IMDB Movie Analysis

Tech-Stack Used- Excel

Analysis done on the following points:

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.)

Your task: Clean the data

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.

Your task: Find the movies with the highest profit?

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!

Your task: Find IMDB Top 250

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.

Your task: Find the best directors

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

Your task: Find popular genres

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.

Your task: Find the critic-favourite and audience-favourite actors

Cleaning the data

This is the most difficult and crucial step of any Data analysis project.

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 need 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 of the duplicate values in the dataset which can be achieved by using the 'Remove Duplicate Values/Cells' available in the 'Data' tab

Google Drive Link for IMDB Movies cleaned data:

https://docs.google.com/spreadsheets/d/1Nbv8EV9BUxMqX1nQ0Ddek9dueowyCwHS/edit?usp=share_link&ouid=116077077614362440241&rtpof=true&sd=true

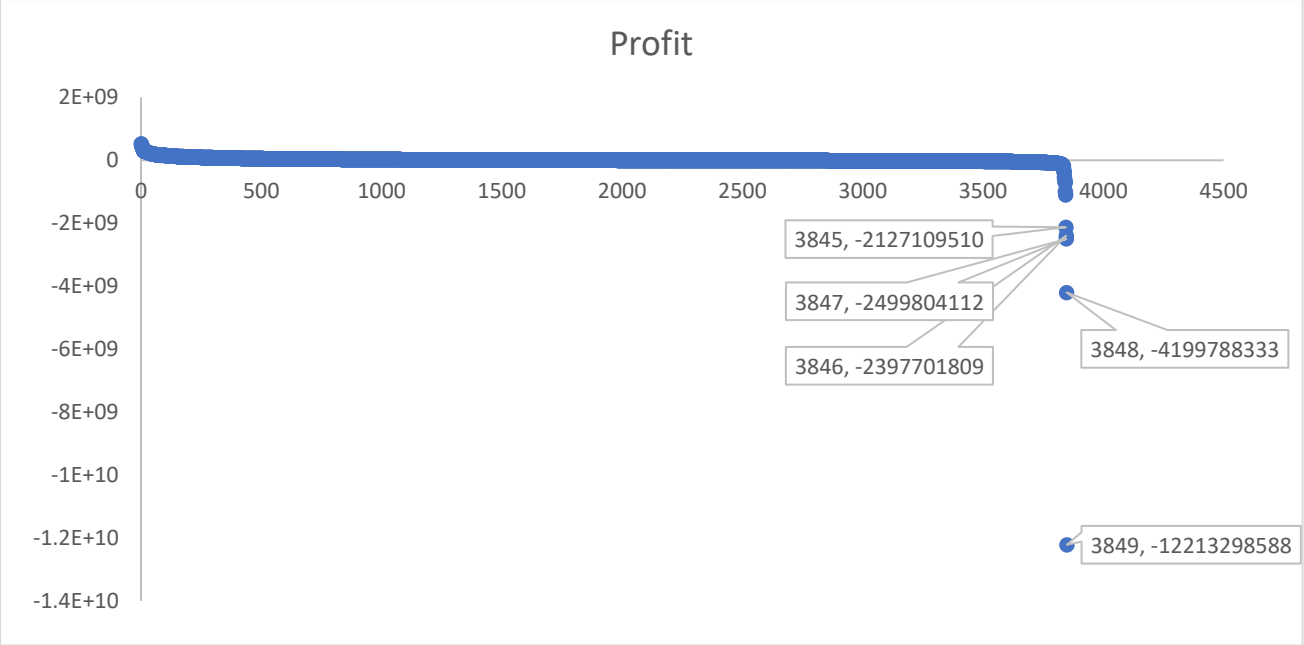
Movies with highest profit

To find the movies with the highest profit: -

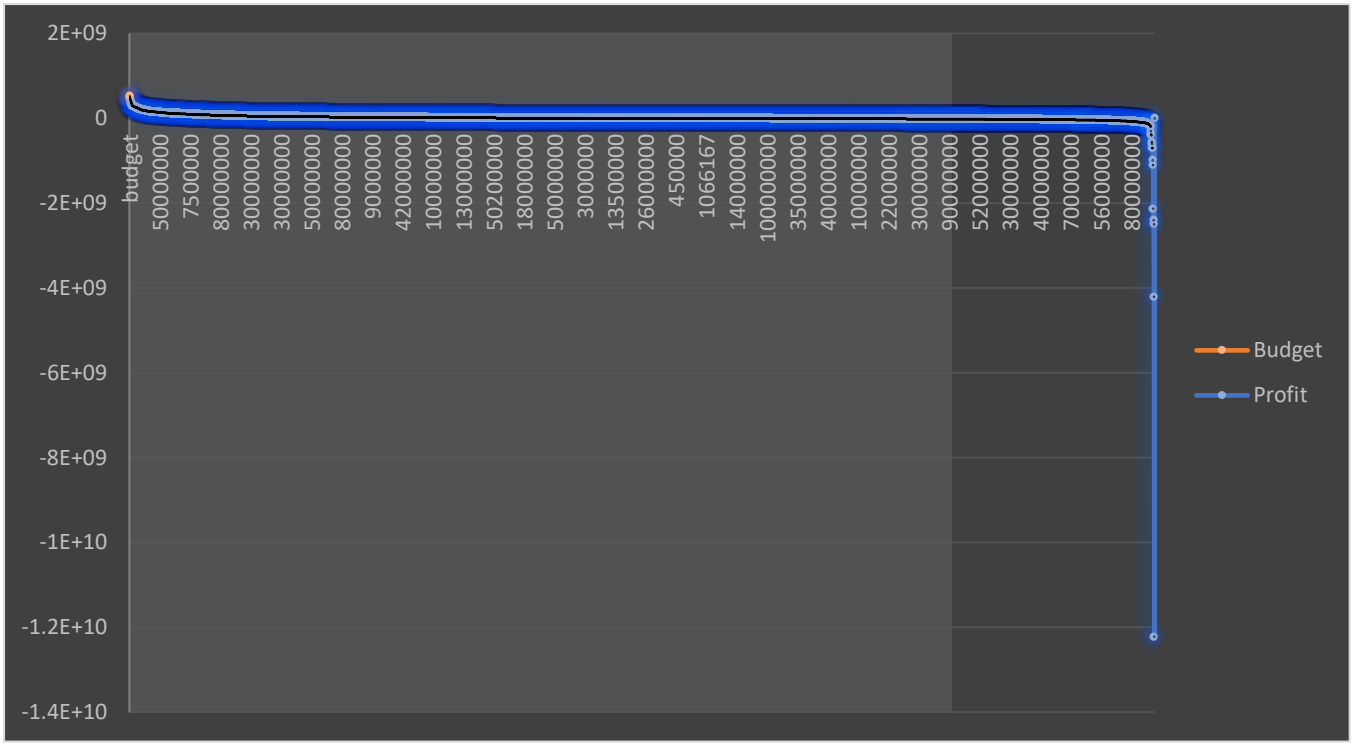
1. First we need to subtract the budget value from the gross value to get the profit.

2. Then, by using the scatter plot option we will plot values of profit(y_axis) and budget(x_axis)

3. Then with the help of graph we will be finding the outliers



The outliers are: -
-12213298588
-4199788333
-2499804112
-2397701809
-2127109510



Top 5 Movies with the Highest Profit are:

genres	actor_1_name	movie_title	num_vo	num_	language	count	conten	budget	title_y	imdb_	Profit
Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	AvatarÂ	886204	3054	English	USA	PG-13	237000000	2009	7.9	523505847
Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic WorldÂ	418214	1290	English	USA	PG-13	150000000	2015	7	502177271
Drama|Romance	Leonardo DiCaprio	TitanicÂ	793059	2528	English	USA	PG-13	200000000	1997	7.7	458672302
Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New HopeÂ	911097	1470	English	USA	PG	11000000	1977	8.7	449935665
Family|Sci-Fi	Henry Thomas	E.T. the Extra-TerrestrialÂ	281842	515	English	USA	PG	10500000	1982	7.9	424449459

Top 250 Movies

To find the IMDB Top 250 we will:-

1. First we will filter out those rows whose num_voted_users > 25000 using the sort and filter option

2. Then we will arrange the dataset on the basis of imdb_score in descending order

3. Then we will select only the top 250 rows for the further analysis

4. Then we will create a new column rank using the RANK() function and using the formula =RANK(N2,$N$2:$N$251,0)+COUNTIFS($N$2:N2,N2)-1

5. Then we will filter out (unselect 'English') from the language column and we will get the desired output.

Google Drive Link for IMDB top 250:
https://docs.google.com/spreadsheets/d/1EEH4AIGauxo8QH7QuMIomnFWaKuitQq6/edit?usp=share_link&ouid=116077077614362440241&rtpof=true&sd=true

Google Drive Link for IMDB top 250(except English):

https://docs.google.com/spreadsheets/d/1Q6ZbgtwzGIWTKhngLmxgK1iYr_cK1Uyv/edit?usp=share_link&ouid=116077077614362440241&rtpof=true&sd=true

Top 5 IMBD movies (all language) are:

movie_title	language	country	content_rating	title_year	imdb_score	Rank
The Shawshank RedemptionÂ	English	USA	R	1994	9.3	1
The GodfatherÂ	English	USA	R	1972	9.2	2
The Dark KnightÂ	English	USA	PG-13	2008	9	3
The Godfather: Part IIÂ	English	USA	R	1974	9	4
The Lord of the Rings: The Return of the KingÂ	English	USA	PG-13	2003	8.9	5

Top 5 IMBD movies (except English) are:

movie_title	language	country	content_rating	budget	title_year	imdb_score	Rank
The Good, the Bad and the UglyÂ	Italian	Italy	Approved	1200000	1966	8.9	8
City of GodÂ	Portuguese	Brazil	R	3300000	2002	8.7	19
Seven SamuraiÂ	Japanese	Japan	Unrated	2000000	1954	8.7	20
Spirited AwayÂ	Japanese	Japan	PG	19000000	2001	8.6	25
The Lives of OthersÂ	German	Germany	R	2000000	2006	8.5	45

Best Directors

Your task: Find the best directors

To find the best top 10 directors on the basis of mean of imdb_score we will:-

1. First select the imdb_score column of the cleaned dataset

2. Then we will click on pivot table

3. We will add director_name into the series section of the pivot table

4. Then we will add average imdb_score into the values section of the pivot table

5. Then we will first sort the data on the basis of average of imdb_score in descending order and then on the basis of director name alphabetically.

Top 10 director having the highest IMDB mean	
Director Name	mean of Imdb Score
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4
Marius A. Markevicius	8.4

Google Drive Link for Top 10 director having highest IMDB mean:-

https://docs.google.com/spreadsheets/d/1vxH9BNIKkE-0ZZMFwU1SUZN6DPIJ5uTg/edit?usp=share_link&ouid=116077077614362440241&rtpof=true&sd=true

Popular Genres

Your task: Find popular genres

To find the Popular Genres we will:-

1. First select the genres column of the cleaned dataset

2. Then we will go for the pivot table option

3. Then we will Select the genres name as row labels

4. Then we will the values as the count of the number of genres and then sort it in descending order on the basis of count of the number of genres

Top 10 Popular genres are:- Genre name	count
Drama	153
Comedy|Drama|Romance	151
Comedy|Drama	147
Comedy	145
Comedy|Romance	135
Drama|Romance	119
Crime|Drama|Thriller	82
Action|Crime|Thriller	55
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	46

Google Drive Link for Top 10 popular genres:-

https://docs.google.com/spreadsheets/d/1XmBCAaMuYFVW9wAFTP5MdXQ1JwMNrNjh/edit?usp=share_link&ouid=116077077614362440241&rtpof=true&sd=true

Charts

Your task: Find the critic-favourite and audience-favourite actors

To find the critic-favourite and audience-favourite actors we will:-

1. First 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 from the actor_1_name column

2. Then we will append the above 3 created columns into 1 column named actor_1_name_combine

3. Then we will group the 3 columns of critic-favourite and audience-favourite actors

4. Then using the pivot table we will find the average, sum and count of critics favourite and audience-favourite actors

Dataset for actor named Meryl Streep:

genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score
Comedy|Drama|Romance	Meryl Streep	It's ComplicatedÂ	69860	214	English	USA	R	85000000	2009	6.6
Action|Adventure|Crime|Thriller	Meryl Streep	The River WildÂ	32544	69	English	USA	PG-13	45000000	1994	6.3
Biography|Drama|Romance	Meryl Streep	Julie & JuliaÂ	79264	277	English	USA	PG-13	40000000	2009	7
Comedy|Drama|Romance	Meryl Streep	The Devil Wears PradaÂ	286178	631	English	USA	PG-13	35000000	2006	6.8
Drama|Thriller|War	Meryl Streep	Lions for LambsÂ	41170	298	English	USA	R	35000000	2007	6.2
Biography|Drama|Romance	Meryl Streep	Out of AfricaÂ	52339	200	English	USA	PG	31000000	1985	7.2
Comedy|Drama|Romance	Meryl Streep	Hope SpringsÂ	34258	178	English	USA	PG-13	30000000	2012	6.3
Drama	Meryl Streep	One True ThingÂ	9283	112	English	USA	R	30000000	1998	7
Drama|Romance	Meryl Streep	The HoursÂ	102123	660	English	USA	PG-13	25000000	2002	7.6
Biography|Drama|History	Meryl Streep	The Iron LadyÂ	82327	350	English	UK	PG-13	13000000	2011	6.4
Comedy|Drama|Music	Meryl Streep	A Prairie Home CompanionÂ	19655	280	English	USA	PG-13	10000000	2006	6.8

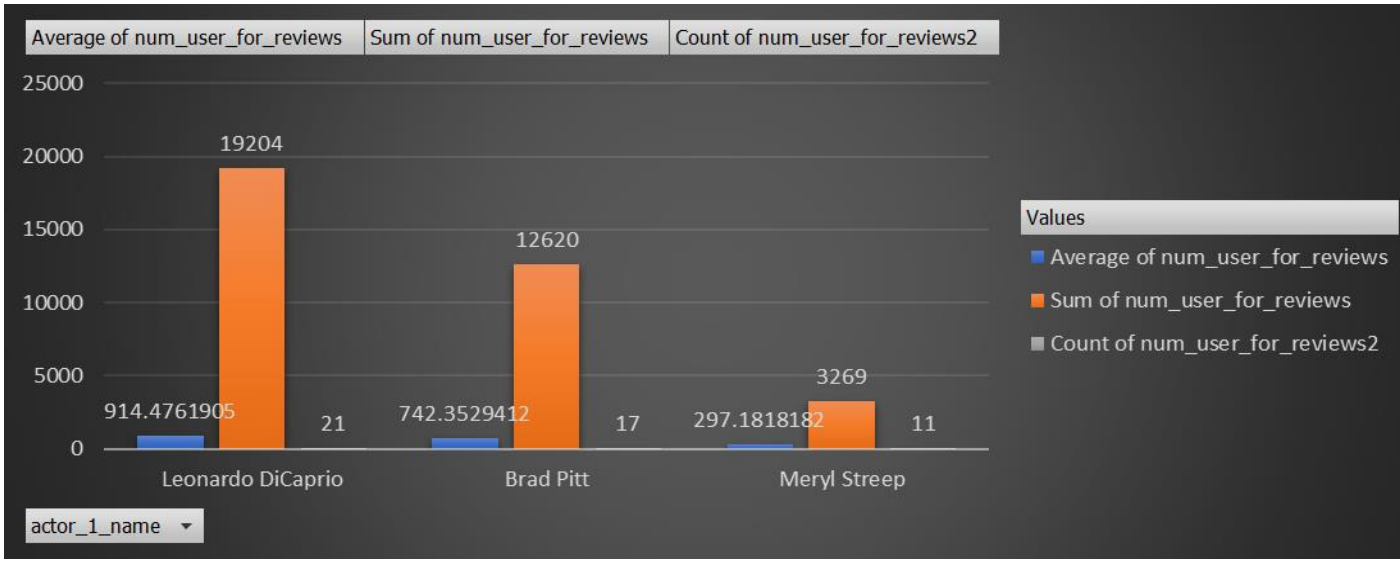
Dataset for actor named Leonardo DiCaprio:

actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score
Leonardo DiCaprio	TitanicÂ	793059	2528	English	USA	PG-13	200000000	1997	7.7
Leonardo DiCaprio	The Great GatsbyÂ	362912	753	English	Australia	PG-13	105000000	2013	7.3
Leonardo DiCaprio	InceptionÂ	1468200	2803	English	USA	PG-13	160000000	2010	8.8
Leonardo DiCaprio	The RevenantÂ	406020	1188	English	USA	R	135000000	2015	8.1
Leonardo DiCaprio	The AviatorÂ	264318	799	English	USA	PG-13	110000000	2004	7.5
Leonardo DiCaprio	Django UnchainedÂ	955174	1193	English	USA	R	100000000	2012	8.5
Leonardo DiCaprio	Blood DiamondÂ	400292	657	English	Germany	R	100000000	2006	8
Leonardo DiCaprio	The Wolf of Wall StreetÂ	780588	1138	English	USA	R	100000000	2013	8.2
Leonardo DiCaprio	Gangs of New YorkÂ	314033	1166	English	USA	R	100000000	2002	7.5
Leonardo DiCaprio	The DepartedÂ	873649	2054	English	USA	R	90000000	2006	8.5
Leonardo DiCaprio	Shutter IslandÂ	786092	964	English	USA	R	80000000	2010	8.1
Leonardo DiCaprio	Body of LiesÂ	174248	263	English	USA	R	70000000	2008	7.1
Leonardo DiCaprio	Catch Me If You CanÂ	525801	667	English	USA	PG-13	52000000	2002	8
Leonardo DiCaprio	The BeachÂ	176169	548	English	USA	R	50000000	2000	6.6
Leonardo DiCaprio	Revolutionary RoadÂ	152591	414	English	USA	R	35000000	2008	7.3
Leonardo DiCaprio	The Man in the Iron MaskÂ	125219	244	English	USA	PG-13	35000000	1998	6.4
Leonardo DiCaprio	J. EdgarÂ	102728	279	English	USA	R	35000000	2011	6.6
Leonardo DiCaprio	The Quick and the DeadÂ	69197	216	English	Japan	R	32000000	1995	6.4
Leonardo DiCaprio	Marvin's RoomÂ	20163	71	English	USA	PG-13	23000000	1996	6.7
Leonardo DiCaprio	Romeo + JulietÂ	167750	506	English	USA	PG-13	14500000	1996	6.8
Leonardo DiCaprio	The Great GatsbyÂ	362933	753	English	Australia	PG-13	105000000	2013	7.3

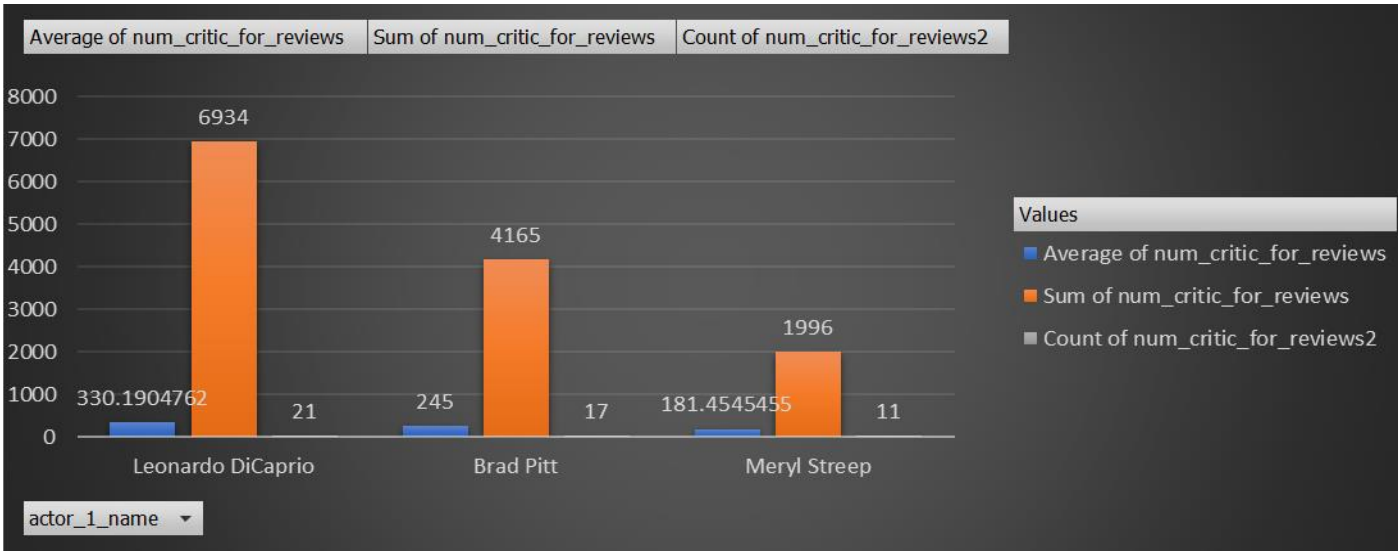
Dataset for actor named Brad Pitt:

director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score
David Fincher	362	127490802	Drama|Fantasy|Romance	Brad Pitt	The Curious Case of Benjamin ButtonÂ	459346	822	English	USA	PG-13	150000000	2008	7.8
Wolfgang Petersen	220	133228348	Adventure	Brad Pitt	TroyÂ	381672	1694	English	USA	R	175000000	2004	7.2
Steven Soderbergh	198	125531634	Crime|Thriller	Brad Pitt	Ocean's TwelveÂ	284852	627	English	USA	PG-13	110000000	2004	6.4
Doug Liman	233	186336103	Action|Comedy|Crime|Romance|Thriller	Brad Pitt	Mr. & Mrs. SmithÂ	348861	798	English	USA	PG-13	120000000	2005	6.5
Tony Scott	142	26871	Action|Crime|Thriller	Brad Pitt	Spy GameÂ	121259	361	English	Germany	R	92000000	2001	7
Steven Soderbergh	186	183405771	Crime|Thriller	Brad Pitt	Ocean's ElevenÂ	402645	845	English	USA	PG-13	85000000	2001	7.8
David Ayer	406	85707116	Action|Drama|War	Brad Pitt	FuryÂ	303185	701	English	USA	R	68000000	2014	7.6
Jean-Jacques Annaud	76	37901509	Adventure|Biography|Drama|History|War	Brad Pitt	Seven Years in TibetÂ	96385	119	English	USA	PG-13	70000000	1997	7
David Fincher	315	37023395	Drama	Brad Pitt	Fight ClubÂ	1347461	2968	English	USA	R	63000000	1999	8.8
Patrick Gilmore	98	26288320	Adventure|Animation|Comedy|Drama|Family|Fantasy|Romance	Brad Pitt	Sinbad: Legend of the Seven SeasÂ	36144	91	English	USA	PG	60000000	2003	6.7
Neil Jordan	120	105264608	Drama|Fantasy|Horror	Brad Pitt	Interview with the Vampire: The Vampire ChroniclesÂ	239752	406	English	USA	R	60000000	1994	7.6
Terrence Malick	584	13303319	Drama|Fantasy	Brad Pitt	The Tree of LifeÂ	136367	975	English	USA	PG-13	32000000	2011	6.7
Andrew Dominik	273	3904982	Biography|Crime|Drama|History|Western	Brad Pitt	The Assassination of Jesse James by the Coward Robert FordÂ	136104	415	English	USA	R	30000000	2007	7.5
Alejandro G. IÃ±Ã¡rritu	285	34300771	Drama	Brad Pitt	BabelÂ	243799	908	English	France	R	25000000	2006	7.5
Angelina Jolie Pitt	131	531009	Drama|Romance	Brad Pitt	By the SeaÂ	7976	61	English	USA	R	10000000	2015	5.3
Andrew Dominik	414	14938570	Crime|Thriller	Brad Pitt	Killing Them SoftlyÂ	111625	369	English	USA	R	15000000	2012	6.2
Tony Scott	122	12281500	Action|Crime|Drama|Romance|Thriller	Brad Pitt	True RomanceÂ	163492	460	English	USA	R	13000000	1993	8

Row Labels	Average of num_user_for_reviews	Sum of num_user_for_reviews	Count of num_user_for_reviews2
Leonardo DiCaprio	914.4761905	19204	21
Brad Pitt	742.3529412	12620	17
Meryl Streep	297.1818182	3269	11
Grand Total	**716.1836735**	**35093**	**49**

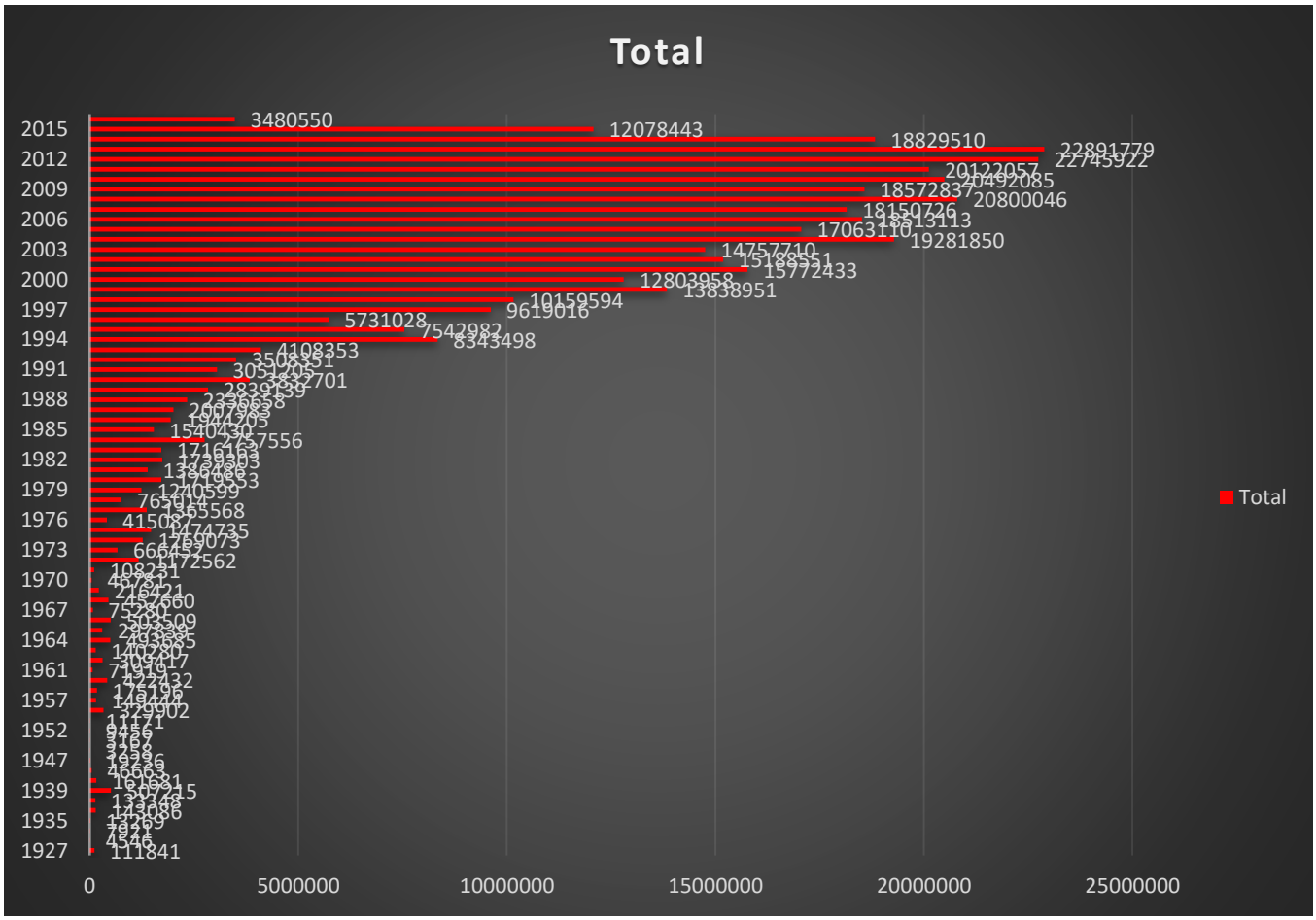


Row Labels	Average of num_critic_for_reviews	Sum of num_critic_for_reviews	Count of num_critic_for_reviews2
Leonardo DiCaprio	330.1904762	6934	21
Brad Pitt	245	4165	17
Meryl Streep	181.4545455	1996	11
Grand Total	**267.244898**	**13095**	**49**

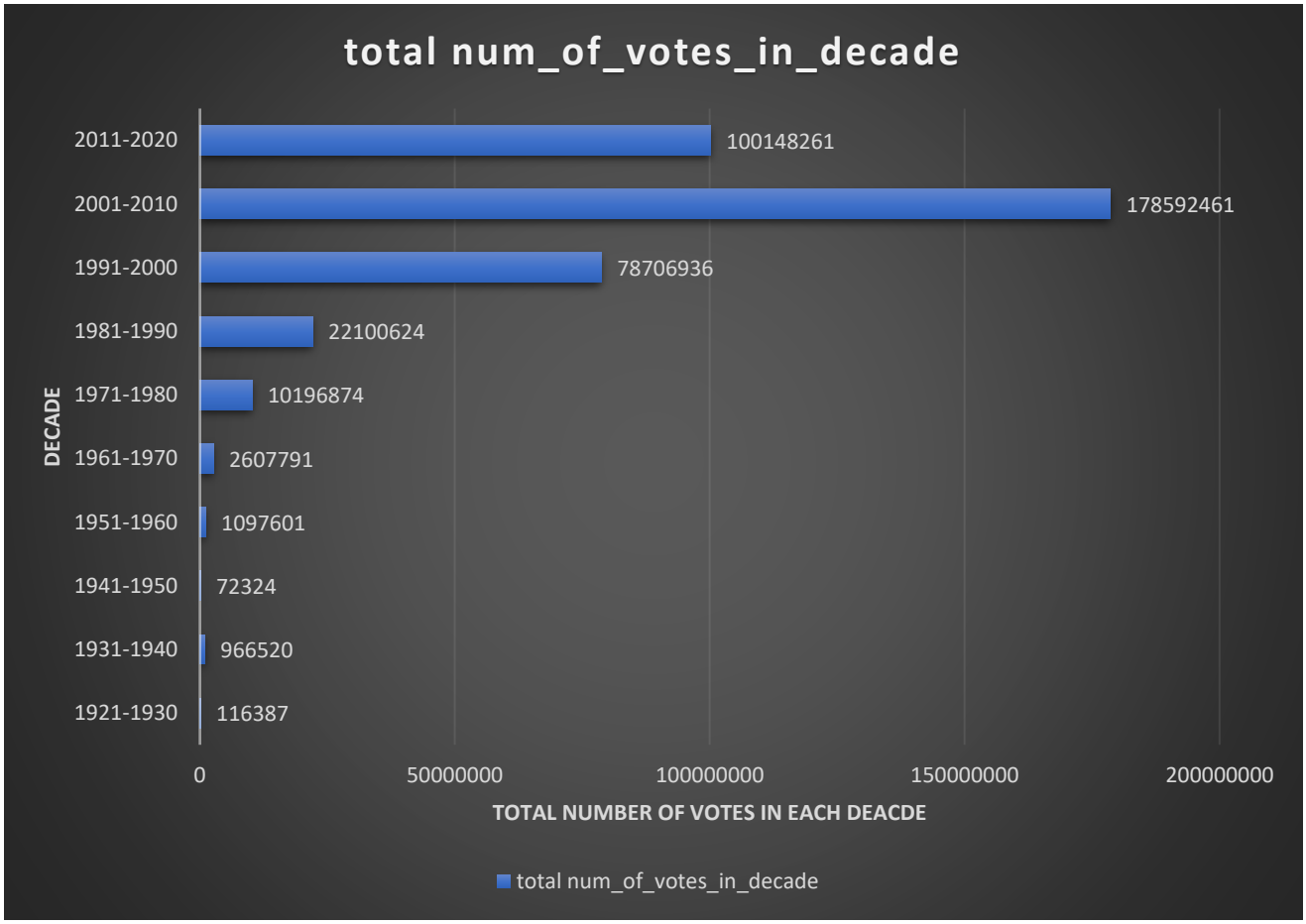


From the above two graphs we can infer that 'Leonardo DiCaprio' was both critics favourite and audience-favourite.

<u>Change in number of voted users over decades using a bar chart</u>



Decade	total num_of_votes_in_decade
1921-1930	116387
1931-1940	966520
1941-1950	72324
1951-1960	1097601
1961-1970	2607791
1971-1980	10196874
1981-1990	22100624
1991-2000	78706936
2001-2010	178592461
2011-2020	100148261
Total	394605779



Google Drive link for highest mean of voted users:

https://docs.google.com/spreadsheets/d/1Hv2KdQVYSeJsMjHjmlejd-xnyimod4Eq/edit?usp=sharing&ouid=116077077614362440241&rtpof=true&sd=true

Google Drive link for number of voted users over decades:

https://docs.google.com/spreadsheets/d/1acfRVH3IPx-tEXC0g7MccvXmml5L_2so/edit?usp=share_link&ouid=116077077614362440241&rtpof=true&sd=true

In this task all the concepts regarding to Excel and statistics like sort, filter, pivot-table, etc. have been implemented using Microsoft Excel.

Google Drive link for all excel sheets:

https://drive.google.com/drive/folders/19LQWasNdlvKJ8eZx-M54kbY3pjHP6I_u?usp=share_link