

# IMDB Movie Analysis

## • DESCRIPTION:

For your Final Project, we are providing you with dataset having various columns of different IMDB Movies. You are required to Frame the problem. For this task, you will need to define a problem you want to shed some light on.

We can do this by asking 'What?' This is where you frame the problem i.e. What is the problem?

Use these questions to guide your thinking:

What do you see happening?

What is your 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?

Answering these questions will help you define a problem you are trying to solve and will allow you to find the right data to solve it.

Once you have defined a problem, clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.

Make sure to use 5 Whys Analysis in your analysis and use this to create a report which conveys a data story.

Once you have framed the problem and gathered initial insights from the data, you can ask the following questions as you dig deeper into your analysis.

What do you see happening?

What are the specific symptoms of the problem?

What is your hypothesis for the cause of the problem?

## • FIVE 'WHYS' APPROACH:

Once you have the problem better defined, you can use 5 Whys technique to determine its root cause by repeatedly asking the question "Why".

It's also called the Root Cause Analysis, developed by Sakichi Toyoda, founder of Toyota Industries. Here's an example of how this technique could be used to figure out the cause of the following problem: A business went over budget on a recent project.

Q: "Why did we go over budget on our project?" A: It took much longer than we expected to complete.

Q: "Why did it take longer than expected to complete?" A: We had to redesign several elements of the product.

Q: "Why did we have to redesign elements of the product?" A: Features of the product were confusing to use.

Q: "Why were the features of the product confusing to use?" A: We made incorrect assumptions about what users wanted.

Q: "Why did we make incorrect assumptions about what users wanted?"

A: Our user experience research team didn't ask effective questions.

As we see above, what looked like a budgeting problem turned out to be a problem with the user experience team not working effectively. While asking Why is easy, what we're interested in is the answer. Each time we answer why, the next time gets more difficult as we must think deeper behind the reasons for this. As we ask why, we may find that you have multiple answers for the same question.

- **APPROACH:**

- **Data Cleaning and Pre-processing:** This step involves removing any irrelevant or missing data from the dataset, and formatting the data so that it can be used for analysis.
- **Feature Engineering:** This step involves creating new features or variables from the existing data that can be used to better understand the movies. For example, creating a new variable that represents the budget of the movie, or creating a new variable that represents the genre of the movie.
- **Exploratory Data Analysis (EDA):** This step involves visualizing the data to understand the distribution of different variables and identify any patterns or trends in the data.
- **Data visualization:** This step involves creating visual representations of the data, such as bar charts, line plots, and scatter plots, to help understand the data and draw insights.

- **Conclusion and Recommendation:** This step will be the final step where the data is presented in an understandable way and conclusion is made based on the analysis.

- **TECH STACK USED:**

- MS-Excel
- MySQL
- Power BI

- **INSIGHTS AND SOLUTION:**

- 1) **Cleaning the data:** This is one of the most important steps to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping Columns, Removing Nulls etc.)

The screenshot shows a Microsoft Excel spreadsheet titled "IMDB Movies". The table contains the following columns and data:

director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_usd
James Cameron	723	760505847	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	886204	
Gore Verbinski	302	309404152	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	
Sam Mendes	602	200074150	Action Adventure Thriller	Christoph Waltz	Spectre	275868	
Christopher Nolan	813	448130642	Action Thriller	Tom Hardy	The Dark Knight Rises	1144337	
Andrew Stanton	461	73058679	Action Adventure Sci-Fi	Dany Sabara	John Carter	212204	
Sam Raimi	392	336530303	Action Adventure Romance	J.K. Simmons	Spider-Man 3	383056	
Nathan Greno	324	200807262	Adventure Animation Comedy Family Fantasy Musical Romance	Brad Garrett	Tangled	294810	
Joss Whedon	635	458991599	Action Comedy Mystery	Chris Hemsworth	Avengers: Age of Ultron	462669	
David Yates	375	301956980	Adventure Family Fantasy Mystery	Alan Rickman	Harry Potter and the Half-Blood Prince	321795	
Zack Snyder	673	330249062	Action Adventure Sci-Fi	Henry Cavill	Batman v Superman: Dawn of Justice	371639	
Bryan Singer	434	200969408	Action Adventure Sci-Fi	Kevin Spacey	Superman Returns	240396	
Marc Forster	403	168368427	Action Adventure	Giancarlo Giannini	Quantum of Solace	330784	
Gore Verbinski	313	423032628	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: Dead Man's Chest	522040	
Gore Verbinski	456	89289910	Action Adventure Western	Johnny Depp	The Lone Ranger	181792	
Zack Snyder	733	291021565	Action Adventure Fantasy Sci-Fi	Henry Cavill	Man of Steel	548573	
Andrew Adamson	258	141614023	Action Adventure Family Fantasy	Peter Dinklage	The Chronicles of Narnia: Prince Caspian	149922	
Joss Whedon	703	623279547	Action Sci-Fi	Chris Hemsworth	The Avengers	995415	
Rob Marshall	448	241063875	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: On Stranger Tides	370704	
Barry Sonnenfeld	451	179020854	Action Adventure Comedy Family Fantasy Sci-Fi	Will Smith	Men in Black 3	268154	
Peter Jackson	422	255108370	Adventure Fantasy	Aidan Turner	The Hobbit: The Desolation of Smaug	354228	
Marc Webb	599	262030663	Action Adventure Fantasy	Emma Stone	The Amazing Spider-Man	451803	
Ridley Scott	343	105219735	Action Adventure Drama History	Mark Addy	Robin Hood	211765	
Peter Jackson	509	258355554	Adventure Fantasy	Aidan Turner	The Hobbit: The Desolation of Smaug	483540	
Chris Weitz	251	70083519	Adventure Family Fantasy	Christopher Lee	The Golden Compass	149019	
Peter Jackson	446	218051260	Action Adventure Drama Romance	Naomi Watts	King Kong	316018	
James Cameron	315	658672302	Drama Romance	Leonardo DiCaprio	Titanic	793059	
Anthony Russo	515	407197282	Action Adventure Sci-Fi	Robert Downey Jr.	Captain America: Civil War	272670	
Peter Berg	377	65173160	Action Adventure Sci-Fi Thriller	Liam Neeson	Battleship	202382	
Colin Trevorrow	644	652177271	Action Sci-Fi Thriller	Bryce Dallas Howard	Jurassic World	418214	
Sam Mendes	750	304360277	Action Adventure Thriller	Albert Finney	Skyfall	522030	
Sam Raimi	300	373377693	Action Fantasy Romance	J.K. Simmons	Spider-Man 2	411164	
Shane Black	608	408992270	Action Adventure Sci-Fi	Robert Downey Jr.	Iron Man 3	557489	
Tim Burton	451	334185206	Action Family Fantasy	Johnny Depp	Alice in Wonderland	306320	
Brett Ratner	334	234360014	Action Adventure Fantasy Sci-Fi Thriller	Hugh Jackman	X-Men: The Last Stand	383427	
Dan Scanlon	376	268488329	Adventure Animation Comedy Family Fantasy	Steve Buscemi	Monsters University	235025	
Michael Bay	366	402076689	Action Adventure Sci-Fi	Glen Morshower	Transformers: Revenge of the Fallen	323207	
Michael Bay	378	245428137	Action Adventure Sci-Fi	Bingbing Li	Transformers: Age of Extinction	242420	

To clean the data, I arranged the columns in the correct format and increased the column width to improve readability. I also removed null values and duplicates by selecting them using the "Go to Special" feature option and deleting the entire row. This helped to ensure that the data was accurate and ready for analysis.

For deleting the Null values used the following steps:

- The "Go To Special" feature:
- Select the range of cells that you want to remove blank cells from.

- Press "Ctrl + G" to open the "Go To" dialog box.
- Select "Special" from the options.
- Select "Blanks" and click "OK".
- Press the "Delete" key to remove the blank cells.

For deleting the Duplicate values used the following steps:

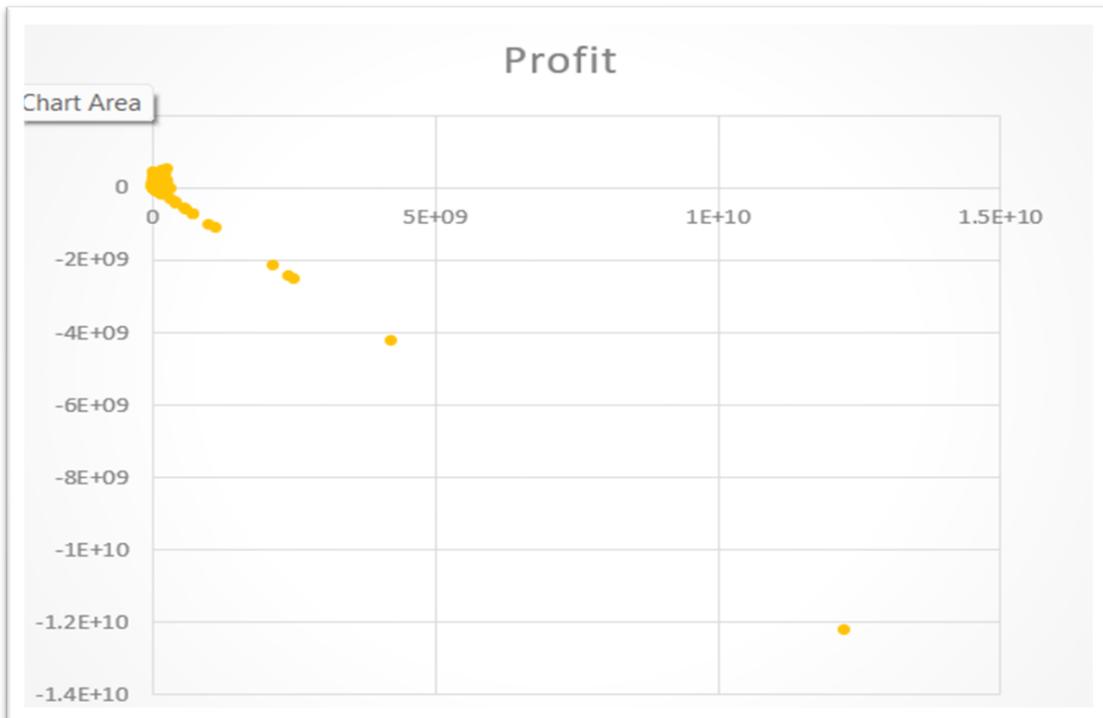
- Select the range of cells that you want to remove blank cells from.
- Go to the "Data" tab in the ribbon.
- Click on "Remove Duplicates" in the "Data Tools" group.
- Make sure that "All" columns are selected, and click "OK".

Before Cleaning: 5044 rows and 28 columns

After Cleaning: 3850 rows and 13 columns

**2) Movies with highest profit:** Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

**Your task:** Find the movies with the highest profit?



movie_title	Sum of Profit
Avatar	523505847
Jurassic World	502177271
Titanic	458672302
Star Wars: Episode IV - A New Hope	449935665
E.T. the Extra-Terrestrial	424449459
The Avengers	403279547
The Lion King	377783777
The Jungle Book	375290282
Star Wars: Episode I - The Phantom Menace	359544677
The Dark Knight	348316061
The Hunger Games	329999255
Twilight	308898950
Deadpool	305024263
The Hunger Games: Catching Fire	294645577
Jurassic Park	293784000
Despicable Me 2	292049635
American Sniper	291323553
Finding Nemo	286838870
Shrek 2	286471036
The Lord of the Rings: The Return of the King	283019252
Star Wars: Episode VI - Return of the Jedi	276625409
Forrest Gump	274691196
Star Wars: Episode V - The Empire Strikes Back	272158751
Juno	271985680
Alice in Wonderland	268370412
Home Alone	267761243
Star Wars: Episode III - Revenge of the Sith	267262555
Spider-Man	264706375
Minions	262029560
The Sixth Sense	253501675
Jaws	252000000
Frozen	250736600
The Secret Life of Pets	248505540
<b>Total</b>	<b>22049537152</b>

In this task we have to first create a new column to store the profit of the movies by taking the difference of the gross and budget

To identify outliers in the data, I plotted a chart and looked for any unusually high or low values. One example of an outlier that I observed was a value of -1.2E+10

I used a tool called Power BI to create this visualization showing that the movie with the highest profit was “Avatar”

**3) Top 250:** Create a new column IMDb\_Top\_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb\_score). Also make sure that for all of these movies, the num\_voted\_users are greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Extract all the movies in the IMDb\_Top\_250 column which are not in the English language and store them in a new column named Top\_Foreign\_Lang\_Film. You can use your own imagination also!

**Your task:** Find IMDB Top 250.

```
SELECT
    row_number() over(order by imdb_score DESC,
    num_voted_users DESC) as ranking,
    imdb_score, num_voted_users, movie_title AS
IMDb_Top_250, language
FROM
    imdb.imdb_movies
WHERE
    num_voted_users > 25000
LIMIT 250;
```

I used an SQL query to identify the top 250 movies with the highest IMDB scores and a minimum of 25,000 voted users. Here is the list:

Ranking	imdb_score	num_voted_user	IMDb_Top_250	language
1	9	1676169	The Dark Knight	English
2	8.9	1215718	The Lord of the Rings: The Return of the King	English
3	8.8	1462004	Inception	English
4	8.8	1347461	Fight Club	English
5	8.8	1251222	Forrest Gump	English
6	8.8	1238746	The Lord of the Rings: The Fellowship of the Ring	English
7	8.8	1214843	Daredevil	English
8	8.7	1217752	The Matrix	English
9	8.7	1100446	The Lord of the Rings: The Two Towers	English
10	8.6	928227	Interstellar	English
11	8.6	881236	Saving Private Ryan	English
12	8.6	159910	Hannibal	English
13	8.5	1144337	The Dark Knight Rises	English
14	8.5	982637	Gladiator	English
15	8.5	955174	Django Unchained	English
16	8.5	873649	The Departed	English
17	8.5	844052	The Prestige	English
18	8.5	782610	The Green Mile	English
19	8.5	744891	Terminator 2: Judgment Day	English
20	8.5	644348	The Lion King	English
21	8.5	497946	The Pianist	English
22	8.5	50391	Outlander	English
23	8.4	736638	Braveheart	English
24	8.4	718837	WALL-E	English
25	8.4	534262	Amélie	French
26	8.4	63982	Stargate SG-1	English
27	8.4	62754	Baahubali: The Beginning	Telugu
28	8.3	980944	Batman Begins	English
29	8.3	885175	Inglourious Basterds	English
30	8.3	665575	Up	English
31	8.3	544884	Toy Story 3	English
32	8.3	515306	Indiana Jones and the Last Crusade	English
33	8.3	414219	L.A. Confidential	English
34	8.3	345198	Inside Out	English
35	8.3	29450	Life	English
36	8.2	791783	V for Vendetta	English
37	8.2	780588	The Wolf of Wall Street	English
38	8.2	692482	Finding Nemo	English
39	8.2	610568	A Beautiful Mind	English
40	8.2	485430	How to Train Your Dragon	English
41	8.2	333542	Casino	English
42	8.2	272670	Captain America: Civil War	English
43	8.2	258078	The Thing	English
44	8.1	995415	The Avengers	English
45	8.1	995415	The Avengers	English
46	8.1	809474	Pirates of the Caribbean: The Curse of the Black Pearl	English
47	8.1	786092	Shutter Island	English
48	8.1	735784	Kill Bill: Vol. 1	English
49	8.1	704766	The Sixth Sense	English
50	8.1	682155	Guardians of the Galaxy	English
51	8.1	667983	The Truman Show	English
52	8.1	588400	Ein Kind ist	German

Ranking	imdb_score	num_voted_user	IMDb_Top_250	language
53	8.1	613473	Jurassic Park	English
54	8.1	585659	Monsters, Inc.	English
55	8.1	569841	Gone Girl	English
56	8.1	552503	Mad Max: Fury Road	English
57	8.1	491077	The Bourne Ultimatum	English
58	8.1	479047	Deadpool	English
59	8.1	472488	The Martian	English
60	8.1	40620	The Revenant	English
61	8.1	383591	Prisoners	English
62	8.1	312629	Rush	English
63	8.1	54057	Solaris	Russian
64	8	525801	Catch Me If You Can	English
65	8	514125	X-Men: Days of Future Past	English
66	8	512749	Kill Bill: Vol. 2	English
67	8	504419	Star Trek	English
68	8	479166	The Incredibles	English
69	8	473887	Ratatouille	English
70	8	470483	Casino Royale	English
71	8	440084	Life of Pi	English
72	8	400292	Blood Diamond	English
73	8	350698	Big Fish	English
74	8	338383	The Pursuit of Happyness	English
75	8	242599	Serenity	English
76	8	241030	Magnolia	English
77	8	148238	Cinderella Man	English
78	8	128455	The Iron Giant	English
79	8	113472	JFK	English
80	7.9	886204	Avatar	English
81	7.9	696338	Iron Man	English
82	7.9	637246	The Hobbit: An Unexpected Journey	English
83	7.9	483541	The Hobbit: The Desolation of Smaug	English
84	7.9	467113	Shrek	English
85	7.9	431620	Edge of Tomorrow	English
86	7.9	407601	The Bourne Identity	English
87	7.9	385871	Toy Story 2	English
88	7.9	361767	Children of Men	English
89	7.9	323353	Captain Phillips	English
90	7.9	279093	Big Hero 6	English
91	7.9	272839	The Hateful Eight	English
92	7.9	221128	How to Train Your Dragon 2	English
93	7.9	207287	Almost Famous	English
94	7.9	149414	Hero	Mandarin
95	7.9	133526	The Insider	English
96	7.8	583341	The Hangover	English
97	7.8	582917	Gravity	English
98	7.8	522030	Skyfall	English
99	7.8	518537	X-Men: First Class	English
100	7.8	496749	Captain America: The Winter Soldier	English
101	7.8	459346	The Curious Case of Benjamin Button	English
102	7.8	402645	Ocean's Eleven	English
103	7.8	395573	Star Trek Into Darkness	English
104	7.8	382255	Harry Potter and the Prisoner of Azkaban	English

Ranking	imdb_score	num_voted_user	IMDb_Top_250	language
105	7.8	348232	The Bourne Supremacy	English
106	7.8	340085	Back to the Future Part II	English
107	7.8	330152	The Girl with the Dragon Tattoo	English
108	7.8	324671	American Gangster	English
109	7.8	294810	Tangled	English
110	7.8	272534	Wreck-It Ralph	English
111	7.8	261069	The Game	English
112	7.8	246698	The Lego Movie	English
113	7.8	237872	3:10 to Yuma	English
114	7.8	236000	Apocalypto	Maya
115	7.8	225122	Donnie Brasco	English
116	7.8	220591	Gattaca	English
117	7.8	213668	The Fugitive	English
118	7.8	199056	Changeling	English
119	7.8	139114	Fantastic Mr. Fox	English
120	7.8	113068	The Last of the Mohicans	English
121	7.8	106072	The Jungle Book	English
122	7.8	64989	The Conjuring	English
123	7.8	38383	3rd Rock from the Sun	English
124	7.8	28276	The Little Prince	English
125	7.7	793059	Titanic	English
126	7.7	607235	300	English
127	7.7	479453	The Social Network	English
128	7.7	452465	Argo	English
129	7.7	399651	Minority Report	English
130	7.7	394317	Cast Away	English
131	7.7	392474	Watchmen	English
132	7.7	385943	Despicable Me	English
133	7.7	343274	The Fifth Element	English
134	7.7	318634	Love Actually	English
135	7.7	317166	The Last Samurai	English
136	7.7	305929	Training Day	English
137	7.7	301279	Zodiac	English
138	7.7	292022	Black Hawk Down	English
139	7.7	266310	Man on Fire	English
140	7.7	240962	True Grit	English
141	7.7	232710	Seven Pounds	English
142	7.7	224671	As Good as It Gets	English
143	7.7	223127	The Blind Side	English
144	7.7	212085	Stardust	English
145	7.7	200359	Road to Perdition	English
146	7.7	189249	Eastern Promises	English

Ranking	imdb_score	num_voted_user	IMDb_Top_250	language
157	7.6	477300	Sherlock Holmes	English
158	7.6	421658	Frozen	English
159	7.6	403836	Rise of the Planet of the Apes	English
160	7.6	385670	Harry Potter and the Goblet of Fire	English
161	7.6	328159	Ice Age	English
162	7.6	317542	Dawn of the Planet of the Apes	English
163	7.6	307029	Kung Fu Panda	English
164	7.6	303185	Fun	English
165	7.6	299258	Die Hard with a Vengeance	English
166	7.6	293662	Collateral	English
167	7.6	283563	Moneyball	English
168	7.6	280228	The Town	English
169	7.6	273108	Inside Man	English
170	7.6	269033	Batman	English
171	7.6	267980	The Godfather: Part III	English
172	7.6	248123	Lord of War	English
173	7.6	243834	Les Misérables	English
174	7.6	239752	Interview with the Vampire: The Vampire Chronicles	English
175	7.6	224013	Moulin Rouge	English
176	7.6	208817	Apollo 13	English
177	7.6	203963	Lone Survivor	English
178	7.6	188887	Enemy at the Gates	English
179	7.6	178118	Bridge of Spies	English
180	7.6	176936	Munich	English
181	7.6	170684	Traffic	English
182	7.6	155496	Grindhouse	English
183	7.6	145270	The Impossible	English
184	7.6	138941	The Thin Red Line	English
185	7.6	136580	The Crow	English
186	7.6	131217	The Abyss	English
187	7.6	97838	Star Trek: First Contact	English
188	7.6	76016	The Hurricane	English
189	7.6	38690	The Flowers of War	Mandarin
190	7.6	25402	The A-Team	English
191	7.5	462669	Avengers: Age of Ultron	English
192	7.5	444683	Harry Potter and the Sorcerer's Stone	English
193	7.5	405973	X-Men 2	English
194	7.5	355137	Harry Potter and the Order of the Phoenix	English
195	7.5	354228	The Hobbit: The Desolation of Smaug	English
196	7.5	338635	Sherlock Holmes: A Game of Shadows	English
197	7.5	321793	Harry Potter and the Half-Blood Prince	English
198	7.5	310383	Game of Thrones	English

Ranking	IMDB_Score	num_voted_user	IMDB_top_250	language
208	7.5	171792	Mulan	English
209	7.5	154487	Sleepers	English
210	7.5	149285	A Nightmare on Elm Street	English
211	7.5	143835	Bram Stoker's Dracula	English
212	7.5	117719	Saving Mr. Banks	English
213	7.5	90360	Pinocchio	English
214	7.5	88270	The Life of David Gale	English
215	7.5	66511	42	English
216	7.5	53607	Star Trek Beyond	English
217	7.5	49049	Sleepy Hollow	English
218	7.5	36919	Constantine	English
219	7.4	452928	X-Men	English
220	7.4	387616	Harry Potter and the Chamber of Secrets	English
221	7.4	375456	Crazy, Stupid, Love.	English
222	7.4	365104	Mission: Impossible - Ghost Protocol	English
223	7.4	313866	Ant-Man	English
224	7.4	283480	Back to the Future Part III	English
225	7.4	259492	The Rock	English
226	7.4	259083	The Simpsons Movie	English
227	7.4	232187	Mission: Impossible - Rogue Nation	English
228	7.4	217480	Law Abiding Citizen	English
229	7.4	216032	Zero Dark Thirty	English
230	7.4	200556	Contact	English
231	7.4	197412	Lincoln	English
232	7.4	184795	The Bucket List	English
233	7.4	177383	The Adventures of Tintin	English
234	7.4	168207	Master and Commander: The Far Side of the World	English
235	7.4	148490	K-PAX	English
236	7.4	142067	The English Patient	English
237	7.4	136973	The Judge	English
238	7.4	124222	Invictus	English
239	7.4	105446	Poltergeist	English
240	7.4	100743	Man on the Moon	English
241	7.4	99558	A Time to Kill	English
242	7.4	96654	The Phantom of the Opera	English
243	7.4	91860	Across the Universe	English
244	7.4	90932	Pirate Radio	English
245	7.4	77394	The Walk	English
246	7.4	67797	Mean Streets	English
247	7.4	54643	Wonder Boys	English
248	7.4	47764	13 Hours	English
249	7.4	36894	Red Cliff	Mandarin
250	7.2	701607	The Hunger Games	English

From this I can also extract the films that are not in English language (Because majority of the movies are in English). And they are stored in a new column named `Top_foreign_lang_film`. Following SQL query used for finding non-English films.

That is, out of the top rated 250 IMDB movies only 31 films are of other languages and they are Italian, German, Hindi, Telugu, Persian, Spanish ...etc and the rest of the movies are in English.

```

1 •  SELECT
2   ↪  row_number() over(order by imdb_score DESC,
3   num_voted_users DESC) as ranking,
4   imdb_score, num_voted_users, movie_title AS
5   Top_foreign_lang_film, language
6   FROM
7   imdb.imdb_movies
8   WHERE
9   language != 'English'
10  LIMIT 250;
11
12

```

ranking	imdb_score	num_voted_users	Top_foreign_lang_film	language
1	8.4	534262	Amélie	French
2	8.4	62756	Baahubali: The Beginning	Telugu
3	8.3	15762	The Returned	French
4	8.2	374	Godzilla Resurgence	Japanese
5	8.2	374	Godzilla Resurgence	Japanese
6	8.1	54057	Solaris	Russian
7	7.9	149414	Hero	Mandarin
8	7.8	236000	Apocalypto	Maya
9	7.8	7630	Oceans	French
10	7.8	5639	Earth	Hindi
11	7.7	62607	A Very Long Engagement	French
12	7.6	38690	The Flowers of War	Mandarin
13	7.4	36894	Red Cliff	Mandarin
14	7.2	24657	Micmacs	French
15	7.2	21912	Ip Man 3	Cantonese
16	7.2	6	10,000 B.C.	
17	7.1	22897	The Warlords	Mandarin
18	7.1	8	Star Wars: Episode VII - The Force Awakens	
19	7	36455	Curse of the Golden Flower	Mandarin
20	6.8	15790	A Monster in Paris	French
21	6.7	18209	The Great Raid	Filipino
22	6.5	24557	The Grandmaster	Mandarin
23	6.4	86152	The Interpreter	Aboriginal
24	6.4	979	Evolution	French
25	6.1	11584	Dragon Blade	Mandarin
26	6	3322	Nomad: The Warrior	Kazakh
27	5.9	71574	The Legend of Zorro	Spanish
28	5.3	4387	Obitae myy ostrov	Russian
29	5.1	20567	Asterix at the Olympic Games	French
30	4.9	590	Animal Kingdom: Let's go Ape	French
31	4.4	230	Top Cat Begins	Spanish

**4) Best Directors:** T-group the column using the director\_name column. Find out the top 10 directors for whom the mean of imdb\_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.

**Your task:** Find the best directors

```
SELECT
    director_name AS top_10_director,
    round(AVG(imdb_score), 2) AS avg_score
FROM
    imdb.imdb_movies
GROUP BY director_name
ORDER BY avg_score DESC , director_name
LIMIT 10
```

	top_10_director	avg_score
▶	Christopher Nolan	8.41
	S.S. Rajamouli	8.4
	Lee Unkrich	8.3
	Pete Docter	8.23
	Hideaki Anno	8.2
	Quentin Tarantino	8.16
	Alejandro G. Iñárritu	8.1
	Andrei Tarkovsky	8.1
	Denis Villeneuve	8.1
	James Gunn	8.1

Based on the data provided, it appears that Christopher Nolan and S.S. Rajamouli are the best director, with an average IMDB score of 8.4 for his movies.

**5) Popular Genres:** Perform this step using the knowledge gained while performing previous steps.

**Your task:** Find popular genres.

```
SELECT genres AS popular_genres, AVG(imdb_score) AS highest_imdb_score
FROM imdb.imdb_movies first
GROUP BY genres
ORDER BY AVG(imdb_score) DESC
LIMIT 10
```

	popular_genres	highest_imdb_score
1	Action Adventure Crime Drama Sci-Fi Thriller	8.8
2	Crime Drama Horror Mystery Thriller	8.6
3	Crime Drama Fantasy Mystery	8.5
4	Action Drama Romance	8.5
5	Drama Western	8.5
6	Adventure Animation Drama Family Musical	8.5
7	Drama Romance Sci-Fi	8.5
8	Action Adventure Drama Fantasy War	8.4
9	Adventure Animation Comedy Drama Family Fantasy	8.3
10	Drama Fantasy Horror Mystery	8.3
11		

Based on the data provided, it appears that the Action|Adventure|Crime|Drama|SciFi|Thriller genre has the highest average IMDB score, indicating that it is a more preferable genre.

**6) Charts:** Create three 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. Use only the actor\_1\_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

Append the rows of all these columns and store them in a new column named Combined.

Group the combined column using the actor\_1\_name column.

Find the mean of the num\_critic\_for\_reviews and num\_users\_for\_review and identify the actors which have the highest mean.

Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title\_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df\_by\_decade.

**Your task:** Find the critic-favourite and audience-favourite actor

```

SELECT actor_1_name,
COUNT(movie_title) AS no_of_movies,
ROUND(AVG(num_user_for_reviews), 2) AS user_reviews,
ROUND(AVG(num_critic_for_reviews), 2) AS critic_reviews
FROM ((SELECT actor_1_name, movie_title,
num_user_for_reviews, num_critic_for_reviews
FROM imdb.imdb_movies
WHERE actor_1_name = 'Meryl Streep')
UNION ALL (SELECT actor_1_name, movie_title,
num_user_for_reviews, num_critic_for_reviews
FROM imdb.imdb_movies
WHERE actor_1_name = 'Leonardo DiCaprio')
UNION ALL (SELECT actor_1_name, movie_title,
num_user_for_reviews, num_critic_for_reviews
FROM imdb.imdb_movies
WHERE actor_1_name = 'Brad Pitt')) c
GROUP BY actor_1_name
ORDER BY user_reviews DESC , critic_reviews DESC

```

	actor_1_name	no_of_movies	user_reviews	critic_reviews
2	Leonardo DiCaprio	20	922.55	322.2
3	Brad Pitt	17	742.35	245
4	Meryl Streep	11	297.18	181.45

Based on the data provided, it appears that Leonardo DiCaprio is the audience favourite and critic favourite actor.

**7) Bar Chart:** Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title\_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade.

**Your task:** Find the number of user votes per decade.

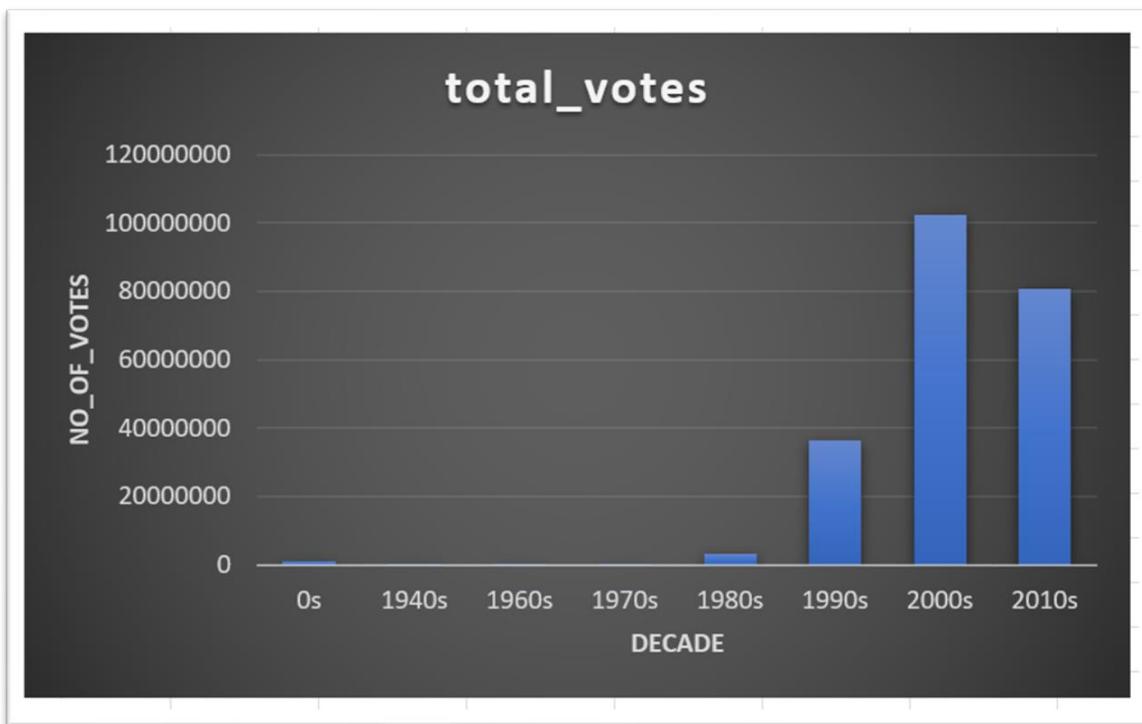
```

SELECT
CONCAT(CONVERT( FLOOR(title_year / 10) * 10 , CHAR),
's') AS decade,
SUM(num_voted_users) AS total_votes
FROM
imdb.imdb_movies
GROUP BY decade
ORDER BY decade

```

During the 2000s, there was a high number of users who voted for movies.

1	decade	total_votes
2	0s	954256
3	1940s	90360
4	1960s	95264
5	1970s	347092
6	1980s	3057135
7	1990s	36486955
8	2000s	102375694
9	2010s	80568844



- **ANALYSIS:**

the dataset provides insights into the popularity of Hollywood movies, the distinction between English and foreign language films in the top 250, the preferences of both audiences and critics towards certain actors, and the increase in the number of voters over time. The process of recording and analyzing the findings from the dataset can be complex, but it can lead to valuable insights and trends that can help understand the film industry and its audience.

- **CONCLUSION:**

this project provides valuable insights into the film industry by analyzing the IMDB ratings of movies, directors, and actors. The results can help filmmakers make informed decisions about future projects, such as which actors and directors to hire or which genres are currently popular. The project also provides practical experience in using Microsoft Excel, including the use of pivot tables and advanced functions. Overall, this project is a useful tool for anyone interested in understanding the trends and patterns in the film industry.