**BUAN 6356**

**Project Report**

**Group 2**

Predicting IMDb Score of Movies

**Submitted by:**

**Abhinav Saluja**

**Richa Kewalramani**

**Ritish Gupta**

**Somya Pandey**

**Xinchao Zhang**

# Table of Contents

[**Table of Contents**](#_5ekmrs9vi5gy) **2**

[**1. Introduction**](#_t1p6xuqo8h6z) **4**

[1.1 Background](#_j7ell4f7vdvn) 4

[1.2 IMDb, Inc.](#_qcth6eh47i6n) 4

[1.3 Business Application or Objectives](#_1hmw33yjn4q1) 4

[**2. Introduction to the Dataset**](#_6xdxosdf32op) **5**

[2.1 Searching for the Dataset and Finalization](#_vsbbbrsrgvt3) 5

[2.2 Dataset Variables and Description](#_d9aqirsogzsx) 5

[**3. Project Approach**](#_jswpparvyf1i) **7**

[3.1 Data Cleaning](#_jpjscv7pdenr) 7

[3.2 Visualizations to be done](#_dlnaohkc69gg) 7

[3.3 Prediction Models to be used](#_aaxb5bvp3zm3) 7

[3.4 Observations and Prediction Results](#_rbxqux2za9hl) 7

[**4. Data Exploration and Cleaning**](#_yvykleaa5396) **8**

[4.1 Loading Data and its Structure](#_9ly4r7hw22bw) 8

[4.2 Checking for Duplication in Data](#_1ojty79if0jt) 9

[4.3 Missing Values in Data](#_e9lwqhuyu7x3) 10

[4.3.1 Delete Some Rows in Data](#_gswc93nb6nwa) 11

[4.3.2 Dealing with Zero Values](#_6t4ril8q0ky3) 12

[**5. Data Visualization**](#_95jm6sl7z568) **13**

[5.1 Libraries used for visualisation.](#_nvf5ugf5fcs1) 13

[5.2 Histogram for frequency of the IMDB score](#_6p4jio9h5s9w) 13

[5.3 Range in years of movies](#_8r43esm36bmr) 15

[5.4 Movies with IMDB score greater than and equal to 8](#_knew3tkp2mir) 15

[5.5 Plotting Correlation between IMDb Score and other Variables using Scatter Plots](#_ej3rcf3itw6v) 16

[5.5.1 Correlation between duration and IMDb Score](#_j1r5jrngdzjp) 16

[5.5.2 Correlation between Number of Critic Reviews and IMDb Score](#_roupnuul5g0s) 16

[5.5.3 Correlation between Face Number in Poster and IMDb Score.](#_of320t8of2t) 17

[5.5.4 Correlation between Budget and IMDb Score](#_jz4sc2edukdg) 18

[5.5.5 Correlation between Number of Users who Voted and IMDb Score](#_a272fogblncq) 19

[5.6 Plotting Correlation between Different Variables](#_jr3pdnbjh5n3) 20

[5.7 Word Cloud](#_6peteg5xhzhb) 21

[5.7.1 Word Cloud of Director Name](#_9kvi91gomhrq) 21

[5.7.2 Word cloud of Protagonist Name in the Movies](#_s8dy57i49hvx) 22

[5.7.3 Word Cloud of Country of the Movies](#_sgz15i1i3plq) 22

[5.8 Movies that have the Most Likes on Facebook](#_yt6hdgsosdft) 23

[5.9 Density of Different Main Variables of the Movies](#_k9iyvf17nro9) 24

[5.10 Plotting Correlation between Different Variables](#_xi01i0dqy8gb) 25

[5.11 Plotting Correlation between Budget and Likes](#_sxa4zizgyduw) 26

[**6. Multiple Regressions Analysis**](#_1u9ubpduo95) **29**

[6.1 Introduction](#_qrmunin0wa04) 29

[6.2 Assumptions](#_qr1bwlkv3xb1) 29

[6.3 Uses](#_yy239pp76kuz) 29

[6.4 Procedure](#_2z26bs80ey84) 30

[6.5 Results](#_yd4hvxbsnu7c) 36

[6.6 Conclusion](#_vscoibax9h3s) 36

[6.6.1. Signs of the Coefficients](#_ncle2rybhtfo) 36

[6.6.2. Significance Level of the Variables](#_2w91rraa64dn) 37

[6.6.3. Efficiency and Accuracy of the Model](#_azkh43r8jzld) 37

[6.6.4. Outliers in the Model](#_uupbw3rs1cpy) 37

[6.6.5. AIC of the model](#_phl5kk8fulxt) 37

[**7. Classification and Regression Tree (CART)**](#_jisslw16tact) **38**

[7.1 Introduction](#_vz8nxp2jxk27) 38

[7.2 Decision tree types](#_n1ng22a7lri) 38

[7.3 The Algorithm behind Decision Trees](#_retfw54p4e00) 39

[7.4 Advantages of Trees](#_zae5lwhfer5p) 39

[7.5 Limitations](#_3xb5pfumka2) 39

[7.6 Algorithm Steps](#_hdcg2qstozoj) and Accuracy Evaluation 39

[**8. Random Forest**](#_9l922c7jfh) **48**

[8.1 Introduction](#_qdjzibcs6d30) 48

[8.2 Algorithm Steps](#_fctd1yyz7vvx) and Accuracy Evaluation 48

**9**[**. Conclusion**](#_83r77dqlv3x4) **50**

**10**[**. Citations**](#_xqi8rjtd7yz8) **51**

# 

# 

# 1. Introduction

## 1.1 Background

Today’s movie watcher uses IMDb ratings as the reference for deciding whether a movie would be worth watching or not. While pre-release marketing can give a movie a huge opening, the collections can significantly drop after a couple of days based on how the audience is perceiving the movie. In this day and age where internet can make or break anything, IMDb ratings can be pivotal to a movie’s failure or success.

## 1.2 IMDb, Inc.

Internet Movie Database (IMDb) originates back to 1990 in UK as a fan-operated website for rating movies and television. At the time, it was known as "rec.arts.movies movie database". The database moved to the world wide web later in 1993 and during the time more features like biographies, trivia, demographics were added to the repository. The company was incorporated in UK in 1996 as Internet Movie Database Ltd with Col Needham as the CEO of the company.

In 1998, Jeff Bezos, Founder and CEO of Amazon, bought IMDb for approximately $55 million and attached it to Amazon as a subsidiary. Since then, IMDb keeps on expanding on features including the launch of a paid service called IMDbPro with services including film production and box office details, as well as a company directory and the ability of subscribers to add personal information pages with details at variance with pages about them appearing in the database.

IMDb has grown into a global phenomenon with titles listed from all countries producing movie, television and digital content like USA, India, Nigeria, Japan, etc.

## 1.3 Business Application or Objectives

We will use the variables to analyze how important they can be to predict the movie’s IMDb rating. This prediction will help production houses and filmmakers to see what the recipe is for getting higher ratings and where they should invest for these higher ratings to be successful at the box office and to build a brand for their respective production houses.

# 2. Introduction to the Dataset

## 2.1 Searching for the Dataset and Finalization

When looking for datasets, we came across many datasets related to IMDb, Rotten Tomatoes, movie reviews, etc and finally decided upon ‘IMDB 5000 Movie Dataset’ from dataworld.com. This dataset has 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 target variable while the other 27 variables are possible predictor variables.

The major reason for going ahead with this dataset was the fact that it takes many factors into consideration which we would not conventionally think can impact the possible movie IMDb rating. These include but are not limited to the actors’ and director’s popularities on facebook, number of keywords that describe the movie plot, aspect ratio of the movie, duration of the movie, country of production and so on. We thought it would be interesting to visualize and analyze the impact of these unconventional factors in addition to the conventional ones to predict the possible outcome of the movie rating.

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

## 2.2 Dataset Variables and Description

|  |  |
| --- | --- |
| Variable Name | Description |
| movie\_title | Title of the Movie |
| duration | Duration of the movie in minutes |
| director\_name | Name of the Director of the Movie |
| director\_facebook\_likes | Director's Facebook Page likes |
| actor\_1\_name | Primary actor starring in the movie |
| actor\_1\_facebook\_likes | Actor\_1's Facebook Page likes |
| actor\_2\_name | Name of the second lead |
| actor\_2\_facebook\_likes | Actor\_2's Facebook Page likes |
| actor\_3\_name | Name of the third lead |
| actor\_3\_facebook\_likes | Actor\_3's Facebook Page likes |
| num\_critic\_for\_reviews | Number of reviews from critics on imdb |
| num\_voted\_users | Number of people who rated the movie |
| cast\_total\_facebook\_likes | Total facebook likes of the entire cast of the movie combined |
| movie\_facebook\_likes | Number of likes on the movie's Facebook page |
| plot\_keywords | Keywords describing the movie plot |
| facenumber\_in\_poster | Number of the actors whose faces feature on the movie poster |
| color | Film colorization. ‘Black and White’ or ‘Color’ |
| genres | Film categorization like ‘Animation’, ‘Comedy’, ‘Romance’, ‘Horror’, ‘Sci-Fi’, ‘Action’, ‘Family’ |
| title\_year | The year in which the movie is released (1916 - 2016) |
| language | Primary language of the movie viz. English, Arabic, Chinese, French, German, Danish, Italian, Japanese, etc. |
| country | Country where the movie is produced |
| content\_rating | Content rating of the movie viz. R, G, PG, PG-13 |
| aspect\_ratio | Aspect ratio the movie was made in viz 16:9, 4:3, etc. |
| movie\_imdb\_link | IMDB link of the movie |
| gross | Gross earnings of the movie in US Dollars |
| budget | Budget of the movie in US Dollars |
| imdb\_score | IMDB rating of the movie |

# 3. Project Approach

During the initial brainstorming sessions of this project, we collectively decided that we would not just focus on one prediction model to conclude our analysis of this dataset. Rather, we would be performing many visualizations and using multiple prediction methods to arrive to our conclusions. This will not only give us a better understanding of the dataset, but also help us in validating the predictions we make from the other models.

## 3.1 Data Cleaning

In this section, we have included preprocessing and basic exploration of the dataset.

## 3.2 Visualizations to be done

* Scatterplot
* Heat map
* Histogram
* Correlation matrix
* Word cloud

## 3.3 Prediction Models to be used

* Linear Regression
* CART Model
* Random Forest

We have not developed a Logistic Regression model because the target variable: imdb\_score is a continuous numerical variable, and Logistic Regression performs well for categorical variables.

## 3.4 Observations and Prediction Results

We will combine the results from the predictions above to arrive to conclusions regarding the dataset.

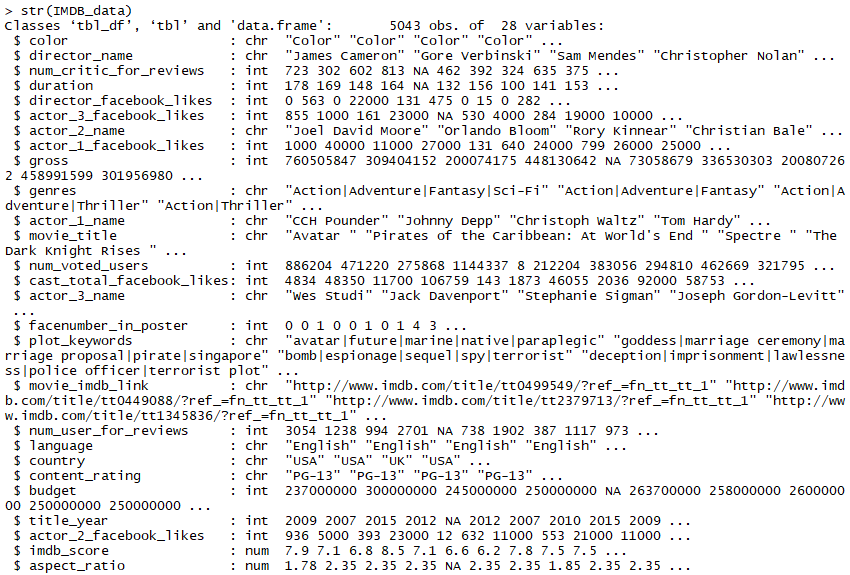
# 4. Data Exploration and Cleaning

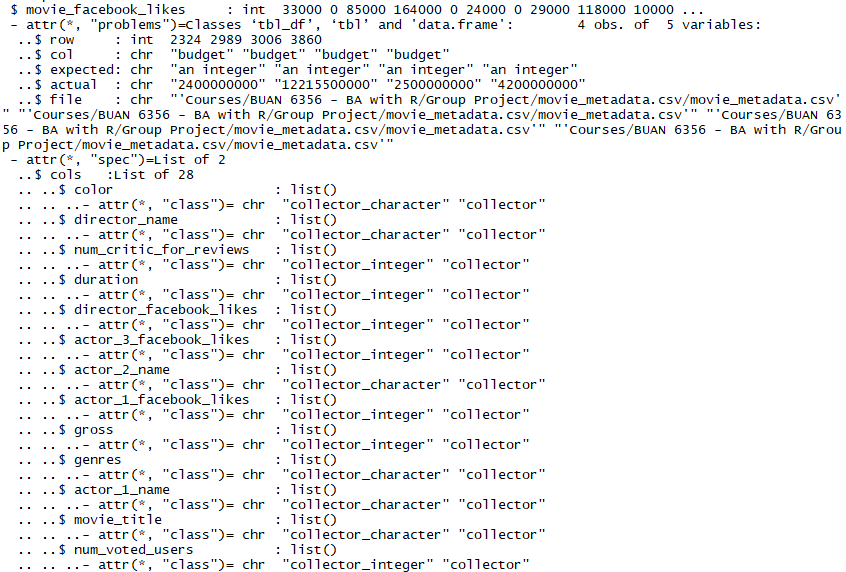
Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a dataset. It is an approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data.

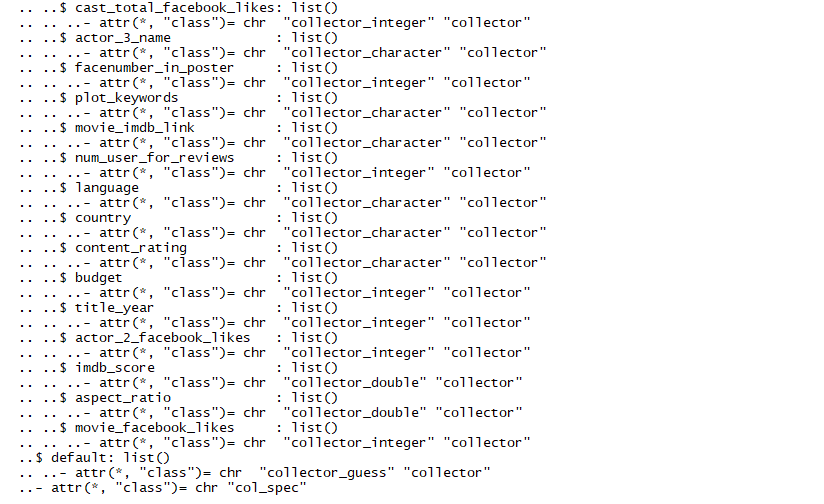
Data cleaning or Data cleansing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the dataand then replacing, modifying, or deleting the dirty or coarse data.

## 4.1 Loading Data and its Structure









## 4.2 Checking for Duplication in Data

Finding the number of duplicate rows in the data:



This returned the following result:



Now deleting the duplicate rows from the dataset:



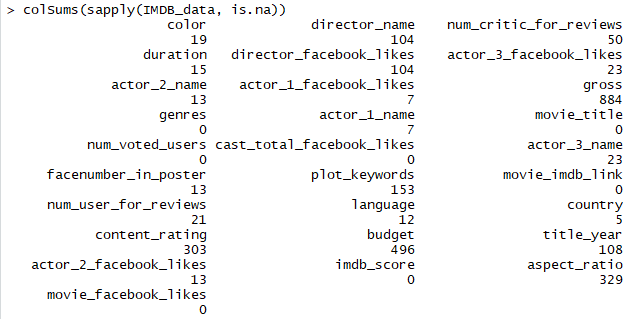
We get 4998 rows after removing the duplicates.

## 4.3 Missing Values in Data

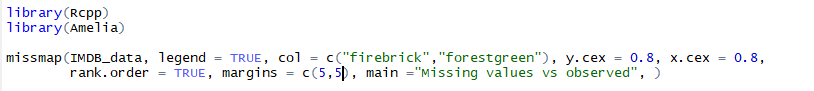
Looking for missing values in the data -

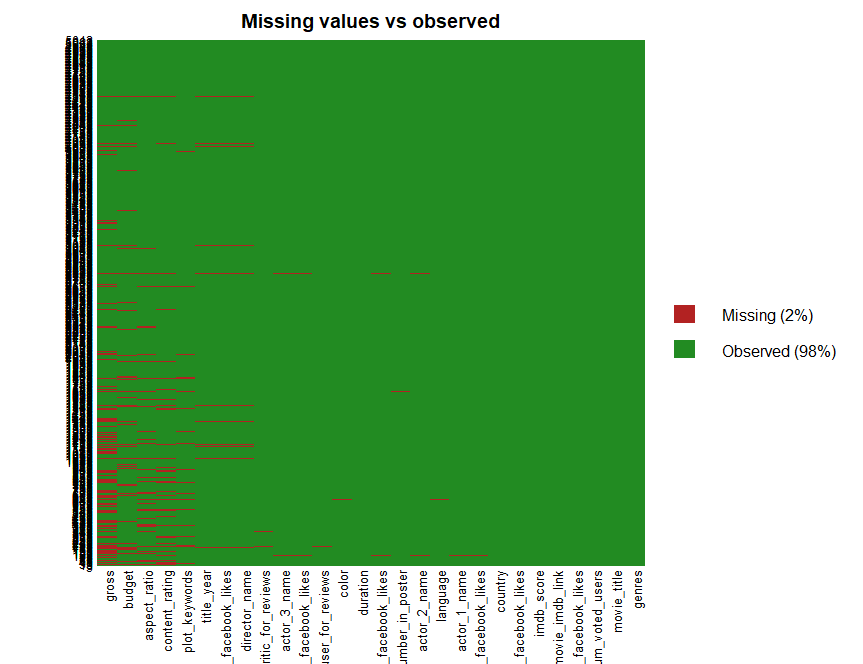


The result:



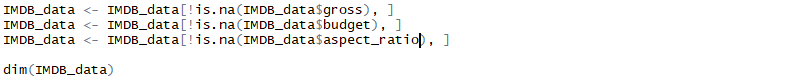
This does show us that some of the predictors have missing values. Now let’s visualize them using a missmap and rank them from highest to lowest missing values. For this we will require libraries Amelia and Rcpp.





### 4.3.1 Delete Some Rows in Data

The graph shows us that there are 2% missing values and maximum number of missing values in are in predictors ‘gross’, ‘budget’ and ‘apect\_ratio’. So, we remove the observations with null values in these three predictors.



After omitting the missing values, we get:



With 3812 observations, the data should now be cleaned with very low number of missing values.

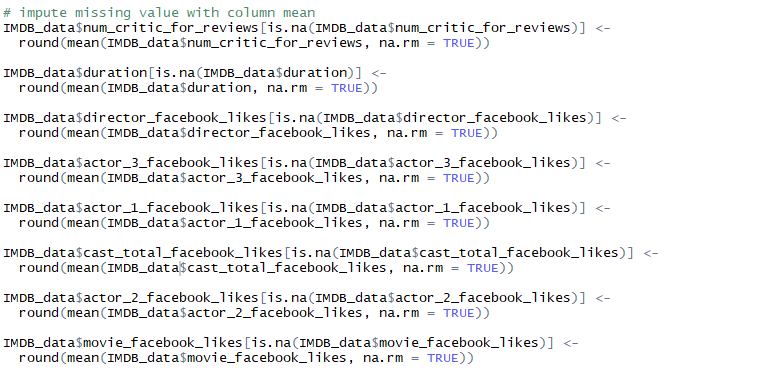
### 

### 4.3.2 Dealing with Zero Values

There are some 0 values which should also be regarded as missing values except for predictor facenumber\_in\_poster. First, we will NA with column average for facenumber\_in\_poster. Then replace 0s in other predictors with NA, and lastly replace all NAs with their respective column mean.







The data is now ready to perform the visualizations and implement prediction algorithms.

# 5. Data Visualization

One of the most important beneﬁts of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Well designed data graphics are usually the simplest and at the same time, the most powerful. What we once saw as dull and mundane, can now be seen as beautiful. Data visualization tools make it quick and easy to create charts and graphs, which can be added to a customizable dashboard. Besides looking beautiful, data visualization tools give us the ability to process information faster and to use that information to boost productivity and results.

Using data visualization communicates information faster than traditional reports. A more concise, more refined language, visualized data can be easily sorted to show a quick overview of whatever information is needed.

## 5.1 Libraries used for visualisation.

library(forecast)

library(tidyverse)

library(gplots)

install.packages("GGally")

library(GGally)

install.packages("scales")

library(scales)

install.packages("mosaic")

library(mosaic)

install.packages("mapproj")

library(mapproj)

install.packages("ggplot2")

library(ggplot2)

## 5.2 Histogram for frequency of the IMDB score

> ggplot(movie, aes(x = imdb\_score)) +

geom\_histogram(aes(fill = ..count..), binwidth =0.5) +

scale\_x\_continuous(name = "IMDB Score", +

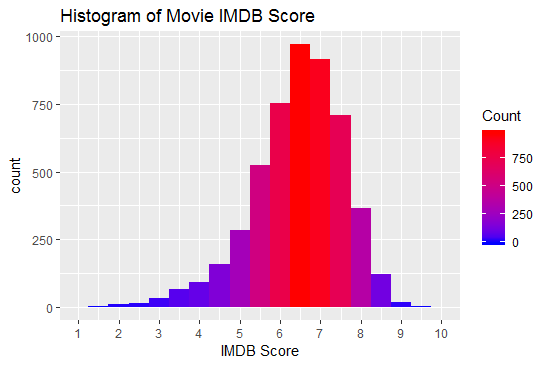
breaks = seq(0,10), +

limits=c(1, 10)) +

ggtitle("Histogram of Movie IMDB Score") +

scale\_fill\_gradient("Count", low = "blue", high = "red")

Output



Figuring out the highest imdb score

> max(movie$imdb\_score)

[1] 9.3

The highest imdb score is 9.3

## 5.3 Range in years of movies

We then looked for the year range of movies using the following code:

> range(movie$title\_year)

[1] 1920 2016

The movies range from 1920-2016

## 5.4 Movies with IMDB score greater than and equal to 8

> sum(with(movie,imdb\_score>=8))

[1] 214

There are 214 movies with imdb score greater than and equal to 8.

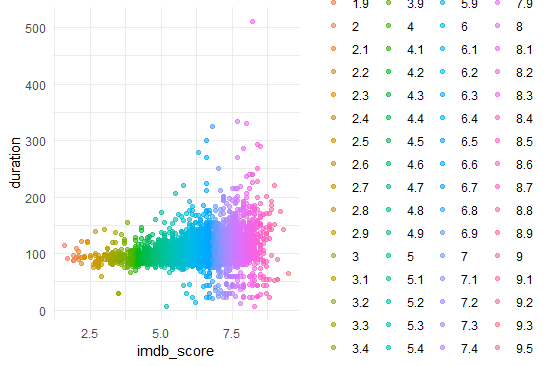
## 5.5 Plotting Correlation between IMDb Score and other Variables using Scatter Plots

### 5.5.1 Correlation between duration and IMDb Score

> ggplot(movie, aes(y = duration, x = imdb\_score, color= factor(imdb\_score))) +

geom\_point(alpha = 0.6) +

theme\_minimal()



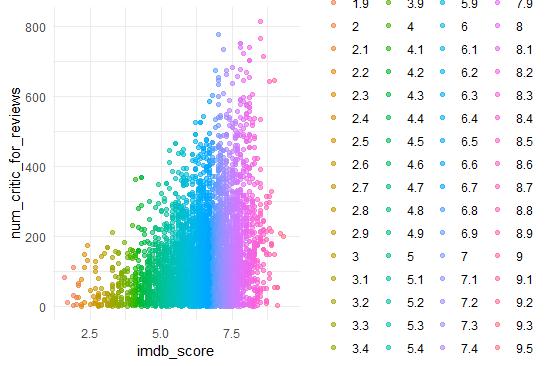
We found out that duration has high correlation with imdb score. More the duration, higher the IMDb score.

### 5.5.2 Correlation between Number of Critic Reviews and IMDb Score

> ggplot(movie, aes(y = num\_critic\_for\_reviews, x = imdb\_score, color= factor(imdb\_score))) +

geom\_point(alpha = 0.6) +

theme\_minimal()



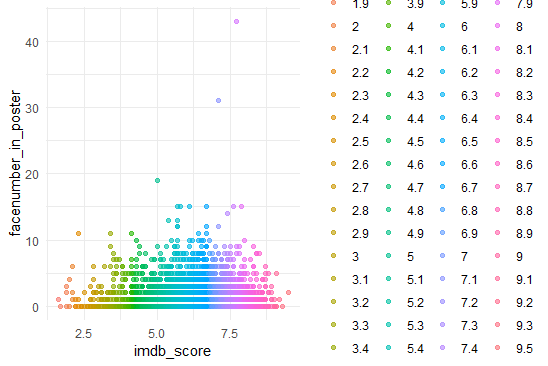
Here again we found high correlation between the variables. Higher the number of crtics for review, higher the IMDb score.

### 5.5.3 Correlation between Face Number in Poster and IMDb Score.

> ggplot(movie, aes(y = facenumber\_in\_poster, x = imdb\_score, color= factor(imdb\_score))) +

geom\_point(alpha = 0.6) +

theme\_minimal()



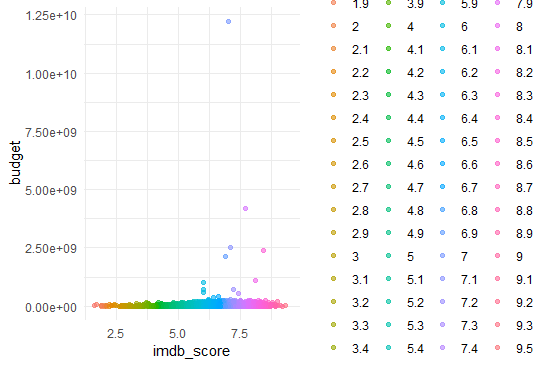
However, we are unable to find out measurable correlation between these variables

### 5.5.4 Correlation between Budget and IMDb Score

ggplot(movie, aes(y = budget, x = imdb\_score, color= factor(imdb \_score))) +

geom\_point(alpha = 0.6) +

theme\_minimal()



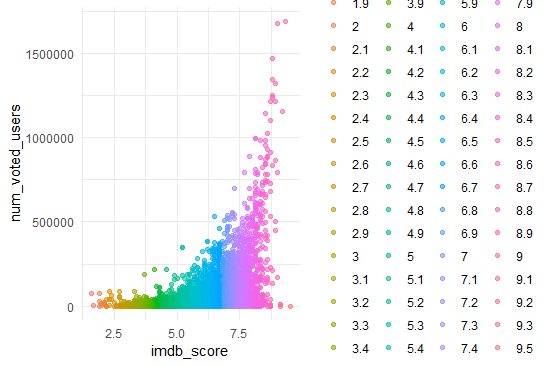
Similarly, we are unable to judge any correlation between these variables.

### 5.5.5 Correlation between Number of Users who Voted and IMDb Score

> ggplot(movie, aes(y = budget, x = num\_voted\_users, color= factor(imdb\_score))) +

geom\_point(alpha = 0.6) +

theme\_minimal()



We can observe high correlation between number of users voted for a movie and imdb score. More the number of voted users, higher is the IMDb score.

## 5.6 Plotting Correlation between Different Variables

We used all the numeric data values to plot correlation plot.

We had to install package corrplot.

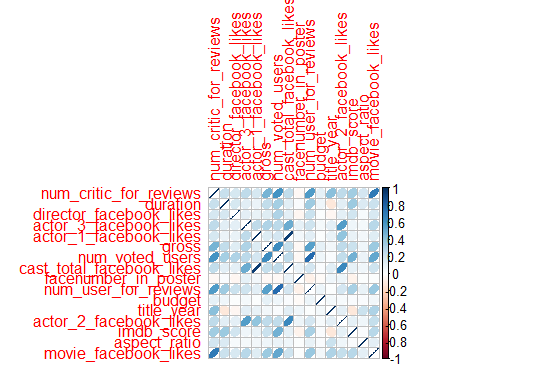
install.packages("corrplot")

library(corrplot)

> num\_movie<- sapply(movie,is.numeric) # select numeric columns

> movie.num<- movies[,nums\_movie]

> corrplot(cor(movie.num),method='ellipse')

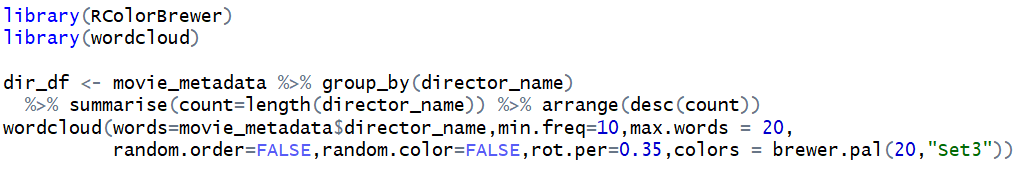


From the correlation plot, we can tell that: Face number in poster has negative correlation with all other predictors. Cast total facebook likes and actor 1 facebook likes has a stronger positive correlation. Interestingly, IMDB scores has strong positive correlation with number of critics for review, which means the more the critics review, the higher the score. Duration and number of voted users also have strong positive correlation with IMDB scores.

## 5.7 Word Cloud

Word cloud is a novel visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

### 5.7.1 Word Cloud of Director Name





Above shows the top 20 director names who make the most movies in the dataset. We can see from the word cloud directly that John, David, Michael made the most movies. And followed with James, Peter, Robert, Richard, Paul.

### 5.7.2 Word cloud of Protagonist Name in the Movies



We can see from the above word cloud that the protagonists have the most movies are Robert, Michael, Tom, Jason, James.

### 5.7.3 Word Cloud of Country of the Movies

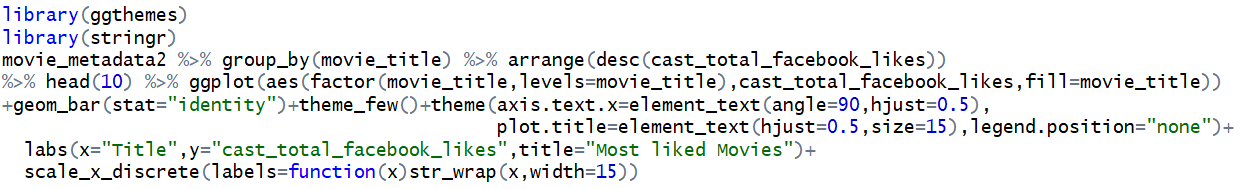


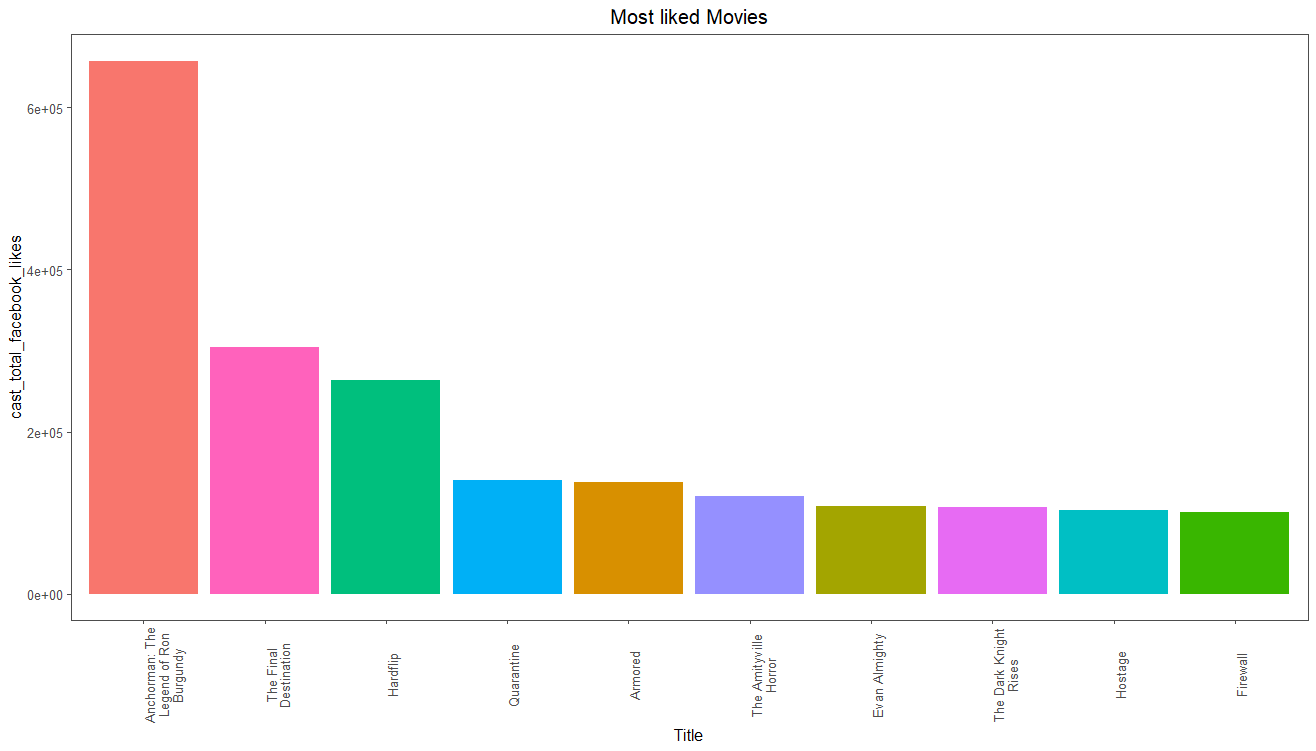


We see the movies in the dataset mostly were made from USA. In the rest of the countries, Canada and France have the most movies.

## 

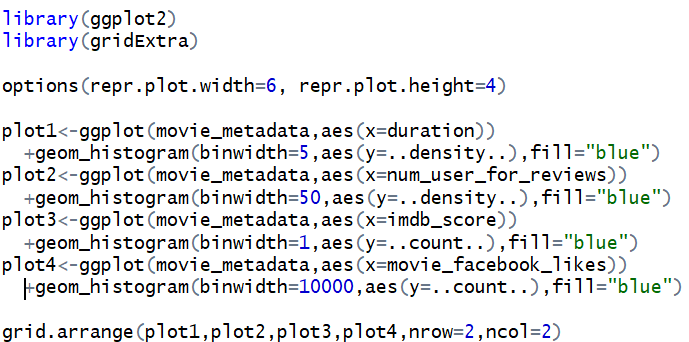
## 5.8 Movies that have the Most Likes on Facebook

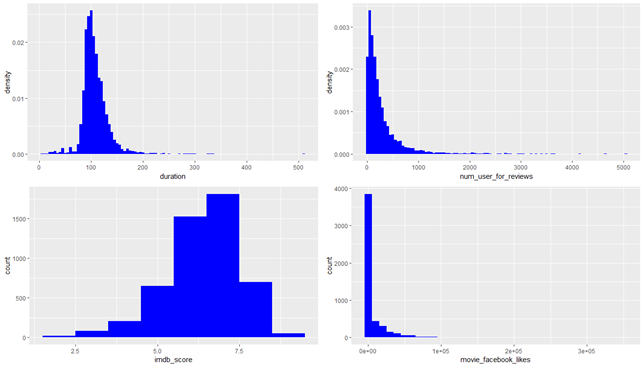




The Anchorman:The Legend of Ron Burgundy is the most liked movie on Facebook. It nearly has two times likes as the second place, The Final Destination, has. Likes for Hardflip are close to those of the Final Destination. The rest of the movies has the close number of likes to each other, but almost the half of the Hardflip.

## 5.9 Density of Different Main Variables of the Movies

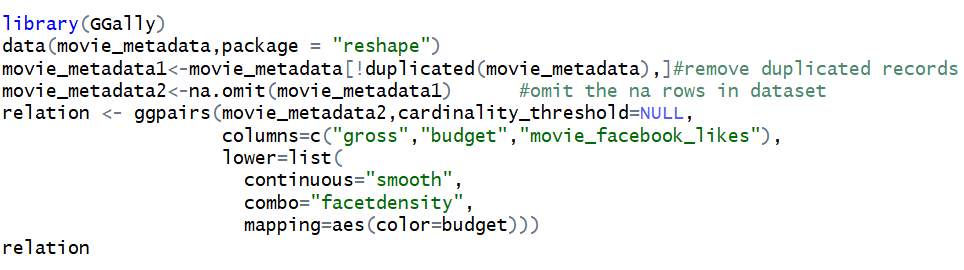


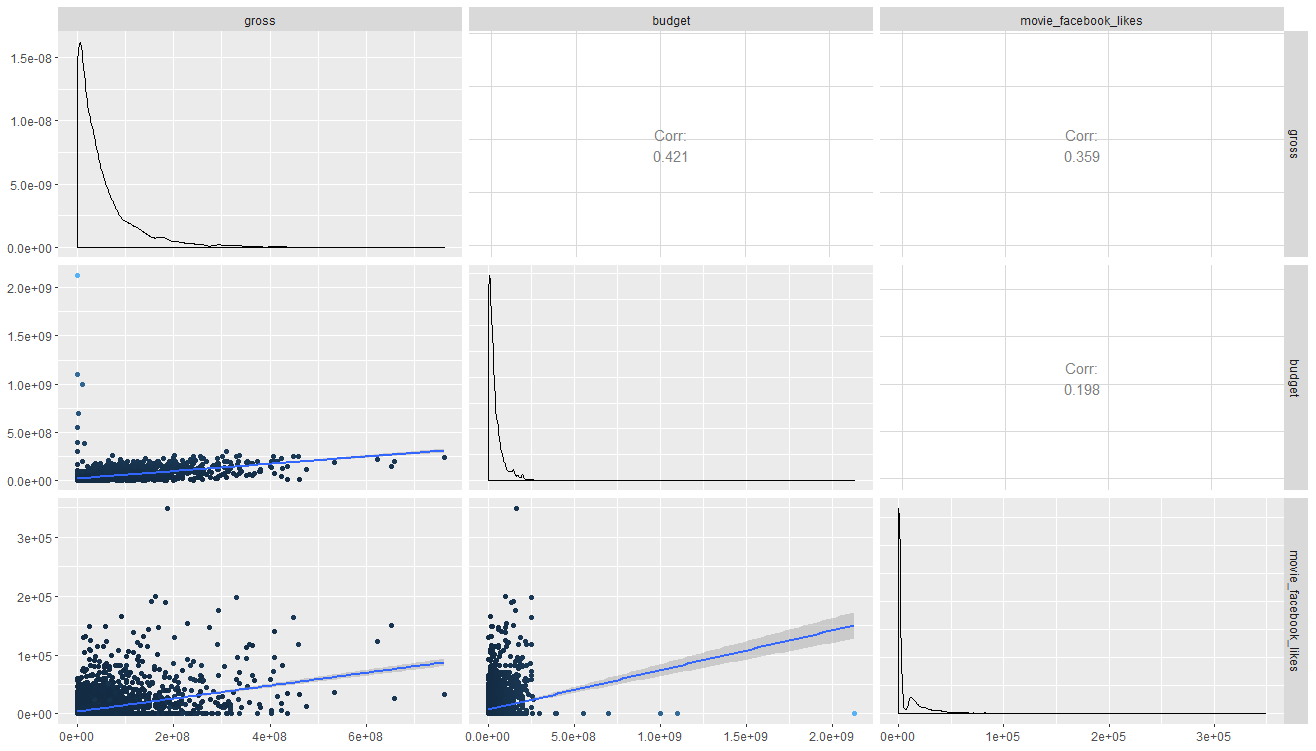


We can see from the plots that the duration of the movies is distributed similar to normal distribution, most of duration is around 100 minutes. The users’ reviews are very imploded. Most of the reviews are given to 500 movies out of 5000 movies in total. However, the likes on the Facebook are even more imploded. Most of the likes given to several movies out of 5000 movies. The IMDB score distribution is left screwed. Most of the scores are below 7.5.

## 

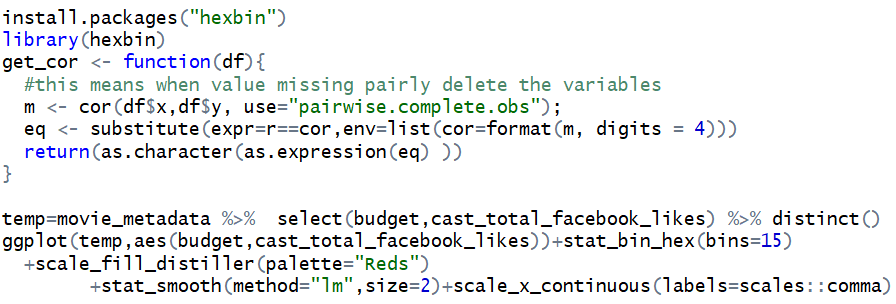
## 5.10 Plotting Correlation between Different Variables

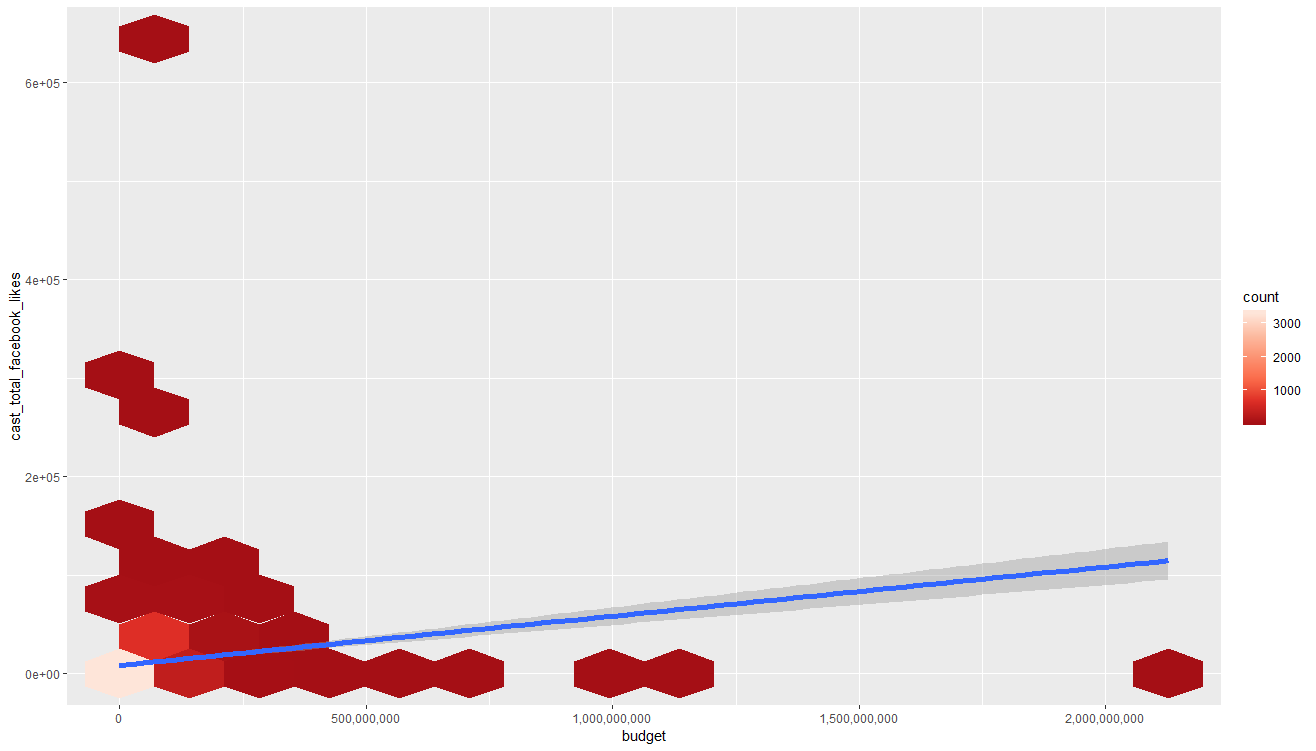




Above matrix plot shows the correlation for gross, budget and numbers of likes on Facebook. The correlation between budget and gross is 0.421, between likes on Facebook and gross is 0.359, between likes and budget is 0.198. So we can only conclude the budget and gross has some correlation.

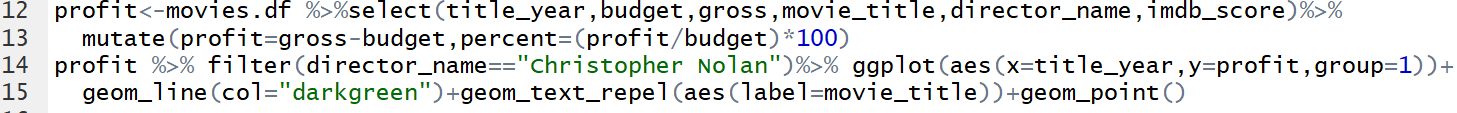
## 5.11 Plotting Correlation between Budget and Likes

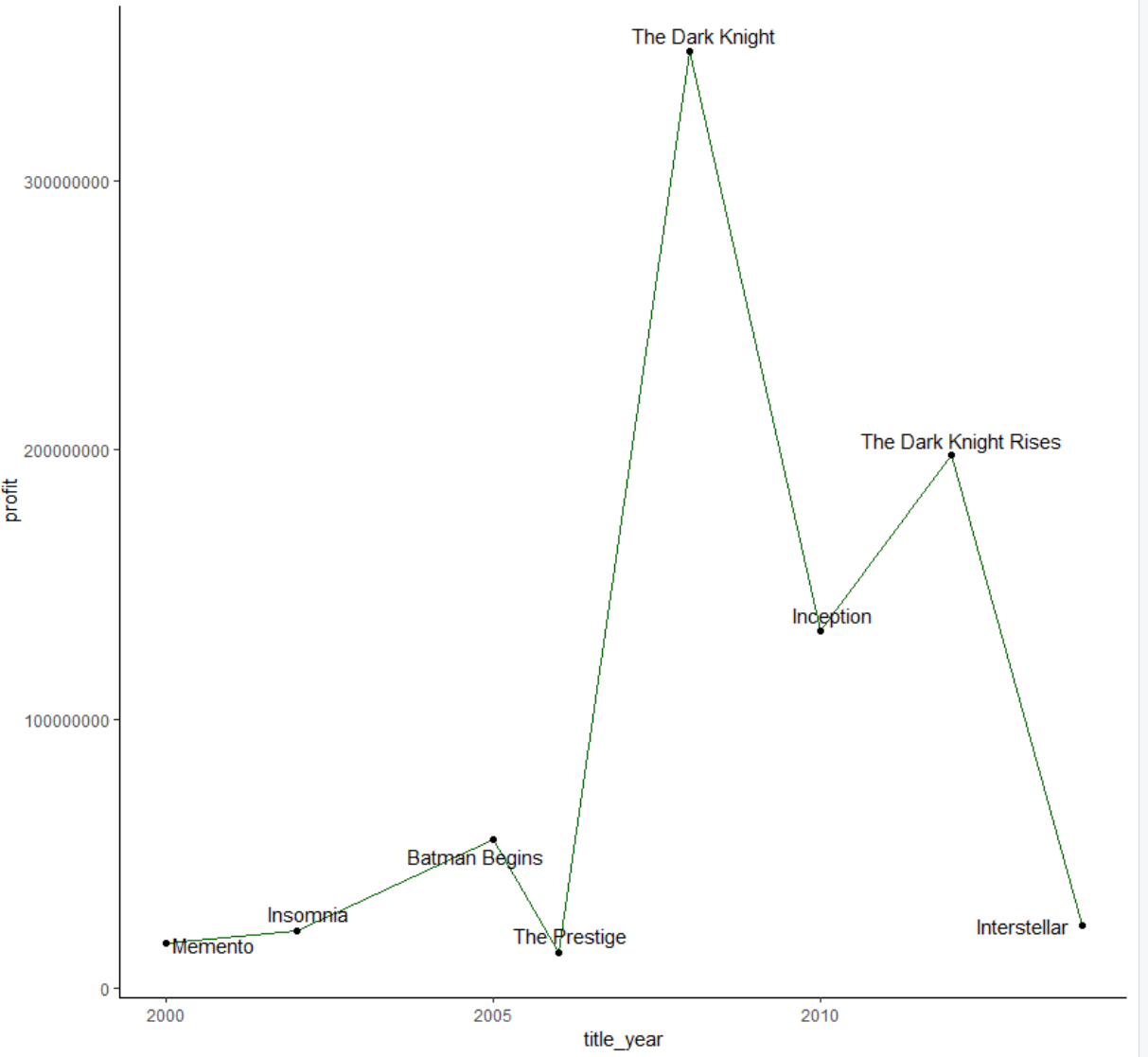




The hex plot can show the number that points at the same position using different colors. In plot above, lighter the color, more points are gathered at the position. Most points concentrated in the left bottom area and we found the correlation can fit a low slope linear relation.

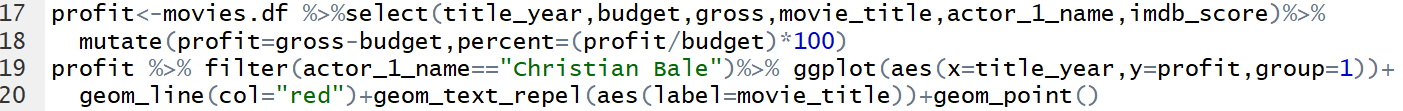
5.12 Plotting Correlation between Profit and Christopher Nolan movies

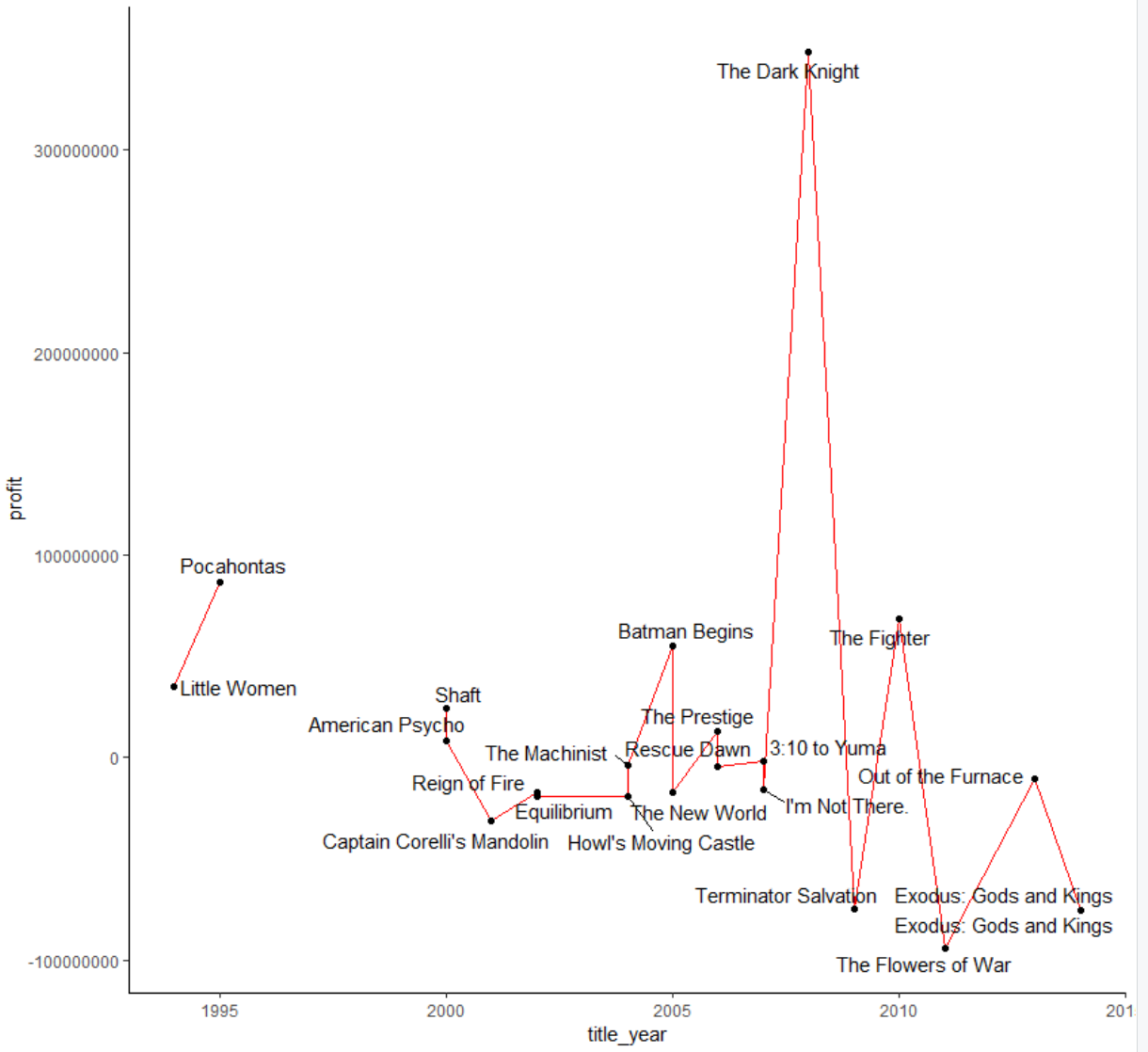




This graph shows the correlation of Christopher Nolan movies with the profit and the year that they were released in. The movie: The Prestige, made almost no profit, while The Dark Knight was the highest grossing Christopher Nolan movie with a profit of more than 300 million USD.

5.13 Plotting Correlation between Profit and Christian Bale movies





The following Christian Bale movies: Reign of Fire, Equilibrium, The New World, I’m Not There, Terminator Salvation, Exodus: Gods and Kings suffered a loss. The Dark Knight was the highest earning Christian Bale movie with a profit of more than 300 million USD.

# 6. Multiple Regressions Analysis

## 6.1 Introduction

As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.  The independent variables can be continuous or categorical (dummy coded as appropriate).

## 6.2 Assumptions

1. Regression residuals are normally distributed.
2. A linear relationship is assumed between the dependent variable(iMDb Rating) and the independent variables
3. The residuals are homoscedastic and approximately rectangular-shaped.
4. Absence of multicollinearity is assumed in the model, meaning that the independent variables are not too highly correlated.
5. At the center of the multiple linear regression analysis is the task of fitting a single line through a scatter plot.  More specifically the multiple linear regression fits a line through a multi-dimensional space of data points.  The simplest form has one dependent and two independent variables.  The dependent variable may also be referred to as the outcome variable or regressand.  The independent variables may also be referred to as the predictor variables or regressors.

## 6.3 Uses

The Multiple linear regression model in this project has been used mainly for 3 purposes

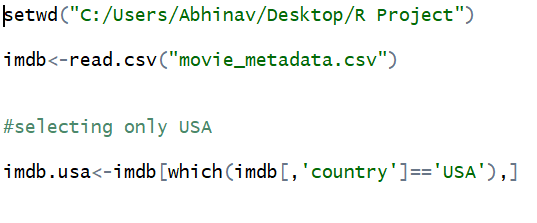
1. It is used to identify the strength of the effect that the independent variables(Gross Earnings,Budget, Language etc.) have on the overall iMDb rating.
2. Second, it is used to forecast effects or impacts of changes.  That is, how much will the iMDb rating change when we change the independent variables.  For instance, a higher gross earnings for a movie should result in a higher iMDb rating for that movie according to basic economic theory.
3. Third, it is used to predict the trends and values. This can be used to answer questions such as, “What would be the iMDb rating for a movie with Gross Earnings of $400,000 and a budget of $200,000?”

## 

## 6.4 Procedure

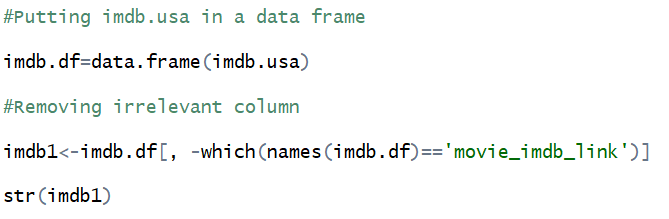
Step 1

1. The overall data set had 66 countries in total but the data set has been used for movies that have their origin in USA. This leaves us with 3005 rows.
2. The reason for such omission is the fact that we wanted the regression model to be country specific. One major reason for that is the Gross Income or Earnings vary from country to country due to varying currency exchange rates.



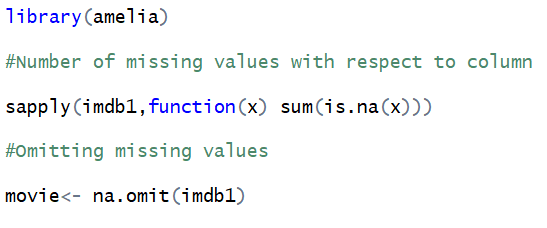
Step 2

1. The column ‘movie\_imdb\_link’ was removed because it didn’t serve the purpose of regression.

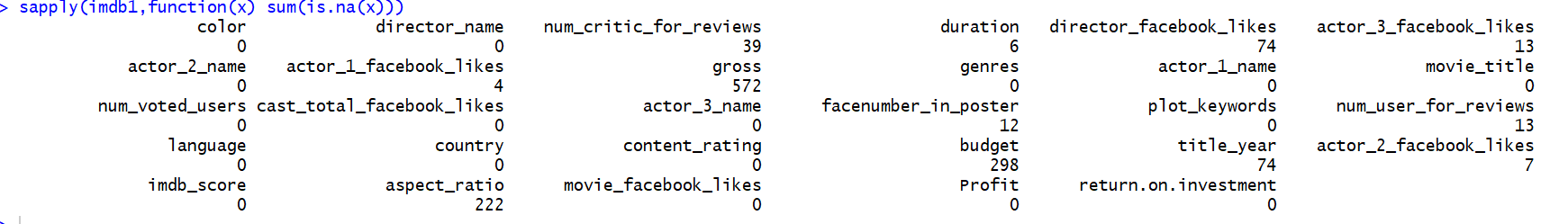


Step 3

1. The row having even a single blank value has been omitted.

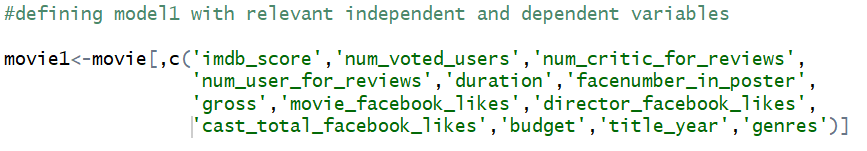


Output ( Step 3)



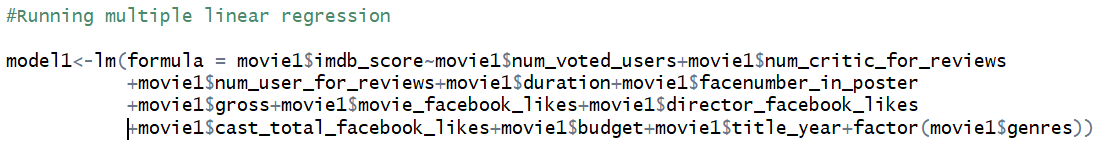
Step 4

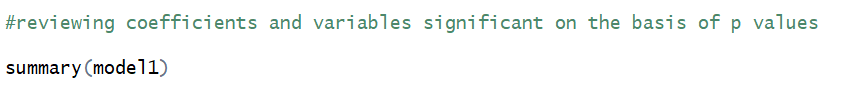
1. Only the variables which could impact the iMDb score were included.

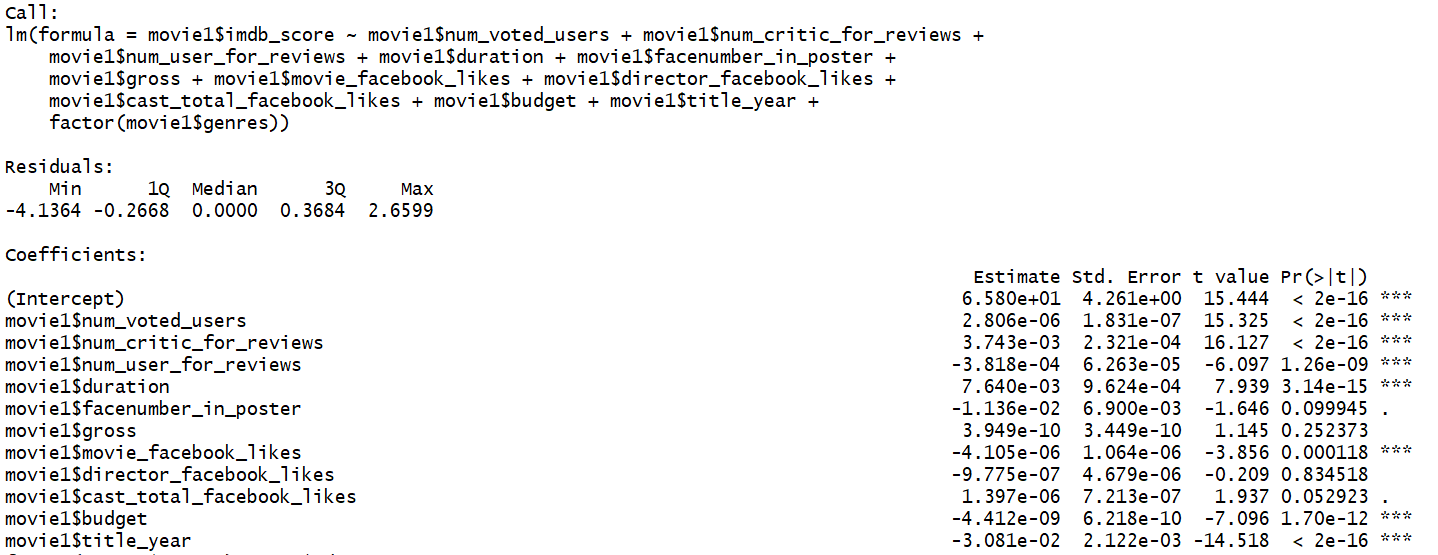


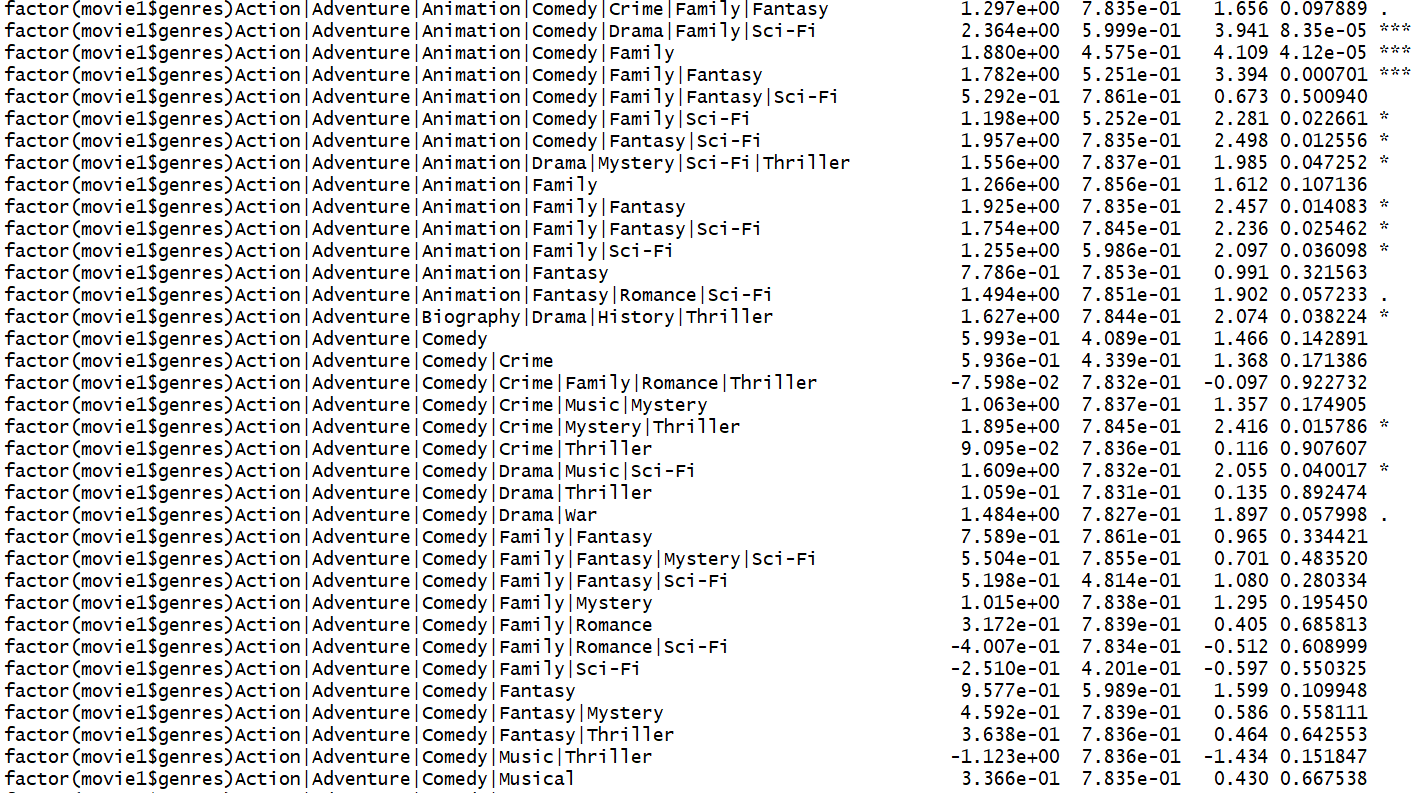
Step 5

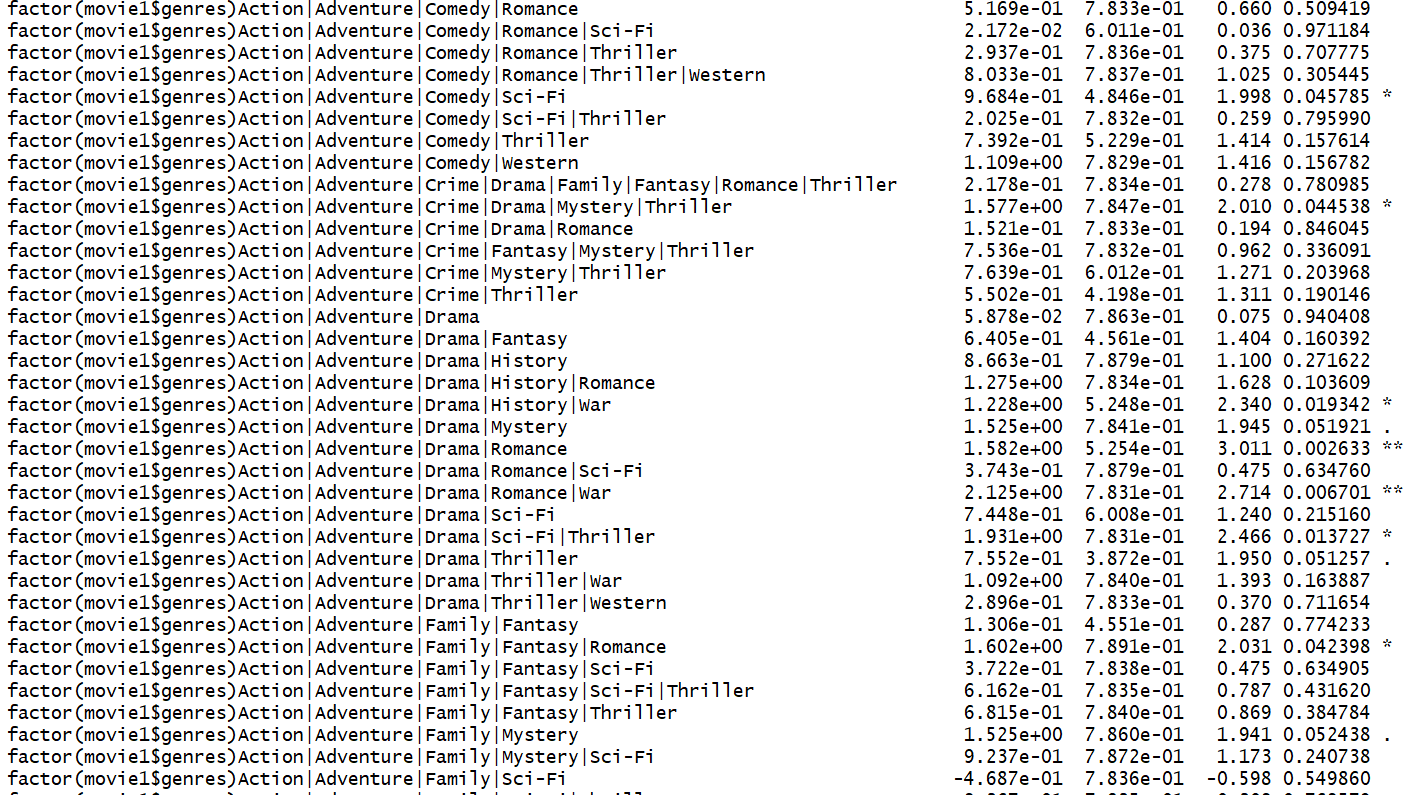
1. Running regression model by taking dependent variable as imdb\_score and independent variables as num\_voted\_users, num\_user\_for\_reviews, duration, facenumber\_in\_poster, gross, movie\_facebook\_likes, director\_facebook\_likes, cast\_facebook\_likes, budget, title\_year and genres.

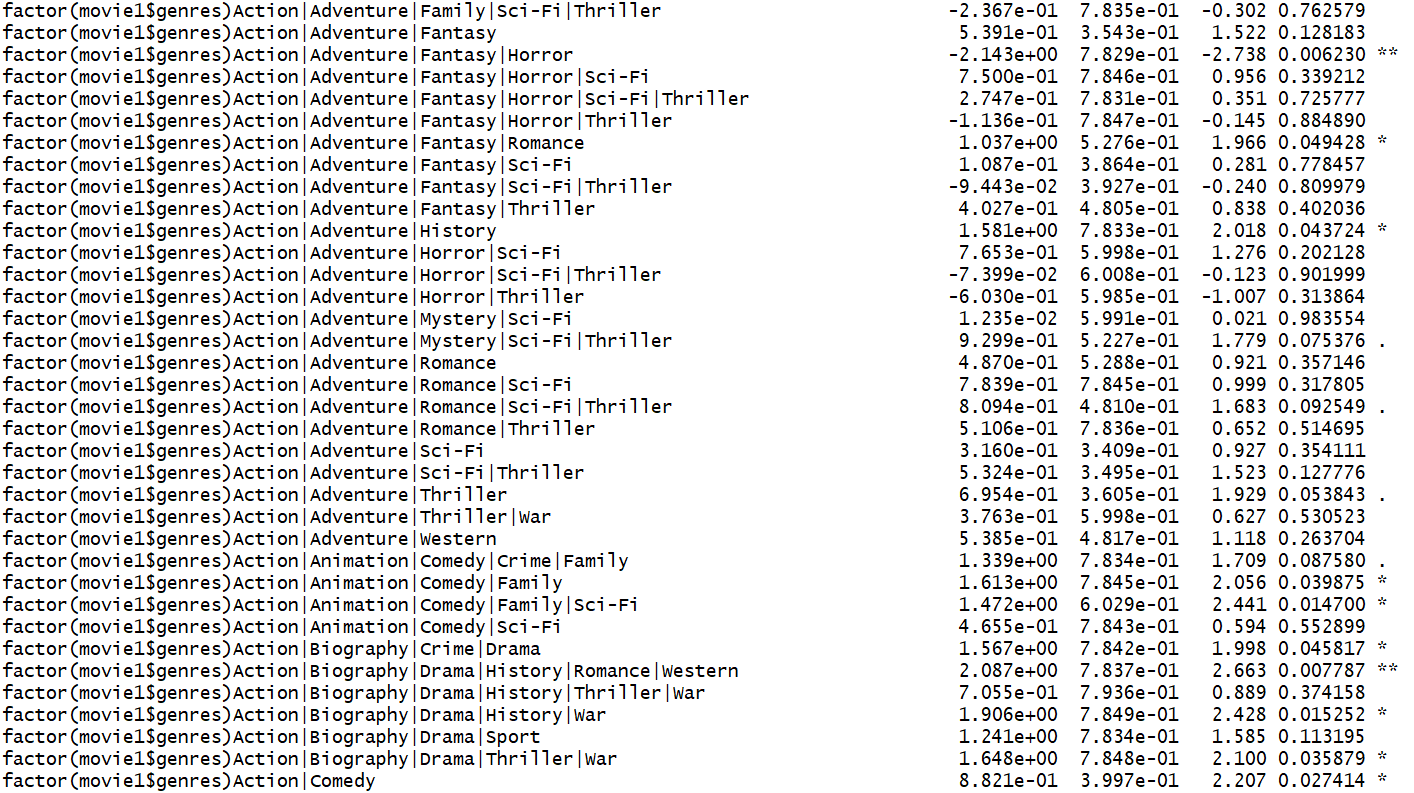


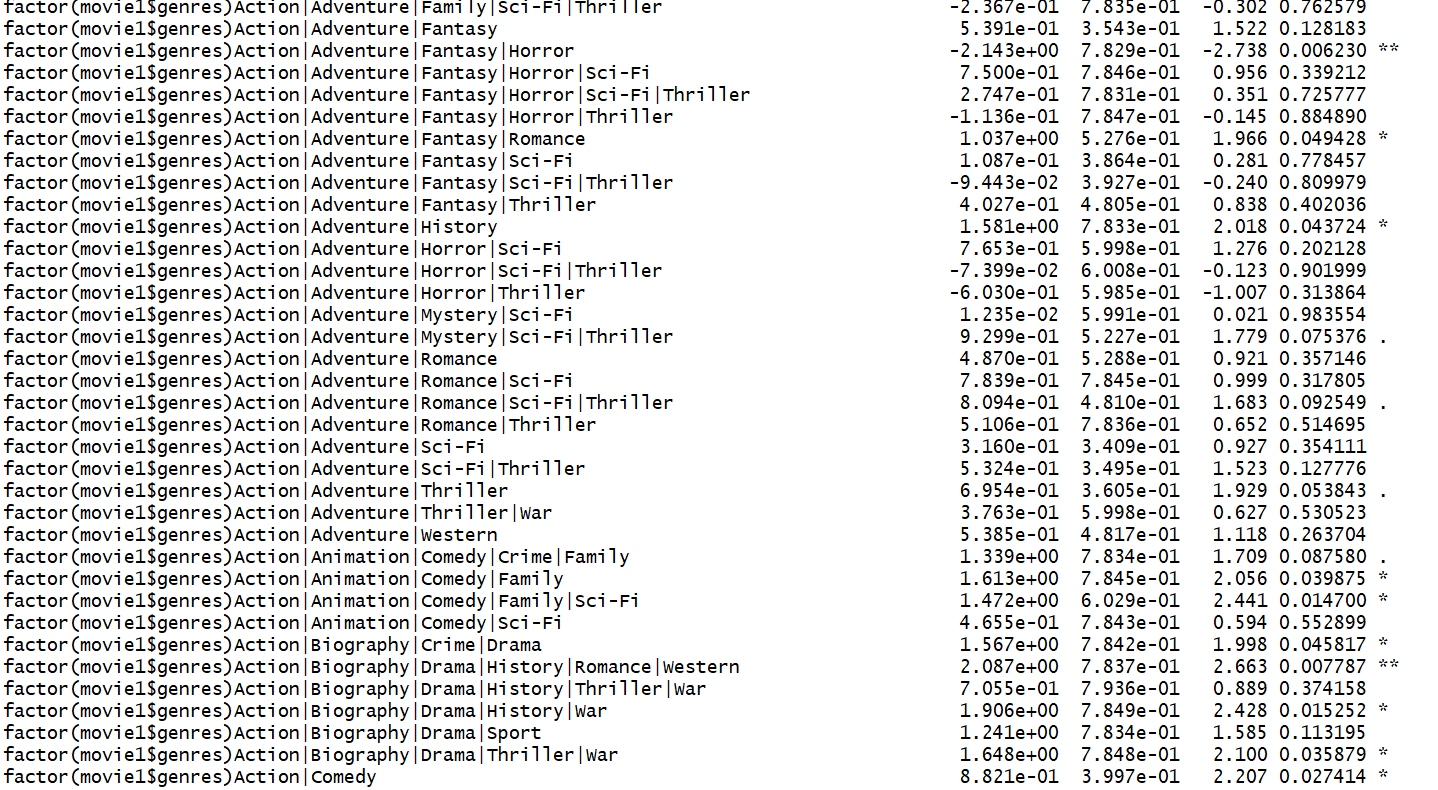


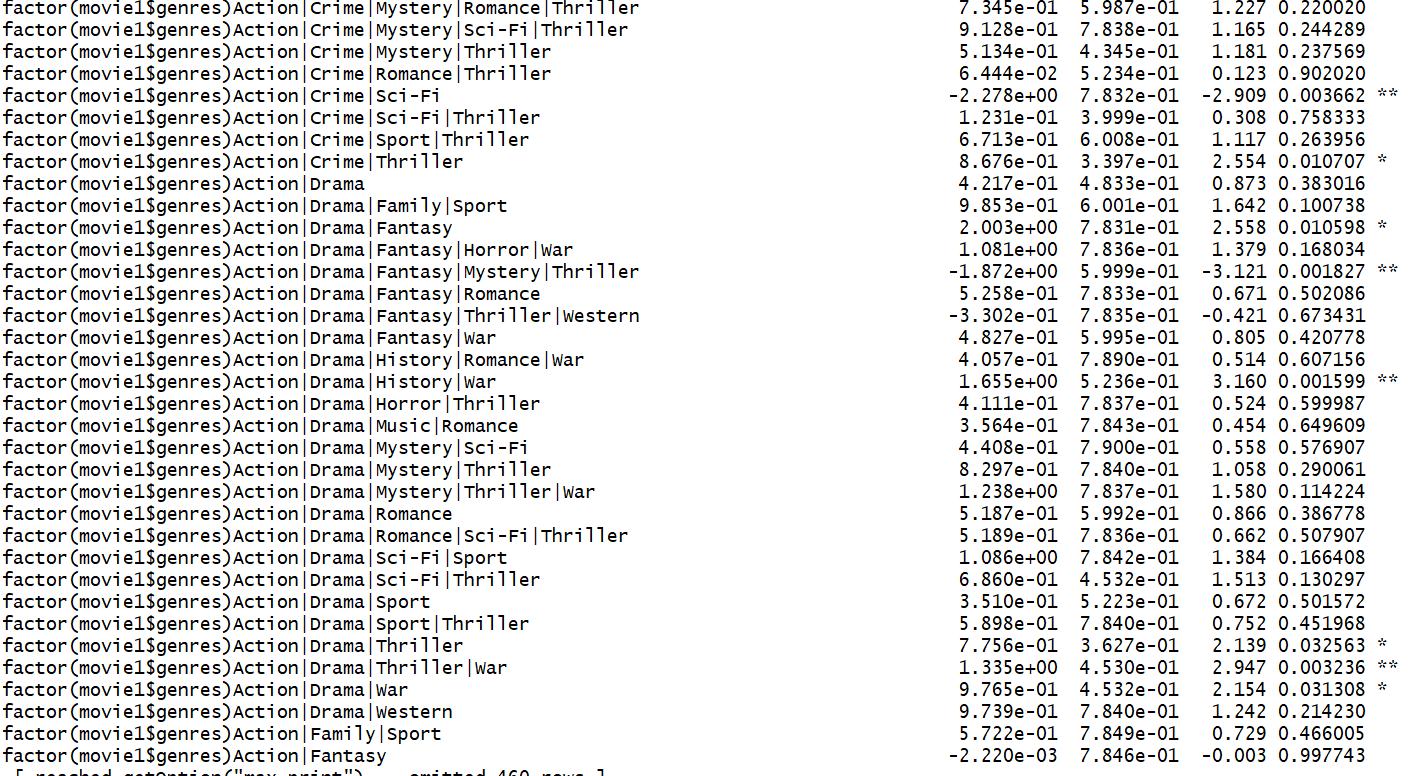




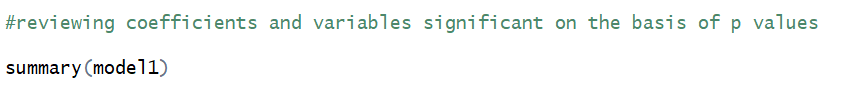


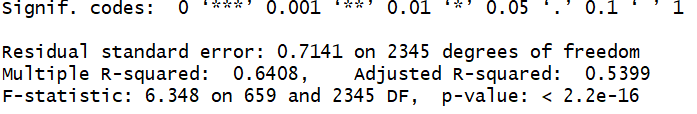






Step 6





RESULTS

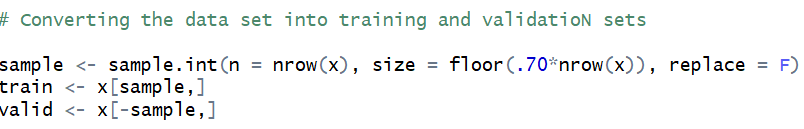
R Squared – 64.08%

F Stat – 6.348

P value – 0.000000000000000022

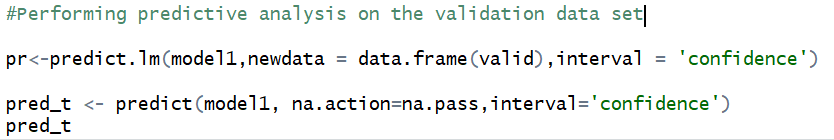
Step 7

1. Converting the Data Sets into Training and Validation Data Sets

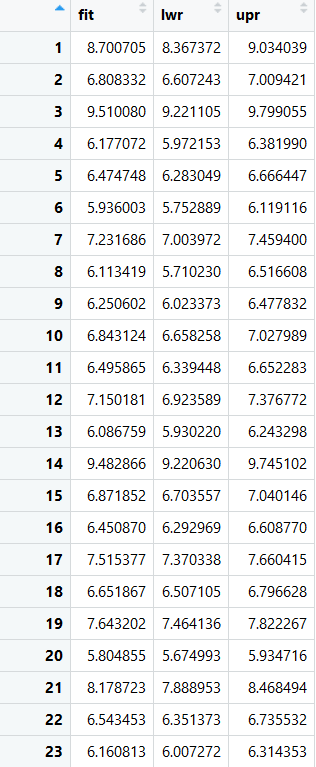


Step 8

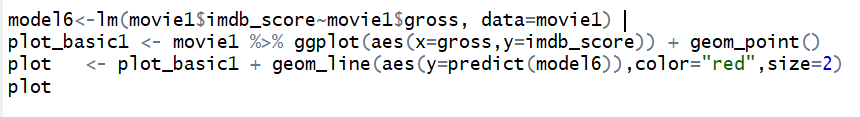
1. Performing predictive analysis using the validation Data Set

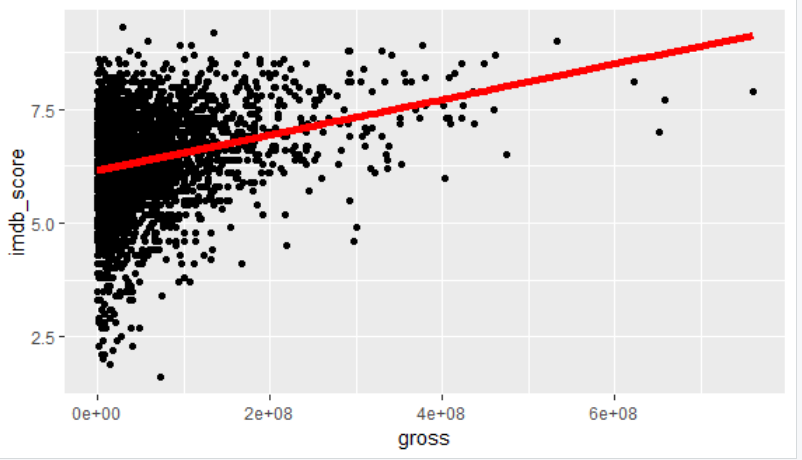


FITTED VALUES WITH THEIR CONFIDENCE INTERVALS



Plotting a relationship between Gross Earnings of the movies and their respective imdb\_score





## 6.5 Results

1. The plot clearly shows that with increase in Gross Earnings, the iMDb score is bound to increase.
2. It also shows that gross earnings of $0 would at least give an iMDb rating of 6.
3. The graph is rational according to the basic economic theory that higher gross earnings would lead to higher iMDb rating and vice versa.

## 6.6 Conclusion

### 6.6.1. Signs of the Coefficients

According to the results of the regression model the variables have the following effect on the IMDb rating

a.) num\_voted\_users – An increase in number of voters increases the iMDb rating.

b.) num\_critic\_for\_reviews-An increase in number of critics for reviews increases the iMDb rating.

c.) num\_user\_for\_reviews- An increase in number of users for reviews decreases the iMdb rating.

d.) duration- An increase in duration of a movie increases the iMDb rating.

e.) facenumber\_in\_poster - A lower face number in a poster of a movie. has a higher iMDb rating.

f.) gross – Higher Gross earnings will lead to a higher iMDb rating because of the reason that more the movie earns, more popular the movie will be.

g.) movie\_facebook\_likes- Total movie facebook likes decreases the overall iMDb rating of the movie.

h.) director\_facebook\_likes- This factor decreases the overall iMDb rating of the movie.

i.) cast\_total\_facebook\_likes - Total Facebook likes of cast has a high positive impact on the iMDb rating.

j.) budget – Greater the budget, less will be the iMDb rating.

k.) title\_year – Older the movie, it is less likely that the movie will have a higher iMDb rating.

### 6.6.2. Significance Level of the Variables

All the variables are significant except for gross, facenumber\_in\_poster, movie\_facebook\_likes and director\_facebook\_likes

### 6.6.3. Efficiency and Accuracy of the Model

The R squared for the model is coming out to be 0.6408. This means that the dependent variables are able to explain 64.08% of the variability in the model.

High and low values of the F-Statistic and p-value is a proof that the model is highly significant.

### 6.6.4. Outliers in the Model

There are some records for which the iMDb score is coming out to be more that 10 which is rationally not possible. These records are called outliers ( an observation point that is distant from other observations). These outliers may be due to variability in the measurement or it may indicate experimental error.

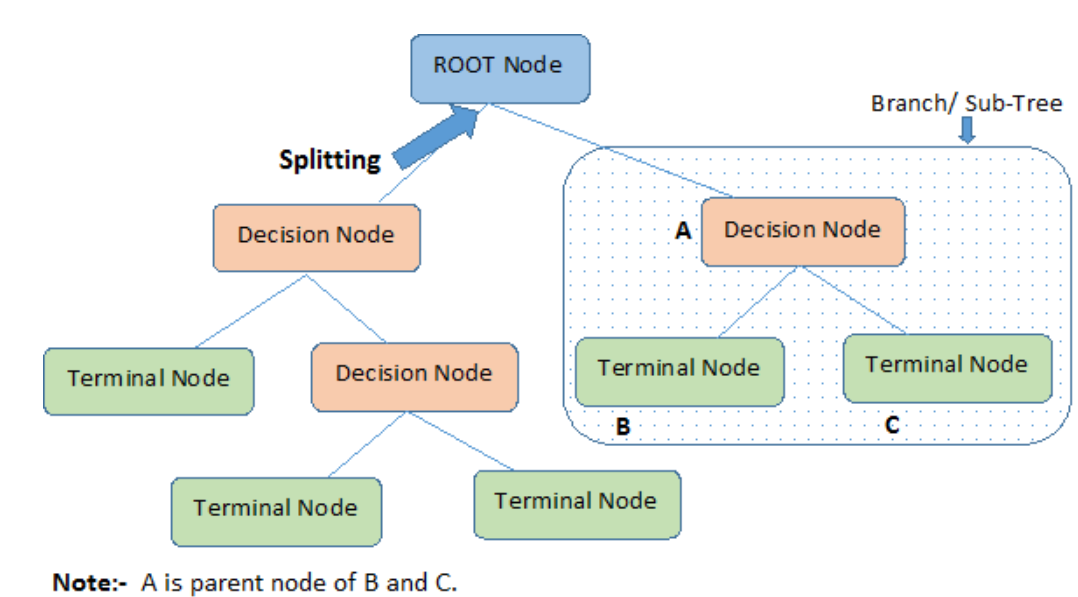
### 6.6.5. AIC of the model

The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data.  The AIC for this model is coming out to be -1448.97. This proves that the model is highly qualitative in nature.

# 7. Classification and Regression Tree (CART)

## 7.1 Introduction

Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems. It works for both categorical and continuous input and output variables.



* Root Node represents the entire population or sample. It further gets divided into two or more homogeneous sets.
* Splitting is a process of dividing a node into two or more sub-nodes.
* When a sub-node splits into further sub-nodes, it is called a Decision Node.
* Nodes that do not split is called a Terminal Node or a Leaf.
* When we remove sub-nodes of a decision node, this process is called Pruning. The opposite of pruning is Splitting.
* A subsection of an entire tree is called Branch.
* A node, which is divided into sub-nodes is called a parent node of the sub-nodes; whereas the sub-nodes are called the child of the parent node.

## 7.2 Decision tree types

Decision trees used in [data mining](https://en.wikipedia.org/wiki/Data_mining) are of two main types:

* [Classification tree](https://en.wikipedia.org/wiki/Classification_tree) analysis is when the predicted outcome is the class to which the data belongs.
* Regression tree analysis is when the predicted outcome can be considered a real number

The term Classification And Regression Tree (CART) analysis is an [umbrella term](https://en.wikipedia.org/wiki/Umbrella_term) used to refer to both of the above procedures, first introduced by [Breiman](https://en.wikipedia.org/wiki/Leo_Breiman) et al. in 1984.

## 7.3 The Algorithm behind Decision Trees

The algorithm of the decision tree models works by repeatedly partitioning the data into multiple sub-spaces, so that the outcomes in each final sub-space is as homogeneous as possible. This approach is technically called *recursive partitioning*. The produced result consists of a set of rules used for predicting the outcome variable, which can be either:

· a continuous variable, for regression trees

· a categorical variable, for classification trees

The decision rules generated by the CART (Classification & Regression Trees) predictive model are generally visualized as a binary tree.

## 7.4 Advantages of Trees

1. Fast computations

2. Invariant under monotone transformations of variables

· Scaling doesn’t matter

· Immune to outliers

3. Resistance to irrelevant variables, so can throw lots of variables into it

4. One tuning parameter (tree size, or cp)

5. Interpretable model representation (easy to read)

6. Extends to categorical outcomes easily

## 7.5 Limitations

1. Trees can be very non-robust. A small change in the [training data](https://en.wikipedia.org/wiki/Training,_test,_and_validation_sets) can result in a large change in the tree and consequently the final predictions.

2. Decision-tree learners can create over-complex trees that do not generalize well from the training data. (This is known as [overfitting](https://en.wikipedia.org/wiki/Overfitting)) Mechanisms such as [pruning](https://en.wikipedia.org/wiki/Pruning_(decision_trees)) are necessary to avoid this problem (with the exception of some algorithms such as the Conditional Inference approach, that does not require pruning).

3. Each subsequent split depends on the previous ones, so an error in a higher split is propagated down.

## 7.6 Algorithm Steps

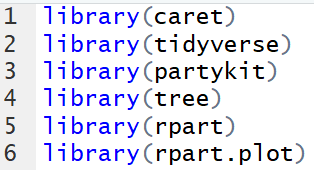
Step (1): Importing the required R libraries:

· tidyverse for easy data manipulation and visualization

· caret for easy machine learning workflow

· rpart and partykit for computing decision tree models

Code:



Step (2): Selecting the variables to be used in the model and processing them [5]

Here, we have chosen the following variables for the decision tree:

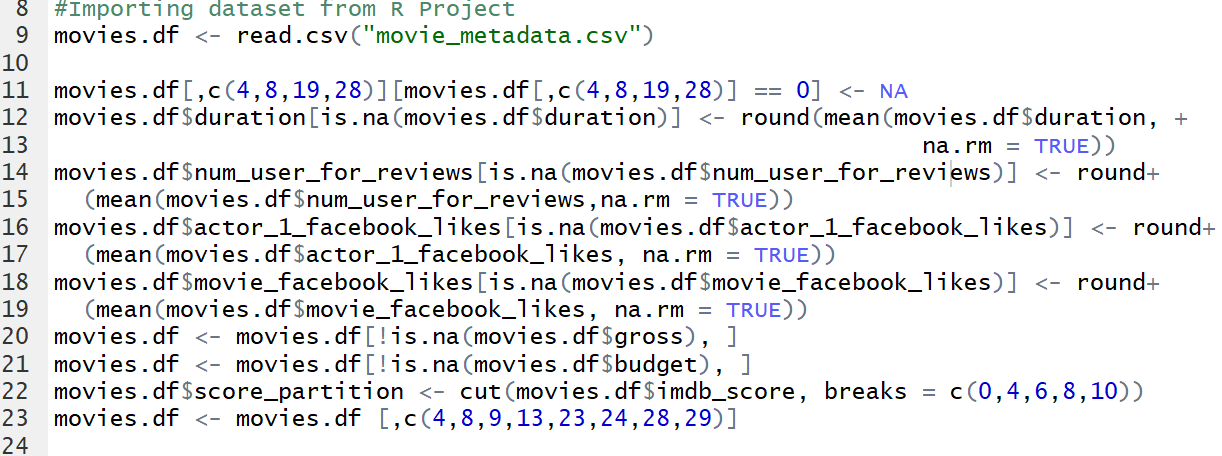
* duration
* actor\_1\_facebook\_likes
* gross
* num\_voted\_users
* budget
* title\_year
* movie\_facebook\_likes
* score\_partition

We have not included num\_critics\_for\_review even though it might be a potential predictor variable, because num\_voted\_users and num\_critics\_for\_review are highly correlated with one another and therefore, may result in multicollinearity.

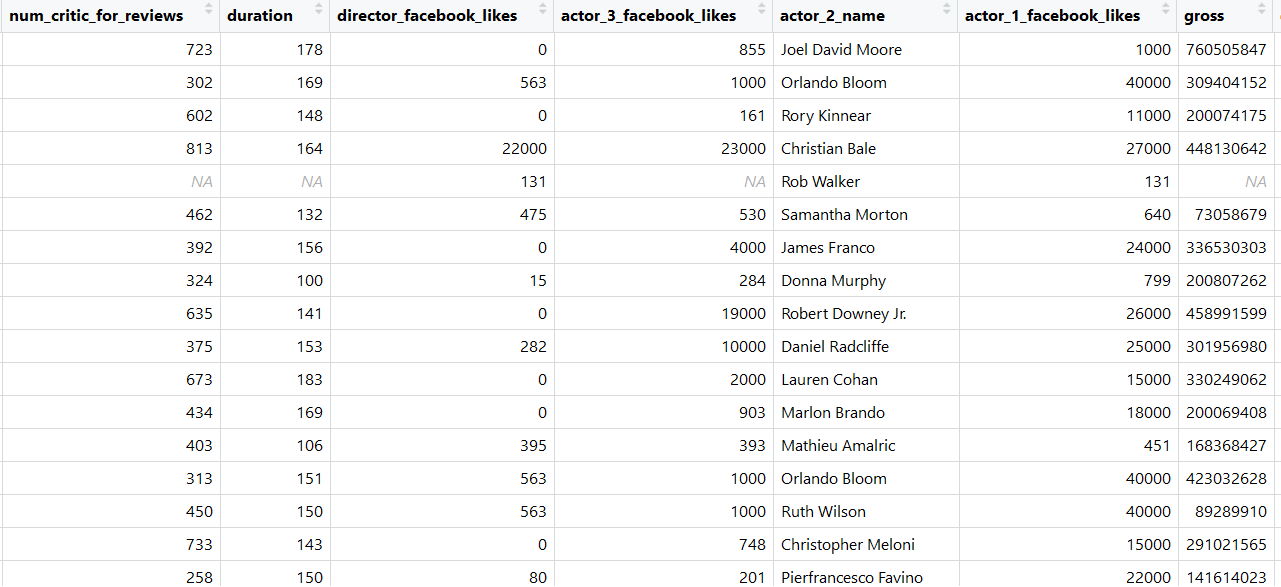
The structure of the data shows that some variables have NA's. Data clean-up is done here as follows:

* The few null values in the gross and budget columns have been deleted.
* The null values in the remaining columns have been replaced by the average of the values for the other observations for that variable. This is because there are a lot of empty data points for these variables, and therefore, keeping them in the dataset by inserting average values might be a better option.
* We have also partitioned the movie score into bins of 0,3,6,8,10 so that it is easy to understand the rating of the movie with 0-3 being a very poorly rated movie and 8-10 being a very highly rated movie.

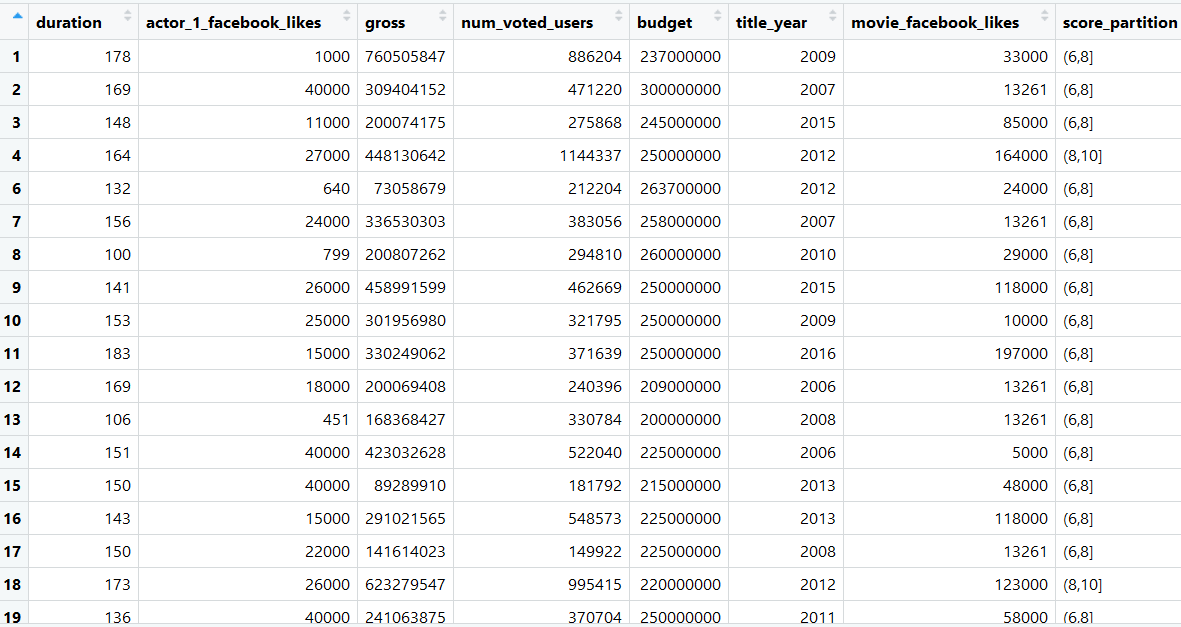
Code:



Output:



Revised Output after Processing:



Step (3): Create training and testing sets

Before we train our model, we need to create a training and testing set: We train the model on the train set and test the prediction on the test set (i.e. unseen data)

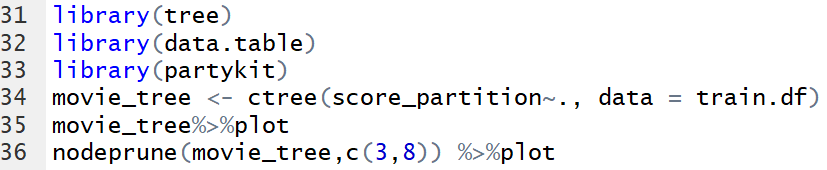
The common practice is to split the data 70/30, 70 percent of the data serves to train the model, and 30 percent to make predictions.

Code:

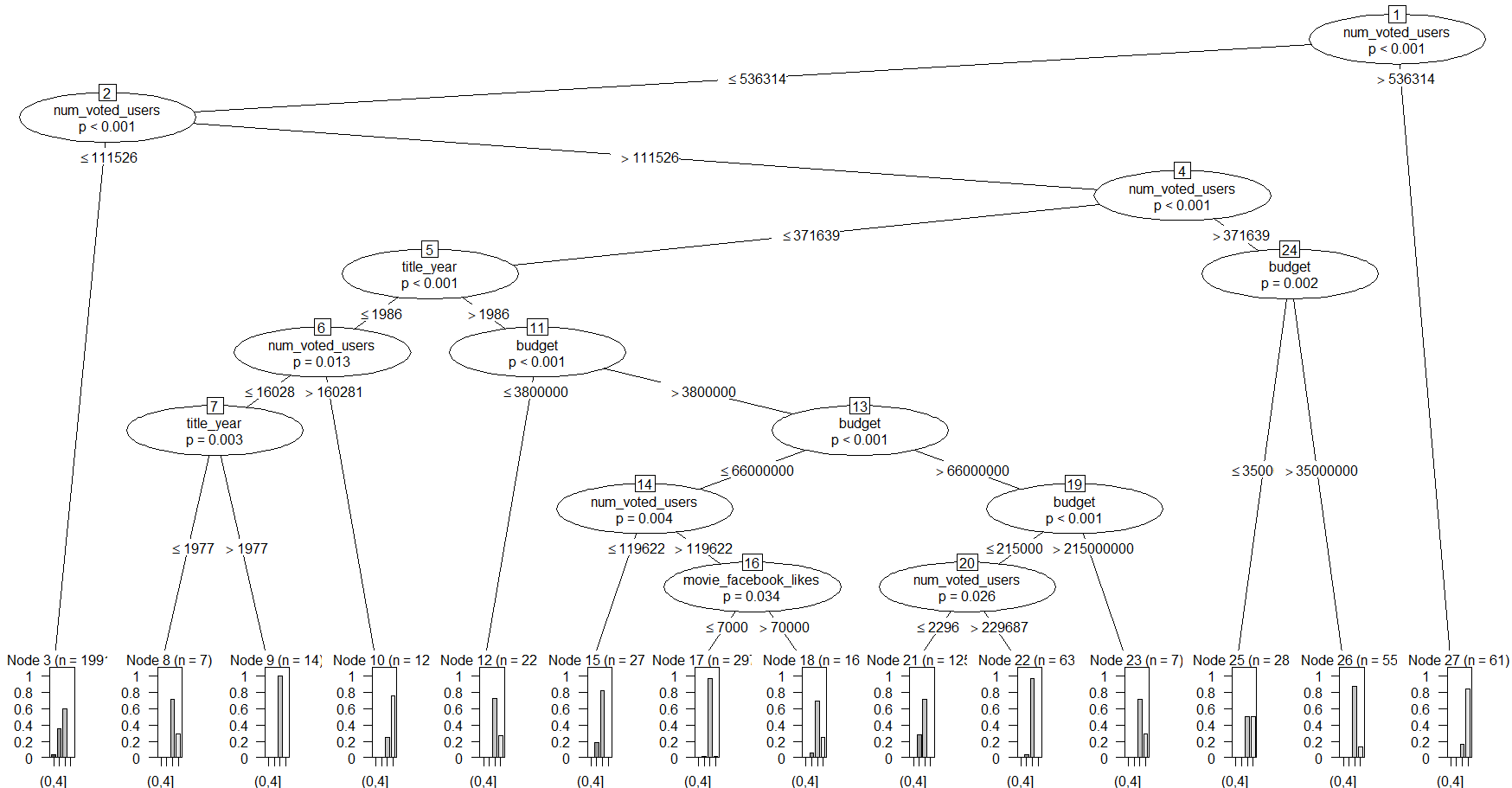


Step (4): Creating a tree

Model 1 Code:



Model 1 Output:



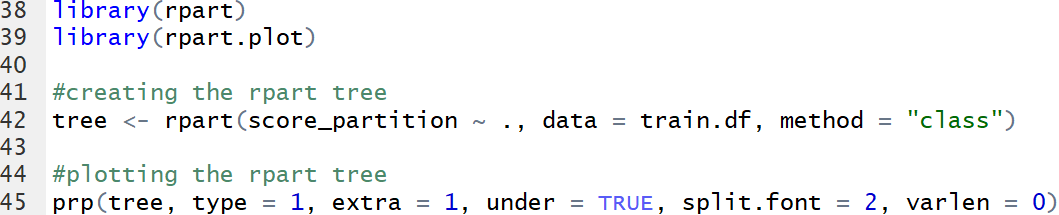
ctree or conditional inference trees (from the partykit package) provides a probability of a movie belonging to a particular movie score partition based on the predictor variables.

Here, we can clearly see that if the number of voted users for a particular movie is more than 5,36,314, it is highly likely for that movie to have a high movie score on a website like IMDb. This could consequently also mean that marketing is an integral factor in score determination. Widespread marketing would lead to more users watching that movie, which could translate to a large number of users voting for that movie.

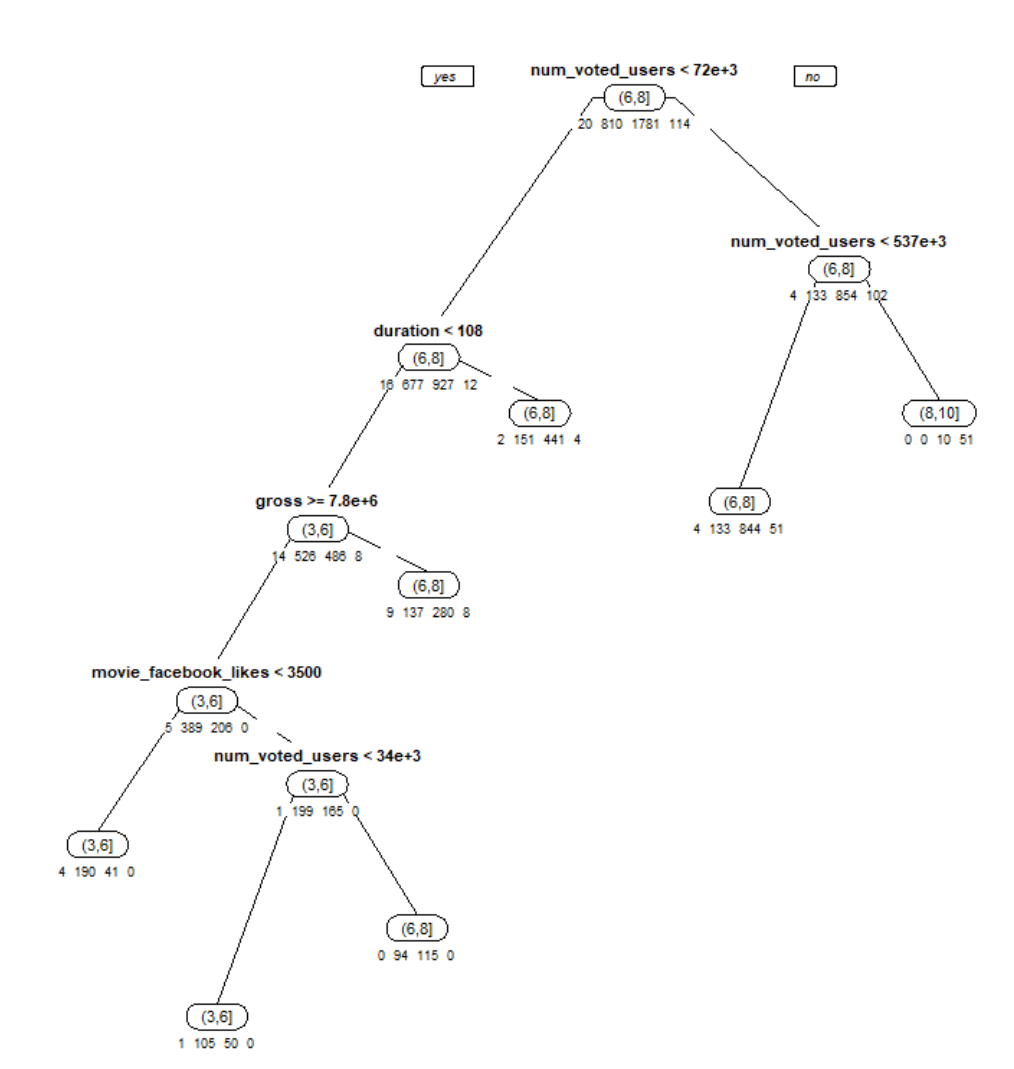
The other factors that might influence the rating of the movie, as picked up by ctree, are the movie year, budget and movie facebook likes. The p-value in the chart shows how statistically significant the value of p is in determining the split between the observations.

However, this model still does not show a clear pattern and therefore, we explore a different package tree model.

Model 2 Code:



Model 2 Output:



This tree is a lot easier to read.

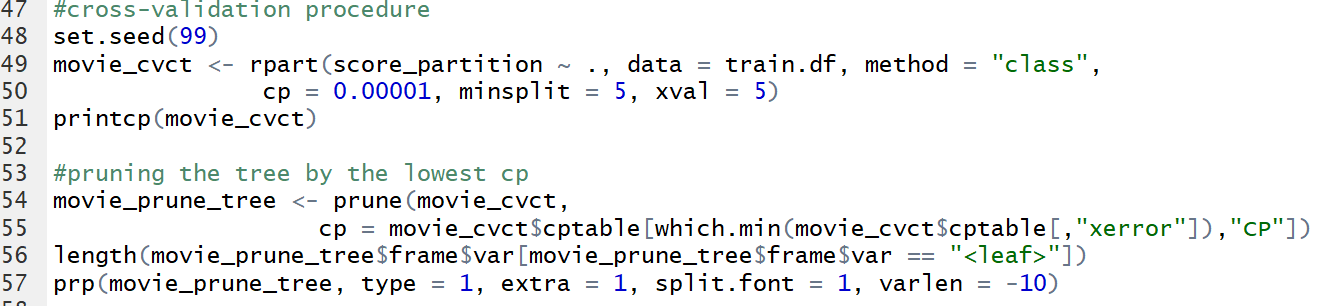
Reading the tree:

A few rules displayed in this tree include:

* If the number of voted users is greater than or equal to 537e+3, the movie will probably have a high rating of between 8 to 10.
* If the number of voted users is greater than 72e+3 but less than 537e+3, it is likely to be in the 6-8 movie score partition bucket.
* If the number of voted users is less than 72e+3 but the duration of the movie is greater than or equal to 108 minutes, it can still achieve a high rating of 6 to 8.
* If the number of voted users is less than 72e+3, the duration of the movie is less than 108 minutes, the gross budget is greater than or equal to 7.8e+6 and the movie Facebook likes are less than 3500, then the movie is likely to be in the 3 to 6 rating bracket.

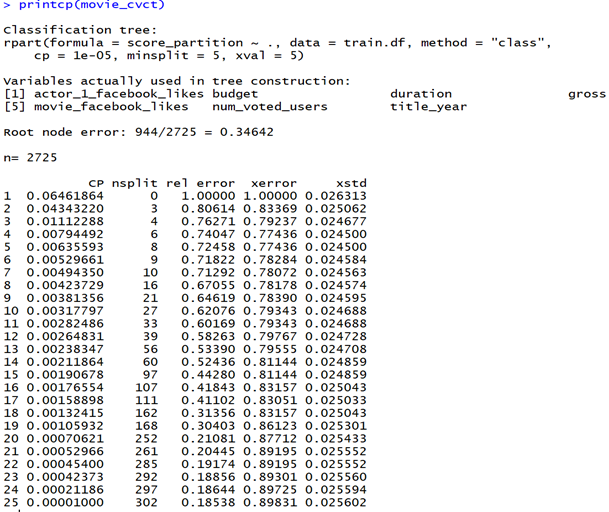
Step 5: Cross validation and complexity parameter:

Code:



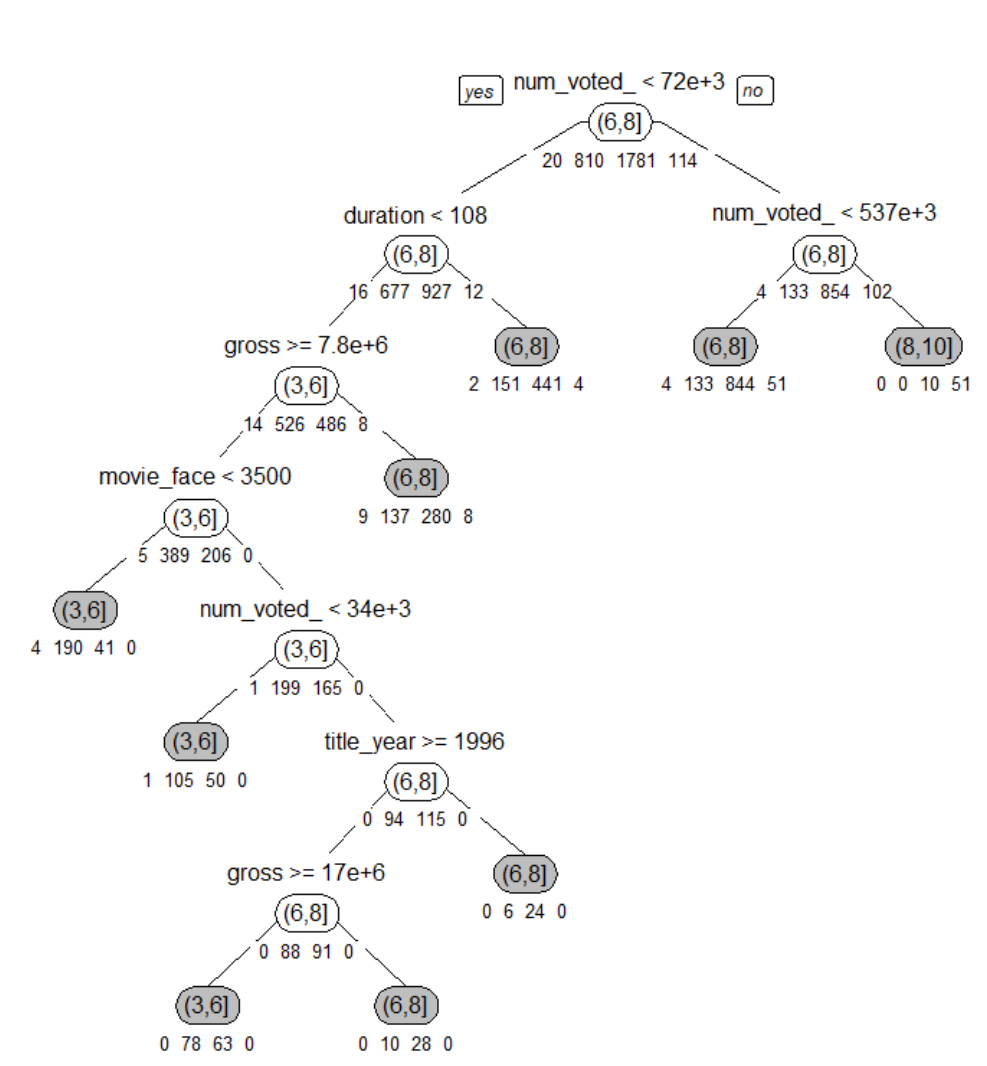
Here, we are trying to choose the complexity parameter (cp) for the tree model that we are creating. A smaller cp value is associated with a better tree.

Output 1:



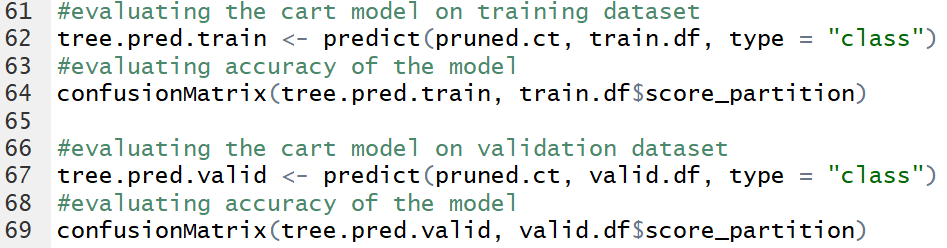
We see that the least xerror value is at cp = 0.0079, which is very small. On choosing the least cp value, we can see the following output:

Output 2:



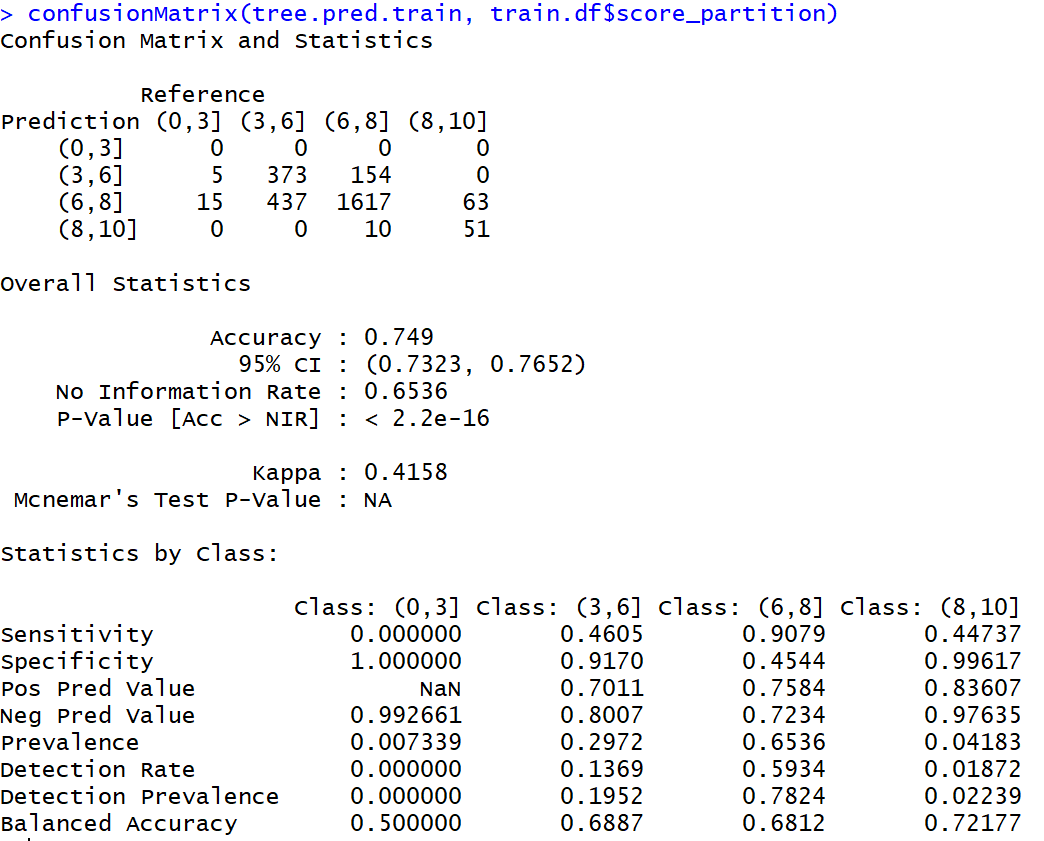
Step 6: Testing the accuracy of the tree that we have created

Code:



First, we apply the model we have created on the training dataset. We can see that the tree fits the training data fairly well with an accuracy of about 75%. The higher the accuracy, the better is the model.

Output 1:

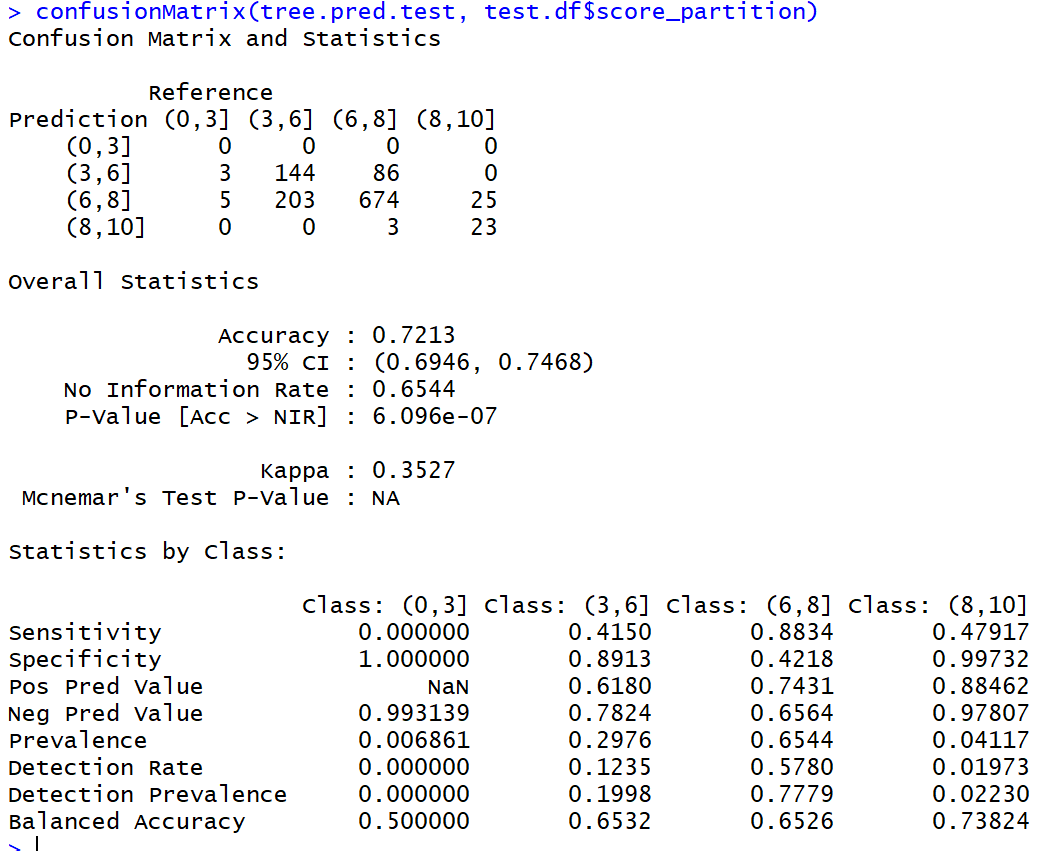


Often, applying the tree on the training dataset provides a higher accuracy simply due to overfitting.

Overfitting is the phenomenon in which the learning system tightly fits the given training data so much that it would be inaccurate in predicting the outcomes of the untrained data. [5]

In decision trees, overfitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. Thus, it ends up with branches with strict rules of sparse data. This affects the accuracy when predicting samples that are not part of the training set. Therefore, to test the real accuracy of the model, we need to apply it on the testing dataset or unseen data.

Output 2:



The application of the tree on the testing dataset also shows an accuracy of 72.13%.

# 8. Random Forest

## 8.1 Introduction

Random Forest is a widely used supervised learning technique in machine learning. It combines the output of multiple decision trees and then finally comes up with its own output. Random Forest works on the same principle as Decision Trees; however, it does not select all the data points and variables in each of the trees. It randomly samples data points and variables in each of the tree that it creates and then combines the output at the end. It removes the bias that a decision tree model might introduce in the system. Also, it improves the predictive power significantly.

## 8.2 Algorithm Steps

Step (1): Loading the appropriate library

Code:



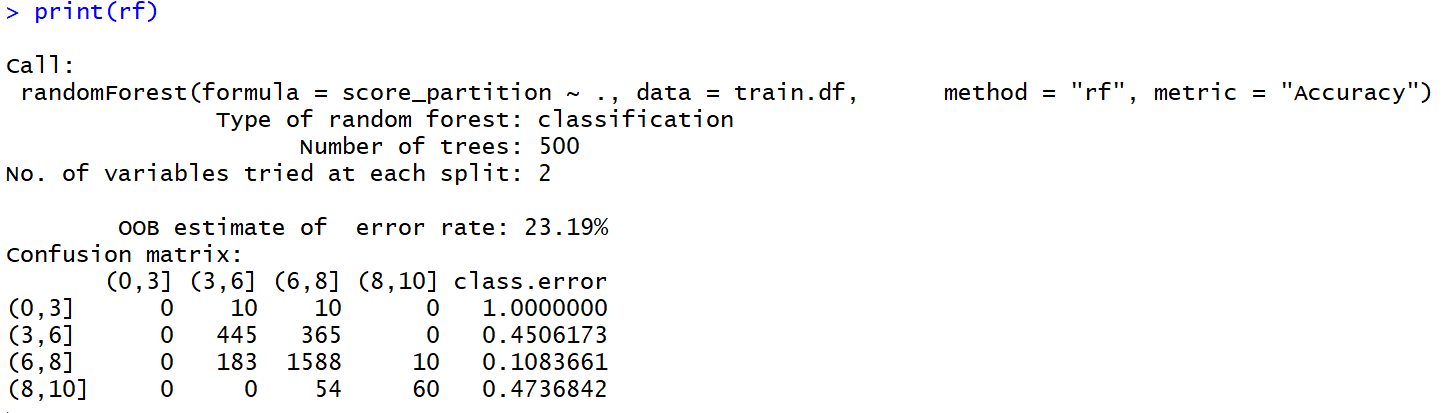
Step (2): Making the random forest model:

Code:



Here too, we are considering exactly the same variables as in the decision tree.

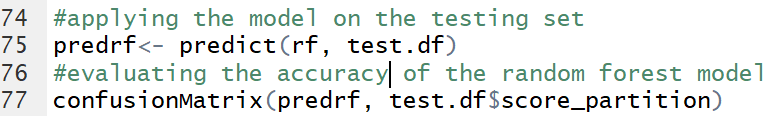
Output:



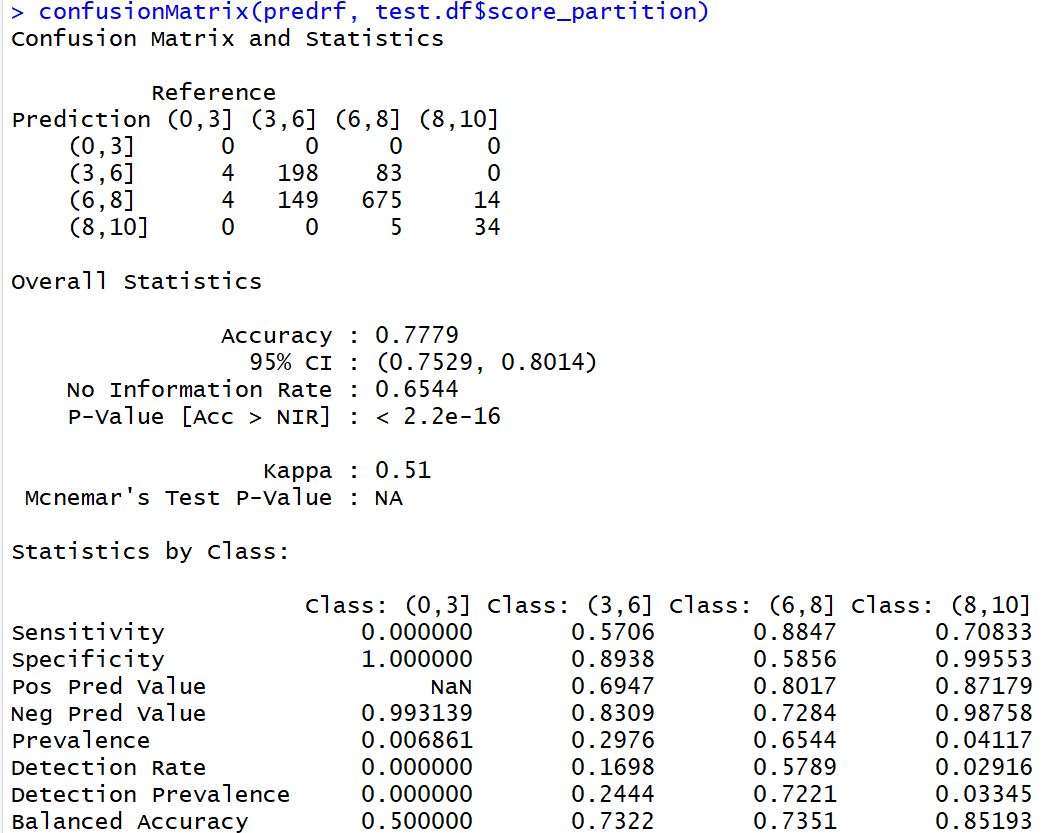
For the highest accuracy, random forest has selected the number of trees (ntree) to be 500.

Step (3): Using the model for predictions on testing dataset and evaluating the accuracy of this model:

Code:



Output:



The accuracy of the random forest model is 78% as compared to the 72% from the decision tree model.

# 9. Conclusion

This was an interesting data set and was fun to play around with. The first step was to understand the data properly. While exploring the data, we found many inconsistencies, which required to be corrected. The very first step was to remove duplicate rows as they may make our results inconsistent. Then we proceeded to find the predictors with the most missing values and then removing those missing values from the predictors. Next we dealt with the zero values in data before progressing to performing visualizations among various predictors.

From the correlation plot, we could interpret that: Face number in poster had negative correlation with all other predictors. Cast total facebook likes and actor 1 facebook likes had a stronger positive correlation. Interestingly, IMDB scores had strong positive correlation with number of critics for review, which means the more the critics review, the higher the score. Duration and number of voted users also had strong positive correlation with IMDB scores.

Based on the regression analysis, we inferred that all the variables were significant except for gross, facenumber\_in\_poster, movie\_facebook\_likes and director\_facebook\_likes. Performing regression analysis on multiple variables, we came up with the model that had the least AIC and highly significant values of R squared,p and F Statistic. We also predicted the values of iMDb score and came with a model that had values nearest to the actual values.

Through the decision tree and the random forest model, we were able to come up with an accuracy of around 78%. The models have worked out pretty well as the predicted scores are very close to the actual scores. However, we could probably be able to improve our model accuracy even more with the inclusion of newer movie observations.

# 10. Citations

* <http://www.secinfo.com/dr643.7kp.c.htm#2ndPage>
* <https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf>
* <https://www.datacamp.com/community/tutorials/decision-trees-R>
* <https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-decision-trees-c24dfd490abb>
* <https://statweb.stanford.edu/~lpekelis/talks/13_datafest_cart_talk.pdf>
* https://www.kaggle.com/carolzhangdc/analyze-imdb-score-with-data-mining-algorithms/notebook
* <https://www.coursera.org/lecture/ml-classification/overfitting-in-decision-trees-XcPVL>
* h[ttp://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decisio](http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decisio)n-tree-essentials/
* https://www.r-bloggers.com/how-to-implement-random-forests-in-r/

**INDIVIDUAL REPORT**

**Experience**

With an innate inclination and knack for subjects such as statistics, I was always drawn towards Regression Analysis and its proximity to real life data Analysis. The data that the group decided to work upon was predicting and analysing IMDb ratings based on the factors that affect it. This was something which really made me ecstatic as I was always curious to know what factors affect the overall rating of a movie. The project was a tremendous learning experience in terms of both technical expertise as well as personal skills and I would love to do such projects in future which make me privy to the real world.

**Group Dynamics**

The group dynamics stages were as follows:

**-Forming:** This involved forming sub-groups and assigning equitable tasks to each member of the group to ensure fairness.

**-Storming:** Brainstorming by producing an idea or way of solving a problem by holding frequent group discussion meetings.

**-Performing**: Performing the assigned tasks, creating group as well as individual reports and finally drawing out inferences from our respective analysis of the data.

**Success**

-After considering multiple models by using permutations and combinations of different variables, I finally came out to a conclusion to choose the best fitted model mentioned in the group report. The icing on the cake was the results which were as follows:

**R Squared = 64.08%**

-This shows that variables included in the model accounted for almost 64% of the variation in the model( Any number greater than 30-40% in considered good for a model)

**F Stat = 6.348 and p-value = 0.000000000000000022**

-The significance level of the model in terms of F-Statistic and p-value was enough for the model to be considered as a good fit.

**AIC = -1448.97.**

-The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data.  The AIC for this model is coming out to be -1448.97. This proves that the model is highly qualitative in nature

**Challenges**

-**Role Uncertainity** : The major challenge was assigning work to the group members because everyone was excited to perform Multiple Linear Regression Analysis because of their comprehensive understanding of the concept.

-**Too Many Independent Variables:** Which variable to consider/omit in the regression model was the most treacherous part since almost every variable was significant according to basic economic theory.

**Individual Performances**

**Somya Pandey** – Somya has worked hard on Data Pre-Processing and overall editing of the data. He is responsible for writing the background, cleaning the data, omitting missing values and checking for duplication in the data.

**Richa –** Richa has done a great job in doing classification and plotting regression trees. She has also plotted decision trees and inferred some great results based on that.

**Ritish Gupta** – Ritish has been really good at Data Visualization. He has plotted multiple correlation plots and heat maps. He has also done a great job in inferring the results.

**Xinchao** – Xinchao being exceptional at Data Visualization, has done some great work on plotting Word Clouds of Directors, Protagonist of the movie and country. They look catchy and at the same time they talk a lot about the data.