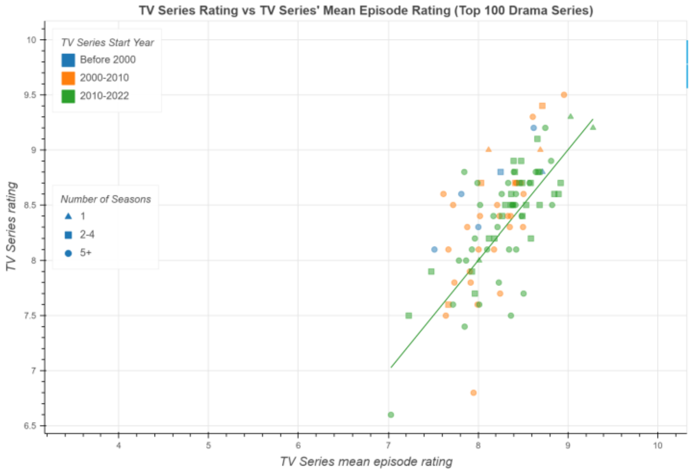

Fig. 14. TV Series Rating vs TV Series' Mean Episode Rating (Top 100 Drama Series)

All of our drama series from the 90's were considered overrated, meaning that their overall rating was higher than their mean episode rating. This might be because people tend to glorify the series from the previous century.

The next step was trying to figure out if drama series tend to drop their rating as the number of seasons increases. We plotted the mean episode rating over number of seasons (Fig.15).

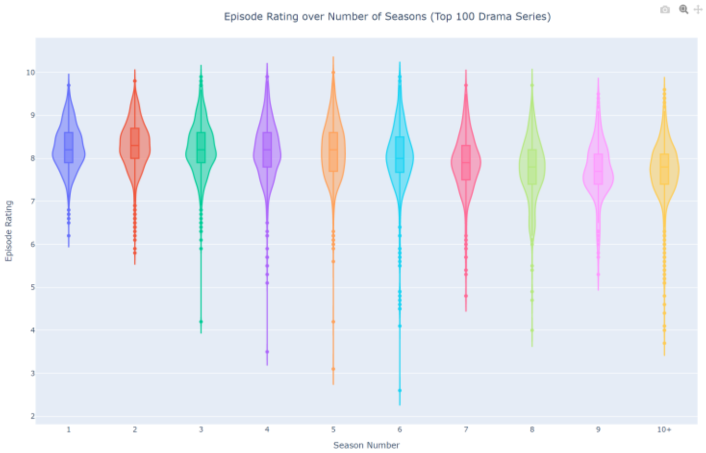

Fig. 15. Episode Rating over Number of Seasons (Top 100 Drama Series)

The mean episode rating is highest around 2-5 seasons and starts falling after that. Based on this plot, generally anywhere between seasons 2 and 5 is the ideal length of a drama TV series.

We also looked at the top 100 drama series' peak ratings separately. We created a plot between final season mean episode rating against mean episode rating amongst all seasons prior to the peak season (Fig. 16). Peak for a series was defined as the season with the highest mean episode rating. Almost all the series with more than 7 seasons had a below average season finale (below the green line), suggesting once again that drama series shouldn't go beyond 5 seasons if they want to keep being successful.

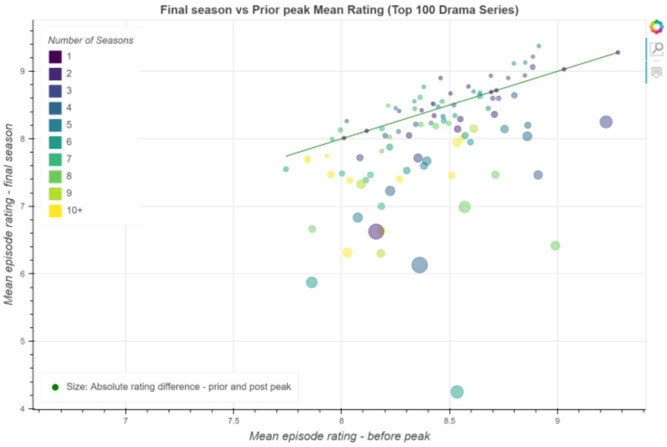

Fig. 16. Final Season vs Prior peak Mean Rating (Top 100 Drama Series

Lastly, we looked for evidence of gender bias in IMDb ratings. We know that TV series targeted towards a female audience tend to have lower ratings on IMDb when compared to male oriented series. We made a plot comparing the rating of drama series with more male leads, more female leads and with an equal number of male and female leads. (Fig. 17)

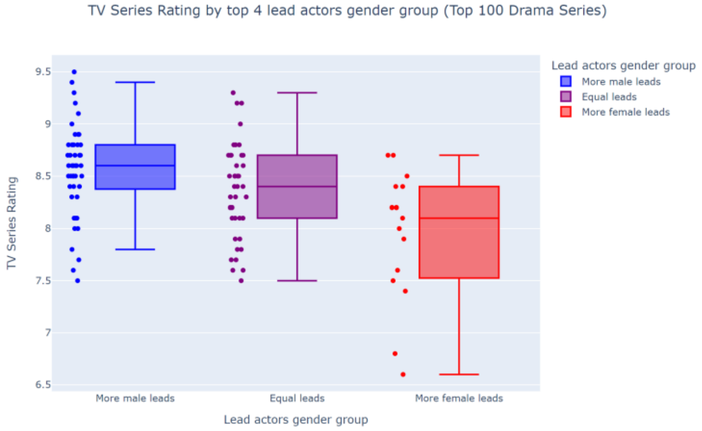

Fig. 17. TV Series Rating by top 4 lead actors gender group (Top 100 Drama Series)

The plot shows that male oriented series are rated higher than neutral and female oriented series, confirming some of the gender bias in rating for female oriented series. It is well documented that men tend to mainly watch series and films made for men by men and a high percentage refuse to watch female oriented series and movies. On the other hand, women are way more open to watching content not necessarily made for them. We made a similar plot but instead of comparing the series' rating, we compared the number of votes. As expected, the results were similar.

V. Conclusions

The goal of this project was to try to understand what makes a TV series successful. We trimmed the data down so that we would only deal with the most popular drama series (top 100). We determined that the most important measurement of a series' success is the number of people who watch it, meaning that a successful series is a popular series. We also determined that, within the top voted drama series, their rating was also a measurement of a series' success.

The hope for this project was to come up with the recipe for the success of a drama TV series with a high number of votes and a high rating on IMDb. We did find the desired length of a successful drama TV series, which is between 2 and 5 seasons if they don't want to drop their rating. We also confirmed the existence of gender bias in IMDb ratings – male oriented series have a higher rating and a higher number of votes than female oriented series. But does this mean that a male oriented drama series with a number of seasons between 2 and 5 will automatically be successful, i.e., will have a high number of votes and a high rating? Our guess is no. It can certainly help its success, but it would take more than that to make it successful.

How a series becomes successful is still a mystery. Several factors must align. Basically, it's like hitting the jackpot, because the truth is, the majority of the most successful series of all time had an unknown cast and crew and brought in something new, something special, something that people had never seen before.