Springboard - Capstone Project Foundations of Data Science

Predicting Movie Quality from Rotten Tomatoes Scores

Storyboard Artist in the Movie Industry

I am a storyboard artist for independent films, working in pre-production.

Storyboards help the director visualize the script into a shot-by-shot sequence.

The cast and the crew know what to do and where to be on each day of the shoot.

Pre-visualization saves time and money for the production.

6.01

7.01

8.01

Movie Review News: Rare but Fascinating

CBS NEWS NEWS SHOWS VIDEO MO

By BOOTIE COSGROVE-MATHER / AP / March 12, 2002, 1:43 PM

Sony Pays For Fake Reviews

Hollywood reel bucks movie money dollars labor / AP

9 Comments / Share / Tweet / Stumble / Email

Sony Pictures Entertainment Inc. has agreed to pay the state $326,000 for using fake reviews attributed to a Connecticut newspaper in promoting its films. Sony also has agreed to stop fabricating movie reviews, and to stop using ads in which Sony employees pose as moviegoers praising films they have just seen, Attorney General Richard Blumenthal said Tuesday.

FiveThirtyEight

Politics Sports Science & Health Economics **Culture**

MAY 18, 2016 AT 3:47 PM

Men Are Sabotaging The Online Reviews Of TV Shows Aimed At Women

By Walt Hickey

Filed under TV

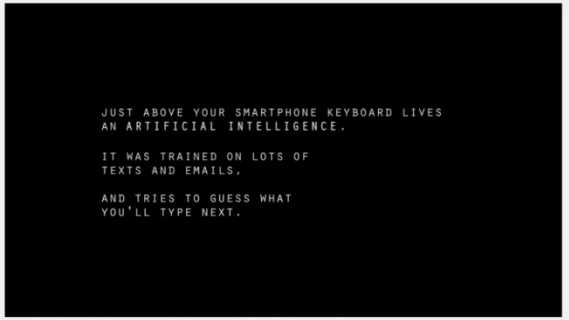
ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE FORUMS

THE MULTIVERSE —

Movie written by algorithm turns out to be hilarious and intense

For *Sunspring*'s exclusive debut on Ars, we talked to the filmmakers about collaborating with an AI.

ANNALEE NEWITZ - 6/9/2016, 3:30 AM



Sunspring, a short science fiction movie written entirely by AI, debuts exclusively on Ars today.

Ars is excited to be hosting this online debut of *Sunspring*, a short science fiction film that's not entirely what it seems. It's about three people living in a weird future, possibly on a space station, probably in a love triangle. You know it's the future because H (played with neurotic gravity by

Good Script Hypothesis

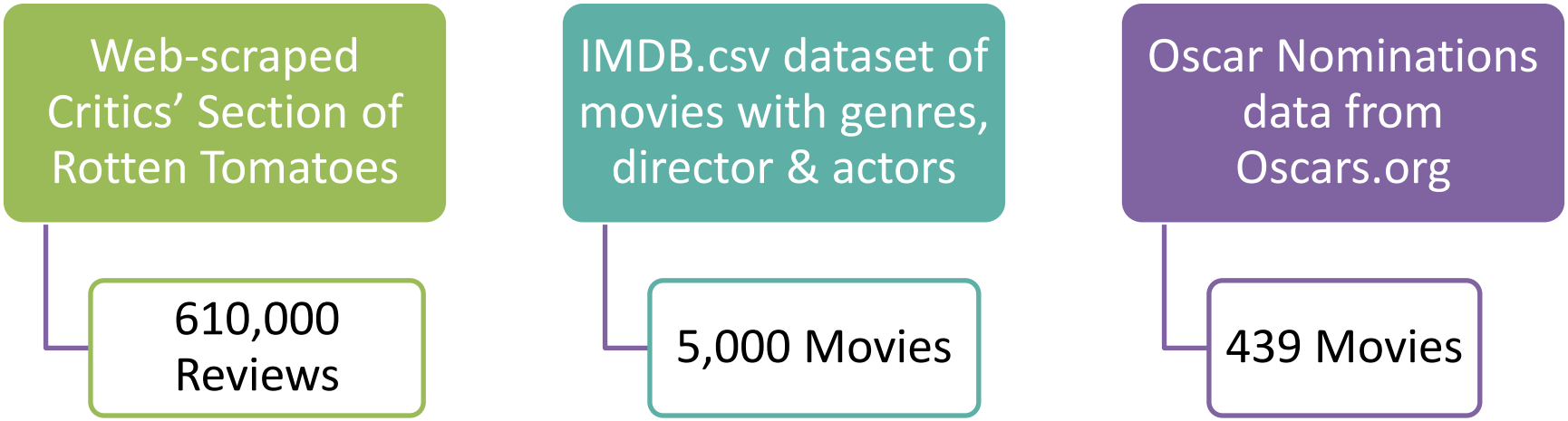
Big name actors and directors do make movies that fail. Star power alone cannot compensate for a bad story

My initial hypothesis is that films with a strong storyline and good writing will be better

Previous Reviews Hypothesis

Can previous good reviews for directors and actors predict future reviews?

Collecting Data

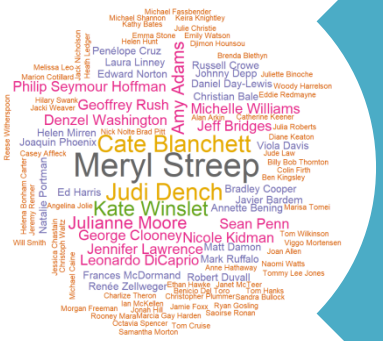


Is Genre Relevant?

Do certain genres or combinations get better reviews or more Oscar nominations?

Are Accolades a Predictor of Quality?

Do famous actors or directors always make acclaimed movies?



Does name recognition bias the critics?

Description of Datasets

Rottentomatoes.com
~1,200 movie critics
~38,000 movie titles
~611,000 reviews

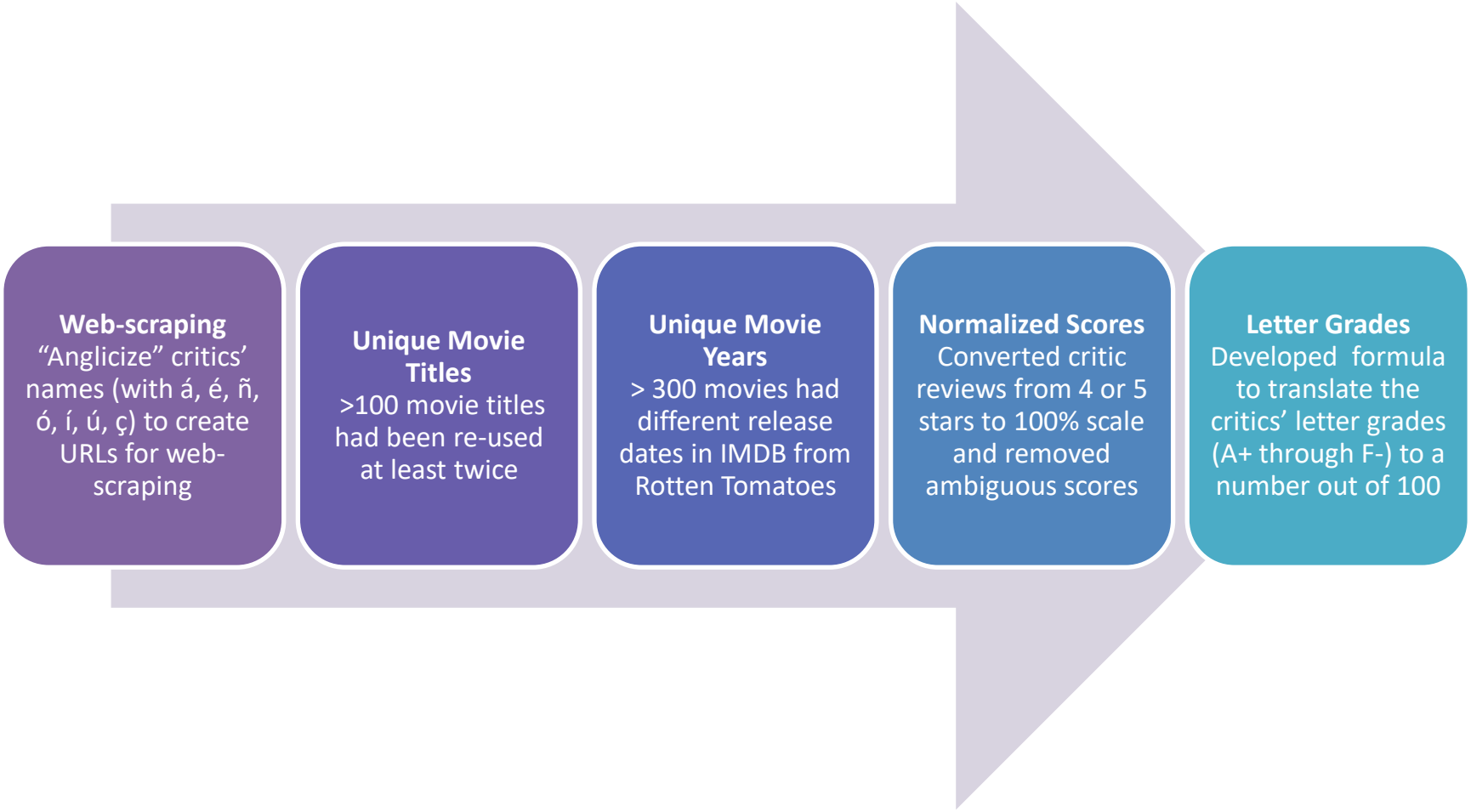
IMDB
3,319 titles: year, genres, director, top three actors, budget and duration

Oscars
21 years of nominations:
439 movie titles
253 actors
74 directors

Training Dataset
1996-2013
2,921 titles

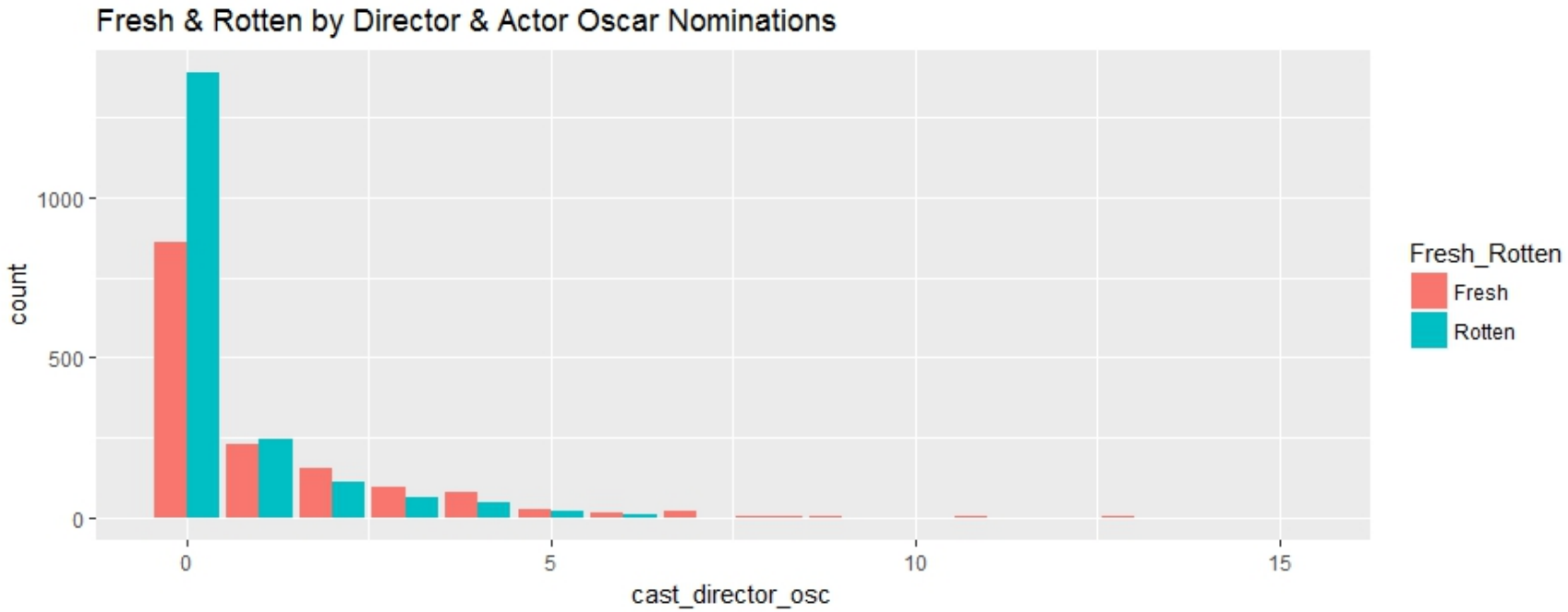
Testing Dataset
2014-2016
398 titles

Data Wrangling



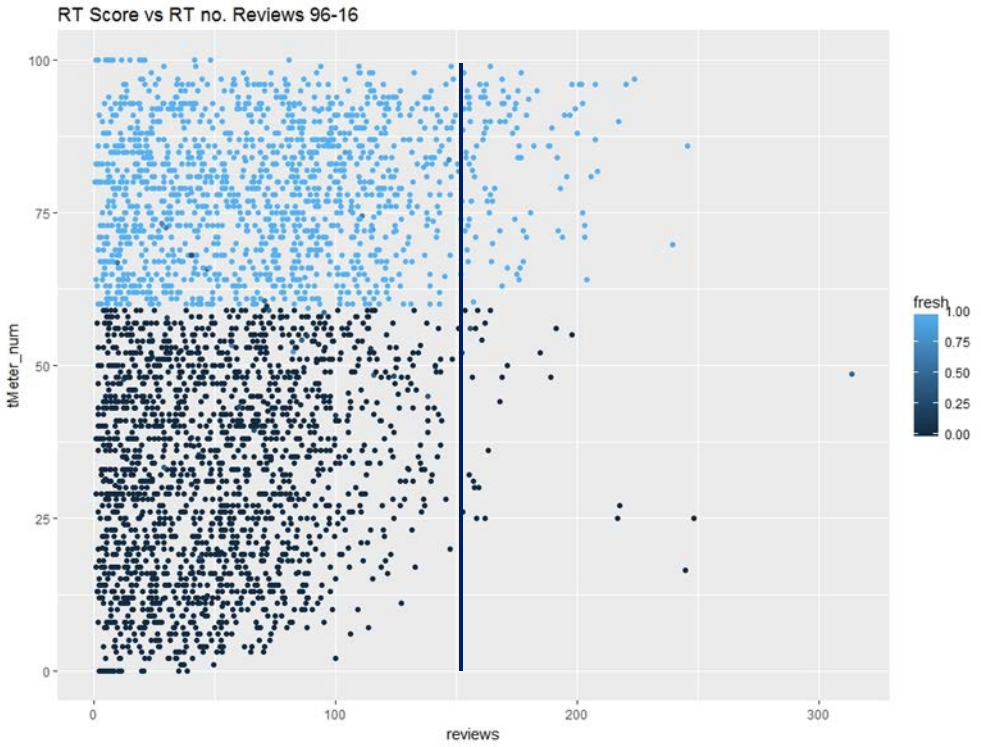
Exploratory Data Analysis

More than 2 career Oscars in a cast and crew suggest a“Fresh” rating *(Tomato Meter Score > 60)*



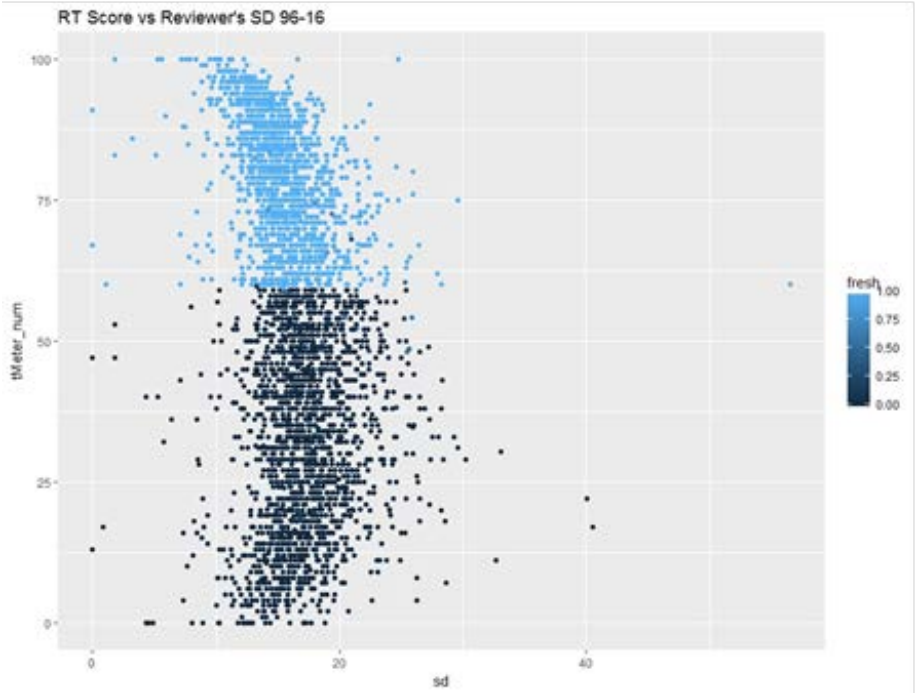
Exploratory Data Analysis

Movies with over 150 reviews have a better chance of having a higher score. Is this a measure of popularity, demand for reviews or publicity?



Exploratory Data Analysis

The standard deviation of movie critic scores is smallest for very good films and very bad films



Modelling
Linear Regression for Tomato Meter Score

Residual standard error:
23.99 on 2681 degrees of freedom

(219 observations deleted due to missingness)

Multiple R-squared: 0.**2243**

Adjusted R-squared: 0.**2174**

F-statistic: 32.31 on 24 and 2681 DF,
p-value: < 2.2e-16

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.318e+01	3.382e+00	9.811	<2e-16	***
Oscars	2.665e+00	2.232e-01	11.938	2e-16	***
budget	-1.893e-09	6.351e-09	-0.298	0.765677	
duration	1.346e-01	2.891e-02	4.657	3.36e-06	***
Action	-6.802e+00	1.426e+00	-4.771	1.93e-06	***
Adventure	2.561e+00	1.548e+00	1.655	0.098106	.
Animation	1.848e+01	2.688e+00	6.875	7.70e-12	***
Biography	4.340e+00	2.163e+00	2.006	0.044926	*
Comedy	-3.727e+00	1.243e+00	-2.997	0.002748	**
Crime	1.194e+00	1.372e+00	0.870	0.384204	
Drama	1.029e+01	1.182e+00	8.709	<2e-16	***
Documentary	3.143e+01	3.545e+00	8.867	<2e-16	***
Family	-1.883e+00	1.943e+00	-0.969	0.332526	
Fantasy	-2.231e-01	1.587e+00	-0.141	0.888252	
History	-5.448e-01	2.852e+00	-0.191	0.848490	
Horror	-3.322e+00	1.814e+00	-1.832	0.067071	.
Romance	-4.462e+00	1.183e+00	-3.771	0.000166	***
Sci_Fi	1.832e+00	1.633e+00	1.122	0.261926	
Sport	-4.837e+00	2.431e+00	-1.990	0.046739	*
Short	4.519e+00	2.432e+01	0.186	0.852606	
Thriller	-4.947e+00	1.333e+00	-3.711	0.000210	***
Musical	4.501e+00	3.425e+00	1.314	0.188816	
Mystery	-2.889e+00	1.627e+00	-1.775	0.075959	.
War	-3.444e+00	2.628e+00	-1.311	0.190092	
Western	-6.044e+00	4.395e+00	-1.375	0.169163	

Modelling
Logistic Regression for “Fresh”

Significant Variables:

- Oscar nominations
- Action Genre
- Animation Genre
- Drama Genre
- Romance Genre
- Thriller Genre

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.391e+00	3.220e-01	-4.319	1.57e-05	***
Oscars	2.654e-01	2.730e-02	9.724	<2e-16	***
budget	1.015e-10	5.452e-10	0.186	0.85225	
duration	8.143e-03	2.762e-03	2.948	0.00319	**
Action	-6.675e-01	1.357e-01	-4.918	8.75e-07	***
Adventure	1.548e-01	1.458e-01	1.062	0.28824	
Animation	1.407e+00	2.449e-01	5.746	9.15e-09	***
Biography	4.849e-01	2.136e-01	2.270	0.02322	*
Comedy	-2.979e-01	1.141e-01	-2.610	0.00905	**
Crime	1.428e-01	1.278e-01	1.117	0.26390	
Drama	6.186e-01	1.089e-01	5.683	1.32e-08	***
Documentary	2.672e+00	4.553e-01	5.867	4.44e-09	***
Family	-3.268e-01	1.848e-01	-1.768	0.07710	.
Fantasy	-1.282e-01	1.487e-01	-0.862	0.38868	
History	-1.681e-01	2.749e-01	-0.612	0.54084	
Horror	-1.146e-01	1.704e-01	-0.673	0.50105	
Romance	-4.528e-01	1.096e-01	-4.132	3.60e-05	***
Sci_Fi	1.894e-01	1.509e-01	1.255	0.20940	
Sport	-2.428e-01	2.281e-01	-1.064	0.28719	
Short	1.052e+01	3.247e+02	0.032	0.97415	
Thriller	-4.973e-01	1.241e-01	-4.008	6.12e-05	***
Musical	1.526e-01	3.082e-01	0.495	0.62049	
Mystery	-1.482e-01	1.504e-01	-0.986	0.32430	
War	-9.609e-02	2.461e-01	-0.390	0.69621	
Western	-1.665e-01	4.191e-01	-0.397	0.69116	

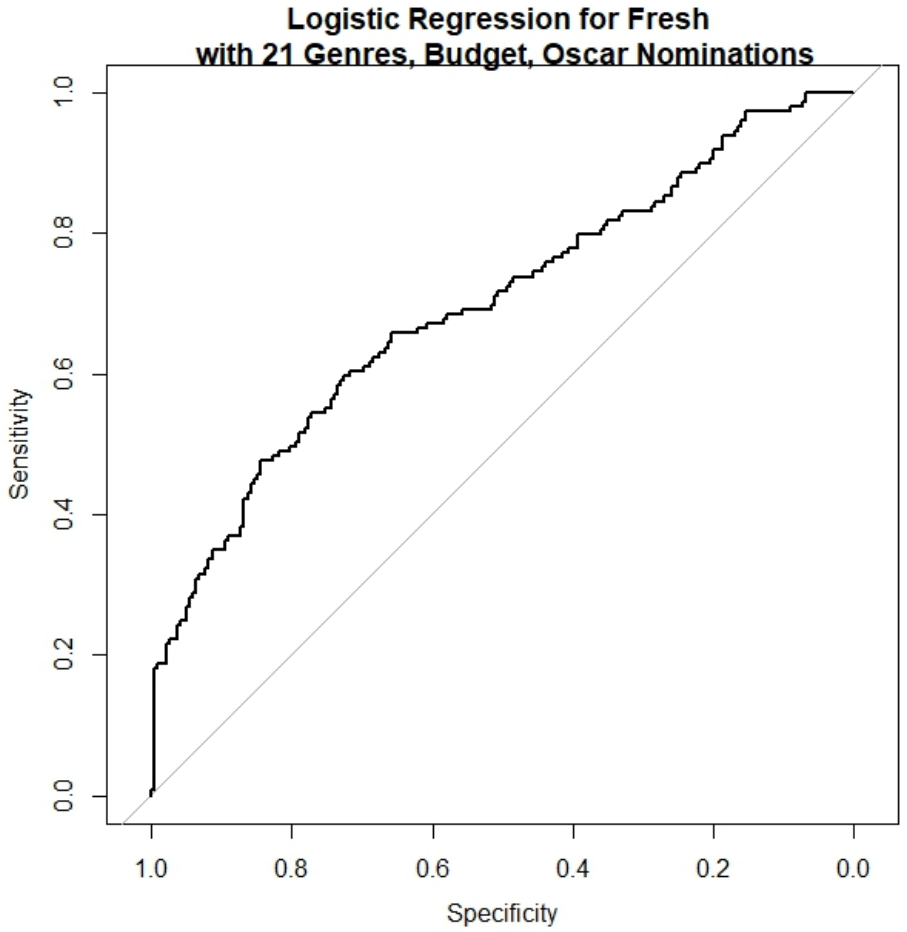
Modelling
Logistic Regression

Logistic Regression model applied to test data Confusion Matrix				
		Target		
		"Fresh"	"Rotten"	
Model	"Fresh"	81	68	"Fresh" Predictive = 81/149 54.4%
	"Rotten"	52	167	"Rotten" Predictive = 167/219 76.3%
		Sensitivity = 81/ 50.9%	Specificity = 167/ 71.1%	Accuracy = 248/368 67.2%

67.2% accuracy against test data

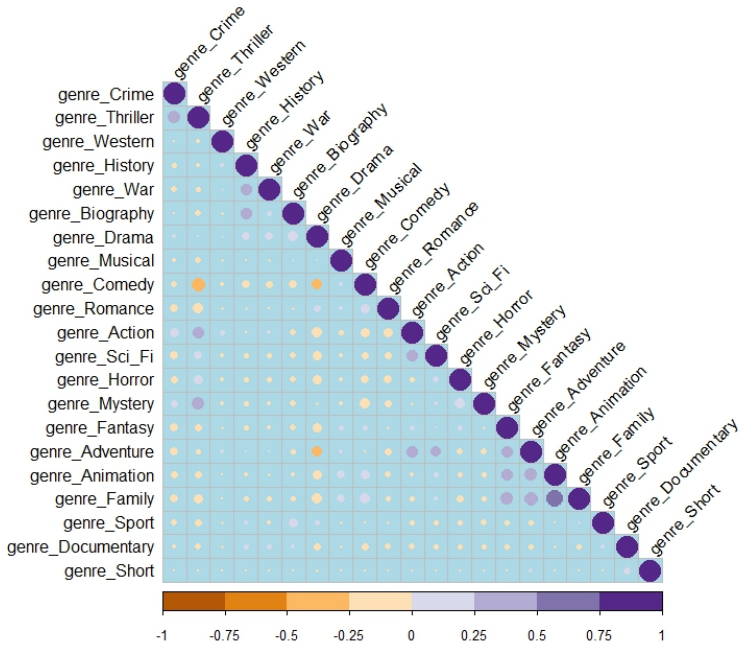
Modelling

ROC curve: using the predictive model from the Logistic Regression the **Area Under the Curve = 0.6965**



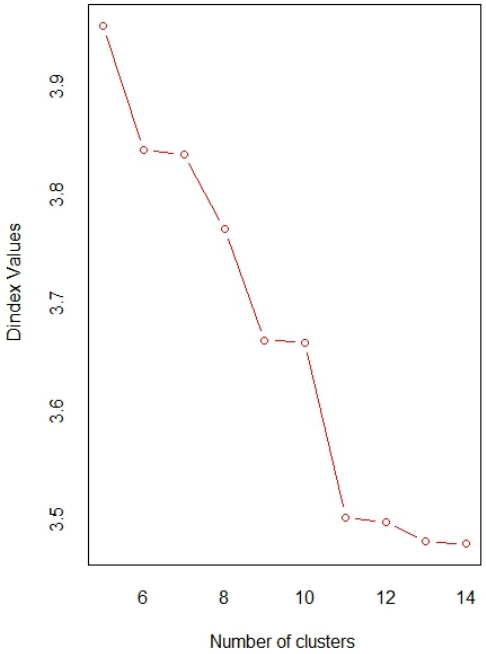
Modelling
Most Frequent Pairing of Genres

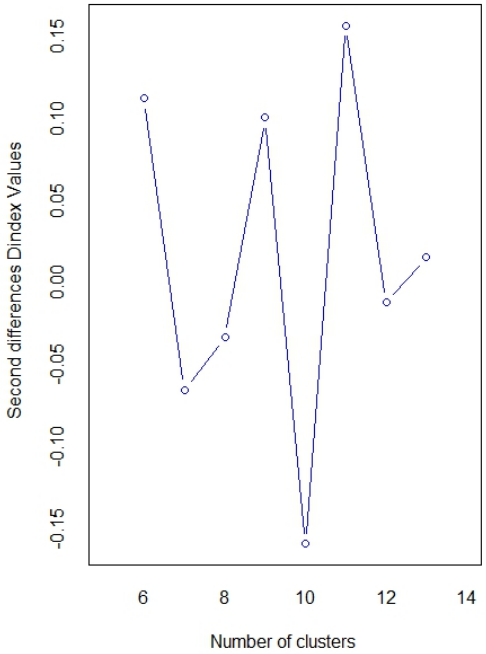
A correlation matrix shows the 14 most common genre pairings



Modelling - Cluster Analysis

The k-means cluster analysis indicates 11 clusters





Modelling
Clustering Genres

Significant genres in each cluster are highlighted

	Cluster Groups										
	1	2	3	4	5	6	7	8	9	10	11
action	0.18	0.32	0	0.01	0.97	0	0	0.22	0	0	0
adventure	0.25	0.02	0	0.02	0.51	0	0	0.82	0	0	0
animation	0.02	0	0	0	0.01	0	0	0.75	0	0	0
biography	0.20	0.01	0	0	0	0	0	0.01	0	0	0
comedy	0.34	0.34	0	0.04	0.13	1	1	0.70	0	1	1
crime	0.07	0.87	0	0	0.11	0	0	0.03	0	0	0
drama	0.66	0.59	1	0.40	0.04	1	0	0.03	1	1	0
documentary	0.08	0	0	0	0	0	0	0	0	0	0
family	0.17	0.01	0	0.01	0.01	0	0	0.96	0	0	0
fantasy	0.17	0.01	0	0.10	0.19	0	0	0.59	0	0	0
history	0.12	0.00	0	0	0	0	0	0	0	0	0
horror	0.05	0.04	0	0.59	0.14	0	0	0.01	0	0	0
romance	0.22	0.20	0	0.00	0.05	1	1	0.06	1	0	0
scifi	0.10	0.01	0	0.19	0.55	0	0	0.19	0	0	0
sport	0.12	0.00	0	0	0.02	0	0	0.02	0	0	0
short	0.00	0	0	0	0	0	0	0	0	0	0
thirller	0.13	0.68	0	0.77	0.54	0	0	0.03	0	0	0
musical	0.04	0.00	0	0	0	0	0	0.10	0	0	0
mystery	0.03	0.23	0	0.44	0.1	0	0	0.07	0	0	0
war	0.12	0.01	0	0	0.01	0	0	0.01	0	0	0
western	0.04	0.00	0	0	0	0	0	0.01	0	0	0
	misc	crime-drama-thriller	drama	thriller-horror-mystery	action-scifi-thriller-adventure	comedy-drama-romance	comedy-romance	adventure-animation-comedy-family	drama-horror	comedy-drama	comedy

Logistic Regression on Clusters

	Estimate	Std Error	z value	Pr(>|z|)	
Cluster 1	-0.03175	0.06735	-0.471	0.63737	
Cluster 2	-0.43619	0.09006	-4.843	1.28e-06	***
Cluster 3	0.50425	0.16216	3.110	0.00187	**
Cluster 4	-0.45042	0.11881	-3.791	0.00015	***
Cluster 5	-0.58027	0.13238	-4.383	1.17e-05	***
Cluster 6	-0.16476	0.16608	-0.992	0.32120	
Cluster 7	-1.64866	0.24416	-6.752	1.46e-11	***
Cluster 8	0.12361	0.15744	0.785	0.43235	
Cluster 9	0.16551	0.19222	0.861	0.38920	
Cluster 10	0.35840	0.17114	2.094	0.03625	*
Cluster 11	-0.94296	0.19379	-4.866	1.14e-06	***

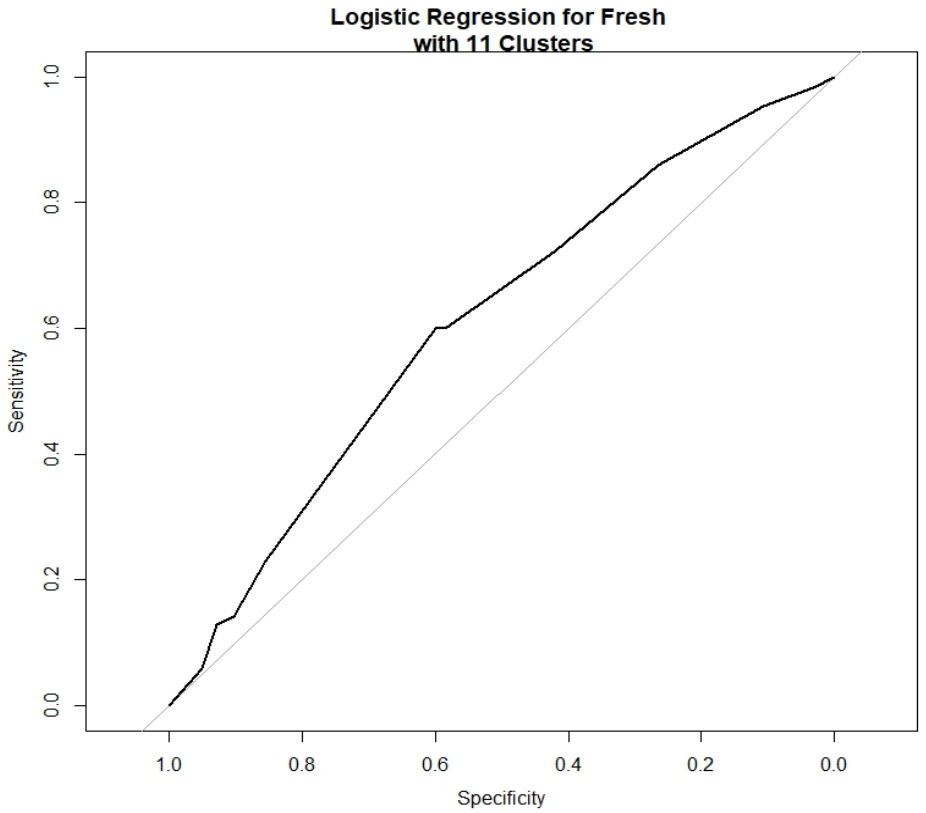
The most significant clusters :

- Group 2: Crime-Drama-Thriller
- Group 3: Drama
- Group 4: Horror-Thriller-Mystery
- Group 5: Action-SciFi-Thriller-Adventure
- Group 7: Romance-Comedy
- Group 11: Comedy

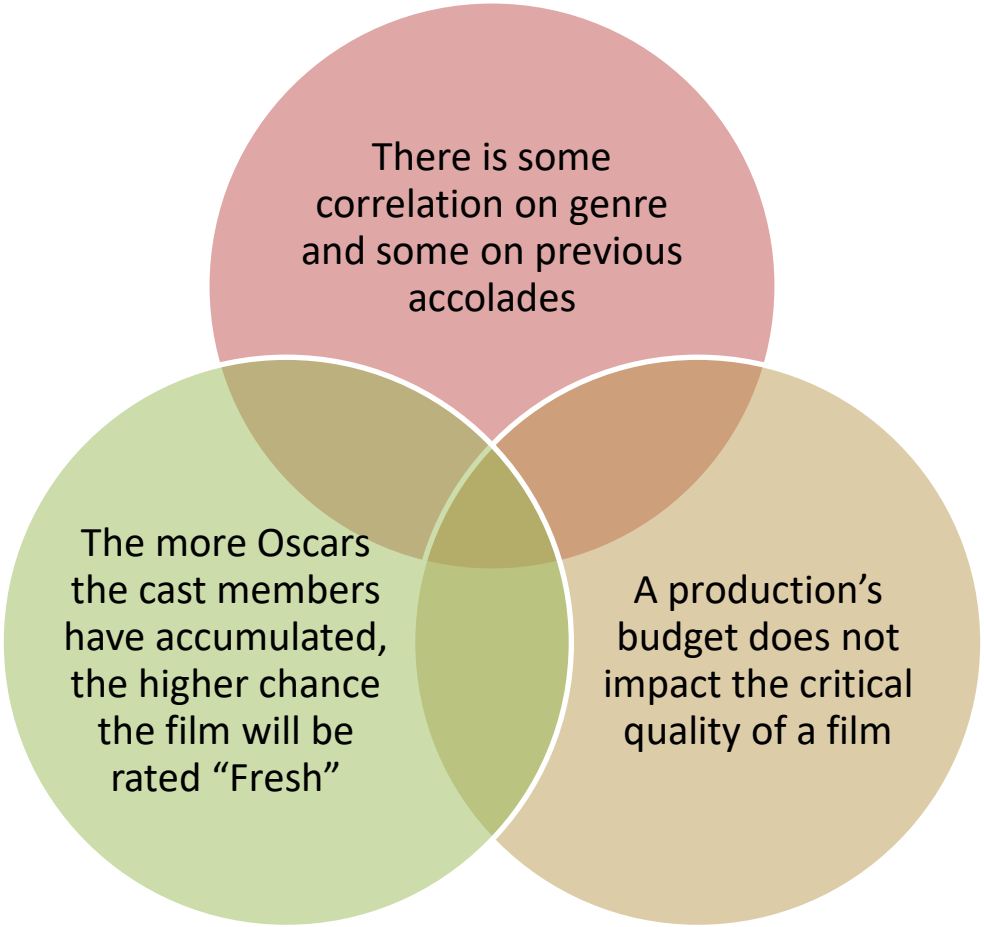
Modelling on Clusters

Overall accuracy of model on test data = 58.7% (test data = 397 movies)

The Area Under the Curve of the model on the test dataset of clusters is **0.6094**



Findings to Date



Learned Info
Next Steps