

IMDB Movie Analysis

Statistics Final Project- 1

Project Description

In this project, we will perform statistical analysis in order to understand IMBD.

Approach

Using MS Excel we will do statistical analysis of data records in IMBD Database

Tech-Stack Used: Microsoft Excel 2016

- MS Excel 2016 is a spreadsheet program where one can record data in the form of tables.
- It is easy to analyse numerical data in an Excel spreadsheet.
- It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA).

Insights:

- A. **Cleaning the data:** This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Soln:

Select and treat all blank cells

1. Selecting all data
2. go to special
3. click blanks and click ok (all the blank cells in data cell will be selected at the same time)
4. Write null in active cell
5. Hit ctrl+ enter
6. Now all empty cell will be field by null

director	actor	genre	title	vote	cast	...
James Cameron	Orlando Bloom	Action	Avatar	8.8	Wes Studi	...
Gore Verbinski	Johnny Depp	Action	Pirates of the Caribbean	7.1	Jack Davenport	...
Sam Mendes	Rory Kinnear	Action	Spectre	6.8	Stephanie	...
Christophe	Christian B	Action	The Dark Knight	9.0	Joseph Go	...
...

Removing duplicate rows

1. Select all
2. Go to data
3. Remove duplicates
4. Select my data has headers
5. Click ok

<

1. Select all
2. Add border
3. Proper spacing to view database properly
4. Making header text bold

IMDB_Movies.csv - Excel (Product Activation Failed)

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...

Calibri 11 A A

Font

Alignment

Number

Conditional Formatting

Normal Bad Good Neutral Calculation Check Call

Styles

Insert Delete Format

AutoSum Fill Sort & Find & Filter - Select - Editing

Clipboard

KB

J.K. Simmons

	K	L	M	N	O	P	Q
	actor_1_name	movie_title	num_voted_users	cast_total_facebook_likes	actor_3_name	facenumber_in_poster	plot_keywords
1	CCOY Pounder	Avatar	886324	4834	Wes Studi	0	pastor [future] marine [native] [paraplegic]
2	Johnny Depp	Pirates of the Caribbean: At World's End	471220	48350	Jack Davenport	0	goddess [marriage ceremony [marriage proposal] pirate] singapore
3	Christopher Waltz	Spectre	275868	11700	Stephanie Sigman	1	bomb [espionage [sequel] spy] terrorist
4	Tom Hardy	The Dark Knight Rises	1144337	106759	Joseph Gordon-Levitt	0	deception [imprisonment] lawlessness [police officer] terrorist plot
5	Doug Walker	Star Wars: Episode VII - The Force Awakens	8	143	NULL	0	NULL
6	Daryl Sabara	John Carter	212204	1873	Polly Walker	1	alien [american civil war [male nipple] [mons] princess
7	J.K. Simmons	Spider-Man 3	383056	46055	Kirsten Dunst	0	sandman [spider man [symbolic] venom] villain
8	Brad Garrett	Tangled	294810	2036	M.C. Gainey	1	17th century [based on fairy tale] Disney [flower] tower
9	Chris Hemsworth	Avengers: Age of Ultron	462669	92000	Scarlett Johansson	4	artificial intelligence [based on comic book] captain america [marvel cinematic universe]
10	Alan Rickman	Harry Potter and the Half-Blood Prince	321795	58753	Rupert Grint	3	blood [book [love] potion] professor
11	Henry Cavill	Batman v Superman: Dawn of Justice	376339	24450	Alan D. Purwin	0	based on comic book [batman] sequel to a reboot [superhero] superhero
12	Kevin Spacey	Superman Returns	240396	29991	Frank Langella	0	crystal [epic lex luthor [lois lane] return to earth
13	Giuseppe Giannini	Quantum of Solace	330784	2023	Rory Kinnear	1	action hero attempted rape [bond girl] [official james bond series] revenge
14	Johnny Depp	Pirates of the Caribbean: Dead Man's Chest	420240	44861	Jack Davenport	2	box office hit [giant squid] [harm] [dice] [monster]
15	Johnny Depp	The Lone Ranger	181792	45757	Tom Wilkinson	1	horse [outlaw] [texas] [texas ranger] train
16	Henry Cavill	Man of Steel	548573	20495	Henry Lennix	0	based on comic book [british actor playing american character] [final battle] origin
17	Peter Dinklage	The Chronicles of Narnia: Prince Caspian	149922	22697	Daniël An Alczjazz	4	brother brother relationship [brother sister relationship] good versus evil [king] [harm]
18	Chris Hemsworth	The Avengers	995415	87697	Scarlett Johansson	3	alien invasion [assassin] battle [iron man] soldier
19	Johnny Depp	Pirates of the Caribbean: On Stranger Tides	370704	54083	Stephen Graham	4	blackbeard [captain] [pirate] revenge [soldier]
20	Will Smith	Men in Black 3	268154	12572	Nicole Scherzinger	1	alien [criminal] [i.b.] [maximum security prison]
21	Aidan Turner	The Hobbit: The Battle of the Five Armies	354228	9152	James Nesbitt	0	army [elf] [hobbit] [middle earth] [orc]
22	Emma Stone	The Amazing Spider-Man	451803	28489	Chris Zylka	0	lizard [outcast] [spider] [spider man] [teenager]
23	Mark Addy	Robin Hood	211765	3244	Scott Grimes	0	1190s [archer] [england] [king of england] [robin hood]
24	Aidan Turner	The Hobbit: The Desolation of Smaug	483540	9152	James Nesbitt	6	dwarf [elf] [lake town] [mountain] [sword] [robbery and sorcery]
25	Christopher Lee	The Golden Compass	149019	24106	Kristin Scott Thomas	2	children [elf] [friend] [girl] quest
26	Nasim Watts	Kong	316018	2123	Evan Parker	0	animal movie in title [ape] about a woman [gorilla] [island] [king kong]
27	Leonardo DiCaprio	Inception	793059	45223	Gloria Stuart	2	artist [love ship] [trick] [wells]
28	Robert Downey Jr.	Captain America: Civil War	272670	64798	Chris Evans	0	based on comic book [knife] [marvel cinematic universe] [returning character killed c

- B. **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.

Soln:

- ✚ Movie name: Avatar
 - Budget: 237000000
 - Gross: 760505847
 - Profit: 523505847
-

- C. **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 is 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!

Soln:

TOP 250 any language:

num_voted_users: filter >25000

Imdb_score: sort Largest to smallest, filter top 250

Rank: ROW() – 1

rank	movie_title	imdb_score	num_voted_users
1	The Shawshank Redemption	9.3	1689764
2	The Godfather	9.2	1155770
3	The Dark Knight	9	1676169
4	The Godfather: Part II	9	790926
5	Fargo	9	170055
6	The Lord of the Rings: The Return of the King	8.9	1215718
7	Pulp Fiction	8.9	1324680
8	Schindler's List	8.9	865020
9	The Good, the Bad and the Ugly	8.9	503509
10	12 Angry Men	8.9	447785
11	Forrest Gump	8.8	1251222
12	Star Wars: Episode V - The Empire Strikes Back	8.8	837759
13	The Lord of the Rings: The Fellowship of the Ring	8.8	1238746
14	Inception	8.8	1468200

15	DaredevilÂ	8.8	213483
16	It's Always Sunny in PhiladelphiaÂ	8.8	133415
17	Fight ClubÂ	8.8	1347461
18	Star Wars: Episode IV - A New HopeÂ	8.7	911097
19	The Lord of the Rings: The Two TowersÂ	8.7	1100446
20	The MatrixÂ	8.7	1217752
21	One Flew Over the Cuckoo's NestÂ	8.7	680041
22	GoodfellasÂ	8.7	728685
23	City of GodÂ	8.7	533200
24	Friday Night LightsÂ	8.7	42746
25	Seven SamuraiÂ	8.7	229012
26	Saving Private RyanÂ	8.6	881236
27	The Silence of the LambsÂ	8.6	887467
28	Se7enÂ	8.6	1023511

29	InterstellarÂ	8.6	928227
30	The Usual SuspectsÂ	8.6	740918
31	HannibalÂ	8.6	159910
32	LutherÂ	8.6	70568
33	Spartacus: War of the DamnedÂ	8.6	173172
34	Once Upon a Time in the WestÂ	8.6	216005
35	It's a Wonderful LifeÂ	8.6	275720
36	CasablancaÂ	8.6	387508
37	American History XÂ	8.6	782437
38	Modern TimesÂ	8.6	143086
39	Spirited AwayÂ	8.6	417971
40	The Lion KingÂ	8.5	644348
41	Raiders of the Lost ArkÂ	8.5	661017
42	The Dark Knight RisesÂ	8.5	1144337
43	Back to the FutureÂ	8.5	732212
44	Terminator 2: Judgment DayÂ	8.5	744891
45	GladiatorÂ	8.5	982637
46	The Green MileÂ	8.5	782610
47	AlienÂ	8.5	563827
48	Django UnchainedÂ	8.5	955174
49	Apocalypse NowÂ	8.5	450676
50	The DepartedÂ	8.5	873649
51	PsychoÂ	8.5	422432
52	MementoÂ	8.5	845580
53	The PrestigeÂ	8.5	844052
54	WhiplashÂ	8.5	399138
55	The Lives of OthersÂ	8.5	259379
56	Children of HeavenÂ	8.5	27882
57	OutlanderÂ	8.5	50391

57	Outlander	8.5	50391
58	Entourage	8.5	135643
59	Airlift	8.5	30977
60	Dr. Strangelove or: How I Learned to Stop Wo	8.5	342585
61	The Pianist	8.5	497946
62	Star Wars: Episode VI - Return of the Jedi	8.4	681857
63	American Beauty	8.4	822500
64	Aliens	8.4	488537
65	WALL-E	8.4	718837
66	A Separation	8.4	151812
67	Braveheart	8.4	736638
68	Reservoir Dogs	8.4	664719
69	Stargate SG-1	8.4	63982
70	The Shining	8.4	610333
71	Veronica Mars	8.4	55524
72	The Inbetweeners	8.4	55987
73	Rang De Basanti	8.4	70233
74	Psych	8.4	67315
75	M*A*S*H	8.4	36265
76	Batman: The Dark Knight Returns, Part 2	8.4	30434
77	To Kill a Mockingbird	8.4	215086
78	Oldboy	8.4	356181
79	Requiem for a Dream	8.4	573541
80	Das Boot	8.4	168203
81	Lawrence of Arabia	8.4	192775
82	Baahubali: The Beginning	8.4	62756
83	Once Upon a Time in America	8.4	221000
84	Amélie	8.4	534262

85	Princess Mononoke	8.4	221552
86	Toy Story 3	8.3	544884
87	Inside Out	8.3	345198
88	Toy Story	8.3	623757
89	The Sting	8.3	175607
90	Indiana Jones and the Last Crusade	8.3	515306
91	Good Will Hunting	8.3	604904
92	Up	8.3	665575
93	Unforgiven	8.3	277505
94	Batman Begins	8.3	980946
95	Inglourious Basterds	8.3	885175
96	2001: A Space Odyssey	8.3	427357
97	Amadeus	8.3	270790
98	L.A. Confidential	8.3	414219
99	Snatch	8.3	600996
100	Some Like It Hot	8.3	175196
101	Scarface	8.3	537442
102	Eternal Sunshine of the Spotless Mind	8.3	666937
103	Room	8.3	161288
104	Monty Python and the Holy Grail	8.3	382240

105	Life	8.3	29450
106	The Great Escape	8.3	165638
107	The Apartment	8.3	109335
108	Judgment at Nuremberg	8.3	44457
109	Singin' in the Rain	8.3	150020
110	Inside Job	8.3	55382
111	Taxi Driver	8.3	507063

112	The Hunt	8.3	170155
113	Metropolis	8.3	111841
114	Downfall	8.3	248354
115	Raging Bull	8.3	235133
116	Finding Nemo	8.2	692482
117	Gone with the Wind	8.2	215340
118	Captain America: Civil War	8.2	272670
119	Gran Torino	8.2	561773
120	A Beautiful Mind	8.2	610568
121	Die Hard	8.2	592582
122	How to Train Your Dragon	8.2	485430
123	The Bridge on the River Kwai	8.2	149444
124	Pan's Labyrinth	8.2	467234
125	The Secret in Their Eyes	8.2	131831
126	The Wolf of Wall Street	8.2	780588
127	V for Vendetta	8.2	791783
128	Trainspotting	8.2	469561
129	On the Waterfront	8.2	100890
130	Into the Wild	8.2	426359
131	Lock, Stock and Two Smoking Barrels	8.2	414976
132	The Big Lebowski	8.2	537419
133	Incendies	8.2	80429
134	The Deer Hunter	8.2	232577
135	Buffy the Vampire Slayer	8.2	101902
136	The Elephant Man	8.2	161972
137	Lage Raho Munna Bhai	8.2	27569
138	Mr. Smith Goes to Washington	8.2	77392

139	Rebecca	8.2	87424
140	It Happened One Night	8.2	64888
141	Blade Runner	8.2	461609
142	The Thing	8.2	258078
143	Casino	8.2	333542
144	Warrior	8.2	332276
145	Howl's Moving Castle	8.2	214091
146	The Avengers	8.1	995415
147	Deadpool	8.1	479047
148	Jurassic Park	8.1	613473
149	The Sixth Sense	8.1	704766
150	Monsters, Inc.	8.1	585659
151	Pirates of the Caribbean: The Curse of the Black Pearl	8.1	809474

151	Pirates of the Caribbean: The Curse of the Black Pearl	8.1	809474
152	Guardians of the Galaxy	8.1	682155
153	The Help	8.1	318955
154	Platoon	8.1	291603
155	The Martian	8.1	472488
156	The Bourne Ultimatum	8.1	491077
157	Rocky	8.1	375240
158	Gone Girl	8.1	569841
159	Butch Cassidy and the Sundance Kid	8.1	152089
160	The Imitation Game	8.1	467613
161	Million Dollar Baby	8.1	482064
162	The Truman Show	8.1	667983
163	Groundhog Day	8.1	437418
164	No Country for Old Men	8.1	612060
165	The Revenant	8.1	406020

166	Shutter Island	8.1	786092
167	Stand by Me	8.1	271794
168	Kill Bill: Vol. 1	8.1	735784
169	12 Years a Slave	8.1	439176
170	Annie Hall	8.1	192940
171	Sin City	8.1	656640
172	The Grand Budapest Hotel	8.1	475518
173	The Terminator	8.1	600266
174	Spotlight	8.1	195333
175	The Best Years of Our Lives	8.1	40359
176	The Wizard of Oz	8.1	291875
177	There Will Be Blood	8.1	372990
178	Prisoners	8.1	383591
179	The Princess Bride	8.1	294163
180	Hotel Rwanda	8.1	264533
181	Mad Max: Fury Road	8.1	552503
182	Amores Perros	8.1	173551
183	Before Sunrise	8.1	183288
184	The Celebration	8.1	65951
185	Solaris	8.1	54057
186	Gandhi	8.1	171726
187	Hachi: A Dog's Tale	8.1	155249
188	Barry Lyndon	8.1	101627
189	Network	8.1	103493
190	A Christmas Story	8.1	104908
191	The Man Who Shot Liberty Valance	8.1	53741
192	Cat on a Hot Tin Roof	8.1	33741

193	Touching the Void	8.1	26926
194	High Noon	8.1	80193
195	Donnie Darko	8.1	580999
196	Elite Squad	8.1	81644
197	The Sea Inside	8.1	64556
198	Rush	8.1	312629
199	Tae Guk Gi: The Brotherhood of War	8.1	31943
200	Akira	8.1	106160
201	Jaws	8	412454
202	The Exorcist	8	284252
203	Aladdin	8	260939
204	The Incredibles	8	479166
205	Dances with Wolves	8	186485
206	The Sound of Music	8	148172
207	Rain Man	8	383784
208	Slumdog Millionaire	8	641997
209	The King's Speech	8	503631
210	Catch Me If You Can	8	525801
211	Star Trek	8	504419
212	The Pursuit of Happyness	8	338383
213	Doctor Zhivago	8	55816
214	Black Swan	8	551363
215	District 9	8	531737
216	Young Frankenstein	8	112671
217	Dead Poets Society	8	277451
218	Mystic River	8	338415
219	Ratatouille	8	473887

220	Fiddler on the Roof	8	29839
221	Kill Bill: Vol. 2	8	512749
222	X-Men: Days of Future Past	8	514125
223	JFK	8	113472
224	The Artist	8	190030
225	Sling Blade	8	72443
226	Dallas Buyers Club	8	326494
227	Boyhood	8	266020
228	Bowling for Columbine	8	123090
229	Casino Royale	8	470483
230	Casino Royale	8	470501
231	Sicko	8	66610
232	Shaun of the Dead	8	395921
233	Life of Pi	8	440084
234	The Perks of Being a Wallflower	8	351274
235	A Fistful of Dollars	8	147566
236	Before Sunset	8	168398

237	Central Station	8	28951
238	Her	8	355126
239	Waltz with Bashir	8	46107
240	The Return	8	31589
241	The Diving Bell and the Butterfly	8	89906
242	Patton	8	76398
243	The Wild Bunch	8	63192
244	Night of the Living Dead	8	87978
245	Rosemary's Baby	8	140527
246	Days of Heaven	8	37594

247	The Hustler	8	62860
248	A Streetcar Named Desire	8	78454
249	The Man from Earth	8	129799
250	True Romance	8	163492

Top_Foreign_Lang_Film:

rank	Top_Foreign_Lang_Film	imdb_score	num_voted_users	language
1	The Good, the Bad and the Ugly	8.9	503509	Italian
2	City of God	8.7	533200	Portuguese
3	Seven Samurai	8.7	229012	Japanese
4	Spirited Away	8.6	417971	Japanese
5	The Lives of Others	8.5	259379	German
6	Children of Heaven	8.5	27882	Persian
7	Airlift	8.5	30977	Hindi
8	A Separation	8.4	151812	Persian
9	Rang De Basanti	8.4	70233	Hindi
10	Oldboy	8.4	356181	Korean
11	Das Boot	8.4	168203	German
13	Baahubali: The Beginning	8.4	62756	Telugu
14	Amélie	8.4	534262	French
15	Princess Mononoke	8.4	221552	Japanese
16	The Hunt	8.3	170155	Danish
17	Metropolis	8.3	111841	German
18	Downfall	8.3	248354	German
19	Pan's Labyrinth	8.2	467234	Spanish
20	The Secret in Their Eyes	8.2	131831	Spanish
21	Incendies	8.2	80429	French
22	Lage Raho Munna Bhai	8.2	27569	Hindi
23	Howl's Moving Castle	8.2	214091	Japanese
24	Amores Perros	8.1	173551	Spanish
25	The Celebration	8.1	65951	Danish
26	Solaris	8.1	54057	Russian
27	Elite Squad	8.1	81644	Portuguese
28	The Sea Inside	8.1	64556	Spanish
29	Tae Guk Gi: The Brotherhood of War	8.1	31943	Korean

30	Akira	8.1	106160	Japanese
31	A Fistful of Dollars	8	147566	Italian
32	Central Station	8	28951	Portuguese
33	Waltz with Bashir	8	46107	Hebrew
34	The Return	8	31589	Russian
35	The Diving Bell and the Butterfly	8	89906	French

D. **Best Directors:** 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.

Soln:

Query:

```
use ibdm;

select director_name as top10director, avg(imdb_score) as imdb_score
from ibdm.imdb_movies
group by top10director
order by imdb_score desc
limit 10;
```

Result:

top10director	imdb_score
Christopher Nolan	8.414285714285715
S.S. Rajamouli	8.4
Lee Unkrich	8.3
Pete Docter	8.233333333333334
Hideaki Anno	8.2
Quentin Tarantino	8.16
Denis Villeneuve	8.1
Tim Miller	8.1
Andrei Tarkovsky	8.1
Alejandro G. Iñárritu	8.1

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

Soln:

Query:

```
select distinct genres, count(genres) as genresCount
from ibdm.imdb_movies
group by genres
order by genresCount desc
limit 10;
```

Result:

genres	genresCount
Action Adventure Sci-Fi	44
Comedy Romance	39
Comedy	35
Action Crime Thriller	31
Comedy Drama Romance	31
Action Adventure Thriller	29
Crime Drama Thriller	27
Action Adventure Sci-Fi Thriller	26
Adventure Animation Comedy Family Fantasy	24
Action Adventure Fantasy	22

-
- F. **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.

Query 1: Critic favourite

```
select actor_1_name critic_favorite, avg(num_critic_for_reviews) as criticReviewCount
from ibdm.imdb_movies
group by critic_favorite
order by criticReviewCount desc
limit 1;
```

Output:

critic_favorite	criticReviewCount
Albert Finney	750

Query 2: Audience Favourite

```
select actor_1_name audience_favorite, avg(num_user_for_reviews) as UserReviewCount
from ibdm.imdb_movies
group by audience_favorite
order by UserReviewCount desc
limit 1;
```

Output:

	audience_favorite	UserReviewCount
▶	Steve Bastoni	2789

Result:

- In this project I have gain practical hands on knowledge to analyse numerical data in an Excel spreadsheet.

- Learnt various MS Excel functions and formulas that can be used in many companies in day to day analysis data record in their company