# Predicting IMDB score for TV shows

# The Problem

- There have been 38 cancelled TV shows in 2018 from 5 networks alone (ABC, CBS, The CW, Fox, NBC)
- The average cost for a 30-minute comedy pilot is $2 million.
- An hour-long drama averages out at $5.5 million.
- Per ScreenRant, 65% of new shows get cancelled per season
- All this is compounded by the fact that streaming is increasing in popularity, actively taking away from traditional tv-based viewing

# The Question

- Can Hollywood’s bean counters be replaced with machine learning?
- Is it possible to predict a show’s success based on certain features about the show?

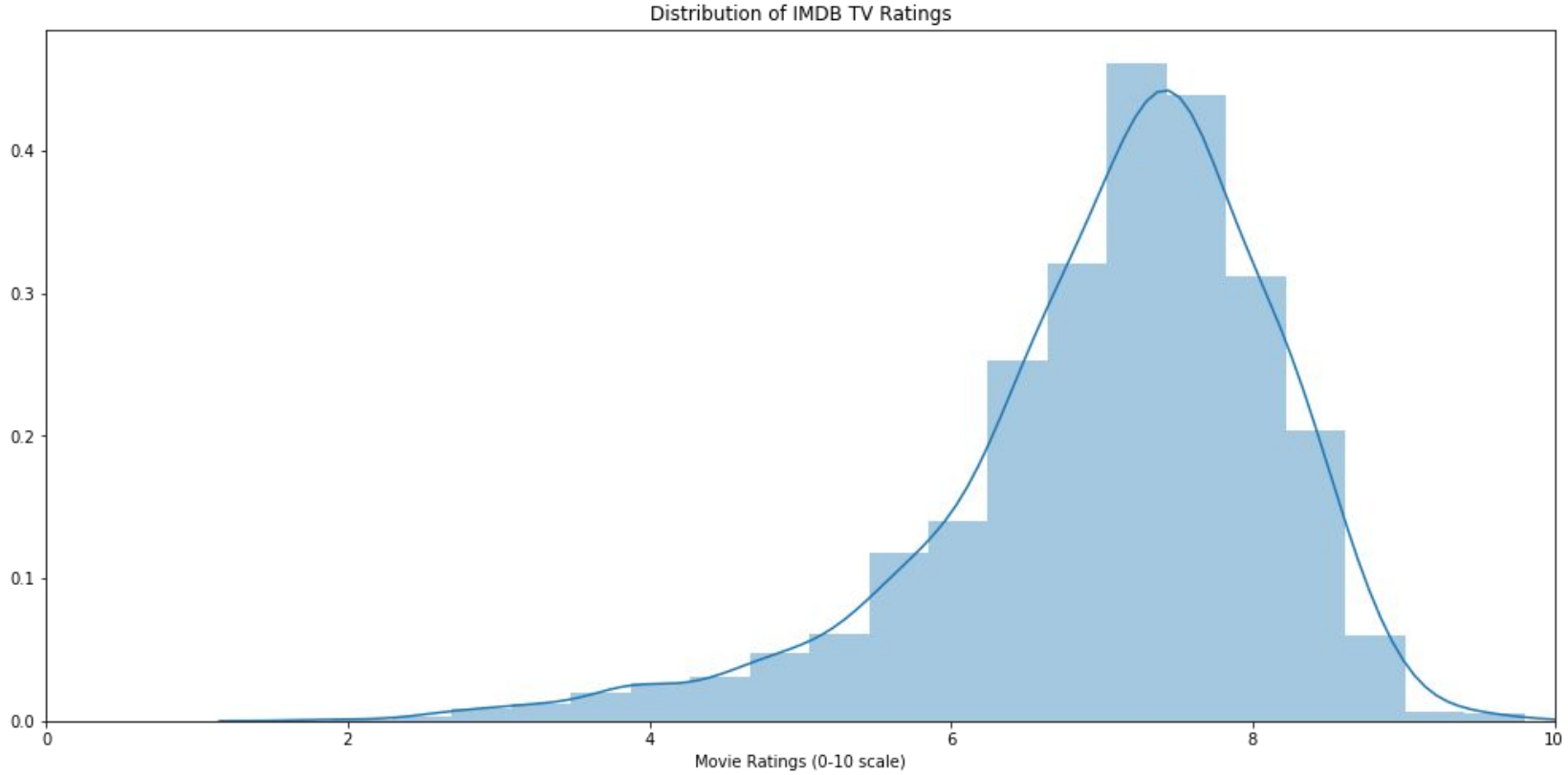
# The Dataset

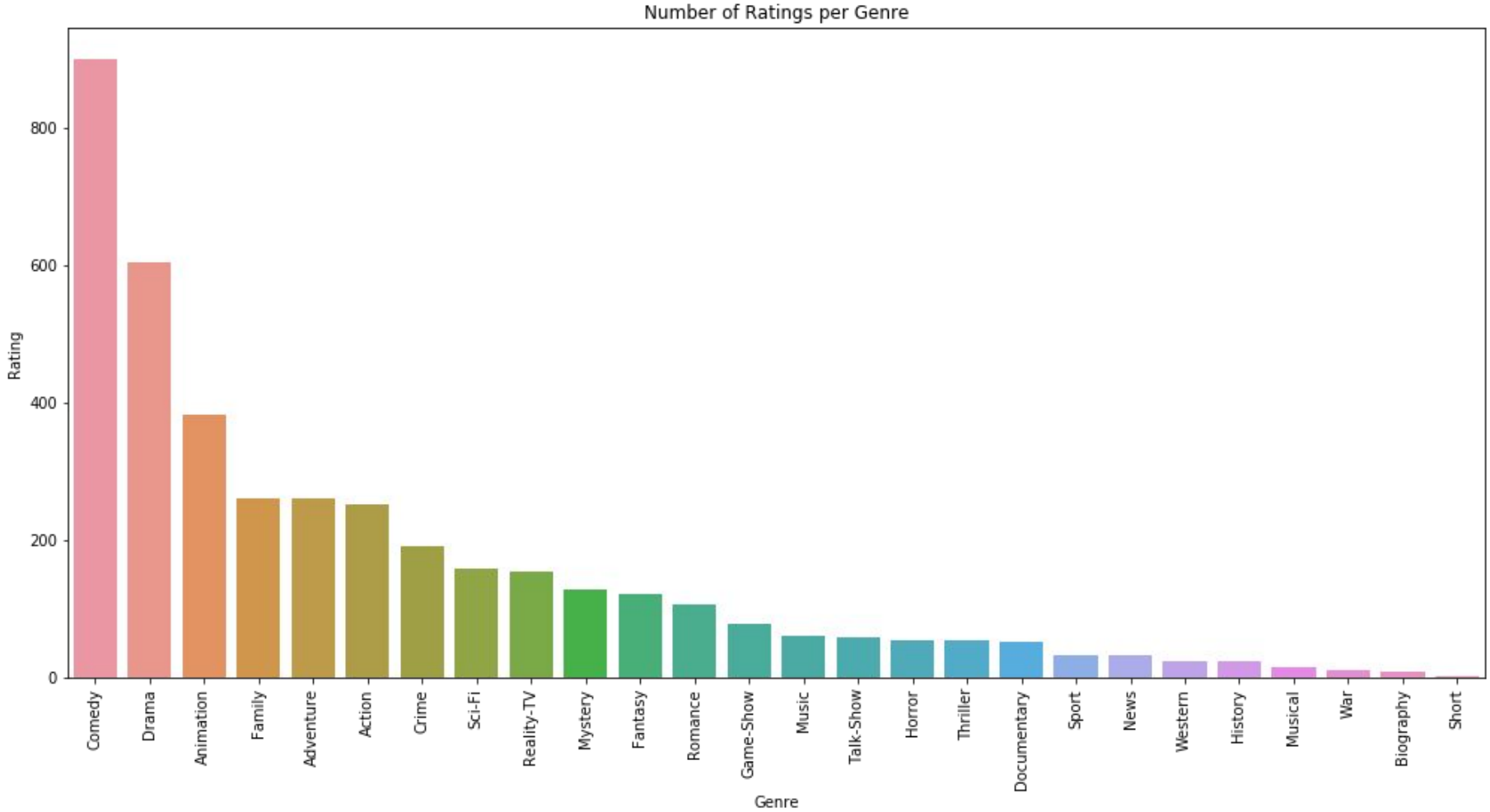
- Data is the modern gold, so there are limited resources on gathering viewership data for tv shows
- Target to predict is IMDB score
- Important features may include:
  - Actors
  - Writers
  - Genre
  - Air time
  - Runtime

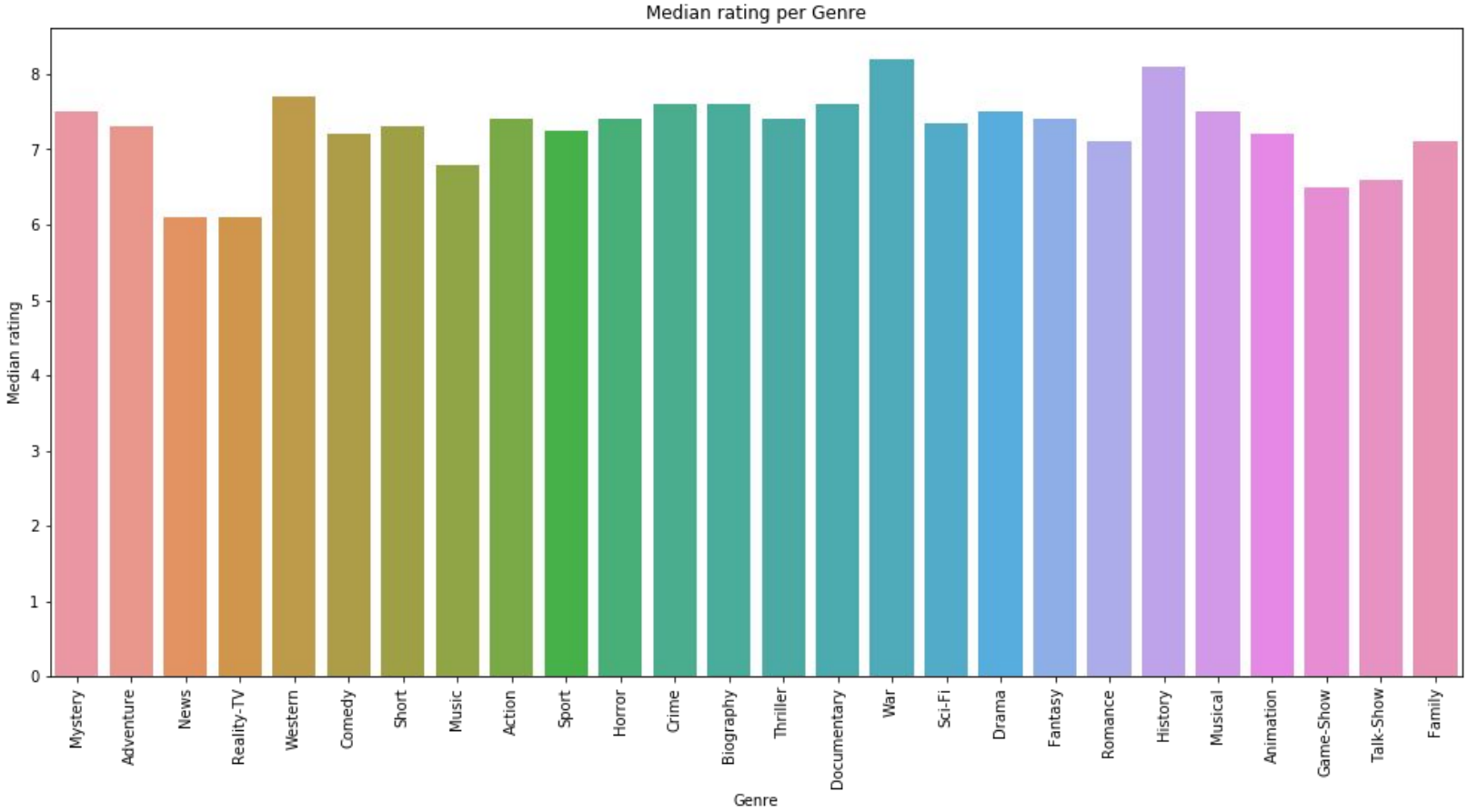
# The Dataset pt 2

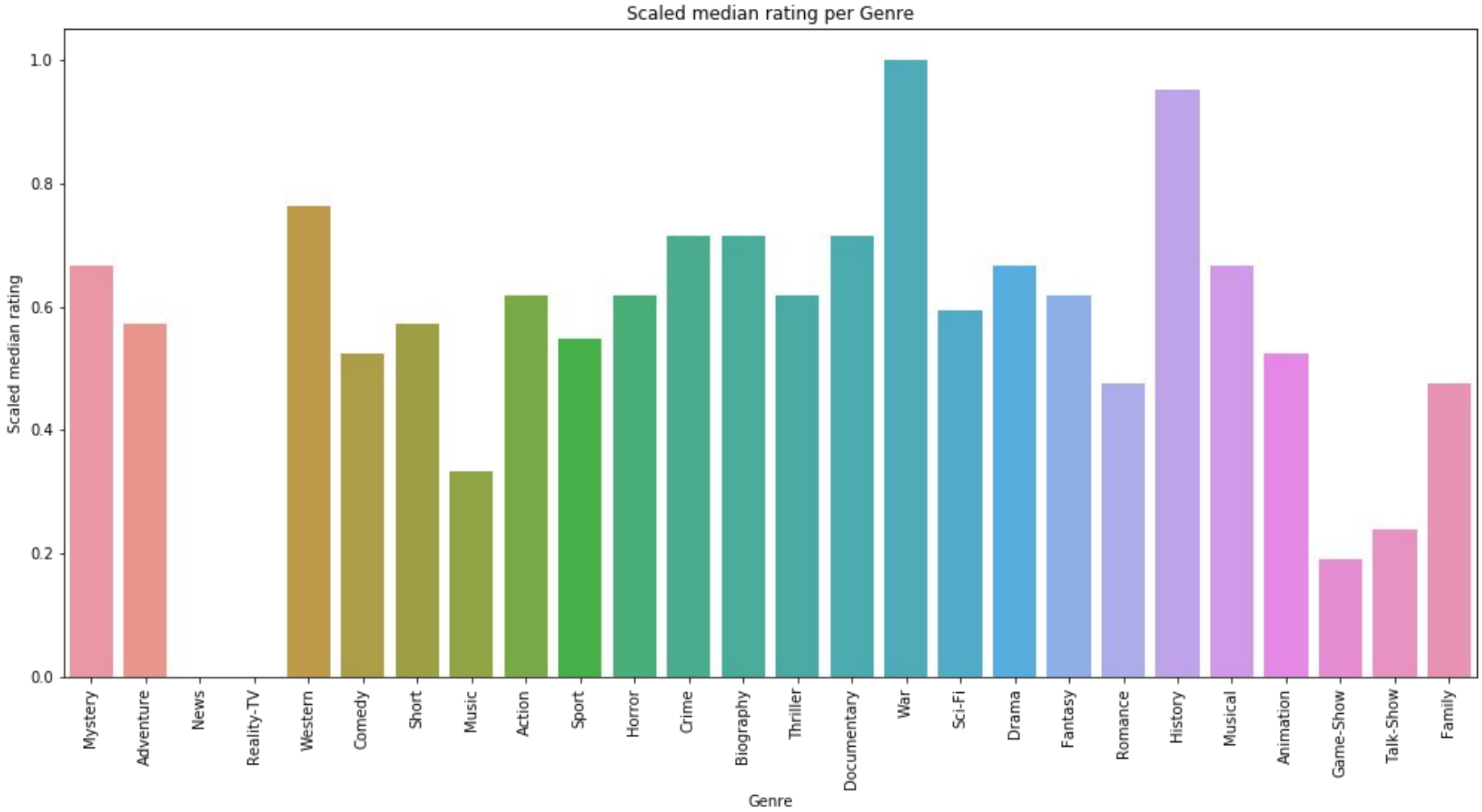
- The intuitive sources:
  - Nielsen - $$$
  - IMDB - No open API
- Network TV shows from Wikipedia
- APIs used instead:
  - The Movie Database (TMDB)
  - The TVDB (TVDB)
  - The Open Movie Database (OMDB)

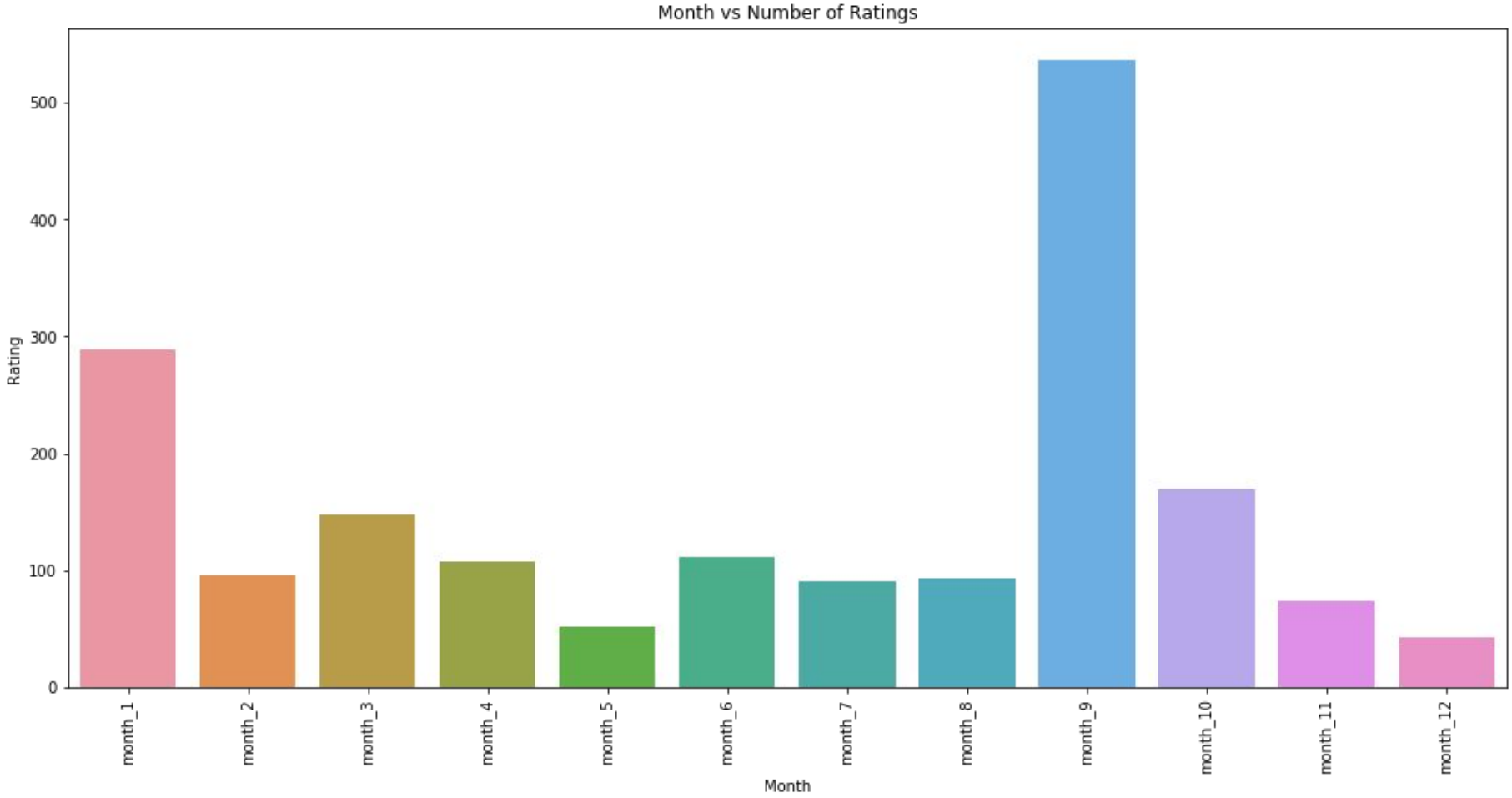
# Initial Data Exploration

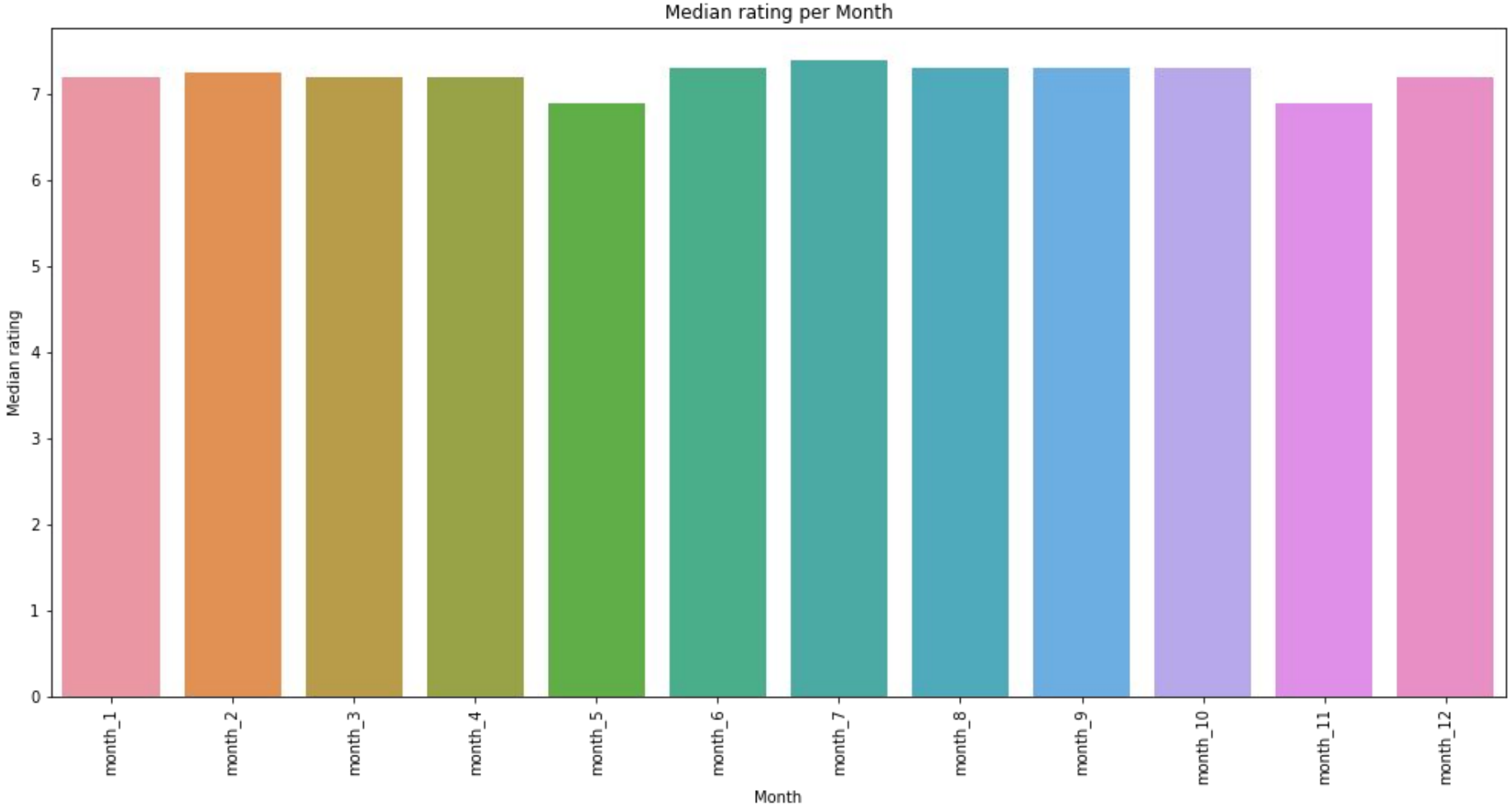


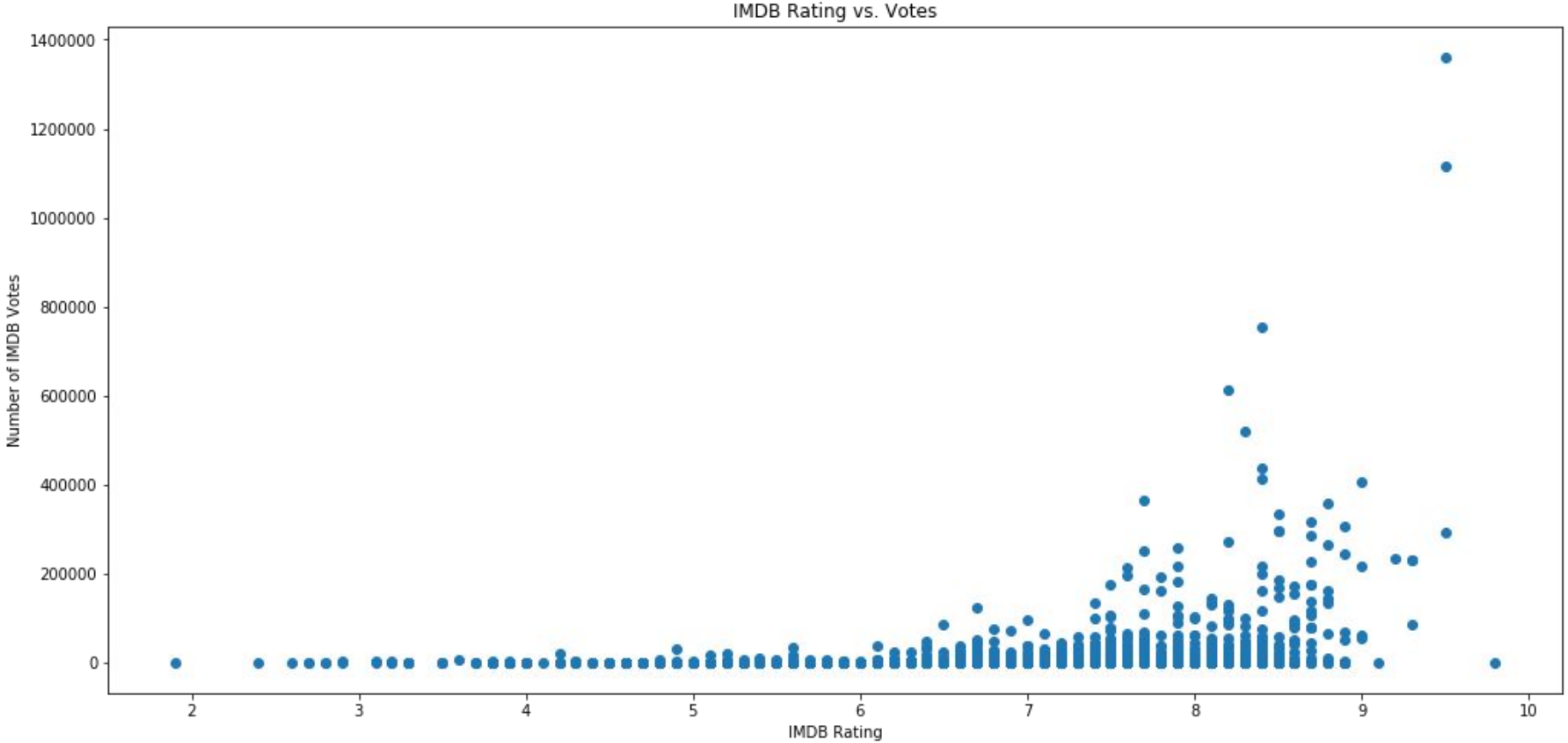

Number of Ratings per Genre
Rating
0
200
400
600
800
Comedy
Drama
Animation
Family
Adventure
Action
Crime
Sci-Fi
Reality-TV
Mystery
Fantasy
Romance
Game-Show
Music
Talk-Show
Horror
Thriller
Documentary
Sport
News
Western
History
Musical
War
Biography
Short
Genre


Median rating per Genre
Median rating
0
1
2
3
4
5
6
7
8
Mystery
Adventure
News
Reality-TV
Western
Comedy
Short
Music
Action
Sport
Horror
Crime
Biography
Thriller
Documentary
War
Sci-Fi
Drama
Fantasy
Romance
History
Musical
Animation
Game-Show
Talk-Show
Family
Genre


Scaled median rating per Genre
Scaled median rating
1.0
0.8
0.6
0.4
0.2
0.0
Mystery
Adventure
News
Reality-TV
Western
Comedy
Short
Music
Action
Sport
Horror
Crime
Biography
Thriller
Documentary
War
Sci-Fi
Drama
Fantasy
Romance
History
Musical
Animation
Game-Show
Talk-Show
Family
Genre

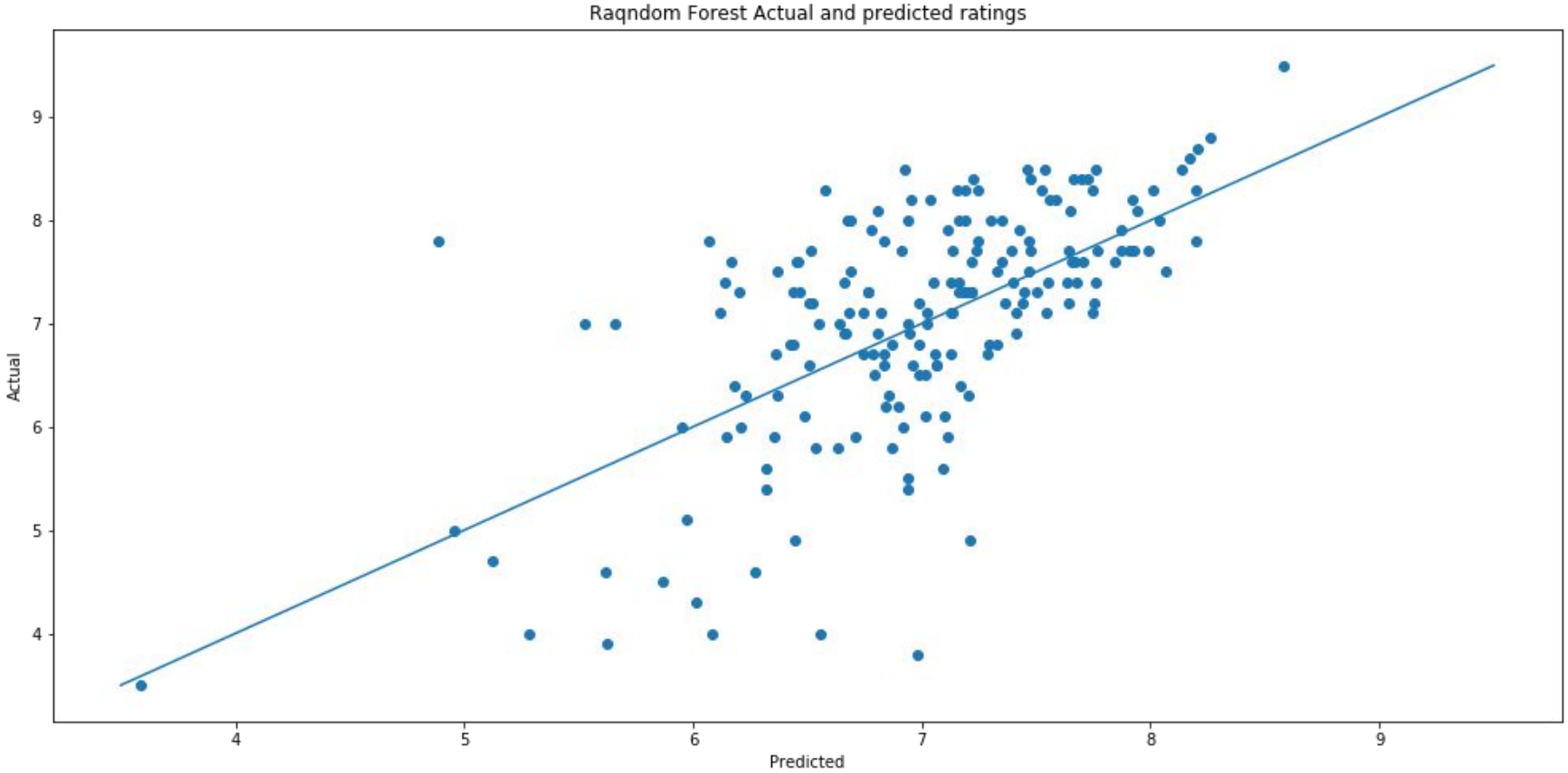

Month vs Number of Ratings
Rating
0
100
200
300
400
500
month_1
month_2
month_3
month_4
month_5
month_6
month_7
month_8
month_9
month_10
month_11
month_12
Month


Median rating per Month
Median rating
0
1
2
3
4
5
6
7
month_1
month_2
month_3
month_4
month_5
month_6
month_7
month_8
month_9
month_10
month_11
month_12
Month


IMDB Rating vs. Votes
Number of IMDB Votes
1400000
1200000
1000000
800000
600000
400000
200000
0
2
3
4
5
6
7
8
9
10
IMDB Rating

# The Approach/Methodologies

- Features engineered:
  - Actors > imbued weights based on frequency of appearance
  - Runtime > half hour/full/special (movies)
  - Timeslot > Morning/Afternoon/Evening/Late Night
  - First episode aired month  > Seasons
  - Awards/nominated > Yes/No
- Models evaluated:
  - Linear Regression
  - Random Forest
  - Gradient Boost


Raqndom Forest Actual and predicted ratings
Actual
Predicted
9
8
7
6
5
4
4
5
6
7
8
9

# Conclusion

- The best model has ~43% of the variance explained by the model.
- The best model has a mean absolute error of .583
  - The mean rating is 7.03
  - One standard deviation is 1.1
- The model will not be able to reliably replace bean counters in the current state
- With better data, the model can be improved for production use

| Model | R2 | Mean Absolute Error |
|---|---|---|
| Linear Regression | 0.200 | 0.715 |
| Random Forest | 0.429 | 0.583 |
| Gradient Boost | 0.401 | 0.599 |

# Next Steps

- Acquire more complete data - no missing values
- Utilize per episode data (rather than per series)
  - Evaluate rating trends between and within seasons
  - Writers per episode
  - Directors per episode
  - Guest Stars Y/N
  - Timeslot
- Possibly change the target variable - viewership numbers may be a better indicator than ratings, which are very subjective