DEREK UPDEGRAFF

PREDICTING PROFITABLE FILMS

(SLIGHTLY) DIFFERENT TASTES

- International

1	Marvel's The Avengers
2	Skyfall
3	The Dark Knight Rises
4	The Hobbit: An Unexpected Journey
5	Ice Age: Continental Drift
6	The Twilight Saga: Breaking Dawn Part 2
7	The Amazing Spider-Man
8	Madagascar 3: Europe's Most Wanted
9	The Hunger Games
10	MIB 3

- Domestic

1	Marvel's The Avengers
2	The Dark Knight Rises
3	The Hunger Games
4	Skyfall
5	The Hobbit: An Unexpected Journey
6	The Twilight Saga: Breaking Dawn Part 2
7	The Amazing Spider-Man
8	Brave
9	Ted
10	Madagascar 3: Europe's Most Wanted

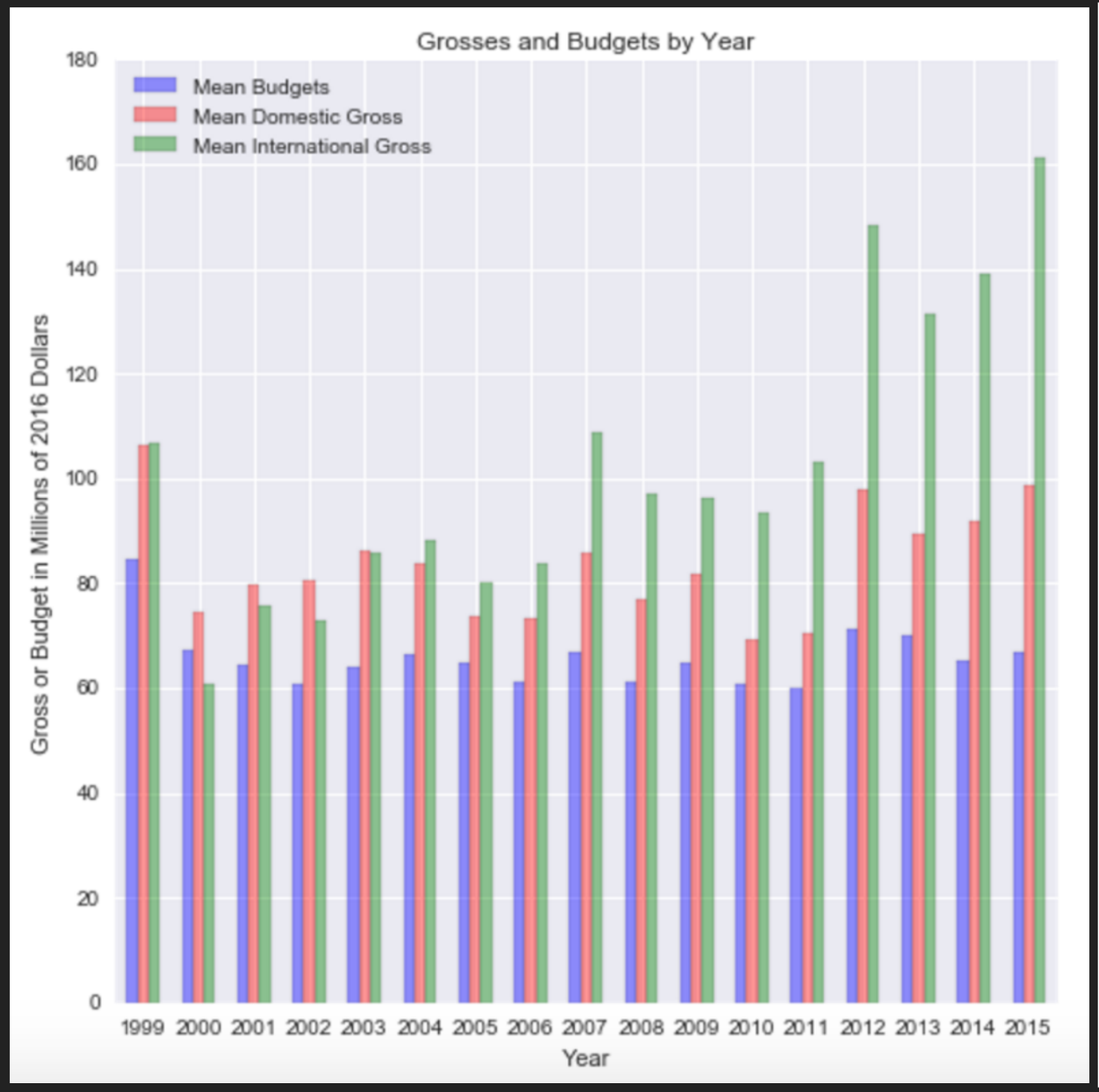
PROCESS

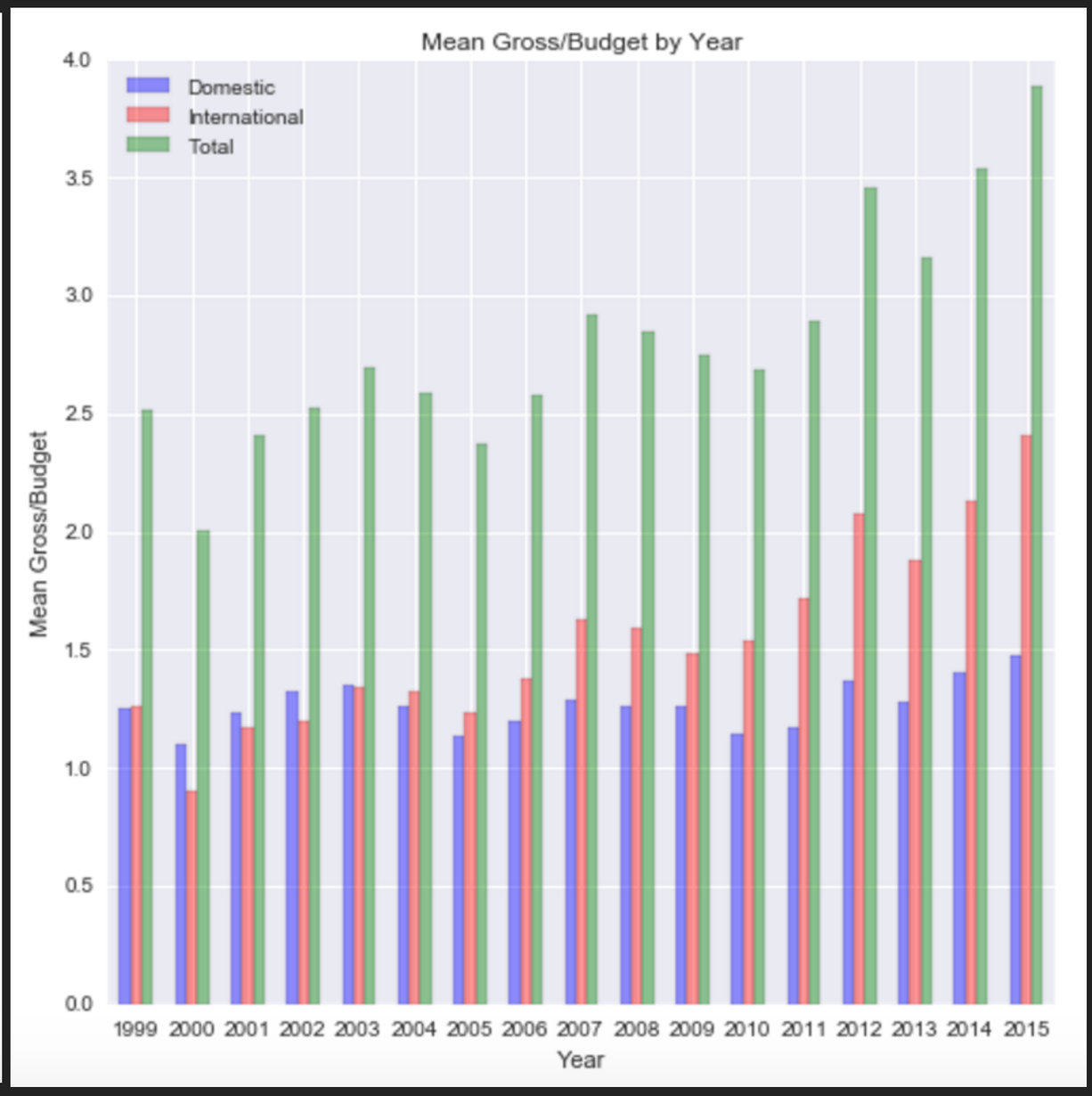
- Scraped >13,000 movies from Boxofficemojo, 1980-2015
  - Got domestic and international gross, MPAA rating, genres
  - MPAA rating used ordinal (G = 1, PG = 2, etc.), genres as binaries
  - Threw out films with budgets < 10 million to avoid skew
    - You only hear about the $50,000 budget movies that make it
  - >2000 movies had all desired features

	Domestic Gross (2016 dollars)	Opening	Open	Year	International Gross (2016 dollars)	Budget (2016 dollars)	Runtime	Rating	action	romantic
0	6.376574e+08	14953367.49	142	1980	7.549033e+08	5.481347e+07	129.0	1	0	0
116	5.779172e+08	22618188.07	164	1981	3.860533e+08	4.901710e+07	115.0	1	0	0
229	9.024340e+08	29734814.64	163	1982	8.989241e+08	2.637983e+07	117.0	1	0	0

MARKET DYNAMICS

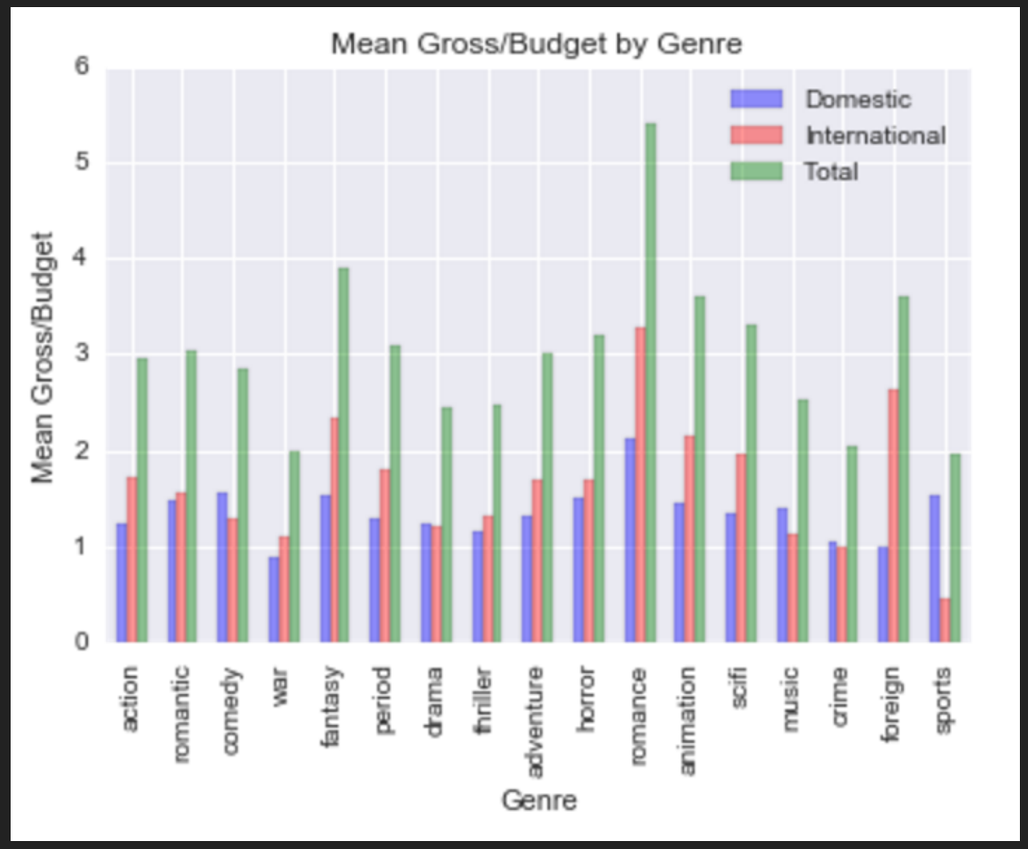
- Films getting more profitable, especially internationally

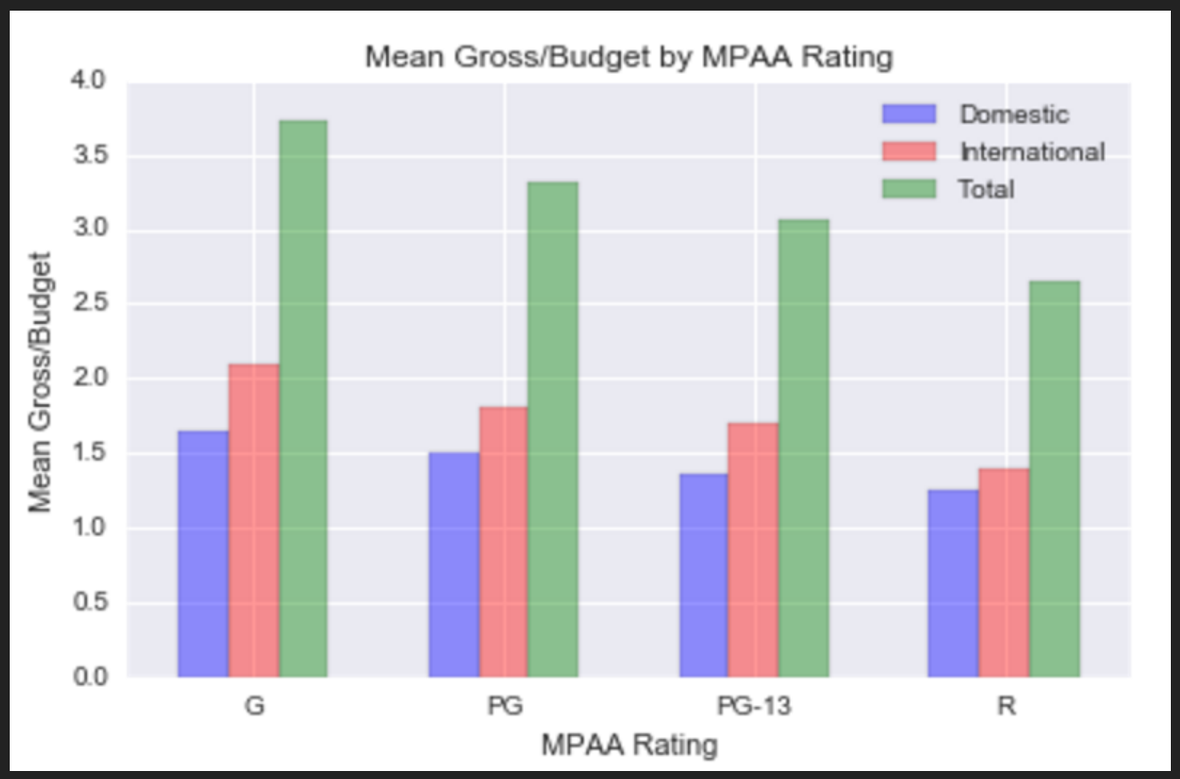


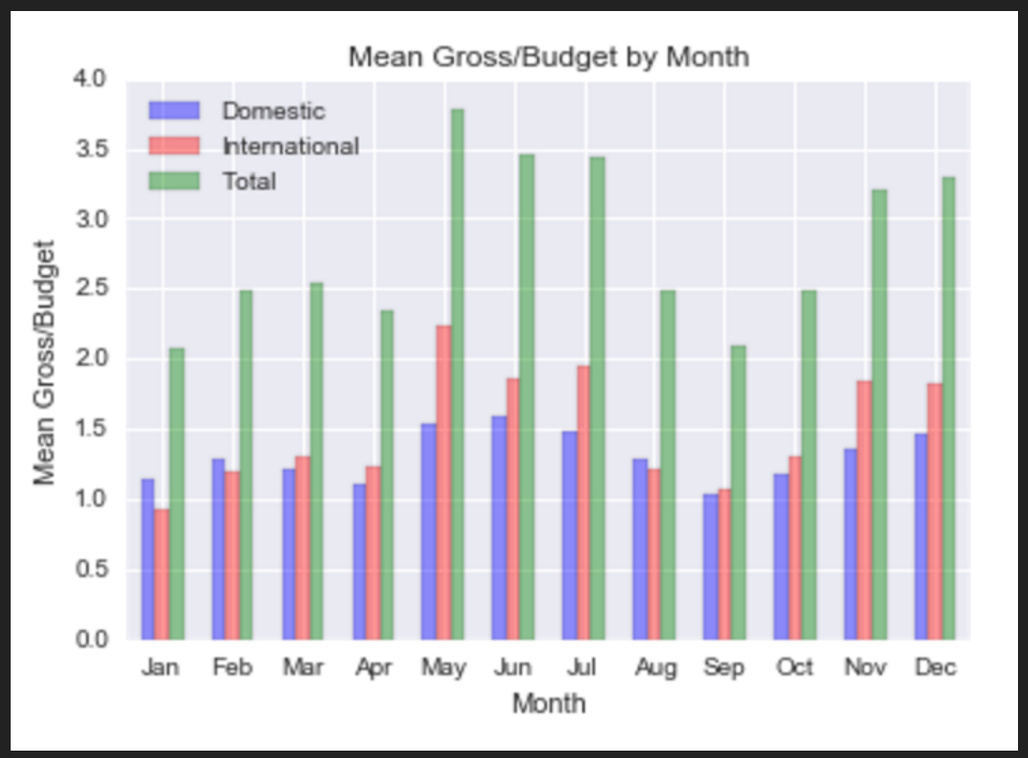


HOW TO MAXIMIZE RETURN

- Romance/Fantasy/Animation
- Release May/Jun/Jul
- Keep rating low





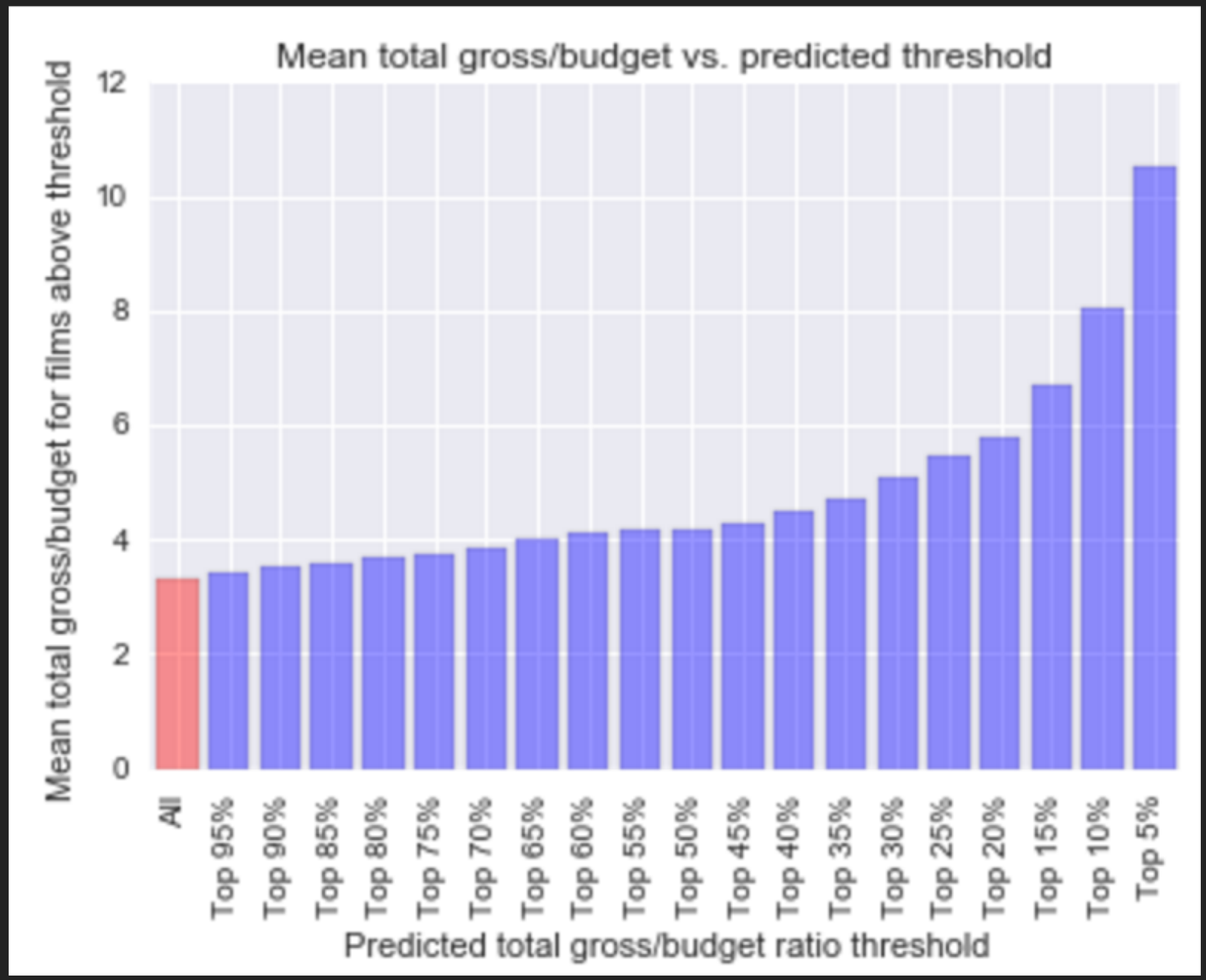


MODEL/RESULTS

- Goal to predict total gross/budget to aid investment
- Built models with grid search on 5-fold cross val, 20% holdout
- Linear Regression, Lasso, Ridge, Elastic Net, Random Forest
- Linear models have R^2 ~0.08, Random Forest 0.23
- Suggests most variance isn’t explained by year/genre/etc.

MODEL/RESULTS

- Model is good enough to use for investments
- Top 5% predicted have actual Gross/ Budget more than 3 times higher than the overall average



MODEL/RESULTS

- Caveats
  - Train set isn’t future, model bakes in assumptions about future market
  - Sample only has films that got made and have good data; skewed
  - Other factors in profit than Box Office/Return ratio

NEXT STEPS

- More and cleaner data
  - Data can still be biased
- Integrate ratings and actors
- Split test/train by time to predict “future”
  - Maybe time series