Merchandising Movies

Avi Grunwald

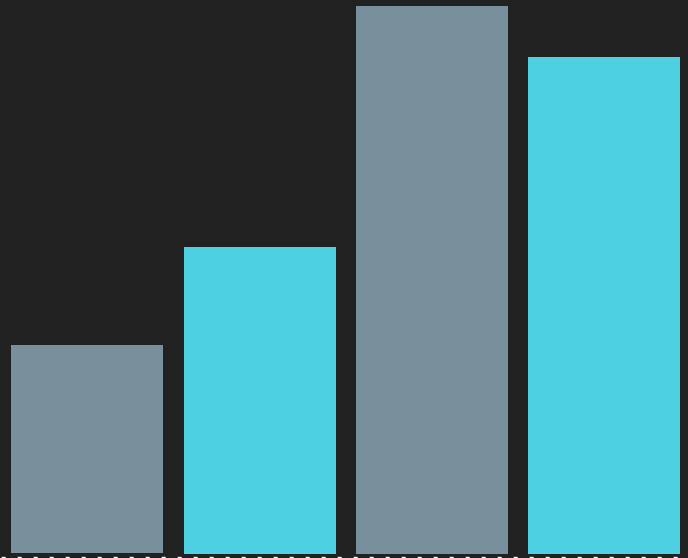
The Products

Pictures: http://www.therichest.com/rich-list/most-popular/10-movies-that-sold-the-most-merchandise/

The problem

Figuring out the perfect number of products to produce is very tough.

Especially if you don’t know how well a movie will perform over time.



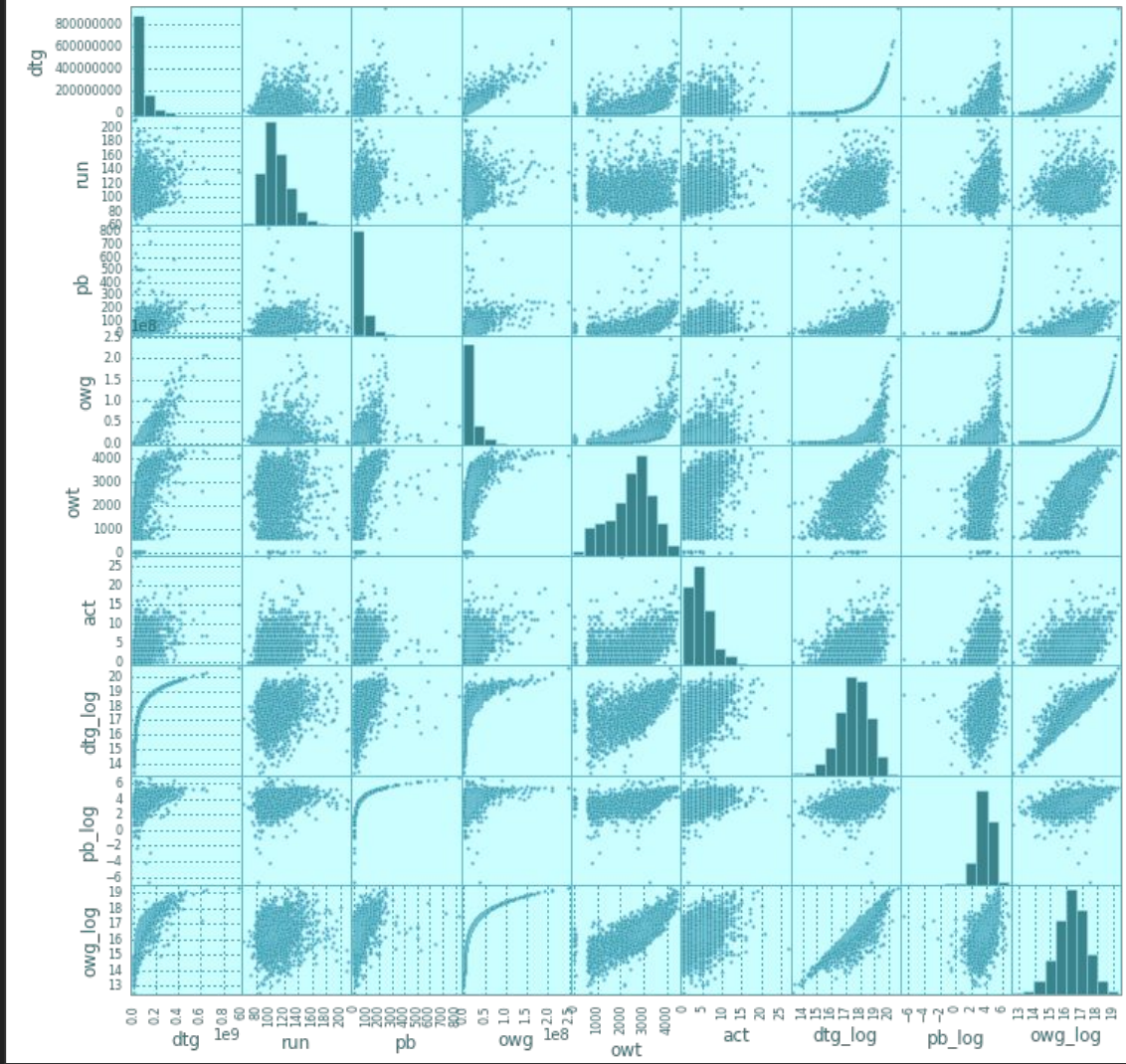
The solution

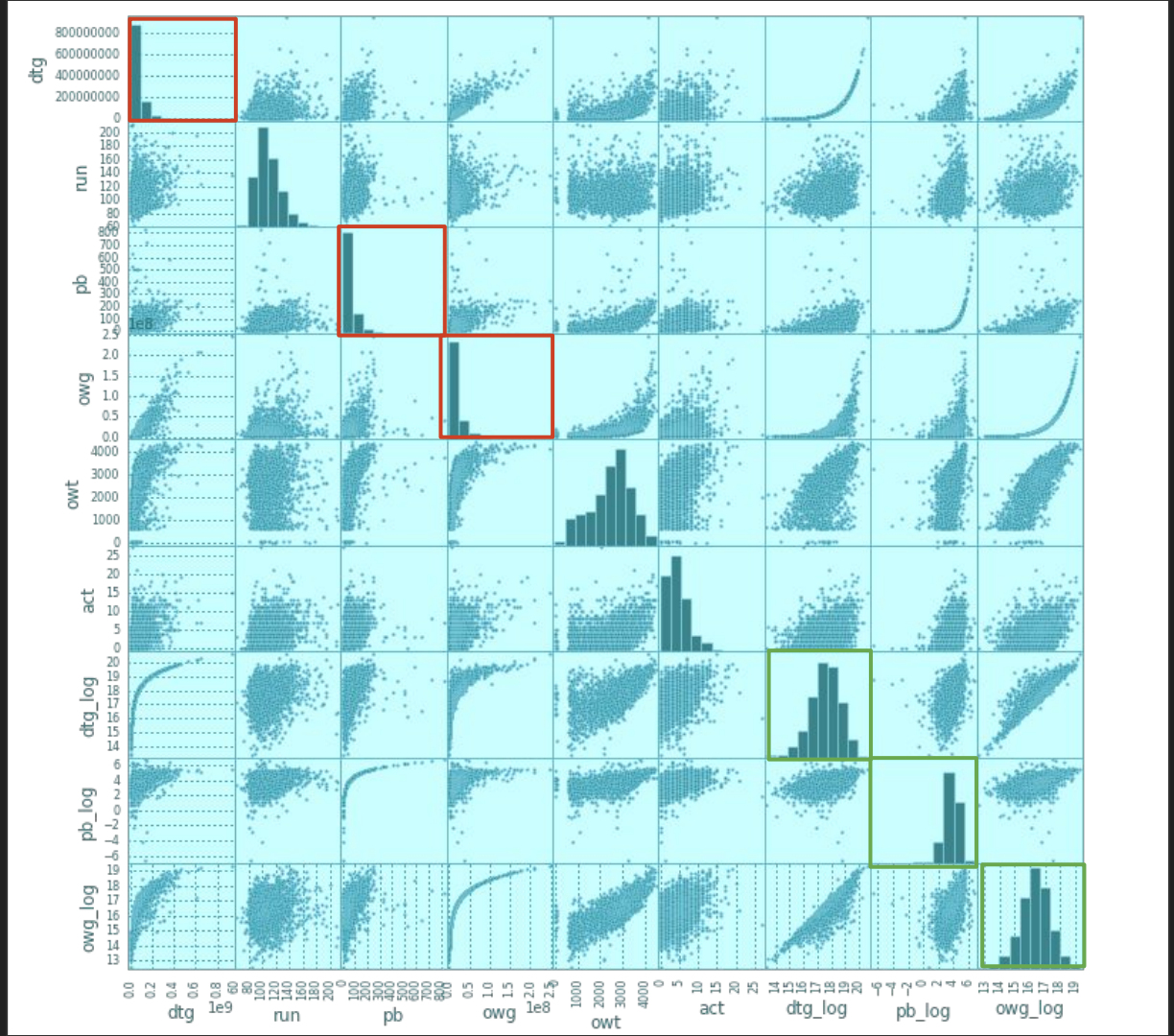
Using available data from Box Office Mojo we can predict the Domestic Total Gross (DTG).

Scrape information on over 4000 movies to make a predictive model.

Data

Title	Domestic Total Gross (millions)	Distributor	Release Date	Runtime	MPAA Rating	Production Budget (millions)	Opening Weekend (millions)	Opening Weekend Theaters	Actors	Series
Star Wars: The Force Awakens	$936	Buena Vista	December 18, 2015	2 hrs. 16 min.	PG-13	$245	$247	4,134	15	Yes
Marvel's The Avengers	$623	Buena Vista	May 4, 2012	2 hrs. 22 min.	PG-13	$220	$207	4,349	13	Yes
Jurassic World	$652	Universal	June 12, 2015	2 hrs. 4 min.	PG-13	$150	$208	4,274	7	Yes
Avengers: Age of Ultron	$459	Buena Vista	May 1, 2015	2 hrs. 21 min.	PG-13	$250	$191	4,276	19	Yes
The Dark Knight	$533	Warner Bros.	July 18, 2008	2 hrs. 30 min.	PG-13	$185	$158	4,366	11	Yes


dtg
run
pb
owg
owt
act
dtg_log
pb_log
owg_log
1e9
1e8


dtg
run
pb
owg
owt
act
dtg_log
pb_log
owg_log
1e9
1e8

OLS Results

OLS Regression Results

Dep. Variable:	dtg_log	**R-squared:**	0.856
Model:	OLS	**Adj. R-squared:**	0.856
Method:	Least Squares	**F-statistic:**	1201.
Date:	Thu, 14 Jul 2016	**Prob (F-statistic):**	0.00
Time:	17:46:59	**Log-Likelihood:**	-2825.4
No. Observations:	4662	**AIC:**	5699.
Df Residuals:	4638	**BIC:**	5854.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.9927	0.123	8.066	0.000	0.751 1.234
dist[T.DreamWorks]	0.1056	0.060	1.755	0.079	-0.012 0.224
dist[T.Paramount]	0.0580	0.024	2.456	0.014	0.012 0.104
dist[T.Universal]	-0.0260	0.021	-1.222	0.222	-0.068 0.016
rdm[T.August]	0.0620	0.023	2.665	0.008	0.016 0.108
rdm[T.December]	0.5432	0.024	22.927	0.000	0.497 0.590
rdm[T.February]	-0.0668	0.028	-2.425	0.015	-0.121 -0.013
rdm[T.January]	-0.1061	0.031	-3.472	0.001	-0.166 -0.046
rdm[T.July]	0.1907	0.025	7.560	0.000	0.141 0.240
rdm[T.June]	0.2025	0.026	7.745	0.000	0.151 0.254
rdm[T.May]	0.0753	0.027	2.745	0.006	0.022 0.129
rdm[T.November]	0.1346	0.025	5.476	0.000	0.086 0.183
rating[T.PG]	-0.1511	0.036	-4.154	0.000	-0.222 -0.080
rating[T.PG-13]	-0.3628	0.036	-10.115	0.000	-0.433 -0.293
rating[T.R]	-0.3197	0.036	-8.946	0.000	-0.390 -0.250
rating[T.Unrated]	-1.0835	0.317	-3.413	0.001	-1.706 -0.461
series[T.Yes]	-0.0145	0.021	-0.695	0.487	-0.055 0.026
run	0.0057	0.000	14.161	0.000	0.005 0.006
owg_log	0.9933	0.008	129.317	0.000	0.978 1.008
act	-0.0110	0.002	-4.416	0.000	-0.016 -0.006

OLS Results

OLS Regression Results

Dep. Variable:	dtg_log	**R-squared:**	0.856
Model:	OLS	**Adj. R-squared:**	0.856
Method:	Least Squares	**F-statistic:**	1201.
Date:	Thu, 14 Jul 2016	**Prob (F-statistic):**	0.00
Time:	17:46:59	**Log-Likelihood:**	-2825.4
No. Observations:	4662	**AIC:**	5699.
Df Residuals:	4638	**BIC:**	5854.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.9927	0.123	8.066	0.000	0.751 1.234
dist[T.DreamWorks]	0.1056	0.060	1.755	0.079	-0.012 0.224
dist[T.Paramount]	0.0580	0.024	2.456	0.014	0.012 0.104
dist[T.Universal]	-0.0260	0.021	-1.222	0.222	-0.068 0.016
rdm[T.August]	0.0620	0.023	2.665	0.008	0.016 0.108
rdm[T.December]	0.5432	0.024	22.927	0.000	0.497 0.590
rdm[T.February]	-0.0668	0.028	-2.425	0.015	-0.121 -0.013
rdm[T.January]	-0.1061	0.031	-3.472	0.001	-0.166 -0.046
rdm[T.July]	0.1907	0.025	7.560	0.000	0.141 0.240
rdm[T.June]	0.2025	0.026	7.745	0.000	0.151 0.254
rdm[T.May]	0.0753	0.027	2.745	0.006	0.022 0.129
rdm[T.November]	0.1346	0.025	5.476	0.000	0.086 0.183
rating[T.PG]	-0.1511	0.036	-4.154	0.000	-0.222 -0.080
rating[T.PG-13]	-0.3628	0.036	-10.115	0.000	-0.433 -0.293
rating[T.R]	-0.3197	0.036	-8.946	0.000	-0.390 -0.250
rating[T.Unrated]	-1.0835	0.317	-3.413	0.001	-1.706 -0.461
series[T.Yes]	-0.0145	0.021	-0.695	0.487	-0.055 0.026
run	0.0057	0.000	14.161	0.000	0.005 0.006
owg_log	0.9933	0.008	129.317	0.000	0.978 1.008
act	-0.0110	0.002	-4.416	0.000	-0.016 -0.006

Comparing Models:
Opening Weekend

OLS Regression Results

Dep. Variable:	dtg_log	**R-squared:**	0.856
Model:	OLS	**Adj. R-squared:**	0.856
Method:	Least Squares	**F-statistic:**	1201.
Date:	Thu, 14 Jul 2016	**Prob (F-statistic):**	0.00
Time:	17:46:59	**Log-Likelihood:**	-2825.4
No. Observations:	4662	**AIC:**	5699.
Df Residuals:	4638	**BIC:**	5854.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.9927	0.123	8.066	0.000	0.751 1.234
dist[T.DreamWorks]	0.1056	0.060	1.755	0.079	-0.012 0.224
dist[T.Paramount]	0.0580	0.024	2.456	0.014	0.012 0.104
dist[T.Universal]	-0.0260	0.021	-1.222	0.222	-0.068 0.016
rdm[T.August]	0.0620	0.023	2.665	0.008	0.016 0.108
rdm[T.December]	0.5432	0.024	22.927	0.000	0.497 0.590
rdm[T.February]	-0.0668	0.028	-2.425	0.015	-0.121 -0.013
rdm[T.January]	-0.1061	0.031	-3.472	0.001	-0.166 -0.046
rdm[T.July]	0.1907	0.025	7.560	0.000	0.141 0.240
rdm[T.June]	0.2025	0.026	7.745	0.000	0.151 0.254
rdm[T.May]	0.0753	0.027	2.745	0.006	0.022 0.129
rdm[T.November]	0.1346	0.025	5.476	0.000	0.086 0.183
rating[T.PG]	-0.1511	0.036	-4.154	0.000	-0.222 -0.080
rating[T.PG-13]	-0.3628	0.036	-10.115	0.000	-0.433 -0.293
rating[T.R]	-0.3197	0.036	-8.946	0.000	-0.390 -0.250
rating[T.Unrated]	-1.0835	0.317	-3.413	0.001	-1.706 -0.461
series[T.Yes]	-0.0145	0.021	-0.695	0.487	-0.055 0.026
run	0.0057	0.000	14.161	0.000	0.005 0.006
owg_log	0.9933	0.008	129.317	0.000	0.978 1.008
act	-0.0110	0.002	-4.416	0.000	-0.016 -0.006

OLS Regression Results

Dep. Variable:	dtg_log	**R-squared:**	0.395
Model:	OLS	**Adj. R-squared:**	0.391
Method:	Least Squares	**F-statistic:**	102.3
Date:	Fri, 15 Jul 2016	**Prob (F-statistic):**	3.69e-243
Time:	11:44:40	**Log-Likelihood:**	-2918.1
No. Observations:	2370	**AIC:**	5868.
Df Residuals:	2354	**BIC:**	5961.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	15.5750	0.110	141.469	0.000	15.359 15.791
dist[T.DreamWorks]	0.2021	0.118	1.707	0.088	-0.030 0.434
dist[T.Newmarket]	0.2349	0.481	0.488	0.625	-0.709 1.178
dist[T.Orion Pictures]	0.0685	0.317	0.216	0.829	-0.553 0.690
rd[T.August]	0.0396	0.062	0.640	0.522	-0.082 0.161
rd[T.December]	0.2928	0.059	4.936	0.000	0.176 0.409
rd[T.July]	0.3974	0.064	6.226	0.000	0.272 0.523
rd[T.June]	0.5193	0.064	8.105	0.000	0.394 0.645
rd[T.May]	0.3590	0.072	4.999	0.000	0.218 0.500
rd[T.November]	0.3318	0.062	5.328	0.000	0.210 0.454
rating[T.PG-13]	-0.1970	0.048	-4.110	0.000	-0.291 -0.103
rating[T.R]	-0.3925	0.050	-7.897	0.000	-0.490 -0.295
series[T.Yes]	0.7954	0.047	17.091	0.000	0.704 0.887
run	0.0066	0.001	6.522	0.000	0.005 0.009
pb_log	0.2637	0.020	13.357	0.000	0.225 0.302
act	0.0510	0.006	8.655	0.000	0.039 0.063

Comparing Models:
Production Budget

OLS Regression Results

Dep. Variable:	dtg_log	**R-squared:**	0.337
Model:	OLS	**Adj. R-squared:**	0.335
Method:	Least Squares	**F-statistic:**	138.9
Date:	Fri, 15 Jul 2016	**Prob (F-statistic):**	0.00
Time:	12:12:34	**Log-Likelihood:**	-6388.6
No. Observations:	4662	**AIC:**	1.281e+04
Df Residuals:	4644	**BIC:**	1.293e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	15.6479	0.102	153.031	0.000	15.447 15.848
dist[T.DreamWorks]	0.5987	0.129	4.647	0.000	0.346 0.851
dist[T.Paramount]	0.3094	0.050	6.129	0.000	0.210 0.408
dist[T.Universal]	0.2550	0.045	5.622	0.000	0.166 0.344
rdm[T.December]	0.4069	0.050	8.201	0.000	0.310 0.504
rdm[T.February]	0.2003	0.058	3.462	0.001	0.087 0.314
rdm[T.January]	0.0648	0.064	1.004	0.315	-0.062 0.191
rdm[T.July]	0.4645	0.053	8.813	0.000	0.361 0.568
rdm[T.June]	0.5870	0.055	10.763	0.000	0.480 0.694
rdm[T.May]	0.2223	0.058	3.853	0.000	0.109 0.335
rdm[T.November]	0.2638	0.051	5.123	0.000	0.163 0.365
rating[T.PG]	-0.3300	0.078	-4.233	0.000	-0.483 -0.177
rating[T.PG-13]	-0.3977	0.077	-5.170	0.000	-0.549 -0.247
rating[T.R]	-0.6593	0.076	-8.624	0.000	-0.809 -0.509
rating[T.Unrated]	-3.3459	0.680	-4.921	0.000	-4.679 -2.013
series[T.Yes]	0.8994	0.042	21.335	0.000	0.817 0.982
run	0.0109	0.001	12.720	0.000	0.009 0.013
act	0.1177	0.005	24.050	0.000	0.108 0.127

Omnibus:	232.536	**Durbin-Watson:**	0.793
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	281.979
Skew:	-0.524	**Prob(JB):**	5.88e-62
Kurtosis:	3.593	**Cond. No.**	5.27e+03

OLS Regression Results

Dep. Variable:	dtg_log	**R-squared:**	0.395
Model:	OLS	**Adj. R-squared:**	0.391
Method:	Least Squares	**F-statistic:**	102.3
Date:	Fri, 15 Jul 2016	**Prob (F-statistic):**	3.69e-243
Time:	11:44:40	**Log-Likelihood:**	-2918.1
No. Observations:	2370	**AIC:**	5868.
Df Residuals:	2354	**BIC:**	5961.
Df Model:	15		
Covariance Type:	nonrobust		

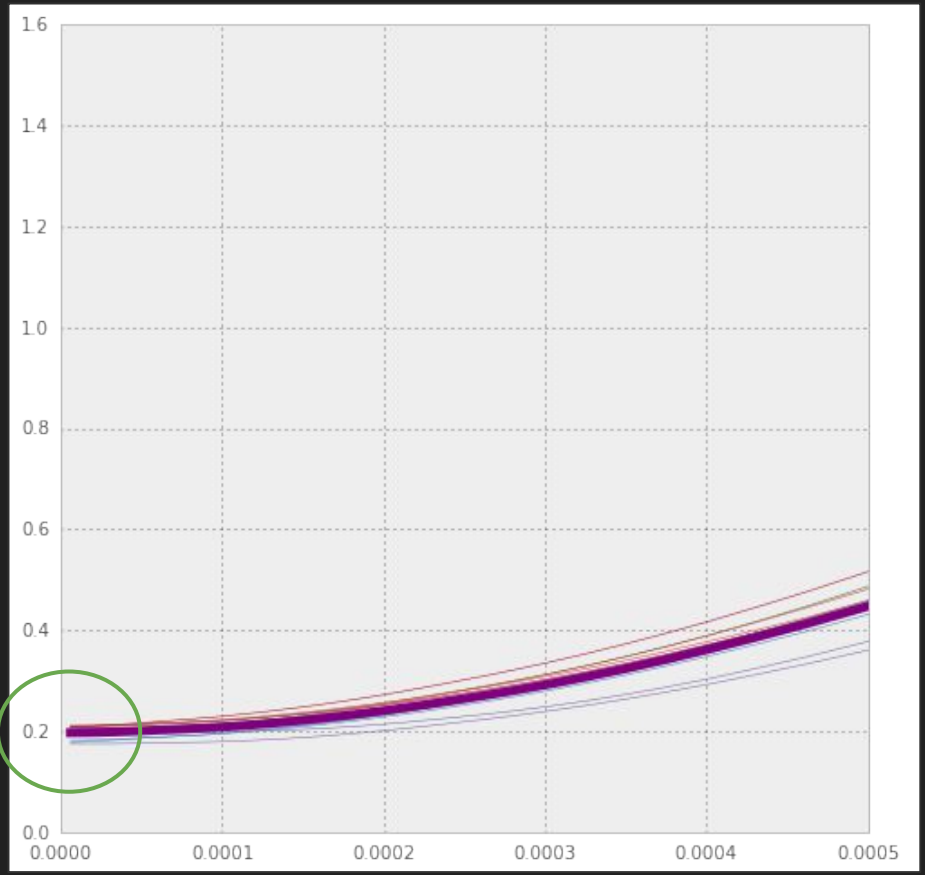
	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	15.5750	0.110	141.469	0.000	15.359 15.791
dist[T.DreamWorks]	0.2021	0.118	1.707	0.088	-0.030 0.434
dist[T.Newmarket]	0.2349	0.481	0.488	0.625	-0.709 1.178
dist[T.Orion Pictures]	0.0685	0.317	0.216	0.829	-0.553 0.690
rd[T.August]	0.0396	0.062	0.640	0.522	-0.082 0.161
rd[T.December]	0.2928	0.059	4.936	0.000	0.176 0.409
rd[T.July]	0.3974	0.064	6.226	0.000	0.272 0.523
rd[T.June]	0.5193	0.064	8.105	0.000	0.394 0.645
rd[T.May]	0.3590	0.072	4.999	0.000	0.218 0.500
rd[T.November]	0.3318	0.062	5.328	0.000	0.210 0.454
rating[T.PG-13]	-0.1970	0.048	-4.110	0.000	-0.291 -0.103
rating[T.R]	-0.3925	0.050	-7.897	0.000	-0.490 -0.295
series[T.Yes]	0.7954	0.047	17.091	0.000	0.704 0.887
run	0.0066	0.001	6.522	0.000	0.005 0.009
pb_log	0.2637	0.020	13.357	0.000	0.225 0.302
act	0.0510	0.006	8.655	0.000	0.039 0.063

Omnibus:	152.617	**Durbin-Watson:**	0.992
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	240.715
Skew:	-0.514	**Prob(JB):**	5.36e-53
Kurtosis:	4.175	**Cond. No.**	3.12e+03

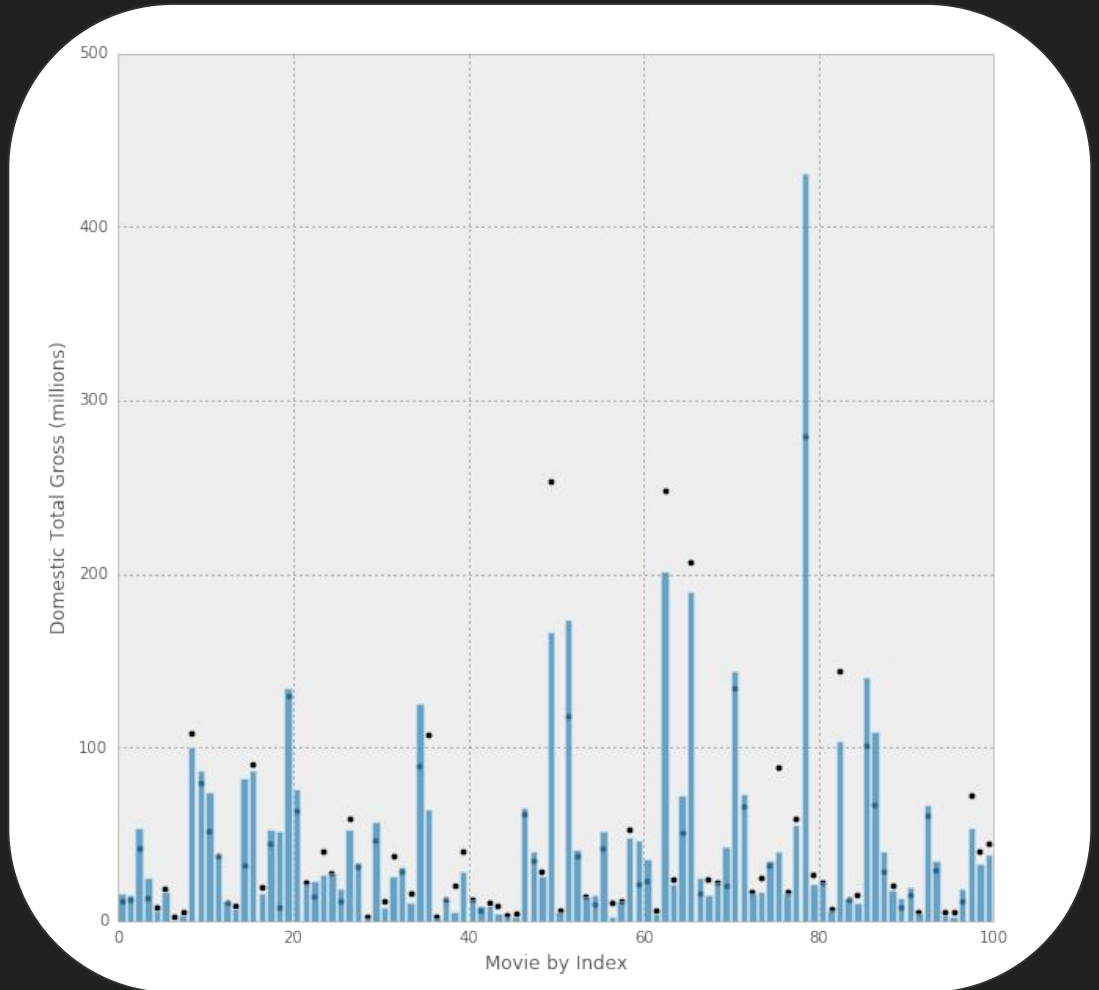
Lasso Regression

Best Alpha:
6.104339012444183e-06

MSE on Test:
0.20509087972043449



Predictions vs Actuals



Further Research

What else can be done in the future

Group Data by Year

Does a Popular Producer Matter

Use Even More Data Points

What Will Make The Model Even Better

Compare Production Budgets

Consider the Effect of Popular Actors

Thank You