Proposal : IMBd movie data analysis report

Group 10

Statement of problem

The Internet Movie Database (IMDb) is one of the largest online databases and authoritative sources for movie, TV, and celebrity content. IMDb's registered users can vote for any movie, and then IMDb takes all the votes to calculate the weighted average rating of a particular movie. In addition, IMDb offers a rating scale that allows users to rate films with a range of different appreciations. The IMDB score method is a 1-10 star movie rating system, where the better the movie, the higher the score is given, such as 1 means that the movie was terrible, and 10 means the users think it is excellent. However, many factors such as storyline, experienced directors, and famous actors are considered for creating good movies but do not ensure a good rating on IMDb. Additionally, the methodology for generating the IMBd movie rating remains unclear, including which factors contribute to the ratings and how these ratings are measured.

Literature Review

Social media and the rapid growth of the film industry worldwide have redefined how consumers get information by offering a variety of reviews such as user ratings, user reviews, and critic comments. Movie rating is a great way to share users' opinions, keep track of movie records, and receive movie recommendations. Also, the ratings are seen as good predictors of the quality of the movie and determine whether or not the movie is worth recommending. Thus, a good understanding of these ratings impact sales and the importance of such impacts would provide practical implications for movie producers, distributors, and consumers.

Agencies and data scientists alike have been trying to extract insights from streaming and behavioral data to target the correct customers and recommend movies at the right time. Machine learning insights can improve the success of streaming companies in many ways. Unfortunately our analysis can’t cover everything but we will try to understand what metrics (user ratings, reviews, and comments) goal of our analysis is to

Research Questions

Upon analyzing this data, we intend to answer the following research questions:
(1) Which factors contribute most to the IMDb score?
(2) What is the impact of votes and revenue on IMDb score?
(3) How do the categories affect rating?
(4) Are there any description keywords that are correlated with a higher IMDb score?

In order to answer these questions, we will create models that will predict the correlation of the factors and how they contribute to the score.

Data description

To obtain a good data set several datasets were reviewed and the team felt collectively the data from data source [IMDB Data from 2006 to 2016 - dataset by promptcloud | data.world] will be a good source to pursue this analysis. The Raw data was downloaded and we followed the following to explore, analyze, clean and finalize the data.

The dataset contains 12 variables for 1000 movies. This is a summary dataset that contains all the data of IMDB from the year 2006 to 2016. The era where the greatest movies were made and what was the start of the Marvel Cinematic Universe and various other movies that went on to change the world. The dataset includes rank, title, genre, description, director, actors, year, runtimes(minutes), rating, votes, revenue(millions) and metascores for reviews that may impact a movie IMDB scores. “imdb_score” is the response variable while the other variables are possible predictors.

Initial exploring of the data found the 2 features Revenue and Metascore have missing values. (please refer to the R code in Appendix A which is used to get the following diagram in Figure 1)

To understand the data we extracted the numeric features in the dataset as shown in figure 1 and generated Correlation heat maps as shown in figure 2 below. (Please refer to the R code in Appendix B)

Figure 1: Numeric features of each variables

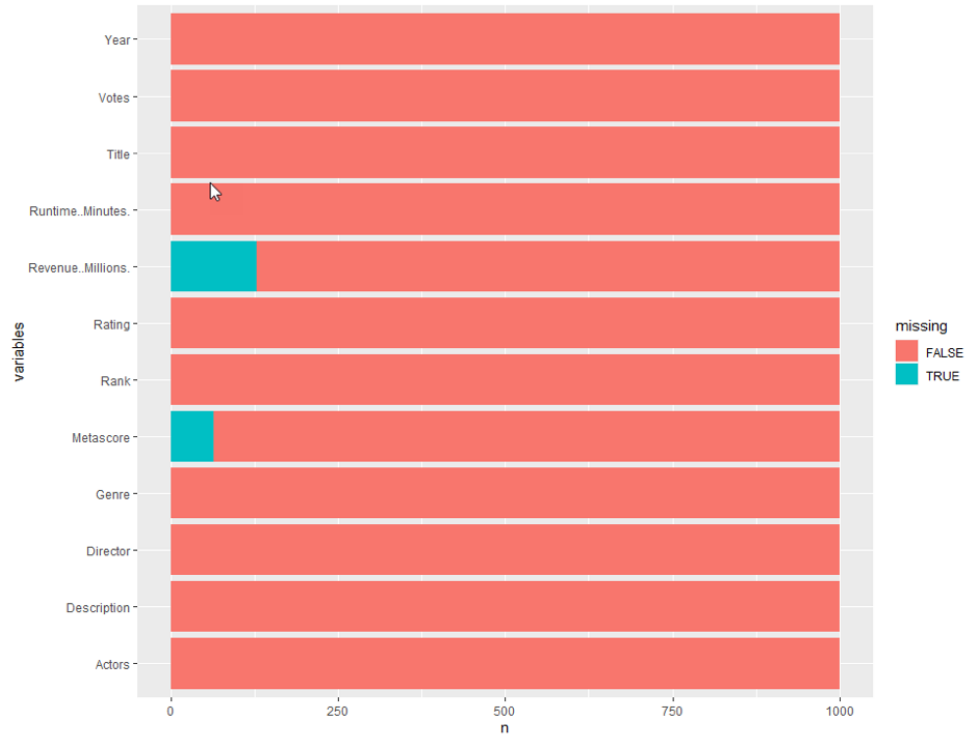


Figure 2 : Correlation heat maps of each variables

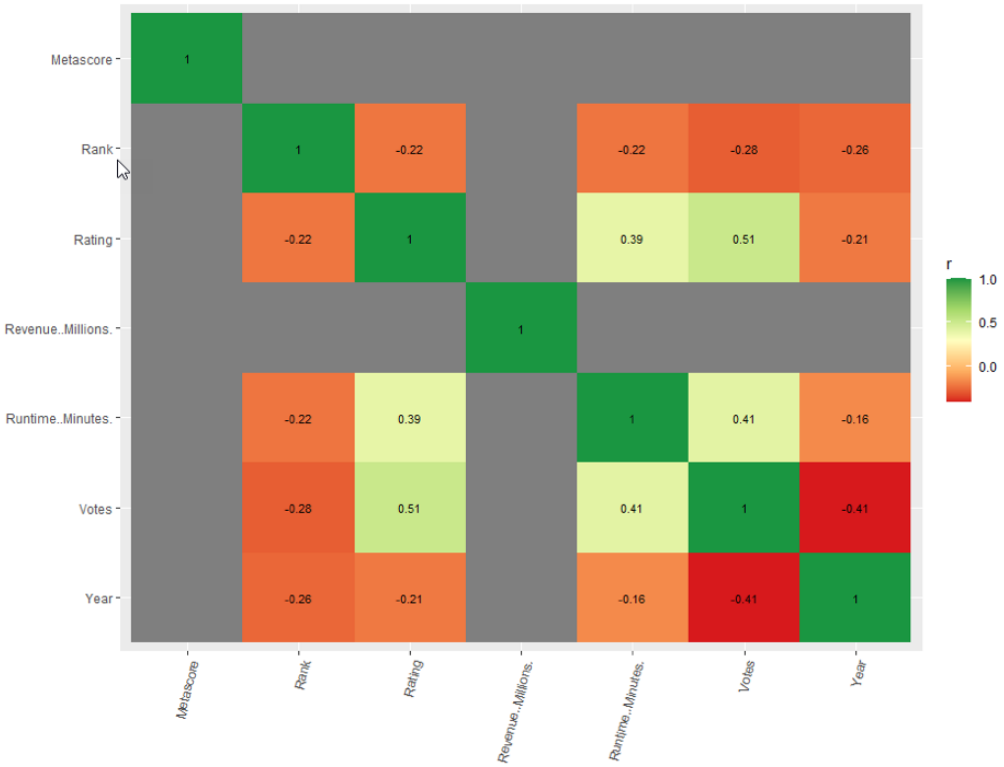
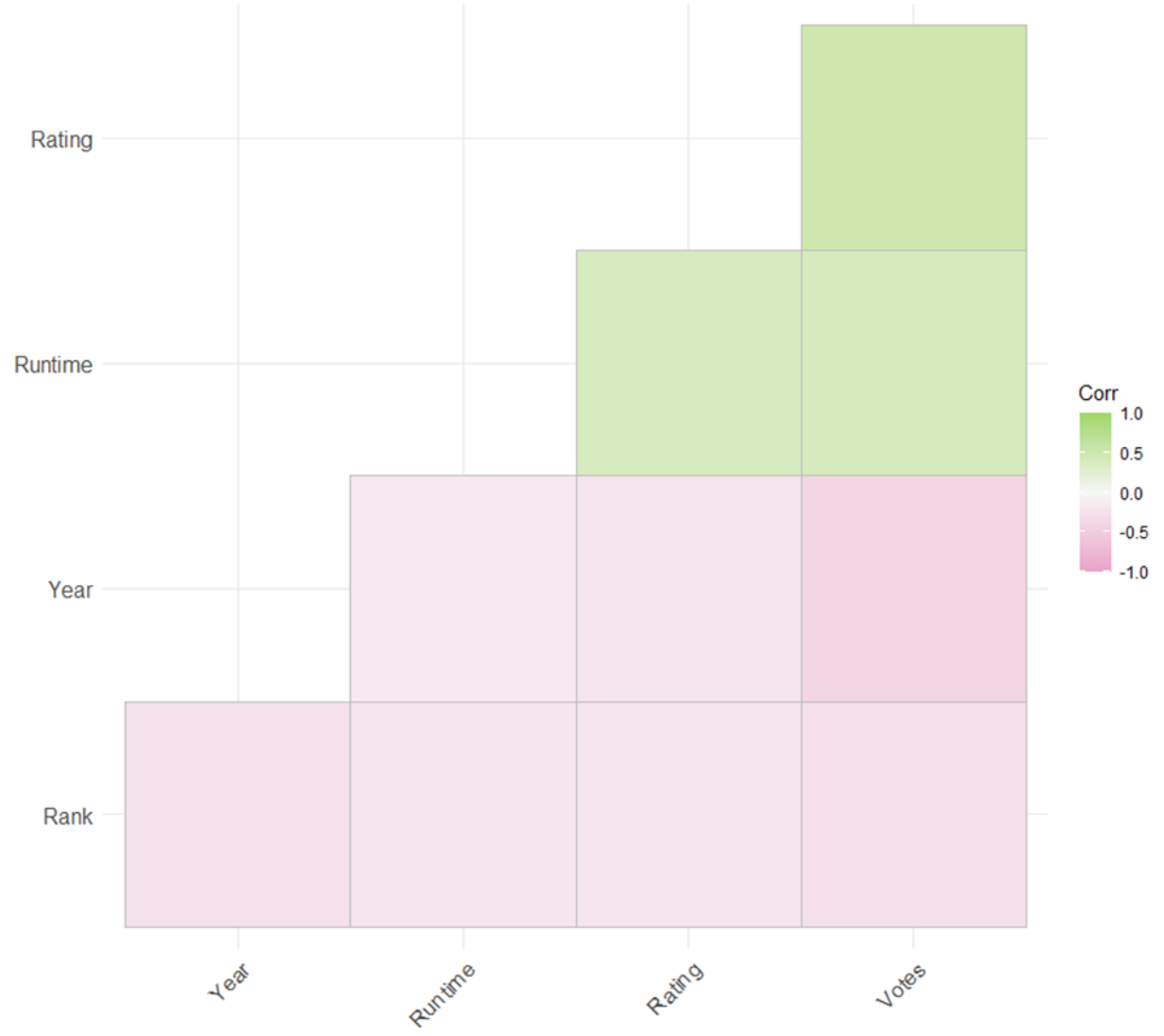
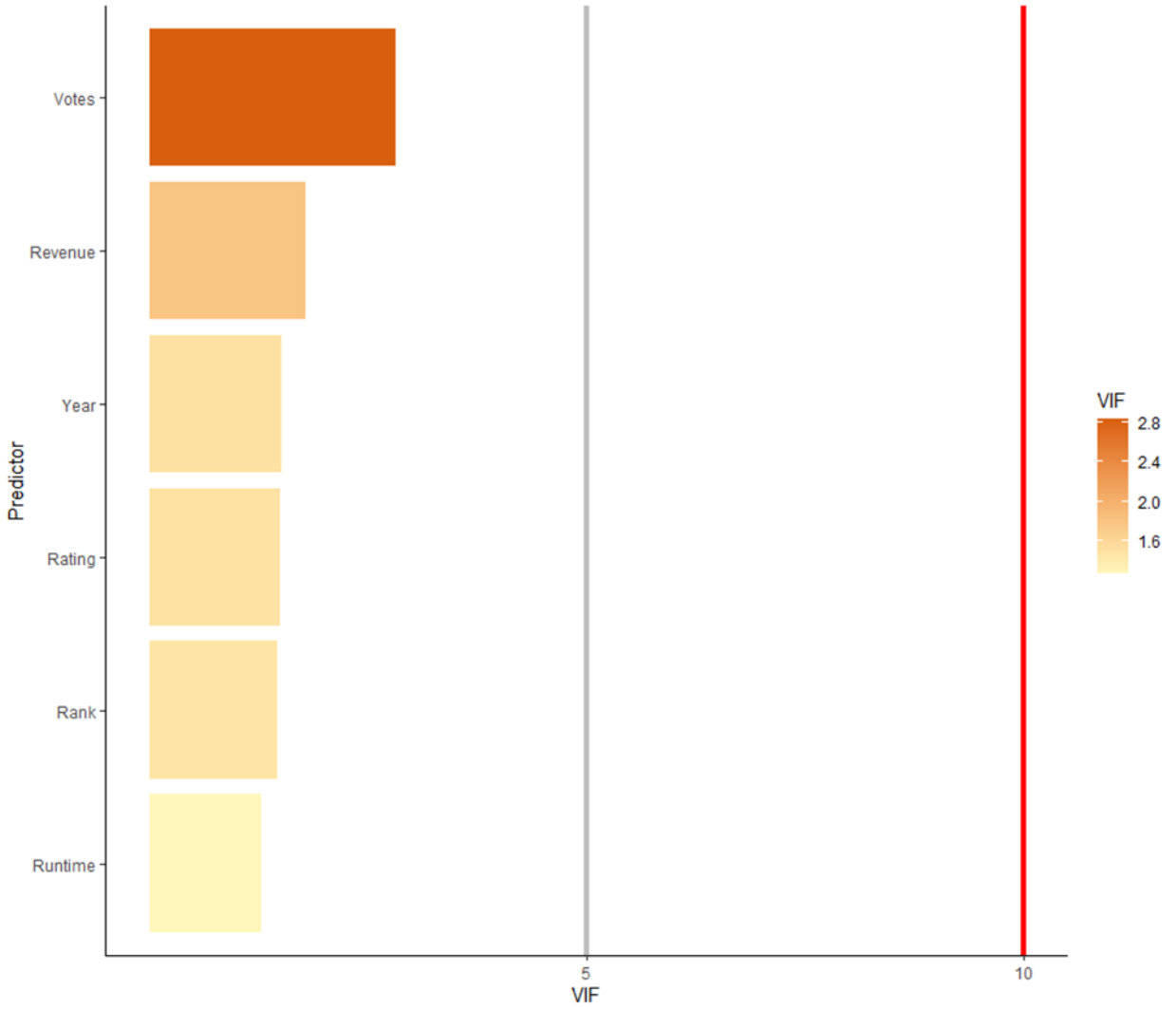


Figure 3 : Correlation heat maps



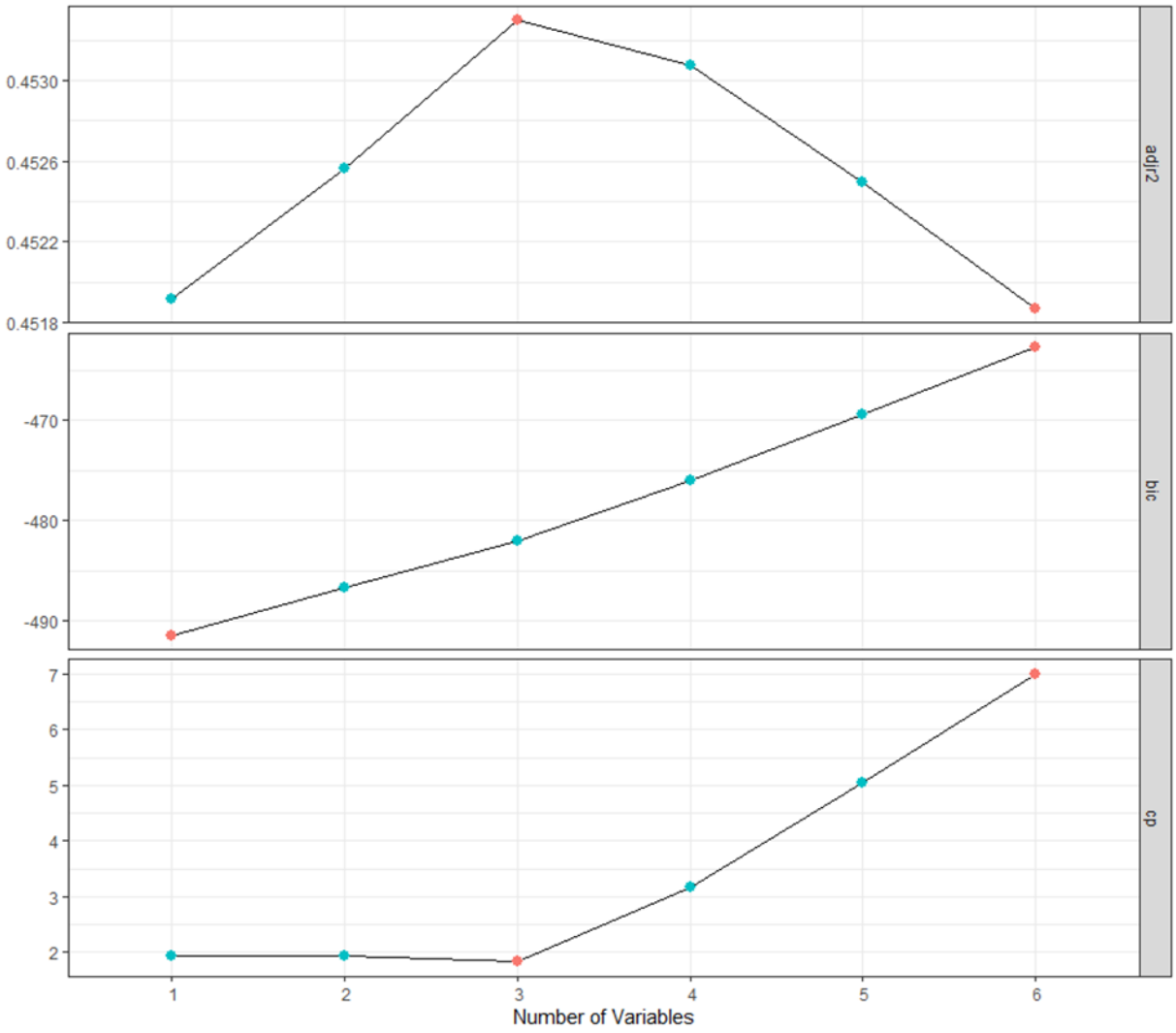
In an effort to understand the data a Variance Inflation Factor (VIF) Map was also created which indicates variance of an independent variable is influenced or inflated, by its correlation with the other independent variables.

Figure 4 : Variance Inflation Factor (VIF) Map



Best-subset selection aims to find a small subset of predictors, so that the resulting linear model is expected to have the most desirable prediction accuracy.

Figure 5 : Best subset selection



Data Preparation (cleaning data)

In the process of cleaning and preparing the data it was identified the following missing values Metascore had 64 NA values . The missing data for Metascore is substituted with the main of Metascore from the movies with same ratingith mean of metascore , based on the similar rating

Missing revenue feature we had 129 rows of missing data. We substituted the revenue with mean revenue for movies with same rating es[Please refrer to R script for code and methods used]

Conclusion

This project aims to see which factors contribute the most to the IMDB score and provide analytical insights into movies. To accomplish our goal, we performed the analytical process as follows:

1. Initial Exploration - Load a dataset and understand the variables
2. Data Preparation - Clean data by removing duplicate and useless columns and inputting the missing value
3. Exploratory Data Analysis & Virtualization - plotting and calculating variables' frequency to understand the relationship better and identify important factors
4. Predict movie rating with Machine Learning
5. Final Conclusion and Recommendations

Appendix A

```
imdb  %>%
  summarise_all(list(~is.na(.)))%>%
  pivot_longer(everything(),
               names_to = "variables", values_to="missing") %>%
  count(variables, missing) %>%
  ggplot(aes(y=variables,x=n,fill=missing))+
  geom_col()
```

Appendix B

```
#-----------------------------------------------------------------------------------------
# Correlation heat maps
#-----------------------------------------------------------------------------------------
library(ggcorrplot)
ggcorrplot(cor(imdb_num),
           method = 'square',
           type = 'lower',
           show.diag = F,
           colors = c('#e9a3c9', '#f7f7f7', '#a1d76a'))
```