



University  
of Glasgow

Group 09  
20-March-2024

# The analysis of variables influencing movie ratings





# CONTENTS

**01**

Introduction

**02**

Data Cleaning & Exploratory Analysis

**03**

Model Fitting

**04**

Model Comparison

**05**

Conclusion

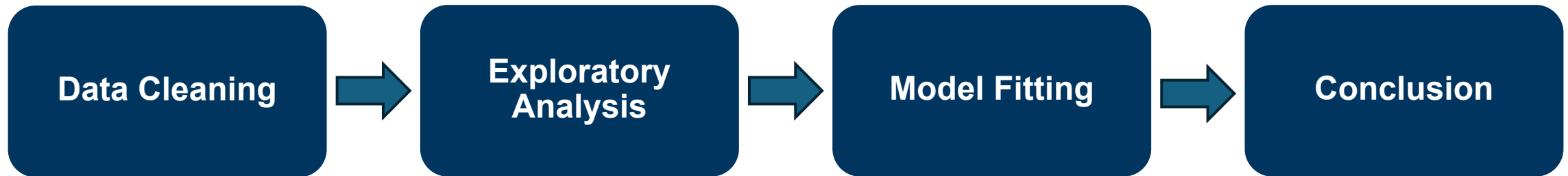
- **Dataset Description**

The IMDB database contains a variety of information on all films that have been released. The following analysis aims to explore the relationship between a set of descriptive variables about a film and its success measured by its respective IMDB rating.

- **Main Research Questions**

Which properties of films influence whether they are rated by IMDB as greater than 7 or not?

- **Analysis Procedure**



## Data Cleaning

As a subset of the IMDB database, the film dataset contains the following variables:

- i. film\_id – The unique identifier for the film
- ii. year – Year of release of the film in cinemas
- iii. length – Duration (in minutes)
- iv. budget – Budget for the films production (in \$1000000s)
- v. votes – Number of positive votes received by viewers
- vi. genre – Genre of the film
- vii. rating – IMDB rating from 0-10


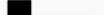



It is now established that film\_id will not be used as an explanatory variable since it is only an identifier for the film, rather than an informative feature about it.

Genre is the only categorical variable contained in the data set.

Year, length, budget, and votes are the numerical explanatory variables to be tested in this analysis.

# Data Cleaning








A tibble: 5 × 11

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>	sd <dbl>	p0 <dbl>	p25 <dbl>	p50 <dbl>	p75 <dbl>	p100 <dbl>	hist <chr>
1	year	0	1.0000000	1975.879040	24.134644	1895.0	1957.00	1983.0	1997.0	2005.0	
2	length	127	0.9576808	81.574461	39.544547	1.0	71.25	90.0	100.0	555.0	
3	budget	0	1.0000000	11.983705	2.969693	1.2	10.10	12.1	14.0	23.4	
4	votes	0	1.0000000	655.833056	3780.096346	5.0	11.00	30.0	118.0	92437.0	
5	rating	0	1.0000000	5.399633	2.072820	0.8	3.70	4.7	7.8	9.2	

5 rows

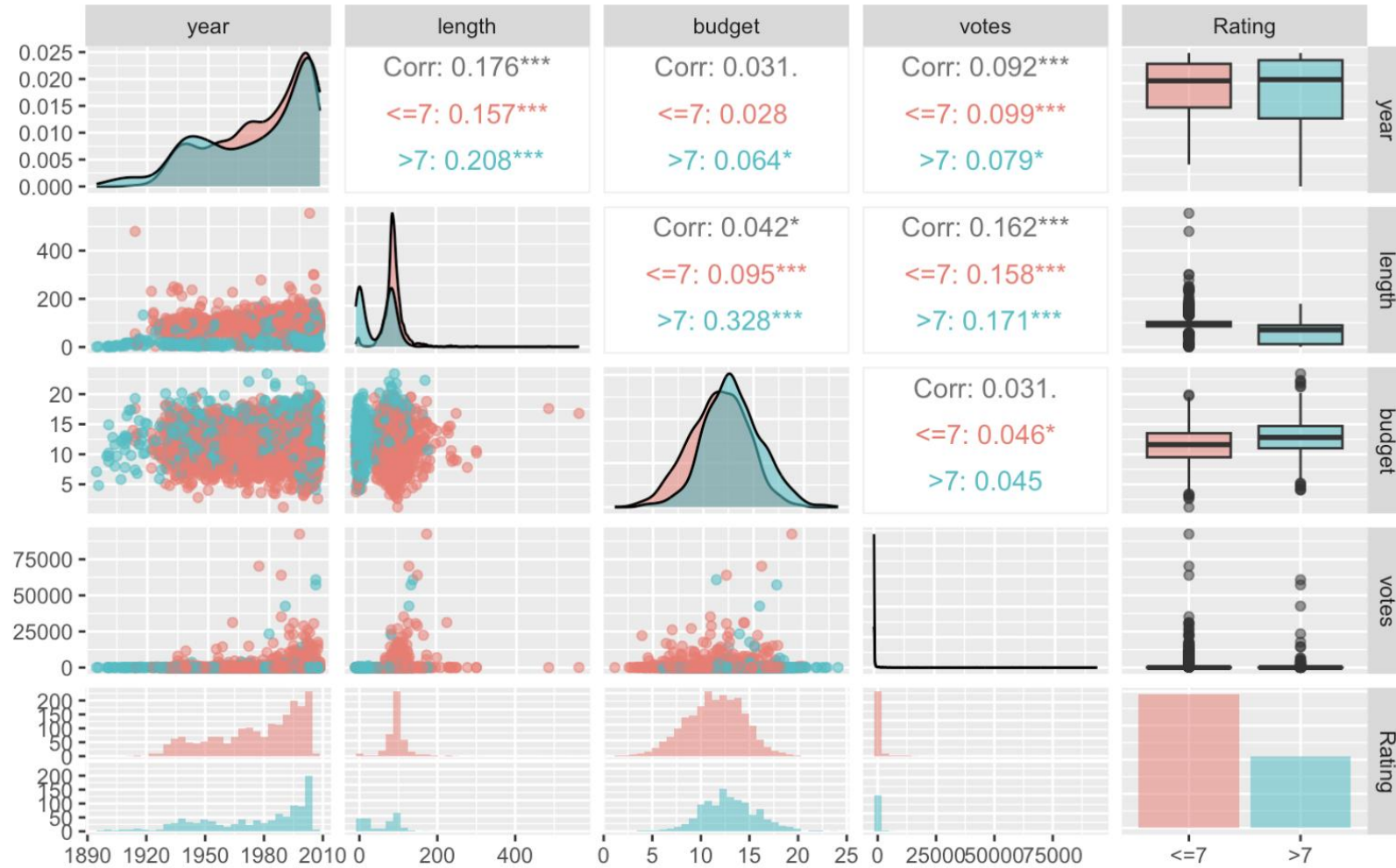
When it comes to the numerical variables, summary statistics above show that there are 127 missing values in length variable. For these missing values, the median film length by genre will be chosen to impute, considering the length distribution is not equal among different film genres.

A tibble: 7 × 12

	skim_variable <chr>	genre <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>	sd <dbl>	p0 <dbl>	p25 <dbl>	p50 <dbl>	p75 <dbl>	p100 <dbl>	hist <chr>
1	length	Action	34	0.9607390	94.66827	29.673118	2	84.00	91	100.0	480	
2	length	Animation	4	0.9787234	13.89130	20.496811	1	7.00	7	8.0	97	
3	length	Comedy	39	0.9466484	82.50000	31.153302	1	78.00	90	100.0	181	
4	length	Documentary	5	0.9728261	70.85475	43.214920	1	43.50	75	90.0	278	
5	length	Drama	40	0.9542334	94.58513	33.274017	4	85.25	96	107.0	555	
6	length	Romance	3	0.9032258	90.21429	44.471678	12	76.75	97	108.5	189	
7	length	Short	2	0.9842520	15.53600	9.117994	2	10.00	13	20.0	44	

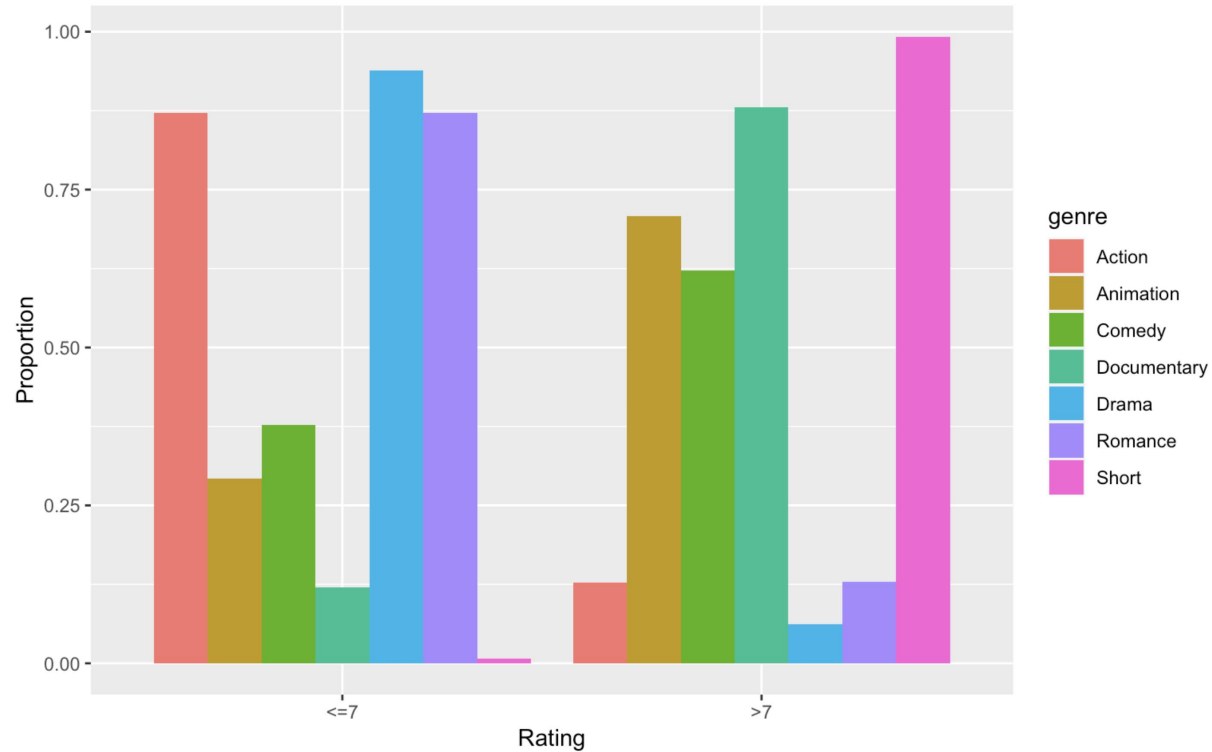
7 rows

# Exploratory Analysis



From this graph, we can see that the correlations between all numerical explanatory variables are very weak. And according to the boxplots, the ratings tends to be influenced by length and budget, while the influence of votes and years appear to be less pronounced.

# Exploratory Analysis



genre	<=7		>7	
Action	87.2%	(755)	12.8%	(111)
Animation	29.3%	(55)	70.7%	(133)
Comedy	37.8%	(276)	62.2%	(455)
Documentary	12.0%	(22)	88.0%	(162)
Drama	93.8%	(820)	6.2%	(54)
Romance	87.1%	(27)	12.9%	(4)
Short	0.8%	(1)	99.2%	(126)

The barplot and table above show the proportion of ratings in different genres. As we can see, the distribution of the proportion varies greatly across genres.

## Model with categorical variable(Genre):

Model 1.1 (Replace missing values with median values in Length)

$$\hat{y}_{Rating} = \hat{\alpha} + \hat{\beta}_{Year} \cdot Year + \hat{\beta}_{Length} \cdot Length + \hat{\beta}_{Budget} \cdot Budget + \hat{\beta}_{Votes} \cdot Votes + \hat{\beta}_{Genre} \cdot \Pi_{Genre}(x)$$

Model 1.2 (Remove missing values in Length)

$$\hat{y}_{Rating} = \hat{\alpha} + \hat{\beta}_{Year} \cdot Year + \hat{\beta}_{Length} \cdot Length + \hat{\beta}_{Budget} \cdot Budget + \hat{\beta}_{Votes} \cdot Votes + \hat{\beta}_{Genre} \cdot \Pi_{Genre}(x)$$

## Model without categorical variable(Genre):

Model 2

$$\hat{y}_{Rating} = \hat{\alpha} + \hat{\beta}_{Year} \cdot Year + \hat{\beta}_{Length} \cdot Length + \hat{\beta}_{Budget} \cdot Budget + \hat{\beta}_{Votes} \cdot Votes$$



## Summary of Model 1.1

Observations	3001
Dependent variable	target
Type	Generalized linear model
Family	binomial
Link	logit
$\chi^2(10)$	2335.40
Pseudo- $R^2$ (Cragg-Uhler)	0.75
Pseudo- $R^2$ (McFadden)	0.60
AIC	1565.91
BIC	1631.98

For **numerical** variables:

- Length( $0.00 < 0.05$ ), budget( $0.00 < 0.05$ ) and votes( $0.00 < 0.05$ ) shows significant.
- Variable year is not significant( $0.08 > 0.05$ ).

	Est.	S.E.	z val.	p
(Intercept)	-13.25	5.58	-2.37	0.02
year	0.00	0.00	1.74	0.08
length	-0.06	0.00	-16.81	0.00
budget	0.52	0.03	17.88	0.00
votes	0.00	0.00	2.90	0.00
genreAnimation	-0.54	0.33	-1.61	0.11
genreComedy	3.34	0.18	19.01	0.00
genreDocumentary	5.31	0.38	13.89	0.00
genreDrama	-1.49	0.23	-6.51	0.00
genreRomance	-0.77	0.86	-0.91	0.36
genreShort	4.28	1.05	4.07	0.00

For **categorical** variables:

- Comedy( $0.00 < 0.05$ ), Documentary( $0.00 < 0.05$ ), Drama( $0.00 < 0.05$ ) and Short( $0.00 < 0.05$ ) shows significant.
- Animation( $0.11 > 0.05$ ) and Romance( $0.36 > 0.05$ ) is not significant.

## Summary of Model 1.2

Observations	2874
Dependent variable	target
Type	Generalized linear model
Family	binomial
Link	logit
$\chi^2(10)$	2254.00
Pseudo-R <sup>2</sup> (Cragg-Uhler)	0.75
Pseudo-R <sup>2</sup> (McFadden)	0.61
AIC	1490.89
BIC	1556.48

For **numerical** variables:

- Length( $0.00 < 0.05$ ), budget( $0.00 < 0.05$ ) and votes( $0.00 < 0.05$ ) shows significant. And variable year is also significant now( $0.02 < 0.05$ ).

	Est.	S.E.	z val.	p
(Intercept)	-16.45	5.73	-2.87	0.00
year	0.01	0.00	2.25	0.02
length	-0.06	0.00	-16.71	0.00
budget	0.52	0.03	17.44	0.00
votes	0.00	0.00	2.89	0.00
genreAnimation	-0.44	0.34	-1.32	0.19
genreComedy	3.37	0.18	18.59	0.00
genreDocumentary	5.32	0.39	13.74	0.00
genreDrama	-1.42	0.23	-6.13	0.00
genreRomance	-0.62	0.90	-0.70	0.49
genreShort	4.29	1.05	4.08	0.00

For **categorical** variables:

- Comedy( $0.00 < 0.05$ ), Documentary( $0.00 < 0.05$ ), Drama( $0.00 < 0.05$ ), Short( $0.00 < 0.05$ ) still shows significant.
- Animation( $0.19 > 0.05$ ), Romance( $0.49 > 0.05$ ) is not significant and p-value become larger.

## Summary of Model 2

Observations	3001
Dependent variable	target
Type	Generalized linear model
Family	binomial
Link	logit
$\chi^2(4)$	1184.45
Pseudo-R <sup>2</sup> (Cragg-Uhler)	0.45
Pseudo-R <sup>2</sup> (McFadden)	0.31
AIC	2704.86
BIC	2734.89

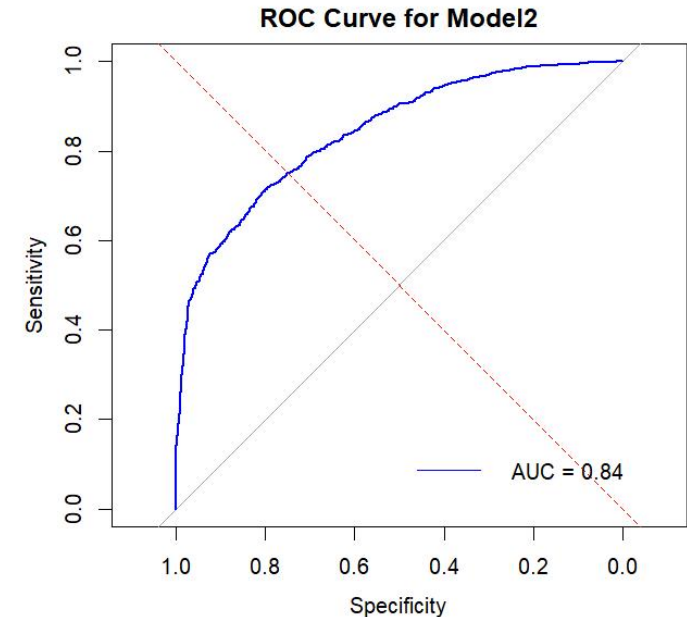
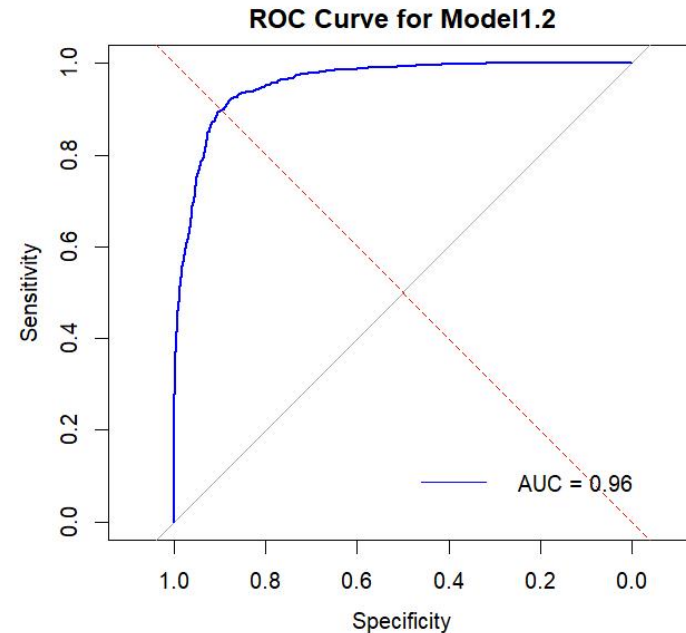
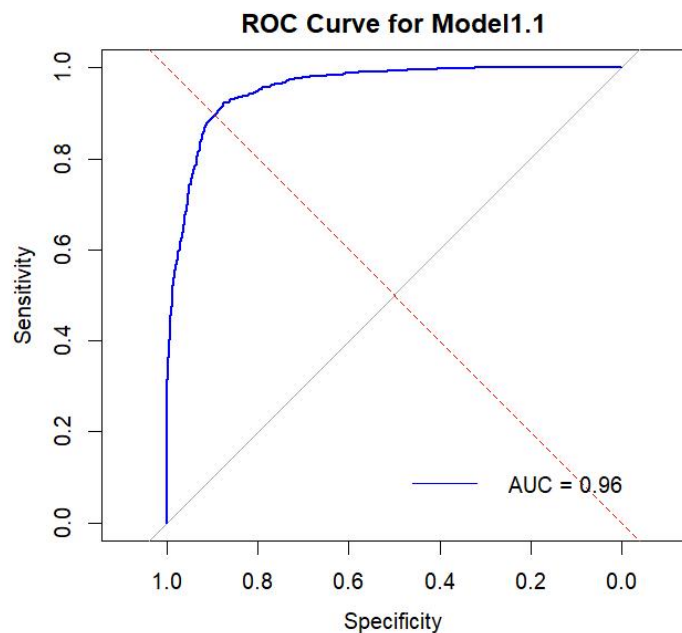
	Est.	S.E.	z val.	p
(Intercept)	-18.87	4.09	-4.61	0.00
year	0.01	0.00	4.40	0.00
length	-0.05	0.00	-24.13	0.00
budget	0.30	0.02	15.91	0.00
votes	0.02	0.01	1.89	0.06

Only **numerical** variables:

- Year( $0.00 < 0.05$ ), budget( $0.00 < 0.05$ ) and length( $0.00 < 0.05$ ) shows significant.
- Votes is not significant( $0.06 > 0.05$ ).

# ROC Plots

To determine the validity of the model more accurately, the ROC curves of the three models need to be plotted and compared. The ROC curve for model 1.1 and model 1.2 show better predictions than model 2



The area under the ROC curve, known as AUC. An AUC value of 0.5 indicates that the predictive model is of no discriminative value. Higher AUC means a better model.



# Confusion Matrix

		Predicted condition	
		Predicted Positive (PP)	Predicted Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Model 1.1

Prediction

		Reference	
		0	1
Prediction	0	1815	178
	1	141	867

Model 1.2

Prediction

		Reference	
		0	1
Prediction	0	1733	168
	1	134	839

Model 2

Prediction

		Reference	
		0	1
Prediction	0	1798	448
	1	161	597

# Precision, Recall and Accuracy

		Predicted condition	
		Predicted Positive (PP)	Predicted Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Positive} + \text{Negative}}$$

## Models Selection

Model	AIC	AUC	Precision	Recall	Accuracy
Model 1.1	1565.906	0.956	0.911	0.928	0.894
Model 1.2	1490.885	0.957	0.912	0.928	0.895
Model 2	2704.857	0.841	0.800	0.918	0.797

- Lower AIC value.
- Better ROC curve and higher AUC value.
- Higher Precision: indicates more reliability in the model's prediction of positive examples.
- Higher recall: indicates that the model is able to identify most of the actual positive examples.
- Higher accuracy: indicates that the model has better overall prediction performance.

### Model 1.2 (Remove missing values in Length)

$$\ln\left(\frac{p}{1-p}\right) = \hat{\alpha} + \hat{\beta}_{Year} \cdot Year + \hat{\beta}_{Length} \cdot Length + \hat{\beta}_{Budget} \cdot Budget + \hat{\beta}_{Votes} \cdot Votes + \hat{\beta}_{Genre} \cdot \Pi_{Genre}(x)$$

## Conclusion

$$odds = e^{\log-odds}$$

For **continuous** variables:

1. It is predicted that an increase in the **length** of the film may **negatively** affect the film's ability to receive a high score of 7 or higher.
2. The higher the **budget** of the film, the higher the probability of the film's rating becoming **higher** improves the **most**.

	variable <chr>	estimate <dbl>
1	(Intercept)	0.00000
2	year	1.00659
3	length	0.93981
4	budget	1.68036
5	votes	1.00004
6	genreAnim...	0.64280
7	genreCome...	29.07030
8	genreDocu...	204.36068
9	genreDrama	0.24291
10	genreRoma...	0.53606

For **categorical** variables:

1. **Animated** films, **dramas** and **romances** are **less likely** to score above 7 than **action films**.
2. **Comedies** and **documentaries** are significantly more likely to receive high scores. **Documentaries** are **significantly more likely** to receive a **high score** than other genres.



# Conclusion

## Implication

- Well-crafted, more costly films are the ones that are clearly more enjoyable to viewers.
- Audiences as well as film raters seem to favour documentaries with more authenticity than other types of films. Also comedy films are well liked by the public.
- Meanwhile drama is perhaps not as popular with audiences due to the form of artistic expression and the time period in which it was made.
- These data and conclusions will give filmmakers and entertainment industry workers some insights into the market and deeper artistic thinking.

# Conclusion

## Validity

*Sample size*

Num of observations:2874

*P-value*

Remove variables that do not have a significant effect.

*Outlier handling*

Models comparison  
Standardisation of data

*Limitations*

- Neglect non-linear relationships
- The influence of time on aesthetics

*Model Selection (GLM)*

- Response variable: binomial distribution
- Explanatory variables are not highly correlated

# Conclusion

## Future Extension

The study examines the trend of change over the years between the audience's aesthetics of the film.

01

Comparison of film genre preferences across age groups.

02

How the cast of a film is put together can increase the film's box office.

03





University  
of Glasgow

Group 09  
20-March-2024

**Thanks for your attention**

