CSP 571 DATA PREPARATION AND ANALYSIS
MOVIE DATA ANALYSIS

GROUP MEMBERS

- Mayur Mehta A20405901
- Abhilash Bhurse A20404893
- Shikha Verma A20408401
- Ayushi Patel A20407392

INTRODUCTION

- A movie is not only for entertaining users, but also for a film company to make great profits. There are lot of factors needed for a movie to be a commercial success.
- **Project Definition and Goal.**

- **Definition**
  - For this project, we take IMDB movie dataset from Kaggle website and analyze what kind of movies are more successful or obtained a higher IMDB score than others.
  - To identify some interesting patterns from the data - derive graphical representations to visualize the information easily and come up with some conclusions such as the countries that produce most movies, profitability analysis, most produced genres of movies and many such patterns.

- **Goal :** The goal of this project is to derive such insights which help making an informed decision for the future generations of movies.

DATA PREPARATION and PRE-PROCESSING

- Load Data from Kaggle site
- Calculate and remove duplicate values
- Remove spurious values from movie title column
- Check and remove n/a values and deal with 0 values - converting 0 to n/a
- Deleting part : unnecessary columns, predictor color, column language
- Adding variable column profit, where profit=gross-budget

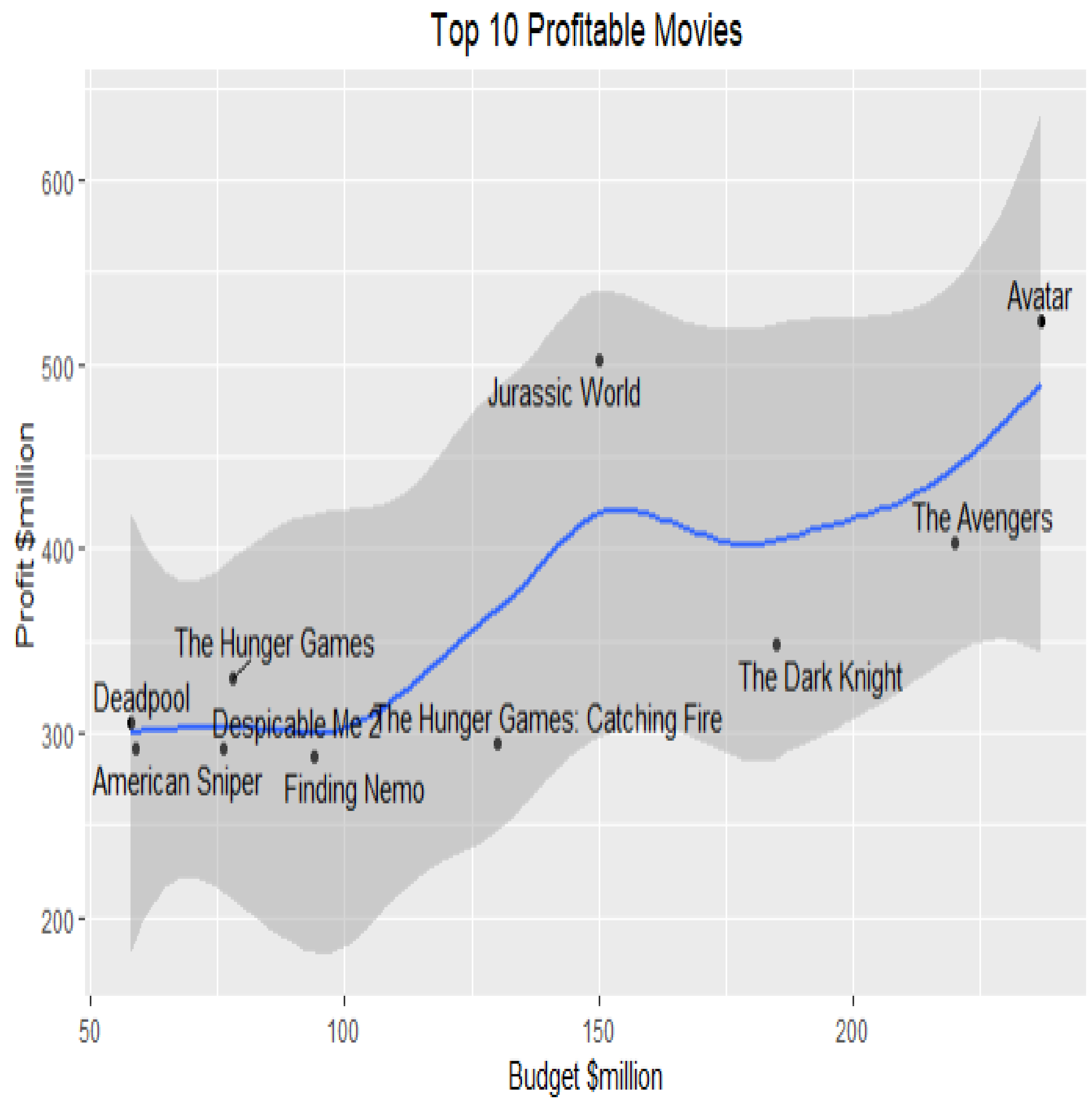
APPROACHES AND METHODOLOGY

- Data Exploration/Visualization

- Genre and Country Analysis
- Modeling
  i. Linear Model Selection : Simple Regression, Multiple Regression
  ii. Non-linear Model Selection : Random Forest

DATA EXPLORATION/VISUALIZATION

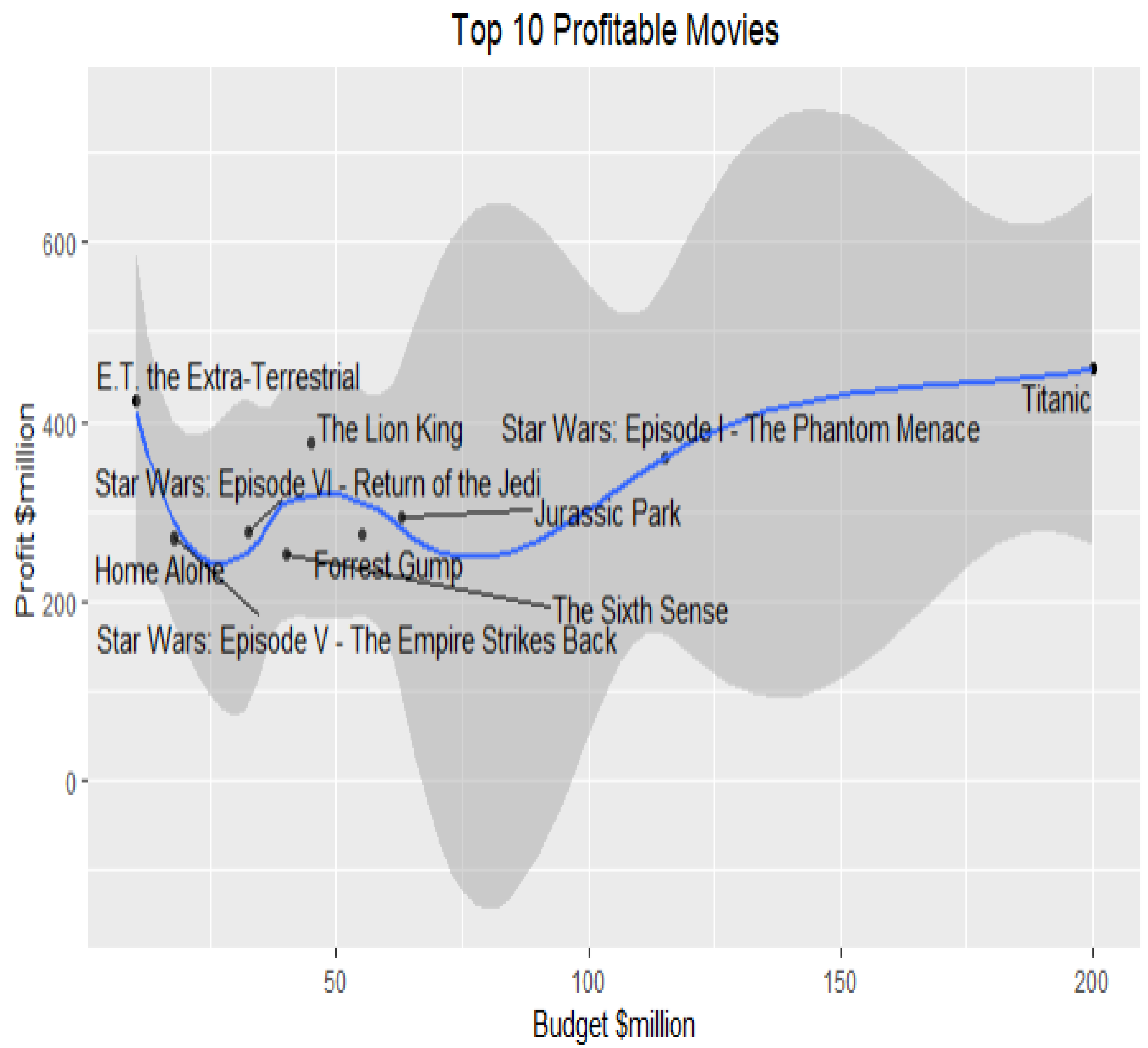
- Top movies based on profit in the decade of 2000-2016.
- We can observe that movies like Avatar and Avengers have made a good profit (less budget comparatively with more profit)



#Here in 21st century, we can observe that movies like Avatar, Avengers, Jurassic world made a good profit.

DATA EXPLORATION/VISUALIZATION

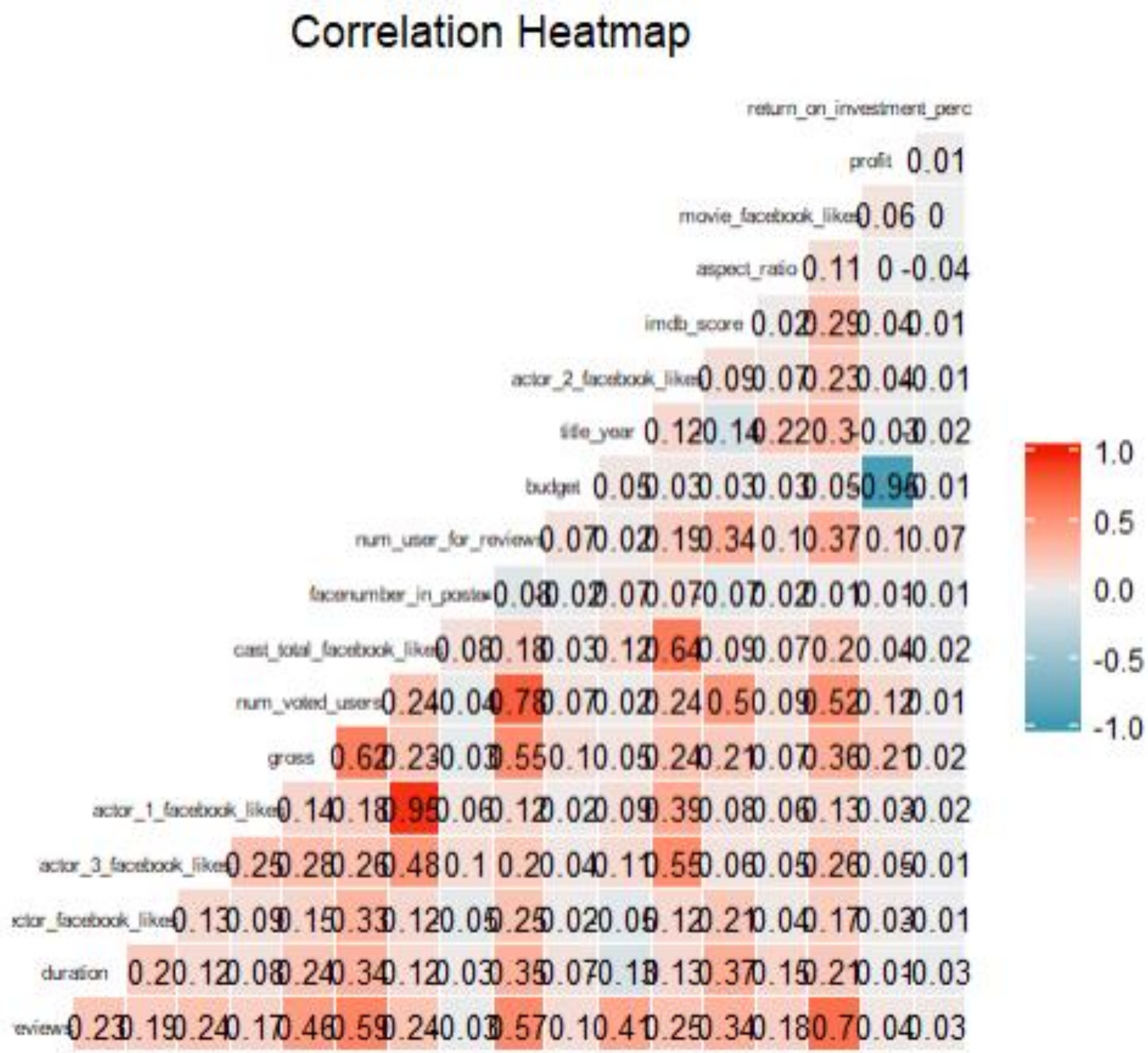
- Top movies based on profit in the decade of 1980-2000.
- We can observe that movies like Titanic and Star wars have made a good profit (less budget comparatively with more profit)



#We can observe The Star wars and Titanic has made a good profit in 90's (less budget comparatively with more profit)

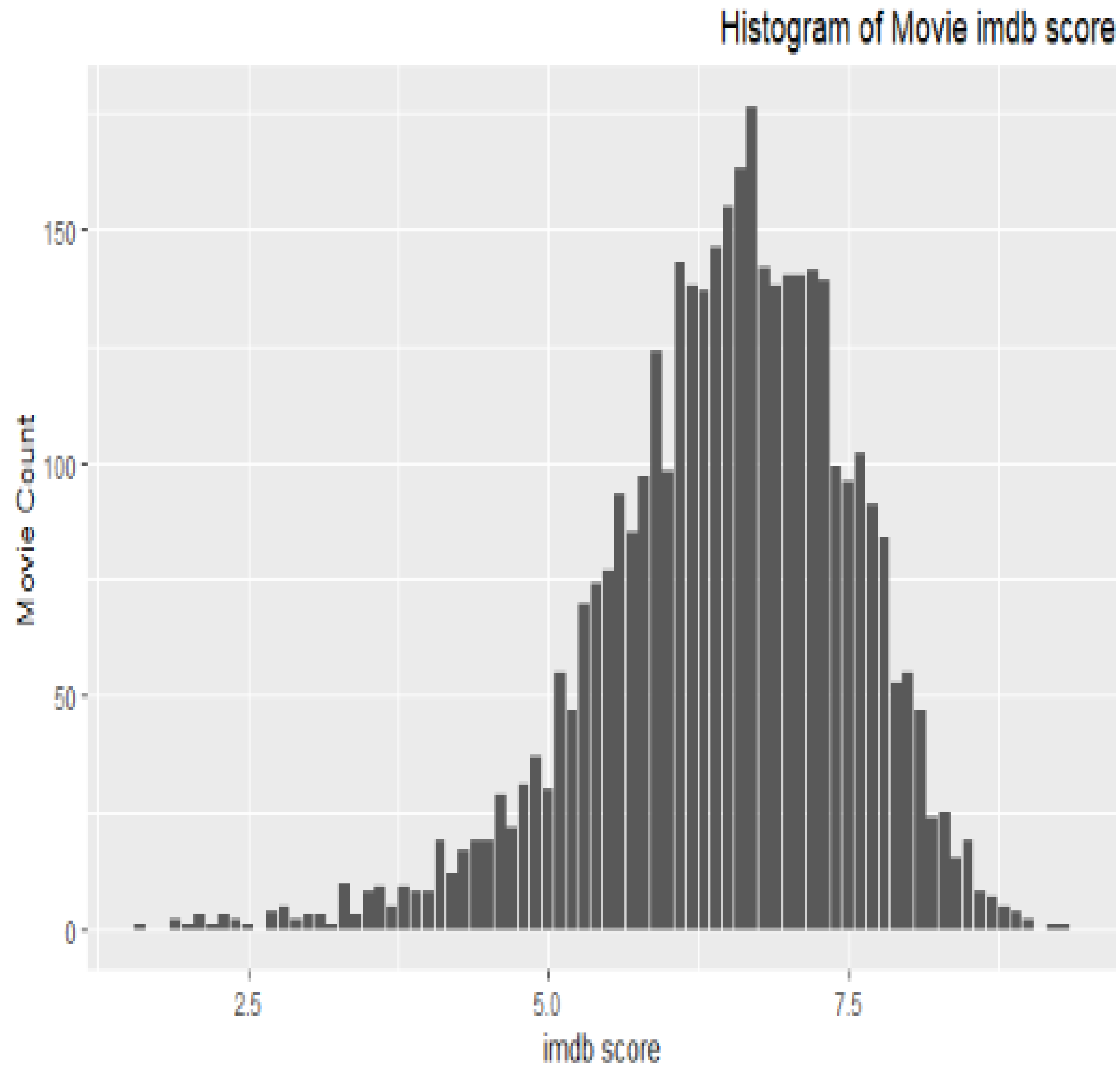
DATA EXPLORATION/VISUALIZATION

- This correlation map is used to tell the association strength of variables in dataset
- On the basis of this heatmap, we can find some high correlations (greater than 0.7) between predictors.
- Here the highest correlation is between actor_1_facebook_like and cast_total_facebook_likes



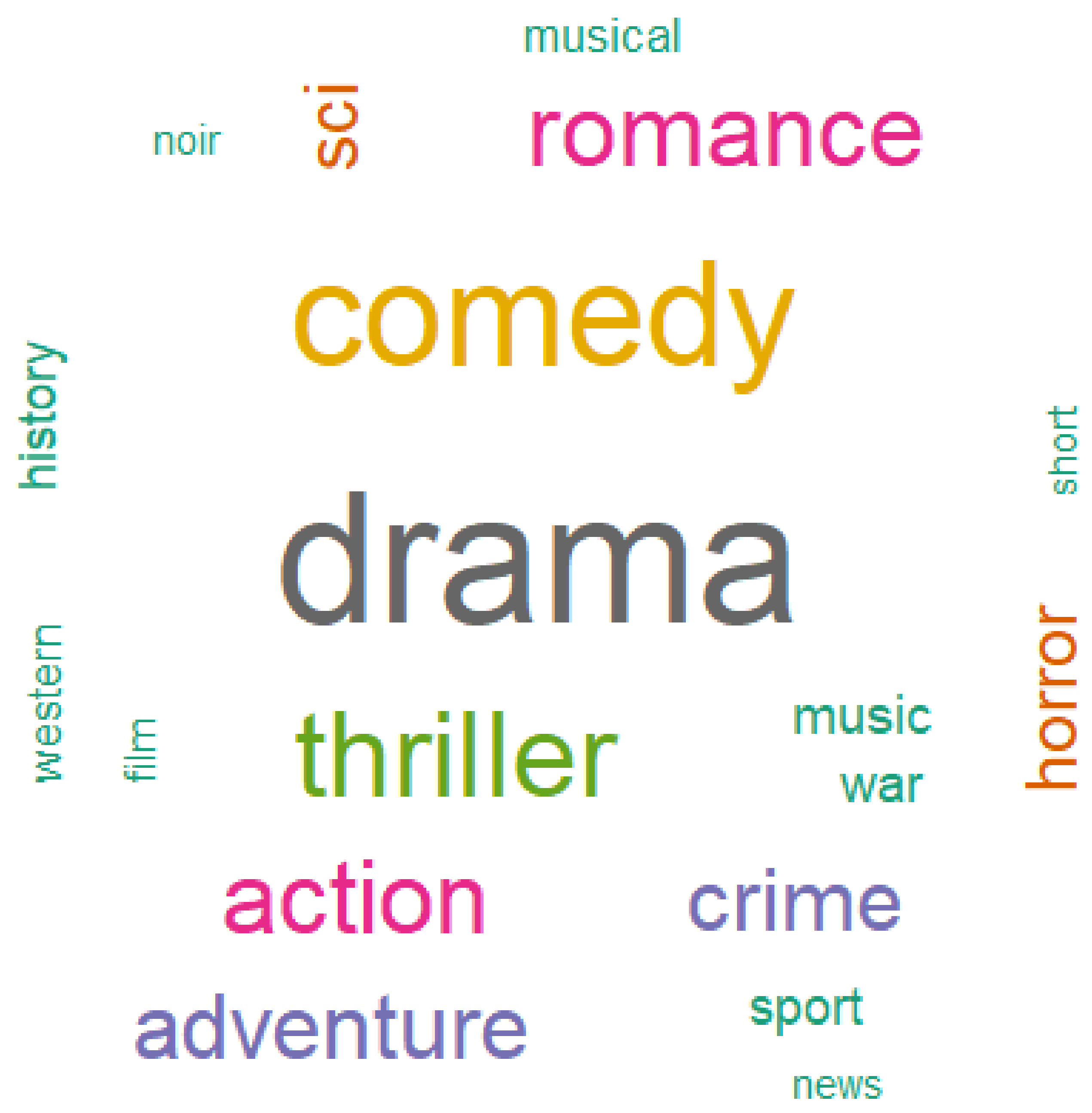
GENERAL ANALYSIS

- Histogram Representation
- We can observe that there are numerous movies with imdb more than 4.5



GENRE ANALYSIS

- Here we determine which genres are used in movie production frequently
- From the word association plot we can observe that DRAMA, COMEDY and Thriller are most used genres



DATA EXPLORATION/VISUALIZATION

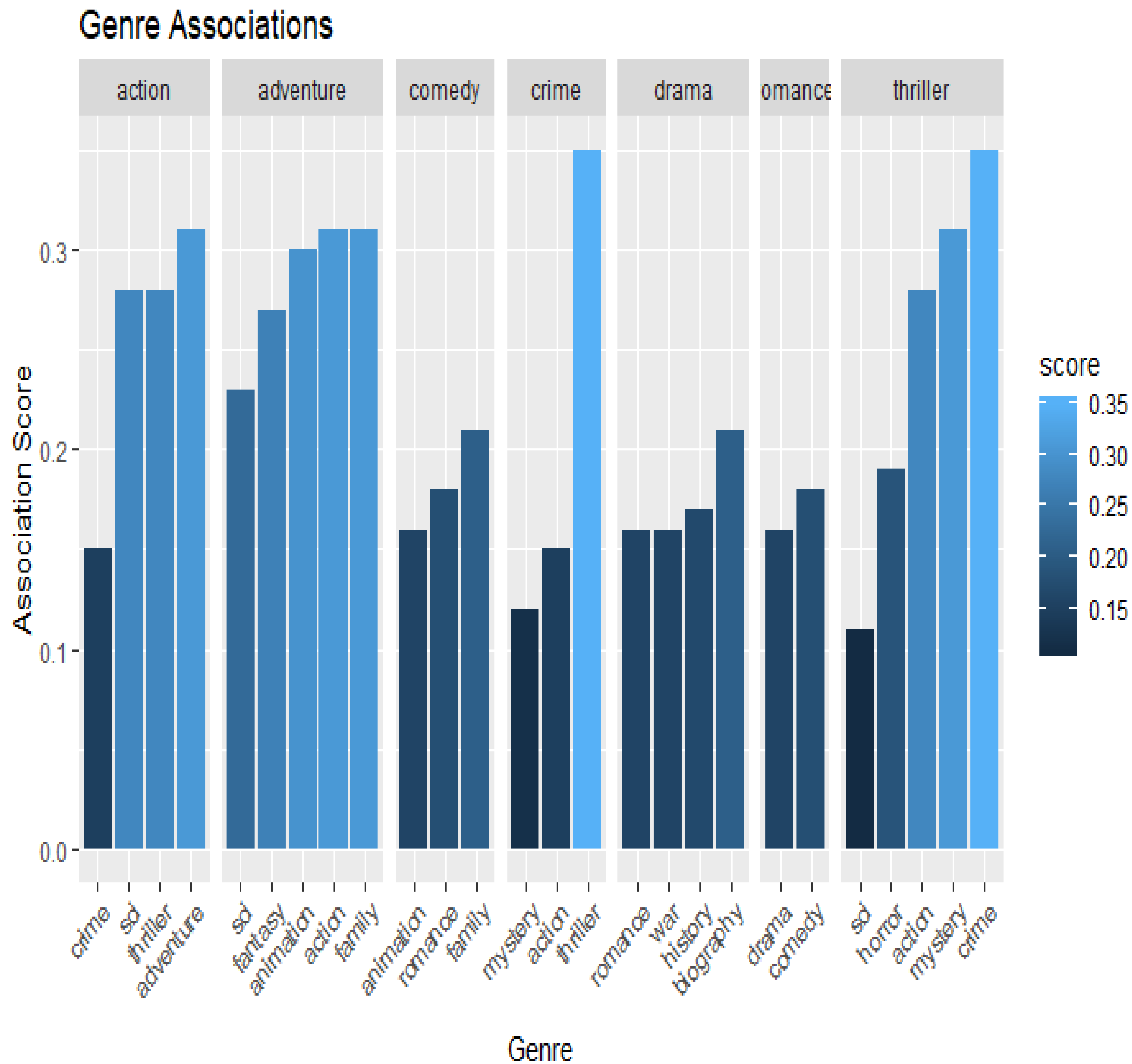
- Top 20 genres based on average IMDB Score
- We can observe Adventure|Animation|Drama|Family|Crime|Drama has good imdb score(avge(8.5))

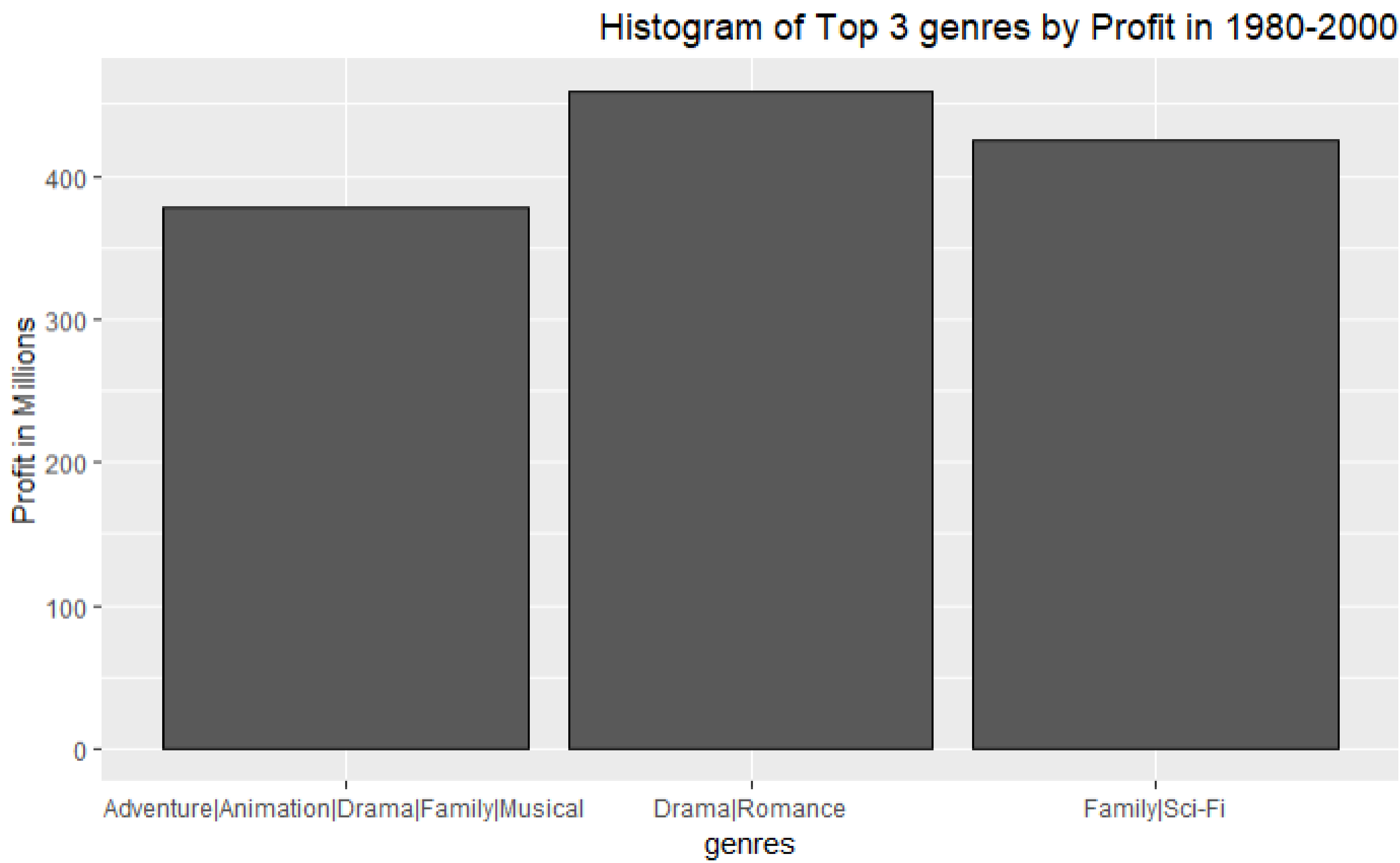
genres	avg_imdb
Adventure|Animation|Drama|Family|Musical	8.50
Crime|Drama|Fantasy|Mystery	8.50
Action|Adventure|Drama|Fantasy|War	8.40
Adventure|Animation|Fantasy	8.40
Adventure|Drama|Thriller|War	8.40
Adventure|Animation|Comedy|Drama|Family|Fantasy	8.30
Biography|Drama|History|Music	8.30
Documentary|Drama|Sport	8.30
Documentary|War	8.30
Adventure|Drama|War	8.25
Biography|Crime|Documentary|History	8.20
Drama|Fantasy|War	8.20
Drama|Mystery|War	8.20
Action|Animation|Sci-Fi	8.10
Adventure|Comedy|Crime|Drama	8.10
Adventure|Drama|Thriller|Western	8.10
Biography|Crime|Drama|History	8.10
Biography|Crime|Drama|Western	8.10
Documentary|History|Music	8.10
Action|Adventure|Animation|Family	8.00
Animation|Biography|Documentary|Drama|History|War	8.00
Animation|Biography|Drama|War	8.00
Biography|Drama|Family|Musical|Romance	8.00
Crime|Documentary|Drama	8.00
Crime|Drama|Musical	8.00

genres	avg_gross
Family|Sci-Fi	434949459
Adventure|Animation|Drama|Family|Musical	422783777
Adventure|Animation|Comedy|Drama|Family|Fantasy	356454367
Action|Biography|Drama|History|Thriller|War	350123553
Action|Adventure|Fantasy|Sci-Fi	296684758
Adventure|Drama|Fantasy|Romance	296481890
Action|Adventure|Fantasy|Romance	289279970
Adventure|Sci-Fi	281666058
Adventure|Family|Fantasy|Mystery	279056317
Action|Adventure|Animation|Family	261437578
Action|Adventure|Comedy|Family|Fantasy	250863268
Adventure|Comedy|Family|Mystery|Sci-Fi	250147615
Animation|Comedy|Family|Sci-Fi	246459955
Adventure|Animation|Comedy|Family|Fantasy|Romance	243310828
Action|Adventure|Family|Fantasy|Romance	241407328
Adventure|Drama|Family|Fantasy	221186651
Adventure|Sci-Fi|Thriller	221050859
Adventure|Animation|Comedy|Family|Fantasy|Musical	220068175
Animation|Comedy|Family|Fantasy|Music	218469864
Drama|Fantasy|Romance|Thriller	217631306

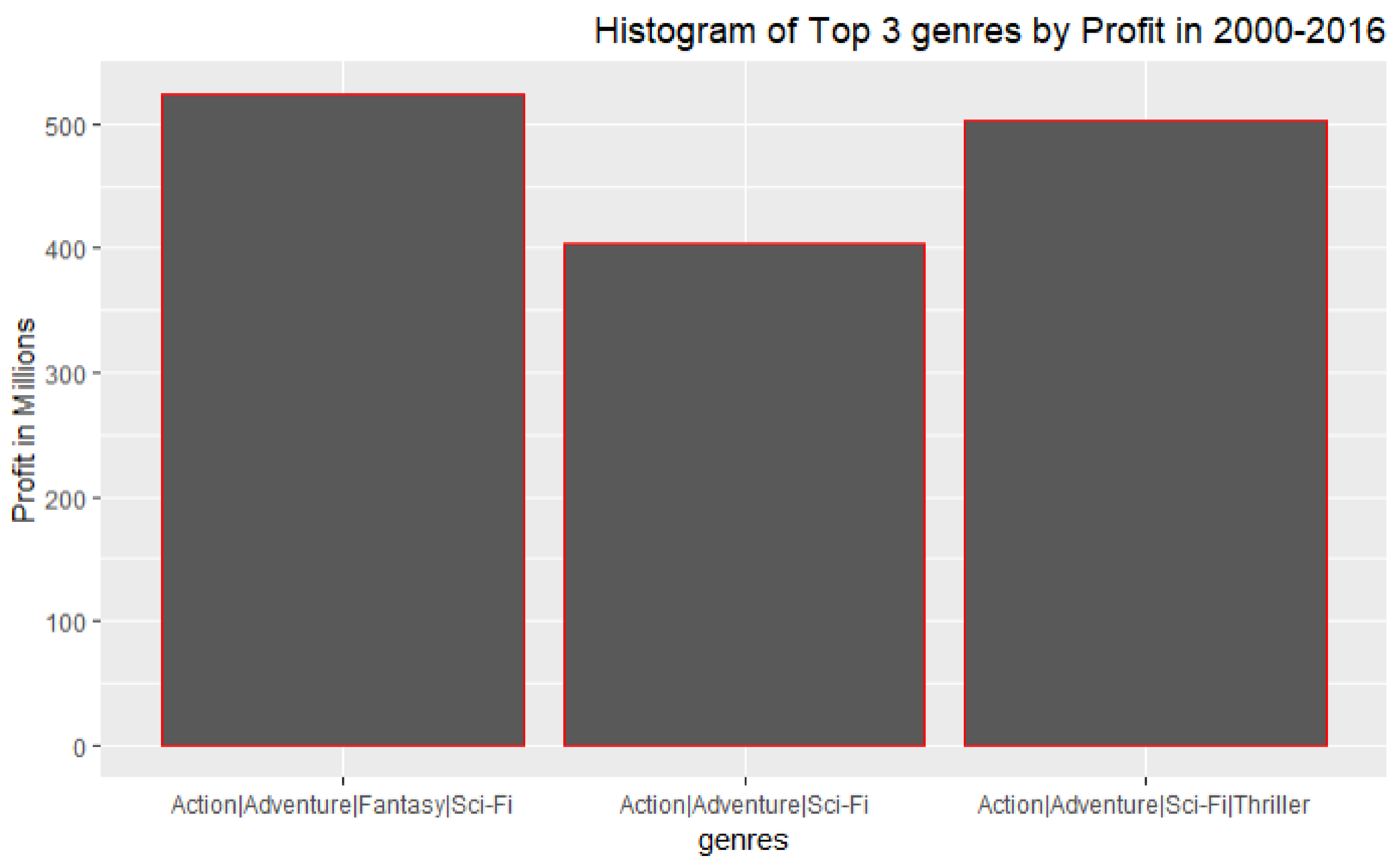
GENRE ANALYSIS

- We also analyzed that how stringly one genre is associated to other genres
- From the analysis graph, Thriller and Crime Genres are closely associated
- Next, Thriller and Mystery are closely associated

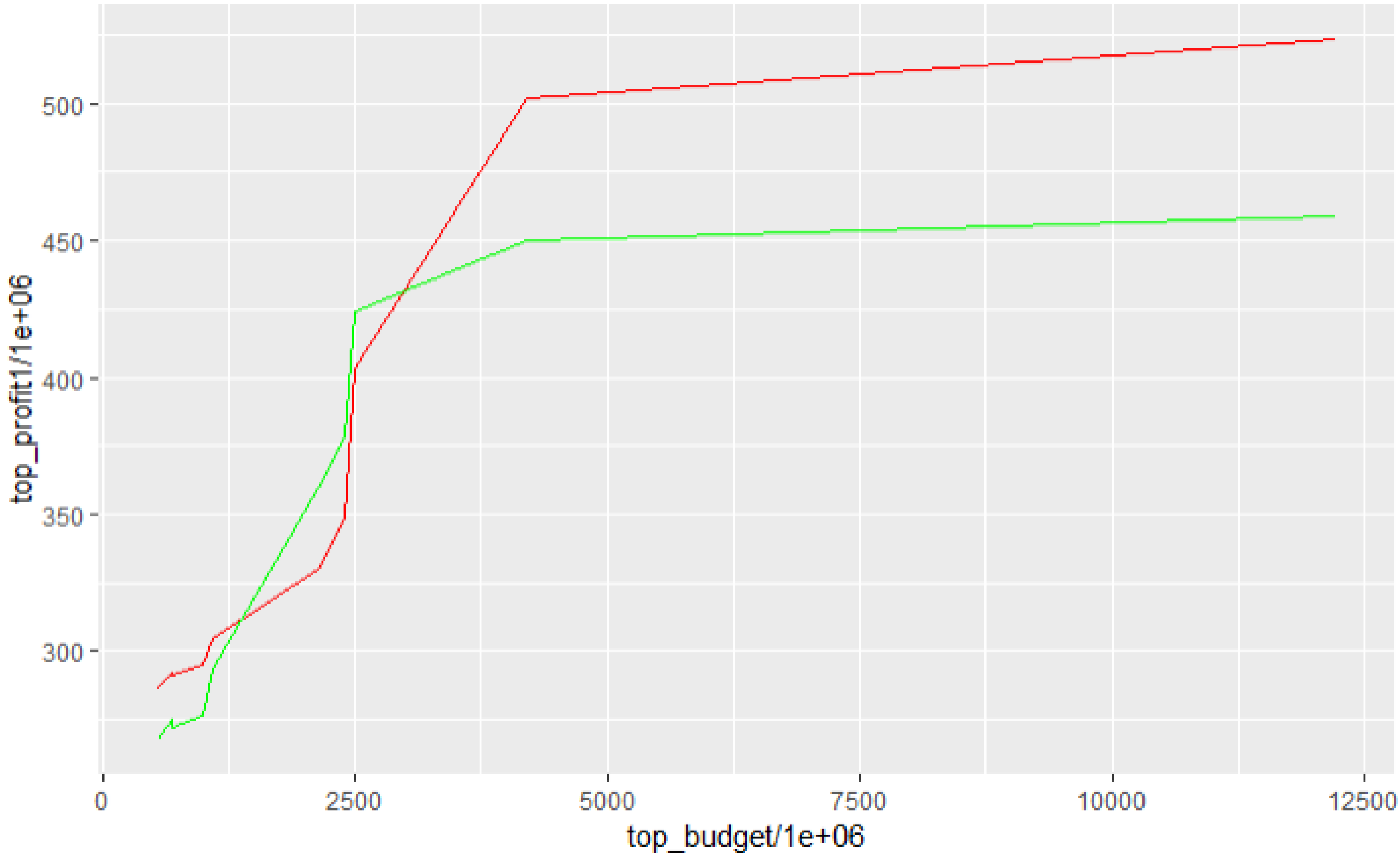





```
#in 90's, Movies having combination of Drama|romance such as 'Titanic' were more popular and made a great
profit in those years.
```



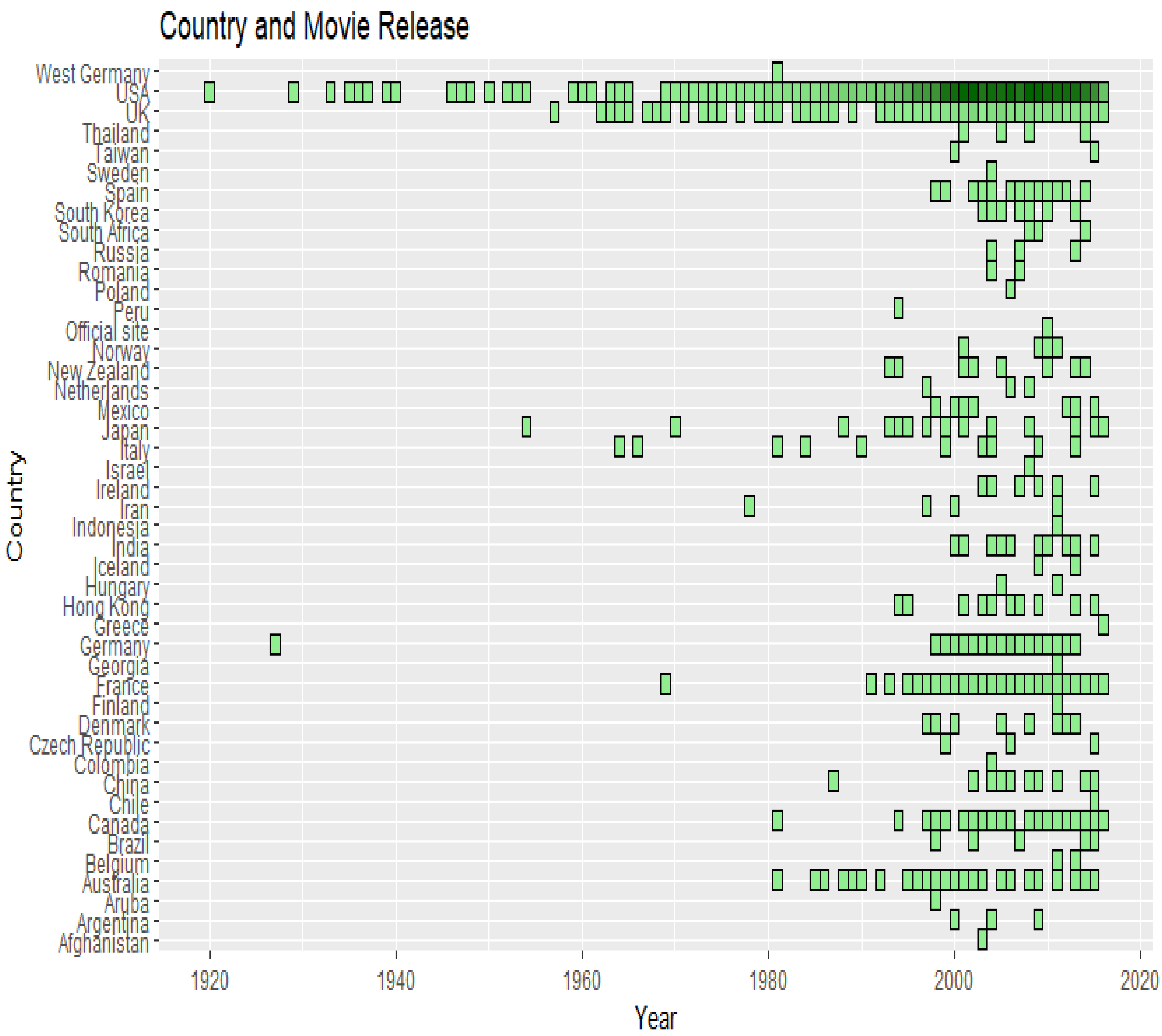
#In 21st century, with the advancement in technology, Movies having combination of Action|Adventure|Fantasy|Sci-Fi such as 'Avengers' made a great profit and are more popular.



#Here a comparision has been made between two profits columns, the green line indicates profit earned by movies released in years between 1980-2000 and the red line indicates profits earned by movies released in years 2000-2016.

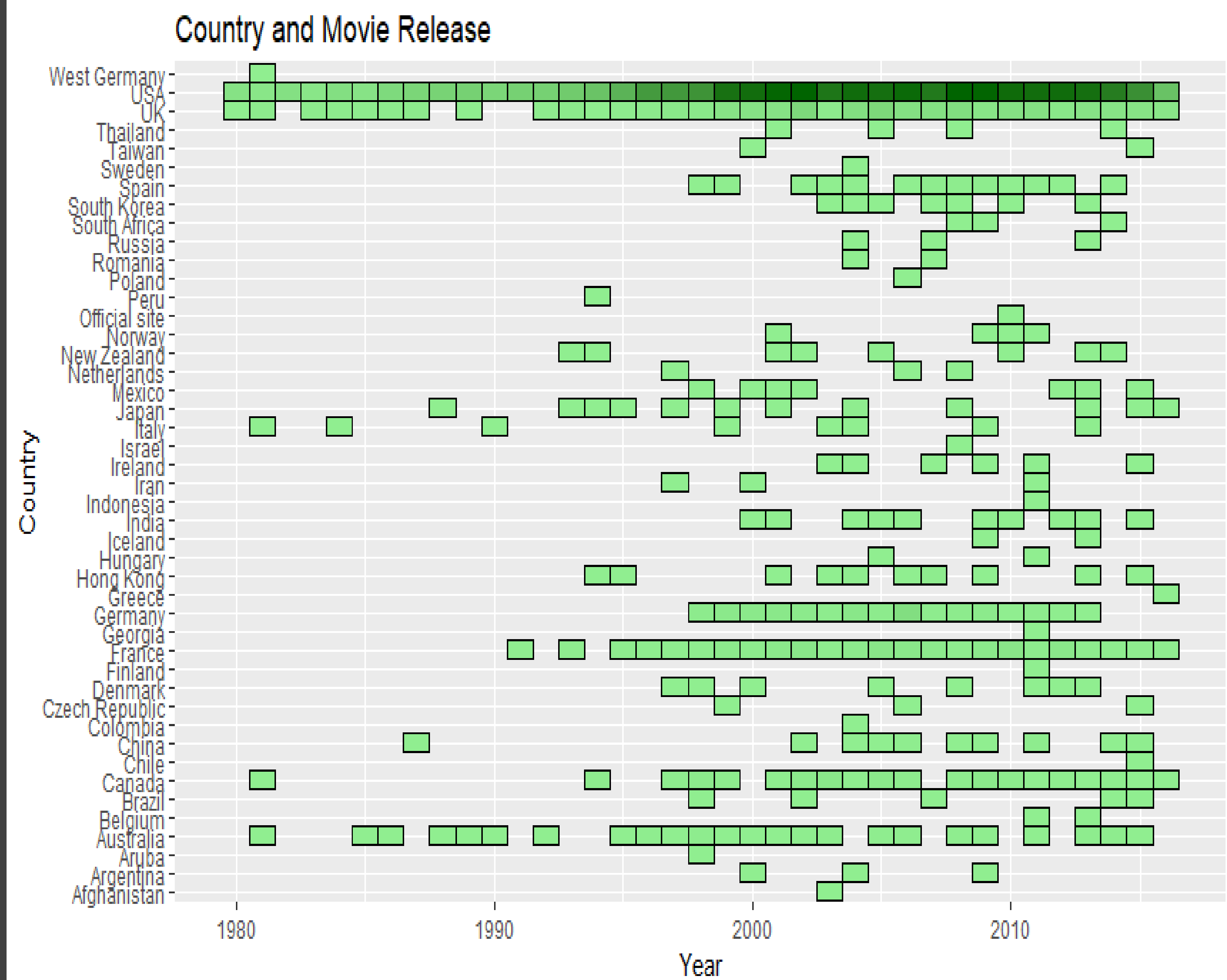
COUNTRY ANALYSIS

- The analysis states which countries has released the highest number of movies
- From the graph it is observed that not many movies released during the period before 1980s and increased in the late 1990s
- Also, it can be observed that highest number of movies are released in West Germany, USA and UK



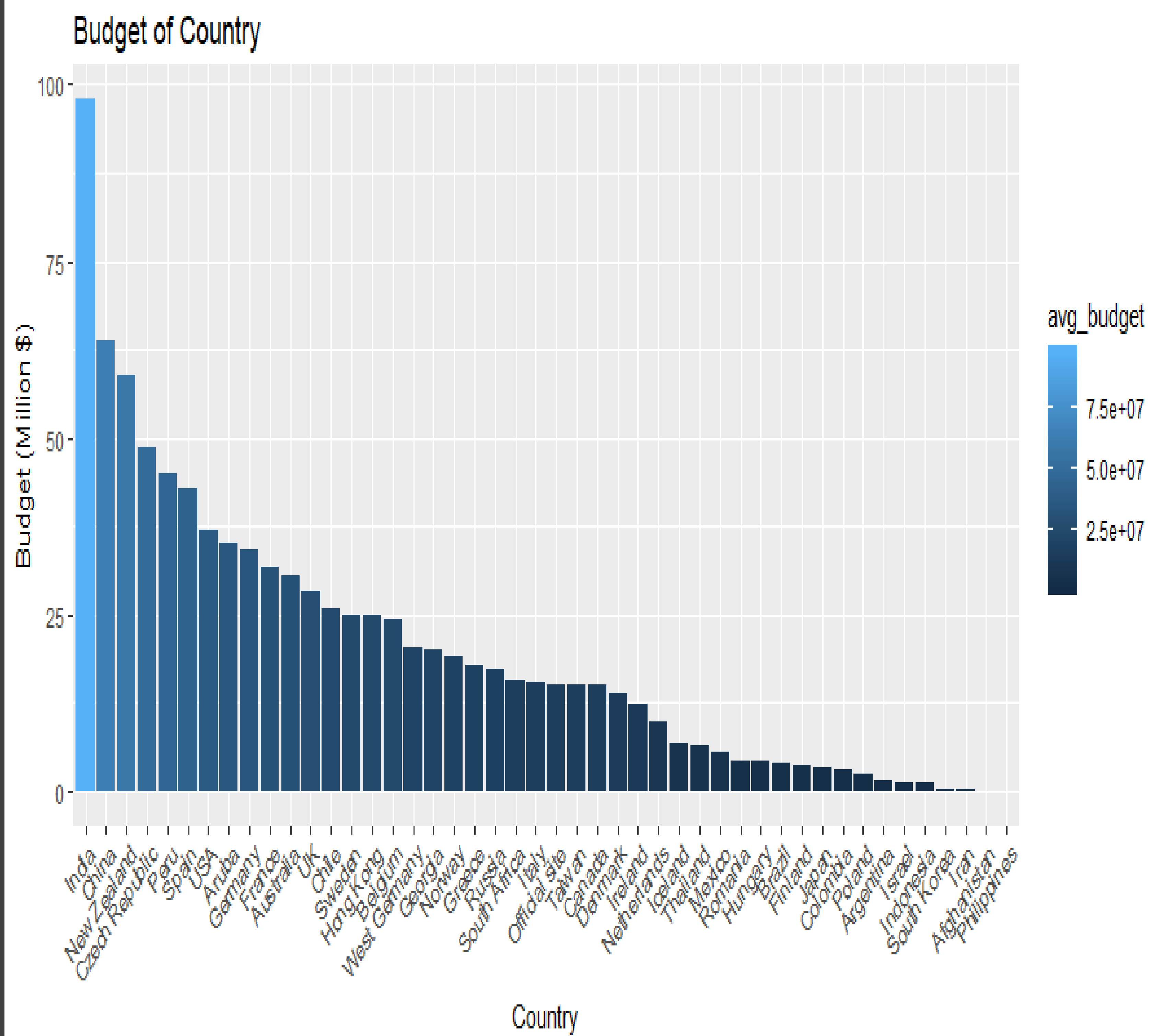
COUNTRY ANALYSIS

- As the movies started releasing in high number after 1980s which was concluded from the above graph, the movies released before 1980s would not play much importance and hence we update the graph



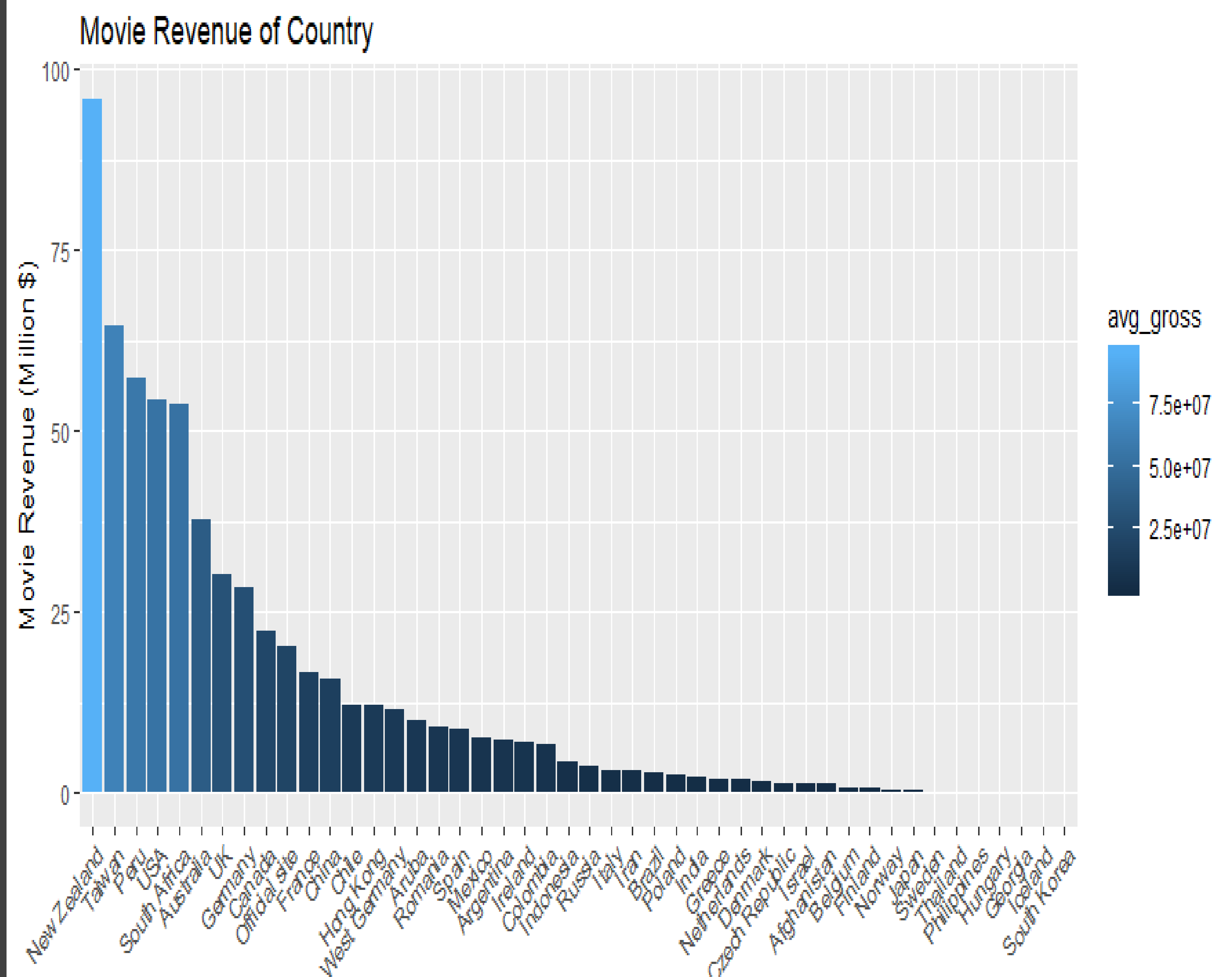
COUNTRY ANALYSIS

- Here the graph represents the analysis of budget of movies in every country
- India has the highest amount of budget spent followed by China and New Zealand



COUNTRY ANALYSIS

- Here the graph represents the analysis of revenue collected by each country from movies
- New Zealand has the highest amount of revenue collected due to movies followed by Taiwan and Peru



SIMPLE REGRESSION

```
> #imdb score vs gross
> sample.reg.model.3 <- lm(gross ~ imdb_score, data = movie)
> summary(sample.reg.model.3)

Call:
lm(formula = gross ~ imdb_score, data = movie)

Residuals:
      Min        1Q    Median        3Q       Max 
-83154381 -43270953 -17615179  17210452 688333320 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -38930255    6887236  -5.653  1.7e-08 ***
imdb_score   14063643    1051175  13.379  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68530000 on 3799 degrees of freedom
Multiple R-squared:  0.045,	Adjusted R-squared:  0.04475 
F-statistic:   179 on 1 and 3799 DF,  p-value: < 2.2e-16

> |
```

But, How Significant is IMDB_SCORE to determine GROSS REVENUE??

```
> #Determining correlation between gross and imdb_score
> cor(movie$gross, movie$imdb_score)
[1] 0.2121244
>
> cat("\nimdb_score is an important predictor, but it alone does not provide bette
r prediction of gross revenue. This means, only a good imdb_score does not indicat
e a higher gross revenue of a movie!!")

imdb_score is an important predictor, but it alone does not provide better predict
ion of gross revenue. This means, only a good imdb_score does not indicate a highe
r gross revenue of a movie!!
>
```

MULTIPLE REGRESSION

PURPOSE : To determine which Predictors are important to Gross Revenue

```
Call:
lm(formula = gross ~ num_critic_for_reviews + duration + director_facebook_likes +

    actor_3_facebook_likes + actor_1_facebook_likes + num_voted_users +
    cast_total_facebook_likes + facenumber_in_poster + num_user_for_reviews +
    budget + title_year + actor_2_facebook_likes + imdb_score +
    aspect_ratio + movie_facebook_likes, data = movie)

Residuals:
       Min         1Q     Median         3Q        Max
-414026940  -23453072   -8099007   13237420  475002637

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                7.474e+08  2.058e+08   3.631 0.000286 ***
num_critic_for_reviews     9.590e+04  1.193e+04   8.036 1.23e-15 ***
duration                   1.233e+05  4.205e+04   2.931 0.003395 **
director_facebook_likes   -1.291e+03  2.868e+02  -4.504 6.88e-06 ***
actor_3_facebook_likes    -1.178e+04  1.272e+03  -9.264  < 2e-16 ***
actor_1_facebook_likes    -1.054e+04  7.658e+02 -13.768  < 2e-16 ***
num_voted_users            2.282e+02  1.041e+01  21.917  < 2e-16 ***
cast_total_facebook_likes  1.052e+04  7.632e+02  13.783  < 2e-16 ***
facenumber_in_poster      -9.386e+05  4.111e+05  -2.283 0.022461 *
num_user_for_reviews       1.156e+04  3.503e+03   3.299 0.000978 ***
budget                     1.307e-02  3.709e-03   3.524 0.000429 ***
title_year                -3.543e+05  1.026e+05  -3.455 0.000557 ***
actor_2_facebook_likes    -1.004e+04  8.092e+02 -12.413  < 2e-16 ***
imdb_score                -7.133e+06  9.679e+05  -7.369 2.10e-13 ***
aspect_ratio              -1.900e+06  2.456e+06  -0.773 0.439293
movie_facebook_likes      -1.121e+02  5.752e+01  -1.949 0.051369 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50980000 on 3785 degrees of freedom
Multiple R-squared:  0.4736,    Adjusted R-squared:  0.4715
F-statistic:   227 on 15 and 3785 DF,  p-value: < 2.2e-16
```

USA Data :

```
Call:
lm(formula = gross ~ num_critic_for_reviews + duration + director_facebook_likes +

    actor_3_facebook_likes + actor_1_facebook_likes + num_voted_users +
    cast_total_facebook_likes + facenumber_in_poster + num_user_for_reviews +
    budget + title_year + actor_2_facebook_likes + imdb_score +
    aspect_ratio + movie_facebook_likes, data = movie.usa)

Residuals:
       Min         1Q     Median         3Q        Max
-342978430  -19965326   -5627919   13411587  438495657

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.419e+09  2.118e+08   6.699 2.50e-11 ***
num_critic_for_reviews     2.434e+04  1.245e+04   1.956  0.05061 .
duration                  -1.547e+05  4.467e+04  -3.464  0.00054 ***
director_facebook_likes   -1.199e+03  2.652e+02  -4.522 6.38e-06 ***
actor_3_facebook_likes    -8.595e+03  1.188e+03  -7.233 5.96e-13 ***
actor_1_facebook_likes    -7.544e+03  7.249e+02 -10.408  < 2e-16 ***
num_voted_users            1.930e+02  1.012e+01  19.063  < 2e-16 ***
cast_total_facebook_likes  7.441e+03  7.235e+02  10.285  < 2e-16 ***
facenumber_in_poster      -1.153e+05  3.989e+05  -0.289  0.77254
num_user_for_reviews       3.291e+03  3.499e+03   0.941  0.34699
budget                     7.674e-01  2.380e-02  32.239  < 2e-16 ***
title_year                -7.002e+05  1.054e+05  -6.642 3.67e-11 ***
actor_2_facebook_likes    -7.474e+03  7.659e+02  -9.758  < 2e-16 ***
imdb_score                 1.600e+06  1.029e+06   1.554  0.12022
aspect_ratio              -6.289e+06  2.343e+06  -2.683  0.00733 **
movie_facebook_likes      -3.663e+01  5.724e+01  -0.640  0.52228
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45640000 on 2989 degrees of freedom
Multiple R-squared:  0.6113,    Adjusted R-squared:  0.6093
F-statistic: 313.3 on 15 and 2989 DF,  p-value: < 2.2e-16
```

Rest of the World Data :

```
Call:
lm(formula = gross ~ num_critic_for_reviews + duration + director_facebook_likes + 
    actor_3_facebook_likes + actor_1_facebook_likes + num_voted_users + 
    cast_total_facebook_likes + facenumber_in_poster + num_user_for_reviews + 
    budget + title_year + actor_2_facebook_likes + imdb_score + 
    aspect_ratio + movie_facebook_likes, data = movie.row)

Residuals:
       Min         1Q     Median         3Q        Max 
-146701597  -15417235   -3752454    7559870  298619477 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                1.657e+08  3.256e+08   0.509 0.611118    
num_critic_for_reviews     3.733e+04  1.943e+04   1.921 0.055036 .  
duration                   7.324e+04  6.411e+04   1.142 0.253636    
director_facebook_likes   -7.739e+02  1.059e+03  -0.731 0.465129    
actor_3_facebook_likes    -9.007e+03  3.694e+03  -2.438 0.014976 *  
actor_1_facebook_likes    -7.864e+03  2.105e+03  -3.735 0.000202 ***
num_voted_users            7.927e+01  2.318e+01   3.420 0.000659 ***
cast_total_facebook_likes  7.725e+03  2.068e+03   3.736 0.000201 ***
facenumber_in_poster      -2.082e+06  8.191e+05  -2.542 0.011203 *  
num_user_for_reviews       4.750e+04  6.324e+03   7.511 1.61e-13 ***
budget                     1.457e-05  2.832e-03   0.005 0.995896    
title_year                -7.258e+04  1.630e+05  -0.445 0.656195    
actor_2_facebook_likes    -5.954e+03  2.127e+03  -2.799 0.005245 ** 
imdb_score                -5.850e+06  1.576e+06  -3.711 0.000221 ***
aspect_ratio               3.278e+06  5.582e+06   0.587 0.557258    
movie_facebook_likes       2.506e+02  1.035e+02   2.421 0.015715 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38290000 on 780 degrees of freedom
Multiple R-squared:  0.4559,	Adjusted R-squared:  0.4454 
F-statistic: 43.57 on 15 and 780 DF,  p-value: < 2.2e-16
```

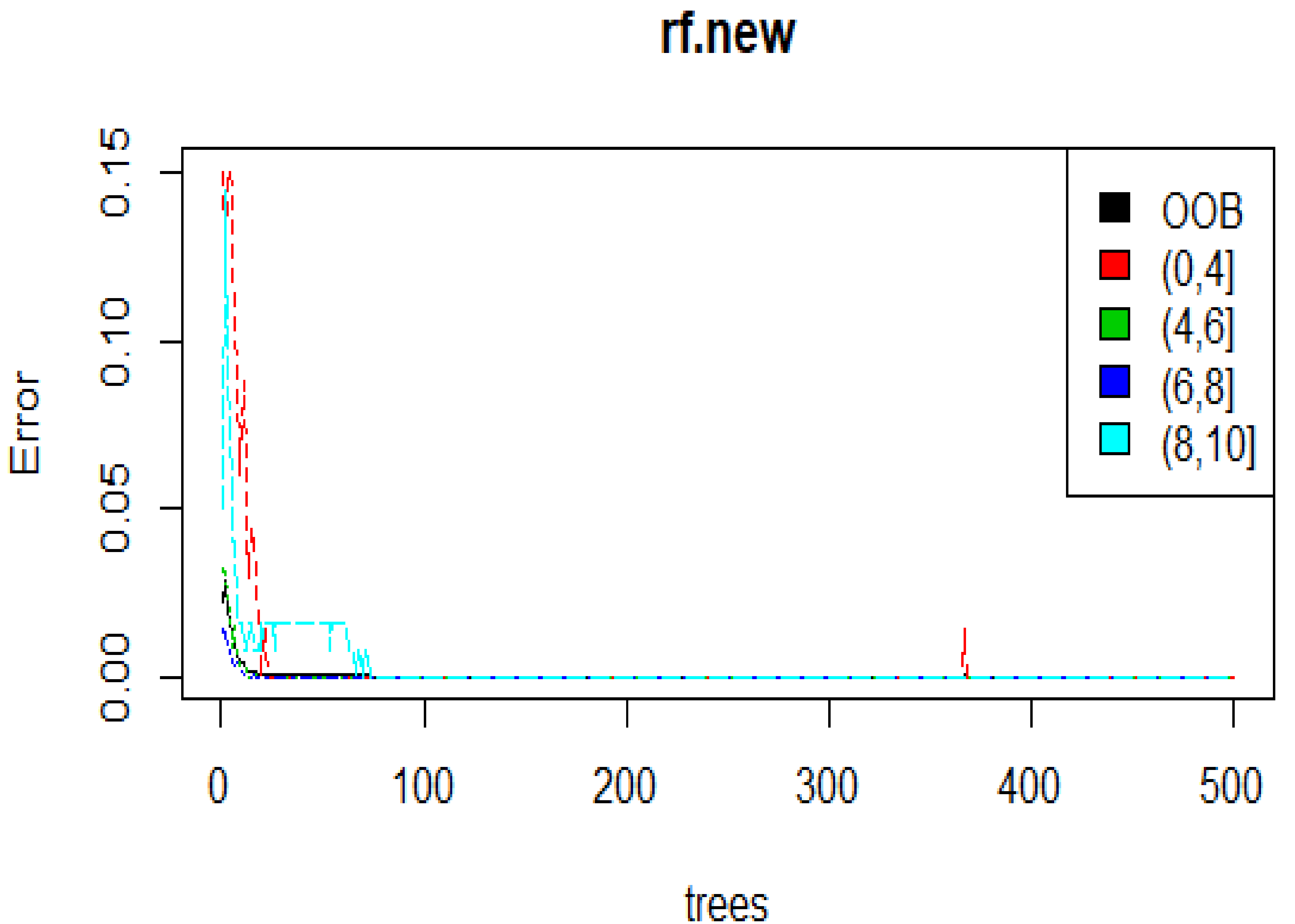
```
Call:
lm(formula = gross ~ num_critic_for_reviews + director_facebook_likes +
    actor_3_facebook_likes + actor_1_facebook_likes + num_voted_users +
    cast_total_facebook_likes + num_user_for_reviews + budget +
    actor_2_facebook_likes + imdb_score, data = train.data)

Residuals:
       Min         1Q     Median         3Q        Max
-426607038  -23476975   -8767344   13360676  467461200

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.531e+07  6.483e+06   5.447 5.52e-08 ***
num_critic_for_reviews     6.708e+04  9.721e+03   6.900 6.30e-12 ***
director_facebook_likes   -9.243e+02  3.168e+02  -2.917  0.00356 **
actor_3_facebook_likes    -1.331e+04  1.416e+03  -9.399  < 2e-16 ***
actor_1_facebook_likes    -1.109e+04  8.487e+02 -13.064  < 2e-16 ***
num_voted_users            2.057e+02  1.120e+01  18.375  < 2e-16 ***
cast_total_facebook_likes  1.104e+04  8.397e+02  13.144  < 2e-16 ***
num_user_for_reviews       2.244e+04  3.707e+03   6.054 1.58e-09 ***
budget                     1.067e-02  3.772e-03   2.829  0.00471 **
actor_2_facebook_likes    -1.048e+04  8.881e+02 -11.801  < 2e-16 ***
imdb_score                -5.123e+06  1.027e+06  -4.990 6.39e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51570000 on 3030 degrees of freedom
Multiple R-squared:  0.4742,	Adjusted R-squared:  0.4725
F-statistic: 273.3 on 10 and 3030 DF,  p-value: < 2.2e-16

>
```

```
Linear Regression

3041 samples
  10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2737, 2737, 2737, 2737, 2736, 2737, ...
Resampling results:

  RMSE      Rsquared   MAE
  53133922  0.4625299  33306027

Tuning parameter 'intercept' was held constant at a value of TRUE
>
> #test data performance for cross validation
> model.pred.cv <- predict(model.cross.valid, newdata = test.data)
>
> cat("\nThe Test MSE value for the cross validated model is :\n")

The Test MSE value for the cross validated model is :
> mean((model.pred.cv - test.data$gross)^2)
[1] 2.477372e+15
> cat("\nThe Test RMSE value for the cross validated model is :\n")

The Test RMSE value for the cross validated model is :
> sqrt(mean((model.pred.cv - test.data$gross)^2))
[1] 49773203
>
> cat("\nThe Cross Validated Model has a lower RMSE for Test Data set. This indica
tes that the model is a good one!")

The Cross Validated Model has a lower RMSE for Test Data set. This indicates that
the model is a good one!
>
```

RANDOM FOREST

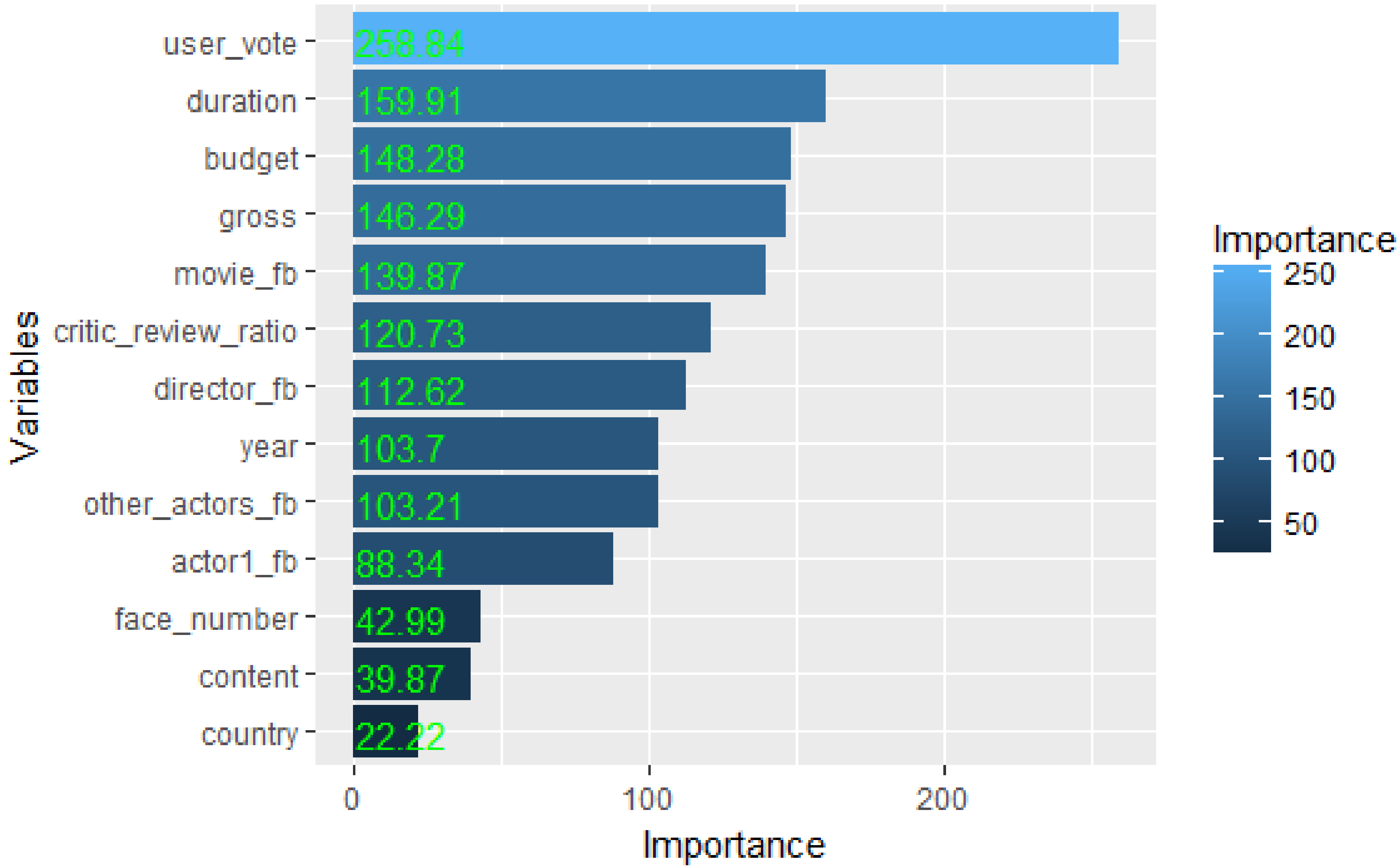
Created a new column- Movie_Quality where we divided the movies into 4 groups namely **BELOW AVERAGE, AVERAGE, GOOD and EXCELLENT** respectively based on their IMDB score.

```
IMDB$Movie_Quality <- cut(IMDB$imdb_score, breaks = c(0,4,6,8,10))

library(randomForest)
set.seed(53)
rf.new <- randomForest(Movie_Quality ~ . -imdb_score, data = train.new,
                       mtry = 5)
#Model Error Plot
plot(rf.new)
legend('topright', colnames(rf.new$err.rate), col=1:5, fill=1:5)
```



rf.new
Error
0.00
0.05
0.10
0.15
0
100
200
300
400
500
trees
OOB
(0,4]
(4,6]
(6,8]
(8,10]

IMPORTANT VARIABLES



CONCLUSION

- From Visualization : A good profitable movie and a good imdb score.
- From Genre analysis : DRAMA, COMEDY and Thriller are most used genres.
- From Country analysis : The highest number of movies are released in West Germany, USA and UK
- Identify significant predictors for gross revenue through regression.
- Based on the analysis of random forest, we found that the accuracy for test data set was 0.7454
- We are in the process of running KNN model to find the accuracy of KNN for test dataset and come-up with the better model based on test dataset.

Thank You