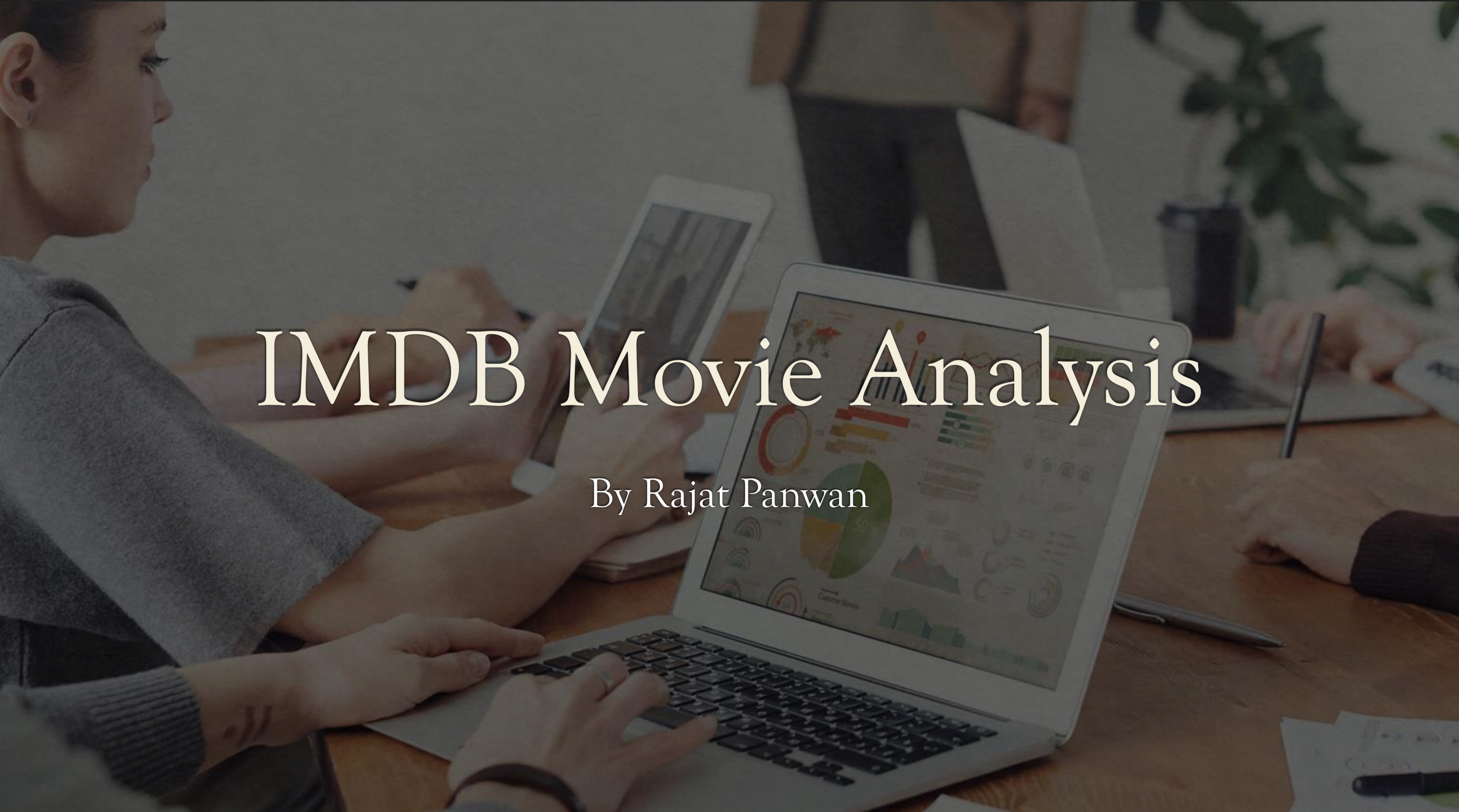


IMDB Movie Analysis

By Rajat Panwan

Project Description

The aim of this project is to perform an analysis on a dataset containing information on various movies from IMDB. The dataset includes columns such as the director name, gross, genres, movie title, num voted users, plot keywords, num user for reviews, language, rating, budget, IMDB score etc. The main objective is to extract useful insights from the data and identify any trends or patterns that can be useful for decision-making.

Approach

The project involved several steps including data cleaning, data visualization, and statistical analysis. Initially, the dataset was explored to identify any missing values, outliers, or errors. The data was then cleaned using various techniques such as removing duplicates and correcting errors. Data visualization tools were used to create charts, graphs, and histograms to analyze the data.

Tech-Stack Used

The project was performed using Excel version 2021. Excel was chosen for its powerful data analysis and visualization capabilities. It is also widely used in the industry and provides a familiar environment for users.

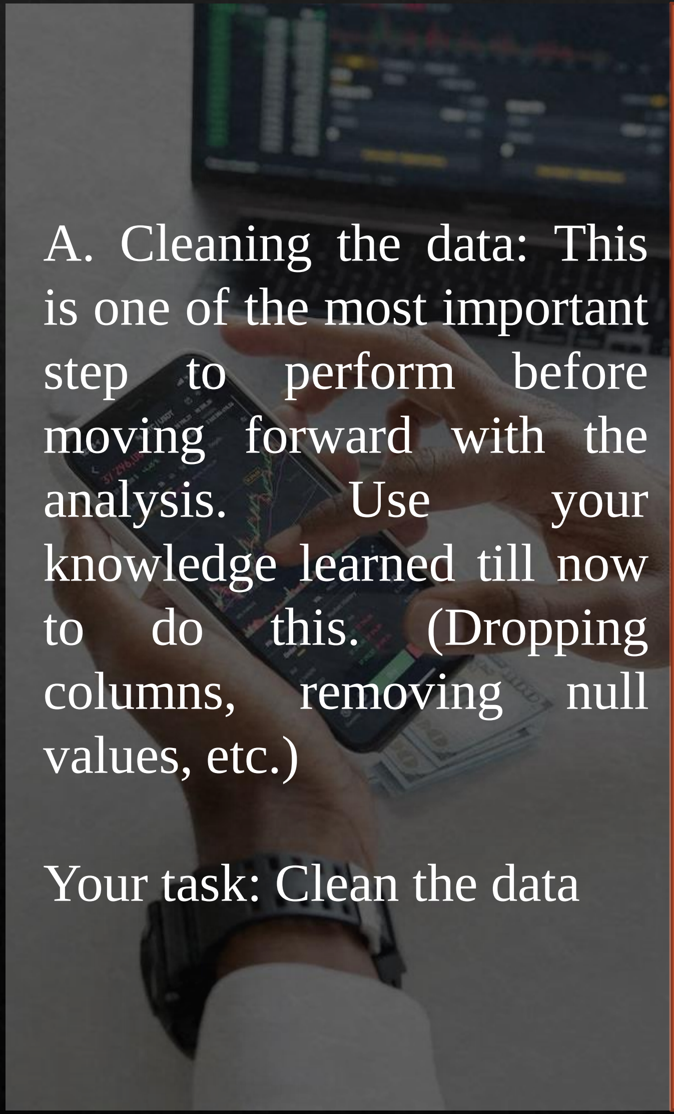
Insights

Several useful insights were obtained from the analysis. For example, Jurassic World made the highest profit . The analysis also showed that movie Shawshank Redemption is highest rated movie on IMDb. Another interesting finding was that movies directed by Cary Bell and Akira Kurosawa tended to have higher ratings than others. The analysis also identified certain genres that were more popular than others.

Result

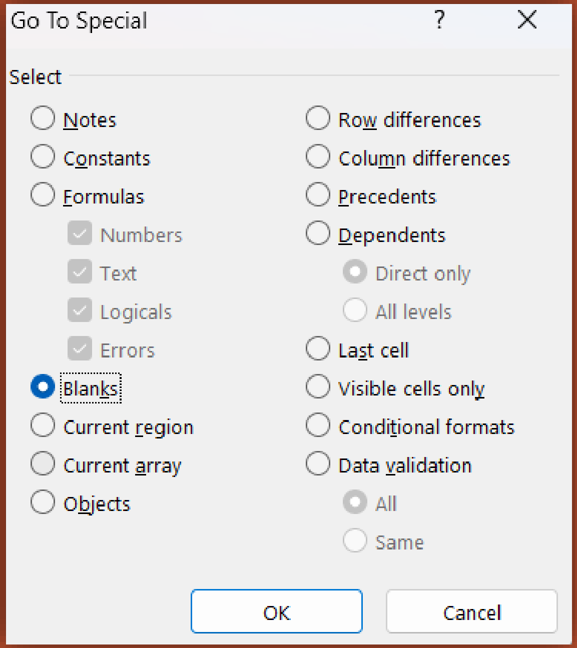
The project was successful in identifying several useful insights from the data. These insights can be useful for decision-making in the movie industry. Results are on next pages.

Result



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.)

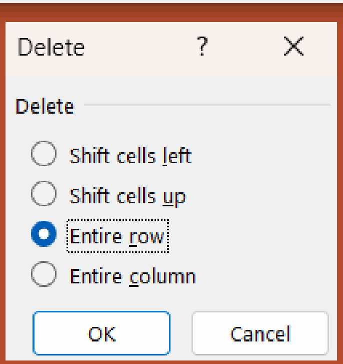
Your task: Clean the data



Select

○ Notes
○ Constants
○ Formulas
☑ Numbers
☑ Text
☑ Logicals
☑ Errors
● Blanks
○ Current region
○ Current array
○ Objects

○ Row differences
○ Column differences
○ Precedents
○ Dependents
● Direct only
○ All levels
○ Last cell
○ Visible cells only
○ Conditional formats
○ Data validation
● All
○ Same

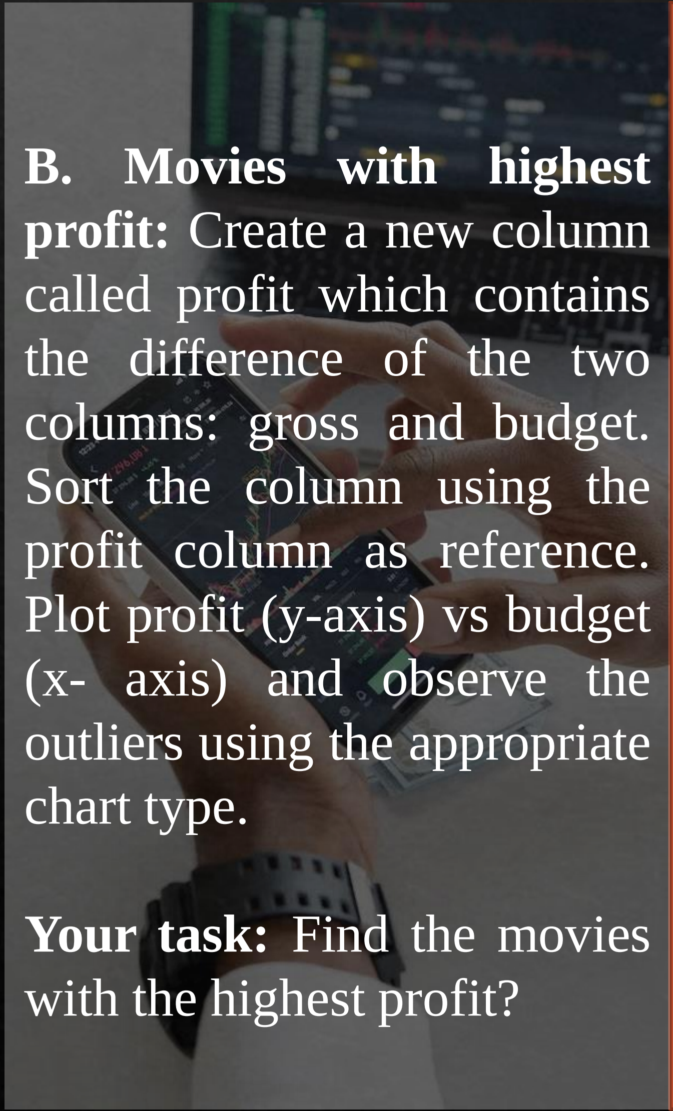


Delete

○ Shift cells left
○ Shift cells up
● Entire row
○ Entire column

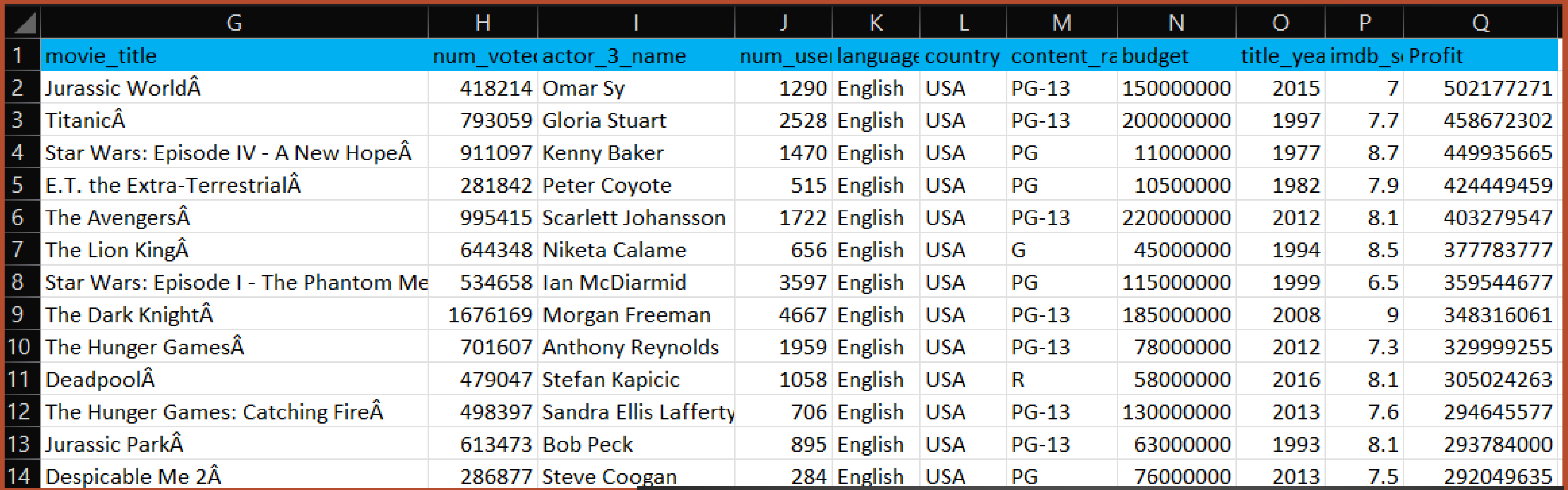
- Deleted the column which are not required in our analysis.
- Then delete all the rows that have bank values in any rows.
- After deleting only 3956 rows are left.

Result

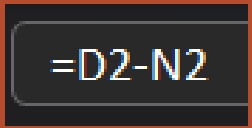


B. Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

Your task: Find the movies with the highest profit?



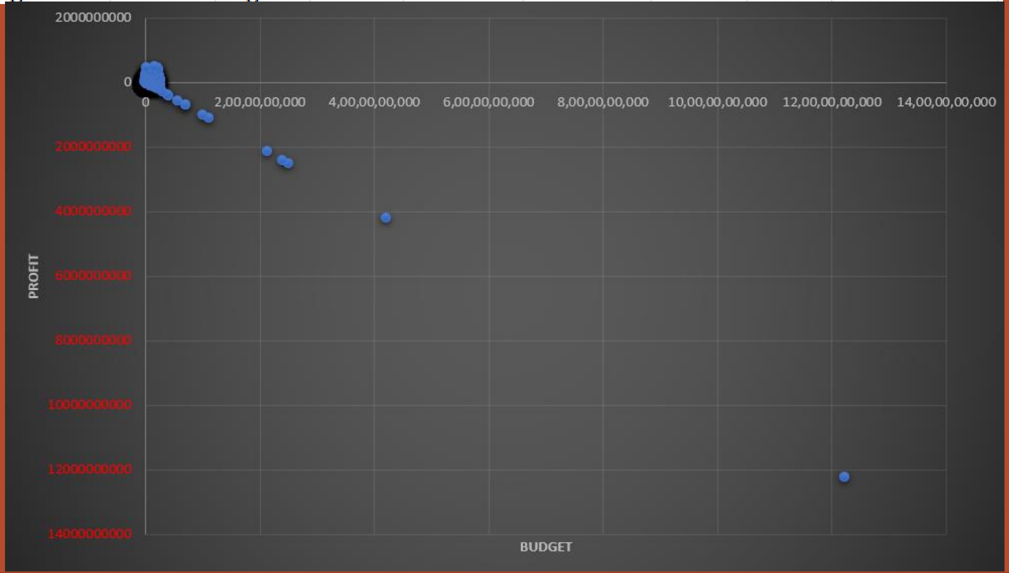
	G	H	I	J	K	L	M	N	O	P	Q
1	movie_title	num_votec	actor_3_name	num_user	language	country	content_ra	budget	title_yea	imdb_s	Profit
2	Jurassic WorldÂ	418214	Omar Sy	1290	English	USA	PG-13	150000000	2015	7	502177271
3	TitanicÂ	793059	Gloria Stuart	2528	English	USA	PG-13	200000000	1997	7.7	458672302
4	Star Wars: Episode IV - A New HopeÂ	911097	Kenny Baker	1470	English	USA	PG	11000000	1977	8.7	449935665
5	E.T. the Extra-TerrestrialÂ	281842	Peter Coyote	515	English	USA	PG	10500000	1982	7.9	424449459
6	The AvengersÂ	995415	Scarlett Johansson	1722	English	USA	PG-13	220000000	2012	8.1	403279547
7	The Lion KingÂ	644348	Niketa Calame	656	English	USA	G	45000000	1994	8.5	377783777
8	Star Wars: Episode I - The Phantom Me	534658	Ian McDiarmid	3597	English	USA	PG	115000000	1999	6.5	359544677
9	The Dark KnightÂ	1676169	Morgan Freeman	4667	English	USA	PG-13	185000000	2008	9	348316061
10	The Hunger GamesÂ	701607	Anthony Reynolds	1959	English	USA	PG-13	78000000	2012	7.3	329999255
11	DeadpoolÂ	479047	Stefan Kapicic	1058	English	USA	R	58000000	2016	8.1	305024263
12	The Hunger Games: Catching FireÂ	498397	Sandra Ellis Lafferty	706	English	USA	PG-13	130000000	2013	7.6	294645577
13	Jurassic ParkÂ	613473	Bob Peck	895	English	USA	PG-13	63000000	1993	8.1	293784000
14	Despicable Me 2Â	286877	Steve Coogan	284	English	USA	PG	76000000	2013	7.5	292049635



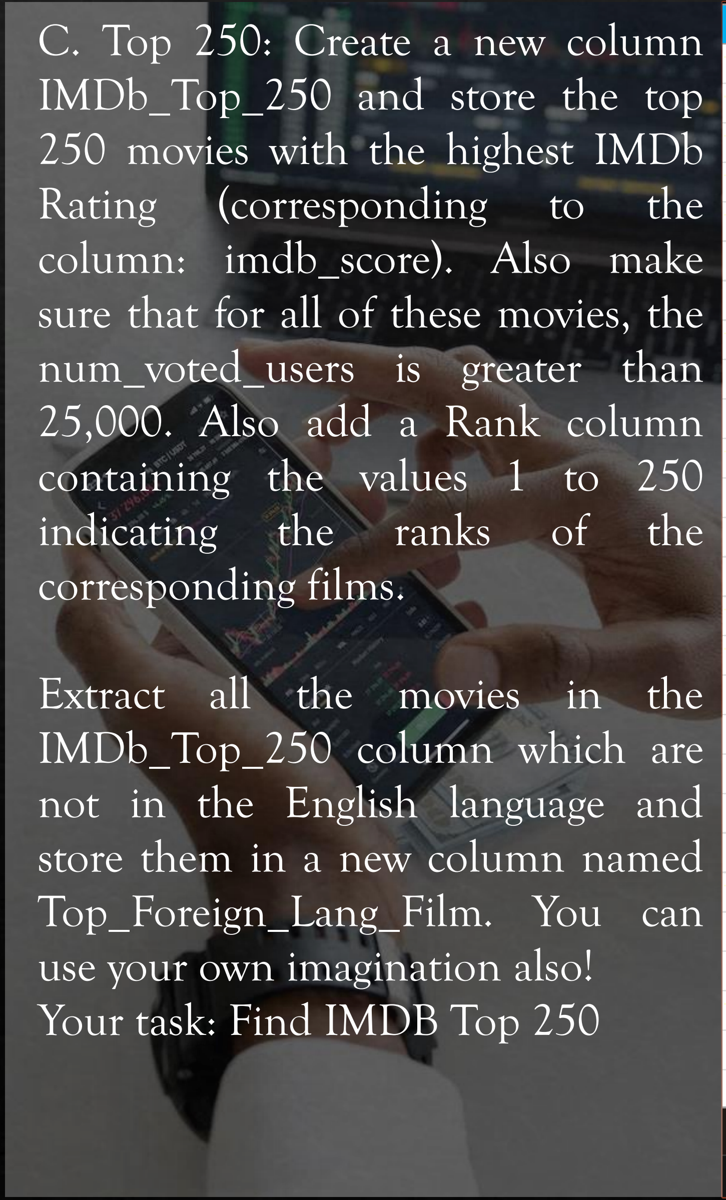
=D2-N2

This formula used to get profit

- Jurassic World has made the maximum profit.
- The host is the outlier with budget 12,215M and loss of 12,213M



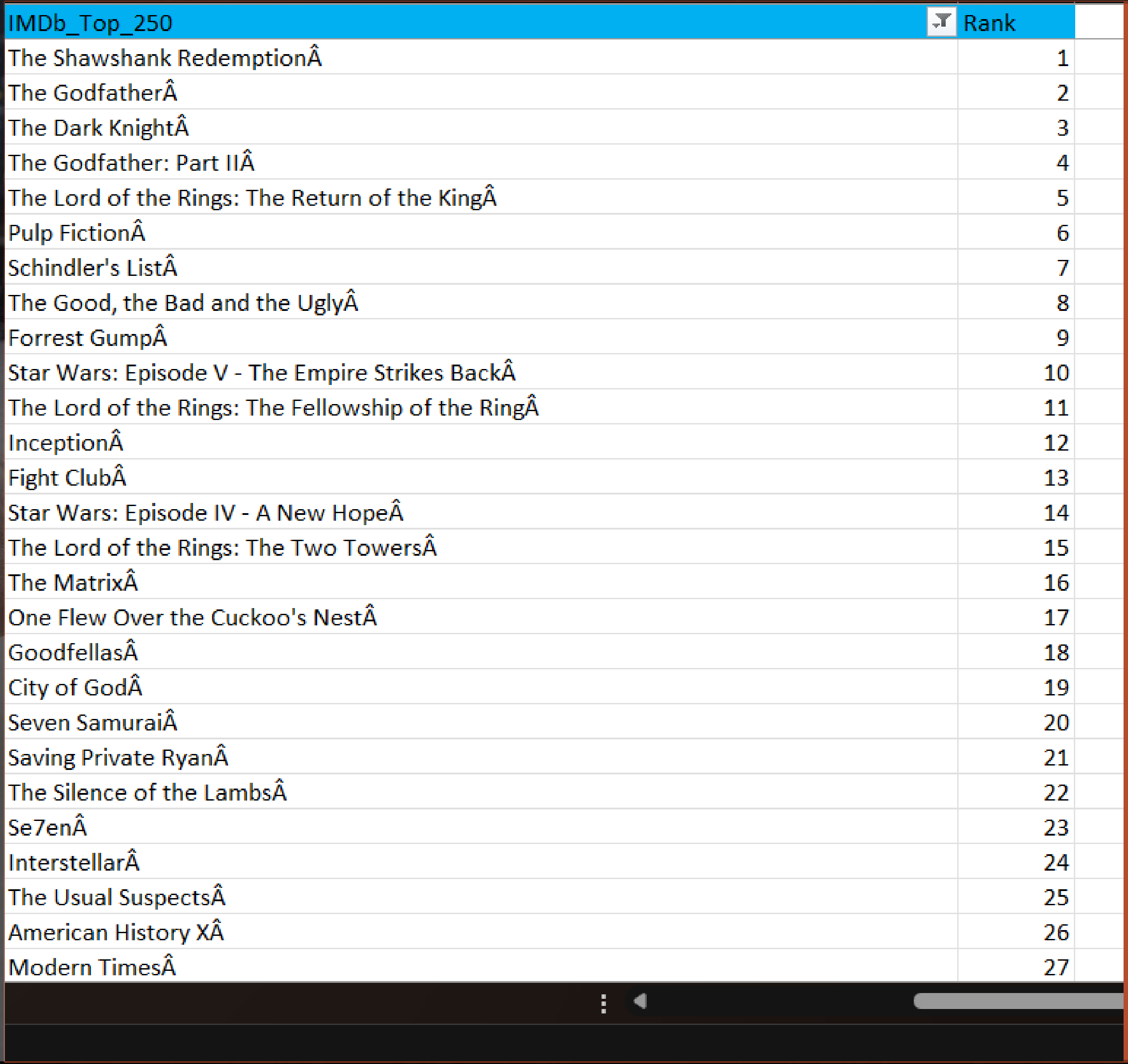
Result



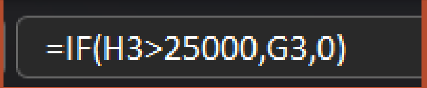
C. Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the 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

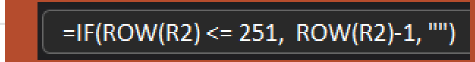


IMDb_Top_250	Rank
The Shawshank RedemptionÂ	1
The GodfatherÂ	2
The Dark KnightÂ	3
The Godfather: Part IIÂ	4
The Lord of the Rings: The Return of the KingÂ	5
Pulp FictionÂ	6
Schindler's ListÂ	7
The Good, the Bad and the UglyÂ	8
Forrest GumpÂ	9
Star Wars: Episode V - The Empire Strikes BackÂ	10
The Lord of the Rings: The Fellowship of the RingÂ	11
InceptionÂ	12
Fight ClubÂ	13
Star Wars: Episode IV - A New HopeÂ	14
The Lord of the Rings: The Two TowersÂ	15
The MatrixÂ	16
One Flew Over the Cuckoo's NestÂ	17
GoodfellasÂ	18
City of GodÂ	19
Seven SamuraiÂ	20
Saving Private RyanÂ	21
The Silence of the LambsÂ	22
Se7enÂ	23
InterstellarÂ	24
The Usual SuspectsÂ	25
American History XÂ	26
Modern TimesÂ	27



```
=IF(H3>25000,G3,0)
```

The above formula is for getting the name of the movies that have num_voted_users is greater than 25000 and than filter out movies that doen't come under the criteria.

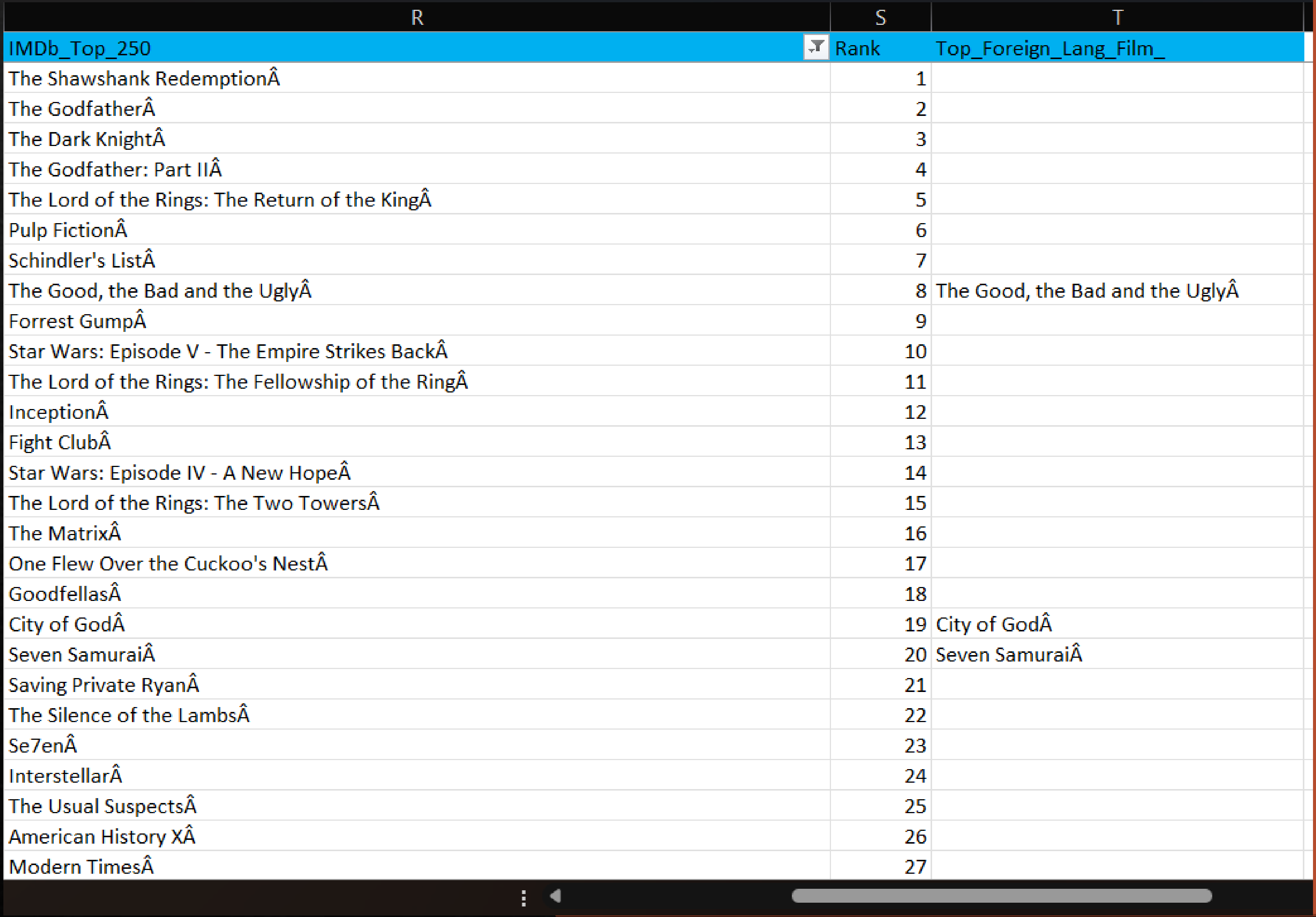


```
=IF(ROW(R2) <= 251,  ROW(R2)-1, "")
```

For ranking row function is used after sorting the imdb_score highest to lowest.

Cont.

Result

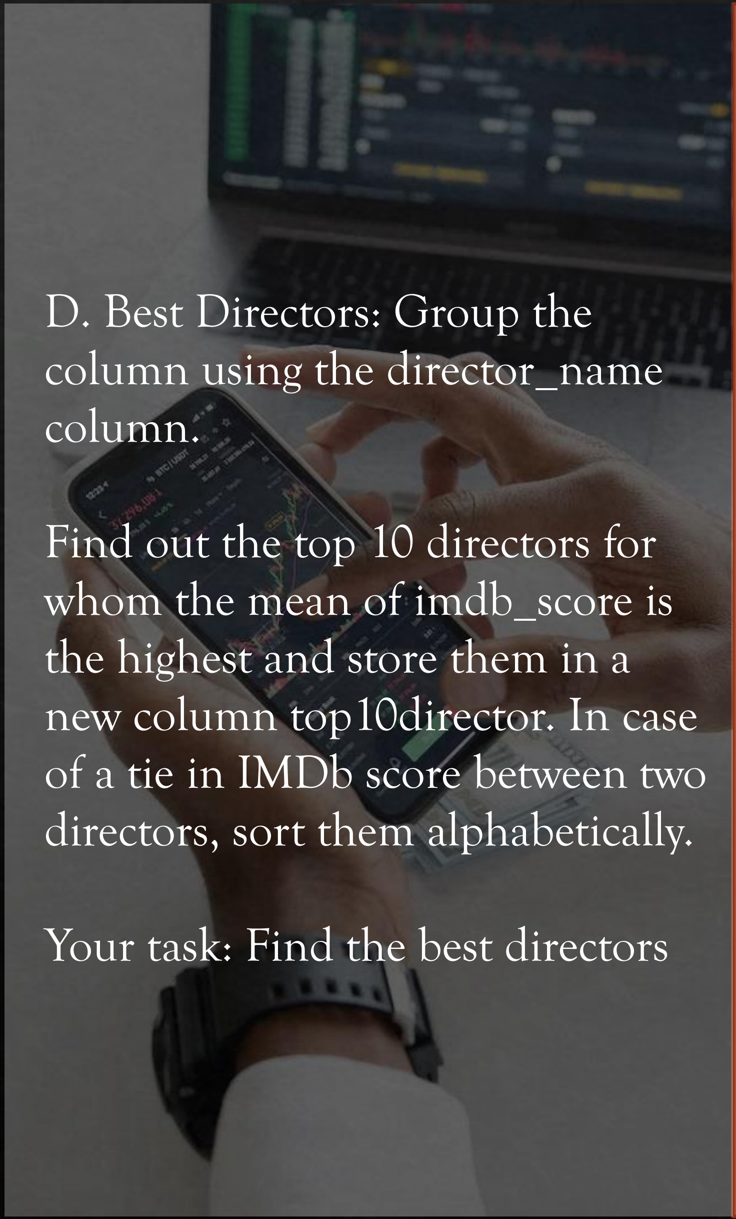


R	S	T
IMDb_Top_250	Rank	Top_Foreign_Lang_Film_
The Shawshank RedemptionÂ	1	
The GodfatherÂ	2	
The Dark KnightÂ	3	
The Godfather: Part IIÂ	4	
The Lord of the Rings: The Return of the KingÂ	5	
Pulp FictionÂ	6	
Schindler's ListÂ	7	
The Good, the Bad and the UglyÂ	8	The Good, the Bad and the UglyÂ
Forrest GumpÂ	9	
Star Wars: Episode V - The Empire Strikes BackÂ	10	
The Lord of the Rings: The Fellowship of the RingÂ	11	
InceptionÂ	12	
Fight ClubÂ	13	
Star Wars: Episode IV - A New HopeÂ	14	
The Lord of the Rings: The Two TowersÂ	15	
The MatrixÂ	16	
One Flew Over the Cuckoo's NestÂ	17	
GoodfellasÂ	18	
City of GodÂ	19	City of GodÂ
Seven SamuraiÂ	20	Seven SamuraiÂ
Saving Private RyanÂ	21	
The Silence of the LambsÂ	22	
Se7enÂ	23	
InterstellarÂ	24	
The Usual SuspectsÂ	25	
American History XÂ	26	
Modern TimesÂ	27	

```
=IF(AND(K2<>"English",S2<250), R2, "")
```

In adjacent picture column T have the movies name that is extracted from Imdb_Top_250 which are not in English Language.

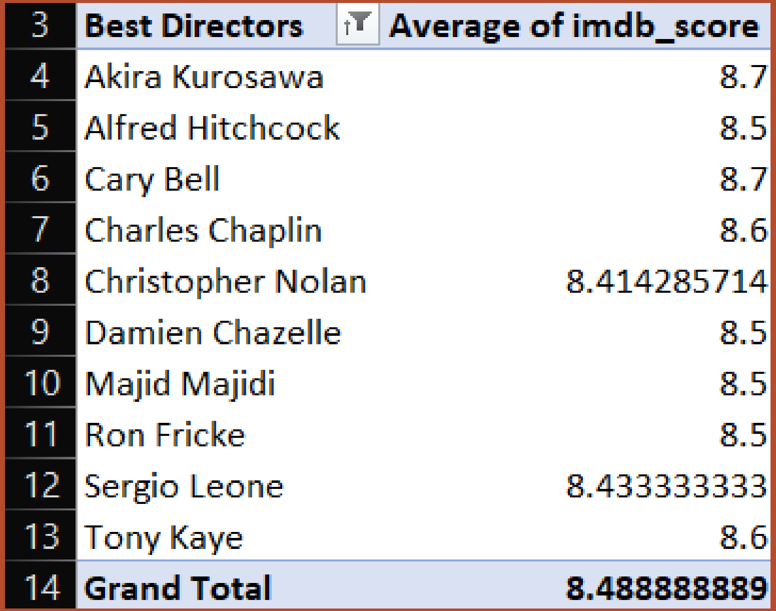
Result



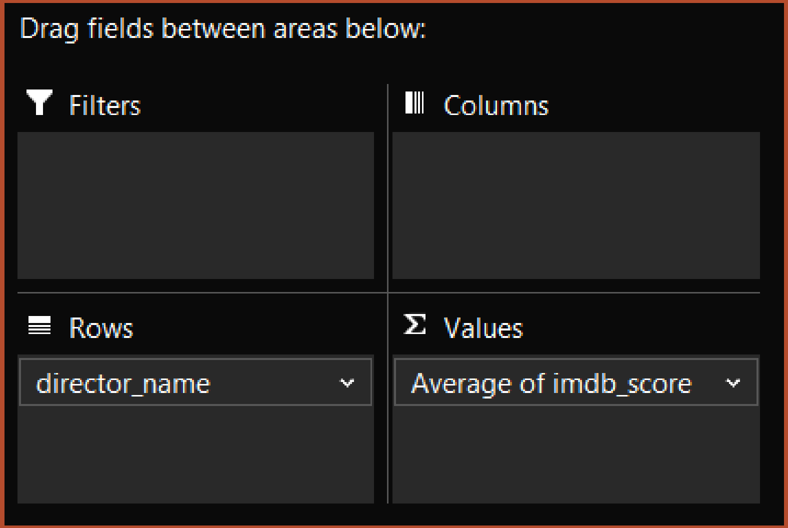
D. Best Directors: Group the column using the director_name column.

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



3	Best Directors	Average of imdb_score
4	Akira Kurosawa	8.7
5	Alfred Hitchcock	8.5
6	Cary Bell	8.7
7	Charles Chaplin	8.6
8	Christopher Nolan	8.414285714
9	Damien Chazelle	8.5
10	Majid Majidi	8.5
11	Ron Fricke	8.5
12	Sergio Leone	8.433333333
13	Tony Kaye	8.6
14	**Grand Total**	**8.488888889**



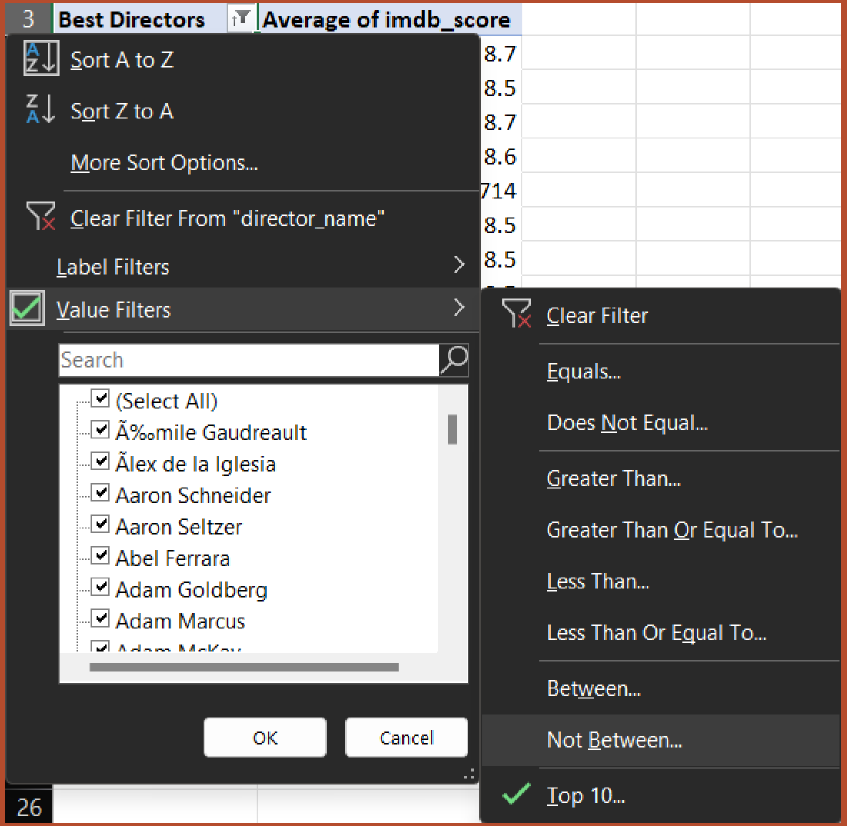
Drag fields between areas below:

Filters

Columns

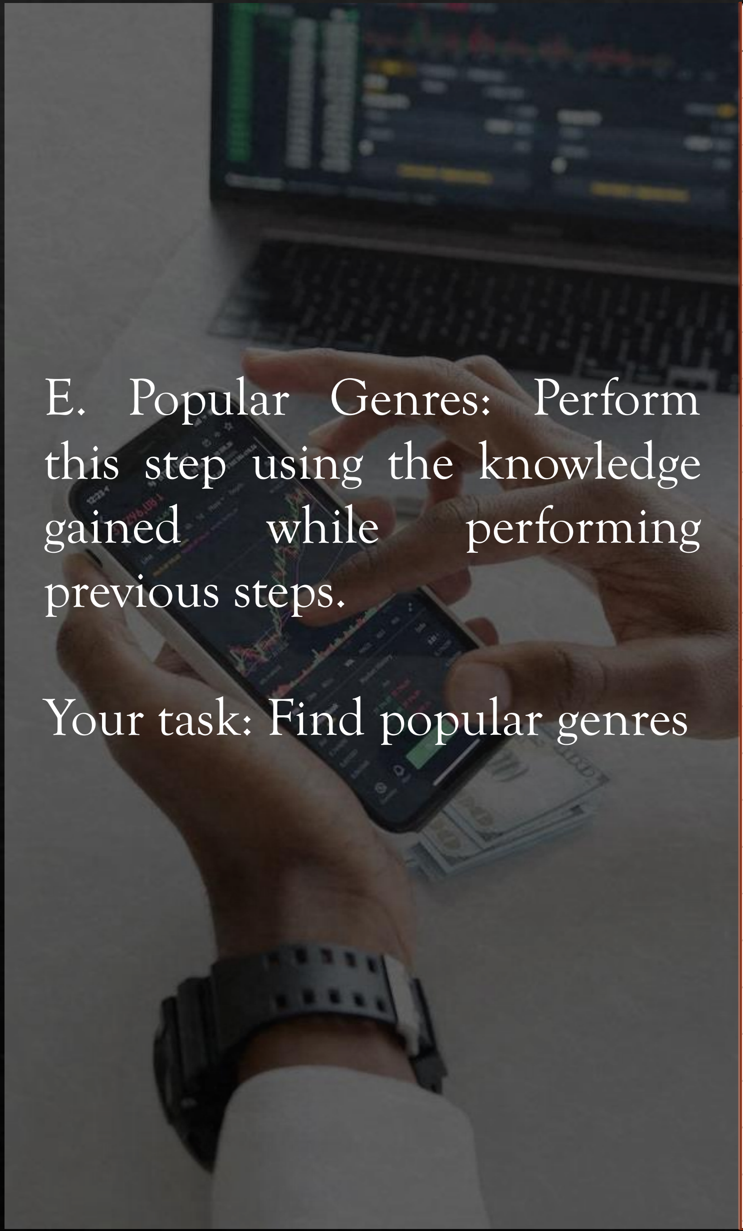
Rows: director_name

Values: Average of imdb_score



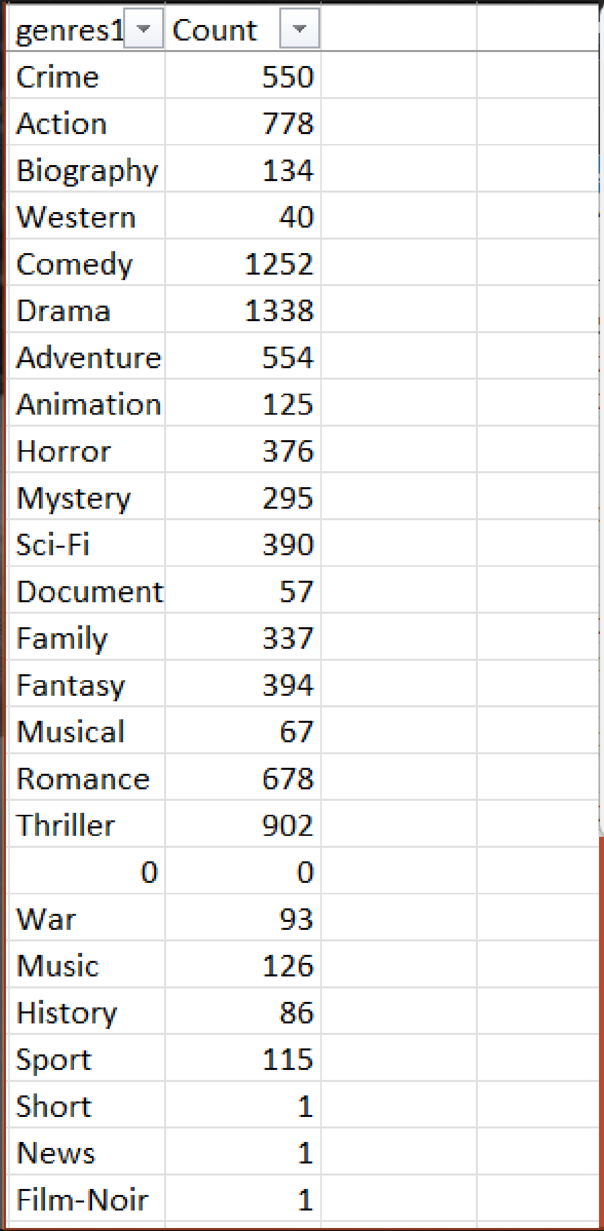
- This could be done using Pivot Table where director_name is selected as rows and Average of imdb_score is selected as Values. Than Top 10 is selected from filter.

Result

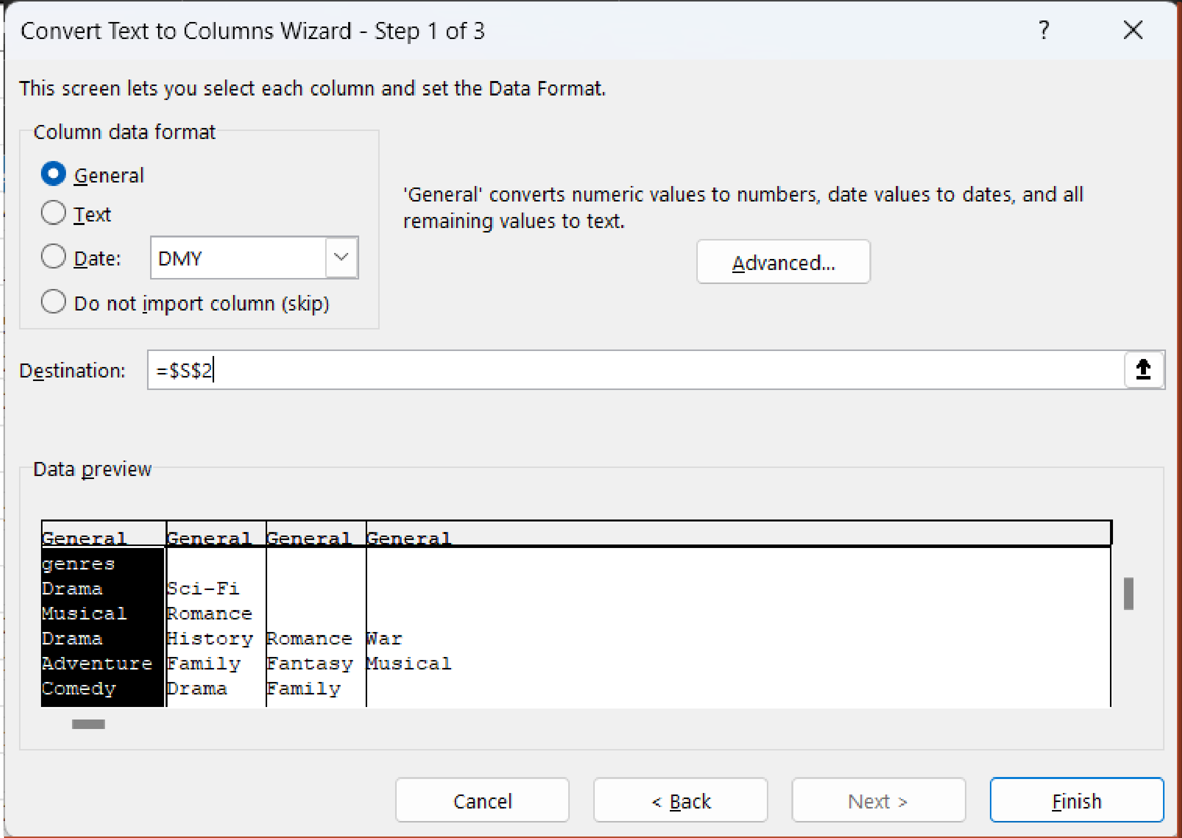


E. Popular Genres: Perform this step using the knowledge gained while performing previous steps.

Your task: Find popular genres



genres1	Count
Crime	550
Action	778
Biography	134
Western	40
Comedy	1252
Drama	1338
Adventure	554
Animation	125
Horror	376
Mystery	295
Sci-Fi	390
Document	57
Family	337
Fantasy	394
Musical	67
Romance	678
Thriller	902
0	0
War	93
Music	126
History	86
Sport	115
Short	1
News	1
Film-Noir	1



Convert Text to Columns Wizard - Step 1 of 3

This screen lets you select each column and set the Data Format.

Column data format

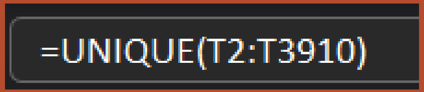
● General
○ Text
○ Date: DMY
○ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Destination: =S2

Data preview

```
General    General  General  General
genres
Drama      Sci-Fi
Musical    Romance
Drama      History  Romance  War
Adventure  Family   Fantasy  Musical
Comedy     Drama    Family
```



=UNIQUE(T2:T3910)

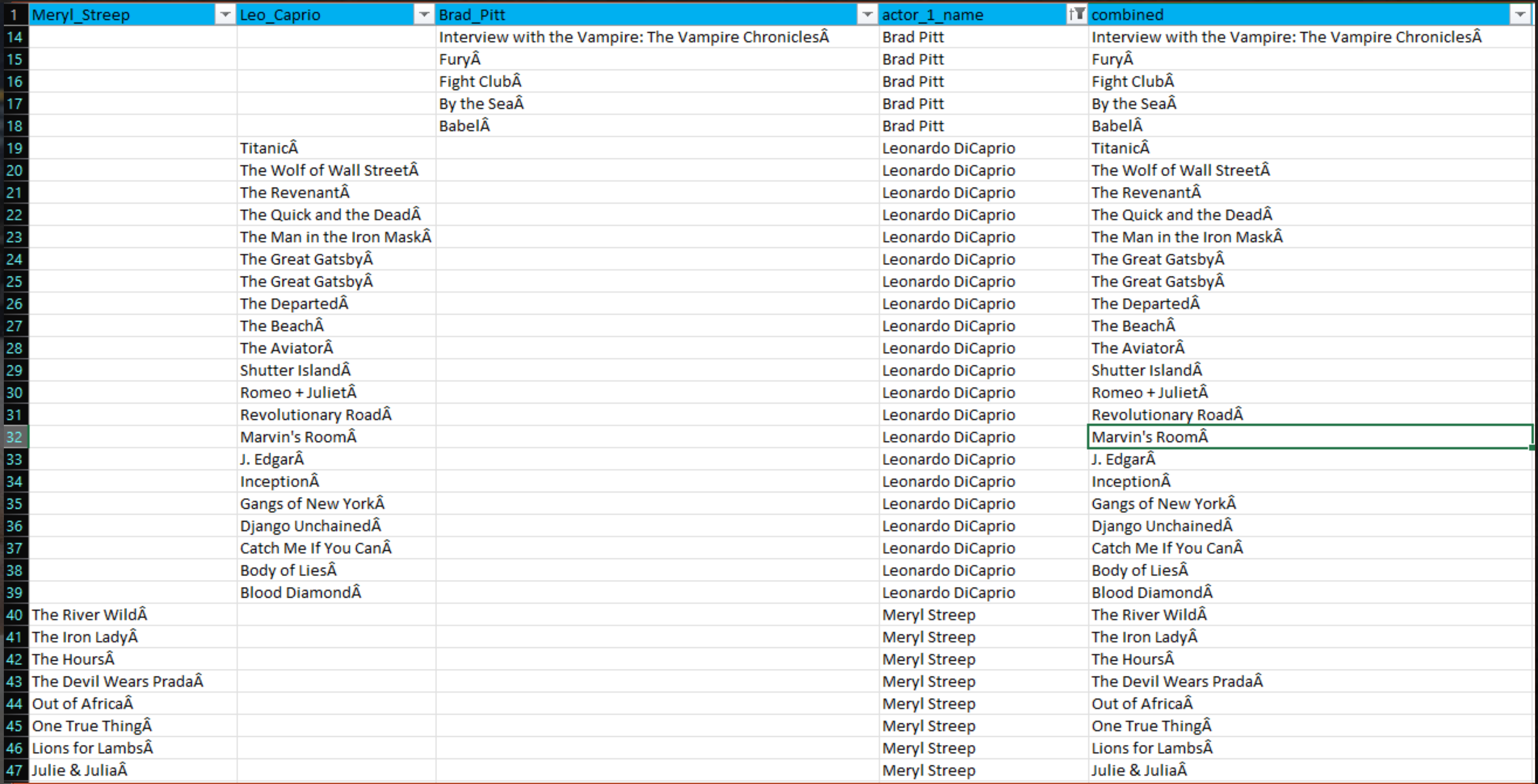
So the most popular genre is Drama which is used 1338 times in genre.

Result

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.



1	Meryl_Streep	Leo_Caprio	Brad_Pitt	actor_1_name	combined
14			Interview with the Vampire: The Vampire ChroniclesÂ	Brad Pitt	Interview with the Vampire: The Vampire ChroniclesÂ
15			FuryÂ	Brad Pitt	FuryÂ
16			Fight ClubÂ	Brad Pitt	Fight ClubÂ
17			By the SeaÂ	Brad Pitt	By the SeaÂ
18			BabelÂ	Brad Pitt	BabelÂ
19		TitanicÂ		Leonardo DiCaprio	TitanicÂ
20		The Wolf of Wall StreetÂ		Leonardo DiCaprio	The Wolf of Wall StreetÂ
21		The RevenantÂ		Leonardo DiCaprio	The RevenantÂ
22		The Quick and the DeadÂ		Leonardo DiCaprio	The Quick and the DeadÂ
23		The Man in the Iron MaskÂ		Leonardo DiCaprio	The Man in the Iron MaskÂ
24		The Great GatsbyÂ		Leonardo DiCaprio	The Great GatsbyÂ
25		The Great GatsbyÂ		Leonardo DiCaprio	The Great GatsbyÂ
26		The DepartedÂ		Leonardo DiCaprio	The DepartedÂ
27		The BeachÂ		Leonardo DiCaprio	The BeachÂ
28		The AviatorÂ		Leonardo DiCaprio	The AviatorÂ
29		Shutter IslandÂ		Leonardo DiCaprio	Shutter IslandÂ
30		Romeo + JulietÂ		Leonardo DiCaprio	Romeo + JulietÂ
31		Revolutionary RoadÂ		Leonardo DiCaprio	Revolutionary RoadÂ
32		Marvin's RoomÂ		Leonardo DiCaprio	Marvin's RoomÂ
33		J. EdgarÂ		Leonardo DiCaprio	J. EdgarÂ
34		InceptionÂ		Leonardo DiCaprio	InceptionÂ
35		Gangs of New YorkÂ		Leonardo DiCaprio	Gangs of New YorkÂ
36		Django UnchainedÂ		Leonardo DiCaprio	Django UnchainedÂ
37		Catch Me If You CanÂ		Leonardo DiCaprio	Catch Me If You CanÂ
38		Body of LiesÂ		Leonardo DiCaprio	Body of LiesÂ
39		Blood DiamondÂ		Leonardo DiCaprio	Blood DiamondÂ
40	The River WildÂ			Meryl Streep	The River WildÂ
41	The Iron LadyÂ			Meryl Streep	The Iron LadyÂ
42	The HoursÂ			Meryl Streep	The HoursÂ
43	The Devil Wears PradaÂ			Meryl Streep	The Devil Wears PradaÂ
44	Out of AfricaÂ			Meryl Streep	Out of AfricaÂ
45	One True ThingÂ			Meryl Streep	One True ThingÂ
46	Lions for LambsÂ			Meryl Streep	Lions for LambsÂ
47	Julie & JuliaÂ			Meryl Streep	Julie & JuliaÂ

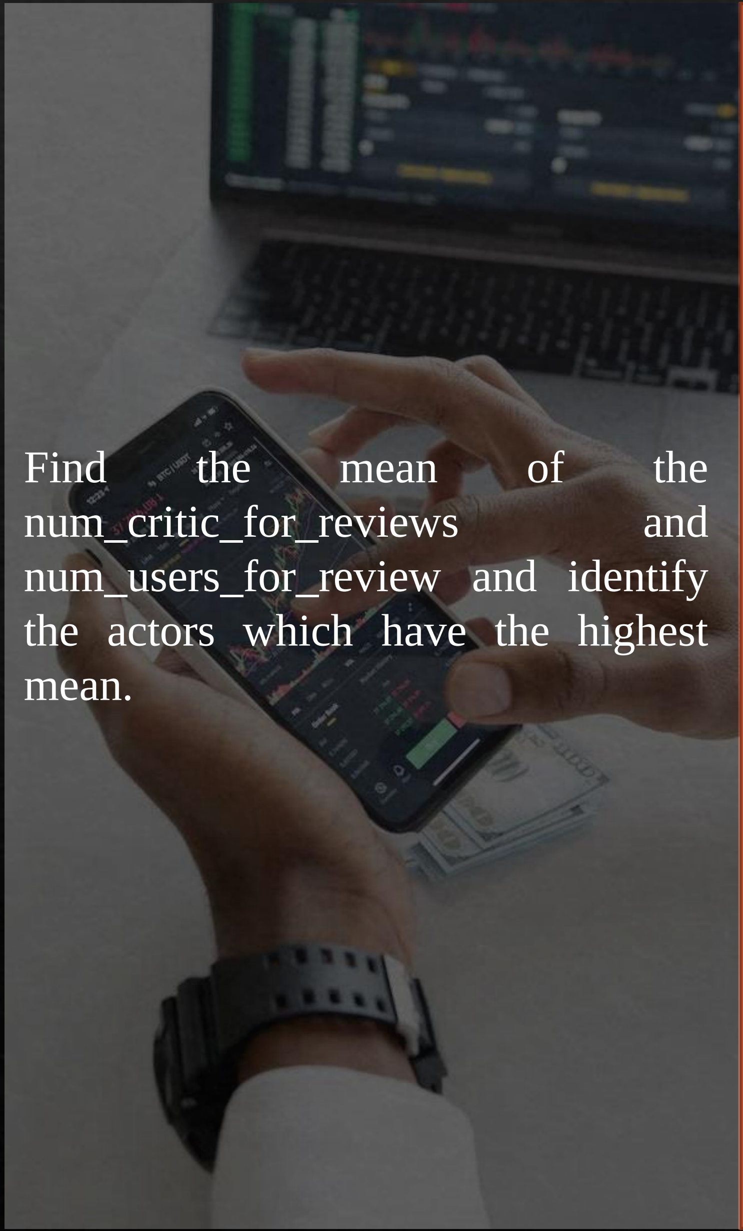
```
=IF(F14="Meryl Streep",G14,"")
```

```
=IF(F15="Leonardo DiCaprio",G15,"")
```

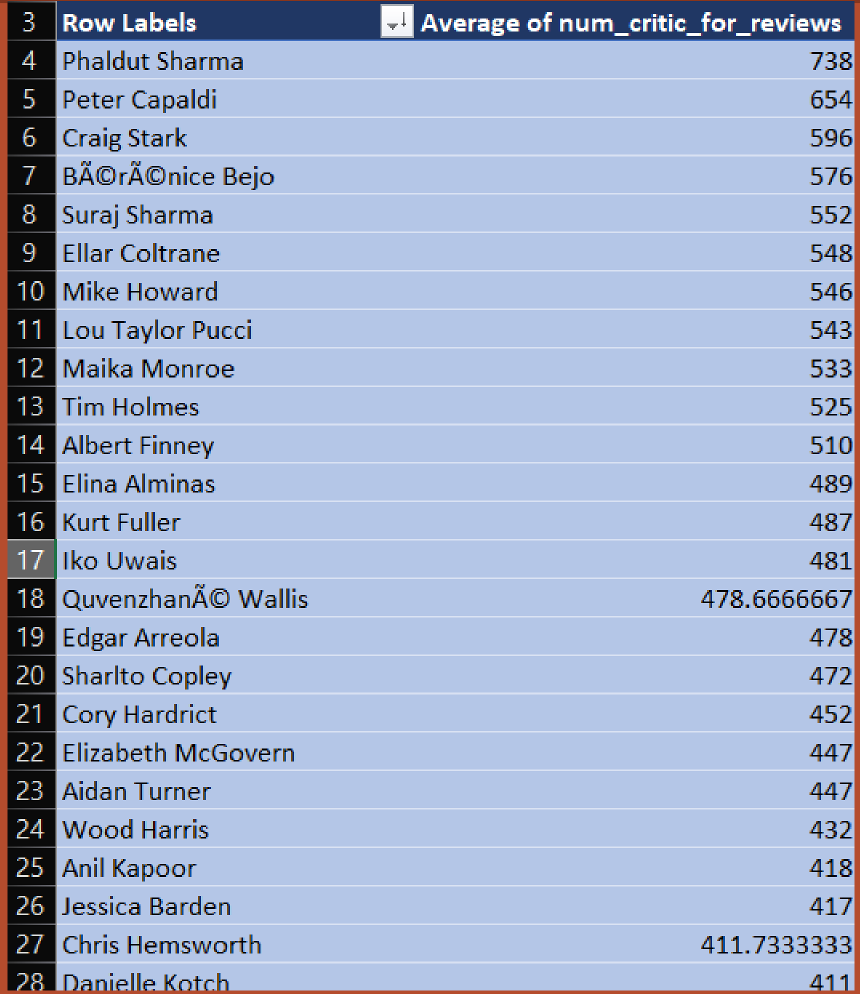
```
=IF(F14="Brad Pitt",G14,"")
```

```
=IF(OR(F14="Meryl Streep", F14="Leonardo DiCaprio", F14= "Brad Pitt"),G14,"")
```

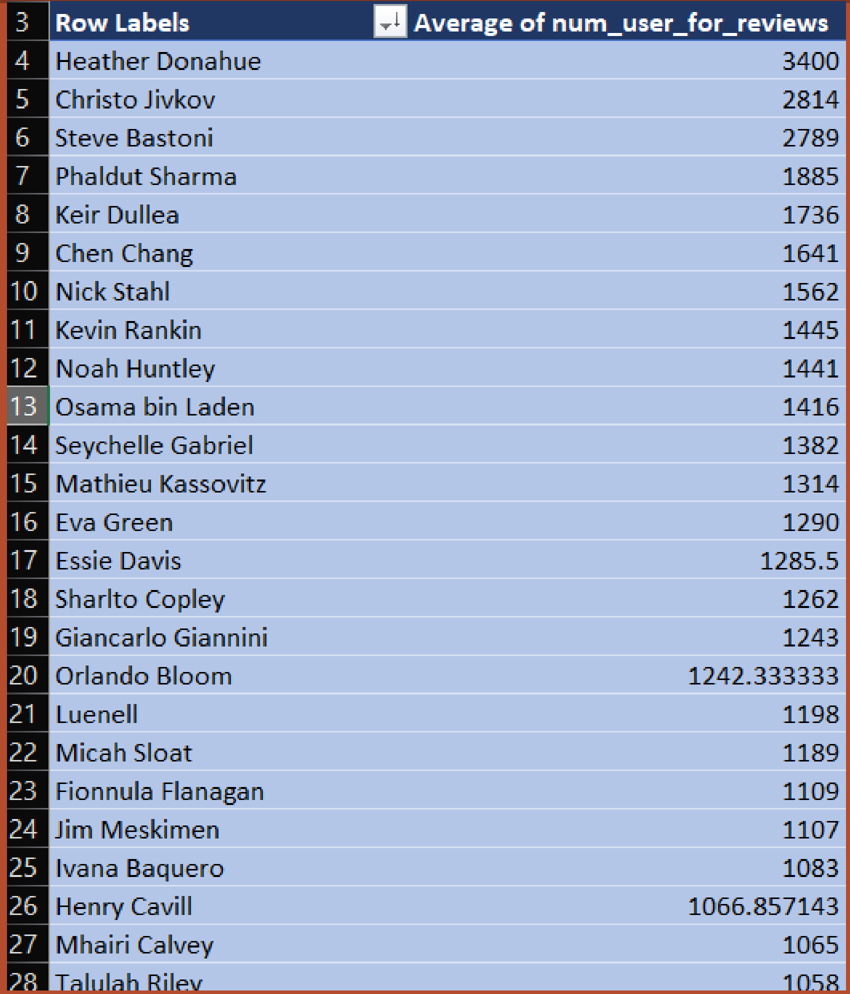
Result



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



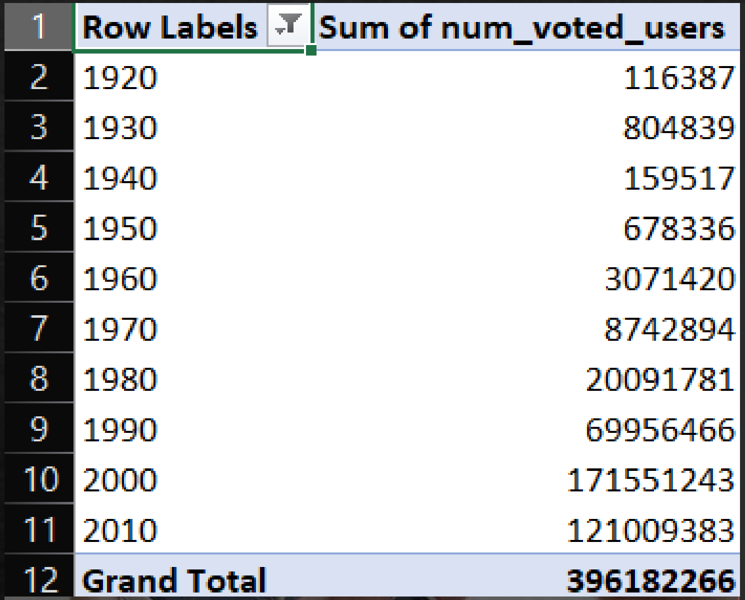
3	Row Labels	Average of num_critic_for_reviews
4	Phaldut Sharma	738
5	Peter Capaldi	654
6	Craig Stark	596
7	BÃ©rÃ©nice Bejo	576
8	Suraj Sharma	552
9	Ellar Coltrane	548
10	Mike Howard	546
11	Lou Taylor Pucci	543
12	Maika Monroe	533
13	Tim Holmes	525
14	Albert Finney	510
15	Elina Alminas	489
16	Kurt Fuller	487
17	Iko Uwais	481
18	QuvenzhanÃ© Wallis	478.6666667
19	Edgar Arreola	478
20	Sharlto Copley	472
21	Cory Hardrict	452
22	Elizabeth McGovern	447
23	Aidan Turner	447
24	Wood Harris	432
25	Anil Kapoor	418
26	Jessica Barden	417
27	Chris Hemsworth	411.7333333
28	Danielle Kotch	411



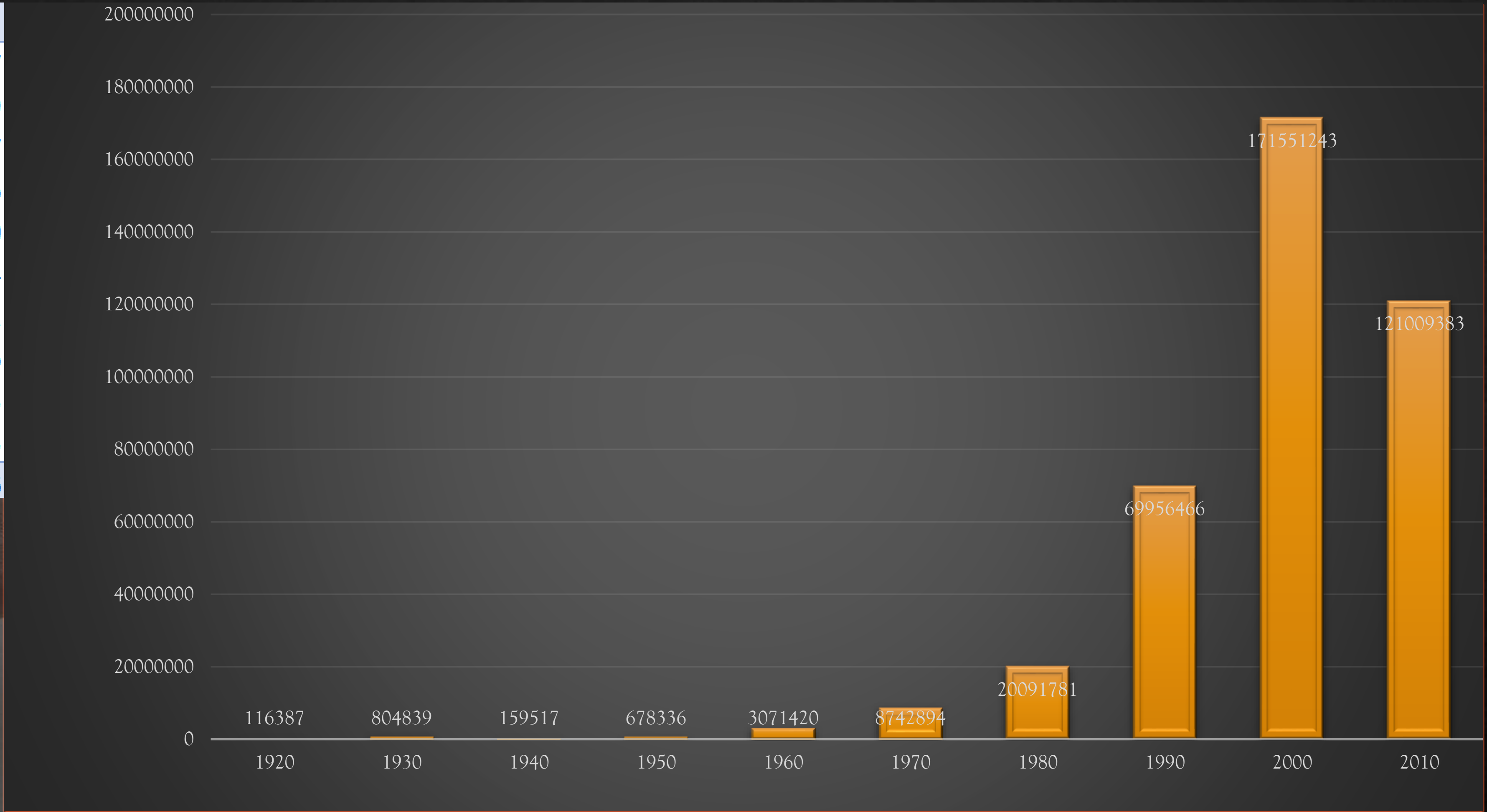
3	Row Labels	Average of num_user_for_reviews
4	Heather Donahue	3400
5	Christo Jivkov	2814
6	Steve Bastoni	2789
7	Phaldut Sharma	1885
8	Keir Dullea	1736
9	Chen Chang	1641
10	Nick Stahl	1562
11	Kevin Rankin	1445
12	Noah Huntley	1441
13	Osama bin Laden	1416
14	Seychelle Gabriel	1382
15	Mathieu Kassovitz	1314
16	Eva Green	1290
17	Essie Davis	1285.5
18	Sharlto Copley	1262
19	Giancarlo Giannini	1243
20	Orlando Bloom	1242.333333
21	Luenell	1198
22	Micah Sloat	1189
23	Fionnula Flanagan	1109
24	Jim Meskimen	1107
25	Ivana Baquero	1083
26	Henry Cavill	1066.857143
27	Mhairi Calvey	1065
28	Talulah Riley	1058

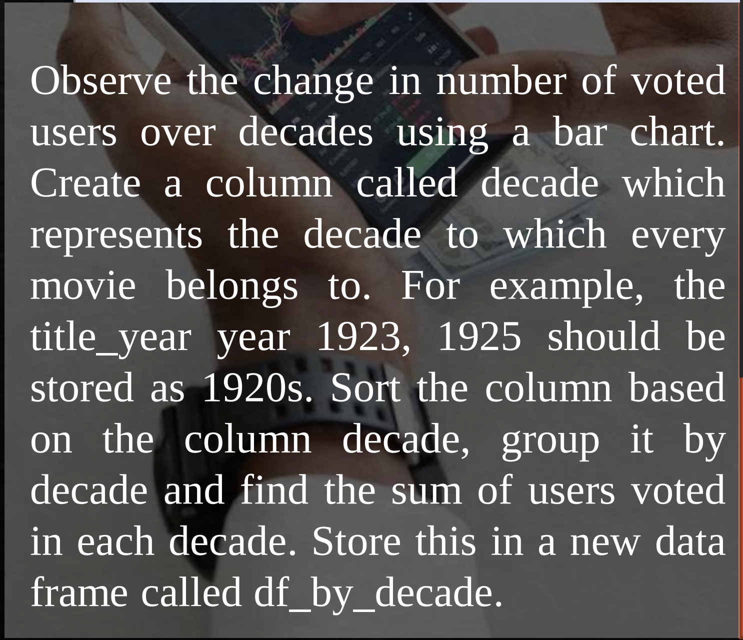
Both the means are calculated using pivot table:
Actor who has the highest mean of num_critic_for_reviews is “Phaldut Sharma”.
Actor who has the highest mean of num_user_for_reviews is “Heather Donahue”.

Result

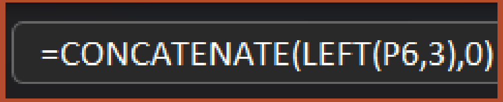


Row Labels	Sum of num_voted_users
1920	116387
1930	804839
1940	159517
1950	678336
1960	3071420
1970	8742894
1980	20091781
1990	69956466
2000	171551243
2010	121009383
Grand Total	**396182266**





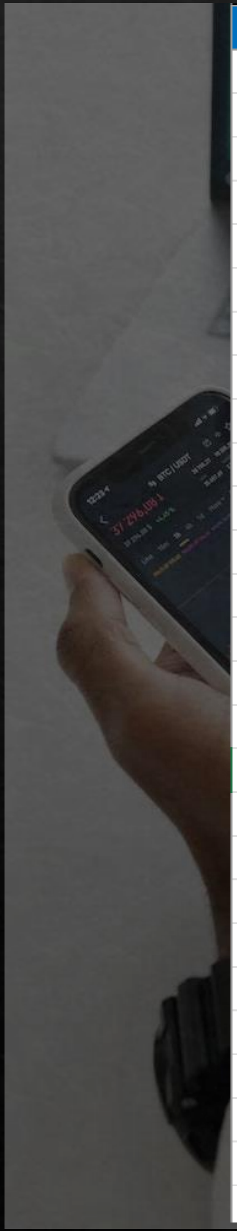
Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade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

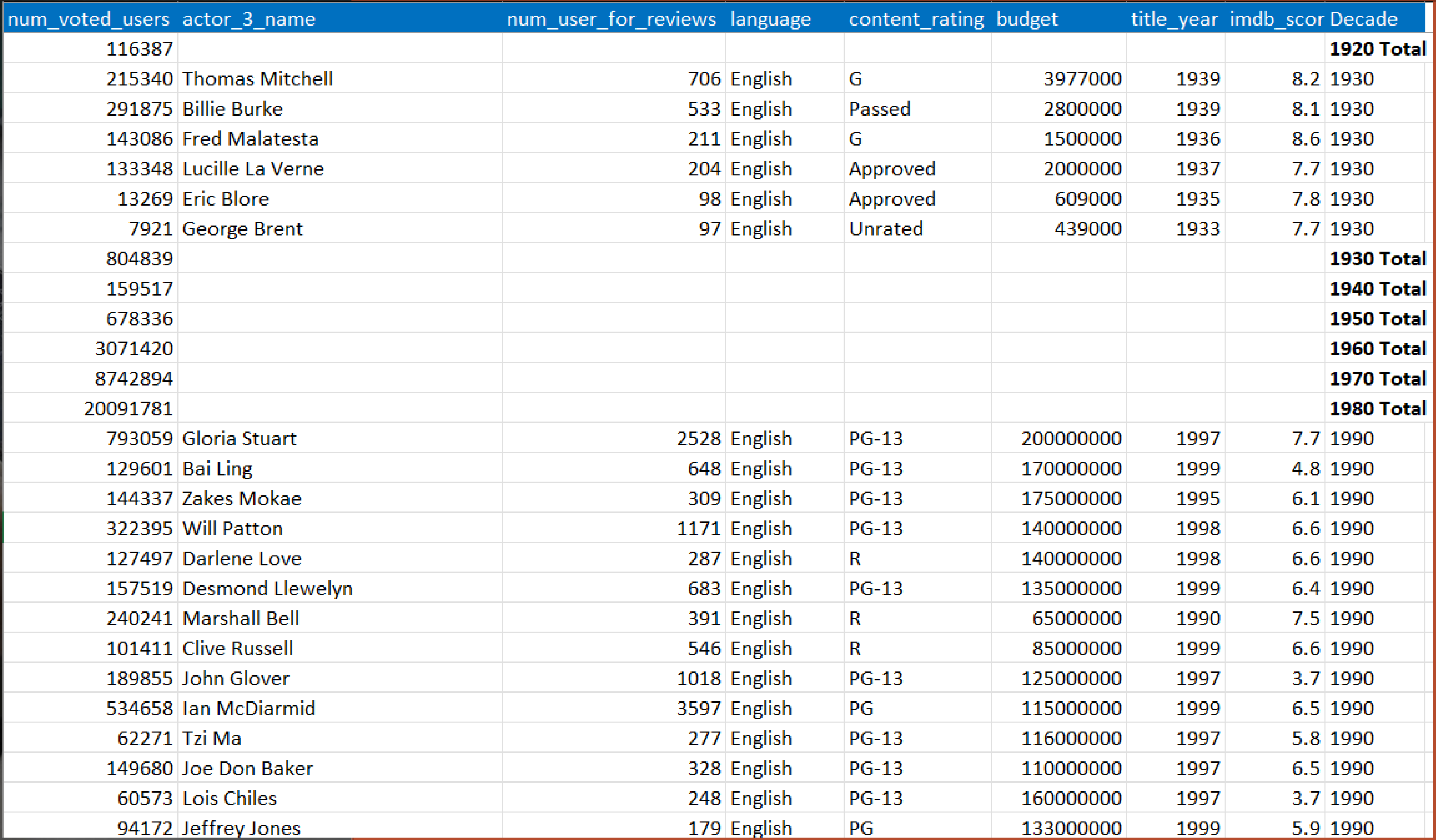


```
=CONCATENATE(LEFT(P6,3),0)
```

- Adjacent formula is used to calculate the decade.

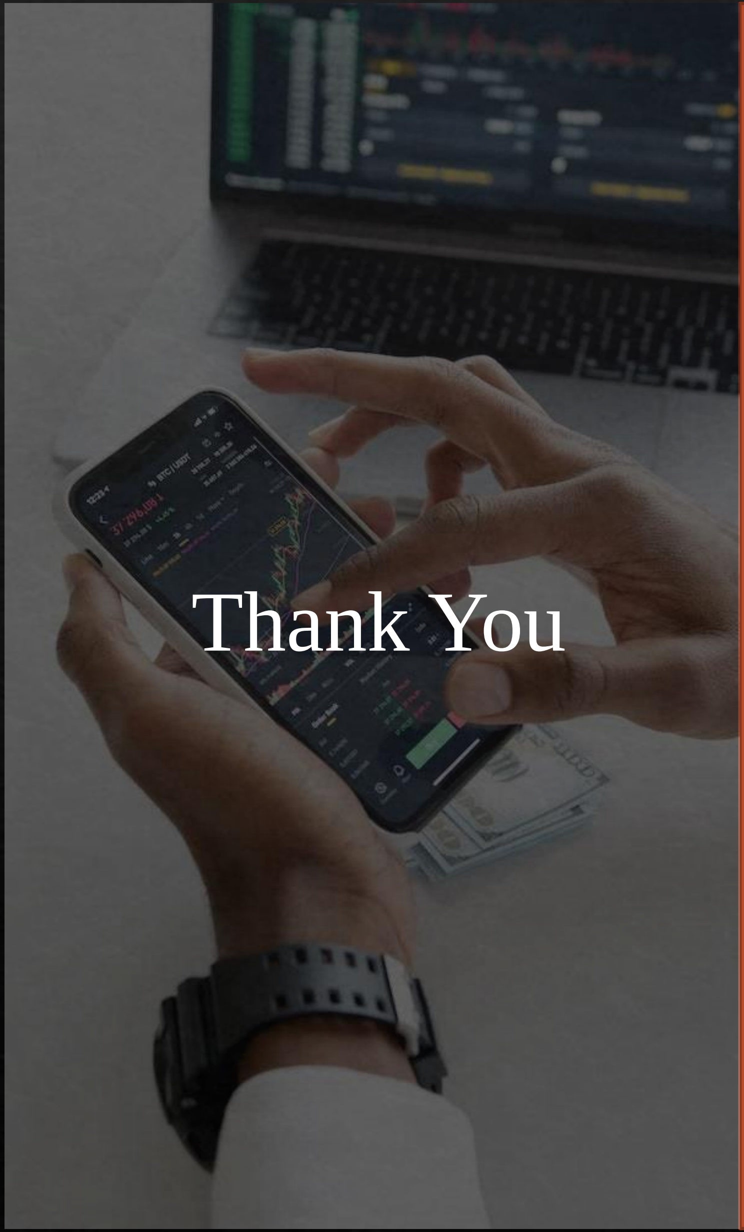
Result





num_voted_users	actor_3_name	num_user_for_reviews	language	content_rating	budget	title_year	imdb_scor	Decade
116387								**1920 Total**
215340	Thomas Mitchell	706	English	G	3977000	1939	8.2	1930
291875	Billie Burke	533	English	Passed	2800000	1939	8.1	1930
143086	Fred Malatesta	211	English	G	1500000	1936	8.6	1930
133348	Lucille La Verne	204	English	Approved	2000000	1937	7.7	1930
13269	Eric Blore	98	English	Approved	609000	1935	7.8	1930
7921	George Brent	97	English	Unrated	439000	1933	7.7	1930
804839								**1930 Total**
159517								**1940 Total**
678336								**1950 Total**
3071420								**1960 Total**
8742894								**1970 Total**
20091781								**1980 Total**
793059	Gloria Stuart	2528	English	PG-13	200000000	1997	7.7	1990
129601	Bai Ling	648	English	PG-13	170000000	1999	4.8	1990
144337	Zakes Mokae	309	English	PG-13	175000000	1995	6.1	1990
322395	Will Patton	1171	English	PG-13	140000000	1998	6.6	1990
127497	Darlene Love	287	English	R	140000000	1998	6.6	1990
157519	Desmond Llewelyn	683	English	PG-13	135000000	1999	6.4	1990
240241	Marshall Bell	391	English	R	65000000	1990	7.5	1990
101411	Clive Russell	546	English	R	85000000	1999	6.6	1990
189855	John Glover	1018	English	PG-13	125000000	1997	3.7	1990
534658	Ian McDiarmid	3597	English	PG	115000000	1999	6.5	1990
62271	Tzi Ma	277	English	PG-13	116000000	1997	5.8	1990
149680	Joe Don Baker	328	English	PG-13	110000000	1997	6.5	1990
60573	Lois Chiles	248	English	PG-13	160000000	1997	3.7	1990
94172	Jeffrey Jones	179	English	PG	133000000	1999	5.9	1990

Result



Thank You

Connect me on:
rajatpawan@gmail.com