Capstone Project - Predicting RottenTomato.com Movie Ratings

Hypothesis and Background

I have worked part time with local movie directors for the last 3 years in pre-production as a storyboard artist ; pre-visualizing the scenes for directors when setting up their shots. My interest in how films and TV shows are reviewed peaked following a series of news articles about the Hollywood writers' strikes, the fake Sony movie critic, male reviewer bias against female cast TV shows, and an article about AI writing TV show scripts. Big name actors and directors do make movies that fail to get critical praise and their star power alone cannot compensate for a bad story, a contrived plot or predictable dialogue.

My initial hypothesis is that films with strong storyline and good writing will be a better indicator of quality than merely star recognition and high production values. Movies, unlike books, are often marketed by their star actors or directors such as Scarlett Johansson, Ridley Scott or Christopher Nolan.

These people are definitely the faces associated most with movies, and definitely bring them to life. They did not however write the story. The role and even the names of the screen writers, editors, cinematographers, art directors and producers are usually not on the public's radar. Yet much of the final version of the film depends on them as well as how the actors play their roles or the director shots the scenes.

Is the track record of Oscar nominations or high review scores for a director and the actors a reliable indicator of a future film's quality?

Rotten Tomatoes produces a "tMeter" (Tomato Meter) score, which "represents the percentage of positive professional reviews for films". Is it possible to use the rating history of Movie Critics' reviews that are published in Rotten Tomatoes to determine the "Fresh" or "Rotten" status of a new release or predicting chances of being nominated for Oscars, with some critics better predictors than others?

Methods & Modeling

RottenTomatoes.com has a section specifically of professional Movie Critics reviews that I was able to web scrape to collect each critic's movie review scores. I chose to limit the analysis to professional critics section since they are more likely to see a broad range of films, as opposed to audience members who self-select the movies they see and possibly only rate ones they love or hate. I also downloaded the IMDB 5000 movie title database that contains information about the actors, director and other elements of each title. From the official Oscars website I downloaded the data of the Oscar nominations 1996-2016.

<u>Data Cleaning: IMDB dataset</u>

Within the IMDB movie dataset where the actor names, director names, release year and budget for over 5000 movies. Some titles were duplicated and had to be removed, this is a different problem from two movies having the same title. With this data I also extrapolated out each of the genres to their own column to do more detailed analysis.

The most common occurring combinations
of genres from the IMDB dataset

<u>Data Cleaning: Critic Name conversion</u>

To collect the critics names and then web-scrape their respective URL pages for the movie reviews I first had to “Anglicize” critics whose names had special characters and accents from other languages such as the letters (á, é, ñ, ó, í, ú and ç).

<u>Data Cleaning: Movie Title Duplicate releases</u>

I limited my movie date range to 1996-2016. Over that 21 year span there were over 100 movies that had duplicate titles, such as *The Revenant* (2015) starring Leonardo DiCaprio and *The Revenant* (2009). In all but one case I was able to identify and remove the movie I did not have actor and director data from the IMDB dataset. These were typically movies that were smaller independent releases, were adapted from books that kept the title, or where also international releases whose titles translated to the same name as a release from a Hollywood studio. In the case were the title was same and I was unable to determine after earlier data joining which and review collating which data was associated with which release I removed the movie from the dataset.

An interesting observation was that there were over 300 movies in the IMDB dataset whose release year was different from the web-scraped data from Rotten Tomatoes. In these cases I took the IMDB dataset to be the more accurate of the two. This seemed to be associated primarily with international titles, including British movies, with different release dates in the US or films that had a delayed wide release to theaters and critic review after completion the previous year.

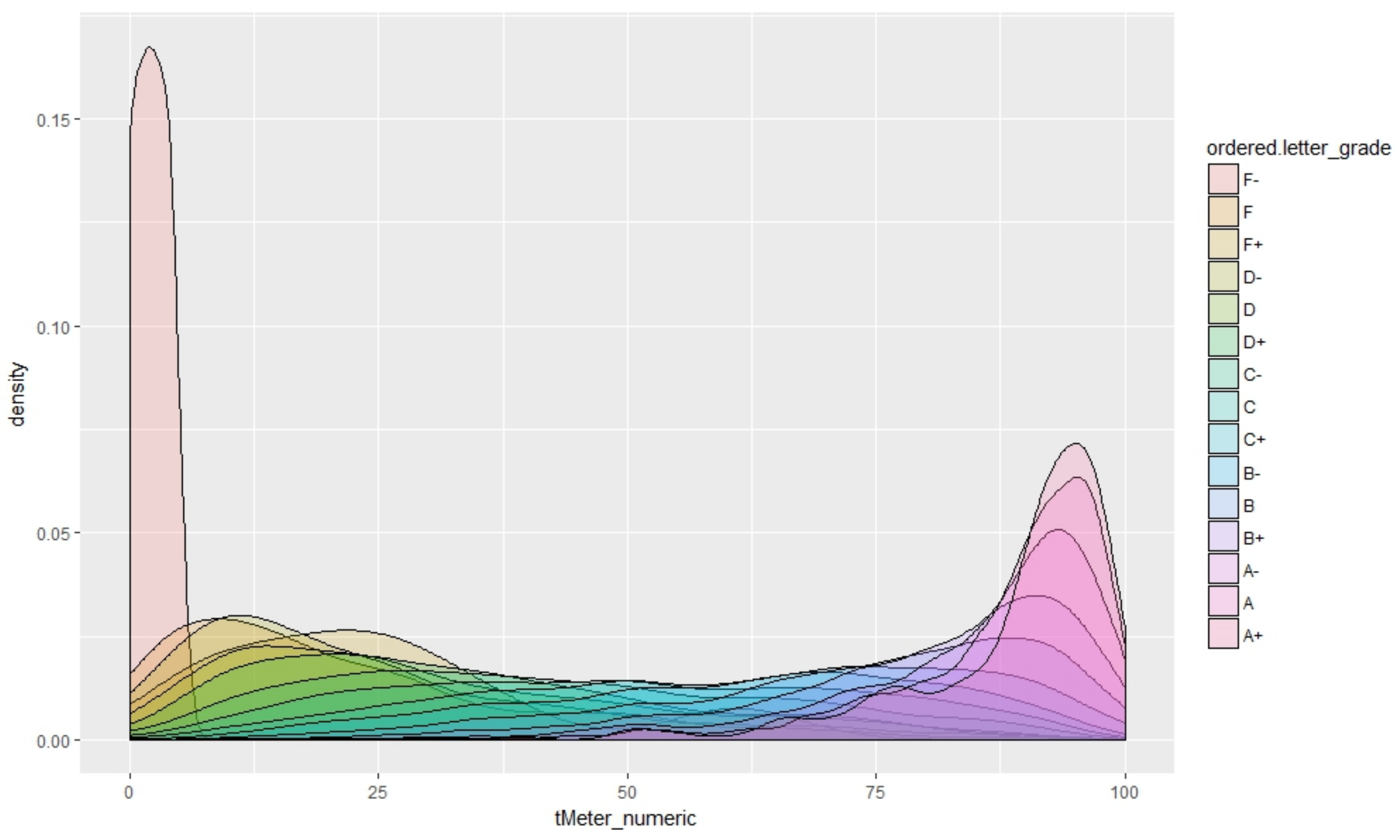
Data Cleaning Critic Review Scores:

The dataset associates critic reviews with movie titles; however the Tomato Meter (tMeter) score is given as a percentage. I developed a formula to translate the various critics' grading methods to a numeric field, translating those that give letter grades (A+ through F-) and those that give marks out of 4, or 5, or 10 to the same metric.

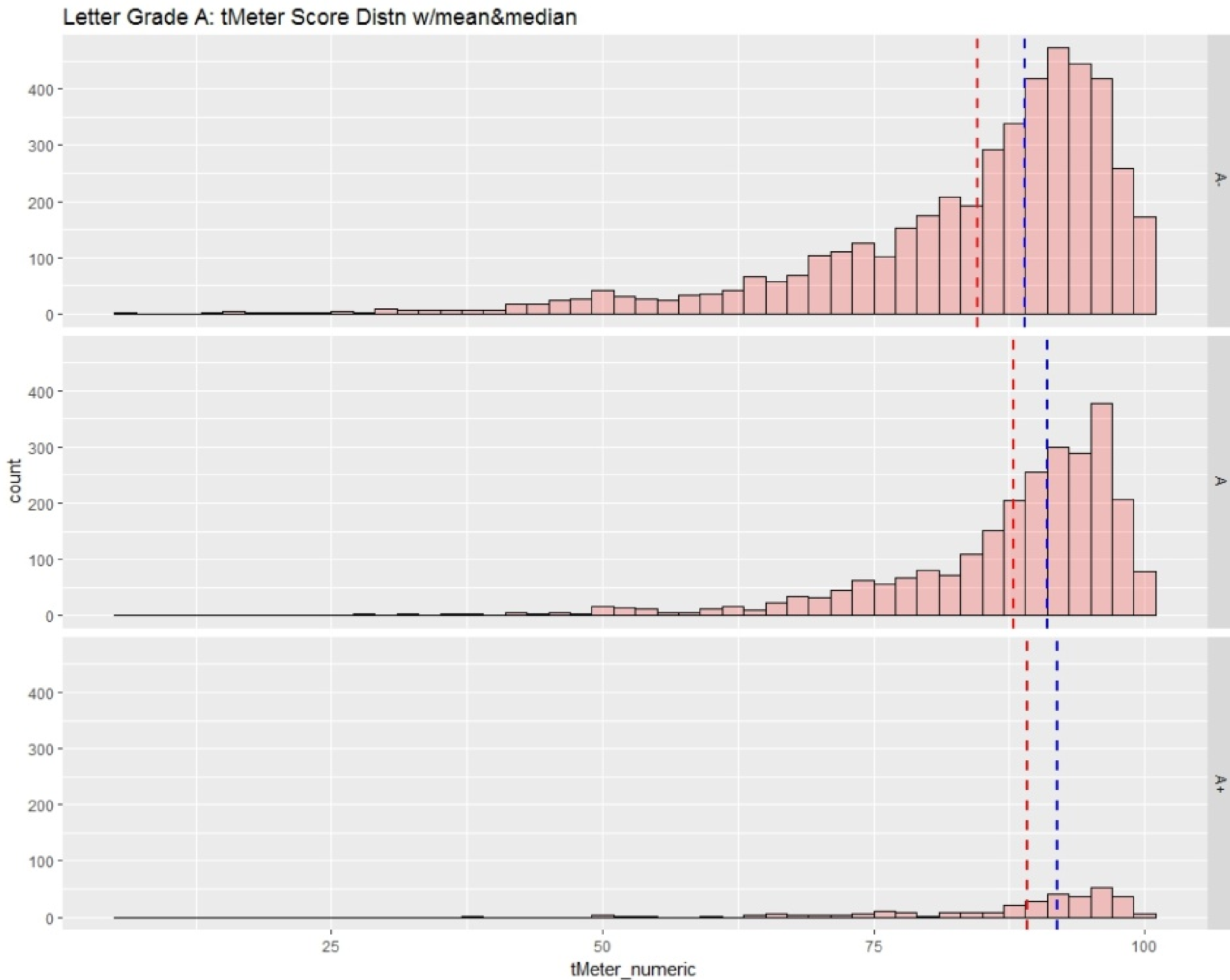
For example, with a score out of 4 or 5 stars I multiplied them out to give a score out of 100. For example a 3.5/4 stars was converted to 87.5%. Sometimes a review omitted the possible range and all that was available was "3 stars". In most instances there were other reviews the same critic gave that provided the missing denominator. Some critics had reviews that were both out of 4 and out of 5 stars. In these cases I had to remove the data as I was unable to determine what the critic's intention was.

For movie reviews with letter grade scores, I selected movies from the last decade. There were over 53,000 movie reviews in from 2006 onwards where letter grades where used.

The distribution below is a visualization of the letter grade scores with the corresponding tMeter score.

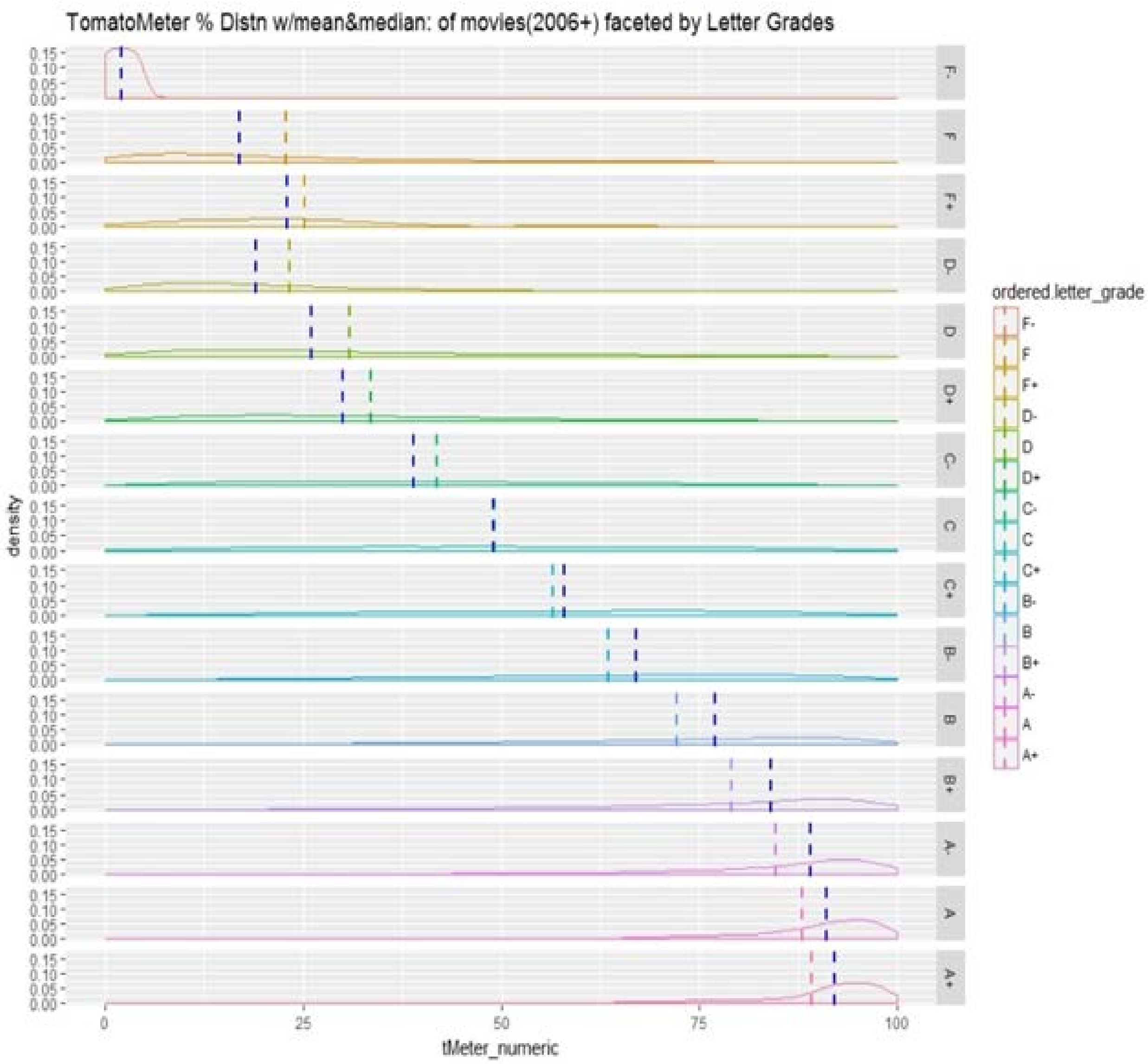


For example the Letter Grades A+, A and A- show a good correlation with a high tMeter score.



Aggregating the tMeter scores associated for films that each by letter grade and using the median is a substitute for the letter grade, and allows for outliers.

Distribution of Letter Grade reviews
against Tomato Meter Score



The replacement values for A+s movies to become 92% while for F- it is 2%.

Letter Grades Conversion Table

Critic's Letter Grade	Median Score	# of Reviews	Distribution
A+	92	299	0.56%
A	91	2,560	4.78%
A-	89	4,561	8.52%
B+	84	8,651	16.15%
B	77	10,317	19.26%
B-	67	6,807	12.71%
C+	58	5,049	9.43%
C	49	5,882	10.98%
C-	39	3,971	7.41%
D+	30	1,653	3.09%
D	26	2,482	4.63%
D-	19	8	0.01%
F+	23	686	1.28%
F	17	634	1.18%
F-	2	3	0.01%
TOTAL		**53,563**	**100.00%**

Oscar Nominations:

From the Academy website I pulled data about all Oscar nominations from 1996-2016. In Excel I associated each film with the directors and actors and joined it to each of the 3 named actors and director who received nominations. I also flagged how many Oscars that actor or director had received in their career in the training dataset of 1996-2013. This latter data may be useful to determine if a track record of Oscar nominations among the cast and crew implied a better scored movie. However my IMDB dataset is missing writer and editor information and additional crew info. Some categories of Oscar are for multiple individuals for a film; For example “Best Picture” Oscars usually are awarded to 2 or more producers and “Art” is awarded to both Art Direction and Production Design roles. One big flaw in the IMDB dataset is that it is limited to 3 named actors, where there might be a cast of several with top billing other names are omitted. My assumption is that the Total Oscars nominated variable will help account for it.

Word-cloud of Oscar nominations 1996-2013 for Actors

Word-cloud of Oscar nominations 1996-2013 for Directors

Description of Datasets

Rottentomatoes.com Critic's pages
~1,200 movie critics with 25 or more reviews on Rotten Tomatoes
~38,000 movie titles where reviewed – dating from 1898 thru Feb 2017
~611,000 reviews were collected

Oscars:
21 years of academy award nominations (1996-2016)
439 movie titles
420 nominations for 253 actors
105 nominations for 74 directors
1187 total nominations when counting additional categories (writing, producing, editing, art, design, and cinematography)

IMDB:
The IMDB dataset of 5000 movies lists the title, release year, genres, director and three actors, budget and film length.
Collating and cleaning these 3 sources produced a dataset of:
3,320 titles (1,439 are rated "Fresh")
202,011 movie critic reviews (before duplicate titles are removed
196,000 critic reviews after duplicate titles are removed.

Training Dataset 1996-2013:
2,921 titles in the train dataset

Testing Dataset 2014-2016:
397 titles in the testing dataset

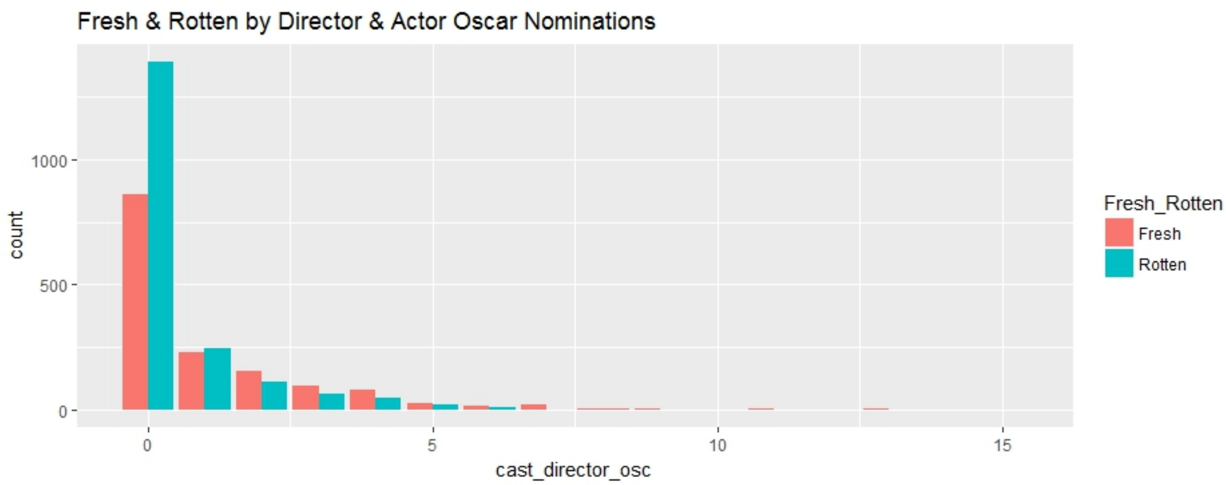
The variables in the dataset
Movie title, Movie year, Actors(3), Director, Budget, Duration, Genres (21 types), total Oscar nominations for a title and for each actor and director, the tMeter score from Rotten Tomatoes, the critic's reviews for each movie.

Data Limitations:
This data does not contain information about the remainder of the crew that received nominations nor does it identify the specific main genre for a film. The data does not account for accolades a movie, cast member or crew received from other sources such as BAFTA.

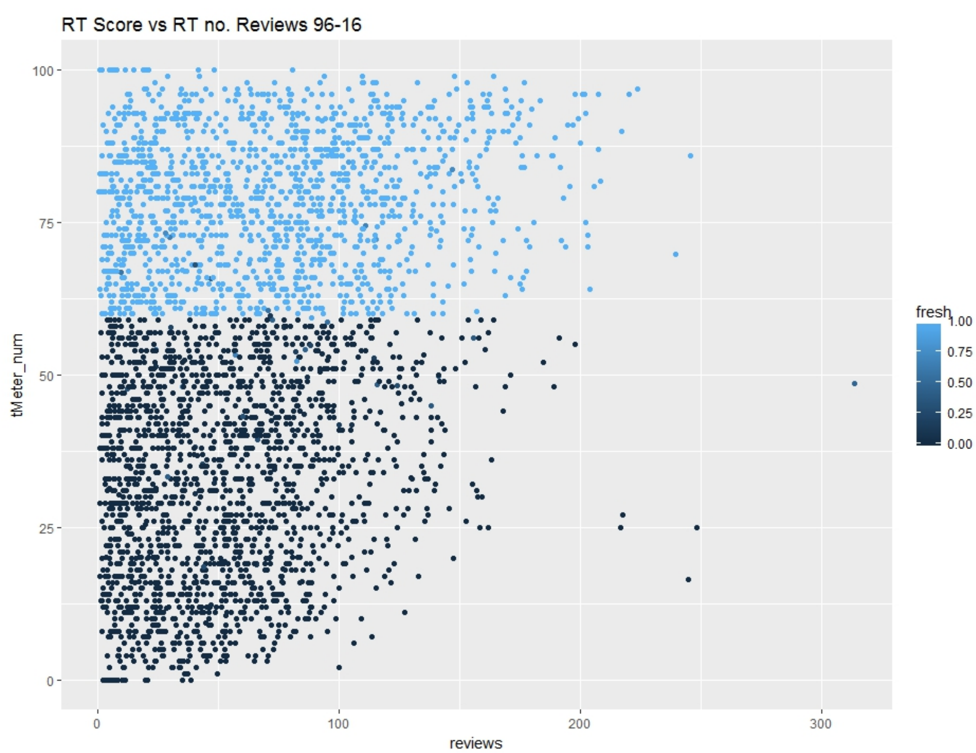
Exploratory Data Analysis Visualizations

Merging the Oscar data with the film review scores I did some visualizations to see if there were any patterns or surprises.

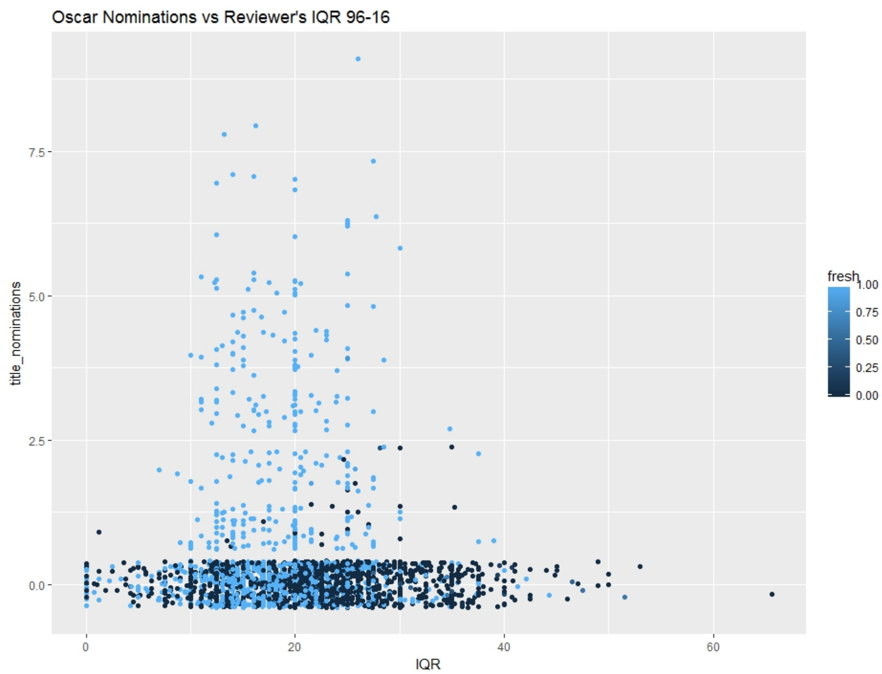
The likelihood of a movie having a "Fresh" rating increases if the cast & crew have received at least 2 Oscar nominations.



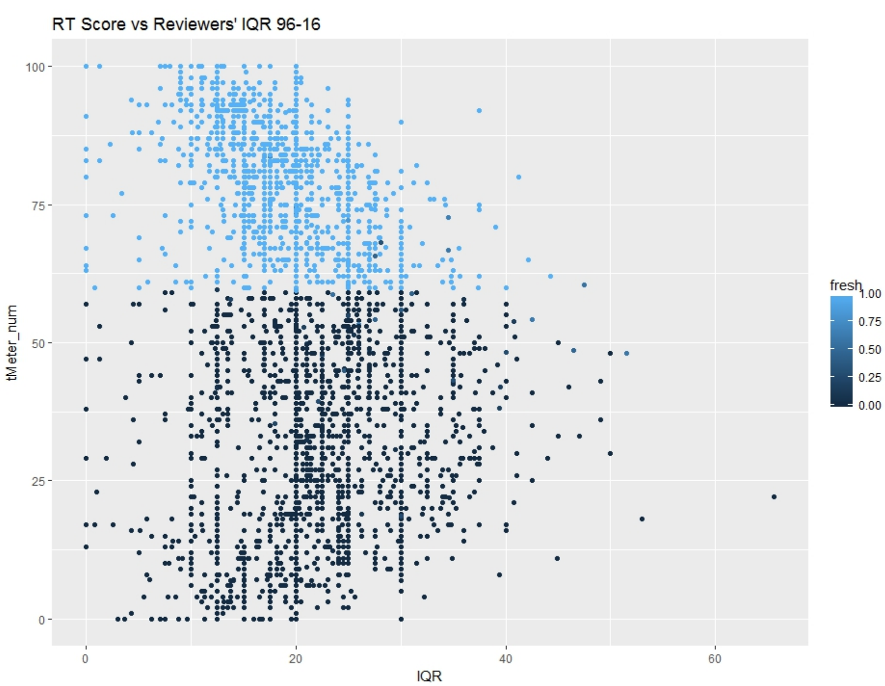
Looking at the quantity of reviews, movies with over 150 reviews have a better chance of having a higher score (tMeter). Is this a measure of popularity, demand for reviews or publicity?



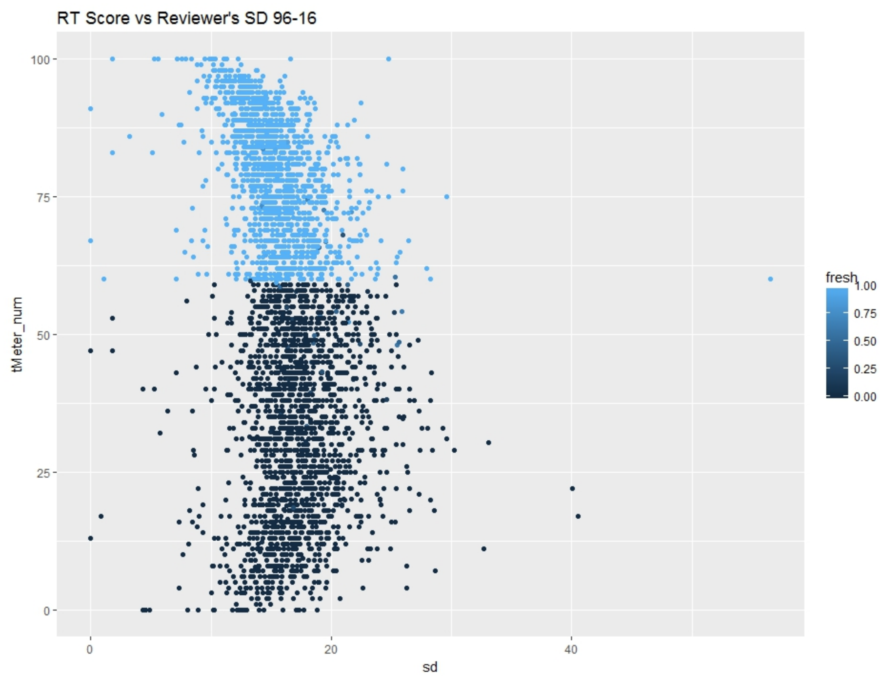
Further exploring Oscar nominated movies I calculated the Interquartile Range (IQR) for the films. The IQR between the reviewers is 30 or less for FRESH films, and 10-20 for films that receive 1 or more Oscar nominations.



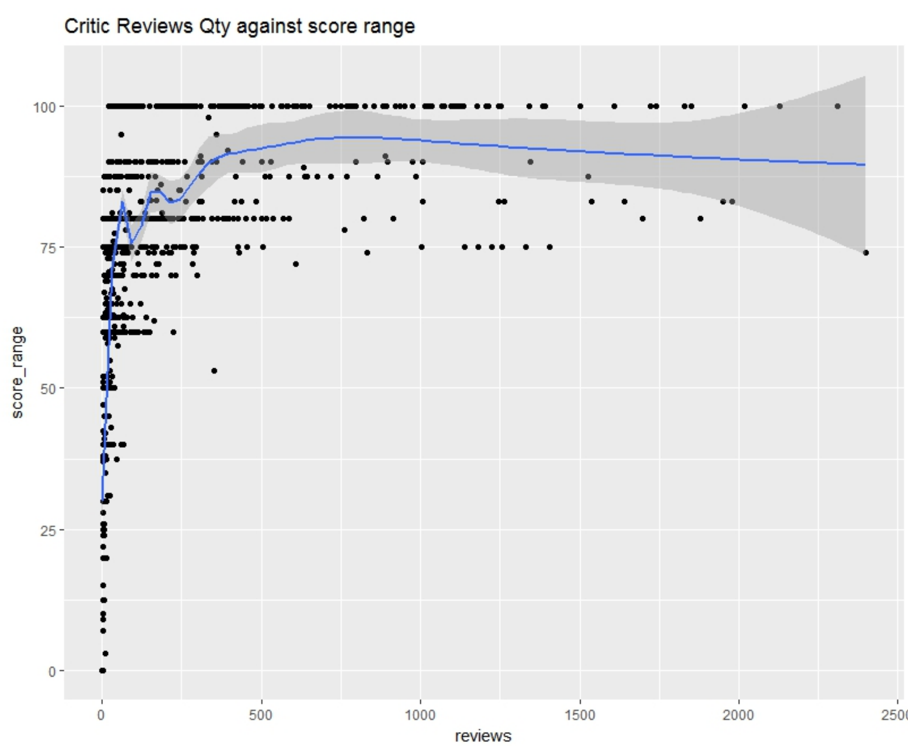
Looking at the relationship of the IQR with a films tMeter rating, the crescent shape seems to indicate that there is a consensus among reviewers for the very good and the very bad films.



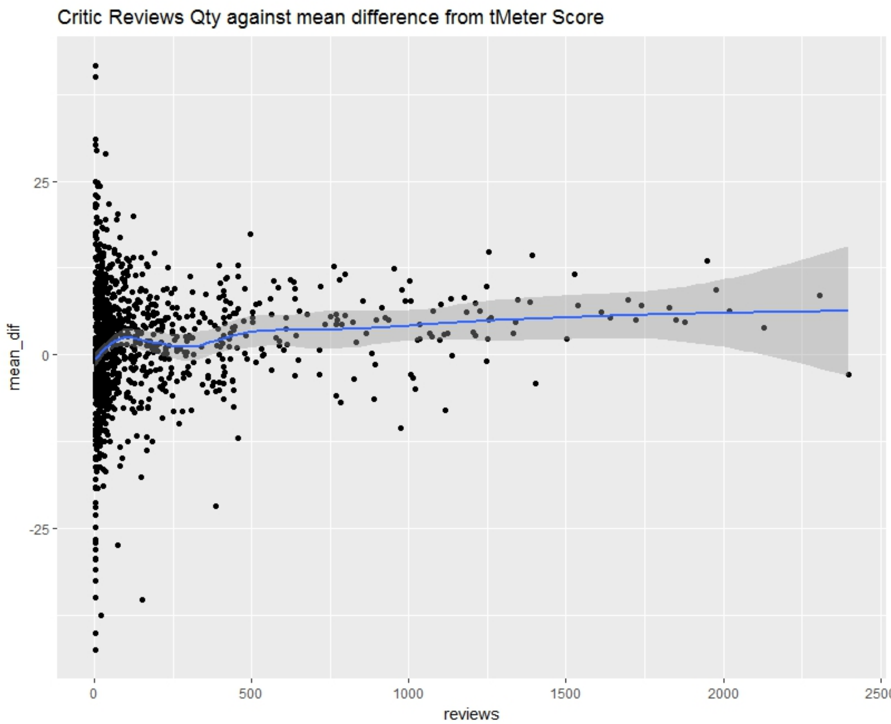
When comparing the Standard Deviation of a movie's reviews there is also a consensus among reviewers for the very good and the very bad films.



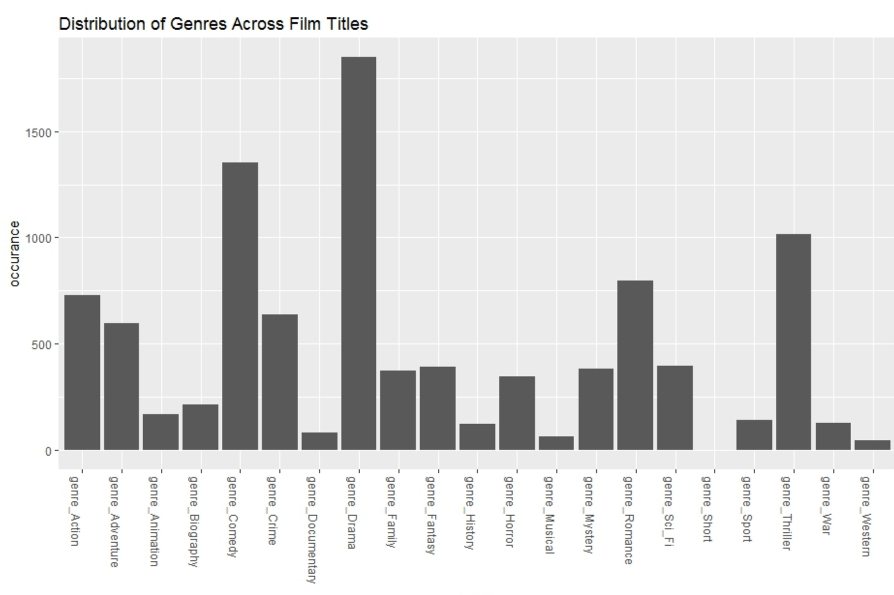
I calculated the range of each critic's ratings and charted it against the number of reviews a critic has given. Critics that have provided more than 250 reviews do give a wider range spanning 75 points to 100 points.



Visualizing the more reviews a critic has published shows the closer their mean difference is to the Tomato Meter score.



The spread of genres indicates that the 5 most popular are: Drama, Comedy, Thriller, Action, and Romance



Confusion Matrix

Using the existing reviews, with the training data set I calculated the mean and median score for **each** actor and director. In the testing dataset there were 366 movies that had at least one actor or the director also in the training set. With a lookup table I substituted their previous scores and extrapolated a value for “Freshness” and calculated a confusion Matrix on these results.

The overall accuracy between using the mean and median scores of the actors and directors remains at **62.6%.**

***Mean* tMeter Score Predictions - For *partial* director and actor list**
Test dataset has 366 films

		Target		
		“Fresh”	“Rotten”	
Model	“Fresh”	69	54	“Fresh” Predictive = 69/123 **56.1%**
	“Rotten”	83	160	“Rotten” Predictive = 160/243 **65.8%**
		Sensitivity = 69/152 **45.4%**	Specificity = 160/214 **74.8%**	Accuracy = 229/366 **62.6%**

***Median* tMeter Score Predictions - For *partial* director and actor list**
Test dataset has 366 films

		Target		
		“Fresh”	“Rotten”	
Model	“Fresh”	78	63	“Fresh” Predictive = 78/141 **55.3%**
	“Rotten”	74	151	“Rotten” Predictive = 151/225 **67.1%**
		Sensitivity = 78/152 **51.3%**	Specificity = 151/214 **70.6%**	Accuracy = 229/366 **62.6%**

I filtered the training dataset to include only films where *all* 3 actors and the director are in the test. This reduced the testing data to 104 films. In this case the accuracy using mean scores improves to **65.4%**

***Mean* tMeter Score Predictions - For complete director and actor list Test dataset has 104 films**				
		Target		
		"Fresh"	"Rotten"	
Model	"Fresh"	25	18	"Fresh" Predictive = 25/43 **58.1%**
	"Rotten"	18	43	"Rotten" Predictive = 43/61 **70.5%**
		Sensitivity = 25/43 **58.1%**	Specificity = 43/61 **70.5%**	Accuracy = 66/104 **65.4%**

***Median* tMeter Score Predictions - For complete director and actor list Test dataset has 104 films**				
		Target		
		"Fresh"	"Rotten"	
Model	"Fresh"	21	16	"Fresh" Predictive = 21/37 **56.8%**
	"Rotten"	22	45	"Rotten" Predictive = 45/67 **67.2%**
		Sensitivity = 21/43 **48.8%**	Specificity = 45/61 **74.8%**	Accuracy = 66/104 **63.5%**

Modelling

Based on my analysis of the data available this evolved to predicting the Fresh/Rotten chance of a movie, using Linear Regression and Logistic regression models.

From the dataset I will be looking to see if a Movies score shows any correlation to the number of Oscars the cast and crew have had in their career (limited to the last 20 years).

Linear Regression on Training Data for predictability of tMeter based on historical Oscars

Coefficients:

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	48.7795	0.5387	90.55	<2e-16	***
Oscars	3.5602	0.2140	16.64	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.97 on 2923 degrees of freedom

Multiple R-squared: 0.08652, Adjusted R-squared: 0.0862

F-statistic: 276.8 on 1 and 2923 DF, p-value: < 2.2e-16

Linear Regression on Training Data for predictability of tMeter based on historical Oscars and budget

Coefficients:

	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	4.781e+01	6.041e-01	79.136	<2e-16	***
Oscars	3.726e+00	2.156e-01	17.281	<2e-16	***
budget	-2.063e-09	6.208e-09	-0.332	0.74	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.74 on 2703 degrees of freedom

(219 observations deleted due to missingness)

Multiple R-squared: 0.09976, Adjusted R-squared: 0.09909

F-statistic: 149.8 on 2 and 2703 DF, p-value: < 2.2e-16

Linear Regression on Training Data for predictability of "**tMeter**" based on historical **Oscars**, **budget**, duration and 21 **genres**

Coefficients:

	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	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

<u>Linear Regression</u> on Training Data for predictability of **tMeter** based on historical **Oscars**, **duration** and **seven** selected genres

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	33.07394	3.06003	10.808	<2e-16	***
Oscars	2.66821	0.21804	12.237	<2e-16	***
duration	0.12594	0.02661	4.732	2.33e-06	***
Action	-5.48887	1.25184	-4.385	1.20e-05	***
Animation	18.85654	2.18967	8.612	<2e-16	***
Comedy	-2.96206	1.12310	-2.637	0.0084	**
Drama	11.66317	1.03981	11.217	<2e-16	***
Documentary	33.63199	2.99164	11.242	<2e-16	***
Romance	-4.42130	1.11016	-3.983	6.98e-05	***
Thriller	-4.95409	1.14893	-4.312	1.67e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.13 on 2915 degrees of freedom

Multiple R-squared: **0.2133**
Adjusted R-squared: **0.2108**
F-statistic: 87.79 on 9 and 2915 DF, p-value: < 2.2e-16

<u>Logistic Regression</u> with 21 genres and other variables predicting **"Fresh"**

Coefficients:

	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	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

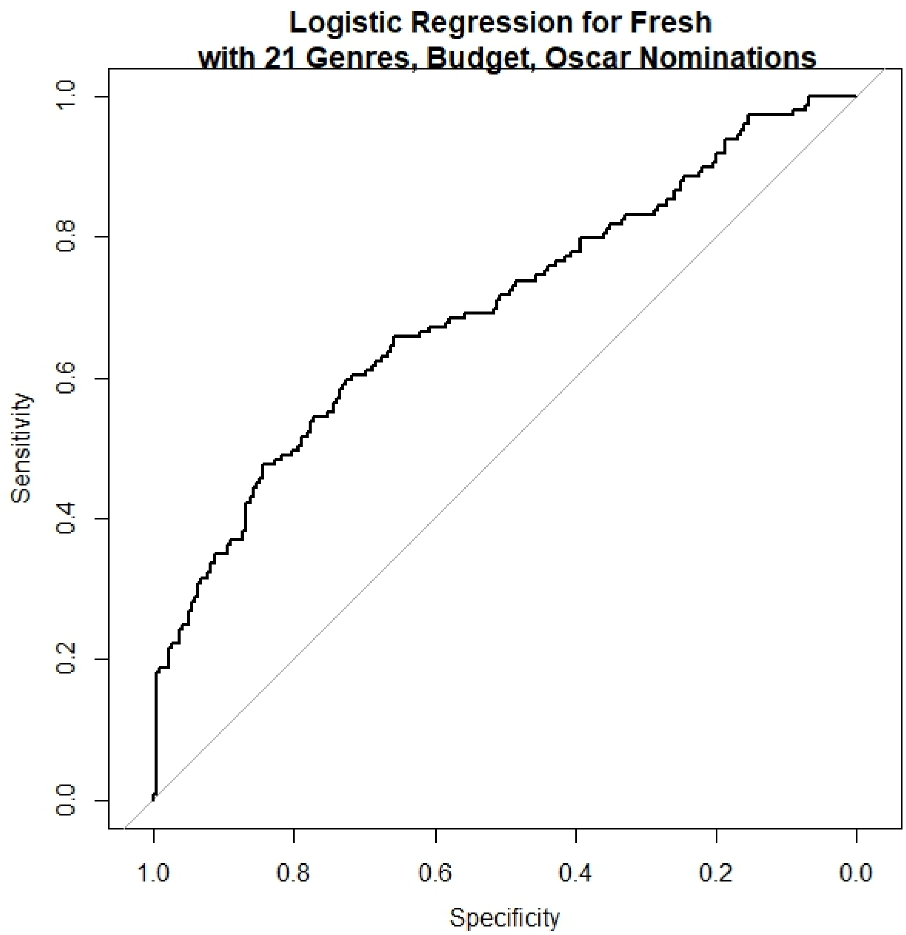
Null deviance: 3703.3 on 2705 degrees of freedom

Residual deviance: 3179.5 on 2681 degrees of freedom

AIC: 3229.5

<u>ROC curve</u>

Using the predictive model from the Logistic Regression the **Area Under the Curve = 0.6965**



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%**

<u>Logistic Regression</u> to variables with significance reduces the equation to **8** genres, Oscars and duration.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.361787	0.287120	-4.743	2.11e-06	***
Oscars	0.254676	0.026075	9.767	<2e-16	***
duration	0.007136	0.002507	2.846	0.004423	**
Action	-0.549728	0.117275	-4.688	2.77e-06	***
Animation	1.197638	0.192357	6.226	4.78e-10	***
Biography	0.392957	0.193966	2.026	0.042775	*
Comedy	-0.250681	0.102130	-2.455	0.014107	*
Drama	0.696326	0.094545	7.365	1.77e-13	***
Documentary	2.862479	0.411867	6.950	3.65e-12	***
Romance	-0.418032	0.101067	-4.136	3.53e-05	***
Thriller	-0.407080	0.105433	-3.861	0.000113	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

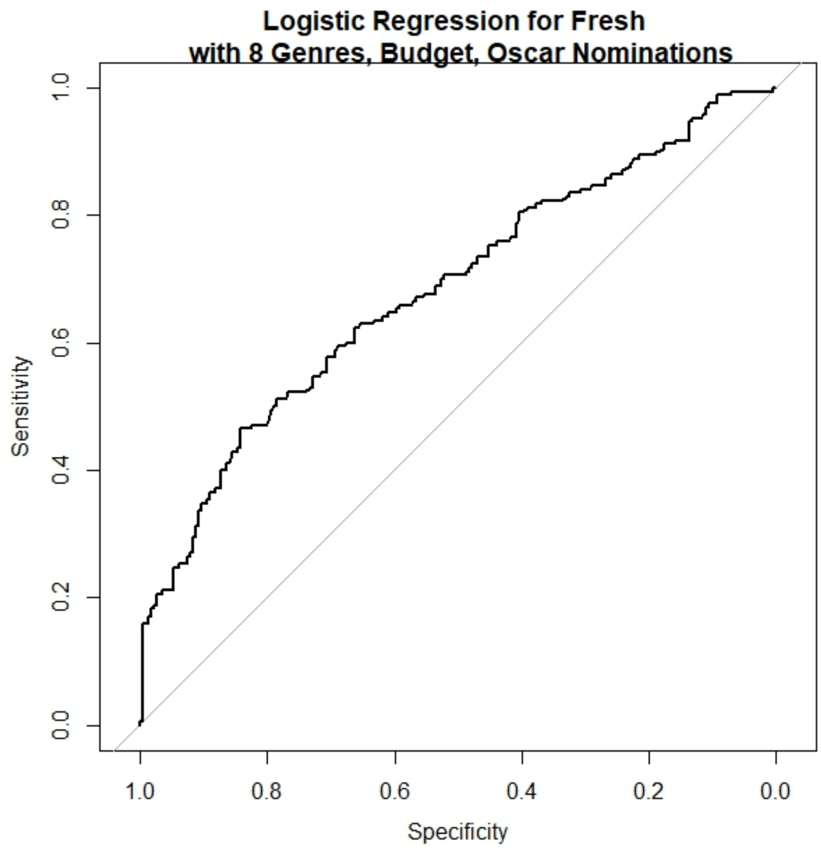
Null deviance: 4017.8 on 2924 degrees of freedom

Residual deviance: 3486.8 on 2914 degrees of freedom

AIC: 3508.8

ROC curve

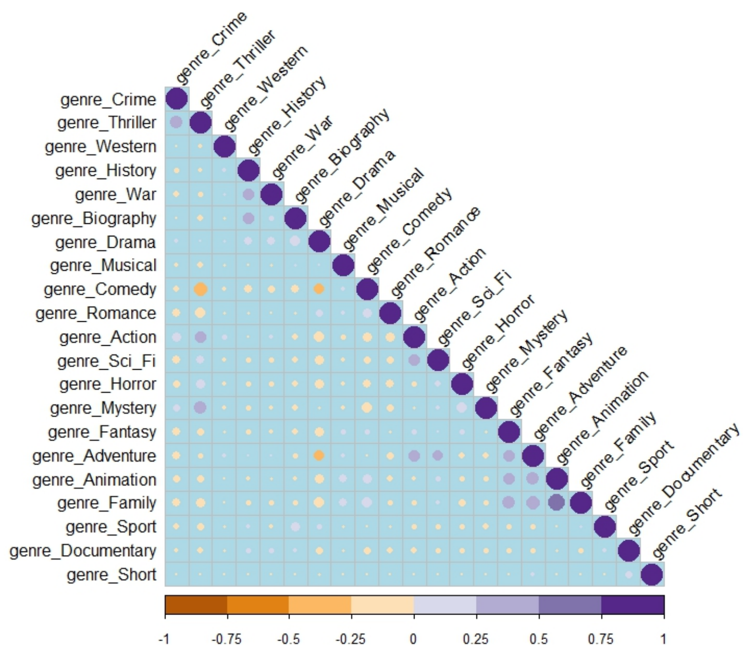
In this second model from the Logistic Regression the **Area Under the Curve is reduced to 0.6824** and a reduced accuracy as shown in the confusion matrix below from **67.2% to 65.3%**



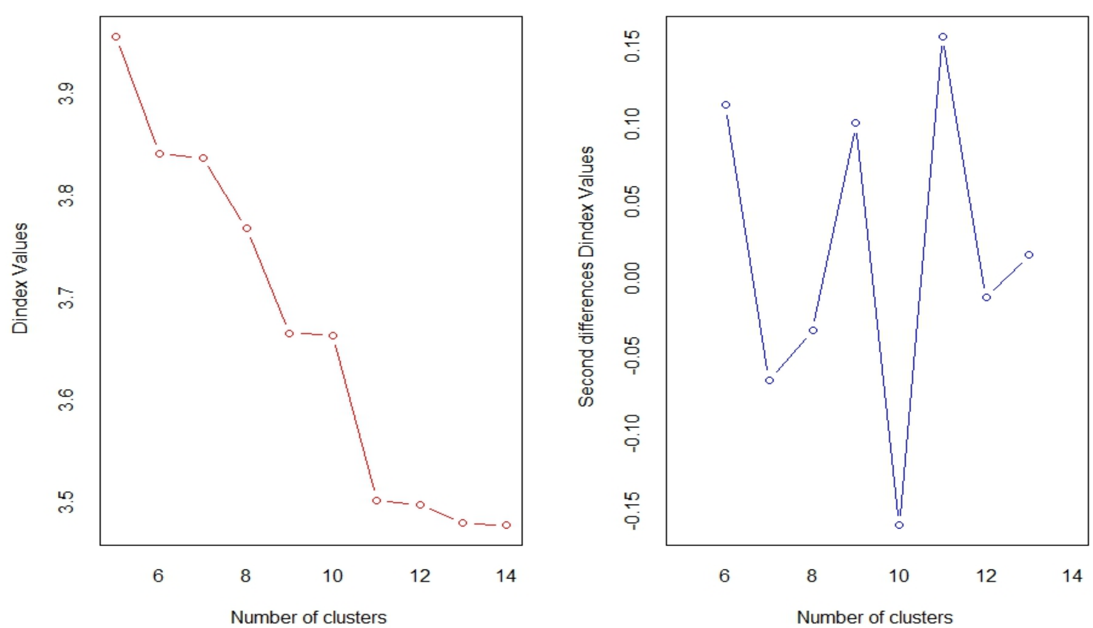
		***Logistic Regression* model applied to test data Confusion Matrix**		
		Target		
		"Fresh"	"Rotten"	
Model	"Fresh"	89	81	"Fresh" Predictive = 89/170 **52.4%**
	"Rotten"	57	171	"Rotten" Predictive = 171/228 **75%**
		Sensitivity = 89/146 **61%**	Specificity = 171/252 **67.9%**	Accuracy = 260/398 **65.3%**

Cluster Analysis

A correlation matrix of the most frequently occurring pairs of genres shows 14 most common pairings.



The k-means cluster analysis indicates 11 clusters.



The genres can be clustered together in the following 11 groups with the significant genres in each cluster group highlighted in yellow.

					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**

For example; Cluster 10 are Romance-Comedies, Cluster 3 are Dramas and Cluster 1 are Miscellaneous combinations of genres that do not fit one category more than any other.

<u>Logistic Regression</u> on 11 clusters:

Coefficients:

	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	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 4049.4 on 2921 degrees of freedom

Residual deviance: 3884.9 on 2910 degrees of freedom

AIC: 3906.9

The most significant cluster groups were :

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**

Confusion Matrix for Logistical Model Against 11 Genres- Training Data

		Target		
		“Fresh”	“Rotten”	
Model	“Fresh”	329	245	“Fresh” Predictive = 329/574 **57.3 %**
	“Rotten”	966	1381	“Rotten” Predictive =1381/2347 **58.8.%**
		Sensitivity =329/1295 **25.4 %**	Specificity =1381/1626 **84.9 %**	Accuracy = 1710/2921 **58.5%**

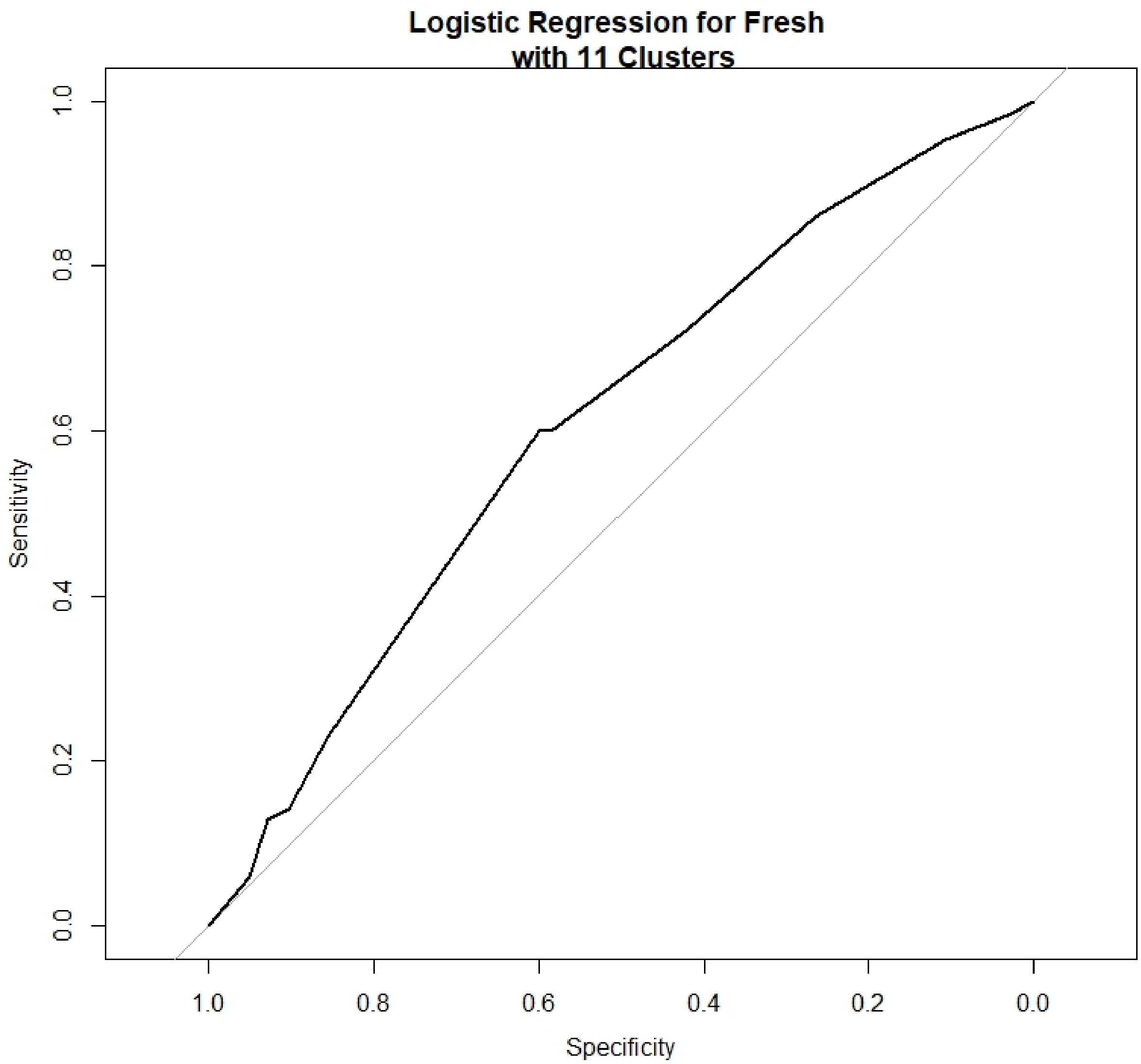
Model accuracy on training data is 58.5% (training data = 2,921 movies)

Confusion Matrix for Logistical Model
Against 11 Genres- Test Data

		Target		
		“Fresh”	“Rotten”	
Model	“Fresh”	39	33	“Fresh” Predictive = 39/72 **54.1 %**
	“Rotten”	131	194	“Rotten” Predictive =194/325 **59.7 %**
		Sensitivity = 39/170 **22.9 %**	Specificity = 194/227 **85.5 %**	Accuracy = 233/397 **58.7 %**

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

So far my linear models are producing some correlation on genre and some on previous accolades. A production's budget does not seem to impact the quality of a film. There is some evidence that the more Oscars cast members have accumulated the higher chance the film with be rated "Fresh" and the confusion matrix shows some accuracy predicting future "Fresh" ratings based on actor's and director's previous scores. Grouping the genres into clusters did not create a better model.

Learned info, Next steps

Some next steps could be to see if any critics are "super critics" and are of themselves able to predict Fresh/Rotten. Do they have a genre specialty?

A more extensive list of cast and crew with additional web-scraping would allow a better use of previous experience and reviews upon future performance. A next step would be to incorporate more screen credit nominations to include writers, producers, art directors and cinematography for the films in the time period, if possible.

I applied the Tomato Meter median to Letter Grade scores; however the Tomato Meter is an indicator of positive reviews and not an average of reviews. A better calculation would be to actually calculate the average reviews and use that to interpret the Letter Grade scores, or what numeric or grade determines the favorable or not indicator from a reviewer.

There are a great number of good movies that do not make it to the finalist for Oscars but have received accolades from other sources such as BAFTA that also are shown in US theaters.