IMDB Film Prediction Summary

By Obi-Win

1. Introduction

IMDb ratings for movies have emerged as a clear and immediate representation of a movie's quality. For many audiences, these ratings have become an important factor, influencing their choice to watch a particular movie or give it a miss. Therefore, predicting these ratings is useful for different stakeholders, as it can drive decisions related to production, marketing strategies, and so on.

As we approach November 2023, it is exciting to meet the impending release of twelve anticipated blockbusters. We want to apply the predictive data analytics skills that we learned in class to forecast the IMDb ratings of these upcoming movies. We have data from roughly 2,000 movies on IMDb, which include encompassing ratings, actors, genres, and several other crucial variables. Based on our understanding and the professional knowledge of the film industry, we removed some irrelevant columns before starting exploration. Then based on the preliminary data processing result, we selected the most influential variables to construct our statistical model. After that, we used feature engineering and selection methods to build the model. In the end, the results showcase the rating for each film and the winner is Napoleon! The result section describes the overall performance of the model and the significance of each predictor. Also, by evaluating the model's R-squared value and Mean Squared Error, the report conveys the efficacy of the model in predicting IMDb scores.

In summary, this report illustrates our approach to forecasting IMDb movie ratings using statistical analysis. The model also holds potential for predicting scores of other upcoming films in the future.

2. Data Description

Here are the steps we took for feature engineering:

- Construct running averages and running film counts for the director, cinematographer, and three actors. Note that here we calculate all the films an actor has participated in, instead of just their actor number (the same actor who has appeared as actor 1 in one film and actor 2 in another will have 2 film counts). We fill the first film by the filmmaker with an average score of 6 across the board and add the max score of a director.
- Standardize IMDB_pro_meter and transform nb_news_articles to log.
- Only train on English films by the UK and USA.
- For distributors, construct film counts and average film score features.
- Add inflation effect to the movie budget, applying a 4% increase for each year before 2023. Standardize the new budget feature.
- Build winter months dummy features (if a release month is in Oct, Nov, Dec, then it equals 1).
- Change films by Pixar to Disney. Construct Disney*animation and Universal*Animation (as animations by these two distributors tend to have distinct characteristics).
- Add dummy variables for each film decade.
- Add comedy, documentary, and biography categories.
- Bin the durations into three new dummy features: films with duration <100 mins, films with duration >=100 and < 150 mins, and films with duration >=150 mins.
- Construct dummy variables based on maturity rating. Here we pre-grouped all films into just G, PG, and R.

actor_star_meter: Initial EDA on the **actor1_star_meter** shows that there are outliers in the datasets. After removing the outliers, we draw a scatter plot and fail to observe any correlation between this variable and the imdb_score.

actor1_film_running_avaerge; actor2_film_running_avaerge; actor3_film_running_avaerge: These three variables represent the effect of actors on the overall performance of a movie. Our research shows that the effect of these three variables are almost equivalent with a positive correlation with the target variable.

Keywords: We excluded the "**Keywords**" column from the IMDb plot of the film because it appears to be different for every single film, making it challenging to analyze using methods like dummy variables or data aggregation methods.

movie_meter_IMDBpro: The correlation between "**movie_meter_IMDBpro**" and the movie rating is not very strong. Consequently, we have opted to standardize this variable before constructing the model.

nb_faces: Regarding the number of faces in the main movie poster, we conducted simple linear regressions between the rating and this variable. However, we observed that the regression between the number of faces ("**nb_faces**") and the IMDb score had a very low R-squared value, and a negative and weak correlation. Consequently, we decided to remove this column from our analysis.

Poly(director_film_running_avaerge, 2): The director_film_running_average stands for the average score before the production of a target movie. We are convinced that this predictor can represent the quality of a production since the styles and skills of director is determinant. Also, you can see we used a poly here, which is used for fully capturing the complexity of this variable. As you can see from the image in the appendix, a second degree polynomial regression model perfectly explains the relationship between these two variables.

Information about Cinematographer, Distributor, and Production: As for the **cinematographer**, we firstly computed the average scores of the cinematographer. Then we integrated these mean scores into the original dataframe, associating them with the respective cinematographers. Following this, we conducted a regression analysis between the IMDb score and the mean score. We obtained a satisfactory R-squared value and a significantly low p-value. In the cases of both **distributor company** and **production company**, we employed a method similar to that used for cinematographers. We constructed a simple linear regression model between the movie rating and the aggregated mean scores of the distributor company/production company. In both models, we identified reasonable R-squared values and notably low p-values as well. For cinematographer, we choose not to employ second degree polynomial regression for cinematographer variable. There are two main reasons: the influence of a cinematographer is relatively weak compared to director, and the complexity of the influence of cinematographers is lower, given the fact that while director is a more multidimensional position compared to them. As a result, we chose to use a simple regression.

Genres: In the film industry, Drama and Comedy are the most produced genres, indicating broad appeal, while Action and Thriller are also prominent, reflecting a taste for dynamic storytelling. Horror and Fantasy have fewer productions, suggesting niche audiences or cautious investment. Genres like Drama are evenly distributed across films, but Western and Animation are rarer. Action, Sci-Fi, and Romance show moderate right-skewness, commonly featured but not dominant. Genres like Action and Thriller often coexist, while Drama and Horror rarely do. Low VIF values indicate minimal multicollinearity among genres, each contributing uniquely to the model. Correlations between genre and IMDb score are weak, except for Drama, which shows a moderate positive correlation. Linear models suggest genre modestly influences IMDb scores, with Drama, War, and Crime linked to higher scores, and Action, Sci-Fi, Thriller, and Horror to lower ratings. However, a movie's rating is more influenced by a combination of elements than genre alone.

Disney_animation; Universal_animation: These variables indicate whether a film can be classified as an animation movie released by The Walt Disney Company and the Universal. They are deemed as valuable because animation films from these two distributors have different characteristics based on past experience. Our exploration towards these variables indicate that they have major positive effects on the IMDB score.

nb_news_articles_log: The indicator, nb_news_articles, captures the level of public attention and discussion. This metric is highly correlated with public interest and sentiment, so we have decided to retain it. However, because the data itself is right-skewed, we have applied a logarithmic transformation to preprocess the data, resulting in the variable

country_dummy_USA: The **country** of movie production can reflect unique cultural, social, and artistic characteristics in the filmmaking process.The dataset includes movies produced by 34 countries, and 89.74% of the movies are produced by the USA and UK. In these movies, only 2 movies are not English.

duration_0_100, duration_100_150: The **duration** of a movie often reflects its narrative pacing and complexity, and it can influence audience immersion and expectations. Therefore, we consider dividing the duration into three segments: 0-100 minutes, 100-150 minutes, and over 150 minutes to examine their varying effects on movie scores.

maturity_rating_G & maturity_rating_PG: When it comes to the level of the film, we classified all the movies into three categories, R, G and PG. The result of our research proved our hypothesis, showing that G is positively correlated with IMDB score while PG has a negative effect on it.

3. Model Selection

As we observed no particular non-linear relationship (except for running director average score) in our EDA, we decided to focus mainly on **feature engineering and selection** to build our final model. After exploring the datasets, we found there are many categorical variables in the original datasets. We decided that instead of constructing them into dummy variables, we will instead conduct mean aggregation and film counts grouped by these variables and construct new variables accordingly. The justification for this method is because prior number of films and film average score will indicate future film performance like what we have seen in real life (films by Ridley Scott in 2023 should have a higher probability of getting high scores than first time directors). We also conducted single variable regression for these variables on the imdb score, and as expected, they all demonstrated positive correlation to a certain degree.

After constructing all the datasets, we first ran a regression on all dependent variables and reduced the number of variables based on the resulting significance levels. During this process, we also decided to drop the decades features to avoid overfitting, as the data in the training datasets appeared to have selection bias (older films tend to show a higher score even though this is probably caused by survivor bias).

We kept all of the features in linear form, except for the average director score, and applied a degree 2 polynomial function to it. This was based on the scatter plot between it and the IMDB score. The correlation heatmap is in the appendix.

4. Results

4.1 Predictions for the 12 movies

Movie Title	Predicted Score
Pencils vs Pixels	6.346877
The Dirty South	6.044617
The Marvels	6.609539
The Holdovers	7.633899
Next Goal Wins	6.635942
Thanksgiving	6.061710
The Hunger Games: The Ballad of Songbirds and Snakes	7.527531
Trolls Band Together	6.298546
Leo	6.830714
Dream Scenario	6.877393
Wish	7.447020
Napoleon	7.652941

Predictive scores for upcoming movies indicate varying audience receptions. "The Holdovers" and "Napoleon" show strong potential with scores above 7.6, suggesting high favorability. In contrast, "The Dirty South" and "Thanksgiving" have more modest scores around 6.0, indicating mixed responses upon release, while films like "The Marvels" and "Leo" fall in the mid-range, hinting at moderate success.

4.2 Model R-Squared

The R-squared value of the model is approximately 0.5087. This metric denotes the proportion of the variance in the dependent variable that is predictable from the independent variables. The value suggests that the model explains about 50.87% of the variability of the target variable, which is a moderate level of explanatory power. In the context of movie ratings, this means that the predictors in the model account for slightly more than half of the variations we see in IMDb scores.

4.3 Predictive Power

The model's Mean Squared Error (MSE) is approximately 0.5781. In simpler terms, the MSE gives us an idea of how close our predictions are to the actual IMDb scores. A lower MSE means our predictions are closer to the real scores, whereas a higher MSE suggests there might be a significant difference between our predictions and the actual scores.

With an MSE of 0.5781, if we think of it in terms of IMDb ratings which are on a scale of 1 to 10, our predictions, on average, are off by a margin of roughly 0.76 (since the square root of 0.5781 is approximately 0.76). This means that if our model predicts an IMDb score of 8 for a movie, the actual score could likely be between 7.24 and 8.76.

4.4 Significance of Each Predictor

nb_news_articles_log: This predictor has a positive coefficient, suggesting movies with more news articles tend to have higher IMDb scores. Given its highly significant p-value (***), this predictor plays a crucial role in our model.

movie_meter_IMDBpro_norm: While included in the model, it doesn't seem to be statistically significant in predicting IMDb scores, given its high p-value. But its coefficient value is large enough to impact our predicted results. Therefore, we decided to keep it.

Genres (action, thriller, musical, horror, drama, animation, crime, comedy): Genres like action, thriller, horror, and comedy have negative coefficients, suggesting movies of these genres tend to have lower IMDb scores compared to the reference genre. Conversely, drama and animation have positive coefficients, implying they tend to score higher. Most of these genre predictors are statistically significant, especially action, thriller, horror, drama, and comedy (***).

Other genres: Several other predictors are in the model, but some, like ‘documentary', ‘biography’, ‘disney_animation’, and ‘animation_universal’, are not statistically significant, suggesting they might not be crucial in predicting in the training dataset. However, this insignificance can be contributed by the lack of samples in our training dataset. We decided to keep these features for our testing dataset because they show particular impact on our research.

avg_distributor_imdb_score: This predictor has a strong positive relationship with the IMDb score. It's highly significant (***), indicating movies from distributors with higher average IMDb scores also tend to have higher IMDb scores.

country_dummy_USA: Movies from the USA tend to have slightly lower IMDb scores. This predictor is statistically significant (*).

duration_0_100 and duration_100_150: These predictors show the impact of movie duration on its IMDb score. Both are significant, with movies of duration 0-100 having a notably negative impact on the IMDb score compared to the reference duration category (***).

director_film_running_avaerge (polynomial term): This predictor has a non-linear relationship with IMDb score. Both terms are statistically significant, especially the first term (***).

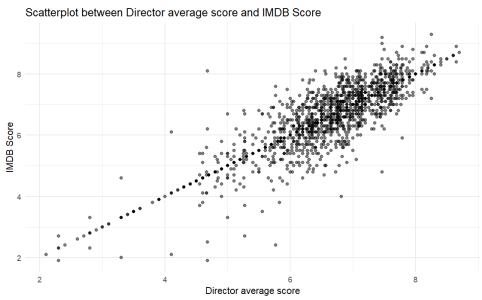
cinematographer_film_running_avaerge and actor1_film_running_avaerge: Both have a positive relationship with IMDb score. The first predictor is statistically significant at the 0.05 level (*), while the second is significant at the 0.001 level (**).

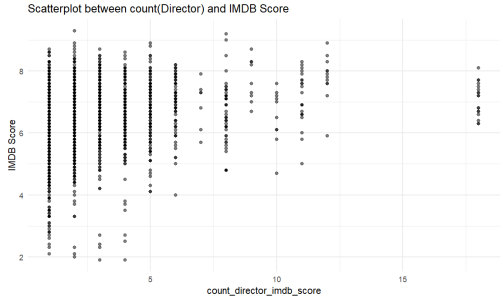
maturity_rating_PG and maturity_rating_G: Movies with a PG rating tend to score lower than the reference maturity rating. This predictor is highly significant (***). Movies with a G rating show no significant relationship.

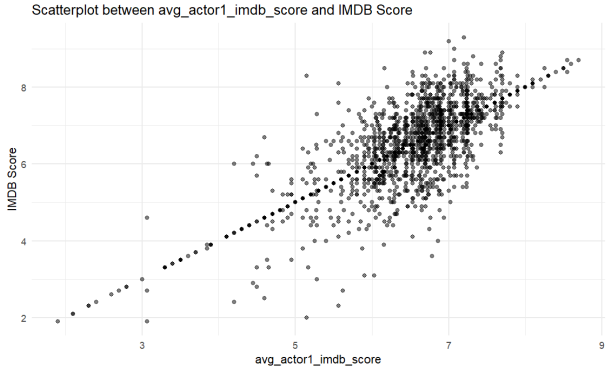
director_max: This predictor, representing the maximum rating a director has received, has a positive coefficient. It's statistically significant (***), suggesting that movies directed by directors with higher maximum IMDb scores in the past tend to have higher IMDb scores.

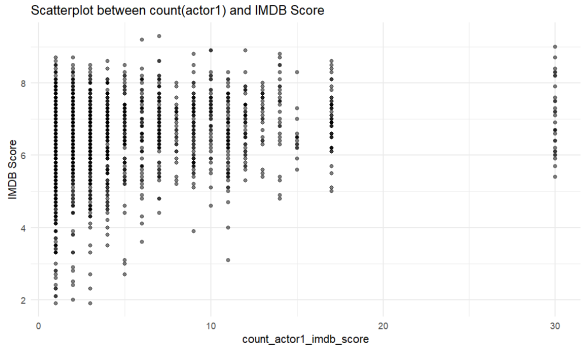
Appendix

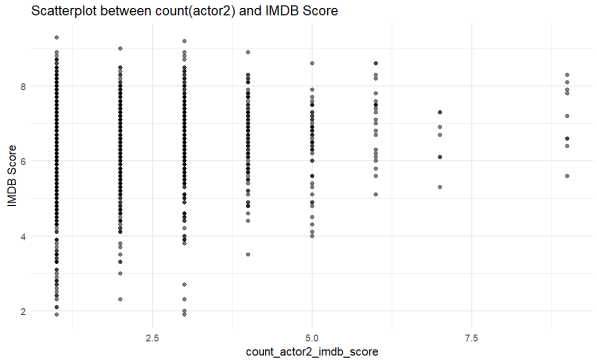
Actors and Director:

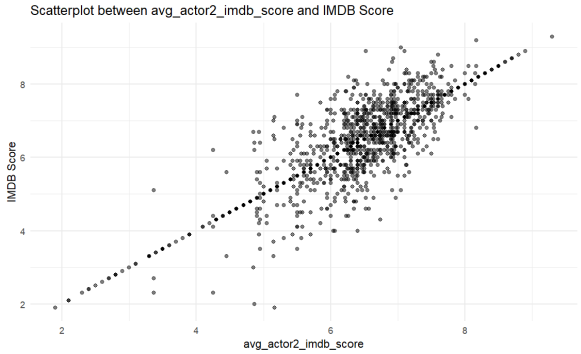


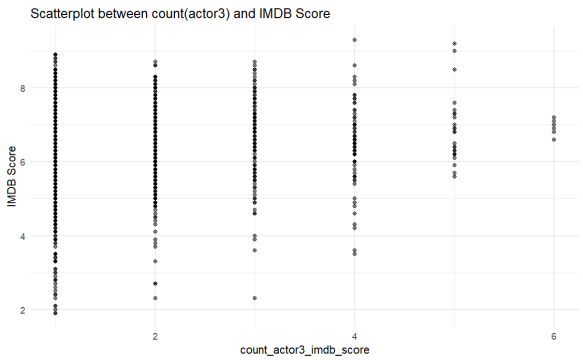


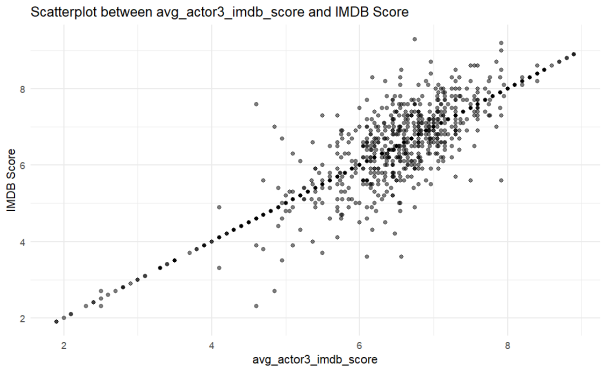












Regression of imdb_scores on Actor/Director average and film counts:

```
Call:
lm(formula = imdb_score ~ avg_director_imdb_score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3667 -0.1333  0.0000  0.1667  3.4250

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -6.728e-14  7.913e-02    0.00        1
avg_director_imdb_score  1.000e+00  1.202e-02   83.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5135 on 1928 degrees of freedom
Multiple R-squared:  0.7822,    Adjusted R-squared:  0.7821
F-statistic:  6924 on 1 and 1928 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = imdb_score ~ avg_actor1_imdb_score, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-3.267 -0.200  0.000  0.250  3.150

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -2.200e-14  1.132e-01    0.00        1
avg_actor1_imdb_score  1.000e+00  1.723e-02   58.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6639 on 1928 degrees of freedom
Multiple R-squared:  0.636,     Adjusted R-squared:  0.6358
F-statistic:  3368 on 1 and 1928 DF,  p-value: < 2.2e-16
```

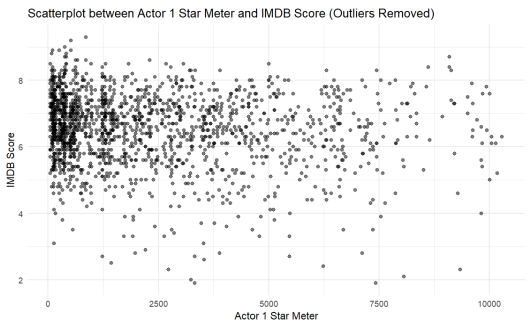
```
Call:
lm(formula = imdb_score ~ avg_actor2_imdb_score, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-3.267  0.000  0.000  0.000  2.375

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.952e-14  8.245e-02    0.00        1
avg_actor2_imdb_score 1.000e+00  1.253e-02   79.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5302 on 1928 degrees of freedom
Multiple R-squared:  0.7678,    Adjusted R-squared:  0.7677
F-statistic:  6374 on 1 and 1928 DF,  p-value: < 2.2e-16

>
```

```
Call:
lm(formula = imdb_score ~ count_director_imdb_score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7024 -0.6224  0.1376  0.7576  2.8576

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.282412   0.036552 171.878   <2e-16 ***
count_director_imdb_score 0.080005   0.009431   8.483   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.08 on 1928 degrees of freedom
Multiple R-squared:  0.03598,   Adjusted R-squared:  0.03548
F-statistic: 71.96 on 1 and 1928 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = imdb_score ~ count_actor2_imdb_score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6663 -0.6155  0.1329  0.7353  2.8353

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.41393    0.04465 143.637  < 2e-16 ***
count_actor2_imdb_score  0.05078    0.01920   2.646  0.00822 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.098 on 1928 degrees of freedom
Multiple R-squared:  0.003617,  Adjusted R-squared:  0.0031
F-statistic: 6.999 on 1 and 1928 DF,  p-value: 0.008223
```

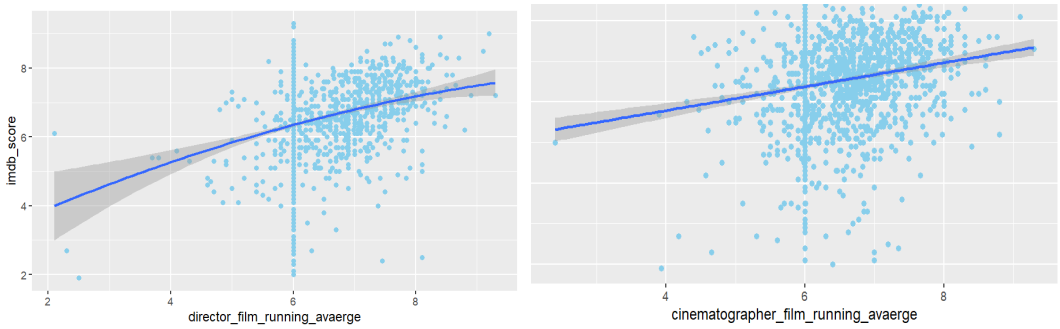
```
Call:
lm(formula = imdb_score ~ count_actor1_imdb_score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5437 -0.5802  0.1237  0.7371  2.7102

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.334055   0.033609 188.461  < 2e-16 ***
count_actor1_imdb_score 0.036536   0.004693   7.785 1.13e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.083 on 1928 degrees of freedom
Multiple R-squared:  0.03047,   Adjusted R-squared:  0.02997
F-statistic:  60.6 on 1 and 1928 DF,  p-value: 1.131e-14
```

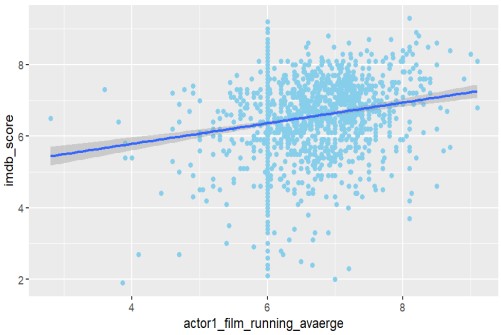
Actor_star_meter:

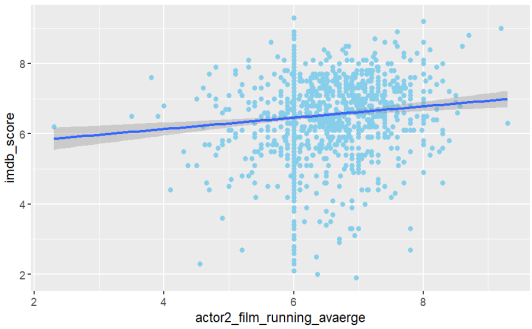


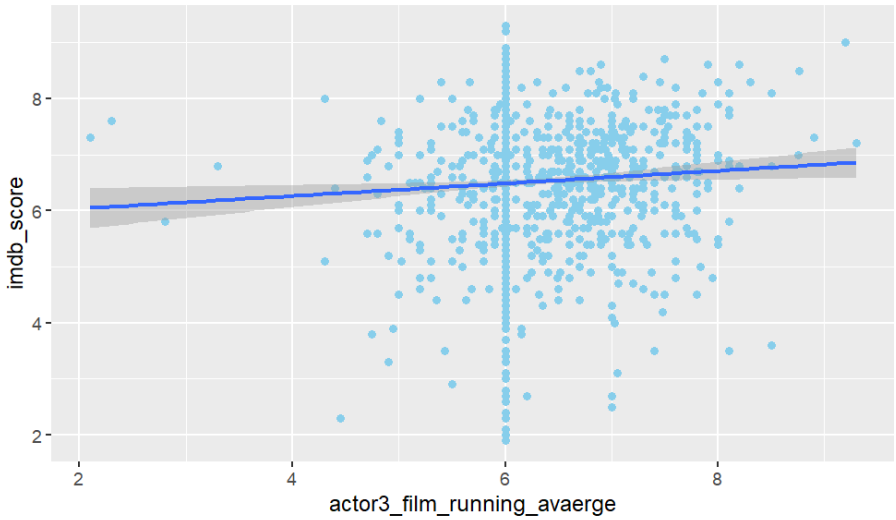

```
Call:
lm(formula = imdb_score ~ standardized_actor1_star, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6103 -0.6097  0.0905  0.7900  2.7904

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.51181    0.02504 260.102   <2e-16 ***
standardized_actor1_star  0.03182    0.02504   1.271    0.204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.1 on 1928 degrees of freedom
Multiple R-squared:  0.0008368, Adjusted R-squared:  0.0003186
F-statistic: 1.615 on 1 and 1928 DF,  p-value: 0.204
```

	Dependent variable:
	imdb_score
nb_news_articles_log	0.092***
	(0.009)
movie_meter_IMDBpro_norm	0.014
	(0.028)
action	-0.268***
	(0.056)
thriller	-0.236***
	(0.053)
musical	-0.272***
	(0.080)
horror	-0.256***
	(0.070)
drama	0.186***
	(0.047)
animation	0.616***
	(0.222)
crime	0.115**
	(0.054)
comedy	-0.254***
	(0.049)
animation_universal	-0.354
	(0.794)
avg_distributor_imdb_score	0.655***
	(0.038)
country_dummy_USA	-0.157**
	(0.064)
duration_0_100	-0.902***
	(0.116)
duration_100_150	-0.572***
	(0.110)
disney_animation	0.787
	(0.796)
academy_months	0.003
	(0.043)
poly(director_film_running_avaerge, 2)1	5.261***
	(0.811)
poly(director_film_running_avaerge, 2)2	-2.078***
	(0.763)
cinematographer_film_running_avaerge	0.056**
	(0.029)
actor1_film_running_avaerge	0.088***
	(0.028)
actor2_film_running_avaerge	0.015
	(0.029)
actor3_film_running_avaerge	0.014
	(0.032)
maturity_rating_G	0.073
	(0.151)
maturity_rating_PG	-0.180***
	(0.042)
director_max	0.060***
	(0.018)
biography	0.144*
	(0.074)
documentary	0.167
	(0.270)
Constant	1.025**
	(0.462)
Observations	1,557
Log Likelihood	-1,751.942
Akaike Inf. Crit.	3,561.885
Note:	*p<0.1; **p<0.05; ***p<0.01

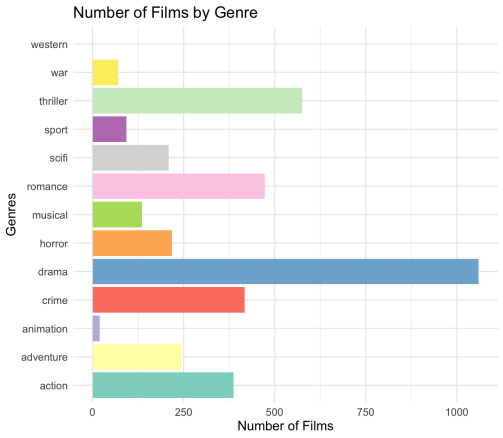


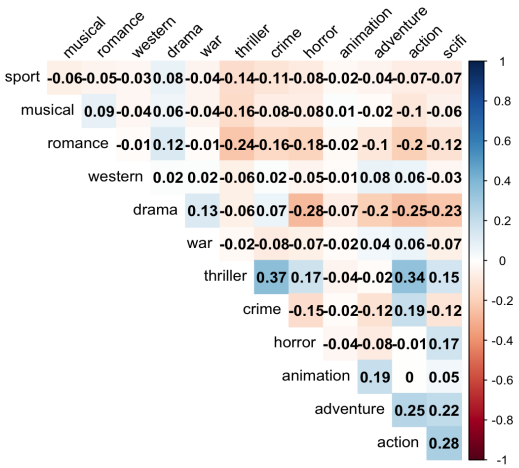

imdb_score
actor1_film_running_avaerge


imdb_score
actor2_film_running_avaerge


imdb_score
actor3_film_running_avaerge

Genres:



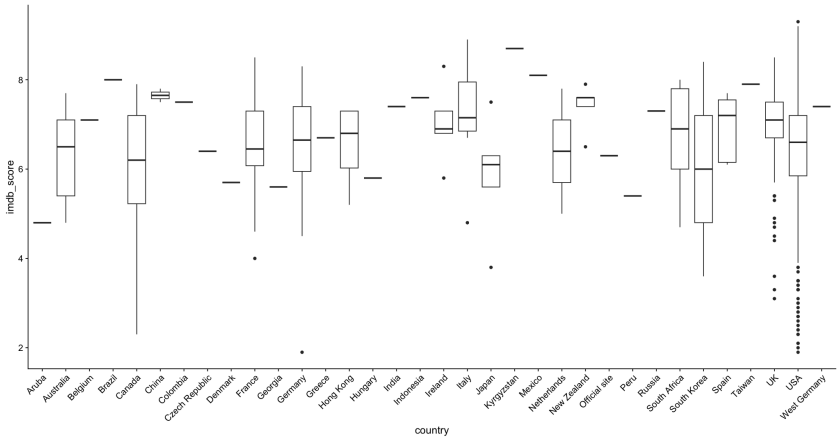



```
> interpretation
              Genre Correlation Direction Strength
action       action -0.15905761  Negative     Weak
adventure adventure -0.06689042  Negative     Weak
scifi         scifi -0.09380293  Negative     Weak
thriller   thriller -0.08003569  Negative     Weak
musical     musical -0.02265507  Negative     Weak
romance     romance -0.01488314  Negative     Weak
western     western  0.06553298  Positive     Weak
sport         sport  0.05500145  Positive     Weak
horror       horror -0.16607140  Negative     Weak
drama         drama  0.33820387  Positive Moderate
war             war  0.10854429  Positive     Weak
animation animation  0.01657983  Positive     Weak
crime         crime  0.06144428  Positive     Weak
```

```
[[1]]
                Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    6.5994167 0.02765483 238.635205 0.000000e+00
imdb[, genre] -0.4368844 0.06175814  -7.074119 2.096054e-12

[[2]]
                Estimate Std. Error    t value    Pr(>|t|)
(Intercept)    6.5397983 0.02673719 244.595544 0.000000000
imdb[, genre] -0.2213557 0.07519683  -2.943684 0.003282087

[[3]]
                Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    6.5477629 0.02640638 247.961437 0.000000e+00
imdb[, genre] -0.3319735 0.08024432  -4.137034 3.669494e-05

[[4]]
                Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    6.5691513 0.02979568 220.473298 0.0000000000
imdb[, genre] -0.1924556 0.05458810  -3.525597 0.0004323968

[[5]]
                 Estimate Std. Error     t value  Pr(>|t|)
(Intercept)    6.51867336 0.02597142 250.9941212 0.0000000
imdb[, genre] -0.09734983 0.09783739  -0.9950166 0.3198531

[[6]]
                 Estimate Std. Error     t value  Pr(>|t|)
(Intercept)    6.52113933 0.02882309 226.2470843 0.0000000
imdb[, genre] -0.03805265 0.05822219  -0.6535763 0.5134628

[[7]]
                Estimate Std. Error    t value    Pr(>|t|)
(Intercept)   6.5021624 0.02521533 257.865461 0.000000000
imdb[, genre] 0.5478376 0.18997823   2.883686 0.003974098

[[8]]
                Estimate Std. Error   t value   Pr(>|t|)
(Intercept)   6.4982036 0.02563338 253.50552 0.00000000
imdb[, genre] 0.2824416 0.11677317   2.41872 0.01566731

[[9]]
                Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    6.576986 0.02622367 250.803401 0.000000e+00
imdb[, genre] -0.576986 0.07802682  -7.394714 2.102335e-13

[[10]]
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   6.1012644 0.03510605 173.79523 0.000000e+00
imdb[, genre] 0.7475092 0.04737047  15.78007 7.482679e-53

[[11]]
                Estimate Std. Error    t value     Pr(>|t|)
(Intercept)   6.4886559 0.02536228 255.838816 0.000000e+00
imdb[, genre] 0.6384869 0.13317354   4.794398 1.756616e-06

[[12]]
                Estimate Std. Error     t value  Pr(>|t|)
(Intercept)   6.5099476 0.02517341 258.6041620 0.0000000
imdb[, genre] 0.1800524 0.24728948   0.7281036 0.4666386

[[13]]
                Estimate Std. Error    t value    Pr(>|t|)
(Intercept)   6.4763384 0.02823433 229.378124 0.000000000
imdb[, genre] 0.1641892 0.06074187   2.703064 0.006930645
```

Country:

```
> table(imdb$country)

        Aruba     Australia       Belgium        Brazil        Canada
            1            23             1             1            38
        China      Colombia Czech Republic       Denmark        France
            2             1             1             1            40
      Georgia       Germany        Greece     Hong Kong       Hungary
            1            34             1             4             1
        India     Indonesia       Ireland         Italy         Japan
            1             1             5             8             5
   Kyrgyzstan        Mexico   Netherlands   New Zealand Official site
            1             1             2             5             1
         Peru        Russia  South Africa   South Korea         Spain
            1             1             5             2             7
       Taiwan            UK           USA  West Germany
            1           177          1555             1
```

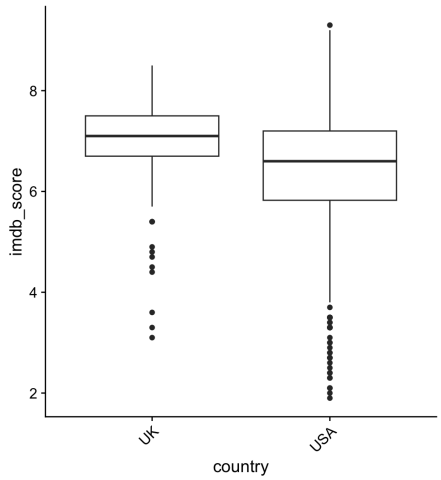


Language:

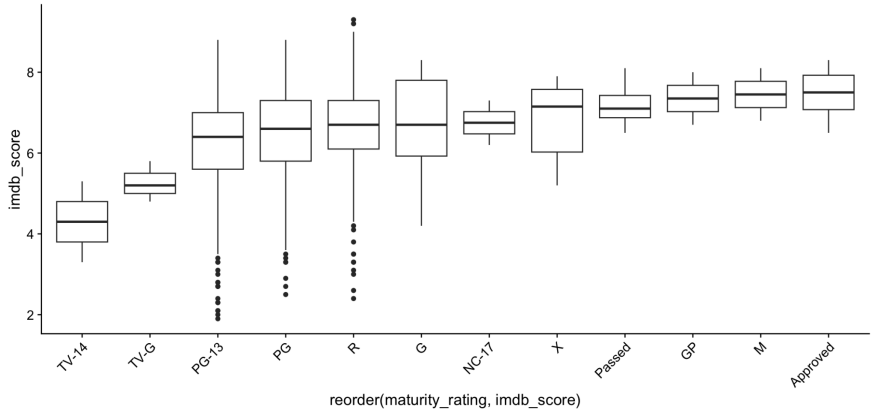
```
> table(imdb$language)

Aboriginal    Aramaic  Cantonese       Dari      Dutch    English     French     German      Hindi Indonesian
         1          1          2          1          1       1892          7          3          1          1
   Italian   Japanese     Korean   Mandarin  Mongolian       None Portuguese    Spanish       Zulu
         3          2          1          2          1          2          1          6          2
```

If we keep English films only, the distribution will be like:

```
> table(imdb_english$country)

        Aruba     Australia       Belgium        Canada         China
            1            22             1            35             1
Czech Republic       Denmark        France      Georgia       Germany
            1             1            35             1            31
       Greece     Hong Kong       Hungary       Ireland         Italy
            1             2             1             5             6
        Japan    Kyrgyzstan   Netherlands   New Zealand Official site
            3             1             1             5             1
         Peru  South Africa   South Korea         Spain            UK
            1             4             1             5           175
          USA  West Germany
         1550             1
```



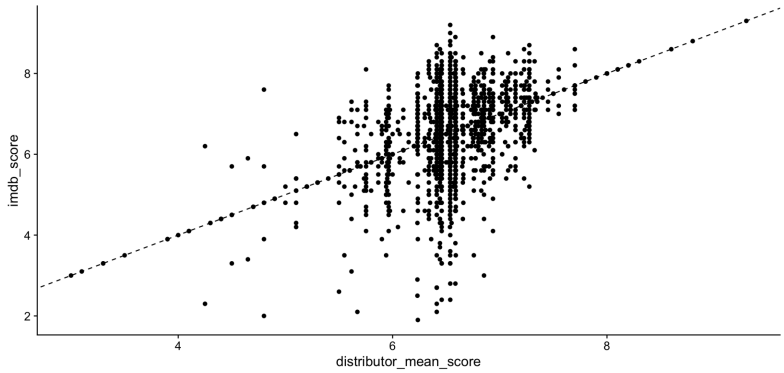
Maturity_rating:




```
> table(imdb_new$maturity_rating)

Approved        G       GP        M    NC-17   Passed       PG    PG-13        R    TV-14     TV-G        X
      20       28        2        2        2        4      233      527      894        2        3        8
```

```
> summary(lm_maturity_rating)

Call:
lm(formula = imdb_score ~ maturity_rating, data = imdb_new)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3142 -0.5660  0.1253  0.7340  2.6340

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             7.4950     0.2372  31.592  < 2e-16 ***
maturity_ratingG       -0.7771     0.3106  -2.502 0.012447 *
maturity_ratingGP      -0.1450     0.7869  -0.184 0.853817
maturity_ratingM       -0.0450     0.7869  -0.057 0.954401
maturity_ratingNC-17   -0.7450     0.7869  -0.947 0.343868
maturity_ratingPassed  -0.2950     0.5811  -0.508 0.611776
maturity_ratingPG      -1.0203     0.2472  -4.127 3.85e-05 ***
maturity_ratingPG-13   -1.2808     0.2417  -5.299 1.32e-07 ***
maturity_ratingR       -0.8290     0.2399  -3.456 0.000562 ***
maturity_ratingTV-14   -3.1950     0.7869  -4.060 5.12e-05 ***
maturity_ratingTV-G    -2.2283     0.6569  -3.392 0.000709 ***
maturity_ratingX       -0.6950     0.4438  -1.566 0.117565
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.061 on 1713 degrees of freedom
Multiple R-squared:  0.05352,   Adjusted R-squared:  0.04744
F-statistic: 8.806 on 11 and 1713 DF,  p-value: 2.21e-15
```

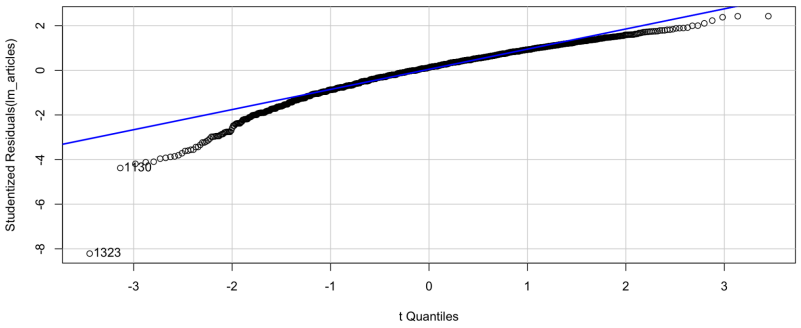
Distributor:




```
> summary(lm_distributor_mean_score)

Call:
lm(formula = imdb_score ~ distributor_mean_score, data = imdb_new)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3355 -0.4500  0.0000  0.5634  2.8000

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.296e-14  2.787e-01    0.00        1
distributor_mean_score 1.000e+00  4.267e-02   23.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9469 on 1723 degrees of freedom
Multiple R-squared:  0.2417,    Adjusted R-squared:  0.2413
F-statistic: 549.3 on 1 and 1723 DF,  p-value: < 2.2e-16
```

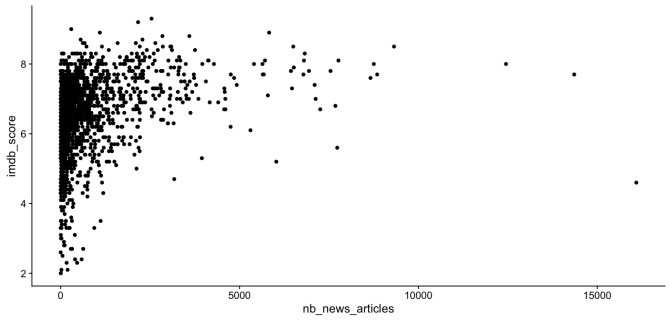
Nb_news_articles:

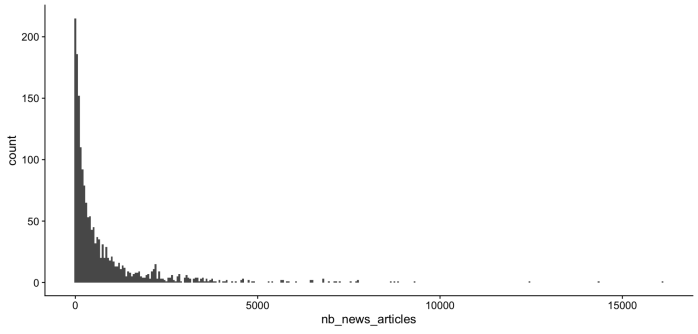
```
> summary(imdb_new$nb_news_articles)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    82.0   295.0   800.7   862.0 60620.0
```




```
> outlierTest(lm_articles)  # Those with p < 0.05 are reported
      rstudent unadjusted p-value Bonferroni p
1323 -8.214517         4.1354e-16   7.1335e-13
1130 -4.378771         1.2653e-05   2.1826e-02
```

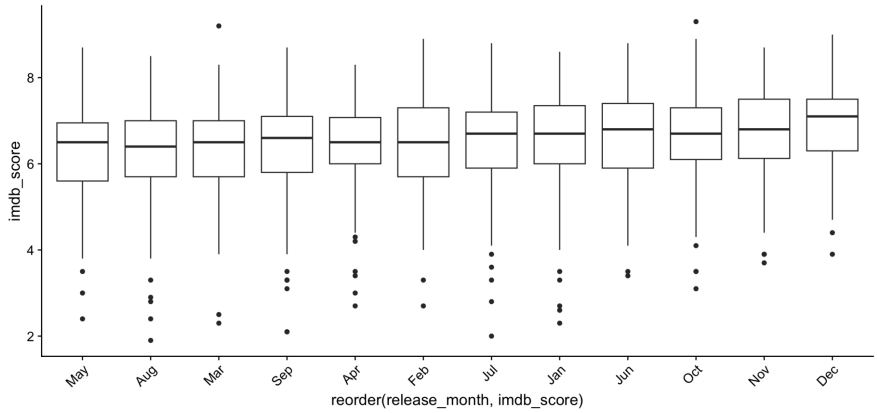
```
> imdb_new[c(1130, 1323), c("movie_title", "nb_news_articles", "imdb_score")]
                          movie_title nb_news_articles imdb_score
1130                     Disaster Movie              725        1.9
1323 Star Wars: Episode IV - A New Hope            60620        8.7
```





(right skewed)

Academy_month:


imdb_score
8
6
4
2
May
Aug
Mar
Sep
Apr
Feb
Jul
Jan
Jun
Oct
Nov
Dec
reorder(release_month, imdb_score)