Predicting Total Movie Grosses

Using Regression Modeling to Predict Total Movie Theater Grosses during the Calendar Year

Shannon McDonnell

Why is predicting total movie grosses important?

- It can help producers decide which month or season is best to release a film
- Allows the movie budget to be adjusted based on expected revenues
- Can help decide how many theaters would produce the most gross for the movie

AGENDA:

1. **Approach**
2. **Methodology**
3. **The Data**
4. **Feature Engineering**
5. **Modeling**
6. **Results**
7. **Next Steps**

1. **Approach:** What makes a movie successful?

Let's look at total movie grosses to predict the best parameters to find maximum gross.

Target Data: Total Gross Revenue
Total data points: 3,191

Features: Monthly movie gross
Number of theaters
Release Month
Release Year (2019-2020)

Data Source:

All data was scraped from IMDb's BoxOfficeMojo

Tools Used:

2. Methodology

1. Check for missing data
2. Create dummy variables
3. Run feature distributions
4. Split the data
5. Plot all of the feature distributions
6. Remove all features with a correlation > 0.1
7. Train-test split
8. Perform feature scaling and normalization
9. Regression Modeling

3. The Data: Coefficients

Negative

Categorical Features:

Rank

Release Date

Release Name (movie title)

Distributor

Positive

Quantitative Features:

Gross Estimate $

Total Estimate $

Theaters (total amount)

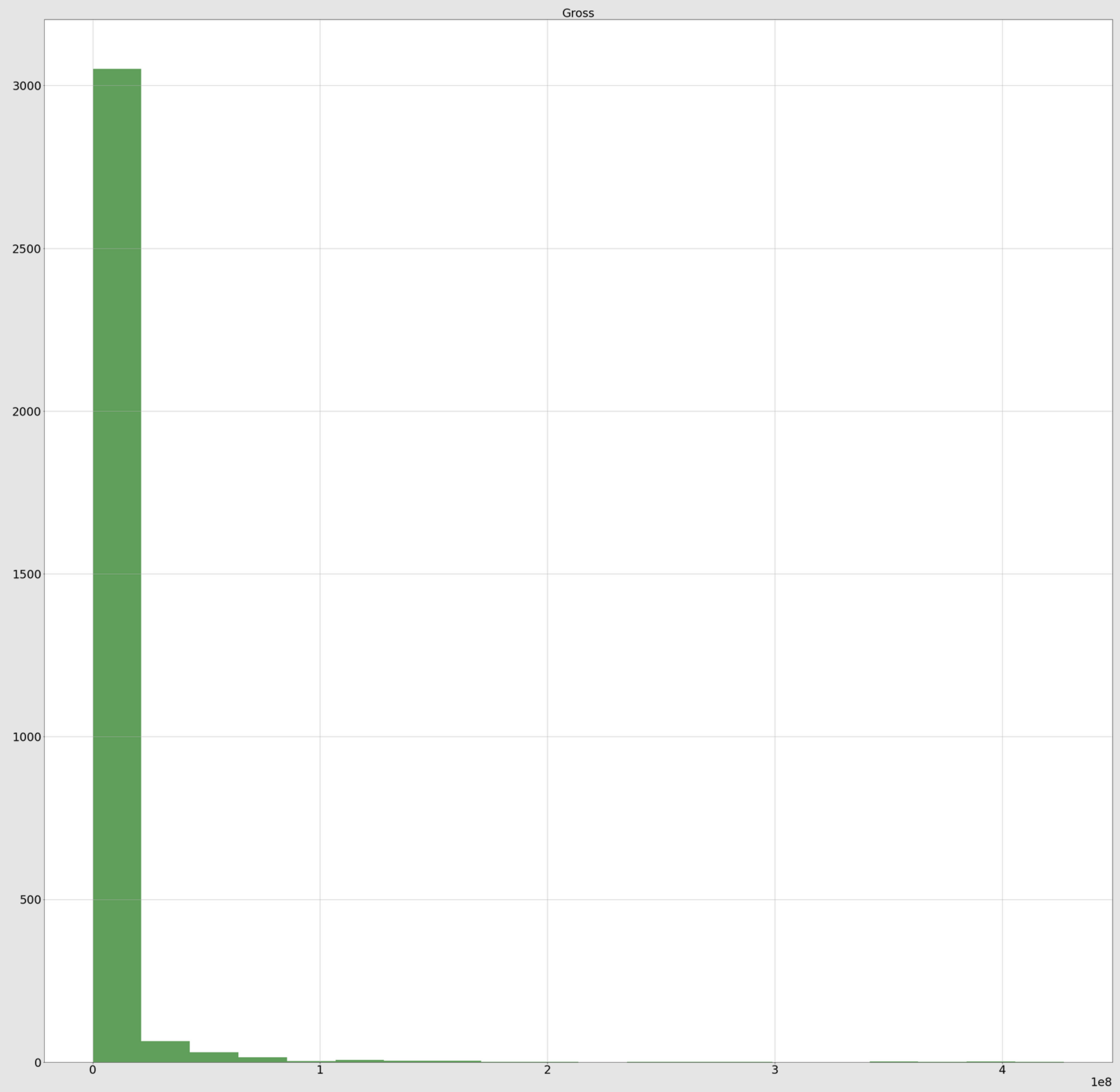
Categorical Features:

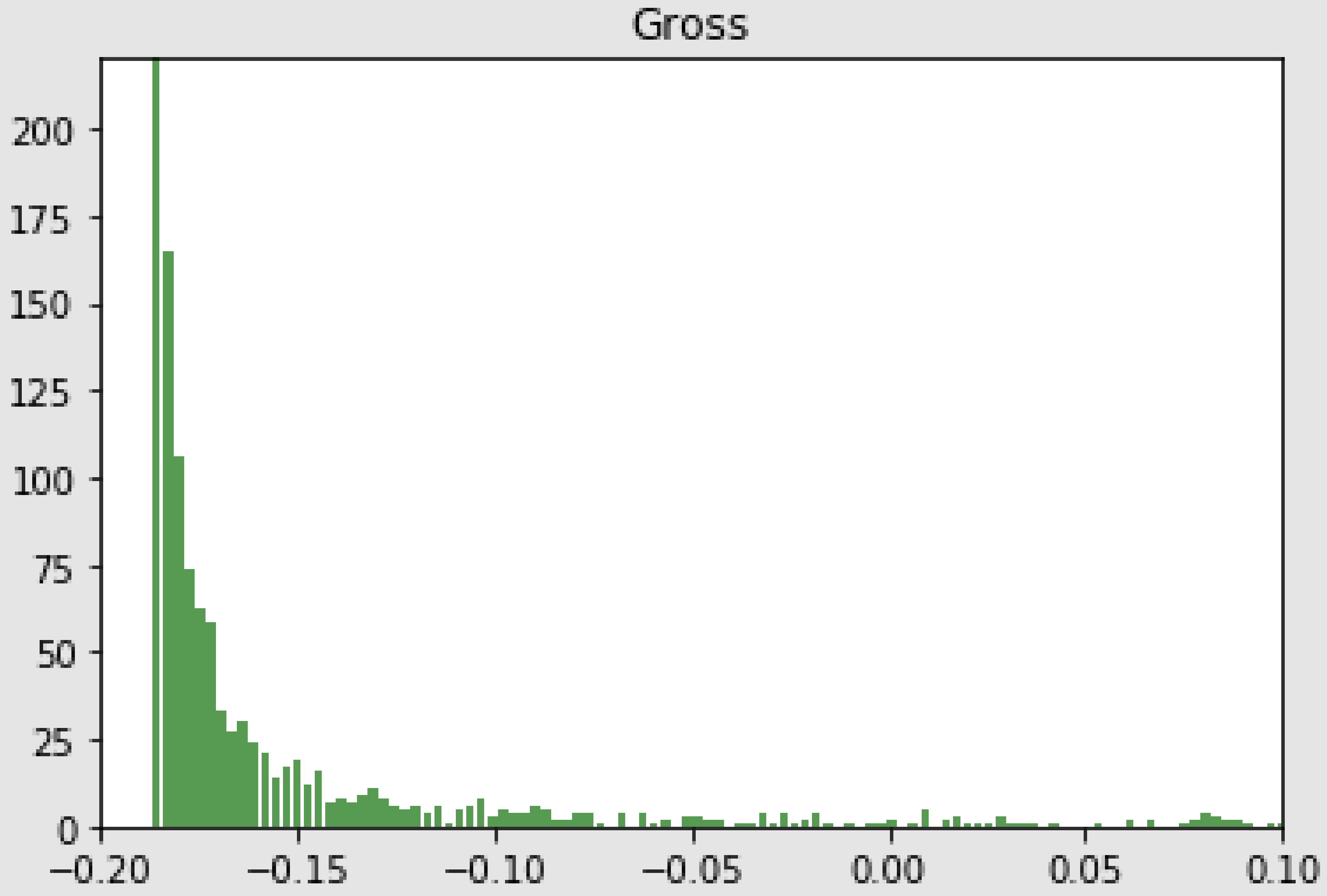
Months (1-12)

Years (2019-2020)

4. Feature Engineering

Feature distribution before and after normalization: **Gross**

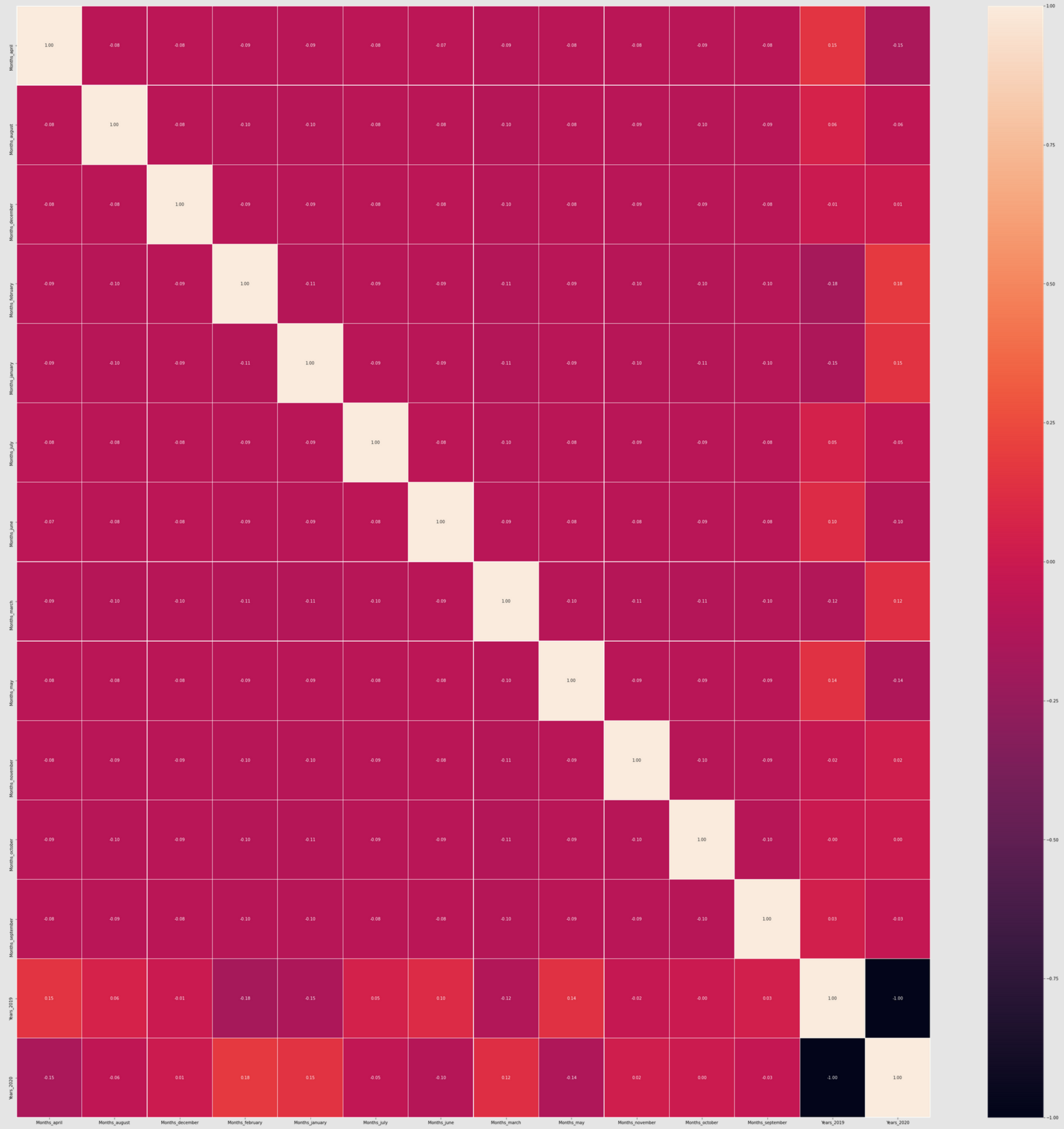




****See Appendix 'B' and 'C' for more images on this****

Examine Features

WEAK CORRELATIONS



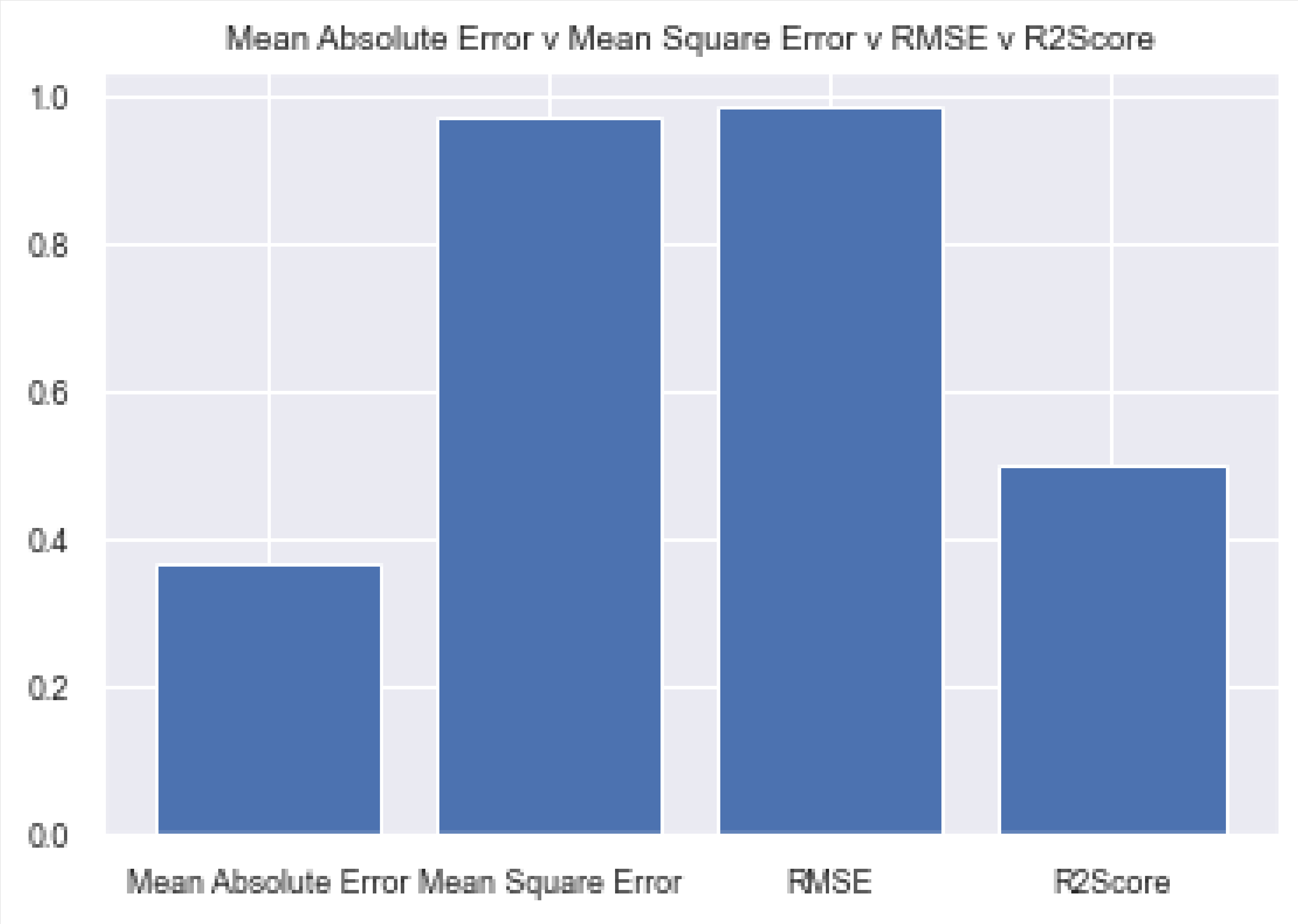
5. MODELING: Linear Regression

R^2 = 0.5011

MSE = 0.9703

MAE = 0.3681

RMSE = 0.9850



MODELING: Lasso Regression

SLIGHTLY WORSE

R^2 = 0.5009

MSE = 0.9701

MAE = 0.3674

RMSE = 0.9851

LASSO COEFFICIENTS:

('Gross', 1.0223928138889704),
('Theaters', 30182.39836992574),
('Months_april', 3823821.8376515703),
('Months_august', 4827778.961552109),
('Months_december', -5519543.147972001),
('Months_february', -322084.72346213204),
('Months_january', -2062942.0395195826),
('Months_july', 6301807.080629837),
('Months_june', 367673.1278506973),
('Months_march', 895340.4868220205),
('Months_may', -1805823.8099229618),
('Months_november', -1769041.533069785),
('Months_october', -2597390.9879438146),
('Months_september', 3695347.566792137),
('Years_2019', 3171720.5583540807),
('Years_2020', -0.0)

MODELING: Ridge Regression

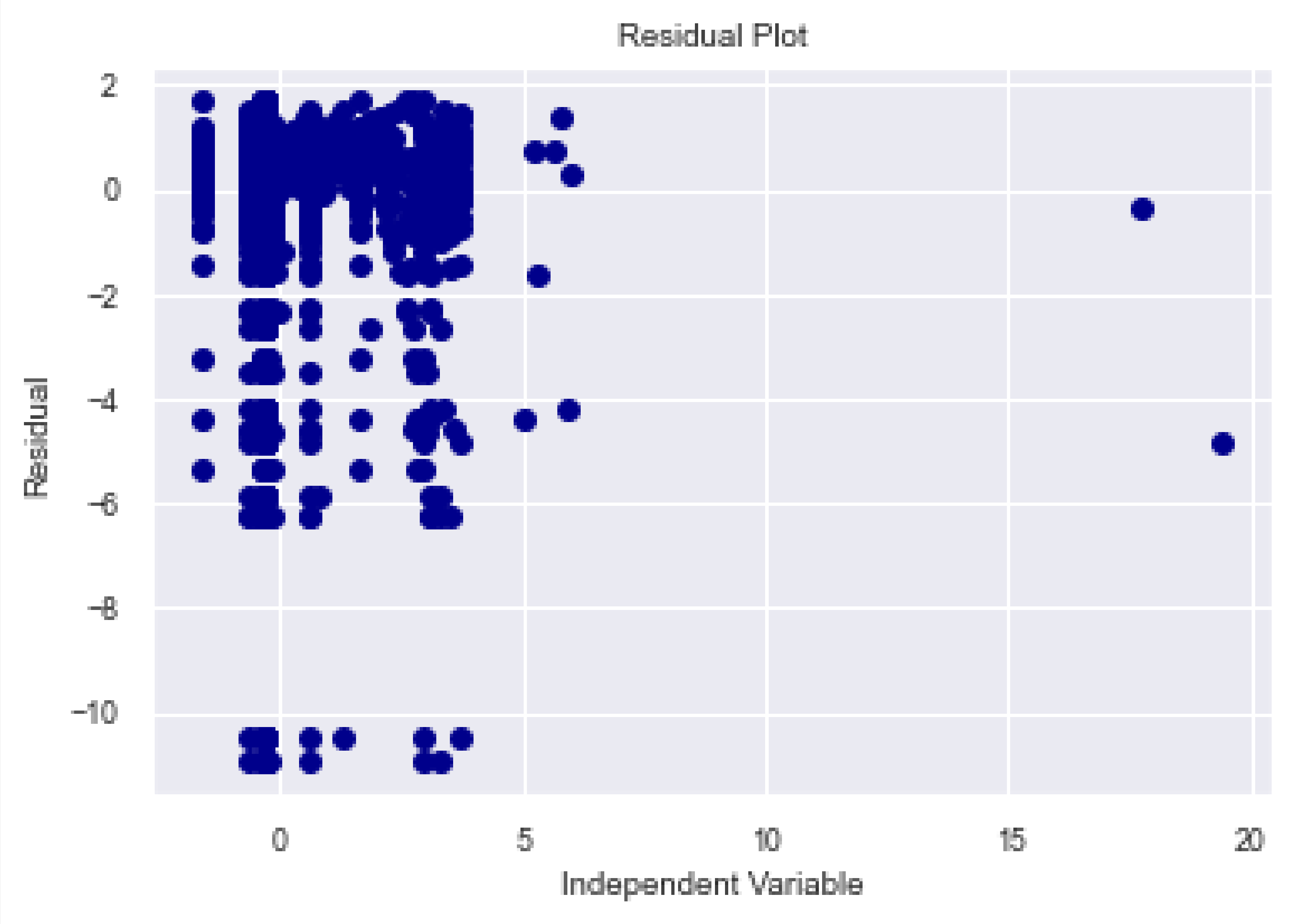
R^2 = 0.4988

MSE = 0.9756

MAE = 0.3639

RMSE = 0.9872

Cross Validation Score: 0.4552



6. **Results:** Final Model

Based off of this data the best predictive model for movie grosses is:

Linear Regression Model

R^2 = 0.50

In this case, simple is best.

7. Next Steps

1. Scrape multiple websites to increase the data set
   - -IMDB, Rotten Tomatoes, AllMovie
2. Include additional features and regressions to observe the model's performance over more time
   - -features like MPAA ratings, reviews, runtime
   - -polynomial regressions
3. Look more specifically into genre
   - -see if this is seasonally correlated
   - -example: horror movies in October
4. Take into account location, holidays and major seasonal events

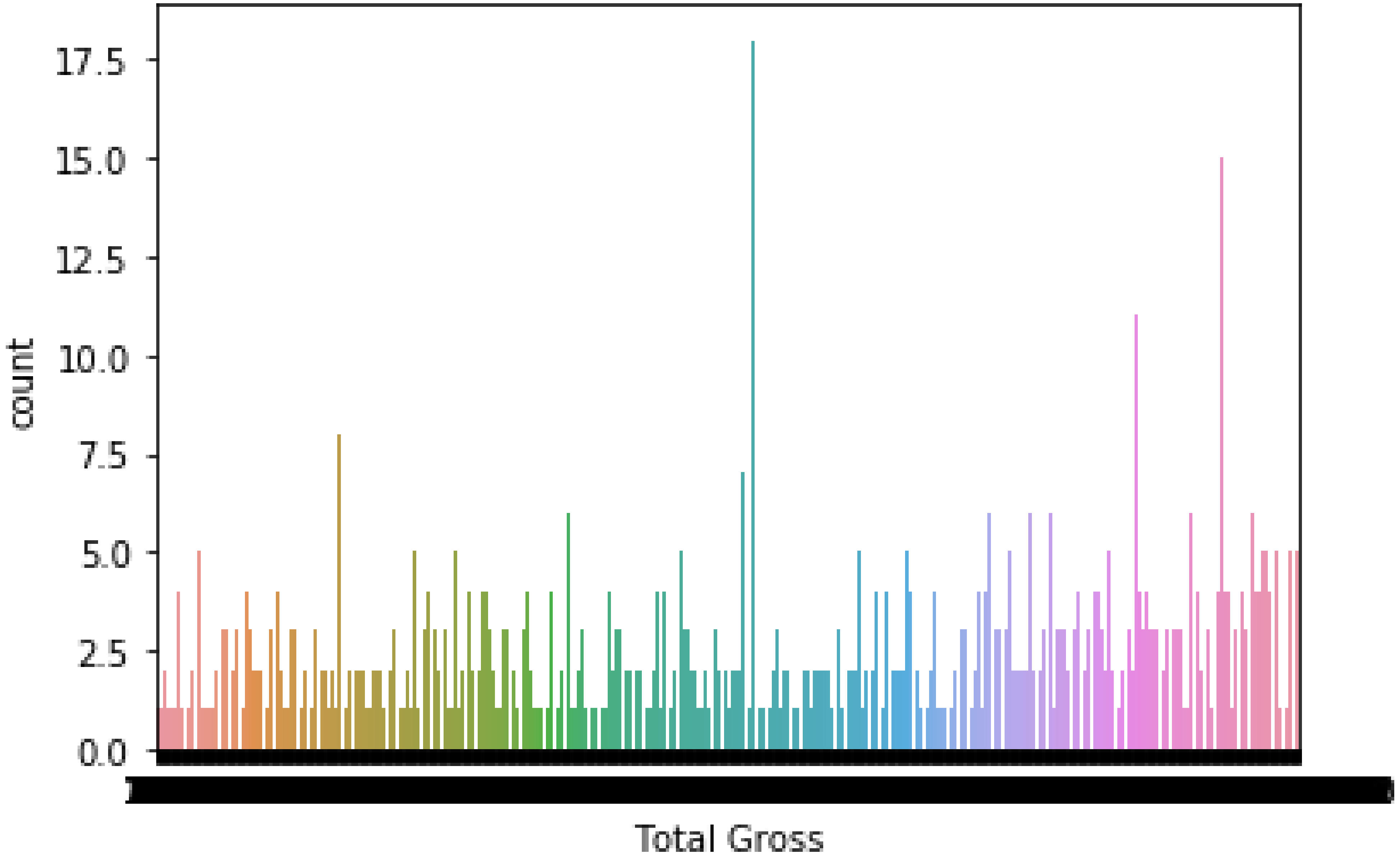
Thank you!

Questions?

Appendix

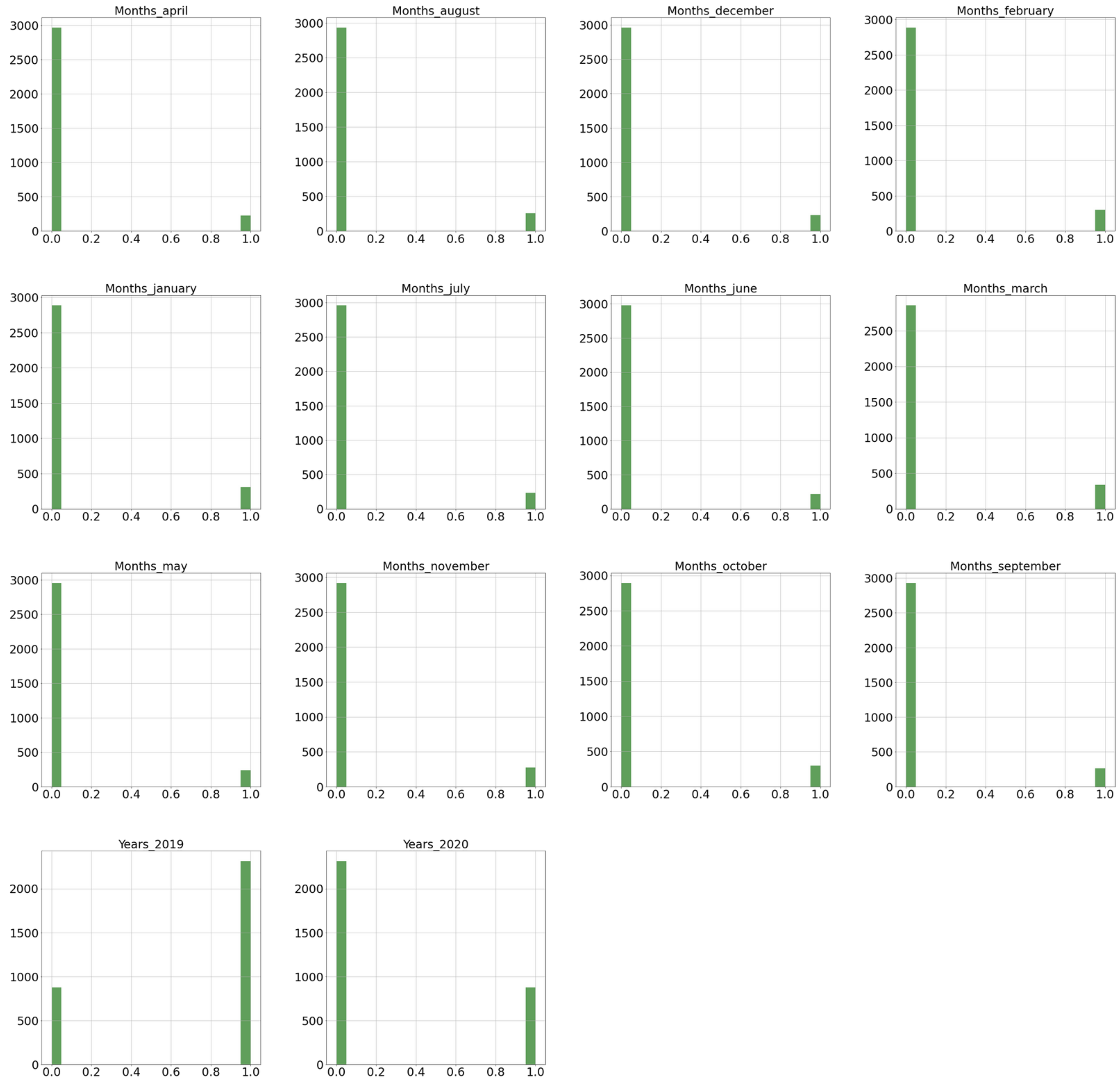
More information and more graphical representations are available via code.

A. Movies with total gross counts



Appendix

B. Feature Distributions



Appendix

C. Standard Scalar Data (transformed)