IMDB Movie Analysis

- Priyanshu Kamal

- callmepriyanshu4@gmail.com

Project Description

The project is on the analysis of the dataset of the IMDB movies where I will study the data properly and ask various questions about it to draw attention to that problem.

After the problem is detected, I will do analysis and try to find the reason behind the problem and draw insights about it.

Approach

The approach of this project is to first clean the data by removing duplicates, dropping null value, dropping columns etc. and then use the Root Cause Analysis, developed by Sakichi Toyoda which is also known as the Five 'Whys' Approach and do analysis according to it using Pivot table and various Excel function such as AVERAGE, COUNTIF, etc.

Tech-Stack Used

I have used Microsoft Excel to do the analysis of the given data.

I have used Microsoft Excel because it is a spreadsheet developed for various platforms. It has calculation or computation capabilities with graphing tools, pivot tables, and also has a macro programming language called Visual Basic for Applications which would make doing the analysis and visualisation effortless.

Data Cleaning

First the data is imported to excel from the csv file after importing the data the duplicate data are checked and deleted using the data tools. After that we check for the gaps in the columns using the trim function, then we look for blank cells using the find special and delete them as it would be of no help for analysis. After that we check the columns and according to our basic understandings we delete the columns which will be of no use for the analysis and it would make analysis easy.

Movies with highest profit

The Top 20 movies with highest profit are given in the table

S.No	Movie	Profit	S.No	Movie	Profit
1	Avatar	$ 523,505,847	11	Deadpool	$ 305,024,263
2	Jurassic World	$ 502,177,271	12	The Hunger Games: Catching Fire	$ 294,645,577
3	Titanic	$ 458,672,302	13	Jurassic Park	$ 293,784,000
4	Star Wars: Episode IV - A New Hope	$ 449,935,665	14	Despicable Me 2	$ 292,049,635
5	E.T. the Extra-Terrestrial	$ 424,449,459	15	American Sniper	$ 291,323,553
6	The Avengers	$ 403,279,547	16	Finding Nemo	$ 286,838,870
7	The Lion King	$ 377,783,777	17	Shrek 2	$ 286,471,036
8	Star Wars: Episode I - The Phantom Menace	$ 359,544,677	18	The Lord of the Rings: The Return of the King	$ 283,019,252
9	The Dark Knight	$ 348,316,061	19	Star Wars: Episode VI - Return of the Jedi	$ 276,625,409
10	The Hunger Games	$ 329,999,255	20	Forrest Gump	$ 274,691,196

INSIGHTS

Avatar is the highest profitable movie having profit of $523,505,847 followed by *Jurassic Park* and then *Titanic. The Hunger Games* is at 10th position and *Forest Gump at 20th position.*
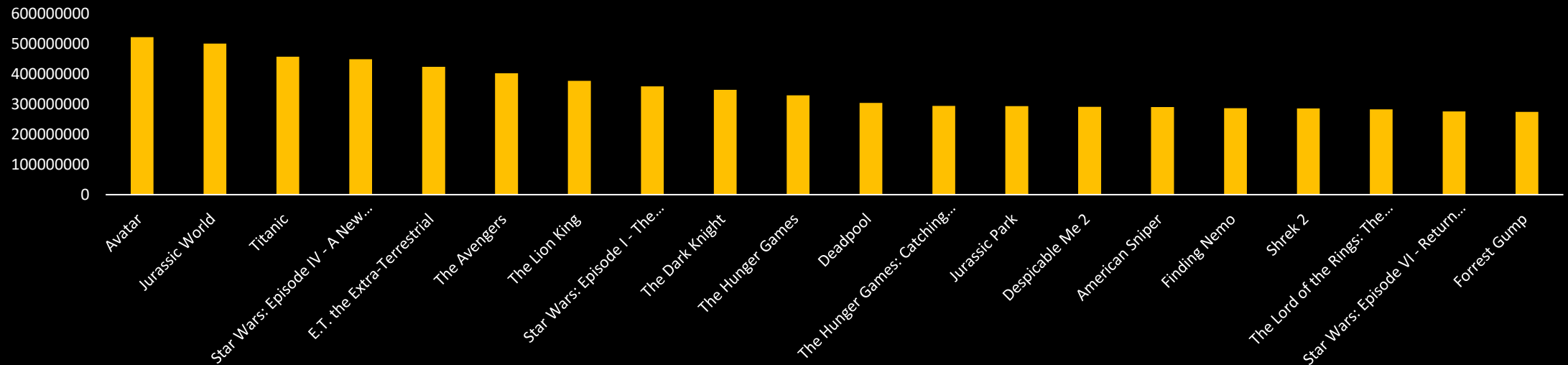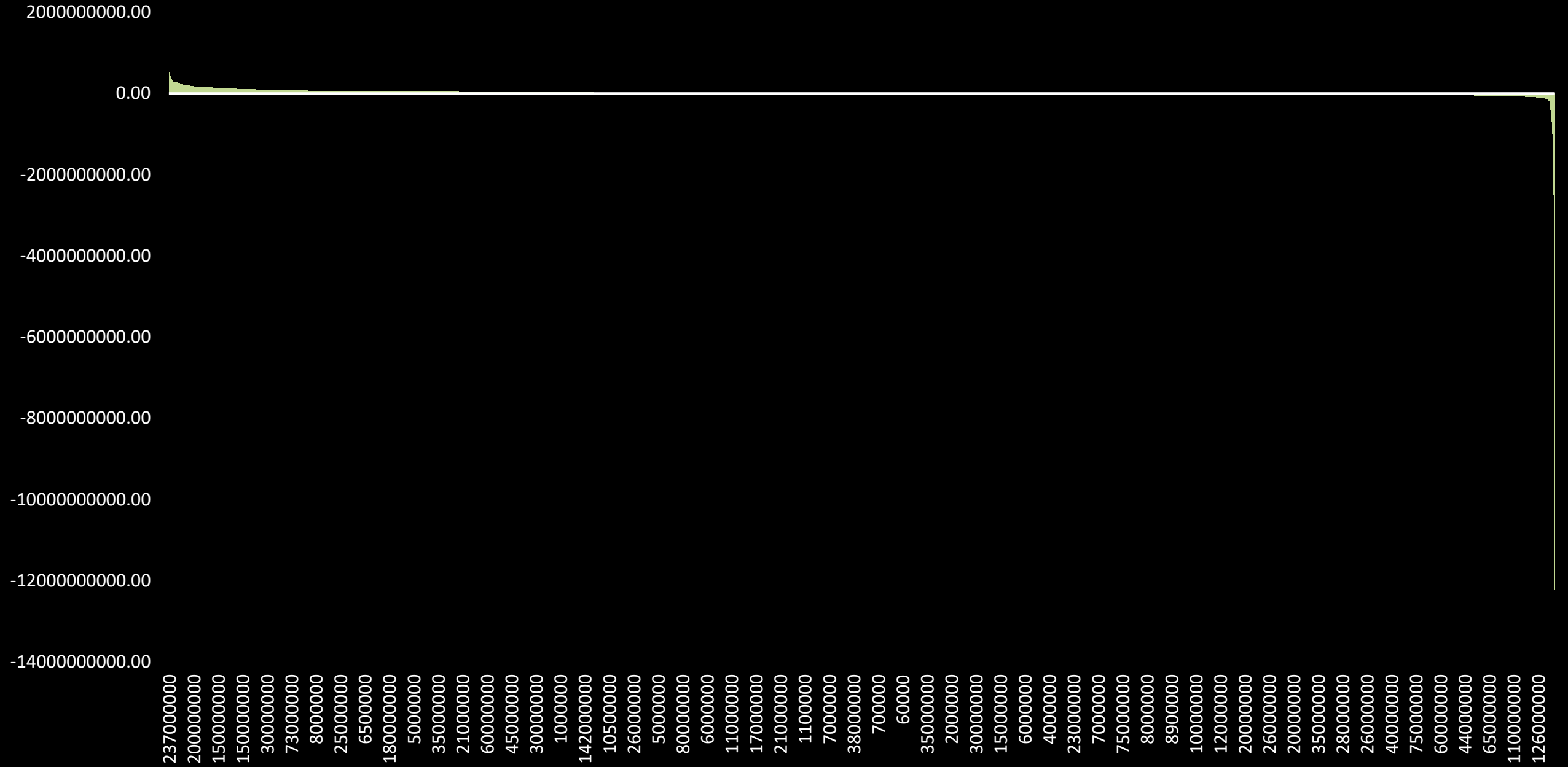
EXPLAINATION

To Extract the desired data a new column named Profit is made which contains the difference of Gross and Budget of each movie. After that, the Profit column is sorted as highest to lowest and then the top movies are extracted for the analysis and various charts are plotted.

BUDGET TO PROFIT CHART

■ Profit

IMDB Top 250

1	Towering Inferno	9.5	31	Queen of the Mountains	8.7	61	Raiders of the Lost Ark	8.5
2	The Shawshank Redemption	9.3	32	Seven Samurai	8.7	62	Psycho	8.5
3	The Godfather	9.2	33	Butterfly Girl	8.7	63	Back to the Future	8.5
4	Dekalog	9.1	34	Interstellar	8.6	64	Alien	8.5
4	Dekalog	9.1	35	Hannibal	8.6	65	Memento	8.5
6	The Dark Knight	9	36	Saving Private Ryan	8.6	66	Airlift	8.5
7	The Godfather: Part II	9	37	Luther	8.6	67	Samsara	8.5
8	Fargo	9	38	Se7en	8.6	68	Whiplash	8.5
9	The Lord of the Rings: The Return of the King	8.9	39	The Silence of the Lambs	8.6	69	The Lives of Others	8.5
10	Schindler's List	8.9	40	Spirited Away	8.6			
11	Pulp Fiction	8.9	41	Spartacus: War of the Damned	8.6	70	Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb	8.5
12	The Good, the Bad and the Ugly	8.9	42	American History X	8.6	71	Happy Valley	8.5
13	12 Angry Men	8.9	43	The Usual Suspects	8.6	72	Children of Heaven	8.5
14	Inception	8.8	44	Once Upon a Time in the West	8.6	73	WALL·E	8.4
15	The Lord of the Rings: The Fellowship of the Ring	8.8	45	Running Forever	8.6	74	Braveheart	8.4
16	Daredevil	8.8	46	It's a Wonderful Life	8.6	75	Stargate SG-1	8.4
17	Fight Club	8.8	47	Modern Times	8.6	76	Amélie	8.4
18	Forrest Gump	8.8	48	Casablanca	8.6	77	Baahubali: The Beginning	8.4
19	It's Always Sunny in Philadelphia	8.8	49	The Dark Knight Rises	8.5	78	Lion of the Desert	8.4
20	Star Wars: Episode V - The Empire Strikes Back	8.8	50	Gladiator	8.5	79	Star Wars: Episode VI - Return of the Jedi	8.4
21	The Lord of the Rings: The Two Towers	8.7	51	Terminator 2: Judgment Day	8.5	80	Once Upon a Time in America	8.4
22	The Matrix	8.7	52	Django Unchained	8.5	81	Princess Mononoke	8.4
23	Friday Night Lights	8.7	53	The Departed	8.5	82	The Shining	8.4
24	The Honeymooners	8.7	54	The Lion King	8.5	83	Aliens	8.4
25	Goodfellas	8.7	55	The Green Mile	8.5	84	American Beauty	8.4
26	Star Wars: Episode IV - A New Hope	8.7	56	Outlander	8.5	85	Lawrence of Arabia	8.4
27	Gomorrah	8.7	57	The Prestige	8.5	86	U2 3D	8.4
28	One Flew Over the Cuckoo's Nest	8.7	58	The Pianist	8.5	87	Das Boot	8.4
29	City of God	8.7	59	Apocalypse Now	8.5	88	Veronica Mars	8.4
30	A Beginner's Guide to Snuff	8.7	60	Entourage	8.5	89	The Inbetweeners	8.4
						90	Rang De Basanti	8.4

#	Title	Rating
91	Psych	8.4
92	Requiem for a Dream	8.4
93	Anne of Green Gables	8.4
94	M*A*S*H	8.4
95	Batman: The Dark Knight Returns, Part 2	8.4
96	Oldboy	8.4
97	To Kill a Mockingbird	8.4
98	Reservoir Dogs	8.4
99	A Separation	8.4
100	The Other Dream Team	8.4
101	A Charlie Brown Christmas	8.4
102	Toy Story 3	8.3
103	Up	8.3
104	Inside Out	8.3
105	Batman Begins	8.3
106	Life	8.3
107	Inglourious Basterds	8.3
108	Indiana Jones and the Last Crusade	8.3
109	The Returned	8.3
110	L.A. Confidential	8.3
111	Toy Story	8.3
112	Scarface	8.3
113	Shaun the Sheep	8.3
114	Eternal Sunshine of the Spotless Mind	8.3
115	Amadeus	8.3
116	Raging Bull	8.3
117	Metropolis	8.3
118	Unforgiven	8.3
119	Downfall	8.3
120	Room	8.3
121	Snatch	8.3
122	2001: A Space Odyssey	8.3
123	Good Will Hunting	8.3
124	Preacher	8.3
125	Lake of Fire	8.3
126	The Sting	8.3
127	The Great Escape	8.3
128	The Hunt	8.3
129	The Apartment	8.3
130	Judgment at Nuremberg	8.3
131	Some Like It Hot	8.3
132	Singin' in the Rain	8.3
133	Inside Job	8.3
134	No End in Sight	8.3
135	Taxi Driver	8.3
136	Hoop Dreams	8.3
137	The Sound and the Shadow	8.3
138	Monty Python and the Holy Grail	8.3
139	Peace, Propaganda & the Promised Land	8.3
140	Captain America: Civil War	8.2
141	How to Train Your Dragon	8.2
142	Godzilla Resurgence	8.2
143	The Wolf of Wall Street	8.2
144	Finding Nemo	8.2
145	A Beautiful Mind	8.2
146	Casino	8.2
147	V for Vendetta	8.2
148	The Thing	8.2
149	Trapped	8.2
150	Die Hard	8.2
151	Blade Runner	8.2
152	Gran Torino	8.2
153	Warrior	8.2
154	Howl's Moving Castle	8.2
155	Into the Wild	8.2
156	Pan's Labyrinth	8.2
157	The Deer Hunter	8.2
158	The Big Lebowski	8.2
159	Buffy the Vampire Slayer	8.2
160	Incendies	8.2
161	War & Peace	8.2
162	Emma	8.2
163	The Elephant Man	8.2
164	Gone with the Wind	8.2
165	The Secret in Their Eyes	8.2
166	Trainspotting	8.2
167	The Bridge on the River Kwai	8.2
168	Lage Raho Munna Bhai	8.2
169	Mr. Smith Goes to Washington	8.2
170	Growing Up Smith	8.2
171	Lock, Stock and Two Smoking Barrels	8.2
172	Rebecca	8.2
173	The Act of Killing	8.2
174	On the Waterfront	8.2
175	Rise of the Entrepreneur: The Search for a Better Way	8.2
176	Home Movies	8.2
177	Cries & Whispers	8.2
178	It Happened One Night	8.2
179	The Brain That Sings	8.2
180	The Last Waltz	8.2

INSIGHTS

'*Towering Inferno*' is the highest rated movie on IMDB with 9.5 rating. A total of 8 movies are rated above 9, 240 movies are rated between 8-9 and 2 movies are rated 8.

EXPLAINATION
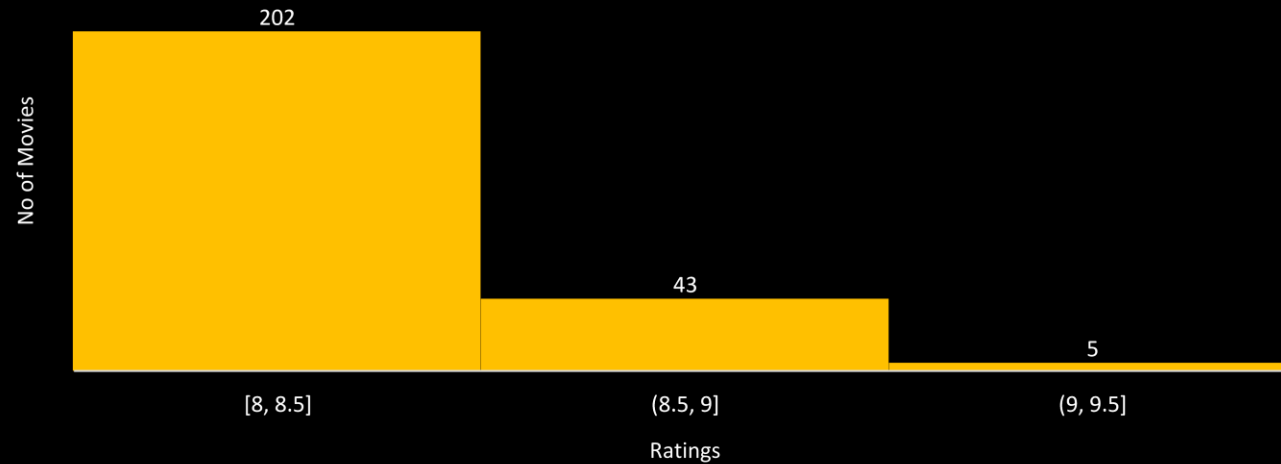
To get the desired results, sort the 'imdb_score' column from highest to lowest and then extract the data accordingly for Non-English movies, sort the 'language' column by deselecting the English language.

MORE INSIGHTS

Non-English movies in IMDB top 250

Movie	IMDB Rating	Movie	IMDB Rating	Movie	IMDB Rating
A Separation	8.4	Godzilla Resurgence	8.2	Spirited Away	8.6
Airlift	8.5	Gomorrah	8.7	Tae Guk Gi: The Brotherhood of War	8.1
Akira	8.1	Howl's Moving Castle	8.2	The Act of Killing	8.2
Amélie	8.4	Incendies	8.2	The Brain That Sings	8.2
Amores Perros	8.1	Lage Raho Munna Bhai	8.2	The Celebration	8.1
Baahubali: The Beginning	8.4	Metropolis	8.3	The Conformist	8.1
Children of Heaven	8.5	Oldboy	8.4	The Good, the Bad and the Ugly	8.9
City of God	8.7	Ordet	8.1	The Hunt	8.3
Cries & Whispers	8.2	Pan's Labyrinth	8.2	The Lives of Others	8.5
Das Boot	8.4	Princess Mononoke	8.4	The Returned	8.3
Dekalog	18.2	Rang De Basanti	8.4	The Sea Inside	8.1
Del 1 - Män som hatar kvinnor	8.1	Samsara	8.5	The Secret in Their Eyes	8.2
Downfall	8.3	Seven Samurai	8.7	The Square	8.1
Elite Squad	8.1	Solaris	8.1	Trapped	8.2

Top 10 language with highest avg rating

Rank	Language	Avg of IMDB Rating
1	Telugu	8.40
2	Polish	8.25
3	None	7.95
4	Indonesian	7.90
5	Maya	7.80
6	Hebrew	7.58
7	Persian	7.58
8	Icelandic	7.55
9	Danish	7.50
10	Dari	7.50

Hindi movies in IMDB top 250 movies

Movie	IMDB Rating
Airlift	8.5
Lage Raho Munna Bhai	8.2
Rang De Basanti	8.4

Best Directors

Best 10 directors with highest average IMDB ratings are:

| Rank	Director	Average IMDB score
1	John Blanchard	9.50
2	Cary Bell	8.70
3	Mitchell Altieri	8.70
4	Sadyk Sher-Niyaz	8.70
5	Charles Chaplin	8.60
6	Mike Mayhall	8.60
7	Damien Chazelle	8.50
8	Majid Majidi	8.50
9	Raja Menon	8.50
10	Ron Fricke	8.50

Avg Director Rating
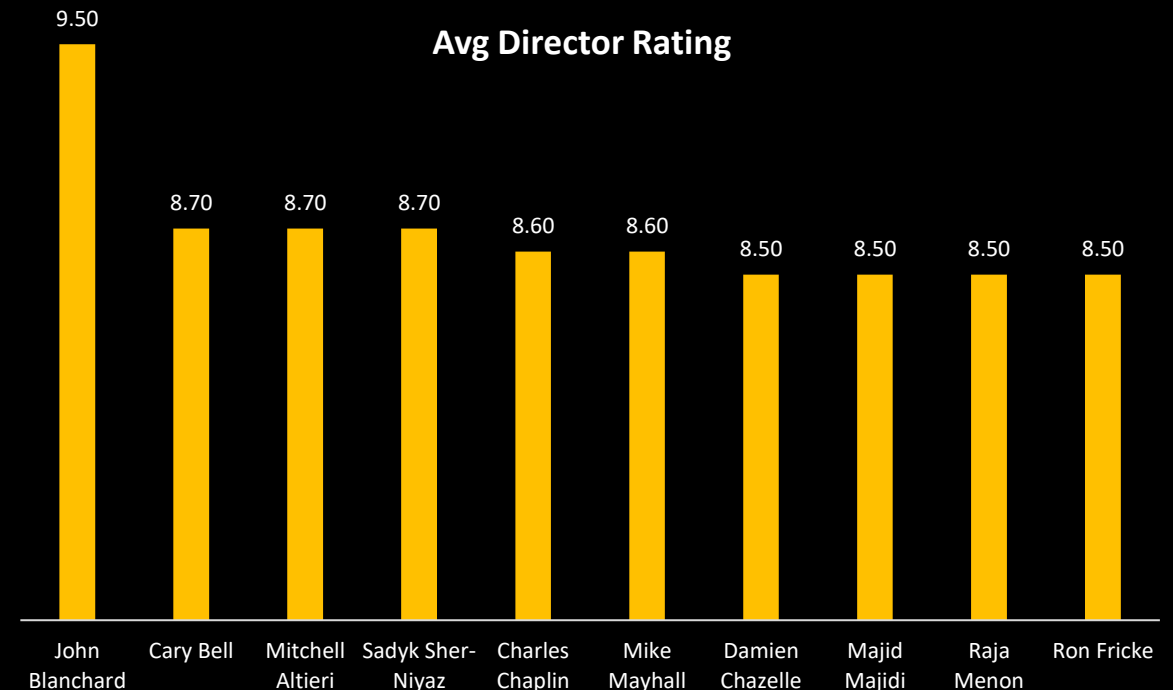
INSIGHTS

'John Blanchard' is the best Director in the IMDB list with average rating of 9.5, *'Ron Fricke'* is at the 10th position with Avg. IMDB rating of 8.5. The Average rating of Top 10 directors is 8.68.

EXPLAINATION

To get the desired result a pivot table is made in which 'director_name' column is put into rows and 'imdb_score' column is put into sum where the value is changed to average. Datas from the Pivot Table is extracted and then new tables are made from the data and then the table with average imdb score is sorted from high to low. Then we get the desired result and can use it for further visual analysis.

Popular Genre

Rank	Genre	No of movies
1	Drama	235.00
2	Comedy	205.00
3	Comedy|Drama	189.00
4	Comedy|Drama|Romance	187.00
5	Comedy|Romance	158.00
6	Drama|Romance	151.00
7	Crime|Drama|Thriller	100.00
8	Horror	70.00
9	Action|Crime|Drama|Thriller	67.00
10	Action|Crime|Thriller	64.00
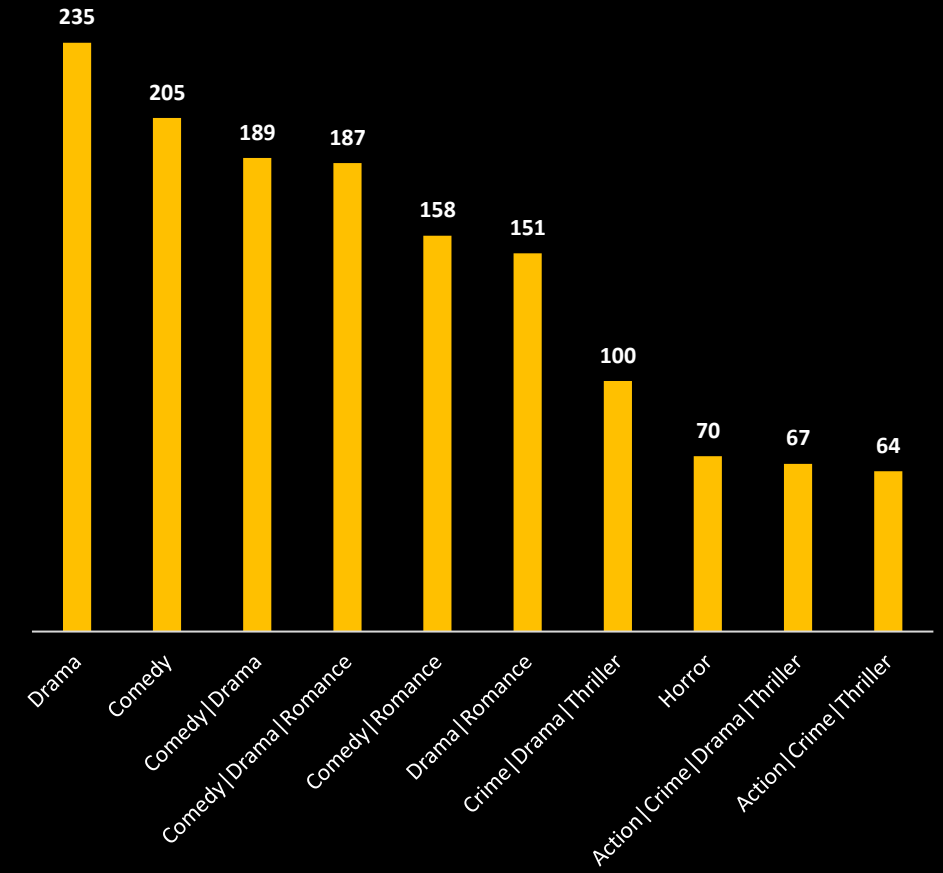
INSIGHTS

'Drama' is the most popular genre on IMDB with most numbers of movies i.e. 235 followed by 'Comedy' with 205 movies.
'Action|Crime|Thriller' is the 10th most popular genre with 64 movies.

EXPLAINATION

To get the desired result a Pivot Table is made from the data in which 'genre' column is set into rows and 'imdb_score' is set into values changed to count to get the count of the genres or movies. After that the data is extracted and new table is made and the count column is sorted from largest to smallest thus getting the desired results for analysis.

Favorite Actors

| Actor	Avg no of user for reviews	Avg no of critic for reviews
Brad Pitt	702.44	231.94
Leonardo DiCaprio	914.48	330.19
Meryl Streep	257.31	163.15
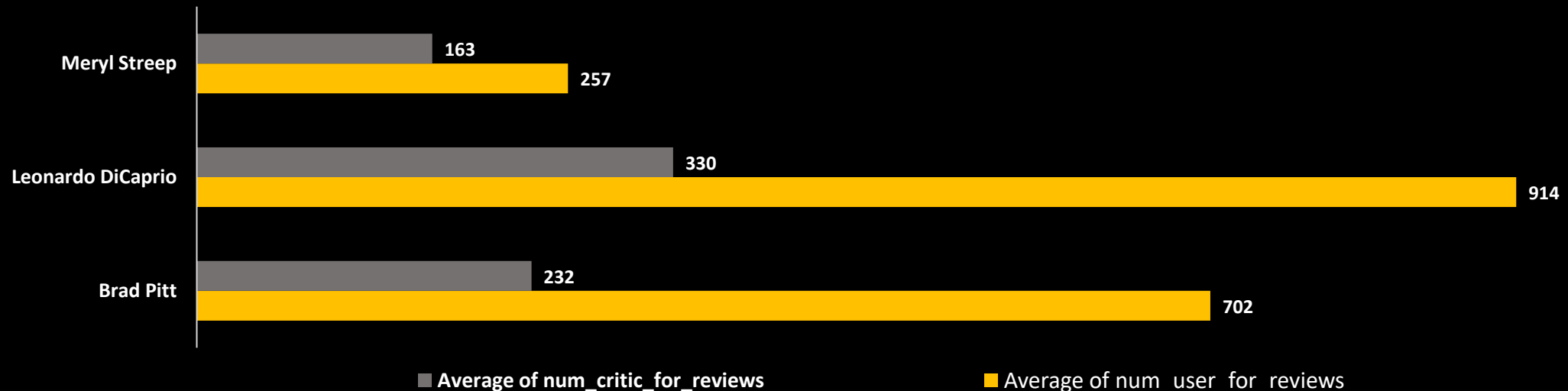
INSIGHTS

According to the data *'Leonardo DiCaprio'* is the favorite actor for user as he got the most numbers of reviews by users followed by *'Bradd Pitt'* and then *'Meryl Streep'*. The same pattern is followed in the case of favorite actor for critics where *'Leonardo DiCaprio'* is reviewed most by the critics followed by *'Bradd Pitt'* and then *'Meryl Streep'*.

EXPLAINATION

To get the desired data 3 new columns are made named after the 3 actors containing their movies, then these data are appended into new column along with their number of reviews for users and critics separately. After that the mean of the reviews are calculated using the AVERAGEIF function of excel of all three actors and then the data is used for analysis and chart preparation.

Change in number of voted users over decades

Decade	No. of reviews
1910	88
1920	614
1930	2745
1940	4254
1950	5231
1960	17394
1970	32513
1980	64724
1990	210484
2000	686738
2010	319863
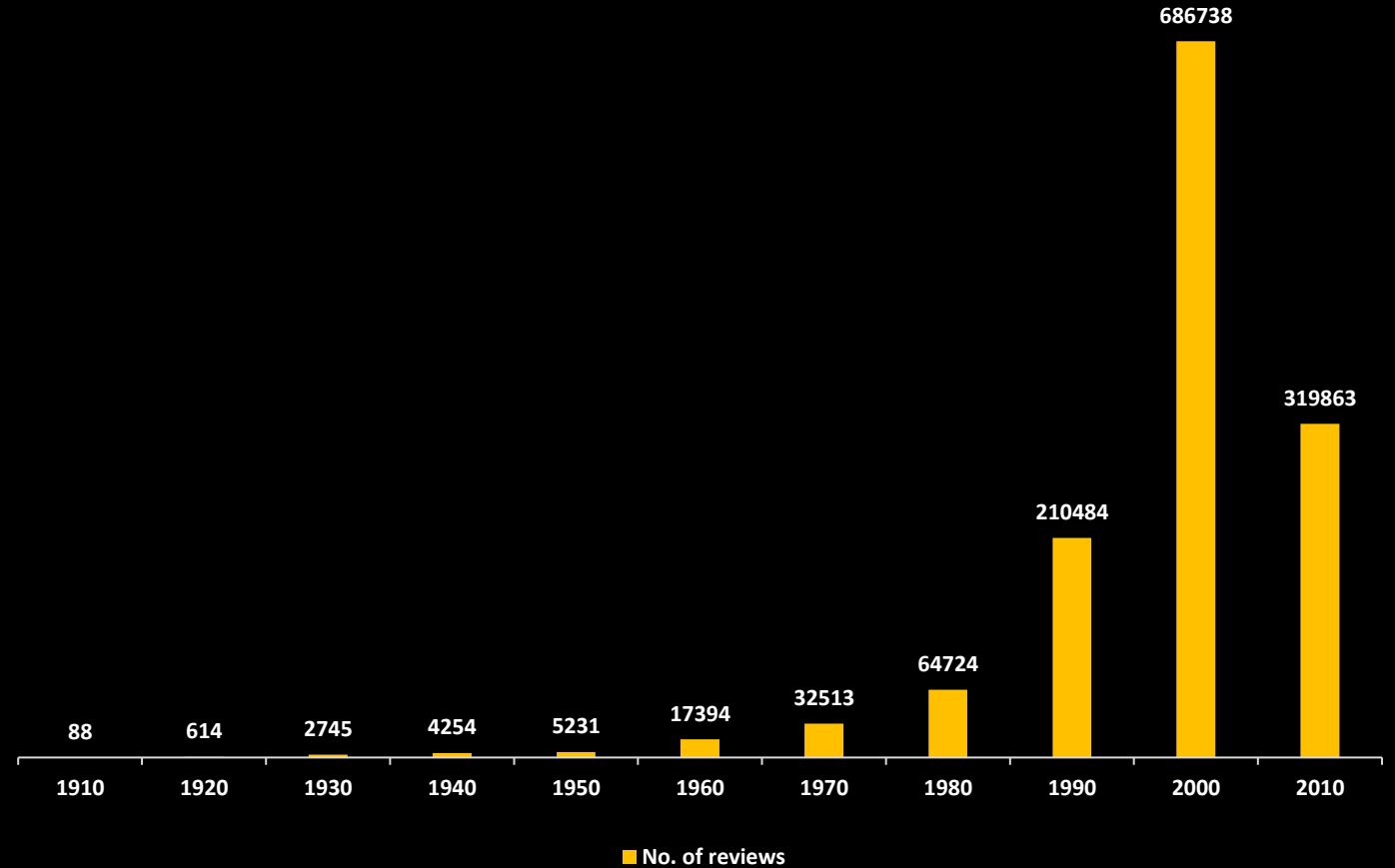
INSIGHTS

As we can see on the bar graph, the number of reviews has increased exponentially over the decades till 2000s and then the number of review decreases in the 2010s.

EXPLAINATION

To get the desired data we first find the decades of the year by using the formula =INT(year/10)*10 after acquiring the decades we use the SUMIF function for different decades and then draw table out of it and charts for visual analysis.

RESULT

I got to understand how different problems are framed from the dataset and how to define the problem. And then doing its analysis using the 5 why Analysis developed by Sakichi Toyoda and then draw insights out of it.

THANK YOU

Priyanshu Kamal

callmepriyanshu4@gmail.com