IMDB Rating Predictions

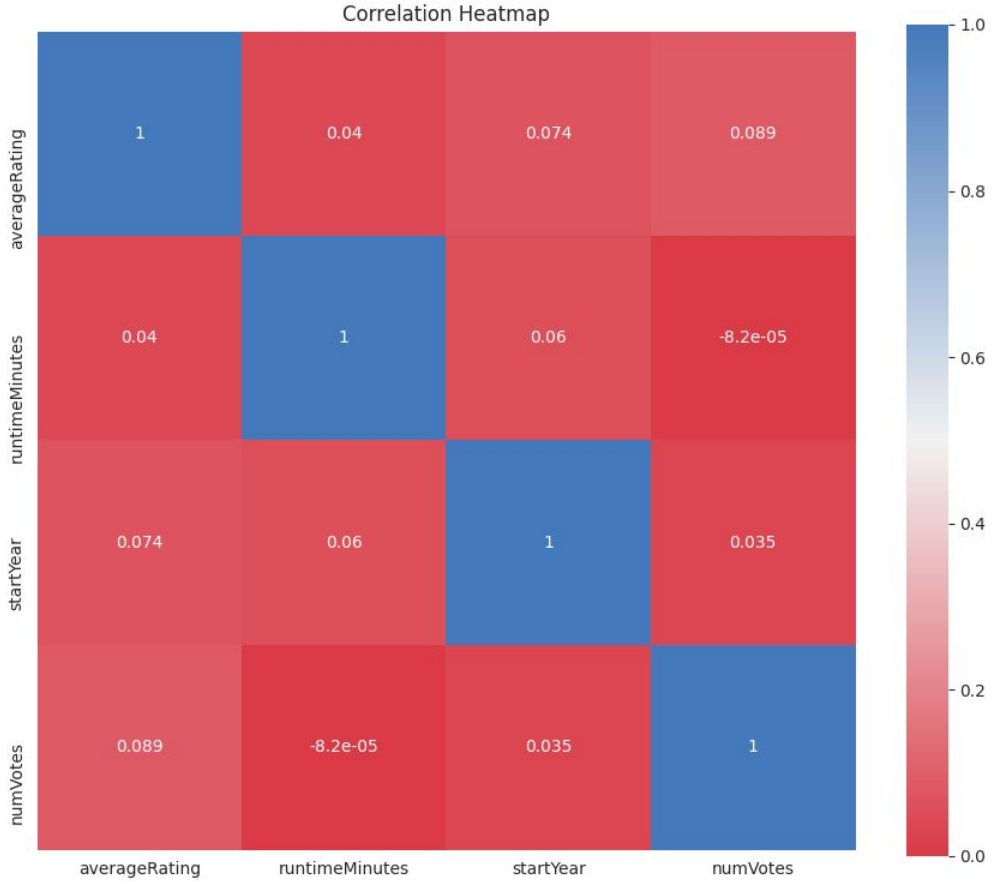
Eli Oceanak

Goal

- Given a movie/show’s basic IMDB stats other than it’s average rating, I want to be able to predict the rating.
- Use things like duration, and year released to come up with a rating without having to actually watch a movie or think about it at all.

Initial plan

- Tried to do analysis on TV shows instead of movies due to fewer people having already done analysis on them
- Datasets that contained TV shows had incomplete data, and what I ended up with had very little actual usability as predictors.

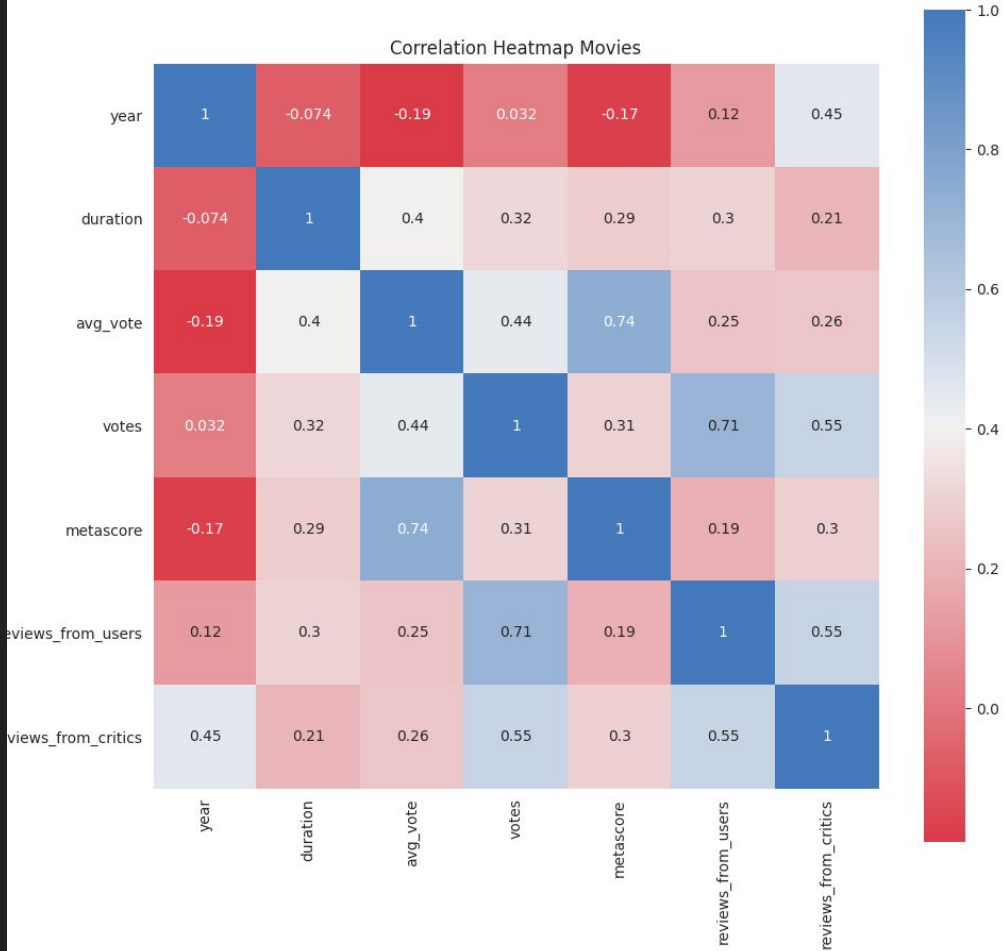

Correlation Heatmap
averageRating
runtimeMinutes
startYear
numVotes
1
0.04
0.074
0.089
0.04
1
0.06
-8.2e-05
0.074
0.06
1
0.035
0.089
-8.2e-05
0.035
1
averageRating
runtimeMinutes
startYear
numVotes
1.0
0.8
0.6
0.4
0.2
0.0

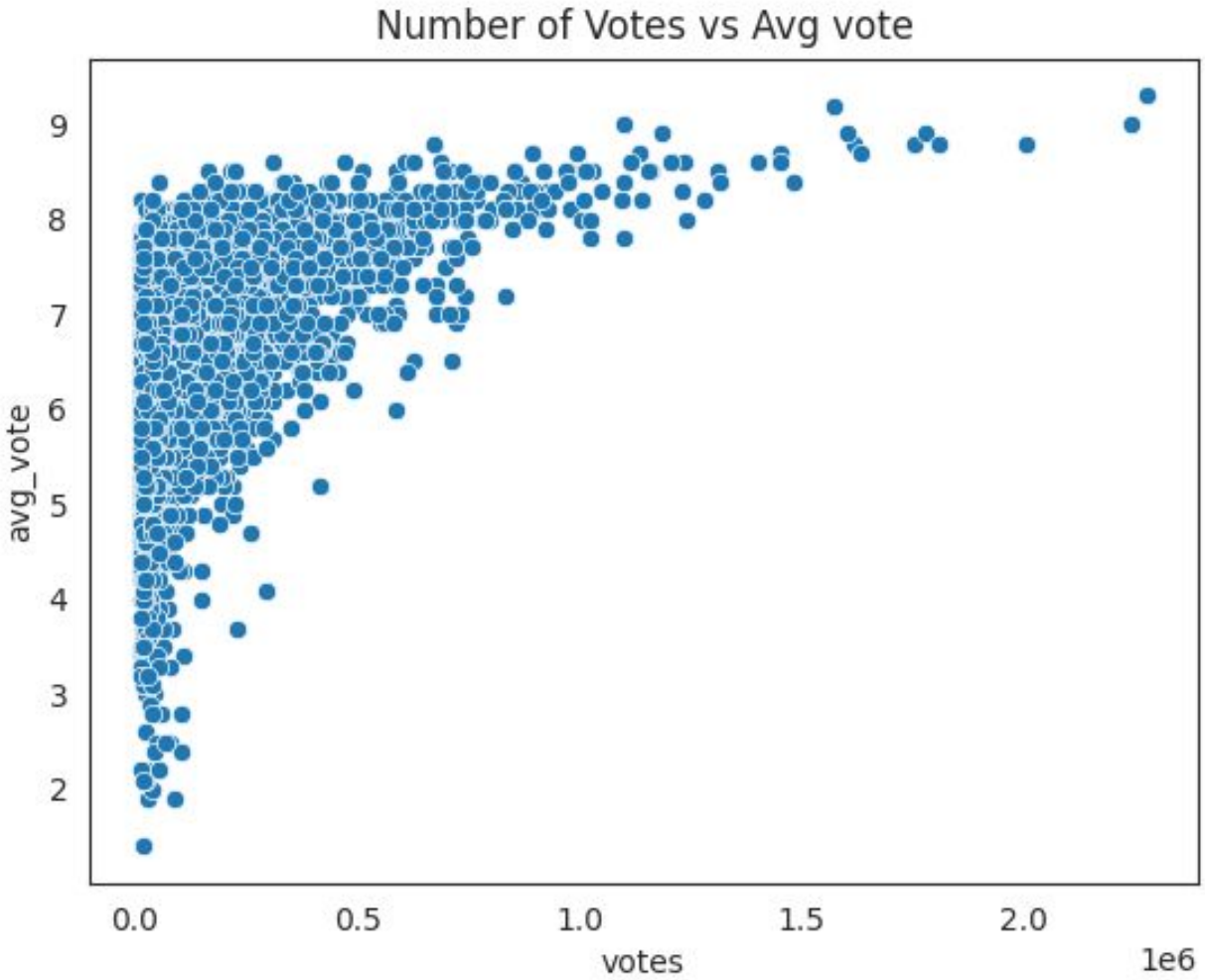
Switched to Movies

- Used this dataset - https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset
- More data, more neat, and in larger quantities.
- Better predictors

Cleaning the data

- Removed all rows with null data
- Removed all movies with < 10,000 total votes cast
- Removed a few outliers with massive values that were obviously incorrect
- Ended up with just over 5,000 movies which I split 70/30 into training and testing sets.


Correlation Heatmap Movies
year
duration
avg_vote
votes
metascore
eviews_from_users
views_from_critics
1
-0.074
-0.19
0.032
-0.17
0.12
0.45
-0.074
1
0.4
0.32
0.29
0.3
0.21
-0.19
0.4
1
0.44
0.74
0.25
0.26
0.032
0.32
0.44
1
0.31
0.71
0.55
-0.17
0.29
0.74
0.31
1
0.19
0.3
0.12
0.3
0.25
0.71
0.19
1
0.55
0.45
0.21
0.26
0.55
0.3
0.55
1
year
duration
avg_vote
votes
metascore
reviews_from_users
reviews_from_critics
1.0
0.8
0.6
0.4
0.2
0.0


Number of Votes vs Avg vote
avg_vote
9
8
7
6
5
4
3
2
0.0
0.5
1.0
1.5
2.0
votes
1e6

Choosing a Model

- Tried a few different models, ended up with a linear regression model.
- Well actually two different linear models
- Average error of only .65 points from just number of votes and movie duration!

```
Intercept:  4.865305933280882
Coefficients:
[('votes', 1.9001086557617272e-06), ('duration', 0.01327254030034391)]
mean squared error:  0.726093638519896
mean absolute error:  0.6474434307452884
```

```
Intercept:  3.7887129069597956
Coefficients:
[('votes', 1.0690068504436637e-06), ('duration', 0.00716504923866635), ('metasco
re', 0.03312036750110403)]
mean squared error:  0.3555701397621731
mean absolute error:  0.4400589079545811
```