Project Proposal

Regression Rockstars - James Cai, Steph Reinke, Sarah Wu, Michael Zhou

Introduction

Scooby-Doo is a popular animated TV show that follows a group of teenagers and a talking Great Dane, Scooby-Doo, as they solve mysteries involving supernatural monsters and creatures. Each episode typically involves seeking and scheming to find the villain, ending with a dramatic unmasking of the monster. The show focuses on themes of friendship and teamwork. The show originally aired on CBS from 1969 - 1976, but there has been many subseries and reboots since.

We are interested in researching Scoody-Doo IMBD ratings because we all enjoyed Scooby-Doo in our childhoods. We also think that finding certain predictors of animated TV series ratings is useful for the entertainment industry. Specifically, our findings could be useful to anyone looking to create an animated TV series and wanting to know what aspects make up a successful episode. In the paper, "Determining and Evaluating The Most Popular Cartoons Among Children Between 4 and 6 Years of Age" published in 2017, the authors criticize the use of violence, vulgar language, and horror music in Scooby-Doo (Başal et. al. 2017), yet we can't ignore the huge impact and popularity of Scooby-Doo. In 2013, Scooby-Doo was ranked the fifth greatest cartoon of all time (TVGuide 2013). If Scooby-Doo continues to create spin-off shows, our findings about what makes a successful episode could inform their future episodes as well.

Our primary research question is what factors best explain the variability in the IMBD scores of Scooby-Doo episodes? In other words, what elements tend to contribute to a successful episode? We want to investigate how predictor variables like `monster.amount`, `engagement`, character that unmasks the villain (`unmask.fred`, and such for all five characters part of the main group), `network`, and more, adequately explain the variability in IMBD ratings. We hypothesize that episodes with a higher monster count will have a better rating, since we think that there is more action and suspense in episodes with more monsters. We also hypothesize that the higher engagement, the worse the rating will be. This is because we think that people are more likely to write a review online when they dislike the episode rather than if they liked the episode. We think that episodes where Fred unmasked the villain will have a higher rating since he is the leader of the group and thus, we think that people will be more drawn to him. Finally, we think that episodes that aired on Cartoon Network will have a better rating, since

we think that Cartoon Network has the ability to generate more positive responses since they specialize in cartoons and are pretty well-known. In our analysis, we would like to explore the interaction between these variables as well as consider other predictors in the dataset.

Data description

This Scooby-Doo data was found on the TidyTuesday database on Github. The data originally comes from Kaggle and was manually aggregated by user Plummye in 2021. The curator took roughly one year to watch every Scooby-Doo iteration and track every variable in this dataset. It is noted that some of the values are subjective by nature of watching, but the original data data curator tried to keep the data collection consistent across the different episodes.

Each observation represents an episode from a rendition of the Scooby-Doo franchise up until February 25, 2021, including movies and specials. The variables that were measured include the series and episode name (which we will not use as predictor variables), network aired on, IMDB score, engagement (represented by number of reviews on IMDB), and many details about what happened in each episode, such as how many monsters appeared, which character captured and unmasked the monster, the terrain of the episode, and more. There is a mix of both numerical and categorical characteristics.

Initial exploratory data analysis

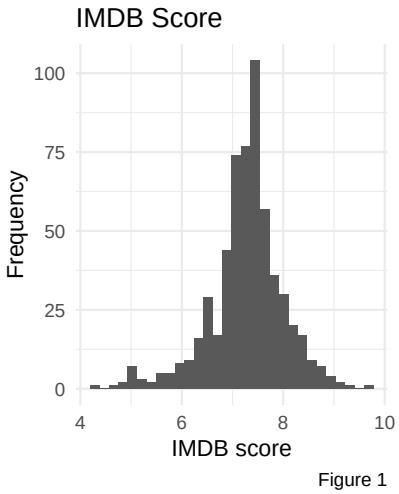

Figure 1

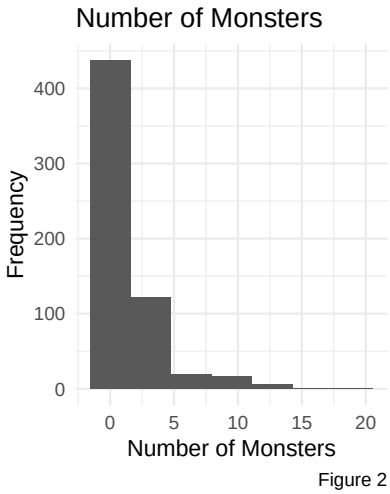

Figure 2

Figure 1: The distribution of our response variable IMDB scores, `imdb`, is unimodal and roughly symmetrical. The mean is 7.278 and the standard deviation is 0.732. The minimum is 4.2 and the maximum is 9.6. There does not seem to be any significant outliers.

Figure 2: The distribution of the number of monsters, `monster.amount`, is unimodal and right skewed. The median is 1 monster and the IQR is 1 monster. The minimum is 0 monsters and the maximum is 19 monsters. There are a few episodes with notably high amounts of monsters, with 5 episodes having 13 or more monsters.

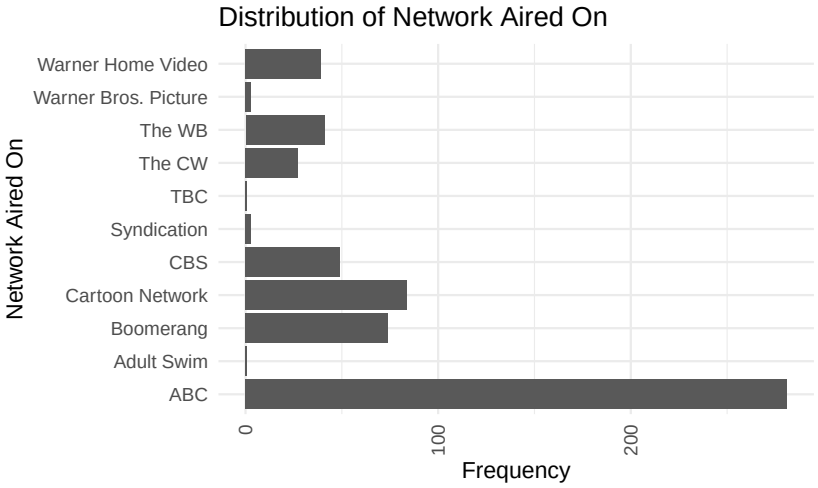

Figure 3

Figure 3: The distribution of the network the episode aired on, `network`, shows that a good majority of the episodes aired on ABC. There were also considerable amounts of episodes that aired on Cartoon Network and Boomerang, while there are also networks that aired very few episodes, such as TBC and Adult Swim.

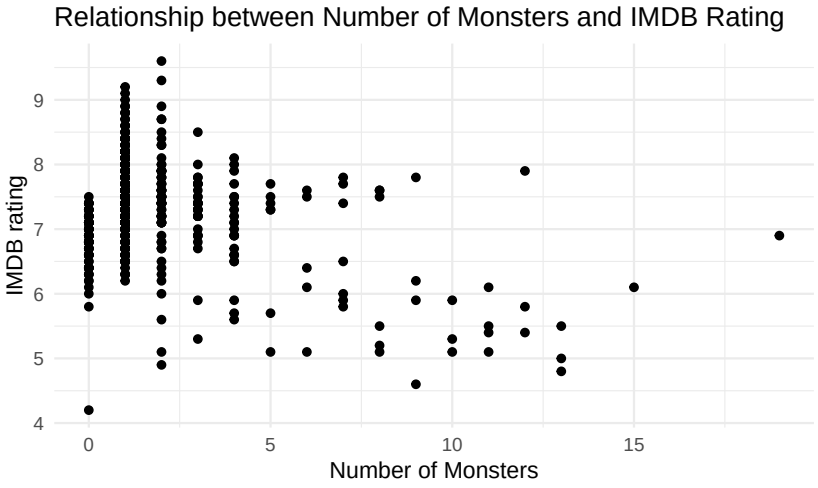

Figure 4

Figure 4: The relationship between the number of monsters and the IMDB score is moderate, negative, and linear. Omitting NULL values in `monster.amount` and `imdb`, the correlation is -0.350. It seems that as the number of monsters increase, the IMDB score tends to decrease,

on average. However, as seen earlier, the median of the distribution of the number of monsters is 1, so as we increase the number of monsters, there are less and less observations, which makes the relationship between the two variables hard to observe as the number of monsters increase.

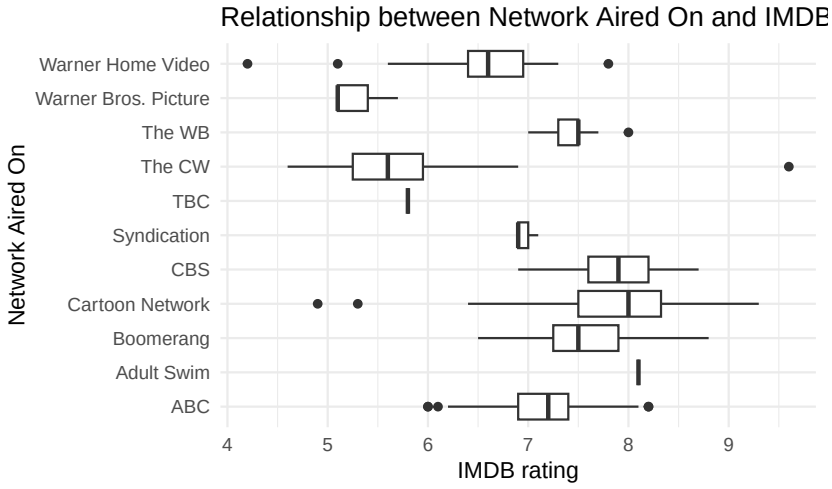

Figure 5

Figure 5: From the distribution of the different boxplots for each network, we observe that many of the interquartile intervals of the boxplots overlap, meaning that their IMDB ratings are quite similar. It seems that Cartoon Network generally received the best ratings, while Warner Bros. Picture and The CW generally received the worst ratings. We also observe a few outliers in the distribution of IMDB ratings for some networks, such as Warner Home Video and ABC. Many of the networks have IMDB ratings that are pretty symmetrical, as the line representing the median is close to the middle of the box, such as in the case of CBS and The CW, but some are pretty skewed, such as in the case of Syndication, The WB, and Warner Bros. Picture.

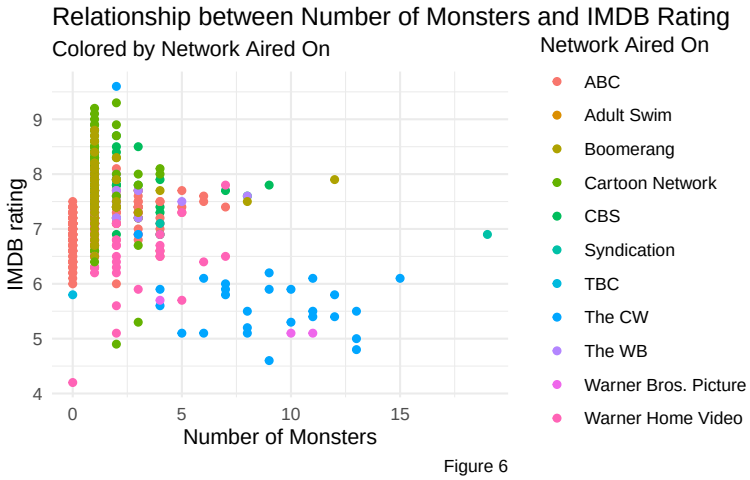

Figure 6

Figure 6: In particular, we tried to see if we could find signs of interactions between the number of monsters in an episode and the network that the episode aired on. For example, were there any networks that particularly included more monsters or less monsters in each episode, and how did this affect the IMDB rating? From the scatterplot, it seems that The CW, in particular, included a higher amount of monsters in each episode, but these episodes did not receive the highest ratings and the ratings were pretty consistent throughout. Boomerang seemed to have many episodes with 1 monster and these episodes received a range of ratings from around 6 to 9. CBS seemed to have the most success when it limited the number of monsters to 2. We are interested in further exploring this interaction in our analysis.

For further data cleaning, we plan on removing the observations that are NULL for predictors that we are interested in. For the variables we are interested in, many of the NULL values appear for the variables surrounding the monster (`unmask.fred`, `unmask.daphnie`, and more), which occur when `monster.amount` is equal to 0 . Since there are 603 total observations in our dataset, we are not too worried about the reduction in sample size after removing null observations. We also plan on creating a categorical variable, `villain_unmask`, that sums up who unmasked the monster in one singular column instead of having separate columns for each main character. This variable would take values of Fred, Daphnie, Velma, Shaggy, and Scooby–the five main characters in the group.

Analysis approach

We are currently planning to use four potential predictors: `monster.amount`, a quantitative variable for the number of monster in the episode, `engagement`, a quantitative variable for the number of reviews on IMDB for an episode, `villain_unmask`, a categorical variable that we will make that takes the values Fred, Daphnie, Velma, Shaggy, or Scooby and represents the character who unmasked the villain, and `network`, a categorical variable that represents the network the episode aired on, to predict the IMDB score of each episode, which is a quantitative variable. We are open to exploring other variables as well, depending on what we find from these three variables first.

We plan to use multiple linear regression for our analysis as we hope to incorporate multiple variables in our model. We plan on testing different models, experimenting with interaction terms and number of included terms, and we will choose our final model based on metrics such as R^2_{adj}, AIC/BIC, and more, while ensuring that there is no multicollinearity in our model using VIF.

Data dictionary

The data dictionary can be found here.