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 episodes 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 aired on CBS from 1969 - 1976, but there has been many subseries and reboots.

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 (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.

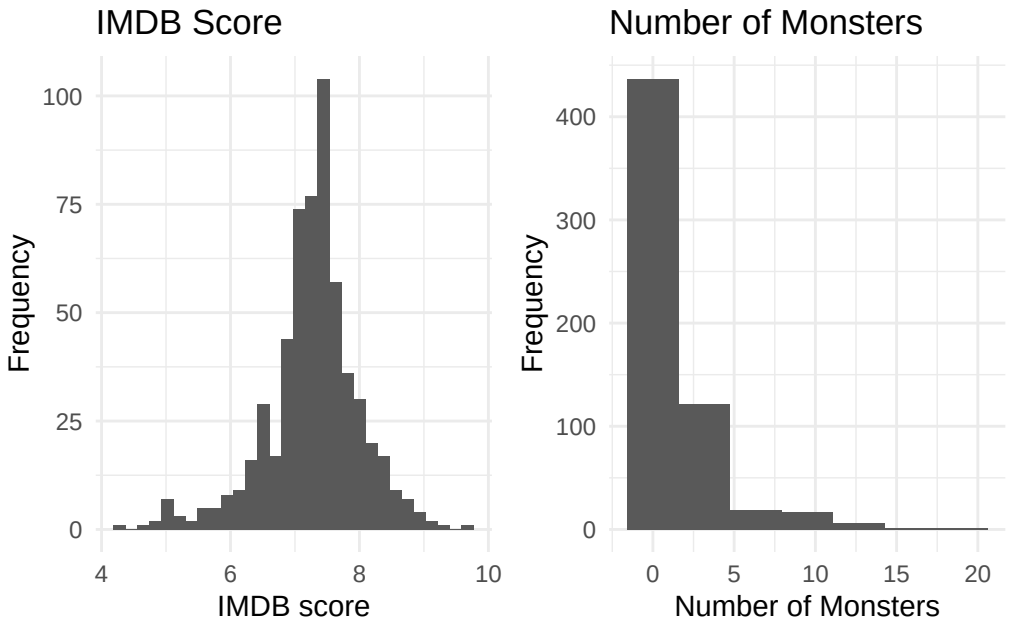
Our primary research question is what factors best explain the variability in the IMBD scores of Scooby-Doo episodes? What elements tend to contribute to a successful episode? We want to investigate how predictor variables like `monster_amount`, `engagement`, character that unmasks the villain (`unmask.fred`, and so on for all characters part of the main group), and such 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 do not like something rather than if they liked the episode. Finally, 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. We would like to explore the interaction between these variables as well, 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.

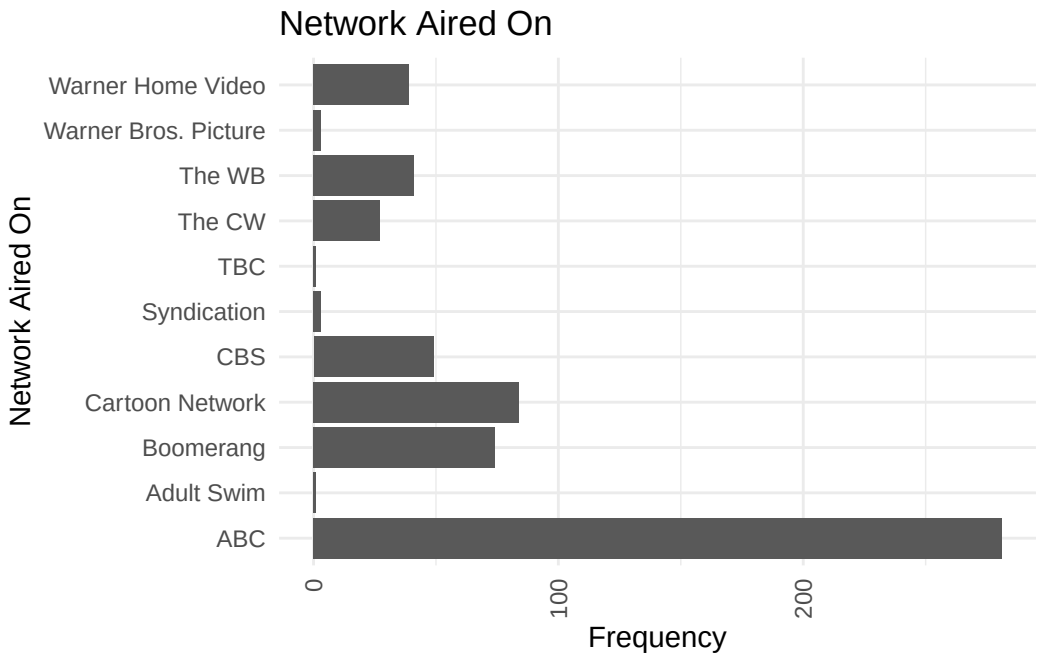
Each observation represents an episode from a rendition of the Scooby-Doo franchise up until February 25, 2021, including movies and specials. The variables that were measured include the series and episode name, network aired on, IMDB score, engagement (represented by number of reviews on IMDB), and many details about what happens in each episode itself, such as how many monsters appeared, which character captures and unmasks the monster, the terrain of the episode, and more. There is a mix of both numerical and categorical characteristics.

Initial exploratory data analysis

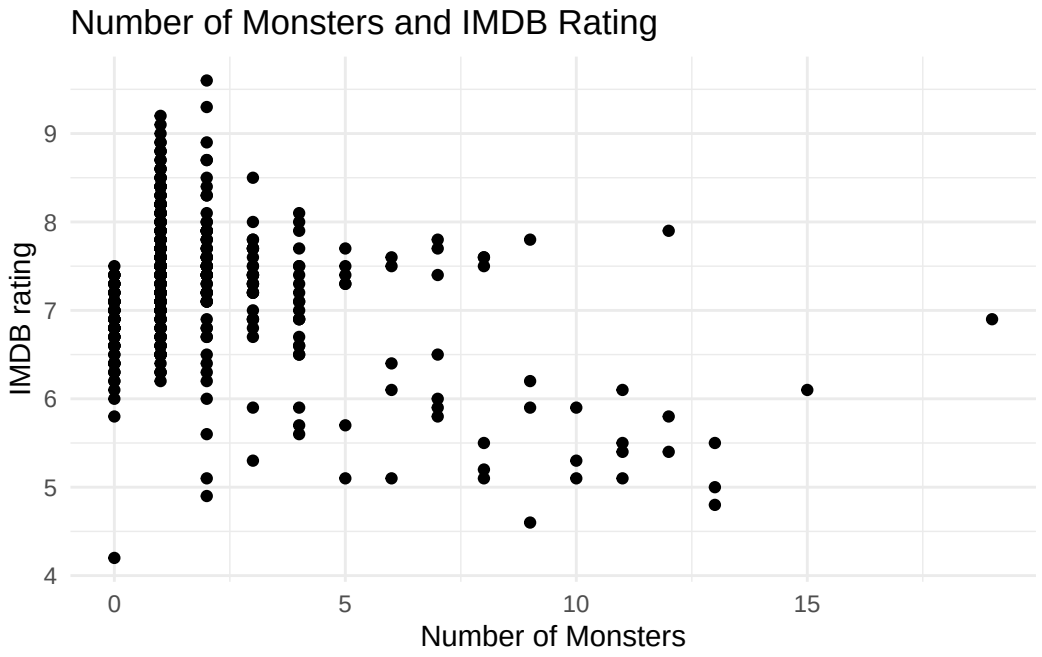


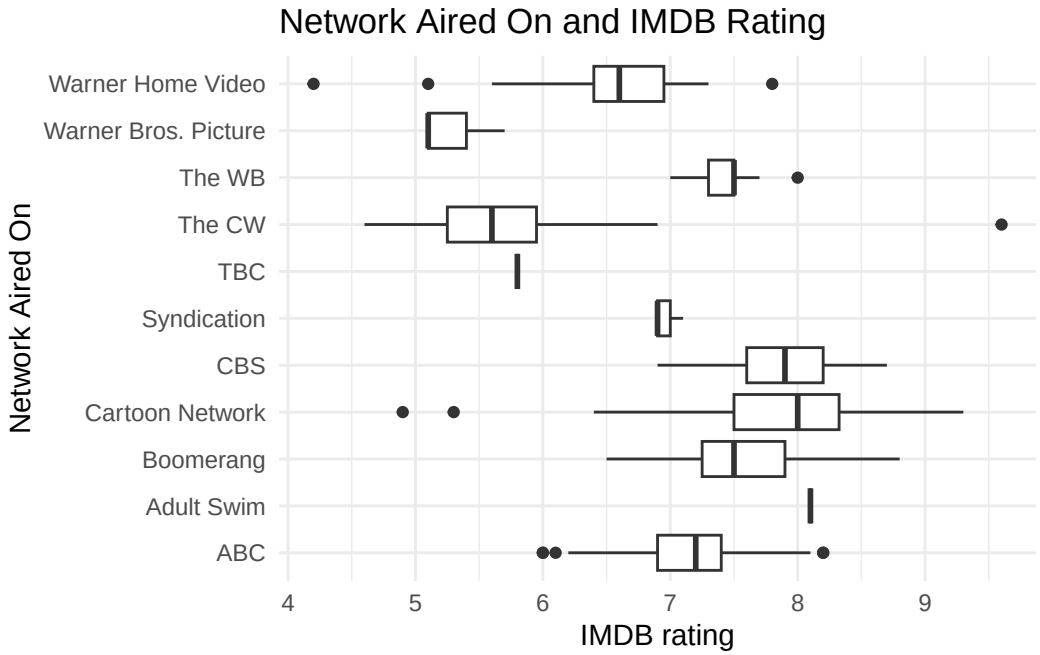
The distribution of IMDB scores is unimodal, roughly symmetrical and has a range from 4 to 10.

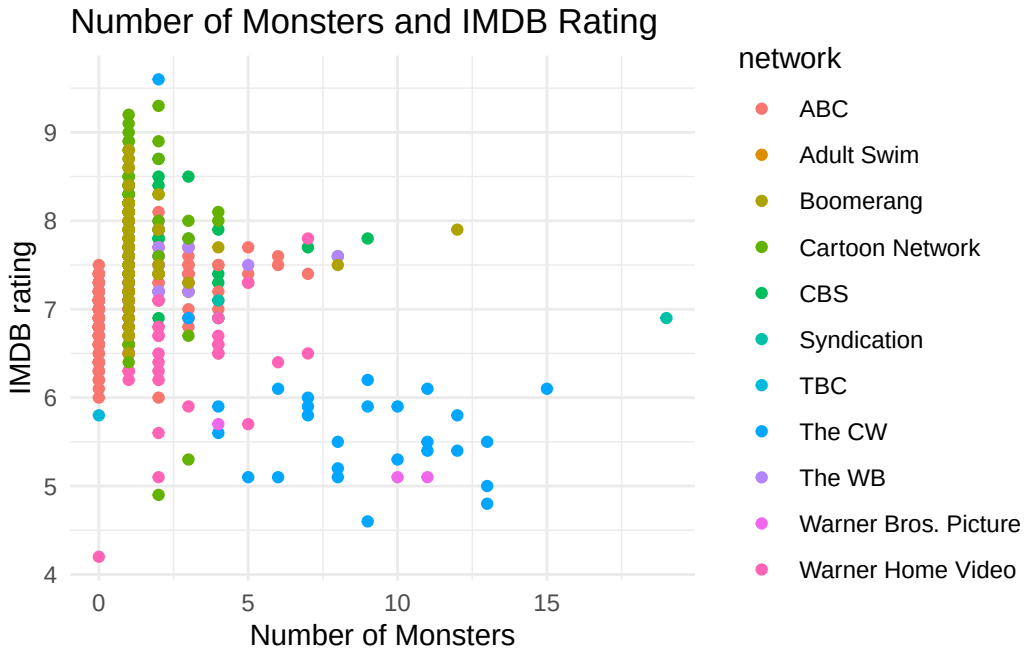
The distribution of the number of monsters is right skewed with a center around 0.



The distribution of Networks aired on shows that there is a clear majority with ABC.







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 three 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; and `villain_unmask`, a categorical variable that we will make that takes the value Fred, Daphnie, Velma, Shaggy, and Scooby for the character who unmasks the villain, 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.