IMDB Movie Dataset

Stat Padders:

Camden Reeves, Toma Shigaki-Than, CJ Frederickson

Subject and Motivation

- Online reviews heavily influence consumer decisions, especially in entertainment.
- Film audiences consult both professional critics (artistic/technical focus) and amateur audiences (personal enjoyment/entertainment value).
- This dual-review dynamic often results in diverging evaluations of the same film.
- Studios and marketers balance audience preferences with critical appeal for success
- Movies are expensive, consumers heavily rely on evaluations before purchasing tickets

Research Question

- What factors in a film influence IMDb user ratings and critic MetaScores; how do differences in these scores relate to movie characteristics such as gross earnings, number of votes online, decade released, runtime, and certificate of censorship?

INTRO TO DATASET

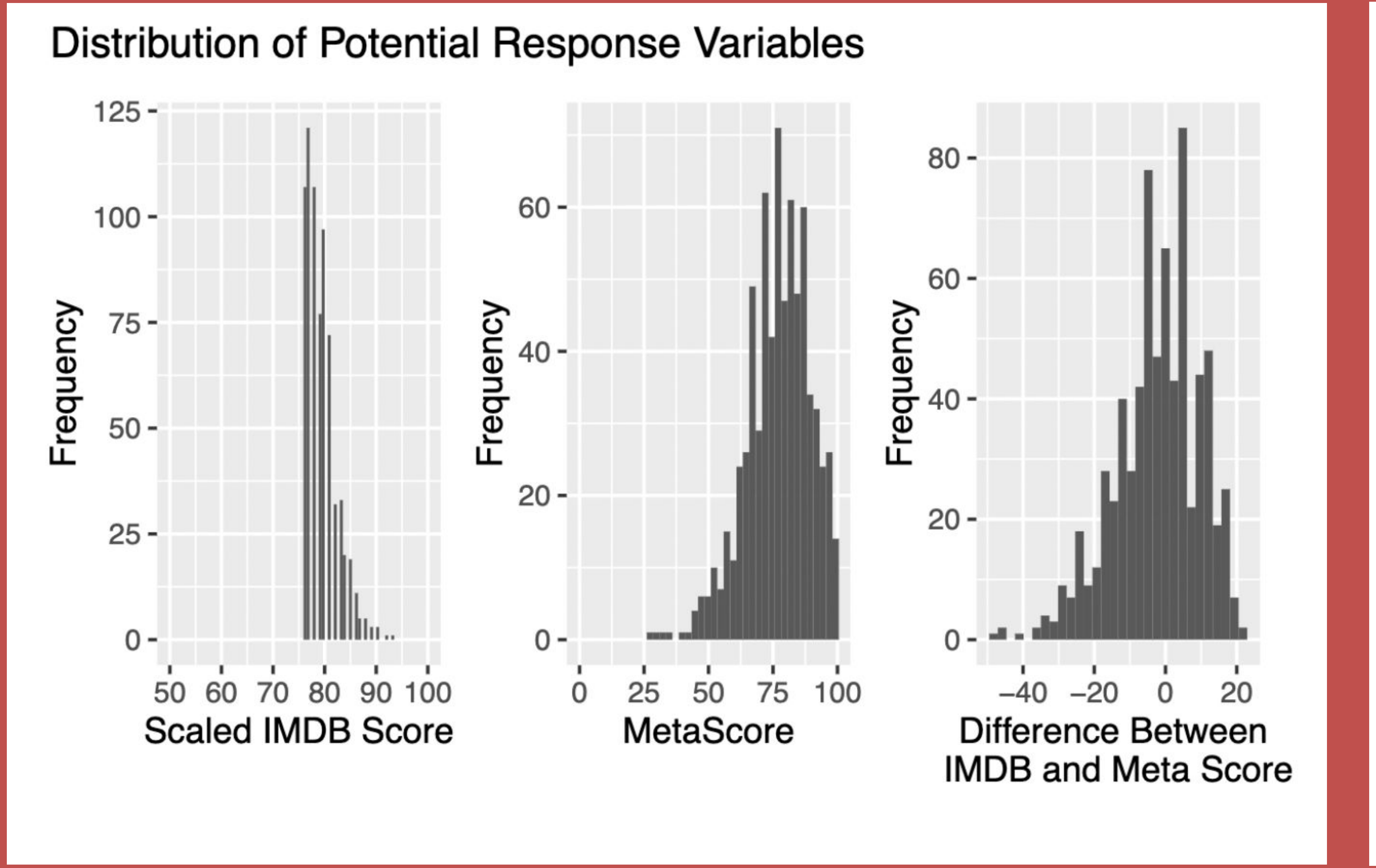
- Dataset from Kaggle scraped from IMDB Website
- observations are from the top 1000 rated movies from the last century, 1930s until 2020.

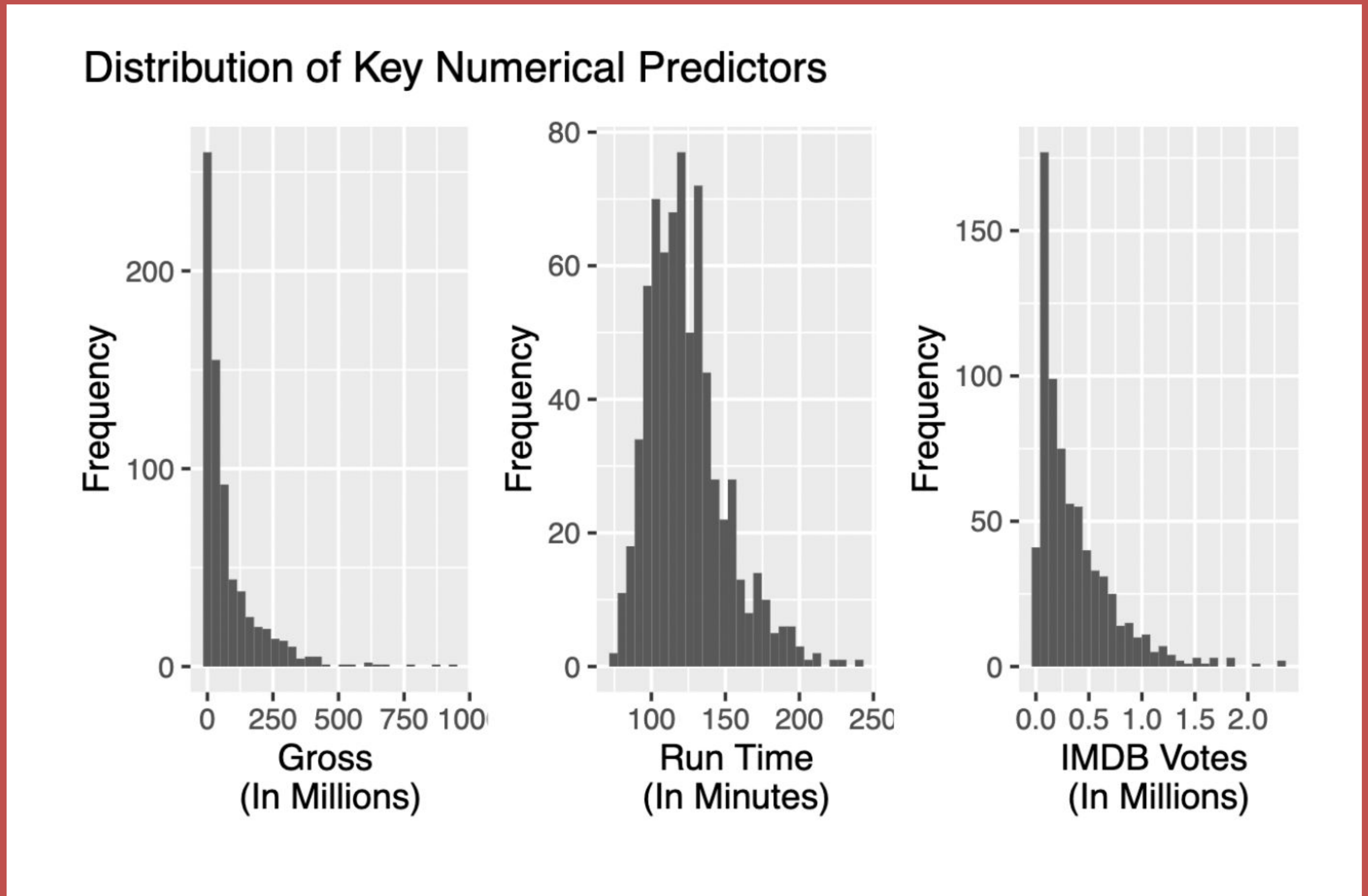
Potential Predictors:

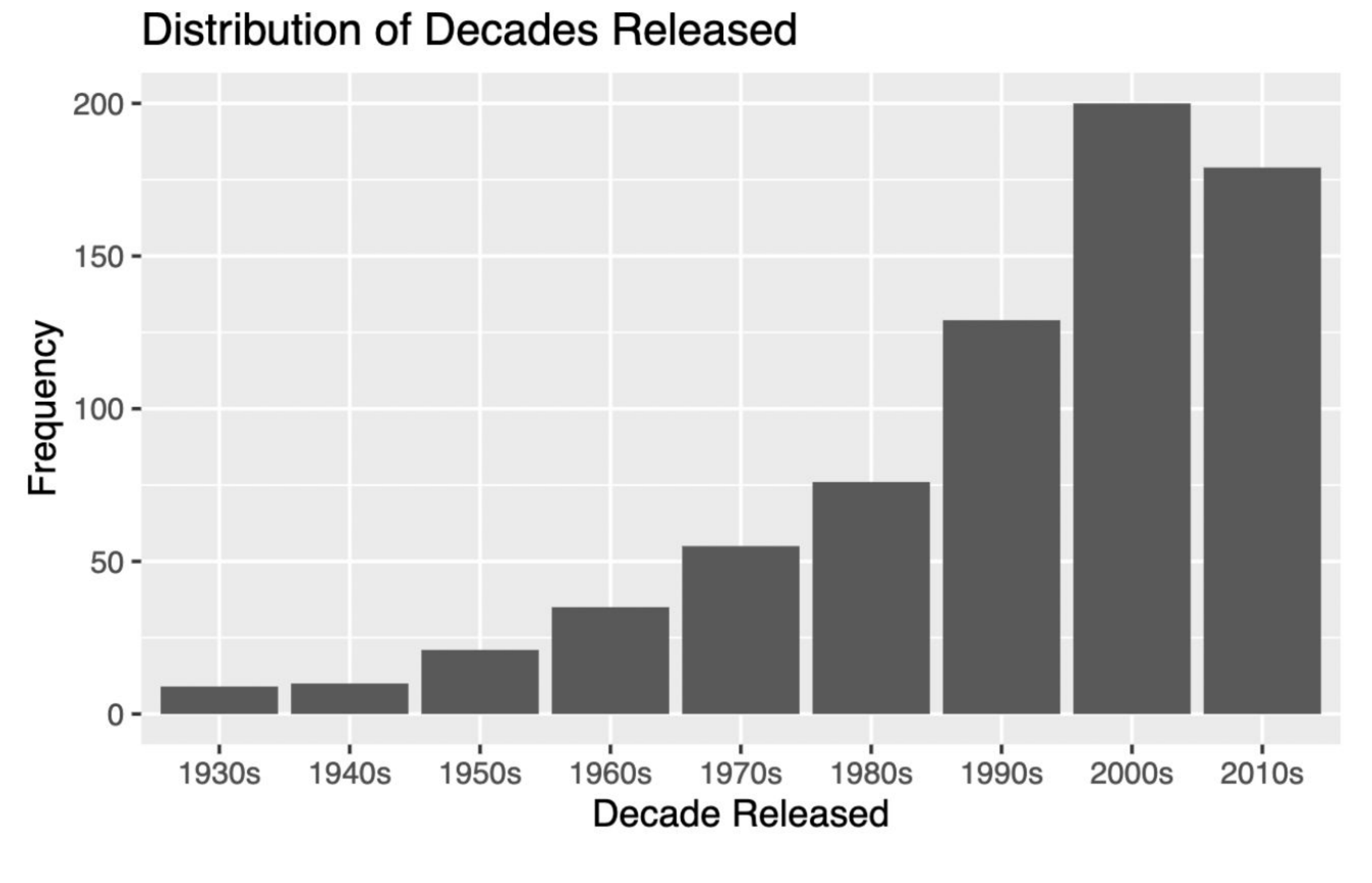
- Runtime (numerical)
- Gross Revenue (numerical)
- Certificate (categorical)
- Decade Released (categorical)
- Number of Votes (numerical)

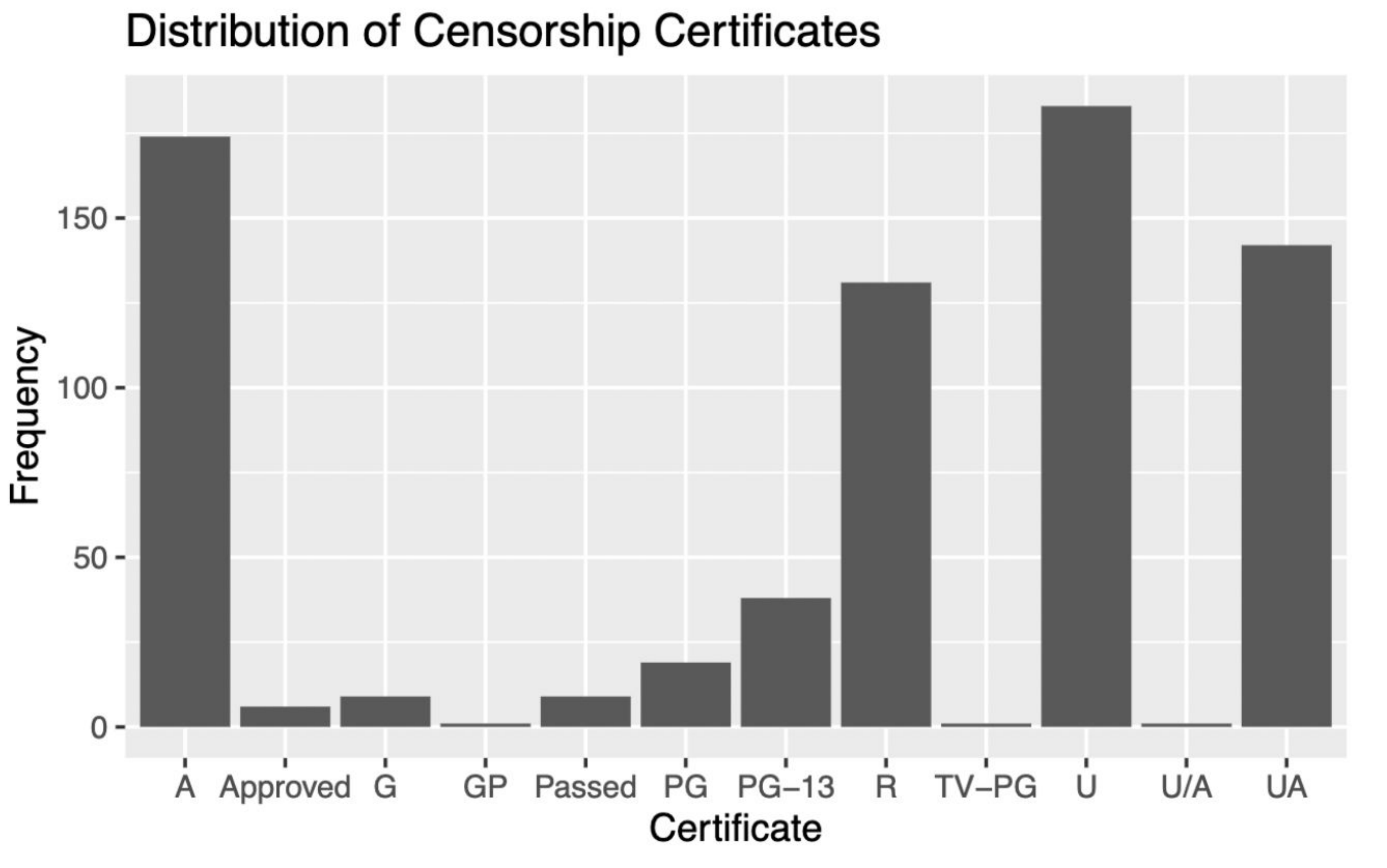
Response:

- IMDB Score
- Meta-Score (scaled)
- Difference (quantified by how much the Meta-Score differs from IMDB score)


Distribution of Potential Response Variables

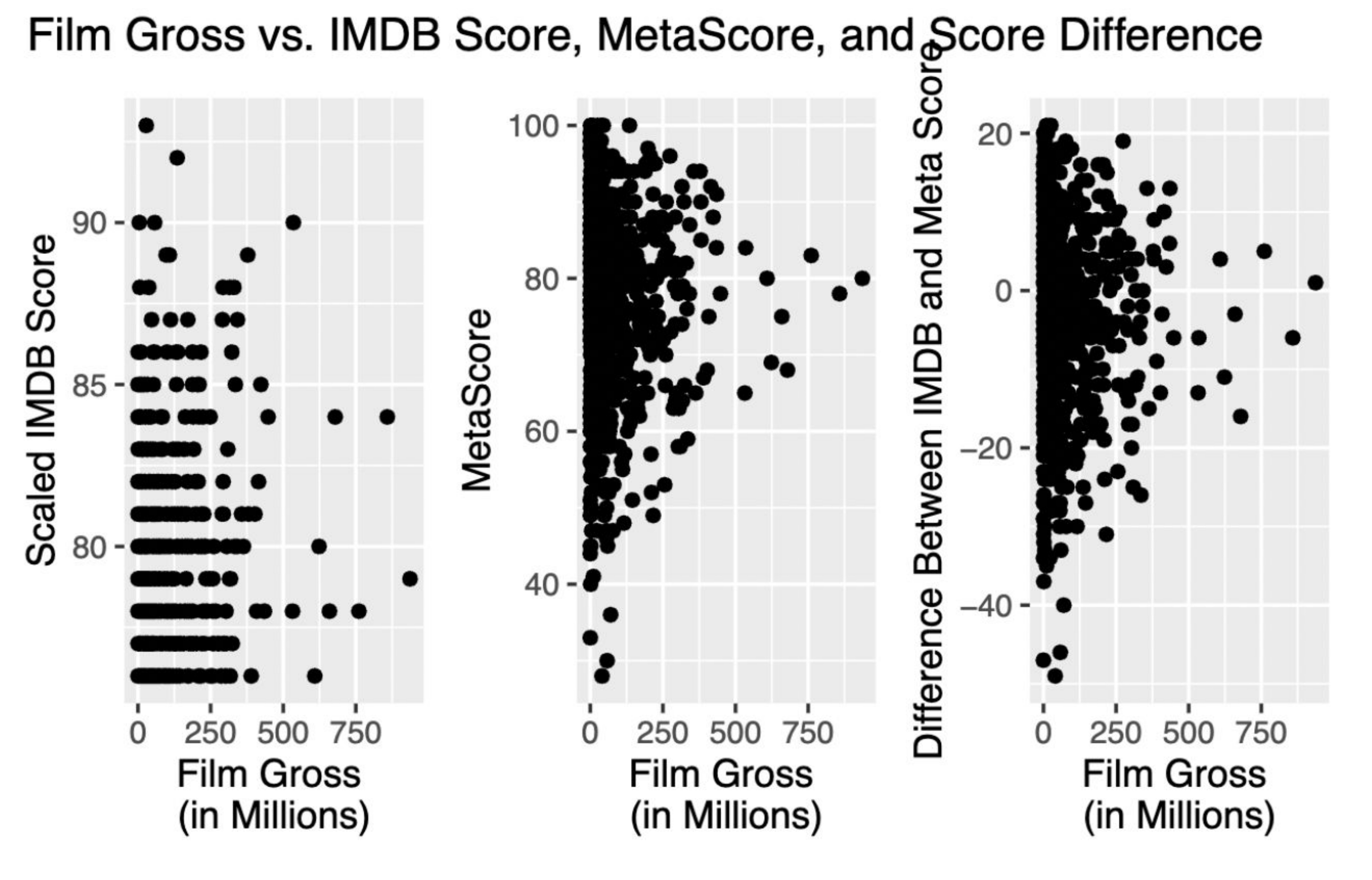

Distribution of Key Numerical Predictors

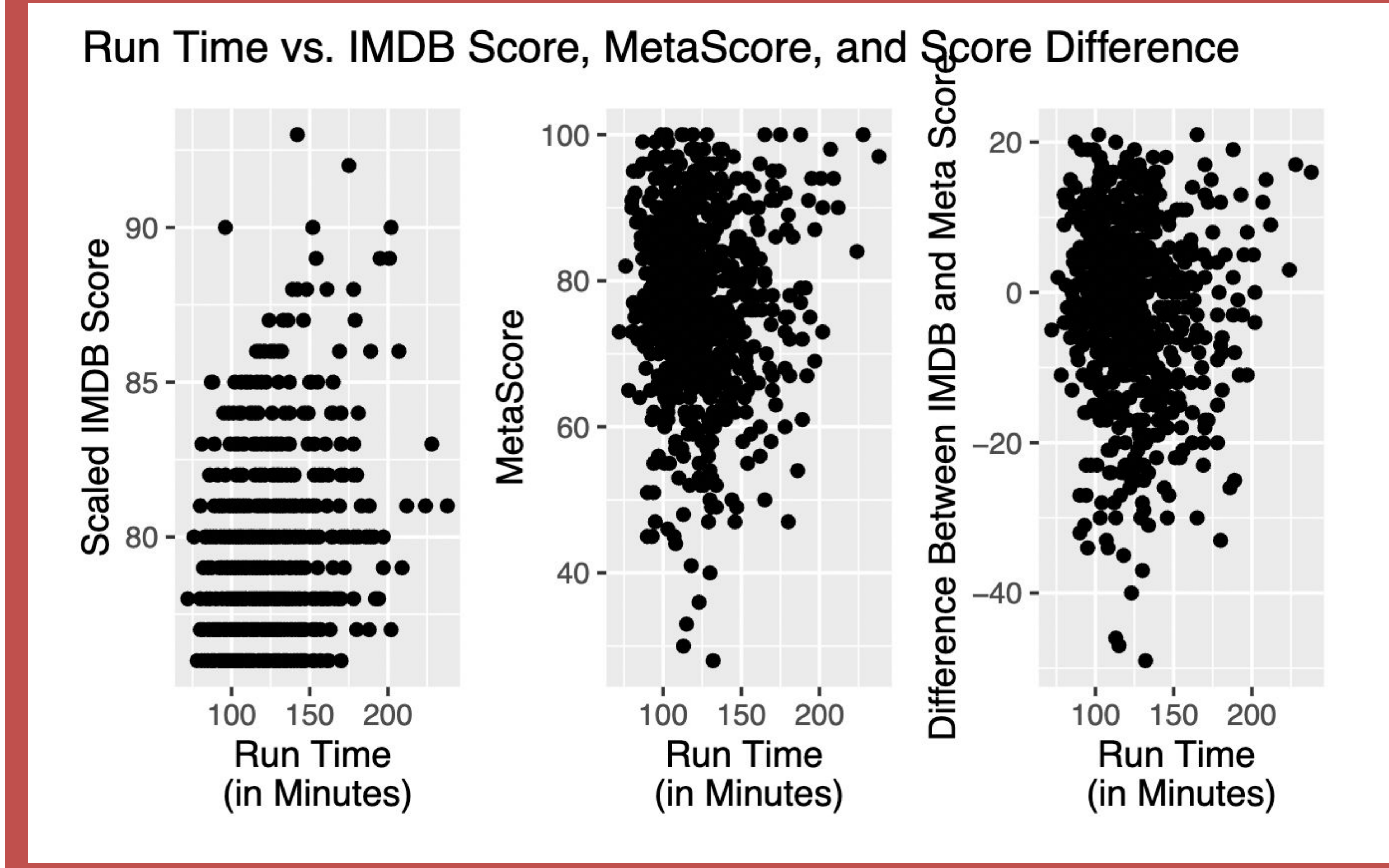

Distribution of Decades Released


Distribution of Censorship Certificates

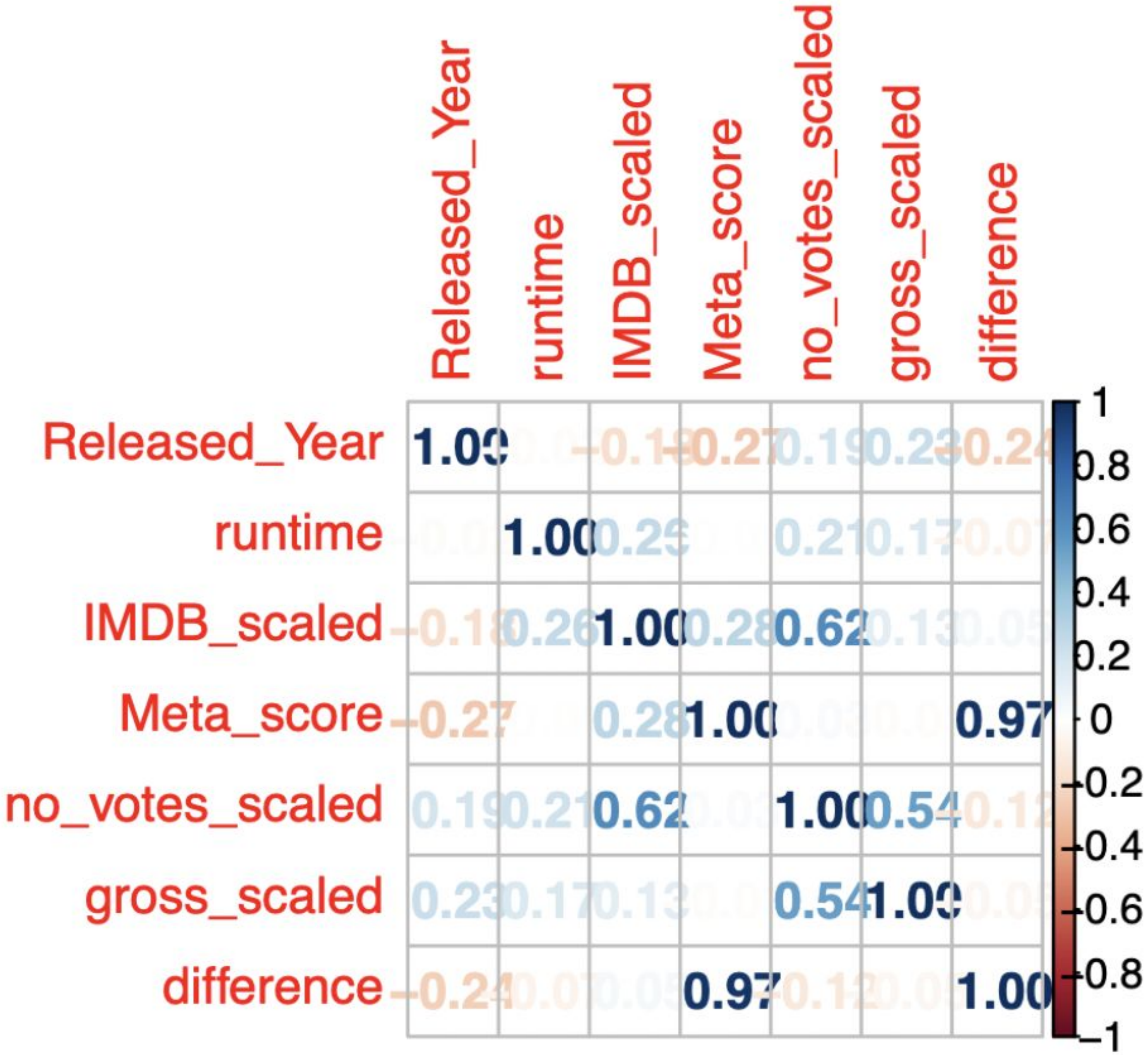
Film Gross/Runtime vs. Predictors

- Plotted gross revenue and run time for all three of our potential variables
  - Non-linear, potential variable transformations?





Potential Multicollinearity



	Released_Year	runtime	IMDB_scaled	Meta_score	no_votes_scaled	gross_scaled	difference
Released_Year	1.00	[illegible]	−0.18	−0.27	0.19	0.23	−0.24
runtime	[illegible]	1.00	0.26	[illegible]	0.21	0.17	−0.07
IMDB_scaled	−0.18	0.26	1.00	0.28	0.62	0.13	0.05
Meta_score	−0.27	[illegible]	0.28	1.00	[illegible]	[illegible]	0.97
no_votes_scaled	0.19	0.21	0.62	[illegible]	1.00	0.54	−0.12
gross_scaled	0.23	0.17	0.13	[illegible]	0.54	1.00	[illegible]
difference	−0.24	−0.07	0.05	0.97	−0.12	[illegible]	1.00

Questions Going Forward

- What response variable approach is best? Should we focus on the difference variable or plot both IMDB and Meta-Score response variables and compare?
- How do we go about our nonlinear predictors?