NAME-AKSHAY ANAND

PROJECT DESCRIPTION: Dataset having various columns of different IMDB Movies.

APPROACH: WE USED MS EXCEL FOR DATA CLEANING AND FUNCTIONS TO CALCULATE REQUIRED PARAMETERS FROM THE DATASET PROVIDED.

TECH-STACK USED: MS-EXCEL

INSIGHTS: EXPLAINED WITH EACH PLOT.

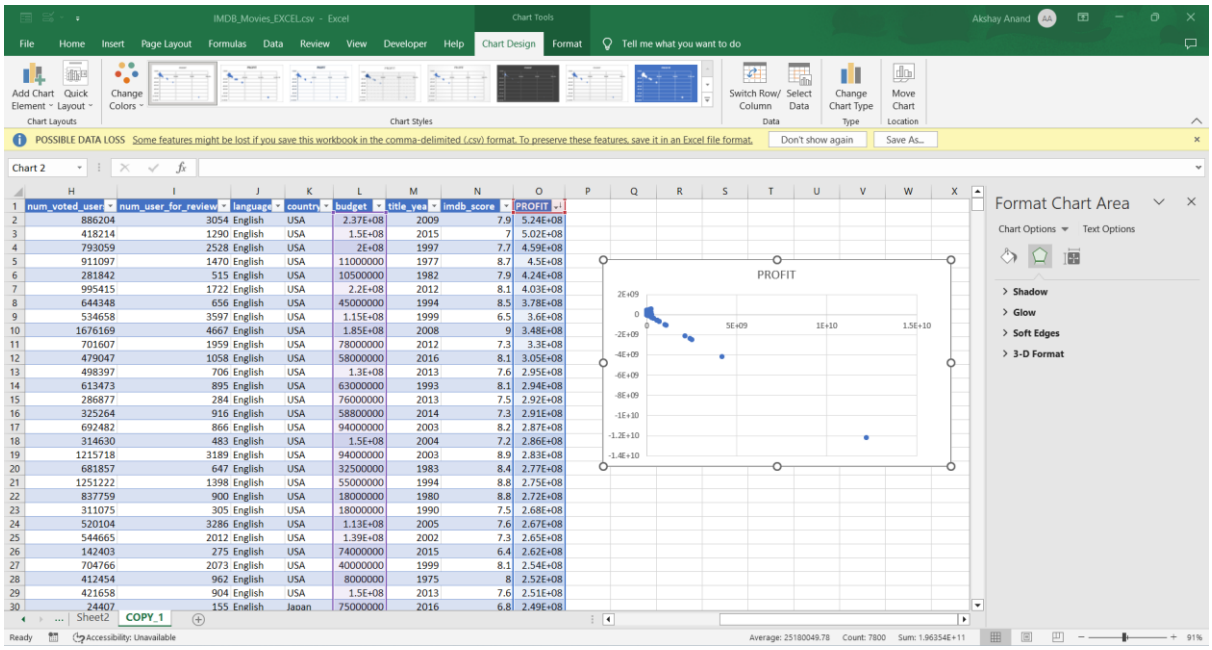
Cleaning the data:: This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
Your task: Clean the data

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 which have been dropped. Also dropped rows having null values. At the end we got rid of duplicate values using Remove Duplicates in data tab.

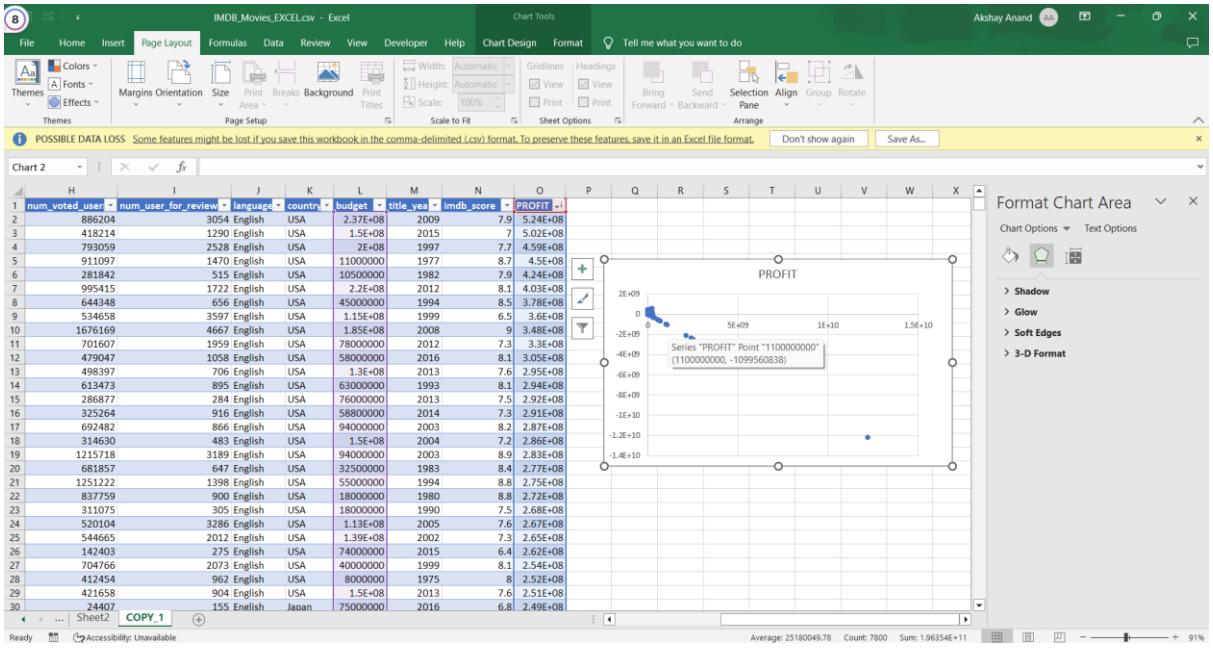
Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference.
Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

OUTLIERS->

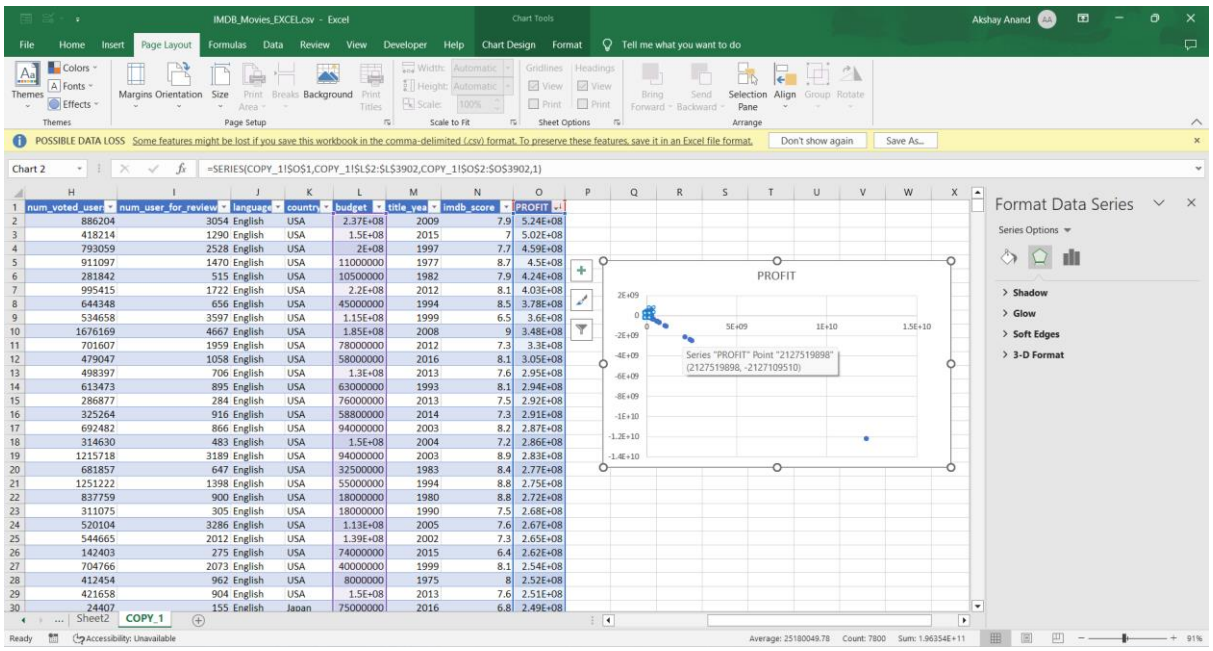
-1099560838, -2127109510, -2499804112, -4199788333, -12213298588



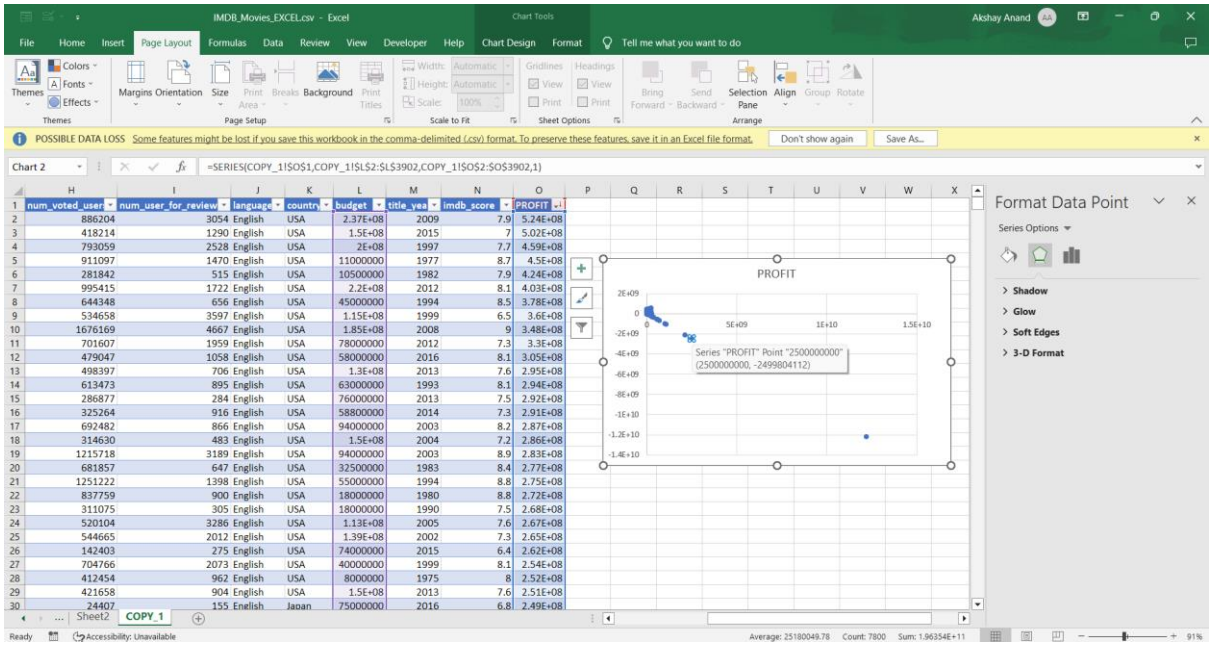
POINTS THAT ARE SIGNIFICANTLY AWAY FROM THE EXPECTED RANGE ARE CALLED OUTLIERS. FROM THE GRAPH WE CAN CLEARLY SEE OUTLIERS.



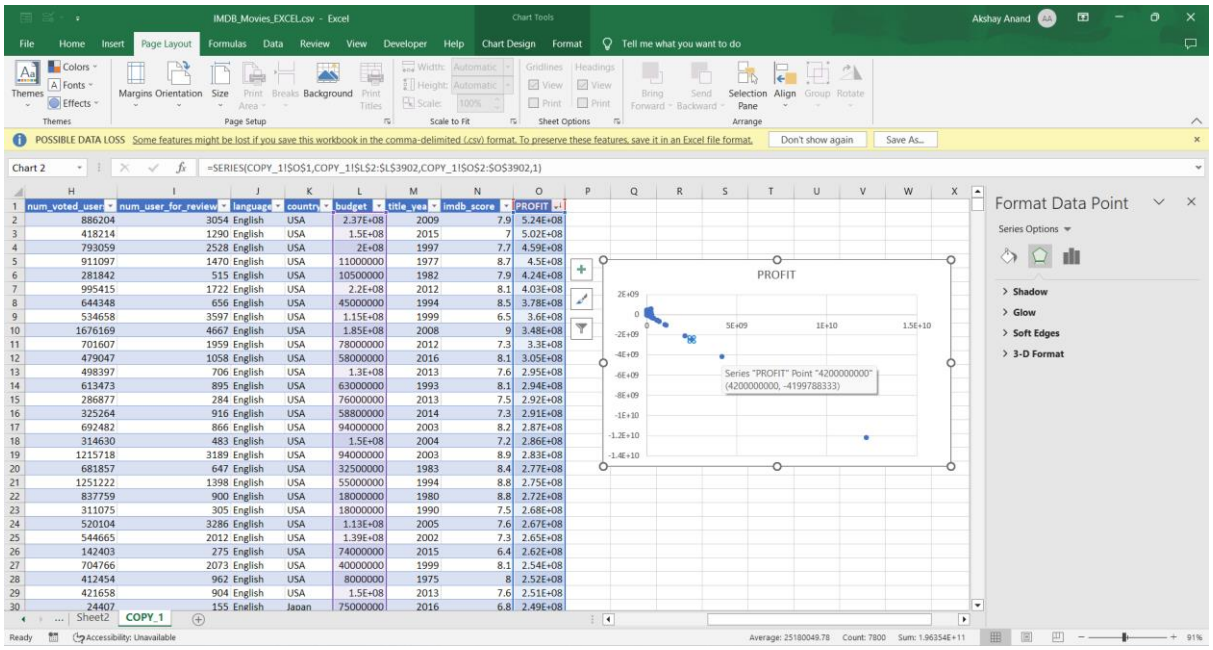
num_voted_user	num_user_for_review	language	countr	budget	title_yea	imdb_score	PROFIT
886204	3054	English	USA	2.37E+08	2009	7.9	5.24E+08
418214	1290	English	USA	1.5E+08	2015	7	5.02E+08
793059	2528	English	USA	2E+08	1997	7.7	4.59E+08
911097	1470	English	USA	11000000	1977	8.7	4.5E+08
281842	515	English	USA	10500000	1982	7.9	4.24E+08
995415	1722	English	USA	2.2E+08	2012	8.1	4.03E+08
644348	656	English	USA	45000000	1994	8.5	3.78E+08
534658	3597	English	USA	1.15E+08	1999	6.5	3.6E+08
1676169	4667	English	USA	1.85E+08	2008	9	3.48E+08
701607	1959	English	USA	78000000	2012	7.3	3.3E+08
479047	1058	English	USA	58000000	2016	8.1	3.05E+08
498397	706	English	USA	1.3E+08	2013	7.6	2.95E+08
613473	895	English	USA	63000000	1993	8.1	2.94E+08
286877	284	English	USA	76000000	2013	7.5	2.92E+08
325264	916	English	USA	58800000	2014	7.3	2.91E+08
692482	866	English	USA	94000000	2003	8.2	2.87E+08
314630	483	English	USA	1.5E+08	2004	7.2	2.86E+08
1215718	3189	English	USA	94000000	2003	8.9	2.83E+08
681857	647	English	USA	32500000	1983	8.4	2.77E+08
1251222	1398	English	USA	55000000	1994	8.8	2.75E+08
837759	900	English	USA	18000000	1980	8.8	2.72E+08
311075	305	English	USA	18000000	1990	7.5	2.68E+08
520104	3286	English	USA	1.13E+08	2005	7.6	2.67E+08
544665	2012	English	USA	1.39E+08	2002	7.3	2.65E+08
142403	275	English	USA	74000000	2015	6.4	2.62E+08
704766	2073	English	USA	40000000	1999	8.1	2.54E+08
412454	962	English	USA	8000000	1975	8	2.52E+08
421658	904	English	USA	1.5E+08	2013	7.6	2.51E+08
24407	155	English	Japan	75000000	2016	6.8	2.49E+08



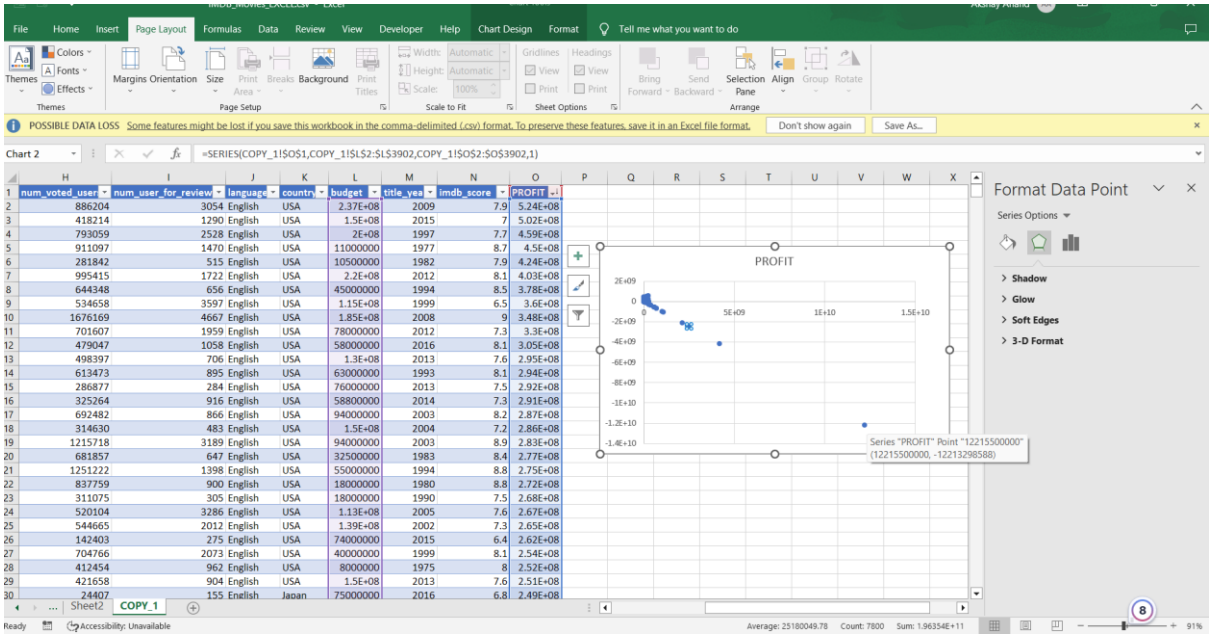
num_voted_user	num_user_for_review	language	countr	budget	title_yea	imdb_score	PROFIT
886204	3054	English	USA	2.37E+08	2009	7.9	5.24E+08
418214	1290	English	USA	1.5E+08	2015	7	5.02E+08
793059	2528	English	USA	2E+08	1997	7.7	4.59E+08
911097	1470	English	USA	11000000	1977	8.7	4.5E+08
281842	515	English	USA	10500000	1982	7.9	4.24E+08
995415	1722	English	USA	2.2E+08	2012	8.1	4.03E+08
644348	656	English	USA	45000000	1994	8.5	3.78E+08
534658	3597	English	USA	1.15E+08	1999	6.5	3.6E+08
1676169	4667	English	USA	1.85E+08	2008	9	3.48E+08
701607	1959	English	USA	78000000	2012	7.3	3.3E+08
479047	1058	English	USA	58000000	2016	8.1	3.05E+08
498397	706	English	USA	1.3E+08	2013	7.6	2.95E+08
613473	895	English	USA	63000000	1993	8.1	2.94E+08
286877	284	English	USA	76000000	2013	7.5	2.92E+08
325264	916	English	USA	58800000	2014	7.3	2.91E+08
692482	866	English	USA	94000000	2003	8.2	2.87E+08
314630	483	English	USA	1.5E+08	2004	7.2	2.86E+08
1215718	3189	English	USA	94000000	2003	8.9	2.83E+08
681857	647	English	USA	32500000	1983	8.4	2.77E+08
1251222	1398	English	USA	55000000	1994	8.8	2.75E+08
837759	900	English	USA	18000000	1980	8.8	2.72E+08
311075	305	English	USA	18000000	1990	7.5	2.68E+08
520104	3286	English	USA	1.13E+08	2005	7.6	2.67E+08
544665	2012	English	USA	1.39E+08	2002	7.3	2.65E+08
142403	275	English	USA	74000000	2015	6.4	2.62E+08
704766	2073	English	USA	40000000	1999	8.1	2.54E+08
412454	962	English	USA	8000000	1975	8	2.52E+08
421658	904	English	USA	1.5E+08	2013	7.6	2.51E+08
24407	155	English	Japan	75000000	2016	6.8	2.49E+08

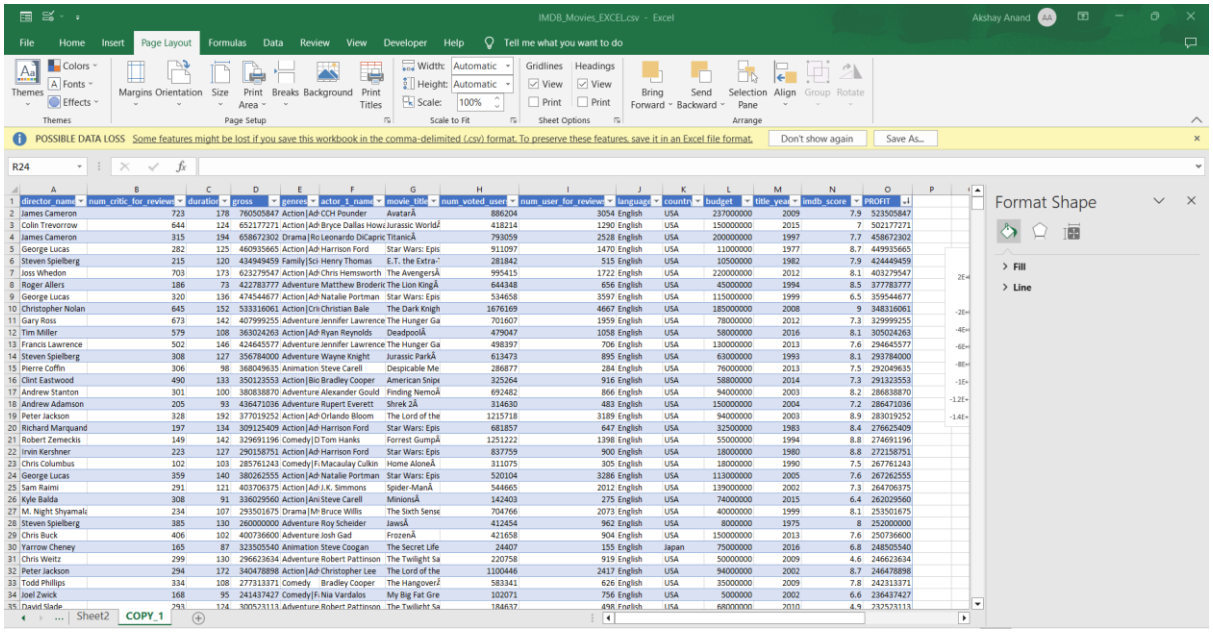


num_voted_user	num_user_for_review	language	countr	budget	title_yea	imdb_score	PROFIT
886204	3054	English	USA	2.37E+08	2009	7.9	5.24E+08
418214	1290	English	USA	1.5E+08	2015	7	5.02E+08
793059	2528	English	USA	2E+08	1997	7.7	4.59E+08
911097	1470	English	USA	11000000	1977	8.7	4.5E+08
281842	515	English	USA	10500000	1982	7.9	4.24E+08
995415	1722	English	USA	2.2E+08	2012	8.1	4.03E+08
644348	656	English	USA	45000000	1994	8.5	3.78E+08
534658	3597	English	USA	1.15E+08	1999	6.5	3.6E+08
1676169	4667	English	USA	1.85E+08	2008	9	3.48E+08
701607	1959	English	USA	78000000	2012	7.3	3.3E+08
479047	1058	English	USA	58000000	2016	8.1	3.05E+08
498397	706	English	USA	1.3E+08	2013	7.6	2.95E+08
613473	895	English	USA	63000000	1993	8.1	2.94E+08
286877	284	English	USA	76000000	2013	7.5	2.92E+08
325264	916	English	USA	58800000	2014	7.3	2.91E+08
692482	866	English	USA	94000000	2003	8.2	2.87E+08
314630	483	English	USA	1.5E+08	2004	7.2	2.86E+08
1215718	3189	English	USA	94000000	2003	8.9	2.83E+08
681857	647	English	USA	32500000	1983	8.4	2.77E+08
1251222	1398	English	USA	55000000	1994	8.8	2.75E+08
837759	900	English	USA	18000000	1980	8.8	2.72E+08
311075	305	English	USA	18000000	1990	7.5	2.68E+08
520104	3286	English	USA	1.13E+08	2005	7.6	2.67E+08
544665	2012	English	USA	1.39E+08	2002	7.3	2.65E+08
142403	275	English	USA	74000000	2015	6.4	2.62E+08
704766	2073	English	USA	40000000	1999	8.1	2.54E+08
412454	962	English	USA	8000000	1975	8	2.52E+08
421658	904	English	USA	1.5E+08	2013	7.6	2.51E+08
24407	155	English	Japan	75000000	2016	6.8	2.49E+08



num_voted_user	num_user_for_review	language	countr	budget	title_yea	imdb_score	PROFIT
886204	3054	English	USA	2.37E+08	2009	7.9	5.24E+08
418214	1290	English	USA	1.5E+08	2015	7	5.02E+08
793059	2528	English	USA	2E+08	1997	7.7	4.59E+08
911097	1470	English	USA	11000000	1977	8.7	4.5E+08
281842	515	English	USA	10500000	1982	7.9	4.24E+08
995415	1722	English	USA	2.2E+08	2012	8.1	4.03E+08
644348	656	English	USA	45000000	1994	8.5	3.78E+08
534658	3597	English	USA	1.15E+08	1999	6.5	3.6E+08
1676169	4667	English	USA	1.85E+08	2008	9	3.48E+08
701607	1959	English	USA	78000000	2012	7.3	3.3E+08
479047	1058	English	USA	58000000	2016	8.1	3.05E+08
498397	706	English	USA	1.3E+08	2013	7.6	2.95E+08
613473	895	English	USA	63000000	1993	8.1	2.94E+08
286877	284	English	USA	76000000	2013	7.5	2.92E+08
325264	916	English	USA	58800000	2014	7.3	2.91E+08
692482	866	English	USA	94000000	2003	8.2	2.87E+08
314630	483	English	USA	1.5E+08	2004	7.2	2.86E+08
1215718	3189	English	USA	94000000	2003	8.9	2.83E+08
681857	647	English	USA	32500000	1983	8.4	2.77E+08
1251222	1398	English	USA	55000000	1994	8.8	2.75E+08
837759	900	English	USA	18000000	1980	8.8	2.72E+08
311075	305	English	USA	18000000	1990	7.5	2.68E+08
520104	3286	English	USA	1.13E+08	2005	7.6	2.67E+08
544665	2012	English	USA	1.39E+08	2002	7.3	2.65E+08
142403	275	English	USA	74000000	2015	6.4	2.62E+08
704766	2073	English	USA	40000000	1999	8.1	2.54E+08
412454	962	English	USA	8000000	1975	8	2.52E+08
421658	904	English	USA	1.5E+08	2013	7.6	2.51E+08
24407	155	English	Japan	75000000	2016	6.8	2.49E+08

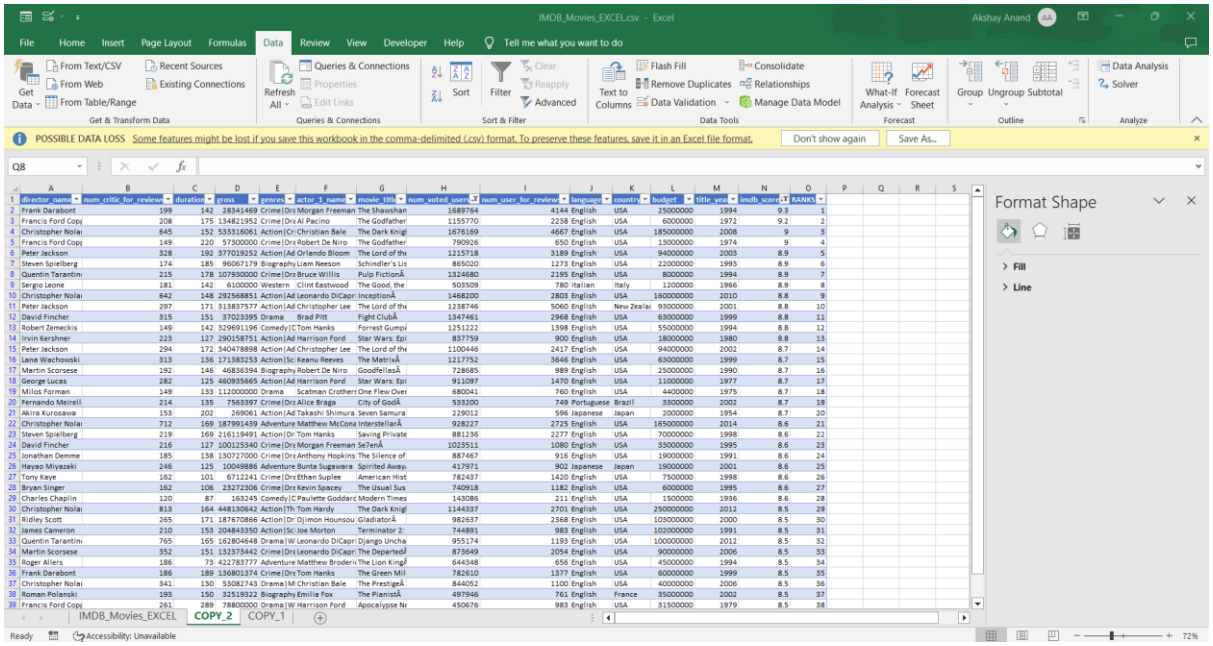




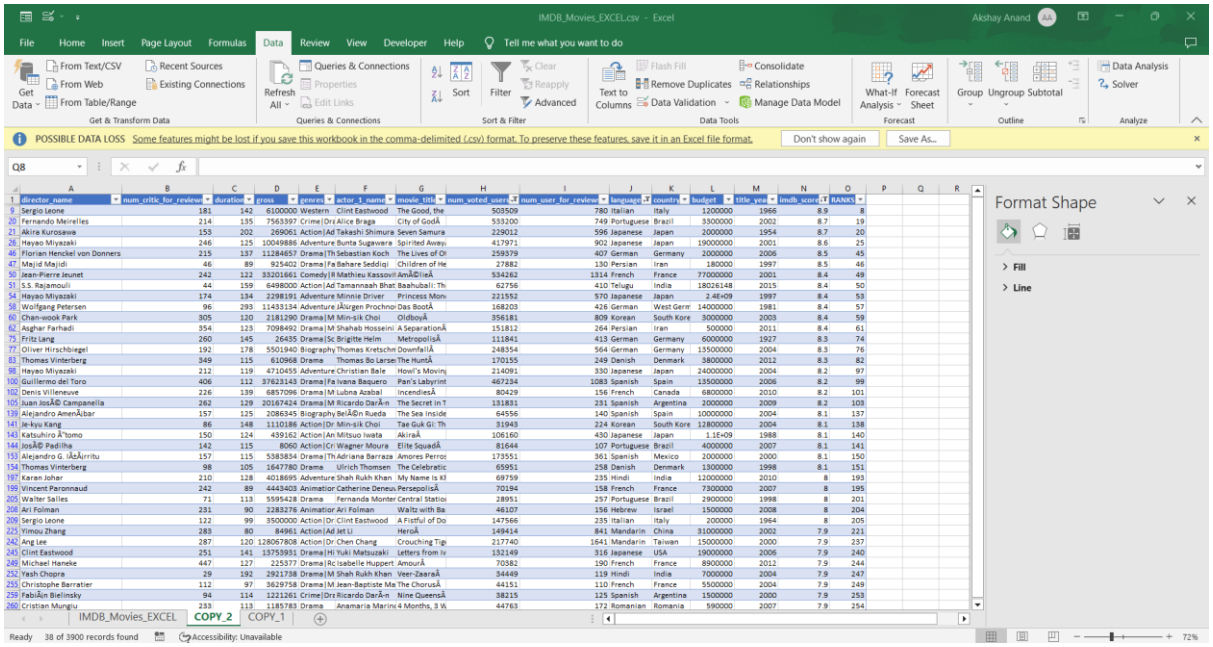
THE MOVIE WITH THE HIGHEST PROFIT IS AvatarÂ.

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

TOP 250



TOP 250 NOT IN ENGLISH



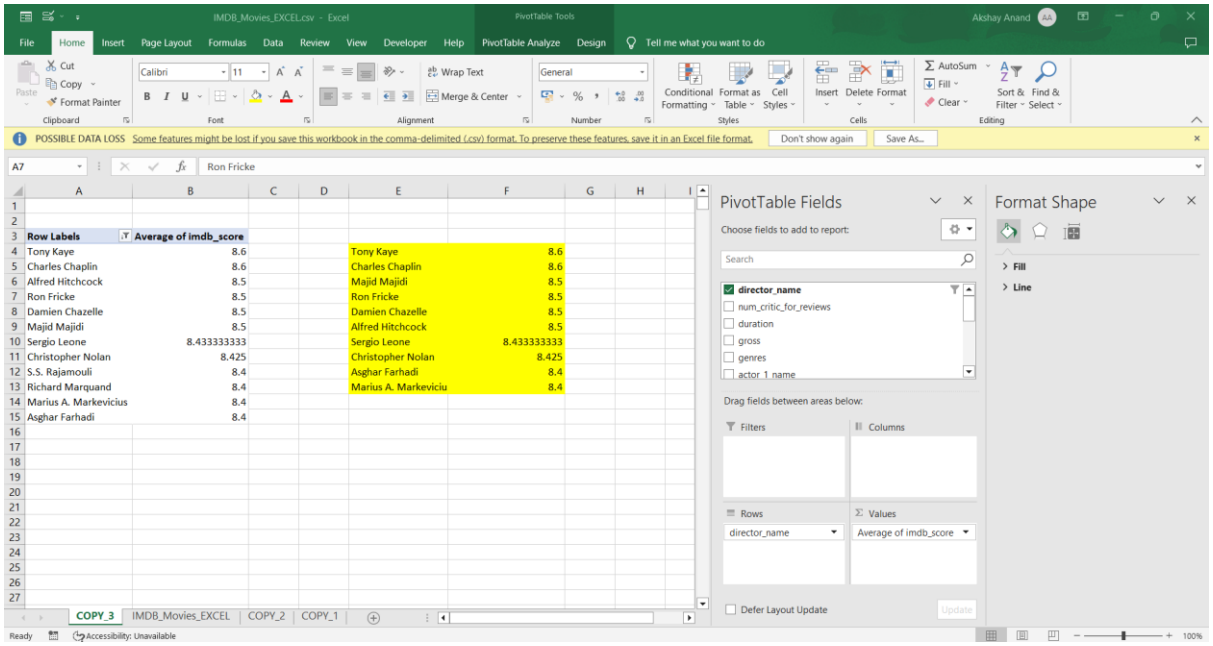
RANKING DONE USING FORMULA:

=RANK([@[imdb_score]],$N$2:$N$262,0)+COUNTIFS($N$2:N9,N9)-1.

WE USED SORTING AND FILTERS TO GET DESIRED OUTPUT.

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

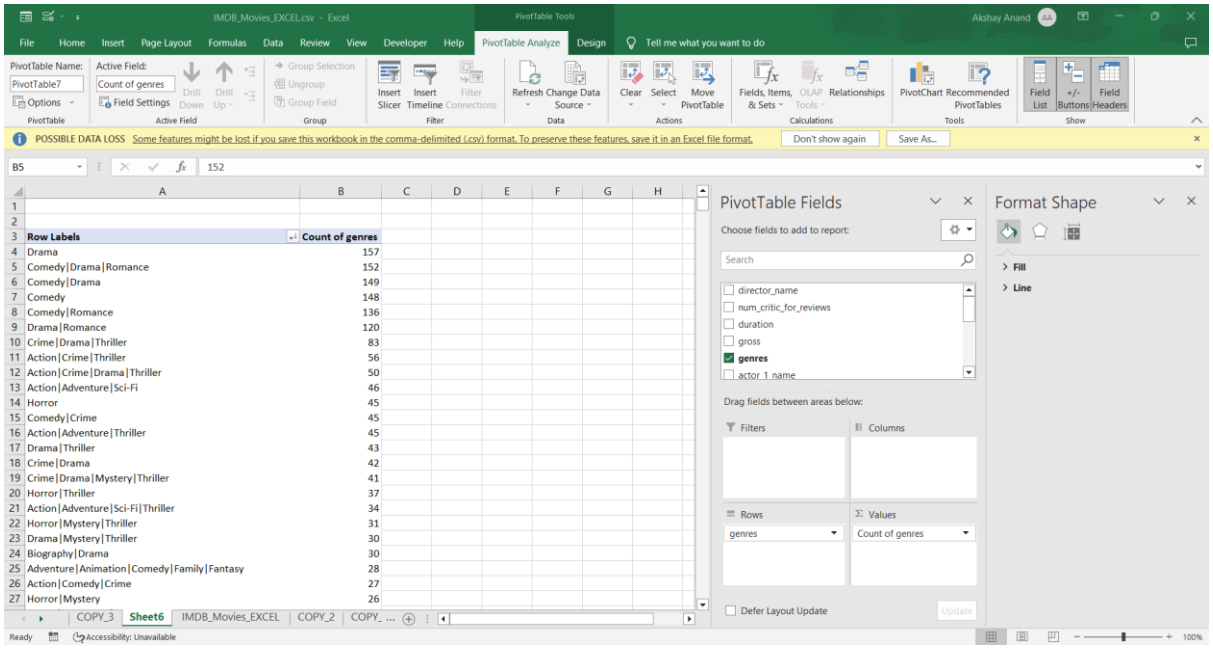


Row Labels	Average of imdb_score
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
S.S. Rajamouli	8.4
Richard Marquand	8.4
Marius A. Markevicius	8.4
Asghar Farhadi	8.4

Tony Kaye	8.6
Charles Chaplin	8.6
Majid Majidi	8.5
Ron Fricke	8.5
Damien Chazelle	8.5
Alfred Hitchcock	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4
Marius A. Markeviciu	8.4

WE USED PIVOT TABLES TO SORT DATA BASED ON IMDB SCORE.

Popular Genres: Perform this step using the knowledge gained while performing previous steps.
Your task: Find popular genres



Row Labels	Count of genres
Drama	157
Comedy|Drama|Romance	152
Comedy|Drama	149
Comedy	148
Comedy|Romance	136
Drama|Romance	120
Crime|Drama|Thriller	83
Action|Crime|Thriller	56
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	46
Horror	45
Comedy|Crime	45
Action|Adventure|Thriller	45
Drama|Thriller	43
Crime|Drama	42
Crime|Drama|Mystery|Thriller	41
Horror|Thriller	37
Action|Adventure|Sci-Fi|Thriller	34
Horror|Mystery|Thriller	31
Drama|Mystery|Thriller	30
Biography|Drama	30
Adventure|Animation|Comedy|Family|Fantasy	28
Action|Comedy|Crime	27
Horror|Mystery	26

WE USED PIVOT TABLES TO SORT DATA BASED ON THE COUNT OF GENRES.

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.

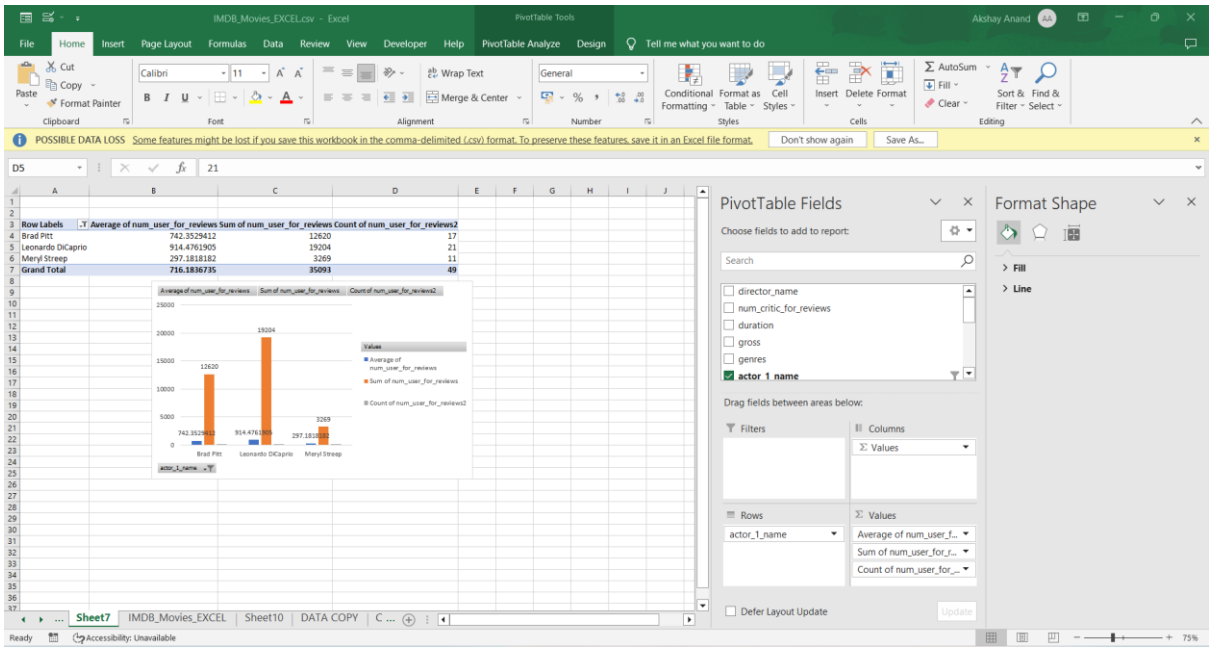
Append the rows of all these columns and store them in a new column named Combined.

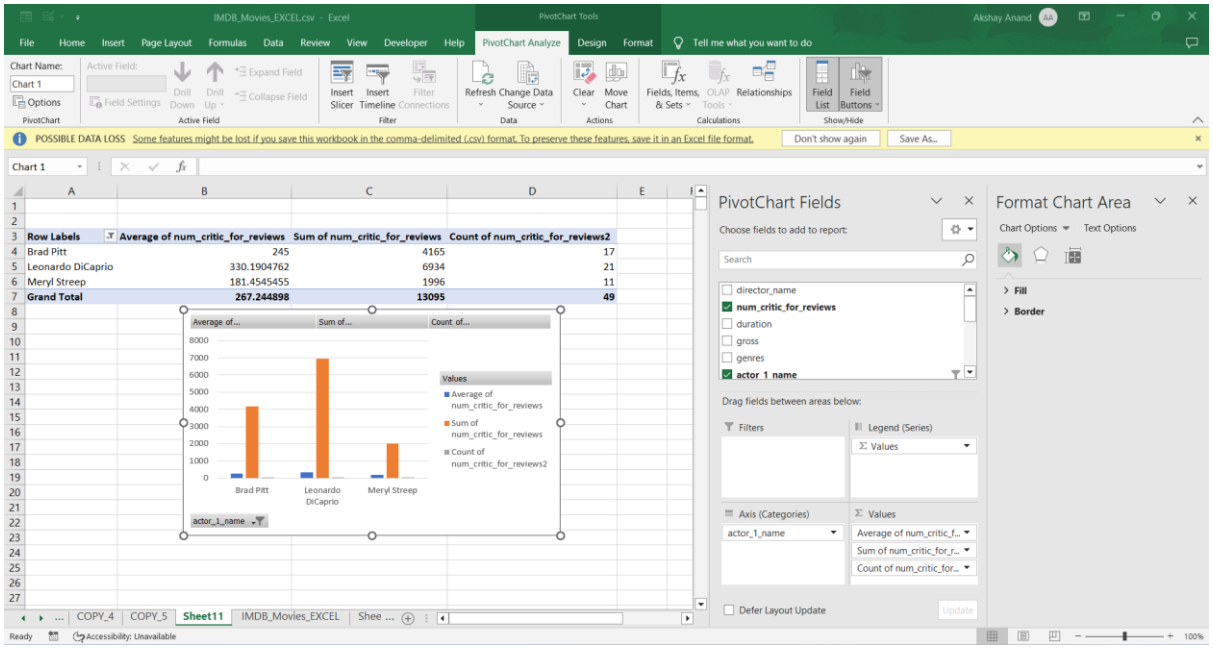
Group the combined column using the actor_1_name column.

Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

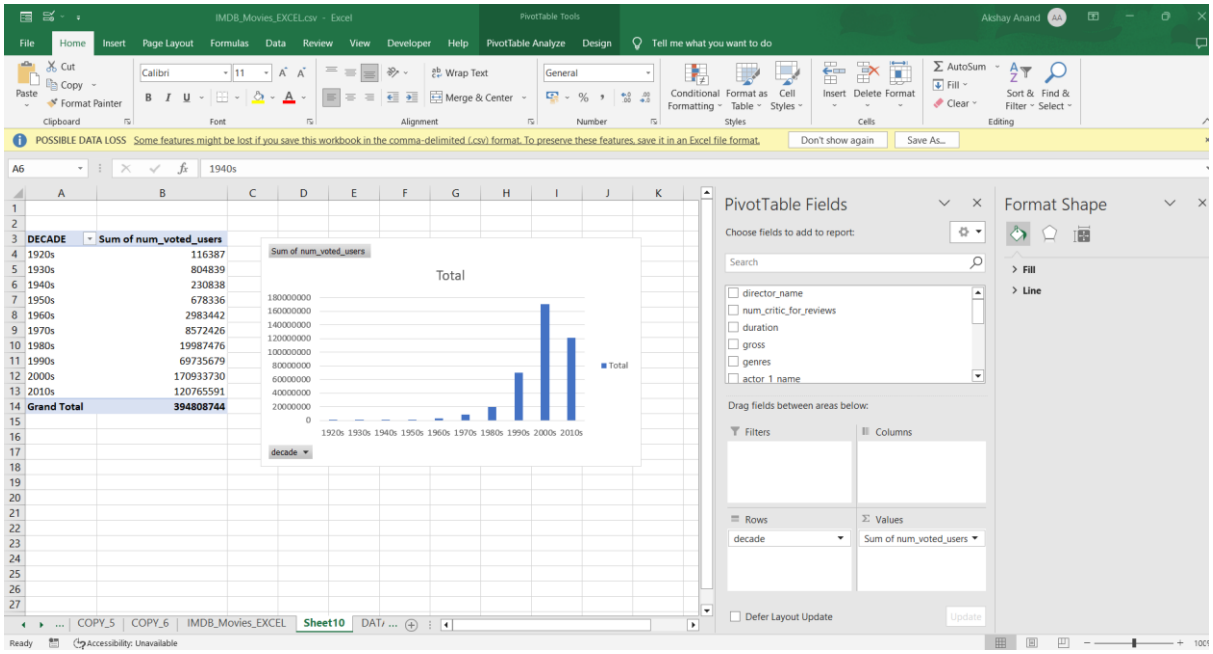
Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade 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





FROM ABOVE IT IS CLEAR THAT LEONARDO DICAPRIO IS BOTH THE CRITIC FAVOURITE AND THE AUDIENCE FAVOURITE ACTOR.



DECADES HAVE BEEN MADE BY: =FLOOR(M2,10) & "s"

WE USED PIVOT TABLES TO DRAW/PLOT GRAPHS.

RESULT: WE USED MS EXCEL FUNCTIONS SUCH AS FORMULAE, SORT, FILTER, PIVOT TABLES ETC TO PERFORM AN ANALYSIS OF THE DATASET AND GET INSIGHTS.