<u>Project 5 : By Ramana Bansal</u>

<u>IMDB Movies Analysis</u>

Excel Worksheet Link:

https://docs.google.com/spreadsheets/d/1A5qgkcDipouQvBPcIa8JAuFOOQ9nANrq

Project Description: The project deals with IMDB movie data from year 1927 to 2016. The data is being used to answer the question:

What movies or the kind of movies does a Youtube channel for movie discussions and analysis need to focus on which will generate the most traffic, and hence revenue.

Approach: The data was processed and analyzed using Microsoft Excel.

The following pointers were used to find popular movies:

1. Movies with highest profit.
2. English movies with highest IMDB scores till now
3. Foreign language movies with highest IMDB scores till now.
4. Best directors and their movies.
5. Popular genre: Based on IMDB as well as Gross.
6. Most favoured actors: By critics and by audience.
7. The ratings for which decade could be most relied on based on number of votes.

Some other questions that could have been used fro the same data were:

1. Movies which were outliers, with huge budgets, huge profits and huge losses.
2. Popular actors based on facebook likes.
3. Popular movies based on facebook likes.
4. Popular directors based on facebook likes.

Tech-Stack Used: Microsoft Excel 2010. The online version of MS Excel 365 was also used for a while due to device issues.

Insights: The analysis aided in identifying a number of popular movies that can be used to create further content. It also helped in identifying the popular directors, actors, genres etc. This may help to make data-driven decision regarding the selection of current or upcoming movies to be focused on.

Result: The project made me feel a little more confident with Microsoft Excel and its use in data cleaning and statistical analysis. It also helped me to understand and apply various tools and formulae, especially filtering and index-match functions. It also gave me a little confidence boost regarding working with bigger data sets.

A. Cleaning the data

The data was formatted into a table for ease of handling.

1. Removing irrelevant columns/data

Various columns that were not required in the analysis such as colors, duration, facenumber_in_poster, plot_keywords, cast_total_likes, actor_3-name, actor_3_facebook_likes, movie_imdb_link, content_rating and aspect_ratio etc. were deleted. Eventually only 14 columns remained.

2. Handling Typos and Inaccurate Values

The following names in directors_name column had special characters, which needed to be removed except ~A. Using Control+F,

‘Ã©’ was replaced with ‘e’. ‘Ã±’ was replaced with ‘n’. ‘Ã³’ replaced with ‘o’. ‘Ã¥’ replaced with ‘a’. ‘Ã¶’ replaced with o. ‘Ã¡’ replaced with ‘a’. ‘Ã§’ was replaced with ‘c’. ‘Ã”’ replaced with ‘O’. ‘Ã»’ replaced with ‘u’. ‘Ã-’ replaced with ‘i’. ‘Ã¤’ was replaced with ‘a’. ‘Ã‰’ replaced with ‘E’. ‘Ã…’ replaced with ‘A’. ‘Ã¨’ replaced with ‘e’. ‘Ã˜’ replaced with ‘O’. ‘Ã¦’ replaced with ‘ae’.

Roland JoffÃ©	FrÃ©dÃ©ric Auburtin	GÃ©rard Krawczyk
JosÃ© Padilha	Lasse HallstrÃ¶m	AndrÃ©s Muschietti
FrÃ©dÃ©ric Forestier	Lasse HallstrÃ¶m	Jean-Marc VallÃ©e
AndrÃ©s Couturier	Lasse HallstrÃ¶m	RyÃ»hei Kitamura
MÃ¥ns MÃ¥rlind	Lasse HallstrÃ¶m	Gabe IbÃ¡Ã±ez
Mikael HÃ¥fstrÃ¶m	JÃ©rÃ´me Deschamps	JosÃ© Padilha
Jorge R. GutiÃ©rrez	Mikael HÃ¥fstrÃ¶m	StÃ©phane Aubier
Roland JoffÃ©	Lasse HallstrÃ¶m	Lasse HallstrÃ¶m
Mark A.Z. DippÃ©	Mikael HÃ¥fstrÃ¶m	Jaume BalaguerÃ³
NimrÃ³d Antal	Juan JosÃ© Campanella	LÃ©a Pool
Jean-Marie PoirÃ©	NimrÃ³d Antal	Timothy BjÃ¶rklund
Mikael HÃ¥fstrÃ¶m	Katsuhiro Ã”tomo	FranÃ§ois Girard
Roland JoffÃ©	Jean-Marie PoirÃ©	AndrÃ© TÃ©chinÃ©
Lasse HallstrÃ¶m	IstvÃ¡n SzabÃ³	MaÃ¯wenn
Jean-Marc VallÃ©e	NimrÃ³d Antal	FranÃ§ois Ozon
Roland JoffÃ©	Rodrigo CortÃ©s	Marc SchÃ¶lermann
Joachim RÃ¸nning	Gaspar NoÃ©	Katsuhiro Ã”tomo
Jean-FranÃ§ois Richet	Lasse HallstrÃ¶m	FranÃ§ois Ozon
Lasse HallstrÃ¶m	JÃ©rÃ´me Salle	Rodrigo GarcÃa

Rodrigo GarcÃa
Jean-Marie PoirÃ©
Max FÃ¤rberbÃ¶ck
JosÃ© Padilha
Jaume BalaguerÃ³
Ã‰mile Gaudreault
Aki KaurismÃ¤ki
Gaspar NoÃ©
Jean-Marc VallÃ©e
Gonzalo LÃ³pez-Gallego
Jaume BalaguerÃ³
JirÃ Menzel
Jonas Ã…kerlund
RenÃ© FÃ©ret
Jorge RamÃrez SuÃ¡rez
Fernando LeÃ³n de Aranoa
Karim AÃ¯nouz
Juan JosÃ© Campanella
AndrÃ© Ã˜vredal
Nnegest LikkÃ©
ThorbjÃ¸rn Christoffersen
Ã‰ric Tessier
Jaume BalaguerÃ³
Jonas Ã…kerlund
Rodrigo CortÃ©s
LluÃs QuÃlez
FranÃ§ois Truffaut
FabiÃ¡n Bielinsky
LÃ©a Pool
JosÃ© Luis Valenzuela
Ã‰tienne Faure
EugÃ¨ne LouriÃ©

3. **Handling Duplicate Values**

This step involves identifying and removing duplicate values in the data. Duplicate values can cause errors and distort the analysis results. 112 duplicates were removed. Only major data was used to find duplicates.

4. **Checking for missing data**

There are total 4998 rows after duplicate removal. There were 874 blanks in gross column, which had to be removed as they could skew our data analysis.

After this, 4124 rows were left. Similarly, 267 blanks from the budget column were also removed. 3 blanks as per language column were also removed.

We are left with 3854 rows to perform our analysis. We make sure that certain necessary columns such as director_name, gross, budget, movie_title, genre and imdb_score do not have blanks.

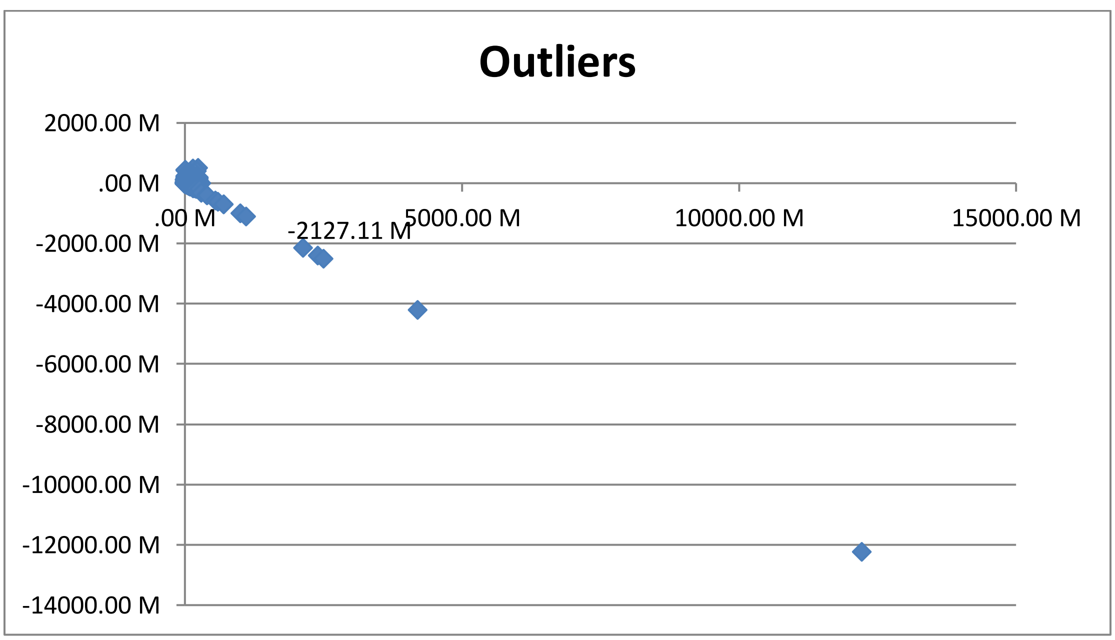
B. Movies with highest profit:

Profit was calculated as a difference of gross and budget. Profit was then further formatted into Number (Millions) using (#,,.00 “M”). The data was sorted based on profit column and a rank was assigned to each based on profit using Rank().

The top 10 movies with highest profit were filtered out.

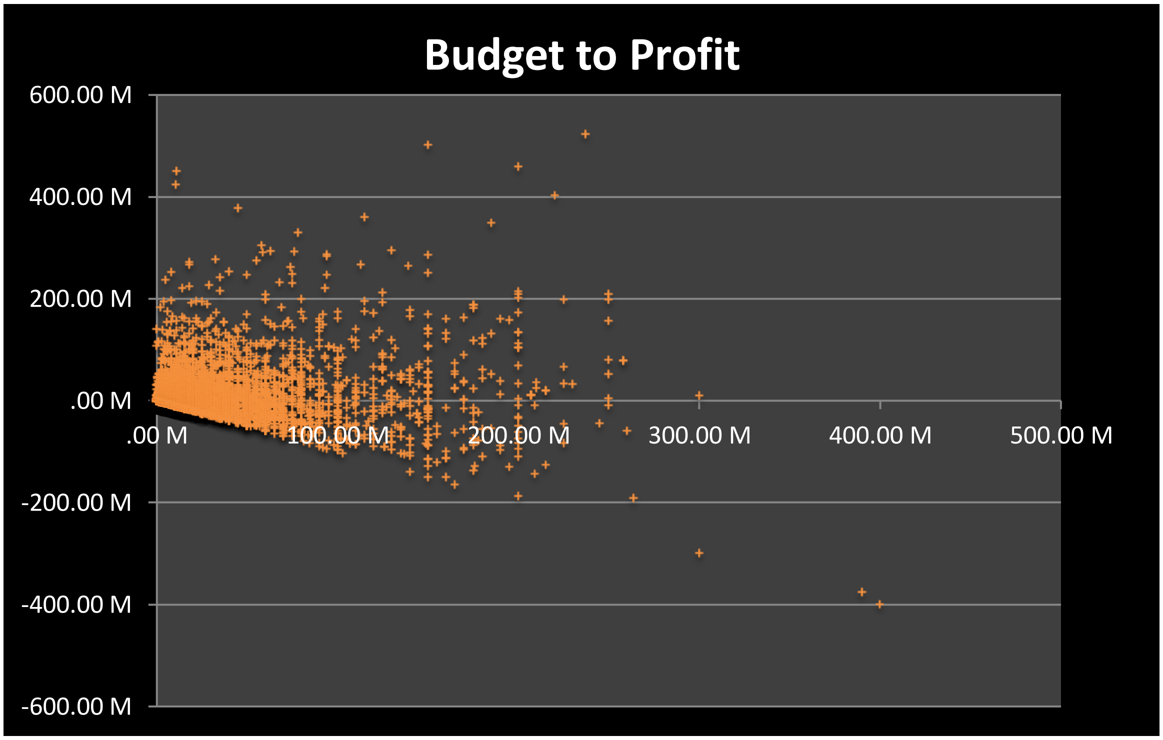
Movies with Highest Profit	Profit
Avatar	523.51 M
Pirates of the Caribbean: At World's End	502.18 M
Spectre	458.67 M
The Dark Knight Rises	449.94 M
John Carter	424.45 M
Spider-Man 3	403.28 M
Tangled	377.78 M
Avengers: Age of Ultron	359.54 M
Harry Potter and the Half-Blood Prince	348.32 M
Batman v Superman: Dawn of Justice	330.00 M

A graph was plotted to observe outliers. Since box-plot is not available in Microsoft Excel 2010, a scatter plot was used.



The movies with budget more than 500M or loss more than 600M were considered outliers and removed before further analyses.

The following graph between budget and profit gives an idea regarding how most movies fare at market.



C. Top 250:

Using columns movie_name, imdb_score, num_voted_users and language, the data was formatted into a table and filters were added. Using filters, the data was filtered based on num_voted_users greater than 25000. Further, the data was sorted as per num_voted_users. Then, the data was sorted as per imdb_score. A new column rank was added and the following formula was used to give unique ranks. Preference was given to higher num_voted_users.

=RANK(C2, $C$2:$C$251,0)+COUNTIF($C$2:C2,C2)-1

Column C contained imdb scores.

The list was truncated at rank 250.

The Top 250 movies are:

1	The Shawshank Redemption
2	The Godfather
3	The Dark Knight
4	The Godfather: Part II
5	Pulp Fiction
6	The Lord of the Rings: The Return of the King
7	Schindler's List
8	The Good, the Bad and the Ugly
9	Inception
10	Fight Club
11	Forrest Gump
12	The Lord of the Rings: The Fellowship of the
13	Star Wars: Episode V - The Empire Strikes Back
14	The Matrix
15	The Lord of the Rings: The Two Towers
16	Star Wars: Episode IV - A New Hope
17	Goodfellas
18	One Flew Over the Cuckoo's Nest
19	City of God
20	Seven Samurai
21	Se7en
22	Interstellar
23	The Silence of the Lambs
24	Saving Private Ryan
25	American History X
26	The Usual Suspects
27	Spirited Away
28	Modern Times
29	The Dark Knight Rises
30	Gladiator
31	Django Unchained
32	The Departed
33	Memento
34	The Prestige
35	The Green Mile
36	Terminator 2: Judgment Day
37	Back to the Future
38	Raiders of the Lost Ark
39	The Lion King
40	Alien
41	The Pianist
42	Apocalypse Now
43	Psycho
44	Whiplash
45	The Lives of Others
46	Children of Heaven
48	American Beauty
49	Braveheart
50	WALL·E
51	Star Wars: Episode VI - Return of the Jedi
52	Reservoir Dogs
53	Requiem for a Dream
54	Amelie
55	Aliens
56	Oldboy
57	Princess Mononoke
58	Once Upon a Time in America
59	Lawrence of Arabia
60	Das Boot
61	A Separation
62	Baahubali: The Beginning
64	Batman Begins
65	Inglourious Basterds
66	Eternal Sunshine of the Spotless Mind
67	Up
68	Toy Story
69	Good Will Hunting
70	Snatch
71	Toy Story 3
72	Scarface
73	Indiana Jones and the Last Crusade
74	2001: A Space Odyssey
75	L.A. Confidential
76	Monty Python and the Holy Grail
77	Inside Out
78	Unforgiven

79	Amadeus	125	Donnie Darko
80	Downfall	126	Gone Girl
81	Raging Bull	127	Mad Max: Fury Road
82	The Sting	128	The Bourne Ultimatum
83	Some Like It Hot	129	Million Dollar Baby
84	The Hunt	130	Deadpool
85	Room	131	The Grand Budapest Hotel
86	Metropolis	132	The Martian
89	V for Vendetta	133	The Imitation Game
90	The Wolf of Wall Street	134	12 Years a Slave
91	Finding Nemo	135	Groundhog Day
92	A Beautiful Mind	136	The Revenant
93	Die Hard	137	Prisoners
94	Gran Torino	138	Rocky
95	The Big Lebowski	139	There Will Be Blood
96	How to Train Your Dragon	140	The Help
97	Trainspotting	141	Rush
98	Pan's Labyrinth	142	The Princess Bride
99	Blade Runner	143	The Wizard of Oz
100	Into the Wild	144	Platoon
101	Lock, Stock and Two Smoking Barrels	145	Stand by Me
102	Casino	146	Hotel Rwanda
103	Warrior	147	Spotlight
104	Captain America: Civil War	148	Annie Hall
105	The Thing	149	Before Sunrise
106	Gone with the Wind	150	Amores Perros
107	Howl's Moving Castle	151	Butch Cassidy and the Sundance Kid
108	The Bridge on the River Kwai	152	Akira
109	The Secret in Their Eyes	153	Elite Squad
110	On the Waterfront	154	The Celebration
111	Incendies	155	The Sea Inside
113	The Avengers	156	The Best Years of Our Lives
114	Pirates of the Caribbean: The Curse of the Black Pearl	157	Tae Guk Gi: The Brotherhood of War
115	Shutter Island	161	Slumdog Millionaire
116	Kill Bill: Vol. 1	162	Black Swan
117	The Sixth Sense	163	District 9
118	Guardians of the Galaxy	164	Catch Me If You Can
119	The Truman Show	165	X-Men: Days of Future Past
120	Sin City	166	Kill Bill: Vol. 2
121	Jurassic Park	167	Star Trek
122	No Country for Old Men	168	The King's Speech
123	The Terminator	169	The Incredibles
124	Monsters, Inc.	170	Ratatouille

171	Casino Royale
172	Life of Pi
173	Jaws
174	Blood Diamond
175	Shaun of the Dead
176	Rain Man
177	Her
178	The Perks of Being a Wallflower
179	Big Fish
180	Mystic River
181	The Pursuit of Happyness
182	Dallas Buyers Club
183	In Bruges
184	The Exorcist
185	Dead Poets Society
186	Boyhood
187	Aladdin
188	Serenity
189	Magnolia
190	Mulholland Drive
191	The Artist
192	Dances with Wolves
193	Before Sunset
194	True Romance
195	Brazil
196	Cinderella Man
197	The Sound of Music
198	A Fistful of Dollars
199	The Iron Giant
200	Bowling for Columbine
201	JFK
202	Young Frankenstein
203	Dancer in the Dark
204	Sling Blade
205	Persepolis
206	My Name Is Khan
207	Sicko
208	The Straight Story
209	Doctor Zhivago
210	Waltz with Bashir

211	Fiddler on the Roof
212	Central Station
215	Avatar
216	Iron Man
217	The Hobbit: An Unexpected Journey
218	Taken
219	The Hobbit: The Desolation of Smaug
220	Shrek
221	Edge of Tomorrow
222	The Bourne Identity
223	The Notebook
224	Toy Story 2
225	Children of Men
226	Crash
227	Edward Scissorhands
228	Little Miss Sunshine
229	Hot Fuzz
230	Captain Phillips
231	Nightcrawler
232	E.T. the Extra-Terrestrial
233	Big Hero 6
234	The Fighter
235	The Hateful Eight
236	Moon
237	The Wrestler
238	How to Train Your Dragon 2
239	The Untouchables
240	Crouching Tiger, Hidden Dragon
241	Almost Famous
242	Boogie Nights
243	Walk the Line
244	Halloween
245	Hero
246	The Blues Brothers
247	Ed Wood
248	The Insider
249	Letters from Iwo Jima
250	Straight Outta Compton

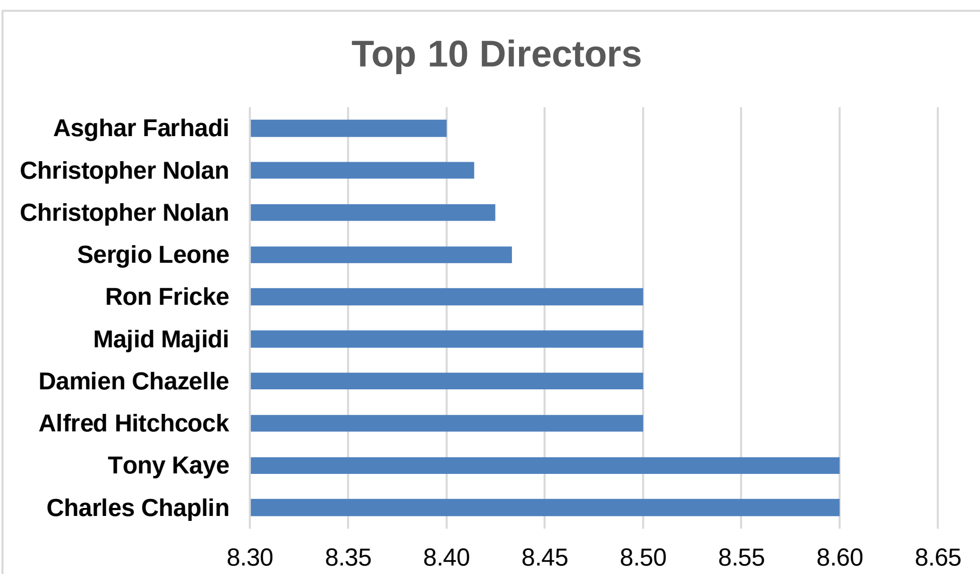
To extract the movies in the IMDb_Top_250 column which are not in the English language, again filter on language is used.

Rank	Top Foreign Language Movies	imdb_score	language
1	The Good, the Bad and the Ugly	8.9	Italian
2	City of God	8.7	Portuguese
3	Seven Samurai	8.7	Japanese
4	Spirited Away	8.6	Japanese
5	The Lives of Others	8.5	German
6	Children of Heaven	8.5	Persian
7	Amelie	8.4	French
8	Oldboy	8.4	Korean
9	Princess Mononoke	8.4	Japanese
10	Das Boot	8.4	German
11	A Separation	8.4	Persian
12	Baahubali: The Beginning	8.4	Telugu
13	Downfall	8.3	German
14	The Hunt	8.3	Danish
15	Metropolis	8.3	German
16	Pan's Labyrinth	8.2	Spanish
17	Howl's Moving Castle	8.2	Japanese
18	The Secret in Their Eyes	8.2	Spanish
19	Incendies	8.2	French
20	Amores Perros	8.1	Spanish
21	Akira	8.1	Japanese
22	Elite Squad	8.1	Portuguese
23	The Celebration	8.1	Danish
24	The Sea Inside	8.1	Spanish
25	Tae Guk Gi: The Brotherhood of War	8.1	Korean
26	A Fistful of Dollars	8	Italian
27	Persepolis	8	French
28	My Name Is Khan	8	Hindi
29	Waltz with Bashir	8	Hebrew
30	Central Station	8	Portuguese
31	Crouching Tiger, Hidden Dragon	7.9	Mandarin
32	Hero	7.9	Mandarin
33	Letters from Iwo Jima	7.9	Japanese

D. Best Directors:

Average imdb score for each director was computed and stored in a new column using formula averageif().
=AVERAGEIF(A2:A3791,A2,B2:B3791)
Here, Column A had director_names and coulumn B had imdb_scores.
The data was sorted on two levels.
First level: average_imdb_score of each director
Second level: Director name in alphabetical order in case of same mean_imdb_score.
The names of top 10 directors were extracted from the data and stored in a separate table.

Top 10 Directors	Mean_imdb_score
Charles Chaplin	8.60
Tony Kaye	8.60
Alfred Hitchcock	8.50
Damien Chazelle	8.50
Majid Majidi	8.50
Ron Fricke	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Christopher Nolan	8.41
Asghar Farhadi	8.40



E. Popular Genres:

The data from Top_Imdb_250 as well as Gross is used. The genre column is added from main worksheet using formula:

=INDEX(IMDB2[genres],MATCH(Genres!B2,IMDB2[movie],0))

Here, B2 column stored the name of movies which were used to match their genres.

After the genres were extracted, the genres for each movie were separated into columns using 'Text to Columns' and delimiter '|'. Then the count of each genre was obtained from the newly created columns using countif().

A new table with genres and their count was created, which was used to plot the graph.

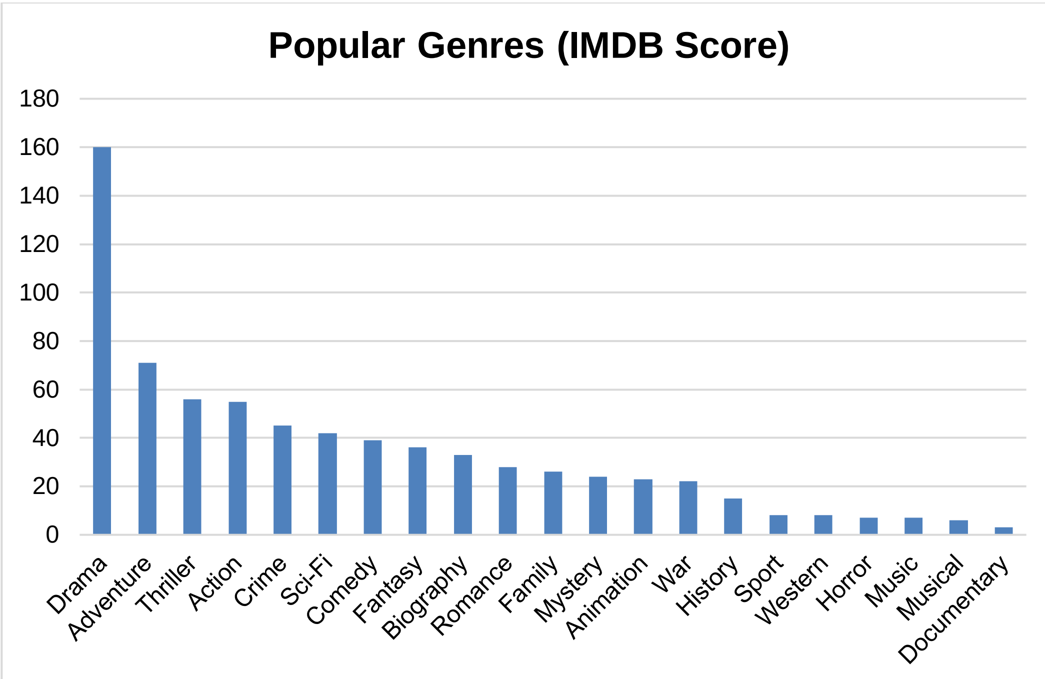
The process was done twice, once for Top_IMDB_250 based on IMDB_scores. Then for, 250 of the highest grossing movies.

Based on IMDB scores (Top 250):

Popular Genres based on IMDB scores

Genre	Count
Drama	160
Adventure	71
Thriller	56
Action	55
Crime	45
Sci-Fi	42
Comedy	39
Fantasy	36
Biography	33
Romance	28
Family	26

Mystery	24
Animation	23
War	22
History	15
Sport	8
Western	8
Horror	7
Music	7
Musical	6
Documentary	3

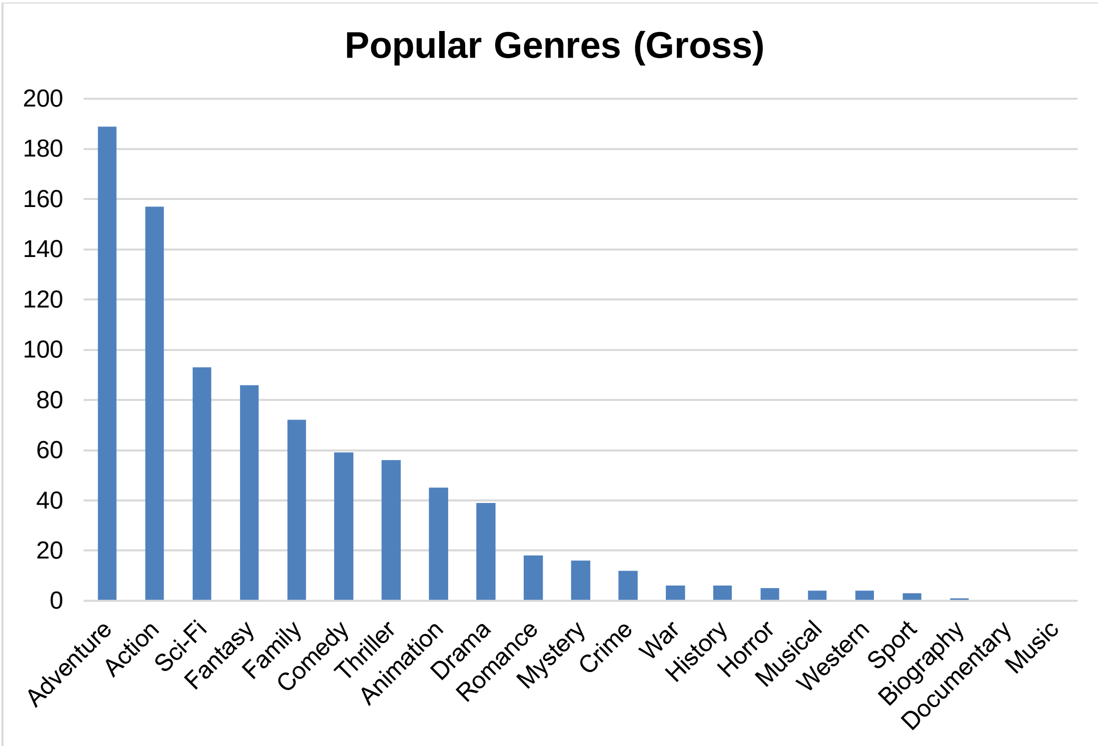


Based on Gross (Top 250)

Gross was used instead of profit since it would better indicate the user interest than profit, which can vary with budget.

Gross based popular genres

Genres	Count
Adventure	189
Action	157
Sci-Fi	93
Fantasy	86
Family	72
Comedy	59
Thriller	56
Animation	45
Drama	39
Romance	18
Mystery	16
Crime	12
War	6
History	6
Horror	5
Musical	4
Western	4
Sport	3
Biography	1
Documentary	0
Music	0


Popular Genres (Gross)
200
180
160
140
120
100
80
60
40
20
0
Adventure
Action
Sci-Fi
Fantasy
Family
Comedy
Thriller
Animation
Drama
Romance
Mystery
Crime
War
History
Horror
Musical
Western
Sport
Biography
Documentary
Music

F. Charts:

The movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors were filtered into three different columns using Filter tool on actor_1_name column.

Meryl_Streep	Leo_Caprio	Brad_Pitt
It's Complicated	Titanic	The Curious Case of Benjamin Button
The River Wild	The Great Gatsby	Troy
Julie & Julia	Inception	Ocean's Twelve
The Devil Wears Prada	The Revenant	Mr. & Mrs. Smith
Lions for Lambs	The Aviator	Spy Game
Out of Africa	Django Unchained	Ocean's Eleven
Hope Springs	Blood Diamond	Fury
One True Thing	The Wolf of Wall Street	Seven Years in Tibet
The Hours	Gangs of New York	Fight Club
The Iron Lady	The Departed	Sinbad: Legend of the Seven Seas
A Prairie Home Companion	Shutter Island	Interview with the Vampire: The Vampire Chronicles
	Body of Lies	The Tree of Life
	Catch Me If You Can	The Assassination of Jesse James by the Coward Robert Ford
	The Beach	Babel
	Revolutionary Road	By the Sea
	The Man in the Iron Mask	Killing Them Softly
	J. Edgar	True Romance
	The Quick and the Dead	
	Marvin's Room	
	Romeo + Juliet	

The 3 columns were then combined into a single column using vstack().

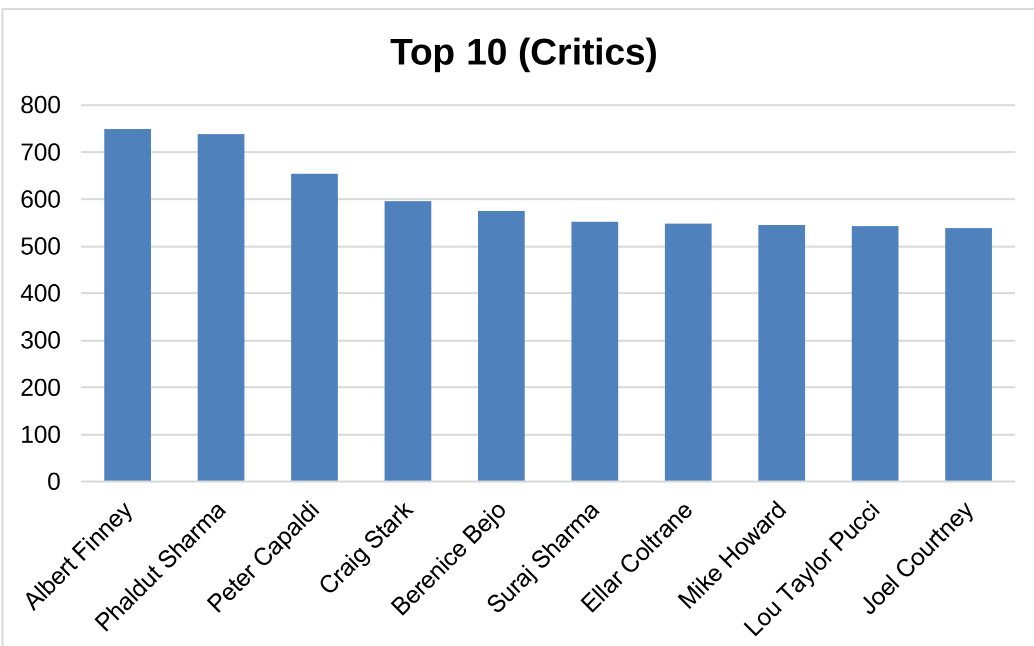
Combined
It's Complicated
The River Wild
Julie & Julia
The Devil Wears Prada
Lions for Lambs
Out of Africa
Hope Springs
One True Thing
The Hours
The Iron Lady
A Prairie Home Companion
Titanic
The Great Gatsby
Inception
The Revenant
The Aviator
Django Unchained

Blood Diamond	Mr. & Mrs. Smith
The Wolf of Wall Street	Spy Game
Gangs of New York	Ocean's Eleven
The Departed	Fury
Shutter Island	Seven Years in Tibet
Body of Lies	Fight Club
Catch Me If You Can	Sinbad: Legend of the Seven Seas
The Beach	Interview with the Vampire: The Vampire Chronicles
Revolutionary Road	The Tree of Life
The Man in the Iron Mask	The Assassination of Jesse James by the Coward Robert Ford
J. Edgar	Babel
The Quick and the Dead	By the Sea
Marvin's Room	Killing Them Softly
Romeo + Juliet	True Romance
The Curious Case of Benjamin Button	
Troy	
Ocean's Twelve	

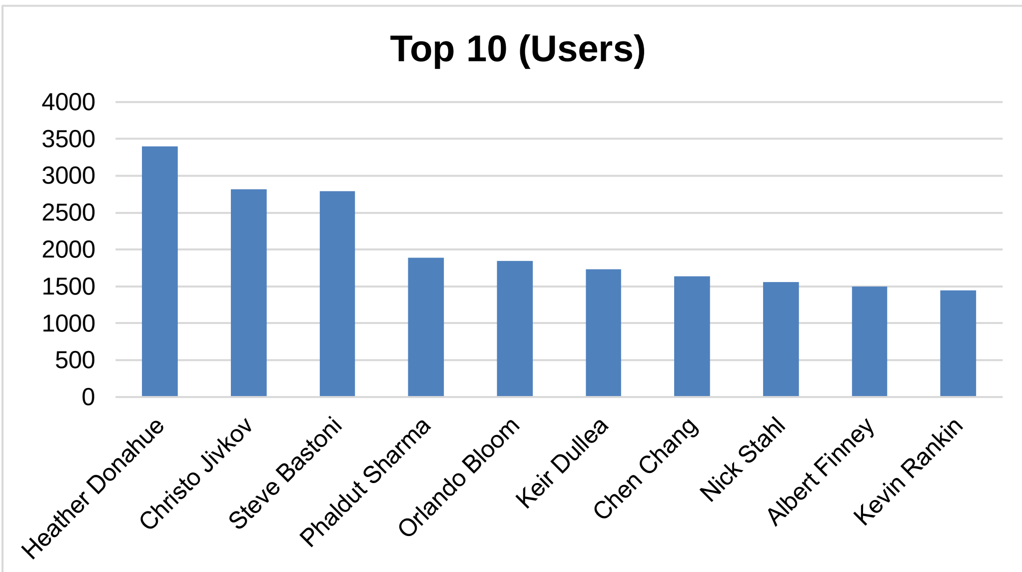
Actors which have the highest reviews.

The means of the num_critic_for_reviews as well as num_users_for_review were calculated separately to identify favourite actors of critics and users respectively. Pivot tables were used for the process.

Top 10 Actors (Critics)	Average of Critic Reviews
Albert Finney	750
Phaldut Sharma	738
Peter Capaldi	654
Craig Stark	596
Berenice Bejo	576
Suraj Sharma	552
Ellar Coltrane	548
Mike Howard	546
Lou Taylor Pucci	543
Joel Courtney	539



Top 10 Actors (Users)	Average of user reviews
Heather Donahue	3400
Christo Jivkov	2814
Steve Bastoni	2789
Phaldut Sharma	1885
Orlando Bloom	1842
Keir Dullea	1736
Chen Chang	1641
Nick Stahl	1562
Albert Finney	1498
Kevin Rankin	1445



Number of voters per decade

Pivot table was used to group the years into a decade and calculate the sum of num_voted_users for the years in a particular decade.

Voting per Decade

Years	Decade	Sum of num_voted_users
1921-1930	1920s	.12 M
1931-1940	1930s	.97 M
1941-1950	1940s	.07 M
1951-1960	1950s	1.10 M
1961-1970	1960s	2.61 M
1971-1980	1970s	9.90 M
1981-1990	1980s	21.50 M
1991-2000	1990s	78.61 M
2001-2010	2000s	172.75 M
2011-2020	2010s	96.72 M