IMDB Movie Analysis

By Thotamchetty Ranjith

Description:

We are required to provide a detailed report for the given data record of movies from the IMDB mentioning the answers of the questions that follows:

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***: TGroup 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.

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

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.

Tech Stack Used:

Microsoft Excel 2019

Approach:

1.Data Cleaning: Clean the dataset by removing irrelevant columns, handling null values, and correcting data types.

2.Movies with highest profit: Create a new column called profit by subtracting budget from gross, sort the dataset by the profit column, and plot the profit vs budget to identify outliers.

3.IMDB Top 250: Create a new column called IMDb_Top_250 that contains the top 250 movies with the highest IMDb rating (imdb_score), and where num_voted_users is greater than 25,000. Also, add a Rank column with values 1 to 250 for each movie.

4.Top Foreign Language Films: Extract the non-English movies from the IMDb_Top_250 column and store them in a new column called Top_Foreign_Lang_Film.

5.Best Directors: Group the dataset by director_name, find the top 10 directors with the highest mean imdb_score (in case of a tie, sort them alphabetically), and store them in a new column called top10director.

6.Popular Genres: Identify the most popular genres by analyzing the frequency of each genre in the dataset.

7.Lead Actor Analysis: Create new columns Meryl_Streep, Leo_Caprio, and Brad_Pitt with movies where these actors are the lead actor. Combine the rows of these columns and group them by actor_1_name to find the mean of num_critic_for_reviews and num_users_for_reviews for each lead actor.

8. Decade Analysis: Create a new column called decade that represents the decade to which each movie belongs. Sort the dataset by the decade column, 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.

After following the steps, we have to document the analysis and insights to give information to the stakeholders

A) Data Cleaning:

- From the given data we have removed the columns such as the ‘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’ as the columns not very much relevant to our analysis so we removed them.
- Then we used filter to remove blank values in the columns of ‘director_name’, ‘language’, ‘actor_1_name’, ‘budget’ etc
- **Note**: we didn’t remove the blank values in the column ‘rating’
- After that we have removed duplicate rows containing similar values by using “remove duplicates” after that we if duplicates still exists then use “conditional formatting” to highlight the duplicate movie names and removes them manually
- Clean the data using clean command and replace special character similar to this Ã in the movie_title column correct the names of actors and directors according their languages

- Here we have also encountered a special case of movie_title which different with different everything has a same which is “The Host” we have both Korean and English movies named as same so to separate them we used the “The Host (K)” to represent the Korean movie.
- The following is the drive link for ***cleaned data:***

https://docs.google.com/spreadsheets/d/1kwTThJVnrvnRm7pP6GOPXfwHvaP7BiQZ/edit?usp=share_link&ouid=112622933348215332431&rtpof=true&sd=true

B) Most Profitable movies

To identify the profitable movies, we calculated the profit by subtracting the budget from the gross box office collections using the formula Profit = (gross – budget). We created a pivot table with new fields for profit and return on investment (ROI) % and filtered out the movies that made a profit. We then created a scatter graph to plot budget vs. profit and analyzed the insights. We also calculated the ROI % for each movie to identify which movies made the most money relative to their budget. Additionally, we identified the outliers in the budget vs. profit plot.

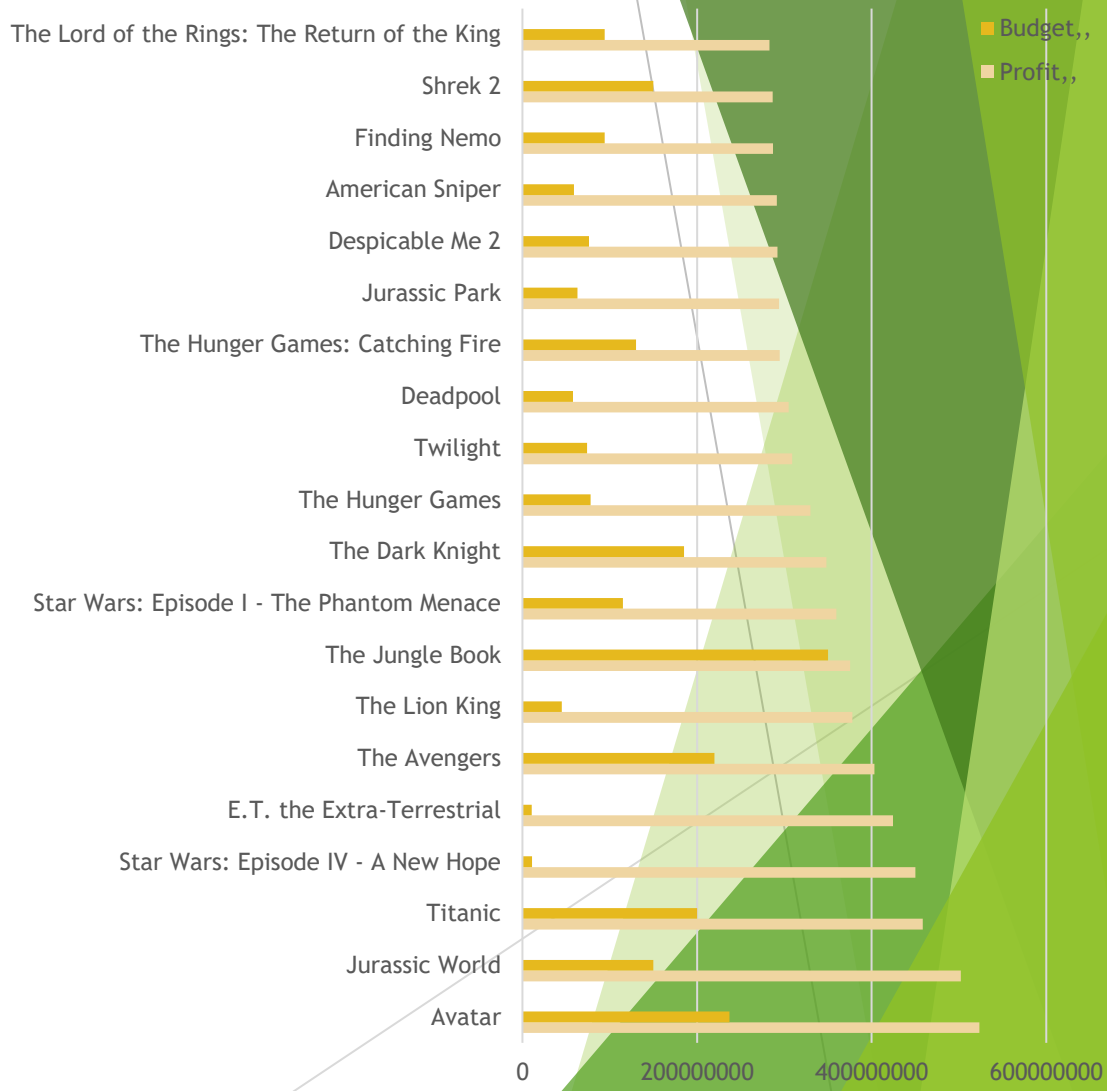
We have also plotted the movies with most ROI%

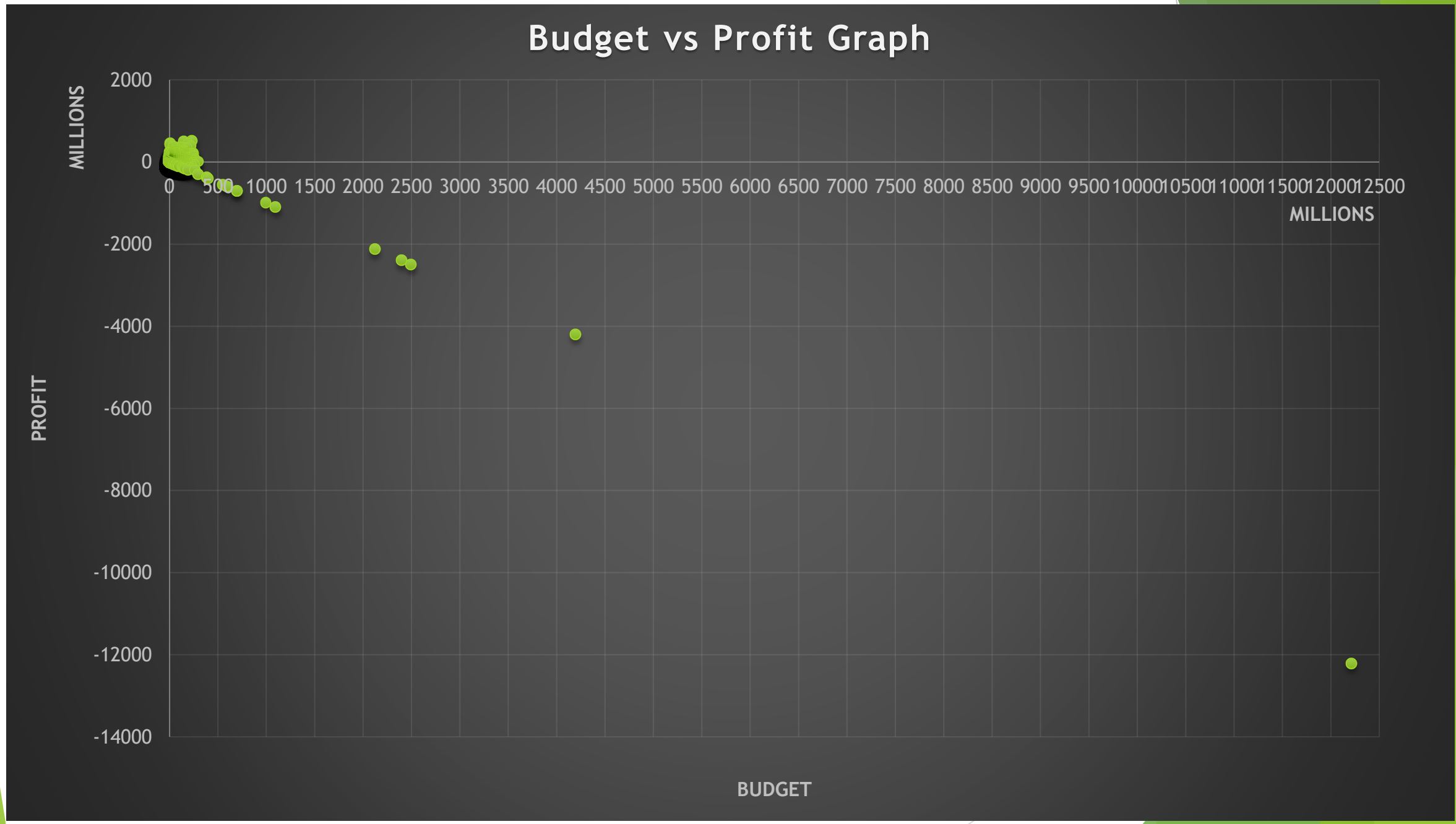
The following is drive link for excel workbook in which have done the analysis:

https://docs.google.com/spreadsheets/d/1qI2BlJ-pfWVapxIcDJsGyq-9BgSa5agM/edit?usp=share_link&ouid=112622933348215332431&rtpof=true&sd=true

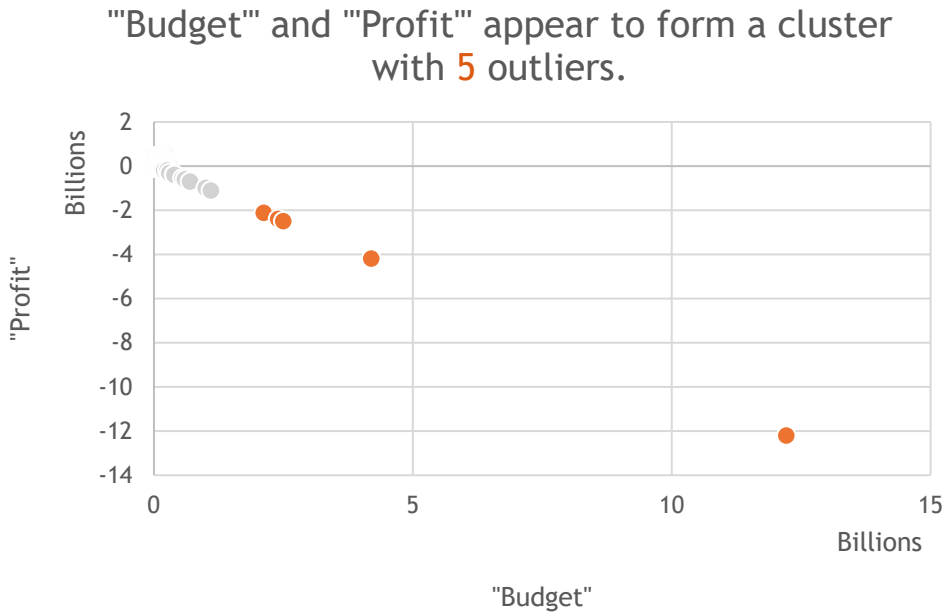
Insights on Most profitable movies

MOVIE_TITLE	PROFIT,,	BUDGET,,	ROI
Avatar	523505847	237000000	220.888543
Jurassic World	502177271	150000000	334.7848473
Titanic	458672302	200000000	229.336151
Star Wars: Episode IV - A New Hope	449935665	11000000	4090.324227
E.T. the Extra-Terrestrial	424449459	10500000	4042.3758
The Avengers	403279547	220000000	183.308885
The Lion King	377783777	45000000	839.5195044
The Jungle Book	375290282	350000000	107.2257949
Star Wars: Episode I - The Phantom Menace	359544677	115000000	312.6475452
The Dark Knight	348316061	185000000	188.2789519
The Hunger Games	329999255	78000000	423.0759679
Twilight	308898950	74000000	417.4310135
Deadpool	305024263	58000000	525.9039017
The Hunger Games: Catching Fire	294645577	130000000	226.6504438
Jurassic Park	293784000	63000000	466.3238095
Despicable Me 2	292049635	76000000	384.2758355
American Sniper	291323553	58800000	495.4482194
Finding Nemo	286838870	94000000	305.147734
Shrek 2	286471036	150000000	190.9806907
The Lord of the Rings: The Return of the King	283019252	94000000	301.0843106




Budget vs Profit Graph
MILLIONS
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
PROFIT
0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000 5500 6000 6500 7000 7500 8000 8500 9000 9500 10000 10500 11000 11500 12000 12500
MILLIONS
BUDGET

Insights about outliers in the profit:

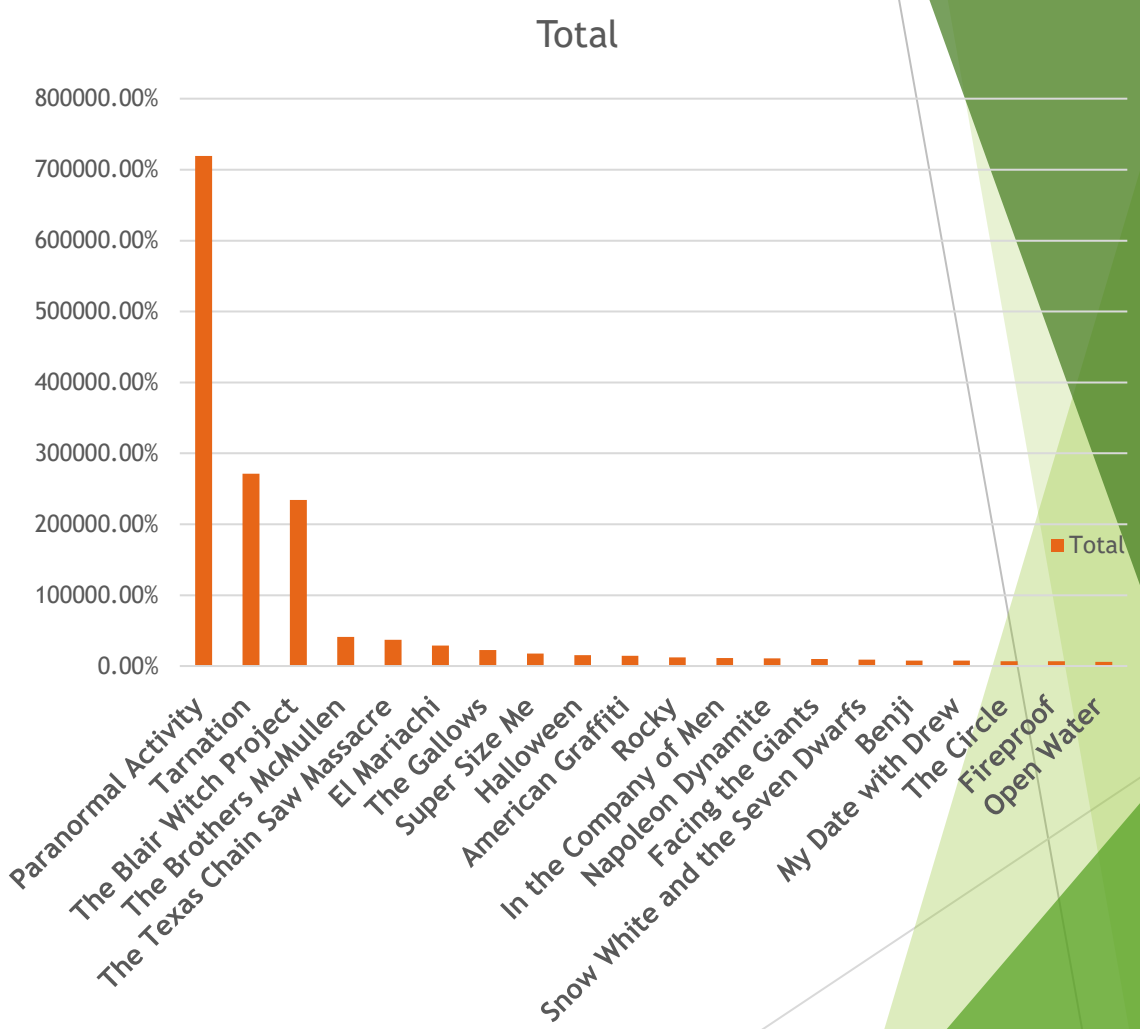


Movie Title	Budget	Profit
Steamboy	2127519898	-2127109510
Princess Mononoke	2400000000	-2397701809
Fateless	2500000000	-2499804112
Lady Vengeance	4200000000	-4199788333
The Host (K)	12215500000	-12213298588

The anomalies detected in the analysis are attributed to the variation in the currency used to report the budget and gross box office collections, resulting in negative profits. Unfortunately, we do not have the necessary information to determine the actual profits of these movies. Notably, most of these anomalies are from Japan, China, and South Korea, all of which use different currencies from the US dollar.

Movies with Highest Returns on Investment:

Movie_title	"ROI"%
Paranormal Activity	719348.55%
Tarnation	271466.06%
The Blair Witch Project	234116.86%
The Brothers McMullen	40886.40%
The Texas Chain Saw Massacre	36842.73%
El Mariachi	29056.00%
The Gallows	22657.82%
Super Size Me	17637.49%
Halloween	15566.67%
American Graffiti	14700.51%
Rocky	12112.00%
In the Company of Men	11326.49%
Napoleon Dynamite	11035.24%
Facing the Giants	10074.66%
Snow White and the Seven Dwarfs	9146.27%
Benji	7810.52%
My Date with Drew	7647.45%
The Circle	6637.80%
Fireproof	6590.30%
Open Water	6000.18%



C) Top Ranked Movies:

- To find the top-ranked movies, we used a formula to assign a rank based on the IMDB score and sort the data.
- We have also removed the rows in the data whose movies has less than 25000 voted for them
- This below formula is used to rank the movies according to their IMDB score:

=RANK(N2,$N$2:$N$2544,0)+COUNTIFS($N$2:N2,N2)-1

is used to rank values in column N (IMDB scores) in descending order, i.e. from highest to lowest. The result of this formula is the rank of the current value (in cell N2) relative to all values in the range N2:N2544.

The RANK function assigns the rank to the current value (in cell N2) by comparing it with all values in the range $N$2:$N$2544. The third argument of the RANK function, which is set to 0, specifies that the rank should be assigned in descending order, i.e. from highest to lowest.

The COUNTIFS function is used to handle the case where there are ties in the rankings, i.e. when multiple movies have the same IMDB score. The COUNTIFS function counts the number of occurrences of the current value (in cell N2) up to the current row. This number is then subtracted by 1 to adjust for the initial rank of 1.

Overall, the formula calculates the rank of the current movie's IMDB score among all the movies in the dataset, taking into account ties.

- Additionally, we sorted the data alphabetically by movie title for movies with the same IMDB score. Then, we filtered the data to show only the top 250 movies and hide unnecessary columns.
- For identifying non-English movies, we filtered out the data with "English" in the language column or used a text filter to select languages other than English. We copied the filtered data into a new sheet.
- Although we could have used the FIND/SEARCH formulas to extract non-English movies, we chose the above method as the dataset was relatively small.

The drive links for both data is given below:

https://docs.google.com/spreadsheets/d/1MOk14iEV7BkMgeyxfkNX8E3lQAOJdjho/edit?usp=share_link&ouid=112622938334821533243 1&rtpof=true&sd=true

These are top 20 ranked movies in the top250 ranked movies

director_name	actor_1_name	movie_title	imdb_score	Rank
Frank Darabont	Morgan Freeman	**The Shawshank Redemption**	9.3	1
Francis Ford Coppola	Al Pacino	**The Godfather**	9.2	2
Christopher Nolan	Christian Bale	**The Dark Knight**	9	3
Francis Ford Coppola	Robert De Niro	**The Godfather: Part II**	9	4
Quentin Tarantino	Bruce Willis	**Pulp Fiction**	8.9	5
Steven Spielberg	Liam Neeson	**Schindler's List**	8.9	6
Sergio Leone	Clint Eastwood	**The Good, the Bad and the Ugly**	8.9	7
Peter Jackson	Orlando Bloom	**The Lord of the Rings: The Return of the King**	8.9	8
David Fincher	Brad Pitt	**Fight Club**	8.8	9
Robert Zemeckis	Tom Hanks	**Forrest Gump**	8.8	10
Christopher Nolan	Leonardo DiCaprio	**Inception**	8.8	11
Irvin Kershner	Harrison Ford	**Star Wars: Episode V - The Empire Strikes Back**	8.8	12
Peter Jackson	Christopher Lee	**The Lord of the Rings: The Fellowship of the Ring**	8.8	13
Fernando Meirelles	Alice Braga	**City of God**	8.7	14
Martin Scorsese	Robert De Niro	**Goodfellas**	8.7	15
Milos Forman	Scatman Crothers	**One Flew Over the Cuckoo's Nest**	8.7	16
Akira Kurosawa	Takashi Shimura	**Seven Samurai**	8.7	17
George Lucas	Harrison Ford	**Star Wars: Episode IV - A New Hope**	8.7	18
Peter Jackson	Christopher Lee	**The Lord of the Rings: The Two Towers**	8.7	19
Lana Wachowski	Keanu Reeves	**The Matrix**	8.7	20

Non English movies in Top250

These are the Top 20 Non English movies that are in top 250 movies

director_name	actor_1_name	movie_title	imdb_score	Rank
Sergio Leone	Clint Eastwood	**The Good, the Bad and the Ugly**	8.9	7
Fernando Meirelles	Alice Braga	**City of God**	8.7	14
Akira Kurosawa	Takashi Shimura	**Seven Samurai**	8.7	17
Hayao Miyazaki	Bunta Sugawara	**Spirited Away**	8.6	26
Majid Majidi	Bahare Seddiqi	**Children of Heaven**	8.5	32
Florian Henckel von Donnersmarck	Sebastian Koch	**The Lives of Others**	8.5	43
Asghar Farhadi	Shahab Hosseini	**A Separation**	8.4	47
Jean-Pierre Jeunet	Mathieu Kassovitz	**AmÃ©lie**	8.4	49
S.S. Rajamouli	Tamannaah Bhatia	**Baahubali: The Beginning**	8.4	51
Wolfgang Petersen	JÃ¼rgen Prochnow	**Das Boot**	8.4	53
Chan-wook Park	Min-sik Choi	**Oldboy**	8.4	55
Hayao Miyazaki	Minnie Driver	**Princess Mononoke**	8.4	57
Oliver Hirschbiegel	Thomas Kretschmann	**Downfall**	8.3	65
Fritz Lang	Brigitte Helm	**Metropolis**	8.3	72
Thomas Vinterberg	Thomas Bo Larsen	**The Hunt**	8.3	79
Hayao Miyazaki	Christian Bale	**Howl's Moving Castle**	8.2	94
Denis Villeneuve	Lubna Azabal	**Incendies**	8.2	95
Guillermo del Toro	Ivana Baquero	**Pan's Labyrinth**	8.2	99
Juan JosÃ© Campanella	Ricardo DarÃn	**The Secret in Their Eyes**	8.2	102
Katsuhiro Ã”tomo	Mitsuo Iwata	**Akira**	8.1	109

There were some movies that had comparable IMDB scores but were not included in the list as they were not ranked high enough due to the sorting criteria based on the alphabetical order of the items.

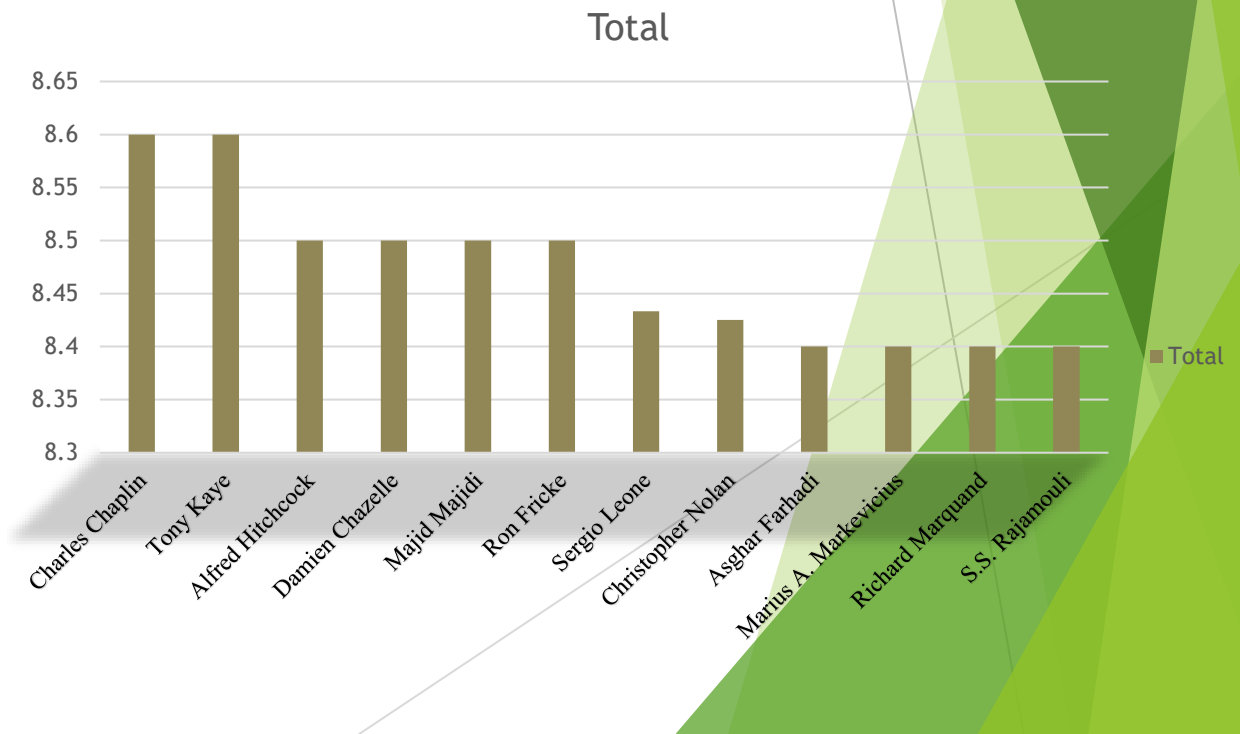
director_name	actor_1_name	movie_title	imdb_score	Rank
John Lasseter	Tom Hanks	**Toy Story 2**	7.9	251
Yash Chopra	Shah Rukh Khan	**Veer-Zaara**	7.9	252
James Mangold	Sandra Ellis Lafferty	**Walk the Line**	7.9	253

D)Top Directors

The objective of this task is to identify the top 10 directors based on their average IMDB score. To achieve this, we will use the pivot table and select two columns - director name and average IMDB score for each director. Then, we will use the value filter option to narrow down the selection to the top 10 directors.

To present our findings, we will use a pivot chart. It's important to note that in cases where multiple directors have the same average IMDB score, the table may display more than 10 directors.

Director Name	Average 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
Richard Marquand	8.4
S.S. Rajamouli	8.4
Grand Total	8.452380952

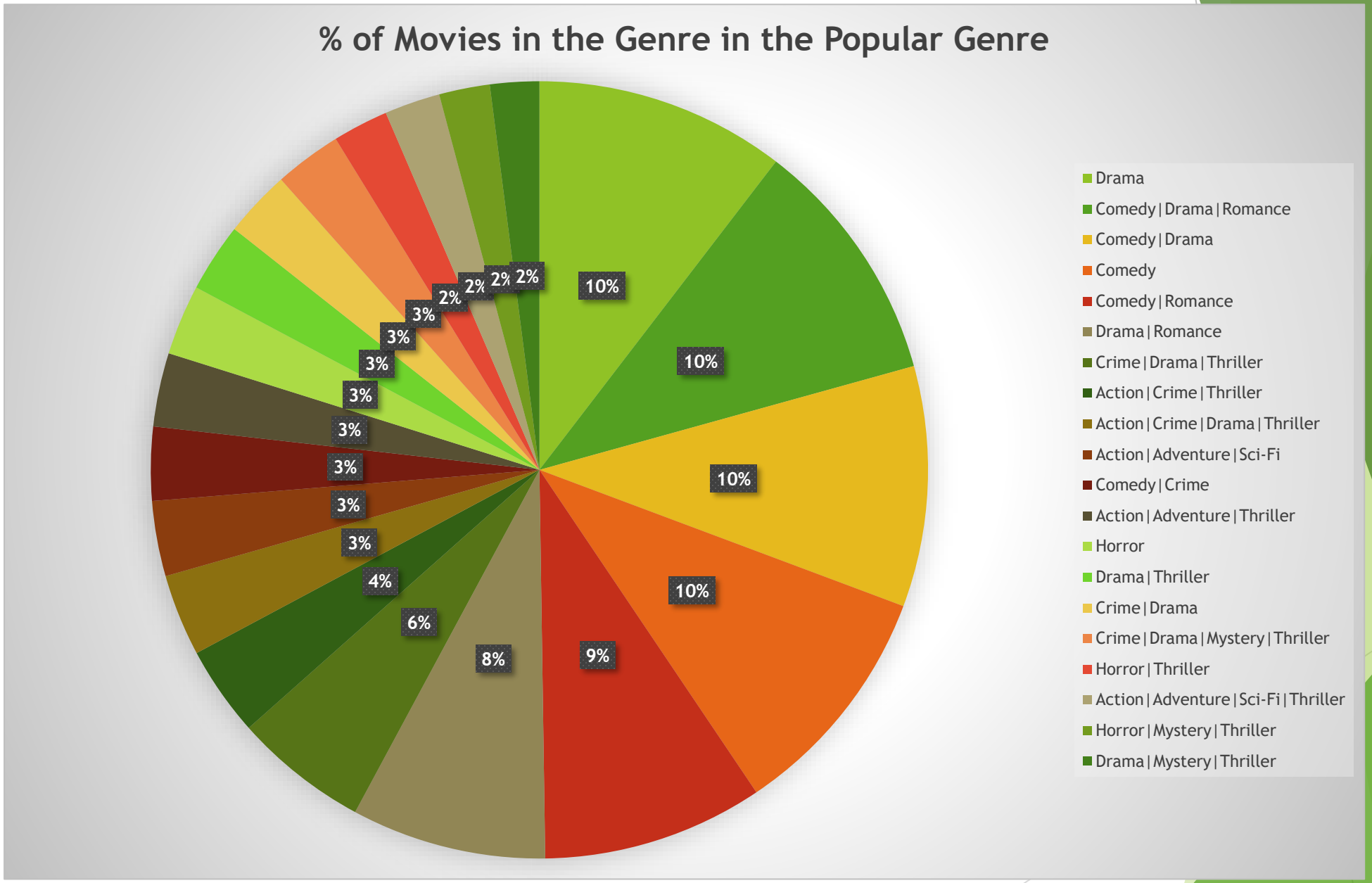


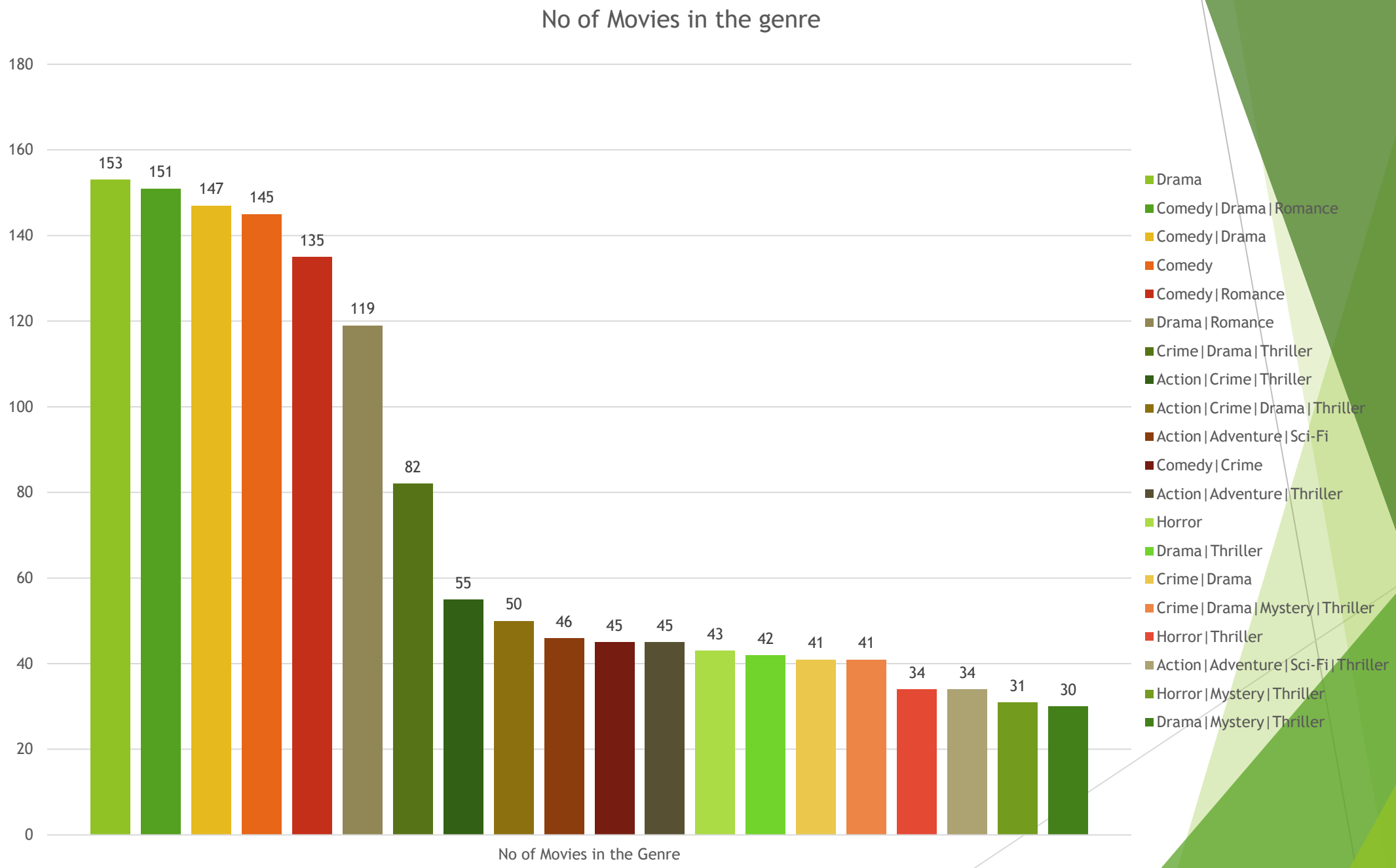
E) Popular Genre

Our main focus was to determine the most popular movie genre using the data available in the genre column. To achieve this, we started by selecting the pivot table command and choosing the genre and movie title columns. The pivot table allowed us to count the number of movie titles for each genre.

After creating the pivot table, we used the pivot chart command to represent our findings visually. To provide a more comprehensive analysis, we decided to use both pie chart and column graph to represent the insights.

Genre	No of Movies in the Genre
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
Action |Adventure |Thriller	45
Comedy |Crime	45
Horror	43
Drama |Thriller	42
Crime| Drama	41
Crime| Drama| Mystery| Thriller	41
Action| Adventure| Sci-Fi| Thriller	34
Horror| Thriller	34
Horror| Mystery| Thriller	31
Biography| Drama	30
Drama| Mystery| Thriller	30


% of Movies in the Genre in the Popular Genre
10%
10%
10%
10%
9%
8%
6%
4%
3%
3%
3%
3%
3%
3%
3%
2%
2%
2%
2%
Drama
Comedy | Drama | Romance
Comedy | Drama
Comedy
Comedy | Romance
Drama | Romance
Crime | Drama | Thriller
Action | Crime | Thriller
Action | Crime | Drama | Thriller
Action | Adventure | Sci-Fi
Comedy | Crime
Action | Adventure | Thriller
Horror
Drama | Thriller
Crime | Drama
Crime | Drama | Mystery | Thriller
Horror | Thriller
Action | Adventure | Sci-Fi | Thriller
Horror | Mystery | Thriller
Drama | Mystery | Thriller


No of Movies in the genre
180
160
140
120
100
80
60
40
20
0
153
151
147
145
135
119
82
55
50
46
45
45
43
42
41
41
34
34
31
30
No of Movies in the Genre
Drama
Comedy | Drama | Romance
Comedy | Drama
Comedy
Comedy | Romance
Drama | Romance
Crime | Drama | Thriller
Action | Crime | Thriller
Action | Crime | Drama | Thriller
Action | Adventure | Sci-Fi
Comedy | Crime
Action | Adventure | Thriller
Horror
Drama | Thriller
Crime | Drama
Crime | Drama | Mystery | Thriller
Horror | Thriller
Action | Adventure | Sci-Fi | Thriller
Horror | Mystery | Thriller
Drama | Mystery | Thriller

In addition to our previous analysis, we have further explored the movies dataset by examining the individual basic genres present in the columns. We accomplished this by using the TEXTSPLIT formula and analyzing the resulting data using the "Analyze data" button. The most basic genres we identified include Action, Adventure, Drama, and Comedy.

Genre1	Count of Genre	Count of Genre2	Count of Genre3	Count of Genre4	Count of Genre5	Count of Genre6	Count of Genre7	Count of Genre8	Number of films
Action	962	956	881	528	200	45	11	2	3585
Adventure	375	373	335	224	114	28	8	1	1458
Crime	257	257	199	77	5				795
Drama	691	538	271	93	9				1602
Documentary	40	19	4						63
Animation	46	46	46	38	22	8			206
Biography	207	207	176	72	15	1			678
Comedy	1029	884	462	123	23	4			2525
Family	3	3	2						8
Fantasy	37	37	20	9	1	1			105
Horror	160	117	52	5					334
Musical	2	2							4
Mystery	23	23	7						53
Romance	3	1	1						5
Sci-Fi	8	7							15
Thriller	3								3
Western	3								3

F) Charts:

We have copied the cleaned data into a newssheet where inserted a new column ‘deacdes’

Then we created new columns Meryl_Streep, Leo_Caprio, and Brad_Pitt and fill the columns with the movies they acted in two methods. They are the following:

1.By using formulae:

```
=IFERROR(INDEX(Table96[movie_title], SMALL(IF(Table96[actor_1_name]="Meryl Streep",
ROW(Table96[actor_1_name])-ROW(Table96[[#Headers],[actor_1_name]])), ROWS($A$1:A1))), "")

=IFERROR(INDEX(Table96[movie_title], SMALL(IF(Table96[actor_1_name]="Leonardo DiCaprio",
ROW(Table96[actor_1_name])-ROW(Table96[[#Headers],[actor_1_name]])), ROWS($A$1:A1))), "")

=IFERROR(INDEX(Table96[movie_title], SMALL(IF(Table96[actor_1_name]="Brad Pitt",
ROW(Table96[actor_1_name])-ROW(Table96[[#Headers],[actor_1_name]])), ROWS($A$1:A1))), "")
```

Then we have created a new column ‘Combined’ in which we have appended all the movies in a single column

Note: we performed this method in “sheet 1”

Here is how the formula works:

The INDEX function returns a value from a table or range of cells specified by the user. In this case, it is returning the values in the "movie_title" column of Table96.

The IF function is used to check if the "actor_1_name" column of Table96 matches the name "Meryl Streep". If it does, the formula returns the row number of that cell using the ROW function. If it doesn't, it returns a blank cell.

The ROWS function keeps track of the number of times the formula is being dragged down, and increases by one each time. This allows the SMALL function to return the next smallest value from the list of row numbers generated by the IF function.

The SMALL function returns the nth smallest value from a list of values, where n is the row number of the cell in which the formula is entered. The first time the formula is entered, it returns the smallest value from the list generated by the IF function. The second time, it returns the second smallest value, and so on.

The formula is wrapped in an IFERROR statement, which returns a blank cell if an error occurs, such as when the formula runs out of values to return.

Overall, the formula searches for movies starring Meryl Streep in the "actor_1_name" column of Table96 and returns the titles of those movies in a list, without any blank spaces.

2. By using filter:

Here we use the filter option in the table to find the rows containing name of the respective order then we hide the columns except actor name and movie title and sort them using their names and group them

Then we move the respective data into respective columns and combined column Note: We performed this in "BML appended" sheet. The below table contains the columns we need to create and fill we have used different colour for filling the movie title of different actors movies. The following is link for this analysis:

https://docs.google.com/spreadsheets/d/17aFpH9QEX9i0FURGggz_8vcxU2WwWW_Z/edit?usp=share_link&ouid=112622933348215332431&rtpof=true&sd=true

Meryl_Streep	Leo_Caprio	Brad_Pitt	Combined
The Hours	Inception	Fight Club	The Hours
Out of Africa	Django Unchained	True Romance	Out of Africa
Julie & Julia	The Departed	Ocean's Eleven	Julie & Julia
One True Thing	The Wolf of Wall Street	The Curious Case of Benjamin Button	One True Thing
A Prairie Home Companion	Shutter Island	Fury	A Prairie Home Companion
The Devil Wears Prada	The Revenant	Interview with the Vampire: The Vampire Chronicles	The Devil Wears Prada
It's Complicated	Blood Diamond	Babel	It's Complicated
The Iron Lady	Catch Me If You Can	The Assassination of Jesse James by the Coward Robert Ford	The Iron Lady
Hope Springs	Titanic	Troy	Hope Springs
The River Wild	Gangs of New York	Seven Years in Tibet	The River Wild
Lions for Lambs	The Aviator	Spy Game	Lions for Lambs
	The Great Gatsby	Sinbad: Legend of the Seven Seas	Inception
	Revolutionary Road	The Tree of Life	Django Unchained
	Body of Lies	Mr. & Mrs. Smith	The Departed
	Romeo + Juliet	Ocean's Twelve	The Wolf of Wall Street

https://docs.google.com/spreadsheets/d/17aFpH9QEX9i0FURGggz_8vcxU2WwWW_Z/edit?usp=share_link&ouid=112622933348215332431&rtpof=true&sd=true

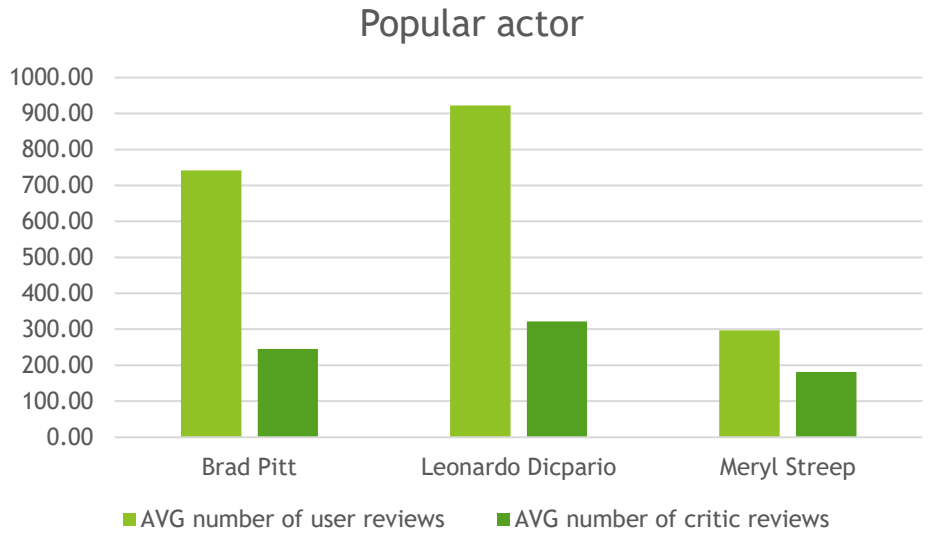
Meryl_Streep	Leo_Caprio	Brad_Pitt	Combined
	Marvin's Room	Killing Them Softly	Shutter Island
	J. Edgar		The Revenant
	The Beach		Blood Diamond
	The Man in the Iron Mask		Catch Me If You Can
	The Quick and the Dead		Titanic
			Body of Lies
			Romeo + Juliet
			Marvin's Room
			J. Edgar
			The Beach
			The Man in the Iron Mask
			The Quick and the Dead
			Fight Club
			True Romance
			Ocean's Eleven
			The Curious Case of Benjamin Button
			Fury
			Interview with the Vampire: The Vampire Chronicles
			Babel
			The Assassination of Jesse James by the Coward Robert Ford
			Troy
			Seven Years in Tibet
			Spy Game
			Sinbad: Legend of the Seven Seas
			The Tree of Life
			Mr. & Mrs. Smith
			Ocean's Twelve
			Killing Them Softly

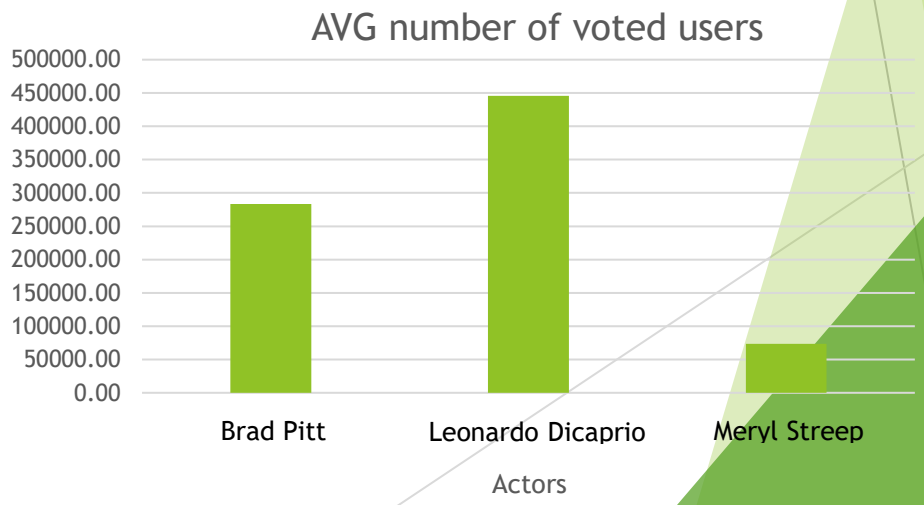
We utilized various columns, including 'num_critic_for_reviews', 'num_user_for_reviews', and 'num_voted_users', to calculate the mean value of our actors. This was achieved by applying filter functions on the BML appended sheet and unhiding the relevant columns.

Actor Name	AVG number of user reviews	AVG number of critic reviews	AVG number of voted users
Brad Pitt	742.35	245.00	283583.82
Leonardo Dicpario	922.55	322.20	445911.20
Meryl Streep	297.18	181.45	73545.55

We have mainly used AVERAGE formula for the above table to obtain th necessary data

Then by using pivot chart we draw the below graphs

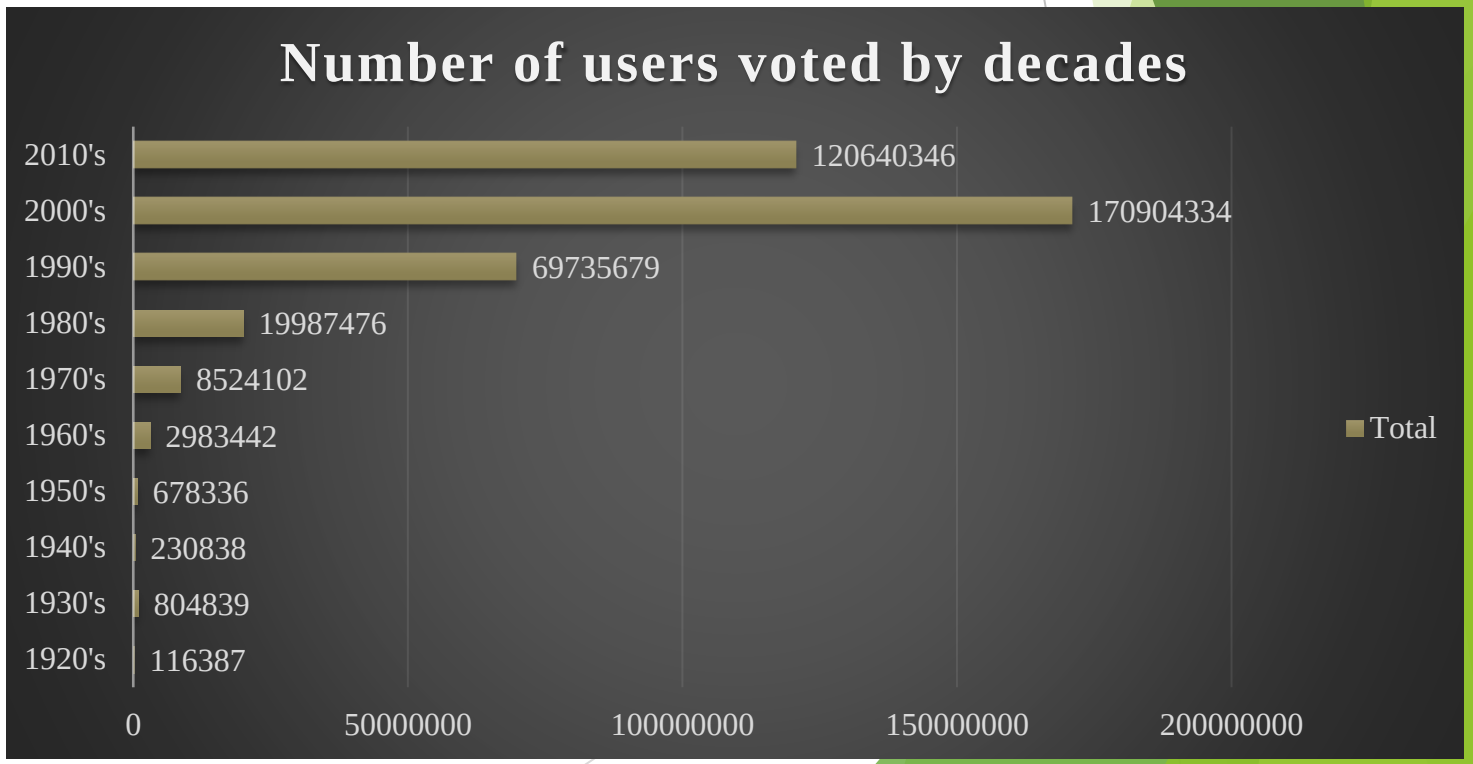




Df_by_decade:

In this task, we have added a new column named "decades" to sheet 1 using the CONCAT and LEFT formula as follows: =CONCAT(LEFT(M2,3),"0's") This formula provides the decade in a format such as 1920's, 1930's, and so on.We have used this column to analyze how the number of voted users per decade has been changing by creating a pivot chart. The pivot chart represents the insights visually, making it easier to observe the changes in the number of voted users over the decades. At last, we have stored the data in a new sheet named "df-by_decade"

Decades	Sum of num_voted_users
1920's	116387
1930's	804839
1940's	230838
1950's	678336
1960's	2983442
1970's	8524102
1980's	19987476
1990's	69735679
2000's	170904334
2010's	120640346
Grand Total	**394605779**



To find actors with high average number of critic and user reviews, we utilized pivot tables. We selected the actor name column along with the columns 'num_critic_for_reviews' and 'num_user_for_reviews' to get the required data. We then used the 'Value Filters' option to filter out the actors with good reviews. The top 20 actors were selected based on their average critic and user reviews.

Top 20 actors by average num of critic reviews

Actor Name	Average of num_critic_for_reviews
Albert Finney	750
Phaldut Sharma	738
Peter Capaldi	654
Craig Stark	596
Bérénice Bejo	576
Suraj Sharma	552
Ellar Coltrane	548
Mike Howard	546
Lou Taylor Pucci	543
Joel Courtney	539
Maika Monroe	533
Tim Holmes	525
Elina Alminas	489
Kurt Fuller	487
Iko Uwais	481
Quvenzhané Wallis	478.6666667
Edgar Arreola	478
Sharlto Copley	472
Cory Hardrict	452
Matt Frewer	451

Top 20 actors by average num of user reviews

Actor Name	Average of num_user_for_reviews
Heather Donahue	3400
Christo Jivkov	2814
Steve Bastoni	2789
Phaldut Sharma	1885
Orlando Bloom	1842
Keir Dullea	1736
Chen Chang	1641
Nick Stahl	1562
Albert Finney	1498
Kevin Rankin	1445
Noah Huntley	1441
Osama bin Laden	1416
Eva Green	1412
Seychelle Gabriel	1382
Mathieu Kassovitz	1314
Essie Davis	1285.5
Sharlto Copley	1262
Giancarlo Giannini	1243
Christopher Lee	1237.142857
Matt Frewer	1229

Result:

These solutions required to answer about the data set has been developed and insights have been provided In this paragraph, we have further explained how the 5 why's helped us in the analysis:

The 5 why questions provided a structured approach to problem-solving and helped in developing and framing used in this project. By asking "what do you see happening?" we were able to identify the issue that needed to be solved. "What is your hypothesis for the cause of the problem?" helped in brainstorming potential causes and led to the discovery that the currency of the budget and gross box office collections was different. "What is the impact of the problem on stakeholders?" helped us understand the significance of the issue and why it needed to be resolved. "What is the impact of the problem not being solved?" helped in recognizing the consequences of not addressing the issue, which could result in misleading financial information. Overall, the 5 questions provided a comprehensive framework for problem-solving and contributed to the development of the analysis in this project.

Link for Analysis:
https://docs.google.com/spreadsheets/d/16z8y8GXT_EvnOEpvuE5ZrMtRr_w5JoN5/edit?usp=share_link&ouid=112622933348215332431&rtpof=true&sd=true

Thank You