Project Description

We are giving you a dataset with multiple columns of different IMDB movies for your final project. It is necessary for you to frame the issue. You must identify the issue you wish to illuminate for this activity.

Once an issue has been identified, you should clean the data as appropriate before using data analysis techniques to examine the data set and draw conclusions. Use the five Whys study in your study to build a report that tells a story with the facts.

Analysis done on the following points:-

1. **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
2. **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?
3. **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
4. **Best Directors:** T-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
5. **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.
   Your task: Find popular genres
6. **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. 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

Approach

I'm going to review the data first, then tidy it up for analysis.When cleaning data, we start by looking for outliers and null values. If there are any null values, either delete that row or replace them with the mean, median, and mode. We should also eliminate any outliers and the useless column.

Tech-Stack Used

Here I am using Microsoft Excel 2016, I will be able to clean the data and develop a pivot table, which is useful for data analysis. We can visualize data using graphs in Excel as well.

Insights

1.Your task: Clean the data

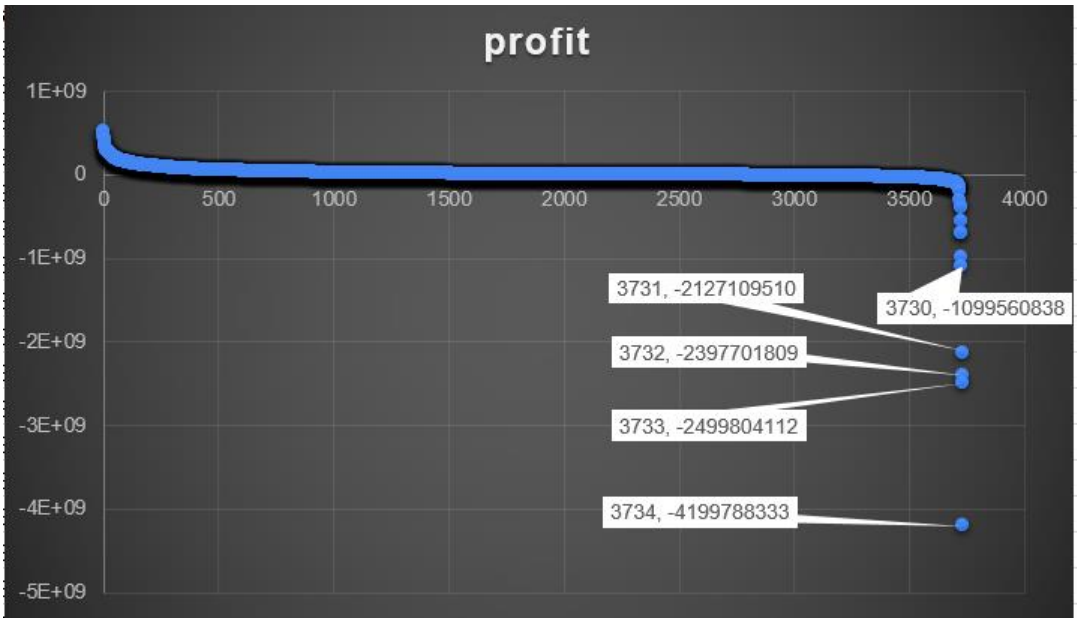
The hardest and most important phase in every data analysis effort is this one.The steps that make up this phase differ from one question and Dataset to another. The dataset will be cleaned by:-

1. First, delete any columns that won't be used in the analysis we'll be performing.
2. The columns with irrelevant information for the given analysis tasks include "Colour," "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,". Thus, it is necessary to remove these columns.
3. After eliminating the unnecessary columns, we now need to eliminate any rows from the dataset whose column values are blank or NULL.
4. Then, we must remove the duplicate values from the dataset using the 'Remove Duplicate Values/Cells' option found in the 'Data' tab.

Output : Task1: Cleaned data link

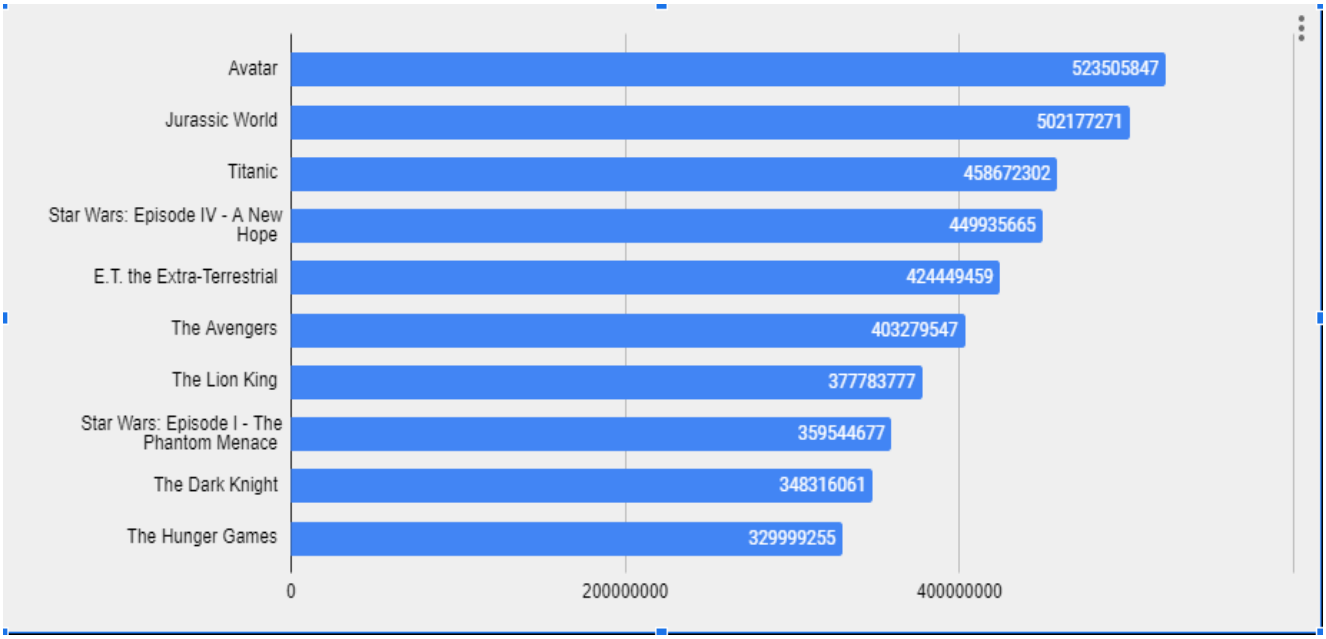
2. Your task: Find the movies with the highest profit?

1. To calculate the profit, we must first remove the budgeted amount from the gross sum.
2. Then, we will plot the profit (y_axis) and budget (x_axis) data using the scatter plot option. Afterward, we will use the graph to identify any outliers.



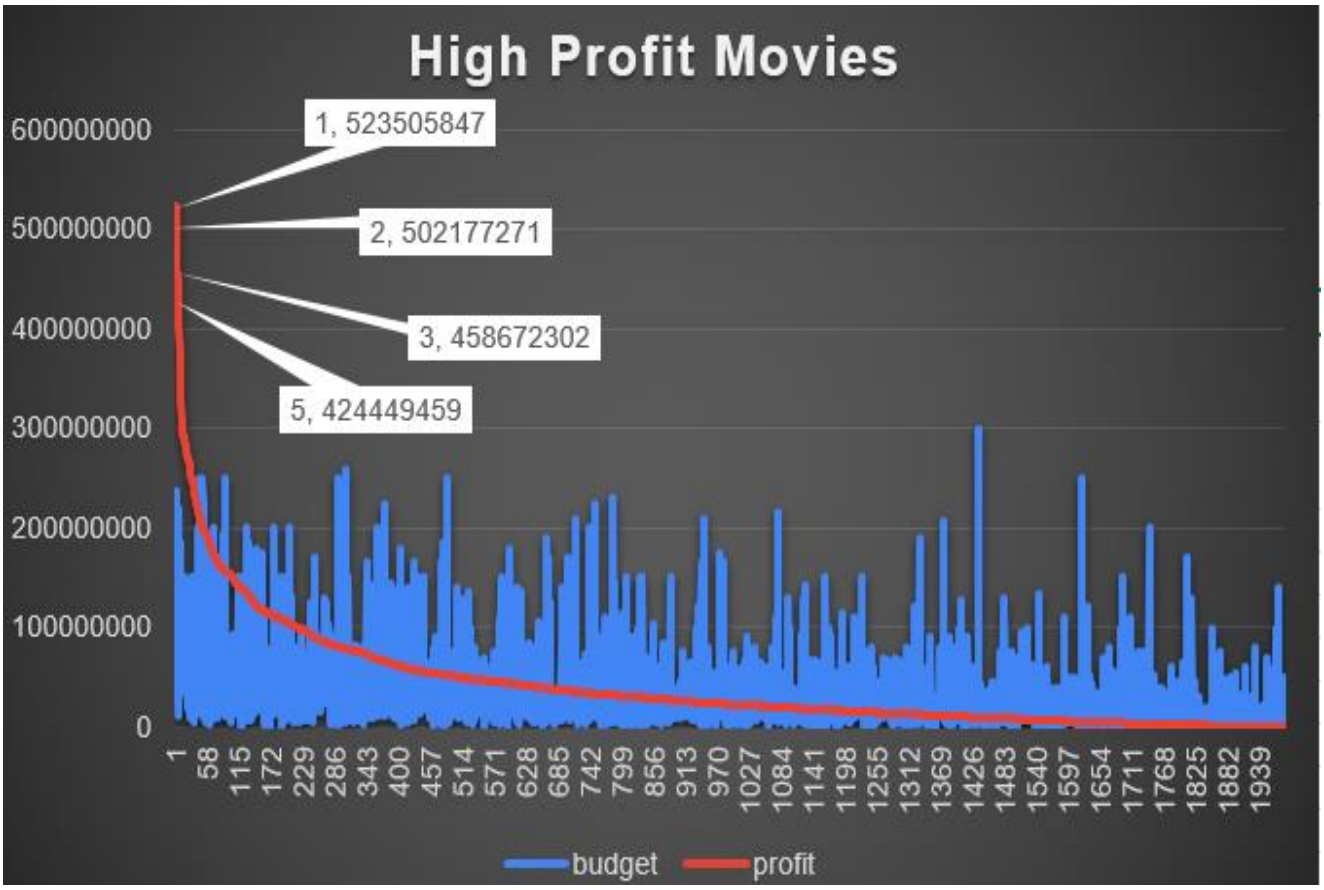
Outliers are
-4199788333
-2499804112
-2397701809
-2127109510
-989962610
-1099560838

Top 10 High Profit Movies



Output : [Task2: Top 10 High profit movies]

Google drive link : [Google drive link for all results]



3.Your task: Find IMDB Top 250

In order to locate the IMDB Top 250, we will:

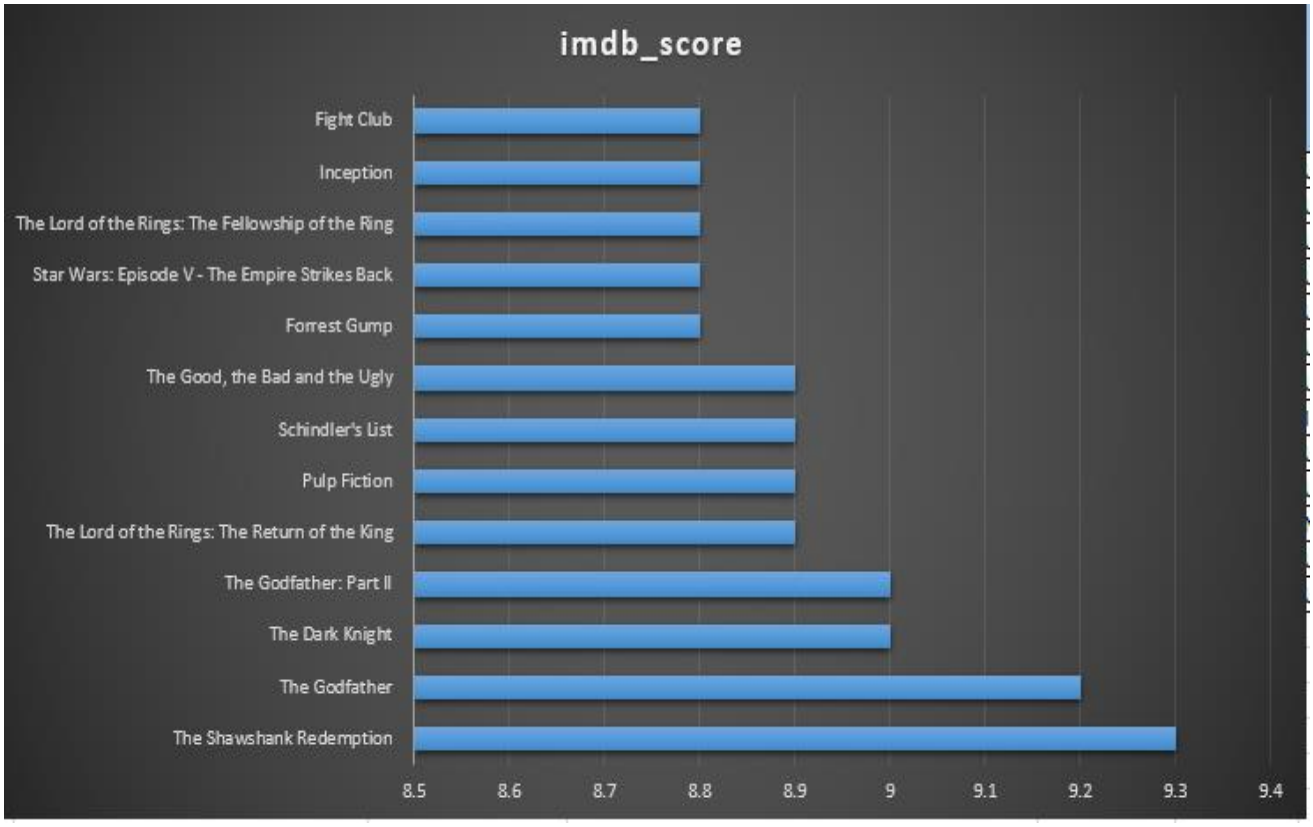
1. Using the sort and filter option, we will first remove any rows with num_voted_users more than 30000.
2. Next, the dataset will be arranged in decreasing order using the imdb_score.
3. Only the top 250 rows will be chosen for further research.
4. Next, we'll use the RANK() function and the formula to generate a new column for rank. =RANK(O2,$O$2:$O$251,0)+COUNTIFS($O$2:O2,O2)-1
5. The desired output will then be obtained once we filter out (unselect "English" from the language column).

Output : Task3: Top 250 Movies

Google drive link : Google drive link for all results

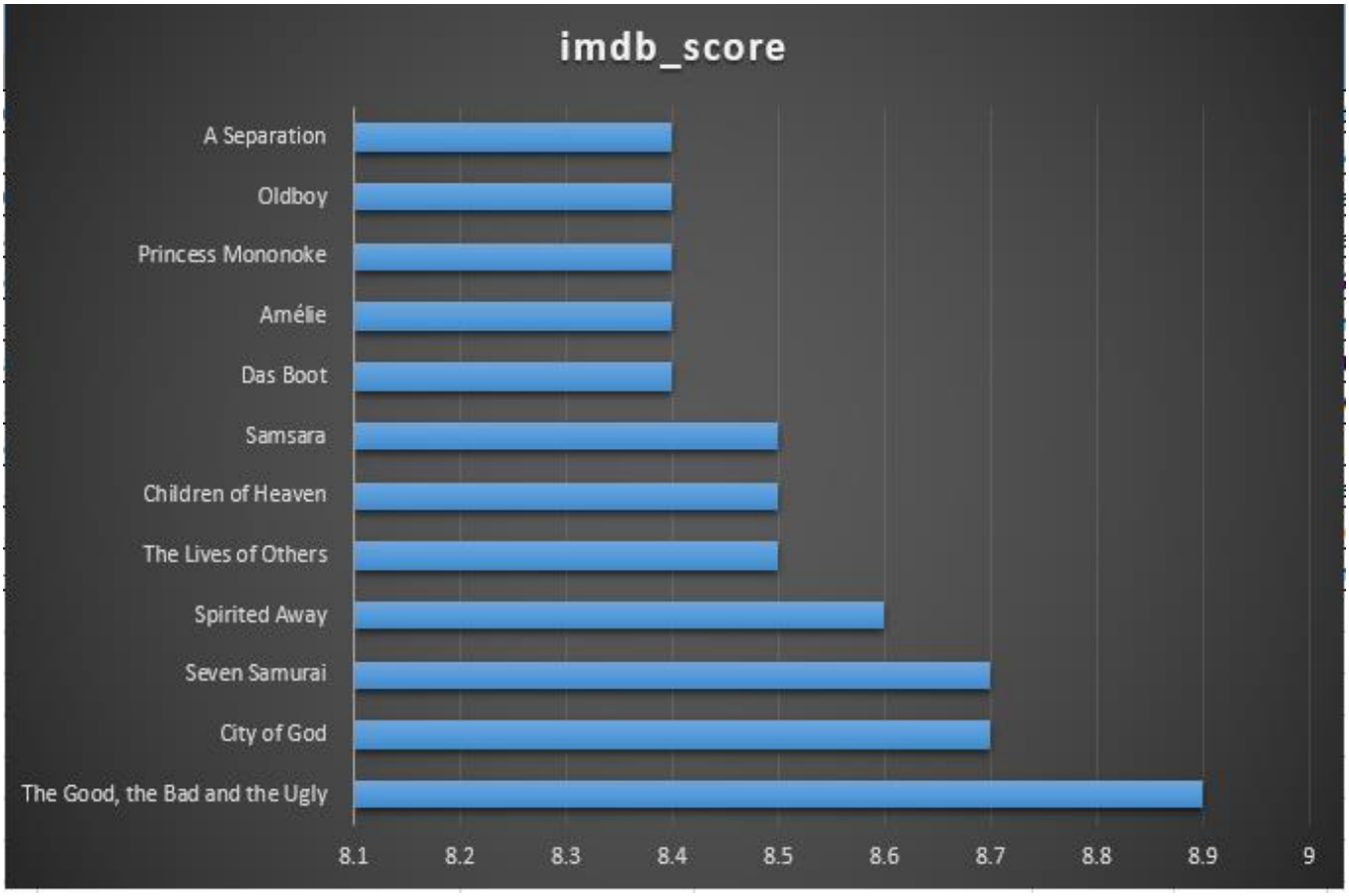
Top 13 IMDB Movies(all language) are:-

director_name	num_critic_for_reviews	duration	gross	genres	actor_1_name	movie_title	num_voted_users	language	country	budget	title_year	imdb_score
Frank Darabont	199	142	28341469	Crime|Drama	Morgan Freeman	The Shawshank Redemption	1689764	English	USA	25000000	1994	9.3
Francis Ford Coppola	208	175	134821952	Crime|Drama	Al Pacino	The Godfather	1155770	English	USA	6000000	1972	9.2
Christopher Nolan	645	152	533316061	Action|Crime|Drama|Thriller	Christian Bale	The Dark Knight	1676169	English	USA	185000000	2008	9
Francis Ford Coppola	149	220	57300000	Crime|Drama	Robert De Niro	The Godfather: Part II	790926	English	USA	13000000	1974	9
Peter Jackson	328	192	377019252	Action|Adventure|Drama|Fantasy	Orlando Bloom	The Lord of the Rings: The Return of the King	1215718	English	USA	94000000	2003	8.9
Quentin Tarantino	215	178	107930000	Crime|Drama	Bruce Willis	Pulp Fiction	1324680	English	USA	8000000	1994	8.9
Steven Spielberg	174	185	96067179	Biography|Drama|History	Liam Neeson	Schindler's List	865020	English	USA	22000000	1993	8.9
Sergio Leone	181	142	6100000	Western	Clint Eastwood	The Good, the Bad and the Ugly	503509	Italian	Italy	1200000	1966	8.9
Robert Zemeckis	149	142	329691196	Comedy|Drama	Tom Hanks	Forrest Gump	1251222	English	USA	55000000	1994	8.8
Irvin Kershner	223	127	290158751	Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode V - The Empire Strikes Back	837759	English	USA	18000000	1980	8.8
Peter Jackson	297	171	313837577	Action|Adventure|Drama|Fantasy	Christopher Lee	The Lord of the Rings: The Fellowship of the Ring	1238746	English	New Zeala	93000000	2001	8.8
Christopher Nolan	642	148	292568851	Action|Adventure|Sci-Fi|Thriller	Leonardo DiCaprio	Inception	1468200	English	USA	160000000	2010	8.8
David Fincher	315	151	37023395	Drama	Brad Pitt	Fight Club	1347461	English	USA	63000000	1999	8.8



Top 12 IMDB Movies(all languages(except English)) are:-

director_name	num_critic_for_review	duration	gross	genres	actor_1_name	movie_title	num_voted_users	language	country	budget	title_year	imdb_score
Sergio Leone	181	142	6100000	Western	Clint Eastwood	The Good, the Bad and the Ugly	503509	Italian	Italy	1200000	1966	8.9
Fernando Meirelles	214	135	7563397	Crime|Drama	Alice Braga	City of God	533200	Portuguese	Brazil	3300000	2002	8.7
Akira Kurosawa	153	202	269061	Action|Adventure|Drama	Takashi Shimura	Seven Samurai	229012	Japanese	Japan	2000000	1954	8.7
Hayao Miyazaki	246	125	10049886	Adventure|Animation|Family|Fantasy	Bunta Sugawara	Spirited Away	417971	Japanese	Japan	19000000	2001	8.6
Florian Henckel von Donnersmarck	215	137	11284657	Drama|Thriller	Sebastian Koch	The Lives of Others	259379	German	Germany	2000000	2006	8.5
Majid Majidi	46	89	925402	Drama|Family	Bahare Seddiqi	Children of Heaven	27882	Persian	Iran	180000	1997	8.5
Ron Fricke	115	102	2601847	Documentary|Music	Collin Alfredo St. Dic	Samsara	22457	None	USA	4000000	2011	8.5
Wolfgang Petersen	96	293	11433134	Adventure|Drama|Thriller|War	Jürgen Prochnow	Das Boot	168203	German	West Germany	14000000	1981	8.4
Jean-Pierre Jeunet	242	122	33201661	Comedy|Romance	Mathieu Kassovitz	Amélie	534262	French	France	77000000	2001	8.4
Hayao Miyazaki	174	134	2298191	Adventure|Animation|Fantasy	Minnie Driver	Princess Mononoke	221552	Japanese	Japan	2400000000	1997	8.4
Chan-wook Park	305	120	2181290	Drama|Mystery|Thriller	Min-sik Choi	Oldboy	356181	Korean	South Korea	3000000	2003	8.4
Asghar Farhadi	354	123	7098492	Drama|Mystery	Shahab Hosseini	A Separation	151812	Persian	Iran	500000	2011	8.4



4.Your task: Find the best directors

In order to determine the top 10 directors based on the average imdb_score, we will:

1. First, choose the cleaned dataset's imdb_score column.
2. Next, we'll select the pivot table.
3. We will include director_name in the pivot table's series section.
4. Next, we'll add the average imdb_score to the pivot table's values section. The information will then be sorted alphabetically by director name after being sorted first based on average imdb_score in descending order.

Row Labels	Average of imdb_score
Akira Kurosawa	8.7
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

5.Your task: Find popular genres

To identify the most popular genres, we will:

1. Choose the genres column in the cleaned dataset first.
2. After that, we'll choose the pivot table option.
3. Next, we'll choose the genre names for the row labels.
4. Next, we will use the values as the total number of genres and sort the data in decreasing order using the total number of genres.

Row Labels	Count of genres
Comedy|Drama|Romance	148
Drama	147
Comedy	143
Comedy|Drama	142
Comedy|Romance	135
Drama|Romance	117
Crime|Drama|Thriller	79
Action|Crime|Thriller	54
Action|Crime|Drama|Thriller	48
Action|Adventure|Sci-Fi	45
Comedy|Crime	45
Grand Total	**1103**

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

To identify the critics' and audiences' favourite actors, we will:

1. The first three new columns, Meryl Streep, Leonardo DiCaprio, and Brad Pitt, comprise the films in which the actors: "Meryl Streep," "Leonardo DiCaprio," and "Brad Pitt" are the principal actors from the actor_1_name column.
2. Next, we'll combine the three newly produced columns (actor_1_name_combine, etc.) into one column.
3. Next, we'll combine the three columns of actors who are popular with critics and viewers.
4. Next, we will use the pivot table to calculate the average, total, and count of actors who are popular with both the critics and the audience.

Dataset for actor name Meryl_Streep

director_name	num_critic_for_reviews	duration	gross	genres	Meryl Streep	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score
Stephen Daldry	174	114	41597830	Drama|Romance	Meryl Streep	The Hours	102123	660	English	USA	PG-13	25000000	2002	7.6
Sydney Pollack	66	161	87100000	Biography|Drama|Romance	Meryl Streep	Out of Africa	52339	200	English	USA	PG	31000000	1985	7.2
Nora Ephron	252	123	94125426	Biography|Drama|Romance	Meryl Streep	Julie & Julia	79264	277	English	USA	PG-13	40000000	2009	7
David Frankel	208	109	124732962	Comedy|Drama|Romance	Meryl Streep	The Devil Wears Prada	286178	631	English	USA	PG-13	35000000	2006	6.8
Robert Altman	211	105	20338609	Comedy|Drama|Music	Meryl Streep	A Prairie Home Companion	19655	280	English	USA	PG-13	10000000	2006	6.8
Nancy Meyers	187	120	112703470	Comedy|Drama|Romance	Meryl Streep	It's Complicated	69860	214	English	USA	R	85000000	2009	6.6
Phyllida Lloyd	331	105	29959436	Biography|Drama|History	Meryl Streep	The Iron Lady	82327	350	English	UK	PG-13	13000000	2011	6.4
David Frankel	234	100	63536011	Comedy|Drama|Romance	Meryl Streep	Hope Springs	34258	178	English	USA	PG-13	30000000	2012	6.3
Curtis Hanson	42	111	46815748	Action|Adventure|Crime|Thriller	Meryl Streep	The River Wild	32544	69	English	USA	PG-13	45000000	1994	6.3
Carl Franklin	64	127	23209440	Drama	Meryl Streep	One True Thing	9283	112	English	USA	R	30000000	1998	7
Robert Redford	227	92	14998070	Drama|Thriller|War	Meryl Streep	Lions for Lambs	41170	298	English	USA	R	35000000	2007	6.2

Dataset for actor named Leonardo_DiCaprio

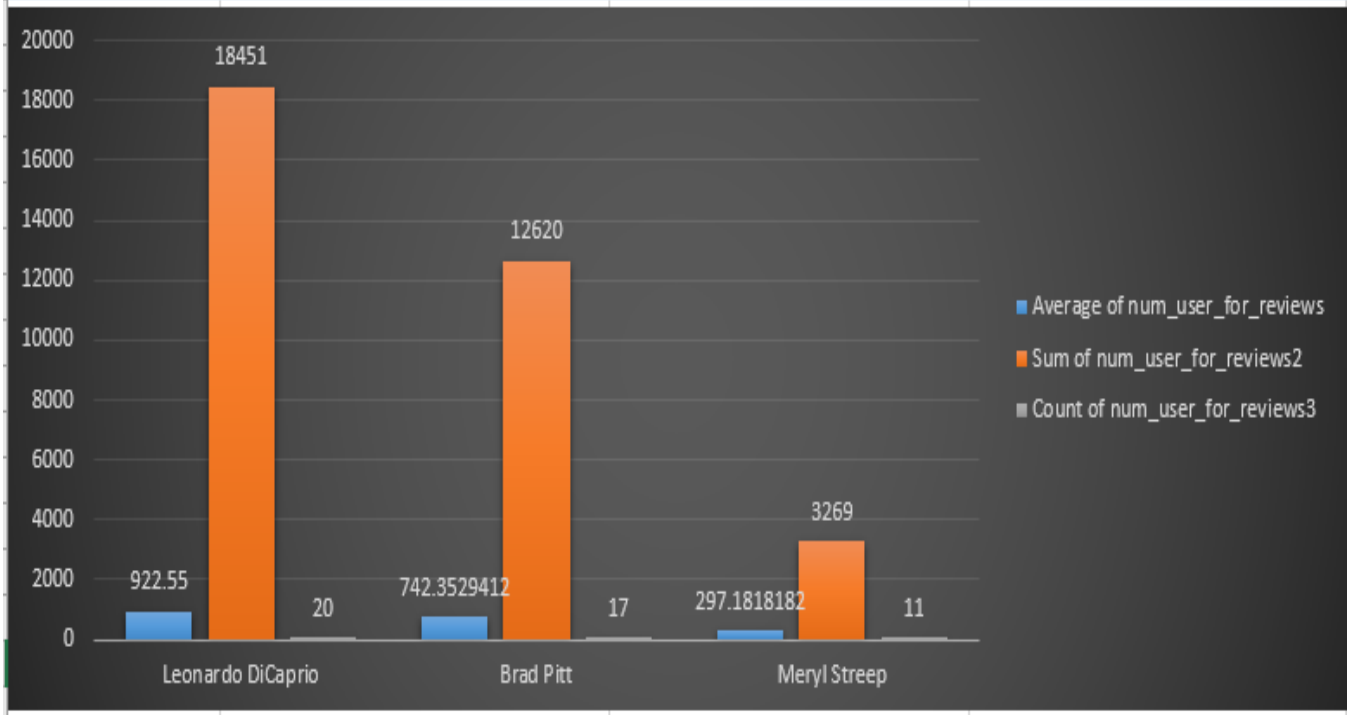
director_name	num_critic_for_reviews	duration	gross	genres	Leonardo DiCaprio	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score
Christopher Nolan	642	148	292568851	Action|Adventure|Sci-Fi|Thriller	Leonardo DiCaprio	Inception	1468200	2803	English	USA	PG-13	160000000	2010	8.8
Quentin Tarantino	765	165	162804648	Drama|Western	Leonardo DiCaprio	Django Unchained	955174	1193	English	USA	R	100000000	2012	8.5
Martin Scorsese	352	151	132373442	Crime|Drama|Thriller	Leonardo DiCaprio	The Departed	873649	2054	English	USA	R	90000000	2006	8.5
Martin Scorsese	606	240	116866727	Biography|Comedy|Crime|Dram	Leonardo DiCaprio	The Wolf of Wall Street	780588	1138	English	USA	R	100000000	2013	8.2
Alejandro G. Iñárr	556	156	183635922	Adventure|Drama|Thriller|Weste	Leonardo DiCaprio	The Revenant	406020	1188	English	USA	R	135000000	2015	8.1
Martin Scorsese	490	138	127968405	Mystery|Thriller	Leonardo DiCaprio	Shutter Island	786092	964	English	USA	R	80000000	2010	8.1
Steven Spielberg	194	141	164435221	Biography|Crime|Drama	Leonardo DiCaprio	Catch Me If You Can	525801	667	English	USA	PG-13	52000000	2002	8
Edward Zwick	166	143	57366262	Adventure|Drama|Thriller	Leonardo DiCaprio	Blood Diamond	400292	657	English	Germany	R	100000000	2006	8
James Cameron	315	194	658672302	Drama|Romance	Leonardo DiCaprio	Titanic	793059	2528	English	USA	PG-13	200000000	1997	7.7
Martin Scorsese	267	170	102608827	Biography|Drama	Leonardo DiCaprio	The Aviator	264318	799	English	USA	PG-13	110000000	2004	7.5
Martin Scorsese	233	216	77679638	Crime|Drama	Leonardo DiCaprio	Gangs of New York	314033	1166	English	USA	R	100000000	2002	7.5
Baz Luhrmann	490	143	144812796	Drama|Romance	Leonardo DiCaprio	The Great Gatsby	362912	753	English	Australia	PG-13	105000000	2013	7.3
Sam Mendes	323	119	22877808	Drama|Romance	Leonardo DiCaprio	Revolutionary Road	152591	414	English	USA	R	35000000	2008	7.3
Ridley Scott	238	128	39380442	Action|Drama|Thriller	Leonardo DiCaprio	Body of Lies	174248	263	English	USA	R	70000000	2008	7.1
Baz Luhrmann	106	120	46338728	Drama|Romance	Leonardo DiCaprio	Romeo + Juliet	167750	506	English	USA	PG-13	14500000	1996	6.8
Clint Eastwood	392	137	37304950	Biography|Crime|Drama	Leonardo DiCaprio	J. Edgar	102728	279	English	USA	R	35000000	2011	6.6
Danny Boyle	118	119	39778599	Adventure|Drama|Thriller	Leonardo DiCaprio	The Beach	176169	548	English	USA	R	50000000	2000	6.6
Randall Wallace	83	132	56876365	Action|Adventure	Leonardo DiCaprio	The Man in the Iron Mask	125219	244	English	USA	PG-13	35000000	1998	6.4
Sam Raimi	63	107	18636537	Action|Thriller|Western	Leonardo DiCaprio	The Quick and the Dead	69197	216	English	Japan	R	32000000	1995	6.4
Jerry Zaks	45	98	12782508	Drama	Leonardo DiCaprio	Marvin's Room	20163	71	English	USA	PG-13	23000000	1996	6.7

Dataset for actor named Brad_Pitt

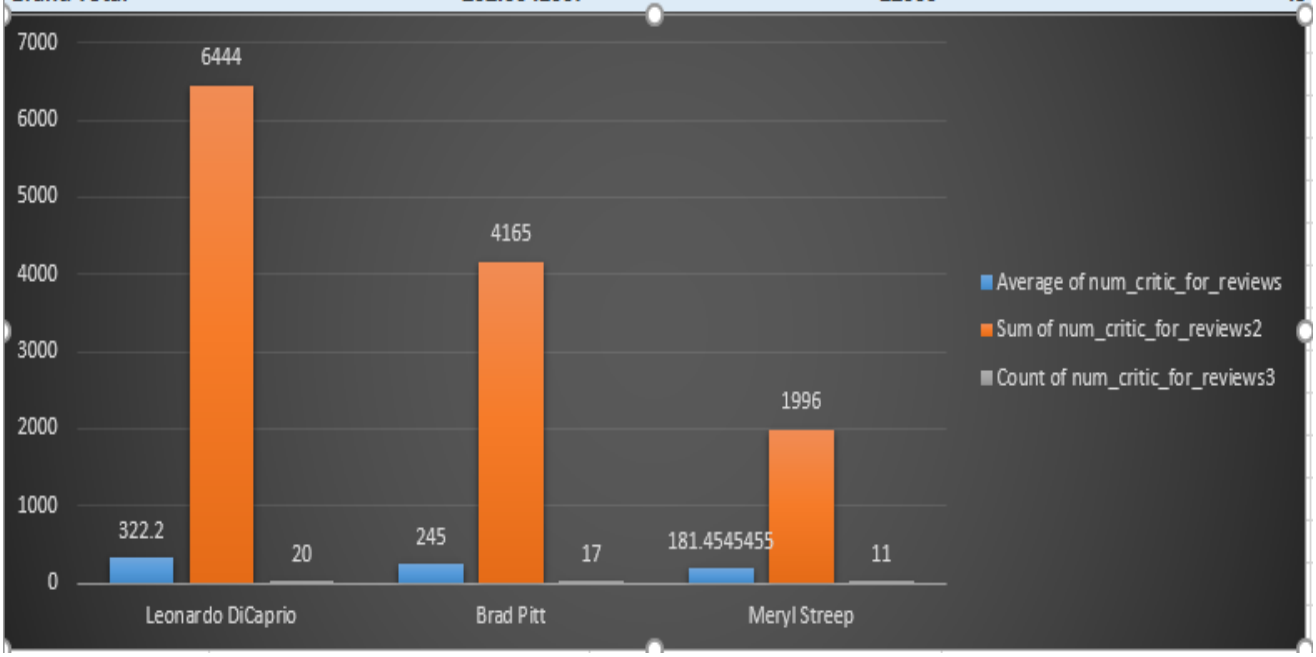
director_name	num_critic_for_reviews	duration	gross	genres	Brad Pitt	movie_title	num_voted_users	num_user_for_reviews	language	country	content_rating	budget	title_year	imdb_score
David Fincher	315	151	37023395	Drama	Brad Pitt	Fight Club	1347461	2968	English	USA	R	63000000	1999	8.8
Tony Scott	122	121	12281500	Action|Crime|Drama|Romance|T	Brad Pitt	True Romance	163492	460	English	USA	R	13000000	1993	8
Steven Soderberg	186	116	183405771	Crime|Thriller	Brad Pitt	Ocean's Eleven	402645	845	English	USA	PG-13	85000000	2001	7.8
David Fincher	362	166	127490802	Drama|Fantasy|Romance	Brad Pitt	The Curious Case of Benjam	459346	822	English	USA	PG-13	150000000	2008	7.8
Neil Jordan	120	123	105264608	Drama|Fantasy|Horror	Brad Pitt	Interview with the Vampire: T	239752	406	English	USA	R	60000000	1994	7.6
David Ayer	406	134	85707116	Action|Drama|War	Brad Pitt	Fury	303185	701	English	USA	R	68000000	2014	7.6
Alejandro G. Iñárr	285	143	34300771	Drama	Brad Pitt	Babel	243799	908	English	France	R	25000000	2006	7.5
Andrew Dominik	273	160	3904982	Biography|Crime|Drama|History	Brad Pitt	The Assassination of Jesse	136104	415	English	USA	R	30000000	2007	7.5
Wolfgang Peterse	220	196	133228348	Adventure	Brad Pitt	Troy	381672	1694	English	USA	R	175000000	2004	7.2
Jean-Jacques Anr	76	136	37901509	Adventure|Biography|Drama|His	Brad Pitt	Seven Years in Tibet	96385	119	English	USA	PG-13	70000000	1997	7
Tony Scott	142	114	26871	Action|Crime|Thriller	Brad Pitt	Spy Game	121259	361	English	Germany	R	92000000	2001	7
Terrence Malick	584	139	13303319	Drama|Fantasy	Brad Pitt	The Tree of Life	136367	975	English	USA	PG-13	32000000	2011	6.7
Patrick Gilmore	98	85	26288320	Adventure|Animation|Comedy|D	Brad Pitt	Sinbad: Legend of the Sever	36144	91	English	USA	PG	60000000	2003	6.7
Doug Liman	233	126	186336103	Action|Comedy|Crime|Romance	Brad Pitt	Mr. & Mrs. Smith	348861	798	English	USA	PG-13	120000000	2005	6.5
Steven Soderberg	198	125	125531634	Crime|Thriller	Brad Pitt	Ocean's Twelve	284852	627	English	USA	PG-13	110000000	2004	6.4
Andrew Dominik	414	97	14938570	Crime|Thriller	Brad Pitt	Killing Them Softly	111625	369	English	USA	R	15000000	2012	6.2
Angelina Jolie Pitt	131	122	531009	Drama|Romance	Brad Pitt	By the Sea	7976	61	English	USA	R	10000000	2015	5.3

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

Row Labels	Average of num_user_for_reviews	Sum of num_user_for_reviews2	Count of num_user_for_reviews3
Leonardo DiCaprio	922.55	18451	20
Brad Pitt	742.3529412	12620	17
Meryl Streep	297.1818182	3269	11
Grand Total	**715.4166667**	**34340**	**48**



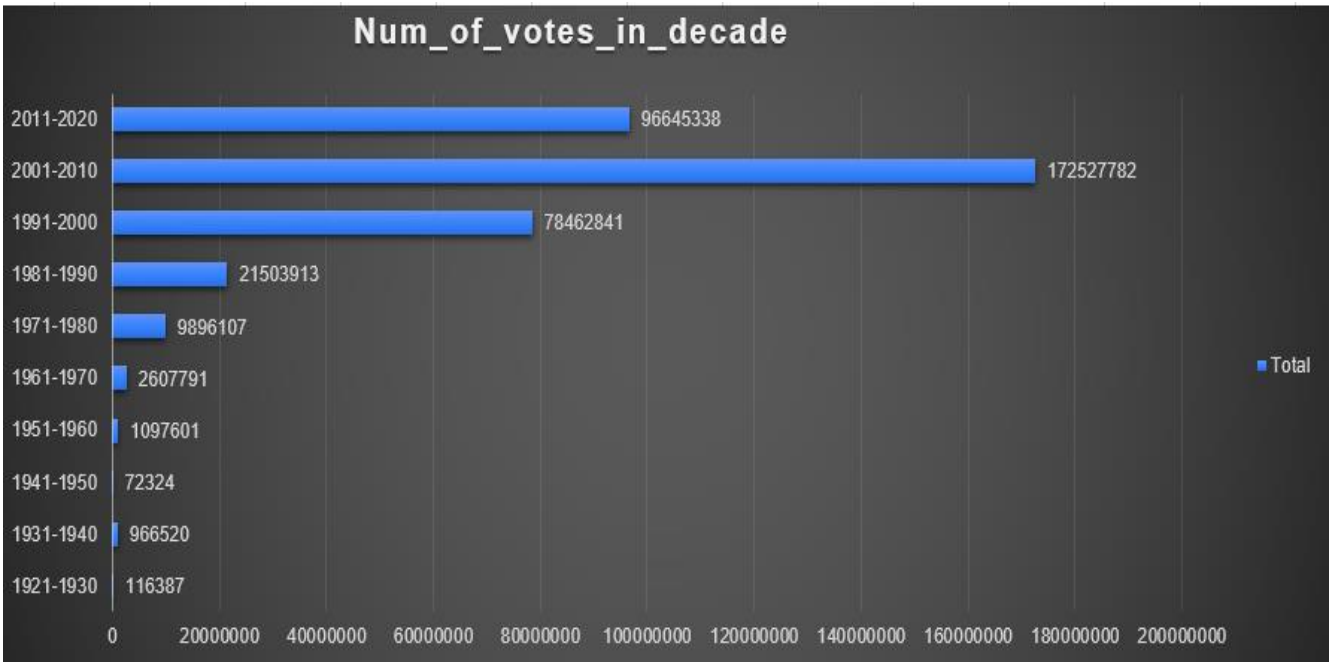
Row Labels	Average of num_critic_for_reviews	Sum of num_critic_for_reviews2	Count of num_critic_for_reviews3
Leonardo DiCaprio	322.2	6444	20
Brad Pitt	245	4165	17
Meryl Streep	181.4545455	1996	11
Grand Total	**262.6041667**	**12605**	**48**



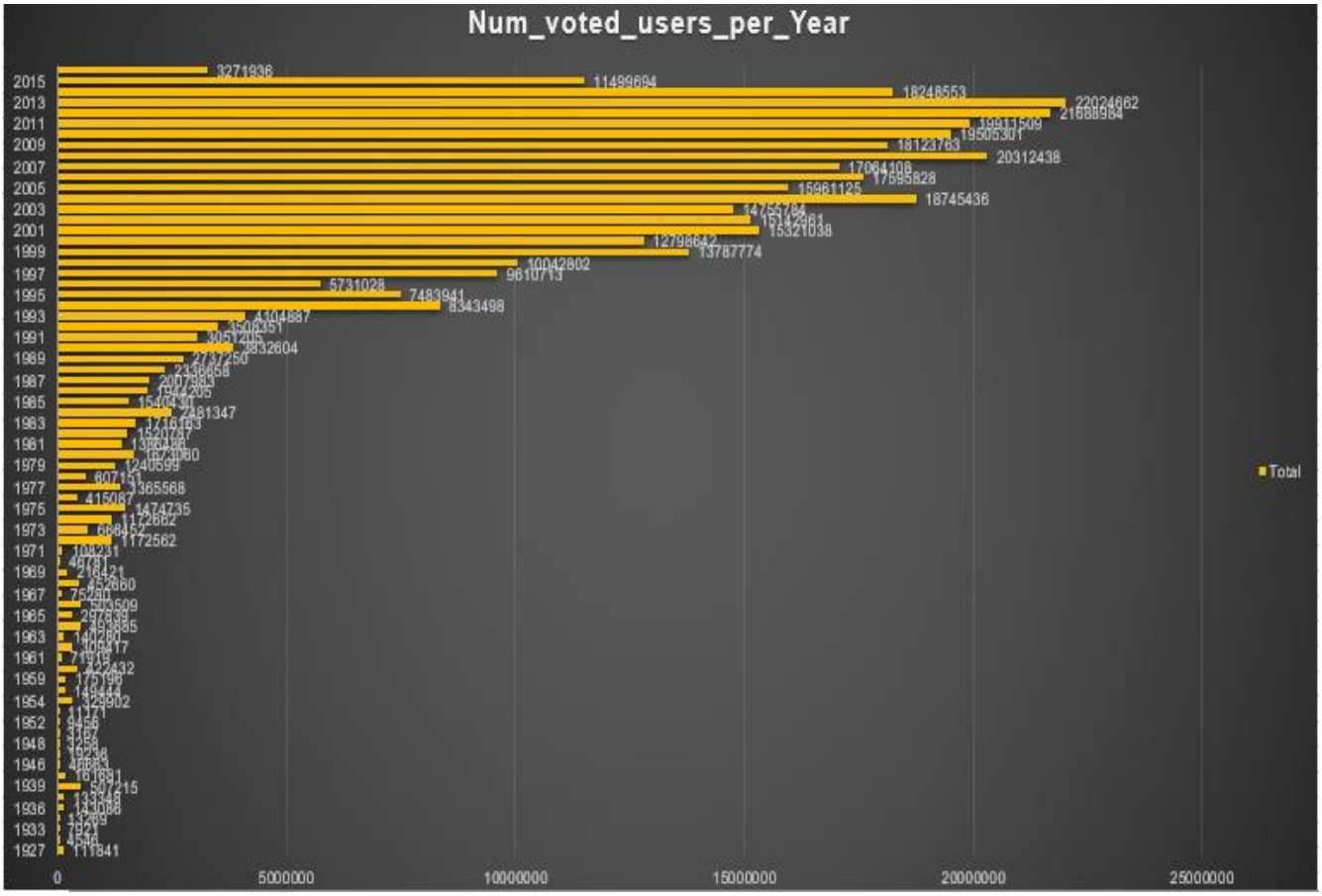
Observation : From the above two graphs we can infer that 'Leonardo DiCaprio' was both critic favorite and audience-favorite actor

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

Row Labels	Sum of num_voted_users
1921-1930	116387
1931-1940	966520
1941-1950	72324
1951-1960	1097601
1961-1970	2607791
1971-1980	9896107
1981-1990	21503913
1991-2000	78462841
2001-2010	172527782
2011-2020	96645338
Grand Total	**383896604**



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



Output Google drive link : [Google drive link for all results]

Result

1. The dataset was first cleaned, which involved removing unnecessary columns, outliers, and null values.
2. Second, by adding a new column on profit that lists which films were successful and brought in more money, I am able to learn more about high-profit films.
3. Third, I locate the list of top IMD-rated films, which reveals which films had the most user involvement.
4. After that, I determined the top 10 directors based on their IMD scores. Here, we learn about well-known filmmakers.
5. I then discovered prominent movie genres, which demonstrate the current popularity of various movie subgenres.
6. After that, I learned which actors were popular with critics and voters alike, as well as how voting patterns had changed throughout the years.

Note : To avoid this, I'm giving the URL to my Google Drive folder, which contains associated output files. Some share links are not accessible in browser due to technical issues.

Google drive link for all results