**IMDB 5000 movies Data Set Analysis**

-Anurag Maji, M10726841

**Dataset Description** :

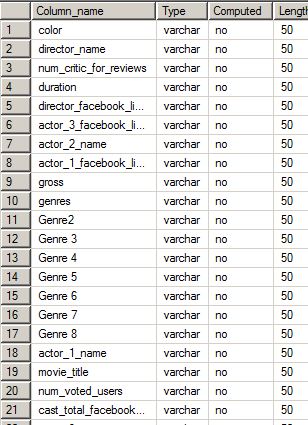
I have picked the IMDB Rating dataset from

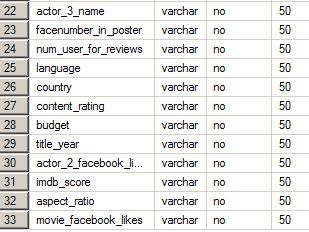
<https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>.

It contains information on movies, their directors, actors, popularity of actors and directors based on Facebook likes, IMDb rating of movies, budget and gross earnings of movies, genre of the movie, plot keywords, language and country of origin.

**Columns and data stored :**

**For SQL Analysis**





**For R and Tableau analysis**

|  |
| --- |
| 'data.frame': 4916 obs. of 36 variables:  $ color : Factor w/ 3 levels ""," Black and White",..: 3 3 3 3 3 3 3 3 3 3 ...  $ Director.Name : Factor w/ 2397 levels "","A. Raven Cruz",..: 928 800 2025 380 109 2028 1651 1227 554 2392 ...  $ num\_critic\_for\_reviews : int 723 302 602 813 462 392 324 635 375 673 ...  $ duration : int 178 169 148 164 132 156 100 141 153 183 ...  $ director\_facebook\_likes : int 0 563 0 22000 475 0 15 0 282 0 ...  $ actor\_3\_facebook\_likes : int 855 1000 161 23000 530 4000 284 19000 10000 2000 ...  $ actor\_2\_name : Factor w/ 3030 levels "","50 Cent","A. Michael Baldwin",..: 1408 2216 2486 534 2546 1228 801 2437 653 1703 ...  $ actor\_1\_facebook\_likes : int 1000 40000 11000 27000 640 24000 799 26000 25000 15000 ...  $ gross : int 760505847 309404152 200074175 448130642 73058679 336530303 200807262 458991599 301956980 330249062 ...  $ genre.1 : Factor w/ 21 levels "Action","Adventure",..: 1 1 1 1 1 1 2 1 2 1 ...  $ Genre.2 : Factor w/ 25 levels "","Adventure",..: 2 2 2 23 2 2 3 2 9 2 ...  $ Genre.3 : Factor w/ 24 levels "","Animation",..: 9 9 22 1 19 18 4 19 9 19 ...  $ Genre.4 : Factor w/ 18 levels "","Comedy","Crime",..: 14 1 1 1 1 1 5 1 11 1 ...  $ Genre.5 : Factor w/ 17 levels "","Crime","Drama",..: 1 1 1 1 1 1 5 1 1 1 ...  $ Genre.6 : Factor w/ 12 levels "","Family","Fantasy",..: 1 1 1 1 1 1 5 1 1 1 ...  $ Genre.7 : Factor w/ 9 levels "","Fantasy","Musical",..: 1 1 1 1 1 1 5 1 1 1 ...  $ Genre.8 : Factor w/ 3 levels "","Romance","Thriller": 1 1 1 1 1 1 1 1 1 1 ...  $ actor\_1\_name : Factor w/ 2095 levels "","50 Cent","A.J. Buckley",..: 305 981 355 1966 443 785 223 338 35 739 ...  $ movie\_title : Factor w/ 4916 levels "#Horror ","[Rec] 2 ",..: 398 2731 3279 3706 1961 3289 3458 399 1631 461 ...  $ num\_voted\_users : int 886204 471220 275868 1144337 212204 383056 294810 462669 321795 371639 ...  $ cast\_total\_facebook\_likes: int 4834 48350 11700 106759 1873 46055 2036 92000 58753 24450 ...  $ actor\_3\_name : Factor w/ 3520 levels "","50 Cent","A.J. Buckley",..: 3440 1394 3132 1770 2712 1969 2162 3016 2939 57 ...  $ facenumber\_in\_poster : int 0 0 1 0 1 0 1 4 3 0 ...  $ plot\_keywords : Factor w/ 4758 levels "","10 year old|dog|florida|girl|supermarket",..: 1320 4281 2076 3482 651 4743 29 1142 2005 1564 ...  $ movie\_imdb\_link : Factor w/ 4916 levels "http://www.imdb.com/title/tt0006864/?ref\_=fn\_tt\_tt\_1",..: 2964 2721 4531 3754 2476 2526 2458 4544 2551 4688 ...  $ num\_user\_for\_reviews : int 3054 1238 994 2701 738 1902 387 1117 973 3018 ...  $ language : Factor w/ 48 levels "","Aboriginal",..: 13 13 13 13 13 13 13 13 13 13 ...  $ country : Factor w/ 66 levels "","Afghanistan",..: 65 65 63 65 65 65 65 65 63 65 ...  $ content\_rating : Factor w/ 19 levels "","Approved",..: 10 10 10 10 10 10 9 10 9 10 ...  $ budget : num 2.37e+08 3.00e+08 2.45e+08 2.50e+08 2.64e+08 ...  $ title\_year : int 2009 2007 2015 2012 2012 2007 2010 2015 2009 2016 ...  $ actor\_2\_facebook\_likes : int 936 5000 393 23000 632 11000 553 21000 11000 4000 ...  $ imdb\_score : num 7.9 7.1 6.8 8.5 6.6 6.2 7.8 7.5 7.5 6.9 ...  $ aspect\_ratio : num 1.78 2.35 2.35 2.35 2.35 2.35 1.85 2.35 2.35 2.35 ...  $ movie\_facebook\_likes : int 33000 0 85000 164000 24000 0 29000 118000 10000 197000 ...  $ Profits : num 5.24e+08 9.40e+06 -4.49e+07 1.98e+08 -1.91e+08 ... |
|  |
| |  | | --- | | > | |

**\*Please note that the data set used for SQL analysis was slightly different since a few columns had to be updated/removed as there were complications in uploading process.**

**Normalization**

The given data set is not normalized as can be observed from the large variances in the gross, profits and budget column and relatively less variance in other numeric columns.

Also, since the entire data is from one data set, normalization by dividing into separate tables does not seem to be necessary. It could be normalized by putting it into separate tables for actors directors etc and creating a searate table for movies and its fatures like genre, gross revenue, earnings etc.

**Problems in the data set**

The data had a lot of bad column names. Many of the fields are blank, especially for budget, director and actor names. Also, the number of Facebook likes were given on a scale of 1000, but not mentioned (many values seem to be not too exact). Also, the ‘genre’ and ‘plot’ columns had a lot of keywords, out of which each keyword in the genre column has been separated for the purpose of analysis.

**Statistics generated using SQL**

The following statistics have been calculated using SQL. Many more statistics can be calculated, but these seemed to be a few prominent ones considering the data :

C:\Users\Anurag Maji\AppData\Local\Microsoft\Windows\INetCacheContent.Word\Capture3.jpg

C:\Users\Anurag Maji\AppData\Local\Microsoft\Windows\INetCacheContent.Word\Capture4.jpg

/\*a. Categorize films by running time into Short, Medium, Long and Very Long durations.\*/

SELECT

movie\_title

, Duration

, CASE

WHEN Duration <=90 THEN 'Short'

WHEN Duration<=150 THEN 'Medium'

WHEN Duration<=180 THEN 'Long'

ELSE 'Super Long'

END AS movie\_Duration

FROM

dbo.movie\_metadata;

/\*B. Movies with the same person as actor and director \*/

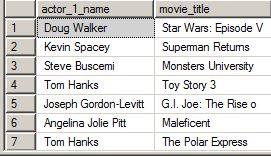
SELECT

actor\_1\_name, movie\_title, director\_name

FROM

dbo.movie\_metadata

WHERE actor\_1\_name = director\_name ;



/\*C Movies with greater than 100 million gross earnings and not produced in the US \*/

SELECT movie\_title, gross, country

from dbo.movie\_metadata

where gross > 100000000 and country <> 'USA';

/\*D. People who have been Actors and directors with the list of films they have acted in \*/

SELECT actor\_1\_name, movie\_title

FROM

dbo.movie\_metadata

WHERE actor\_1\_name in (

SELECT director\_name

FROM dbo.movie\_metadata

)

;

/\*E. Listing directors by the number of movies they have directed.

We can see that 104 movies do not have director names mentioned in the table.

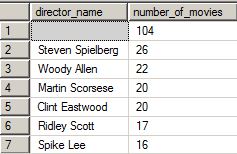
\*/

SELECT director\_name, count(movie\_title) as number\_of\_movies, sum(gross) as gross\_dir\_earning

from dbo.movie\_metadata

group by director\_name

order by count(movie\_title) DESC;



/\*F. Top 10 actors with the highest gross of individual movies, number of movies they have worked in,

and their respective gross of least earning movie \*/

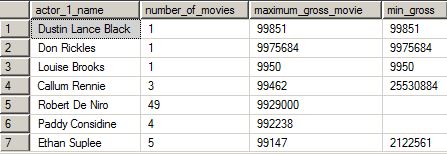
SELECT top 10 with ties actor\_1\_name, count(movie\_title) as number\_of\_movies,

max(gross) as maximum\_gross\_movie, min(gross) as min\_gross

from dbo.movie\_metadata

group by actor\_1\_name

order by max(gross) DESC;

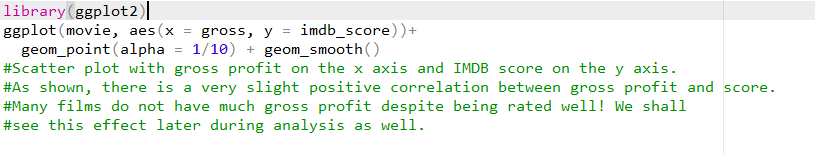


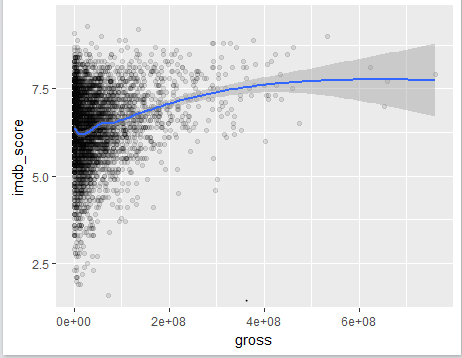
The codes for all of the basic SQL queries have been attached in a another file that will be submitted along with this file.

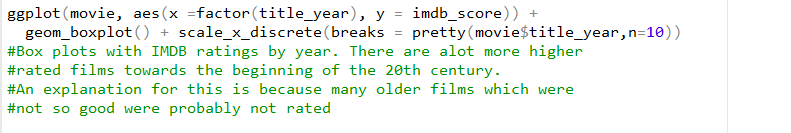
**Analysis on R**

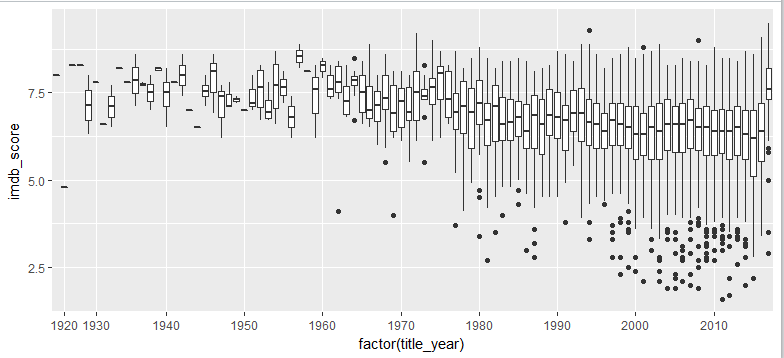
The commentary below is self explanatory.



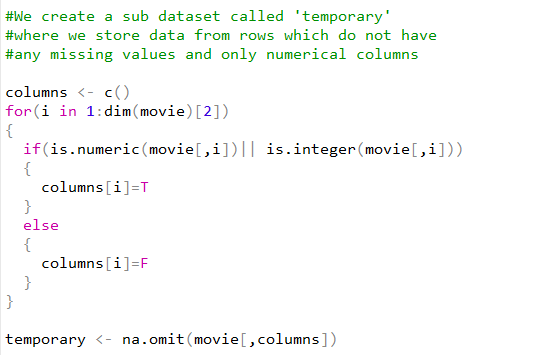


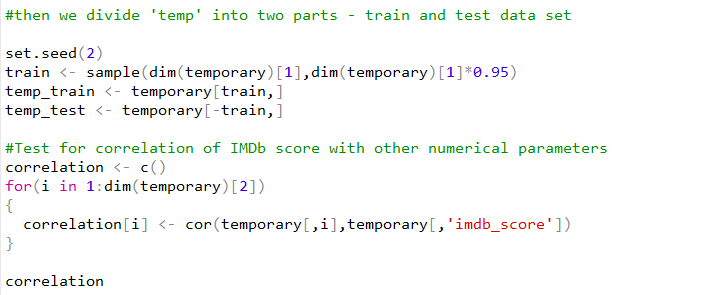


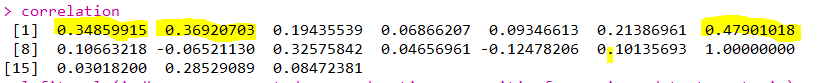




**Now, we try to form predictions for IMDb score**



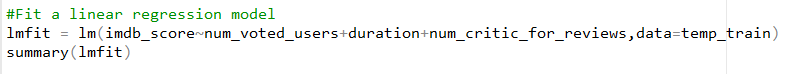


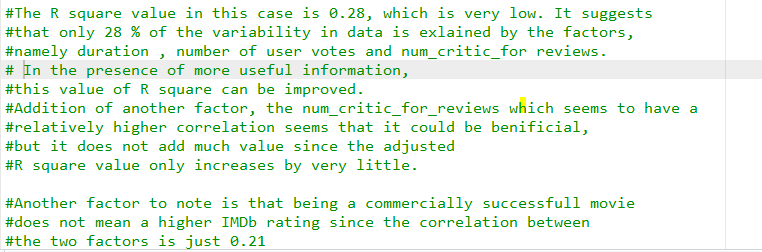


Variable 1 , Variable 2 and Variable 7, i.e ‘num\_critic\_for\_reviews’ , ‘Duration’, and ‘number of voted users’ seem slightly significant,

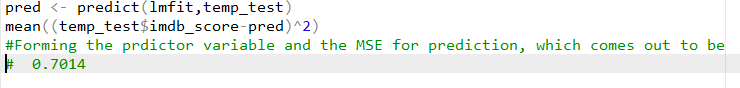
Although not much considering their correlation is not strong with **IMDb score.**

Hence a linear regression is performed considering these three variables.



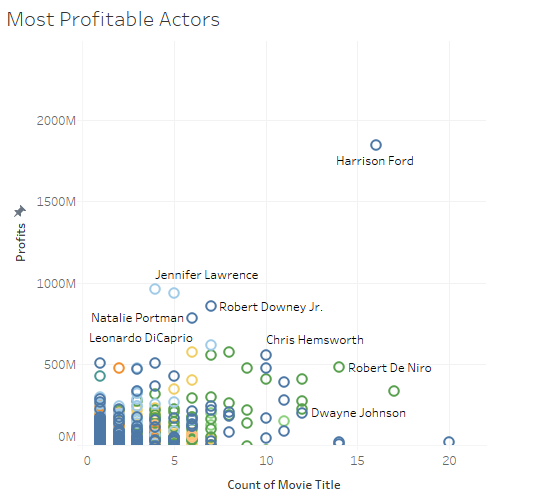
All the P values obtained are significant. But the R squared values are not much, showing these donot provide much explanation for the variance in the IMDb Scores.



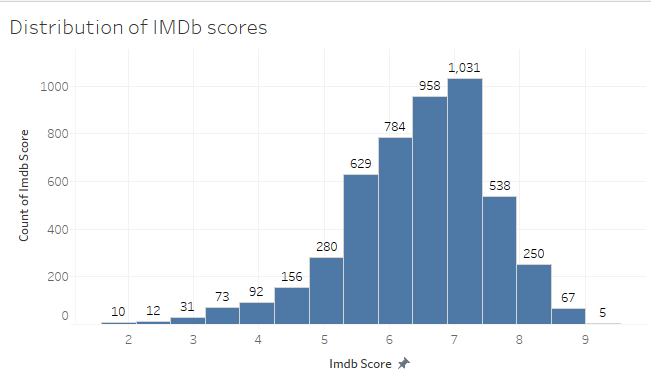


From the analysis above, it can be concluded that the variables present in the dataset are not good indicators of IMDb rating of a movie.

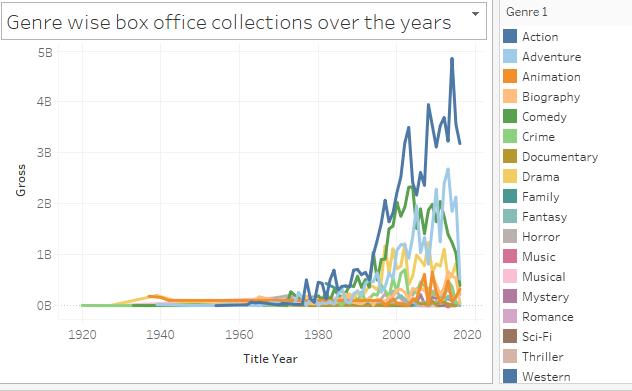
**Visual Analysis of Data on Tableau**



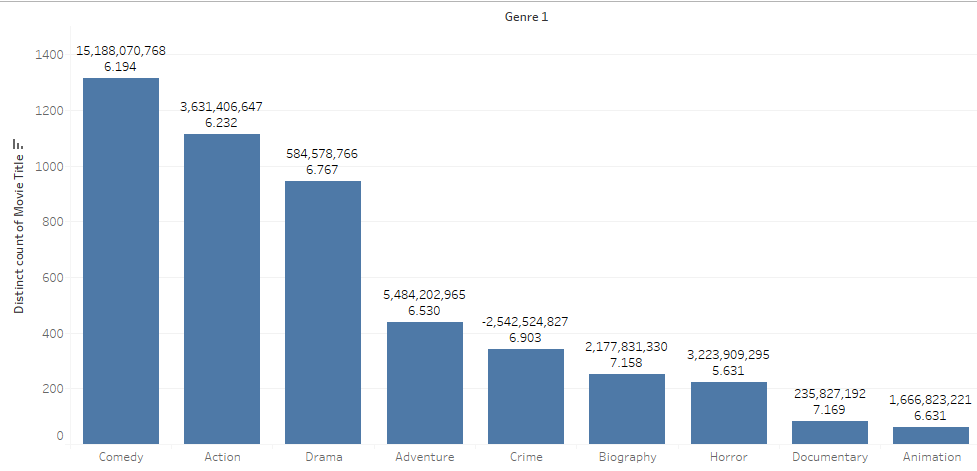
There are a few significant outliers in terms of number of movies acted in and the combined profits earned for such movies. We can see the most prominent ones in the visualization.



We see a distribution of IMDb scores from the graph. As observed, the majority of the scores are around 7 and the graph follows an approximately normal distribution.



In the genre wise box office collections over the years, we see that the genre ‘Action’ clearly dominates over all other genres. Also, the difference between Action and other genres has been increasing over the years.



Comparison of top genres (with total genre wise wise profits and average IMDb rating) shows that although the earnings for action movies has increase seignifiantly as observed above, due to very high budgets, it has been able to amass only 3 billion of profits over time as compared to 15 billion of profits from Comedy movies.

Also, Crime has one of the highest average IMDB rating but has negative profits!

**Insights**

From the overall analysis of the IMDb data set using Tableau, R and SQL queries the following key points can be noted

1. The correlation between IMDb ratings and other variables such as Box office collections, budget of the movie, popularity of the actors and directors ( using FB likes as the criteria for popularity) is very low. Hence none of these individually and even a combination of these do not form a reliable method for prediction of IMDb ratings.
2. Variation in IMDb ratings have increased over the years as seen from the boxplot using R. A possible reason might be the fact that more number of recently made movies are reviewed (both good and bad) as compared to older classics (mostly good)

This is also indicated by the fact that older movies have a higher IMDb rating in general.

1. The IMDb ratings for movies follow an approximately normal curve, with mostly lying close to 7.0.
2. Action dominates all other genres in term of total box office collections.

But in terms of number of movies produced and profitability, ‘Comedy’ leads all Genres.

**Summary-**

Although it is difficult ot break down movie making into a science, some recommendations that be given to a production company are that

* If you want higher box office returns, go for Action movies. They would make huge amounts of money, but the budgets would be equally high.
* For lower budget movies which make significant amount of money, Comedy is a preferable genre.
* The box office collections do not play a vital role in deciding the IMDb rating. Even if your movie does not make a significant amount of money, it might be have a good rating, i.e for the classes and not the masses.
* If you make an average effort, you should expect your movie to be rated arounf 6.8 or 7.0 on IMDb.

**Challenges faced**

* Data Cleaning was a major challenge since this data set ha a lot of unclean data which had to be formatted properly.
* It was difficult to choose predictors for IMDb score since all parameters had a very low correlation with IMDb score.

Although ultimately, even the parameters chosen were not enough to rovide significant predictions.