Group 23

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

Research Topic

In this study, an analysis of 1000 IMDb movies (16 variables) was conducted to investigate the relationship between IMDb rating, year of release, meta score, number of votes, gross and run time. A predictive model was constructed to understand how these variables influence the overall success and reception of movies. By examining IMDb ratings alongside with predictors variables, insights were gained into the factors that contribute to a movie's popularity in the film industry. (Kaggle)

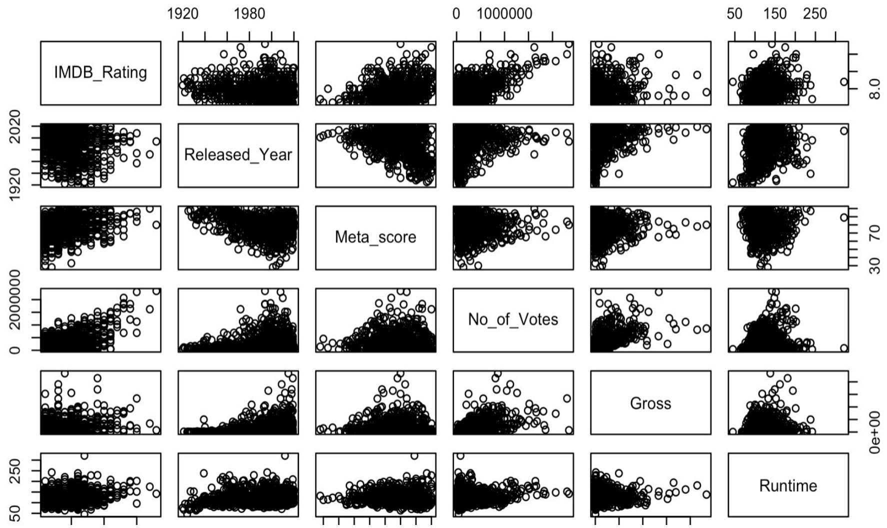
Dataset

Observations: 1000   Predictor variables: 16

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 a 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 in U.S. dollars
Predictor Variable 5: X_5	Runtime	Duration of the movie

Scatter Plot Matrix

```{r}
library(dplyr)
library(corrplot)
library(stringr)
imdb <- read.csv("imdb_top_1000.csv")
attach(imdb)
head(imdb)
imdb$Released_Year <- as.numeric(as.character(imdb$Released_Year)) # convert Released_Year into numeric variable
imdb$Runtime <- as.numeric(str_extract(imdb$Runtime, "[0-9]{2,3}")) # convert Runtime into numeric variable
numeric_imdb <- select_if(imdb, is.numeric) # Original dataset
attach(numeric_imdb)
pairs(IMDB_Rating ~ Released_Year + Runtime + Meta_score + No_of_Votes + Gross)
# This is the scatter plot matrix
```



Full model + Anova

```{r}
Am1 <- lm(IMDB_Rating ~ Released_Year + Runtime +
           Meta_score + No_of_Votes + Gross, data = imdb)
plot(Am1)
summary(Am1)
```

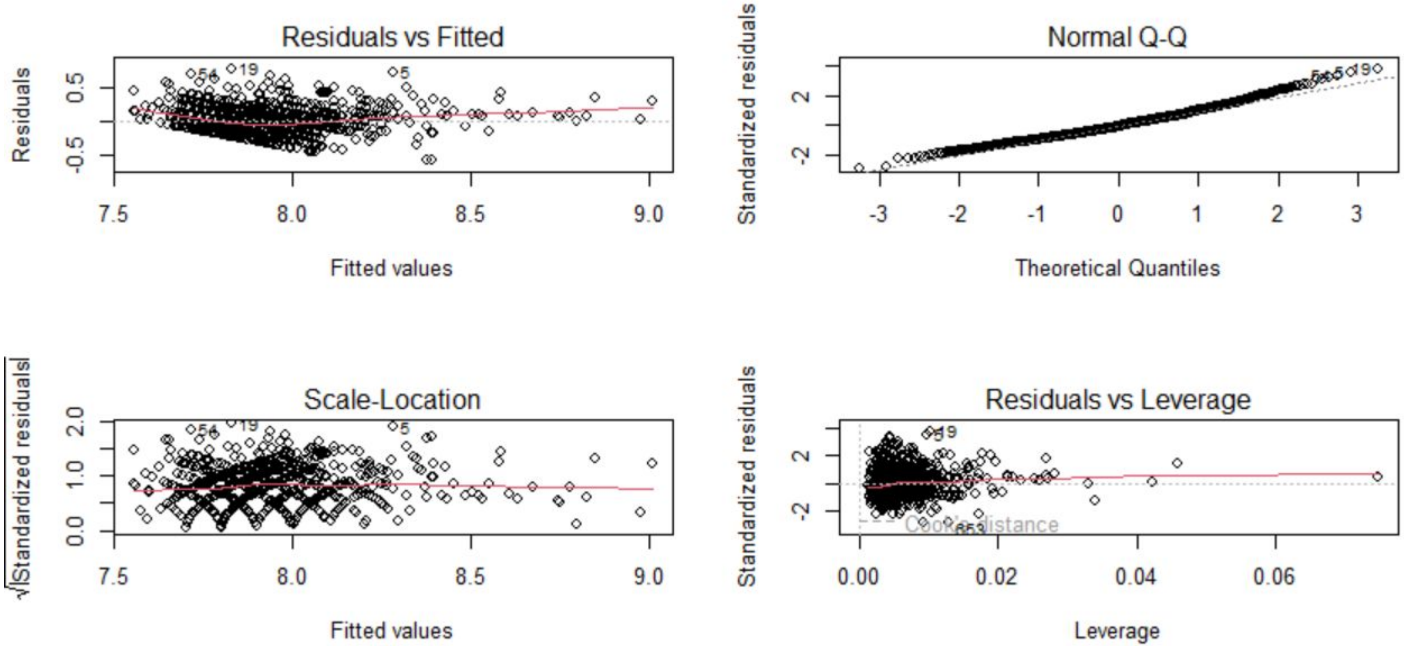
```
Call:
lm(formula = IMDB_Rating ~ Released_Year + Runtime + Meta_score +
    No_of_Votes + Gross, data = imdb)

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
```

```
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
```

Diagnostic plots

❖ **plot(Am1)**



Model assumptions and Interpretation

- We can see 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 (trying transformation). All residuals between -2 and 2, which suggests no outliers.
- The normal QQ plot is linear (straight line) which suggest the the error term has a normal distribution
- There are leverage points but none of them are bad leverage points.

❖ 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)
```

Box-Cox Transformation

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

```

```
bcPower Transformations to Multinormality
              Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
IMDB_Rating     -7.6005       -7.60      -8.7955      -6.4055
Meta_score       2.2647        2.00       1.8957       2.6336
No_of_Votes      0.1914        0.19       0.1327       0.2501
Released_Year   47.1213       47.12      39.2984      54.9443
Gross            0.1919        0.19       0.1650       0.2188
Runtime         -0.6260       -0.50      -0.9193      -0.3327

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

	LRT <dbl>	df <int>	pval <chr>
LR test, lambda = (0 0 0 0 0 0)	747.2549	6	< 2.22e-16

	LRT <dbl>	df <int>	pval <chr>
LR test, lambda = (1 1 1 1 1 1)	2875.69	6	< 2.22e-16

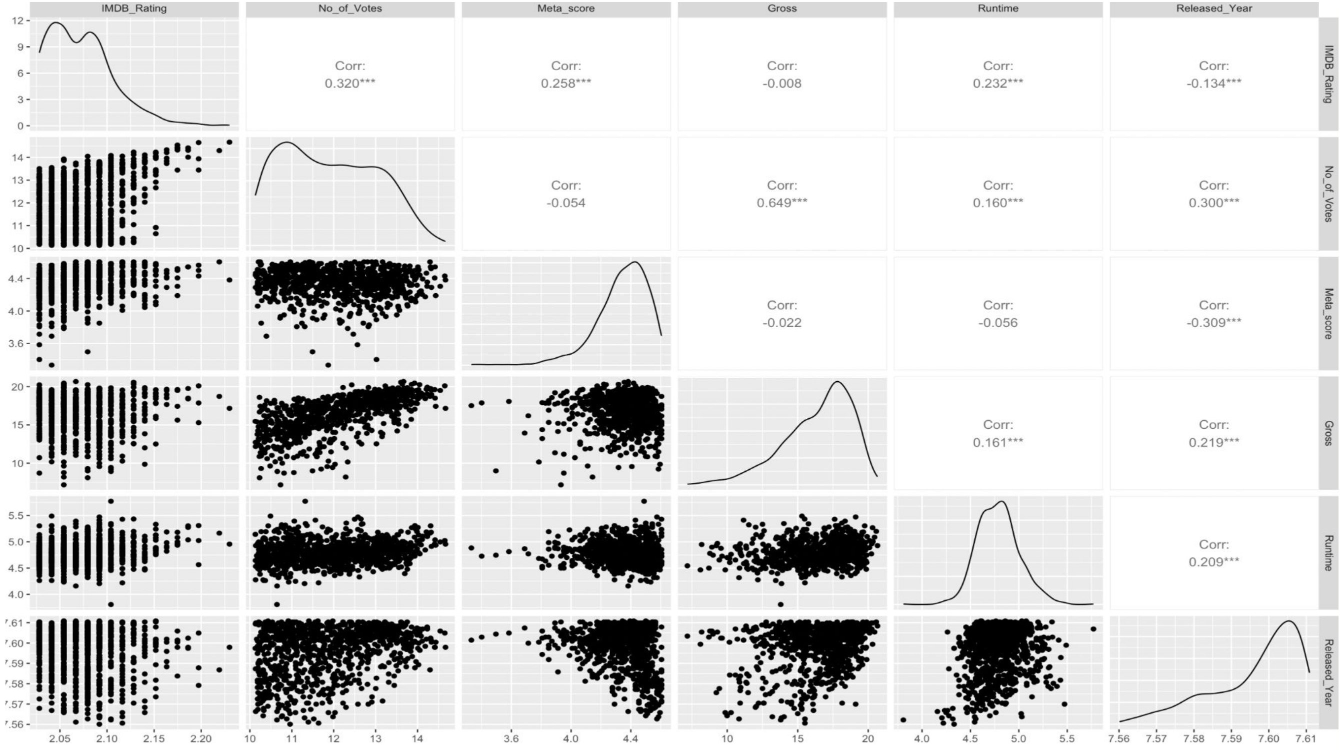
Am2 <- lm(log(IMDB_Rating) ~ log(Released_Year) + log(Meta_score) + log(No_of_Votes) + log(Runtime), data = imdb)

summary(Am2)

R2 is 0.3475 for the transformed model.

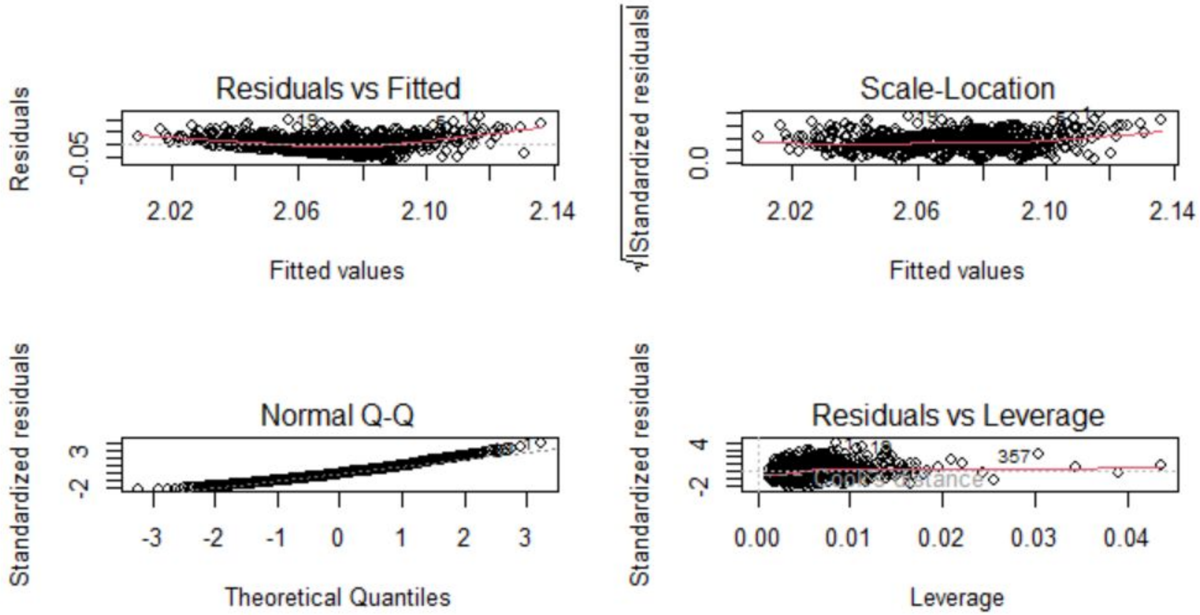
```
Call:
lm(formula = log(IMDB_Rating) ~ log(Released_Year) + log(Meta_score) +
    log(No_of_Votes) + log(Runtime), data = imdb)

Residuals:
     Min       1Q   Median       3Q
-0.06223 -0.01970 -0.00239  0.01688
     Max
 0.11310

Coefficients:
                    Estimate Std. Error
(Intercept)        7.6777609  0.7593688
log(Released_Year) -0.8073907  0.0994698
log(Meta_score)    0.0435662  0.0059014
log(No_of_Votes)   0.0156789  0.0009625
log(Runtime)       0.0308243  0.0048244
                   t value Pr(>|t|)
(Intercept)         10.111  < 2e-16 ***
log(Released_Year)  -8.117 1.70e-15 ***
log(Meta_score)      7.382 3.76e-13 ***
log(No_of_Votes)    16.290  < 2e-16 ***
log(Runtime)         6.389 2.76e-10 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1

Residual standard error: 0.0284 on 837 degrees of freedom
  (158 observations deleted due to missingness)
Multiple R-squared:  0.3475,    Adjusted R-squared:  0.3444
F-statistic: 111.4 on 4 and 837 DF,  p-value: < 2.2e-16
```

```r
imdb$Released_Year <- log(imdb$Released_Year)
imdb$Runtime <- log(imdb$Runtime)
imdb$Gross <- log(imdb$Gross)
imdb$Meta_score <- log(imdb$Meta_score)
imdb$IMDB_Rating <- log(imdb$IMDB_Rating)
imdb$No_of_Votes <- log(imdb$No_of_Votes)

ggpairs(imdb[, c("IMDB_Rating", "No_of_Votes", "Meta_score", "Gross", "Runtime", "Released_Year")])
```



Diagnostic plots for the transformed model-

The standardized residual plot may show more random distribution, but since the R2 decreased, we revert back to the original model.



Variable Selection

```
backAIC <- step(Am1, direction = "backward", data = imdb)

Start:  AIC=-2673.12

IMDB_Rating ~ Released_Year + Meta_score + No_of_Votes + Runtime

                Df Sum of Sq    RSS     AIC

<none>                       34.785 -2673.1

- Runtime        1    1.3455 36.130 -2643.2

- Meta_score     1    2.4949 37.279 -2616.8

- Released_Year  1    3.1057 37.890 -2603.1

- No_of_Votes    1   23.1557 57.940 -2245.5
```

So, the best model is with all the predictors.

Final Model

$$Y = 13.5 - (3.153e\text{-}03)X_1 + (1.558e\text{-}03)X_2 + (4.6993e\text{-}03)X_3 + (6.375e\text{-}07)X_4 - (6.862e\text{-}10)X_5$$

```
{r}
Am1 <- lm(IMDB_Rating ~ Released_Year + Runtime +
            Meta_score + No_of_Votes + Gross, data = imdb)
plot(Am1)
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
```

Chunk 2

Interpret Slope

- All of the p-value 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.

Challenges? / Conclusion Analysis

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 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.

Thank You

Questions?