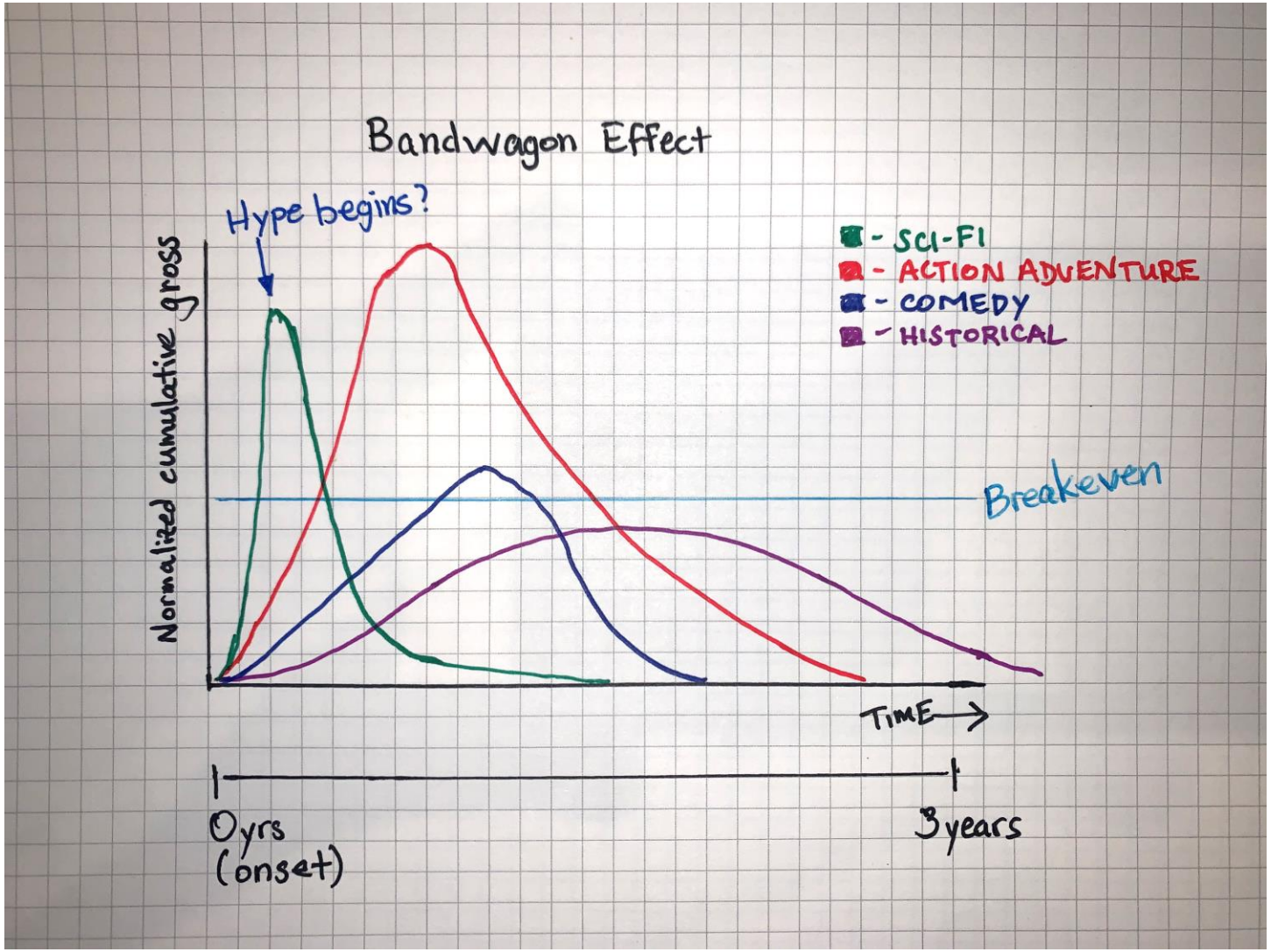
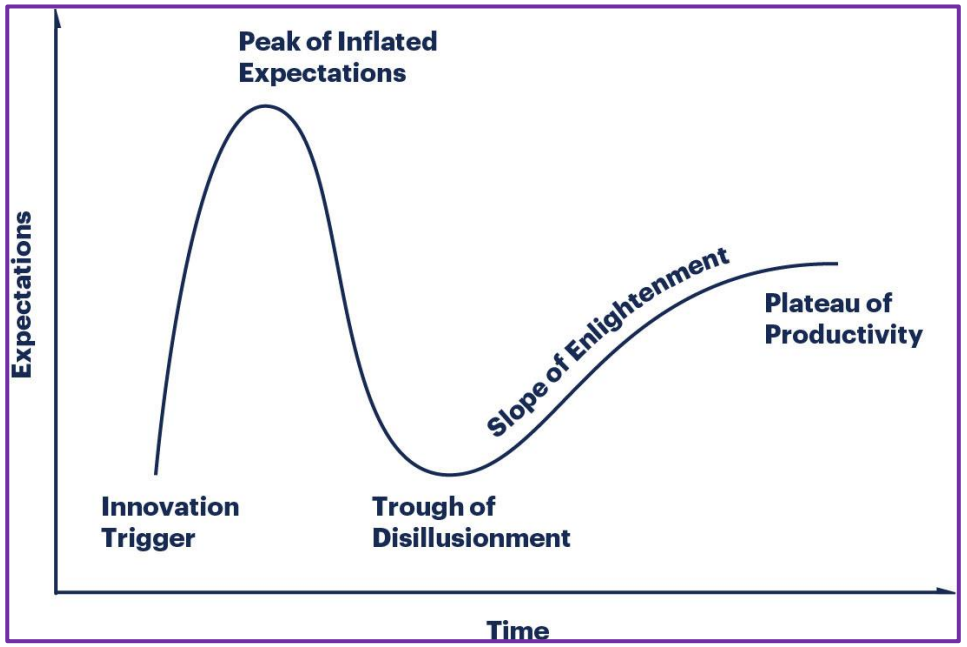
WHEN WILL IT ALL END?!

MOVIE STUDIO INVESTMENTS: PREDICTING GENRE-BASED HYPE CURVE DURATION

WHAT MIGHT A MOVIE ‘HYPE’ LOOK LIKE?

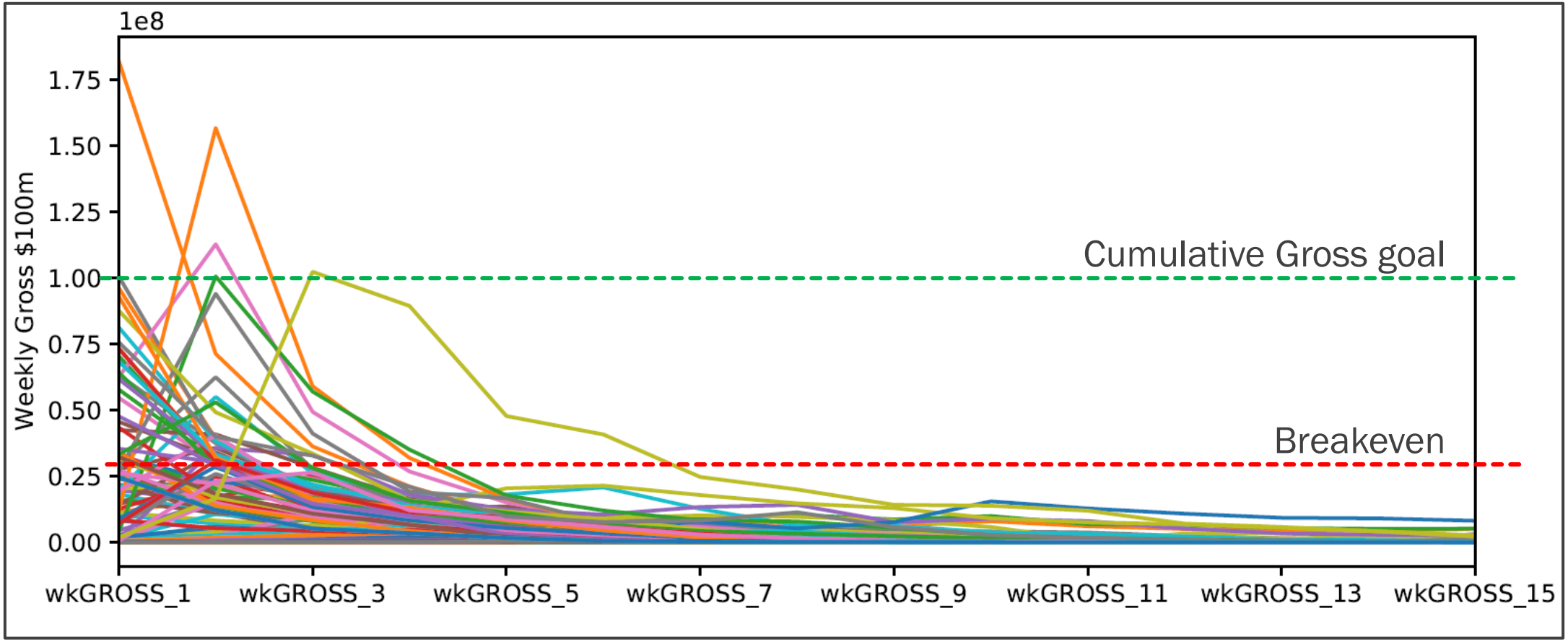




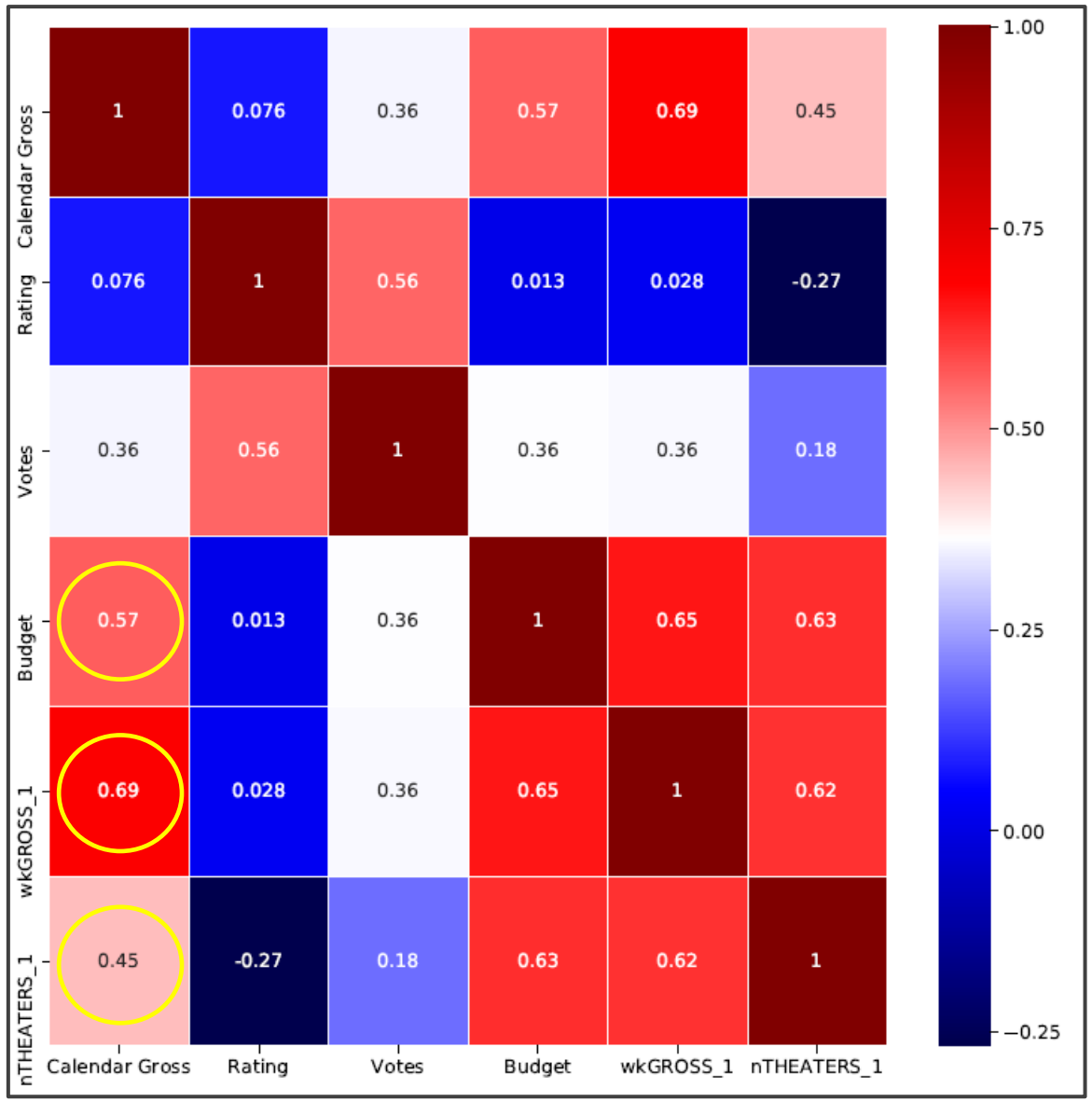
https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

HOW LONG UNTIL WE'RE PROFITABLE?

- By extension: What genres (over X years) do we invest in to increase **likelihood of sustained profit**?



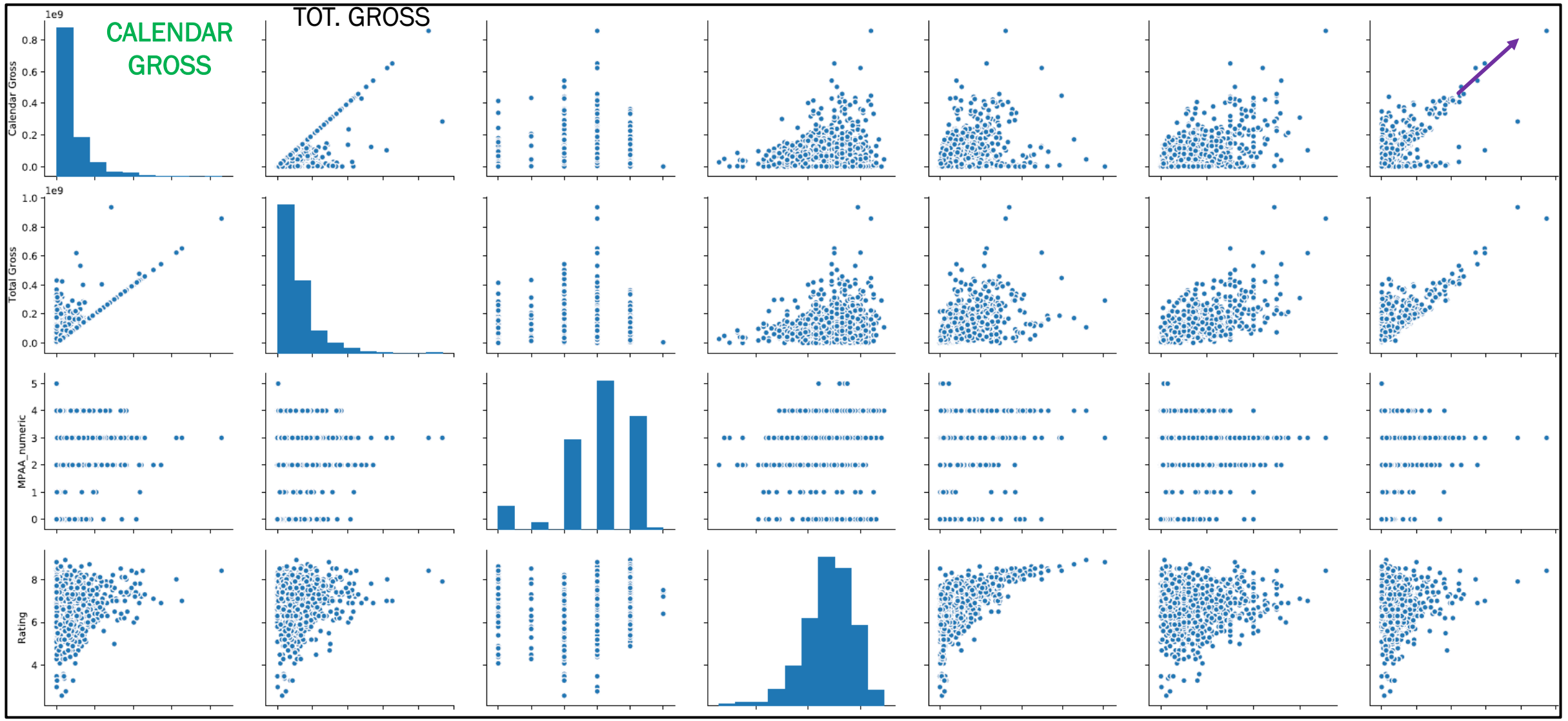
CORRELATIONS



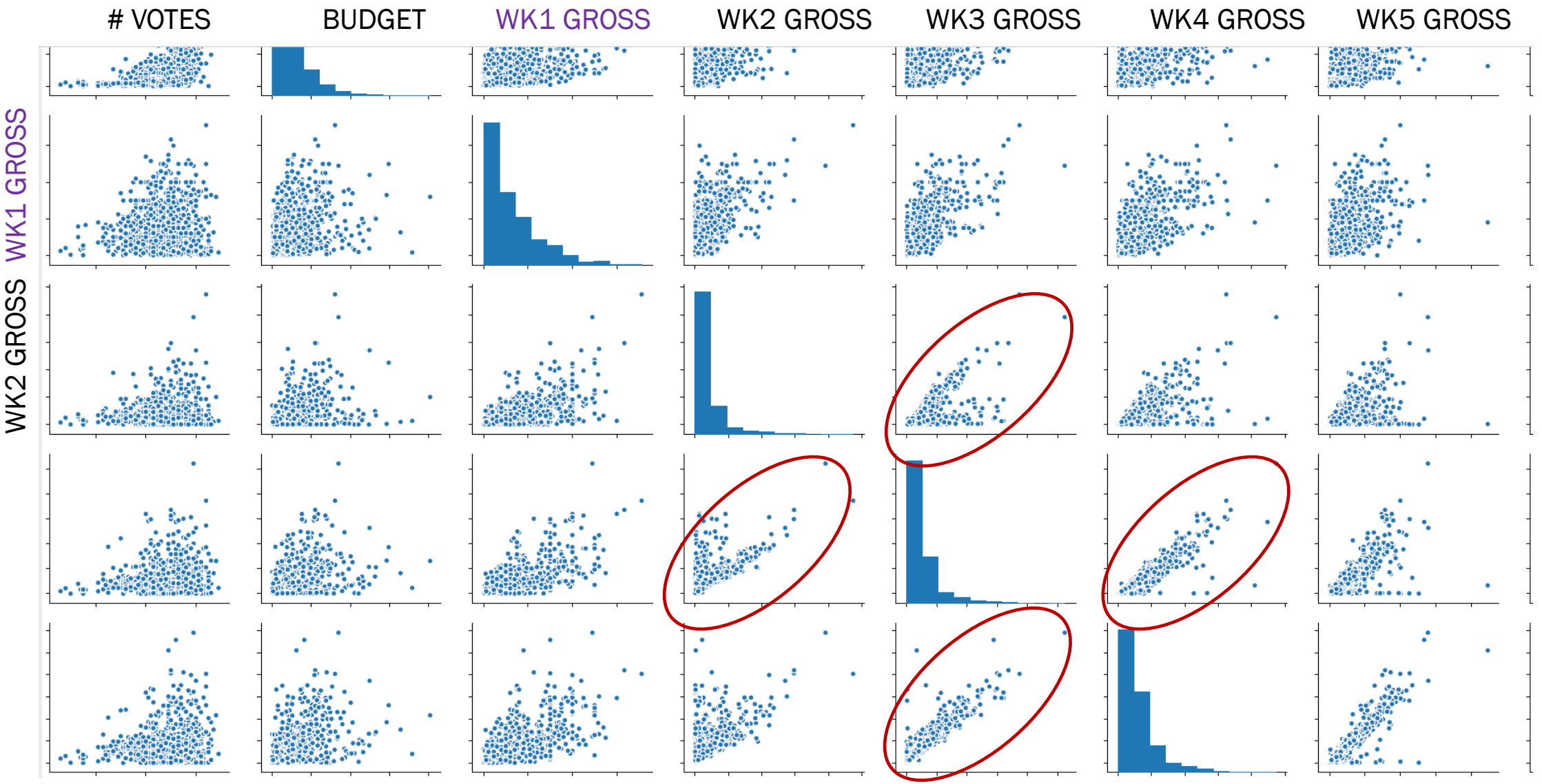
Considerations / Observations

- 19,355 movies scraped, 4341 with budget info
  - 1121 movies used (has weekly gross data)
- 83 features considered, reduced to 5
- Data scraped: BoxOfficeMojo, IMDB, and Statista
- Budget, Week 1 gross, Number of opening theaters: largest initial correlation to annual gross
- Surprisngly, IMDB ratings and number of votes appear to be poorly correlated to a film's success
  - Sample-size related?

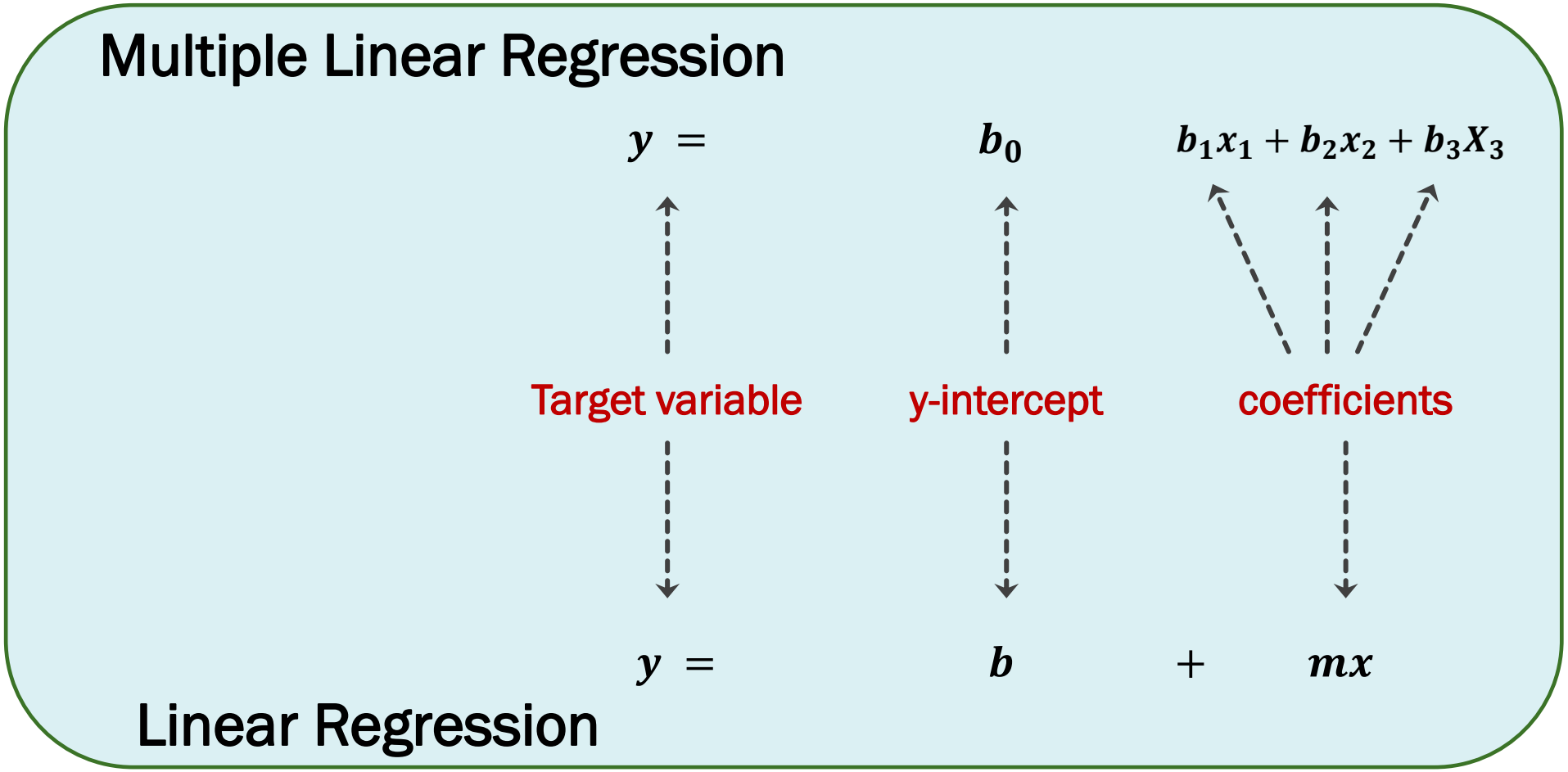
EXAMINING DATA



COLLINEARITY



LINEAR REGRESSION



INITIAL PASS (OLS): WHAT'S HAPPENING

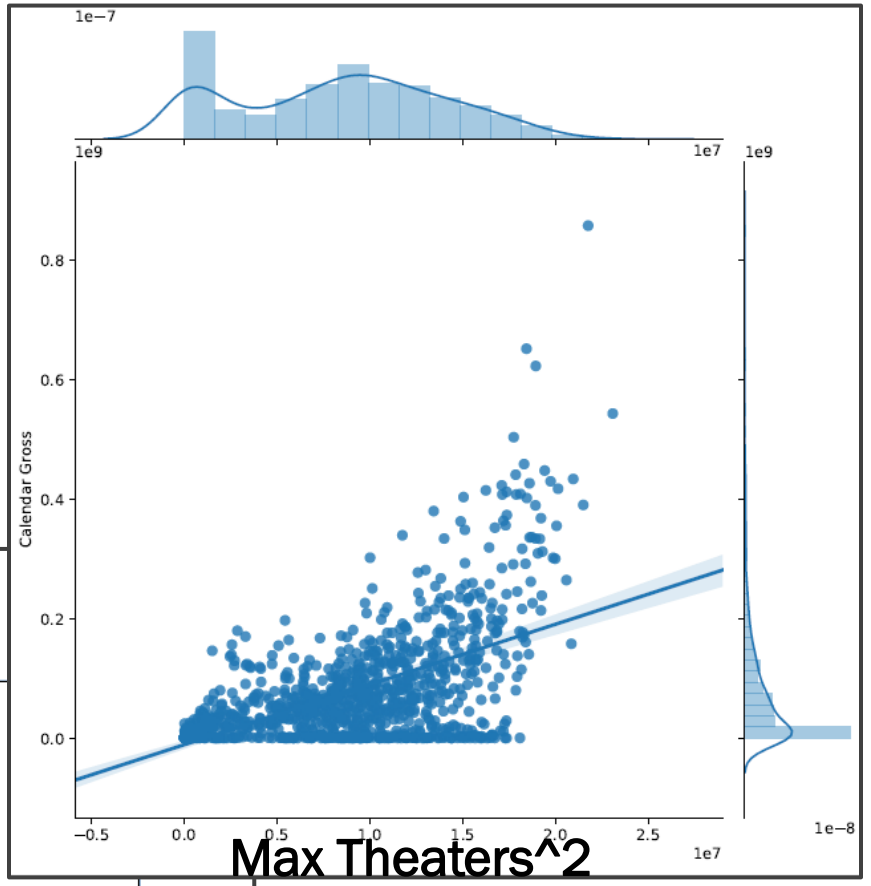
OLS Regression Results

Dep. Variable:	Calendar Gross	R-squared:	0.493
Model:	OLS	Adj. R-squared:	0.492
Method:	Least Squares	F-statistic:	543.6
Date:	Fri, 09 Oct 2020	Prob (F-statistic):	1.20e-165
Time:	03:29:47	Log-Likelihood:	-21795.
No. Observations:	1121	AIC:	4.360e+04
Df Residuals:	1118	BIC:	4.361e+04
Df Model:	2		
Covariance Type:	nonrobust		

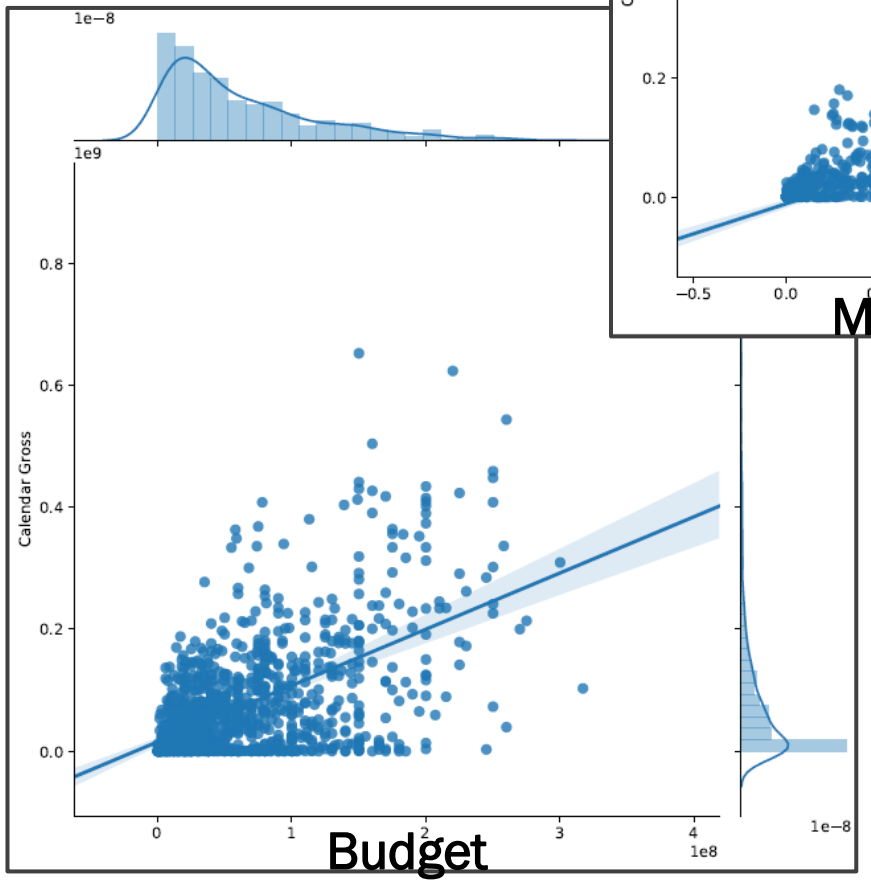
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.21e+07	6.46e+06	-1.873	0.061	-2.48e+07	5.74e+05
Max Theaters sqrt	9.045e+05	1.41e+05	6.407	0.000	6.27e+05	1.18e+06
wkGROSS_1	1.2856	0.052	24.700	0.000	1.184	1.388

Omnibus:	137.053	Durbin-Watson:	1.904
Prob(Omnibus):	0.000	Jarque-Bera (JB):	964.335
Skew:	0.304	Prob(JB):	3.96e-210
Kurtosis:	7.503	Cond. No.	1.77e+08

$R^2 = 0.493$



Max Theaters^2



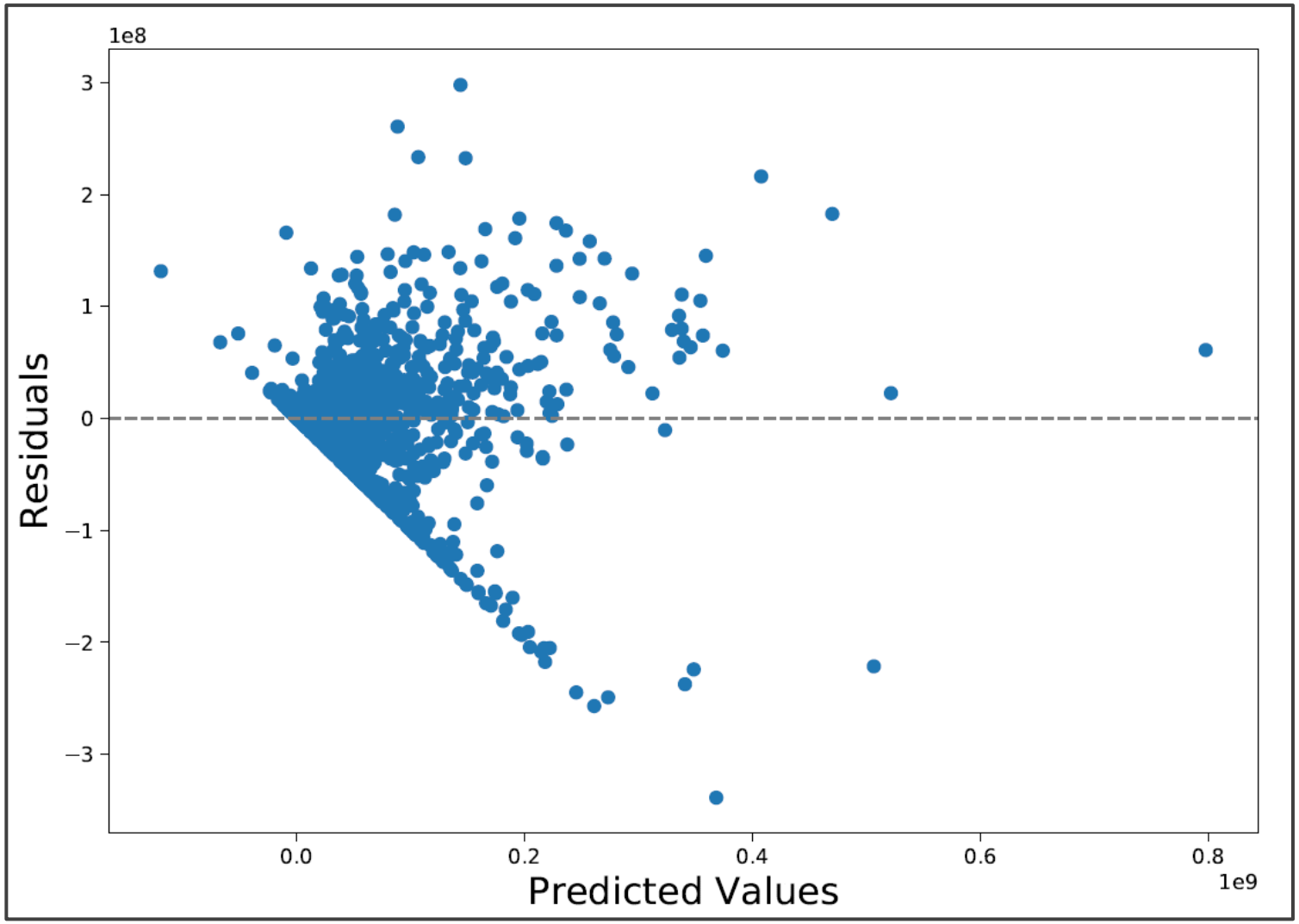
Budget

POLYNOMIAL REGRESSION

Features: Max Theaters, Rating, Votes, Budget, Week 1 Gross

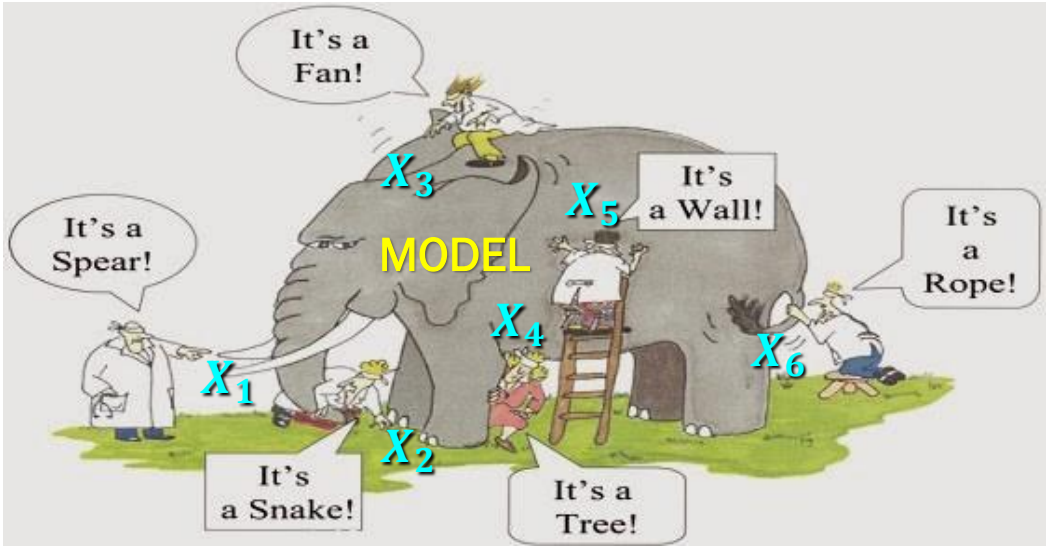
Dep. Variable:	Calendar Gross	R-squared:	0.578
Model:	OLS	Adj. R-squared:	0.572
Method:	Least Squares	F-statistic:	88.89
Date:	Fri, 09 Oct 2020	Prob (F-statistic):	9.60e-193
Time:	08:19:40	Log-Likelihood:	-21692.
No. Observations:	1121	AIC:	4.342e+04
Df Residuals:	1103	BIC:	4.351e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	31.2779	10.678	2.929	0.003	10.326	52.230
x1	4.128e+04	1.41e+04	2.932	0.003	1.37e+04	6.89e+04
x2	135.9299	46.125	2.947	0.003	45.428	226.432
x3	623.8616	143.038	4.362	0.000	343.205	904.519
x4	-0.5528	0.564	-0.981	0.327	-1.659	0.553
x5	-4.9005	0.837	-5.854	0.000	-6.543	-3.258
x6	-7.3128	2.337	-3.129	0.002	-11.898	-2.727
x7	-1539.8816	1818.592	-0.847	0.397	-5108.172	2028.409
x8	0.0094	0.011	0.832	0.406	-0.013	0.032
x9	0.0004	9.32e-05	3.882	0.000	0.000	0.001
x10	0.0008	0.000	6.118	0.000	0.001	0.001
x11	317.6986	104.678	3.035	0.002	112.307	523.090
x12	-71.3154	19.175	-3.719	0.000	-108.938	-33.693
x13	-0.0684	0.078	-0.879	0.380	-0.221	0.084



RECOMMENDATIONS

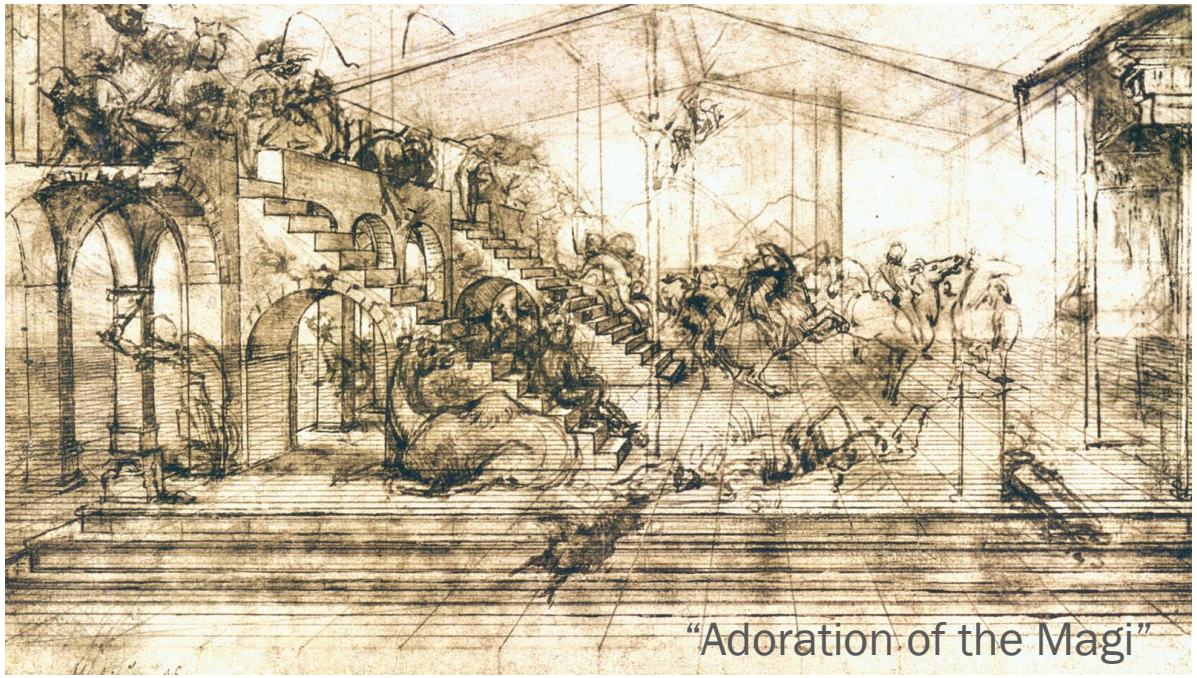
- Feature-engineering could be critical
  - There's no smoking-gun approach with real-world data
  - Better features give our model more flexibility



- Advanced quality control or data constraints
  - Removing bias through Lasso/Ridge regressions

- More iterations
  - Going back to look at assumptions during EDA

- Sentiment analysis
  - In particular: Momentum (rate) of initial reviews


"Adoration of the Magi"

QUESTIONS