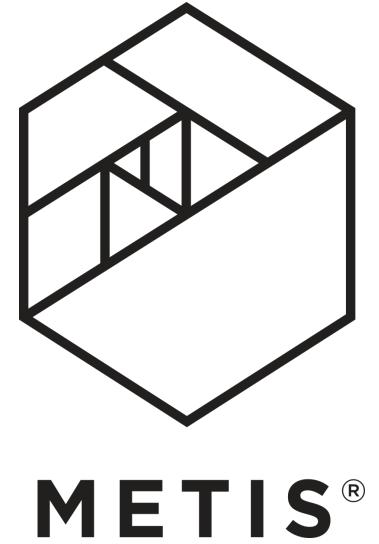


IMDB TV Rating Predictor

METIS Data Science and Machine Learning Bootcamp

by Krystian Krystkowiak, 2022

Introduction

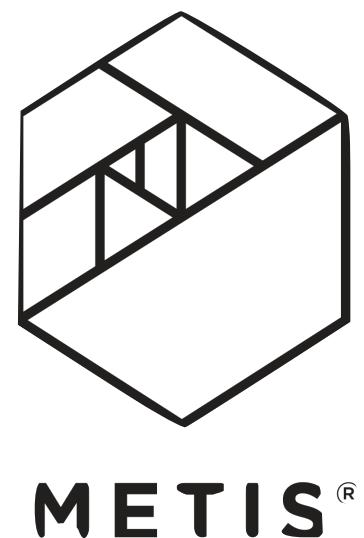
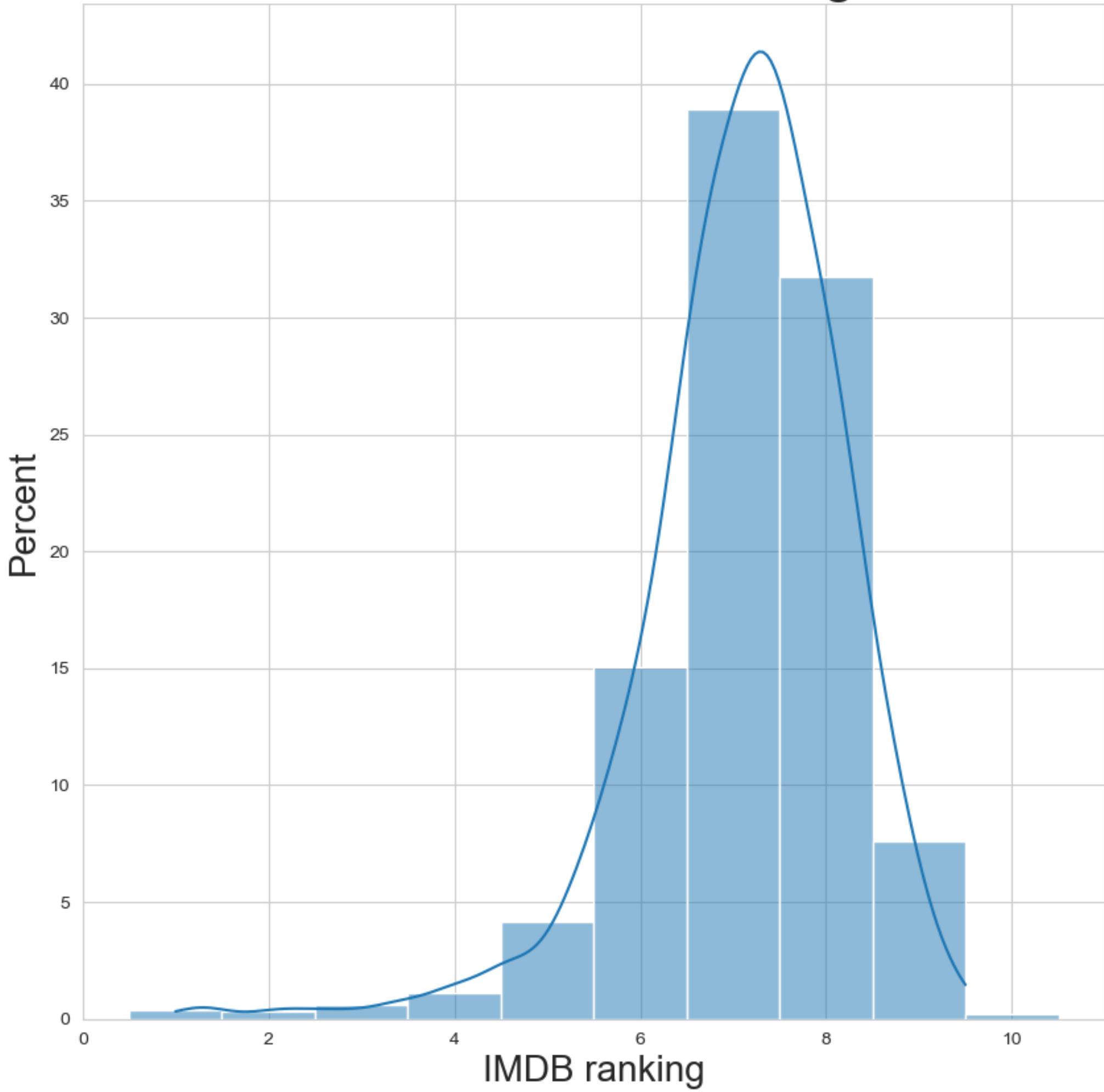


- 2000 - now: **Golden Age of television**
- To create a model that can **predict the reception of a TV show** by viewers.
- GOAL: Identify key **factors that influence a show's success** to help producers during production planning

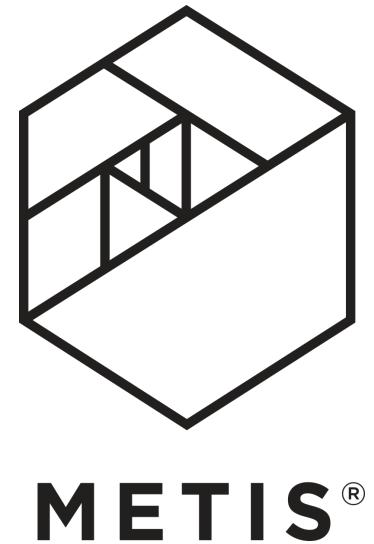
Data

- **IMDB data**
- Shows from **2019-2021** with **at least 1,000 ratings selected**
- Anonymous scraping
- 1335 TV shows
- Target: **IMDB rating**

2019/21 TV shows IMDB ranking distribution



Data

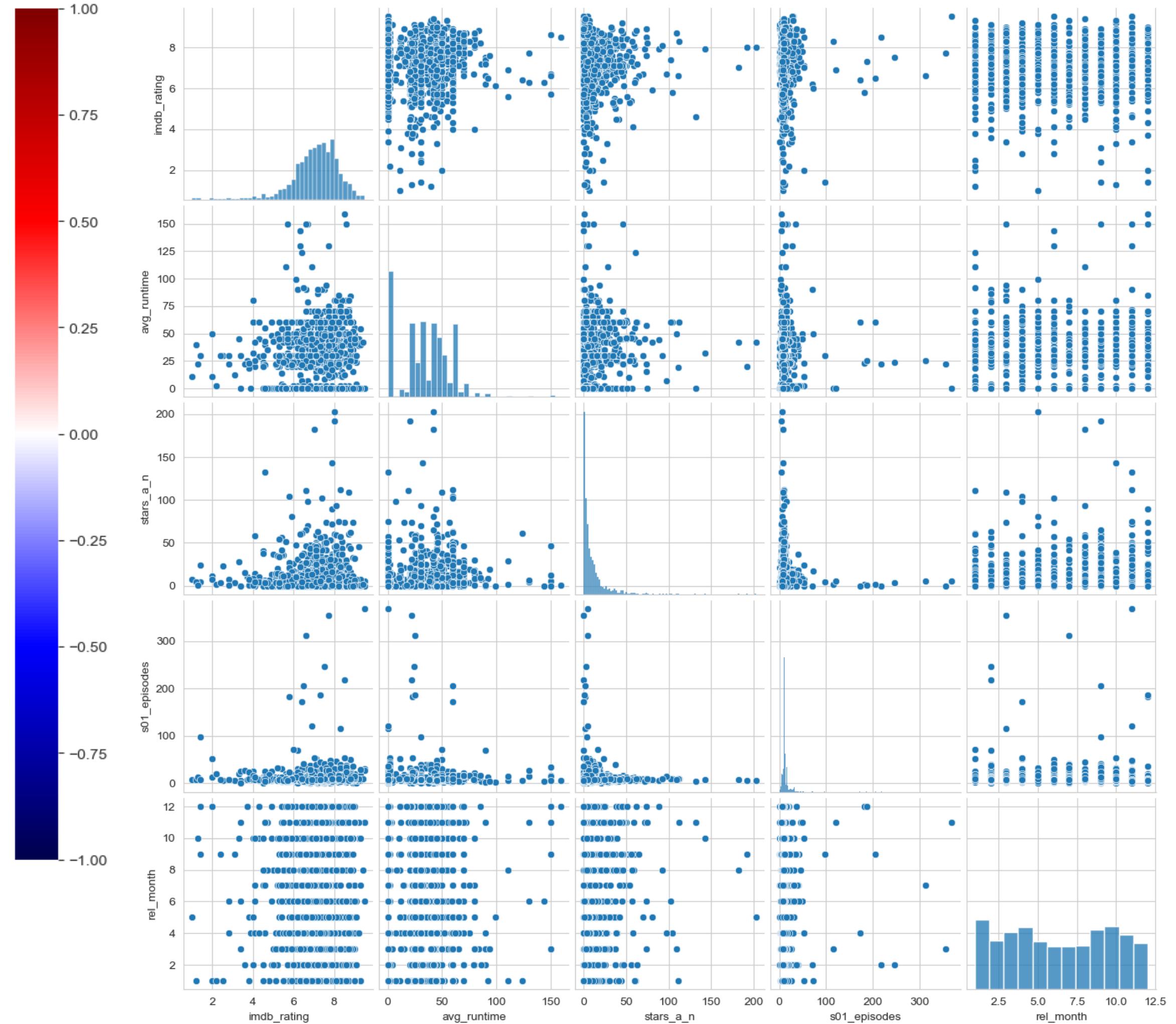
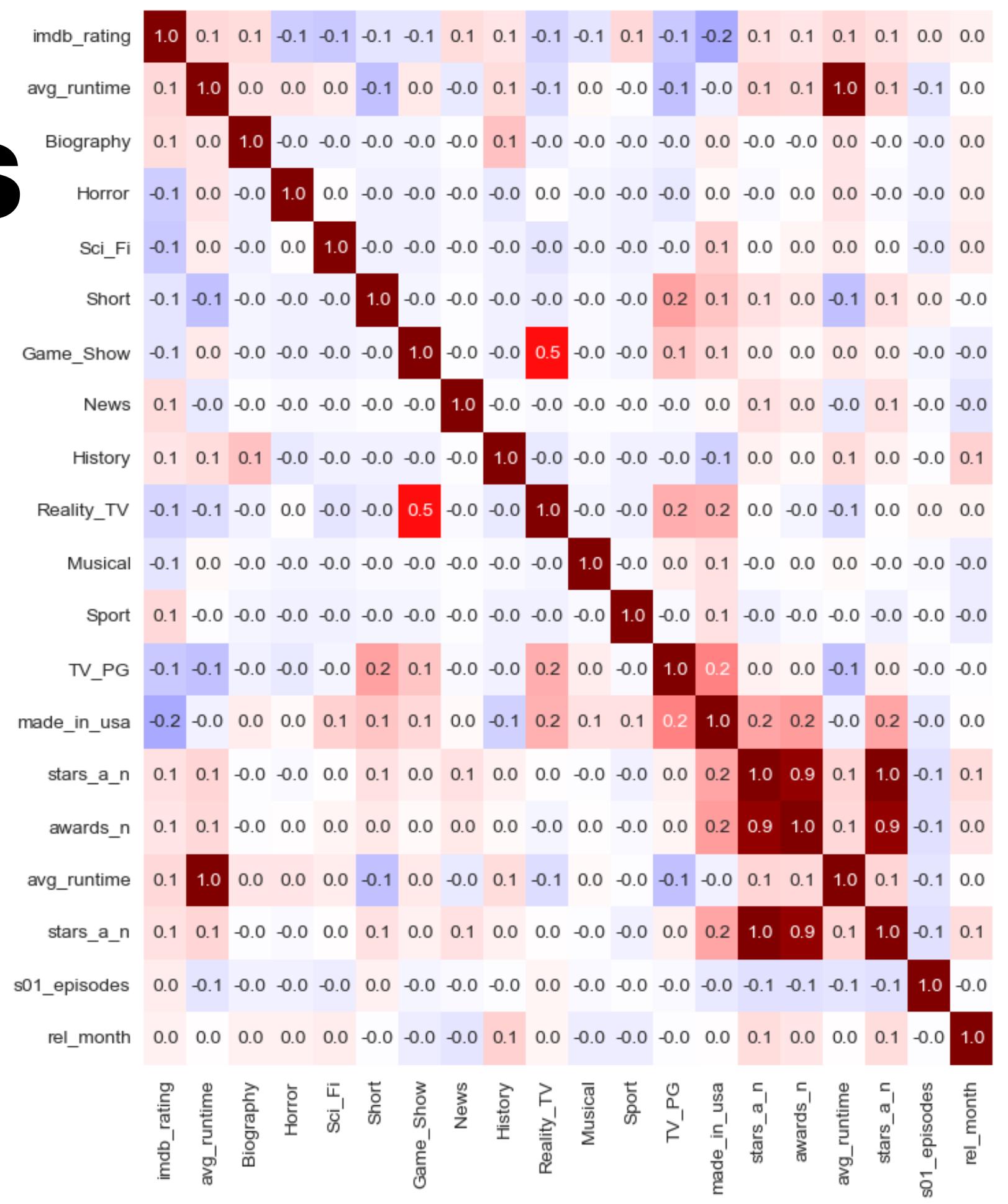


- Features: 11

imdb_rating	title	s01_episodes	avg_runtime	genres	rel_date	certification	origin	company	creators	creators_a	stars	stars_a
5.8	The Masked Singer	10	60	[Game-Show, Music, Reality-TV]	2019-01-02	TV-PG	United States	[Smart Dog Media, Fox Alternative Entertainment...]	[]	[Jenny McCarthy-Wahlberg, Ken Jeong, Nicole Scherzinger]	[3, 5, 1]	
6.2	Siempre Bruja	11	40	[Drama, Fantasy]	2019-01-01	TV-14	Colombia	[Caracol]	[]	[Sofía Araujo Mejía, Angely Gaviria, Sofía Arango]	[0, 0, 0]	
6.6	Tidying Up with Marie Kondo	8	40	[Reality-TV]	2019-01-01	TV-PG	United States	[Netflix, The Jackal Group]	[Marie Kondo]	[Marie Kondo, Charlotte Hervieux, Marie Iida]	[1, 0, 0]	

- The dataset consists of a **majority of categorical features**

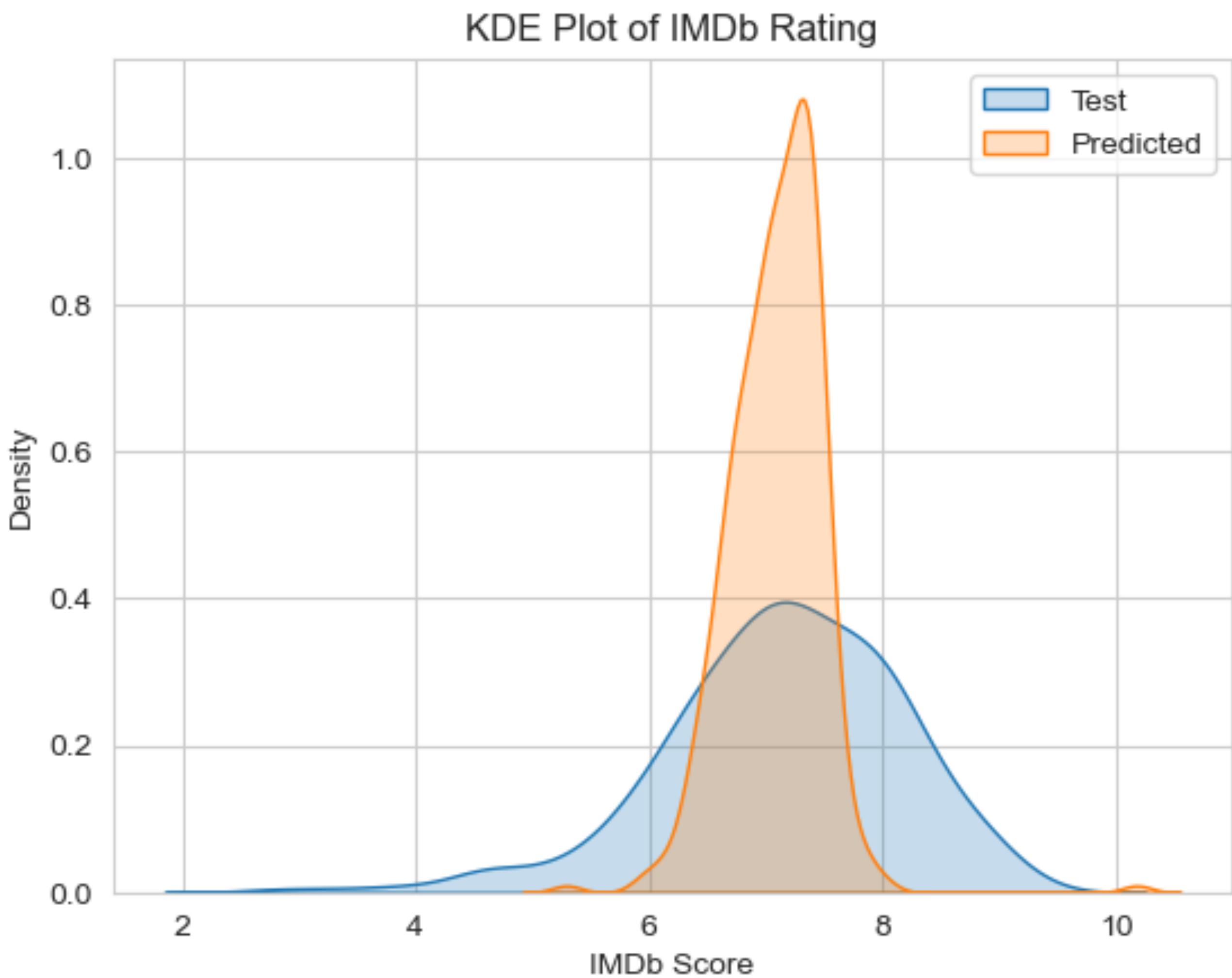
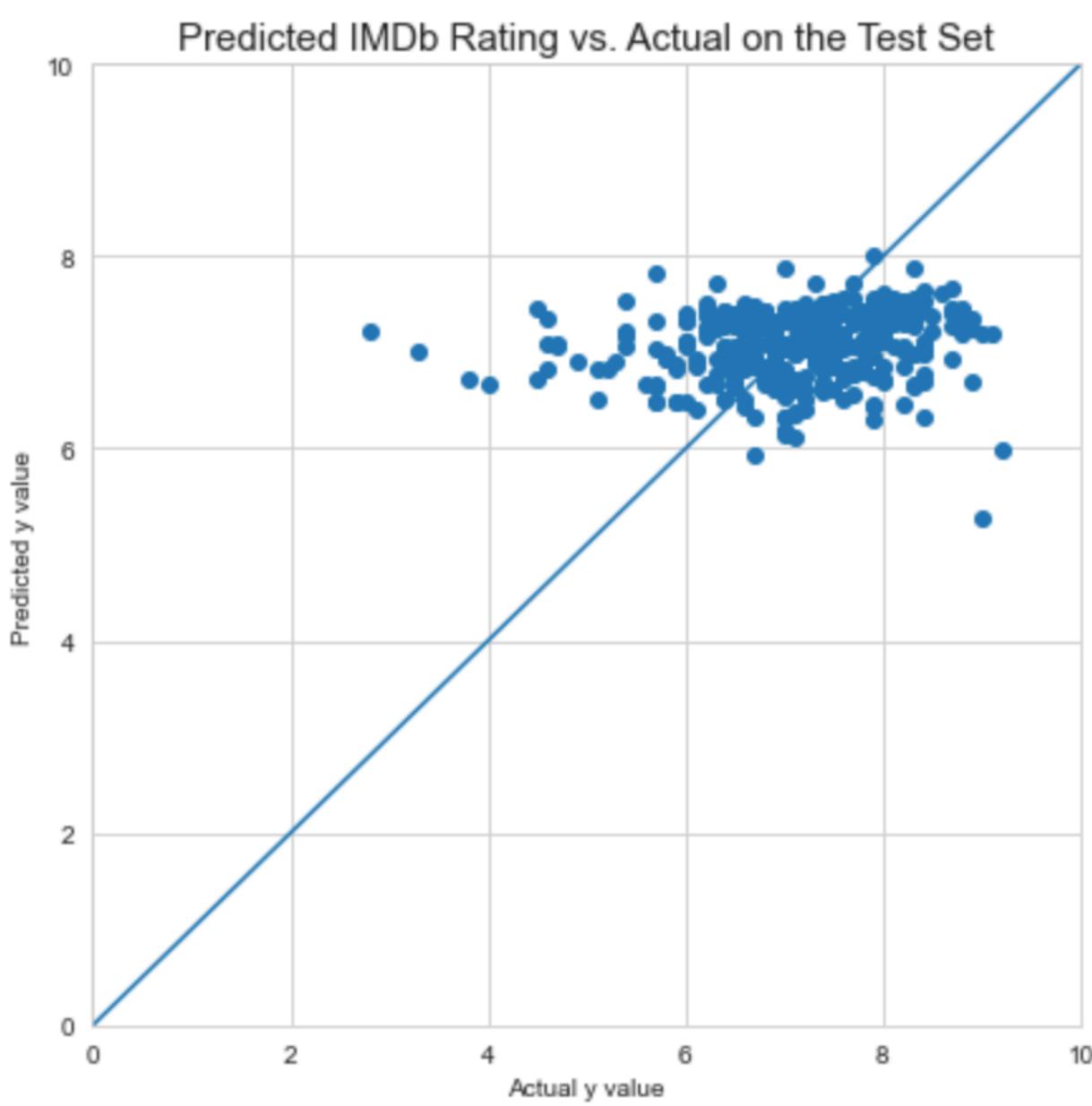
Results



- Features were **not highly correlated** with the target
- Additional data scraped. **New features** were engineered (combinations and counters)

Results

- **Linear Regression**
- 75train/25test split
- 0.096/-0.021 R-squared
- Better predictions around median.



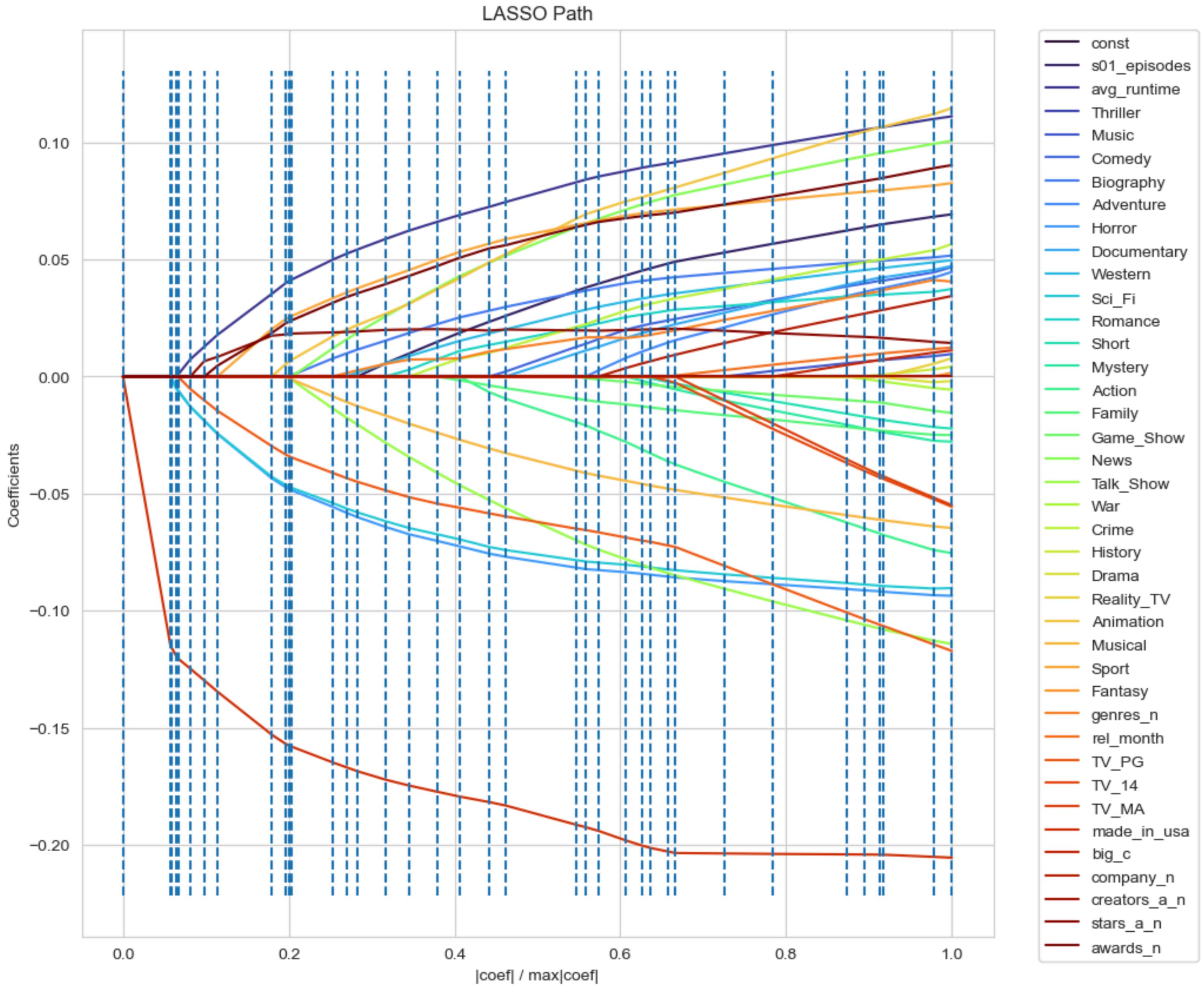
Results

- **Ridge:**

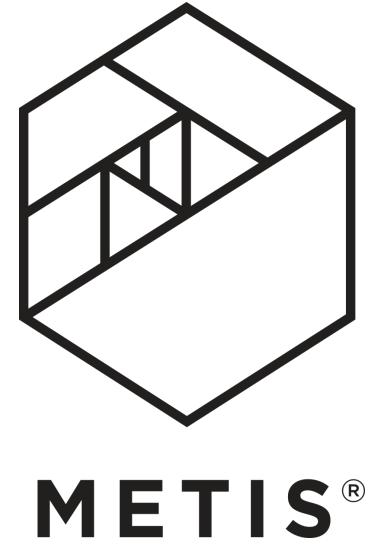
MAE 0.781, R-squared 0

- **Lasso:**

MAE 0.782, R-squared -



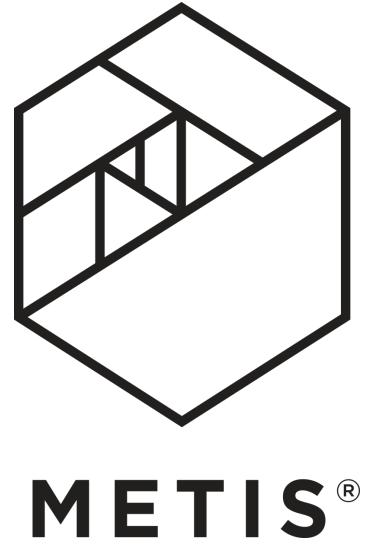
Conclusions



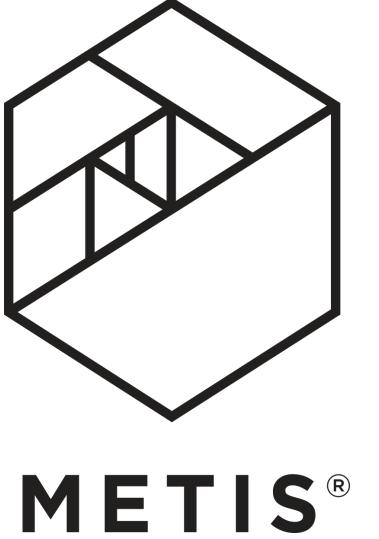
- Animation and news/sport-related shows with awarded main cast tend to be **highly rated**
 - Horror, sci-fi, talk shows, US productions, and PG-rated materials tend to have **lower ratings**
 - The world is **complex**
 - Features were **not strong** predictors of the target variable



Future Work



- **More data** to improve model accuracy
- Limited information available on TV show budgets compared to movie budgets -> **Twitter NLP**
- Investigate **actors' social media** presence
- Examining **top TV show** lists to identify common success factors



Thank you!

Questions?

project for Metis EDA Bootcamp

by Krystian Krystkowiak, 2022