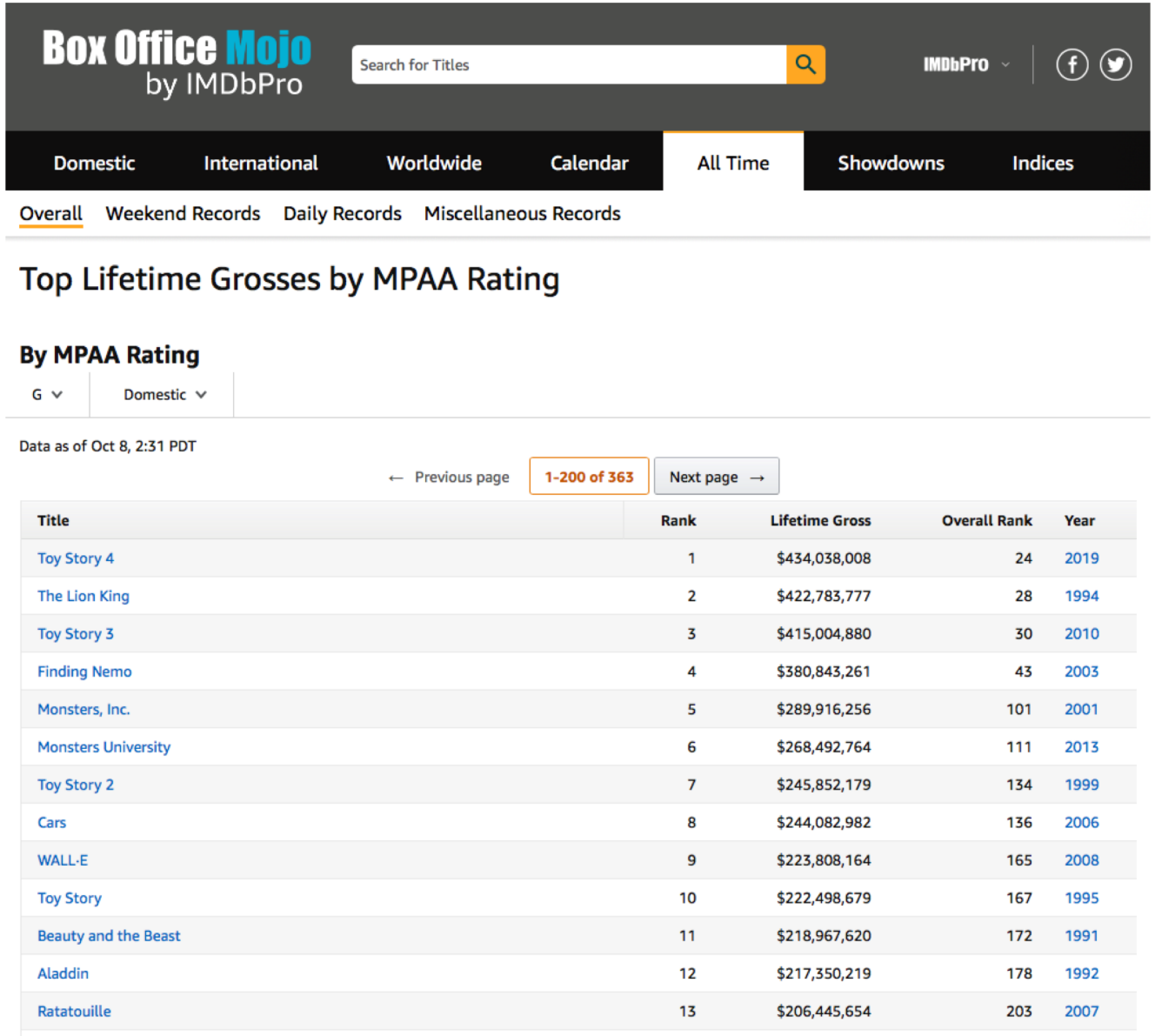
Modeling Domestic Earnings for Movies with Linear Regression

Jason Pizzollo

October 9, 2020

Background



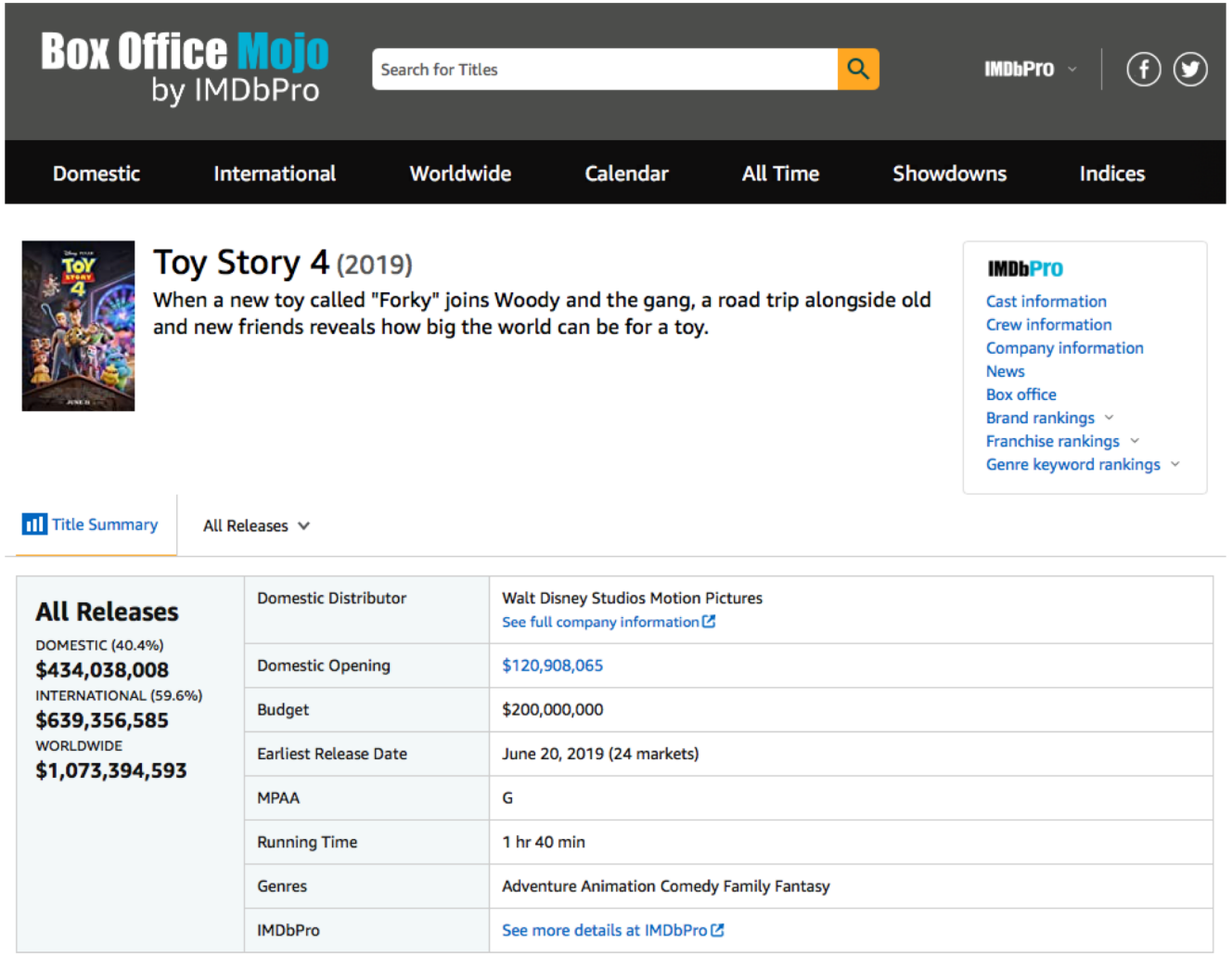
Title	Rank	Lifetime Gross	Overall Rank	Year
Toy Story 4	1	$434,038,008	24	2019
The Lion King	2	$422,783,777	28	1994
Toy Story 3	3	$415,004,880	30	2010
Finding Nemo	4	$380,843,261	43	2003
Monsters, Inc.	5	$289,916,256	101	2001
Monsters University	6	$268,492,764	111	2013
Toy Story 2	7	$245,852,179	134	1999
Cars	8	$244,082,982	136	2006
WALL·E	9	$223,808,164	165	2008
Toy Story	10	$222,498,679	167	1995
Beauty and the Beast	11	$218,967,620	172	1991
Aladdin	12	$217,350,219	178	1992
Ratatouille	13	$206,445,654	203	2007

Data collection

BoxOfficeMojo.com

- Thousands of movies
- Lifetime earnings ~ MPAA rating

Background



Data collection

BoxOfficeMojo.com

- Summary features
  - Total Domestic Gross
  - Domestic distributor (Studio)
  - Opening weekend earnings
  - Budget
  - Earliest release date
  - MPAA rating
  - Runtime
  - Genre

Data format and cleaning

	Studio	Opening	Budget	Release	Rating	Runtime	Genre	Domestic
0	Walt Disney Studios Motion Pictures	$120,908,065	$200,000,000	June 20, 2019	G	1 hr 40 min	[Adventure, Animation, Comedy, Family, Fantasy]	$434,038,008
1	Walt Disney Studios Motion Pictures	$1,586,753	$45,000,000	June 15, 1994	G	1 hr 28 min	[Adventure, Animation, Drama, Family, Musical]	$422,783,777
2	Walt Disney Studios Motion Pictures	$110,307,189	$200,000,000	June 16, 2010	na	1 hr 43 min	[Adventure, Animation, Comedy, Family, Fantasy]	$415,004,880
3	Walt Disney Studios Motion Pictures	$70,251,710	$94,000,000	May 30, 2003	na	1 hr 40 min	[Adventure, Animation, Comedy, Family]	$380,843,261
4	Walt Disney Studios Motion Pictures	$62,577,067	$115,000,000	November 2, 2001	G	1 hr 32 min	[Adventure, Animation, Comedy, Family, Fantasy]	$289,916,256
...	...	...	...	...	...	...	...	...
3358	Paramount Pictures	na	na	September 24, 1975	na	1 hr 57 min	[Mystery, Thriller]	$27,476,252
3359	New Line Cinema	$6,589,341	$3,500,000	April 28, 1995	R	1 hr 31 min	[Comedy, Drama]	$27,467,564
3360	Screen Gems	$10,302,846	$6,000,000	March 23, 2001	R	1 hr 46 min	[Comedy, Drama]	$27,457,409
3361	A24	$6,560,030	$9,000,000	July 3, 2019	R	2 hr 28 min	[Drama, Horror, Mystery, Thriller]	$27,426,361
3362	Focus Features	$260,865	$5,000,000	November 1, 2013	R	1 hr 57 min	[Biography, Drama]	$27,298,285

3363 rows × 8 columns

Numeric Features

- Opening
- Budget
- Domestic
- Release Year

Categorical Features (dummy variables)

- Studio
- Release Month
- Rating
- Genre

Drop rows “na” for Opening and Budget

- 1935 rows remaining

Feature engineering

	Studio	Opening	Budget	Release	Rating	Runtime	Genre	Domestic
0	Walt Disney Studios Motion Pictures	$120,908,065	$200,000,000	June 20, 2019	G	1 hr 40 min	[Adventure, Animation, Comedy, Family, Fantasy]	$434,038,008
1	Walt Disney Studios Motion Pictures	$1,586,753	$45,000,000	June 15, 1994	G	1 hr 28 min	[Adventure, Animation, Drama, Family, Musical]	$422,783,777
2	Walt Disney Studios Motion Pictures	$110,307,189	$200,000,000	June 16, 2010	na	1 hr 43 min	[Adventure, Animation, Comedy, Family, Fantasy]	$415,004,880
3	Walt Disney Studios Motion Pictures	$70,251,710	$94,000,000	May 30, 2003	na	1 hr 40 min	[Adventure, Animation, Comedy, Family]	$380,843,261
4	Walt Disney Studios Motion Pictures	$62,577,067	$115,000,000	November 2, 2001	G	1 hr 32 min	[Adventure, Animation, Comedy, Family, Fantasy]	$289,916,256
...	...	...	...	...	...	...	...	...
3358	Paramount Pictures	na	na	September 24, 1975	na	1 hr 57 min	[Mystery, Thriller]	$27,476,252
3359	New Line Cinema	$6,589,341	$3,500,000	April 28, 1995	R	1 hr 31 min	[Comedy, Drama]	$27,467,564
3360	Screen Gems	$10,302,846	$6,000,000	March 23, 2001	R	1 hr 46 min	[Comedy, Drama]	$27,457,409
3361	A24	$6,560,030	$9,000,000	July 3, 2019	R	2 hr 28 min	[Drama, Horror, Mystery, Thriller]	$27,426,361
3362	Focus Features	$260,865	$5,000,000	November 1, 2013	R	1 hr 57 min	[Biography, Drama]	$27,298,285

3363 rows × 8 columns

Studio (n = 61)

- 17 studios account for 90% of movies
- Replace remainder with “Other”

Impute missing ratings

- Movies sorted by rating

Bin Runtime

- Short: < 80 min
- Medium: 80 – 110 min
- Long: 110 – 140 min
- Xlong: > 140 min

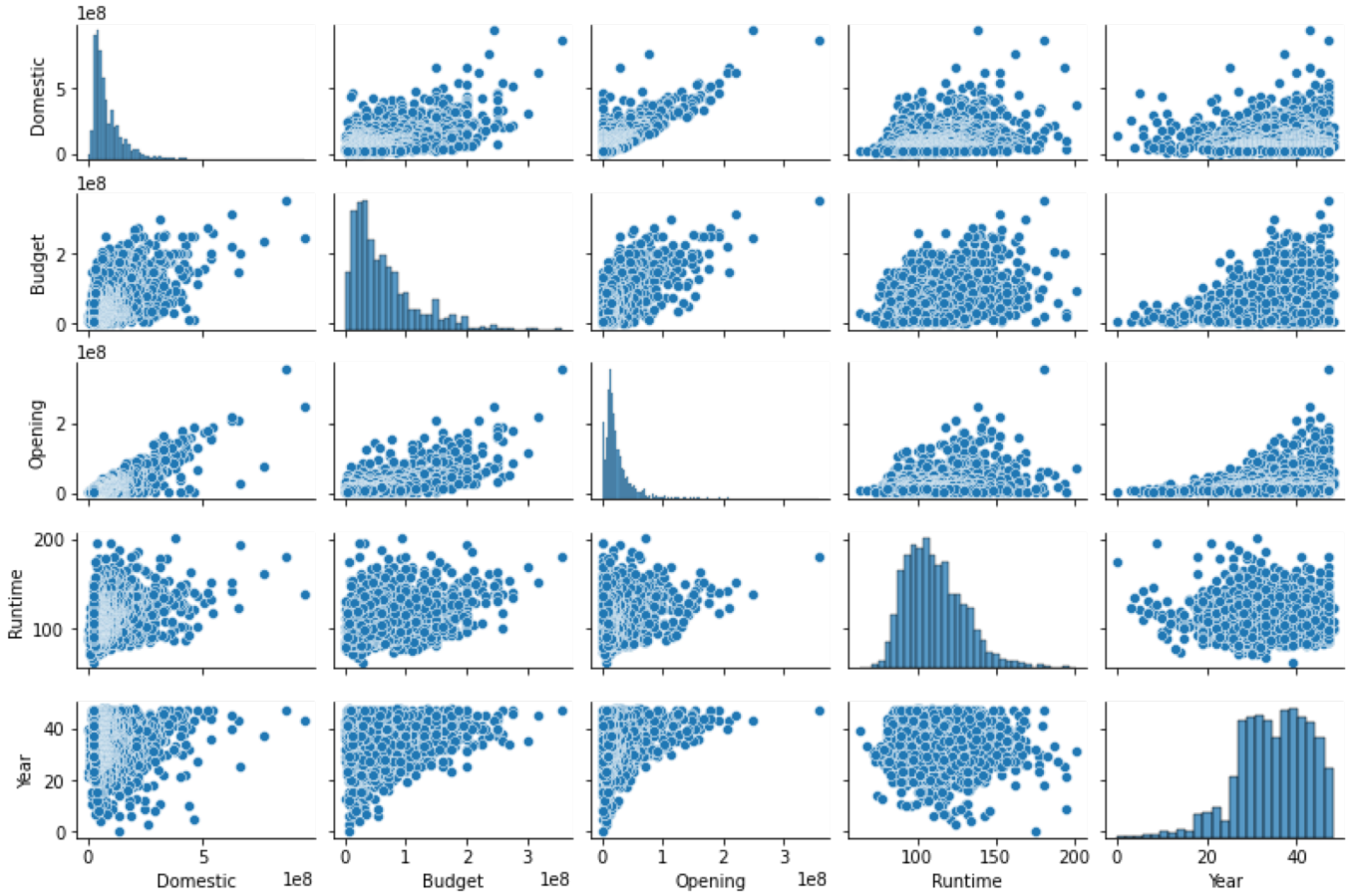
Parse Release Date

- Numeric “Year” = ReleaseYear – min(ReleaseYear)
- Categorical “Month”

Create Dummy Variables

- Total Features = 62

Exploring (domestic ~ feature) relationships

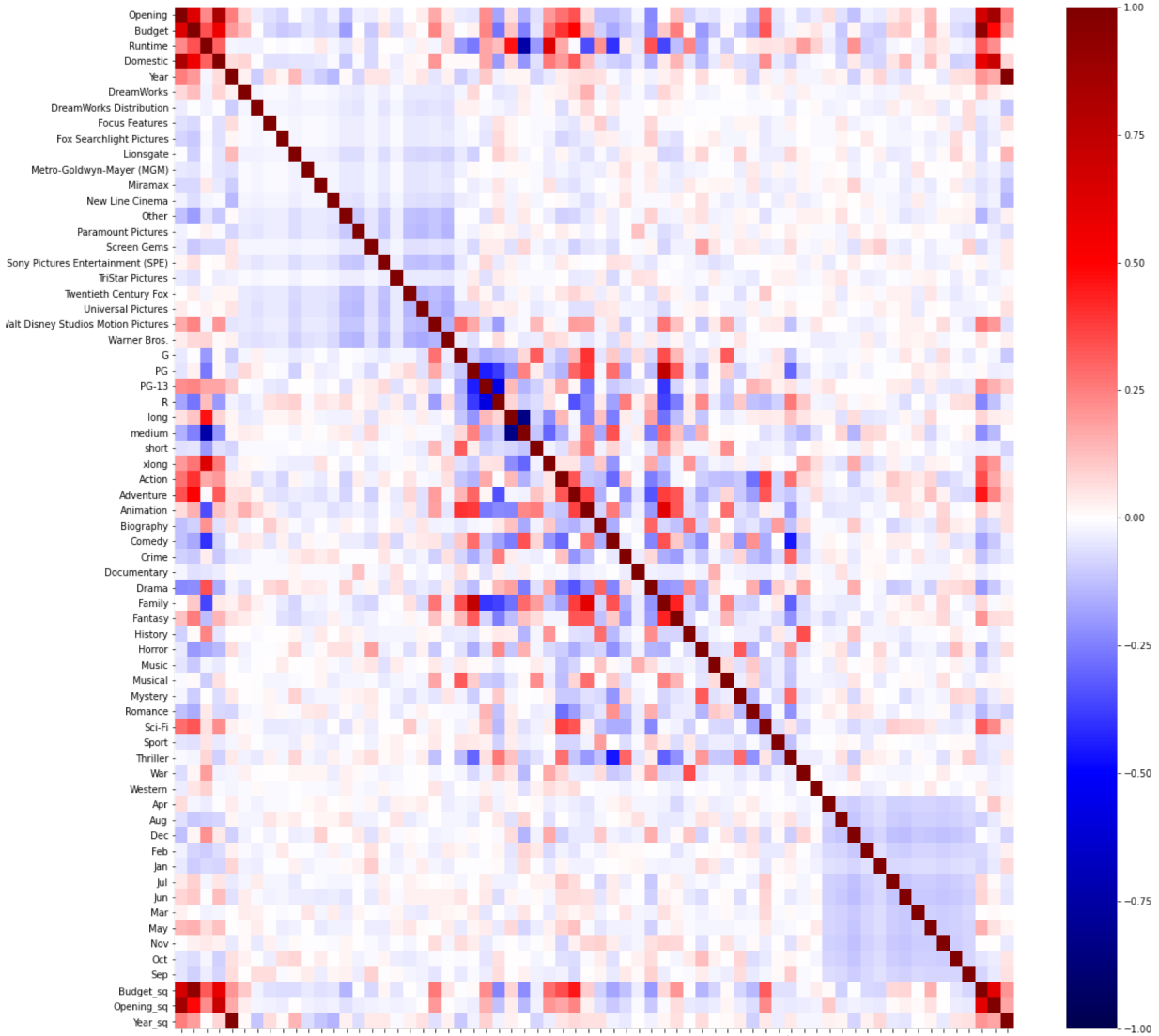


No obvious non-linear relationship

Maybe… try squaring Budget, Opening, Year

- Try regression with and without squared terms

Exploring (domestic: feature) correlations



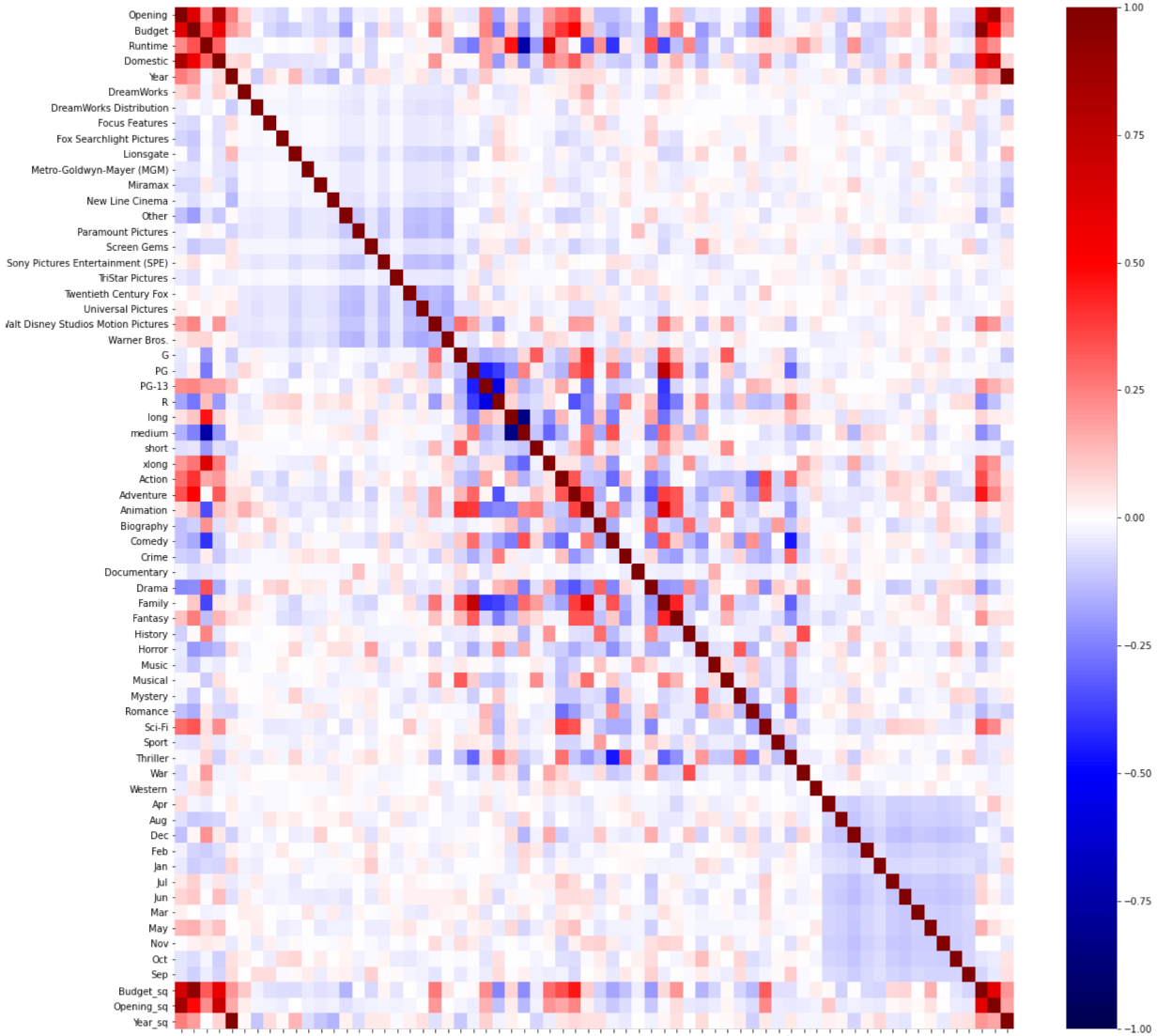
Giant correlation plots provide minimal information

- In some places correlations are strong
- In some places correlations are weak

	Opening	Budget	Runtime	Domestic	Year
Opening	1.000000	0.640224	0.229619	0.831127	0.253289
Budget	0.640224	1.000000	0.337755	0.568783	0.200211
Runtime	0.229619	0.337755	1.000000	0.305928	-0.018023
Domestic	0.831127	0.568783	0.305928	1.000000	0.063880
Year	0.253289	0.200211	-0.018023	0.063880	1.000000

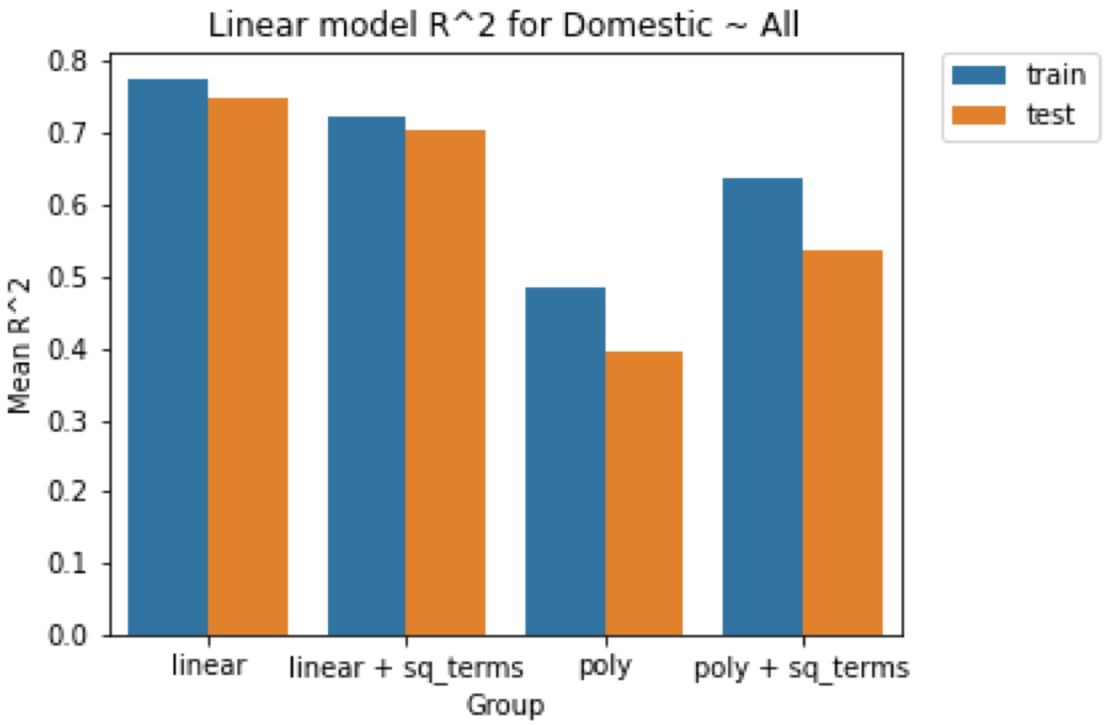
- Max correlation between features = 0.64

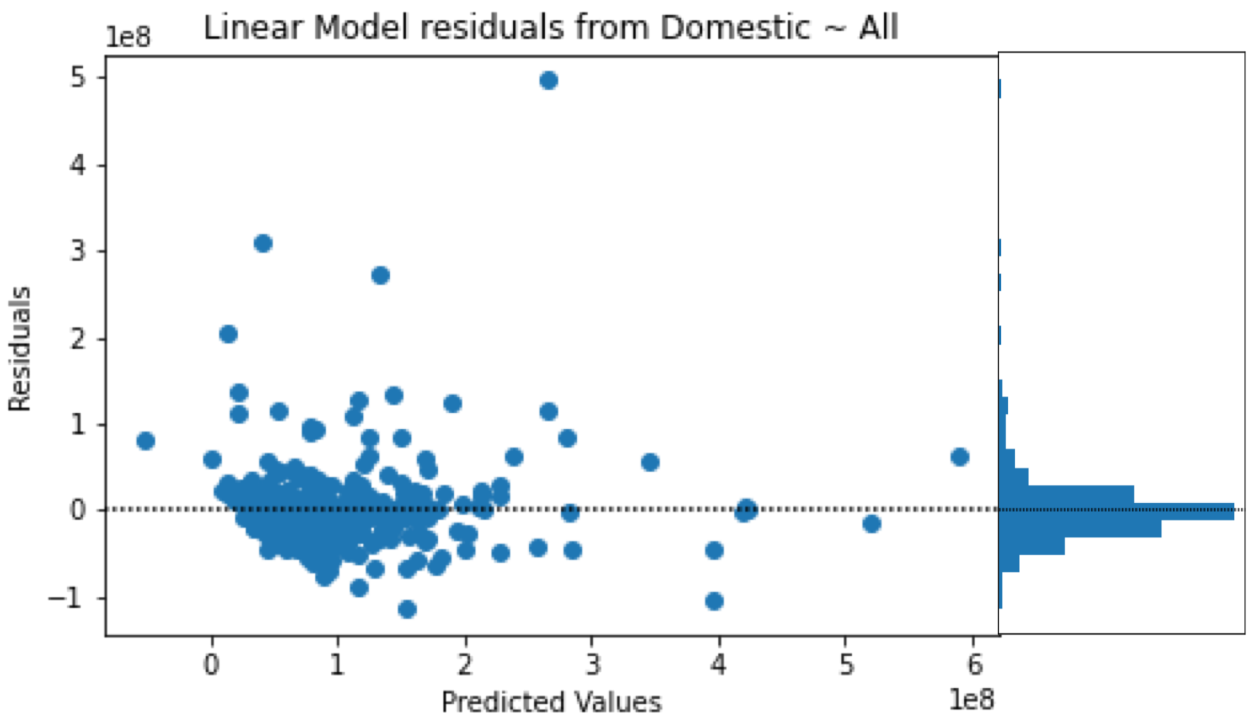
Exploring (domestic: feature) correlations



Feature	Correlation
Opening	0.831
Opening_sq	0.716
Budget_sq	0.592
Budget	0.569
Adventure	0.323
Runtime	0.306
xlong	0.258
Sci-Fi	0.225
medium	-0.207
Disney	0.205
Action	0.201
R	-0.198
PG-13	0.176
Drama	-0.159
Fantasy	0.146
Horror	-0.133
Jun	0.127
May	0.121
Animation	0.119
Other	-0.107
Crime	-0.107

Basic linear model with all features



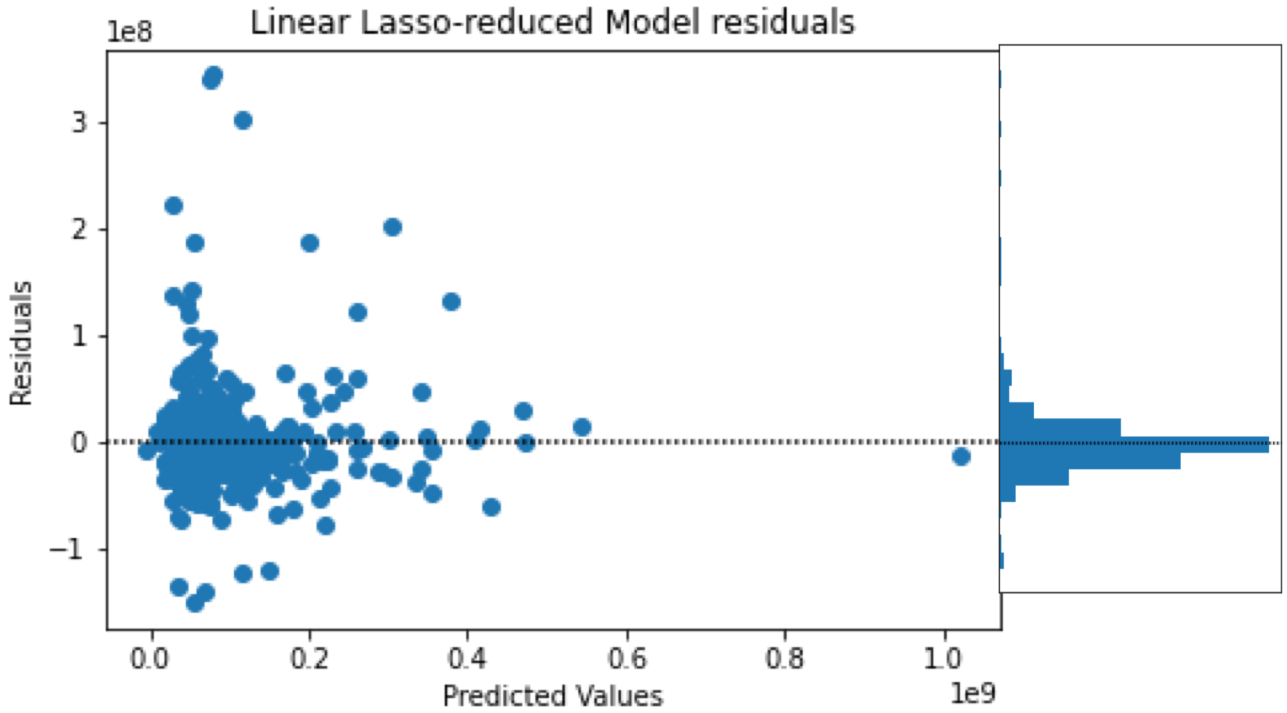


- K=5 KFold R^2 measures on train/ test
- Polynomial features for interaction terms only
- Best R^2 (0.747) from linear model with no squared or interaction terms

- Residuals from first linear model
- Distribution skews positive, but mostly normal

LassoCV regularization

- Linear model with 62 features — R^2 = 0.747 (mean k-fold R^2 on test set)
- Lasso model with 62 features — R^2 = 0.685 (n = 1 test set)
- Linear model with 41 (lasso-reduced) features — R^2 = 0.753 (mean k-fold R^2 on test set)



Conclusions

- Basic linear model performs very well
  - 62 features
  - R^2 = 0.747
- Linear model with 41 (lasso-reduced) features performs marginally better
  - R^2 = 0.753
- Addition of polynomial and interaction terms reduces R^2

Next Steps

- Scale and identify primary predictors of total domestic gross
- Impute missing opening/ budget values
- Lasso polynomial/ interaction terms
- K-fold residuals
- Additional data:
  - Director, Actors, Oscar nominations

Questions?