Predicting A Movie’s Total Foreign Gross

Christine Maroti

Two Questions...

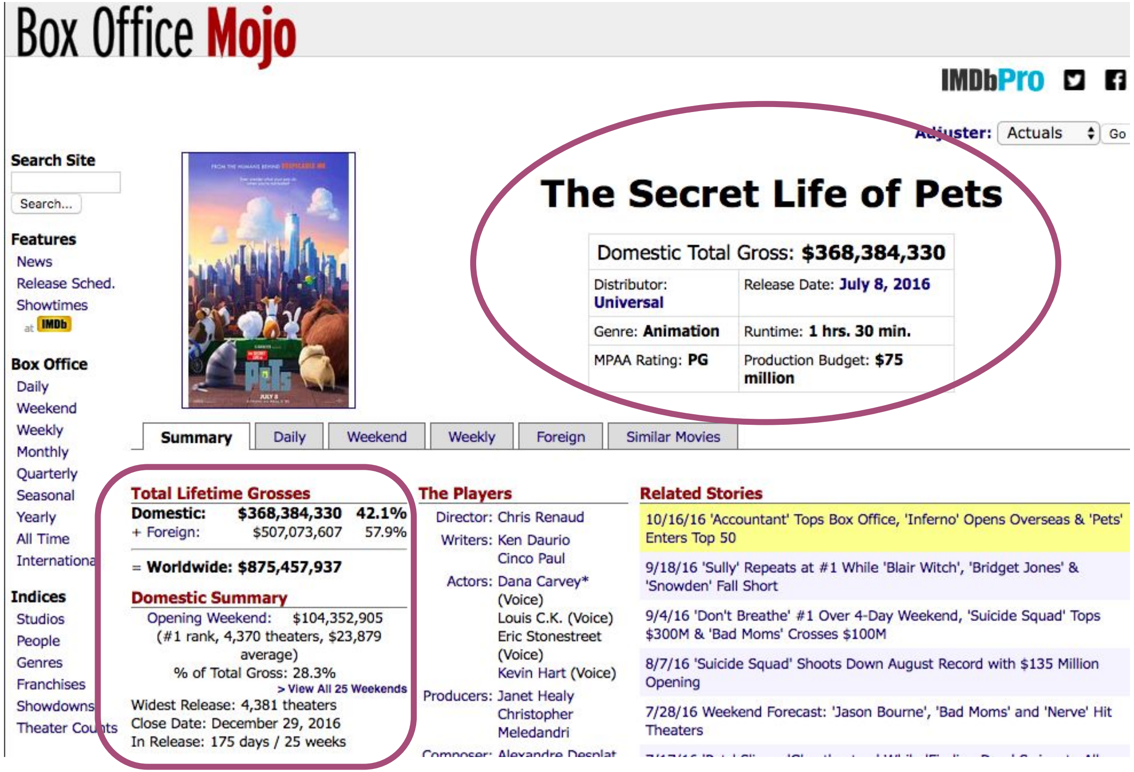
1. Can we predict foreign box office numbers?

2. Can we tell which movie genres are more popular in other countries?

Data

Web Scraping: BoxOfficeMojo

- Title
- Domestic Total Gross
- Distributor
- Release Date
- Genre
- Runtime
- Budget
- Total Foreign Gross
- Opening Weekend
- Widest Release
- Days In Theaters



Box Office Mojo

IMDbPro

Adjuster: Actuals Go

Search Site

Search...

Features
News
Release Sched.
Showtimes at IMDb

Box Office
Daily
Weekend
Weekly
Monthly
Quarterly
Seasonal
Yearly
All Time
International

Indices
Studios
People
Genres
Franchises
Showdowns
Theater Counts

The Secret Life of Pets

Domestic Total Gross: **$368,384,330**

Distributor: **Universal**	Release Date: **July 8, 2016**
Genre: **Animation**	Runtime: **1 hrs. 30 min.**
MPAA Rating: **PG**	Production Budget: **$75 million**

Summary | Daily | Weekend | Weekly | Foreign | Similar Movies

Total Lifetime Grosses

Domestic:	**$368,384,330**	**42.1%**
+ Foreign:	$507,073,607	57.9%

= **Worldwide: $875,457,937**

Domestic Summary

Opening Weekend: $104,352,905
(#1 rank, 4,370 theaters, $23,879 average)
% of Total Gross: 28.3%
> View All 25 Weekends

Widest Release: 4,381 theaters
Close Date: December 29, 2016
In Release: 175 days / 25 weeks

The Players

Director: Chris Renaud
Writers: Ken Daurio
Cinco Paul
Actors: Dana Carvey* (Voice)
Louis C.K. (Voice)
Eric Stonestreet (Voice)
Kevin Hart (Voice)
Producers: Janet Healy
Christopher Meledandri
Composer: Alexandre Desplat

Related Stories

10/16/16 'Accountant' Tops Box Office, 'Inferno' Opens Overseas & 'Pets' Enters Top 50

9/18/16 'Sully' Repeats at #1 While 'Blair Witch', 'Bridget Jones' & 'Snowden' Fall Short

9/4/16 'Don't Breathe' #1 Over 4-Day Weekend, 'Suicide Squad' Tops $300M & 'Bad Moms' Crosses $100M

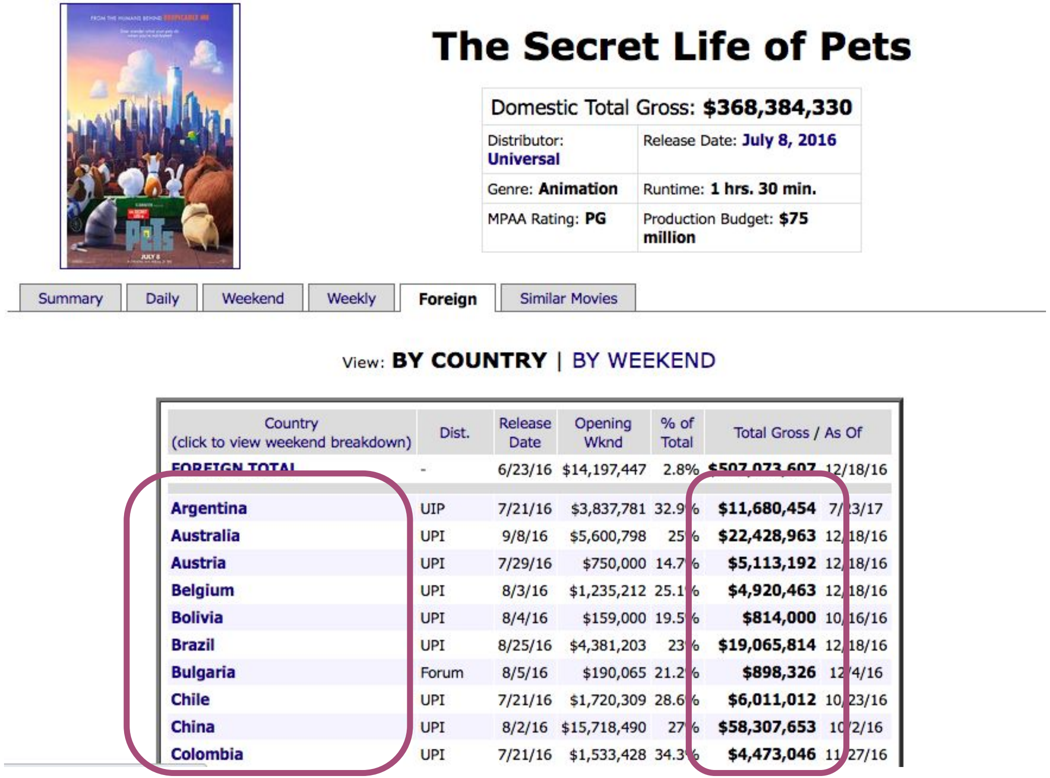
8/7/16 'Suicide Squad' Shoots Down August Record with $135 Million Opening

7/28/16 Weekend Forecast: 'Jason Bourne', 'Bad Moms' and 'Nerve' Hit Theaters

Web Scraping: BoxOfficeMojo

- Countries
- Gross for Each Country

- Scraped top 200 grossing movies (domestically) for 2010-2016 (7 years)



The Secret Life of Pets

Domestic Total Gross: **$368,384,330**	
Distributor: **Universal**	Release Date: **July 8, 2016**
Genre: **Animation**	Runtime: **1 hrs. 30 min.**
MPAA Rating: **PG**	Production Budget: **$75 million**

View: **BY COUNTRY** | BY WEEKEND

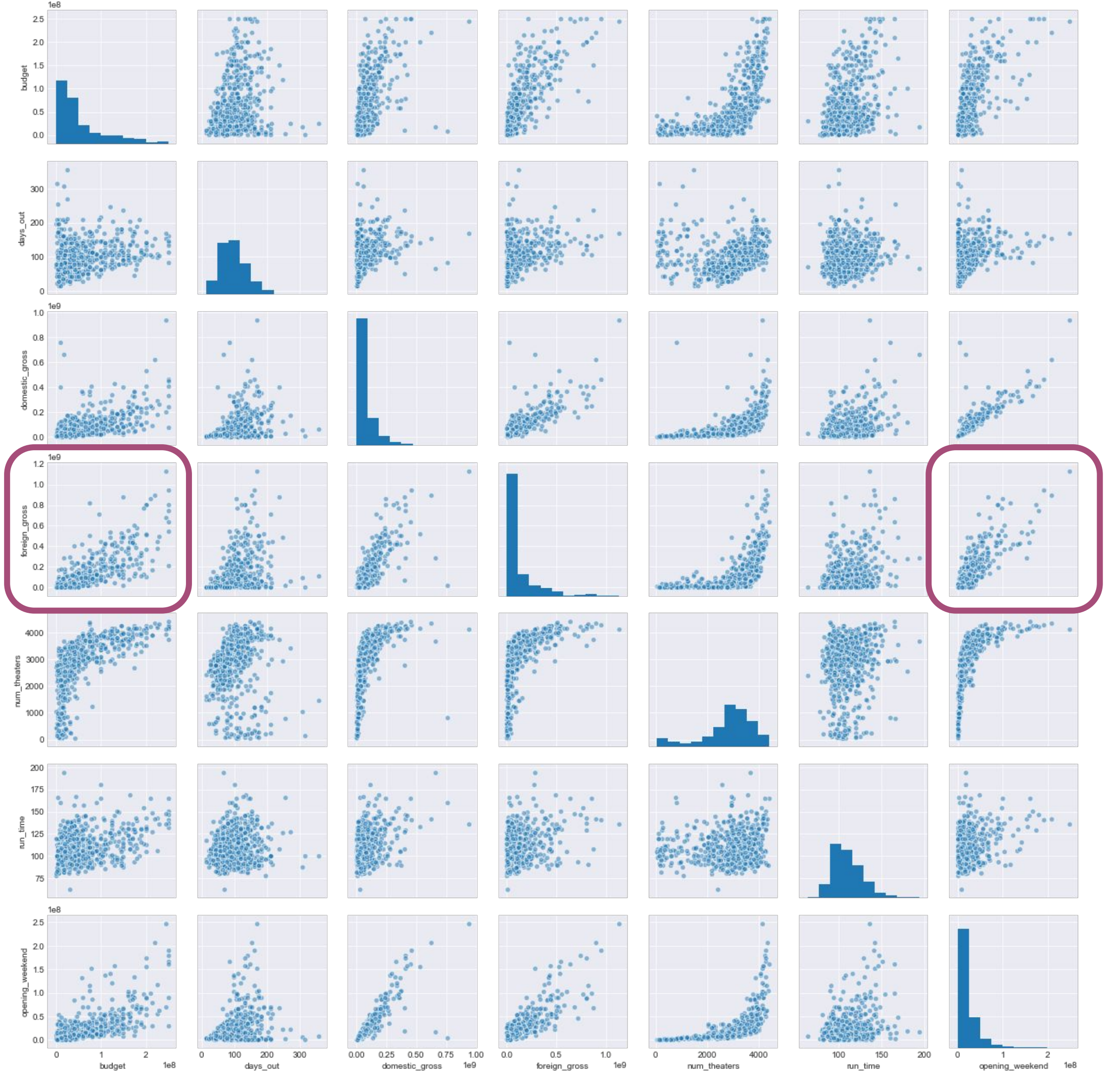
Country (click to view weekend breakdown)	Dist.	Release Date	Opening Wknd	% of Total	Total Gross / As Of	
FOREIGN TOTAL	-	6/23/16	$14,197,447	2.8%	**$507,073,607**	12/18/16
Argentina	UIP	7/21/16	$3,837,781	32.9%	**$11,680,454**	7/23/17
Australia	UPI	9/8/16	$5,600,798	25%	**$22,428,963**	12/18/16
Austria	UPI	7/29/16	$750,000	14.7%	**$5,113,192**	12/18/16
Belgium	UPI	8/3/16	$1,235,212	25.1%	**$4,920,463**	12/18/16
Bolivia	UPI	8/4/16	$159,000	19.5%	**$814,000**	10/16/16
Brazil	UPI	8/25/16	$4,381,203	23%	**$19,065,814**	12/18/16
Bulgaria	Forum	8/5/16	$190,065	21.2%	**$898,326**	12/4/16
Chile	UPI	7/21/16	$1,720,309	28.6%	**$6,011,012**	10/23/16
China	UPI	8/2/16	$15,718,490	27%	**$58,307,653**	10/2/16
Colombia	UPI	7/21/16	$1,533,428	34.3%	**$4,473,046**	11/27/16

Exploratory Analysis

Pairplot of Variables

- Not very normally distributed
- Remove days out, domestic gross, number of theaters, and runtime

- (See appendix for log analysis)



Add Categorical Variables

Genre (8)

Distributor (9)

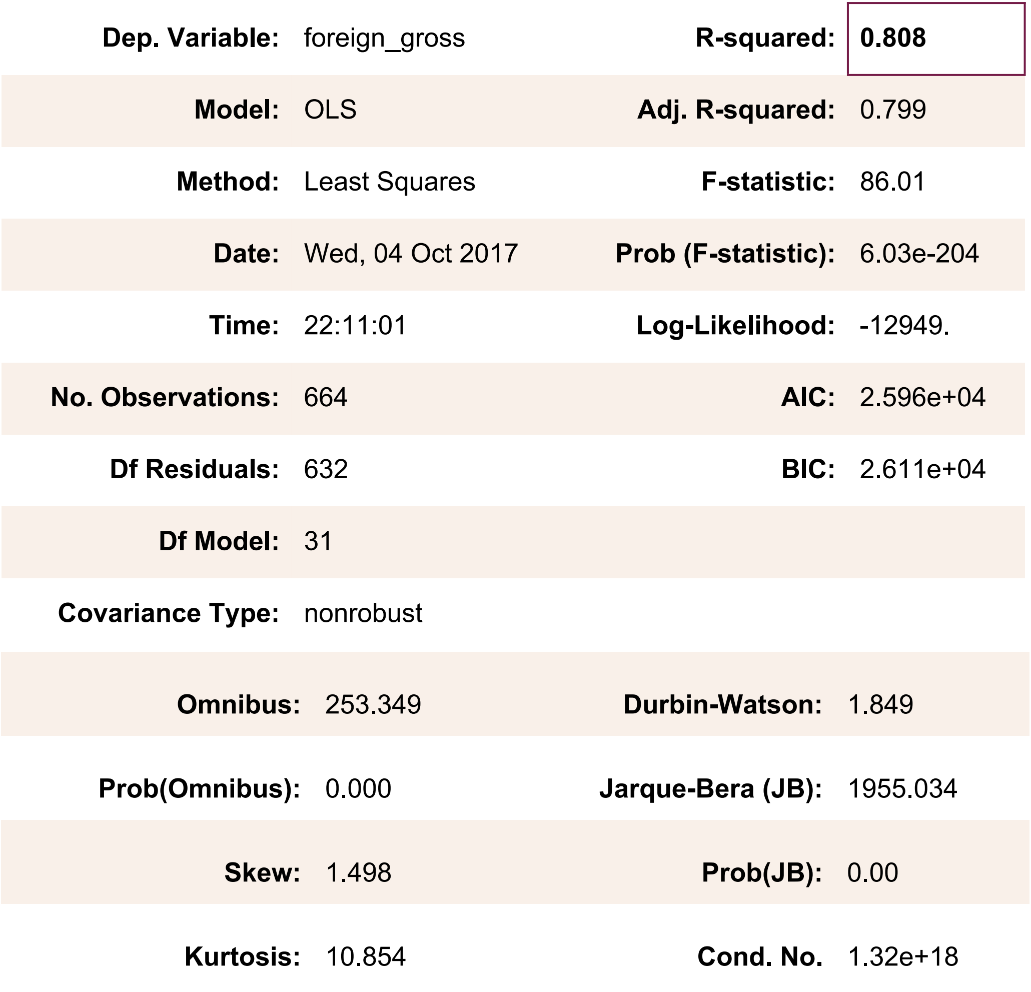
MPAA Rating (4)

Month of Release (12)

First Model

Ordinary Least Squares Regression

R^2 of 0.808



Dep. Variable:	foreign_gross	**R-squared:**	**0.808**
Model:	OLS	**Adj. R-squared:**	0.799
Method:	Least Squares	**F-statistic:**	86.01
Date:	Wed, 04 Oct 2017	**Prob (F-statistic):**	6.03e-204
Time:	22:11:01	**Log-Likelihood:**	-12949.
No. Observations:	664	**AIC:**	2.596e+04
Df Residuals:	632	**BIC:**	2.611e+04
Df Model:	31		
Covariance Type:	nonrobust		

Omnibus:	253.349	**Durbin-Watson:**	1.849
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1955.034
Skew:	1.498	**Prob(JB):**	0.00
Kurtosis:	10.854	**Cond. No.**	1.32e+18

Ordinary Least Squares Regression

Some high p-values for certain categories

	coef	P>|t|
const	-1.617e+07	0.004
budget	1.0306	0.000
opening_weekend	3.1082	0.000
genre[T.Comedy]	-2.057e+07	0.013
genre[T.Documentary]	**-1.417e+07**	**0.688**
genre[T.Drama]	-9.037e+06	0.289
genre[T.Family]	3.304e+07	0.008
genre[T.Horror]	**-2.208e+06**	**0.860**
genre[T.Other]	**5.213e+06**	**0.794**
genre[T.Thriller]	-1.202e+07	0.318
genre[T.Act-Adventure]	**3.581e+06**	**0.709**
rating[T.PG]	**2.643e+06**	**0.765**
rating[T.PG-13]	**1.731e+05**	**0.984**
rating[T.R]	8.125e+06	0.370
rating[T.G]	-2.711e+07	0.227
distributor[T.Fox]	2.21e+07	0.004
distributor[T.Independ.]	**-4.182e+06**	**0.600**
distributor[T.Lionsgate]	-1.785e+07	0.052
distributor[T.Paramount]	1.452e+07	0.112

	coef	P>|t|
distributor[T.Sony]	**-3.421e+06**	**0.643**
distributor[T.Universal]	-6.669e+06	0.378
distributor[T.Warner Bros]	-2.395e+07	0.002
distributor[T.Weinstein]	1.117e+07	0.435
distributor[T.Disney]	-7.89e+06	0.457
month[T.August]	-1.533e+07	0.099
month[T.December]	7.465e+06	0.408
month[T.February]	-1.277e+07	0.211
month[T.January]	-8.434e+06	0.423
month[T.July]	1.6e+07	0.088
month[T.June]	-7.146e+06	0.487
month[T.March]	-3.328e+07	0.001
month[T.May]	1.604e+07	0.142
month[T.November]	1.822e+07	0.048
month[T.October]	7.219e+06	0.429
month[T.September]	**-5.835e+06**	**0.532**
month[T.April]	**1.684e+06**	**0.867**

Feature Selection

Cross Validation

- K-folds cross validation with 10 folds
  - Selecting significant features (p-value < 0.05) for each fold
  - Ran OLS regression on each fold on only the selected features

- **Mean R^2 of 0.756**

Eliminate Features

- Checked adjusted R^2 eliminating different combinations of Rating, Distributor, and Month
  - Ranged from 0.788 to 0.799

- Decided to keep only Budget, Opening Weekend, and Genre as features

Final Model

Can we use a movie’s budget, opening weekend, and genre to predict its foreign gross?

- R^2 of 0.773 (cross-validated)
- Intercept: -$20 million
- Coefficients:
  - **Budget: 1.067**
  - **Opening weekend: 3.093**
  - **Comedy: -$16,414,922**
  - Documentary: -$16,281,281
  - Drama: -$5,895,403
  - **Family: $37,196,306**
  - Horror: $2,471,289
  - Other: $8,575,768
  - Thriller: -$10,026,560
  - Action-Adventure: $374,803

Most Underpredicted Foreign Grosses

Title	Genre	Budget	Opening Weekend	Foreign Gross	Residual
Frozen	Family	$150,000,000	$67,391,326	$875,742,326	$490,294,500
Ice Age: Continental Drift	Family	$95,000,000	$46,629,259	$715,922,939	$453,389,400
Minions	Family	$74,000,000	$115,718,405	$823,352,627	$369,554,300
Transformers: Age of Extinction	Action-Adventure	$210,000,000	$100,038,390	$858,614,996	$344,980,000
Skyfall	Action-Adventure	$200,000,000	$88,364,714	$804,200,736	$337,343,000

Most Overpredicted Foreign Grosses

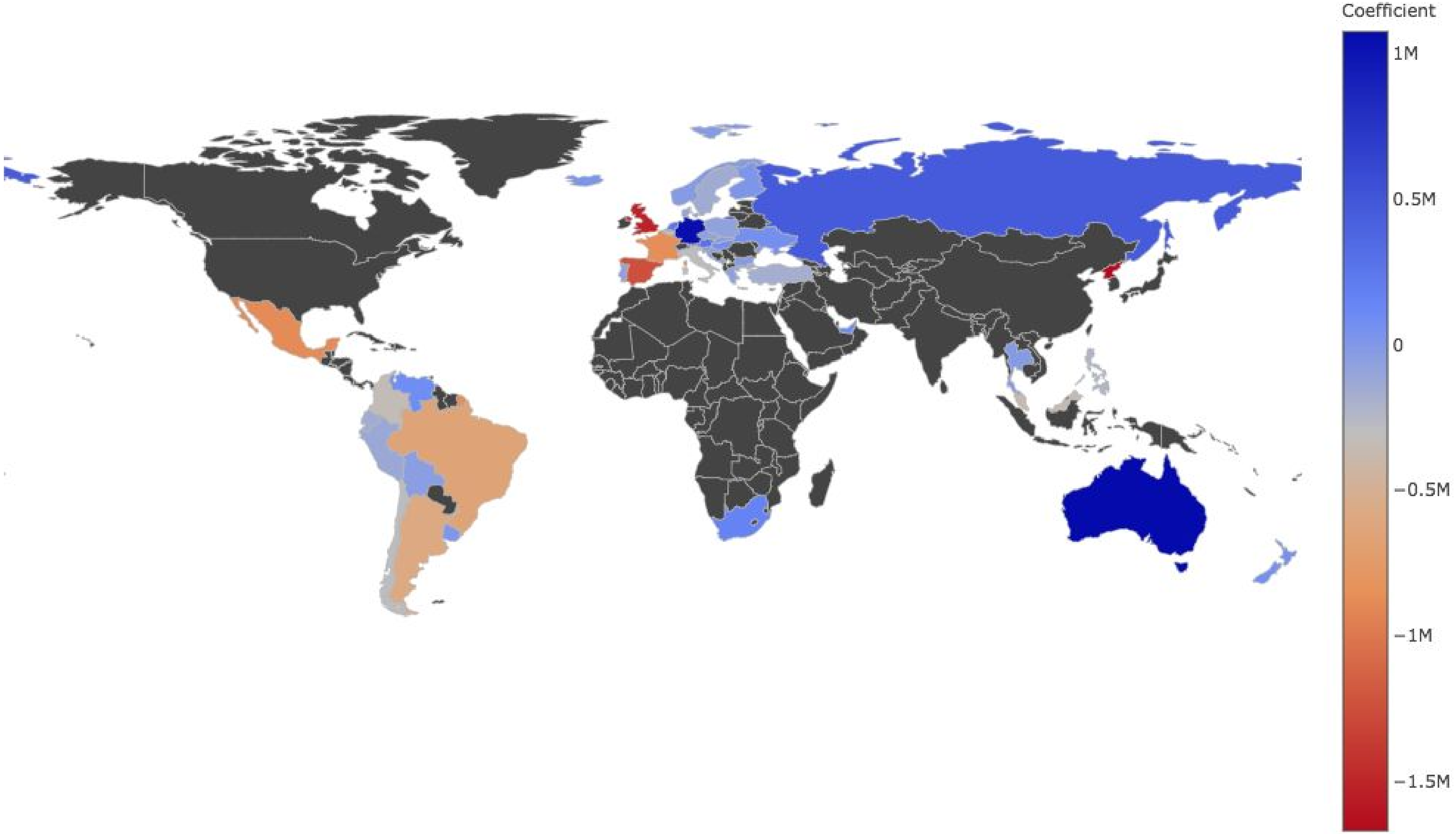
Title	Genre	Budget	Opening Weekend	Foreign Gross	Residual
Oz The Great and Powerful	Action-Adventure	$215,000,000	$79,110,453	$258,400,000	-$195,845,800
Man of Steel	Action-Adventure	$225,000,000	$116,619,362	$377,000,000	-$203,926,000
Batman v Superman: Dawn of Justice	Action-Adventure	$250,000,000	$166,007,347	$542,900,000	-$217,455,200
The Hunger Games	Action-Adventure	$78,000,000	$152,535,747	$286,384,032	-$248,730,300
Green Lantern	Action-Adventure	$200,000,000	$53,174,303	$103,250,000	-$254,771,200

What about by country?

Countries vs. Genres

- Filtered by each country with more than 400 data points, ran OLS regression with a test/train split
- Got R^2 and coefficients for each variable for each country

- Less predictive, mean R^2 of 0.51
- Compared coefficients for comedy



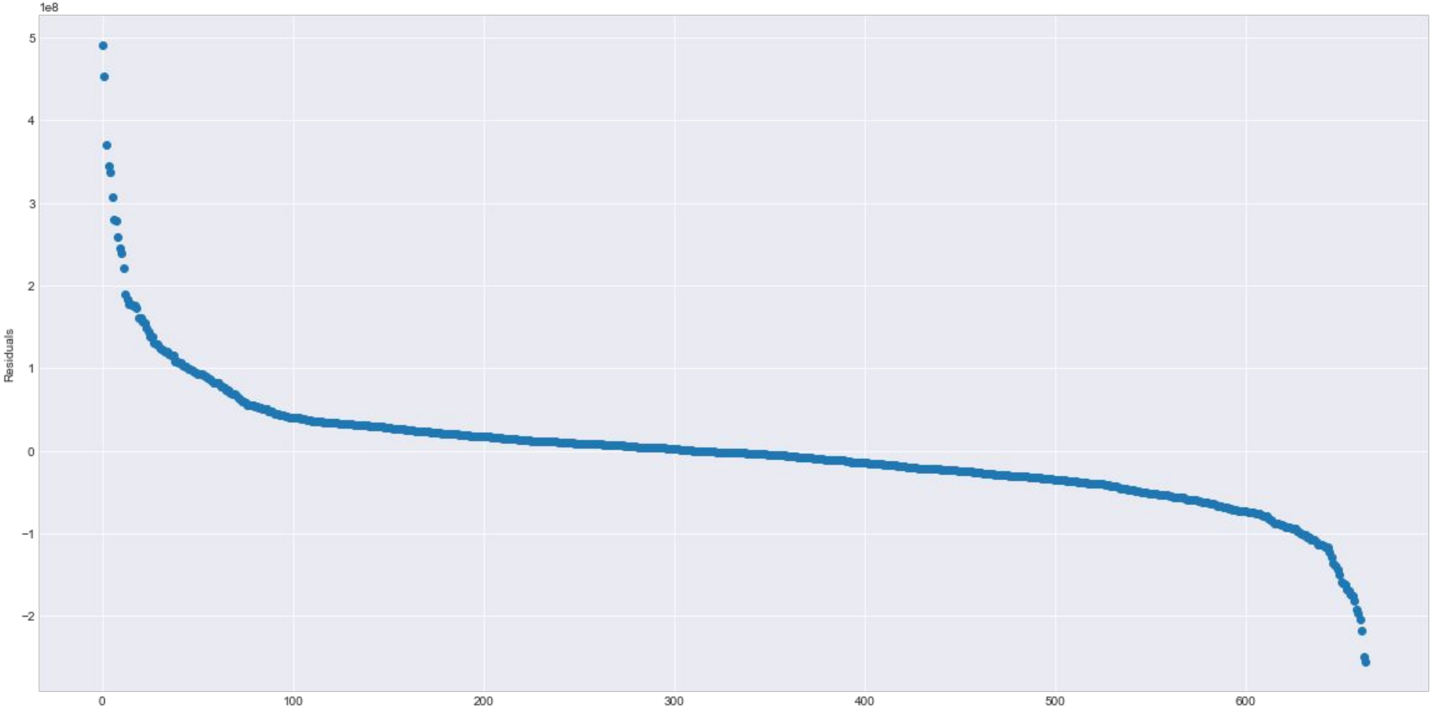
Coefficient of Comedy vs. Box Office by Country

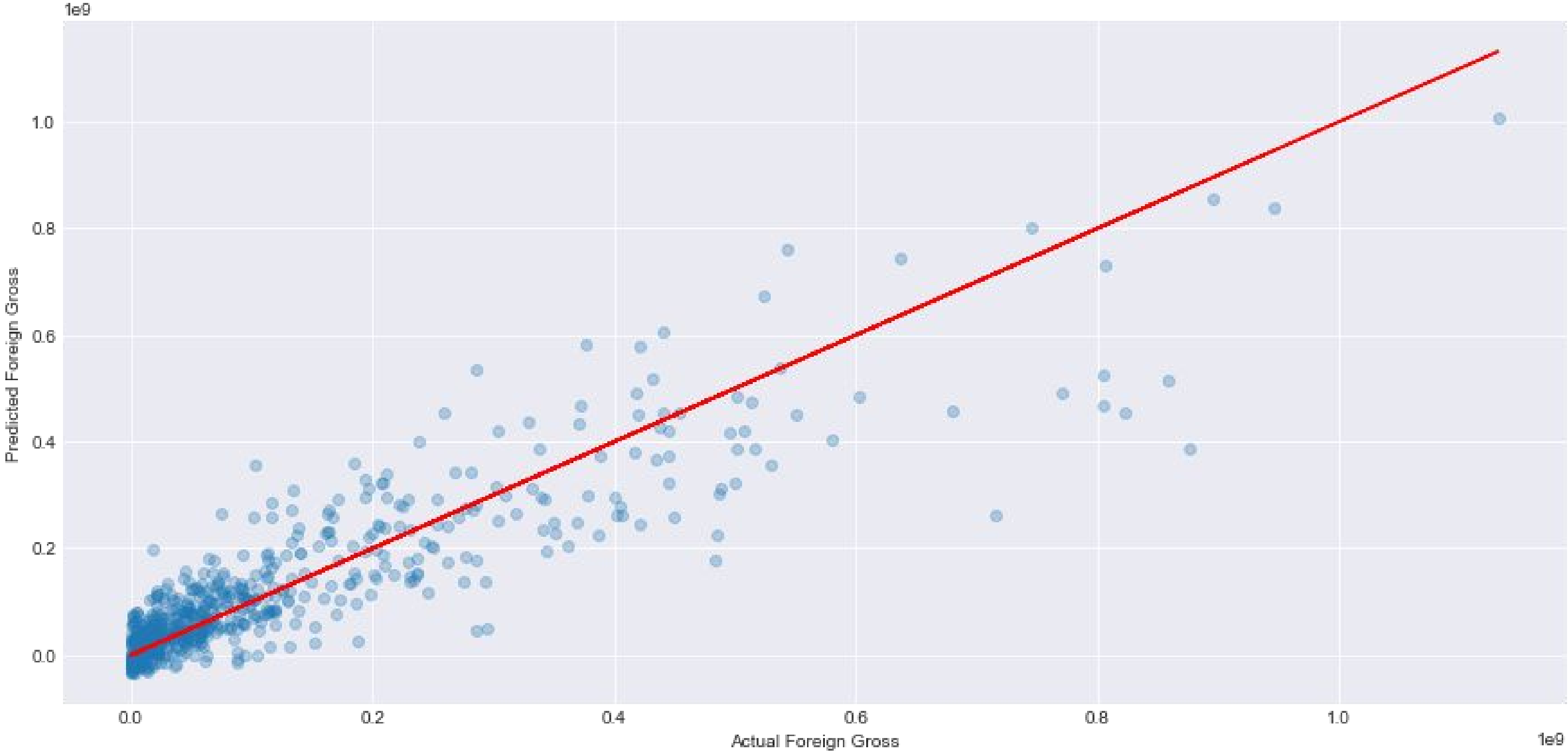
Thank you!

Appendix

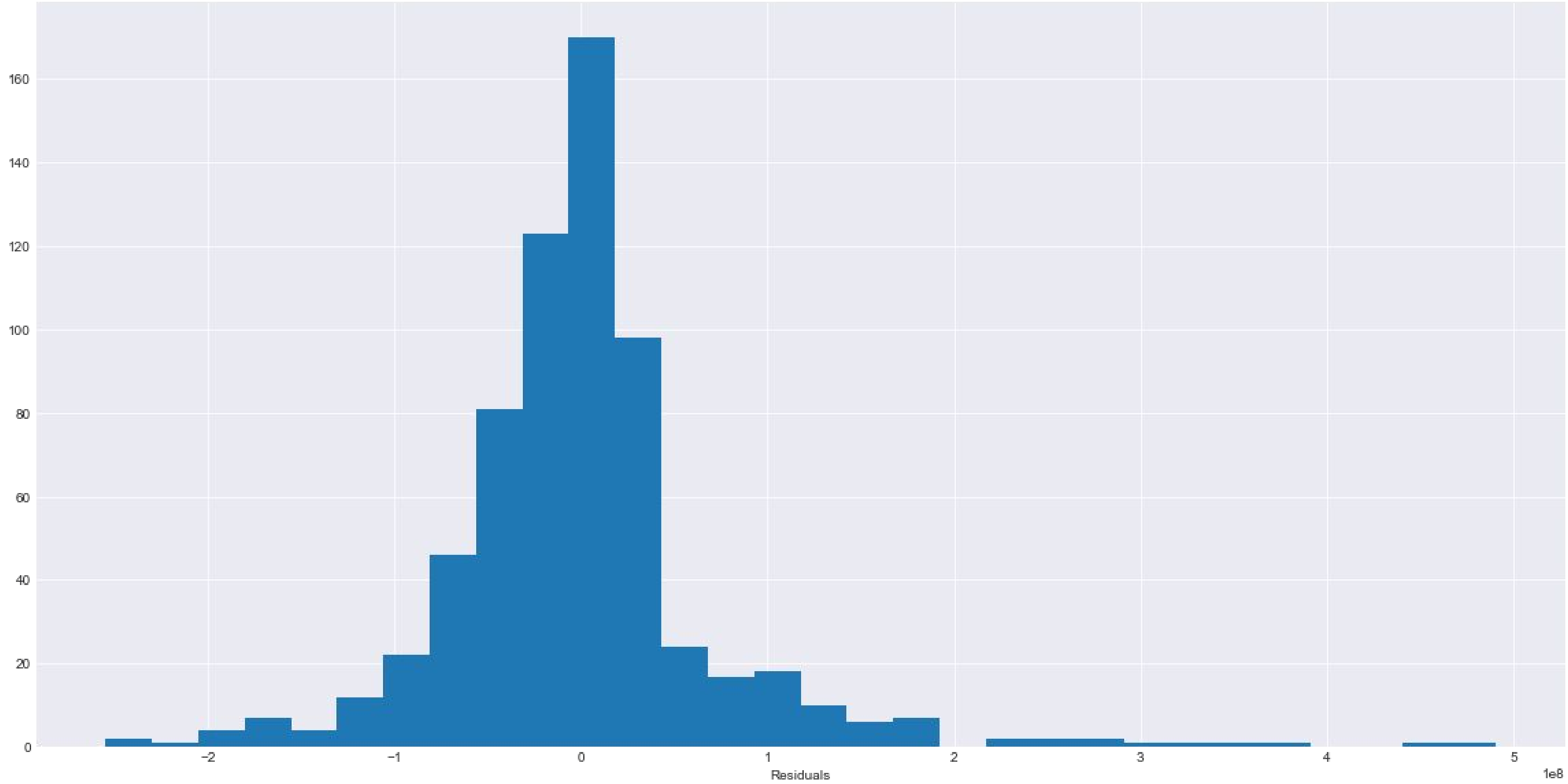
Final Model Additional Charts

Scatter Plot of Sorted Residuals

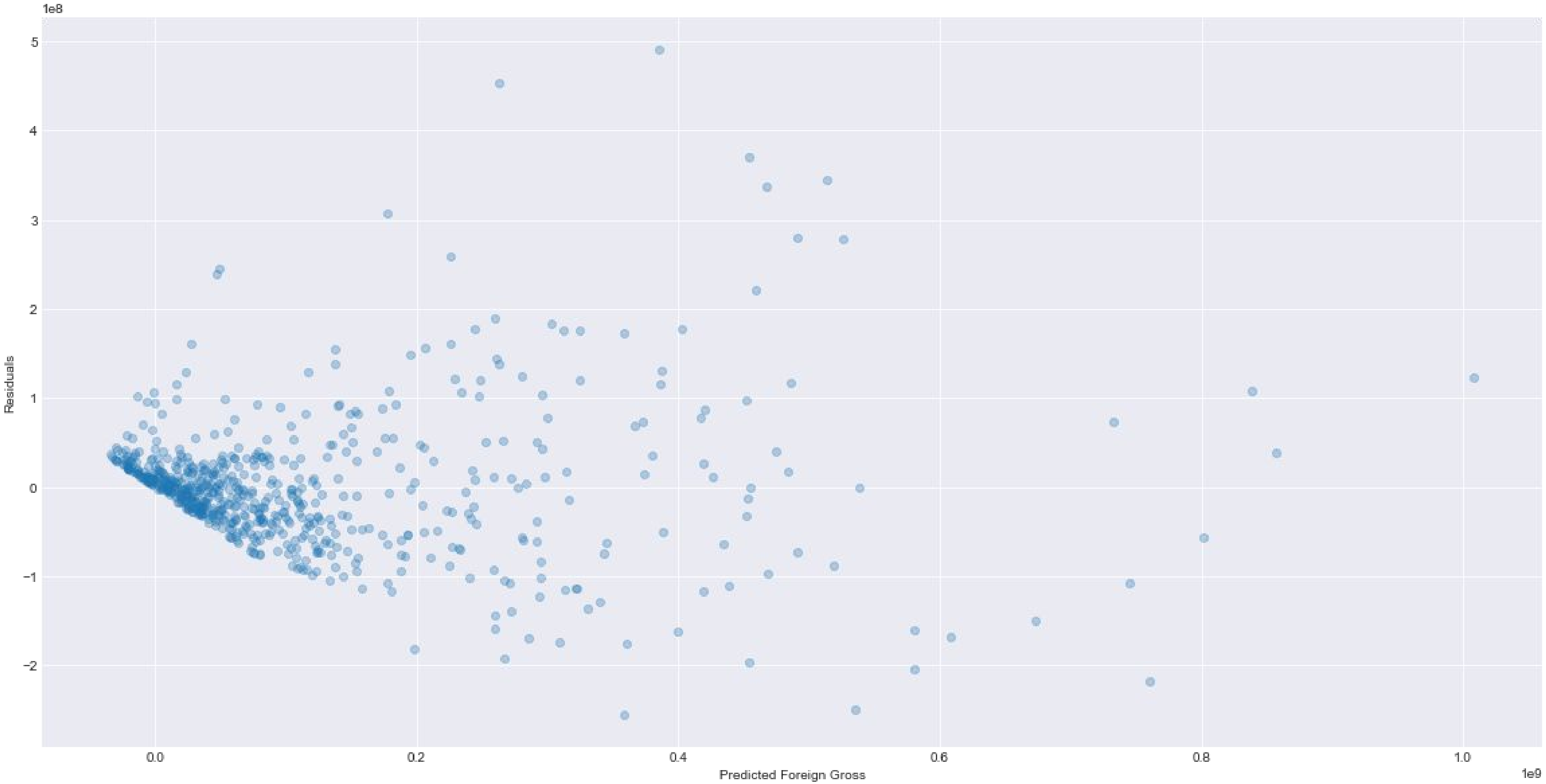




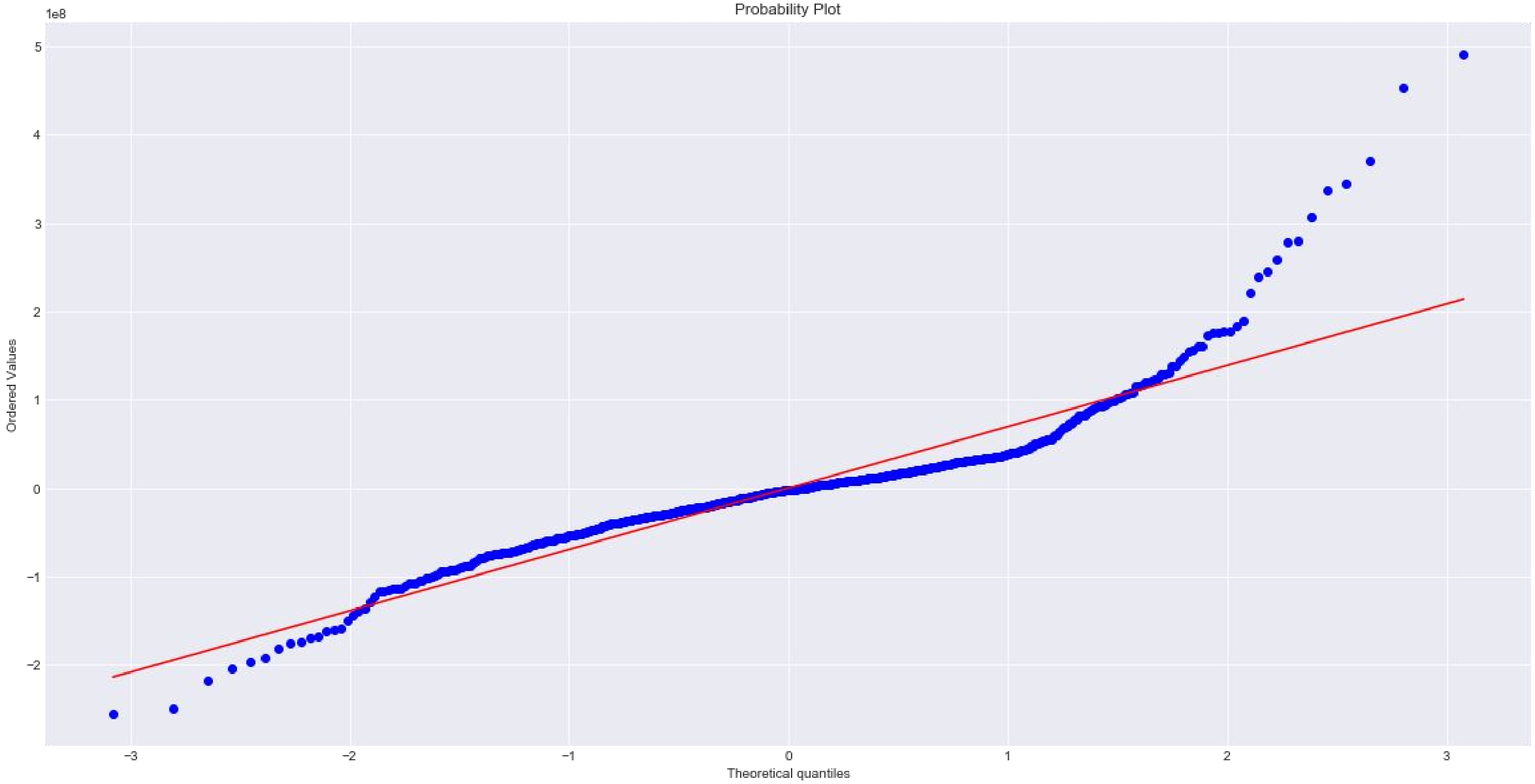
Actual vs. Predicted Foreign Gross



Histogram of Residuals



Residual Plot


Probability Plot
1e8
5
4
3
2
1
0
−1
−2
Ordered Values
−3
−2
−1
0
1
2
3
Theoretical quantiles

Q-Q Plot

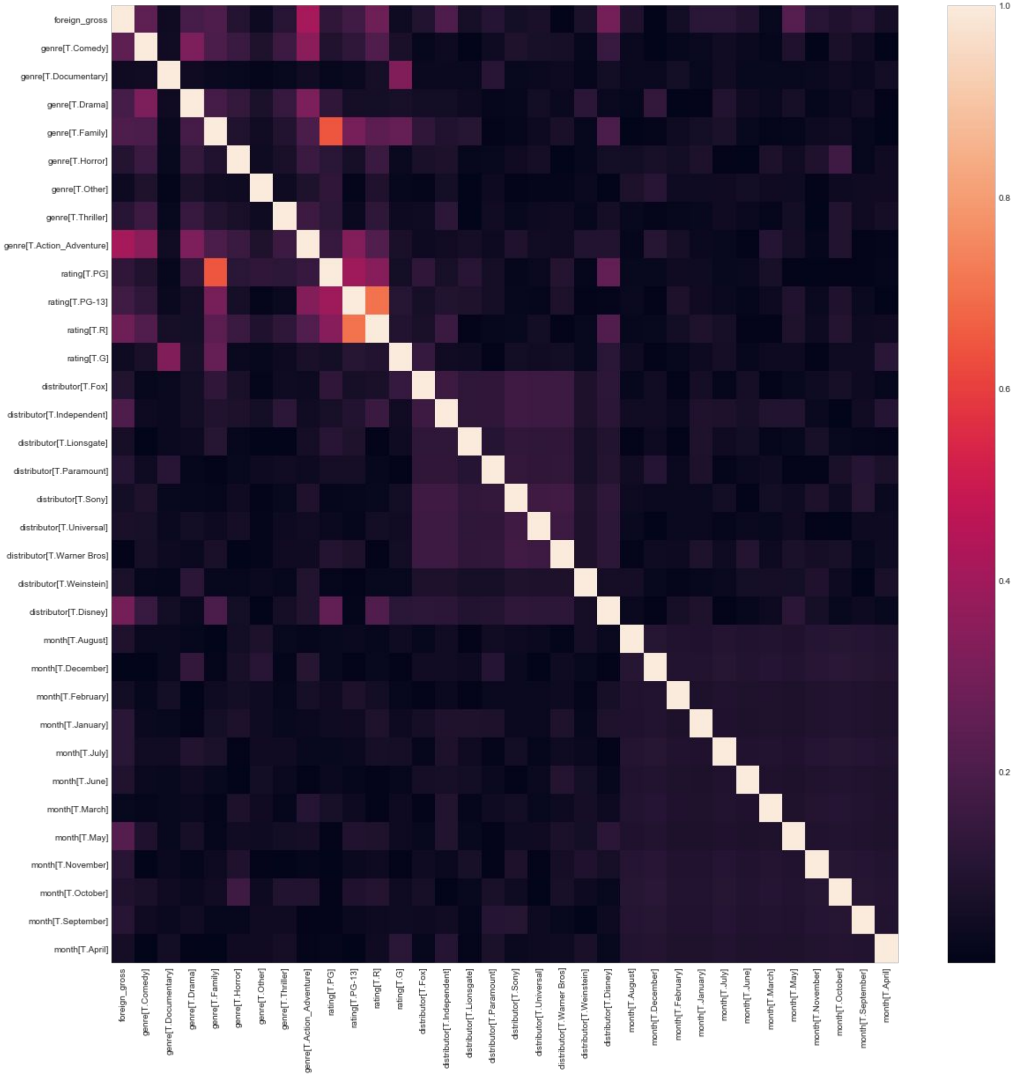
Add Categorical Variables

Genre (8)

Distributor (9)

MPAA Rating (4)

Month of Release (12)



Cross Validation

- Features that were significant for all 10 folds:
  - **Budget**, **opening weekend**
  - **Genres**: Comedy, Drama, Family, Horror, Thriller, Action-Adventure
  - **Ratings**: PG, PG-13, R
  - **Distributors**: Independent, Disney
  - **Months**: January, July, May, November, September

K-Folds Cross Validation with Feature Selection

Mean correlation coefficients for features that were significant in all 10 folds

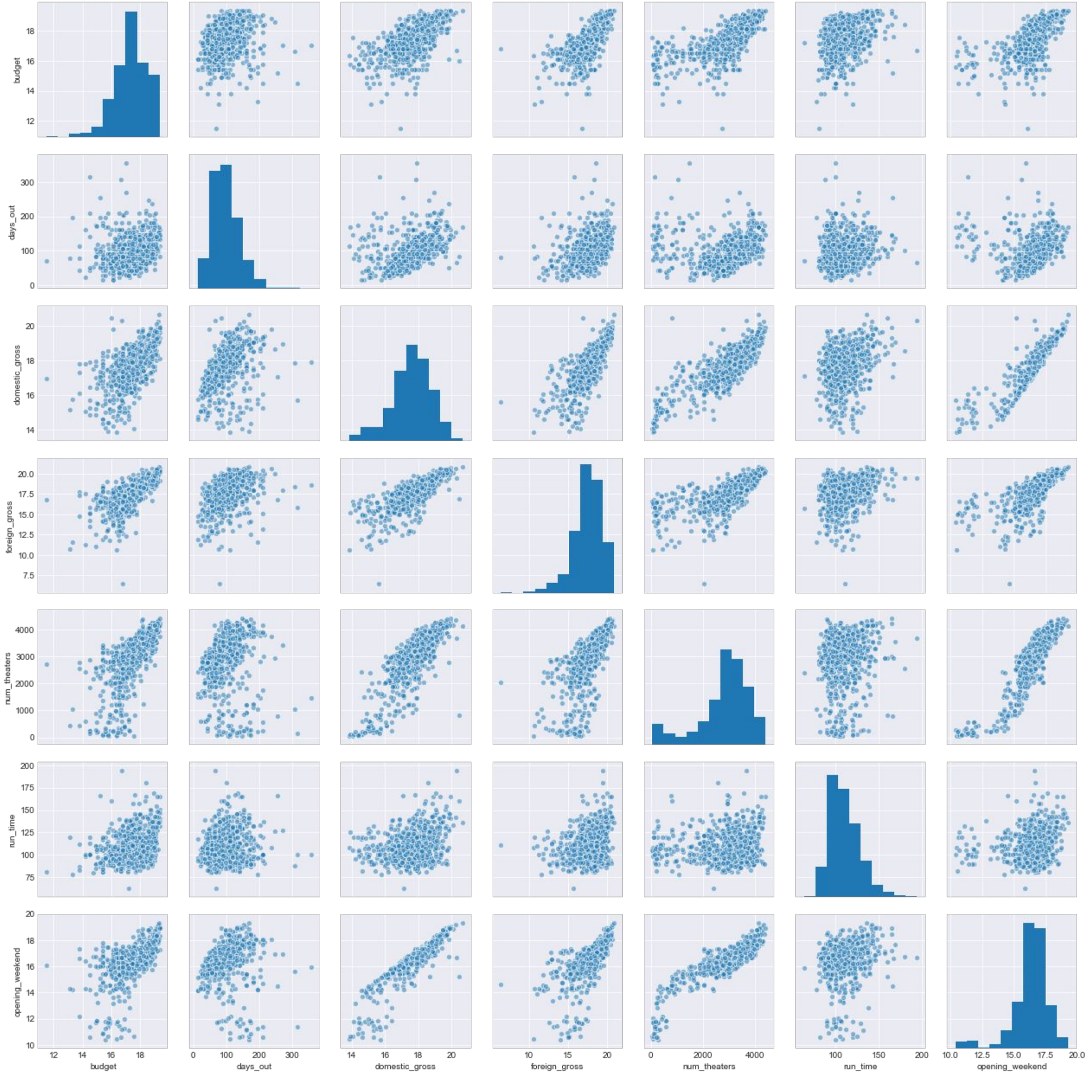
- Budget: 1.025
- Opening weekend: 3.083
- Comedy: -24,730,490
- Drama: -10,681,798
- Family: 31,489,922
- Horror: -3,665,441
- Thriller: -14,988,243
- Action Adventure: -2,976,119
- PG: 27,125,665
- PG-13: 24,158,667
- R: 32,088,963
- Independent: 3,063,612
- Disney: 1,038,925
- January: -1,077,363
- July: 23,181,543
- May: 22,962,655
- November: 26,767,167
- September: 1,411,196

Log Analysis

Log-transforming the data to make the distributions more normal

Pairplot of Variables

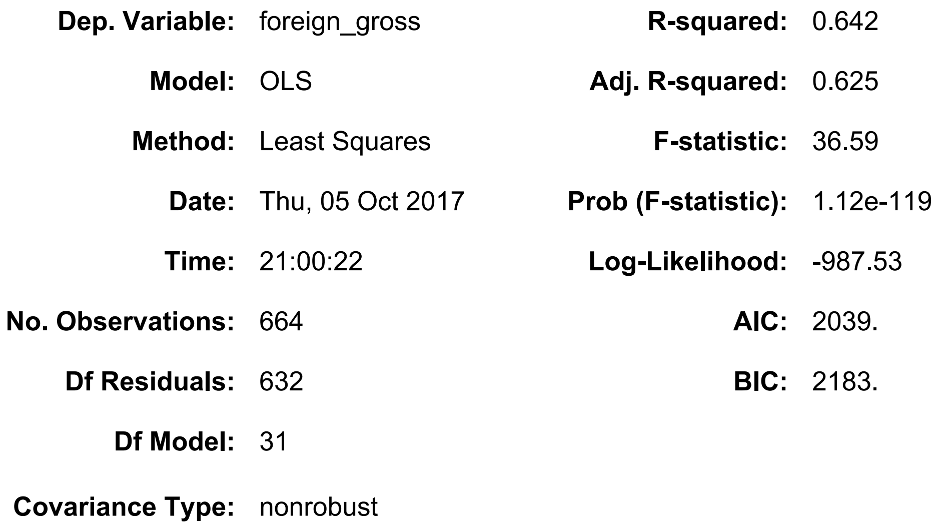
- More normally distributed than without log
- Foreign gross is now slightly left skewed



Ordinary Least Squares Regression on all Variables

R^2 of 0.642

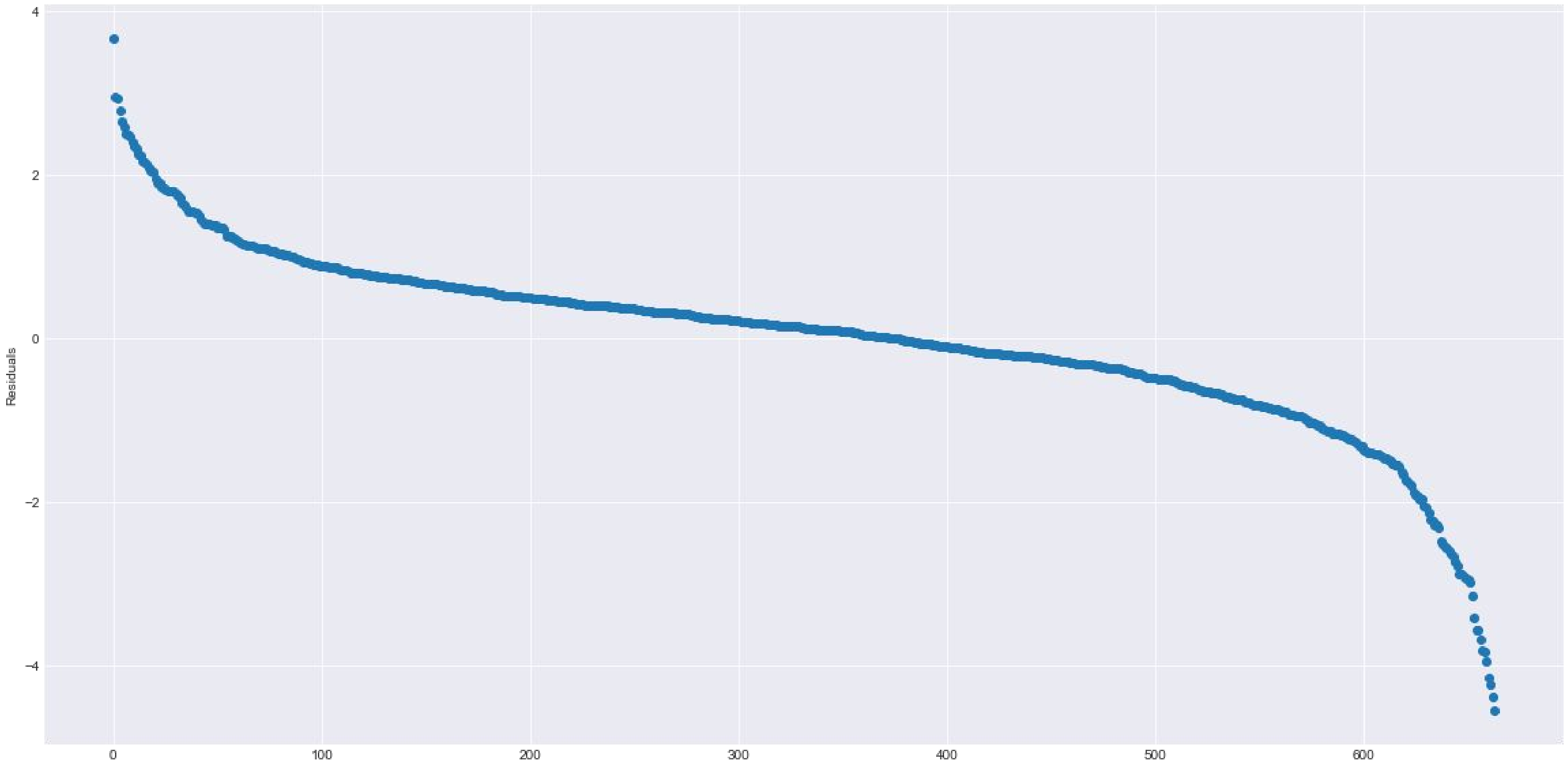
Better AIC and Log-Likelihood



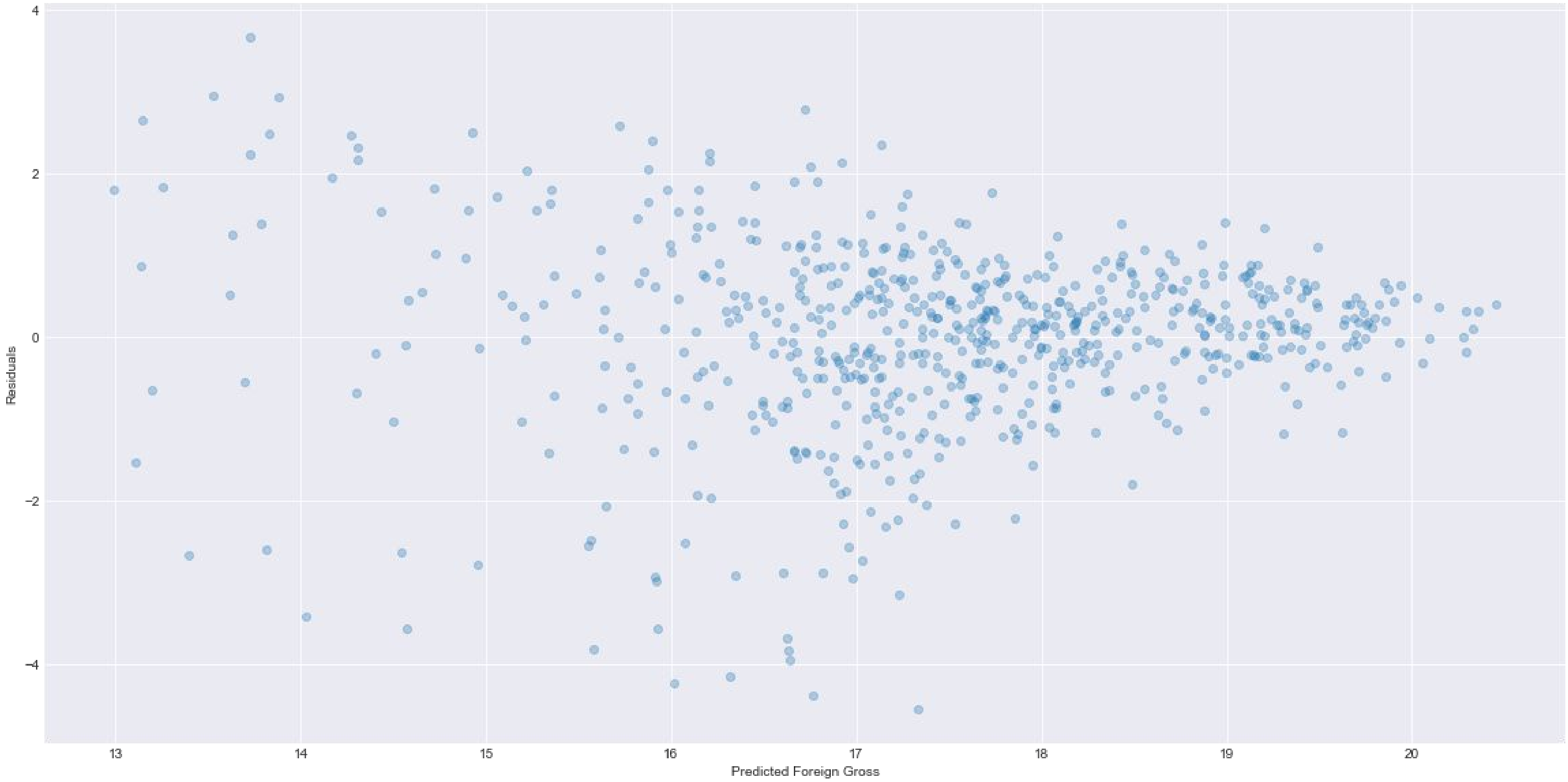
Dep. Variable:	foreign_gross	**R-squared:**	0.642
Model:	OLS	**Adj. R-squared:**	0.625
Method:	Least Squares	**F-statistic:**	36.59
Date:	Thu, 05 Oct 2017	**Prob (F-statistic):**	1.12e-119
Time:	21:00:22	**Log-Likelihood:**	-987.53
No. Observations:	664	**AIC:**	2039.
Df Residuals:	632	**BIC:**	2183.
Df Model:	31		
Covariance Type:	nonrobust		

Final Model: Only Budget, Opening Weekend, Genre

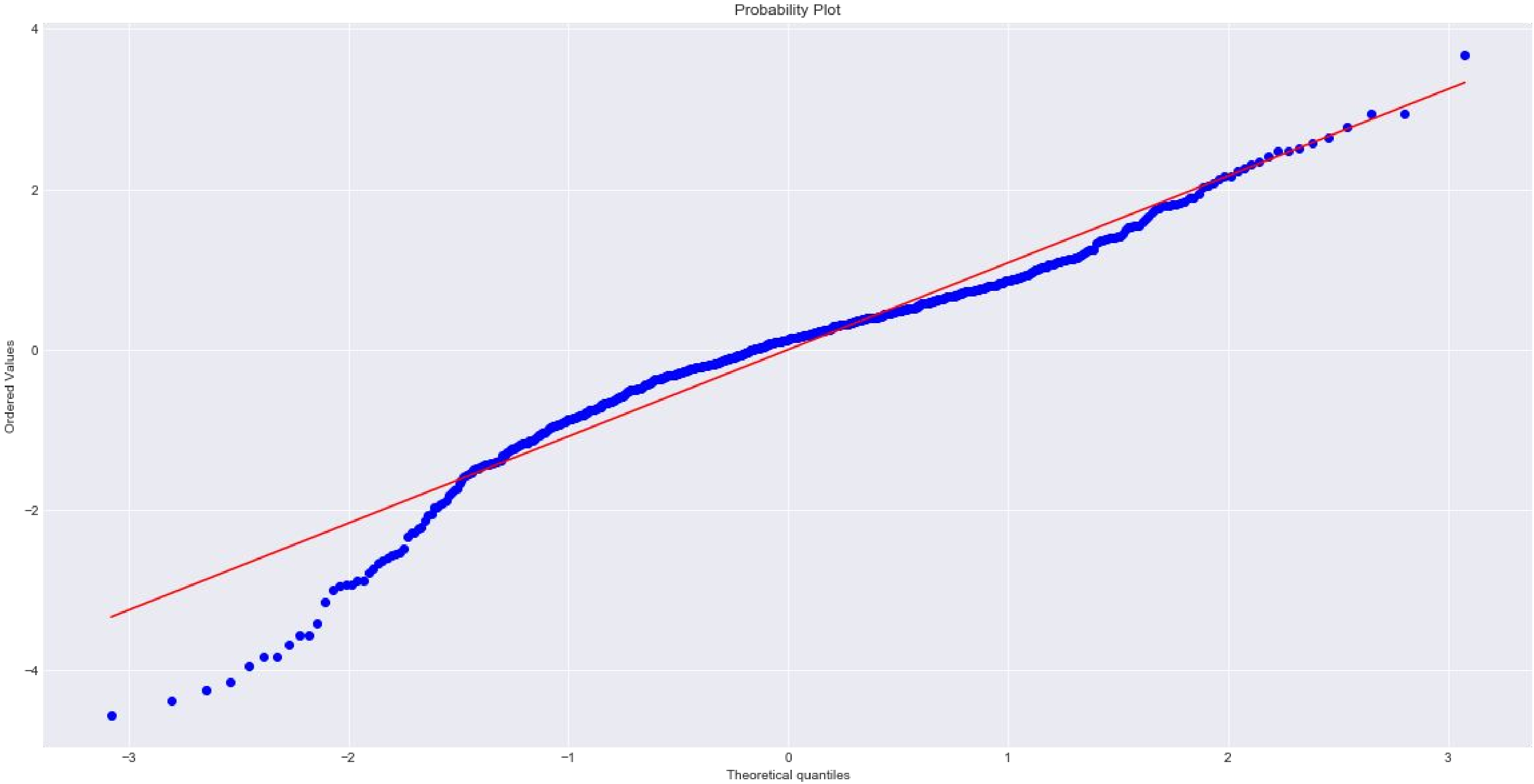
- R^2 of 0.616
- Model tends to overpredict vs. underpredict



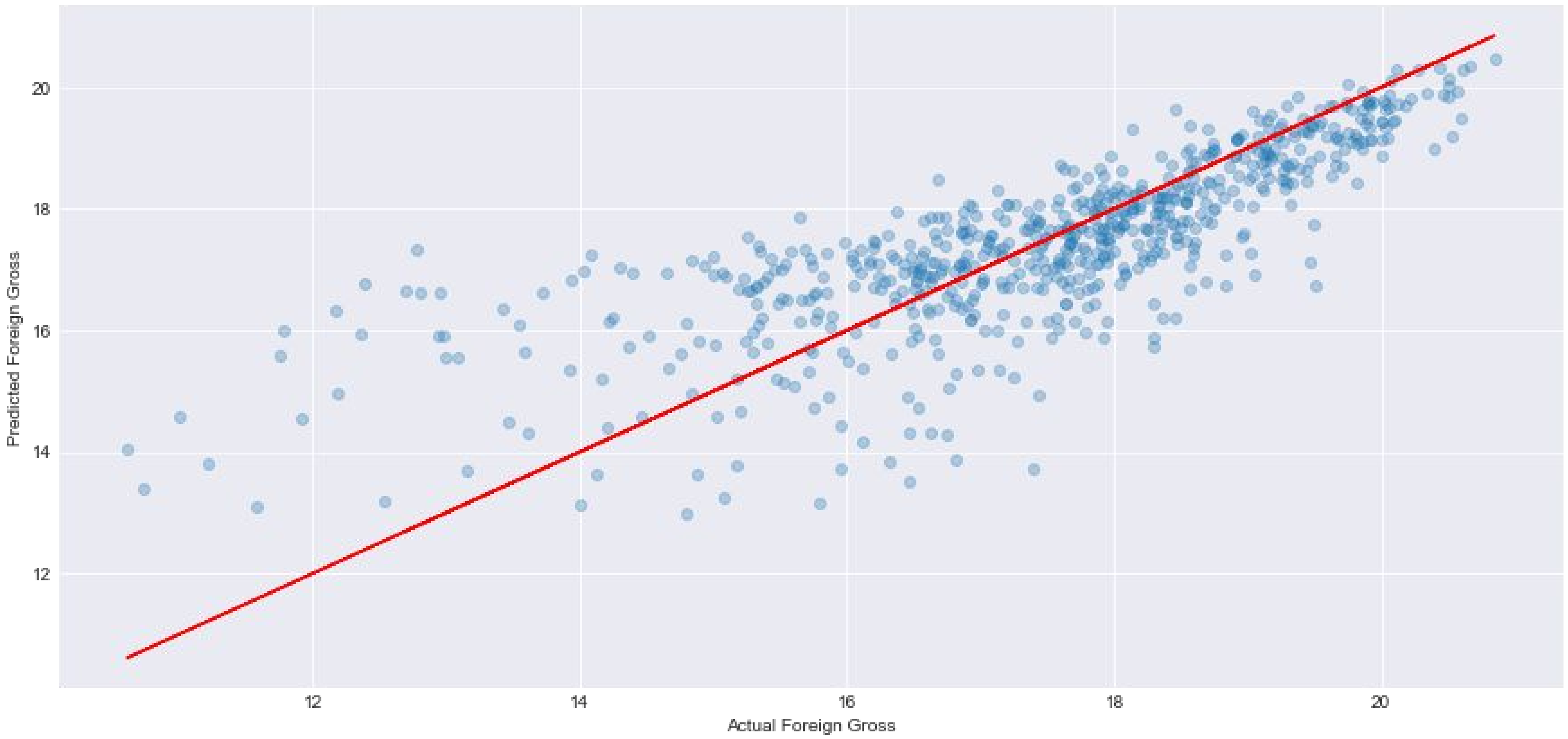
Scatter Plot of Sorted Residuals



Residual Plot (more random)


Probability Plot
Ordered Values
Theoretical quantiles
4
2
0
−2
−4
−3
−2
−1
0
1
2
3

Q-Q Plot



Actual vs. Predicted Foreign Gross