T TRAINITY

PROJECT: IMDB MOVIE ANALYSIS

SUBMITTED BY: MUDIT VYAS

Video Link: Click Here!

Aim: Use Excel Workbench to pre-process data (handle missing values, remove duplicates, data type conversion, etc.), explore the data to understand relation between different variables and provide valuable insights by using visualizations to help tell stories which assist to elaborate the finding in much better way.

Project Overview: People nowadays are pretty cautious before watching a movie. They tend to wait for reviews and decide whether to watch it or not. Many platforms have emerged to help people with ratings, reviews and details about movies and one of the most popular and trusted Sight is IMDB. Getting good reviews and ratings on IMDB is a new focus point for many movie production houses as it became major reason to attract crowd followed by mouth publicity.
Therefore in this project we will focus on how different factors affect IMDB rating and what improvements can be brought to improve the ratings. With the help of descriptive Statistics data and visualisation we shall disclose some possibilities.

Tech-Stack Used: Microsoft Excel 2016 is used however it's recommended to use latest Version of Excel, but the 2016 version fulfilled the requirements as per the project.

Data Pre-processing/Data Cleaning: Before Analysis we check the dataset provided. Certainly it needs to be cleaned and missing values to be handled.

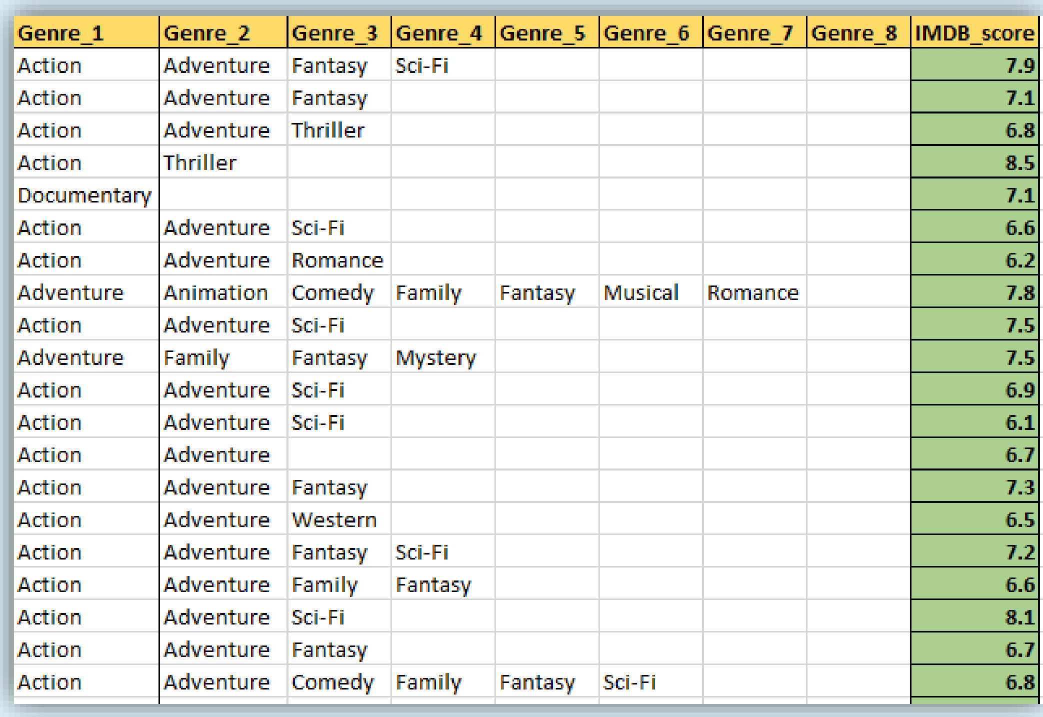
- We observe each column contains unique data and missing values in each column is different in counts.
- However, if we clean the entire data overall, we might omit useful data which is important for a particular analysis.
- So we decide to focus on cleaning the specific data as per the problem statement which shall not affect the analysis of other problems.
- If there is some relation and data might get affected by other columns/factors, those factors are considered while selective data cleaning.
- As overall cleaning, in 'movie_title' column the extra symbol 'Â 'was removed. Columns were rearranged to have a structured data. Some of the rows with almost no or abrupt data were removed. Duplicates rows are also removed.

What type of movie is generally liked by people?

Let’s observe the distribution of IMDB score across genres. Generally a wave of popular genre as per time occurs. While In late 90s and early 2000s household drama used to be trending, now a days people like to watch crime thrillers.

Data Cleaning:

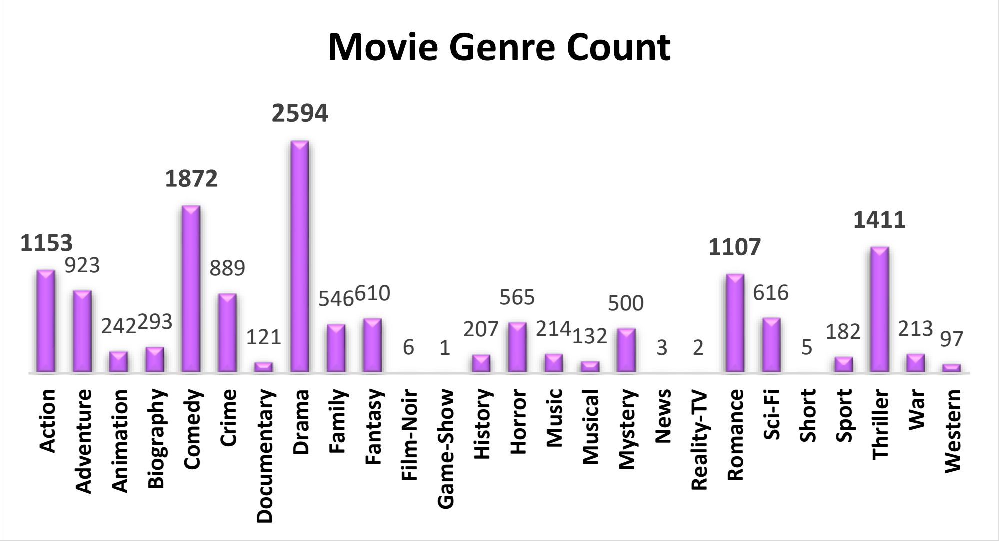
- We use ‘genre’ and ‘imdb_score’ column for this analysis.
- There were no blank data so no requirement to handle the missing data.
- Used ‘Text-to-column’ tab in Data Menu we split ‘genre’ column to get different genre if contained in a single cell.

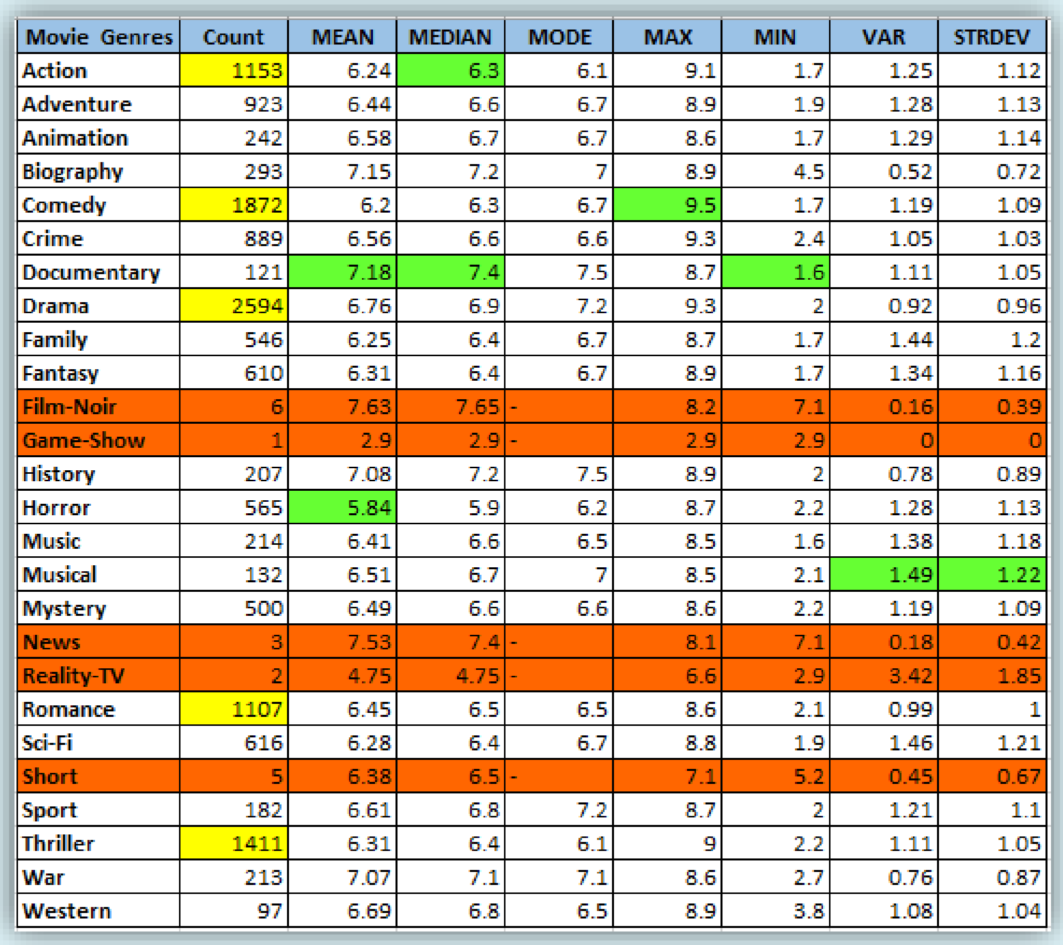


Genre_1	Genre_2	Genre_3	Genre_4	Genre_5	Genre_6	Genre_7	Genre_8	IMDB_score
Action	Adventure	Fantasy	Sci-Fi					7.9
Action	Adventure	Fantasy						7.1
Action	Adventure	Thriller						6.8
Action	Thriller							8.5
Documentary								7.1
Action	Adventure	Sci-Fi						6.6
Action	Adventure	Romance						6.2
Adventure	Animation	Comedy	Family	Fantasy	Musical	Romance		7.8
Action	Adventure	Sci-Fi						7.5
Adventure	Family	Fantasy	Mystery					7.5
Action	Adventure	Sci-Fi						6.9
Action	Adventure	Sci-Fi						6.1
Action	Adventure							6.7
Action	Adventure	Fantasy						7.3
Action	Adventure	Western						6.5
Action	Adventure	Fantasy	Sci-Fi					7.2
Action	Adventure	Family	Fantasy					6.6
Action	Adventure	Sci-Fi						8.1
Action	Adventure	Fantasy						6.7
Action	Adventure	Comedy	Family	Fantasy	Sci-Fi			6.8

Descriptive Analysis:

- Using COUNTIF() function total count per genre is calculated
- Several Statistics parameters such as Mean, Mode, Median, Max, Min and Standard Deviation are calculated with Statistic standard Formulae(such as MAX() , AVERAGE(), MODE.SNGL() ,etc.) incorporating with other functions such as IF(), IFERROR(), ROUND().





Movie Genres	Count	MEAN	MEDIAN	MODE	MAX	MIN	VAR	STRDEV
Action	1153	6.24	6.3	6.1	9.1	1.7	1.25	1.12
Adventure	923	6.44	6.6	6.7	8.9	1.9	1.28	1.13
Animation	242	6.58	6.7	6.7	8.6	1.7	1.29	1.14
Biography	293	7.15	7.2	7	8.9	4.5	0.52	0.72
Comedy	1872	6.2	6.3	6.7	9.5	1.7	1.19	1.09
Crime	889	6.56	6.6	6.6	9.3	2.4	1.05	1.03
Documentary	121	7.18	7.4	7.5	8.7	1.6	1.11	1.05
Drama	2594	6.76	6.9	7.2	9.3	2	0.92	0.96
Family	546	6.25	6.4	6.7	8.7	1.7	1.44	1.2
Fantasy	610	6.31	6.4	6.7	8.9	1.7	1.34	1.16
Film-Noir	6	7.63	7.65	-	8.2	7.1	0.16	0.39
Game-Show	1	2.9	2.9	-	2.9	2.9	0	0
History	207	7.08	7.2	7.5	8.9	2	0.78	0.89
Horror	565	5.84	5.9	6.2	8.7	2.2	1.28	1.13
Music	214	6.41	6.6	6.5	8.5	1.6	1.38	1.18
Musical	132	6.51	6.7	7	8.5	2.1	1.49	1.22
Mystery	500	6.49	6.6	6.6	8.6	2.2	1.19	1.09
News	3	7.53	7.4	-	8.1	7.1	0.18	0.42
Reality-TV	2	4.75	4.75	-	6.6	2.9	3.42	1.85
Romance	1107	6.45	6.5	6.5	8.6	2.1	0.99	1
Sci-Fi	616	6.28	6.4	6.7	8.8	1.9	1.46	1.21
Short	5	6.38	6.5	-	7.1	5.2	0.45	0.67
Sport	182	6.61	6.8	7.2	8.7	2	1.21	1.1
Thriller	1411	6.31	6.4	6.1	9	2.2	1.11	1.05
War	213	7.07	7.1	7.1	8.6	2.7	0.76	0.87
Western	97	6.69	6.8	6.5	8.9	3.8	1.08	1.04

Insights:

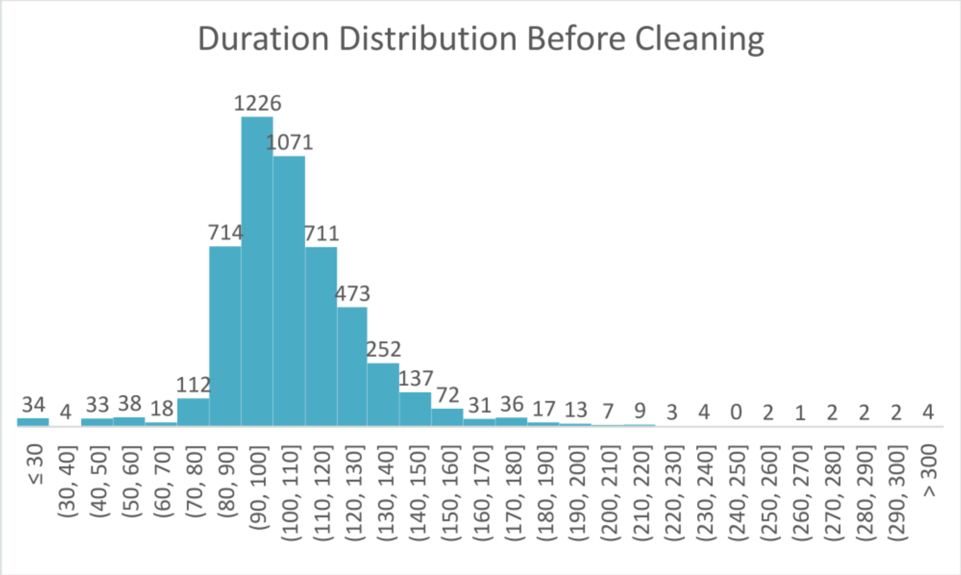
- ‘Drama’ is most preferred genre by movie makers followed by Comedy, Thriller, Action and Romance.
- Genres like Film-Noir, Game-Show, News, Reality-TV and Short have pretty less data which shall not give any general conclusion. Moreover, these are considered mainly as format of content rather than genre. Therefore we omit their analysis.
- ‘Action’ Genre have close Average and Median values, yet there is huge different between Maximum and Minimum Values. Large Variance value shows Action Movies are either liked a lot or rejected straight away. It largely depends upon Lead Actor, Director, action sequence/ choreography and may be storyline.
- ‘Documentary’ Genre have Highest Mean and Median Values. Period Movies or biography of a famous/ historic personality are generally liked by people. As stories as history are available, the creators just have to refine and present it in most glorified intense manner. However, the lowest ranking of 1.6 is also bagged by this genre, means not all stories are good to tell.
- Likewise, other specific genre can also be explained. Now genre selection can also be affected by year in which it was selected, duration of movie, Director and many other factors.

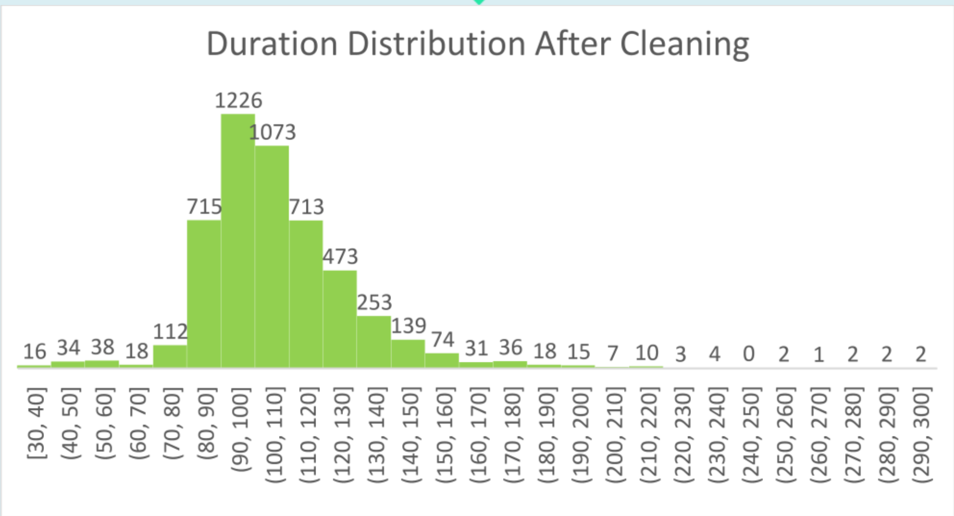
How does Duration of Movie Emphasize Ratings?

Content is the king but sometimes longer duration fails to hold audience with the storyline. While in early 2000s movies used to be longer than 3 hours which included long song sequence and screenplays. Now a days in OTT era people like fast paced movies with engaging content. Let's see how Duration of Movies influence IMDB ratings.

Data Cleaning:

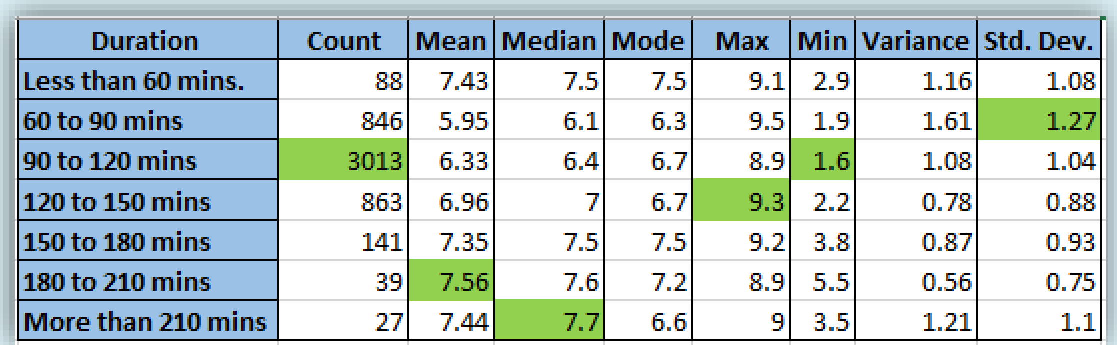
- We use 'Duration' and 'imdb_score' columns for this analysis.
- There are 15 blank columns in 'Duration', which won't effect on overall large scale data. So we remove those rows.
- We check outliers using Histogram. Movies with less than 15 minutes and more than 300 minutes makes no sense. So we remove those data as well.
- For scores with blank duration, since there were unignorable number of records duration were randomly generated using RANDBETWEEN() function.
- In order to Standardise Analysis, we take movie duration between 30 min and 300 minutes. It's certainly unlikely to have movies more than 4 hours long, but in 1980s 1990s there were a few good movies made of such large duration.


Duration Distribution Before Cleaning

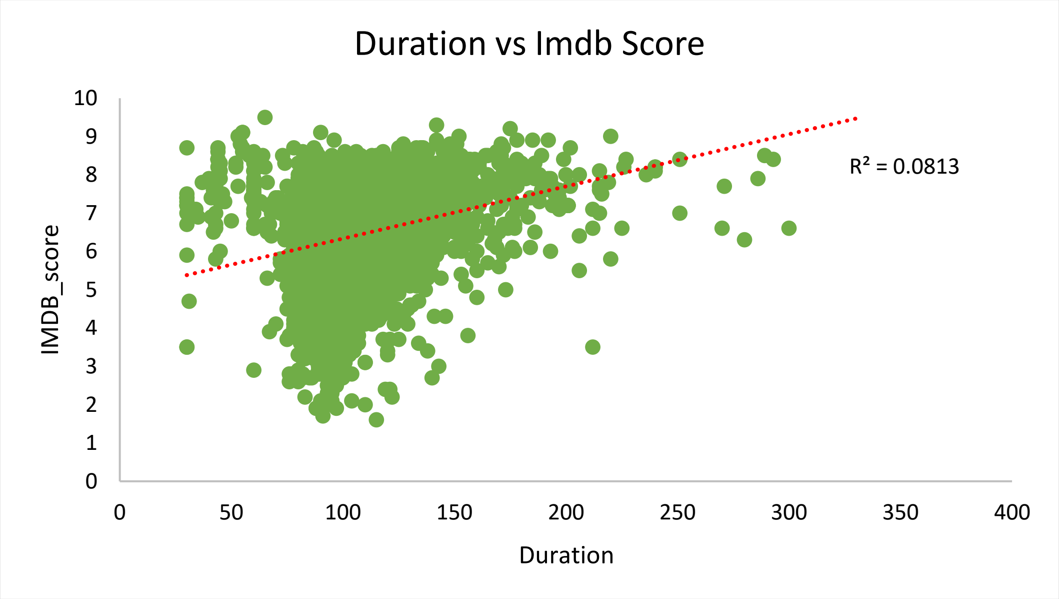

Duration Distribution After Cleaning

Descriptive Analysis:

- We categorize data based on popular time duration. Using COUNTIF() function calculate number of movies in each duration slot.
- Several Statistics parameters such as Mean, Mode, Median, Max, Min and Standard Deviation are calculated with Statistic standard Formulae(such as MAX() , AVERAGE(), MODE.SNGL() ,STDEV.P() etc.) incorporating with other functions such as IF(), IFERROR(), ROUND().



Duration	Count	Mean	Median	Mode	Max	Min	Variance	Std. Dev.
Less than 60 mins.	88	7.43	7.5	7.5	9.1	2.9	1.16	1.08
60 to 90 mins	846	5.95	6.1	6.3	9.5	1.9	1.61	1.27
90 to 120 mins	3013	6.33	6.4	6.7	8.9	1.6	1.08	1.04
120 to 150 mins	863	6.96	7	6.7	9.3	2.2	0.78	0.88
150 to 180 mins	141	7.35	7.5	7.5	9.2	3.8	0.87	0.93
180 to 210 mins	39	7.56	7.6	7.2	8.9	5.5	0.56	0.75
More than 210 mins	27	7.44	7.7	6.6	9	3.5	1.21	1.1



Insights:

- About 60% (3010 in count) of movies are made in duration 90 to 120 minutes. However the variation and lowest average is seen most in this slot itself. With such a short duration movie duration movie needs to be very fine and filtered which either is highly liked by people or straight away rejected.
- Stats for long duration movies are pretty high since the count of such movies pretty low, the stats gathered are such. In such cases, the genre of movies, direction and writing plays a big role. For example, the latest release Adipurush failed miserably due to poor dialogues and screenplay, while movies like Padmaavat, Lord of the Rings were appreciated a lot.
- The Coefficient of Determination (R^2) is calculated as 0.0813 which shows poor interpretation and prediction of data. This shows that by given data it's difficult to predict how movies will perform based on duration. Maybe if we ignore a slot of duration the R^2 value may improve.

Does language have an Impact on Ratings?

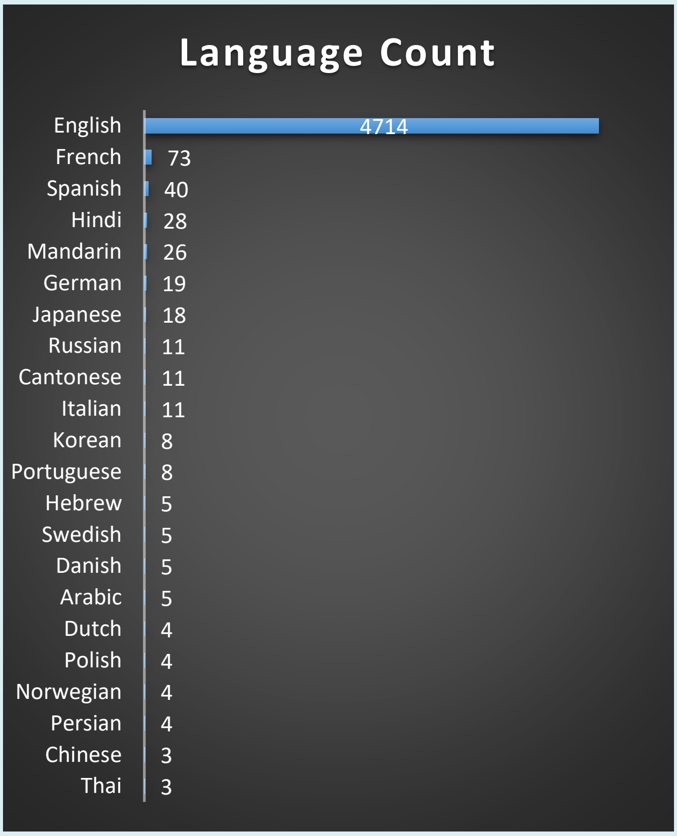
If movie is approachable to larger group of audience the profits might increase, the ratings might improve. Language plays a great role in that context. In earlier decades Movies were limited regional only due to language barrier but now with subtitles, dubbing and approach to learn new languages have made any movie International. Let's observe the effect of Language on ratings.

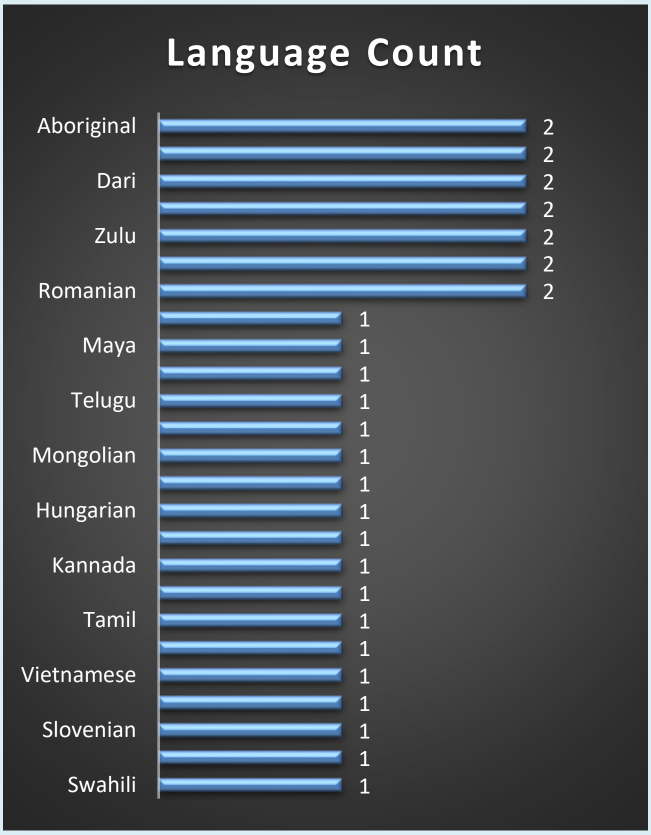
Data Cleaning:

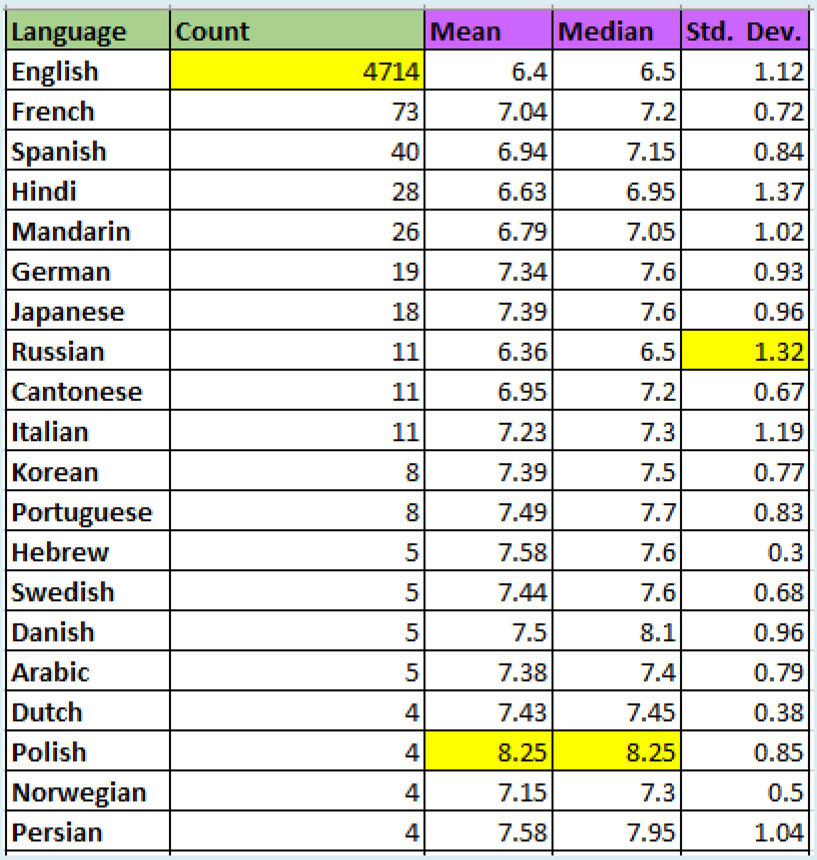
- For this task 'Language' and 'imdb_score' columns are used.
- There looks no text error. For missing cells, we check country the movie is made in. All the missing data are from USA, so we fill them with 'English' Language.

Descriptive Analysis:

- Using COUNTIF() function count per Language is calculated.
- Several Statistics parameters such as Mean, Median, and Standard Deviation are calculated with Statistic standard Formulae(such as AVERAGE(), MEDIAN() ,STDEV.P() etc.) incorporating with other functions such as IF(), IFERROR(), ROUND().
- We compare statistic data only for language with count more than 3, as statistics is much understandable for higher data count.







Language	Count	Mean	Median	Std. Dev.
English	4714	6.4	6.5	1.12
French	73	7.04	7.2	0.72
Spanish	40	6.94	7.15	0.84
Hindi	28	6.63	6.95	1.37
Mandarin	26	6.79	7.05	1.02
German	19	7.34	7.6	0.93
Japanese	18	7.39	7.6	0.96
Russian	11	6.36	6.5	1.32
Cantonese	11	6.95	7.2	0.67
Italian	11	7.23	7.3	1.19
Korean	8	7.39	7.5	0.77
Portuguese	8	7.49	7.7	0.83
Hebrew	5	7.58	7.6	0.3
Swedish	5	7.44	7.6	0.68
Danish	5	7.5	8.1	0.96
Arabic	5	7.38	7.4	0.79
Dutch	4	7.43	7.45	0.38
Polish	4	8.25	8.25	0.85
Norwegian	4	7.15	7.3	0.5
Persian	4	7.58	7.95	1.04

Insights:

- An outstanding Count of English Movies! That's because world's biggest Film industries Such as Hollywood and British Cinema are English based. Due to huge count certainly the average rating went down with large spread in ratings.
- The French and Spanish Cinema have also performed well in ratings. That may be because the genre is mostly action/drama with average duration about 110 minutes.
- Hindi Cinema is much onto Entertainment and Action side, has average pretty less than other languages. With a Deviation of 1.37 it shows that movies are either appreciated a lot or completely thumbs down.
- Although there are just 4 entries for Polish language movies, the data shows symmetric normal distribution with deviation of 0.85, meaning all 4 release are pretty liked by the audience.

How Much Director's Name Influence IMDB Rating?

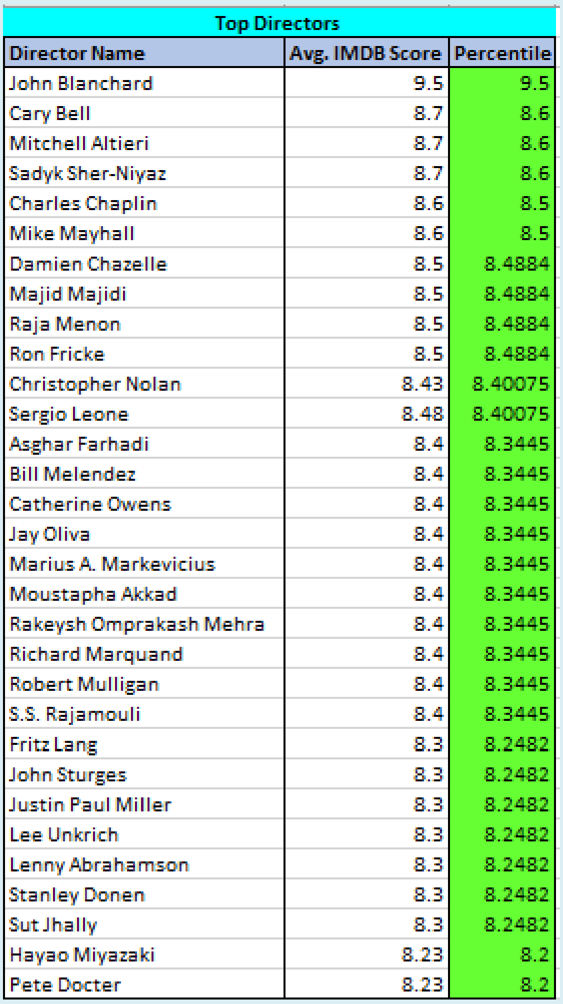
Along with superstar's name and songs released, Director have major Impact on craze of a movie. The reasons could be way of direction, the topics they choose, the intensity, the vision and cinematography one presents. While we see Craze for Nolan's Movie and Social Drama movies from Shoojit Sarcar are highly rated inspite of duration, some directors with vague story and dull comedy scripts fail to entertain people. Let's have a look on top directors.

Data Cleaning:

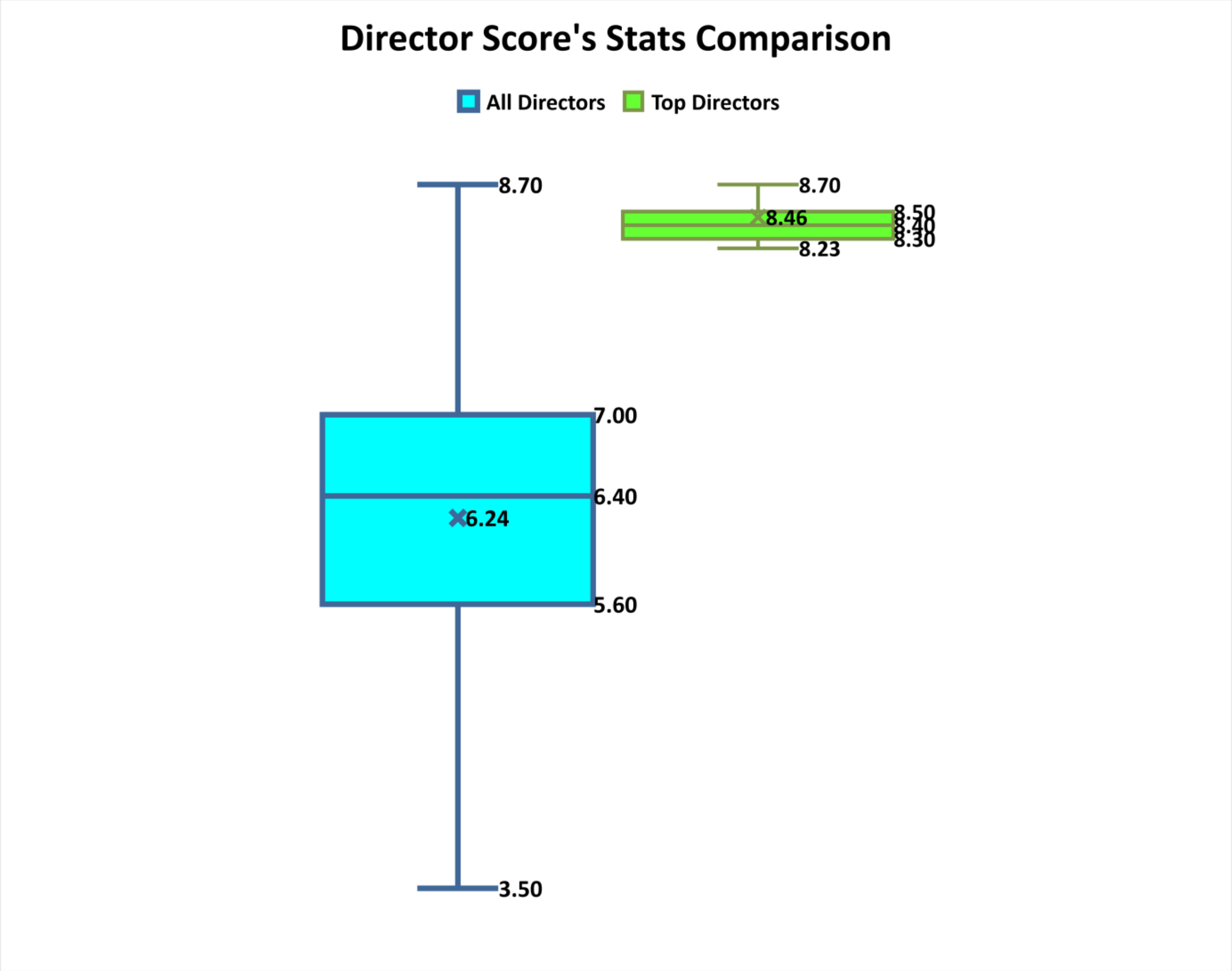
- We choose ‘content rating’, ‘director’ and ‘imdb_score’ columns for analysis.
- We observe that TV shows director names are missing as it’s understood to have multiple directors for different seasons. So we remove those rows from data set.
- This leaves us with few rows with missing directors, whose removal won’t affect the analysis on overall basis, so we leave ‘director’ column with no blanks.

Descriptive Analysis:

- Using AVERAGE() function along with IF() and ROUND() function average rating per director is calculated.
- To find out top Directors PERCENTILE.INC() function is used incorporated with PERCENTRANK.INC() function. Using conditional formatting Top 18 percentile directors based on their average IMDB rating are separated.
- Box plot is used to visualize comparison between all directors and top directors stats.



Top Directors		
Director Name	Avg. IMDB Score	Percentile
John Blanchard	9.5	9.5
Cary Bell	8.7	8.6
Mitchell Altieri	8.7	8.6
Sadyk Sher-Niyaz	8.7	8.6
Charles Chaplin	8.6	8.5
Mike Mayhall	8.6	8.5
Damien Chazelle	8.5	8.4884
Majid Majidi	8.5	8.4884
Raja Menon	8.5	8.4884
Ron Fricke	8.5	8.4884
Christopher Nolan	8.43	8.40075
Sergio Leone	8.48	8.40075
Asghar Farhadi	8.4	8.3445
Bill Melendez	8.4	8.3445
Catherine Owens	8.4	8.3445
Jay Oliva	8.4	8.3445
Marius A. Markevicius	8.4	8.3445
Moustapha Akkad	8.4	8.3445
Rakeysh Omprakash Mehra	8.4	8.3445
Richard Marquand	8.4	8.3445
Robert Mulligan	8.4	8.3445
S.S. Rajamouli	8.4	8.3445
Fritz Lang	8.3	8.2482
John Sturges	8.3	8.2482
Justin Paul Miller	8.3	8.2482
Lee Unkrich	8.3	8.2482
Lenny Abrahamson	8.3	8.2482
Stanley Donen	8.3	8.2482
Sut Jhally	8.3	8.2482
Hayao Miyazaki	8.23	8.2
Pete Docter	8.23	8.2



Insights:

- While top directors have a smaller variation, we see along with other directors the data is pretty spread.
- There is a huge difference in Average values. For all directors the average rating is 6.24 while top directors play around 8.5 rating, with Mean above Median value.
- This shows the impact of top directors overall. With just name or declaration of movie by them the conversation among people begins.
- Hence if a production house is looking forward for a blockbuster movie, they should approach notable directors with good reviews and ratings.

Blockbuster, Superhit or Flop? : The Profit Report

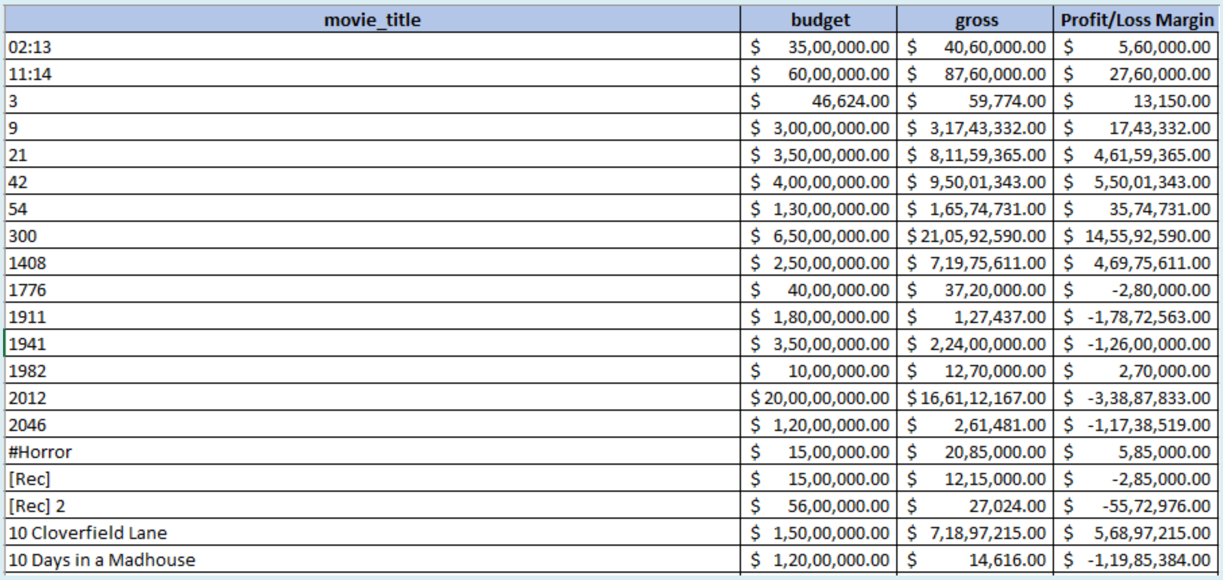
All the combinations, all the hard work boils down to this factor- how much did movie earn? How much Profit did it make? With all the investments and expenses film industry is one of the costliest yet most profitable business in world. And success of a movie is now a days measured in terms if profit they made. Let’s have a look on how movies performed on box office and conclude the Analysis.

Data Cleaning:

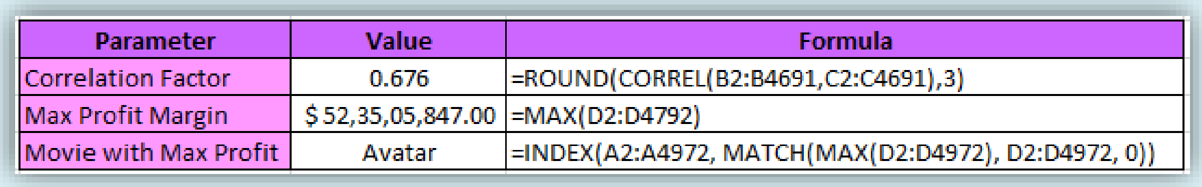
- We choose ‘movie_title’, ‘budget’ and ‘gross’ columns for this analysis.
- Rows with neither of budget and gross data are of no use. We can’t randomly put the data as we don’t know the scale at which movie is made.
- For rows with missing budget data, we assume the budget could be between 85% and 120% of gross collection shows movie might have performed well or failed. So we use RANDBETWEEN() function fill those missing data. Similar approach is taken for gross missing data.
- Some of the budget or gross data is a 3 digit number. It’s assumed that any figure below $7000 is not practical. So we omit those entries as well.
- Format of entries is changed from general to Currency (USD).

Descriptive Analysis:

- Profit/Loss Margin is calculated by subtracting Budget from Gross collection.
- Correlation Coefficient tells us relation between two variables. Using CORREL() function we calculate how strong relation is between Budget and Gross collection.
- Using MAX() function , the highest profit among all movies is calculated.
- To find out movie with Highest profit, INDEX() AND MATCH() functions are used.
  =INDEX(A2:A4972, MATCH(MAX(D2:D4972), D2:D4972, 0))
  Where column A contents movie titles and Column D contains Profit/Loss Margin.



movie_title	budget	gross	Profit/Loss Margin
02:13	$ 35,00,000.00	$ 40,60,000.00	$ 5,60,000.00
11:14	$ 60,00,000.00	$ 87,60,000.00	$ 27,60,000.00
3	$ 46,624.00	$ 59,774.00	$ 13,150.00
9	$ 3,00,00,000.00	$ 3,17,43,332.00	$ 17,43,332.00
21	$ 3,50,00,000.00	$ 8,11,59,365.00	$ 4,61,59,365.00
42	$ 4,00,00,000.00	$ 9,50,01,343.00	$ 5,50,01,343.00
54	$ 1,30,00,000.00	$ 1,65,74,731.00	$ 35,74,731.00
300	$ 6,50,00,000.00	$ 21,05,92,590.00	$ 14,55,92,590.00
1408	$ 2,50,00,000.00	$ 7,19,75,611.00	$ 4,69,75,611.00
1776	$ 40,00,000.00	$ 37,20,000.00	$ -2,80,000.00
1911	$ 1,80,00,000.00	$ 1,27,437.00	$ -1,78,72,563.00
1941	$ 3,50,00,000.00	$ 2,24,00,000.00	$ -1,26,00,000.00
1982	$ 10,00,000.00	$ 12,70,000.00	$ 2,70,000.00
2012	$ 20,00,00,000.00	$ 16,61,12,167.00	$ -3,38,87,833.00
2046	$ 1,20,00,000.00	$ 2,61,481.00	$ -1,17,38,519.00
#Horror	$ 15,00,000.00	$ 20,85,000.00	$ 5,85,000.00
[Rec]	$ 15,00,000.00	$ 12,15,000.00	$ -2,85,000.00
[Rec] 2	$ 56,00,000.00	$ 27,024.00	$ -55,72,976.00
10 Cloverfield Lane	$ 1,50,00,000.00	$ 7,18,97,215.00	$ 5,68,97,215.00
10 Days in a Madhouse	$ 1,20,00,000.00	$ 14,616.00	$ -1,19,85,384.00



Parameter	Value	Formula
Correlation Factor	0.676	=ROUND(CORREL(B2:B4691,C2:C4691),3)
Max Profit Margin	$ 52,35,05,847.00	=MAX(D2:D4792)
Movie with Max Profit	Avatar	=INDEX(A2:A4972, MATCH(MAX(D2:D4972), D2:D4972, 0))

Insights:

- Correlation coefficient comes out to be 0.676 which shows a positive relation between Budget and Gross Earnings. Most of the movies have been towards profit side.
- AVATAR movie earns highest profit of $523.5M. A James Cameron Masterpiece which revolutionized Film Industry that time.

Conclusions

- Film Industry is recognized as Business Industry now and with that comes profit and loss scenarios. While in earlier days, if we specifically talk about Indian Cinema the lead actor-actress pair was more than enough to pull crowd to watch the movies.
- But now since there is so much advancement people think multiple times to watch a movie. That is where movie production went with thorough analysis of data
- From this Analysis we figured out that how even a single factor can impact other factors and overall ratings/viewpoint of cinema.
- Genre selection depends on the target audience. A lot of people goes with Drama, Thriller and comedy yet it might be different with certain language particular cinema where even Sci-Fi or Horror movies work. That largely depends on movie industry and Director.
- Duration shall be chosen based on genre of movie to get maximum impact out of movie. Theatre release somewhere are watchable for more than 3 hours given it is from a prominent director, star cast and with some Action/Drama genre.
- Now a days Profit depends on large scale release of the movie. Most of the Hollywood/Indian Cinema Movies goes with worldwide release in theatres, while OTT release already makes it an international release.

THANK YOU!!

(Looking Forward for Valuable Suggestions and Feedbacks)