Final report

Back ground and final question

From the 1990s movie industry is booming. As a industrial product the profit of a movie is important.

So there are growing interest in analyze the profit of a movie based on the different factors can be scaled.

However as kind of an art we also want to know how these factors can affect the reputation of the movie which can be revealed by the IMDb score.

So as a training project I will use 5000 IMDB data set collected form kaggle to start the project.

Project environment

BUilding data sets

The data set is collected from kaggle

```
## 'data.frame':    5043 obs. of  28 variables:
##  $ color                    : Factor w/ 3 levels ""," Black and White",..: 3 3 3 3 1 3 3 3 3 3 ...
##  $ director_name            : Factor w/ 2399 levels "","A. Raven Cruz",..: 923 796 2023 375 602 101 2
##  $ num_critic_for_reviews   : int  723 302 602 813 NA 462 392 324 635 375 ...
##  $ duration                 : int  178 169 148 164 NA 132 156 100 141 153 ...
##  $ director_facebook_likes  : int  0 563 0 22000 131 475 0 15 0 282 ...
##  $ actor_3_facebook_likes   : int  855 1000 161 23000 NA 530 4000 284 19000 10000 ...
##  $ actor_2_name             : Factor w/ 3033 levels "","50 Cent","A. Michael Baldwin",..: 1407 2218 2
##  $ actor_1_facebook_likes   : int  1000 40000 11000 27000 131 640 24000 799 26000 25000 ...
##  $ gross                    : int  760505847 309404152 200074175 448130642 NA 73058679 336530303 2008
##  $ genres                   : Factor w/ 914 levels "Action","Action|Adventure",..: 107 101 128 288 75
##  $ actor_1_name             : Factor w/ 2098 levels "","50 Cent","A.J. Buckley",..: 301 978 351 1965
##  $ movie_title              : Factor w/ 4917 levels "#Horror ","[Rec] 2 ",..: 397 2729 3278 3706 3331
##  $ num_voted_users          : int  886204 471220 275868 1144337 8 212204 383056 294810 462669 321795
##  $ cast_total_facebook_likes: int  4834 48350 11700 106759 143 1873 46055 2036 92000 58753 ...
##  $ actor_3_name             : Factor w/ 3522 levels "","50 Cent","A.J. Buckley",..: 3439 1390 3131 17
##  $ facenumber_in_poster     : int  0 0 1 0 0 1 0 1 4 3 ...
##  $ plot_keywords            : Factor w/ 4761 levels "","10 year old|dog|florida|girl|supermarket",..
##  $ movie_imdb_link          : Factor w/ 4919 levels "http://www.imdb.com/title/tt0006864/?ref_=fn_tt_
##  $ num_user_for_reviews     : int  3054 1238 994 2701 NA 738 1902 387 1117 973 ...
##  $ language                 : Factor w/ 48 levels "","Aboriginal",..: 13 13 13 13 1 13 13 13 13 13 .
##  $ country                  : Factor w/ 66 levels "","Afghanistan",..: 65 65 63 65 1 65 65 65 65 63
##  $ content_rating           : Factor w/ 19 levels "","Approved",..: 10 10 10 10 1 10 10 9 10 9 ...
##  $ budget                   : num  2.37e+08 3.00e+08 2.45e+08 2.50e+08 NA ...
##  $ title_year               : int  2009 2007 2015 2012 NA 2012 2007 2010 2015 2009 ...
##  $ actor_2_facebook_likes   : int  936 5000 393 23000 12 632 11000 553 21000 11000 ...
##  $ imdb_score               : num  7.9 7.1 6.8 8.5 7.1 6.6 6.2 7.8 7.5 7.5 ...
##  $ aspect_ratio             : num  1.78 2.35 2.35 2.35 NA 2.35 2.35 1.85 2.35 2.35 ...
##  $ movie_facebook_likes     : int  33000 0 85000 164000 0 24000 0 29000 118000 10000 ...
```

Clean the data sets

First check if there are duplicate rows and remove them.

```
## [1] 45
```

The movie title is useless then we remove it from dataset

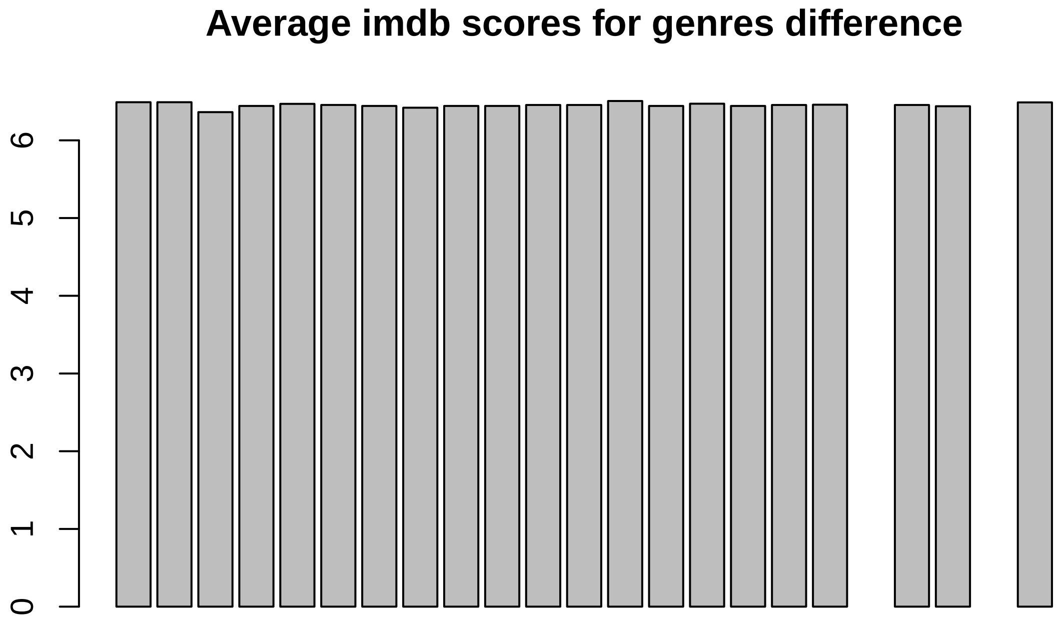
Then check if the genres affect our factors

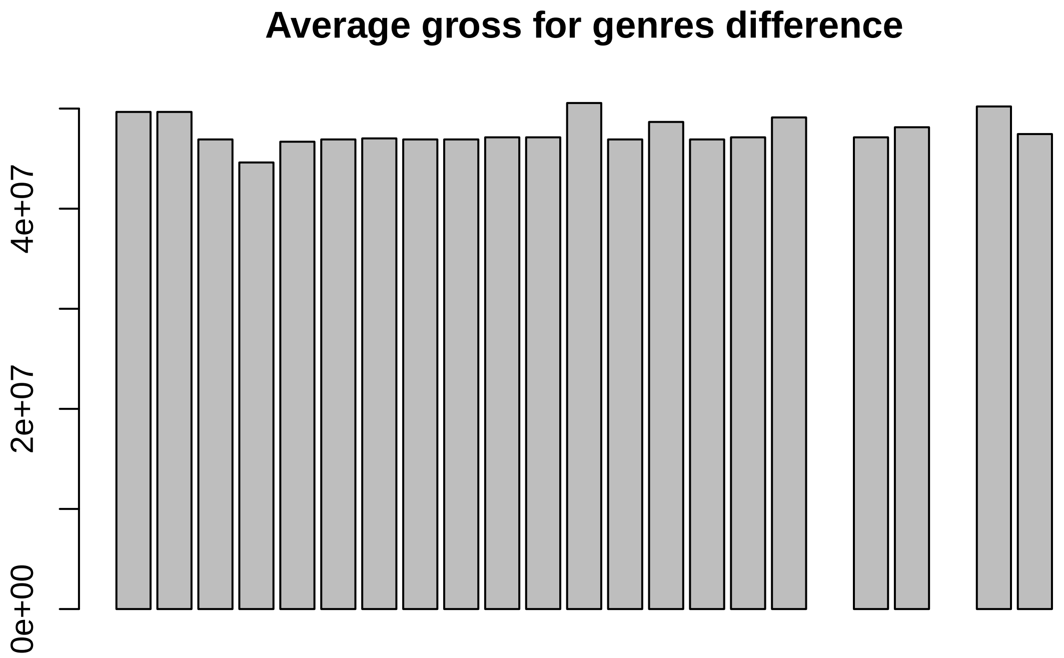
```
## [1] "sepreate the genres in genres_score"
```

```
## [1] "seperate genres in genres_gross"
```

```
## [1] "get the mean of factors for different genres"
```

```
## [1] "plot the means"
```





It’s clearly there is no significante effect. So take them away from data set

Obiviously the name is also useless

And swipe other useless cols

Now try to check is there any missing value in the dataset

```
##                       color    num_critic_for_reviews
##                           0                        49
##                    duration   director_facebook_likes
##                          15                       103
##      actor_3_facebook_likes    actor_1_facebook_likes
##                          23                         7
##                       gross          num_voted_users
##                         874                         0
## cast_total_facebook_likes       facenumber_in_poster
##                           0                        13
##        num_user_for_reviews                    budget
##                          21                       487
##                  title_year    actor_2_facebook_likes
##                         107                        13
##                  imdb_score              aspect_ratio
##                           0                       327
##        movie_facebook_likes
##                           0
```

Treat zero as missing value

Elimate all the rows have missing value

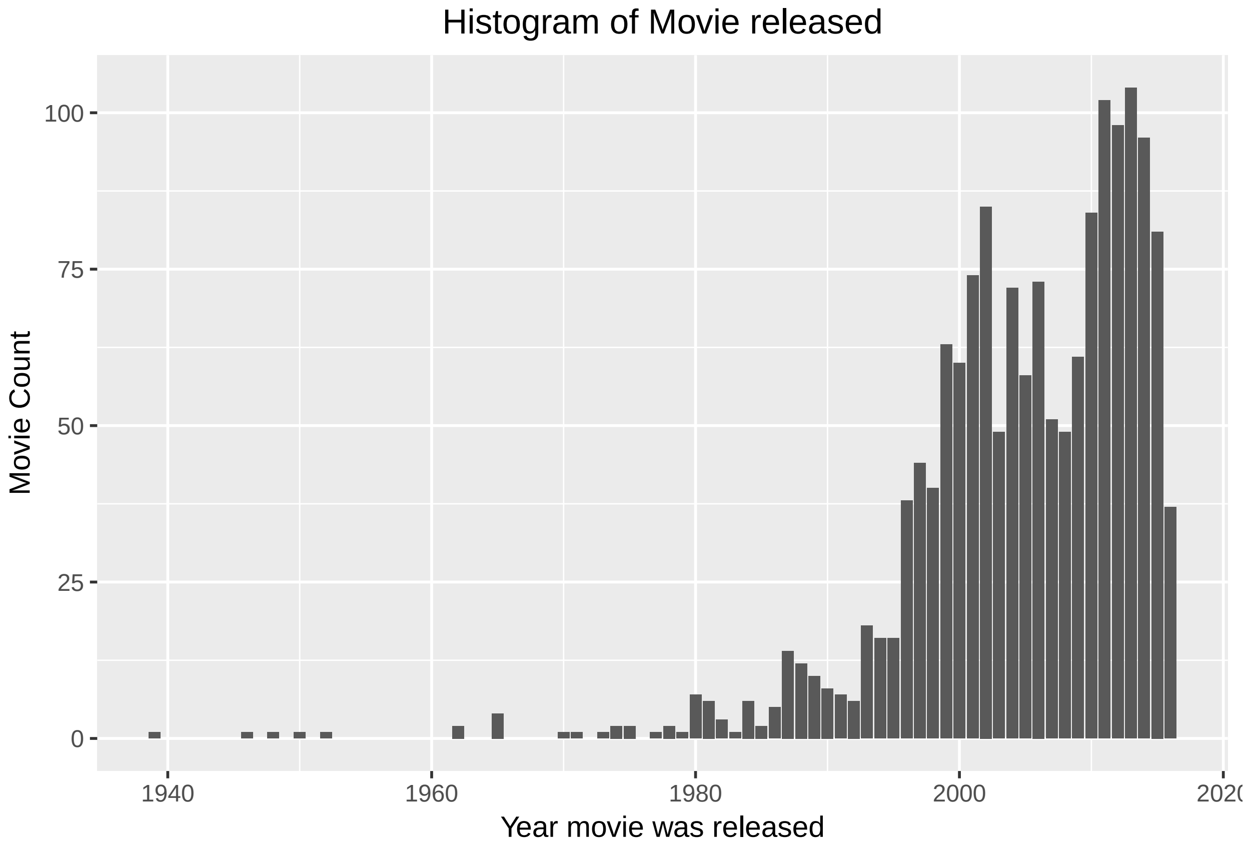
Now recheck the data set

```
##                     color    num_critic_for_reviews
##                         0                         0
##                  duration   director_facebook_likes
##                         0                         0
##    actor_3_facebook_likes    actor_1_facebook_likes
##                         0                         0
##                     gross           num_voted_users
##                         0                         0
## cast_total_facebook_likes      facenumber_in_poster
##                         0                         0
##      num_user_for_reviews                    budget
##                         0                         0
##                title_year    actor_2_facebook_likes
##                         0                         0
##                imdb_score              aspect_ratio
##                         0                         0
##      movie_facebook_likes
##                         0
```

We only take color movie into consideration

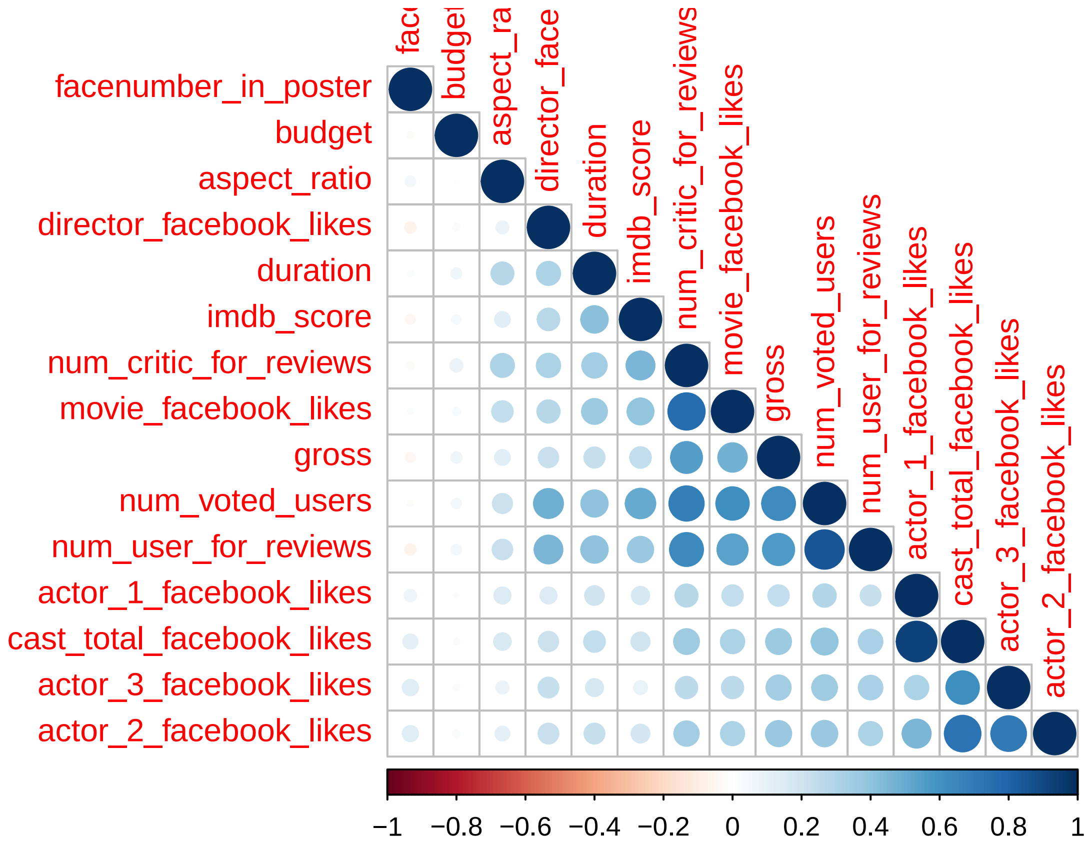
```
##
##                   Black and White            Color
##                0               47             1578
```

Estimate the title-year seperation



Most data is after 1980 SO eliminate the data before 1995

Check the correlation between variables



There is a positive correaltionship between num_critc_for_reviews,movie_facebook_likes/num_voted_users/num_users_fo

There is a positive correaltionship between actor_1_facebook_likes with cast_total_facebook_likes.

There is a positive correaltionship between actor_2_facebook_likes with actor_3_facebook_likes.

combine highly correalte variables.

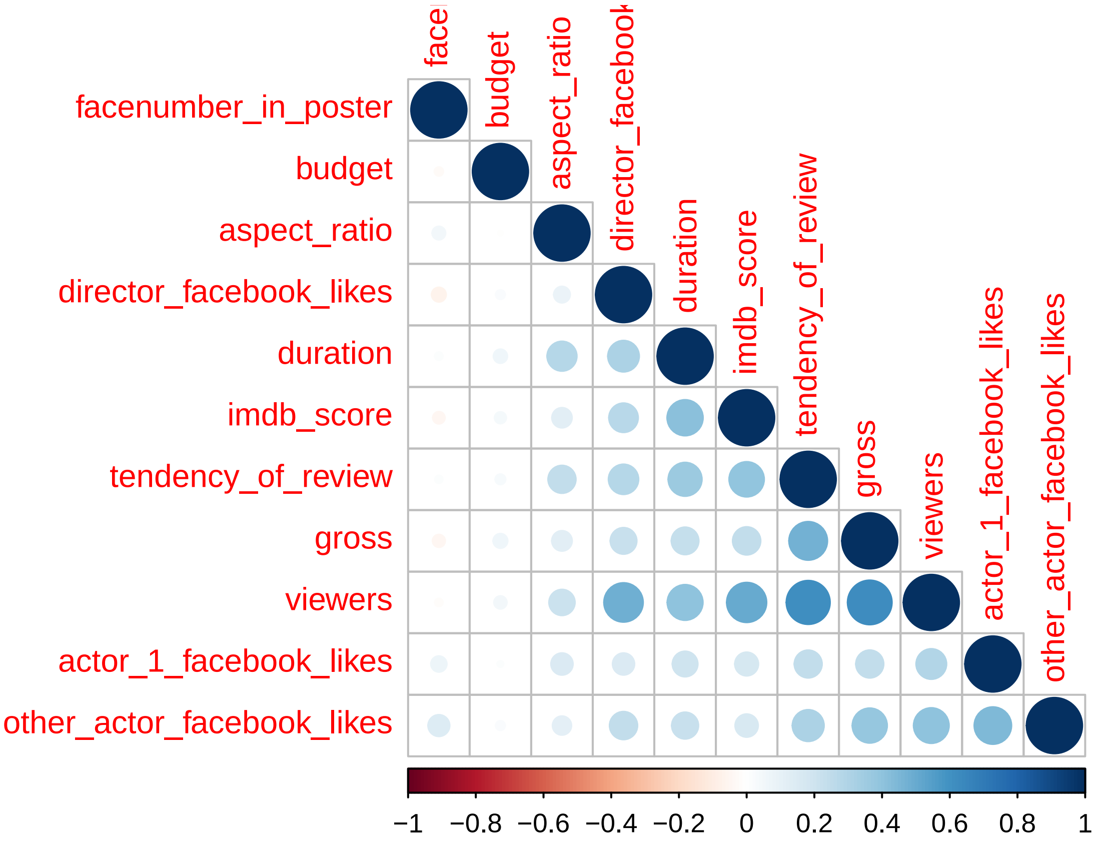
combine actor_2_facebook_likes with actor_3_facebook_likes as other_actor_facebook_likes

combine the num_voted_users/num_users_for_reviews as viewers

combine num_critic_for_reviews and movie_facebook_likes as tendency of review.

generate a new data set

check the correlation

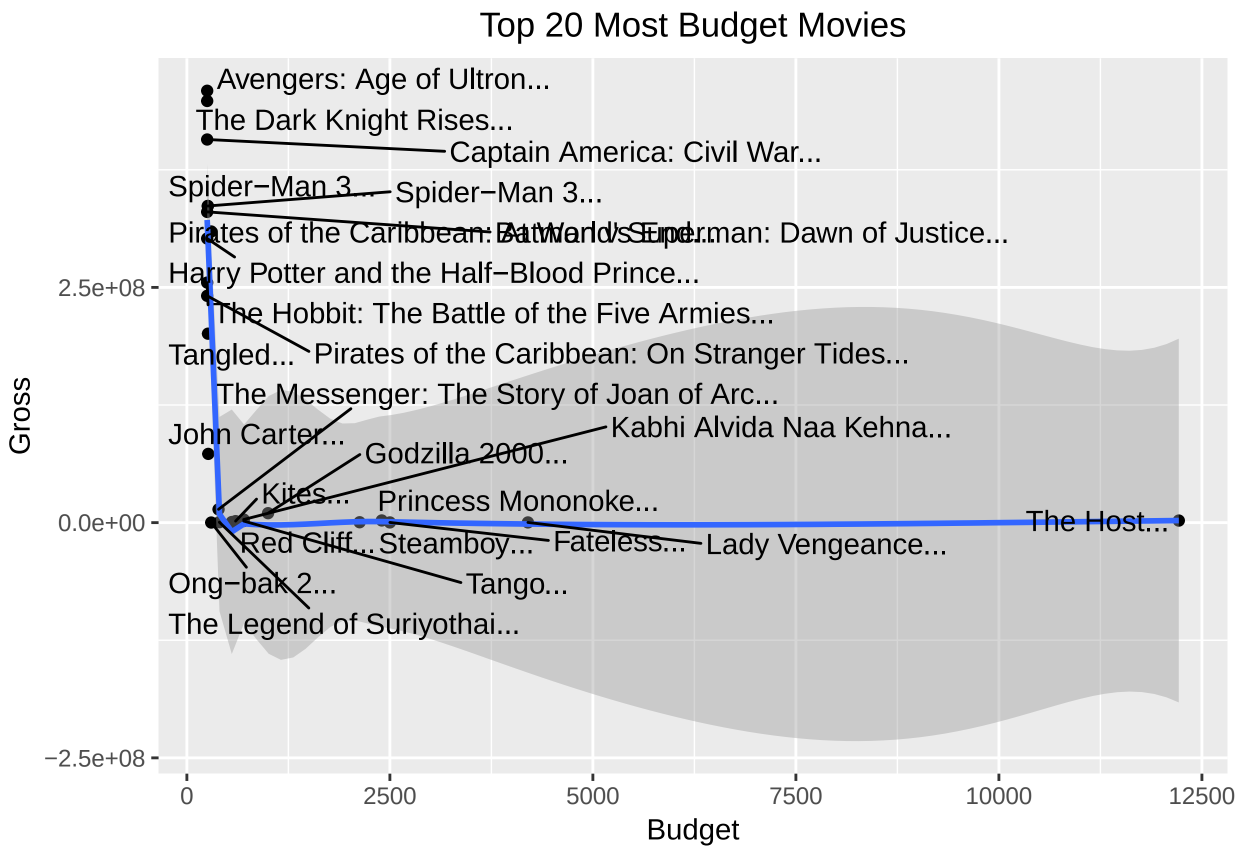


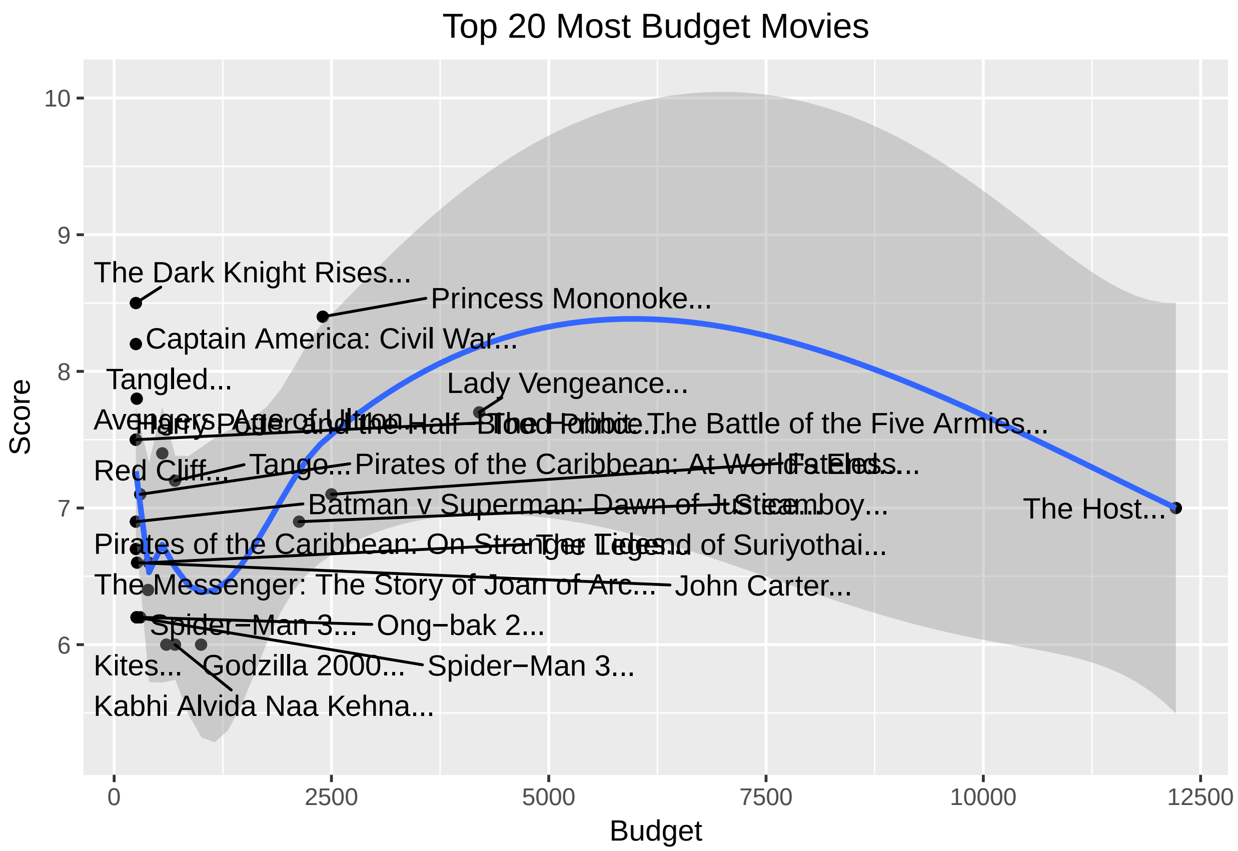
Based on the common sense we can ignore the viewers correlation.

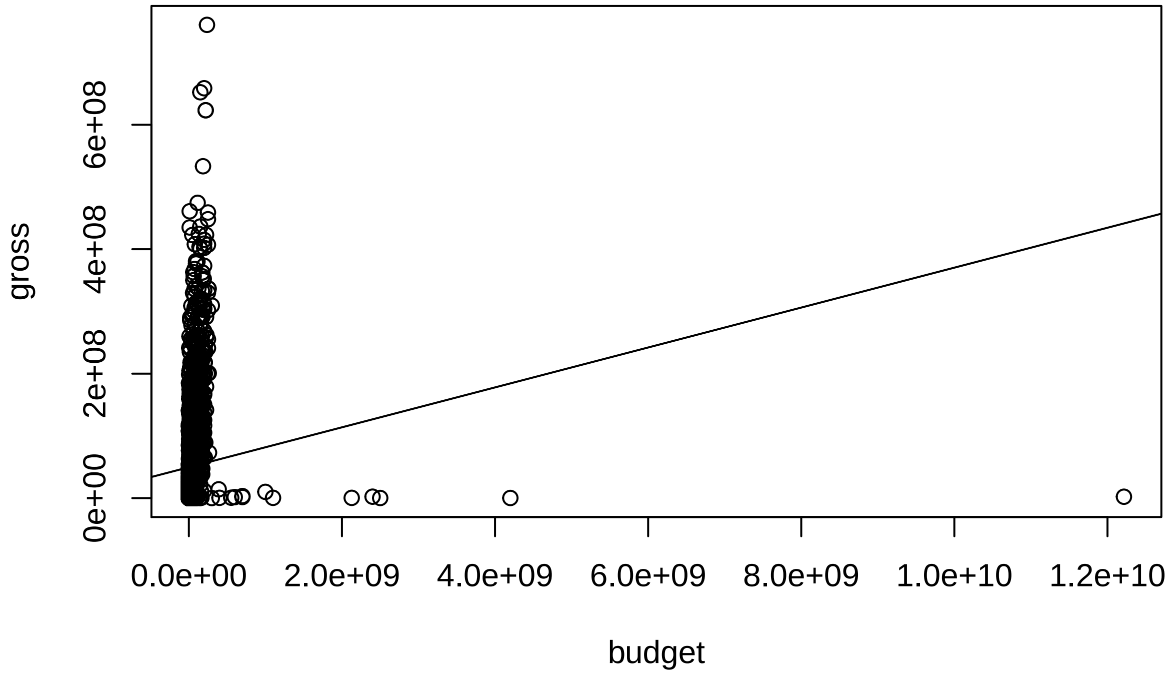
Basic stastical learning

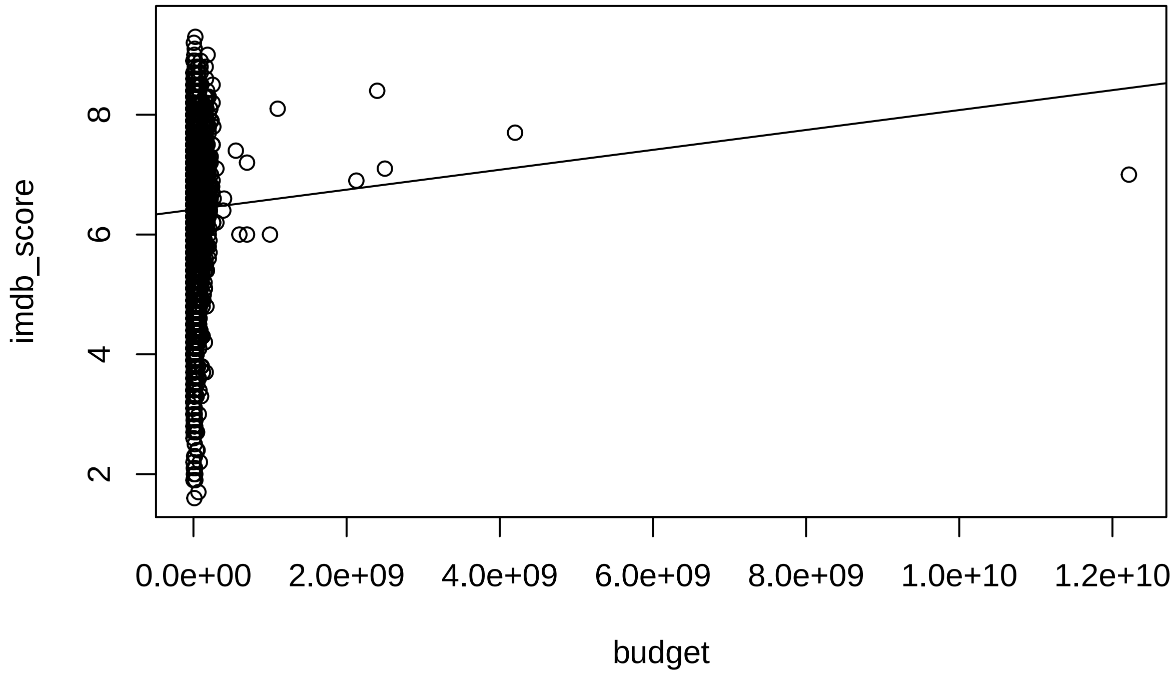
The main target is Gross and IMDB_score.Based on common scence we pick some parameters to do some simple linear test on the data set

Plot of most budget movies






gross
0e+00
2e+08
4e+08
6e+08
0.0e+00
2.0e+09
4.0e+09
6.0e+09
8.0e+09
1.0e+10
1.2e+10
budget



From the plot we can see there are some movie can have good score with not that high budget. However from the plot I think the movie have higher budget have a larger probability to get higher score. This can be conclude when you focus on the middle range of the plot.

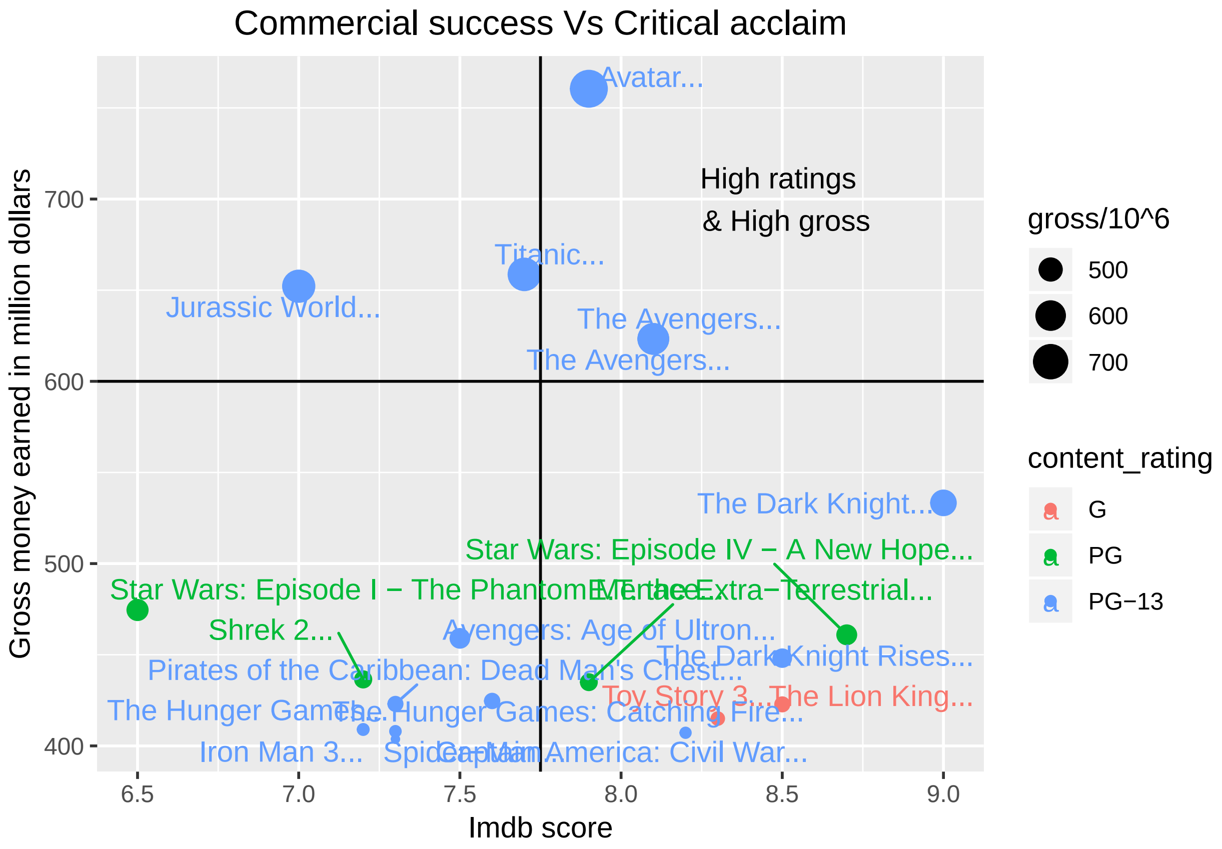
Form the plot I cant see a strong relation for budget to affect the gross. For some reason in those largest budget movies there is a strange trend that larger budget lead to bad commercial performance. If we take all the data into consideration. After fit a linear model we can get a model which is close to our assumption. However most movie have small budget so there are some high leverage outliers can affect the result. So the relation between those factors need more precise modeling technique.

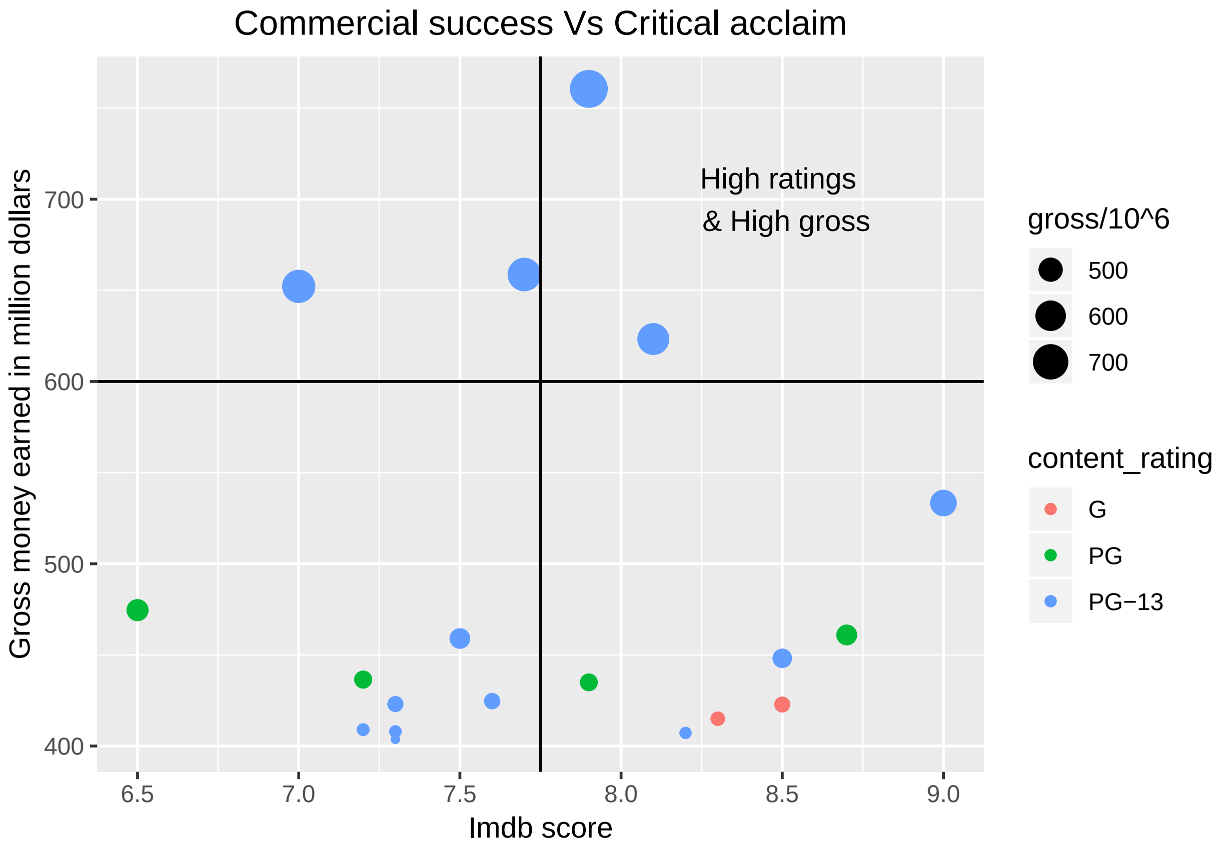
Check the outliers

```
##   duration director_facebook_likes actor_1_facebook_likes   gross
## 1      134                    6000                    893 2298191
## 2      103                      78                    488  410388
## 3      110                     584                    629 2201412
##   facenumber_in_poster      budget imdb_score aspect_ratio
## 1                    0  2400000000        8.4         1.85
## 2                    1  2127519898        6.9         1.85
## 3                    0 12215500000        7.0         1.85
##   other_actor_facebook_likes viewers tendency_of_review
## 1                       1596  222122              10826
## 2                        437   13806                868
## 3                        472   69162               6637
```

Based on the specific data here I think they are reasonable data. However still need to be erased.

Plot for gross vs score

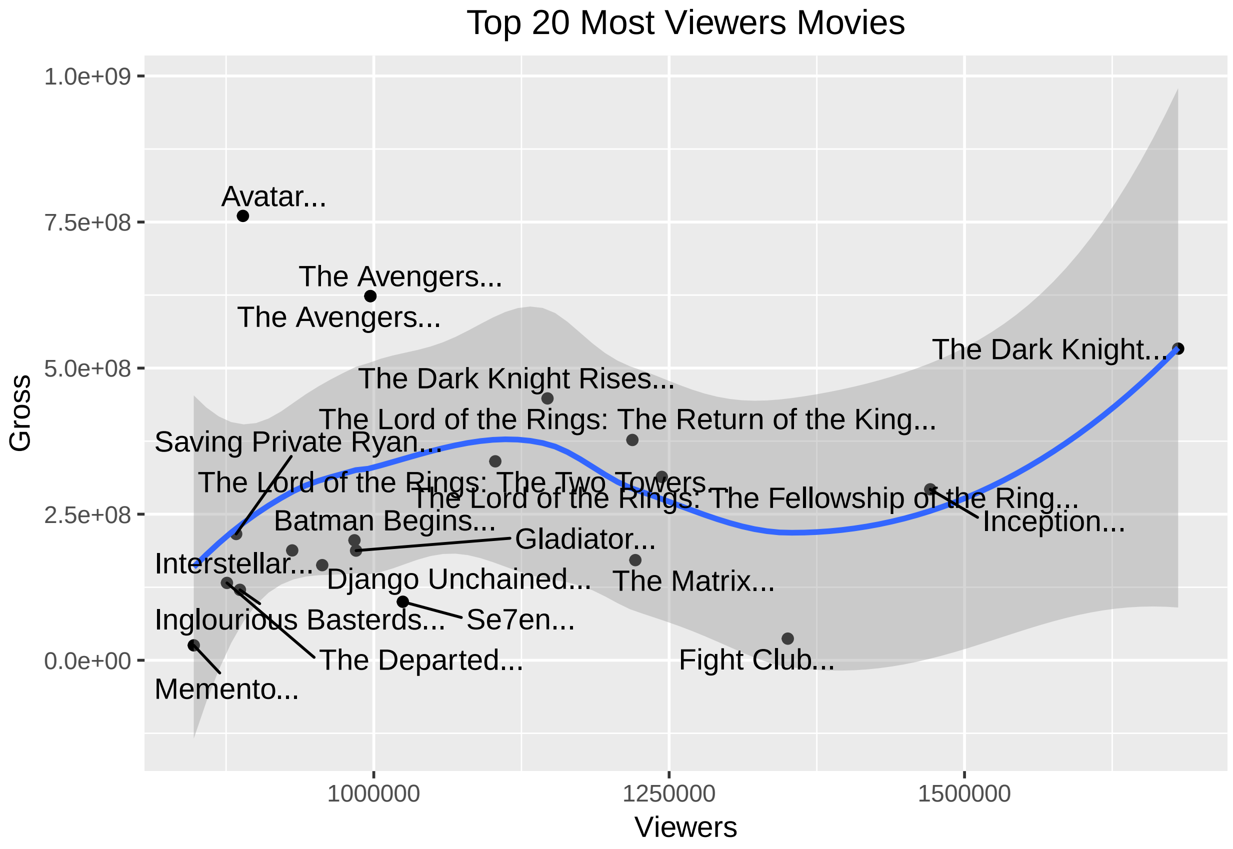


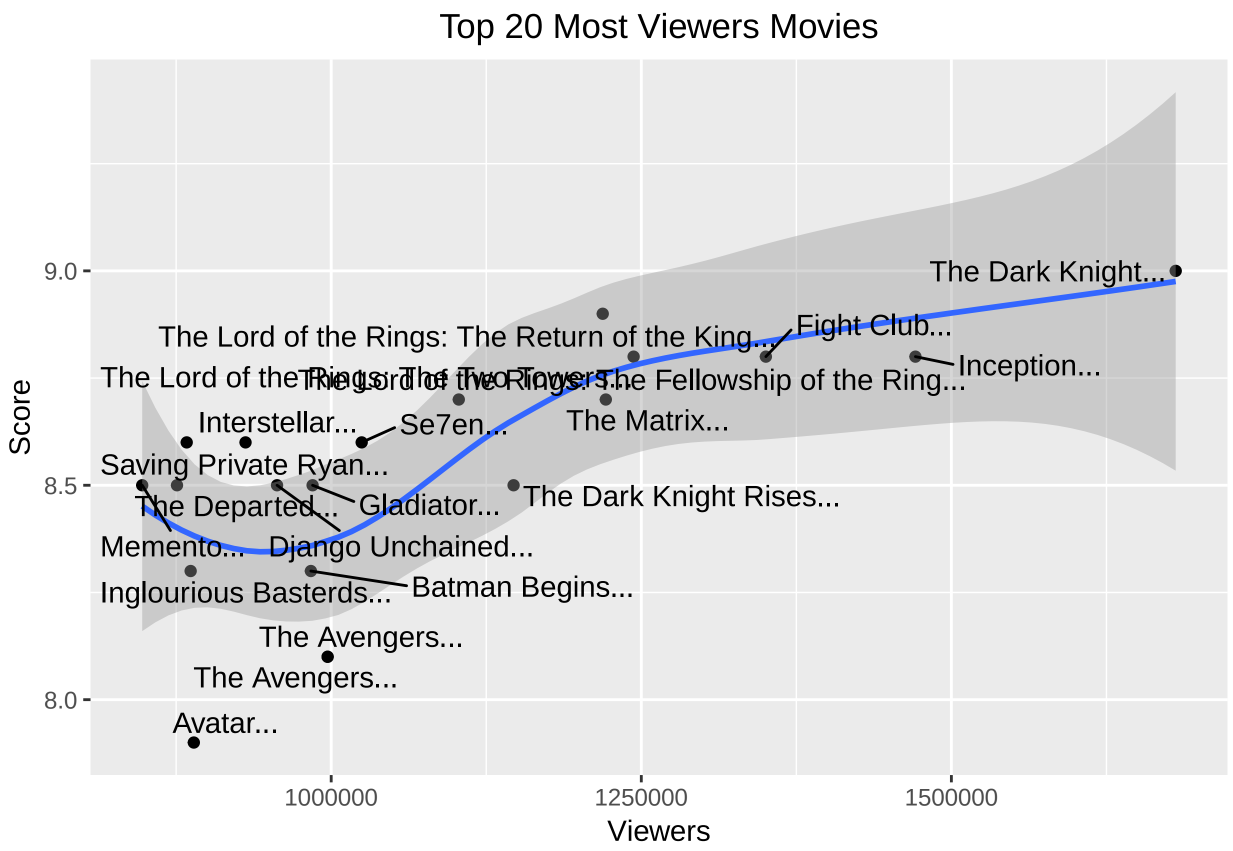


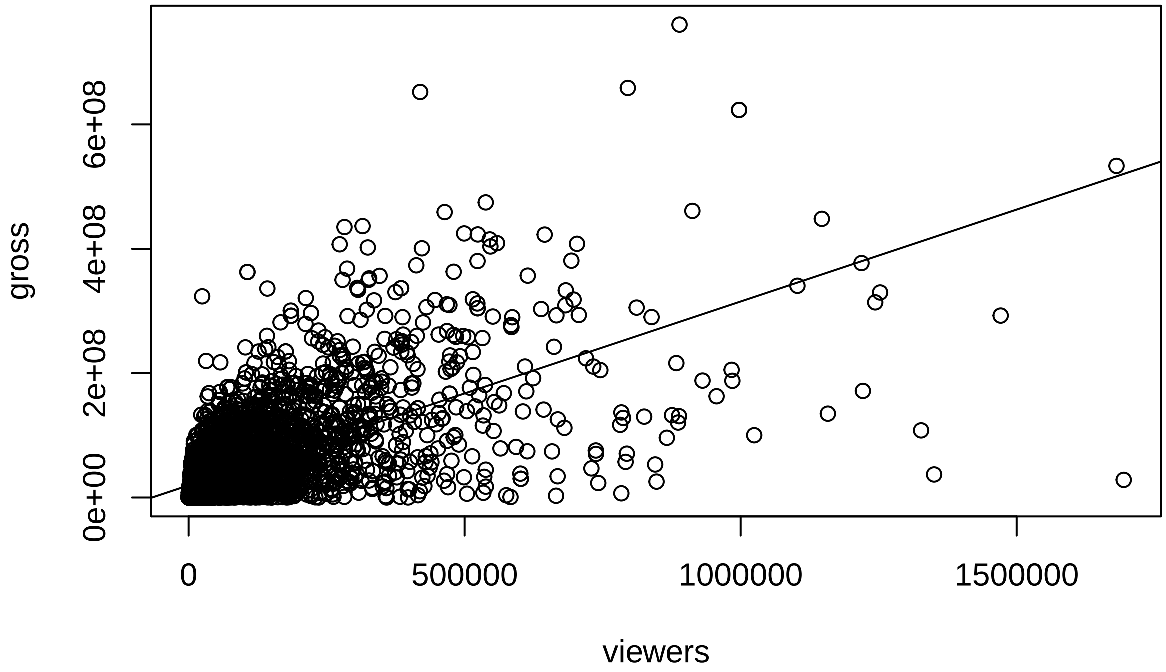
In the plot we can say it's really hard for a movie to both have a high score and high rate. Most movies can only achieved one side.

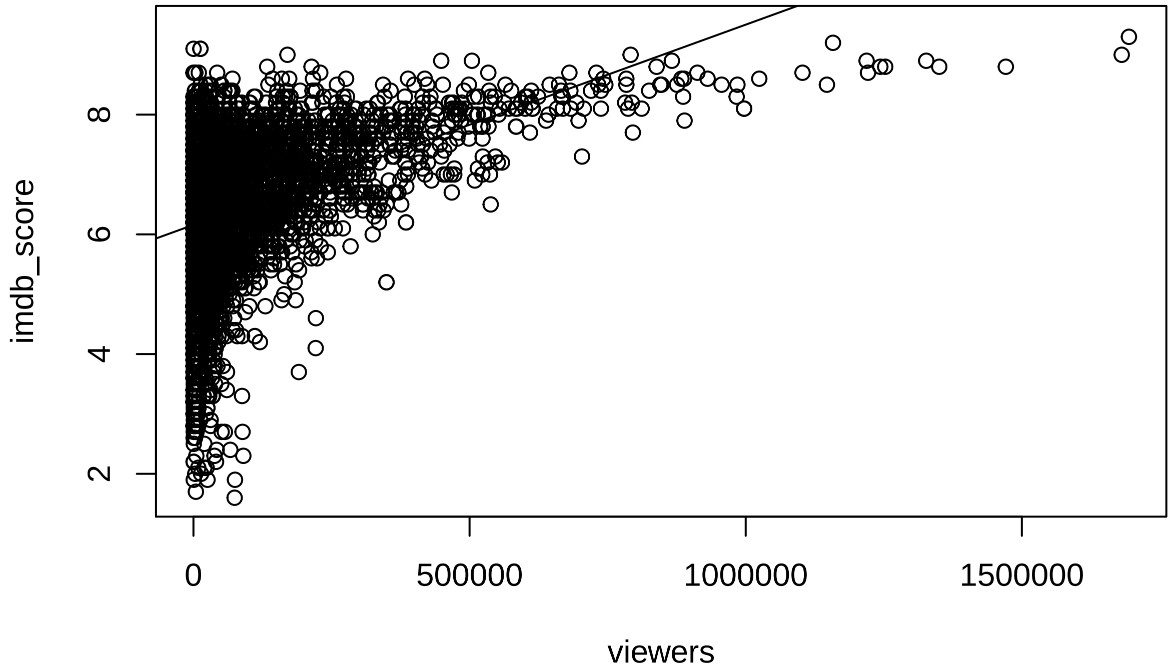
Plot of most viewers movies

Now we plot the Gross of the Most Viewers Movies and the Score of them.




Top 20 Most Viewers Movies
Score
9.0
8.5
8.0
Viewers
1000000
1250000
1500000
The Dark Knight...
The Lord of the Rings: The Return of the King...
Fight Club...
Inception...
The Lord of the Rings: The Two Towers...
The Lord of the Rings: The Fellowship of the Ring...
Interstellar...
Se7en...
The Matrix...
Saving Private Ryan...
The Departed...
Gladiator...
The Dark Knight Rises...
Memento...
Django Unchained...
Inglourious Basterds...
Batman Begins...
The Avengers...
The Avengers...
Avatar...


gross
0e+00
2e+08
4e+08
6e+08
0
500000
1000000
1500000
viewers



Form the Plot we can see for gross there is a weak relation for those most viewers movie to have a large gross. However for the IMDB_score we can say for sure there are relative strong relation for those most viewers movies to have a high score. Based on common sense. I'd like to assume that when a movie have a large enough number of viewers, it's commercial success is destined, the gross is more rely on it's ticket prize or it's budget. If we take all the data into consideration. After fitting a linear model we can see the viewers have positive relation with both score and gross. Such relation is more stable when the viewers reach a high enough number.

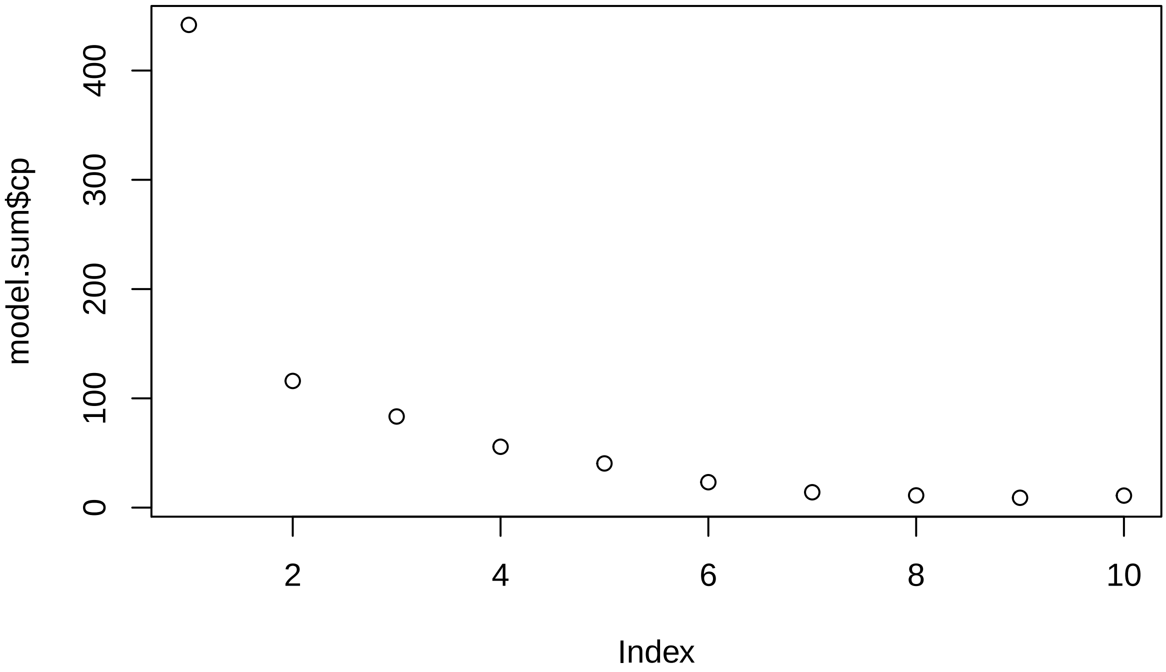
Model fitting

In this part we try to fit varies stastical model on the data. We want to get a deeper understanding of the data.

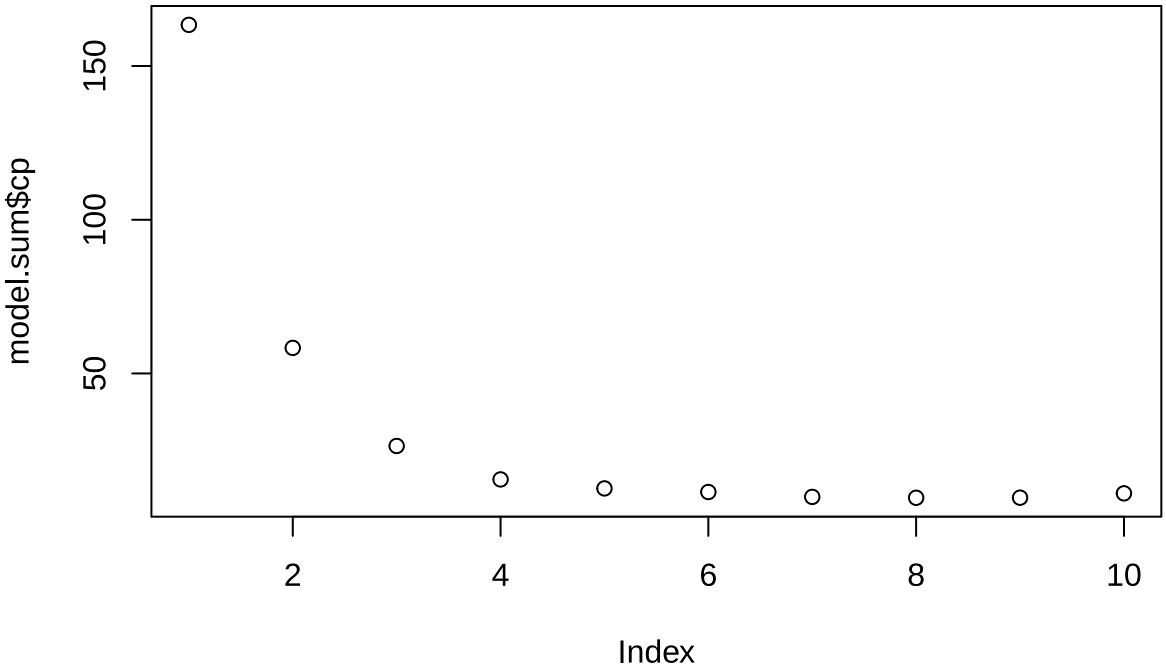
First we want to scale the Data to get rid of magnitude difference effect

Then we split it into Train and Test Dataset

refine the variables



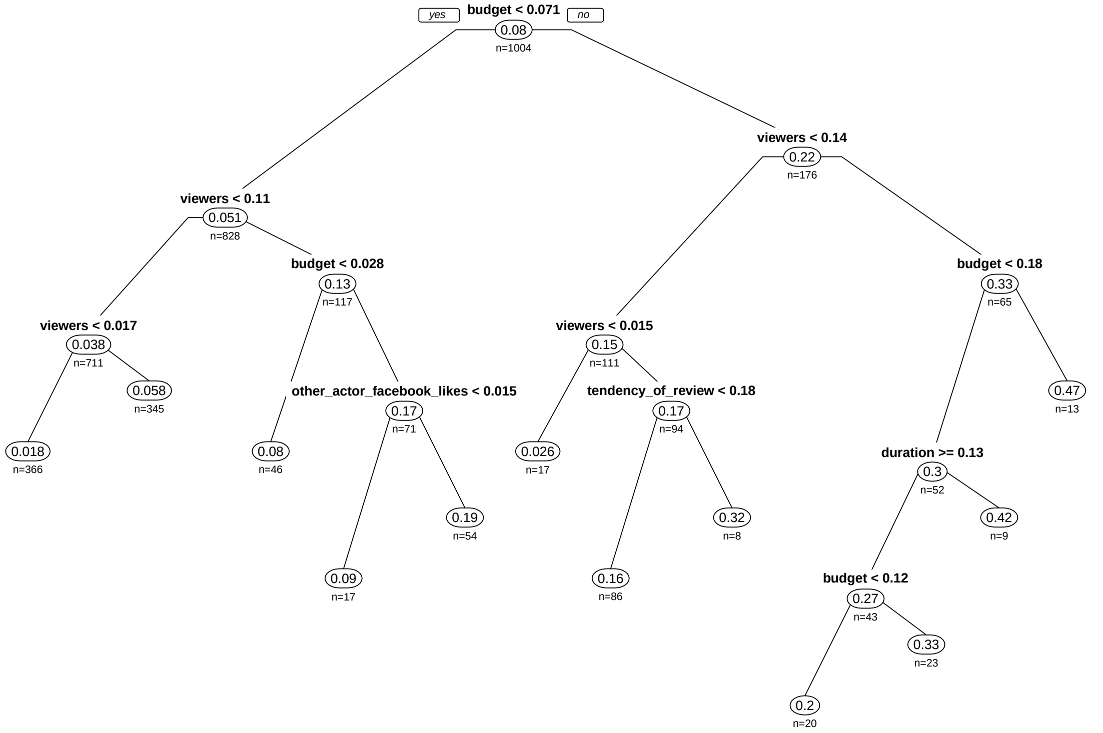
```
##  [1] "(Intercept)"               "duration"
##  [3] "director_facebook_likes"   "facenumber_in_poster"
##  [5] "budget"                    "imdb_score"
##  [7] "aspect_ratio"              "other_actor_facebook_likes"
##  [9] "viewers"                   "tendency_of_review"
```

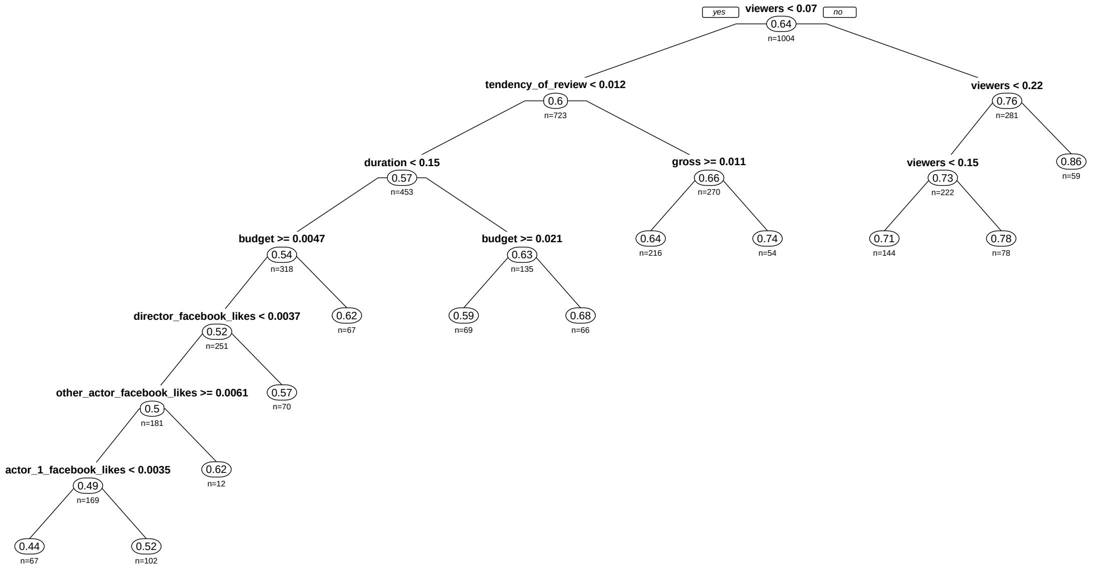


```
## [1] "(Intercept)"                "duration"
## [3] "actor_1_facebook_likes"     "gross"
## [5] "facenumber_in_poster"       "budget"
## [7] "other_actor_facebook_likes" "viewers"
## [9] "tendency_of_review"
```

It tells us the useful parameters in the Data for gross or IMDB score. We can see when parameters is large than 7 there is no large difference so we don't change our variables.

Full Grown Tree





The full grown tree tells us the gross is linked to viewers budget duration and other__acter__facebook__likes.

To be specific when viewers is little the budget and viewers control the gross. When the viewers is large the duation is matters for the small budget movie,the other__acter__facebook__likes is matter with large budget but not that much viewers movie.

The IMDB_score is linked to viewers, budget, duration and tendency of review.

To be specific when viewers is large the main parameter is viewers and budget. When the viewers is little the duration is the second parameter. When duration is short the tendency of review is controling,opposite is budget control.

Random Forest

We use caret to run random forest fitting and use 5-fold crossvalidation.

Now we check the variable importance of the random forest model

```
## rf variable importance
##
##                           Overall
## budget                    100.000
## viewers                    85.631
## tendency_of_review         30.864
## imdb_score                 26.022
## duration                   24.114
## aspect_ratio               18.393
## other_actor_facebook_likes  16.334
## director_facebook_likes    14.025
## actor_1_facebook_likes      6.761
## facenumber_in_poster        0.000
```

```
## rf variable importance
##
##                           Overall
## viewers                   100.000
## duration                   81.084
## tendency_of_review         76.860
## budget                     76.323
## gross                      58.241
## director_facebook_likes    23.645
## other_actor_facebook_likes  23.362
## facenumber_in_poster        9.174
## actor_1_facebook_likes      9.058
## aspect_ratio                0.000
```

Further we check the model performence

```
## [1] 0.01011943
```

```
## [1] 0.003980105
```

The MSE for score is 0.00903

The MSE for gross is 0.00401

The MSE of the model is Very low even close to zero which means the randomforest model prediction has very high accurracy.

Linear regression

We still want to try simple linear regression and ridge regression.

First is the linear regression model and 5-fold crossvalidation

Further we check the model performence

```
## [1] 0.01480953
```

```
## [1] 0.006313239
```

The MSE for score is 0.01251

The MSE for gross is 0.00521

The MSE of the model is Very low even close to zero which means they still have a good enough accuracy.

Second we fit the ridge regression with 5-fold cross-validation

Further we check the model performence

```
## [1] 0.01480962
```

```
## [1] 0.006313165
```

The MSE for score is 0.01252

The MSE for gross is 0.00521

The MSE of the model is Very low even close to zero which means they still have a good enough accuracy. It shows the almost the same result as linear regression.

summary for value prediction

The three model we test all show great accuracy. We choose the best random forest model to predict the outcome of the movie. Further we want to try classification type of modeling.

First we need to change some value in the dataset.

Base on common sense we indicate those moives which have a gross or score higher than the mean value are success movie.

prepare for classification

We set success as 1 fail as 0

Build new test and train data

Bulid a evaluate function

RandomForest(ranger)

Since the random forest give us great accuracy in regression condition, we want use it for classification prediction.

We use caret to run random forest fitting and use 5-fold crossvalidation.

Evaluate the model

```
## [1] 0.8598131
```

```
## [1] 0.796729
```

The accuracy for gross is 0.82944

The accuracy for score is 0.78972

The accuracy for predict the success of movie is moderate, can be use to predict but need improvement.

Linear classifier

We still can use glm to fit a model

Evaluate the model

```
## [1] 0.8621495
```

```
## [1] 0.7266355
```

The accuracy for gross is 0.78972

The accuracy for score is 0.75467

The accuracy for predict the success of movie is worse.

We can try logic regression too

Evaluate the model

```
## [1] 0.8621495
```

```
## [1] 0.7266355
```

The accuracy for gross is 0.78972

The accuracy for score is 0.72897

The accuracy for predict the success of movie is worse.

naive bayes

We try bayes classifier to fit a model

Evaluate the model

```
## [1] 0.8107477
```

```
## [1] 0.7102804
```

The accuracy for gross is 0.77336

The accuracy for score is 0.71729

The accuracy for predict the success of movie is still worse.

SVM

We try SVM model to do classification

Evaluate the model

```
## [1] 0.8551402
```

```
## [1] 0.7640187
```

The accuracy for gross is 0.81776

The accuracy for score is 0.74533

The accuracy for predict the success of movie is still worse than random forest.

Ada boost

Try ada boost

Evaluate the model

```
## [1] 0.8504673
```

```
## [1] 0.796729
```

The accuracy for gross is 0.83177

The accuracy for score is 0.76636

The accuracy for predict the success of movie is worse than random forest.

Deep Neural network

In the end we want to try DNN

Evaluate the model

```
## [1] 0.6915888
```

Evaluate the model

```
## [1] 0.6915888
```

The accuracy for score and gross is the same 0.69159

Which is the worst of all classifier model.

Summary for classification

The random forest model is the best.

The accuracy for gross is 0.82944

The accuracy for score is 0.78972

The accuracy for predict the success of movie is moderate, can be use to predict but need improvement.

Summary for the predictive model

We generate model to predict the value of gross and imdb_score base on specific parameters which have extremely low MSE and model to predict whether a movie can achieve commercial or art success which may not have that high accuracy but is good enough to use.

Summary for the inner relation of the data

In this project we find out there are several parameters can control the gross and score.

To be specific when viewers is little the budget and viewers control the gross. When the viewers is large the duation is matters for the small budget movie,the other_acter_facebook_likes is matter with large budget but not that much viewers movie.

The IMDB_score is linked to viewers, budget, duration and tendency of review.

To be specific when viewers is large the main parameter is viewers and budget. When the viewers is little the duration is the second parameter. When duration is short the tendency of review is controling,opposite is budget control.

The improtant parameters for gross are budget, viewers, imdb_score, tendency_of_review, duration.

The improtant parameters for score are budget, viewers, tendency_of_review, duration, gross.

So it's fit the common sense that a high budget movie is reasonable to be believed to have a commercial and art success. For large budget they are well produced and can arract more viewers which is lead to higher gross and better IMDB_score.

Challenge the result

We can noticed that by using those parameters to predict whether the movie success or not is not that accurate like predict the value of their gross and score. In the project we set the success standard is above the mean value of all the movie which means it's relative low standard for success. So I believe in the real situation the standard for success is much higher and it will take some other ignored parameters in to consideration like the fame of actors and directors which will significantly influence the viewers and the tendency to review.