




# IMDB Movie Analysis

Final Project-1

BY AKSHAT MITTAL

# PROJECT DESCRIPTION

- ▶ The project is about finding out the various insights in IMDB Movies dataset .
- ▶ We analyze this data and answer the questions that follows:
  - A. **Cleaning the data:** This is one of the most important step to perform before moving forward with the analysis.
  - B. **Movies with Highest Profit:** Create a new column called profit which contains the difference of the two columns: gross and budget.
  - C. **Top 250:** Create a new column IMDb\_Top\_250 and store the top 250 movies with the highest IMDb Rating. Also make sure that for all of these movies, the num\_voted\_users is greater than 25,000. 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.

- 
- 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`.
  - E. Popular Genres:** Find popular genres. Perform this step using the knowledge gained while performing previous steps.
  - 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. Append the rows of all these columns and group the combined 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`. Find the critic-favorite and audience-favorite actors



# APPROACH

- ▶ First thing is to accessing the dataset carefully using Excel and understanding that what's happening in the data.
- ▶ Applying functions to remove the null values and the duplicate data.
- ▶ Categorizing the data into different tables to gain insights asked in the project.
- ▶ Creating the graphs and other visualization methods to get more descriptive analysis for easy representation.

# TECH-STACK Used

- ▶ I have used Microsoft Excel to perform statistical analysis on this dataset because allows users to edit, organize, and analyze different types of information.
- ▶ It allows collaborations, and multiple users can edit and format files in real-time, and any changes made to the spreadsheet can be tracked by a revision history.
- ▶ This tool is used to create graphical representation of the results and to understand the result set better.
- ▶ Microsoft Excel enables us to format, organize, data analyses, and calculate data in a spread sheet, and users can make information easier to view as data is added or changed.
- ▶ Microsoft PowerPoint used to make the report to represent the project report clearly.



# INSIGHTS

- ▶ There were several function present in Excel which was quite unknown to me.
- ▶ The data set was quite large and working with such dataset in real time was fun and more over easy learnable.
- ▶ There was critical analysis was involved in this whole process.
- ▶ Creating tables and using charts to analyze something is quite useful and we have also seen its practical approach.

# 1) Cleaning the Data:

- ▶ This is the most difficult and crucial step of any Data analysis project.
- ▶ The process included in this step varies from question to question and Dataset to Dataset.
- ▶ To clean the dataset we will be:-
  1. First dropping the columns which have no use for the analysis that we will be doing.
  2. Columns like 'Color', 'director\_facebook\_likes', 'actor\_3\_facebook\_likes', 'actor\_2\_name', 'actor\_1\_facebook\_likes', 'cast\_total\_facebook\_likes', 'actor\_3\_name', 'facenumber\_in\_posts', 'plot\_keywords', 'movie\_imdb\_link', 'content\_rating', 'actor\_2\_facebook\_likes', 'aspect\_ratio', 'movie\_facebook\_likes' are the columns containing irrelevant data for the analysis tasks provided. So, these columns needs to be dropped.
  3. After dropping the irrelevant columns now we need to remove the rows from the dataset having anyone of its column value as blank/NULL.
  4. Then we need to get rid off the duplicate values in the dataset which can be achieved by using the 'Remove Duplicate Values/Cells' available in the 'Data' tab.

# Cleaned Data:

	director_name	num_critics_for_rev	duration	gross	genres	actor_1_name	movie_title	num_voted_us	num_user_for_rev	language	country	budget	title_year	imdb_score
1	James Cameron	723	178	7.61E+08	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	886204	3054	English	USA	237000000	2009	7.9
2	Gore Verbinski	302	169	3.09E+08	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	1238	English	USA	300000000	2007	7.1
3	Sam Mendes	602	148	2E+08	Action Adventure Thriller	Christoph Waltz	Spectre	275868	994	English	UK	245000000	2015	6.8
5	Christopher Nolan	813	164	4.48E+08	Action Thriller	Tom Hardy	The Dark Knight Rises	1144337	2701	English	USA	250000000	2012	8.5
6	Andrew Stanton	462	132	73058679	Action Adventure Sci-Fi	Daryl Sabara	John Carter	212204	738	English	USA	263700000	2012	6.6
7	Sam Raimi	392	156	3.37E+08	Action Adventure Romance	J.K. Simmons	Spider-Man 3	383056	1902	English	USA	258000000	2007	6.2
8	Nathan Greno	324	100	2.01E+08	Adventure Animation Comedy Family Fantasy Musical Romance	Brad Garrett	Tangled	294810	387	English	USA	260000000	2010	7.8
9	Joss Whedon	635	141	4.59E+08	Action Adventure Sci-Fi	Chris Hemsworth	Avengers: Age of Ultron	462669	1117	English	USA	250000000	2015	7.5
10	David Yates	375	153	3.02E+08	Adventure Family Fantasy Mystery	Alan Rickman	Harry Potter and the Half-Blood Prince	321795	973	English	UK	250000000	2009	7.5
11	Zack Snyder	673	183	3.3E+08	Action Adventure Sci-Fi	Henry Cavill	Batman v Superman: Dawn of Justice	371639	3018	English	USA	250000000	2016	6.9
12	Bryan Singer	434	169	2E+08	Action Adventure Sci-Fi	Kevin Spacey	Superman Returns	240396	2367	English	USA	209000000	2006	6.1
13	Marc Forster	403	106	1.68E+08	Action Adventure	Giancarlo Giannini	Quantum of Solace	330784	1243	English	UK	200000000	2008	6.7
14	Gore Verbinski	313	151	4.23E+08	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: Dead Man's Chest	522040	1832	English	USA	225000000	2006	7.3
15	Gore Verbinski	450	150	89289910	Action Adventure Western	Johnny Depp	The Lone Ranger	181792	711	English	USA	216000000	2013	6.5
16	Zack Snyder	733	143	2.91E+08	Action Adventure Fantasy Sci-Fi	Henry Cavill	Man of Steel	548573	2536	English	USA	225000000	2013	7.2
17	Andrew Adamson	258	150	1.42E+08	Action Adventure Family Fantasy	Peter Dinklage	The Chronicles of Narnia: Prince Caspian	149322	438	English	USA	225000000	2008	6.6
18	Joss Whedon	703	173	6.23E+08	Action Adventure Sci-Fi	Chris Hemsworth	The Avengers	995415	1722	English	USA	220000000	2012	8.1
19	Rob Marshall	448	136	2.41E+08	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: On Stranger Tides	370704	484	English	USA	250000000	2011	6.7
20	Barry Sonnenfeld	451	106	1.79E+08	Action Adventure Comedy Family Fantasy Sci-Fi	Will Smith	Men in Black 3	268154	341	English	USA	225000000	2012	6.8
21	Peter Jackson	422	164	2.55E+08	Adventure Fantasy	Aidan Turner	The Hobbit: The Battle of the Five Armies	354228	902	English	New Zealand	250000000	2014	7.5
22	Marc Webb	539	153	2.62E+08	Action Adventure Fantasy	Emma Stone	The Amazing Spider-Man	451803	1225	English	USA	230000000	2012	7
23	Ridley Scott	343	156	1.05E+08	Action Adventure Drama History	Mark Addy	Robin Hood	211765	546	English	USA	200000000	2010	6.7
24	Peter Jackson	509	186	2.59E+08	Adventure Fantasy	Aidan Turner	The Hobbit: The Desolation of Smaug	483540	951	English	USA	225000000	2013	7.9
25	Chris Weitz	251	113	70083519	Adventure Family Fantasy	Christopher Lee	The Golden Compass	149019	666	English	USA	180000000	2007	6.1
26	Peter Jackson	446	201	2.18E+08	Action Adventure Drama Romance	Naomi Watts	King Kong	316018	2618	English	New Zealand	207000000	2005	7.2
27	James Cameron	315	194	6.59E+08	Drama Romance	Leonardo DiCaprio	Titanic	793059	2528	English	USA	200000000	1997	7.7
28	Anthony Russo	516	147	4.07E+08	Action Adventure Sci-Fi	Robert Downey Jr.	Captain America: Civil War	272670	1022	English	USA	250000000	2016	8.2
29	Peter Berg	377	131	65173160	Action Adventure Sci-Fi Thriller	Liam Neeson	Battleship	202382	751	English	USA	205000000	2012	5.9
30	Colin Trevorrow	644	124	6.52E+08	Action Adventure Sci-Fi Thriller	Bryce Dallas Howard	Jurassic World	418214	1290	English	USA	150000000	2015	7
31	Sam Mendes	750	143	3.04E+08	Action Adventure Thriller	Albert Finney	Skyfall	522030	1498	English	UK	200000000	2012	7.8
32	Sam Raimi	300	135	3.73E+08	Action Adventure Fantasy Romance	J.K. Simmons	Spider-Man 2	411164	1303	English	USA	200000000	2004	7.3
33	Shane Black	608	195	4.09E+08	Action Adventure Sci-Fi	Robert Downey Jr.	Iron Man 3	557489	1187	English	USA	200000000	2013	7.2
34	Tim Burton	451	108	3.34E+08	Adventure Family Fantasy	Johnny Depp	Alice in Wonderland	306320	736	English	USA	200000000	2010	6.5
35	Brett Ratner	334	104	2.34E+08	Action Adventure Fantasy Sci-Fi Thriller	Hugh Jackman	X-Men: The Last Stand	383427	1912	English	Canada	210000000	2006	6.8
36	Dan Scanlon	376	104	2.68E+08	Adventure Animation Comedy Family Fantasy	Steve Buscemi	Monsters University	235025	265	English	USA	200000000	2013	7.3
37	Michael Bay	366	150	4.02E+08	Action Adventure Sci-Fi	Glenn Morshower	Transformers: Revenge of the Fallen	323207	1439	English	USA	200000000	2009	6
38	Michael Bay	378	165	2.45E+08	Action Adventure Sci-Fi	Bingbing Li	Transformers: Age of Extinction	242420	918	English	USA	210000000	2014	5.7
39	Sam Raimi	525	130	2.35E+08	Adventure Family Fantasy	Tim Holmes	Oz the Great and Powerful	175409	511	English	USA	215000000	2013	6.4
40	Marc Webb	495	142	2.03E+08	Action Adventure Fantasy Sci-Fi	Emma Stone	The Amazing Spider-Man 2	321227	1067	English	USA	200000000	2014	6.7
41	Joseph Kosinski	469	125	1.72E+08	Action Adventure Sci-Fi	Jeff Bridges	TRON: Legacy	264183	665	English	USA	170000000	2010	6.8
42	John Lasseter	304	106	1.91E+08	Adventure Animation Comedy Family Sport	Joe Mantegna	Cars 2	101178	283	English	USA	200000000	2011	6.3
43	Martin Campbell	436	123	1.17E+08	Action Adventure Sci-Fi	Ryan Reynolds	Green Lantern	223393	550	English	USA	200000000	2011	5.6
44	Lee Unkrich	453	103	4.15E+08	Adventure Animation Comedy Family Fantasy	Tom Hanks	Toy Story 3	544884	733	English	USA	200000000	2010	8.3
45	McG	422	118	1.25E+08	Action Adventure Sci-Fi	Christian Bale	Terminator Salvation	286095	974	English	USA	200000000	2009	6.6

Cleaned Dataset Link: <https://docs.google.com/spreadsheets>



## 2) Movies with the Highest Profit:

Sr. No.	Movie_Title	Gross	Budget	Profit
1	Avatar	760505847	237000000	523505847
2	Jurassic World	652177271	150000000	502177271
3	Titanic	658672302	200000000	458672302
4	Star Wars: Episode IV - A New Hope	460935665	11000000	449935665
5	E.T. the Extra-Terrestrial	434949459	10500000	424449459
6	The Avengers	623279547	220000000	403279547
7	The Lion King	422783777	45000000	377783777
8	Star Wars: Episode I - The Phantom Menace	474544677	115000000	359544677
9	The Dark Knight	533316061	185000000	348316061
10	The Hunger Games	407999255	78000000	329999255

The outliers are:-

-12213298588

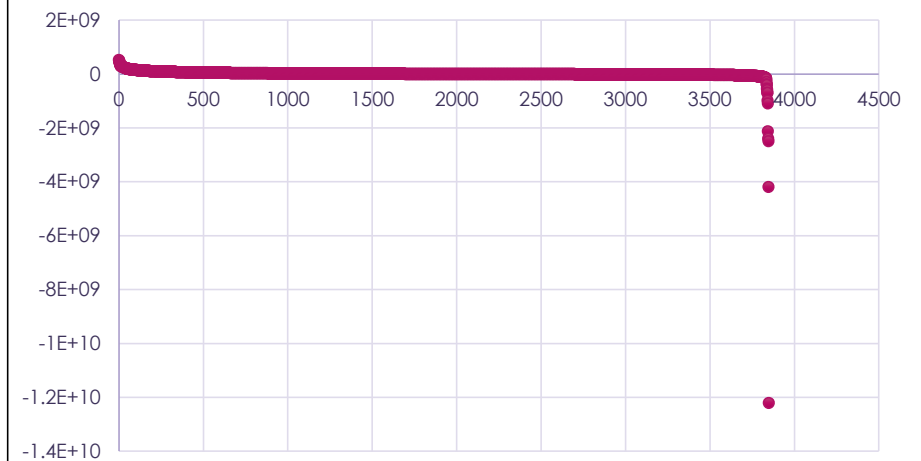
-4199788333

-2499804112

-2127109510

-1099560838

Profit vs Budget



### 3) Top 250: Find IMDB Top 250

#### Top 25 Movies in all Languages:

Rank	IMDb_Top_250	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	The Lord of the Rings: The Return of the King	8.9	1215718
6	Schindler's List	8.9	865020
7	Pulp Fiction	8.9	1324680
8	The Good, the Bad and the Ugly	8.9	503509
9	Inception	8.8	1468200
10	The Lord of the Rings: The Fellowship of the Ring	8.8	1238746
11	Fight Club	8.8	1347461
12	Forrest Gump	8.8	1251222
13	Star Wars: Episode V - The Empire Strikes Back	8.8	837759
14	The Lord of the Rings: The Two Towers	8.7	1100446
15	The Matrix	8.7	1217752
16	Goodfellas	8.7	728685
17	Star Wars: Episode IV - A New Hope	8.7	911097
18	One Flew Over the Cuckoo's Nest	8.7	680041
19	City of God	8.7	533200
20	Seven Samurai	8.7	229012
21	Interstellar	8.6	928227
22	Saving Private Ryan	8.6	881236
23	Se7en	8.6	1023511
24	The Silence of the Lambs	8.6	887467
25	Spirited Away	8.6	417971

#### Top 25 Movies not in English Language:

Rank	Movie_Title	Language	IMDb_Score	Num_Voted_Users
1	The Good, the Bad and the Ugly	Italian	8.9	503509
2	City of God	Portuguese	8.7	533200
3	Seven Samurai	Japanese	8.7	229012
4	Spirited Away	Japanese	8.6	417971
5	The Lives of Others	German	8.5	259379
6	Children of Heaven	Persian	8.5	27882
7	Amélie	French	8.4	534262
8	Baahubali: The Beginning	Telugu	8.4	62756
9	Princess Mononoke	Japanese	8.4	221552
10	Das Boot	German	8.4	168203
11	Oldboy	Korean	8.4	356181
12	A Separation	Persian	8.4	151812
13	Metropolis	German	8.3	111841
14	Downfall	German	8.3	248354
15	The Hunt	Danish	8.3	170155
16	Howl's Moving Castle	Japanese	8.2	214091
17	Pan's Labyrinth	Spanish	8.2	467234
18	Incendies	French	8.2	80429
19	The Secret in Their Eyes	Spanish	8.2	131831
20	The Sea Inside	Spanish	8.1	64556
21	Tae Guk Gi: The Brotherhood of War	Korean	8.1	31943
22	Akira	Japanese	8.1	106160
23	Elite Squad	Portuguese	8.1	81644
24	Amores Perros	Spanish	8.1	173551
25	The Celebration	Danish	8.1	65951

IMDB Top 250 Dataset Link: <https://docs.google.com/spreadsheets/>

## 4) Best Directors:

### Top10 Best Directors:

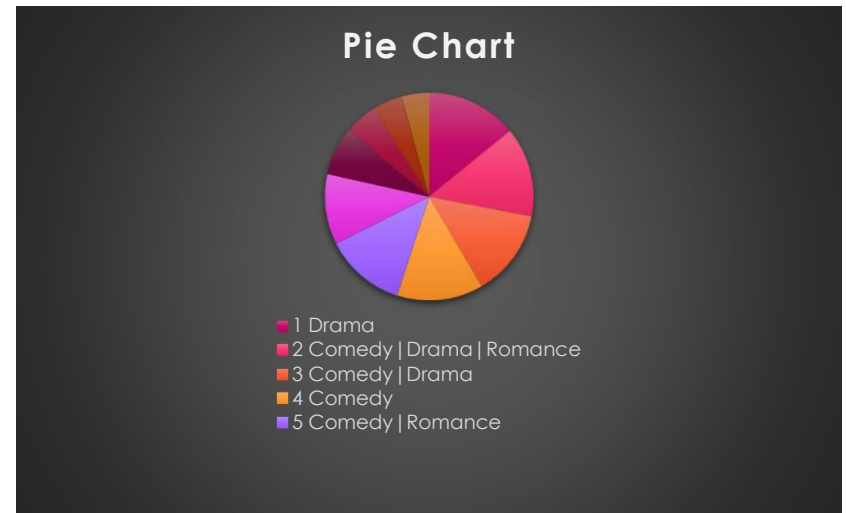
Rank	Top_10_Directors	Mean of IMDb_Score
1	Charles Chaplin	8.6
2	Tony Kaye	8.6
3	Alfred Hitchcock	8.5
4	Damien Chazelle	8.5
5	Majid Majidi	8.5
6	Ron Fricke	8.5
7	Sergio Leone	8.433333333
8	Christopher Nolan	8.425
9	Asghar Farhadi	8.4
10	Marius A. Markevicius	8.4

- One of the best director is **Charles Chaplin** as seen from the observations.

## 5) Popular Genres:

Here are the top 10 popular genres-

Rank	Popular Genres	Count
1	Drama	153
2	Comedy   Drama   Romance	151
3	Comedy   Drama	147
4	Comedy	145
5	Comedy   Romance	136
6	Drama   Romance	119
7	Crime   Drama   Thriller	82
8	Action   Crime   Thriller	55
9	Action   Crime   Drama   Thriller	50
10	Action   Adventure   Sci-Fi	46

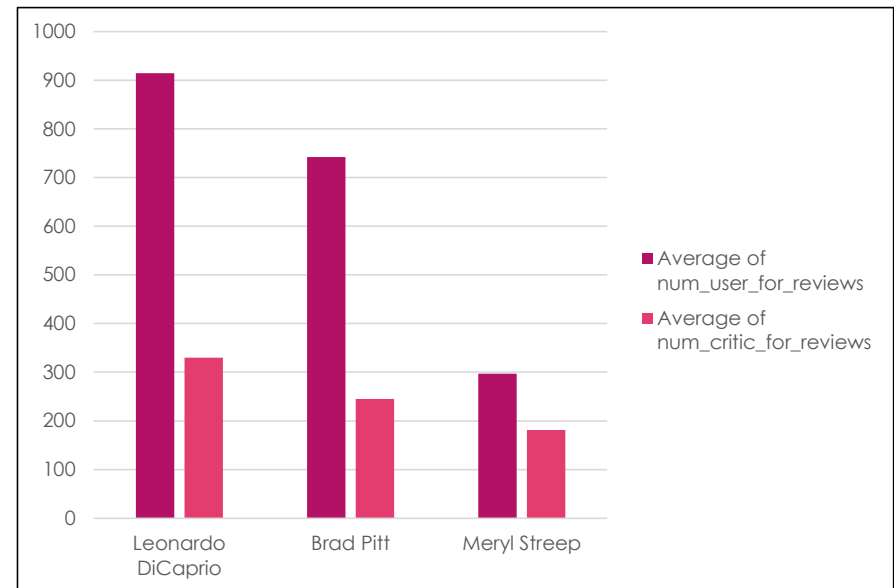


➤ Most popular genre is **Drama** followed by Comedy.

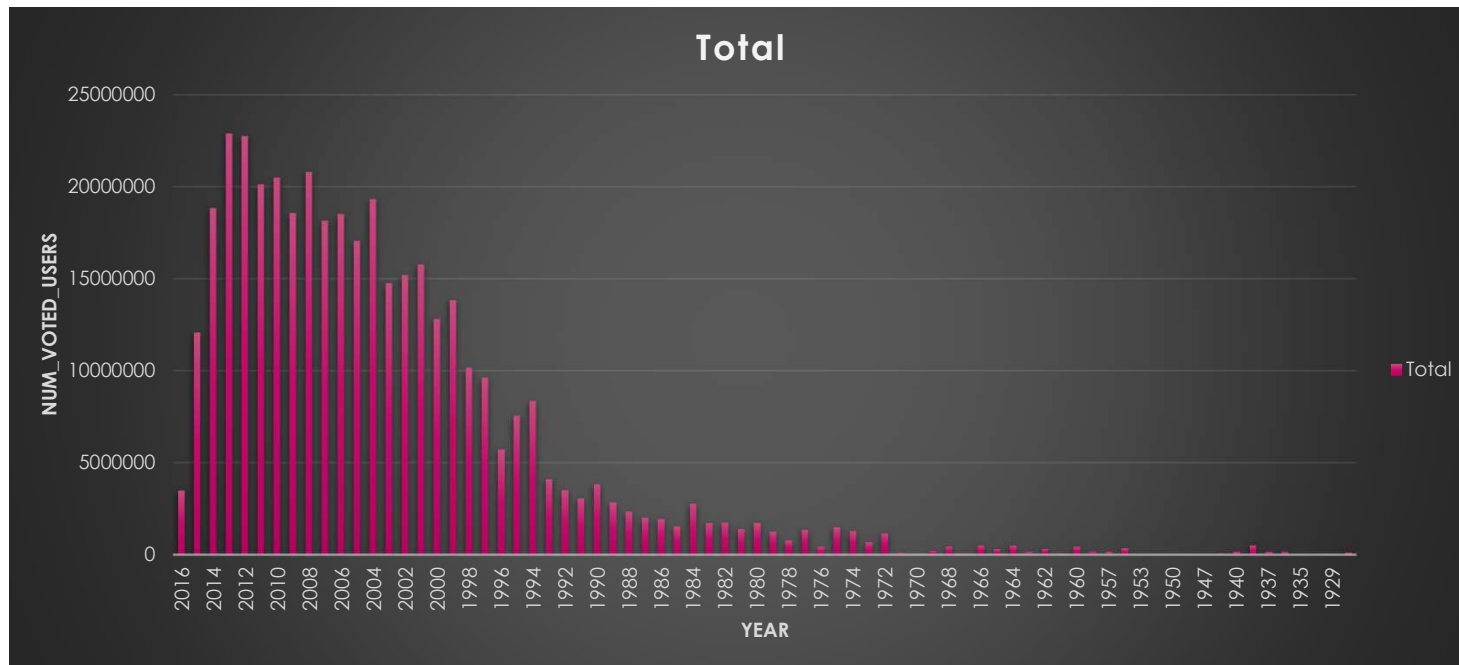
## 6) Charts: Find the critic-favorite and audience-favorite actors

Actor_Name	Average of num_user_for_reviews	Average of num_critic_for_reviews
Leonardo DiCaprio	914.4761905	330.1904762
Brad Pitt	742.3529412	245
Meryl Streep	297.1818182	181.4545455
<b>Grand Total</b>	<b>716.1836735</b>	<b>267.244898</b>

➤ As we can see clearly from the results, Leonardo DiCaprio is the critic-favorite and audience-favorite actor.

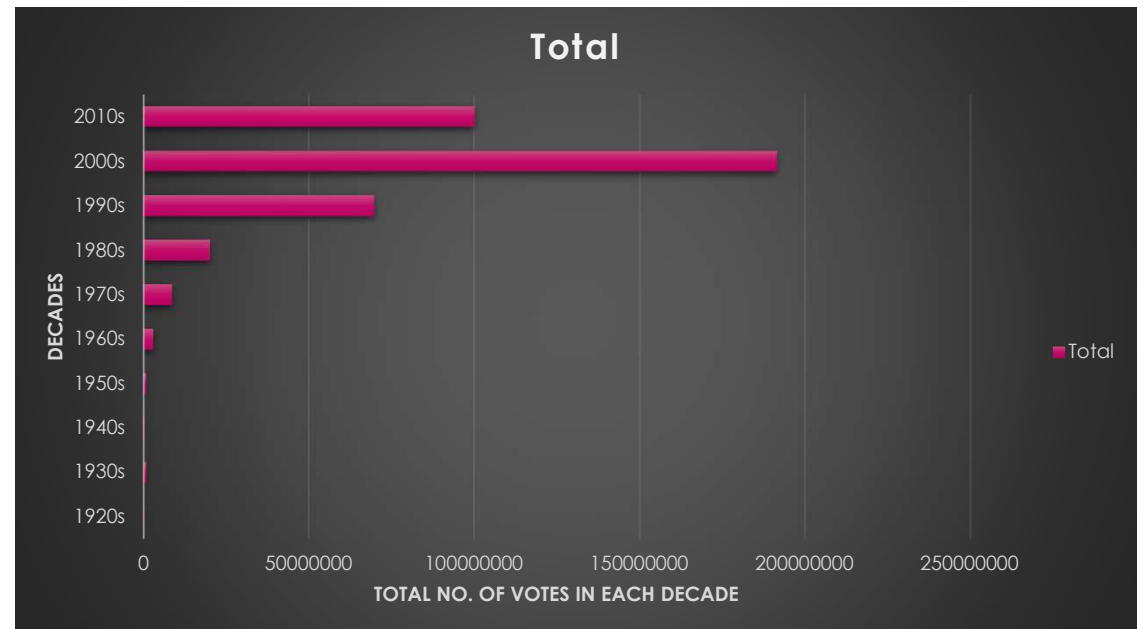


# Change in Number of Voted Users Over Decades



## ❖ Change in Number of Voted Users Over Decades using Bar Chart

Decade	Sum of num_voted_users
1920s	116387
1930s	804839
1940s	230838
1950s	678336
1960s	2983442
1970s	8524102
1980s	19987476
1990s	69735679
2000s	191426254
2010s	100148261
<b>Grand Total</b>	<b>394635614</b>



# Working file Containing all the Analysis of IMDB Movie Dataset:

## **Link:**

[https://docs.google.com/spreadsheets/IMDB\\_Movies\\_Dataset](https://docs.google.com/spreadsheets/IMDB_Movies_Dataset)

- These are the required insights and the result which were asked.
- Hence, all the questions given as a part of Data Analytics IMDB Movie Analysis (Final Project -1) have been provided with answers along with graphs.



# RESULT

- ▶ In this task all the concepts regarding to Excel and statistics like sort, filter, pivot- table, etc. have been implemented using Microsoft Excel.
- ▶ This project has helped me to understand the concept of exploratory data analysis.
- ▶ From this project, I learnt the need of data filtration and how to generate deeper insights for a problem mentioned.
- ▶ I have gained knowledge of various excel sheet functions which helped me to solve the questions asked in this project.
- ▶ I have achieved how to analyze the project, as this helped me to gain more knowledge about Excel to perform better.



THANK YOU