

# IMDB Movie Analysis

## Project Description:

The dataset provided by the company contains various columns of different IMDB Movies. We are required to Frame the problem. For this task, we will need to define a problem we want to shed some light on.

We can do this by asking the following 'What?' :

- What do we see happening?
- What is our hypothesis for the cause of the problem? (this will be broadly based on intuition initially)
- What is the impact of the problem on stakeholders?
- What is the impact of the problem not being solved?

How to handle the things:

- Clean the Data
- Use the Data Analysis skills to explore the data set.
- Derive insights.

# IMDB Movie Analysis

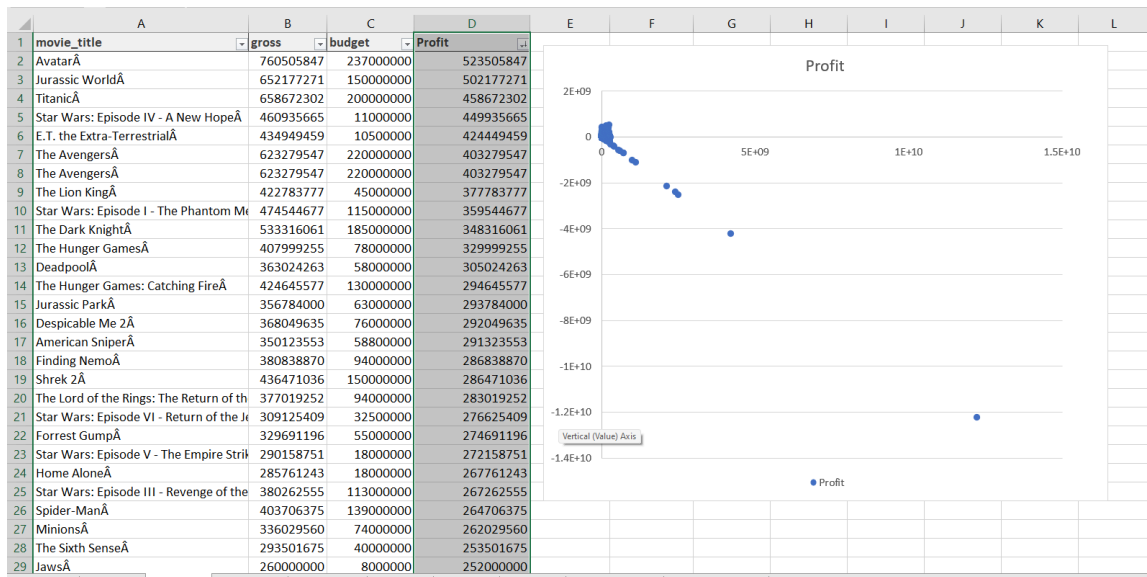
## Approach:

1. Task: Clean the data: This is one of the most important step to perform before moving forward with the analysis.

- First, we dropped the columns which have no use for the analysis.
- Second, we dropped the rows which are blank/null.
- Third, removed the duplicate row values.

2. Task: Find the movies with the highest profit?

- First, we created a new column 'Profit' by subtracting budget column from gross column.
- Second, we sorted the columns using the profit column as reference from the largest to the smallest.
- Third, we plotted the budget and profit in XY Scatter chart to find the outliers.
- There are as many as 5 outliers in the profit columns.
- The movie with the highest profit is 'Avatar' followed by 'Jurassic World' and 'Titanic' and so on



# IMDB Movie Analysis

### 3. Task: Find IMDB Top 250

- First, we filtered the 'num\_voted\_users' column greater than 25,000.
- Second, created a new column named 'IMDb\_Top\_250' and stored the top 250 movies with the highest IMDb Rating (sorted the 'imdb\_score' column from the largest to the smallest).
- Third, added a 'Rank' containing the values 1 to 250 using the RANK() function + COUNTIFS() function.
- Fourth, extracted all the movies in the IMDb\_Top\_250 column by filtering the 'language' column (unselecting English language) and stored them in a new column named 'Top\_Foreign\_Lang\_Film'.

[illegible]

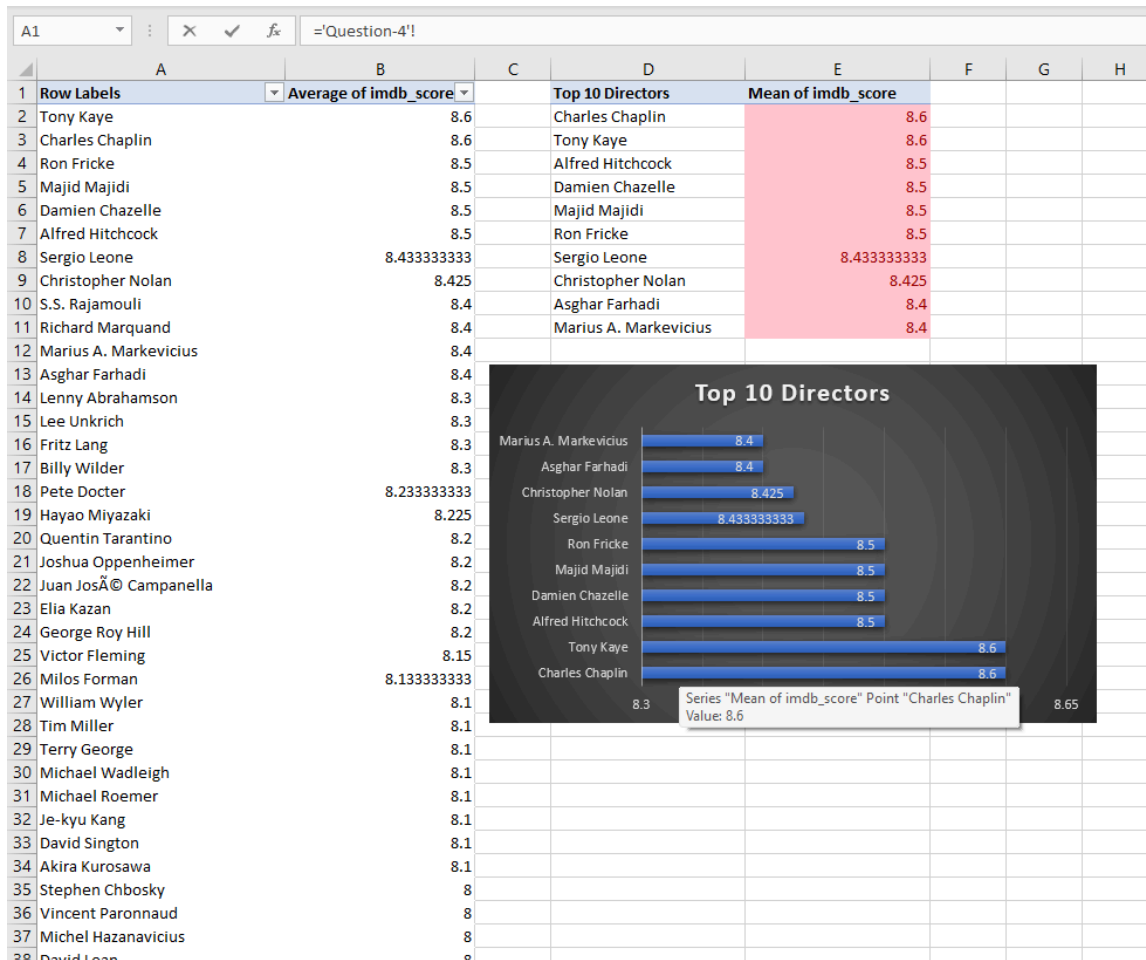
# IMDB Movie Analysis

Top_Foreign_Lang_Film					
	A	B	C	D	E
1	Top_Foreign_Lang_Film	num_voted_users	imdb_score	language	
2	The Good, the Bad and the Ugly	503509	8.9	Italian	
3	City of God	533200	8.7	Portuguese	
4	Seven Samurai	229012	8.7	Japanese	
5	Spirited Away	417971	8.6	Japanese	
6	Children of Heaven	27882	8.5	Persian	
7	The Lives of Others	259379	8.5	German	
8	A Separation	151812	8.4	Persian	
9	Amélie	534262	8.4	French	
10	Baahubali: The Beginning	62756	8.4	Telugu	
11	Das Boot	168203	8.4	German	
12	Oldboy	356181	8.4	Korean	
13	Princess Mononoke	221552	8.4	Japanese	
14	Metropolis	111841	8.3	German	
15	The Hunt	170155	8.3	Danish	
16	Unforgiven	248354	8.3	German	
17	Pan's Labyrinth	80429	8.2	French	
18	The Bridge on the River Kwai	131831	8.2	Spanish	
19	The Thing	467234	8.2	Spanish	
20	Warrior	214091	8.2	Japanese	
21	Annie Hall	81644	8.1	Portuguese	
22	In the Shadow of the Moon	65951	8.1	Danish	
23	Sling Blade	106160	8.1	Japanese	
24	Tae Guk Gi: The Brotherhood of War	64556	8.1	Spanish	
25	The Best Years of Our Lives	173551	8.1	Spanish	
26	The Imitation Game	31943	8.1	Korean	
27	Bowling for Columbine	28951	8	Portuguese	
28	Jaws	70194	8	French	
29	True Romance	69759	8	Hindi	
30	Waltz with Bashir	147566	8	Italian	
31	Young Frankenstein	46107	8	Hebrew	

## 4. Task: Find the best directors

- First, we selected the cleaned dataset done in Task 1 and created a pivot table.
- Second, we put the 'director\_name' into the Rows and took average of 'imdb\_score' in the Values section.
- Third, we sorted the 'director\_name' in ascending order and then sorted the 'average of imdb\_score' (largest to smallest).
- Then we selected the top 10 directors and their mean of imdb\_score in other columns.
- Next, we made a bar chart of the top 10 directors for the better insights.

# IMDB Movie Analysis

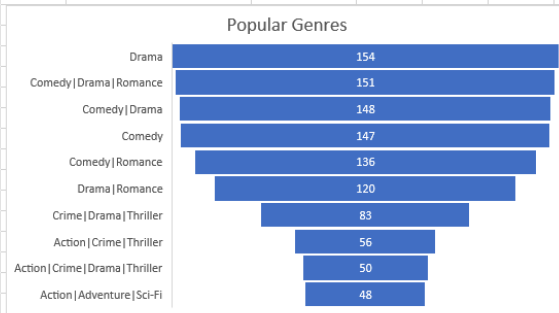


## 5. Task: Find popular genres

- First, we selected the 'genres' column from the cleaned dataset done in Task 1 and created a pivot table.
- Second, we put the 'genres' into the Rows and took count of 'genres' in the Values section.
- Third, we sorted the 'Count of genres' in descending order.
- Then we copied the top 10 genres and their count and pasted it in the other columns.
- Next, we made a funnel chart of the top 10 genres for the better insights.

# IMDB Movie Analysis

	A	B	C	D	E	F	G	H
1	Row Labels	Count of genres		Popular Genres	Count of Genres			
2	Drama	154		Drama	154			
3	Comedy Drama Romance	151		Comedy Drama Romance	151			
4	Comedy Drama	148		Comedy Drama	148			
5	Comedy	147		Comedy	147			
6	Comedy Romance	136		Comedy Romance	136			
7	Drama Romance	120		Drama Romance	120			
8	Crime Drama Thriller	83		Crime Drama Thriller	83			
9	Action Crime Thriller	56		Action Crime Thriller	56			
10	Action Crime Drama Thriller	50		Action Crime Drama Thriller	50			
11	Action Adventure Sci-Fi	48		Action Adventure Sci-Fi	48			
12	Comedy Crime							
13	Action Adventure Thriller							
14	Horror							
15	Drama Thriller							
16	Crime Drama							
17	Crime Drama Mystery Thriller							
18	Horror Thriller							
19	Action Adventure Sci-Fi Thriller							
20	Horror Mystery Thriller							
21	Drama Mystery Thriller							
22	Biography Drama							
23	Adventure Animation Comedy Family Fantasy							
24	Action Comedy Crime							
25	Horror Mystery							
26	Action Adventure Fantasy							
27	Documentary							
28	Biography Drama Sport							
29	Drama Sport							
30	Drama Music							
31	Action Thriller							
32	Adventure Animation Comedy Family							
33	Action Comedy Crime Thriller							
34	Comedy Family							
35	Biography Drama History							
36	Action Crime Drama Mystery Thriller							
37	Action Drama Thriller							
38	Adventure Family Fantasy							



6. Task: Find the critic-favorite and audience-favorite actors

- First, we created 3 new columns namely, Meryl\_Streep, Leo\_Caprio, and Brad\_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors using the 'actor\_1\_name' column.
- Second, appended the rows of all these columns and stored them in a new column named 'Combined'.
- We grouped the column by the actor's name: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt'.
- Then, we selected the cleaned dataset done in Task 1 and created a pivot table.
- Next, we put the 'actor\_1\_name' into the Rows and took mean/average of 'num\_users\_for\_review' in the Values section.
- Sorted the column from largest to smallest by mean of 'num\_users\_for\_review'.
- Then, we made a pivot chart (bar chart) of the mean of 'num\_users\_for\_review'.
- Again, we did the same above process for the mean of 'num\_critic\_for\_review'.

# IMDB Movie Analysis

#	B	C	D	E	F	G	H	I	J	K	L	M
1	num_critc_for_reviews	num_user_for_reviews	num_voted_users	movie_title	title_year	Meryl Streep	Leo Caprio	Brad Pitt	Combined	Group By		
2	813	2701	1144337	The Dark Knight Rises	2012	Lucy	Django Unchained	Divergent	Lucy	Meryl Streep		
3	775	2326	456260	Soul Food	2012	Maleficent	Inception	Shanghai Knights	Maleficent	Meryl Streep		
4	765	1193	955174	Django Unchained	2012	Transcendence	How to Train Your Dragon	The Circle	Transcendence	Meryl Streep		
5	750	1498	522048	This Is England	2012	Down and Out with the Dolls	Pirates of the Caribbean: The Curse of the Black Pearl	42nd Street	Down and Out with the Dolls	Meryl Streep		
6	750	1498	522030	Saw	2012	Machine Gun Preacher	The Sessions	Fight Club	Machine Gun Preacher	Meryl Streep		
7	739	1588	552303	Guardians of the Galaxy	2015	Like Crazy	Frozen River	Scott Pilgrim vs. the World	Like Crazy	Meryl Streep		
8	738	1885	582917	Weekend	2013	Brown Sugar	The Martian	Mulan	Brown Sugar	Meryl Streep		
9	733	2536	548573	Green Room	2013	Bridge of Spies	Lethal Weapon 4	Leap Year	Bridge of Spies	Meryl Streep		
10	723	3054	886204	Maurice	2009	True Lies	The Departed	The Kids Are All Right	True Lies	Meryl Streep		
11	712	2725	928227	Interstellar	2014	Hellboy II: The Golden Army	The Hunger Games	Connie and Carla	Hellboy II: The Golden Army	Meryl Streep		
12	703	1722	995415	A Fistful of Dollars	2012	Gerry	Animal House	Star Trek Into Darkness	Gerry	Meryl Streep		
13	703	1722	995415	Akira	2012		The Squid and the Whale	Sunshine State	Django Unchained	Leonardo DiCaprio		
14	682	678	245333	A Room with a View	2011		Dream with the Fishes	Zookeeper	Inception	Leonardo DiCaprio		
15	676	1264	431578	Black Book	2011		Transamerica	Black Swan	How to Train Your Dragon	Leonardo DiCaprio		
16	673	3018	371639	Sliding Doors	2016		The Iron Giant	Kung Fu Panda	Pirates of the Caribbean: The Curse of the Black Pearl	Leonardo DiCaprio		
17	673	1959	701607	Tarnation	2012		Casino Royale	The Puffy Chair	The Sessions	Leonardo DiCaprio		
18	669	1140	551363	In Bruges	2010		Hurricane Streets	It Follows	Frozen River	Leonardo DiCaprio		
19	663	269	70336	The Queen	2012		The Devil Wears Prada		The Martian	Leonardo DiCaprio		
20	656	695	452465	Cast Away	2012		Ocean's Twelve		Lethal Weapon 4	Leonardo DiCaprio		
21	654	985	465019	Sunshine Cleaning	2013		The Longest Yard		The Departed	Leonardo DiCaprio		
22	653	1097	682155	The Avengers	2014		The Family Man		The Hunger Games	Leonardo DiCaprio		
23	645	1367	637246	Gangster's Paradise: Jerusalem	2012				Animal House	Leonardo DiCaprio		
24	645	4667	1676169	The Dark Knight	2008				The Squid and the Whale	Leonardo DiCaprio		
25	644	1290	418214	Wild Target	2015				Dream with the Fishes	Leonardo DiCaprio		
26	642	2803	1468200	Inception	2010				Transamerica	Leonardo DiCaprio		
27	635	1117	462669	Laver Cake	2015				The Iron Giant	Leonardo DiCaprio		
28	634	986	277172	Rio	2012				Casino Royale	Leonardo DiCaprio		
29	608	1187	557489	Secretary	2013				Hurricane Streets	Leonardo DiCaprio		
30	606	1138	780588	How to Train Your Dragon	2013				The Devil Wears Prada	Leonardo DiCaprio		
31	602	994	275868	Machete	2015				Ocean's Twelve	Leonardo DiCaprio		
32	599	1225	451803	The Rose	2012				The Longest Yard	Leonardo DiCaprio		
33	597	695	439176	There Will Be Blood	2013				The Family Man	Leonardo DiCaprio		
34	596	1018	272839	Iron Man	2015				Divergent	Brad Pitt		
35	590	1171	395573	You Can Count on Me	2013				Shanghai Knights	Brad Pitt		
36	589	667	428916	I Served the King of England	2012				The Circle	Brad Pitt		
37	588	1101	395087	O Brother, Where Art Thou?	2014				42nd Street	Brad Pitt		
38	585	741	431620	Murderball	2014				Fight Club	Brad Pitt		
39	584	975	136367	Divergent	2011				Scott Pilgrim vs. the World	Brad Pitt		
40	579	1058	479047	The Truman Show	2016				Mulan	Brad Pitt		
41	576	742	496749	Dead Men's Shoes	2014				Leap Year	Brad Pitt		
42	576	583	190030	Slumdog Millionaire	2011				The Kids Are All Right	Brad Pitt		

#	A	B	C	D	E	F	G	H	I	J	K	L
2												
3												
4												
5												
6												
7												
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												
25												

Mean of num\_user\_for\_reviews

actor\_1\_name

- Leonardo DiCaprio
- Brad Pitt
- Meryl Streep

Mean of num\_critc\_for\_reviews

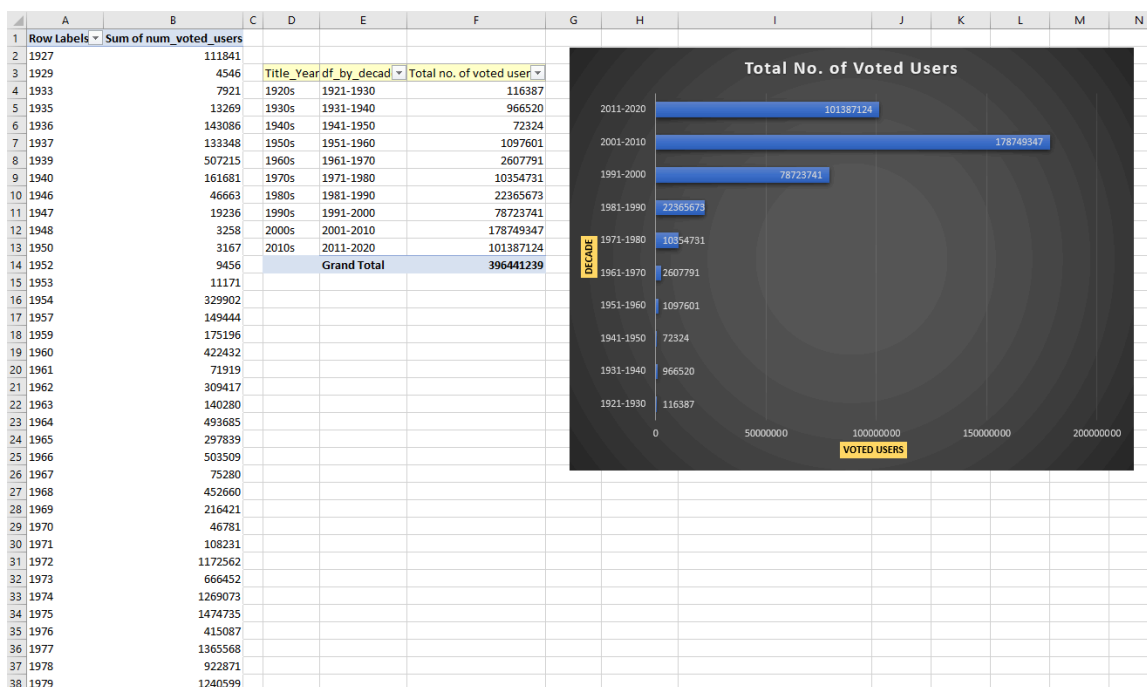
actor\_1\_name

- Leonardo DiCaprio
- Brad Pitt
- Meryl Streep

# IMDB Movie Analysis

Second part(change in number of voted users over decades):

- First, we selected the cleaned dataset done in Task 1 and created a pivot table.
- Second, we put the 'title\_year' into the Rows and took the sum of 'num\_voted\_users' in the Values section.
- Third, we grouped the title\_year by decade and stored in df\_by\_decade column.
- Lastly, we plotted the total no. of voted users against the decade in a bar chart.





# IMDB Movie Analysis

## Tech-Stack Used:

- **Microsoft Excel 365:** It enables users to format, organize and calculate data in a spreadsheet. It organize data in an easy-to-navigate way. We need not to perform any complex mathematical functions. And it turn piles of data into helpful graphics and charts.
- **Microsoft Word 2021:** It is used to make a report (PDF) to be presented to the leadership team.

## Insights:

- There are as many as 5 outliers in the profit columns.
- The movie with the highest profit is 'Avatar' followed by 'Jurassic World' and 'Titanic' and so on.
- The Shawshank Redemption is the top-most movie with the highest IMDB rating.
- The Good, the Bad and the Ugly (Italian) is the top-most foreign language movie.
- Charles Chaplin is the top-most director followed by Tony Kaye.
- The most popular genres is Drama followed by Comedy.
- 'Leonardo DiCaprio' is the critic-favorite as well as the audience-favorite actor.
- The most users voted in the decade 2000s and the least in the decade 1940s.

## Results:

- In this project, I applied the basic and advance Excel concepts. The concepts related to statistics and EDA have been implemented here by using MS Excel.
- In this task, the concepts regarding the sort, filter, pivot table, charts, different functions like rank, etc have been implemented.
- I learned to implement the learning of Excel in the real-time project.
- I learned how to frame the problem by asking 'what' looking at the dataset.
- It helped me in learning the '5 Why Analysis' to determine the root cause of the problem.

# IMDB Movie Analysis

- I learned how a data analyst think deeper and deeper to generate the valuable insights.
- It was a great learning experience while doing this project and it was challenging too while asking the different questions and finding their answers.

Excel Sheet Link:

[open in excel software for better visuals](#)

PPT Presentation Link

[PPT](#)

Video/Tutorial presentation link

[watch here](#)