MGSC 661: Multivariate Statistics (Fall 2022)

Midterm Project

Professor: Juan Camilo Serpa

Group: We R Boba

Dreama Wang - 261112206

Iris Liu - 260795028

Sangwoon Park - 261121800

Tristan Wang - 261093294

Table of Contents:

Introduction

IMDb, the internet movie database, has the world's most authorized source of information about movies, TV shows, and video games, as well as other industry professionals, including directors, actors, and producers (Lavery, 2017). IMDb allows its user to rate specific content from 1 to 10 stars. Individual votes will then be aggregated and summarized as a single IMDb rating, visible on the title's main page (IMDb 2022). The process of searching and giving ratings is free of charge.

For this project, our focus is solely on movie-related content. The dataset we are going to analyze contains historical data of more than 19,000 movies, including their information about language, country, actors, etc., in a total of 39 different variables. Then, based on this dataset, our goal is to build a model which can generate the most accurate prediction of the IMDb score of any given movie.

Building the model requires three steps. The first step is analyzing each variable and determining their usabilities. Suppose we decide to keep a particular variable. In that case, we will move to the second step - cleaning the data by regrouping the categorical variables and changing the range of continuous variables or other methods to best fit for our purpose. Lastly, using the cleaned dataset, we seek to build a model which will balance the tradeoff between large adjusted R-squared and low MSE when performing cross-validation.

To validate our final model, we will then utilize it to predict the IMDb scores of 12 movies releasing in November 2022. Our ultimate objective is to achieve the lowest predictive error for these 12 movies using the model we built and win movie tickets and popcorn.

Data Description

1) **Continuous Variables:**

The 6 columns of the table below represent:
Distribution - Identify the distribution according to the boxplot and histogram.
Outlier - Whether the variable contains outliers according to the boxplot. Yes or No.

Heteroscedasticity - Whether the variable contains heteroscedasticity

Relationship - Variable's relationship with imdbScore. Positive or Negative.

P-value - Whether the model between the variable and imdbScore has a p-value less than 0.05 or not. Yes or No.

R-squared - R-squared of the model between the variable and imdbScore.

#	Variable	Distribution	Outlier	Heteroscedasticity	Relationship	P-value	R-squared
1	imdbScore	Left_skewed	Yes	-	-	-	-
2	movieBudget	Right_skewed	No	Yes	Negative	Yes	0.0062
3	releaseYear	Left_skewed	Yes	No	Negative	Yes	0.0380
4	duration	Right_skewed	Yes	Yes	Positive	Yes	0.1686
5	nbNewsArticles	Right_skewed	Yes	Yes	Positive	Yes	0.0508
6	actor1_starMeter	Right_skewed	Yes	No	Positive	No	0.0008
7	actor2_starMeter	Right_skewed	Yes	No	Positive	No	0.0015
8	actor3_starMeter	Right_skewed	Yes	No	Negative	No	0.0001
9	nbFaces	Right_skewed	Yes	No	Negative	Yes	0.0080
10	movieMeter_IMDBpro	Right_skewed	Yes	Yes	Negative	Yes	0.0081

2) Categorical Variable (See Appendix A)

3) Data Preprocessing

- movieBudget: Since the value of money has been inflated with time, we need to adjust the budget for each movie to its value as in 2022. To do this, an inflation data, which is calculated based on the consumer price index, was collected online. The based year for this dataset is 1980. We calculated the inflation-adjusted budget based on the following equation. Suppose the budget for a specific movie released in 1999 is $27,000,000, to get the equivalent value in 2022:

$$\frac{360.2}{202.18} \times \$27,000,000 \approx \$48,102,681,$$

where 360.2 is the value in 2022, and 202.18 is the value in 1999. The output is $48,102,681, we repeated this process for all movies.

- releaseDay: Drop this variable.
- releaseMonth: This variable does not require any processing.
- releaseYear: Since most data are clustered at the higher end, we decided to drop any movie released before 1980.
- Duration: Since movies with too long and too short duration are all very rare, we only kept movies with duration between 60 and 200 minutes.
- Language: Since the majority of movies are in English, we regrouped it into “English” and “Others”.
- country: Since the majority of movies are produced in the USA and the UK, we regrouped it into “USA”, “UK”, and “Others”.
- maturityRating: Change TV-14 to PG-13, TV-G to G, and NC-17 to R, because they have similar age limits.
- aspectRatio: Irrelevant information. Drop this variable.
- distributor: Hard to include it in our model. Drop this variable.
- nbNewsArticles: Drop this variable.
- director: Drop this variable
- actor1: Hard to include it in our model. Drop this variable.
- actor1_starMeter: Due to the extremely right-skewed pattern (Figure A6.2), we decided to keep the entries which are below 3rd quantile.
- actor2_starMeter: Due to the extremely right-skewed pattern (Figure A6.2), we decided to keep the entries which are below 3rd quantile.
- actor3_starMeter: Due to the extremely right-skewed pattern (Figure A6.2), we decided to keep the entries which are below 3rd quantile.
- colourFilm:This variable does not require any processing.
- Genres: Drop this variable
- nbFaces: Since movies with too many faces on the poster are very rare, we only kept movies with less than 15 faces.
- plotKeywords: Difficult to analyze this variable based on the techniques that we had learned so far. Drop this variable.
- genres (dummy variables, including action, adventure, etc.): This variable does not require any processing.

- movieMeter_IMDBpro: Since most data is clustered at the lower end (Figure A8.3), we wanted to keep data below the 3rd quartile, similar to the starMeter. However, in the test dataset, there's one movie that has a movieMeter of 14223, which is more than the 3rd quantile of the overall dataset. We decided to drop any movie with movieMeter more than 20000 instead to avoid extrapolation.
- cinematographer: Hard to include it in our model. Drop this variable.
- productionCompany: Hard to include it in our model. Drop this variable.

Our final cleaned dataset contains 29 variables: movieTitle, movieID, imdbScore, releaseYear, movieBudget_inflated, releaseMonth, duration, language2, country2, maturityRating, actor1_starMeter, actor2_starMeter, actor1_starMeter, colourFilm, nbFaces, action, adventure, scifi, thriller, musical, romance, western, sport, horror, drama, war, animation, crime, movieMeter_IMDBpro.

Model Selection

Step 1: To build a reasonable model, we created dummy variables for all categorical predictors. Then, we created Model 1, which includes all predictors except moveTitle and movieID in the cleaned dataset.

Step 2: To detect nonlinearity, we plotted residual plots for every variable against the target variable imdbScore (Figure B1). We found that duration, nbFaces, movieMeter, and releaseYear were not linear since these variables had curved patterns. We will further analyze these variables in the later steps. Moreover, according to the Tukey Test, the p-value 0.072 is greater than 0.05. Therefore, the overall model is reasonably linear.

Step 3: To detect heteroskedasticity, we first drew a residual plot for Model 1. The pattern displays a funnel shape (Figure B2). To further confirm our visual analysis, we conducted ncvTest. Since the p-value was close to 0, we concluded heteroskedasticity is present in our model (Table B1). Moreover, to resolve this issue, we corrected Model 1 using coeftest.

Step 4: We conducted an outlier test to detect potential outliers. There are 4 outliers in the model, and we removed those entries to obtain Model2. Compared to Model 1, the adjusted R-squared of Model 2 increased from 0.474 to 0.49

Step 5: We checked collinearity based on the variance inflation factor (Table B2). All predictors have VIF less than 4; thus, there is no sign of collinearity Model 2.

Step 6: Referring to Step 2, there are four non-linear variables - duration, nbFaces, movieMeter, and releaseYear. We tried creating multiple polynomial and spline regression models to select the best degree. We tried to build 4 polynomial regression models for the variable duration with degrees 1 to 5. According to the ANOVA test between the 4 models (Table B3), duration with degree 2 is the best selection because when compared to degree 1, it has a p-value of 0.02658, which was significantly better at alpha level = 0.05. Compared to degree 3, the p-value was 0.36809, which was not significant. Therefore, we added a quadratic variable for the duration, obtaining Model 3. Compared to Model 2, the adjusted R-squared of Model 3 increased from 0.49 to 0.4923. Trying to find another way to resolve non-collinearity for the duration, we also tried to create 5 spline regression models with degree 1 to 5 and knot at 140. According to the plots, degree 2 and 3 might be the better choices (Figure B3). However, we did not include a spline for the variable duration to avoid overfitting.

Step 7: For the nbFaces, we ran ANOVA test to find the optimal polynomial model. Since the p-values for degree 2 were larger than 0.05, we determine that the polynomial with degree 1 works better than degree 2. By looking through the plot of nbFaces (Figure A 9.3), most of the observations were between 0 and 8 and should be integers. We stated that spline regression was not the best model to capture the nbFaces variable's shape. Thus, we decided to use linear regression for nbFaces and keep Model 3 with adjusted R-squared as 0.4923.

Step 8: Since movieMeter_IMDBpro is not linear, we checked the polynomial regression model from degree 1 to degree 5 through ANOVA test. We detected that the cubic regression is the optimized model for movieMeter_IMDBpro because the ANOVA test state the p-value for degree 3 is less than 0.5, which is significant to have polynomial regression with degree=3 for movieMeter_IMDBpro. We also tried spline functional form to see if it can improve the model since the movieMeter_IMDBpro has a unique pattern (see Figure A 7.3). Based on the scatterplot, we chose to include 3 knots at 5000,10000,15000 (Figure B4) and visualized them using ggplot and decided to choose the spline regression at degree 5 to fit in our model called Model 4 with adjusted R-squared of 0.5188.

Step 9: We also ran the ANOVA test for releaseYear to determine the optimal polynomial model from degree 1 to degree 5. Since the quadratic regression had a p-value of 0.016 and the cubic

regression had the p-value of 0.82 which was larger than 0.05, we chose quadratic polynomial regression for releaseYear. Since the plot (Figure A 3.3) between releasesYear and the imdbScore did not show a unique pattern, we decided not to use spline regression for this variable andd define the model with polynomial regression at degree 2 as Model 5 with adjusted R-squared of 0.5214.

Step 10: For the final model, Model 6, we checked the significance for each predictor and removed these variables,releaseMonth_Oct, releaseMonth_May,releaseMonth_Sep, war, releaseMonth_Apr, that have p-value larger than 0.7. End with a model that has an adjusted R-squared of 0.5241.

Results

---

1) Our final regression model was selected using 25 predictor variables from a total of 39 variables (original dataset) to predict IMDb scores of 12 different movies, which will be released in November. The result showed that our model is **statistically significant** ($p <$ .001, *Multiple R-squared* = .5472, *Adjusted R-squared* = .5241).
2) As our final goal is to predict, cross-validation was crucial to avoid an overfitting issue. We have initially conducted 2 tests – Validation set and LOOCV – to measure the model's predictive power (i.e., out-of-sample performance). The mean squared error (MSE) we obtained for **the LOOCV test was 0.451**, and **0.341 for the Validation set test**, both being rounded to 3 decimals; it's important to note that the result obtained from the validation set test is not trustworthy because of its drawbacks: highly variable MSE and wasted observations. Additionally, we also ran a K-fold test using K=10 folds and obtained **an average mse of 0.454**.
3) The analysis on our final model's predictors shows that releaseYear, movieBudget_inflated, duration, country, nbFaces, specific movie genres, and movieMeter_IMDBPro were **significant predictors for IMDb score**.
    a) For one unit increase in each of the continuous variable, we can expect that IMDb score will be increased by its coefficient below:

Predictor	Coefficient
releaseYear (Polynomial)	Interpretability Issue
movieBudget_inflated	-8.296e-09
Duration (Polynomial)	Interpretability Issue
nbFaces	-0.0286
movieMeter_IMDBPro (Polynomial Splines)	Interpretability Issue

b) For categorical predictor country, movies produced in the USA ($p > .05$) have a 0.04982 higher IMDb score than movies produced in other countries; movies produced in UK ($p < .05$) has 0.3577 higher IMDb score than movies produced in other countries.

c) For binary variable drama, romance, horror and animation, drama movies have a 0.3575 higher IMDb score compared to non drama movies, 0.0137 lower for romance movies, 0.5262 lower for horror movies, and 1.217 higher for animation movies.

4) Several categorical and continuous variables – language, maturityRating, actor1_starMeter, actor2_starMeter, actor3_starMeter, releaseMonth, and colourFilm_USA – **did not show statistical significance** ($p > .05$) for our final model. We still included these variables in our final model because they seem to be influencing factors. They just look not significantly important for our IMDB dataset. If we have more recent variables – for example, rank of production company, popularity rank of the type of movie – and their data, these variables may become significantly important. For the future project, we could try to use machine learning models to forecast movie rates, resulting in a more accurate prediction using such variables.

5) Predictions for the 12 movies using our final model

Movie Title	Predicted IMDb Score
Falling for Christmas	6.5026
Black Panther: Wakanda Forever	6.0246
Spirited	5.8098
Paradise City	5.1124
Poker Face	4.7788
¡Que viva México!	6.0478
Slumberland	5.9838
Blue’s Big City Adventure	6.5104
The Menu	5.9942
The Fabelmans	7.5063
Devotion	6.2011
Strange World	6.7039

References

(2022) IMDb. Available at: https://hel
p.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#

Lavery, T. (2017) What is internet movie database (imdb)?, WhatIs.com. TechTarget. Available at: https://www.techtarget.com/whatis/definition/Internet-Movie-Database-IMDb

Appendix A

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.900	5.900	6.600	6.512	7.300	9.300

Table A1.1 - Summary of IMDb rating

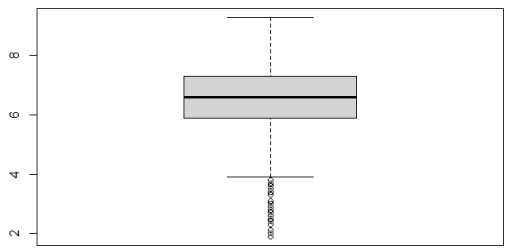


Figure A1.1 - Boxplot of IMDb rating

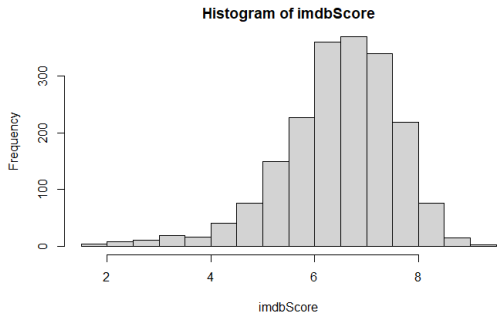
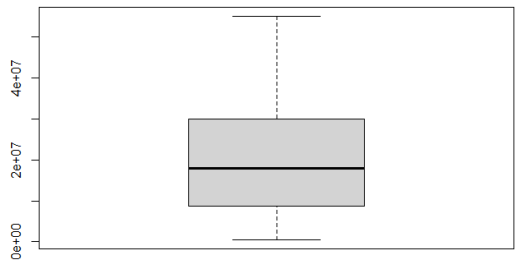


Figure A1.2 - Histogram of IMDb rating

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
560000	8725000	18000000	20973774	30000000	55000000

Table A2.1 - Summary of movies' budget



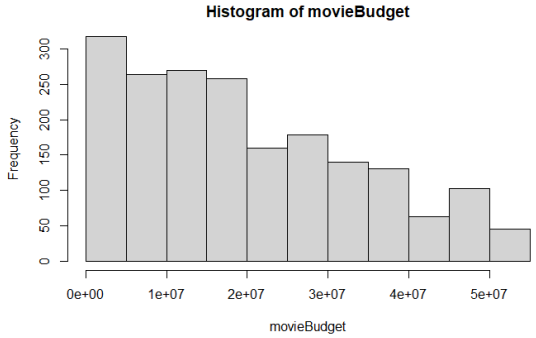


Figure A2.1 - Boxplot of movies' budget

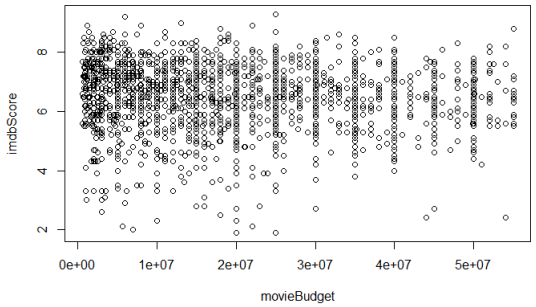


Figure A2.2 - Histogram of movies' budget

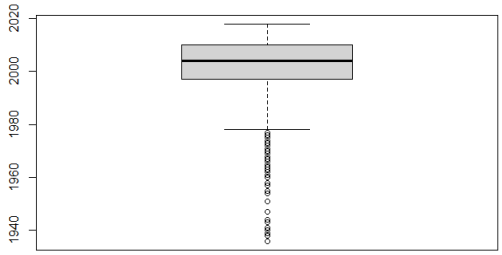
```
lm(formula = imdbScore ~ movieBudget)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6176 -0.5879  0.1303  0.7528  2.8120

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.636e+00  4.365e-02 152.010  < 2e-16 ***
movieBudget -5.916e-09  1.707e-09  -3.465 0.000542 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.097 on 1928 degrees of freedom
Multiple R-squared:  0.006189,  Adjusted R-squared:  0.005673
F-statistic: 12.01 on 1 and 1928 DF,  p-value: 0.0005418
```

Figure A2.3 - Scatterplot of movies' budget and IMDb rating

Table A2.2 - Summary of linear regression model between IMDb rating and movie's budget

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1936	1997	2004	2001	2010	2018

Table A3.1 - Summary of release year



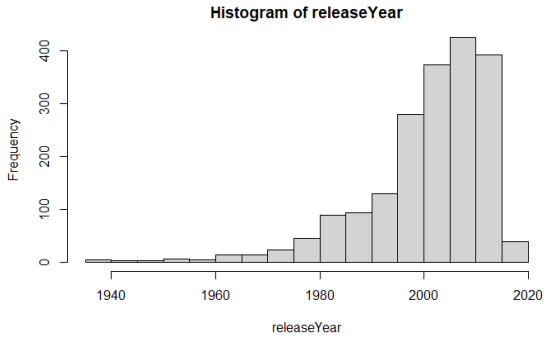
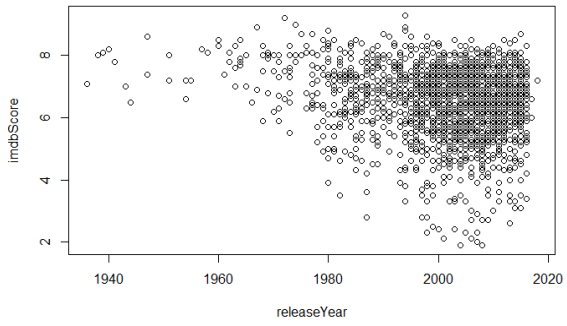


Figure A3.1 - Boxplot of release year

Figure A3.2 - Histogram of release year




```
lm(formula = imdbScore ~ releaseYear)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5643 -0.6006  0.1448  0.7629  2.6537

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.935761   4.176398  10.281   <2e-16 ***
releaseYear -0.018199   0.002087  -8.722   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.079 on 1928 degrees of freedom
Multiple R-squared:  0.03796,   Adjusted R-squared:  0.03746
F-statistic: 76.07 on 1 and 1928 DF,  p-value: < 2.2e-16
```

Figure A3.3 - Scatterplot of release year and IMDb rating

Table A3.2 - Summary of linear regression model between IMDb rating and release year

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
37.0	96.0	106.0	109.7	118.0	330.0

Table A4.1 - Summary of duration of films

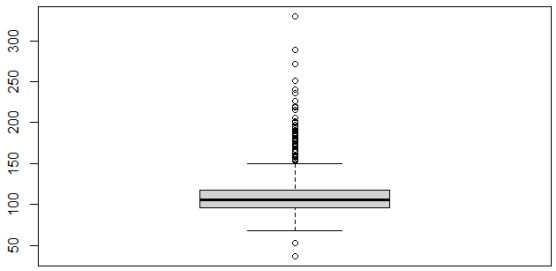


Figure A4.1 - Boxplot of duration of films

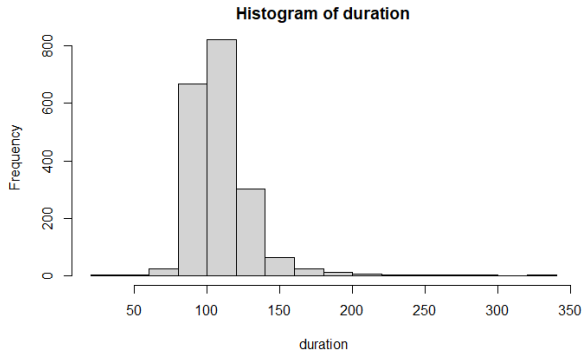


Figure A4.2 - Histogram of duration of films

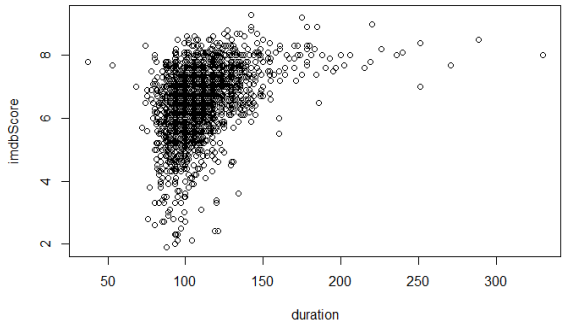


Figure A4.3 - Scatterplot of duration of films and IMDb rating

```
lm(formula = imdbScore ~ duration)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3521 -0.5568  0.0979  0.6812  2.8339

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.179462   0.120134   34.79   <2e-16 ***
duration    0.021261   0.001075   19.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.003 on 1928 degrees of freedom
Multiple R-squared:  0.1686,    Adjusted R-squared:  0.1682
F-statistic: 391.1 on 1 and 1928 DF,  p-value: < 2.2e-16
```

Table A4.2 - Summary of linear regression model between IMDb rating and duration of films

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    78.0   286.0   770.6   845.5 60620.0
```

Table A5.1 - Summary of number of articles in news

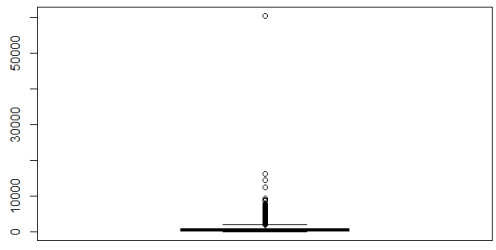


Figure A5.1 - Boxplot of number of articles in news

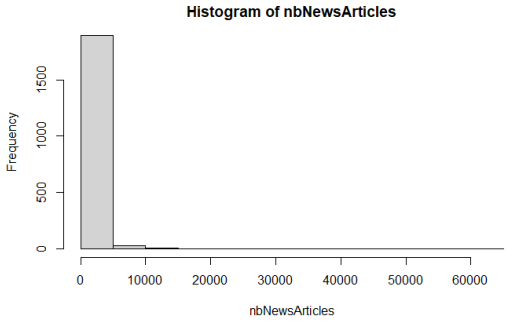


Figure A5.2 - Histogram of number of articles in news

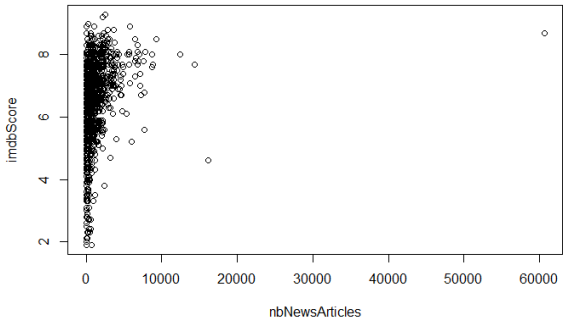


Figure A5.3 - Scatterplot of number of articles in news and IMDb rating

```
lm(formula = imdbScore ~ nbNewsArticles)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7900 -0.5381  0.1259  0.7430  2.5523

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.409e+00  2.641e-02  242.65   <2e-16 ***
nbNewsArticles 1.333e-04  1.312e-05   10.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072 on 1928 degrees of freedom
Multiple R-squared:  0.05083,   Adjusted R-squared:  0.05034
F-statistic: 103.2 on 1 and 1928 DF,  p-value: < 2.2e-16
```

Table A5.2 - Summary of linear regression model between IMDb rating and number of articles

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      9     505    1888   21190    4665 8342201
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      3    1895    3986   17114    7667 5529461
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      8    3075    5856   35469   12250 6292982
```

Table A6.1 - Summary of the 2022 ranking of actors/actresses 1, 2 and 3 made by IMDbPro

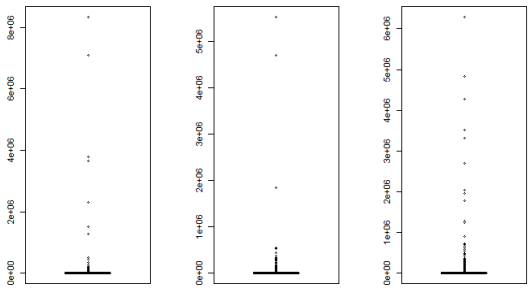


Figure A6.1 - Boxplot of the 2022 ranking of actors/actresses 1, 2 and 3

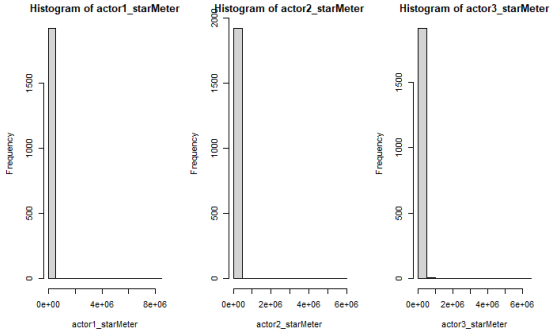


Figure A6.2 - Histogram of the 2022 ranking of actors/actresses 1, 2 and 3

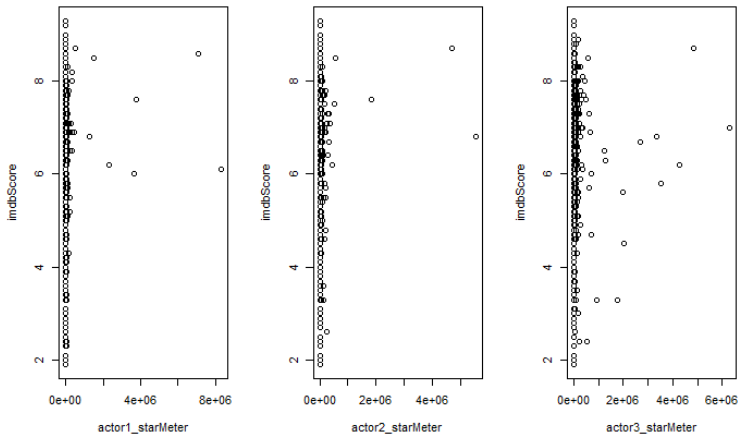


Figure A6.3 - Scatterplot of the ranking of actors/actresses and IMDb rating

```
lm(formula = imdbScore ~ actor1_starMeter)

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)     6.509e+00  2.510e-02 259.297   <2e-16 ***
actor1_starMeter 1.113e-07  8.756e-08   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

lm(formula = imdbScore ~ actor2_starMeter)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6105 -0.6086  0.0920  0.7904  2.7905

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.508e+00  2.515e-02 258.758   <2e-16 ***
actor2_starMeter 2.430e-07  1.445e-07   1.682   0.0928 .
---
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:  0.001465,  Adjusted R-squared:  0.0009471
F-statistic: 2.829 on 1 and 1928 DF,  p-value: 0.09276

lm(formula = imdbScore ~ actor3_starMeter)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6124 -0.6123  0.0877  0.7876  2.7877

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.512e+00  2.528e-02 257.640   <2e-16 ***
actor3_starMeter -1.720e-08  9.620e-08  -0.179    0.858
---
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:  1.657e-05, Adjusted R-squared:  -0.0005021
F-statistic: 0.03195 on 1 and 1928 DF,  p-value: 0.8582
```

Table A6.2 - Summary of linear regression model between IMDb rating and the ranking of actors/actresses

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.00	1.00	1.44	2.00	31.00

Table A7.1 - Summary of nbFaces

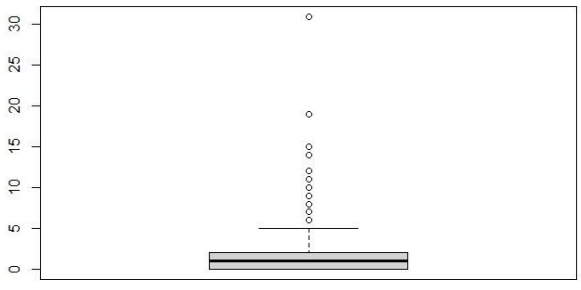


Figure A7.1 - Boxplot of the nbFaces

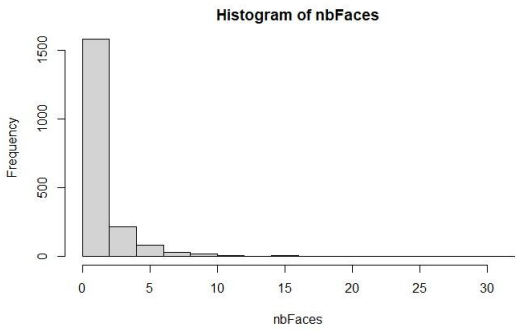


Figure A7.2 - Histogram of nbFaces

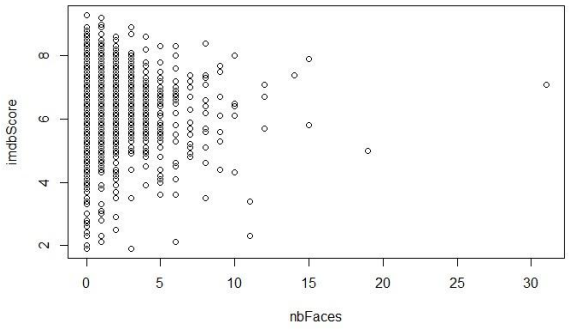


Figure A7.3 - Scatterplot of nbFaces

```
lm(formula = imdbScore ~ nbFaces)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6805 -0.5851  0.1195  0.7627  2.7195

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.58055    0.03044 216.201  < 2e-16 ***
nbFaces     -0.04773    0.01211  -3.941 8.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.096 on 1928 degrees of freedom
Multiple R-squared:  0.007992,  Adjusted R-squared:  0.007478
F-statistic: 15.53 on 1 and 1928 DF,  p-value: 8.395e-05
```

Table A7.2 - Summary of linear regression between IMDb rating and nbFaces

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     71    2836    5406   11612   10198  849550
```

Table A8.1 - Summary of the 2022 ranking of movies made by IMDbPro

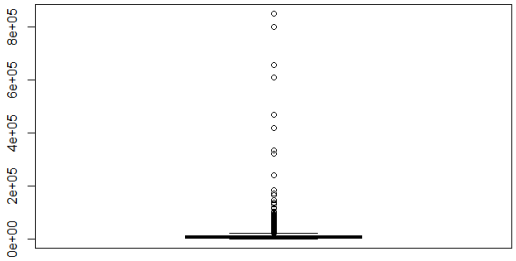


Figure A8.1 - Boxplot of the 2022 ranking of movies

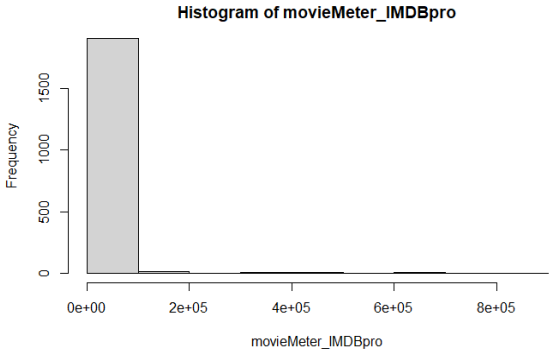


Figure A8.2 - Histogram of the 2022 ranking of movies

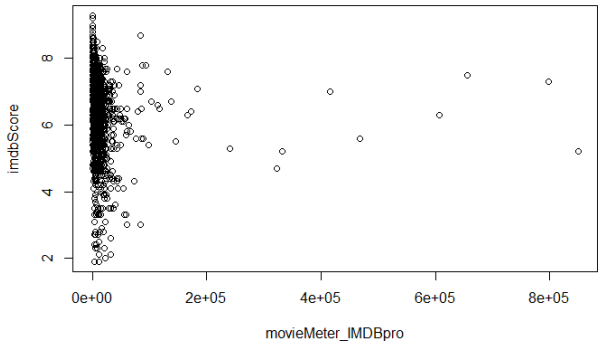


Figure A8.3 - Scatterplot of the 2022 ranking of movies and IMDb rating

```
lm(formula = imdbScore ~ movieMeter_IMDBpro)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6311 -0.6207  0.1009  0.7667  2.7602

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.540e+00  2.597e-02 251.877  < 2e-16 ***
movieMeter_IMDBpro -2.457e-06  6.210e-07  -3.956  7.9e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.096 on 1928 degrees of freedom
Multiple R-squared:  0.008052,  Adjusted R-squared:  0.007537
F-statistic: 15.65 on 1 and 1928 DF,  p-value: 7.897e-05
```

Table A8.2 - Summary of linear regression model between IMDb rating and the 2022 ranking of movies made by IMDbPro

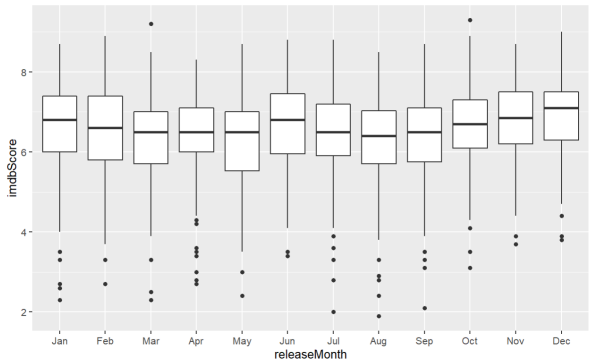


Figure A9.1 - Boxplot of releaseMonth

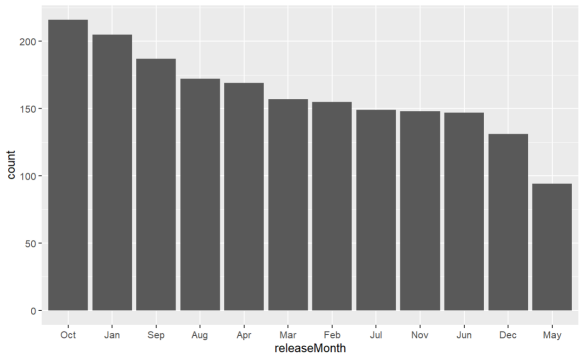


Figure A9.2 - Bar graph of releaseMonth

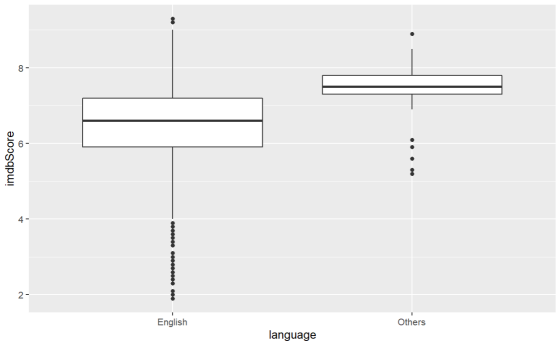


Figure A10.1 - Boxplot of language

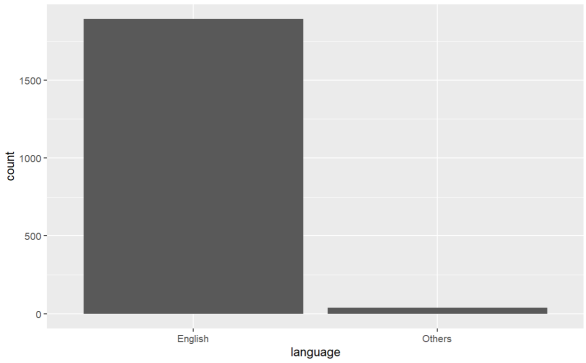


Figure A10.2 - Bar graph of language

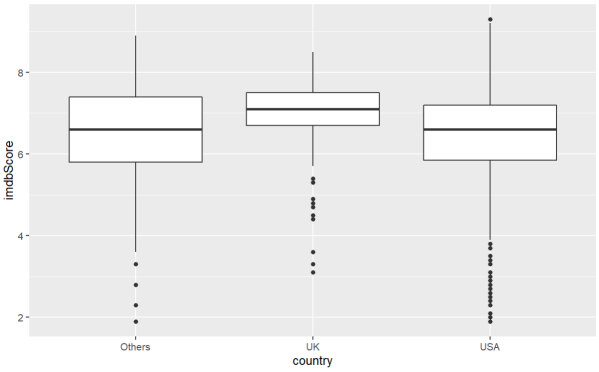


Figure A11.1 - Boxplot of country

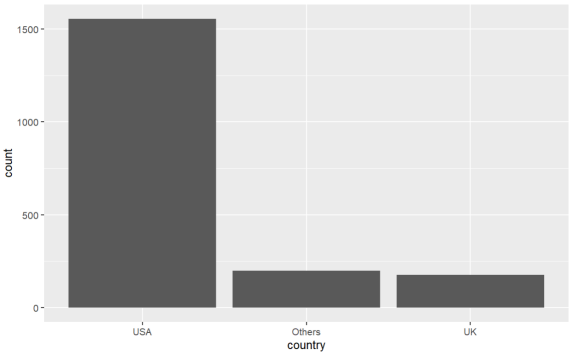


Figure A11.2 - Bar graph of country

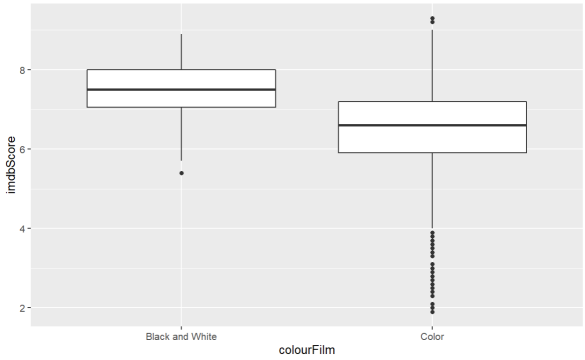


Figure A12.1 - Boxplot of colourFilm

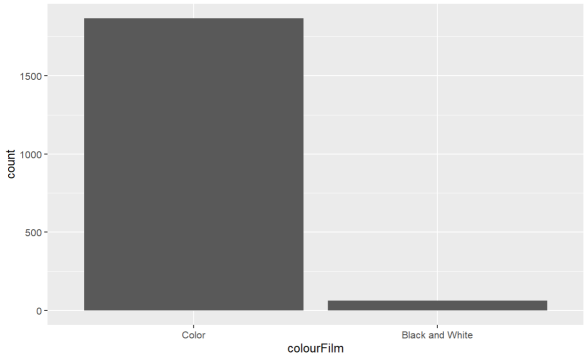


Figure A12.2 - Bar graph of colourFilm

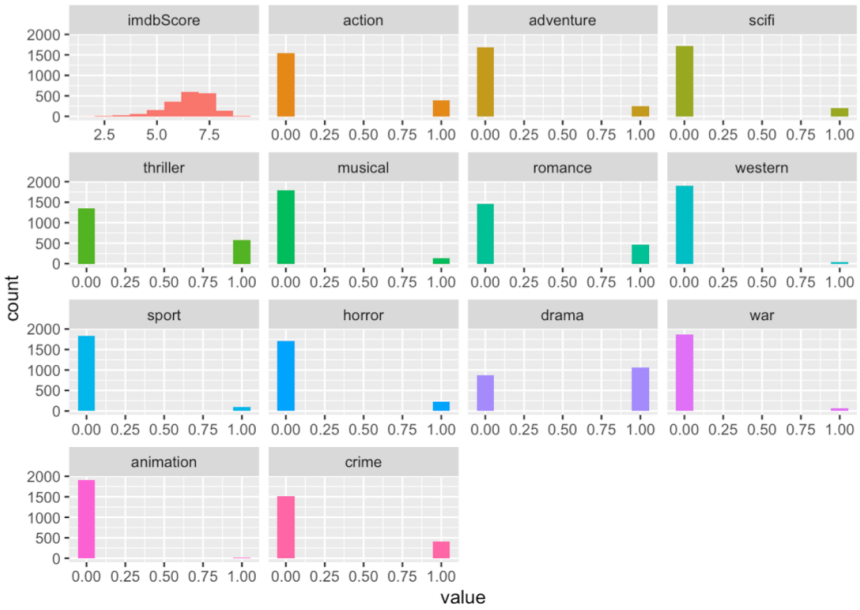
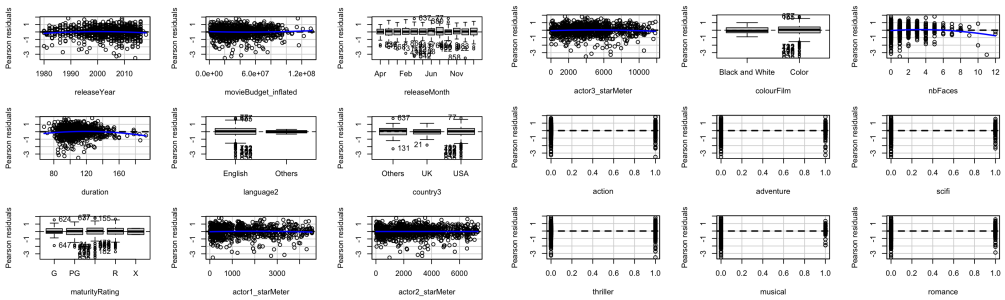


Figure A13.2 - Combined bar graphs of all genres

Appendix B



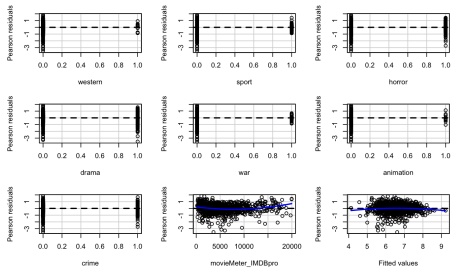


Figure B1 - Residual plots for each variable against the target variable imdbScore

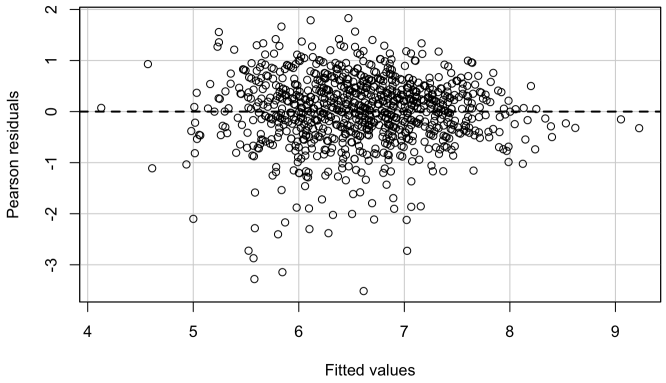


Figure B2 - Residual plot for Model 1

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 48.28682, Df = 1, p = 3.6822e-12
```

Table B1 - NcvTest for Model 1

```
##                          GVIF Df GVIF^(1/(2*Df))
## releaseYear          1.349579  1        1.161714
## movieBudget_inflated 1.501929  1        1.225532
## releaseMonth         1.624893 11        1.022311
## duration             1.524051  1        1.234525
## language2            1.123003  1        1.059719
## country2             1.184901  2        1.043327
## maturityRating       2.046022  4        1.093613
## actor1_starMeter     1.257360  1        1.121321
## actor2_starMeter     1.213970  1        1.101803
## actor3_starMeter     1.173720  1        1.083384
## colourFilm           1.078529  1        1.038523
## nbFaces              1.159369  1        1.076740
## action               1.500791  1        1.225068
## adventure            1.400983  1        1.183631
## scifi                1.247875  1        1.117083
## thriller             1.622814  1        1.273897
## musical              1.096732  1        1.047250
## romance              1.305134  1        1.142424
## western              1.054906  1        1.027086
## sport                1.214751  1        1.102157
## horror               1.422827  1        1.192823
## drama                1.504059  1        1.226401
## war                  1.112333  1        1.054672
## animation            1.247101  1        1.116737
## crime                1.559346  1        1.248738
## movieMeter_IMDBpro   1.172732  1        1.082927
```

Table B2 - Variance inflation factor of Model 2

```
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    867 393.41
## 2    866 391.19  1   2.22465 4.9350 0.02658 *
## 3    865 390.82  1   0.36556 0.8109 0.36809
## 4    864 389.51  1   1.31762 2.9229 0.08769 .
## 5    863 389.03  1   0.47656 1.0572 0.30415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table B3 - ANOVA test between the 5 models with different degree for variable duration

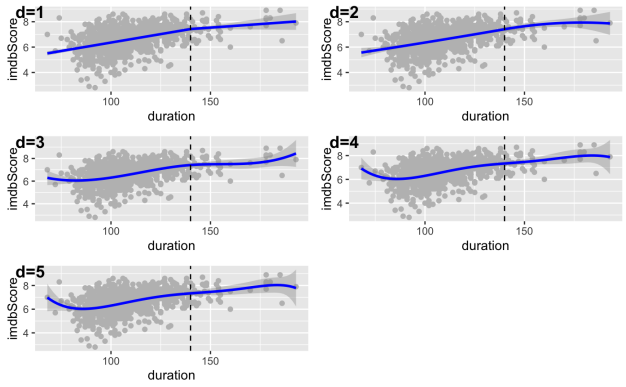


Figure B3 - Spline model for variable duration with degree from 1 to 5

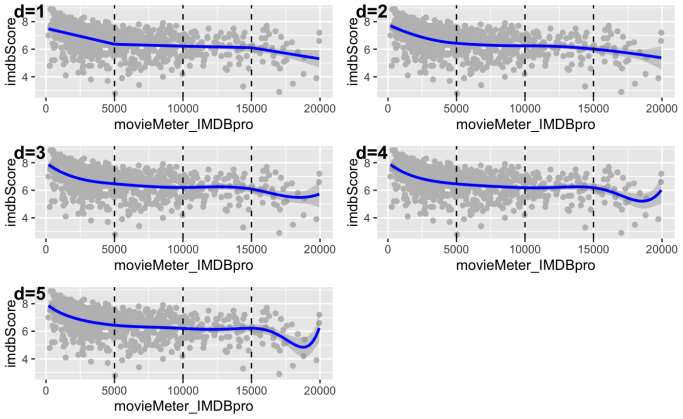


Figure B4 - Spline model for variable movieMeter_IMDBpro with degree from 1 to 5