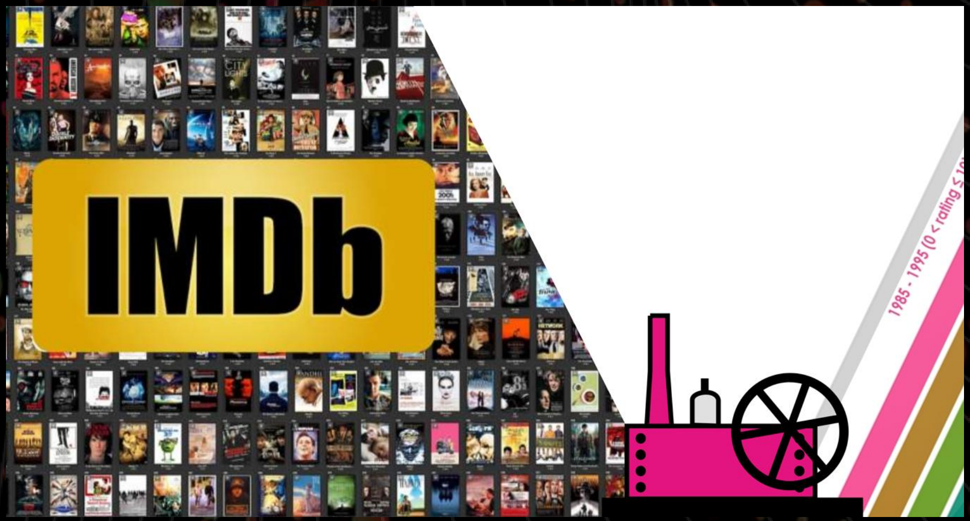


IMDB Movie Analysis

SOURAV PATTANAYAK

Table of contents

Project description

A designated data analyst has an important role in evaluating the processes and progresses of organizations, with the help of given datasets a data analyst performs few steps to reach the insight.

In this project I was provided with IMDb movie dataset containing various columns for my final project and by applying the 'Five whys' approach and data analysis skills the answers of asked queries were explained.

The entire dataset was cleaned before taking it any further to reach the required insights.

Project objectives

- To inspect the **Root Cause Analysis**.
- To clean the data.
- To find the movies with highest profit.
- To find IMDb top 250 movies based on the number of voted users and extracting the non-english movies from the list.
- To find the best directors.
- To find popular movie genres.
- To find the critic favourite and audience favourite actors.
- To find the decade with highest votes.

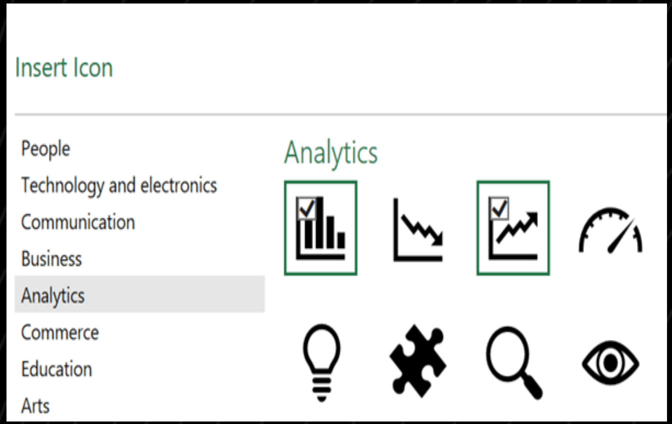
Approach

Basic and advanced data analysis methods were used to calculate the queries, inspect the queries and reach the desired insights, data was cleaned before the ‘Five whys’ approach was used extensively to understand the Root cause analysis.

Different tools were used to visualize and represent the results and later the information were gathered and loopholes were found to jot them down.

Methodology and Tech-Stack used

- ✓ Microsoft excel tool was used to complete the entire project.
- ✓ To evaluate the queries Microsoft advanced excel formulas were used.
- ✓ Some of the queries were visualized using Microsoft excel charts for better explanation.
- ✓ Microsoft Powerpoint presentation was used to prepare the presentation.



Insights

1.Data cleaning :

- Data cleaning is one of the most important step that needs to be taken before analyzing the dataset.
- In the given dataset the data was cleaned after revising the queries by deleting unnecessary columns, null values etc.
- Initially I had 5044 row and 28 columns and after cleaning I have 3785 rows with 14 columns.
- Basic excel methods(e.g:filter columns, sorting) were used to exclude null values and sort the data based on requirements.
- The duplicate values were removed using 'Data' tab → Remove duplicate values/cells.
- *To go to the cleaned dataset after the operations were done, click here.*

Insights

Preview before cleaning:

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	director_r	num_criti	duration	director_f	actor_3_fa	actor_2_n	actor_1_fa	gross	genres	actor_1_n	movie_tit	num_vote	cast_tota	actor_3_n	facenumb	plot_keyw	movie_im	num_user	language	country	content_r	budget	title_year	actor_2_fa	imdb_sco	aspect_ra	movie_fa
2	James Cai	723	178	0	855	Joel Davic	1000	7.6E+08	Action|Ad	CCH Poun	AvatarÂ	886204	4834	Wes Studi	0	avatar|fu	http://ww	3054	English	USA	PG-13	2.4E+08	2009	936	7.9	1.78	33000
3	Gore Verb	302	169	563	1000	Orlando B	40000	3.1E+08	Action|Ad	Johnny De	Pirates of	471220	48350	Jack Dave	0	goddess|	http://ww	1238	English	USA	PG-13	3E+08	2007	5000	7.1	2.35	0
4	Sam Menc	602	148	0	161	Rory Kinn	11000	2E+08	Action|Ad	Christoph	SpectreÂ	275868	11700	Stephanie	1	bomb|esp	http://ww	994	English	UK	PG-13	2.5E+08	2015	393	6.8	2.35	85000
5	Christoph	813	164	22000	23000	Christian	27000	4.5E+08	Action|Th	Tom Hard	The Dark K	1144337	106759	Joseph Gc	0	deception	http://ww	2701	English	USA	PG-13	2.5E+08	2012	23000	8.5	2.35	164000
6	Doug Walker			131		Rob Walke	131		Documen	Doug Wal	Star Wars	8	143		0		http://ww							12	7.1		0
7	Andrew St	462	132	475	530	Samantha	640	7.3E+07	Action|Ad	Daryl Sabi	John Carte	212204	1873	Polly Wall	1	alien|ame	http://ww	738	English	USA	PG-13	2.6E+08	2012	632	6.6	2.35	24000
8	Sam Raim	392	156	0	4000	James Fra	24000	3.4E+08	Action|Ad	J.K. Simm	Spider-Ma	383056	46055	Kirsten Du	0	sandman	http://ww	1902	English	USA	PG-13	2.6E+08	2007	11000	6.2	2.35	0
9	Nathan Gr	324	100	15	284	Donna Mu	799	2E+08	Adventure	Brad Garr	TangledÂ	294810	2036	M.C. Gain	1	17th cent	http://ww	387	English	USA	PG	2.6E+08	2010	553	7.8	1.85	29000
10	Joss Whe	635	141	0	19000	Robert Do	26000	4.6E+08	Action|Ad	Chris Hem	Avengers:	462669	92000	Scarlett Jo	4	artificial i	http://ww	1117	English	USA	PG-13	2.5E+08	2015	21000	7.5	2.35	118000
11	David Yat	375	153	282	10000	Daniel Ra	25000	3E+08	Adventure	Alan Rickr	Harry Pott	321795	58753	Rupert Gr	3	blood|bo	http://ww	973	English	UK	PG	2.5E+08	2009	11000	7.5	2.35	10000
12	Zack Snyd	673	183	0	2000	Lauren Co	15000	3.3E+08	Action|Ad	Henry Cav	Batman v	371639	24450	Alan D. Pu	0	based on	http://ww	3018	English	USA	PG-13	2.5E+08	2016	4000	6.9	2.35	197000
13	Bryan Sing	434	169	0	903	Marlon Br	18000	2E+08	Action|Ad	Kevin Spa	Supermar	240396	29991	Frank Lan	0	crystal|ep	http://ww	2367	English	USA	PG-13	2.1E+08	2006	10000	6.1	2.35	0
14	Marc Fors	403	106	395	393	Mathieu A	451	1.7E+08	Action|Ad	Giancarlo	Quantum	330784	2023	Rory Kinn	1	action her	http://ww	1243	English	UK	PG-13	2E+08	2008	412	6.7	2.35	0
15	Gore Verb	313	151	563	1000	Orlando B	40000	4.2E+08	Action|Ad	Johnny De	Pirates of	522040	48486	Jack Dave	2	box office	http://ww	1832	English	USA	PG-13	2.3E+08	2006	5000	7.3	2.35	5000
16	Gore Verb	450	150	563	1000	Ruth Wils	40000	8.9E+07	Action|Ad	Johnny De	The Lone I	181792	45757	Tom Wilki	1	horse|out	http://ww	711	English	USA	PG-13	2.2E+08	2013	2000	6.5	2.35	48000
17	Zack Snyd	733	143	0	748	Christoph	15000	2.9E+08	Action|Ad	Henry Cav	Man of Ste	548573	20495	Harry Len	0	based on	http://ww	2536	English	USA	PG-13	2.3E+08	2013	3000	7.2	2.35	118000
18	Andrew Ad	258	150	80	201	Pierfrance	22000	1.4E+08	Action|Ad	Peter Dinl	The Chron	149922	22697	DamiÃ¡n A	4	brother bı	http://ww	438	English	USA	PG	2.3E+08	2008	216	6.6	2.35	0
19	Joss Whe	703	173	0	19000	Robert Do	26000	6.2E+08	Action|Ad	Chris Hem	The Aveng	995415	87697	Scarlett Jo	3	alien inva	http://ww	1722	English	USA	PG-13	2.2E+08	2012	21000	8.1	1.85	123000
20	Rob Marsl	448	136	252	1000	Sam Clafli	40000	2.4E+08	Action|Ad	Johnny De	Pirates of	370704	54083	Stephen G	4	blackbear	http://ww	484	English	USA	PG-13	2.5E+08	2011	11000	6.7	2.35	58000
21	Barry Son	451	106	188	718	Michael S	10000	1.8E+08	Action|Ad	Will Smith	Men in Bla	268154	12572	Nicole Sch	1	alien|crin	http://ww	341	English	USA	PG-13	2.3E+08	2012	816	6.8	1.85	40000
22	Peter Jack	422	164	0	773	Adam Bro	5000	2.6E+08	Adventure	Aidan Turi	The Hobbi	354228	9152	James Ne	0	army|elf|	http://ww	802	English	New Zeala	PG-13	2.5E+08	2014	972	7.5	2.35	65000
23	Marc Web	599	153	464	963	Andrew G	15000	2.6E+08	Action|Ad	Emma Sto	The Amazi	451803	28489	Chris Zylka	0	lizard|out	http://ww	1225	English	USA	PG-13	2.3E+08	2012	10000	7	2.35	56000
24	Ridley Sco	343	156	0	738	William H	891	1.1E+08	Action|Ad	Mark Add	Robin Hoc	211765	3244	Scott Grim	0	1190s|arc	http://ww	546	English	USA	PG-13	2E+08	2010	882	6.7	2.35	17000
25	Peter Jack	509	186	0	773	Adam Bro	5000	2.6E+08	Adventure	Aidan Turi	The Hobbi	483540	9152	James Ne	6	dwarf|elf	http://ww	951	English	USA	PG-13	2.3E+08	2013	972	7.9	2.35	83000
26	Chris Wei	251	113	129	1000	Eva Green	16000	7E+07	Adventure	Christoph	The Golde	149019	24106	Kristin Sco	2	children|	http://ww	666	English	USA	PG-13	1.8E+08	2007	6000	6.1	2.35	0
27	Peter Jack	446	201	0	84	Thomas K	6000	2.2E+08	Action|Ad	Naomi Wa	King Kong	316018	7123	Evan Park	0	animal na	http://ww	2618	English	New Zeala	PG-13	2.1E+08	2005	919	7.2	2.35	0
28	James Cai	315	194	0	794	Kate Wins	29000	6.6E+08	Drama|Rc	Leonardo	TitanicÂ	793059	45223	Gloria Stu	0	artist|lov	http://ww	2528	English	USA	PG-13	2E+08	1997	14000	7.7	2.35	26000
29	Anthony R	516	147	94	11000	Scarlett Jo	21000	4.1E+08	Action|Ad	Robert Do	Captain A	272670	64798	Chris Evar	0	based on	http://ww	1022	English	USA	PG-13	2.5E+08	2016	19000	8.2	2.35	72000
30	Peter Berg	377	131	532	627	Alexander	14000	6.5E+07	Action|Ad	Liam Nee	Battleship	202382	26679	Tadanobu	0	box office	http://ww	751	English	USA	PG-13	2.1E+08	2012	10000	5.9	2.35	44000
31	Colin Trev	644	124	365	1000	Judy Gree	3000	6.5E+08	Action|Ad	Bryce Dall	Jurassic W	418214	8458	Omar Sy	0	dinosaur|	http://ww	1290	English	USA	PG-13	1.5E+08	2015	2000	7	2	150000

IMDB_Movies(Original)

Insights

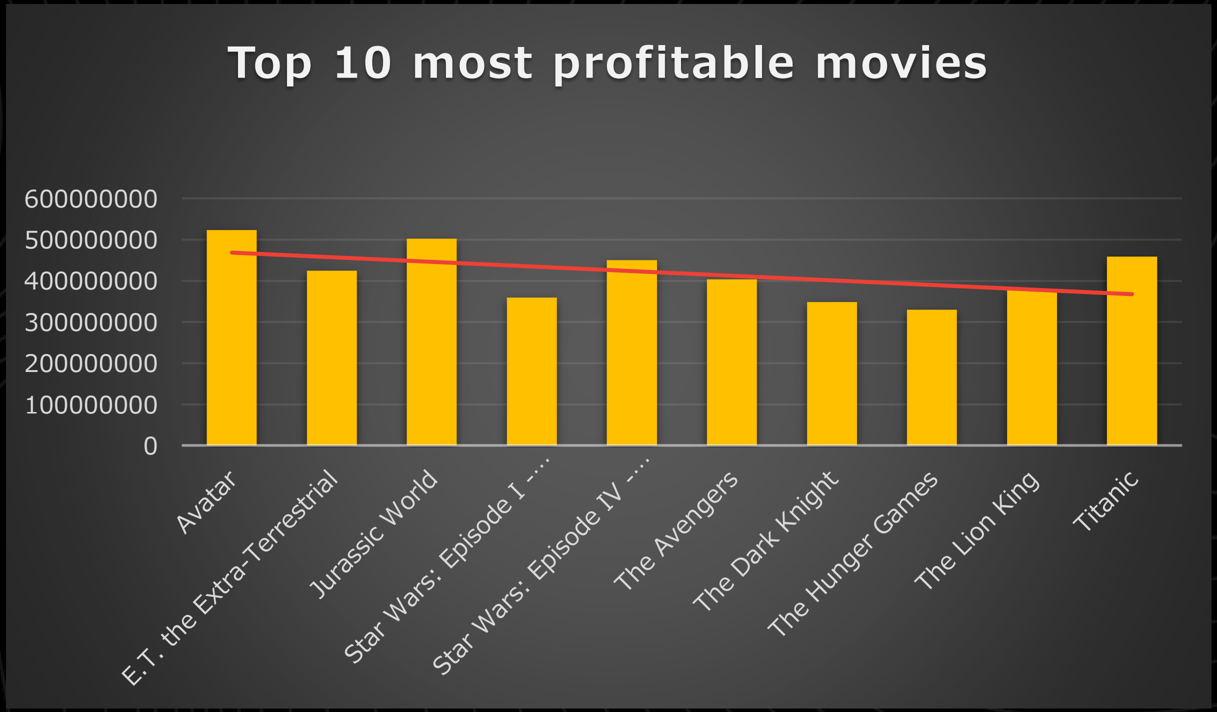
Preview after cleaning:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	director_name	num_critic_for_reviews	duration	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	budget	title_year	imdb_score
2	James Cameron	723	178	7.6E+08	Action|Adventur	CCH Pounder	Avatar	886204	3054	English	USA	2.4E+08	2009	7.9
3	Gore Verbinski	302	169	3.1E+08	Action|Adventur	Johnny Depp	Pirates of the Caribbean: At World's End	471220	1238	English	USA	3E+08	2007	7.1
4	Sam Mendes	602	148	2E+08	Action|Adventur	Christoph Waltz	Spectre	275868	994	English	UK	2.5E+08	2015	6.8
5	Christopher Nolan	813	164	4.5E+08	Action|Thriller	Tom Hardy	The Dark Knight Rises	1144337	2701	English	USA	2.5E+08	2012	8.5
7	Andrew Stanton	462	132	7.3E+07	Action|Adventur	Daryl Sabara	John Carter	212204	738	English	USA	2.6E+08	2012	6.6
8	Sam Raimi	392	156	3.4E+08	Action|Adventur	J.K. Simmons	Spider-Man 3	383056	1902	English	USA	2.6E+08	2007	6.2
9	Nathan Greno	324	100	2E+08	Adventure|Anim	Brad Garrett	Tangled	294810	387	English	USA	2.6E+08	2010	7.8
10	Joss Whedon	635	141	4.6E+08	Action|Adventur	Chris Hemsworth	Avengers: Age of Ultron	462669	1117	English	USA	2.5E+08	2015	7.5
11	David Yates	375	153	3E+08	Adventure|Famil	Alan Rickman	Harry Potter and the Half-Blood Prince	321795	973	English	UK	2.5E+08	2009	7.5
12	Zack Snyder	673	183	3.3E+08	Action|Adventur	Henry Cavill	Batman v Superman: Dawn of Justice	371639	3018	English	USA	2.5E+08	2016	6.9
13	Bryan Singer	434	169	2E+08	Action|Adventur	Kevin Spacey	Superman Returns	240396	2367	English	USA	2.1E+08	2006	6.1
14	Marc Forster	403	106	1.7E+08	Action|Adventur	Giancarlo Giannir	Quantum of Solace	330784	1243	English	UK	2E+08	2008	6.7
15	Gore Verbinski	313	151	4.2E+08	Action|Adventur	Johnny Depp	Pirates of the Caribbean: Dead Man's Chest	522040	1832	English	USA	2.3E+08	2006	7.3
16	Gore Verbinski	450	150	8.9E+07	Action|Adventur	Johnny Depp	The Lone Ranger	181792	711	English	USA	2.2E+08	2013	6.5
17	Zack Snyder	733	143	2.9E+08	Action|Adventur	Henry Cavill	Man of Steel	548573	2536	English	USA	2.3E+08	2013	7.2
18	Andrew Adamson	258	150	1.4E+08	Action|Adventur	Peter Dinklage	The Chronicles of Narnia: Prince Caspian	149922	438	English	USA	2.3E+08	2008	6.6
19	Joss Whedon	703	173	6.2E+08	Action|Adventur	Chris Hemsworth	The Avengers	995415	1722	English	USA	2.2E+08	2012	8.1
20	Rob Marshall	448	136	2.4E+08	Action|Adventur	Johnny Depp	Pirates of the Caribbean: On Stranger Tides	370704	484	English	USA	2.5E+08	2011	6.7
21	Barry Sonnenfeld	451	106	1.8E+08	Action|Adventur	Will Smith	Men in Black 3	268154	341	English	USA	2.3E+08	2012	6.8
22	Peter Jackson	422	164	2.6E+08	Adventure|Fanta	Aidan Turner	The Hobbit: The Battle of the Five Armies	354228	802	English	New Zealai	2.5E+08	2014	7.5
23	Marc Webb	599	153	2.6E+08	Action|Adventur	Emma Stone	The Amazing Spider-Man	451803	1225	English	USA	2.3E+08	2012	7
24	Ridley Scott	343	156	1.1E+08	Action|Adventur	Mark Addy	Robin Hood	211765	546	English	USA	2E+08	2010	6.7
25	Peter Jackson	509	186	2.6E+08	Adventure|Fanta	Aidan Turner	The Hobbit: The Desolation of Smaug	483540	951	English	USA	2.3E+08	2013	7.9
26	Chris Weitz	251	113	7E+07	Adventure|Famil	Christopher Lee	The Golden Compass	149019	666	English	USA	1.8E+08	2007	6.1
27	Peter Jackson	446	201	2.2E+08	Action|Adventur	Naomi Watts	King Kong	316018	2618	English	New Zealai	2.1E+08	2005	7.2
28	James Cameron	315	194	6.6E+08	Drama|Romance	Leonardo DiCapri	Titanic	793059	2528	English	USA	2E+08	1997	7.7

Insights

2A.Top 10 movies with highest profit:

- We have gross and budget columns in our dataset, the profit was calculated by subtracting budget from the gross amount and was shown as a new column.
- The profit column was sorted with top 10 after selecting the column and this has given us the top 10 movies with highest profit.
- A graph of Profit vs Top 10 movies was plotted for better understanding.



Insights

2B. Observing the outliers:

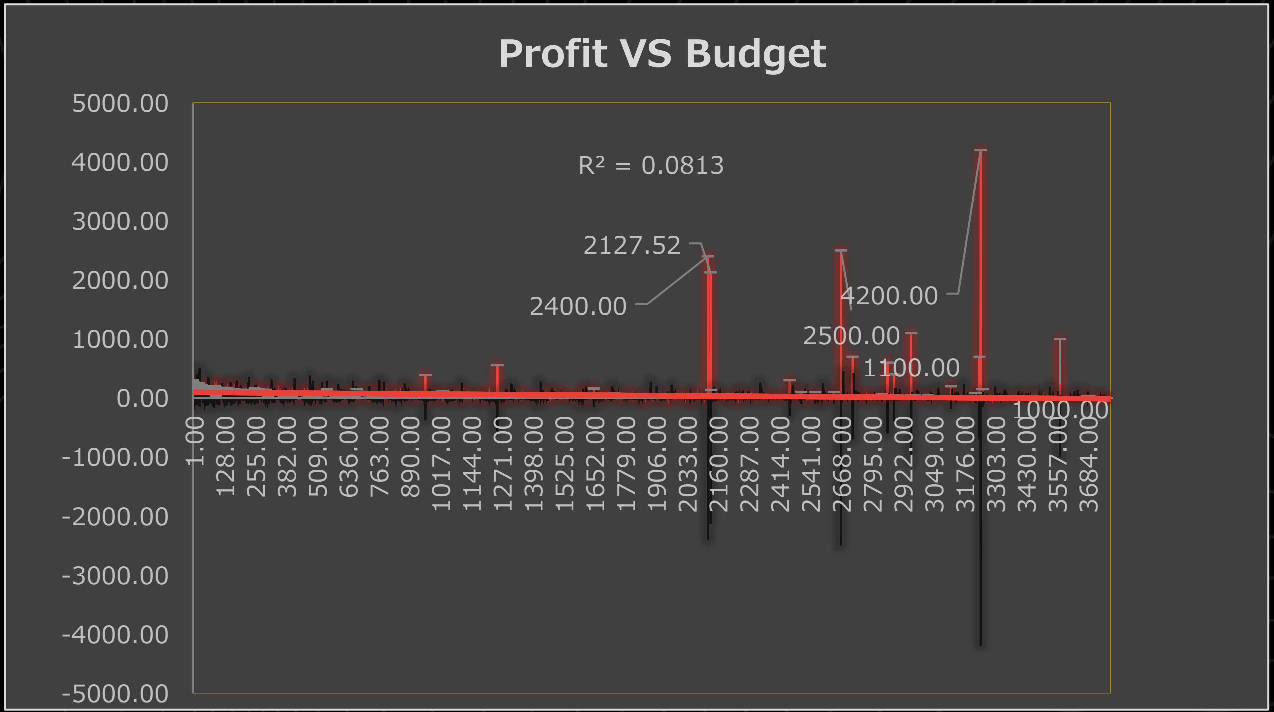
- ✓ Major outliers are on points:

  2127.52,2400,2500,1100 and 4200

  having the R2 Score 0.0813.

- ✓ The profit and budget were divided by 100000

  For easier calculation.



IMDb

The Internet Movie Database

TOP 250

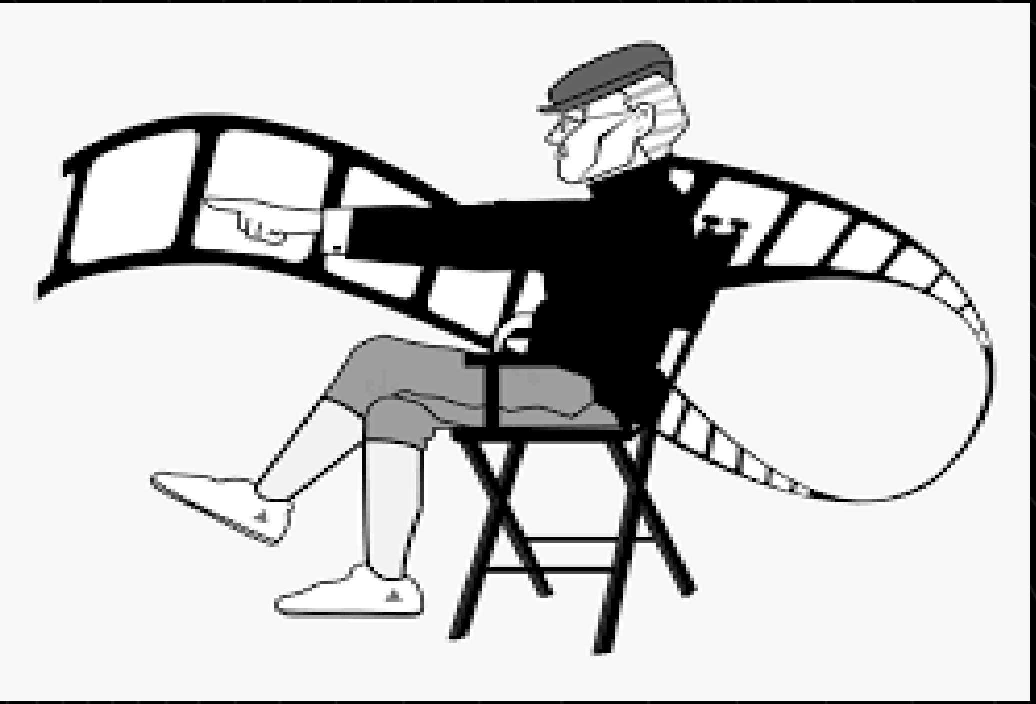
Insights

3. IMDb Top 250 movies:

- ✓ Using sort and filter method,the movies >25000 num_voted_users were filtered.
- ✓ The IMDb scores were arranged in a decreasing order and only top 250 rows were chosen for evaluation.
- ✓ A new Rank column was added to rank the movies from 1 to 250(Using cell+1,drag and fill method)
- ✓ By unselecting the ‘English’ language from language column we get the movies of foreign_language under IMDbs top 250 list.

Top 250 of IMDb preview:

	A	B	C	D	E
1	Rank	IMDb_Top_250	imdb_score	num_voted_users	Top_Foreign_Lang_Film
2	1	The Shawshank Redemption	9.3	1689764	The Good, the Bad and the Ugly
3	2	The Godfather	9.2	1155770	City of God
4	3	The Dark Knight	9	1676169	Seven Samurai
5	4	The Godfather: Part II	9	790926	Spirited Away
6	5	The Lord of the Rings: The Return of the King	8.9	1215718	Samsara
7	6	Schindler's List	8.9	865020	
8	7	Pulp Fiction	8.9	1324680	
9	8	The Good, the Bad and the Ugly	8.9	503509	
10	9	Inception	8.8	1468200	
11	10	The Lord of the Rings: The Fellowship of the Ring	8.8	1238746	
12	11	Fight Club	8.8	1347461	
13	12	Forrest Gump	8.8	1251222	
14	13	Star Wars: Episode V - The Empire Strikes Back	8.8	837759	
15	14	The Lord of the Rings: The Two Towers	8.7	1100446	
16	15	The Matrix	8.7	1217752	
17	16	Goodfellas	8.7	728685	
18	17	Star Wars: Episode IV - A New Hope	8.7	911097	
19	18	One Flew Over the Cuckoo's Nest	8.7	680041	
20	19	City of God	8.7	533200	
21	20	Seven Samurai	8.7	229012	




DIRECTOR

Insights

4.Finding the best directors:

To find the best directors based on the average imdb scores,

- ✓ A pivot table of Director_name and imdb_score was created.
- ✓ In the values section of the pivot table the mean(average) of imdb_score was calculated.
- ✓ The values were sorted in descending order.

3	Director_name	Average of imdb_score
4	Tony Kaye	8.60
5	Charles Chaplin	8.60
6	Alfred Hitchcock	8.50
7	Ron Fricke	8.50
8	Damien Chazelle	8.50
9	Majid Majidi	8.50
10	Sergio Leone	8.43
11	Christopher Nolan	8.43
12	S.S. Rajamouli	8.40
13	Richard Marquand	8.40
14	Marius A. Markevicius	8.40
15	Asghar Farhadi	8.40

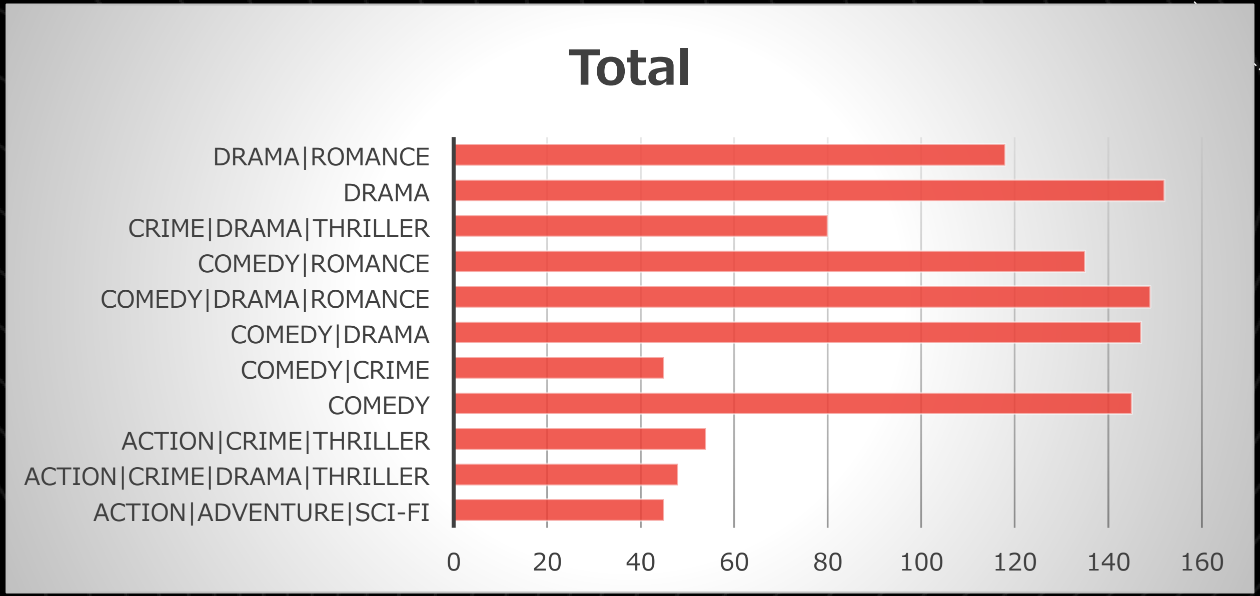

Comedy
History
Sci-Fi
Romance
Thriller
Mystery
Drama
Horror
War
Action
Musical
Superhero
Animation
Road Movie
Fantasy
Movie

Insights

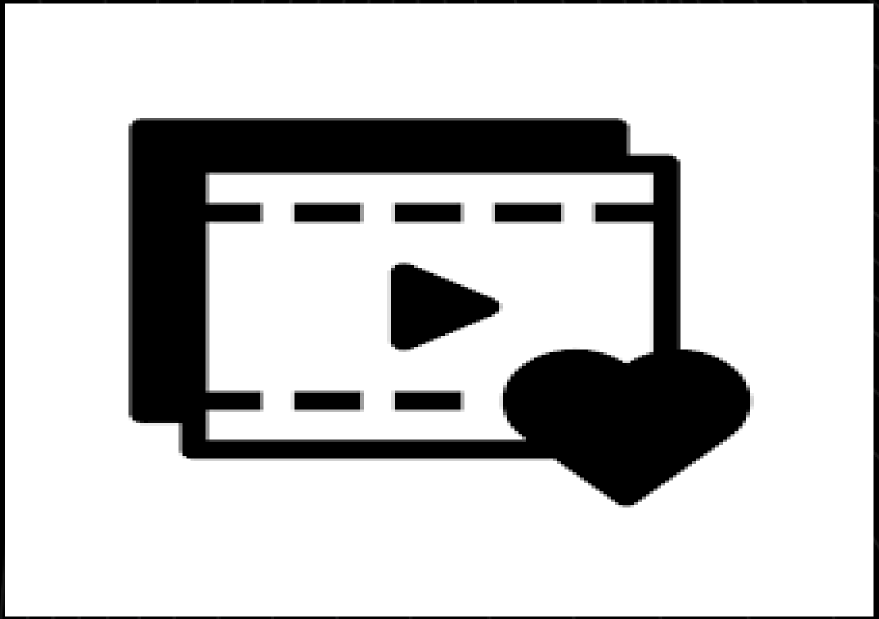
5.Finding popular genres:

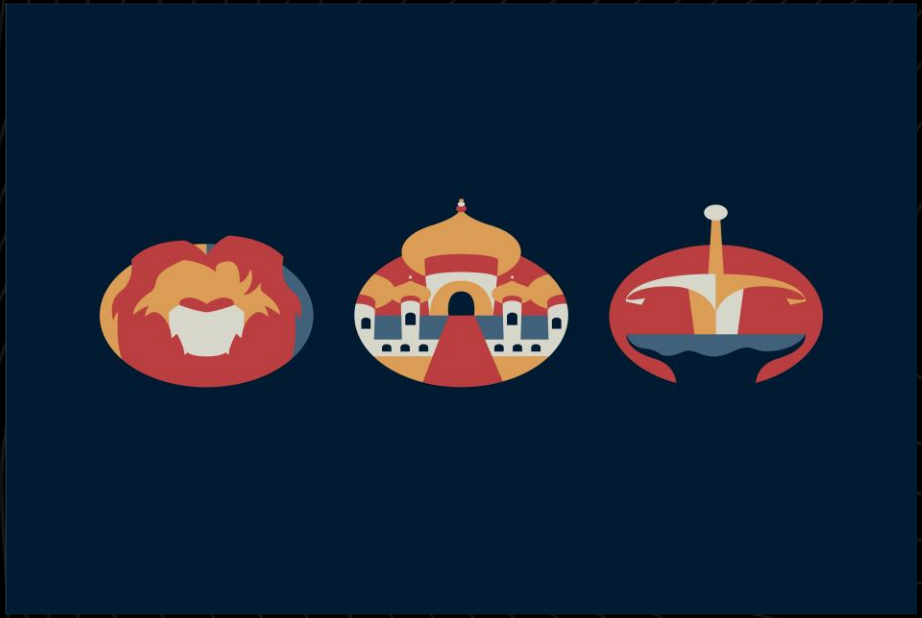
To find the popular genre based on the average imdb scores,

- ✓ A pivot table of genres and imdb_score was created.
- ✓ In the values section of the pivot table the mean(average) of imdb_score was calculated.
- ✓ The values were sorted in descending order.
- ✓ For better understanding, genre-wise count and average profit was also calculated using pivot table.



	A	B	C	D
1	Genre	Average imdb_score	genres count	Average Profit
2	Action|Adventure|Sci-Fi	6.67	45	41416969.69
3	Action|Crime|Drama|Thriller	6.52	48	-1516065.438
4	Action|Crime|Thriller	6.40	54	8055116.611
5	Comedy	5.84	145	21034780.79
6	Comedy|Crime	6.04	45	16373313.49
7	Comedy|Drama	6.58	147	8344097.81
8	Comedy|Drama|Romance	6.50	149	10352425.22
9	Comedy|Romance	5.90	135	20002148.13
10	Crime|Drama|Thriller	6.87	80	7115124.925
11	Drama	7.04	152	1314534.941
12	Drama|Romance	6.95	118	9366491.754
13				





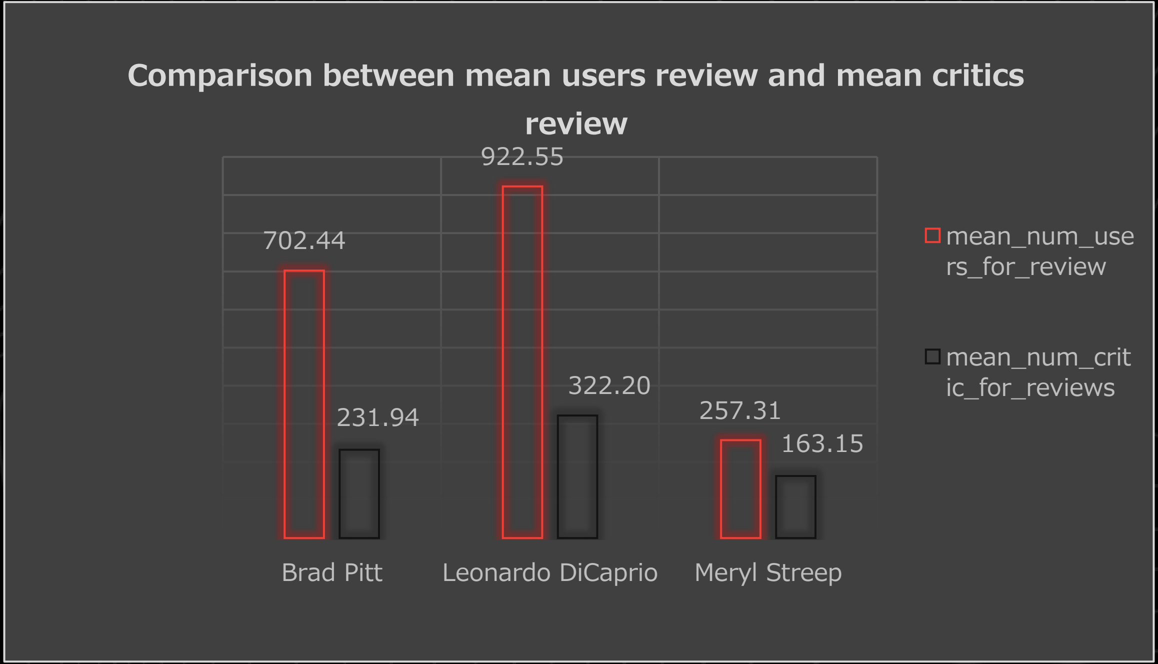
Insights

6A. Inspecting the critic and audience favourite actors:

Since the inspection was asked on specific three actors,

- ✓ Three separate columns Meryl_Streep, Leo_Caprio, and Brad_Pitt were created.
- ✓ These three columns contain movies where the said actors were lead actors , I have used pivot table for this action.
- ✓ The actor_1_name column containing the lead actor names under respective movies was used for extracton.
- ✓ In the values section of pivot table, mean of num_critic_for_reviews and mean of num_user_for_reviews were put against the actor names.

2			
3	Row Labels	mean_num_users_for_review	mean_num_critic_for_reviews
4	Brad Pitt	702.44	231.94
5	Leonardo DiCaprio	922.55	322.20
6	Meryl Streep	257.31	163.15
7			
8			



THE OSCARS

Insights

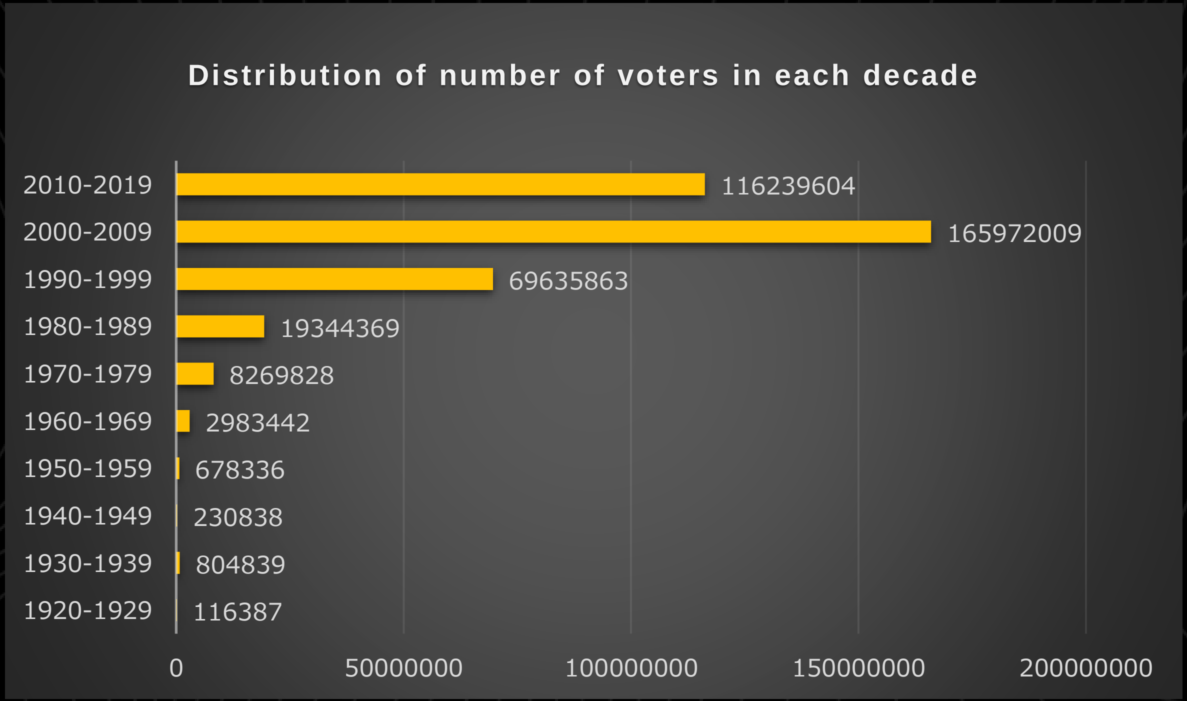
6B. Calculating decade-wise voted users:

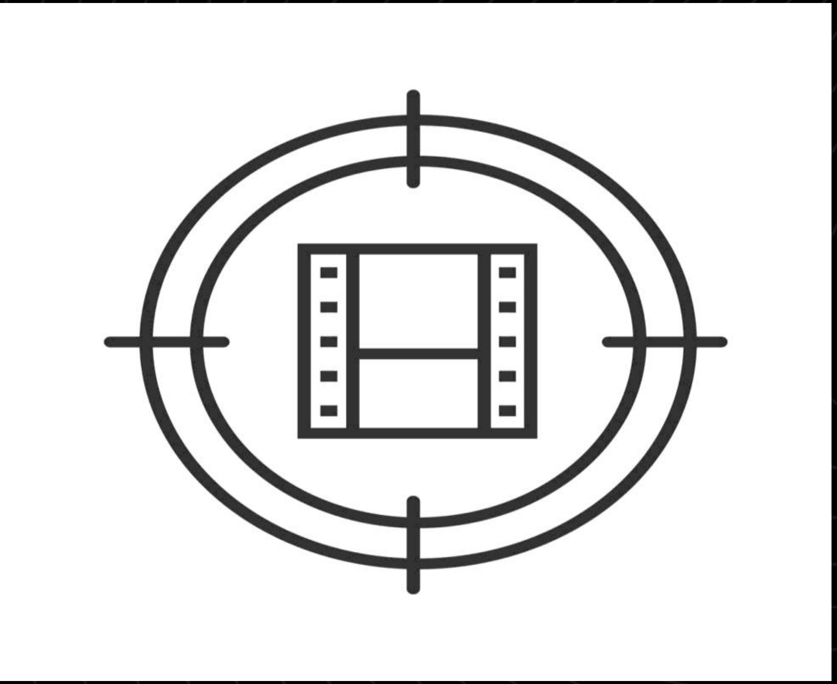
- ✓ For this calculation I have created one new column names ‘Decade’ that divides the title_year into decades of 10 years(e.g: 1920-1929).

Used code for creating the decade column:

=CONCATENATE(LEFT(A3, 3), "9-", LEFT(A3, 3), "0")

- ✓ Dragged and filled the rows using this formula/code.
- ✓ Created a pivot table, dragged the Decade column and sorted it , in the values section calculated the sum of num_voted_users to get Decade wise number of voted users.





Key findings

- The movie with highest profit is **Avatar(**523505847) followed by **Jurassic world(502177271) ,Titanic(458672302)** and seven other movies.
- **The Shawshank Redemption** is the highest rated movie and also number one movie in IMDb and IMDb top 250 list with highest number of voted users.
- **The foreign language movies in IMDbs top 250 list are:** 1.The Good the Bad and the Ugly 2. Seven Samurai 3.Samsara 4.City of God 5.Spirited Away
- **Tony Kaye and Charles Chaplin** are the top two Best directors amount the list of ten holding the same average IMDb score(8.60)
- **Drama** is the most popular movie genre holding 152 counts.
- **Leonardo DiCaprio is** the highest user and critic rated actor , the most popular actor.
- The highest number of movie voters were seen in the **2000-2009 decade** holding 165972009 votes.

Achievements

- **Data handling**: The entire project has helped me in learning methods to handle large dataset and tract records of each column.
- **Data cleaning**: This project has helped in understanding the data cleaning process after inspecting required queries.
- **Uses of Microsoft excel as business tool**: learned how Microsoft excel is being used with it’s entire diversity in the process of statistical calculations.
- **Excel filters , sorting and formulas:** Learned different methods of filtering and sorting data based on requirements and implementation of excel formulas when needed.
- **The power of Powerpoint**: Learned how Powerpoint can strongly help explaining complex business queries in simple ways and slides.

THANK YOU