IMDB Movie Analysis

Kushagra Singh
March 2023

Description:

We have here dataset for IMDB Movies. I have framed the problem. For this task, I have define a problem I want to shed light on.

I did this by asking 'What?' i.e. What is the problem? Eg. What do I see happening? What is my hypothesis for the cause of the problem? What is the impact of the problem on stakeholders? What is the impact of the problem not being solved?

Answering these questions helped define problem I was trying to solve and will allowed me to find the right data to solve it.
Now next step I cleaned the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.

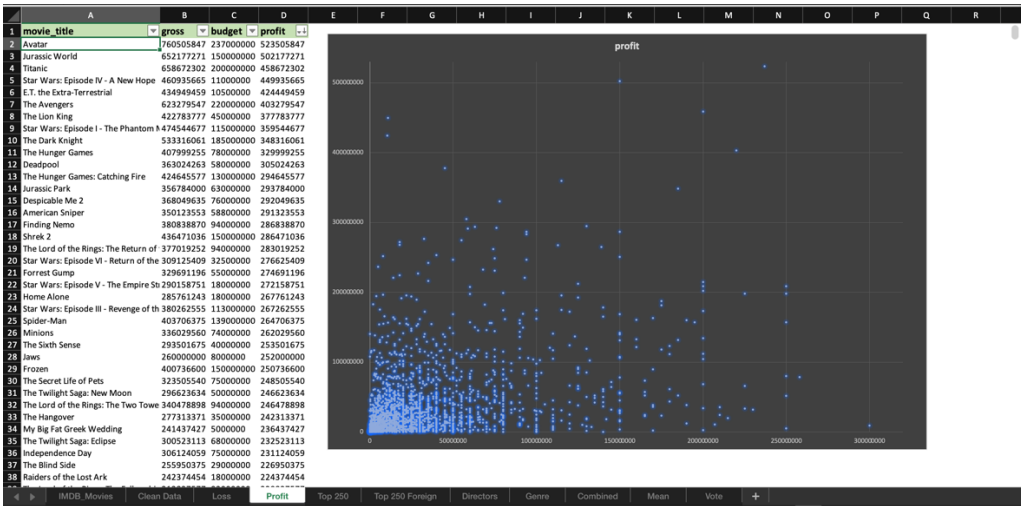
Using Advanced Excel and Statistical concepts we have found out movies with the highest profit, top movies as per imdb rating, top directors, most popular genres, top foreign language films and more.

Data Cleaning:

Before analysing anything in data I have performed Data Cleaning.
Where I performed three things

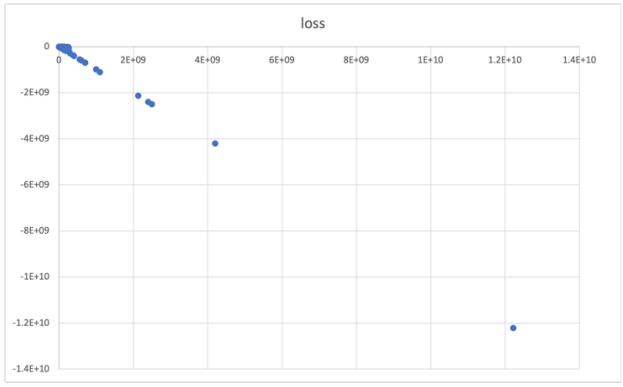
- First, I dropped the columns which have no use for the analysis.
- Second, I dropped the rows which are blank/null.
- Third, removed the duplicate row values.

Q1. Which movie had the highest profit?



movie_title	gross	budget	profit
Avatar	760505847	237000000	523505847
Jurassic World	652177271	150000000	502177271
Titanic	658672302	200000000	458672302
Star Wars: Episode IV - A New Hope	460935665	11000000	449935665
E.T. the Extra-Terrestrial	434949459	10500000	424449459
The Avengers	623279547	220000000	403279547
The Lion King	422783777	45000000	377783777
Star Wars: Episode I - The Phantom M	474544677	115000000	359544677
The Dark Knight	533316061	185000000	348316061
The Hunger Games	407999255	78000000	329999255
Deadpool	363024263	58000000	305024263
The Hunger Games: Catching Fire	424645577	130000000	294645577
Jurassic Park	356784000	63000000	293784000
Despicable Me 2	368049635	76000000	292049635
American Sniper	350123553	58800000	291323553
Finding Nemo	380838870	94000000	286838870
Shrek 2	436471036	150000000	286471036
The Lord of the Rings: The Return of	377019252	94000000	283019252
Star Wars: Episode VI - Return of the	309125409	32500000	276625409
Forrest Gump	329691196	55000000	274691196
Star Wars: Episode V - The Empire Str	290158751	18000000	272158751
Home Alone	285761243	18000000	267761243
Star Wars: Episode III - Revenge of th	380262555	113000000	267262555
Spider-Man	403706375	139000000	264706375
Minions	336029560	74000000	262029560
The Sixth Sense	293501675	40000000	253501675
Jaws	260000000	8000000	252000000
Frozen	400736600	150000000	250736600
The Secret Life of Pets	323505540	75000000	248505540
The Twilight Saga: New Moon	296623634	50000000	246623634
The Lord of the Rings: The Two Towe	340478898	94000000	246478898
The Hangover	277313371	35000000	242313371
My Big Fat Greek Wedding	241437427	5000000	236437427
The Twilight Saga: Eclipse	300523113	68000000	232523113
Independence Day	306124059	75000000	231124059
The Blind Side	255950375	29000000	226950375
Raiders of the Lost Ark	242374454	18000000	224374454

Avatar movie made the highest profit



Some observed loss too removing some outlier values gives us better idea

movie_title	gross	budget	loss
Poseidon	60655503	160000000	-99344497
The Widow of Saint-Pierre	3058380	100000000	-96941620
Tomorrowland	93417865	190000000	-96582135
R.I.P.D.	33592415	130000000	-96407585
Hugo	73820094	170000000	-96179906
The Adventures of Pluto Nash	4411102	100000000	-95588898
Robin Hood	105219735	200000000	-94780265
The Flowers of War	9213	94000000	-93990787
Alice Through the Looking Glass	76846624	170000000	-93153376
Spy Game	26871	92000000	-91973129
The Island	35799026	126000000	-90200974
The Nutcracker in 3D	190562	90000000	-89809438
How Do You Know	30212620	120000000	-89787380
House of Flying Daggers	11041228	100000000	-88958772
Pacific Rim	101785482	190000000	-88214518
The Wolfman	61937495	150000000	-88062505
The BFG	52792307	140000000	-87207693
Cutthroat Island	11000000	98000000	-87000000
The Sorcerer's Apprentice	63143812	150000000	-86856188
Waterworld	88246220	175000000	-86753780
Around the World in 80 Days	24004159	110000000	-85995841
The Alamo	22406362	107000000	-84593638
Flushed Away	64459316	149000000	-84540684
I Served the King of England	617228	84450000	-83832772
Green Lantern	116593191	200000000	-83406809
The Chronicles of Narnia: Prince Ca	141614023	225000000	-83385977
Town & Country	6712451	90000000	-83287549
Kingdom of Heaven	47396698	130000000	-82603302
Australia	49551662	130000000	-80448338
The Little Prince	1339152	81200000	-79860848
Seventh Son	17176900	95000000	-77823100
Edge of Tomorrow	100189501	178000000	-77810499
Transcendence	23014504	100000000	-76985496
White House Down	73103784	150000000	-76896216
Point Break	28772222	105000000	-76227778
Speed Racer	43929341	120000000	-76070659
In the Heart of the Sea	24985612	100000000	-75014388
Exodus: Gods and Kings	65007045	140000000	-74992955
Cloud Atlas	27098580	102000000	-74901420
Evan Almighty	100289690	175000000	-74710310

Here we have removed loss above 100000000

Interesting thing is movie Happiness made no profit no loss.

Approach: Created a new column called profit which contains the difference of the two columns: gross and budget. Sorted the column using the profit column as reference. Ploted profit (y-axis) vs budget (x- axis) and observed the outliers using the appropriate chart type.

Q2. What are the top 250 IMDB movies?

	movie_title	num_voted_users	imdb_score	rank
2	The Shawshank Redemption	1689764	9.3	1
3	The Godfather	1155770	9.2	2
4	The Dark Knight	1676169	9	3
5	The Godfather: Part II	790926	9	4
6	Pulp Fiction	1324680	8.9	5
7	The Lord of the Rings: The Return of the King	1215718	8.9	6
8	Schindler's List	865020	8.9	7
9	The Good, the Bad and the Ugly	503509	8.9	8
10	Inception	1468200	8.8	9
11	Fight Club	1347461	8.8	10
12	Forrest Gump	1251222	8.8	11
13	The Lord of the Rings: The Fellowship of the Ring	1238746	8.8	12
14	Star Wars: Episode V - The Empire Strikes Back	837759	8.8	13
15	The Matrix	1217752	8.7	14
16	The Lord of the Rings: The Two Towers	1100446	8.7	15
17	Star Wars: Episode IV - A New Hope	911097	8.7	16
18	Goodfellas	728685	8.7	17
19	One Flew Over the Cuckoo's Nest	680041	8.7	18
20	City of God	533200	8.7	19
21	Seven Samurai	229012	8.7	20
22	Se7en	1023511	8.6	21
23	Interstellar	928227	8.6	22
24	The Silence of the Lambs	887467	8.6	23
25	Saving Private Ryan	881236	8.6	24
26	American History X	782437	8.6	25
27	The Usual Suspects	740918	8.6	26
28	Spirited Away	417971	8.6	27
29	Modern Times	143086	8.6	28
30	The Dark Knight Rises	1144337	8.5	29
31	Gladiator	982637	8.5	30
32	Django Unchained	955174	8.5	31
33	The Departed	873649	8.5	32
34	Memento	845580	8.5	33
35	The Prestige	844052	8.5	34
36	The Green Mile	782610	8.5	35
37	Terminator 2: Judgment Day	744891	8.5	36
38	Back to the Future	732212	8.5	37
39	Raiders of the Lost Ark	661017	8.5	38
40	The Lion King	644348	8.5	39
41	Alien	563827	8.5	40

IMDB_Movies | Clean Data | Loss | Profit | Top 250 | Top 250 For

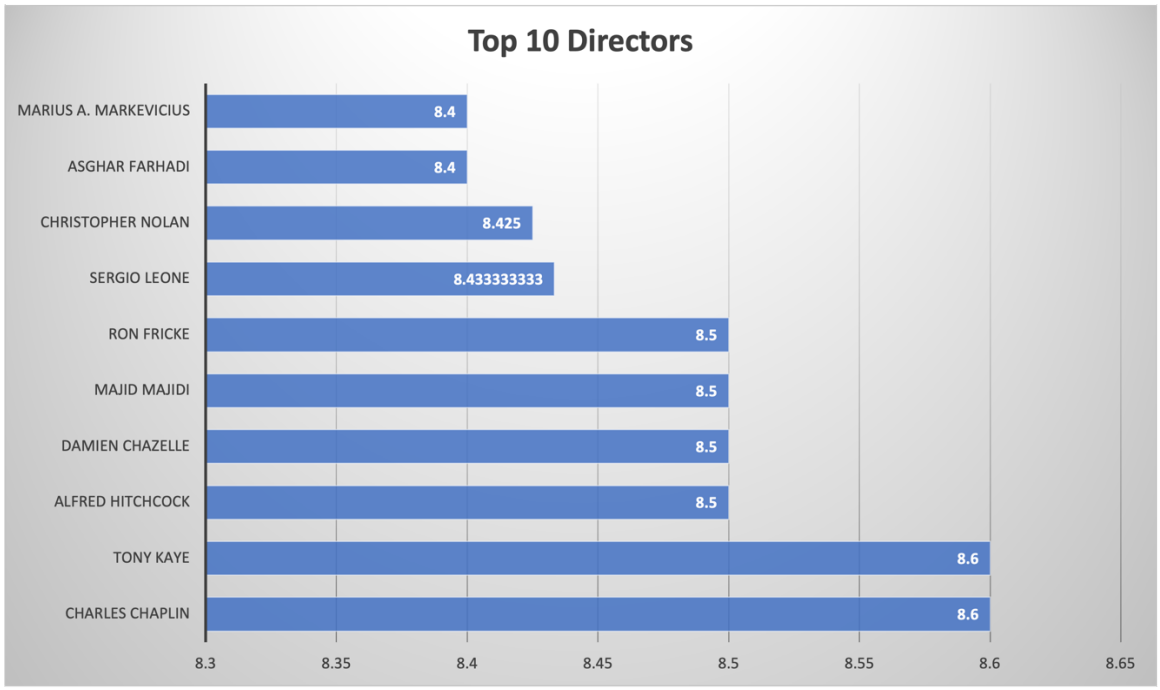
Approach: Created a new column IMDb_Top_250 and stored the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also made sure that for all of these movies, the num_voted_users is greater than 25,000. Also added a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Q3. Top 250 IMDB movies not in English?

	A	B	C	D	E
1	Top_Foreign_Lang_Film	num_voted_users	language	imdb_score	rank
2	The Good, the Bad and the Ugly	503509	Italian	8.9	1
3	City of God	533200	Portuguese	8.7	2
4	Seven Samurai	229012	Japanese	8.7	3
5	Spirited Away	417971	Japanese	8.6	4
6	The Lives of Others	259379	German	8.5	5
7	Children of Heaven	27882	Persian	8.5	6
8	AmvÂ©lie	534262	French	8.4	7
9	Oldboy	356181	Korean	8.4	8
10	Princess Mononoke	221552	Japanese	8.4	9
11	Das Boot	168203	German	8.4	10
12	A Separation	151812	Persian	8.4	11
13	Baahubali: The Beginning	62756	Telugu	8.4	12
14	Downfall	248354	German	8.3	13
15	The Hunt	170155	Danish	8.3	14
16	Metropolis	111841	German	8.3	15
17	Pan's Labyrinth	467234	Spanish	8.2	16
18	Howl's Moving Castle	214091	Japanese	8.2	17
19	The Secret in Their Eyes	131831	Spanish	8.2	18
20	Incendies	80429	French	8.2	19
21	Amores Perros	173551	Spanish	8.1	20
22	Akira	106160	Japanese	8.1	21
23	Elite Squad	81644	Portuguese	8.1	22
24	The Celebration	65951	Danish	8.1	23
25	The Sea Inside	64556	Spanish	8.1	24
26	Tae Guk Gi: The Brotherhood of War	31943	Korean	8.1	25
27	A Fistful of Dollars	147566	Italian	8	26
28	Persepolis	70194	French	8	27
29	My Name Is Khan	69759	Hindi	8	28
30	Waltz with Bashir	46107	Hebrew	8	29
31	Central Station	28951	Portuguese	8	30
32	Crouching Tiger, Hidden Dragon	217740	Mandarin	7.9	31
33	Hero	149414	Mandarin	7.9	32
34	Letters from Iwo Jima	132149	Japanese	7.9	33
35	Amour	70382	French	7.9	34
36	4 Months, 3 Weeks and 2 Days	44763	Romanian	7.9	35
37	The Chorus	44151	French	7.9	36
38	Nine Queens	38215	Spanish	7.9	37
39	Veer-Zaara	34449	Hindi	7.9	38
40	Apocalypto	236000	Maya	7.8	39
41	Run Lola Run	101471	German	7.8	40

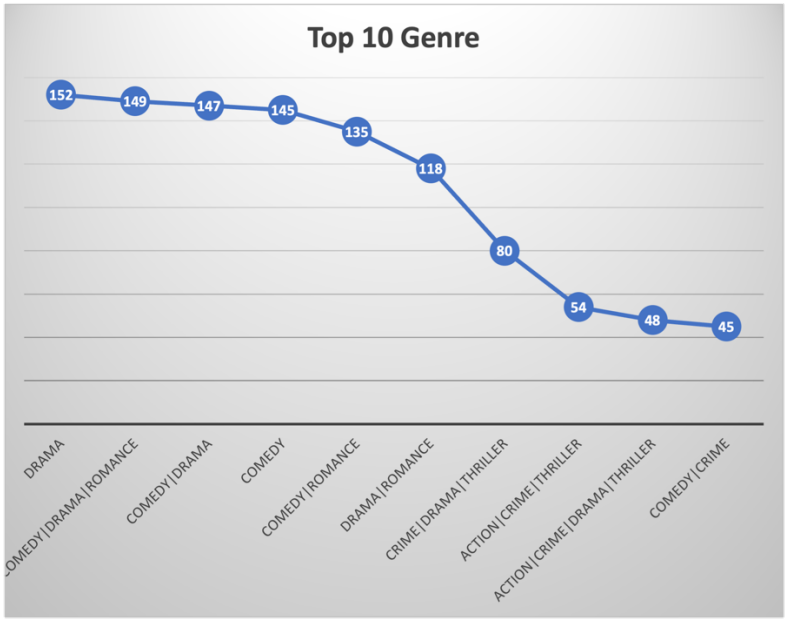
Approach: Extracted all the movies in the IMDb_Top_250 column which are not in the English language and storeed them in a new column named Top_Foreign_Lang_Film.

Q3. Top 10 Directors as per imdb_rating?

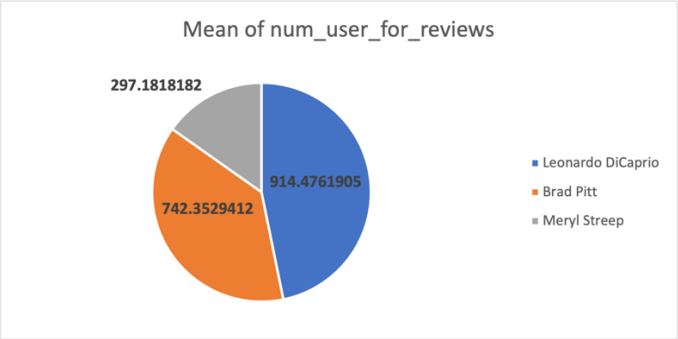


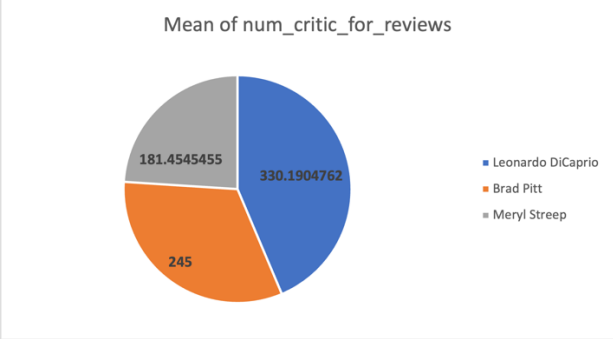
Approach: Grouped the column using the director_name column. Found out the top 10 directors for whom the mean of imdb_score is the highest and stored them in a new column top10director. In case of a tie in IMDb score between two directors, sorted them alphabetically. (graph lowest to higest)

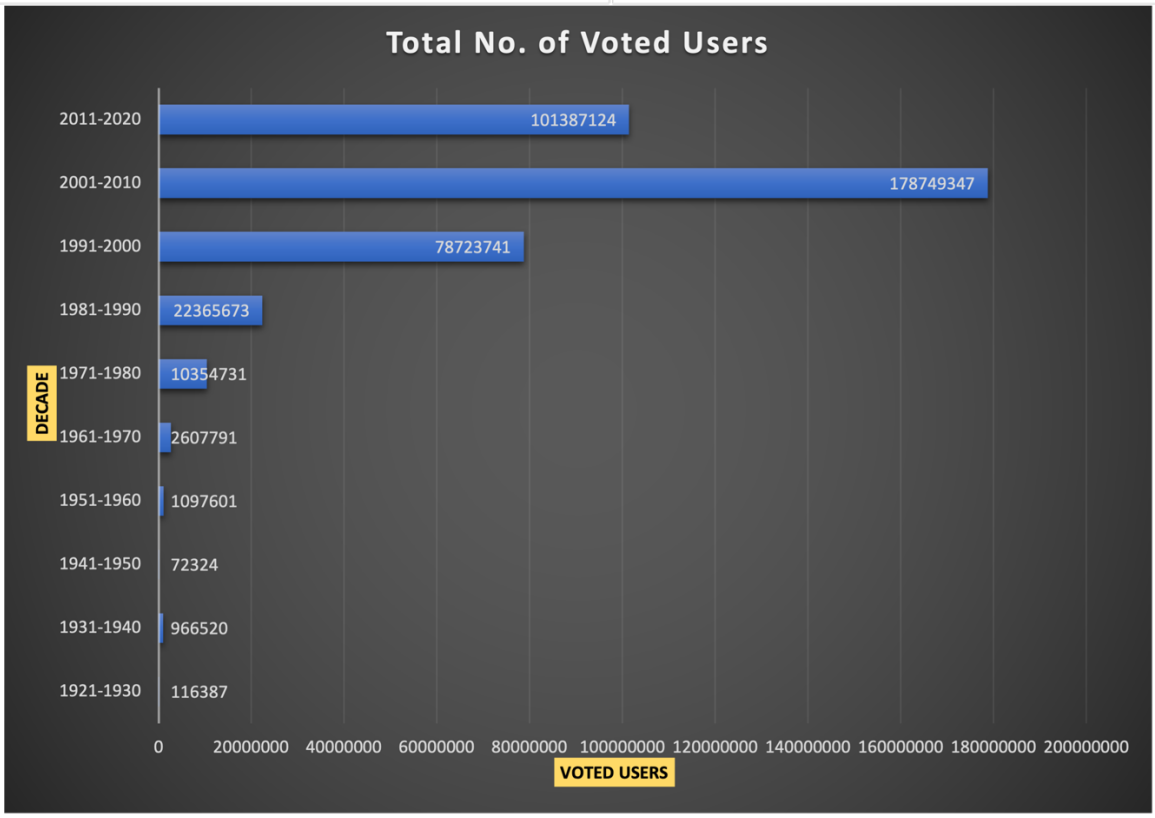
Q4. Popular Genres?



Q5. Find the critic-favorite and audience-favorite actors?







Approach: Created three new columns for 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. Used only the actor_1_name column for extraction.
Appended the rows of all these columns and storeed them in a new column named Combined. Grouped the combined column using the actor_1_name column.
Found the mean of the num_critic_for_reviews and num_users_for_review and identifed the actors which have the highest mean.
Observed the change in number of voted users over decades using a bar chart. Created a column called decade which represents the decade to which every movie belongs

Data set used: https://github.com/ks127d/Data-Analytics-Project/blob/dc7089f1f709c1b8e38250545efbb35a6312e51a/IMDB%20Movie%20Analysis/~$IMDB_Movies.xlsx
(contains analysis too)