Joshua Banks Mailman

Predicting film-director ratings

Questions investigated

Characterizing response to film directors

The investor, sponsor, or granting agency that funds film directors may be interested in...

- Besides profit margins they may be interested in alternative gauges of success of a movie

- Crowd-sourced critical response to the films of prolific acclaimed directors are a valuable resource to exploit to address this.

Predicting the ratings of their films

Based on data freely available to the general public, it may be possible to a build predictive models tailored to various film directors of the past.

- Such types of models can also be applied to directors that are still active.

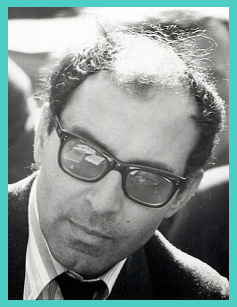
- Most of the film directors examined are still active today.

Who?

Seven acclaimed directors who are prolific

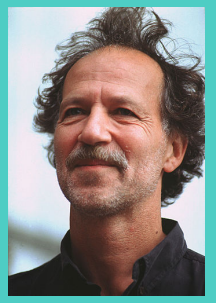
Alfred Hitchcock

Ingmar Bergman

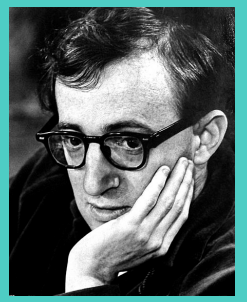


Jean-Luc Godard

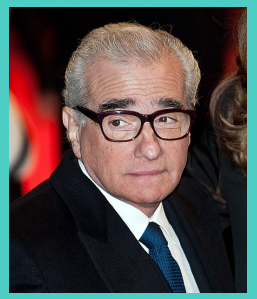
Rainer Fassbinder



Werner Herzog



Woody Allen



Martin Scorsese

Source

IMDB

(Internet Movie Database

- A ***nested*** **web-scraping algorithm** was required, since IMDB lists all of a director's films on one page, but includes cast members and other info on a separate page dedicated to each film

- To improve on the ***worldwide gross*** box office returns data, www.the-numbers.com, was also scraped and joined, but this data was ultimately **discarded** for being incomplete

Target variable

Viewers' Ratings

(scale from 1 to 10)

Target variable

Viewers' Ratings

Features examined

continuous

- Year released
- Duration (in minutes)
- Rating count (total number of times the film was rated on IMDB)
- Budget

categorical

- TV series vs. theatrical release
- Documentary or not
- Cinematographer
- Cast members

7 directors x ~50 films each
x 15 actors & 10 other features

> **9,000**
data points

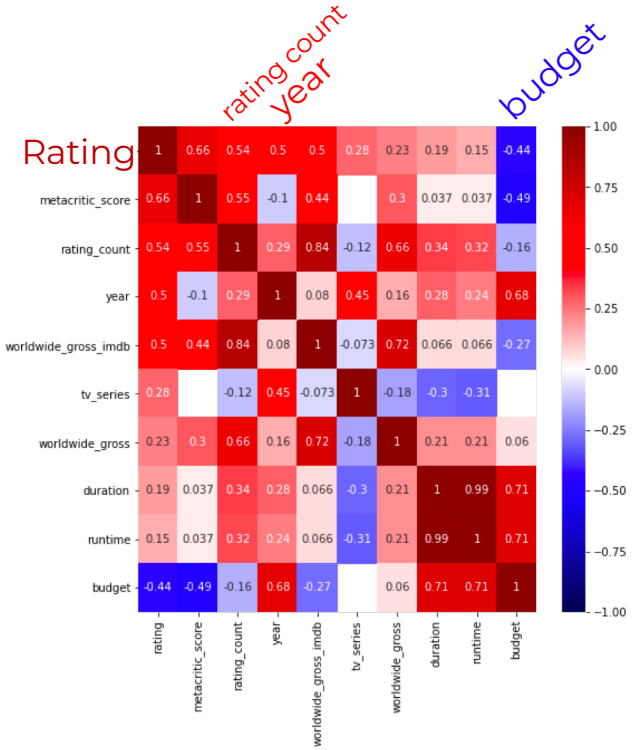
Cleaned and Filtered

As follows:

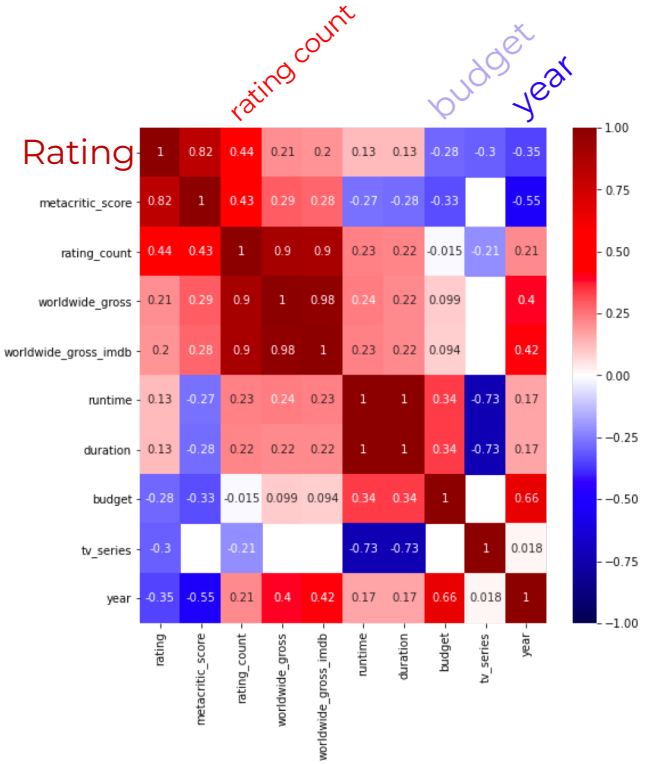
- Exclude actors appearing in < 3 films & one-off cinematographers
- Exclude films < 60 minutes
- Exclude films with crucial missing data and obvious outliers
- Exclude collinear features and those too often missing

Patterns in the data

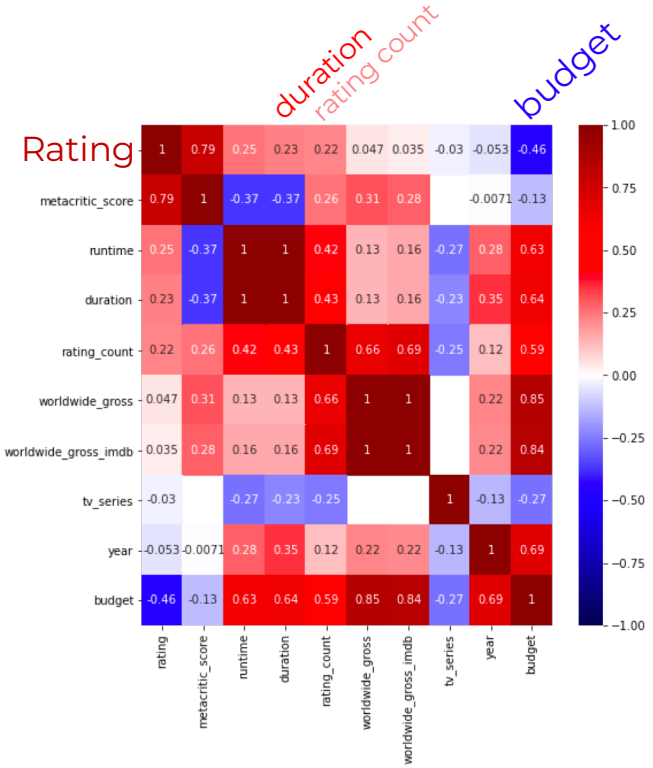
Heatmap sorted by correlation to the target variable



Hitchcock



Allen



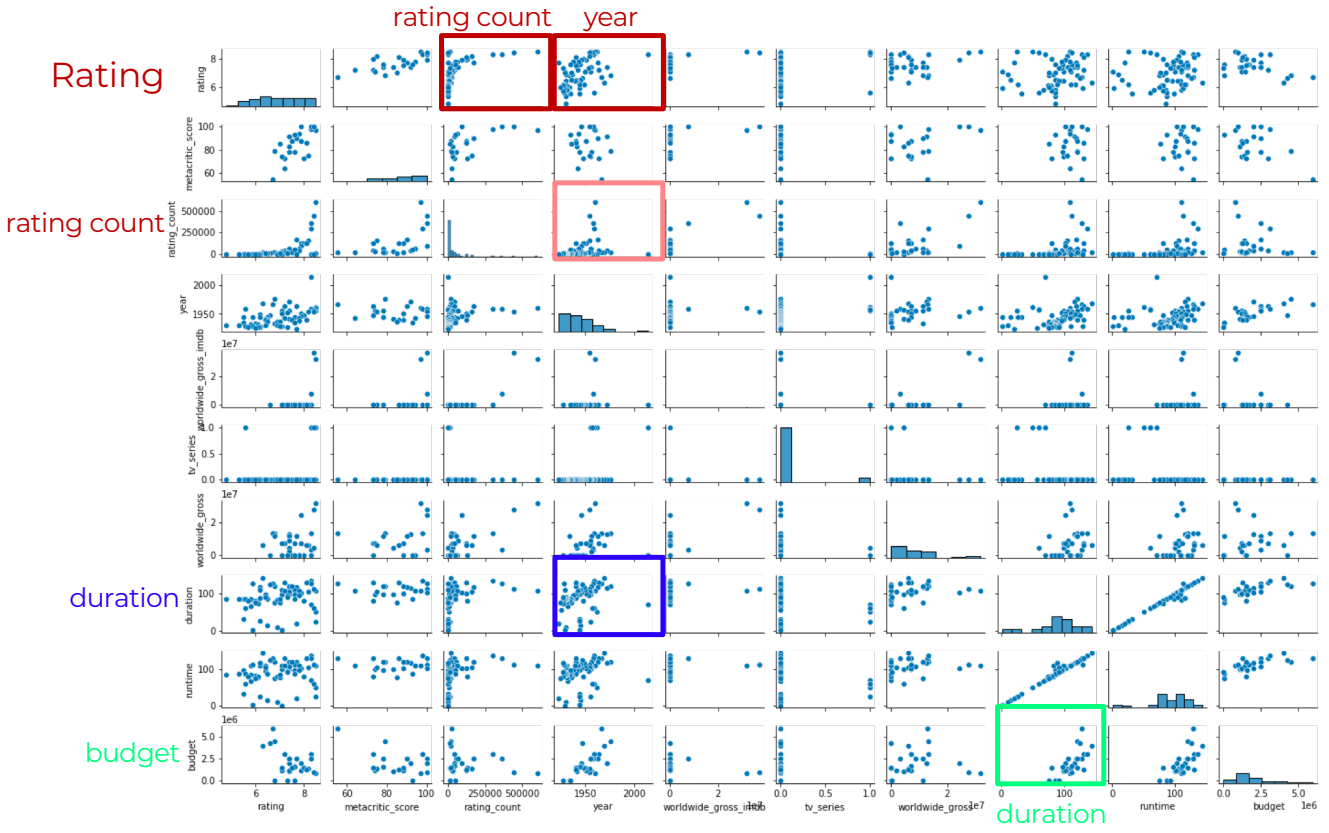
Herzog

Alfred Hitchcock

Patterns in the data

Multicollinearity as expected between features and proxies for them, such as *rating* and *metacritic score*, and *runtime* and *duration*

Shapes of correlation



To predict ratings:

Linear Regression

with train-test split to enable cross-validation

continuous features

- Year released
- Duration (in minutes)
- Rating count (total number of times the film was rated on IMDB)

categorical features

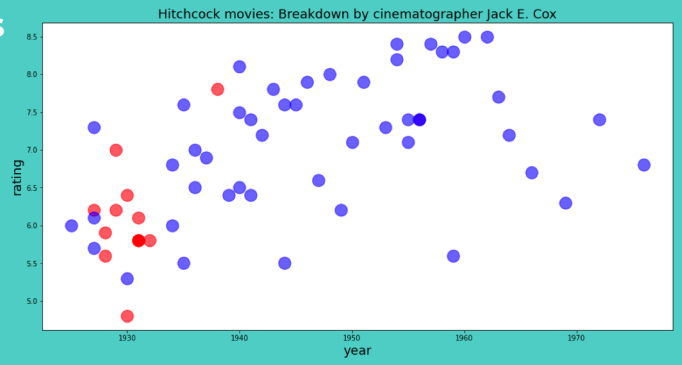
- TV series vs. theatrical release
- Documentary or not
- Cinematographer (excluding those who worked the least number of times with that director)
- Cast members (if appearing in > 3 films)

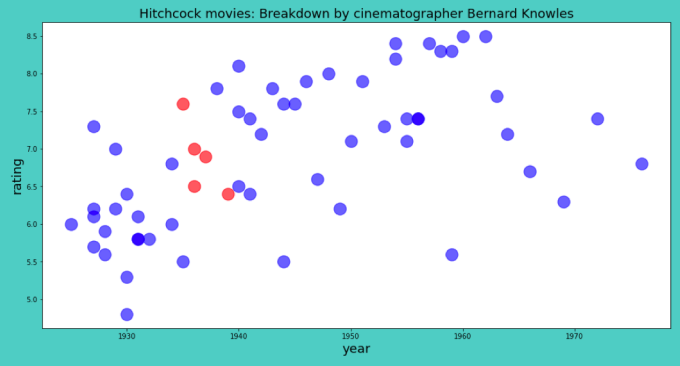
Three of Hitchcock's cinematographers

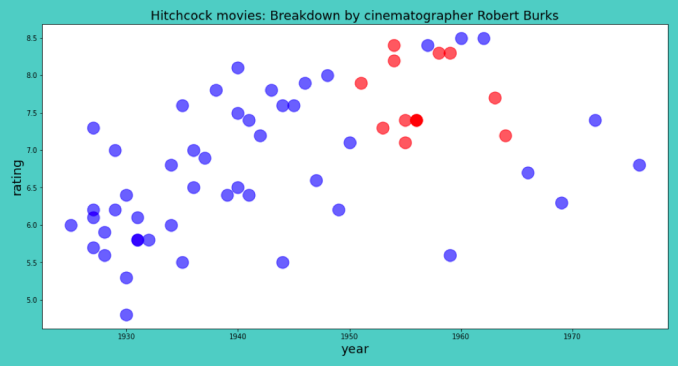
Jack E. Cox
(1920-30s)

Bernard Knowles
(late 1930s)

Robert Burks
(1950s-60s)







Alfred Hitchcock and other directors, use a particular cinematographer during a particular phase of their career

Therefore *cinematographer* somewhat serves as a proxy for the time variable in years

Linear regression models

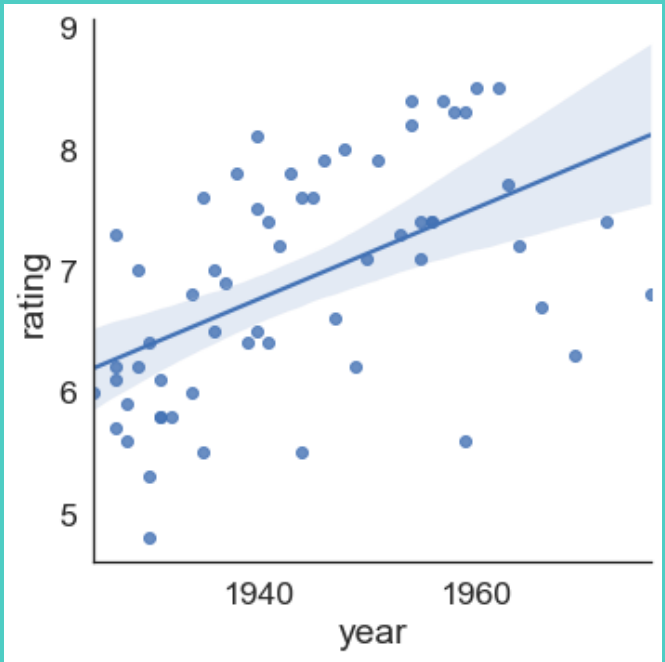
R^2 values on test-set after training

Feature sets	Hitchcock	Bergman	Godard	Allen	*Allen**	Fassbinder	Herzog	Scorsese
All (continuous & categories)	-0.97	-1.41	-0.09	-14.3	*0.21*	-1.26	-0.07	-0.41
Continuous only	0.38	0.15	0.26	**0.51**	0.4	0.27	0.07	0.17
Year	0.28	-0.15	-0.11	0.24	0.04	-0.08	0.04	-0.18
Year & **Rating count**	0.37	-0.1	0.11	**0.51**	**0.46**	0.21	0.06	-0.19
Year & **Duration** only	0.27	0.16	0.25	0.23	-0.04	-0.02	0.03	0.19

* Including two additional Allen films excluded for their genre and duration yielded slightly different results

How Hitchcock's ratings rise

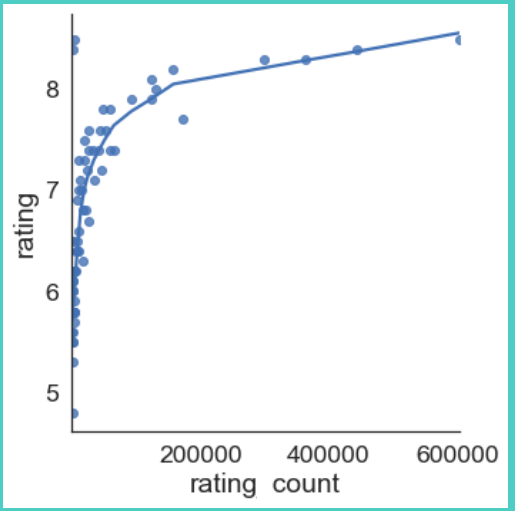
Hitchcock movies tend to be rated higher to the extent they are released later in his career and are rated by many

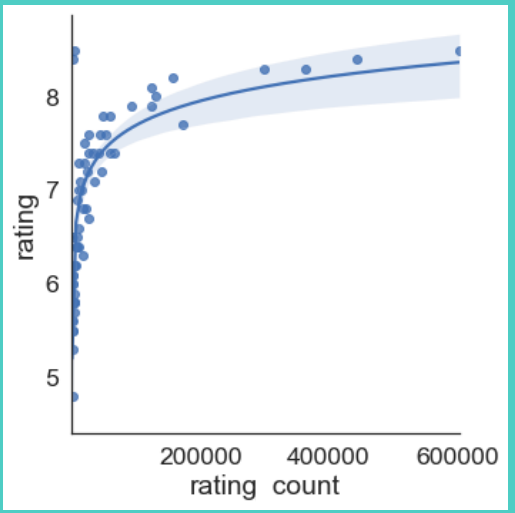


Feature engineering

Regression on log of ratings count

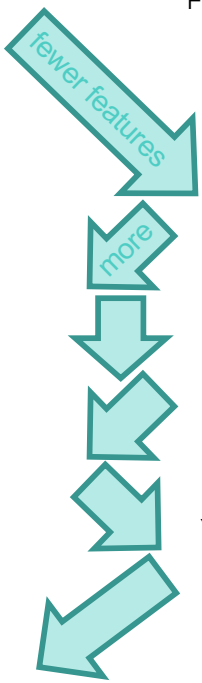
Smoothed moving average *(lowess)*





Linear regression models

R^2 values on test-set after training



Feature sets	Hitchcock	Bergman	Godard	Allen	*Allen**	Fassbinder	Herzog	Scorsese
All (continuous & categories)	-0.97	-1.41	-0.09	-14.3	*0.21*	-1.26	-0.07	-0.41
Continuous only	0.38	0.15	0.26	**0.51**	0.4	0.27	0.07	0.17
Year	0.28	-0.15	-0.11	0.24	0.04	-0.08	0.04	-0.18
Year & **Rating count**	0.37	-0.1	0.11	**0.51**	**0.46**	0.21	0.06	-0.19
Year & **Duration** only	0.27	0.16	0.25	0.23	-0.04	-0.02	0.03	0.19
Year & **Rating count** & their **logs**	**0.8**	**0.84**	0.35	**0.48**	0.37	**0.48**	**0.16**	-0.22
Year & **log(rating count)**	**0.78**	**0.8**	**0.34**	0.36	**0.45**	0.12	0.11	**0.25**
Lasso CV with polynomials on all features	***0.68***			***0.41***				
Include **budget** but w/ smaller dataset								**0.69**

* Including two additional Allen films excluded for their genre and duration yielded slightly different results

Linear regression models

R^2 values on test-set after training

slightly more stable
(less variance, less overfitting)

Hitchcock

Lasso regression on polynomial features, with cross validation

Three trials:	validation	test
	-0.09	0.51
	0.75	0.4
	0.32	0.57
std	0.41851	0.08469

Less stable
(higher variance, overfitting)

Allen

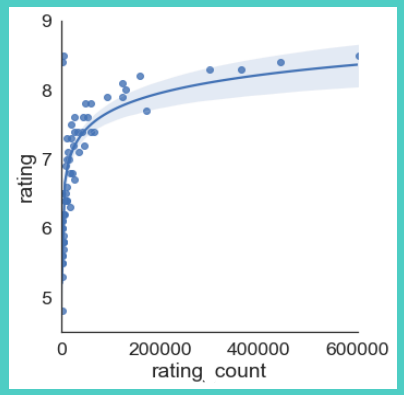
Lasso regression on polynomial features, with cross validation

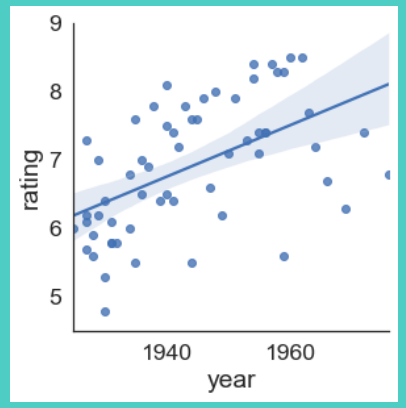
Three trials:	validation	test
	-7.32	0.46
	-0.9	0.29
	-3.15	0.28
std	3.257	0.09987

		year,	*duration*,	*rating count*,	log(*year*),	log(*rating count*)
Hitchcock	Important coefficients:	-50.668,	-0.198,	0.101,	50.892,	0.560

Opposite trends

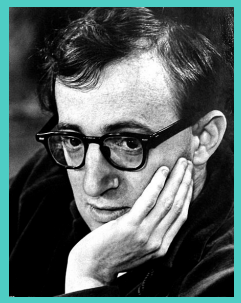
Hitchcock

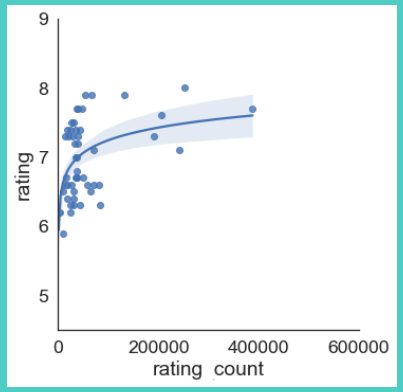


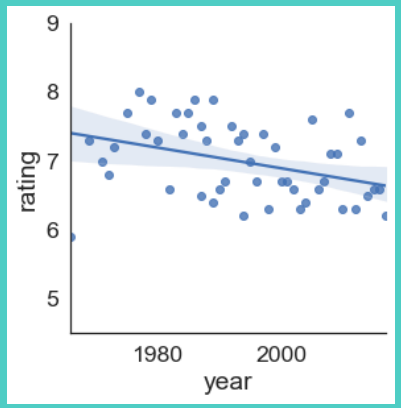


	R^2 on test set
Year & Ratings count	0.37
Year & Duration only	0.28
Year & Ratings count & their **logs**	**0.8**
Year & log(rating count)	**0.78**
Lasso CV with polynomials on all features	***0.68***

Allen







	R^2 on test set
Year & Ratings count	**0.51**
Year & Duration only	0.23
Year & Ratings count & their **logs**	**0.48**
Year & log(rating count)	0.36

R^2 values shown in blue represent the best model for the given director, based on comparisons of *mean absolute error* (*mae*) between different sets of features and rows, and within the same set of features and rows, choosing Lasso vs. Ridge based on lower R^2

* ***Post hoc***, the best model (mae=0.3217; r^2=**0.1573**) for Fassbinder was found with Ridge CV using only *year, duration, budget*, with their logs, and polynomial features. The highest coefficients (all positive) are year^2, year x duration, log(year) x duration, duration, and duration x log(budget).

Linear regression models

A more thorough investigation using Lasso CV on all features

w/ rating count

		Hitchcock	Bergman	Godard	Fassbinder*	Herzog	Allen	Scorsese
Lasso CV	r^2	**0.7979**	**0.4737**	**-0.0246**	**0.0060**	**0.0968**	**0.2947**	**0.3369**
	mae	*0.3070*	0.4795	0.4863	0.3856	0.5787	0.3606	0.3049

Polynomial Features

w/ rating count

		Hitchcock	Bergman	Godard	Fassbinder	Herzog	Allen	Scorsese
Lasso CV	r^2	**0.7695**	**0.2721**	**0.0893**	**-0.4796**	**0.2804**	**0.2571**	**-1.6643**
	mae	0.3297	0.5312	0.4864	0.5080	0.5526	0.3464	0.3483

Contributors
(positive coefficients)

Greatest coefficients*:

Detractors
(negative coefficients)

* Font sizes are roughly proportional to absolute value of coefficients

DOP stands for *director of photography*, which is the cinematographer

Caveat:

Although average *rating* of a movie could vary up or down regardless of the *rating count*, the particular ratings can only exist in so far as there already exist ratings to count up. Therefore there is some question of:

Reflexivity

Therefore, can the ratings be predicted or interpreted *without* using *rating count* as a feature?

* ***Post hoc***, the best model (mae=0.3217; r^2=**0.1573**) for Fassbinder was found with Ridge CV using only *year, duration, budget,* with their logs, and polynomial features. The highest coefficients (all positive) are year^2, year x duration, log(year) x duration, duration, and duration x log(budget).

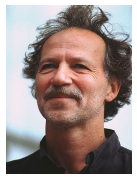
Linear regression models without *rating count* as a feature

A more thorough investigation using Lasso CV on all features

		Hitchcock	Bergman	Godard	Fassbinder *	Herzog	Allen	Scorsese
w/out rating count								
Lasso CV	r^2	**0.1460**	**0.0493**	**0.0756**	**-0.3318**	**-0.0343**	**0.1838**	**-0.0687**
Ridge CV		**0.3806**	**0.2071**	**-0.1495**	**-0.2123**	**0.0726**	**0.1070**	**-0.0527**
Lasso CV	mae	0.5909	0.6981	0.4722	0.4211	0.6497	0.3904	0.3505
Ridge CV		0.5097	0.6138	0.5023	0.4375	0.6325	0.3489	0.3709
Polynomial Features								
w/out rating count								
Lasso CV	r^2	**0.0955**	**0.0522**	**-0.3081**	**-0.5940**	**0.2086**	**0.1925**	**-0.0687**
Ridge CV		**0.3285**	**0.4565**	**-0.0069**	**-0.2809**	**0.1282**	**-0.1373**	**-1.3768**
Lasso CV	mae	0.6239	0.6926	0.5072	0.5122	0.6415	0.3859	0.3585
Ridge CV		0.4904	0.5152	0.4767	0.4494	0.6180	0.3826	0.5560







Pairs of *mean absolute error*s (*mae*) are boxed if they are lower than the corresponding *mae* for a different sized set of features here (and slightly different sets of rows shown on the next slide).

* ***Post hoc***, the best model (mae=0.3217; r^2=**0.1573**) for Fassbinder was found with Ridge CV using only *year, duration, budget,* with their logs, and polynomial features. The highest coefficients (all positive) are year^2, year x duration, log(year) x duration, duration, and duration x log(budget).

Linear regression models

using *budget* as a feature

A more thorough investigation using Lasso CV on all features

w/ budget (w/out rating count)		Hitchcock	Bergman	Godard	Fassbinder *	Herzog	Allen	Scorsese
Lasso CV	r^2	**0.2442**	**-0.3382**	**0.1542**	**-0.1116**		**-1.2286**	**-0.2586**
Ridge CV		**0.2264**	**-0.7183**	**-0.0693**	**-0.0269**	insufficient	**-1.3884**	**-0.1430**
Lasso CV	mae	0.3993	0.7500	0.4608	0.3550	data	0.4518	0.5210
Ridge CV		0.4135	0.8255	0.3717	0.3279		0.4553	0.5690

Polynomial Features w/ budget (w/out rating count)								
Lasso CV	r^2	**0.2905**	**-0.4779**	**0.3768**	**-0.1116**		**-1.3102**	**-4.0672**
Ridge CV		**0.1001**	**-0.1964**	**-0.1767**	**-0.6524**	insufficient	**-1.9036**	**-0.6297**
Lasso CV	mae	0.3862	0.8424	0.3952	0.3550	data	0.4518	0.9911
Ridge CV		0.4289	0.92812	0.4942	0.4118		0.4426	0.6627

Contributors
(positive coefficients)

Greatest coefficients*:

Detractors
(negative coefficients)

Georges Staquet
budget x Raoul Coutard (DOP)
duration x J.P. Belmondo
Anna Karina x André S. Labarthe

year

Pairs of *mean absolute error*s (*mae*) are boxed if they are lower than the corresponding *mae* for a different sized set of features here (and slightly different sets of rows shown on the previous slide).

Further investigations

Probe regularization with polynomial features to improve prediction

- Examine quantile-quantile plots and individual interactions

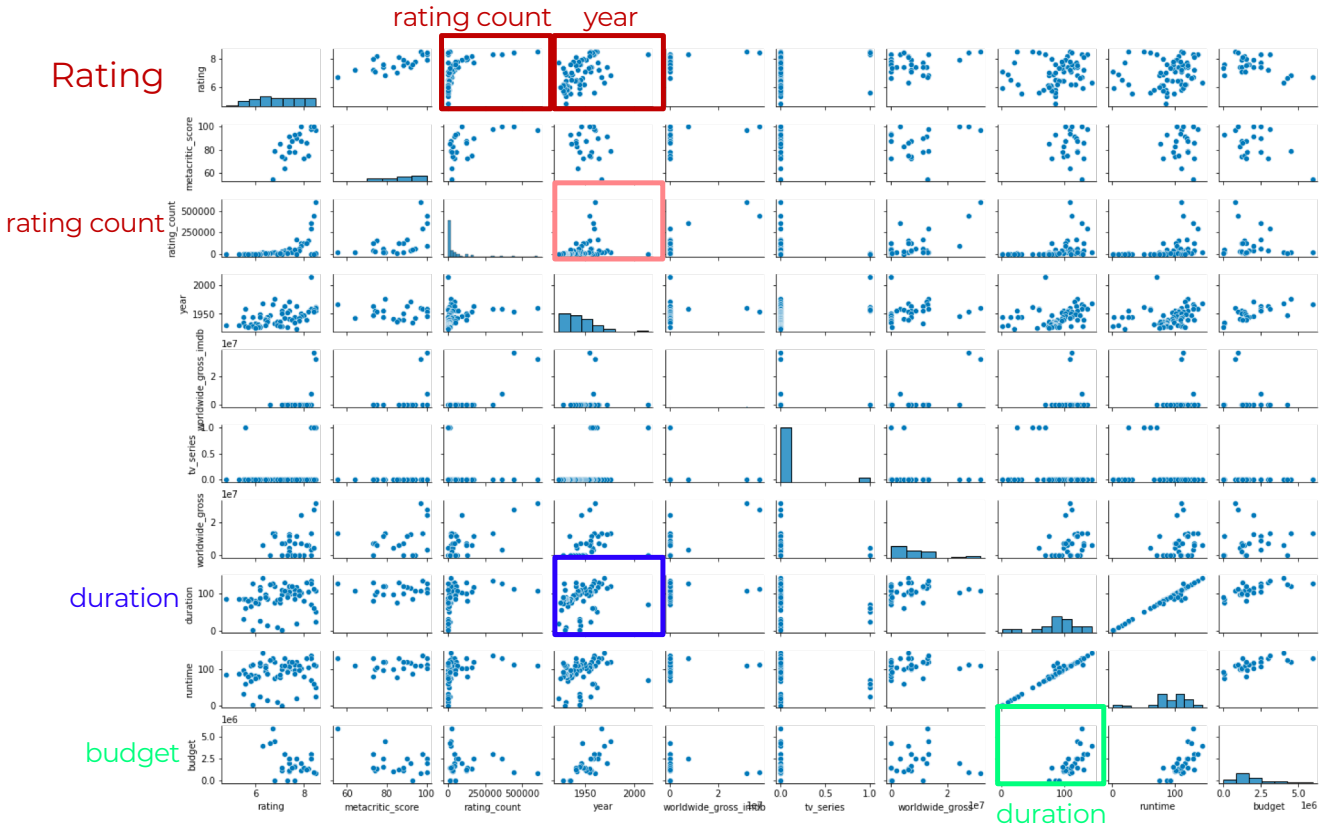
Develop explanation

- Examine **residuals** in tandem with *domain knowledge* to discover and explain **interactions** between features, such as:
  - budget & rating: inversely correlated?
  - duration ~ budget ?
  - more visualizations

Alfred Hitchcock

Further investigations

Shapes of correlation

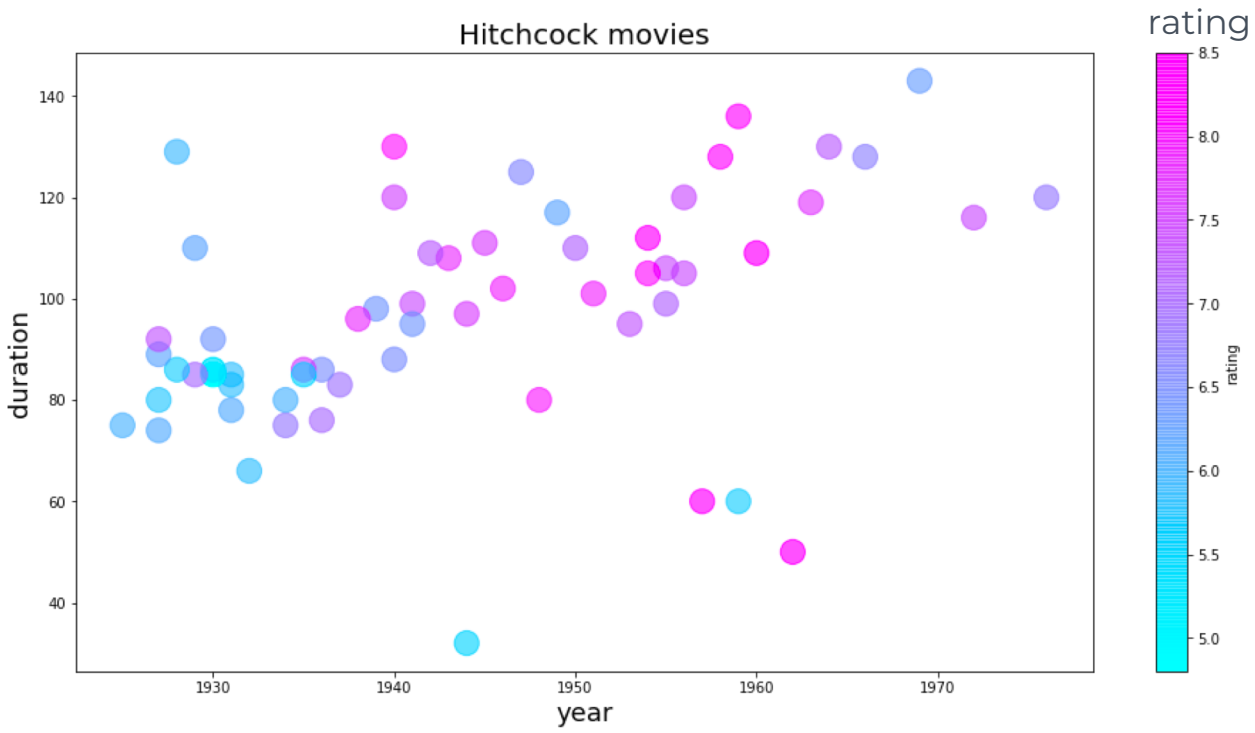


and...

More visualizations for exploratory data analysis

Hitchcock's movies increased in duration

Their ratings have a sweet spot just past the middle



Further investigations

Probe regularization with polynomial features to improve prediction

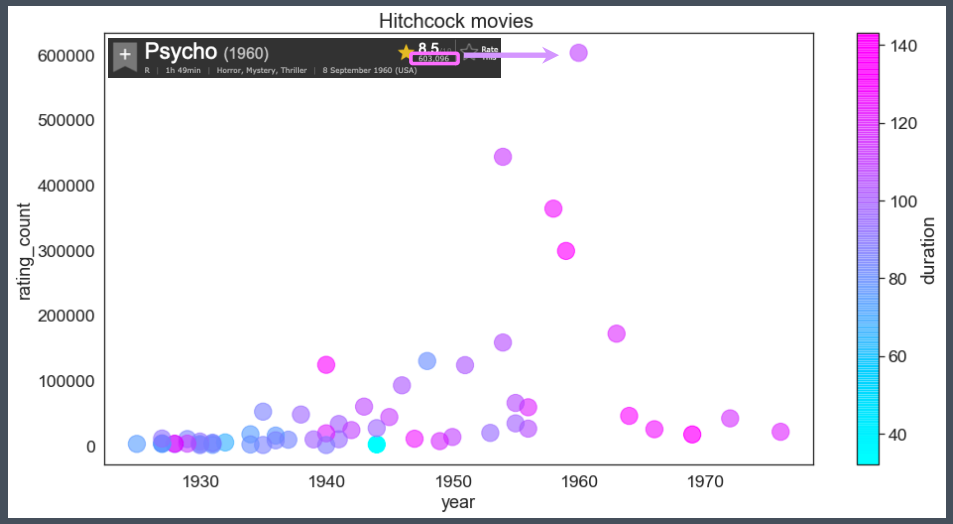
- Examine quantile-quantile plots and individual interactions

Develop explanation

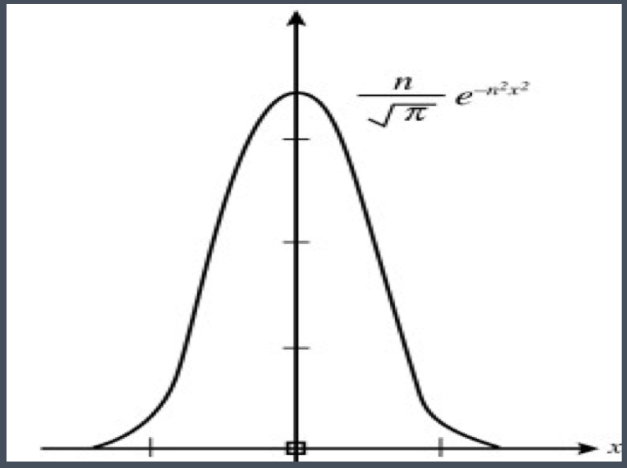
- Examine **residuals** in tandem with *domain knowledge* to discover and explain **interactions** between features, such as:
  - budget & rating: inversely correlated?
  - duration ~ budget ?
  - more visualizations

and...

Feature engineering



Model the spike in rating count, using the *Dirac Delta* function (”impulse” function)



Joshua Banks Mailman

joshuabanksmailman@gmail.com

Thanks!

Predicting film-director ratings