Project Proposal

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

```
library(tidyverse)
library(tidymodels)
library(knitr)
library(patchwork)

scoobydoo <- read_csv("data/Scooby-Doo Completed.csv")
```

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. If Scooby-Doo continues to create spin off shows, this information could inform their future episodes as well–what elements of the show tend to receive better responded and a higher IMBD rating?

Our primary research question is what factors explain the variability in the IMBD scores of Scooby-Doo episodes?

We want to investigate how monster_amount, engagement, and who unmasked the villain (unmask_fred, and so on for all the main group) adequately explain the variability of the IMBD rating. We think that the more monsters the better the rating will be, because we think there is more action and suspense in episodes with more monsters. We also think that the higher engagement the worse the rating will be, because we think that people are more likely to write a review online when they do not like something than if they do. We think that

episodes where Fred unmasked the villian will have a higher rating because we think that he is the leader of the group and thus, we think that people will be more drawn to him.

Data description

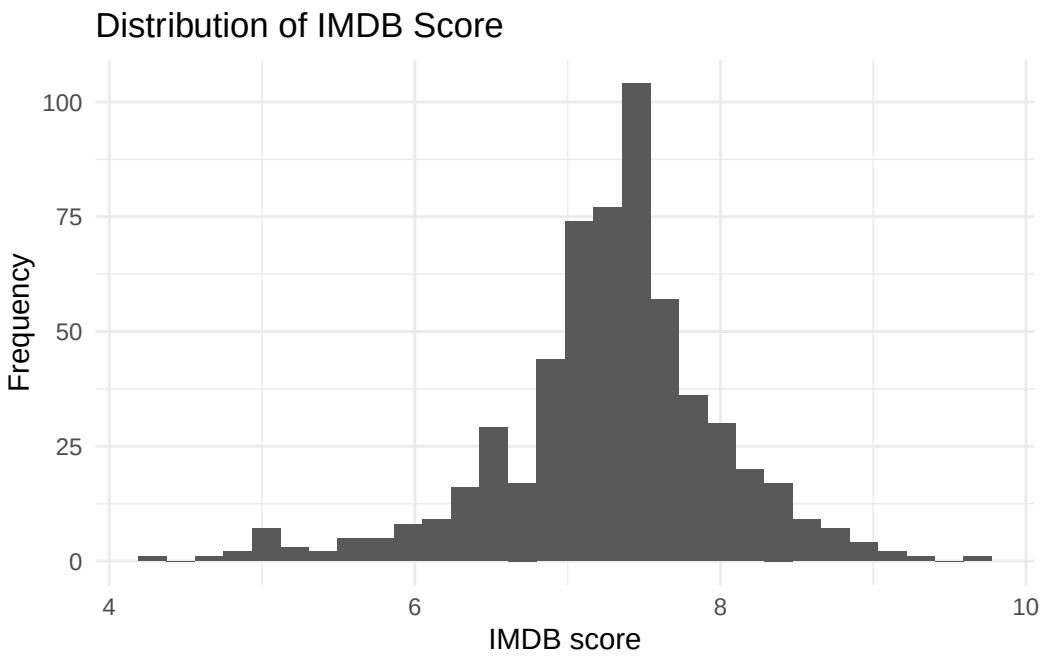
The Scooby-Doo data was found on the TidyTuesday database on Github. The data comes from Kaggle and was manually aggregated by user `plummye` in 2021. The curator took roughly one year to watch every Scooby-Doo iteration and track every variable in this dataset. It is noted that some of the values are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

Each observation is an episode from a rendition of the Scooby-Doo franchise up until February 25, 2021. The variables that were measured include the series/episode name, IMDB score, and many details about what happens in the episode itself, such as how many monsters appeared and which character captures/unmasks them.

Initial exploratory data analysis

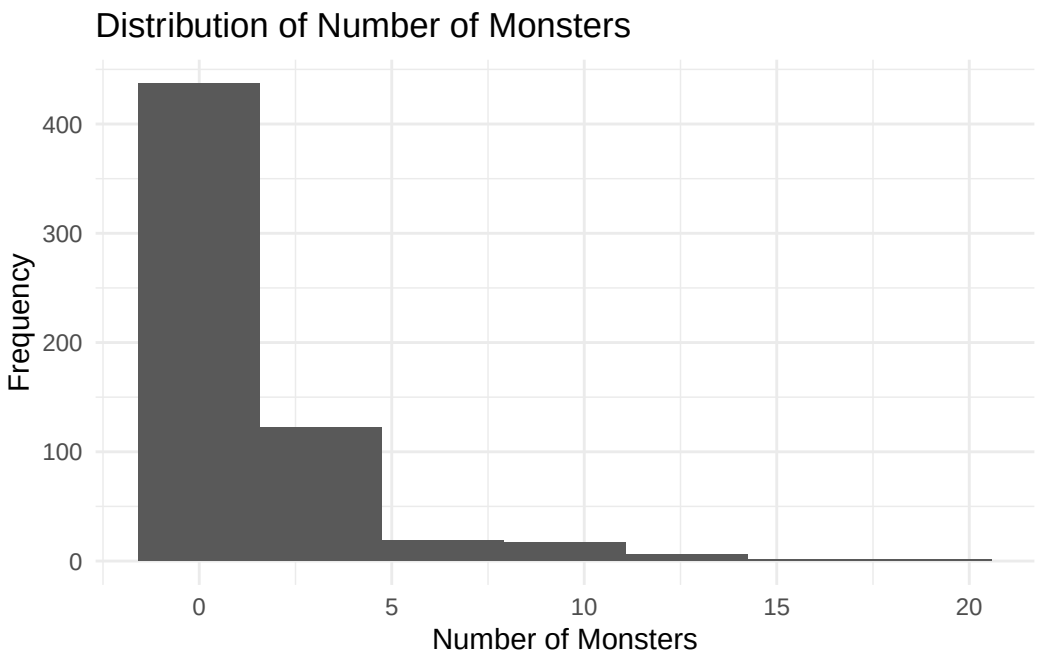
```
scoobydoo <- scoobydoo |>
  mutate(imdb = as.numeric(imdb))

scoobydoo |>
  ggplot(aes(x = imdb)) +
  geom_histogram() +
  labs(x = "IMDB score",
       y = "Frequency",
       title = "Distribution of IMDB Score") +
  theme_minimal()
```



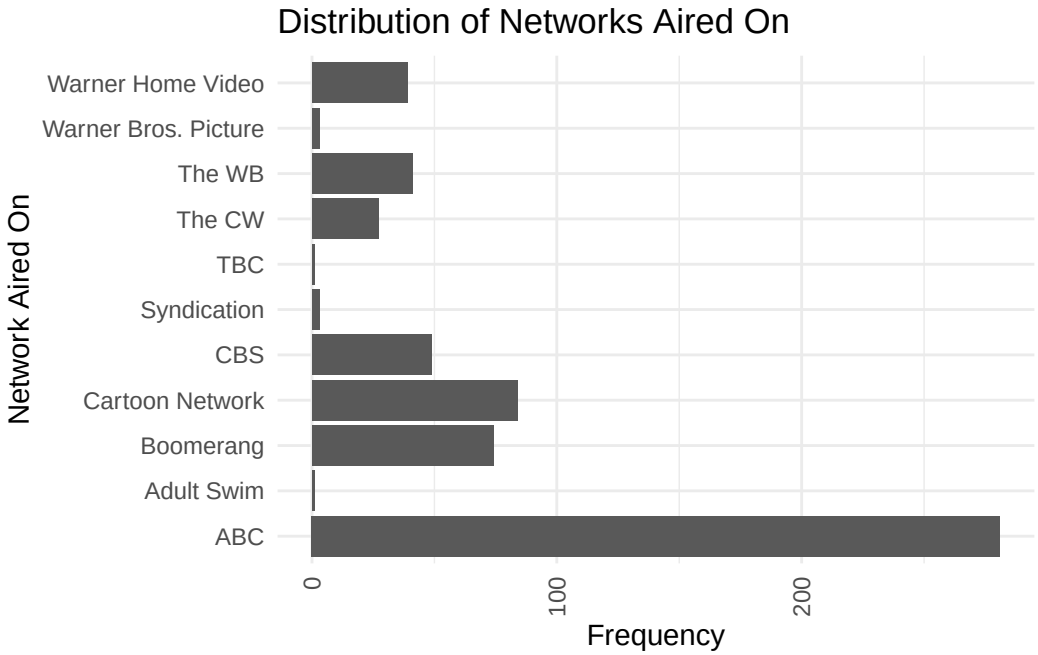
The distribution of IMDB scores is unimodal, roughly symmetrical and has a range from 4 to 10.

```
scoobydoo |>
  ggplot(aes(x = monster.amount)) +
  geom_histogram(bins=7) +
  labs(x = "Number of Monsters",
       y = "Frequency",
       title = "Distribution of Number of Monsters") +
  theme_minimal()
```



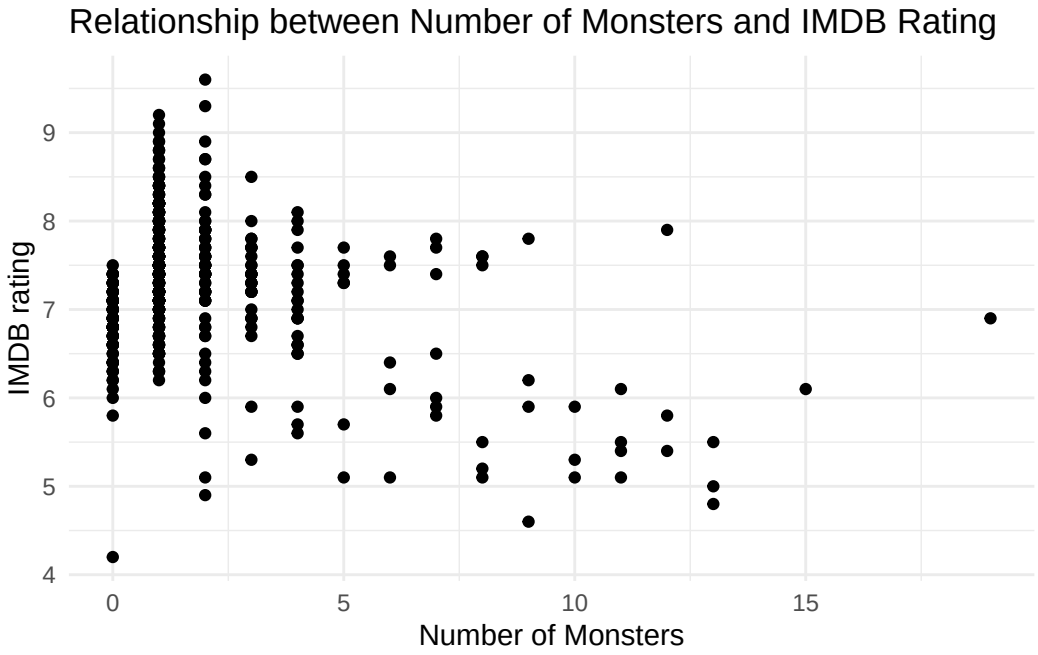
The distribution of the number of monsters is right skewed with a center around 0.

```
scoobydoo |>
  ggplot(aes(x = network)) +
  geom_bar() +
  coord_flip() +
  labs(x = "Network Aired On",
       y = "Frequency",
       title = "Distribution of Networks Aired On") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

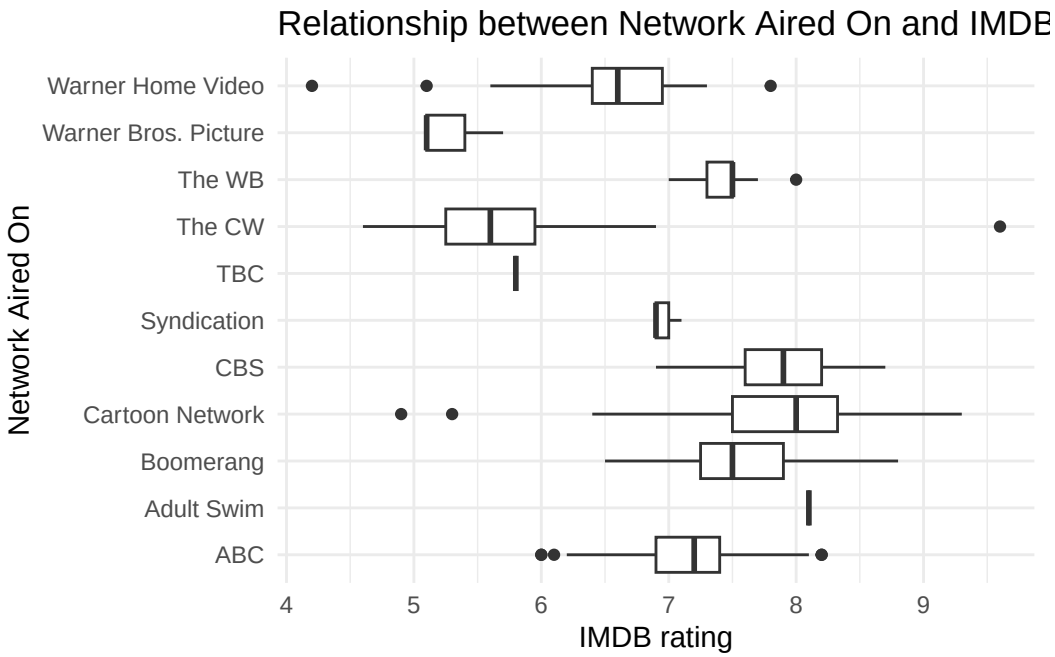


The distribution of Networks aired on shows that there is a clear majority with ABC.

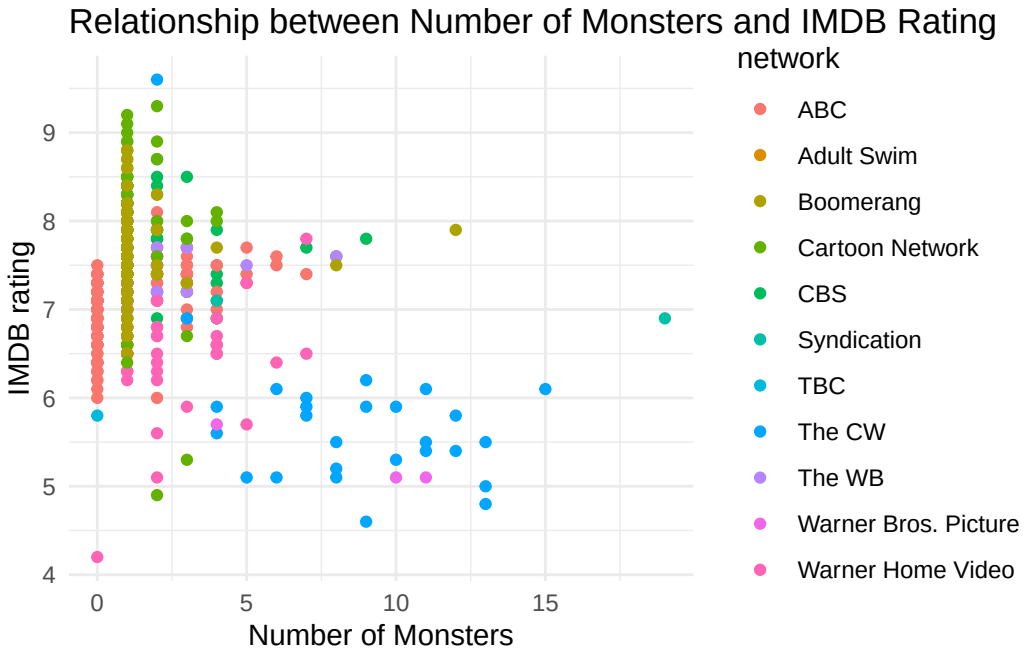
```
scoobydoo |>
  ggplot(aes(x = monster.amount, y = imdb)) +
  geom_point() +
  labs(x = "Number of Monsters",
       y = "IMDB rating",
       title = "Relationship between Number of Monsters and IMDB Rating") +
  theme_minimal()
```




```
scoobydoo |>
  ggplot(aes(x = network, y = imdb)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = "Network Aired On",
       y = "IMDB rating",
       title = "Relationship between Network Aired On and IMDB Rating") +
  theme_minimal()
```




```
scoobydoo |>
  ggplot(aes(x = monster.amount, y = imdb, color = network)) +
  geom_point() +
  labs(x = "Number of Monsters",
       y = "IMDB rating",
       title = "Relationship between Number of Monsters and IMDB Rating") +
  theme_minimal()
```



Analysis approach

We are planning to use three potential predictors: monster_amount, a quantitative variable for the number of monster in the episode; engagement, a quantative variable for the number of reviews on IMDB for an episode; and villain_unmask, a categorical variable 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 obviously is a quantitative variable.

We plan to use multiple linear regression for analyze the data.

Data dictionary

The data dictionary can be found here [Update the link and remove this note!]