Group 23

Ronit Bhatt, Mateo Bianchi, Francis Chan, Sreedeekshita Gorugantu Venkata, Gabriel Pham, Khang Thai

1 Introduction
1.1 Research Topic of Interest

This research delves into the intricate dynamics of movie success by analyzing a dataset comprising 1000 IMDb movies across 16 variables. Focusing on IMDb rating as the primary indicator of a movie's reception, this study investigates the relationships between IMDb rating and critical factors including year of release, Metascore, number of votes, gross revenue, and run time. Through the construction of a predictive model, the research aims to unveil the underlying mechanisms shaping a movie's overall success and reception in the film industry. By scrutinizing IMDb ratings alongside these predictor variables, valuable insights are gleaned into the multifaceted determinants of a movie's popularity, offering nuanced perspectives for filmmakers and industry stakeholders.

1.2 Background and Source of Data Set

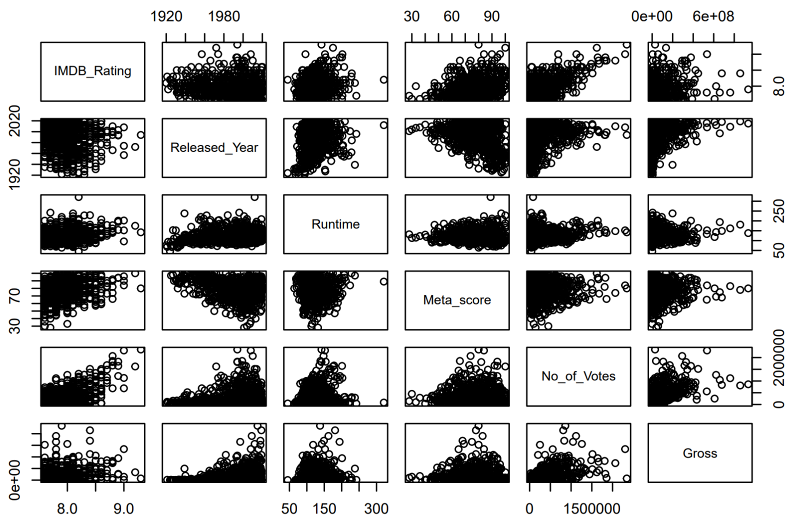
This dataset found from Kaggle contains a total of 1000 observations and 16 variables. After filtering through the variables, we narrowed down our selection to include 5 variables as our final predictor variables. Table below summarizes the response and predictor variables used in this analysis. We chose these predictors since they are the variables that most affect the IMDb ratings of movies.

Variable	Name	Description
Response Variable : Y	IMDB_Rating	Based on scale 1-10, 1 is the worst, 10 is the best
Predictor Variable 1: x_1	Released_Year	Year of the movie is released
Predictor Variable 2: x_2	Meta_Score	Rating of a film. Based on scale 1-100
Predictor Variable 3: x_3	No_of_Votes	Number of registered IMDB members who casted the votes
Predictor Variable 4: x_4	Gross	Gross earning (U.S. Dollars)
Predictor Variable 5: x_5	Runtime	Duration of the movie

Overview- We decided to run the full model with all the predictors, then looked at the transformed model. After comparing, we used the model that resulted in better fit, R-square etc.

2 Scatter plot - Full model - ANOVA

2.1 Scatter Plot Matrix



We can observe that there is an almost positive linear relation between the predictors and the response variable, except for possibly, Runtime and Released year. The relation between the different predictors is not that significant as can be seen from the plots and the correlation coefficients (in the appendix- more about the summary statistics of the variables also).

2.2 Full model (summary and ANOVA)

```
summary(Am1)
```

```
##
## Call:
## lm(formula = IMDB_Rating ~ Released_Year + Runtime + Meta_score +
##     No_of_Votes + Gross, data = imdb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43152 -0.13241 -0.02217  0.11877  0.68715
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.350e+01  7.879e-01  17.130  < 2e-16 ***
## Released_Year -3.150e-03  3.869e-04  -8.141 1.65e-15 ***
## Runtime        1.558e-03  2.737e-04   5.694 1.79e-08 ***
## Meta_score     4.699e-03  5.844e-04   8.040 3.53e-15 ***
## No_of_Votes    6.375e-07  2.417e-08  26.371  < 2e-16 ***
## Gross         -6.862e-10  7.460e-11  -9.198  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1896 on 743 degrees of freedom
##   (251 observations deleted due to missingness)
## Multiple R-squared:  0.5727, Adjusted R-squared:  0.5698
## F-statistic: 199.2 on 5 and 743 DF,  p-value: < 2.2e-16
```

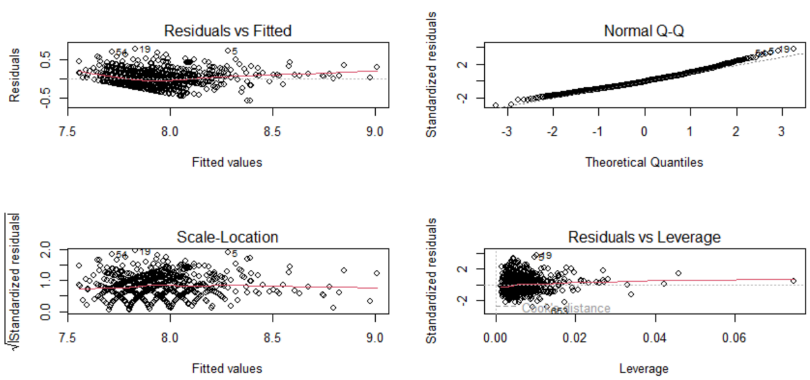
We have an R- square value of 57.27% (56.98% adjusted), which is quite reasonable for a real-life data set.

```r
anova(Am1)
```

```
## Analysis of Variance Table
##
## Response: IMDB_Rating
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## Released_Year   1  2.0037  2.0037  55.757 2.303e-13 ***
## Runtime         1  4.2096  4.2096 117.141 < 2.2e-16 ***
## Meta_score      1  3.6532  3.6532 101.658 < 2.2e-16 ***
## No_of_Votes     1 22.8825 22.8825 636.756 < 2.2e-16 ***
## Gross           1  3.0401  3.0401  84.597 < 2.2e-16 ***
## Residuals     743 26.7005  0.0359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we observe that all predictors are significant.

2.3 Diagnostic plots



- We observe that the residuals are scattered randomly around 0, which suggests randomness and independence of the error term.
- In the standardized residual plot, some could argue that there is a pattern, but it is mostly random. All residuals lie between -2 and 2, which suggests no outliers.
- The normal QQ plot is linear (straight line) which suggests that the error term has a normal distribution.
- There are leverage points but none of them are bad leverage points (see Appendix).

3 Transformation
3.1 Box-Cox Transformation

Box-Cox LR test for log transformation suggests not to use log transformation or no transformation (as seen in the output).
Instead, we use rounded power transformation.

```r
bc <- powerTransform(cbind(IMDB_Rating, Meta_score, No_of_Votes, Released_Year,Gross, Runtime) ~ 1)
summary(bc)
```

```
## bcPower Transformations to Multinormality
##               Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## IMDB_Rating     -7.6176       -7.62     -10.4260      -4.8092
## Meta_score       2.2677        2.00       1.8986       2.6368
## No_of_Votes      0.1914        0.19       0.1309       0.2520
## Released_Year   47.3295       47.33      39.4969      55.1621
## Gross            0.1920        0.19       0.1651       0.2189
## Runtime         -0.6301       -0.50      -0.9235      -0.3367
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                   LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 747.2558  6 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                   LRT df       pval
## LR test, lambda = (1 1 1 1 1 1) 2875.691  6 < 2.22e-16
```

3.2 Transformed model

```
## Call:
## lm(formula = t_IMDB_Rating ~ t_Released_Year + t_Runtime + t_Gross +
##     t_Meta_score + t_No_of_Votes)
```

```
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -7.812e-08 -1.783e-08 -1.442e-09  1.682e-08  6.201e-08
##
## Coefficients:
##                   Estimate  Std. Error t value Pr(>|t|)
## (Intercept)      1.758e-07  1.205e-08  14.591  < 2e-16 ***
## t_Released_Year 1.173e-164  0.000e+00     Inf  < 2e-16 ***
## t_Runtime        7.451e-07  1.046e-07   7.123  2.5e-12 ***
## t_Gross          1.601e-09  1.340e-10  11.949  < 2e-16 ***
## t_Meta_score    -4.612e-12  5.159e-13  -8.940  < 2e-16 ***
## t_No_of_Votes   -1.262e-08  6.026e-10 -20.937  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.526e-08 on 743 degrees of freedom
## Multiple R-squared:  0.4778, Adjusted R-squared:  0.4743
## F-statistic:   136 on 5 and 743 DF,  p-value: < 2.2e-16
```

The R-square value for the transformed model is 47.78% and adjusted is 47.43%. So we may prefer to go with the original model instead.

(Diagnostic plots and anova for the transformed model in the Appendix)

3.3 Variable Selection

This is the Variance Inflation Factor (VIF) for the model, none of the values are found to be greater than 5, so there is no multicollinearity between the predictors.

```
vif(Am1)
```

```
## Released_Year       Runtime     Meta_score    No_of_Votes         Gross
##      1.185761      1.056545      1.111177       1.500300      1.488399
```

Backward AIC way: (more ways in the Appendix)

```
Start:  AIC=-2485.21
IMDB_Rating ~ Released_Year + Meta_score + No_of_Votes + Gross +
    Runtime

                Df Sum of Sq    RSS     AIC
<none>                       26.700 -2485.2
- Runtime        1    1.1649 27.865 -2455.2
- Meta_score     1    2.3232 29.024 -2424.7
- Released_Year  1    2.3819 29.082 -2423.2
- Gross          1    3.0401 29.741 -2406.4
- No_of_Votes    1   24.9901 51.691 -1992.4
```

Variable selection (AIC) suggests that the model with all the predictors is the best model.

4 Final Model

Y = 13.5 - (3.153e-03)Release_Year + (1.558e-03)Runtime + (4.6993e-03)Meta_score + (6.375e-07)No_of_Votes - (6.862e-10)Gross

4.1 Slope Interpretation

All of the p-values are less than 0.05, hence we reject the null hypothesis, conclude that coefficients are significant.

- For a 1-percent increase in __+__ we expect that IMDB ratings have:

+ **Released Year**: A decrease of approximately **0.00315** percent
+ **Runtime**: An increase of approximately **0.001558** percent
+ **Meta-score**: An increase of approximately **0.004699** percent
+ **Number of Votes**: An increase of approximately **6.375e-07** percent
+ **Gross**: A decrease of approximately **-6.862e-10** percent

Evaluate related to reality:

1) Released year: Older movies tend to have slightly lower IMDB ratings, possibly due to changing audience tastes over time.

2) Runtime: Longer movies tend to have higher IMDB ratings; possibly because they offer more depth and engagement for viewers.

3) Meta-Score: Movies with higher Metacritic scores usually have higher IMDB ratings, reflecting critical acclaim.

4) Number of votes: Popular movies with more votes tend to have higher IMDB ratings, which means broader audience appeal.

5) Gross: Surprisingly, higher-grossing movies tend to have slightly lower IMDB ratings. This possibly because commercial success doesn’t always equate to critical or audience acclaim.

5 Discussion

In this study, we examined a dataset comprising 1000 IMDb movies, exploring the intricate interplay between IMDb ratings and an array of crucial variables such as year of release, meta score, number of votes, gross revenue, and run time. The objective was to construct a predictive model that unravels the factors contributing to a movie's success and reception in the film industry.

Our final model does make sense in real-world situations as there are many challenges, factors, and shifting variables that all have to align for a movie to be a hit. Our predictive model captured a few patterns in the dataset, offering some insights into the factors shaping IMDb ratings. However, it's important to note that predicting success remains challenging, as the film industry is inherently unpredictable as there are many factors that may affect its success such as competition and release timings, advertising, economic factors, etc. These factors make a huge difference and are incredibly challenging to calculate and account for, as they are shifting variables for any movie.

In an article, “Study explores what really makes a movie successful” by University of Technology, Sydney they also state some important variables for a movie to be successful. They state the following, “Star power, acting expertise, rousing reviews and public ratings are all key factors that influence our decision to see a movie. Researchers from UTS, HEC Montreal and the University of Cambridge compared these factors across 150 studies to boil down the formula for box office success.” There are so many variables and it is evident even from our findings that movie ratings and success are very difficult to predict as there are variables that can alter or change the odds of any predictive model.

The only limitation of the analysis would be to elongate and broaden the data so that way there can be a more accurate predictive model. There would need to be more data collected for factors outside of cinema such as economic factors, actors ratings, release timings, production quality, directors, marketing, etc.

APPENDIX

Correlation coefficients between the predictors-

```
Released_Year     Runtime  IMDB_Rating          
Released_Year             1           NA          NA
Runtime                  NA   1.00000000   0.2470946
IMDB_Rating              NA   0.24709456   1.0000000
Meta_score               NA  -0.03145197   0.2685308
No_of_Votes              NA   0.21639063   0.5866643
Gross                    NA           NA          NA
               Meta_score No_of_Votes Gross
Released_Year          NA          NA    NA
Runtime       -0.03145197  0.21639063    NA
IMDB_Rating    0.26853084  0.58666434    NA
Meta_score     1.00000000 -0.01850697    NA
No_of_Votes   -0.01850697  1.00000000    NA
Gross                  NA          NA     1
```

More summary statistics-

```
mean(IMDB_Rating)
[1] 7.935247
> mean(Released_Year)
[1] 1995.071
> mean(Runtime)
[1] 123.2804
> mean(Meta_score)
[1] 77.46061
> mean(No_of_Votes)
[1] 342230
> mean(Gross)
[1] 195.9346
> sd(IMDB_Rating)
[1] 0.2890365
> sd(Released_Year)
[1] 19.50906
> sd(Runtime)
[1] 26.03096
> sd(Meta_score)
[1] 12.5023
> sd(No_of_Votes)
[1] 351203.9
> sd(Gross)
[1] 232.9353
```

No bad leverage points for the original model-

- hvalues <- hatvalues(Am1)
- stanresDeviance <- residuals(Am1) / sqrt(1 - hvalues)
- which(hvalues > 2*5 / length(IMDB_Rating))
- which(hvalues > 2*5 /length(IMDB_Rating) & abs(stanresDeviance) > 2)

```
named integer(0)
```

AIC and BIC of Am1-

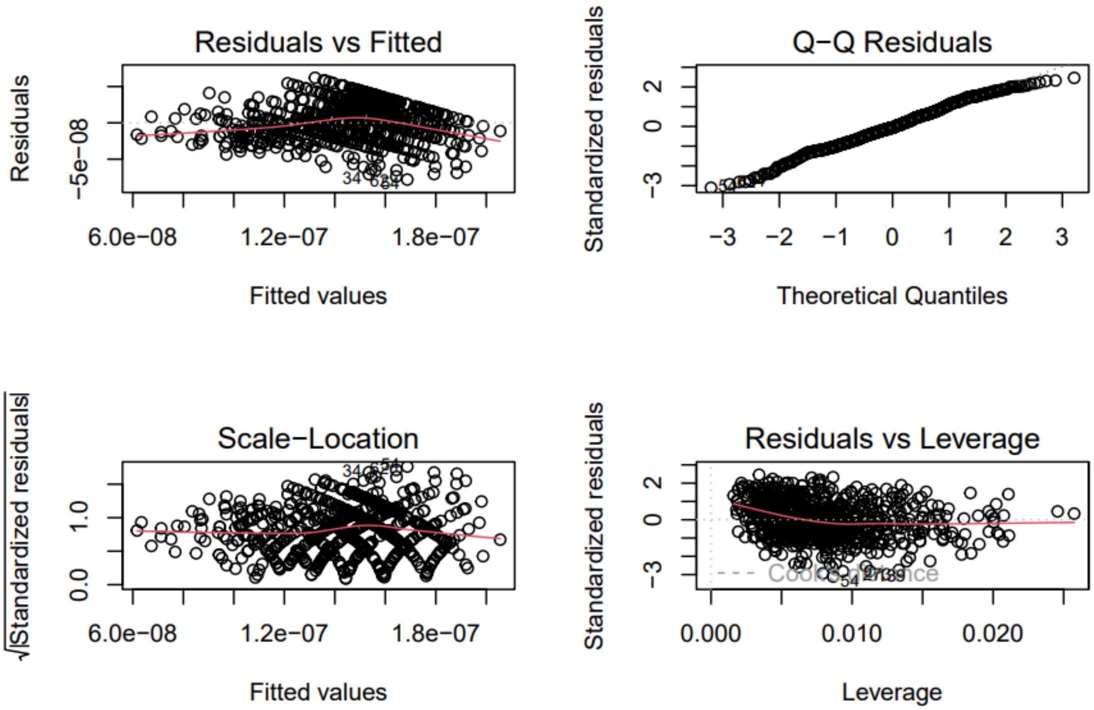
```
n <- length(Am1$residuals)
backAIC <- step(Am1, direction = "backward", data = imdb)
```

```
## Start:  AIC=-2485.21
## IMDB_Rating ~ Released_Year + Runtime + Meta_score + No_of_Votes +
##     Gross
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       26.700 -2485.2
## - Runtime        1    1.1649 27.865 -2455.2
## - Meta_score     1    2.3232 29.024 -2424.7
## - Released_Year  1    2.3819 29.082 -2423.2
## - Gross          1    3.0401 29.741 -2406.4
## - No_of_Votes    1   24.9901 51.691 -1992.4
```

```
backBIC <- step(Am1, direction = "backward", k = log(n), data = imdb)
```

```
## Start:  AIC=-2457.5
## IMDB_Rating ~ Released_Year + Runtime + Meta_score + No_of_Votes +
##     Gross
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       26.700 -2457.5
## - Runtime        1    1.1649 27.865 -2432.1
## - Meta_score     1    2.3232 29.024 -2401.6
## - Released_Year  1    2.3819 29.082 -2400.1
## - Gross          1    3.0401 29.741 -2383.3
## - No_of_Votes    1   24.9901 51.691 -1969.3
```

Diagnostic plots for the transformed model-



There is no significant improvement in the diagnostic plots from the full model.

ANOVA for the transformed-

```
anova(Am2)
```

```
## Analysis of Variance Table
##
## Response: t_IMDB_Rating
##                  Df     Sum Sq    Mean Sq  F value    Pr(>F)
## t_Released_Year   1 2.7140e-14 2.7137e-14  42.5367 1.282e-10 ***
## t_Runtime         1 5.3460e-14 5.3465e-14  83.8045 < 2.2e-16 ***
## t_Gross           1 1.4100e-15 1.4090e-15   2.2082    0.1377
## t_Meta_score      1 7.2080e-14 7.2078e-14 112.9800 < 2.2e-16 ***
## t_No_of_Votes     1 2.7967e-13 2.7967e-13 438.3742 < 2.2e-16 ***
## Residuals       743 4.7401e-13 6.3800e-16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Gross is not significant here.