

Name : ADVAIT GURUNATH CHAVAN

Email : advaitchavan135@gmail.com

Task 5: IMDB Movies Analysis (Final Project - 1), Tech Stack Used: Microsoft 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

Analysis done on the following points:- (Continue)

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-favorite and audience-favorite actors

Cleaning the data

Your task: Clean 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 need to be dropped.
3. After dropping the irrelevant columns now we need to remove the rows from the dataset having any 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 the cleaned dataset

[TASK 5 A IMDB Movies Cleaned Data.xlsx - Google Sheets](#)

Movies with highest profit

Your task: Find the movies with the 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

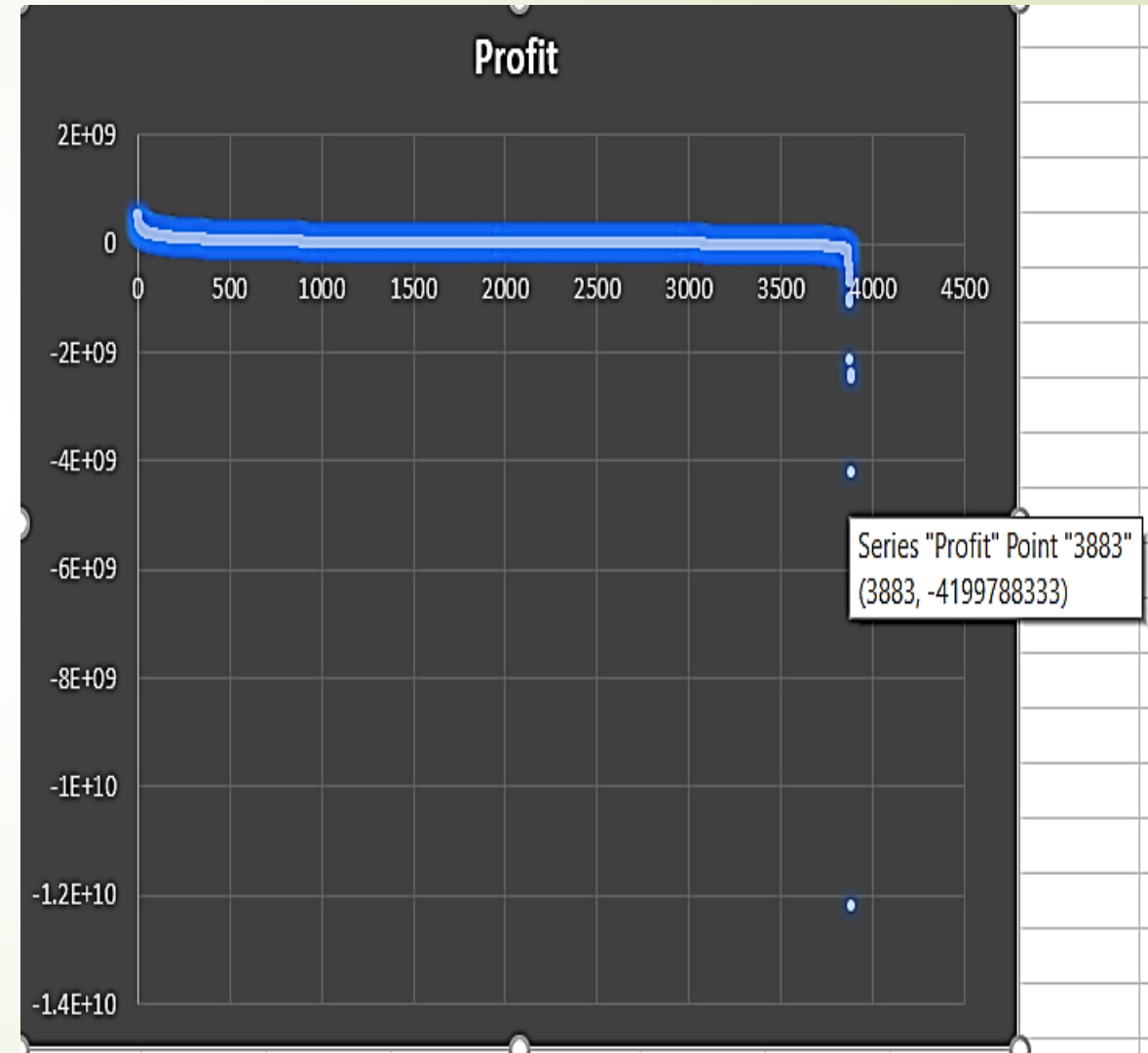
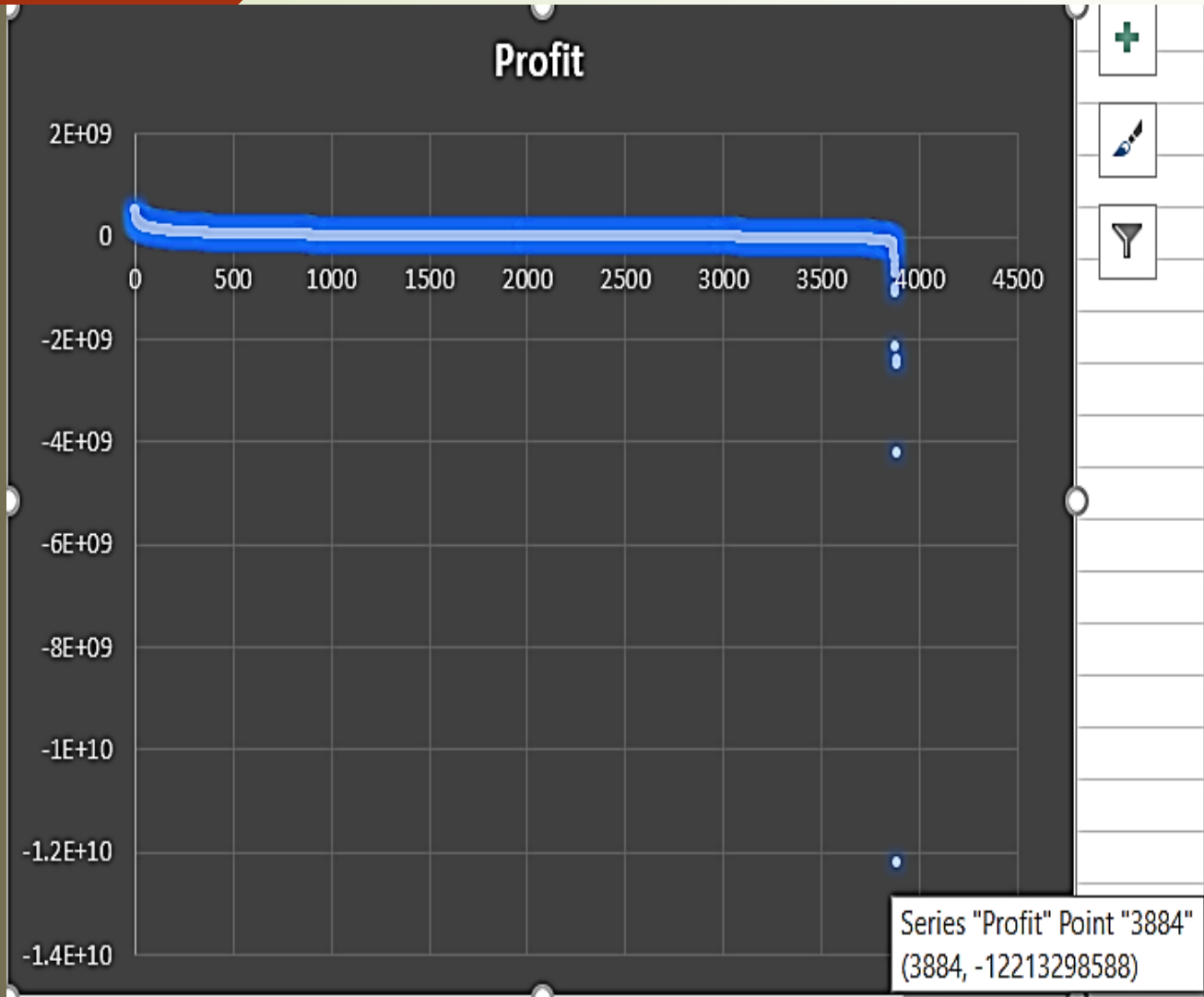
-2499804112

-2127109510

-1099560838

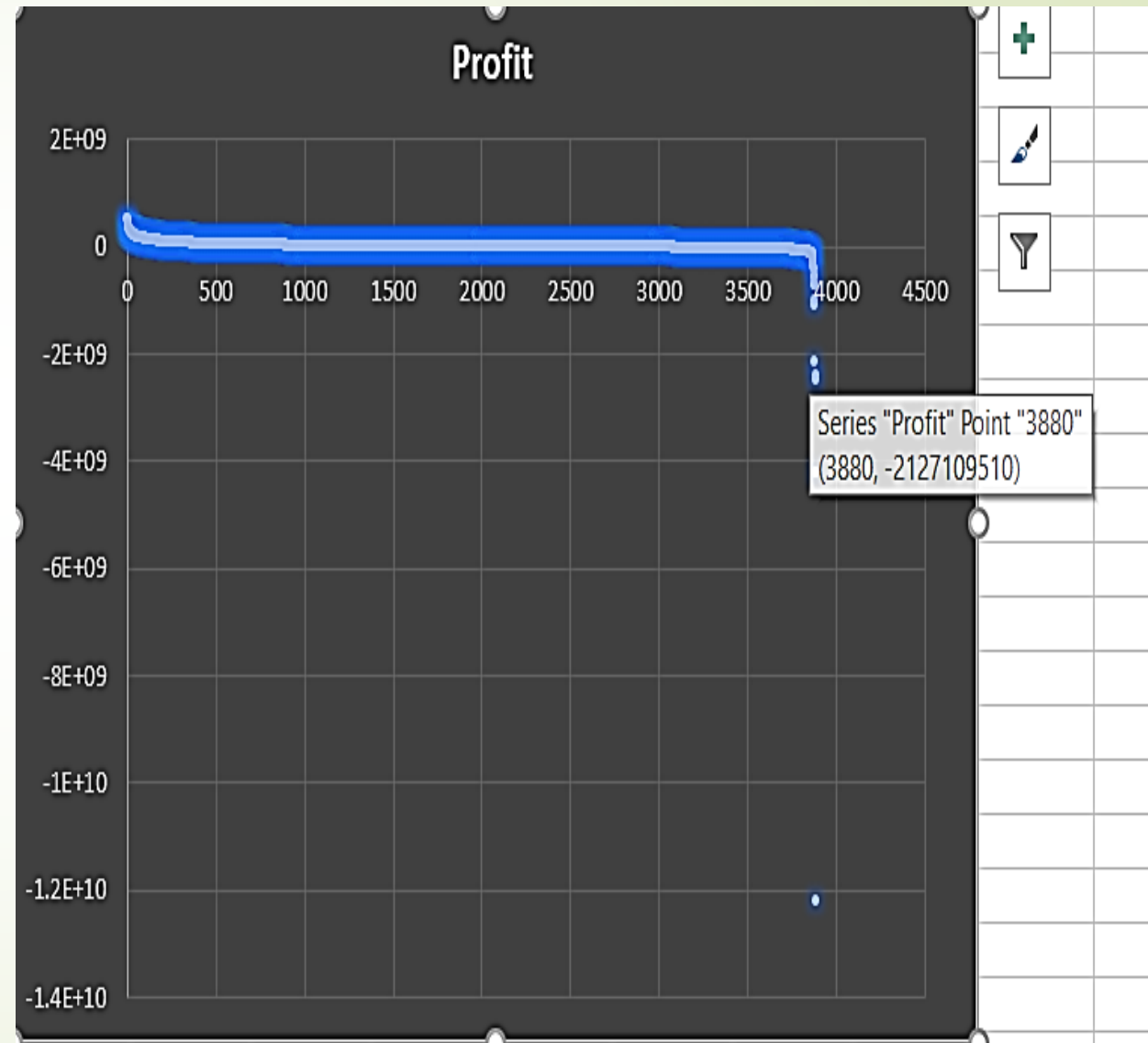
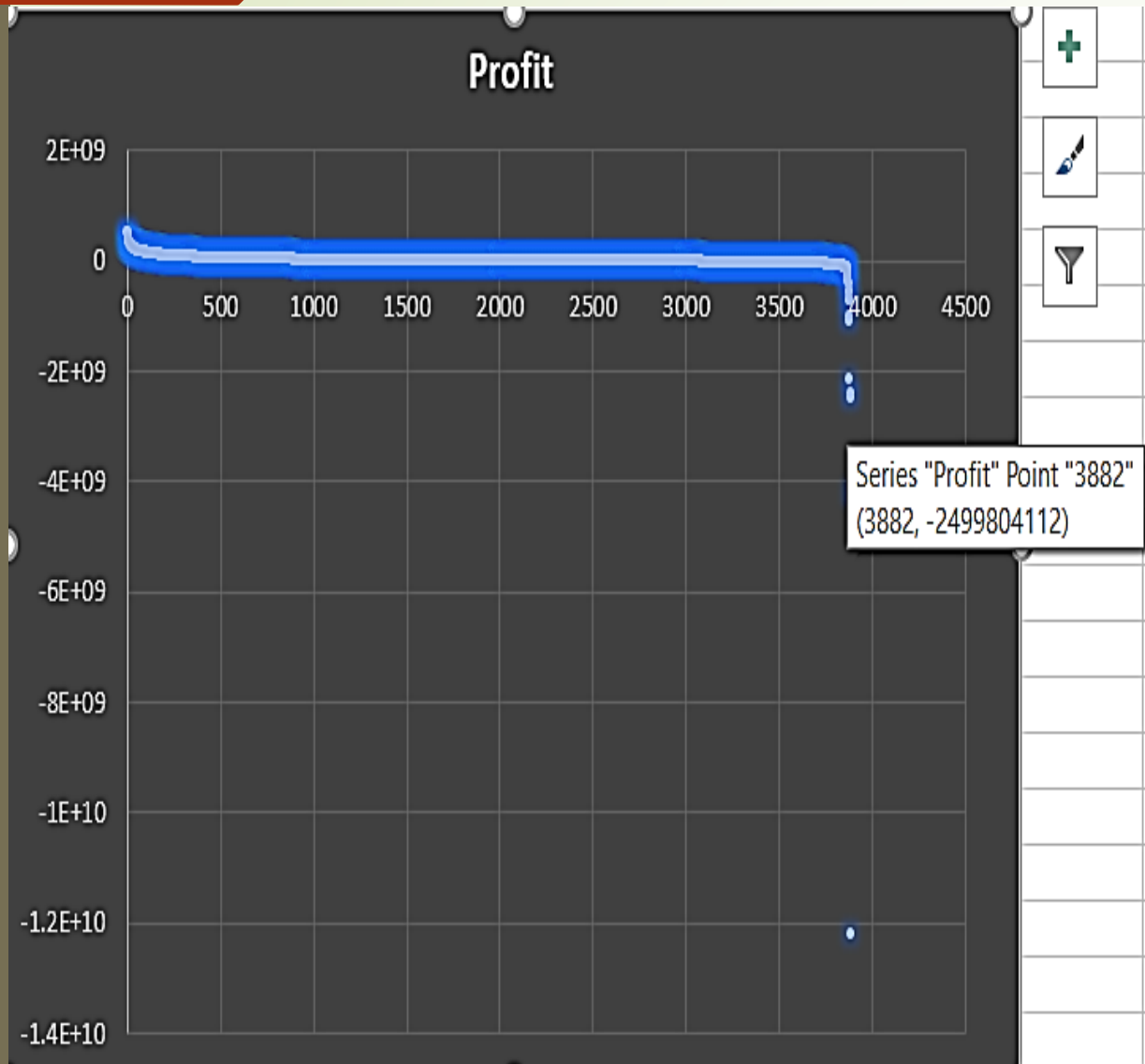
Movies with highest profit

Your task: Find the movies with the highest profit?



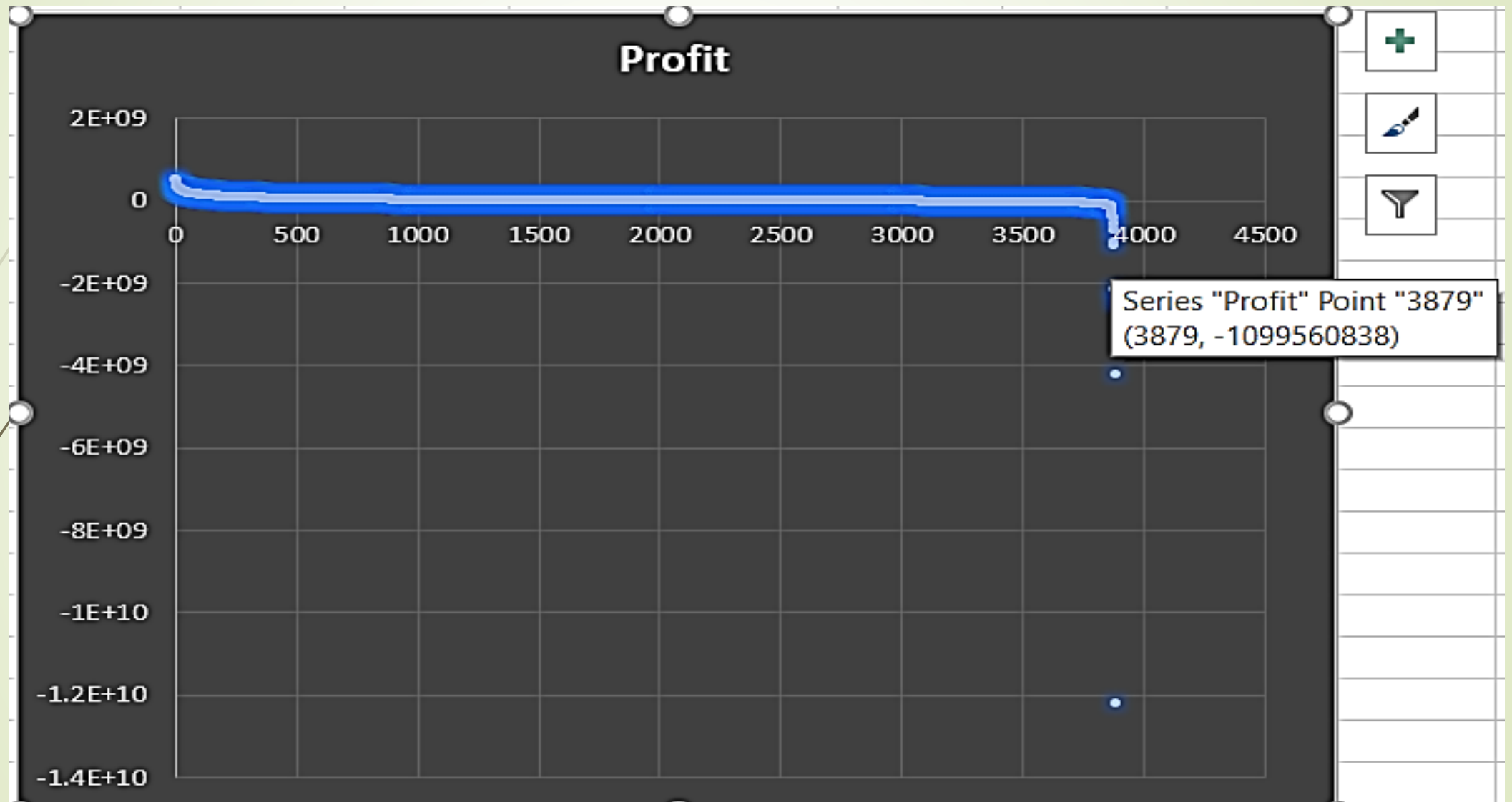
Movies with highest profit

Your task: Find the movies with the highest profit?



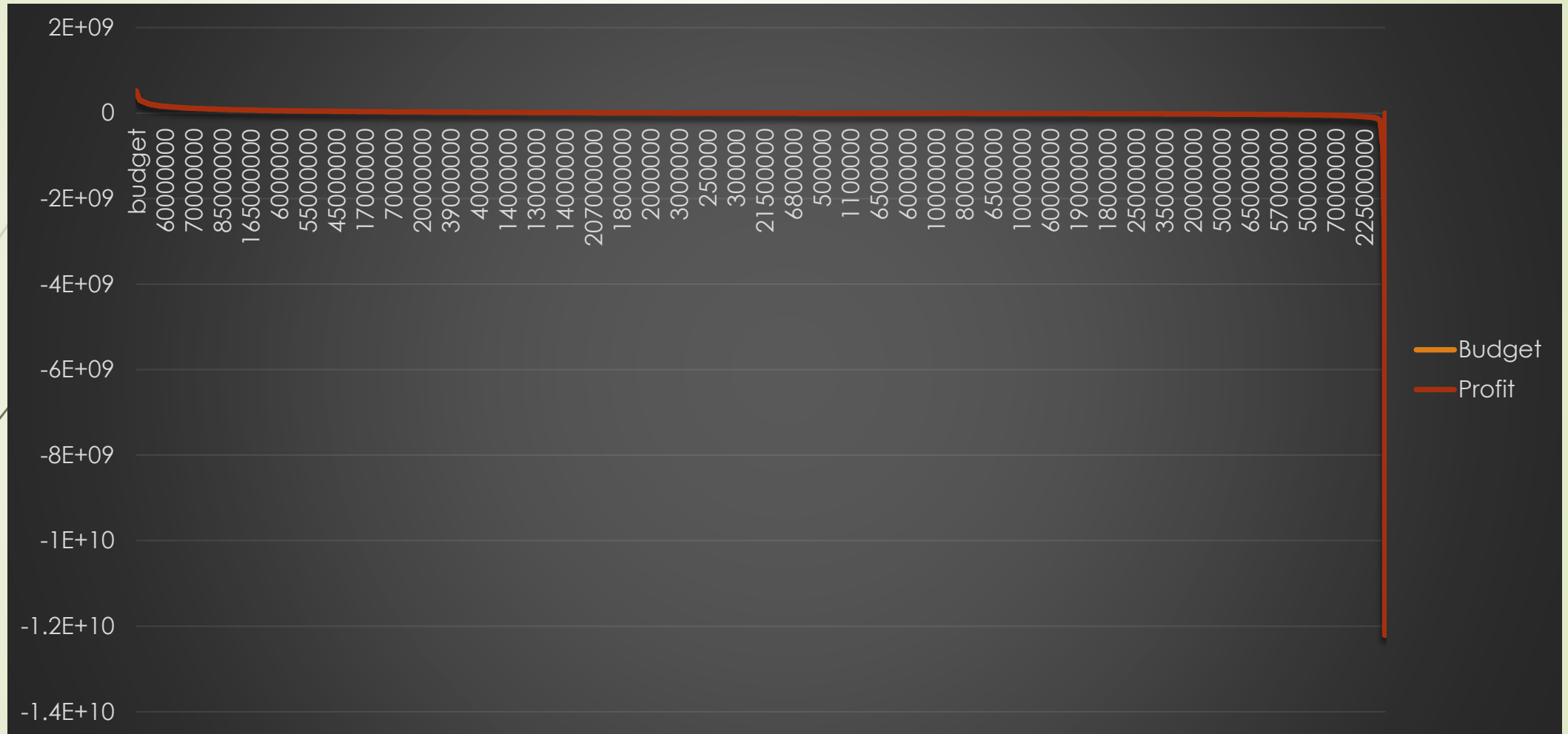
Movies with highest profit

Your task: Find the movies with the highest profit?



Movies with highest profit

Your task: Find the movies with the highest profit?



**Google Drive link for
the excel sheet**

[TASK 5 B IMDB Movies Profit Highest.xlsx - Google Sheets](#)

Movies with highest profit

Your task: Find the movies with the highest profit?

Top 5 Movies with the Highest Profit are

director_name	num_critics_for_review	gross	genres	actor_1_name	movie_title	num_voted_users	num_users_for_review	language	country	content_rating	budget	title_year	imdb_score	Profit
James Cameron	723	760505847	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	886204	3054	English	USA	PG-13	237000000	2009	7.9	523505847
Colin Trevorrow	644	652177271	Action Adventure Sci-Fi Thriller	Bryce Dallas Howard	Jurassic World	418214	1290	English	USA	PG-13	150000000	2015	7.5	2177271
James Cameron	315	658672302	Drama Romance	Leonardo DiCaprio	Titanic	793059	2528	English	USA	PG-13	200000000	1997	7.7	458672302
George Lucas	282	460935665	Action Adventure Fantasy Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New Hope	911097	1470	English	USA	PG	11000000	1977	8.7	449935665
Steven Spielberg	215	434949459	Family Sci-Fi	Henry Thomas	E.T. the Extra-Terrestrial	281842	515	English	USA	PG	10500000	1982	7.9	424449459

Top 250

Your task: Find IMDB Top 250

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
the excel sheet for
imdb_top_250**

**Google Drive link for
the excel sheet for
imdb_top_250_non_e
nglish**

[Task C Top 250 imdb movies.xlsx - Google Sheets](#)

[Task C Top 250 imdb movies non-english.xlsx - Google Sheets](#)

Top 250

Your task: Find IMDB Top 250

Top 5 IMDB Movies(all language) are:-

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	Rank
Frank Darabont	199	28341469	Crime Drama	Morgan Freeman	The Shawshank Redemption	1689764	4144	English	USA	R	25000000	1994	9.3	1
Francis Ford Coppola	208	134821952	Crime Drama	Al Pacino	The Godfather	1155770	2238	English	USA	R	6000000	1972	9.2	2
Christopher Nolan	645	533316061	Action Crime Drama Thriller	Christian Bale	The Dark Knight	1676169	4667	English	USA	PG-13	185000000	2008	9	3
Francis Ford Coppola	149	57300000	Crime Drama	Robert De Niro	The Godfather: Part II	790926	650	English	USA	R	13000000	1974	9	4
Peter Jackson	328	377019252	Action Adventure Drama Fantasy	Orlando Bloom	The Lord of the Rings: The Return of the King	1215718	3189	English	USA	PG-13	94000000	2003	8.9	5

Top 5 IMDB Movies(all languages(except English)) are:-

director_name	num_critic_for_review	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score	Rank
Sergio Leone	181	6100000	Western	Clint Eastwood	The Good, the Bad and the Ugly	503509	780	Italian	Italy	Approved	1200000	1966	8.9	8
Fernando Meirelles	214	7563397	Crime Drama	Alice Braga	City of God	533200	749	Portuguese	Brazil	R	3300000	2002	8.7	19
Akira Kurosawa	153	269061	Action Adventure Drama	Takashi Shimura	Seven Samurai	229012	596	Japanese	Japan	Unrated	2000000	1954	8.7	20
Hayao Miyazaki	246	10049886	Adventure Animation Family Fantasy	Bunta Sugawara	Spirited Away	417971	902	Japanese	Japan	PG	19000000	2001	8.6	25
Florian Henckel von Donnersmarck	215	11284657	Drama Thriller	Sebastian Koch	The Lives of Others	259379	407	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

Result:

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

Best Directors

Your task: Find the best directors

**Google Drive link for
the excel sheet for
Top_10_directors**

[Task D Top 10 IMDB director mean.xlsx - Google Sheets](#)

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

Result:

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

Popular Genres

Your task: Find popular genres

**Google Drive link for
the excel sheet for
Popular genres**

[Task E Most popular genres.xlsx - Google Sheets](#)

Charts

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

To find the critic-favorite and audience-favorite 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-favorite and audience-favorite actors
4. Then using the pivot table we will find the average, sum and count of critic-favorite and audience-favorite actors

Dataset for actor named Meryl_Streep

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
Nancy Meyers	187	112703470	Comedy Drama Romance	Meryl Streep	It's Complicated	69860	214	English	USA	R	85000000	2009	6.6
Curtis Hanson	42	46815748	Action Adventure Crime Thriller	Meryl Streep	The River Wild	32544	69	English	USA	PG-13	45000000	1994	6.3
Nora Ephron	252	94125426	Biography Drama Romance	Meryl Streep	Julie & Julia	79264	277	English	USA	PG-13	40000000	2009	7
David Frankel	208	124732962	Comedy Drama Romance	Meryl Streep	The Devil Wears Prada	286178	631	English	USA	PG-13	35000000	2006	6.8
Robert Redford	227	14998070	Drama Thriller War	Meryl Streep	Lions for Lambs	41170	298	English	USA	R	35000000	2007	6.2
Sydney Pollack	66	87100000	Biography Drama Romance	Meryl Streep	Out of Africa	52339	200	English	USA	PG	31000000	1985	7.2
David Frankel	234	63536011	Comedy Drama Romance	Meryl Streep	Hope Springs	34258	178	English	USA	PG-13	30000000	2012	6.3
Carl Franklin	64	23209440	Drama	Meryl Streep	One True Thing	9283	112	English	USA	R	30000000	1998	7
Stephen Daldry	174	41597830	Drama Romance	Meryl Streep	The Hours	102123	660	English	USA	PG-13	25000000	2002	7.6
Phyllida Lloyd	331	29959436	Biography Drama History	Meryl Streep	The Iron Lady	82327	350	English	UK	PG-13	13000000	2011	6.4
Robert Altman	211	20338609	Comedy Drama Music	Meryl Streep	A Prairie Home Companion	19655	280	English	USA	PG-13	10000000	2006	6.8

Charts

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

Dataset for actor named Leonardo_DiCaprio

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
James Cameron	315	658672302	Drama Romance	Leonardo DiCaprio	Titanic	793059	2528	English	USA	PG-13	200000000	1997	7.7
Baz Luhrmann	490	144812796	Drama Romance	Leonardo DiCaprio	The Great Gatsby	362912	753	English	Australia	PG-13	105000000	2013	7.3
Christopher Nolan	642	292568851	Action Adventure Sci-Fi Thriller	Leonardo DiCaprio	Inception	1468200	2803	English	USA	PG-13	160000000	2010	8.8
Alejandro G. Iñárritu	556	183635922	Adventure Drama Thriller Western	Leonardo DiCaprio	The Revenant	406020	1188	English	USA	R	135000000	2015	8.1
Martin Scorsese	267	102608827	Biography Drama	Leonardo DiCaprio	The Aviator	264318	799	English	USA	PG-13	110000000	2004	7.5
Quentin Tarantino	765	162804648	Drama Western	Leonardo DiCaprio	Django Unchained	955174	1193	English	USA	R	100000000	2012	8.5
Edward Zwick	166	57366262	Adventure Drama Thriller	Leonardo DiCaprio	Blood Diamond	400292	657	English	Germany	R	100000000	2006	8
Martin Scorsese	606	116866727	Biography Comedy Crime Drama	Leonardo DiCaprio	The Wolf of Wall Street	780588	1138	English	USA	R	100000000	2013	8.2
Martin Scorsese	233	77679638	Crime Drama	Leonardo DiCaprio	Gangs of New York	314033	1166	English	USA	R	100000000	2002	7.5
Martin Scorsese	352	132373442	Crime Drama Thriller	Leonardo DiCaprio	The Departed	873649	2054	English	USA	R	90000000	2006	8.5
Martin Scorsese	490	127968405	Mystery Thriller	Leonardo DiCaprio	Shutter Island	786092	964	English	USA	R	80000000	2010	8.1
Ridley Scott	238	39380442	Action Drama Thriller	Leonardo DiCaprio	Body of Lies	174248	263	English	USA	R	70000000	2008	7.1
Steven Spielberg	194	164435221	Biography Crime Drama	Leonardo DiCaprio	Catch Me If You Can	525801	667	English	USA	PG-13	52000000	2002	8
Danny Boyle	118	39778599	Adventure Drama Thriller	Leonardo DiCaprio	The Beach	176169	548	English	USA	R	50000000	2000	6.6
Sam Mendes	323	22877808	Drama Romance	Leonardo DiCaprio	Revolutionary Road	152591	414	English	USA	R	35000000	2008	7.3
Randall Wallace	83	56876365	Action Adventure	Leonardo DiCaprio	The Man in the Iron Mask	125219	244	English	USA	PG-13	35000000	1998	6.4
Clint Eastwood	392	37304950	Biography Crime Drama	Leonardo DiCaprio	J. Edgar	102728	279	English	USA	R	35000000	2011	6.6
Sam Raimi	63	18636537	Action Thriller Western	Leonardo DiCaprio	The Quick and the Dead	69197	216	English	Japan	R	32000000	1995	6.4
Jerry Zaks	45	12782508	Drama	Leonardo DiCaprio	Marvin's Room	20163	71	English	USA	PG-13	23000000	1996	6.7
Baz Luhrmann	106	46338728	Drama Romance	Leonardo DiCaprio	Romeo + Juliet	167750	506	English	USA	PG-13	14500000	1996	6.8
Baz Luhrmann	490	144812796	Drama Romance	Leonardo DiCaprio	The Great Gatsby	362933	753	English	Australia	PG-13	105000000	2013	7.3

Charts

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

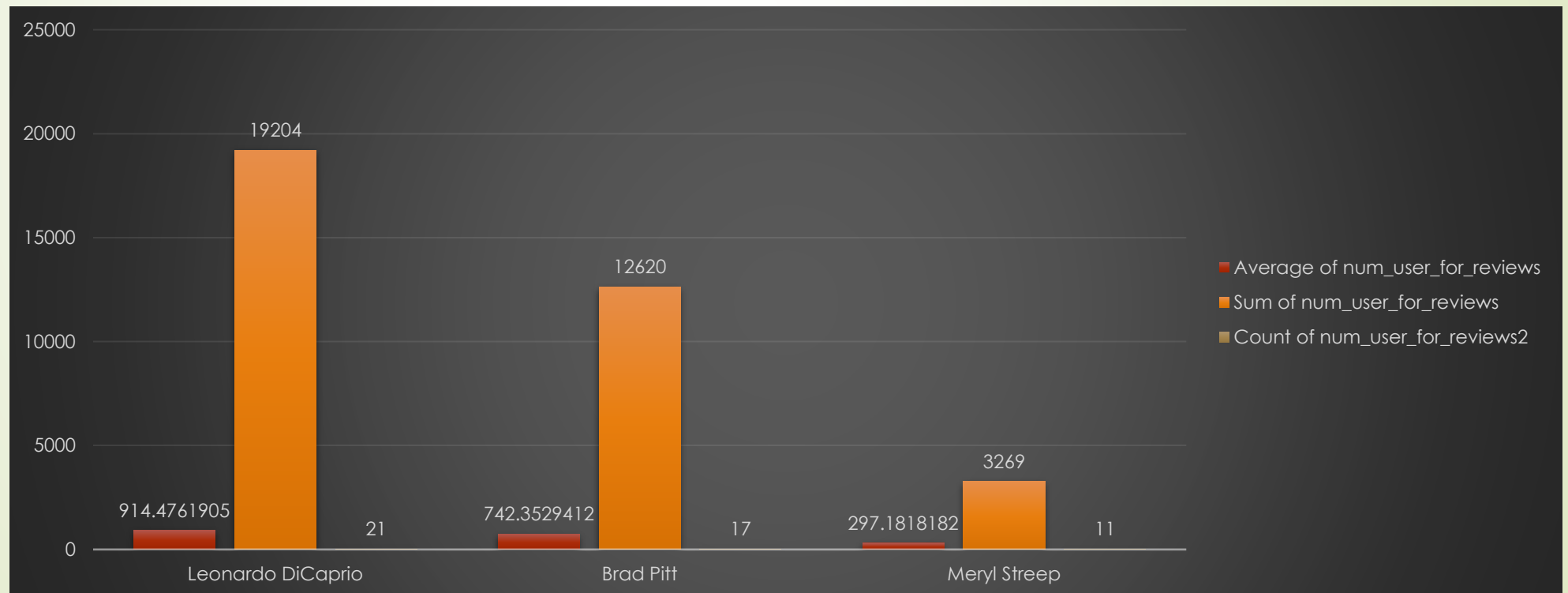
Dataset for actor named Brad_Pitt

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 Button	459346	822	English	USA	PG-13	150000000	2008	7.8
Wolfgang Petersen	220	133228348	Adventure	Brad Pitt	Troy	381672	1694	English	USA	R	175000000	2004	7.2
Steven Soderbergh	198	125531634	Crime Thriller	Brad Pitt	Ocean's Twelve	284852	627	English	USA	PG-13	110000000	2004	6.4
Doug Liman	233	186336103	Action Comedy Crime Romance Thriller	Brad Pitt	Mr. & Mrs. Smith	348861	798	English	USA	PG-13	120000000	2005	6.5
Tony Scott	142	26871	Action Crime Thriller	Brad Pitt	Spy Game	121259	361	English	Germany	R	92000000	2001	7
Steven Soderbergh	186	183405771	Crime Thriller	Brad Pitt	Ocean's Eleven	402645	845	English	USA	PG-13	85000000	2001	7.8
David Ayer	406	85707116	Action Drama War	Brad Pitt	Fury	303185	701	English	USA	R	68000000	2014	7.6
Jean-Jacques Annaud	76	37901509	Adventure Biography Drama History War	Brad Pitt	Seven Years in Tibet	96385	119	English	USA	PG-13	70000000	1997	7
David Fincher	315	37023395	Drama	Brad Pitt	Fight Club	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 Seas	36144	91	English	USA	PG	60000000	2003	6.7
Neil Jordan	120	105264608	Drama Fantasy Horror	Brad Pitt	Interview with the Vampire: The Vampire Chronicles	239752	406	English	USA	R	60000000	1994	7.6
Terrence Malick	584	13303319	Drama Fantasy	Brad Pitt	The Tree of Life	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 Ford	136104	415	English	USA	R	30000000	2007	7.5
Alejandro G. Iñárritu	285	34300771	Drama	Brad Pitt	Babel	243799	908	English	France	R	25000000	2006	7.5
Angelina Jolie Pitt	131	531009	Drama Romance	Brad Pitt	By the Sea	7976	61	English	USA	R	10000000	2015	5.3
Andrew Dominik	414	14938570	Crime Thriller	Brad Pitt	Killing Them Softly	111625	369	English	USA	R	15000000	2012	6.2
Tony Scott	122	12281500	Action Crime Drama Romance Thriller	Brad Pitt	True Romance	163492	460	English	USA	R	13000000	1993	8

Charts

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

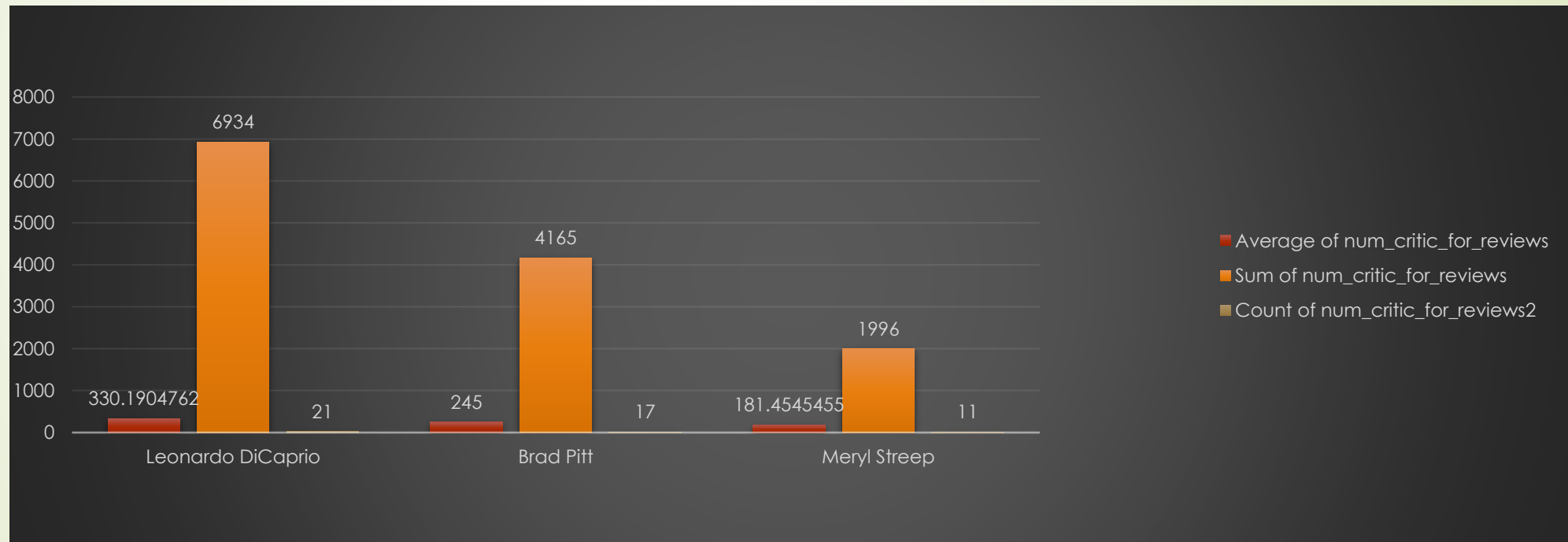
Row Labels	Average of num_user_for_reviews	Sum of num_user_for_reviews	Count of num_user_for_reviews2
Leonardo DiCaprio	914.4761905	19204	21
Brad Pitt	742.3529412	12620	17
Meryl Streep	297.1818182	3269	11
Grand Total	716.1836735	35093	49



Charts

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

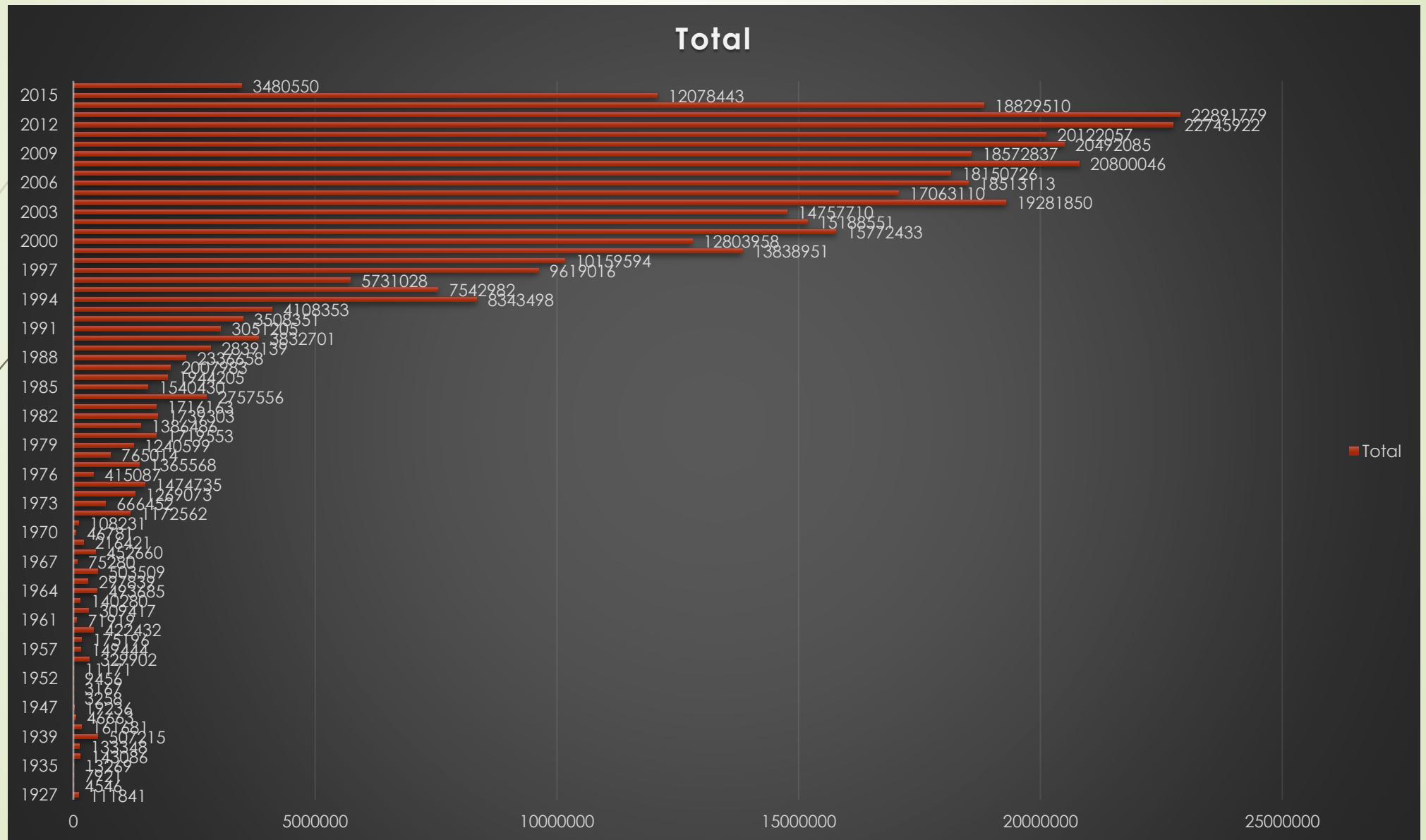
Row Labels	Average of num_critic_for_reviews	Sum of num_critic_for_reviews	Count of num_critic_for_reviews2
Leonardo DiCaprio	330.1904762	6934	21
Brad Pitt	245	4165	17
Meryl Streep	181.4545455	1996	11
Grand Total	267.244898	13095	49



From the above two graphs we can infer that '**Leonardo DiCaprio**' was both critic-favorite and audience-favorite actor

Charts

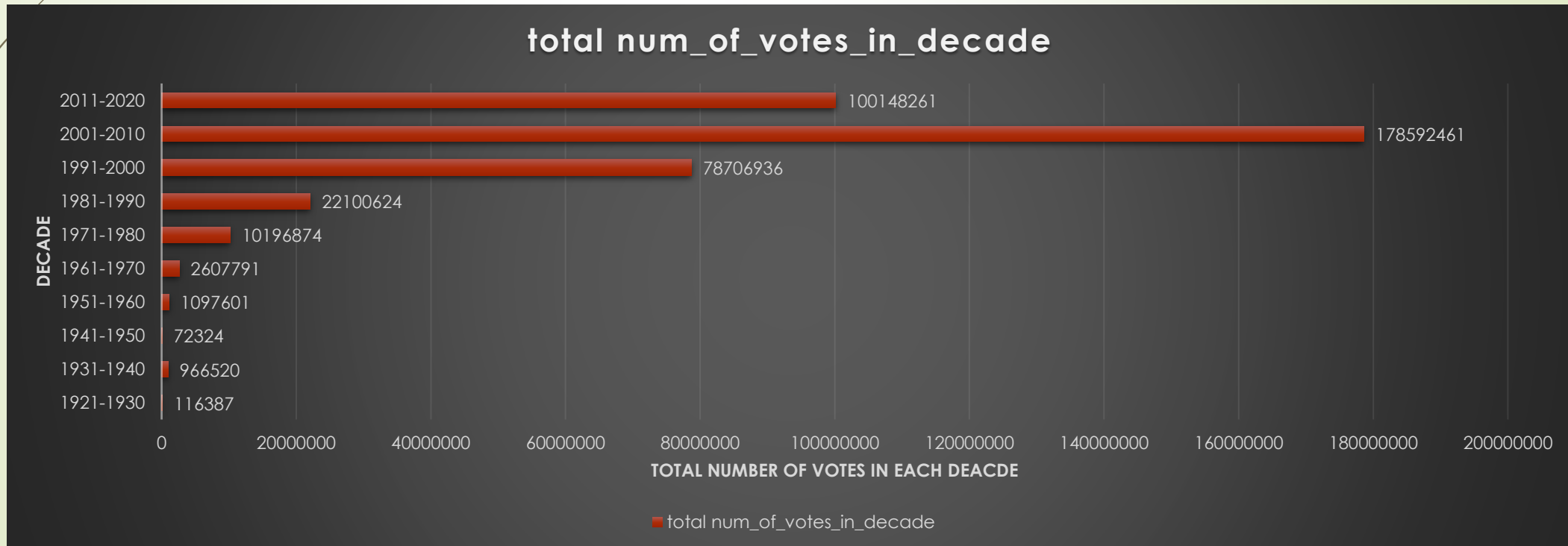
Change in number of voted users over decades using a bar chart



Charts

Change in number of voted users over decades using a bar chart

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



Charts

Change in number of voted users over decades using a bar chart

**Google Drive link for
the excel sheet for
highest mean of
voted users**

[Task F Charts highest mean.xlsx - Google Sheets](#)

**Google Drive link for
the excel sheet for
number of voted
users over decades**

[Task F Charts num of votes by decade.xlsx - Google Sheets](#)



Hence, all the questions given as a part of
Data Analytics Trainee Task 5: IMDB Movie Analysis (Final Project - 1)
have been provided with answers along with graphs

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**

[trainity task 5 final project excel files - Google Drive](#)