FEATURE FILM ROI

Torin Rettig

PROJECT GOAL

Predict ROI on film budget based on data available on IMDb

MOTIVATION

- Filmmaking is an expensive, high-stakes endeavor.
- Film companies and filmmakers put a great deal of focus on financial performance.
- Identifying factors for financial success could provide valuable insights for the business.

METHODOLOGY

- Gather data on films that is available on IMDb and IMDbPro.
- Select films that have completed their box office run.
- Select features related to both financial performance and popularity.

WEB SCRAPING

SELENIUM

Get page of filtered films on IMDb Pro ¶

In [154]:
```
# Create list of movies that have budget and revenue numbers
# Open IMDbPro Log in page
driver = webdriver.Chrome('/Applications/chromedriver')
driver.get('https://pro.imdb.com/login/imdb?u=%2F')

# Enter email and password on login page
driver.find_element_by_xpath('//*[@id="ap_email"]').send_keys('starplatinum87@gmail.com')
driver.find_element_by_xpath('//*[@id="ap_password"]').send_keys('GumoDT7311#')
driver.find_element_by_xpath('//*[@id="signInSubmit"]').click()

# Go to MOVIEmeter page
# driver.get('https://pro.imdb.com/inproduction?ref_=hm_reel_mm_all#sort=ranking') # Skip clicks and go directly t
driver.find_element_by_xpath('//*[@id="meter_type_selections"]/li[2]/a').click()
driver.find_element_by_xpath('//*[@id="meter_headshots"]/div[2]/div[2]/a').click()

# Filter movies on MOVIEmeter page
driver.find_element_by_xpath('//*[@id="type_movie"]').click() # Click Movie checkbox
driver.find_element_by_xpath('//*[@id="status_RELEASED"]').click() # Click Released checkbox
driver.find_element_by_xpath('//*[@id="year_yearMin"]').send_keys('2012') # Set minimum release year to 2012
driver.find_element_by_xpath('//*[@id="year_yearMax"]').send_keys('2017') # Set max release year to 2017
driver.find_element_by_xpath('//*[@id="budget_budgetMin"]').send_keys('0.01') # Set minimum budget to $10k
driver.find_element_by_xpath('//*[@id="gross_grossMin"]').send_keys('0.01') # Set minimum US gross to $10k
driver.find_element_by_xpath('//*[@id="gross"]/ul/li[11]/span/a').click() # Click Go to filter
```

In [10]:
```
# Scroll page until all films are displayed with 3sec delays to allow page to load
for i in range(55):
    driver.execute_script("window.scrollTo(0, 250000)")
    time.sleep(3)
```

Create Master List of of filtered films

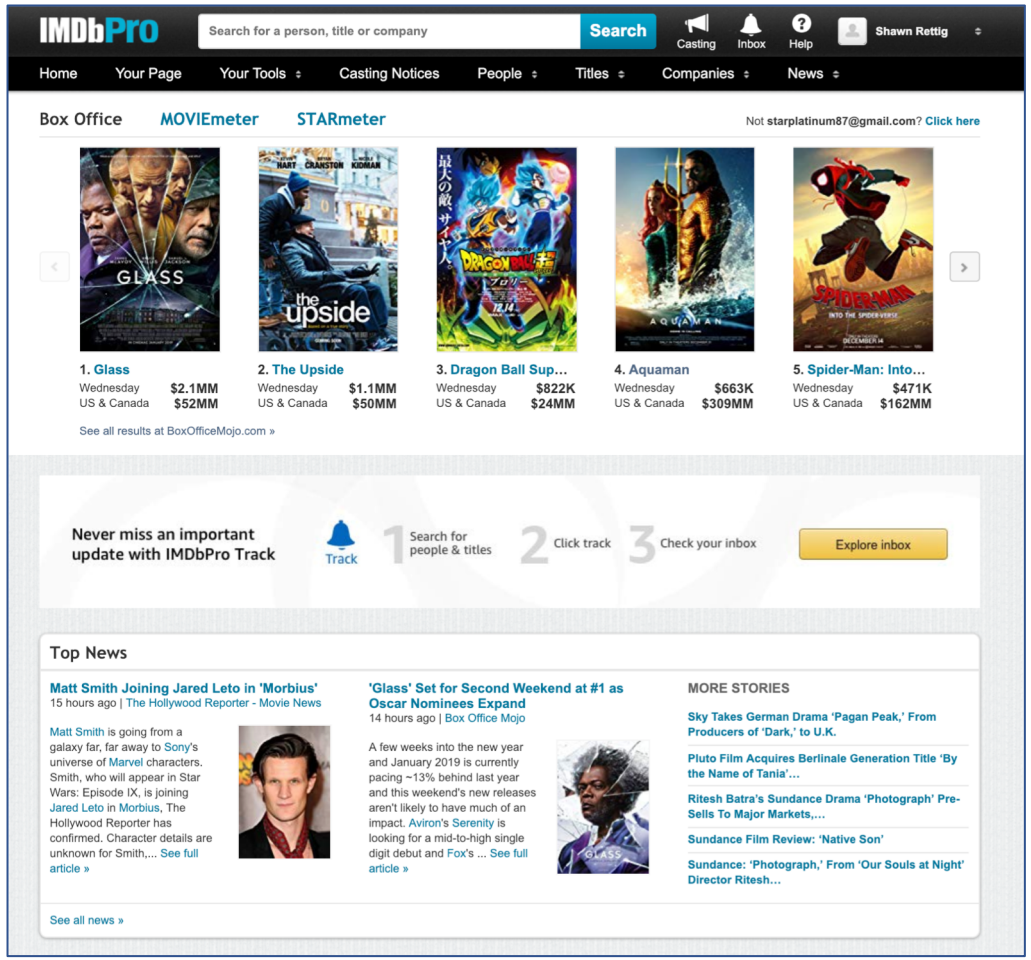
In [ ]:
```
# Grab name, URL, and titleID and put into top_films df

# Get number of results and convert to int
results = driver.find_elements_by_xpath('//*[@id="title"]/div[1]/div/span[1]/span[1]')[0]
results = results.text.replace(',', '')
results = int(results)

# Get movie name, URL and
title_list = []
movie_URL_list = []
MOVIEmeter_list = []
for i in range(1, results+1):
    title_path = '//*[@id="results"]/ul/li[' + str(i) + ']/ul/li[1]/span/a'
    get_title = driver.find_elements_by_xpath(title_path)[0]
    title = get_title.text
    title_list.append(title)

    URL = '//*[@id="results"]/ul/li[' + str(i) + ']/ul/li[1]/span/a'
    movie_URL_element = driver.find_elements_by_xpath(URL)[0]
    movie_URL = movie_URL_element.get_attribute('href')
    movie_URL_list.append(movie_URL)
```

1300+ Pages



DATA COLLECTION

Track

Add to list

Spider-Man: Into the Spider-Verse (2018)

Share | IMDb Visit on IMDb

PG | 117 min | Animation, Action, Adventure

Teen Miles Morales becomes Spider-Man of his reality, crossing his path with five counterparts from other dimensions to stop a threat for all realities.

Read more: Plot summary | Synopsis

Directors	Bob Persichetti | Peter Ramsey | Rodney Rothman
Writers	Phil Lord (screenplay by) (story by) | Rodney Rothman (screenplay by) | Meghan Malloy (story consultant)
Producers	Avi Arad (p.g.a.) | Phil Lord (p.g.a.) | Christopher Miller (p.g.a.) | Amy Pascal (p.g.a.) | Christina Steinberg (p.g.a.)
Composer	Daniel Pemberton
Editor	Robert Fisher Jr.
Casting Director	Mary Hidalgo
Production Designer	Justin Thompson (as Justin K. Thompson)

See all filmmakers & crew (694)

Trending & News

MOVIEmeter 13

News articles 1,044
4 new articles

Box Office as of 01/23/19

Budget	$90,000,000
Opening weekend	$35,363,376
Gross (US & Canada)	$161,917,707
Gross (World)	$327,014,943

See all box office data

Release date
Dec 14, 2018 (United States)

Awards

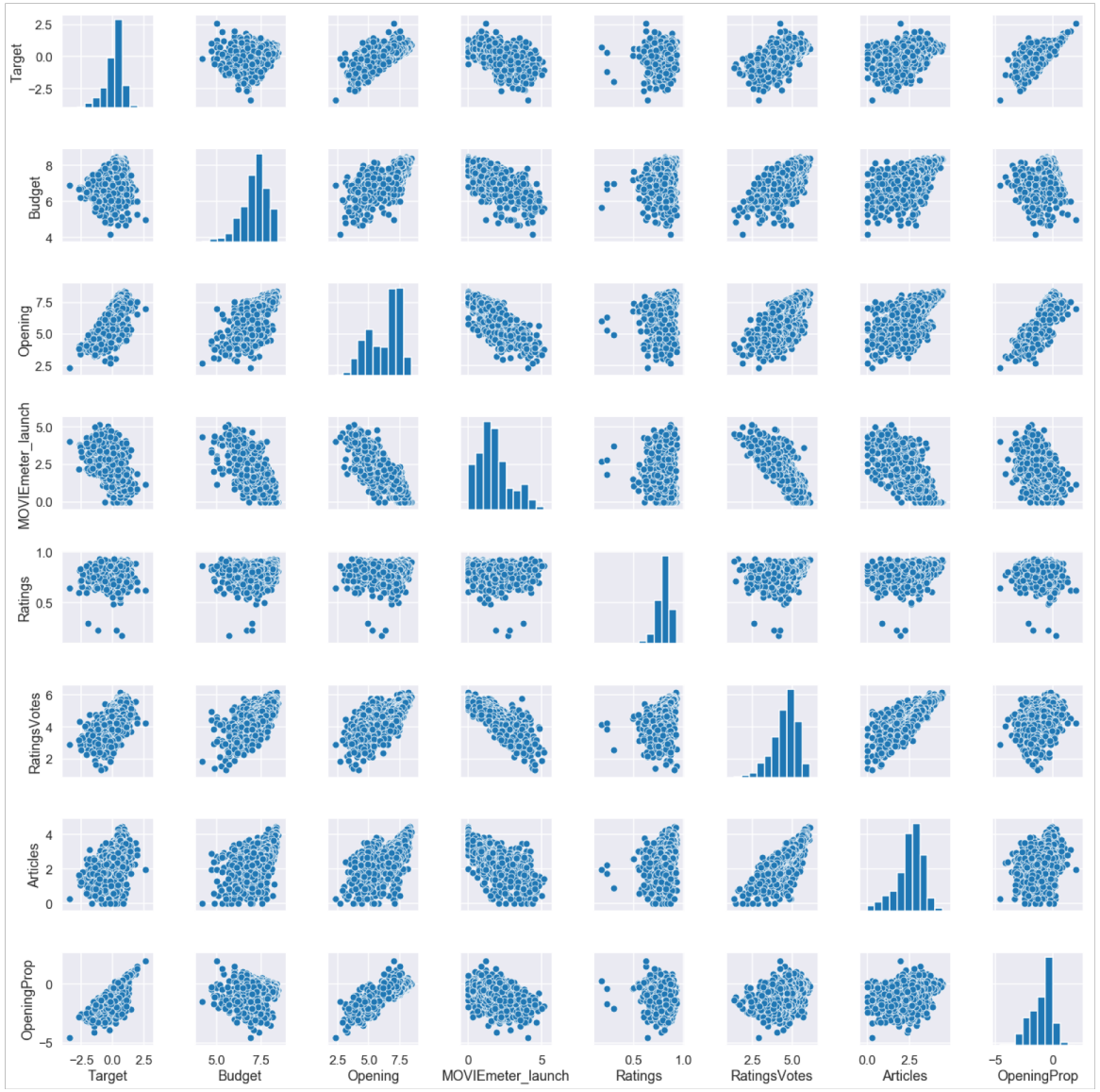
Nominated for 1 Oscar Award.
Another 35 wins & 37 nominations

ANALYSIS

- 1319 Observations: Films
- Target: Gross Revenue / Budget
- 7 Features:
  - Budget
  - Opening Weekend
  - MOVIEmeter
  - Ratings
  - Votes
  - Articles
  - Opening / Budget

ANALYSIS

Log Transform



BASIC TRAIN-VALIDATION-TEST

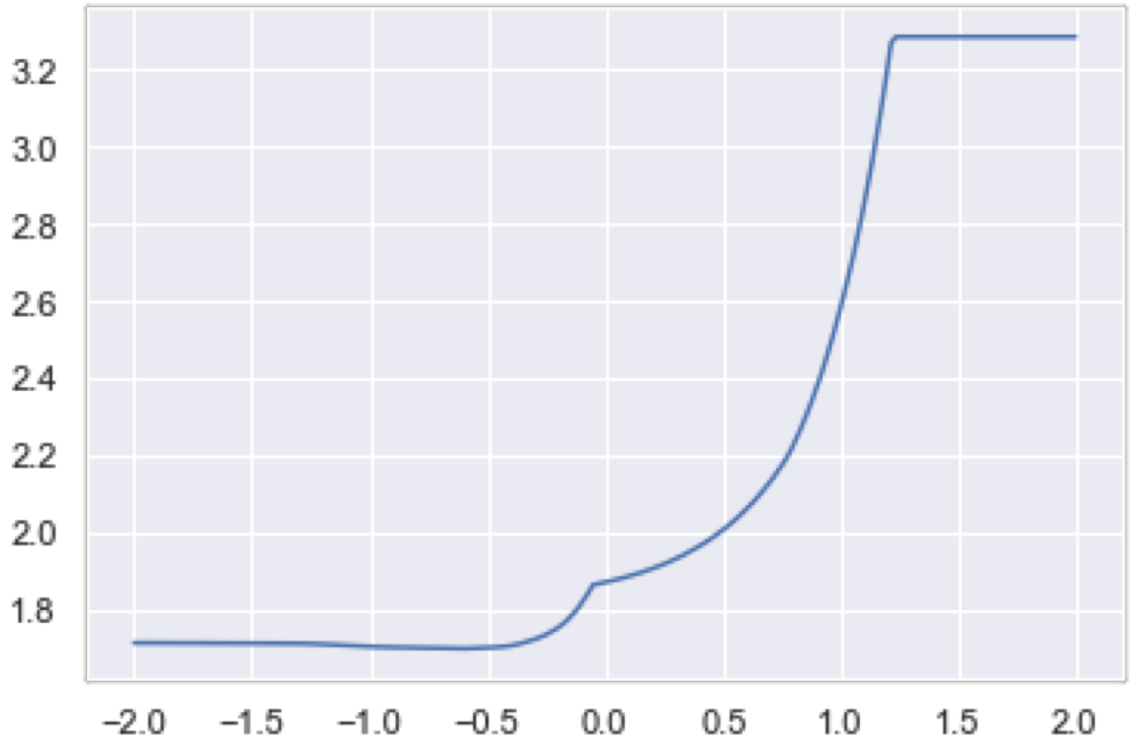
- Standard Scaled 60/20/20

- **Linear Regression: 0.686**
- **Ridge Regression: 0.686**
- **Degree 2 Polynomial: 0.235**

- Keep these values in mind for later

REGULARIZATION – TUNE ALPHA

Validation Error Plot



Alpha Value = **0.23272**

REGULARIZATION - LASSO SCALED

LASSO

60/20/20

- Alpha = **0.2327**

- Budget: **-0.0**
- Opening: **-0.0**
- MOVIEmeter launch: **-0.00**
- Ratings: **0.5671**
- RatingsVotes: **0.4036**
- Articles: **0.0**
- OpeningProp: **16.4261**

REGULARIZATION – AUTOMATED

LASSO CV

Keep features!

- Alpha = **0.01**

- Budget: **-0.0116**
- Opening: **-0.2659**
- MOVIEmeter launch: **-0.1151**
- Ratings: **0.7239**
- RatingsVotes: **0.8271**
- Articles: **-0.1498**
- OpeningProp: **16.6904**

TEST RESULTS

- Basic T/V/T Linear Regression: **0.580**
- Ridge CV Regression: **0.6961**
- LASSO CV Regression: **0.6952**
- Ridge is slightly better, but they are both nearly identical.

CONCLUSIONS

- Certain features seem more strongly predictive than others.

- All features together seem to have cumulative predictive effect that is strongest.

- Both LASSO and Ridge CV were nearly identical, but LASSO is thought to be better suited for models with fewer features, so we'll go with LASSO.

THANK YOU!

Q & A

QUESTIONS

APPENDIX

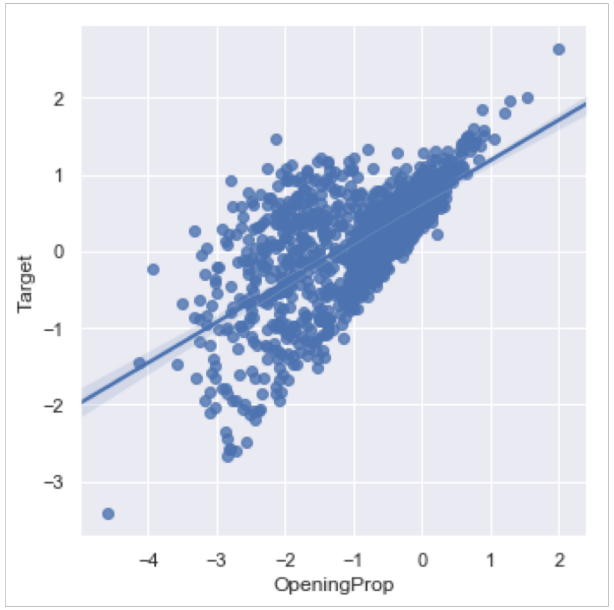
FUTURE WORK

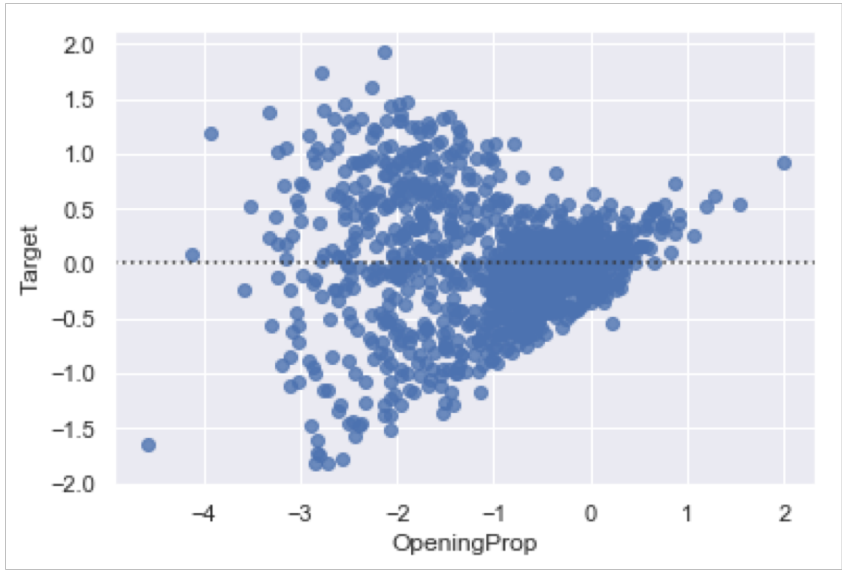
- More Features: Gross +1 Month, Principal Cast Details, Metacritic, Season, Month, Holiday, Genre, etc.

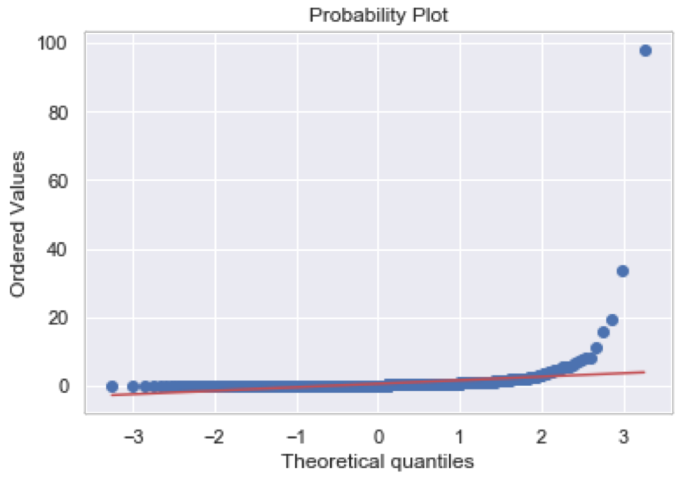
- Add trends information on revenue and other details.

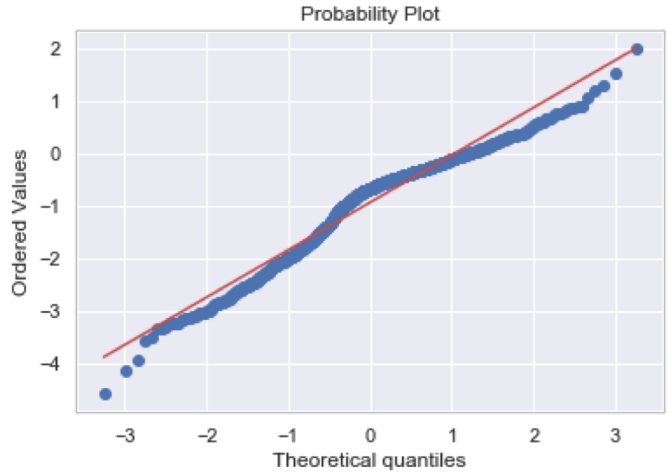
ANALYSIS

Opening | Target





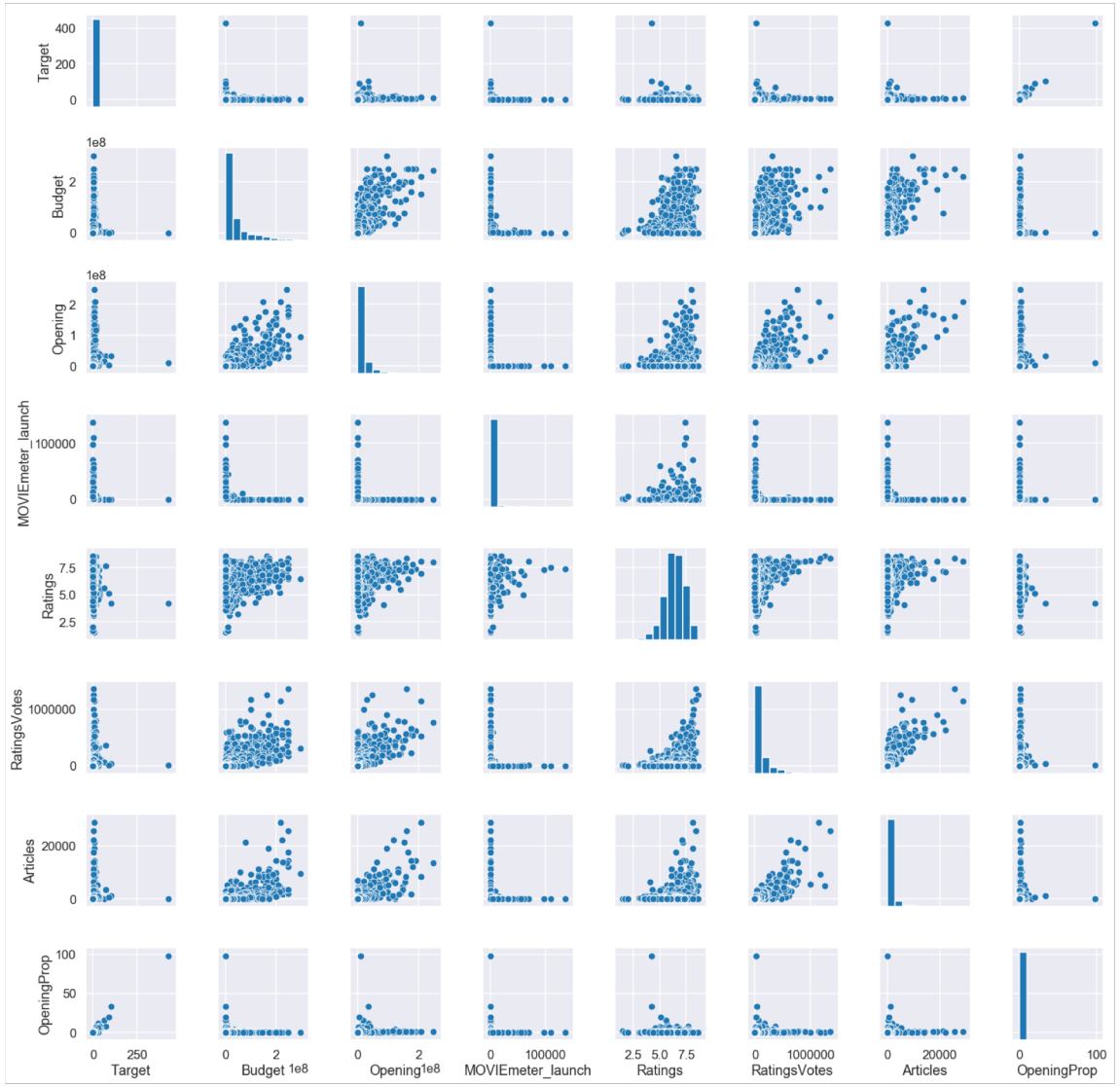




ANALYSIS

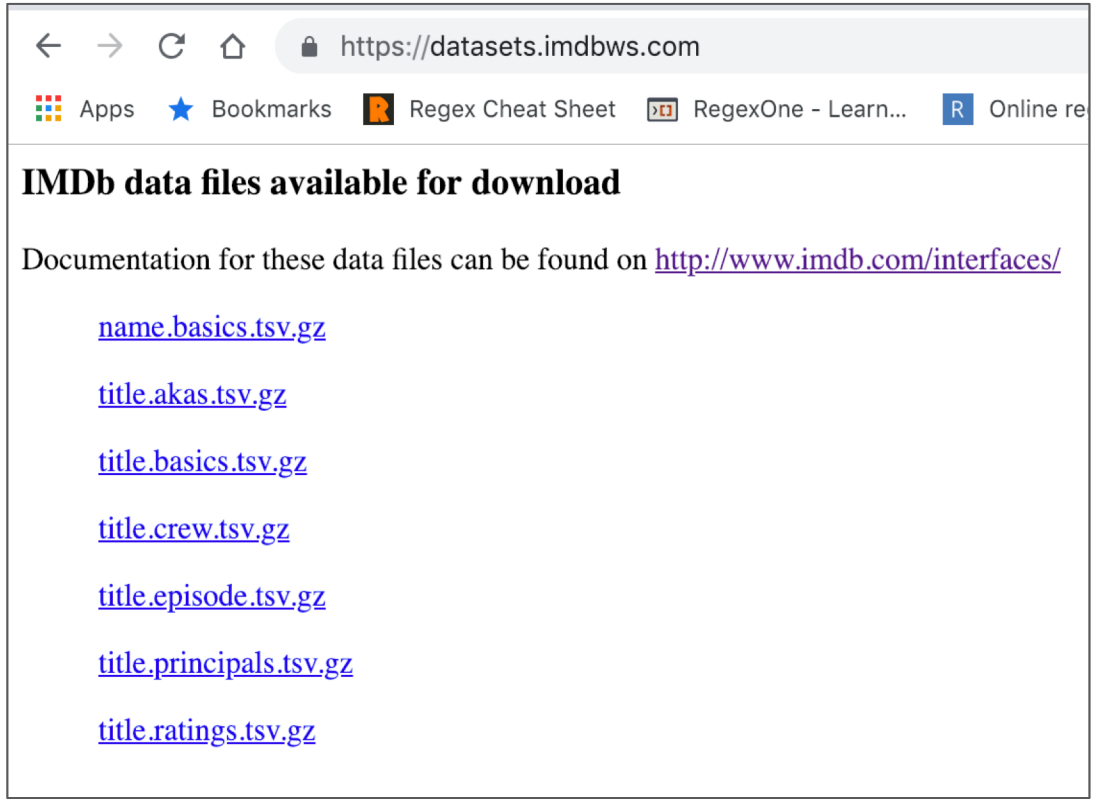
Data at Different Scales

	Target	Budget	Opening	MOVIEmeter_launch	Ratings	RatingsVotes	Articles	OpeningProp
0	1.353692	37500000.0	20355000.0	27.0	5.8	6400.0	246.0	0.542800
1	30.939373	9000000.0	40010975.0	5.0	7.3	318839.0	1197.0	4.445664
2	5.178490	84000000.0	8805843.0	8.0	7.7	184604.0	1044.0	0.104831
3	5.029525	175000000.0	117027503.0	1.0	7.5	386135.0	6349.0	0.668729
4	10.064096	19400000.0	166564.0	12.0	7.4	283003.0	2758.0	0.008586
5	4.438083	20000000.0	14415922.0	7.0	7.3	414261.0	2182.0	0.720796



DATA COLLECTION

IMDb TSVs

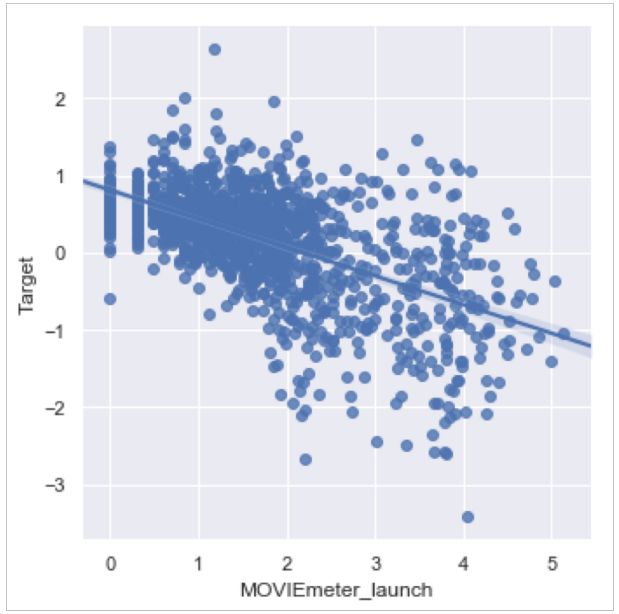


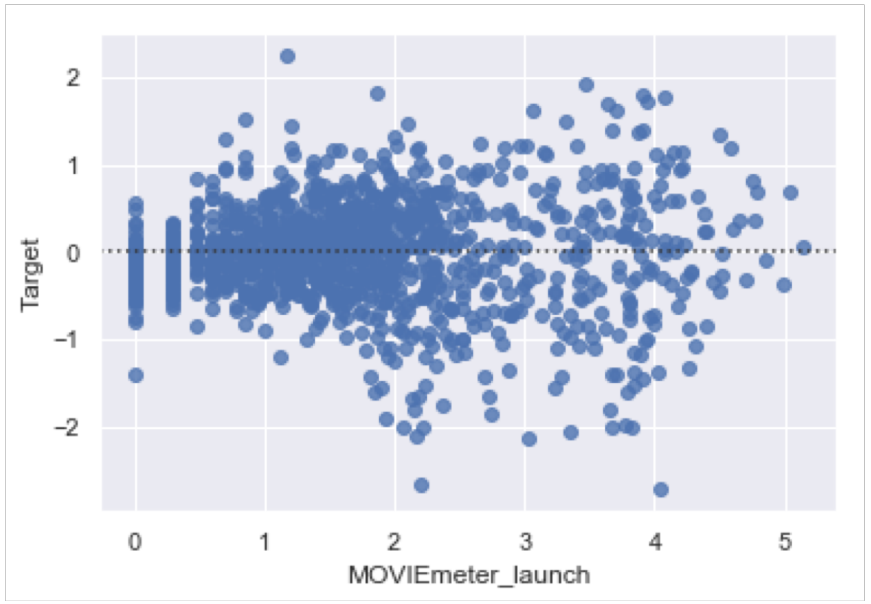
DataFrame

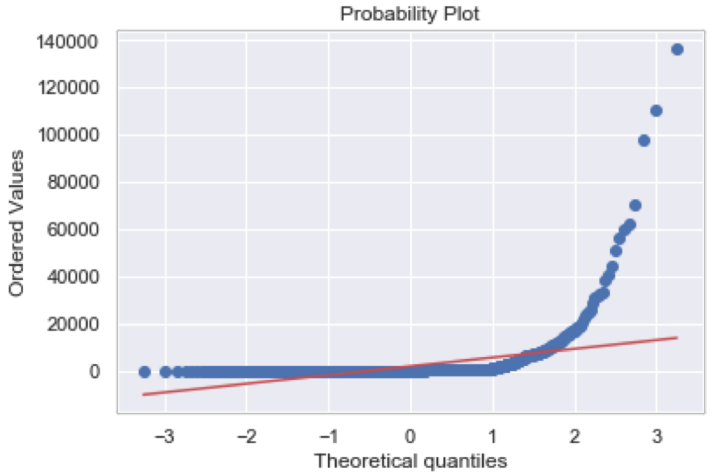
	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
0	tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary,Short
1	tt0000002	short	Le clown et ses chiens	Le clown et ses chiens	0	1892	\N	5	Animation,Short
2	tt0000003	short	Pauvre Pierrot	Pauvre Pierrot	0	1892	\N	4	Animation,Comedy,Romance
3	tt0000004	short	Un bon bock	Un bon bock	0	1892	\N	\N	Animation,Short
4	tt0000005	short	Blacksmith Scene	Blacksmith Scene	0	1893	\N	1	Comedy,Short
5	tt0000006	short	Chinese Opium Den	Chinese Opium Den	0	1894	\N	1	Short
6	tt0000007	short	Corbett and Courtney Before the Kinetograph	Corbett and Courtney Before the Kinetograph	0	1894	\N	1	Short,Sport
7	tt0000008	short	Edison Kinetoscopic Record of a Sneeze	Edison Kinetoscopic Record of a Sneeze	0	1894	\N	1	Documentary,Short
8	tt0000009	movie	Miss Jerry	Miss Jerry	0	1894	\N	45	Romance
9	tt0000010	short	Exiting the Factory	La sortie de l'usine Lumière à Lyon	0	1895	\N	1	Documentary,Short
10	tt0000011	short	Akrobatisches Potpourri	Akrobatisches Potpourri	0	1895	\N	1	Documentary,Short
11	tt0000012	short	The Arrival of a Train	L'arrivée d'un train à La Ciotat	0	1896	\N	1	Documentary,Short
12	tt0000013	short	The Photographical Congress Arrives in Lyon	Neuville-sur-Saône: Débarquement du congrès de...	0	1895	\N	1	Documentary,Short
13	tt0000014	short	The Sprinkler Sprinkled	L'arroseur arrosé	0	1895	\N	1	Comedy,Short
14	tt0000015	short	Autour d'une cabine	Autour d'une cabine	0	1894	\N	2	Animation,Short
15	tt0000016	short	Barque sortant du port	Barque sortant du port	0	1895	\N	1	Documentary,Short
16	tt0000017	short	Italienischer Bauerntanz	Italienischer Bauerntanz	0	1895	\N	1	Documentary,Short
17	tt0000018	short	Das boxende Känguruh	Das boxende Känguruh	0	1895	\N	1	Short
18	tt0000019	short	The Clown Barber	The Clown Barber	0	1898	\N	\N	Comedy,Short
19	tt0000020	short	The Derby 1895	The Derby 1895	0	1895	\N	1	Documentary,Short,Sport
20	tt0000022	short	Blacksmith Scene	Les forgerons	0	1895	\N	1	Documentary,Short
21	tt0000023	short	The Sea	Baignade en mer	0	1895	\N	1	Documentary,Short
22	tt0000024	short	Opening of the Kiel Canal	Opening of the Kiel Canal	0	1895	\N	\N	News,Short

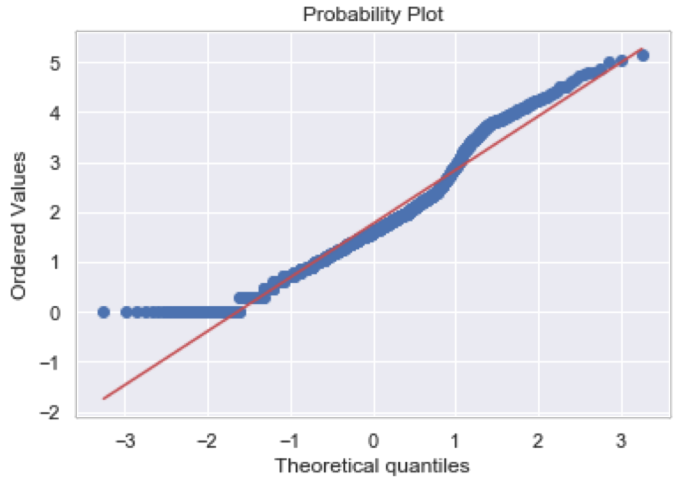
ANALYSIS

MOVIEmeter | Target









REGULARIZATION - LASSO SCALED

LASSO

60/20/20

- Alpha = **0.2** (Arbitrary)

- Budget: **-0.0**
- Opening: **-0.0**
- MOVIEmeter launch: **-0.00**
- Ratings: **0.5956**
- RatingsVotes: **0.4235**
- Articles: **0.0**
- OpeningProp: **16.4626**

REGULARIZATION - LASSO

LASSO

60/20/20

- Alpha = **0.2**
- R^2 Score: **0.6984~**

- Budget: **-0.00**
- Opening: **-0.00**
- MOVIEmeter launch: **-0.00**
- Ratings: **0.72**
- RatingsVotes: **-0.00**
- Articles: **-0.00**
- OpeningProp: **4.22**

Feature Engineering – Interactions

Ratings * Votes

Split & Validate

- R^2 Score: **0.6984~**

- Opening: **-0.00**
- MOVIEmeter launch: **-0.00**
- Ratings: **0.77**
- RatingsVotes: **-0.00**
- Articles: **-0.00**
- OpeningProp: **4.22**
- Ratings_x_Votes: **-0.00**

REGULARIZATION – AUTOMATED

RIDGE CV

- Alpha = **0.01**

- Budget: -0.0246
- Opening: -0.272
- MOVIEmeter launch: -0.1252
- Ratings: 0.7287
- RatingsVotes: 0.8713
- Articles: -0.1844
- OpeningProp: 16.6902