CAPSTONE-PROJECT

BOX-OFFICE PREDICTION

The Science and Art of Successful Film-Making

Raymond Chun

student id: chun1

Contents

What is Covered

Can we predict a successful film?

"In the past 20 years, as we all know, the movie business has changed on all fronts. But the most ominous change has happened stealthily and under cover of night: the gradual but steady elimination of risk."
-Martin Scorsese

Can we predict a successful film?

We all have a favorite film that we look back and cherish. A superhero film, a romance one, an action movie that made our hearts race. It's hard not to be entranced by the magic of movie-making. It is easy to forget that films are made by countless individuals, actors, directors, but also corporations too. These companies are strategically planning on the next big hit that will make you remember that magical feeling. Movies are an artform fueled by box-office expectations. When the great films make great money, you've got a hit. The question is can we predict if a movie will be a hit or not?

Can we predict a successful film?

Successful films are not easy to define. If a $100 indie film makes $10,000 USD, that's terrific. But if a 1 million USD film makes $10,000, that would be a flop. Success in movies, much like life, is in the eye of the beholder. However, there are certain standards in this domain. Most professional film-makers would categorize a successful film with, but not all of the following:

- Produced by a major studio
- Outperforms its budget
- Creates popularity in the media
- Contends for awards



II

The Data

IMDB data set, Oscars Award data set
Links: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset
https://www.kaggle.com/unanimad/the-oscar-award

The first data set is a robust, comprehensive IMDB movie data set from kaggle.com. The information is quite strong, it consists of categorical and numerical columns from as early as 1894 to now. It is quite extensive and consists of both a budget and box-office performance column which are highly integral to this analysis.

The second data set is about the Oscar awards from 1927-2020 also from kaggle. Unlike the previous data set, it is organized by nominated films of each movie category. Each year there are nominated films and also the films that actually won the award. It has a large, expansive amount of data based on movie history.

III Analysis

Box-Office, Ratio, Oscar Prediction

Outline:

Section 1: Predicting Box-office performance

A. Exploratory Analysis: Correlations, Plots, Histograms, etc
B. Models: Linear Regression, Lasso, Ridge, Elastic Net
C. Results

Section 2: Predicting the ratio of Box-Office over Budget Ratio

A. Adjustments
B. Models: Linear Regression, Lasso, Ridge, Elastic Net
C. Results

Section 3: Predicting Oscar Nominations

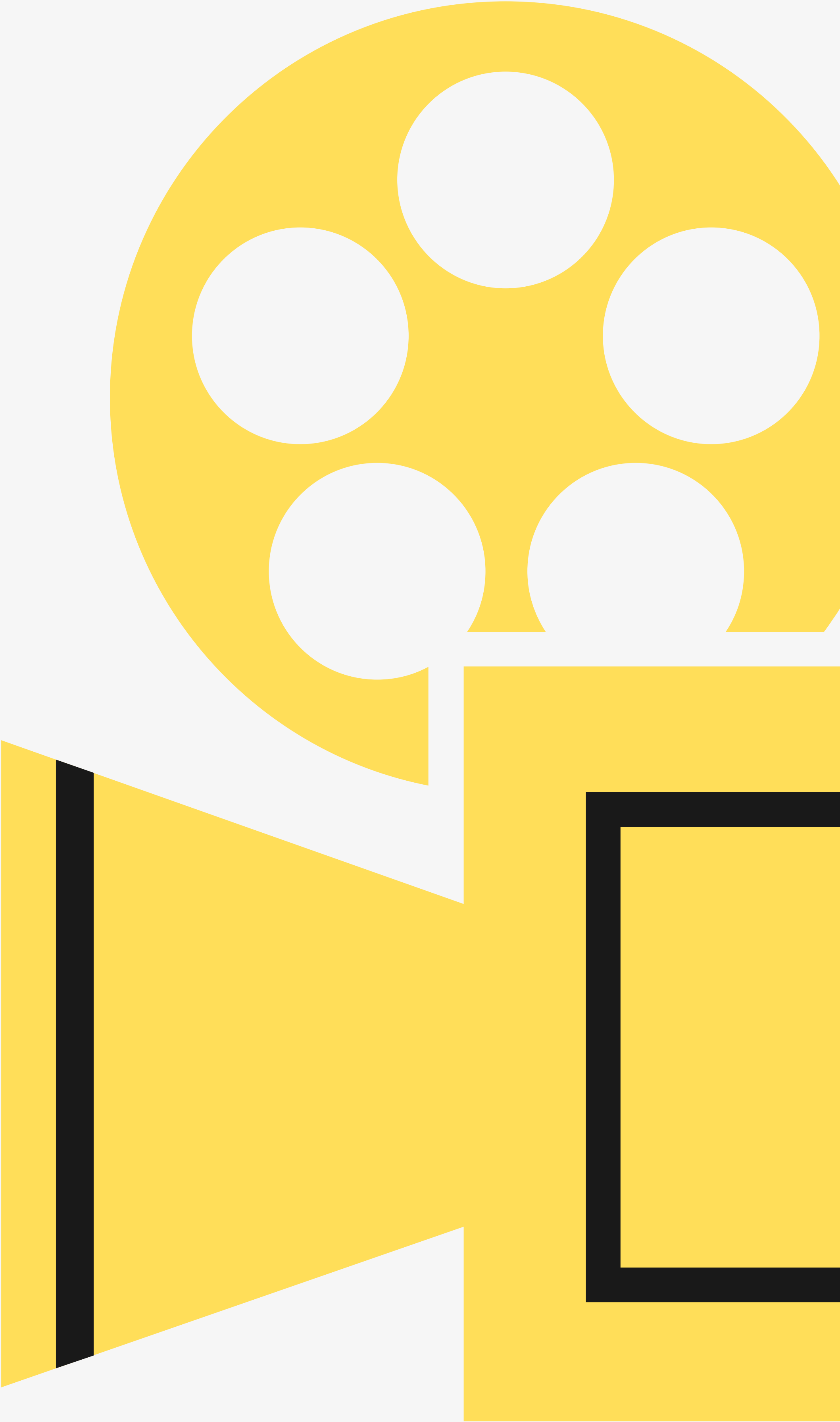
A. Data Filters
B. Models: Lasso, Ridge, Enet, Logistic
C. Results

Analysis Section 1

PREDICTING BOX-OFFICE
PERFORMANCE

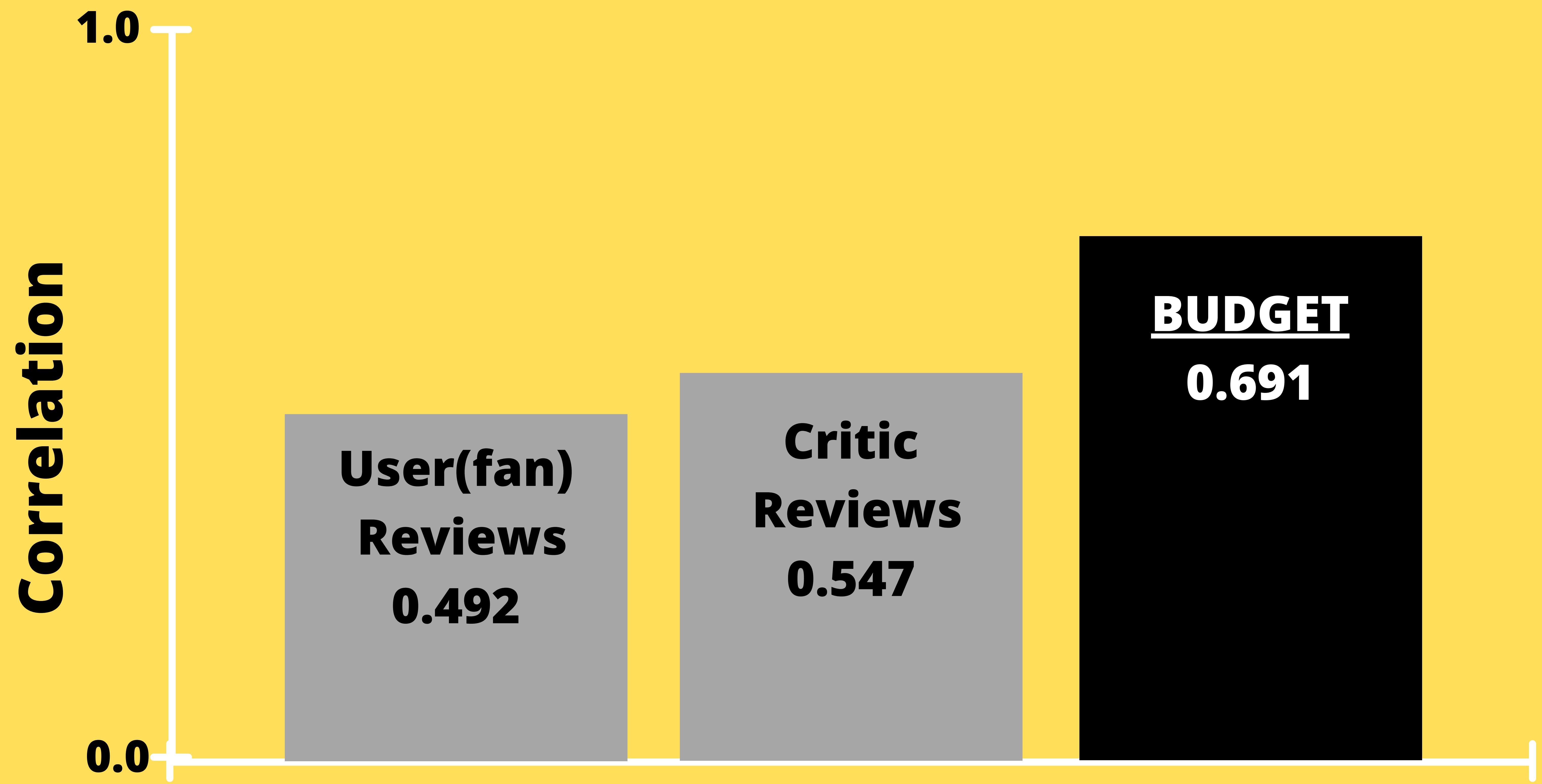
THE PROCESS / ALGORITHM

- Filter the data from the year 1980
- Label the box-office column
- Run models: Lasso, Linear Regression, Ridge, Elastic Net
- Study the results, make adjustments



Exploratory Data

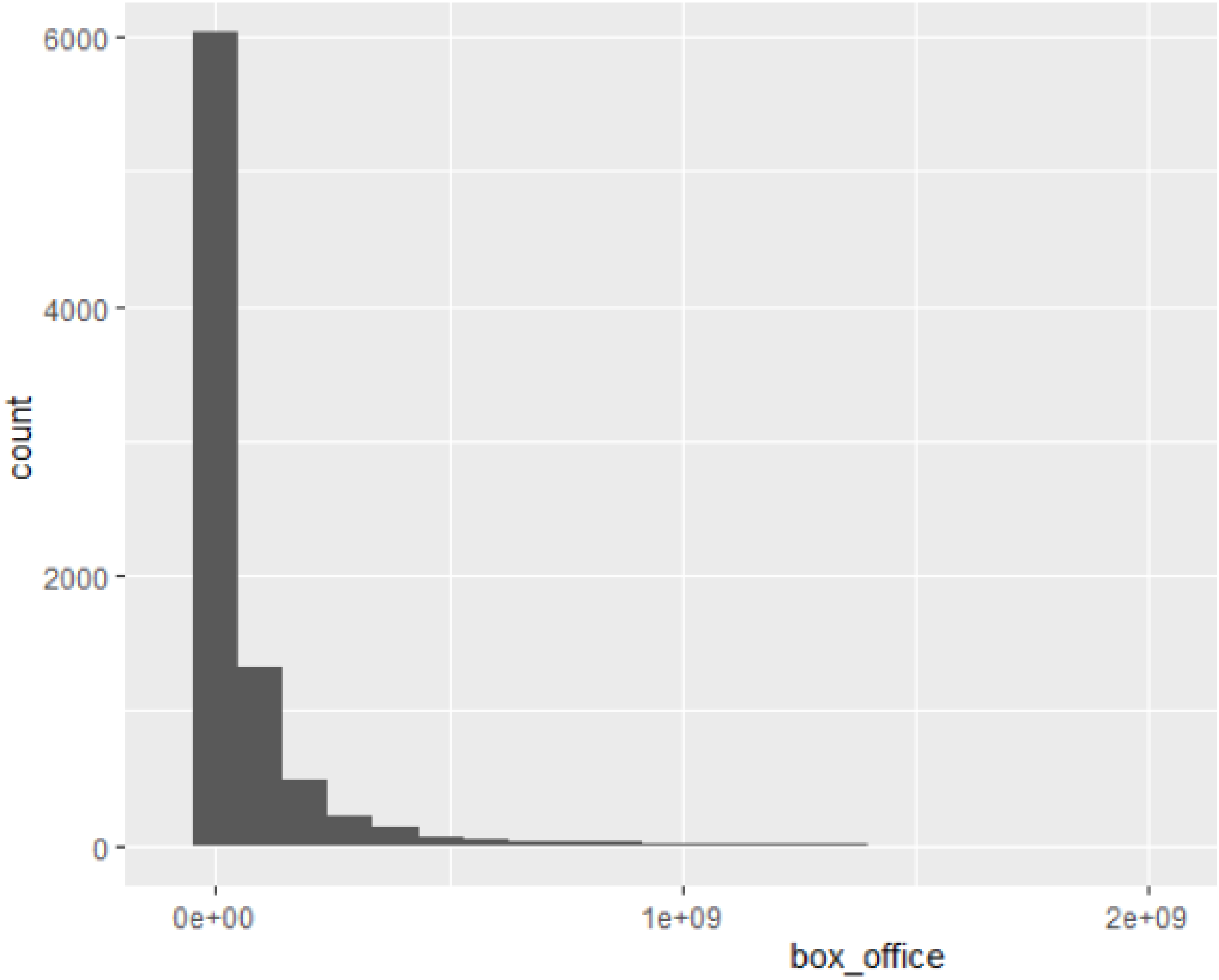
Top 3 numerical variables correlated with Box-office



Exploratory Data

Histogram of Box-Office Numbers

*Highly right-skewed, I later took the log of the column to scale it properly.

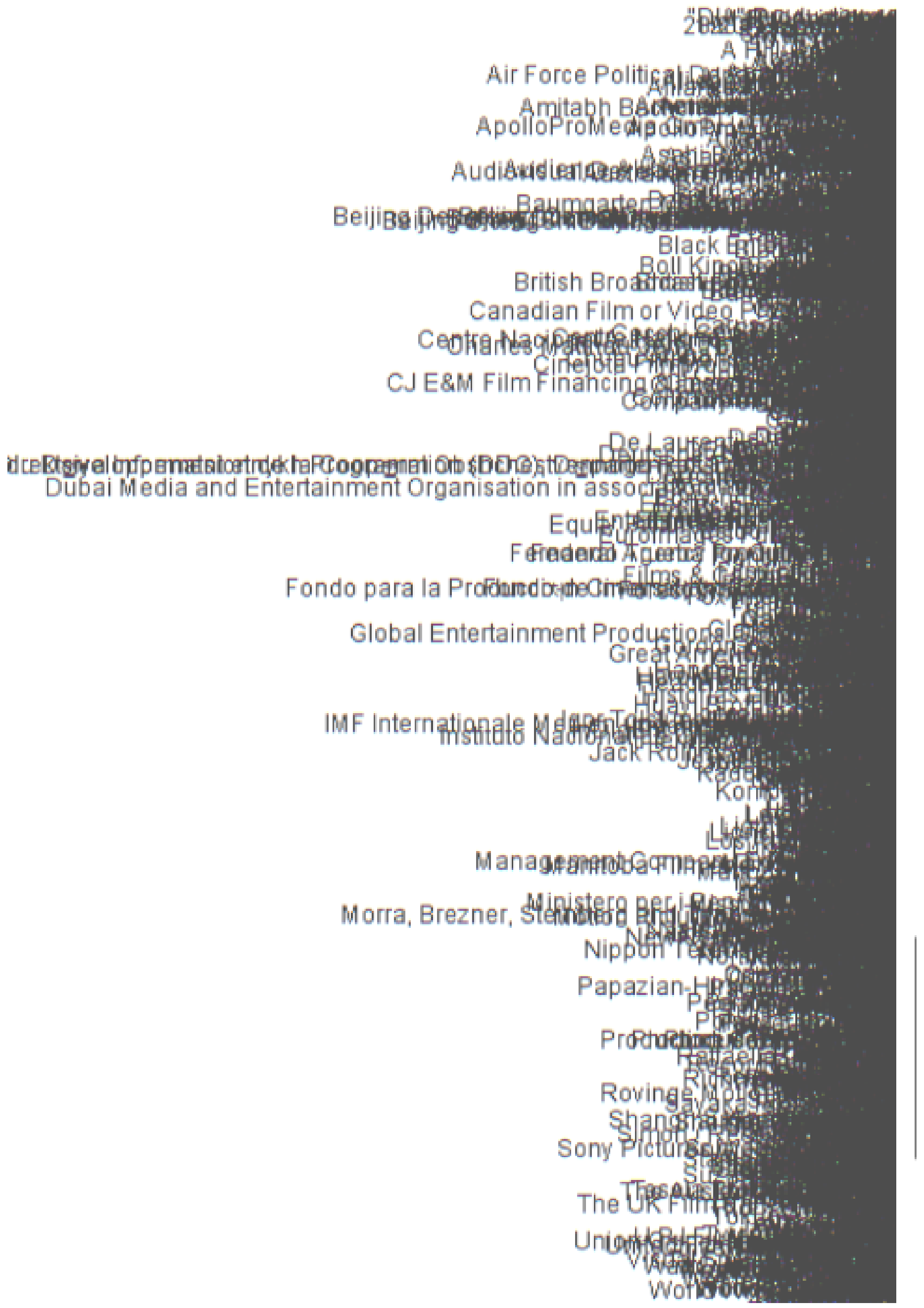


Exploratory Data

Production Companies

*These 55 companies are responsible for nearly ever major film in the US since 1980.

	production_company
1	20th Century Pictures
2	Allied Artists Pictures
3	Artisan Entertainment
4	Cannon Group
5	The Cannon Group
6	Castle Rock Entertainment
7	CBS Films
8	Columbia Pictures
9	DreamWorks
10	DreamWorks Animation
11	FilmDistrict
12	Focus Films
13	Global Road Entertainment
14	Lions Gate Entertainment
15	Lions Gate Films
16	Lucasfilm
17	Marvel Enterprises
18	Marvel Entertainment
19	Marvel Studios
20	Metro-Goldwyn-Mayer (MGM)
21	Metro-Goldwyn-Mayer Animation
22	Metro-Goldwyn-Mayer British Studios
23	Miramax
24	New Line Cinema
25	New World Pictures
26	Orion Pictures
27	Overture Films
28	Paramount Pictures
29	Pixar Animation Studios
30	PolyGram Filmed Entertainment
31	Relativity Media
32	Republic Pictures
33	The Samuel Goldwyn Company
34	Samuel Goldwyn Films
35	Sony Pictures Animation
36	Sony Pictures Classics
37	Sony Pictures Entertainment
38	Sony Pictures Entertainment (SPE)
39	Summit Entertainment
40	Touchstone Pictures
41	TriStar Pictures
42	Turner Pictures (I)
43	Twentieth Century Fox
44	United Artists
45	Universal Pictures
46	Viacom Enterprises
47	Viacom Productions
48	Walt Disney Animation Studios
49	Walt Disney Feature Animation Florida
50	Walt Disney Pictures
51	Walt Disney Productions
52	Warner Bros.
53	Warner Bros. Pictures
54	The Weinstein Company
55	Weintraub Entertainment Group



Filtering the Data

The Super Set of Variables

Filters

- 1980-2020: The 80's are considered the first modern decade artistically, technologically, and through production value.
- Major Studios: I created a join with the modern production studios in America. These studios have consistently produced a vast majority of the major films.
- Box-office: Box-office was not in the original set, I set it to USD and converted it to the total gross world-wide.

Models for Box-Office

Linear Regression, Lasso, Ridge, Elastic Net

LINEAR REGRESSION

predictionsLR= predict(linearRegression, movies_test)
RMSE(predictionsLR, movies_test$box_office)
R2(predictionsLR, movies_test$box_office)

RMSE: 0.6091574
R2: 0.7006303

LASSO

predictionsL <- predict(lasso, movies_test)
RMSE(predictionsL, movies_test$box_office)
R2(predictionsL, movies_test$box_office)

RMSE: 0.6098021
R2: 0.7023074

Models for Box-Office

Linear Regression, Lasso, Ridge, Elastic Net

RIDGE

```
predictionsRidge1 <- predict(ridge,movies_test)
RMSE(predictionsRidge1, movies_test$box_office)
R2(predictionsRidge1, movies_test$box_office)
```

RMSE: 0.6105359
R2: 0.7052497

ENET

```
predictionsElasticNet1 <- predict(enet, movies_test)
RMSE(predictionsElasticNet1, movies_test$box_office)
R2(predictionsElasticNet1, movies_test$box_office)
```

RMSE: 0.6120553
R2: 0.7055012

The Problems

INVESTMENT

The R2 and RMSE results are definitely strong, but the issue is the budget and box-office relationship. Logically, if you invest more money than you should see a greater return. But the sales would still need to be 2-3 times the size of the budget to make a strong profit, especially after marketing.

SKEWNESS

Despite taking the log, the box-office variable is still highly skewed. What this means is that some films are simply outliers to the point where they shape the data and results too much. Some films are also not geared for financial goals, but rather artistic expression.

TIME PERIOD

It's worth noting that film cinema history is unique and important. But it might be more practical to have a modern viewpoint. Films from 1980 may not be as useful as films from more recent decades.

Adjustments

PREDICT A RATIO

Instead of box-office, lets predict a ratio compared to its budget. This will give a more practical metric. This ratio will represent the proportional success a film can have relative to is budget. This will also add some more scale.

REMOVE A FEW KEY COLUMNS

If we use a ratio, we will have remove the usa, worldwide box-office results, along with the budget. This will undoubtedly be less accurate, but the target variable shouldn't be related.

TIME PERIOD

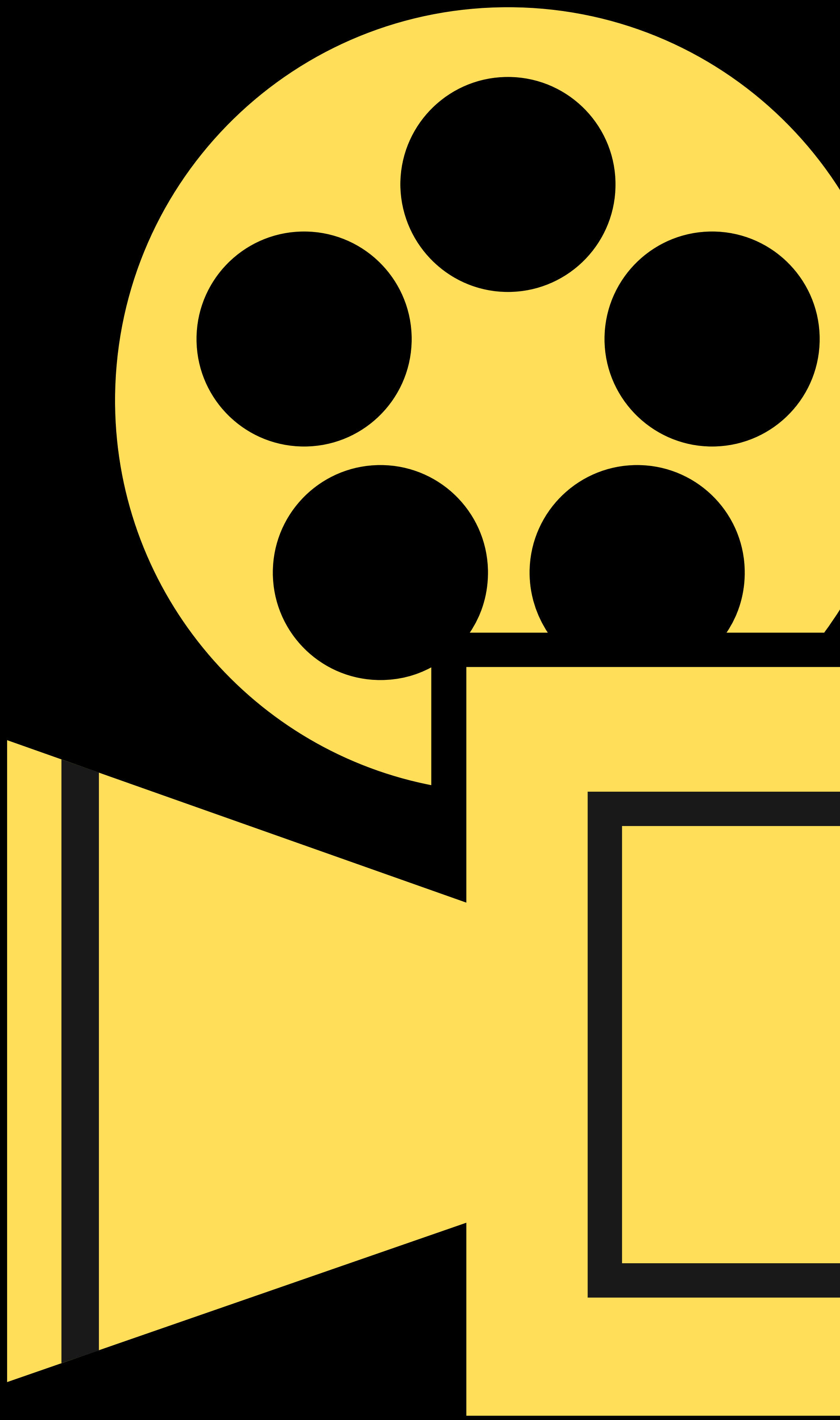
Let's tighten up the time period from 1993 to now. 1993 marks when IMDB went online and when the internet was in the early stages of movie review aggregation. We also see the rise of movie rentals. We cut down on about 800 films by doing this.

Analysis Section 2

PREDICTING THE BOX-OFFICE/BUDGET RATIO

THE PROCESS / ALGORITHM

- Filter the data from the year 1993
- Create a box-office/budget ratio column
- Run models: Lasso, Linear Regression, Ridge, Elastic Net



Models for the RATIO

Linear Regression, Lasso, Ridge, Elastic Net

LINEAR REGRESSION

```
predictionsLR= predict(linearRegression,
movies_test)
RMSE(predictionsLR, movies_test$ratio)
R2(predictionsLR, movies_test$ratio)
```

RMSE: 0.9644851
R2: 0.2484607

LASSO

```
coef(lasso$finalModel,
lasso$bestTune$lambda)
predictionsL <- predict(lasso, movies_test)
RMSE(predictionsL, movies_test$ratio)
R2(predictionsL, movies_test$ratio)
```

RMSE: 0.9623387
R2: 0.2479994

Models for the RATIO

Linear Regression, Lasso, Ridge, Elastic Net

RIDGE

```
predictionsRidge <- predict(ridge,movies_test)
RMSE(predictionsRidge, movies_test$ratio)
R2(predictionsRidge, movies_test$ratio)
```

RMSE: 0.9629903
R2: 0.2444957

ELASTIC NET

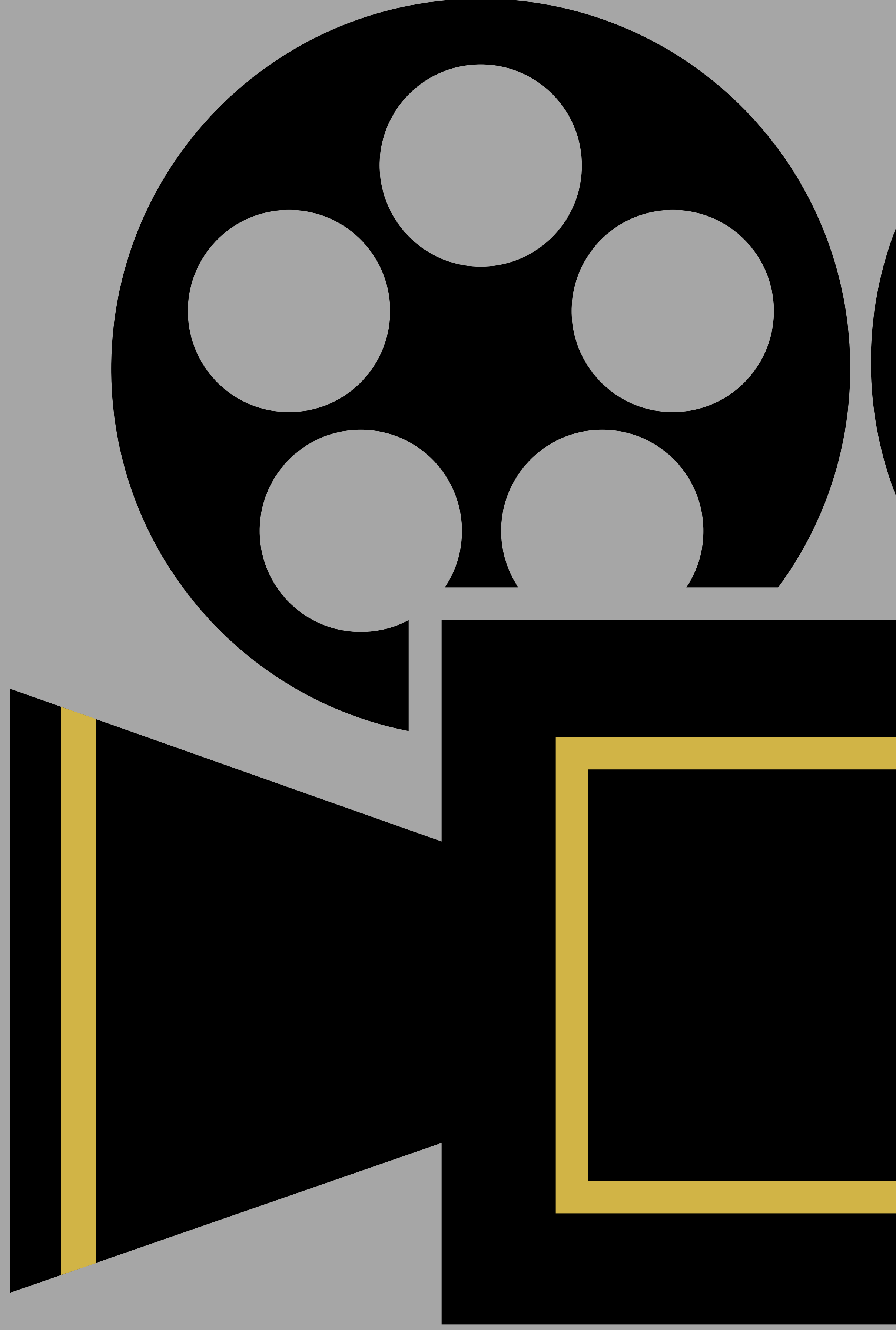
```
predictionsElasticNet <- predict(enet, movies_test)
RMSE(predictionsElasticNet, movies_test$ratio)
R2(predictionsElasticNet, movies_test$ratio)
```

RMSE: 0.9621042
R2: 0.2466875

Analysis Section 3

PREDICT OSCAR NOMINATIONS THROUGH CLASSIFICATION

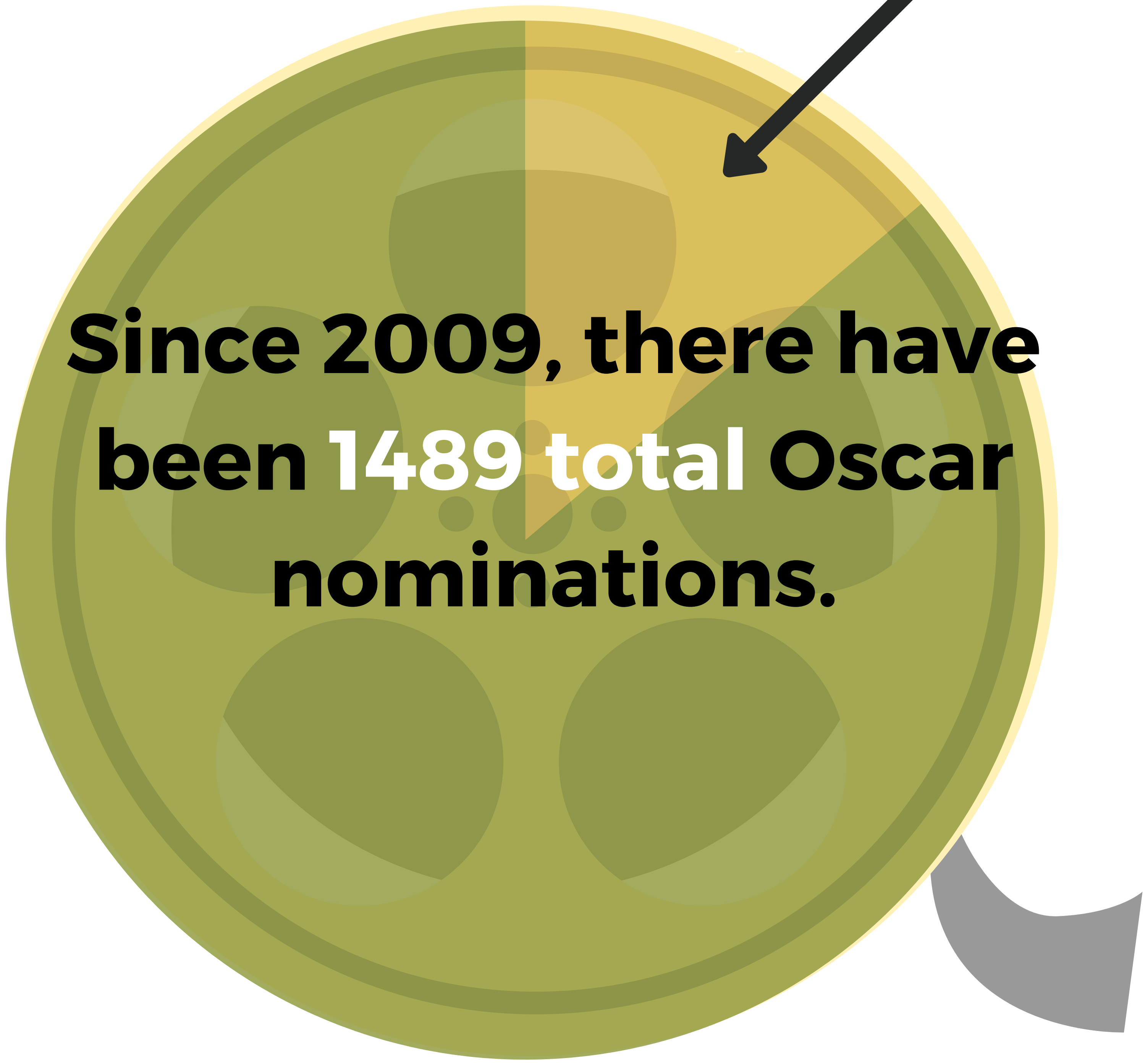
- Filter the data from the year 2009
- Join data from an Oscars data set
- Run models: Lasso, Ridge, Enet, and Logistic for classification



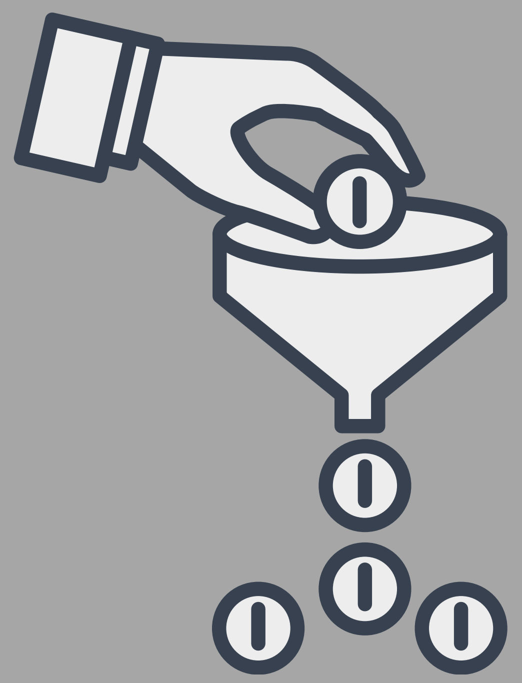
Exploratory Data

*We start from **2009** for recent movies and because the Academy expanded the amount of films nominated for awards this year.

**Keeping in mind, some films get multiple nominations/awards.



FIlters



REMOVE VOTES/REVIEWS COLUMNS

To detect nomination-worthy films, it is prudent to remove the review and vote related columns. Higher rated films would lean closer to the Oscars and skew the data. I took out budgets and box-office but left in the ratio.

JOIN THE DATA

After filtering, the IMDB and Oscars data set were joined. Every nominated film was put in. While not every movie won an award, nominated films still hold great quality. Many great films such as ***The Shawshank Redemption, Toy Story2***, etc were nominated but did not win any Oscars.

Models for Oscars

Lasso, Ridge, Elastic Net, Logistic

LASSO

```
pred_lasso = prediction(lasso_predictions_prob$`1`,
movies_oscars_test$Oscar)
performance(pred_lasso, measure = "auc")@y.values
perf <- performance(pred_lasso, measure = "tpr",
x.measure = "fpr")
```

Area Under the Curve: 0.8469001
Accuracy: 0.895

RIDGE

```
pred_ridge = prediction(ridge_predictions_prob$`1`,
movies_oscars_test$Oscar)
performance(pred_ridge, measure = "auc")@y.values
perfR <- performance(pred_ridge, measure = "tpr",
x.measure = "fpr")
```

Area Under the Curve: 0.8007106
Accuracy: 0.8922

Models for Oscars

Lasso, Ridge, Elastic Net, Logistic

ENET

```
enet_predictions_prob=predict(enet, movies_oscars_test,
type="prob")
pred_enet = prediction(enet_predictions_prob$`1`,
movies_oscars_test$Oscar)
performance(pred_enet, measure = "auc")@y.values
perfE <- performance(pred_enet, measure = "tpr",
x.measure = "fpr")
```

Area Under the Curve: 0.820497
Accuracy: 0.9031

LOGISTIC

```
predict.logistic <- predict(model.logistic, movies_oscars_test,
type="response")
predict.logistic.label = factor(ifelse(predict.logistic > .5, "Yes",
"No"))
actual.label = movies_oscars_test$Oscar
table(actual.label, predict.logistic.label)
```

Area Under the Curve: 0.7415
Accuracy: 0.8867

IV Lessons Learned

- Big Budgets mean Big Results: The correlation numbers aside, the accuracy of the machine learning models go way up when the budget numbers are in. The more money put in, the bigger the money made. However, the ratio of box-office over budget is much more difficult to predict.
- People and Time Matter: The rate of films being produced has definitely increased, especially internationally. However, directors and writers take time between movies. When following the all-time great film makers, one can see they go for quality over quantity. Great films are still outliers and it takes time to make them.
- Awards are an Inner Circle: The accuracy of predicting the Oscar awards seems highly accurate based off the fact that the industry is essentially awarding itself. The awards are made by the big studios who also market and work with the Academy to find and nominate these films.
- Movies are Art: The data does not capture everything about movies, especially their artistic meaning. Even films with the exact same genres can be wildly different. However, they are expensive. This data captures some great info but it doesn't capture the soul of a film.

Improving Business

How to improve the Movie Business

- Screen Films with Streaming: Many studios show screen films to test out the market. These are usually show in secret, but the more eyeballs in advance, the better. Two strongly correlated variables with box-office were critic and fan reviews. Instead of inviting a few people to a private theater, stream it out to a few thousand. More reviews should give better ideas of the quality of the film.
- New Metrics: While IMDB has been around since the 90's, the data it contains is more of a reference guide. I would like to see metrics on the pace of a film, scene changes, film techniques, computer graphics used, etc. There should be new types of data collected to see how these can measure the quality and style of a film.
- Machine Learning Methods: While this machine learning project worked with numerical and categorical variables, it is possible that we could use language processing with scripts and screenplays to compare with past movies. Experimentation in this area would be interesting.
- Several paths to Success: There is more than one way to succeed in movies. Big profits and awards are only a couple ways. Culture can be impacted even if the box-office is not. Fans will have more access to films way past the release date. TV, streaming, can recoup money so the goal should be to make a quality film first.

V

Appendix, Notes

- Full codes and technical details are in the R notebook and pdf of the code. The code in this presentation are only snippets for demonstration.

- There might be some **slight differences** in the R2, Accuracy numbers depending on the random settings of the packages. Some libraries/packages have their own built in random samplers/generators despite seeding, but the numbers should be close.

- Data Sets: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset
  https://www.kaggle.com/unanimad/the-oscar-award

- History of Film and Cinema references:
  https://en.wikipedia.org/wiki/History_of_film
  https://open.lib.umn.edu/mediaandculture/chapter/8-2-the-history-of-movies/

- Movie Rating Sites:
  https://www.imdb.com/
  https://www.rottentomatoes.com/
  https://www.metacritic.com/