Movie Data Analysis

Illinois Institute of Technology, Chicago
by Shikha Verma, Ayushi Patel, Mayur Mehta and Abhilash Bhurse

---

Abstract – Although mainly viewed to be a source of entertainment, Movies are also a source of business. There is a great deal of capital associated with the movie business. Hence, understanding the factors that affect the success of a movie as well as discovering some interesting patterns based of past data would be of great help to derive maximum revenue and output from the movies that get produced. This report serves as a description of the results of a project carried out on the IMDB datasets from Kaggle website. Data obtained from the data source is processed and cleaned up to derive an operable dataset. Various types of visualizations such as graphs, bar-charts, heatmaps, etc are then derived through Exploratory Data Analysis, which help to deduce important information regarding the movies. Certain linear and non-linear models such as Linear Regression and Random Forests are then fit on the data to obtain valuable information regarding various factors that affect the gross revenue of the movies and the IMDB score of the movies. R serves as the language of programming and R-Studio serves as the platform.

Key words : IMDB Dataset, Exploratory Data Analysis, Linear Regression, Random Forests, R, R-studio

1. 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. 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. Knowing which movies are likely to succeed and which are likely to fail before their release could benefit the production houses greatly as it would enable them to better strategize their advertising campaigns, which themselves cost millions of dollars.

It could also help them to know when it is most appropriate to release a movie by looking at the overall market. Hence, the prediction of movie success is of great importance to the industry. Exploratory Data Analysis serves as a simple yet effective way to identify some interesting patterns from the data. Through this project we intend to 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. The goal of this project is to derive such insights which help in making an informed decision for the future generations of movies. To accomplish this task, we use a stepwise data

mining approach - data preprocessing, data integration and transformation, data exploration and visualization, model selection - and analyze and results.

2. Data Processing

The data we obtained is highly susceptible to noise, has missing values for many columns and is inconsistent due to its size and structure. So, we need to remove inconsistency and make it simple to better use it for our analysis. Once the data is suitably cleaned and integrated, it goes through selection and transformation activities, to translate the textual information (where necessary) into numerical information, which is better analyzed by data mining processes. This also discards irrelevant data, and selects a subset of the data to be mined, which is better suited to perform the analysis of our choice.
Below is a description of all the steps that were performed to process and clean the data :

2.1 Load Data from Kaggle:

IMDB <- read.csv ("C:/Users/User/Desktop/SEM III/DPA/project/movie_metadata.csv")

str(IMDB)

```
## '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",..: 927 801 2027 377 603 106 2030 1652
1228 551 ...
##  $ 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 2488 534 2
432 2549 1227 801 2439 653 ...
##  $ actor_1_facebook_likes    : int  1000 40000 11000 27000 131 640 24000 799 26000 25000 ...
##  $ gross                     : int  760505847 309404152 200074175 448130642 NA 73058679 336530303 200807262 458
991599 301956980 ...
##  $ genres                    : Factor w/ 914 levels "Action","Action|Adventure",..: 107 101 128 288 754 126 120
308 126 447 ...
##  $ actor_1_name              : Factor w/ 2098 levels "","50 Cent","A.J. Buckley",..: 302 979 353 1968 526 440 7
85 221 336 32 ...
##  $ movie_title               : Factor w/ 4917 levels "[Rec] ","[Rec] 2 ",..: 398 2731 3279 3708 3332 1961 3291
3459 399 1631 ...
##  $ 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",..: 3442 1392 3134 1769 1 2714
1969 2162 3018 2941 ...
##  $ 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",..: 1320 428
3 2076 3484 1 651 4745 29 1142 2005 ...
##  $ movie_imdb_link           : Factor w/ 4919 levels "http://www.imdb.com/title/tt0006864/?ref_=fn_tt_tt_1",..:
2965 2721 4533 3756 4918 2476 2526 2458 4546 2551 ...
##  $ 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 ...
```

Fig.1

As seen in Fig.1, we have 5043 observations of 28 variables. The response variable "imdb_score" is numerical, and the predictors are mixed with numerical and categorical variables.

2.2 Calculate and remove duplicate values :

We checked for duplicated values and removed them. As seen below, there were 45 duplicated rows and removal of them resulted in 4998 observations being left.

```
sum(duplicated(IMDB))
```

```
[1] 45
```

Fig.2

Remove duplicated values

```
IMDB = IMDB[!duplicated(IMDB), ]
```

```
dim(IMDB)
```

```
[1] 4998   28
```

Fig.3

2.3 Remove spurious values from movie title column :

Remove Spurious characters from movie_title

```
IMDB$movie_title <- gsub("?", "", as.character(factor(IMDB$movie_title)))
```

Fig.4

2.4 Check and remove NA values and deal with 0 values :

We checked for NA values and removed those unnecessary values and got the dimensions as shown below.

```
#Check for na values
```{r}
colSums(sapply(IMDB, is.na))
```
```

```
                    color              director_name    num_critic_for_reviews
                        0                          0                        49
                 duration    director_facebook_likes    actor_3_facebook_likes
                       15                        103                        23
             actor_2_name     actor_1_facebook_likes                     gross
                        0                          7                       874
                   genres               actor_1_name               movie_title
                        0                          0                         0
         num_voted_users  cast_total_facebook_likes              actor_3_name
                        0                          0                         0
      facenumber_in_poster              plot_keywords         movie_imdb_link
                       13                          0                         0
      num_user_for_reviews                   language                  country
                       21                          0                         0
            content_rating                     budget               title_year
                        0                        487                       107
    actor_2_facebook_likes                  imdb_score             aspect_ratio
                       13                          0                       327
       movie_facebook_likes
                        0
```

```
#Remove na values from gross, budget, aspect_ratio, title_year
```{r}
IMDB = IMDB[!is.na(IMDB$gross), ]
IMDB = IMDB[!is.na(IMDB$budget), ]
IMDB = IMDB[!is.na(IMDB$aspect_ratio), ]
IMDB = IMDB[!is.na(IMDB$title_year), ]
```
```

Fig.5

Similarly we checked for the ‘0’ values, and then we converted those ‘0’ values to ‘na’ and removed them.The conversion is displayed in figure below.

```
#Deal with '0' values
```{r}
mean_fnposter=mean(IMDB$facenumber_in_poster, na.rm = TRUE)
IMDB$facenumber_in_poster[is.na(IMDB$facenumber_in_poster)]=round(mean_fnposter)
```
#convert the 0 values to na
```{r}
IMDB[,c(5,6,8,13,24,26)][IMDB[,c(5,6,8,13,24,26)] == 0] <- NA
```
#impute(replace) missing values with column mean for specific columns

#for critic reviews
```{r}
IMDB$num_critic_for_reviews[is.na(IMDB$num_critic_for_reviews)] <- round(mean(IMDB$num_critic_for_reviews,
na.rm = TRUE))
```

#for duration
```{r}
IMDB$duration[is.na(IMDB$duration)] <- round(mean(IMDB$duration, na.rm = TRUE))
```
#for director_facebook like , actor 123 fb likes...
```{r}
IMDB$director_facebook_likes[is.na(IMDB$director_facebook_likes)] <-
round(mean(IMDB$director_facebook_likes, na.rm = TRUE))
IMDB$actor_3_facebook_likes[is.na(IMDB$actor_3_facebook_likes)] <- round(mean(IMDB$actor_3_facebook_likes,
na.rm = TRUE))
IMDB$actor_1_facebook_likes[is.na(IMDB$actor_1_facebook_likes)] <- round(mean(IMDB$actor_1_facebook_likes,
na.rm = TRUE))
IMDB$cast_total_facebook_likes[is.na(IMDB$cast_total_facebook_likes)] <-
round(mean(IMDB$cast_total_facebook_likes, na.rm = TRUE))
IMDB$actor_2_facebook_likes[is.na(IMDB$actor_2_facebook_likes)] <- round(mean(IMDB$actor_2_facebook_likes,
na.rm = TRUE))
IMDB$movie_facebook_likes[is.na(IMDB$movie_facebook_likes)] <- round(mean(IMDB$movie_facebook_likes, na.rm
= TRUE))
```

Fig.6

2.5 Delete Columns :

Unnecessary columns such as color, language are deleted as shown in figures Fig.7, and Fig.8.

delete column language

```
IMDB <- subset(IMDB, select = -c(language))
```

```
dim(IMDB)
```

```
[1] 3783   26
```

Fig.7

generate table for color

```
table(IMDB$color)
```

```
                Black and White            Color
              2             128             3653
```

delete predictor color

```
IMDB <- subset(IMDB, select = -c(color))
```

```
table(IMDB$language)
```

```
      Aboriginal      Arabic     Aramaic     Bosnian   Cantonese     Chinese      Czech      Danish
               2           2           1           1           1           8          0           1           3
            Dari       Dutch    Dzongkha     English    Filipino      French      German       Greek      Hebrew
               2           3           1        3608           1          36          13          0           2
           Hindi   Hungarian  Icelandic  Indonesian     Italian    Japanese     Kannada      Kazakh      Korean
               8           1           1           2           7          12          0           1           5
        Mandarin        Maya    Mongolian        None   Norwegian      Panjabi     Persian      Polish  Portuguese
              14           1           1           1           4          0           3           0           5
        Romanian     Russian    Slovenian     Spanish     Swahili     Swedish       Tamil      Telugu        Thai
               1           1           0          23           0           1           0           1           3
            Urdu  Vietnamese        Zulu
               0           1           1
```

Fig.8
```

2.6 Add Columns :

We added the variable column profit, where profit=gross-budget, as shown in below figure.

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
IMDB <- IMDB %>%
  mutate(profit = gross - budget,
         return_on_investment_perc = (profit/budget)*100)
```

```
table(IMDB$profit)
```

```
-12213298588  -4199788333  -2499804112  -2397701809  -2127109510  -1099560838   -989962610   -698312689
           1            1            1            1            1            1            1            1
  -696724557   -553005191   -399545745   -375868702   -299897945   -190641321   -188094481   -164334574
           1            1            1            1            1            1            1            1
  -149800772   -149237822   -143826840   -139853928   -136702695   -129828140   -128624673   -128620685
           1            1            1            1            1            1            1            1
```

Fig.9

3. Exploratory Data Analysis

Now that the data is processed and cleaned up, we are ready to perform Exploratory Data Analysis on the dataset to derive some interesting patterns and important information regarding the movies using data visualization techniques.
Data Visualization may be viewed as the process of extracting and visualizing the data in a very clear and understandable way without any form of reading or writing by displaying the results in the form of pie charts, bar graphs, statistical representation and through graphical forms as well.

The following section describes the results obtained after performing Exploratory Data Analysis on the cleaned dataset :

3.1 Top movies based on IMDB:

Here we just performed a general analysis and plotted a graph below which depicts the count of the movie relative to imdb score.From below plot, we inferred that our dataset comprises of maximum of number of movies having imdb score more than 3.5.

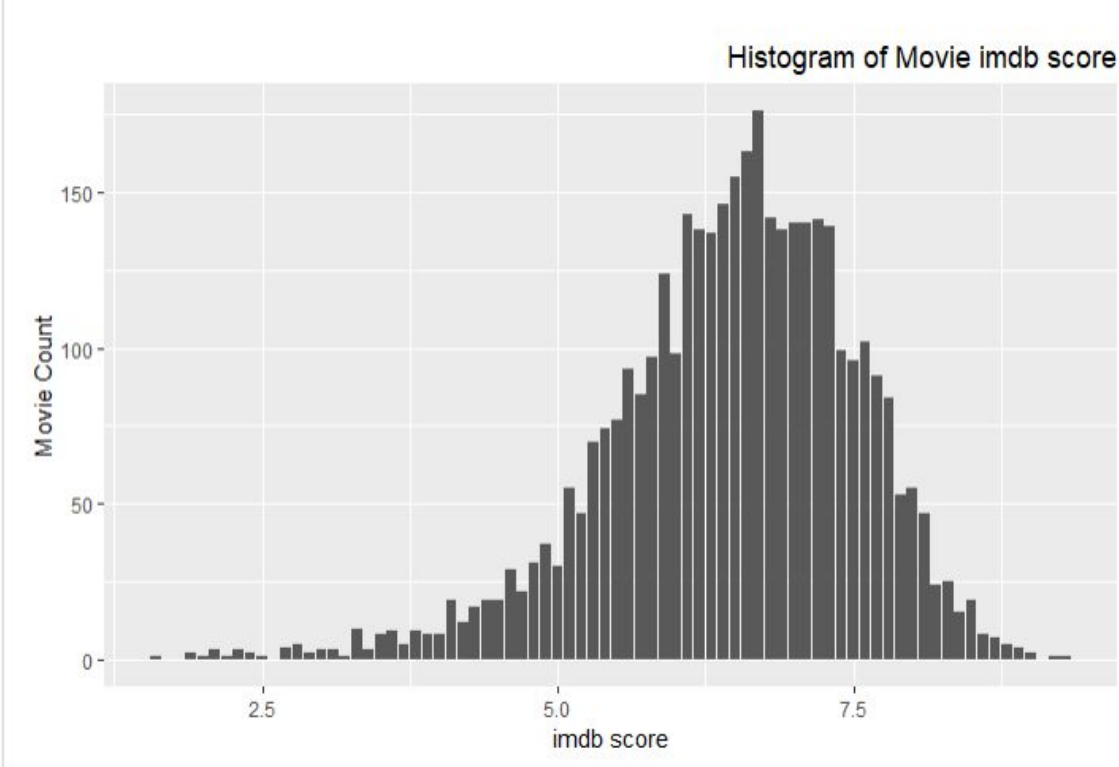


Fig.10

3.2 Genre Analysis :

Fig.11 shows a Word-Association Plot for the Movie Genres that are produced the most. We can see that Drama, Comedy and Thriller are the top 3 genres that are produced.

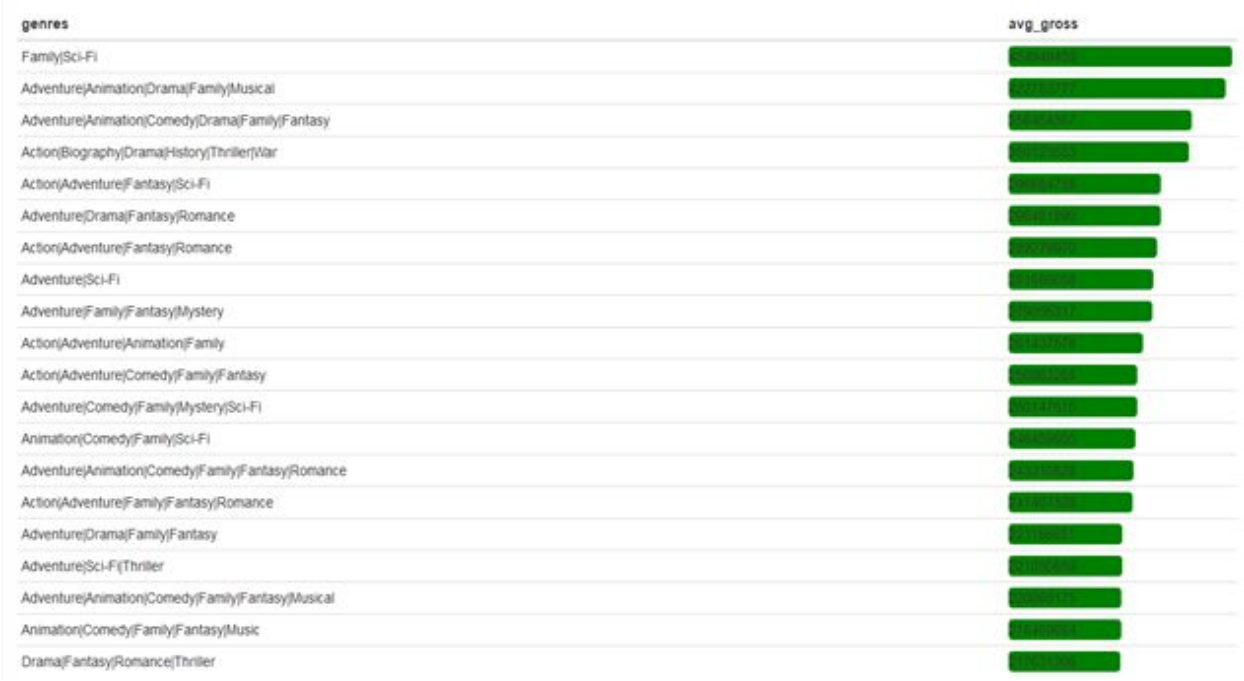
Fig.11

But are these genres the ones that have the highest IMDB score? Or the highest gross revenue? These questions are addressed by the plots of Fig.12 and Fig.13, which show the top genres with respect to IMDB score and Gross Revenue, respectively.

genres	avg_imdb
Adventure|Animation|Drama|Family|Musical	
Crime|Drama|Fantasy|Mystery	
Action|Adventure|Drama|Fantasy|War	
Adventure|Animation|Fantasy	
Adventure|Drama|Thriller|War	
Adventure|Animation|Comedy|Drama|Family|Fantasy	
Biography|Drama|History|Music	
Documentary|Drama|Sport	
Documentary|War	
Adventure|Drama|War	
Biography|Crime|Documentary|History	
Drama|Fantasy|War	
Drama|Mystery|War	
Action|Animation|Sci-Fi	
Adventure|Comedy|Crime|Drama	
Adventure|Drama|Thriller|Western	
Biography|Crime|Drama|History	
Biography|Crime|Drama|Western	
Documentary|History|Music	
Action|Adventure|Animation|Family	
Animation|Biography|Documentary|Drama|History|War	
Animation|Biography|Drama|War	
Biography|Drama|Family|Musical|Romance	
Crime|Documentary|Drama	
Crime|Drama|Musical	

Fig.12

As it can be seen here, a movie that belongs to Adventure, Animation, Drama, Family, Musical Genre has the highest IMDB score, whereas a Family, Sci-Fi movie has the highest Gross Revenue. We can also see that Drama, Comedy and Thriller Genres do feature in the top half of both the plots. Hence, we can infer that the genres that get the highest IMDB score and earn the highest revenues are the ones that get produced the most too.



genres	avg_gross
Family|Sci-Fi	
Adventure|Animation|Drama|Family|Musical	
Adventure|Animation|Comedy|Drama|Family|Fantasy	
Action|Biography|Drama|History|Thriller|War	
Action|Adventure|Fantasy|Sci-Fi	
Adventure|Drama|Fantasy|Romance	
Action|Adventure|Fantasy|Romance	
Adventure|Sci-Fi	
Adventure|Family|Fantasy|Mystery	
Action|Adventure|Animation|Family	
Action|Adventure|Comedy|Family|Fantasy	
Adventure|Comedy|Family|Mystery|Sci-Fi	
Animation|Comedy|Family|Sci-Fi	
Adventure|Animation|Comedy|Family|Fantasy|Romance	
Action|Adventure|Family|Fantasy|Romance	
Adventure|Drama|Family|Fantasy	
Adventure|Sci-Fi|Thriller	
Adventure|Animation|Comedy|Family|Fantasy|Musical	
Animation|Comedy|Family|Fantasy|Music	
Drama|Fantasy|Romance|Thriller	

Fig.13

Next, we determined the top genres, based on the profit earned. Fig.14 and Fig.15 show this analysis in the form of Histograms, for the movies released after year 2000, and the movies released before year 2000, respectively.

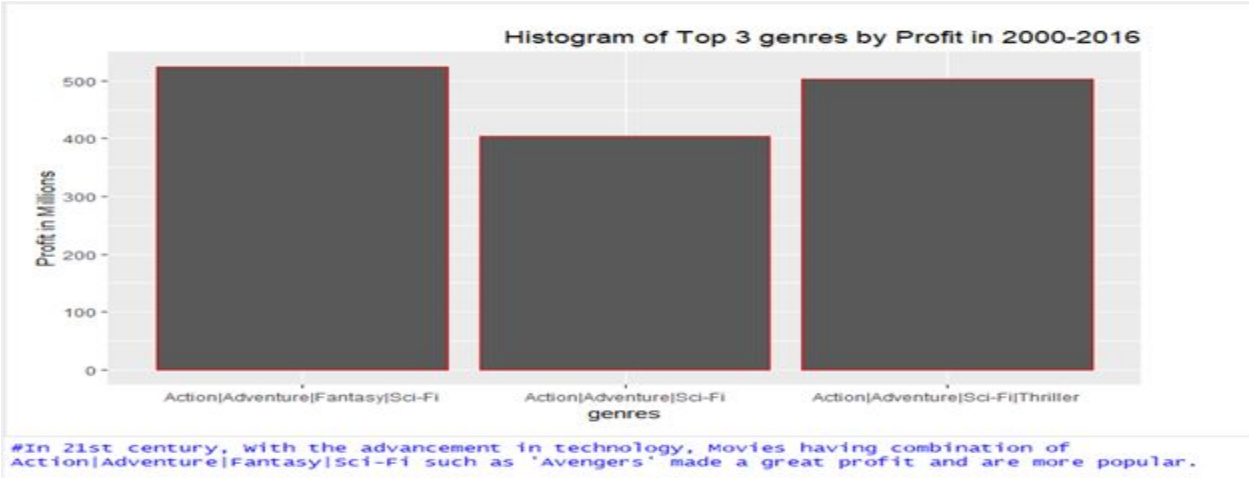
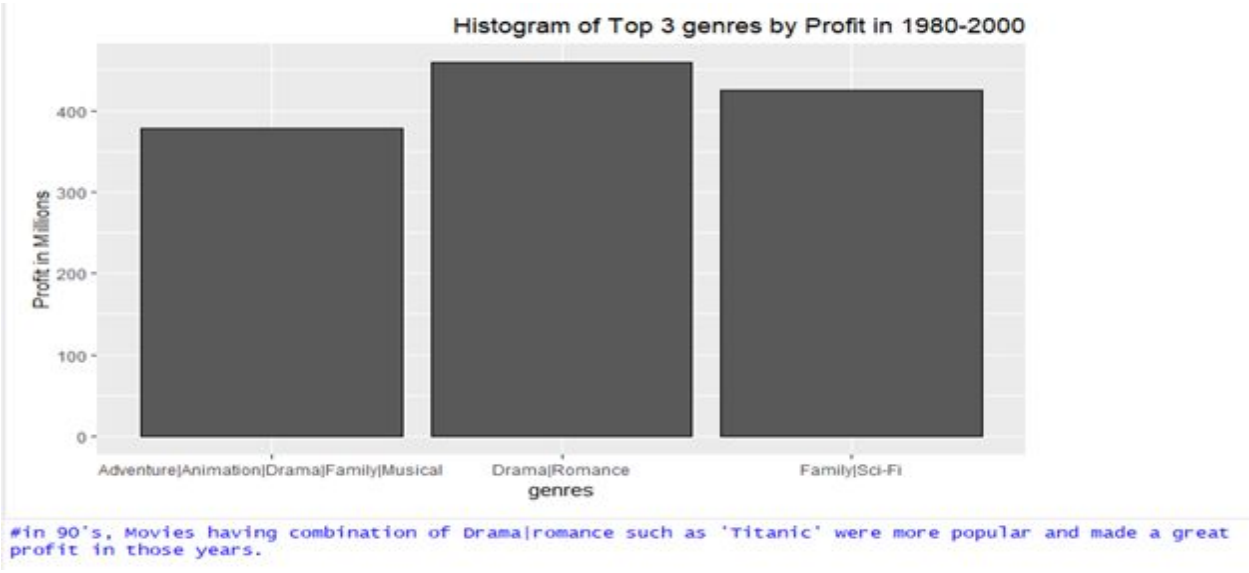


Fig.14




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

Fig.15

3.3 Country Analysis :

Fig.16 shows the budget of the movies in every country. We see that India has the highest amount of budget spent on movies, followed by China and New Zealand.

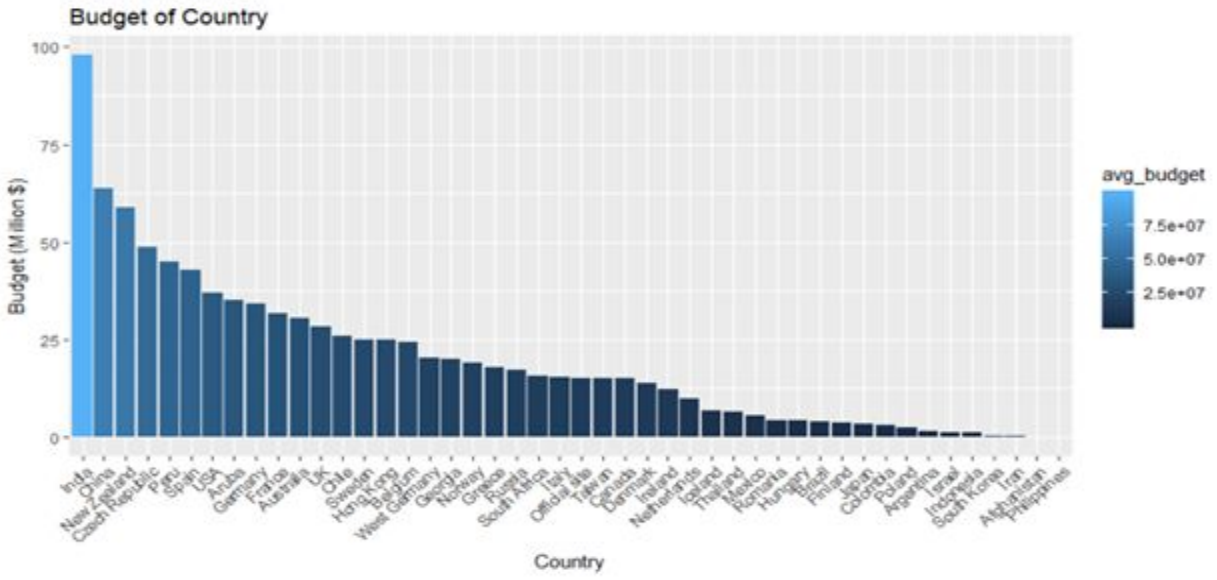


Fig.16

Fig.17 shows the Gross Revenue of the movies in every country. New Zealand leads the revenue charts, followed by Taiwan and Peru.

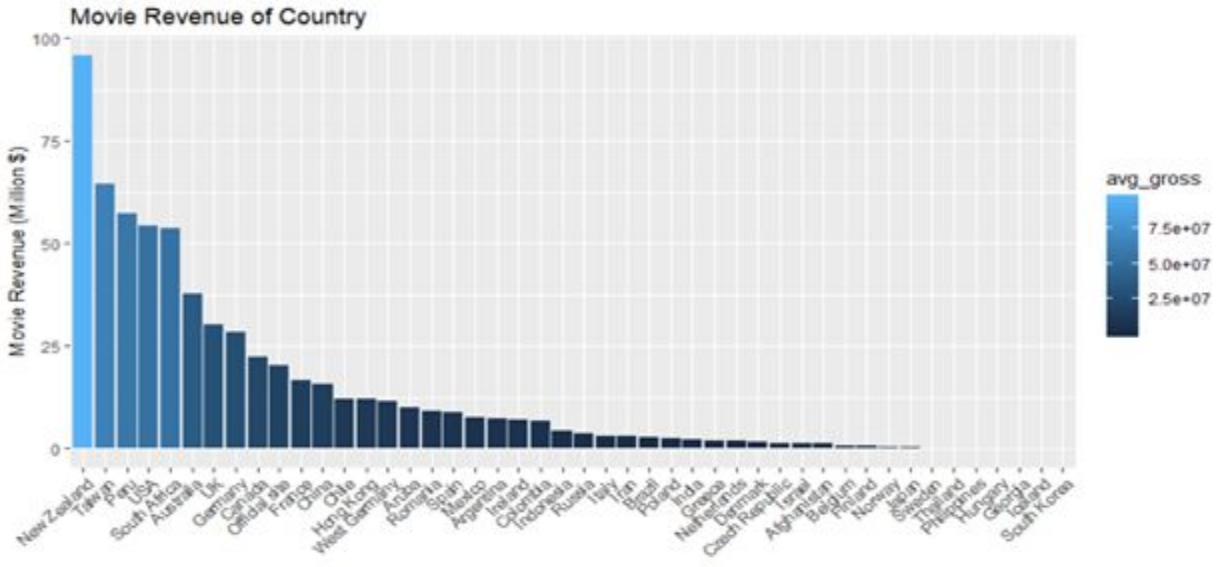
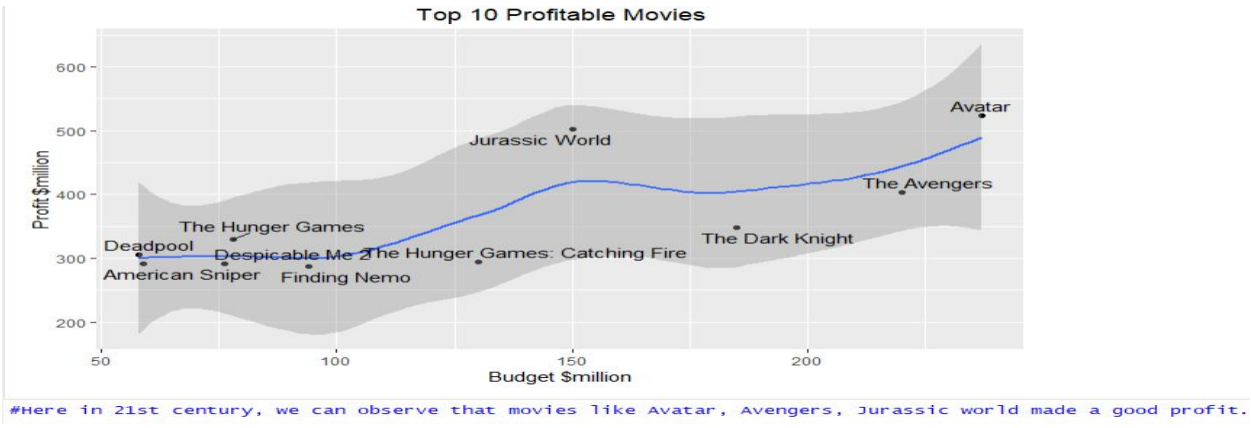


Fig.17

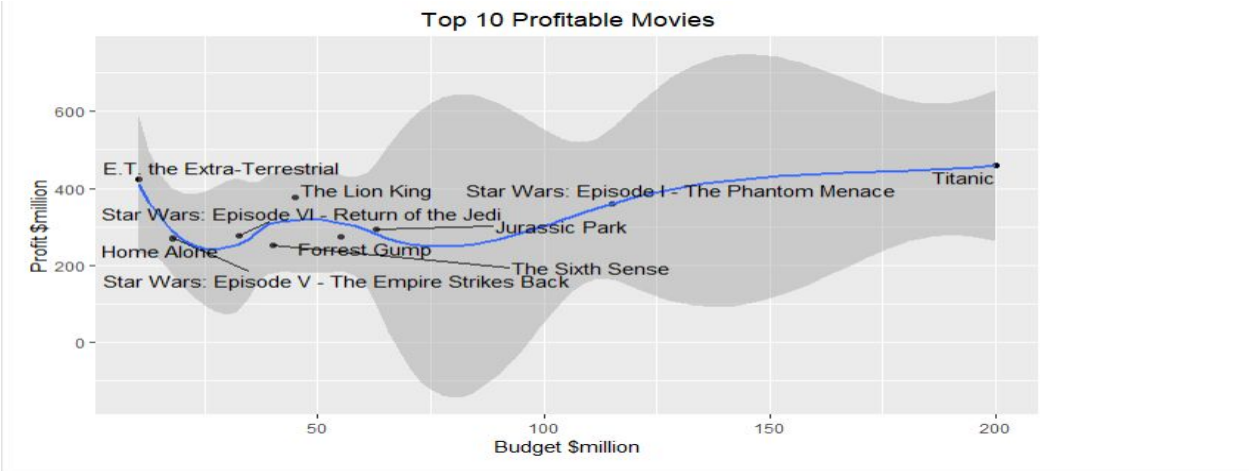
3.4 Profit Analysis :

From Fig.18, we can observe that 21st Century movies such as The Avengers, Avatar and Jurassic World, which are big budget movies, have earned the maximum profit. But a movie such as The Dark Knight which has a comparatively less budget, has earned a comparable profit. This would lead us to conclude that The Dark Knight was a successful movie.




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

Fig. 18




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

Fig.19

Similarly, Fig.19 shows that for the movies released before 2000, Titanic is the movie that has earned the maximum profit, but also had a large budget. Movies such as E.T : the Extra-Terrestrial, The Lion King, and the Star Wars movies have earned a larger profit with a smaller budget, making them successful movies.

Comparing the profits of the movies that were released before 2000, i.e, between 1980-2000 and those released after 2000, i.e, 2000-2016, gives an interesting analysis. As seen in Fig.20, movies released in the 21st Century earn more profit than those released before 2000 used to. This highlights the growing value of the movie business over the years.

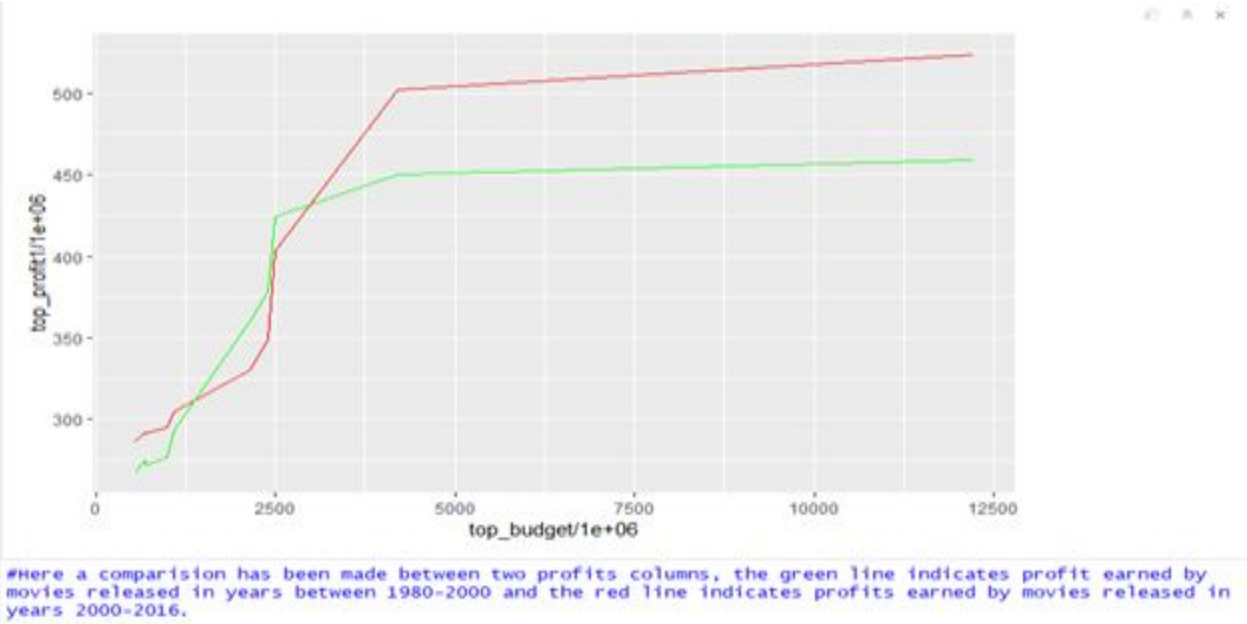


Fig.20

Looking at this trend, we can expect the movies to keep earning higher profits as years go by.

4. Model Selection

The following section describes the results of the various models fitted over the data and the inferences that were derived out them. Simple and Multiple Linear Regression was used as a linear model, whereas Random Forests served as the non-linear model.

4.1 Simple Linear Regression:

Throughout the dataset, IMDB score and Gross Revenue serve as the 2 most influential and important features for the movies. Hence it is important to understand the relationship among these features. More specifically, we intend to discover what is effect of the IMDB score on the gross revenue of a movie.

To answer this question, we fitted a Simple Linear Regression Model on the data, with Gross Revenue serving as the Target variable and IMDB score serving as the single predictor.

As seen in Fig.21, IMDB score is indeed a significant predictor for the gross revenue of the movie. But we observe that the R-Squared value of the fitted value is very small - 0.045, indicating that the model explains only 4% of the variability.

```
> #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

> |
```

Fig. 21

To determine exactly how important IMDB score is significant to gross revenue, we calculated the correlation between these variables. The correlation is 0.21 which is not quite good, thus this indicates only a good IMDB score does not guarantee a good gross revenue, there might be other predictors too.

```
> #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!!
>
```

Fig. 22

4.2 Multiple Linear Regression :

To determine which predictors are important to Gross Revenue, we fit a multiple regression model, with gross revenue as the response variable and all other numeric variables as predictors.

```
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
```

Fig. 23

As seen in Fig.23, we get a total of 10 predictors as significant to gross revenue. These predictors are: num_critic_for_reviews, director_facebook_likes, actor_1_facebook_likes,

actor_2_facebook_likes, actor_3_facebook_likes, cast_total_facebook_likes, num_voted_users, num_user_for_reviews, budget and imdb_score.

The dataset is dominated by movies produced in the USA. Hence, we divided the dataset based on movies produced in USA and movies produced in the Rest of the World. Since the dataset is dominated by movies produced in the USA, we get similar results for the USA data. The only difference being that imdb_score is not a significant predictor.

```
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
```

Fig. 24

However, movies produced outside the USA show different results. We only get 5 predictors as being significant. This indicates that different countries have different factors that affect the gross revenue of a movie.

```
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)

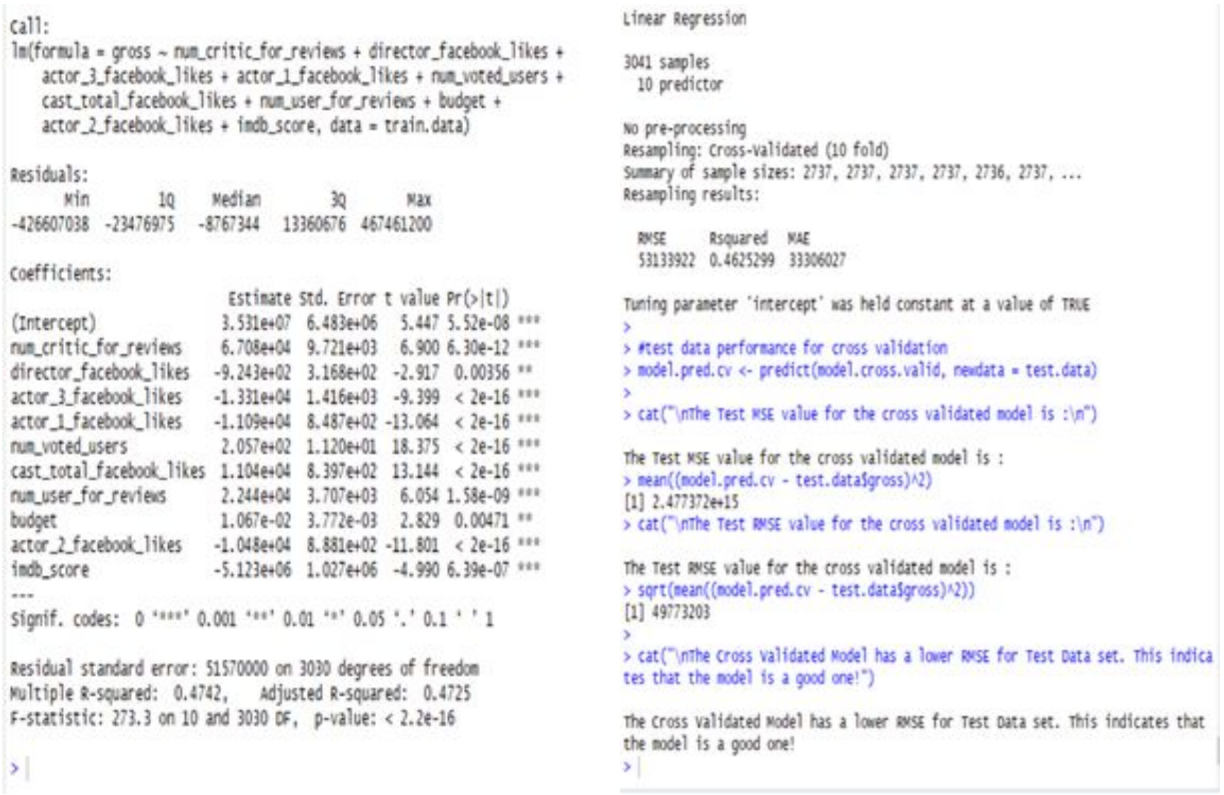
Residuals:
      Min        1Q    Median        3Q       Max
-146701597  -15417235   -3752454    7559870  298619477

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
```

Fig. 25

Next we fit a regression model on the 10 significant predictors that we got on a training data, performed 10-fold Cross Validation, and predicted the response on the test data. We found that the model gives a lower test RMSE than the train MSE. This indicates that the model is a good one.



```
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!
>
```

Fig.26

4.3 Random Forest :

We chose Random Forests as the non-linear model and tried to see whether the model predicts the movie performance, based on the IMDB score.

For this purpose we 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.

We then built a random forests model on this new column and plotted the error rate graph. The results of the graph show that the model is effective in predicting all types of movies with a very less error. For upto 100 trees the error is about 0.15, which is quite a good value. We see that as the number of trees increase this error decreases and becomes stable.

```
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)
```

Fig. 27

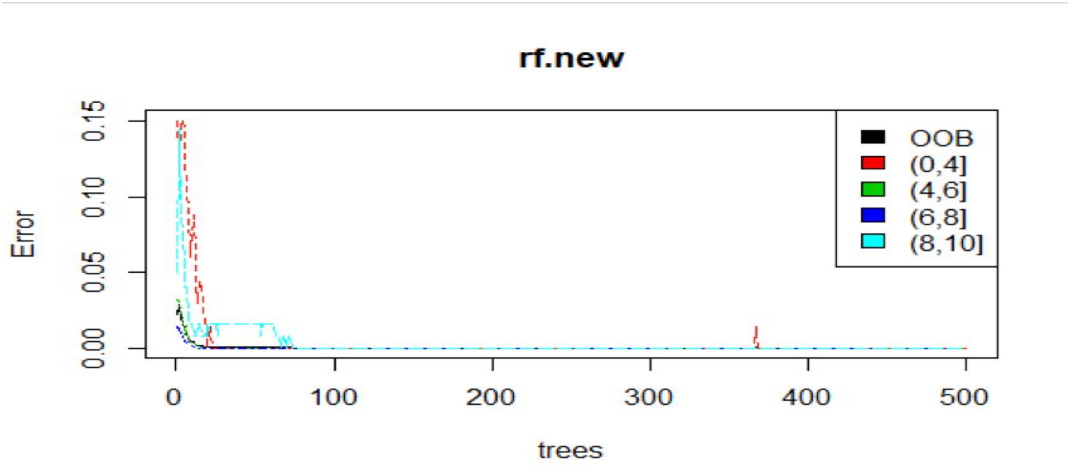


Fig. 28

We then generated a plot for the important variables based on Mean decrease in Gini. We found that for IMDB score, the user_vote, i.e, the number of users who vote or rate a movie is the most important variable, followed by the duration of the movie and budget. Country and content rating are the least important variables.

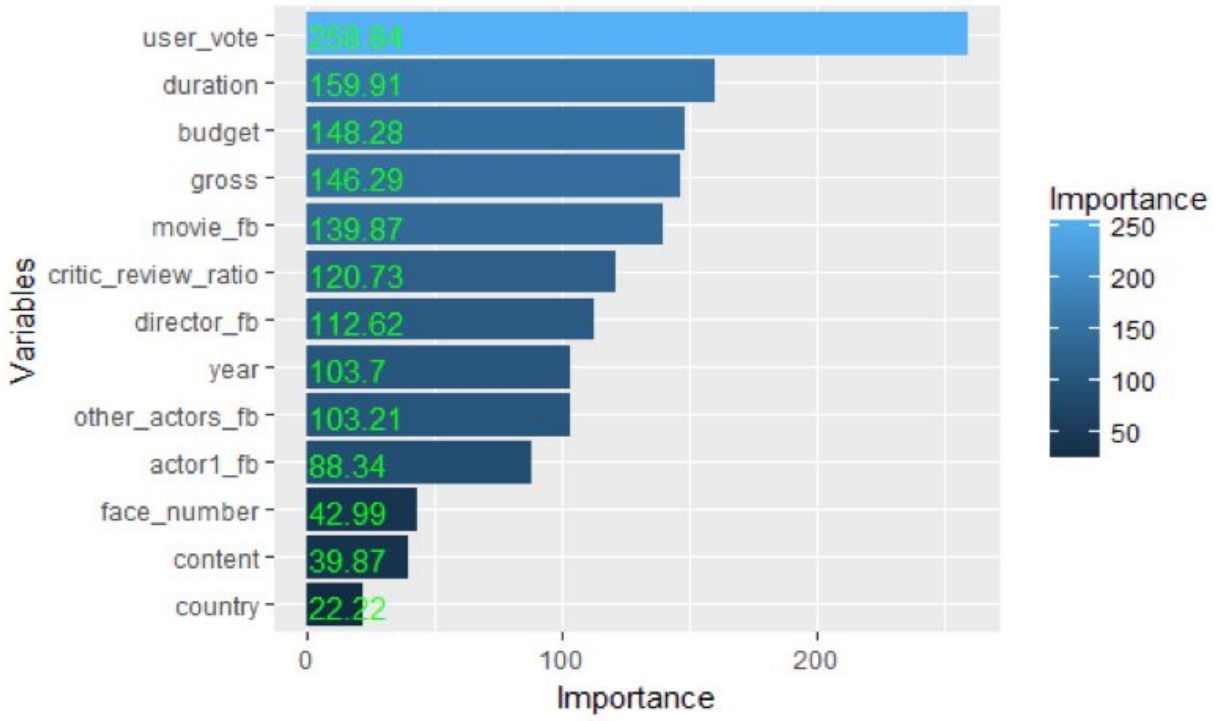


Fig. 29

We ran the model for a test dataset and found that the test accuracy for the Random Forests model was 74.54%.

This concluded the model selection phase of the project.

Conclusion and Future Work

Through this project we were able to derive many valuable insights into the movie dataset. Analysis was based on Genres, Countries and Profitability. This was depicted in the form of different graphs. We were also able to fit regression models on the dataset and determine all the factors that affect the movie revenue in different countries. We found that different factors affect the movie revenue in different parts of the world. We also fit a random forests model to determine whether the model could predict the movie performance. We found that the model predicts the movie performance with a very less error.

Since big budget movies tend to skew the results, it would be more appropriate to separate the big budget movies and the lower budget movies and then fit a regression model and analyze the results. We keep this as future work for the project. We also intend to run a KNN model on the dataset and find the test accuracy of the KNN model. Based on this we would then be able to determine which model has the better performance for our dataset.

Data Sources

The dataset is from Kaggle website. It contains 28 variables for 5043 movies, spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses. “imdb_score” is the response variable while the other 27 variables are possible predictors. Link for the original data is given.

Link : https://data.world/data-society/imdb-5000-movie-dataset

Source Code

Source code is attached in zip file.

Bibliography

[1] https://data.world/data-society/imdb-5000-movie-dataset

[2] https://www.kaggle.com/arillo03/imdb-movie-dataset-analysis

[3] https://www.cyclismo.org/tutorial/R/

[4] https://minimaxir.com/2018/07/imdb-data-analysis/

[5] https://www.tutorialspoint.com/r/index.htm

[6] https://www.kaggle.com/nandys/analysis-of-imdb-dataset

[7] https://pdfs.semanticscholar.org/d75f/1d075cab4c0d77754c1f7ca0fe4f3a998028.pdf

[8] https://pdfs.semanticscholar.org/d75f/1d075cab4c0d77754c1f7ca0fe4f3a998028.pdf