Regression Analyses with Movie Data

Insights with Davis Vance
Box Office Mojo

Data Collection / Features

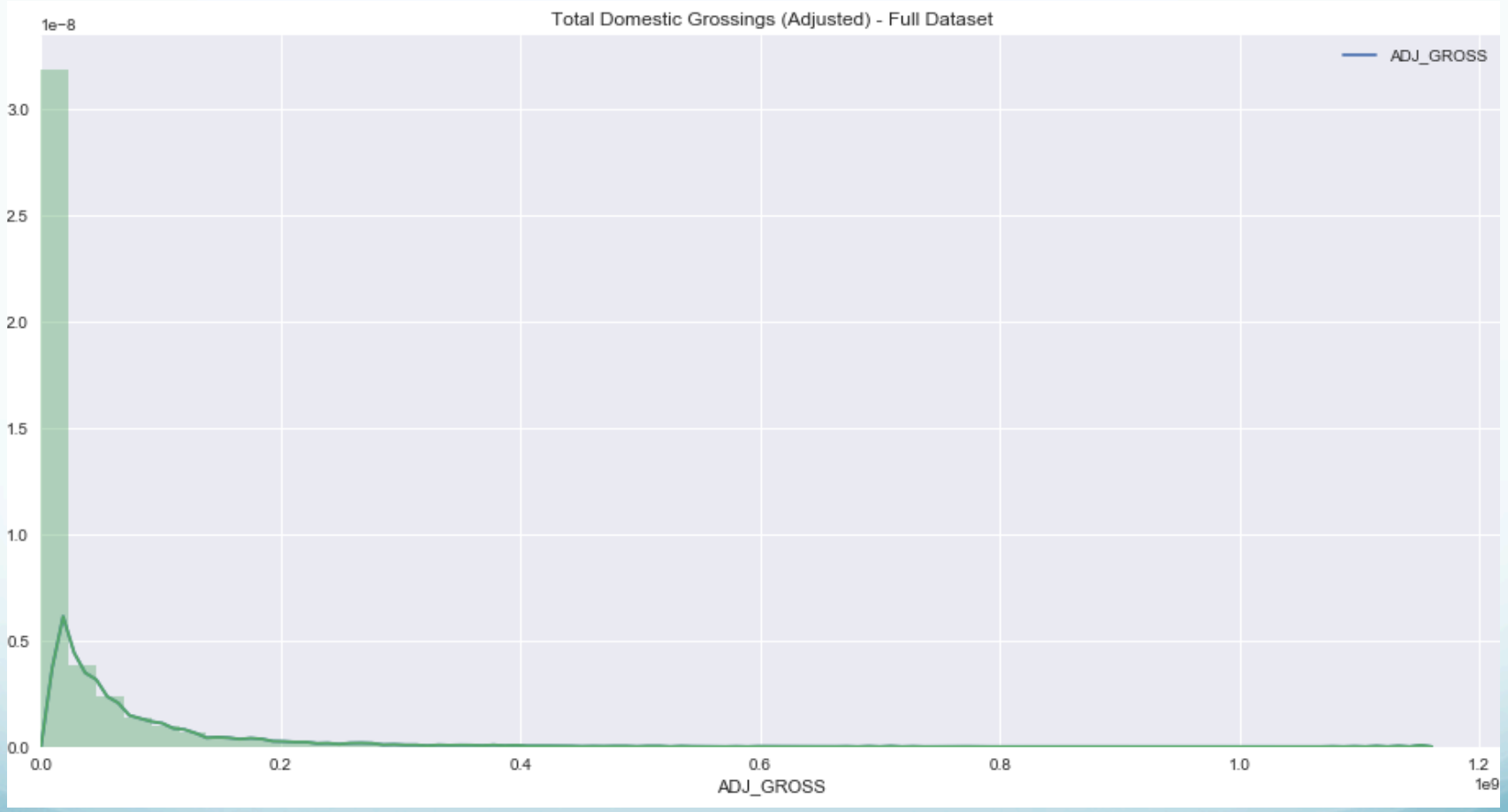
- Box Office Mojo:
  - 15K+ movies
  - 1980 – 2018
- Movie Grossing
  - Total Domestic, Adjusted to 2017 Dollars
- Date of Release
  - Converted to Age as of January 31, 2018 as a Time Delta object
- Major Genres
  - 66 -> 6 categories, including other
- Runtime in Minutes
  - < 300 minutes, Further analysis with stricter range
- MPAA Rating
  - G, PG, PG-13, R, Other
- Budgets, Worldwide Gross, Domestic Unadjusted Gross
  - These features were dropped due to missing data

GENRE	
Comedy	3142
Drama	2315
Adventure_Action	1655
Documentary	1644
Thriller_Horror	1606

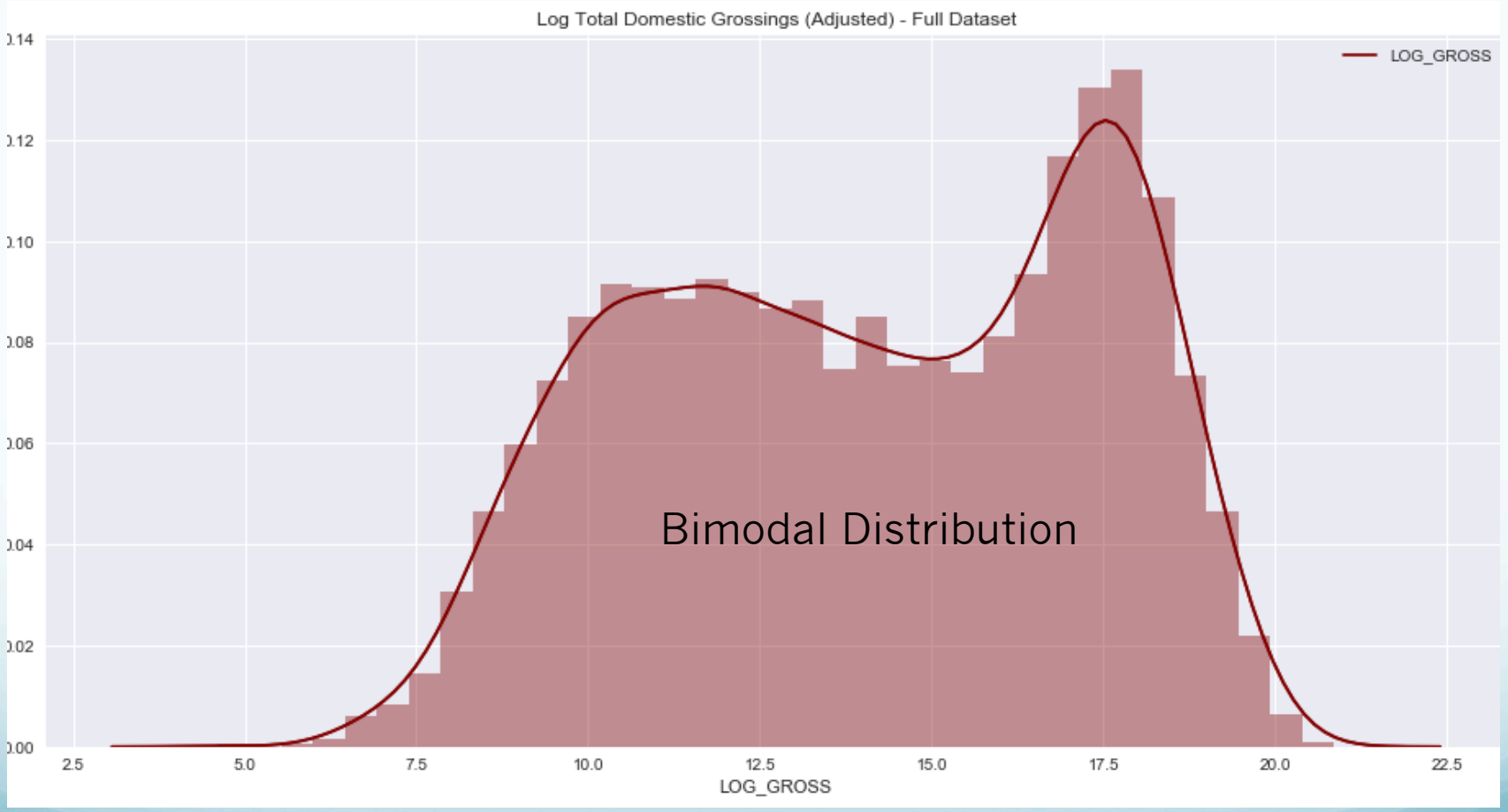
Data Cleaning

- 12K records did not have budgets
  - Important feature for predicting box office revenue
  - Made a subset of 3K movies with budget for analysis
  - Did not prove to be worthy endeavor
- ~500 Duplicates
  - Dropping duplicates was compared. Minor implications overall, but could be cleaned further to combine re-releases

Original Distribution of Domestic Movie Gross (ADJ'17)



Natural Log Transformation of Movie Gross

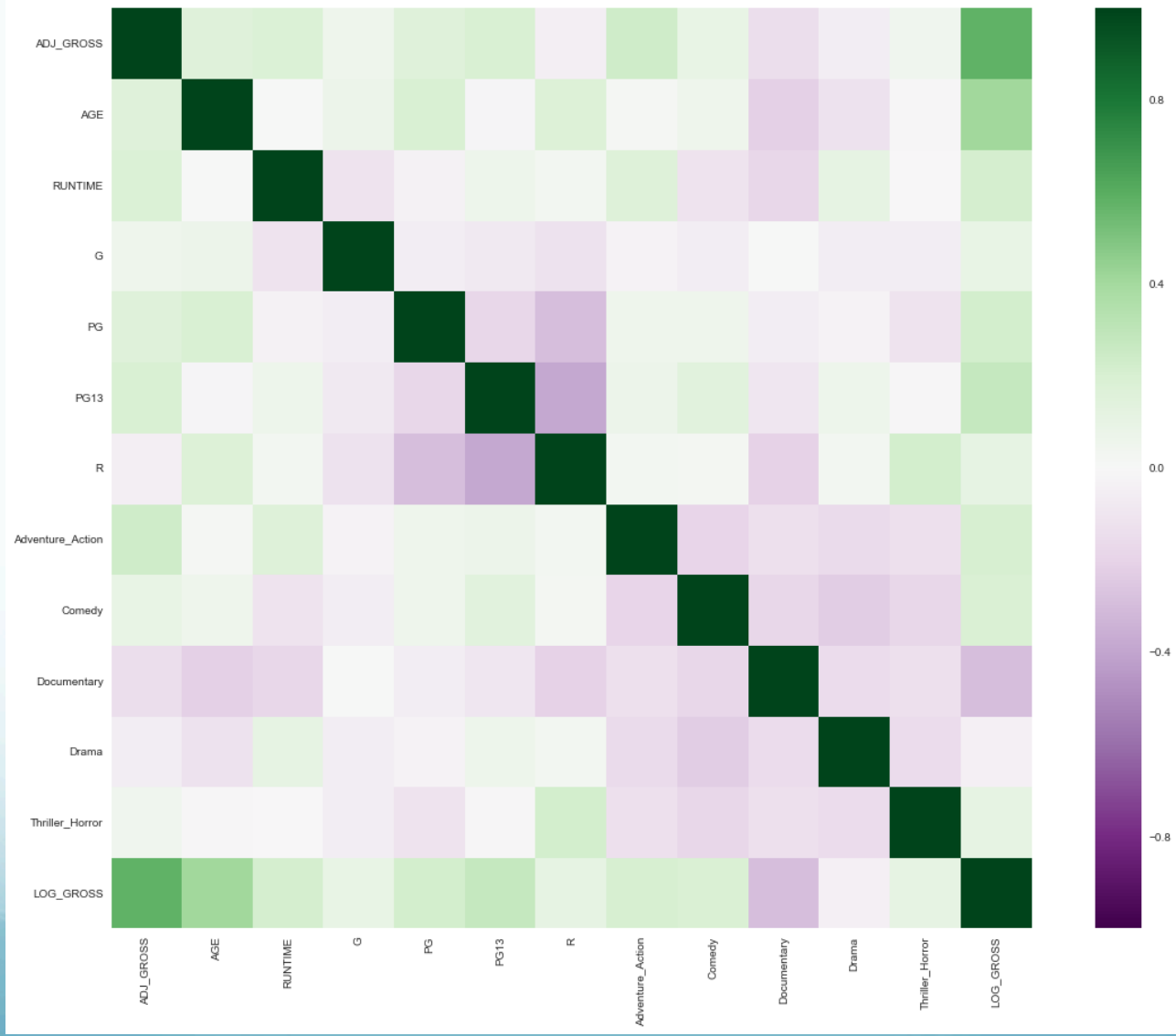


Dependent Variable Log Transformation

- Changes the fundamental question of the regression
- Estimating a different average effect from independent variables
- Basically, a 1 unit increase in our independent variable relates to a (Beta*100) % increase in movie grossing

Correlation Heatmap

No Multi-collinearity



Regression

OLS Regression Results

Dep. Variable:	ADJ_GROSS	**R-squared:**	0.147
Model:	OLS	**Adj. R-squared:**	0.147
Method:	Least Squares	**F-statistic:**	422.3
Date:	Thu, 01 Feb 2018	**Prob (F-statistic):**	0.00
Time:	20:35:03	**Log-Likelihood:**	-2.8420e+05
No. Observations:	14694	**AIC:**	5.684e+05
Df Residuals:	14687	**BIC:**	5.685e+05
Df Model:	6		
Covariance Type:	nonrobust		

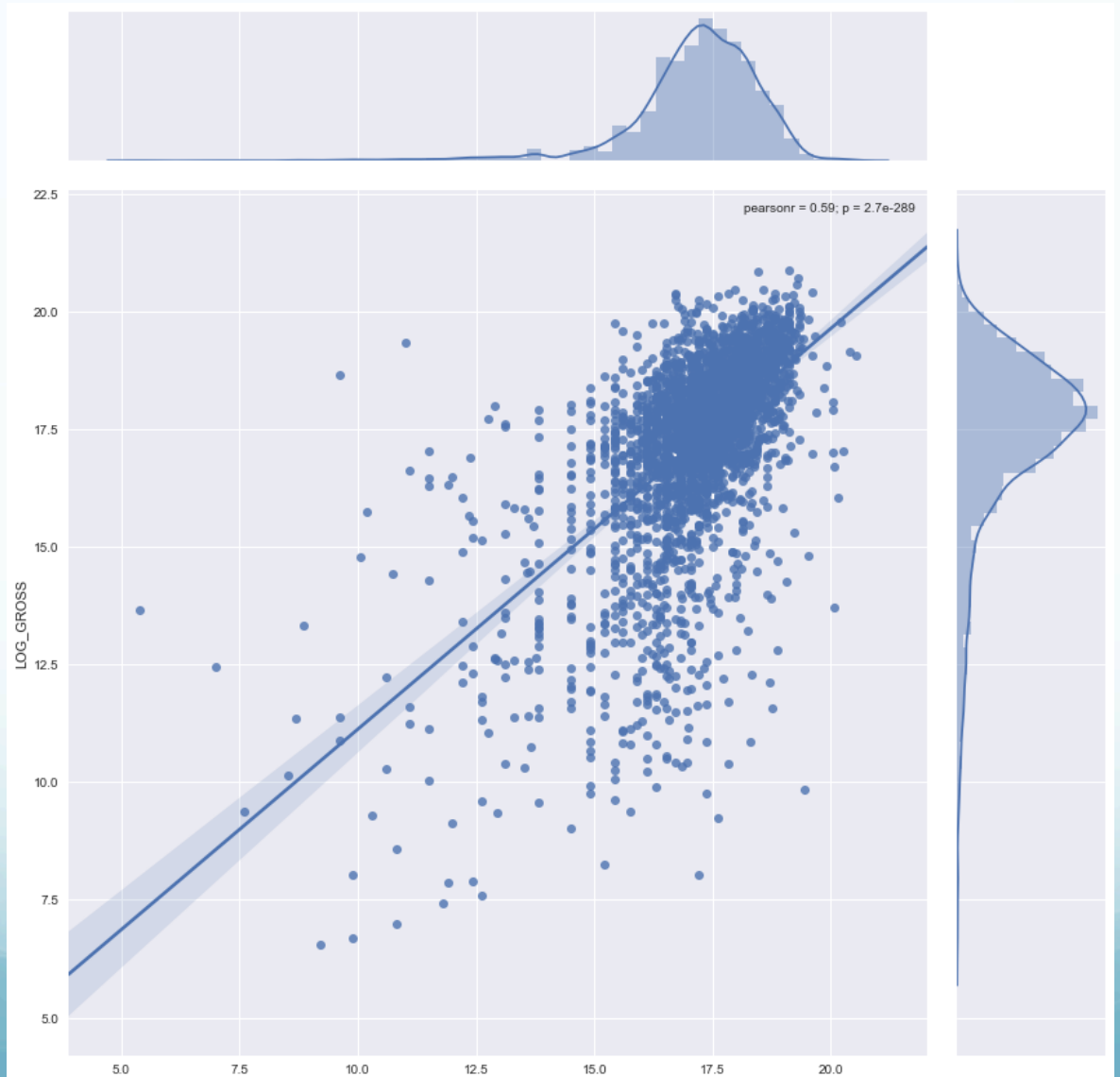
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.654e+07	2.81e+06	-23.699	0.000	-7.2e+07	-6.1e+07
AGE_Y	5.234e+05	5.49e+04	9.531	0.000	4.16e+05	6.31e+05
RUNTIME	6.116e+05	2.54e+04	24.089	0.000	5.62e+05	6.61e+05
G	5.592e+07	3.42e+06	16.329	0.000	4.92e+07	6.26e+07
PG	5.053e+07	1.82e+06	27.742	0.000	4.7e+07	5.41e+07
PG13	4.884e+07	1.5e+06	32.449	0.000	4.59e+07	5.18e+07
R	1.795e+07	1.33e+06	13.550	0.000	1.54e+07	2.06e+07

Omnibus:	14394.391	**Durbin-Watson:**	1.631
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	991609.115
Skew:	4.714	**Prob(JB):**	0.00
Kurtosis:	42.125	**Cond. No.**	763.

- Adjusted Movie Grossing
- Baseline with Age (Y), Runtime, and MPAA Ratings
- R-Sq. = .147
- I decided to log the data before modeling, so I confidently threw this base case out

Linear Relationship

Logged Budget on Logged Grossing



Regression

OLS Regression Results

Dep. Variable:	LOG_GROSS	**R-squared:**	0.444
Model:	OLS	**Adj. R-squared:**	0.444
Method:	Least Squares	**F-statistic:**	1954.
Date:	Thu, 01 Feb 2018	**Prob (F-statistic):**	0.00
Time:	20:35:04	**Log-Likelihood:**	-34072.
No. Observations:	14694	**AIC:**	6.816e+04
Df Residuals:	14687	**BIC:**	6.821e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.9790	0.114	61.414	0.000	6.756	7.202
AGE_Y	0.0819	0.002	36.850	0.000	0.078	0.086
RUNTIME	0.0333	0.001	32.372	0.000	0.031	0.035
G	4.6908	0.139	33.840	0.000	4.419	4.962
PG	4.1244	0.074	55.938	0.000	3.980	4.269
PG13	4.2732	0.061	70.150	0.000	4.154	4.393
R	2.7454	0.054	51.188	0.000	2.640	2.851

Omnibus:	365.722	**Durbin-Watson:**	1.857
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	392.657
Skew:	-0.399	**Prob(JB):**	5.44e-86
Kurtosis:	2.925	**Cond. No.**	763.

- Logged Movie Grossing
- New Baseline with Age, Runtime, and MPAA Ratings
- R-Sq. = .444
- 6 Variables

Regression

OLS Regression Results

Dep. Variable:	LOG_GROSS	**R-squared:**	0.497
Model:	OLS	**Adj. R-squared:**	0.497
Method:	Least Squares	**F-statistic:**	1318.
Date:	Fri, 02 Feb 2018	**Prob (F-statistic):**	0.00
Time:	09:06:25	**Log-Likelihood:**	-33337.
No. Observations:	14694	**AIC:**	6.670e+04
Df Residuals:	14682	**BIC:**	6.679e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5160	0.120	54.205	0.000	6.280	6.752
AGE_Y	0.0904	0.002	41.237	0.000	0.086	0.095
RUNTIME	0.0327	0.001	32.224	0.000	0.031	0.035
G	4.7964	0.133	36.140	0.000	4.536	5.057
PG	3.5706	0.073	49.150	0.000	3.428	3.713
PG13	3.4643	0.063	55.012	0.000	3.341	3.588
R	1.9886	0.056	35.362	0.000	1.878	2.099
Adventure_Action	1.9697	0.070	27.976	0.000	1.832	2.108
Comedy	1.6702	0.059	28.505	0.000	1.555	1.785
Documentary	0.1464	0.072	2.029	0.043	0.005	0.288
Drama	0.5730	0.064	9.019	0.000	0.448	0.698
Thriller_Horror	1.9312	0.072	26.707	0.000	1.789	2.073

- Logged Movie Grossing
- Baseline with Age, Runtime, and MPAA Ratings
- R-Sq. = .497
- 11 Variables, mostly categorical variables
- Still interpretable

Regression

OLS Regression Results

Dep. Variable:	LOG_GROSS	**R-squared:**	0.497
Model:	OLS	**Adj. R-squared:**	0.497
Method:	Least Squares	**F-statistic:**	1318.
Date:	Fri, 02 Feb 2018	**Prob (F-statistic):**	0.00
Time:	09:06:25	**Log-Likelihood:**	-33337.
No. Observations:	14694	**AIC:**	6.670e+04
Df Residuals:	14682	**BIC:**	6.679e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5160	0.120	54.205	0.000	6.280	6.752
AGE_Y	0.0904	0.002	41.237	0.000	0.086	0.095
RUNTIME	0.0327	0.001	32.224	0.000	0.031	0.035
G	4.7964	0.133	36.140	0.000	4.536	5.057
PG	3.5706	0.073	49.150	0.000	3.428	3.713
PG13	3.4643	0.063	55.012	0.000	3.341	3.588
R	1.9886	0.056	35.362	0.000	1.878	2.099
Adventure_Action	1.9697	0.070	27.976	0.000	1.832	2.108
Comedy	1.6702	0.059	28.505	0.000	1.555	1.785
Documentary	0.1464	0.072	2.029	0.043	0.005	0.288
Drama	0.5730	0.064	9.019	0.000	0.448	0.698
Thriller_Horror	1.9312	0.072	26.707	0.000	1.789	2.073

- Once genres were accounted for, the G-rated effect grew a bit
- PG, PG13, and R rated movies effect actually dropped by ~1 coefficient point

Regression

OLS Regression Results

Dep. Variable:	LOG_GROSS	**R-squared:**	0.497
Model:	OLS	**Adj. R-squared:**	0.497
Method:	Least Squares	**F-statistic:**	1318.
Date:	Fri, 02 Feb 2018	**Prob (F-statistic):**	0.00
Time:	09:06:25	**Log-Likelihood:**	-33337.
No. Observations:	14694	**AIC:**	6.670e+04
Df Residuals:	14682	**BIC:**	6.679e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5160	0.120	54.205	0.000	6.280	6.752
AGE_Y	0.0904	0.002	41.237	0.000	0.086	0.095
RUNTIME	0.0327	0.001	32.224	0.000	0.031	0.035
G	4.7964	0.133	36.140	0.000	4.536	5.057
PG	3.5706	0.073	49.150	0.000	3.428	3.713
PG13	3.4643	0.063	55.012	0.000	3.341	3.588
R	1.9886	0.056	35.362	0.000	1.878	2.099
Adventure_Action	1.9697	0.070	27.976	0.000	1.832	2.108
Comedy	1.6702	0.059	28.505	0.000	1.555	1.785
Documentary	0.1464	0.072	2.029	0.043	0.005	0.288
Drama	0.5730	0.064	9.019	0.000	0.448	0.698
Thriller_Horror	1.9312	0.072	26.707	0.000	1.789	2.073

- G rated movies could produce about a third (33%) more grossing than another main movie rating like PG or PG13 movies on average but comes with a bit more variation
- Need to look more into the segmentation to confirm these results

Regression

OLS Regression Results

Dep. Variable:	LOG_GROSS	**R-squared:**	0.497
Model:	OLS	**Adj. R-squared:**	0.497
Method:	Least Squares	**F-statistic:**	1318.
Date:	Fri, 02 Feb 2018	**Prob (F-statistic):**	0.00
Time:	09:06:25	**Log-Likelihood:**	-33337.
No. Observations:	14694	**AIC:**	6.670e+04
Df Residuals:	14682	**BIC:**	6.679e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5160	0.120	54.205	0.000	6.280	6.752
AGE_Y	0.0904	0.002	41.237	0.000	0.086	0.095
RUNTIME	0.0327	0.001	32.224	0.000	0.031	0.035
G	4.7964	0.133	36.140	0.000	4.536	5.057
PG	3.5706	0.073	49.150	0.000	3.428	3.713
PG13	3.4643	0.063	55.012	0.000	3.341	3.588
R	1.9886	0.056	35.362	0.000	1.878	2.099
Adventure_Action	1.9697	0.070	27.976	0.000	1.832	2.108
Comedy	1.6702	0.059	28.505	0.000	1.555	1.785
Documentary	0.1464	0.072	2.029	0.043	0.005	0.288
Drama	0.5730	0.064	9.019	0.000	0.448	0.698
Thriller_Horror	1.9312	0.072	26.707	0.000	1.789	2.073

- Adventurey-Actiony, Thrilling, Scary movies and Comedies are good options rather than Documentaries, serious Dramas, or the “Others”
- I could do some manual best set regressions to derive more meaning from this modeling style

Regression

OLS Regression Results

Dep. Variable:	LOG_GROSS	**R-squared:**	0.497
Model:	OLS	**Adj. R-squared:**	0.497
Method:	Least Squares	**F-statistic:**	1318.
Date:	Fri, 02 Feb 2018	**Prob (F-statistic):**	0.00
Time:	09:06:25	**Log-Likelihood:**	-33337.
No. Observations:	14694	**AIC:**	6.670e+04
Df Residuals:	14682	**BIC:**	6.679e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5160	0.120	54.205	0.000	6.280	6.752
AGE_Y	0.0904	0.002	41.237	0.000	0.086	0.095
RUNTIME	0.0327	0.001	32.224	0.000	0.031	0.035
G	4.7964	0.133	36.140	0.000	4.536	5.057
PG	3.5706	0.073	49.150	0.000	3.428	3.713
PG13	3.4643	0.063	55.012	0.000	3.341	3.588
R	1.9886	0.056	35.362	0.000	1.878	2.099
Adventure_Action	1.9697	0.070	27.976	0.000	1.832	2.108
Comedy	1.6702	0.059	28.505	0.000	1.555	1.785
Documentary	0.1464	0.072	2.029	0.043	0.005	0.288
Drama	0.5730	0.064	9.019	0.000	0.448	0.698
Thriller_Horror	1.9312	0.072	26.707	0.000	1.789	2.073

- Considered dropping Documentaries because of F-stat:
  - Full model F-stat:
    - 1318
  - Reduced model F-stat:
    - 1449
  - Both very significant, but it didn't hurt so I left it in.

Regression

OLS Regression Results

Dep. Variable:	LOG_GROSS	**R-squared:**	0.497
Model:	OLS	**Adj. R-squared:**	0.497
Method:	Least Squares	**F-statistic:**	1318.
Date:	Fri, 02 Feb 2018	**Prob (F-statistic):**	0.00
Time:	09:06:25	**Log-Likelihood:**	-33337.
No. Observations:	14694	**AIC:**	6.670e+04
Df Residuals:	14682	**BIC:**	6.679e+04
Df Model:	11		
Covariance Type:	nonrobust		

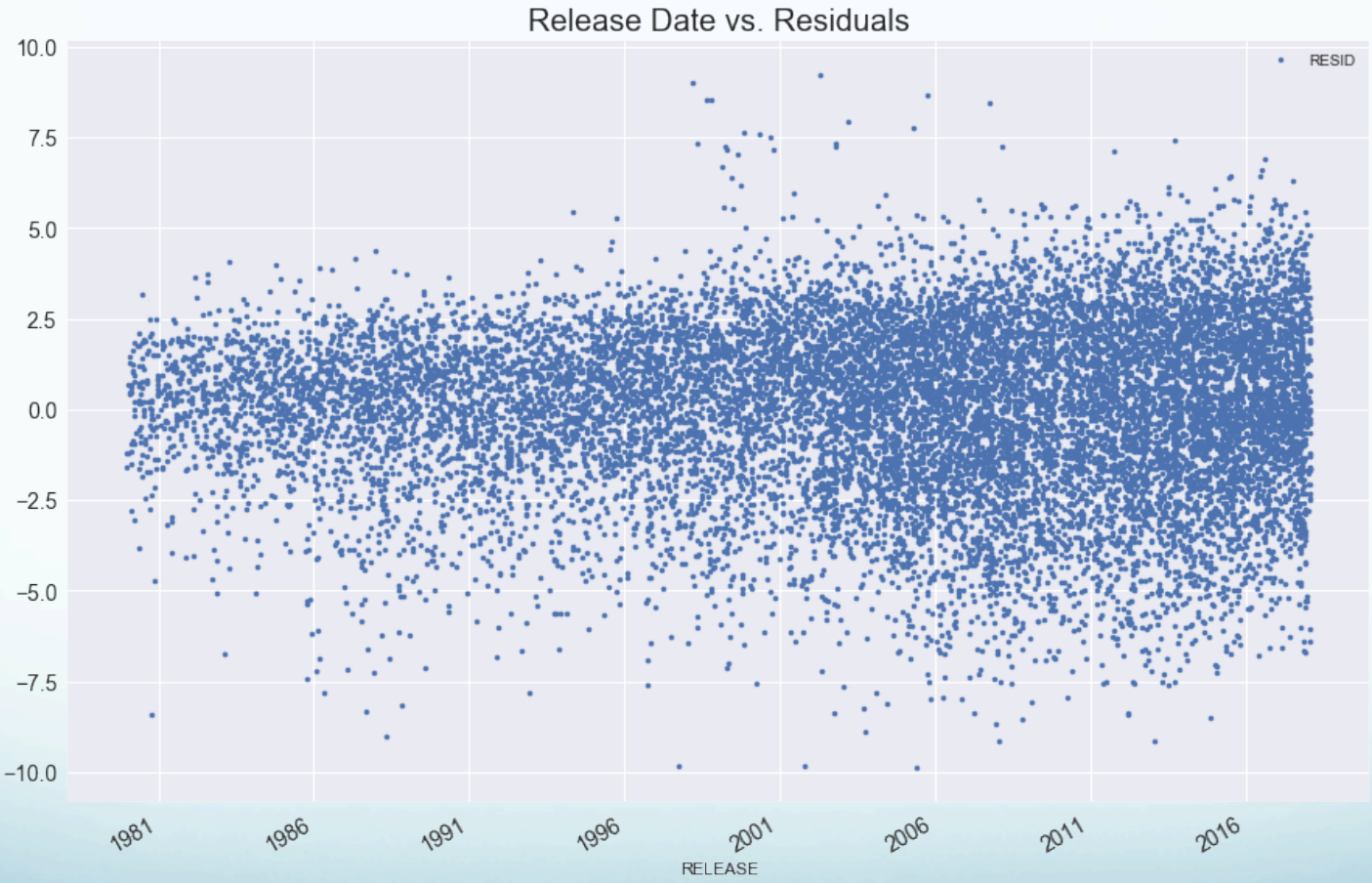
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.5160	0.120	54.205	0.000	6.280	6.752
AGE_Y	0.0904	0.002	41.237	0.000	0.086	0.095
RUNTIME	0.0327	0.001	32.224	0.000	0.031	0.035
G	4.7964	0.133	36.140	0.000	4.536	5.057
PG	3.5706	0.073	49.150	0.000	3.428	3.713
PG13	3.4643	0.063	55.012	0.000	3.341	3.588
R	1.9886	0.056	35.362	0.000	1.878	2.099
Adventure_Action	1.9697	0.070	27.976	0.000	1.832	2.108
Comedy	1.6702	0.059	28.505	0.000	1.555	1.785
Documentary	0.1464	0.072	2.029	0.043	0.005	0.288
Drama	0.5730	0.064	9.019	0.000	0.448	0.698
Thriller_Horror	1.9312	0.072	26.707	0.000	1.789	2.073

- R rated movies are much more risky than the family oriented movies, but can still be fruitful.

Cross Validated Results

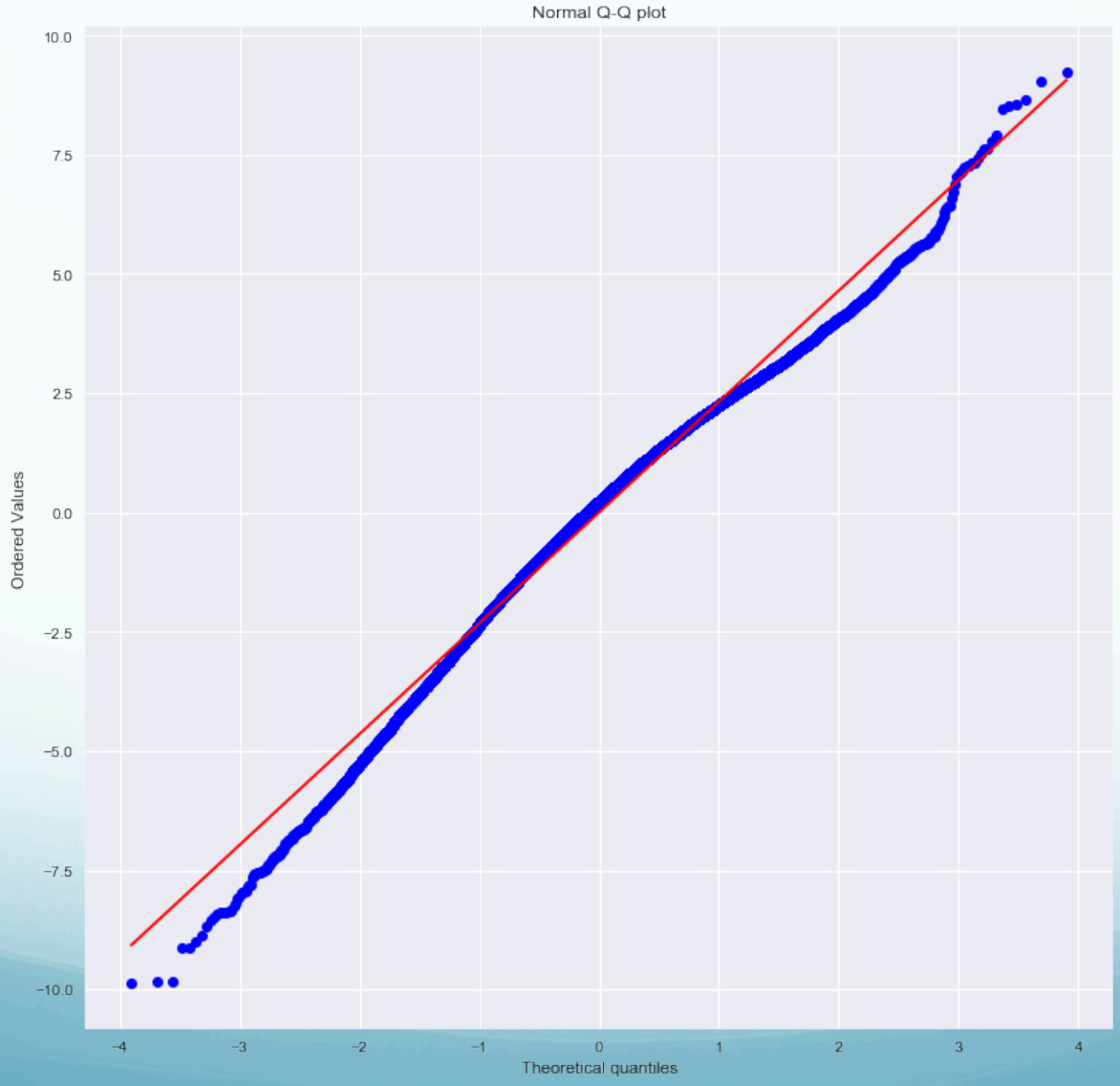
```
OLS regression score val R^2: 0.510

Baseline Coefficients:
 Intercept,  6.573906612683581
[('AGE_Y', 0.08936089546980854),
 ('RUNTIME', 0.032263533145545864),
 ('G', 4.7502111435114145),
 ('PG', 3.5710692976377643),
 ('PG13', 3.4781789023248977),
 ('R', 1.9571410403011749),
 ('Adventure_Action', 1.9588031345231838),
 ('Comedy', 1.6962493508482084),
 ('Documentary', 0.15551388026547674),
 ('Drama', 0.6205693429512681),
 ('Thriller_Horror', 1.9612622999698346)]


OLS regression with interaction terms val R^2: 0.553
OLS regression with interaction3 terms val R^2: 0.555
OLS regression with polynomial terms val R^2: 0.555
OLS regression with polynomial3 terms val R^2: 0.568
Random Forest R^2: 0.508
Gradient Boosting R^2: 0.570
```

Diagnostics

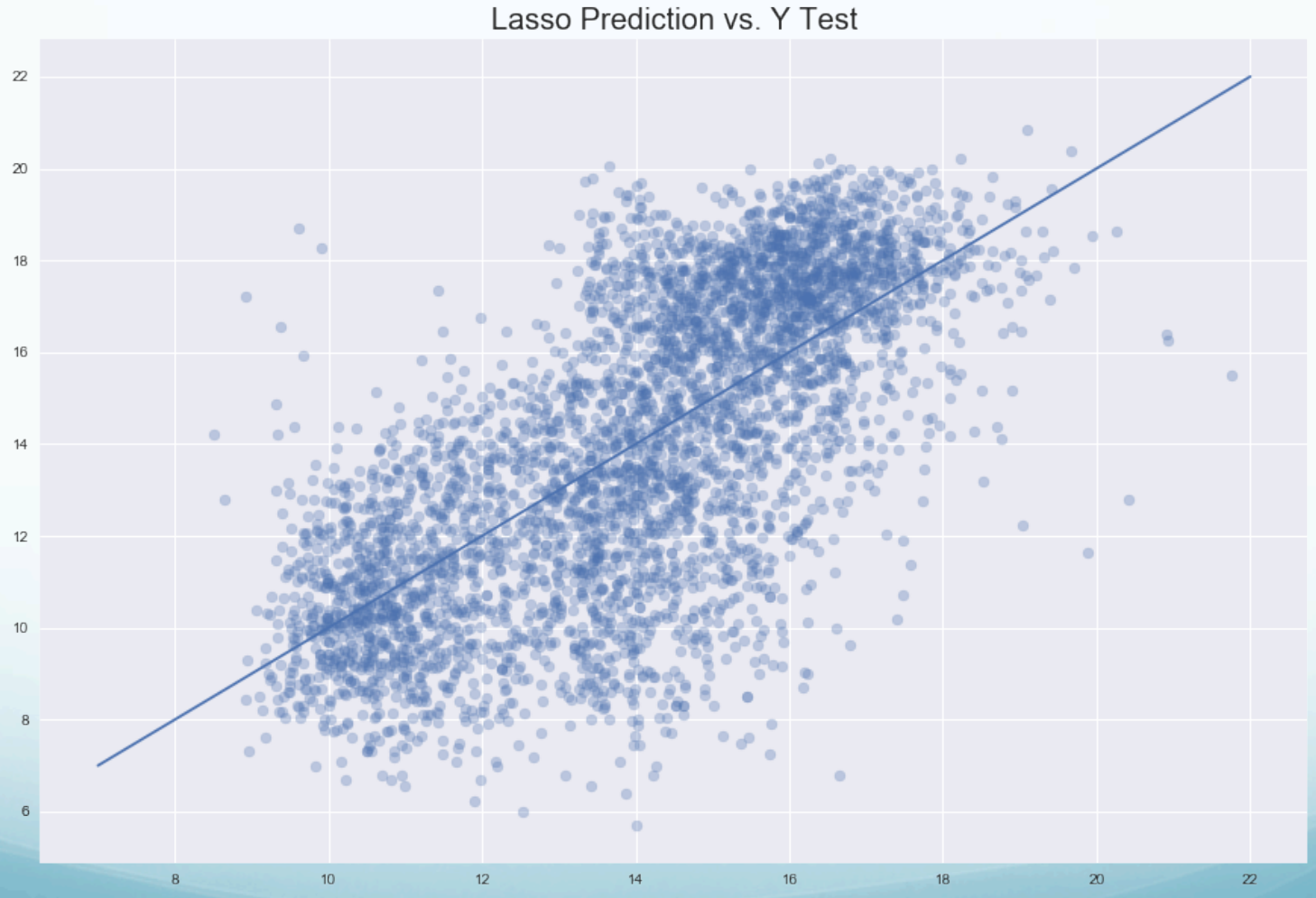


Diagnostics

Diagnostics

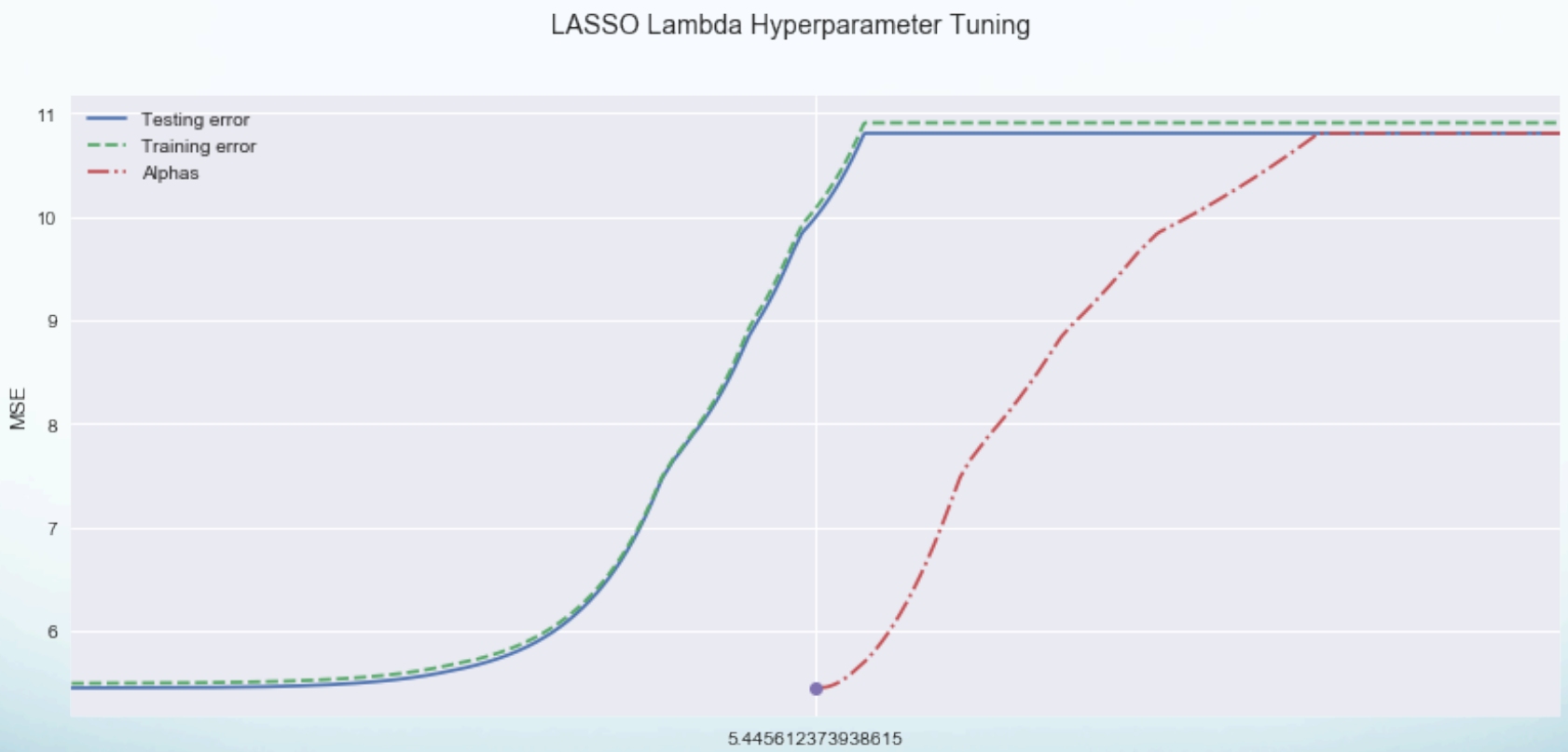


Cross Validated 5-Fold LASSO Results



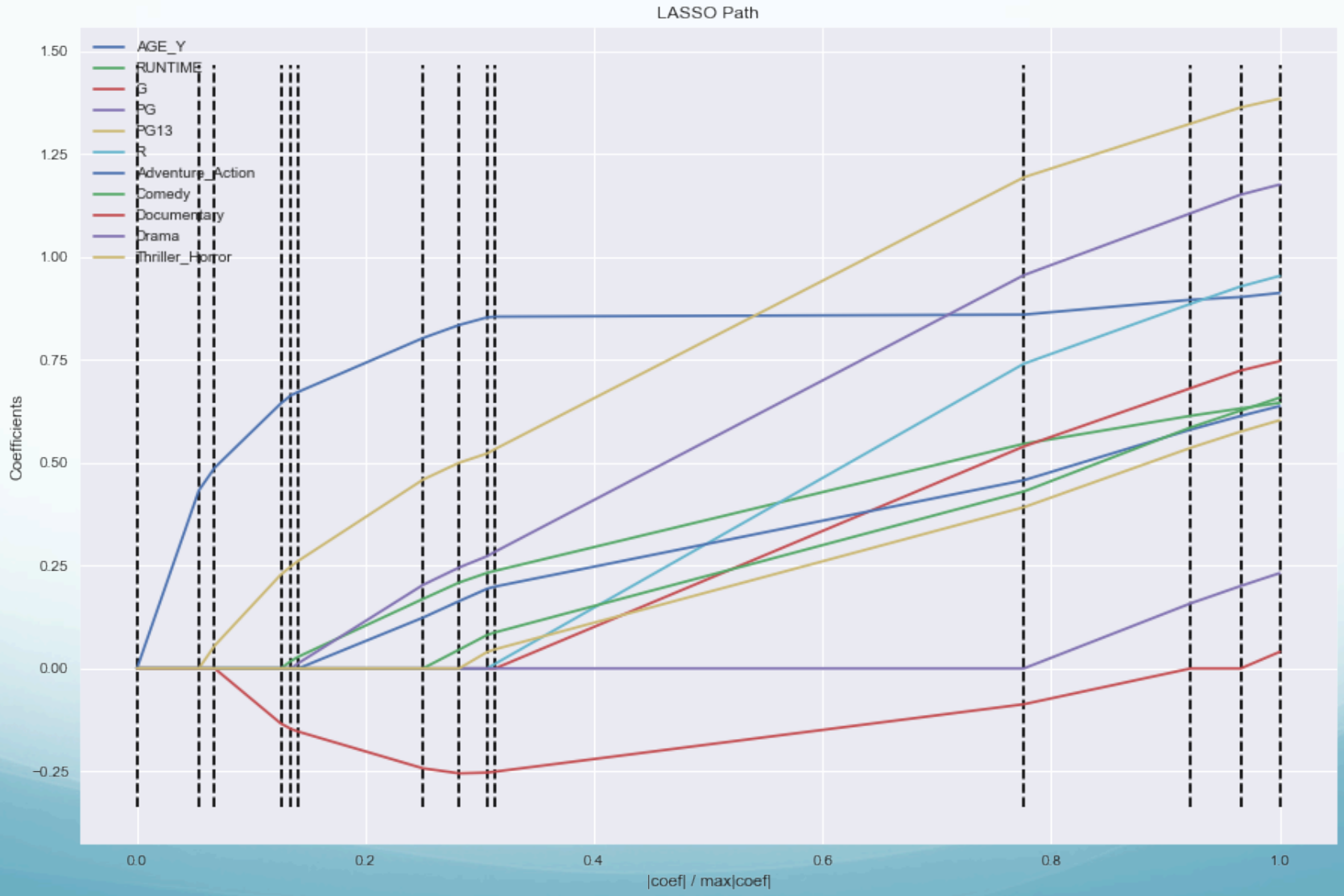
R.Sq = ~0.50

Tuning Lasso



Optimal Lambda = 0.01

Cross Validated 5-Fold LASSO



CV 5-KF LASSO Results

```
[('AGE_Y', 0.9027405463203023),
 ('RUNTIME', 0.6319155890441316),
 ('G', 0.72407863165989),
 ('PG', 1.151432554538883),
 ('PG13', 1.3633193500018281),
 ('R', 0.9283346652616568),
 ('Adventure_Action', 0.6138215741041063),
 ('Comedy', 0.6268045404368685),
 ('Documentary', 0.0002152687849560583),
 ('Drama', 0.2006833159007831 4),
 ('Thriller_Horror', 0.5758535381002862)]
```

R.sq. = ~ .50

Next Steps

- Things to do:
  - Explore worldwide grossing
  - Try budget on movie **revenue** (gross – budget)
  - Decision trees: most important features
  - Look over specific segmentations
  - Manual best subsets regression