Springboard Data Science Career Track

Capstone Project # 1 Milestone Report:

Predicting IMDB Rating of Upcoming Movies

Submitted By:
Koshika Agrawal

February 21, 2019

1. Introduction

1.1. Problem Statement

Today's movie watcher is spoilt for choice with increase in movie streaming platforms, DVD rental services and easy access to movie theatres. Knowing a movie's rating before watching the movie helps reduce the decision fatigue arising out of increased options. IMDb (Internet Movie database) is often a go-to source for knowing a movie's ratings, user reviews, plot, casting and other details before watching that movie. The ratings in IMDb are a function of the ratings given by the users.

But unfortunately, IMDb allows users to provide ratings and reviews only after movie release. So, if you want to select a movie for a first day show, you don't have the ratings to rely on.

1.2. Objective

Develop a model to predict rating of an upcoming movie based on its cast, director, genre and release date

1.3. Client

- Movie watchers
  - Will have a rating even before a movie's release to decide whether to watch the movie or not
- Movie Theatres
  - The buyers representing the theatres can use the model to decide which movie to lease
  - They can further use the model to decide for how many weeks they should show the movie to have an optimum box-office collection
- Movie makers
  - The model will be using factors like cast, director, genre to predict revenue earned. So the movie makers can use this model to find a perfect recipe for a movie

2. Data Acquisition and Wrangling

Following 3 broad steps were involved in wrangling the data:

1. Data Acquisition: Collecting data from source
2. Filtering the dataframes
3. Merging dataframes

2.1. Data Acquisition

The dataset to be used in the project has been acquired from IMDB website from the link: https://www.imdb.com/interfaces/.
IMDB has 6 relevant datasets contained in separate gzipped, tab-separated-values (TSV) formatted files. These datasets have details of all the movies, tv episodes, documentaries etc for all countries of the world.

2.2. Challenges with dataset

Initial loading and inspection of datasets exposed following challenges in it for our study

1. Datasets have redundant columns that are not useful for the problem
   - ➢ These columns were excluded while loading data
2. The missing or null values are present as '\N'
   - ➢ '\N' entries replaced by NaN while loading data
3. Some columns do not have uniform datatype for all entries
   - ➢ dtype was set for non-uniform datatype columns

The aim was to resolve maximum issues during data load.
The shape of all datasets after loading is:

```
title_akas (3486933, 4)
title_basics (5642968, 7)
title_crew (5642968, 3)
title_principals (32301717, 3)
title_ratings (919215, 3)
name_basics (9139818, 5)
```

Fig 2.1

2.3. Filtering the dataframe

For this problem, we are just considering full-length feature films (movies) for the region United States of America released after 1980.

The reason for considering this subset of the dataset is mentioned below:

a. Full-length feature films:
   The goal is to predict the rating of upcoming movies, so taking tv episodes and other videos into consideration is redundant

b. Region: US:
   It would be irrelevant to consider movies from region A to predict ratings of movies from region B. So the model to be developed should use movies from a particular country to predict the ratings of upcoming movies from same country.

c. Movies released after 1980:

To start, we are considering only the movies which have been released after 1980 as the ratings of a movie are greatly expected to be influenced by the cast and crew involved. Most of the crew involved in movie making today, probably would not have any movies to their credit before 1980.

As can be seen above in fig. 1.1, number of entries in almost all the datasets are in millions. So before merging, the datasets arerequired to be filtered out on above parameters so as to keep the merge operation light on system resources.

2.4. Merging Dataframes

All the datasets to be merged have the column 'tconst' which was used as merge key. The dataset name_basics was merged on 'nconst', which is present in other datasets too. The final dataset has 432774 rows and 18 columns.

The wrangled dataset is saved to a csv file.

The code can be accessed on github at:
https://github.com/koshika15/Springboard/blob/master/Capstone%20Project%201/B.%20Data_wrangling.ipynb

3. Data Exploration

Data exploration for this project has been carried out in 3 iterations. First 2 iterations also served to clean and refine the data further, to better aid in our analsis.

3.1. **1st Iteration**: Following visualizations have been plotted:

a. Distribution of movies throughout the years under consideration across different ratings

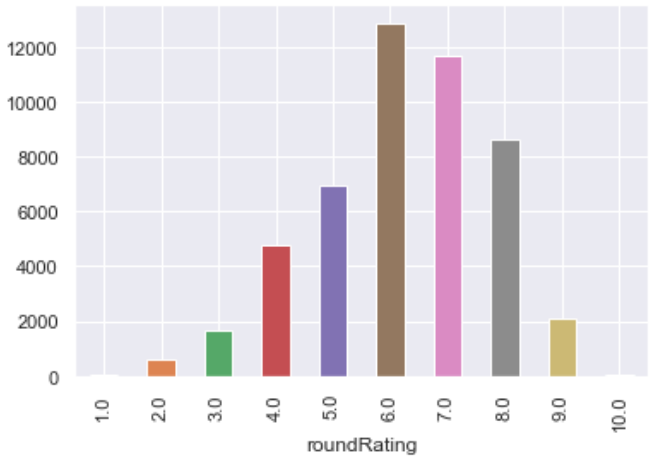

Fig 3.1

Observations:

i. The average rating of the movies seems to be 6.

ii. The movie ratings are normally distributed

b. Year wise trends
   i. Number of movies released each year
   ii. Average runtime of movies
   iii. Average rating of movies
   iv. Number of votes per movie

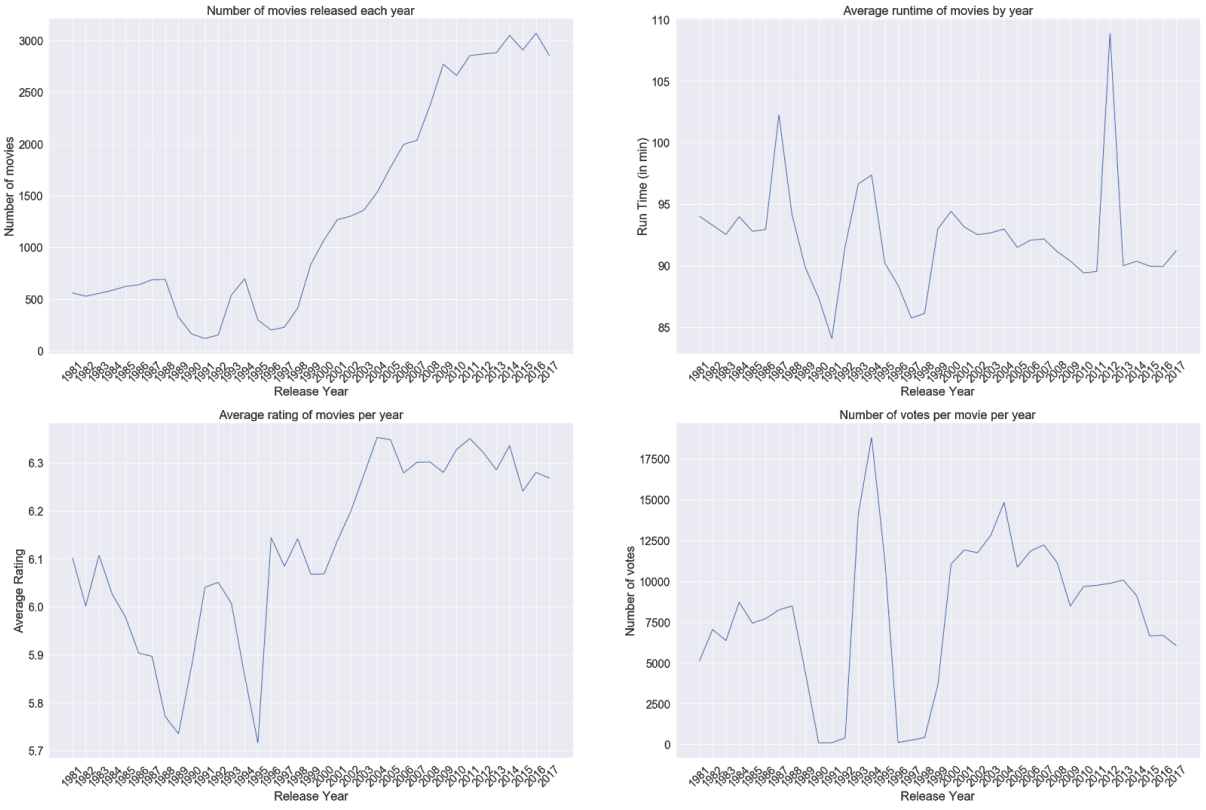

Fig 3.2

In the data wrangling phase, only the movies released after 1980 were considered. Above initial visualizations revealed inconsistent trends before 1999 on multiple features. So, we are considering the movies released after 1999 for the purpose of this project.

c. Trend of maximum runtime of movies over the years
This trend revealed the maximum runtime of one of the movies as 51420 minutes or 857 hours. Most feature films are between 70 and 210 minutes long. (Source: https://en.wikipedia.org/wiki/Feature_film). So, we are considering only movies with less than 240 runtime minutes.

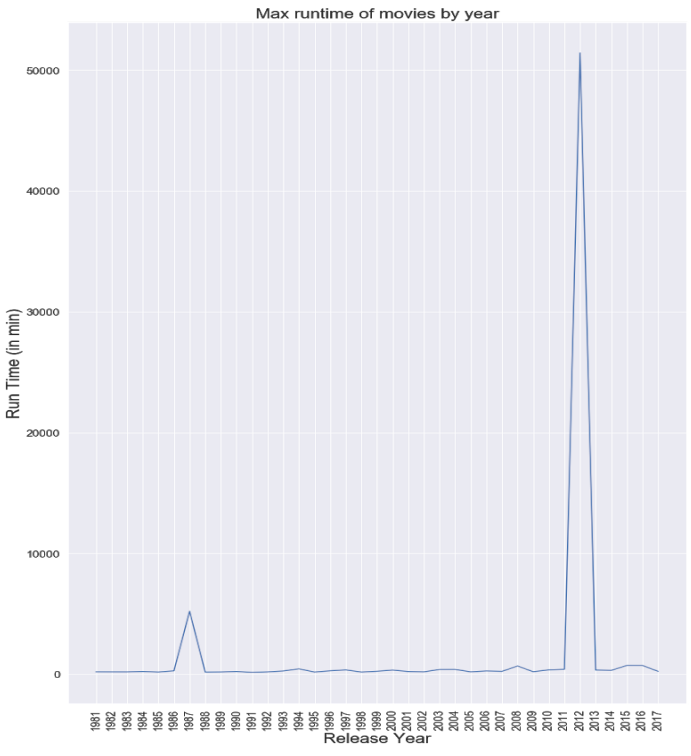

Fig 3.3

The new dataset now has 286930 rows.

	Old dataset	**New dataset**	**% change**
No. of rows	432774	311444	29%
No. of unique titles	49440	34917	30%

3.2. 2nd Iteration:

The dataset has multiple rows for a single movie title. In order to visualize the data, a dataset with unique title entries is required. So another dataset named 'titles' is filtered from the main dataset.

Visualizations and observations:

i. Distribution of titles among the ratings
   a. The distribution is slightly left skewed
   b. Mean rating should be between 6 and 7

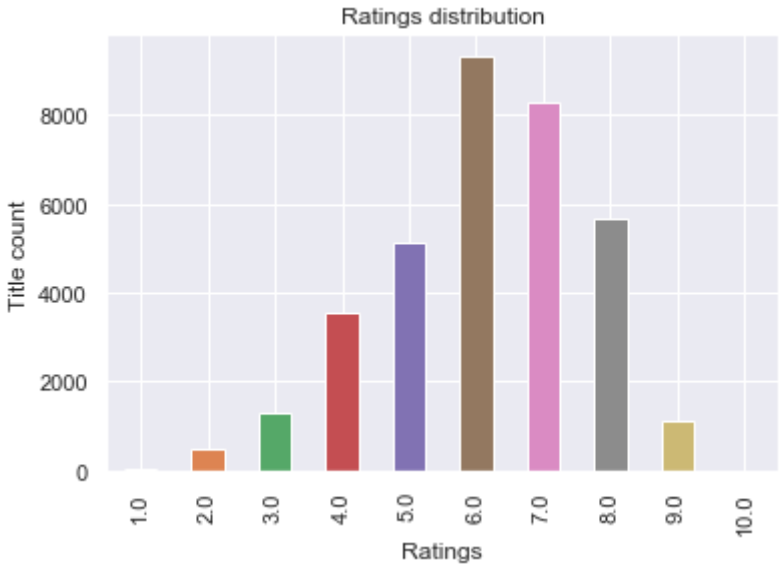

Fig 3.4

ii. Year wise trends

a. Number of movies released have increased over the years

b. Average runtime of movies shows a decreasing trend

Fig 3.5

The code can be accessed on github at:
https://github.com/koshika15/Springboard/blob/master/Capstone%20Project%201/C.%20Data_Story.ipynb

4. Further exploration

We will explore the relationship between dataset columns and the movie ratings.

4.1. Length of the movie

Length of the movies is positively correlated with the ratings. The average runtime of movies is at 102 minutes. Movies with runtime of more than 175 minutes have higher ratings.

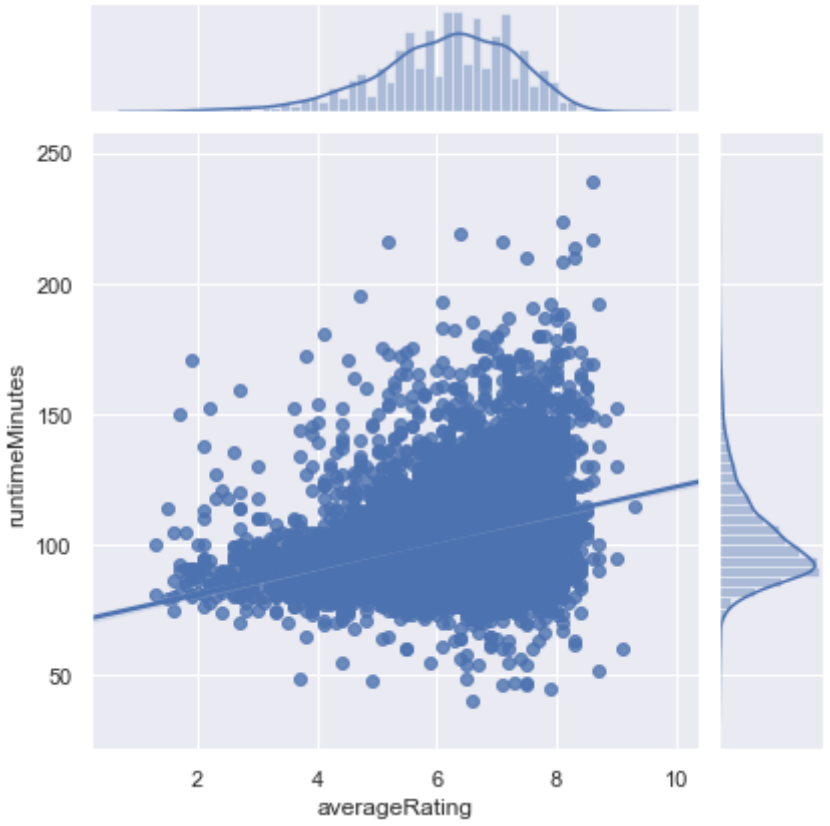

Fig 4.1

4.2. Genre of the movie

Genres with higher than average rating are - history, biography, war, documentary. So it appears that non-fictional movies get a higher rating.

Horror movies are rated much lower than average.

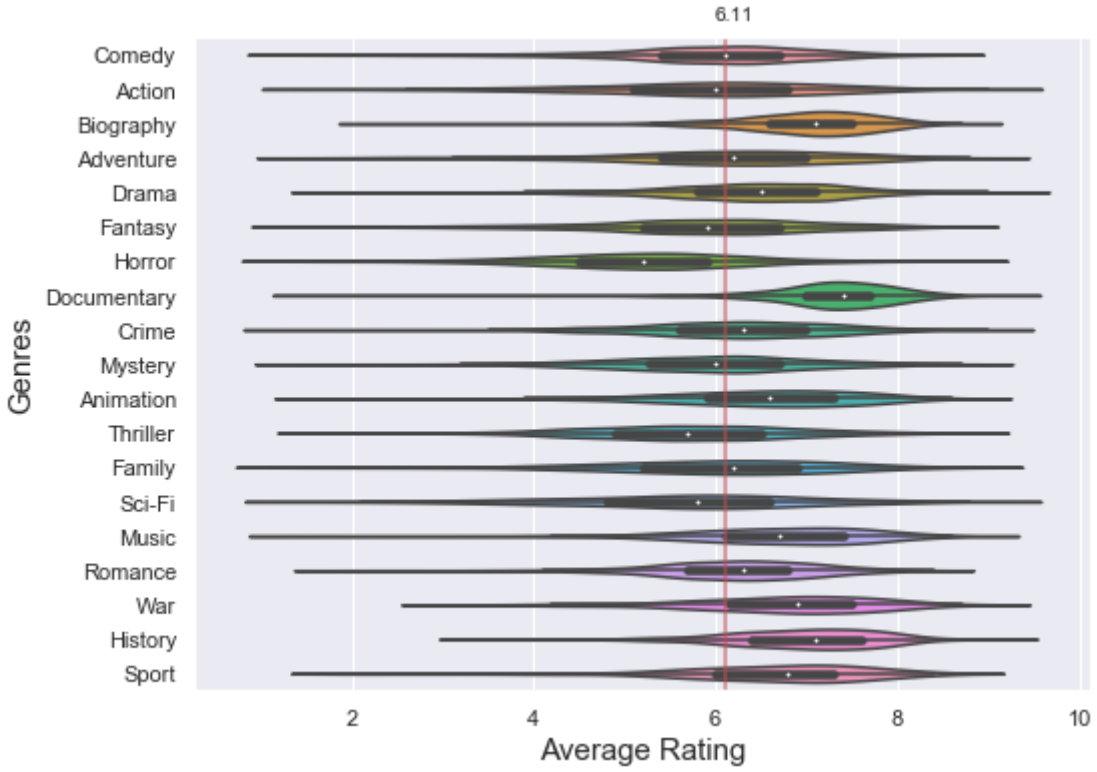

Fig 4.2

4.3. Cast and Crew

We refer to cast as those who have acted in the movie, which means it includes actor, actress and self. Crew includes those who were involved in making the movie - director, producer, editor etc.

We analysed the effect of cast/crew by parameters derived from cast/crew associated with the movies. These parameters are:

a) Number of each category of cast/crew in a movie. Category refers to actor, actress, director and so on.
b) Total count of cast, which means total number of actors, actress and self.
c) Total count of crew
d) Ratio of number of actors to number of actress in a movie

The relation between these parameter can be seen in the heatmap in Fig 4.3.

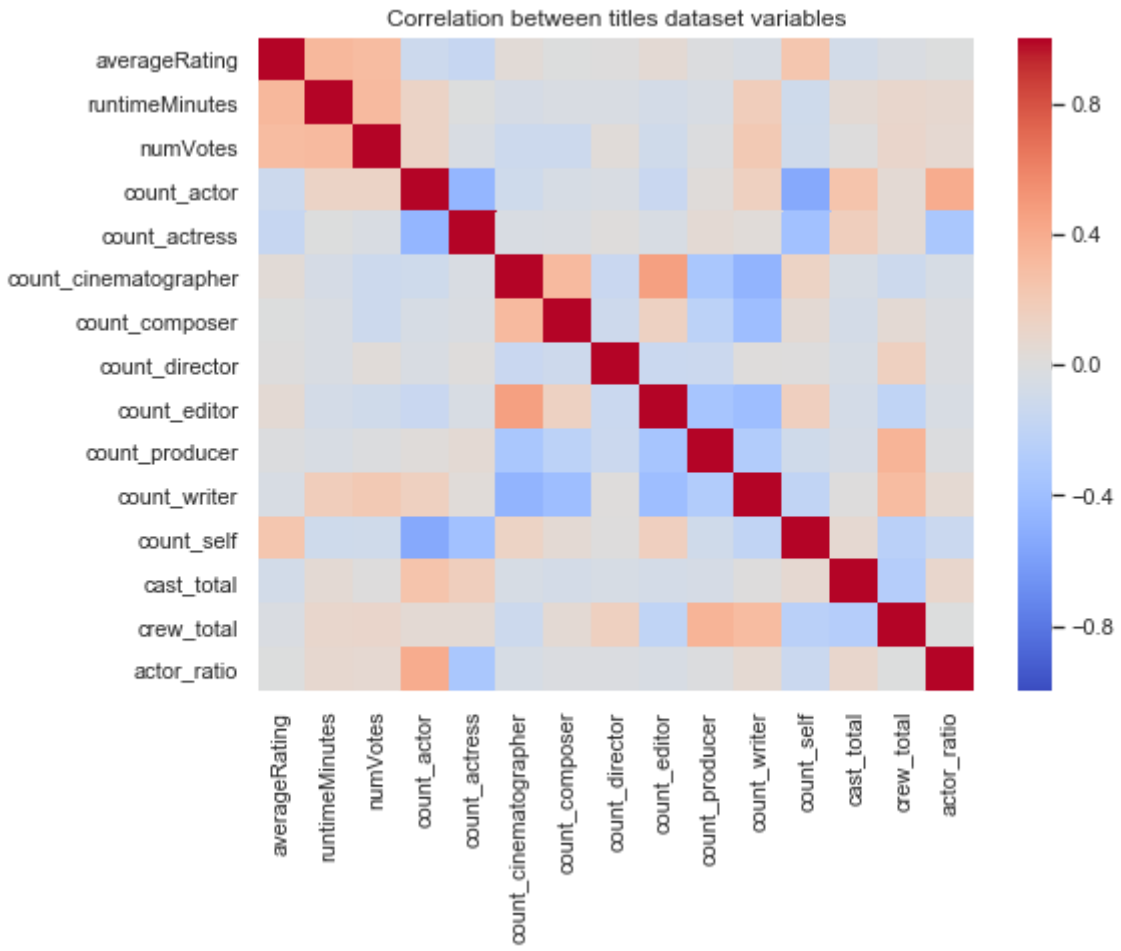

Fig 4.3

From fig 4.3, we can conclude that:

1. The heatmap shows positive correlation between ratings and
    1. runtimeMinutes (length of the movie)
    2. count of self
2. The heatmap shows negative correlation between ratings and
    1. count of actors
    2. count of actress
    3. count of total cast

For each movie's category, the ratings of cast/crew in that category were aggregated to get the rating for that category. For now, the aggregation function used are mean and max. Subsequently, while building our model, we will use the aggregation function that will better serve our model.

Heatmaps of both aggregation models look similar:

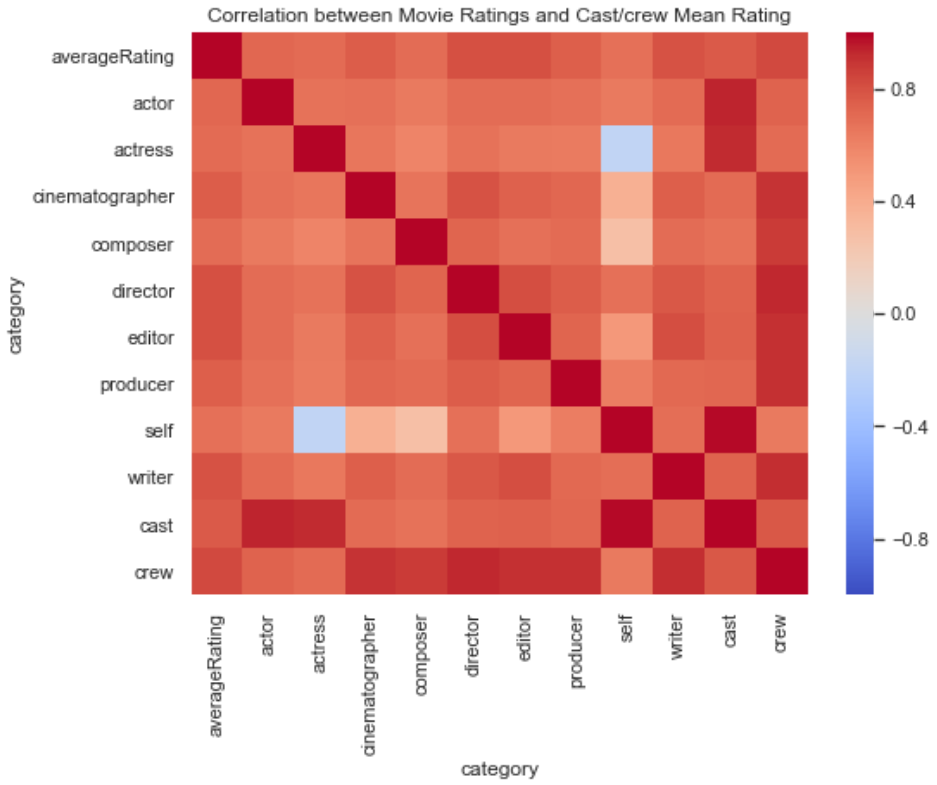
Correlation between Movie Ratings and Cast/crew Mean Rating
averageRating
actor
actress
cinematographer
composer
director
editor
producer
self
writer
cast
crew
category
0.8
0.4
0.0
−0.4
−0.8

Fig 4.4

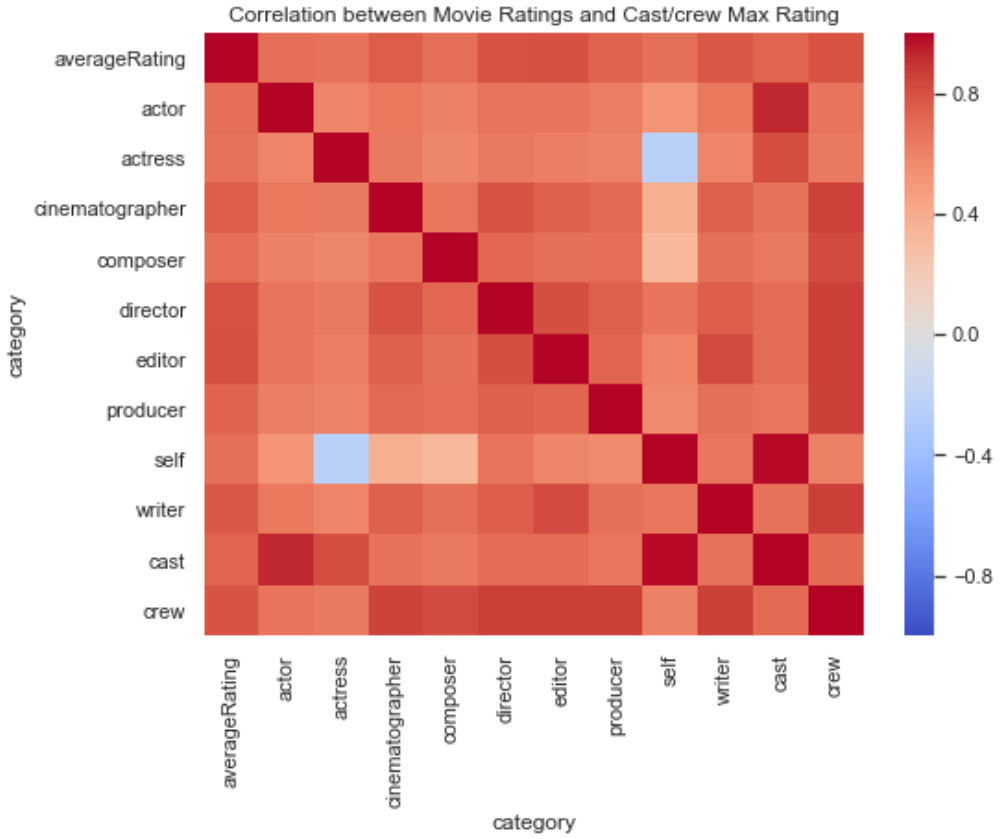
Correlation between Movie Ratings and Cast/crew Max Rating
averageRating
actor
actress
cinematographer
composer
director
editor
producer
self
writer
cast
crew
category
0.8
0.4
0.0
−0.4
−0.8

Fig 4.5

All the variables show a positive correlation with the ratings, probably due to the fact that these variables have been calculated from the ratings itself.

4.4. Number of votes

Number of votes could not be used to predict the ratings, as this will not be available to us during prediction. We have still plotted it to get insights.

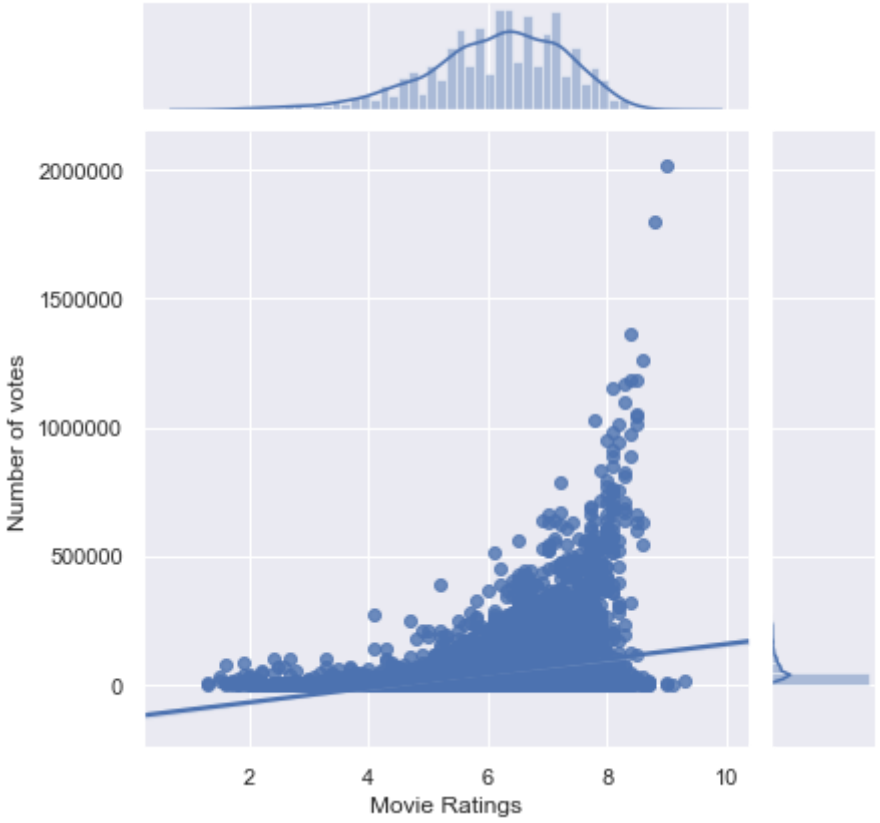

Fig 4.6

1. Number of votes for a movie are positively correlated to the rating.
2. Most of the cases of high vote count are for highly rated movies. From this we can conclude that people are more likely to vote for movies if they like it.

Scope of further study:

The visualizations show a sudden spurt in all the trends for the year 1994. Though this might be a result of bad data but it can be further explored in a separate study to explore interesting insights.