Predicting Opening Weekend Box Office Performance

From IMDB Search Frequency of Principal Cast Members

Christopher Giler

January 26, 2018

Agenda

1. Project Objective
2. Data Overview
3. Key Features
4. Exploratory Data Analysis
5. Feature Engineering
6. Modeling Box Office Sales
7. Conclusion & Next Steps

1

Project Objective

Defining Project Goals

The Problem Statement

- How much does casting affect the hype surrounding a movie's release?
- Evaluate the impact of cast popularity on opening weekend box office ticket sales.

STALLONE STATHAM LI LUNDGREN NORRIS CREWS COUTURE
HEMSWORTH ALSO VAN DAMME WITH WILLIS AND SCHWARZENEGGER
THE EXPENDABLES 2
16
BACK FOR WAR

2
Data Overview

Scraping and Cleaning from Data Sources

2014 ~ 2017

4 years of box office data

994

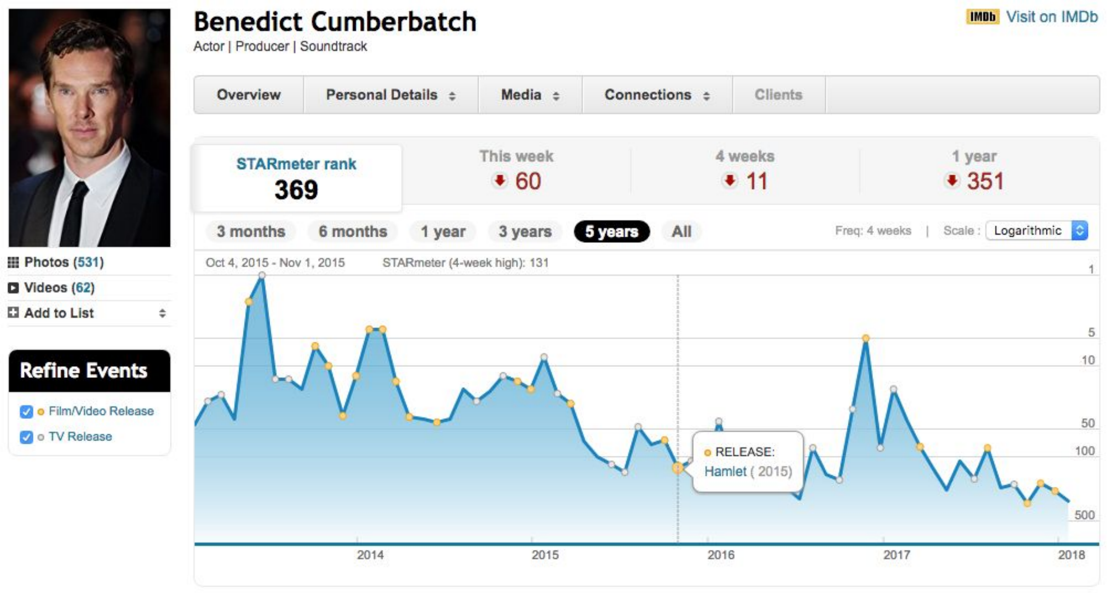
Movies available w/ box office data

8

Total features for model (raw)

- IMDB
  - Release Date
  - Opening Weekend Gross
  - Metacritic / IMDB Reviews
  - Genre
  - MPAA Rating
  - Principal Cast
- IMDB Pro
  - STARmeter Data (next slide)
  - Number Theaters

IMDB Pro’s STARmeter



Benedict Cumberbatch

Actor | Producer | Soundtrack

STARmeter rank 369

This week ↓ 60

4 weeks ↓ 11

1 year ↓ 351

Freq: 4 weeks | Scale : Logarithmic

Oct 4, 2015 - Nov 1, 2015 STARmeter (4-week high): 131

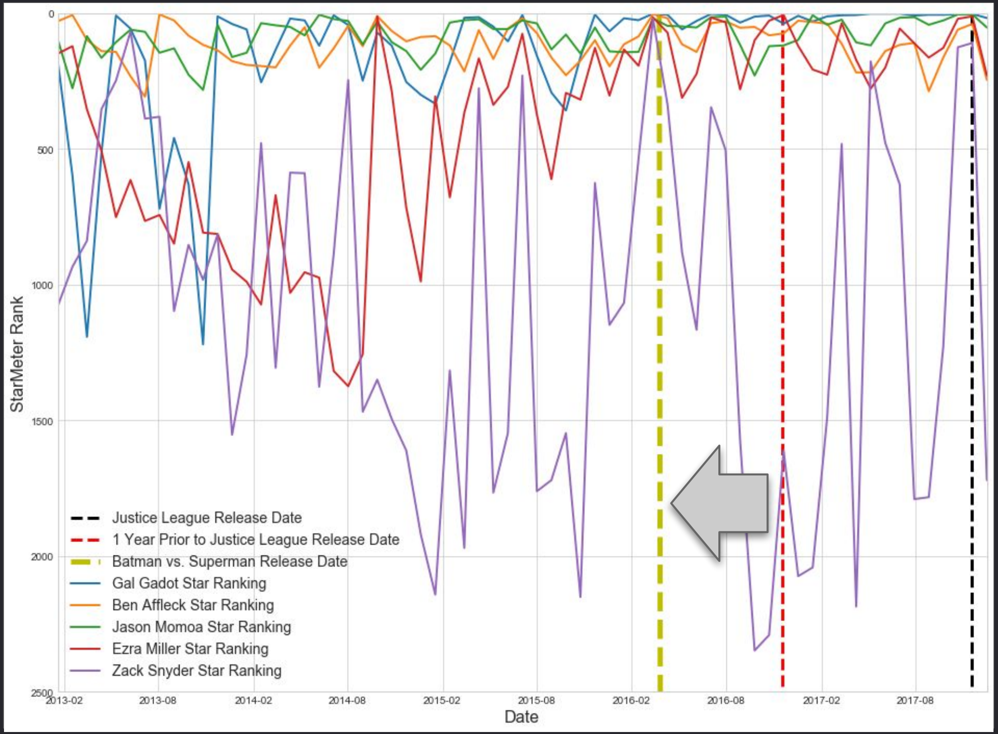
RELEASE: Hamlet (2015)

- Cast / Crew ranking based on IMDB search frequency
- Presented as time series data
- Scraped via Python / Selenium

3
Key Features

From IMDB Pro

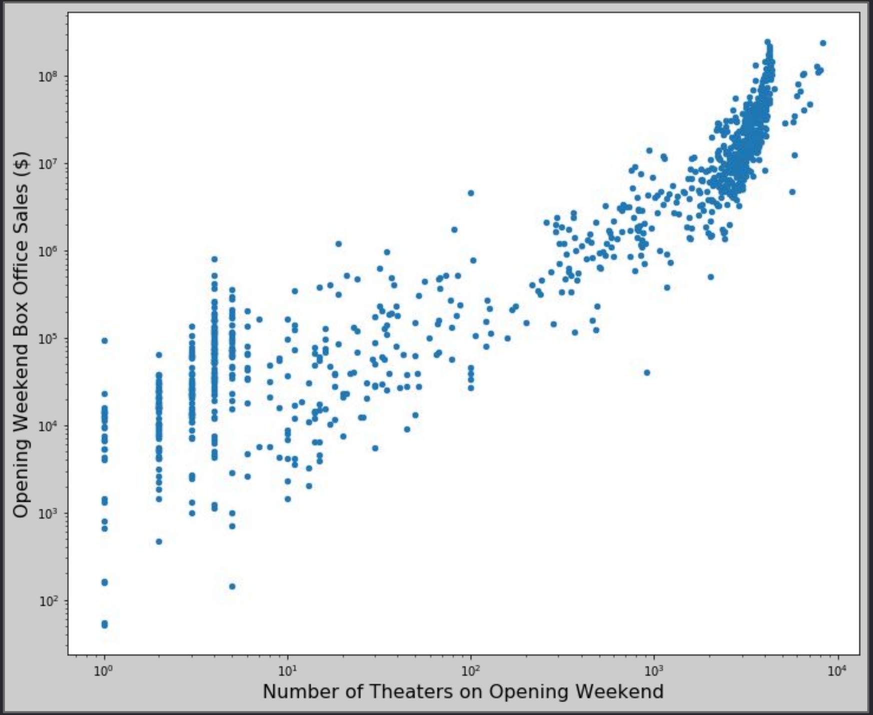
Using IMDB Pro's StarMeter Data



- Feature defined as minimum star ranking prior to 12 months before release date.

- A film's "Star Power" is based on average of top 5 star rankings

Opening Weekend Gross per Theater

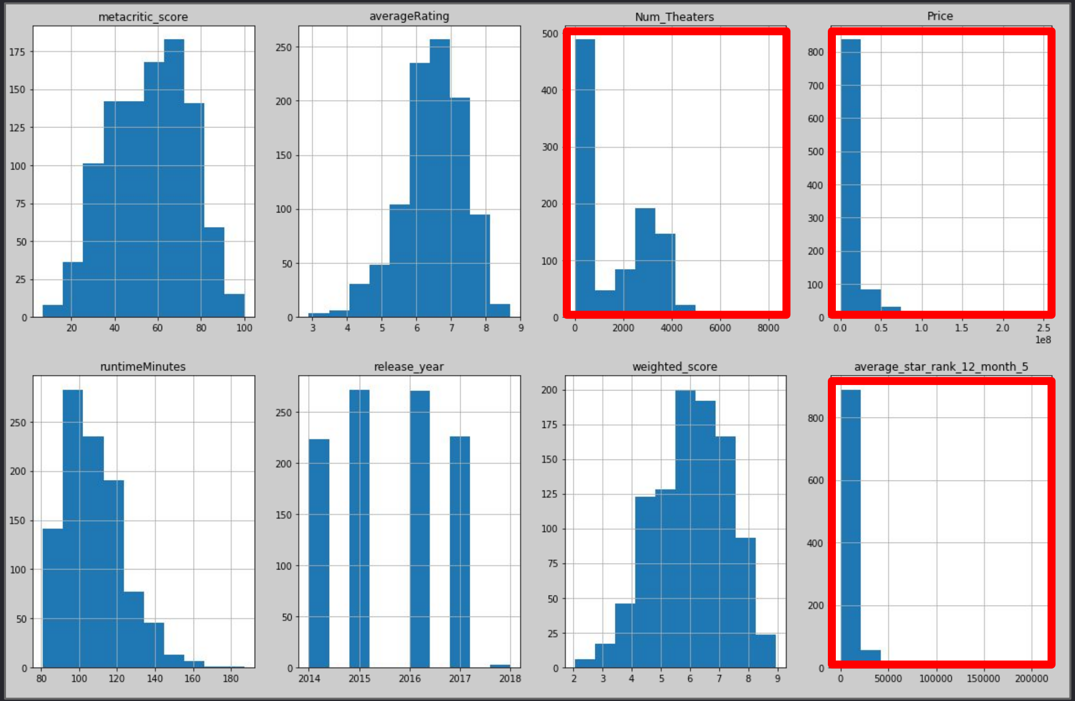


- Number of theaters is highly correlated with overall opening weekend gross

4
Data Analysis

Exploring the Data

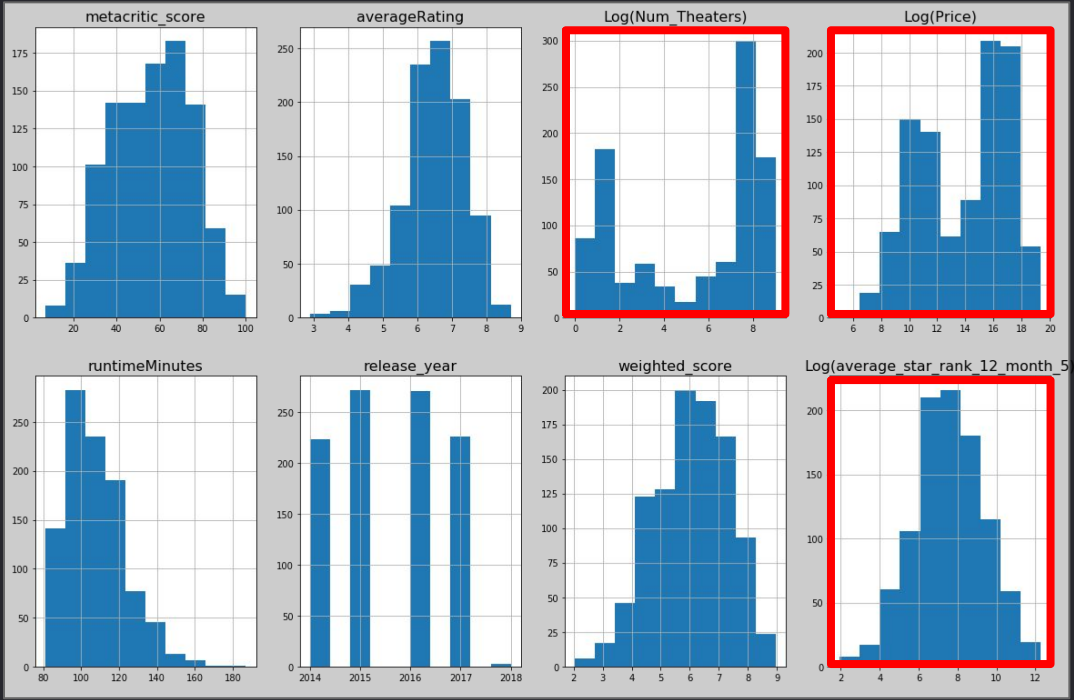
Continuous Features



- All continuous features should have normal distribution.

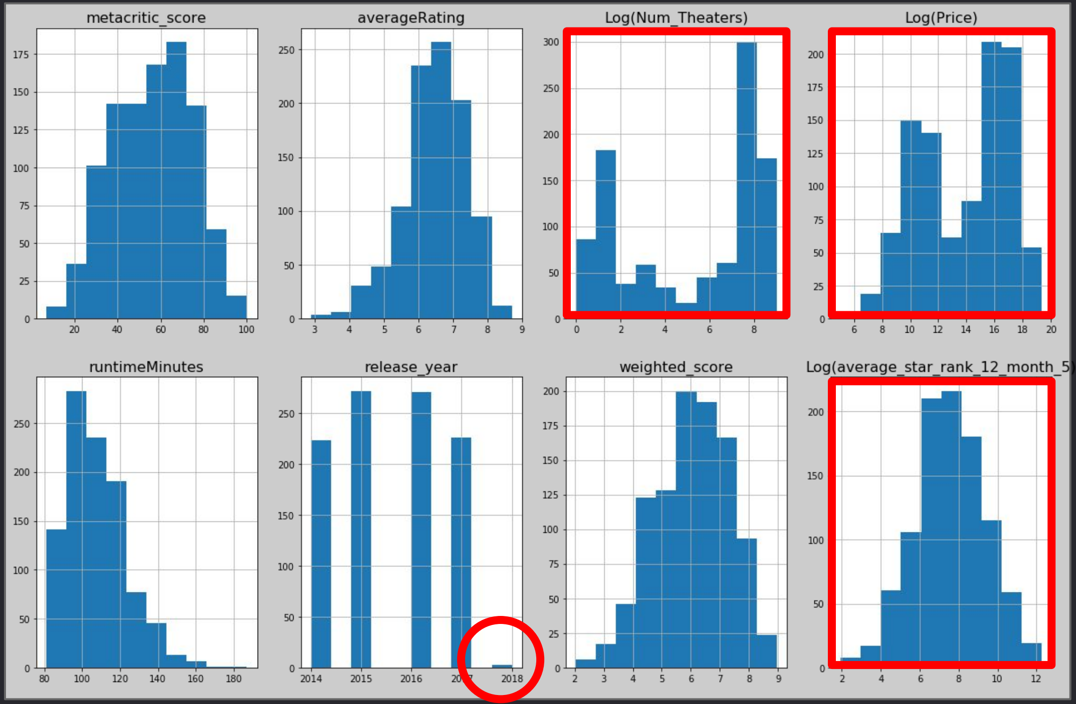
- Issues:
  - Opening Weekend Gross
  - Number of Theaters
  - Star Power

Continuous Features



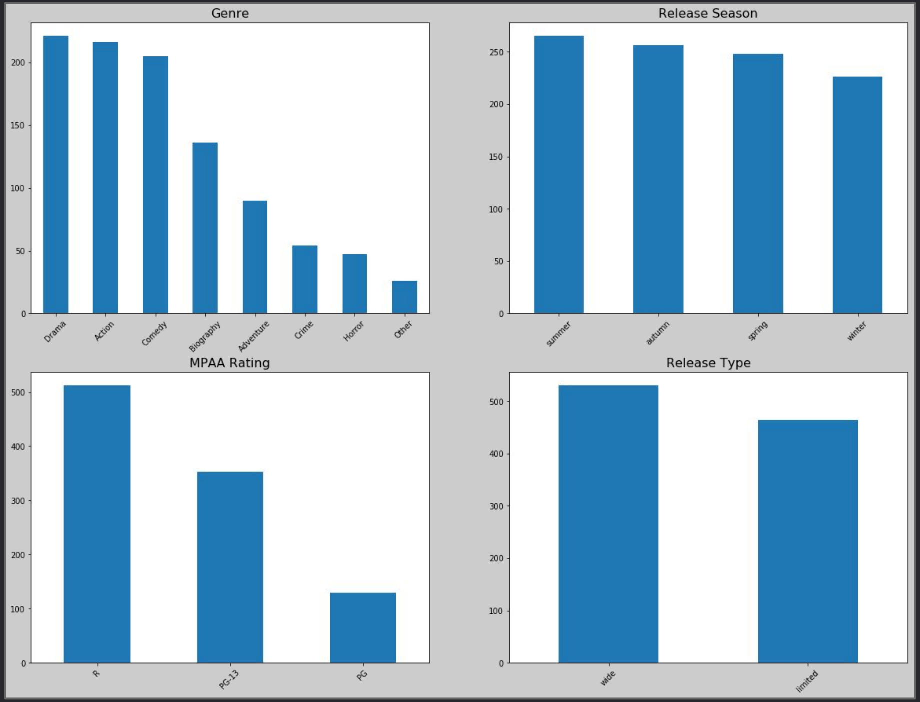
- All continuous features should have normal distribution.
- Logarithmic transform applied to skewed data.

Continuous Features



- All continuous features should have normal distribution.
- Logarithmic transform applied to skewed data.
- Removed all 2018 data points

Categorical Features



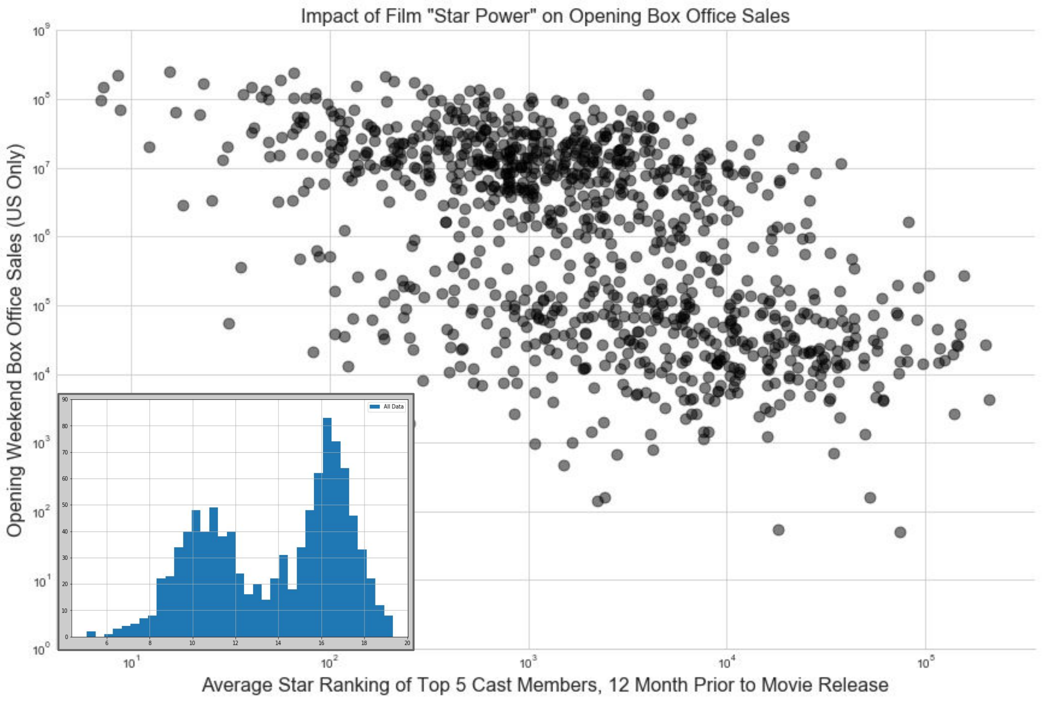
- Movie Genres
  (Action, Drama, Comedy, etc.)
- Release Season
  (Spring, Summer, Autumn, Winter)
- MPAA Rating
  (PG, PG-13, R)
- Film Release Type
  (Wide vs. Limited)

5
Feature Engineering

Modifying Model Features

Star Power

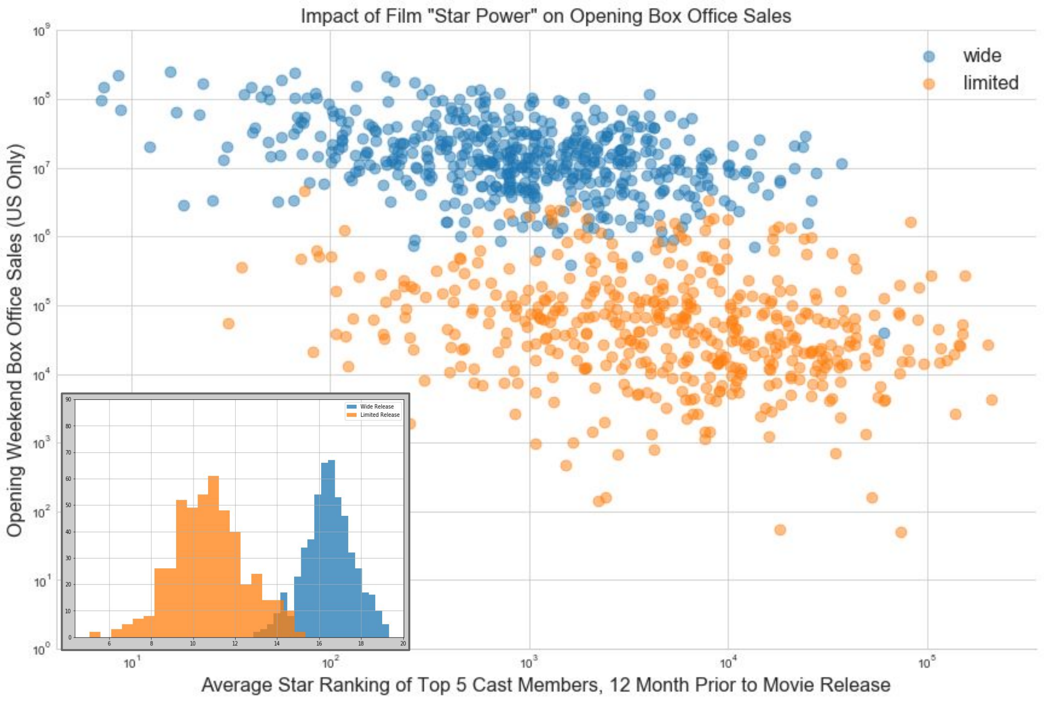
- Opening Weekend Gross vs. Star Ranking alone does not show a clear linear relationship.



Star Power

- Opening Weekend Gross vs. Star Ranking alone does not show a clear linear relationship.

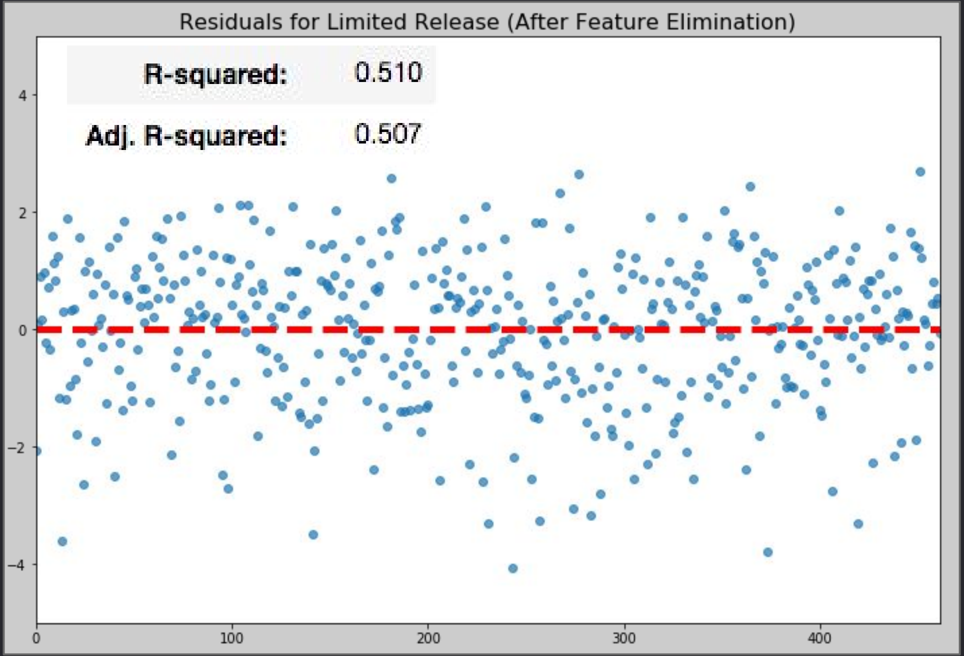
- “Limited Release”: # Theaters < 600



6
Modeling Box Office Sales

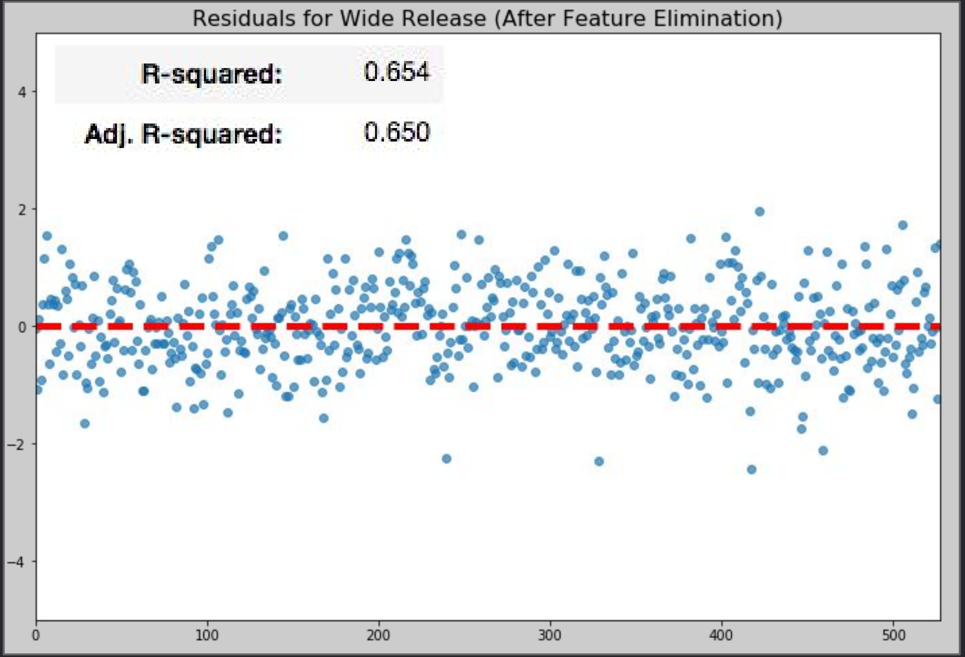
Final Model Results

Limited Release - Feature Selection



- Model trained on all data.
- Features selected based on individual p-values.
- Remaining features:
  - Num_Theaters
  - Runtime_Minutes
  - Star_Power

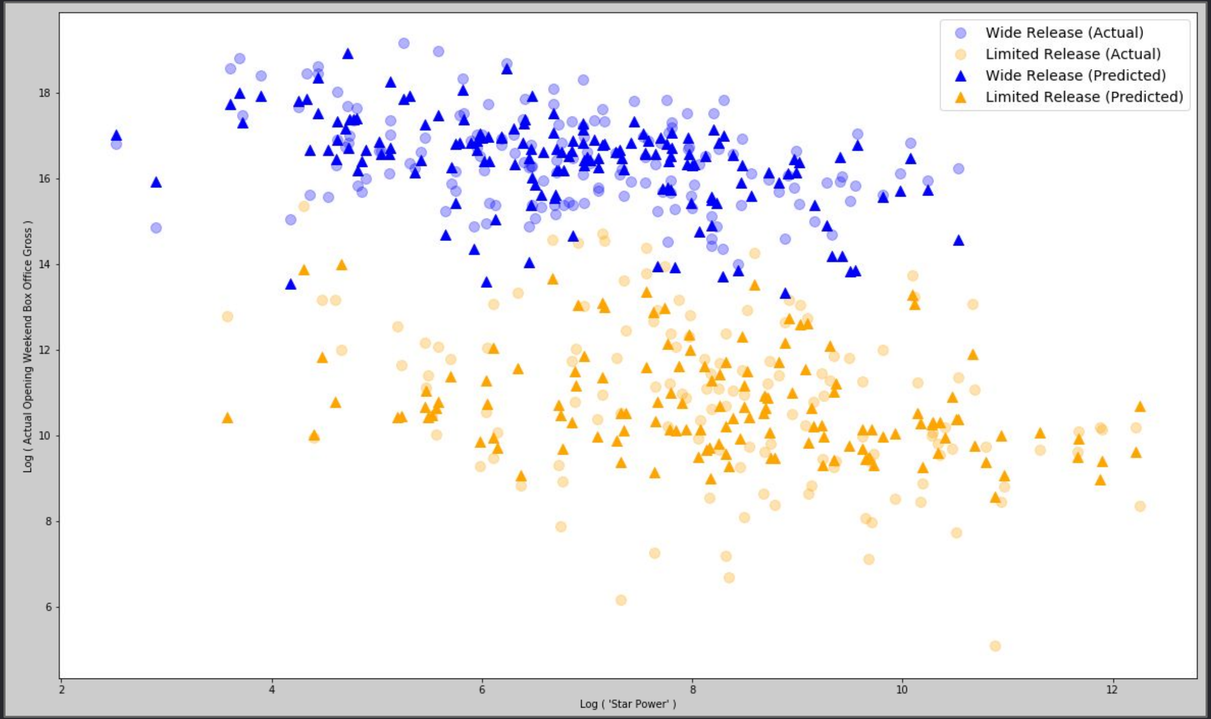
Wide Release - Feature Selection



- Model trained on all data.
- Features selected based on individual p-values.
- Remaining features:
  - Num_Theaters
  - Runtime_Minutes
  - Star_Power
  - Release_Year
  - Rated_R
  - Release_in_Spring

Predicted vs. Actual Opening Weekend Gross

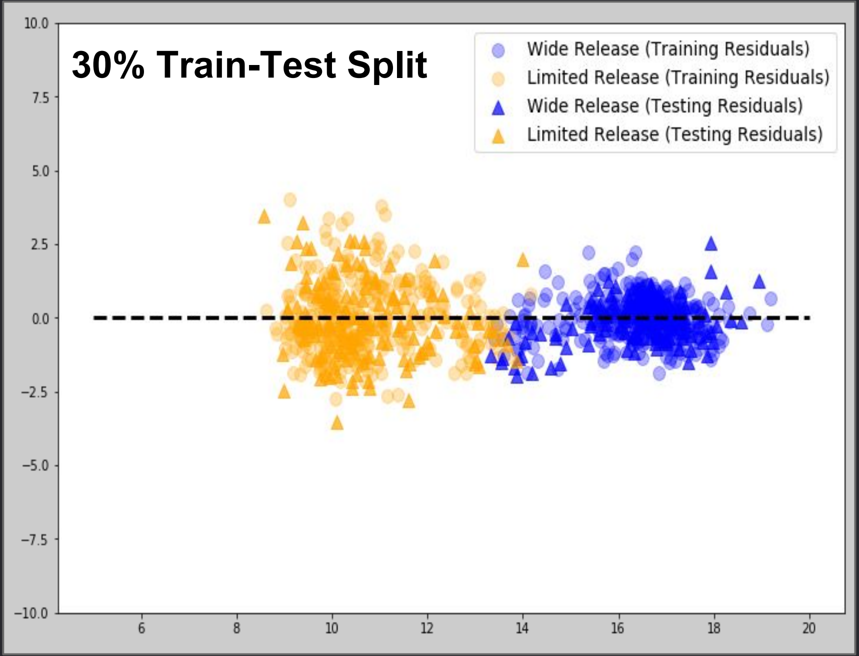
- 70-30 random train-test split
- Test data shown
- Predicted results within expected range



Residuals for Train/Test Split

- Model trained 70% of data, and validated on remaining 30%.
- Cross-Validation score aligns with observed results.

Regression Model	5-Fold Cross-Validation R^2 Score
All Data (Base)	0.16
Limited Release	0.49
Wide Release	0.64



Applying Predictive Model

- 2018 Opening Weekend Box Office Predictions

Movie Title	Release Date	Predicted Gross	Actual Gross	Release Type	% Error
The Post	2018-01-12	$19,783,540	$19,887,979	wide	-0.53 %
Paddington 2	2018-01-12	$18,109,110	$11,001,961	wide	64.60 %

7
Conclusion

And Next Steps

Conclusions

Results

- "Star Power" is a factor in predicting opening weekend box office performance.
- Splitting data into two models proved effective.
- Genre removed as feature

Next Steps

- New features to consider:
  - Box office competition
  - Is Sequel?
    (Title/Brand recognition)
  - Other metrics for cast / crew popularity
    (Google Trends, Twitter)

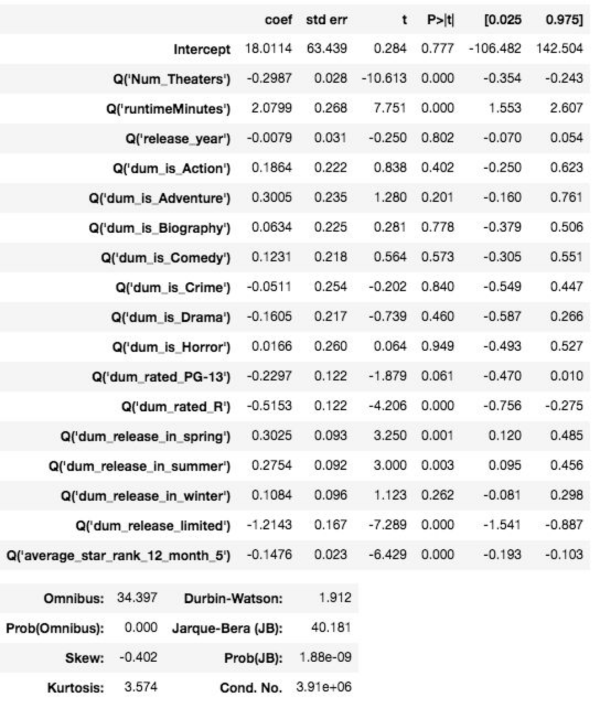
THANKS!

ANY QUESTIONS?

Scraping IMDB Pro’s StarMeter

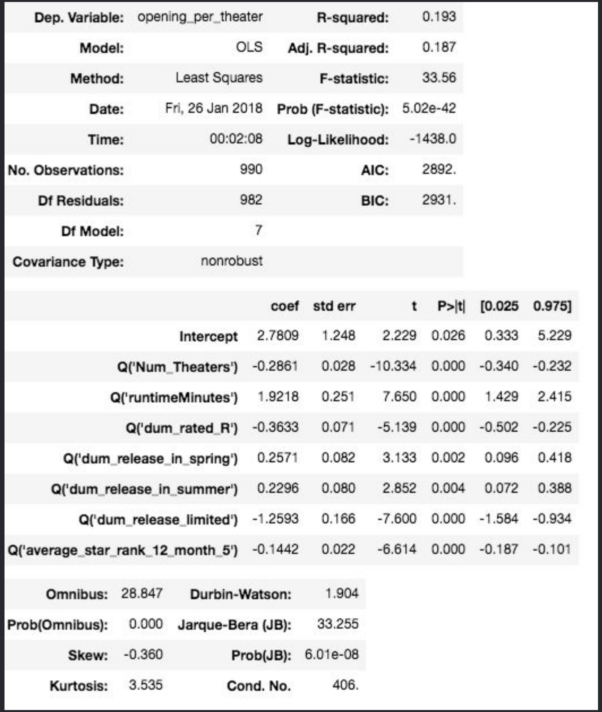
- Visualization uses JavaScript SVG Element
- Scraping using Selenium

All Data - Feature Elimination



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	18.0114	63.439	0.284	0.777	-106.482	142.504
Q('Num_Theaters')	-0.2987	0.028	-10.613	0.000	-0.354	-0.243
Q('runtimeMinutes')	2.0799	0.268	7.751	0.000	1.553	2.607
Q('release_year')	-0.0079	0.031	-0.250	0.802	-0.070	0.054
Q('dum_is_Action')	0.1864	0.222	0.838	0.402	-0.250	0.623
Q('dum_is_Adventure')	0.3005	0.235	1.280	0.201	-0.160	0.761
Q('dum_is_Biography')	0.0634	0.225	0.281	0.778	-0.379	0.506
Q('dum_is_Comedy')	0.1231	0.218	0.564	0.573	-0.305	0.551
Q('dum_is_Crime')	-0.0511	0.254	-0.202	0.840	-0.549	0.447
Q('dum_is_Drama')	-0.1605	0.217	-0.739	0.460	-0.587	0.266
Q('dum_is_Horror')	0.0166	0.260	0.064	0.949	-0.493	0.527
Q('dum_rated_PG-13')	-0.2297	0.122	-1.879	0.061	-0.470	0.010
Q('dum_rated_R')	-0.5153	0.122	-4.206	0.000	-0.756	-0.275
Q('dum_release_in_spring')	0.3025	0.093	3.250	0.001	0.120	0.485
Q('dum_release_in_summer')	0.2754	0.092	3.000	0.003	0.095	0.456
Q('dum_release_in_winter')	0.1084	0.096	1.123	0.262	-0.081	0.298
Q('dum_release_limited')	-1.2143	0.167	-7.289	0.000	-1.541	-0.887
Q('average_star_rank_12_month_5')	-0.1476	0.023	-6.429	0.000	-0.193	-0.103

Omnibus:	34.397	Durbin-Watson:	1.912
Prob(Omnibus):	0.000	Jarque-Bera (JB):	40.181
Skew:	-0.402	Prob(JB):	1.88e-09
Kurtosis:	3.574	Cond. No.	3.91e+06

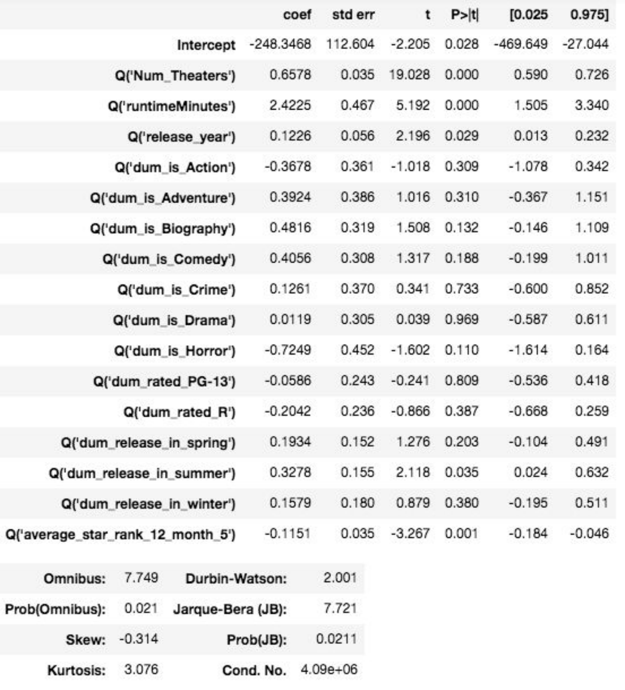


Dep. Variable:	opening_per_theater	R-squared:	0.193
Model:	OLS	Adj. R-squared:	0.187
Method:	Least Squares	F-statistic:	33.56
Date:	Fri, 26 Jan 2018	Prob (F-statistic):	5.02e-42
Time:	00:02:08	Log-Likelihood:	-1438.0
No. Observations:	990	AIC:	2892.
Df Residuals:	982	BIC:	2931.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.7809	1.248	2.229	0.026	0.333	5.229
Q('Num_Theaters')	-0.2861	0.028	-10.334	0.000	-0.340	-0.232
Q('runtimeMinutes')	1.9218	0.251	7.650	0.000	1.429	2.415
Q('dum_rated_R')	-0.3633	0.071	-5.139	0.000	-0.502	-0.225
Q('dum_release_in_spring')	0.2571	0.082	3.133	0.002	0.096	0.418
Q('dum_release_in_summer')	0.2296	0.080	2.852	0.004	0.072	0.388
Q('dum_release_limited')	-1.2593	0.166	-7.600	0.000	-1.584	-0.934
Q('average_star_rank_12_month_5')	-0.1442	0.022	-6.614	0.000	-0.187	-0.101

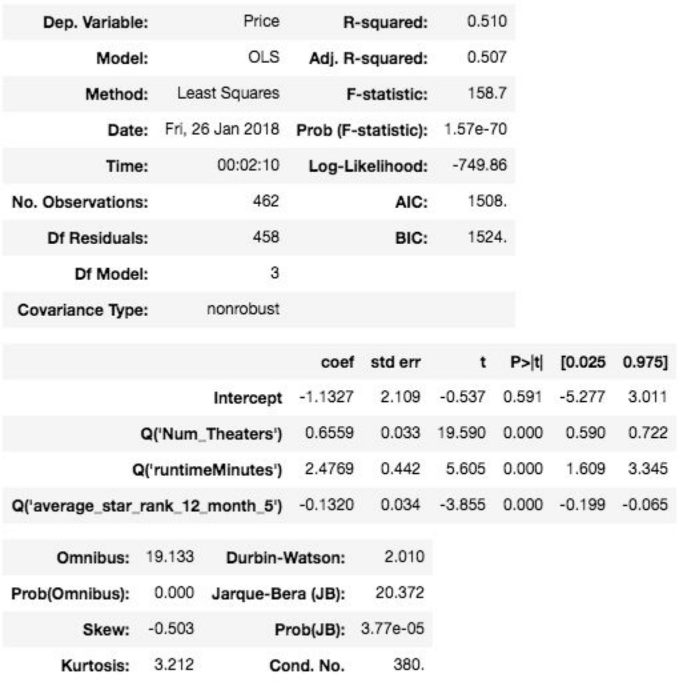
Omnibus:	28.847	Durbin-Watson:	1.904
Prob(Omnibus):	0.000	Jarque-Bera (JB):	33.255
Skew:	-0.360	Prob(JB):	6.01e-08
Kurtosis:	3.535	Cond. No.	406.

Limited Release - Feature Elimination



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-248.3468	112.604	-2.205	0.028	-469.649	-27.044
Q('Num_Theaters')	0.6578	0.035	19.028	0.000	0.590	0.726
Q('runtimeMinutes')	2.4225	0.467	5.192	0.000	1.505	3.340
Q('release_year')	0.1226	0.056	2.196	0.029	0.013	0.232
Q('dum_is_Action')	-0.3678	0.361	-1.018	0.309	-1.078	0.342
Q('dum_is_Adventure')	0.3924	0.386	1.016	0.310	-0.367	1.151
Q('dum_is_Biography')	0.4816	0.319	1.508	0.132	-0.146	1.109
Q('dum_is_Comedy')	0.4056	0.308	1.317	0.188	-0.199	1.011
Q('dum_is_Crime')	0.1261	0.370	0.341	0.733	-0.600	0.852
Q('dum_is_Drama')	0.0119	0.305	0.039	0.969	-0.587	0.611
Q('dum_is_Horror')	-0.7249	0.452	-1.602	0.110	-1.614	0.164
Q('dum_rated_PG-13')	-0.0586	0.243	-0.241	0.809	-0.536	0.418
Q('dum_rated_R')	-0.2042	0.236	-0.866	0.387	-0.668	0.259
Q('dum_release_in_spring')	0.1934	0.152	1.276	0.203	-0.104	0.491
Q('dum_release_in_summer')	0.3278	0.155	2.118	0.035	0.024	0.632
Q('dum_release_in_winter')	0.1579	0.180	0.879	0.380	-0.195	0.511
Q('average_star_rank_12_month_5')	-0.1151	0.035	-3.267	0.001	-0.184	-0.046

Omnibus:	7.749	Durbin-Watson:	2.001
Prob(Omnibus):	0.021	Jarque-Bera (JB):	7.721
Skew:	-0.314	Prob(JB):	0.0211
Kurtosis:	3.076	Cond. No.	4.09e+06

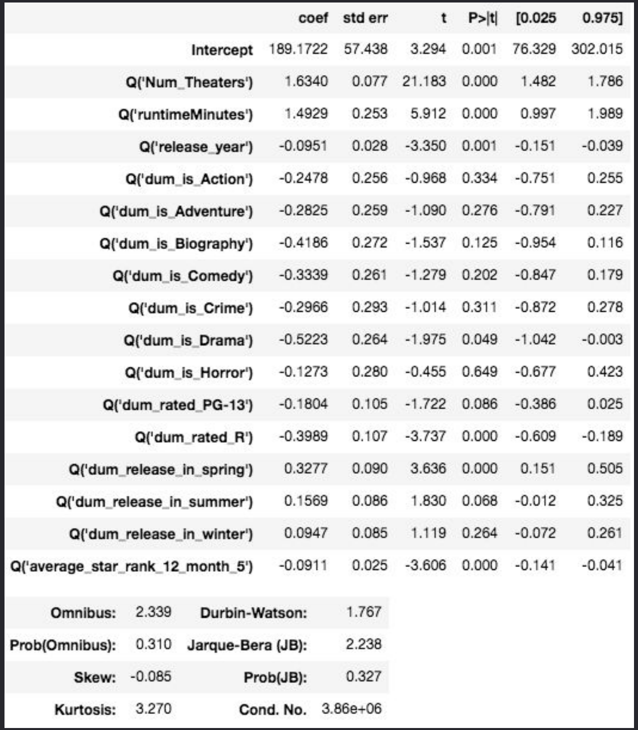


Dep. Variable:	Price	R-squared:	0.510
Model:	OLS	Adj. R-squared:	0.507
Method:	Least Squares	F-statistic:	158.7
Date:	Fri, 26 Jan 2018	Prob (F-statistic):	1.57e-70
Time:	00:02:10	Log-Likelihood:	-749.86
No. Observations:	462	AIC:	1508.
Df Residuals:	458	BIC:	1524.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.1327	2.109	-0.537	0.591	-5.277	3.011
Q('Num_Theaters')	0.6559	0.033	19.590	0.000	0.590	0.722
Q('runtimeMinutes')	2.4769	0.442	5.605	0.000	1.609	3.345
Q('average_star_rank_12_month_5')	-0.1320	0.034	-3.855	0.000	-0.199	-0.065

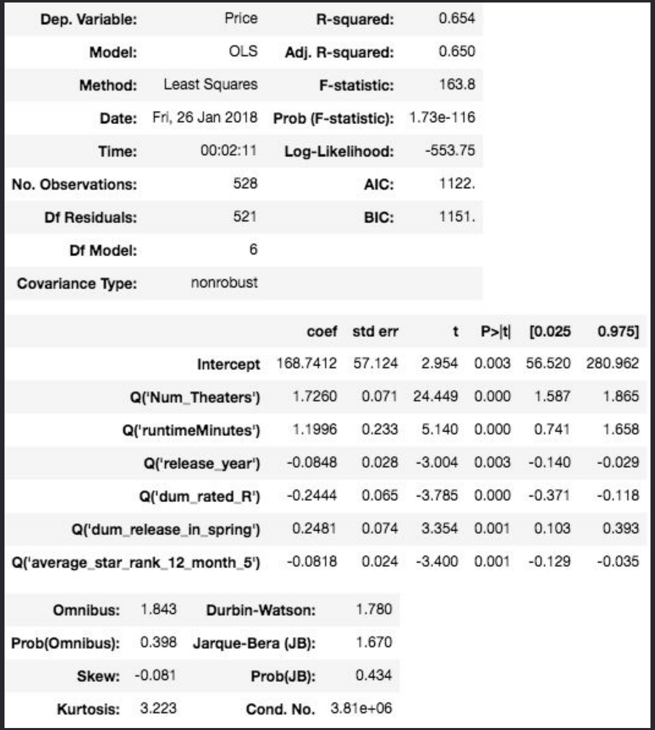
Omnibus:	19.133	Durbin-Watson:	2.010
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20.372
Skew:	-0.503	Prob(JB):	3.77e-05
Kurtosis:	3.212	Cond. No.	380.

Wide Release - Feature Elimination



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	189.1722	57.438	3.294	0.001	76.329	302.015
Q('Num_Theaters')	1.6340	0.077	21.183	0.000	1.482	1.786
Q('runtimeMinutes')	1.4929	0.253	5.912	0.000	0.997	1.989
Q('release_year')	-0.0951	0.028	-3.350	0.001	-0.151	-0.039
Q('dum_is_Action')	-0.2478	0.256	-0.968	0.334	-0.751	0.255
Q('dum_is_Adventure')	-0.2825	0.259	-1.090	0.276	-0.791	0.227
Q('dum_is_Biography')	-0.4186	0.272	-1.537	0.125	-0.954	0.116
Q('dum_is_Comedy')	-0.3339	0.261	-1.279	0.202	-0.847	0.179
Q('dum_is_Crime')	-0.2966	0.293	-1.014	0.311	-0.872	0.278
Q('dum_is_Drama')	-0.5223	0.264	-1.975	0.049	-1.042	-0.003
Q('dum_is_Horror')	-0.1273	0.280	-0.455	0.649	-0.677	0.423
Q('dum_rated_PG-13')	-0.1804	0.105	-1.722	0.086	-0.386	0.025
Q('dum_rated_R')	-0.3989	0.107	-3.737	0.000	-0.609	-0.189
Q('dum_release_in_spring')	0.3277	0.090	3.636	0.000	0.151	0.505
Q('dum_release_in_summer')	0.1569	0.086	1.830	0.068	-0.012	0.325
Q('dum_release_in_winter')	0.0947	0.085	1.119	0.264	-0.072	0.261
Q('average_star_rank_12_month_5')	-0.0911	0.025	-3.606	0.000	-0.141	-0.041

Omnibus:	2.339	Durbin-Watson:	1.767
Prob(Omnibus):	0.310	Jarque-Bera (JB):	2.238
Skew:	-0.085	Prob(JB):	0.327
Kurtosis:	3.270	Cond. No.	3.86e+06



Dep. Variable:	Price	R-squared:	0.654
Model:	OLS	Adj. R-squared:	0.650
Method:	Least Squares	F-statistic:	163.8
Date:	Fri, 26 Jan 2018	Prob (F-statistic):	1.73e-116
Time:	00:02:11	Log-Likelihood:	-553.75
No. Observations:	528	AIC:	1122.
Df Residuals:	521	BIC:	1151.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	168.7412	57.124	2.954	0.003	56.520	280.962
Q('Num_Theaters')	1.7260	0.071	24.449	0.000	1.587	1.865
Q('runtimeMinutes')	1.1996	0.233	5.140	0.000	0.741	1.658
Q('release_year')	-0.0848	0.028	-3.004	0.003	-0.140	-0.029
Q('dum_rated_R')	-0.2444	0.065	-3.785	0.000	-0.371	-0.118
Q('dum_release_in_spring')	0.2481	0.074	3.354	0.001	0.103	0.393
Q('average_star_rank_12_month_5')	-0.0818	0.024	-3.400	0.001	-0.129	-0.035

Omnibus:	1.843	Durbin-Watson:	1.780
Prob(Omnibus):	0.398	Jarque-Bera (JB):	1.670
Skew:	-0.081	Prob(JB):	0.434
Kurtosis:	3.223	Cond. No.	3.81e+06