Shakti prasad Nayak

Data Analytics Project
Bhubaneswar, Odisha
Dataset URL: CLICK HERE

IMDb Movie Analysis

April 13, 2024

Project Description

The project "IMDb movies Analysis" is a data analysis project that aims to explore the trends and patterns of movies ratings, genres, budgets, revenues, and popularity on the IMDb website. The project uses a dataset which contains various columns such as title, year, director, cast, rating, votes, genre, runtime, budget, revenue, and popularity. The project will apply descriptive statistics, data visualization techniques to answer some interesting questions.

Approach

We are going to use various data analysis and visualization techniques to explore and answer questions about a dataset of IMDb movies. The dataset contains information such as title, year, genre, rating, votes, revenue, and runtime for over 1000 movies.

Some of the questions that the project aims to answer are:

- What are the most popular genres?

- What are the top 20 highest profit movies?

- Which directors and actors have the most movies and the highest ratings?

- Which Movies have the highest number of votes?

- Which decade has the highest number of votes?

To answer these questions, the project will use Excel functions and formulas to manipulate and analyze the data, such as sorting, filtering, conditional formatting, pivot tables and charts. The project will also use Excel features such as slicers, timelines, and data validation to create interactive dashboards and reports that can present the findings in a clear and engaging way. The project will follow the data analysis process of defining the problem, collecting and cleaning the data, exploring and analyzing the data, and communicating the results.

Tech-Stack Used

Microsoft Excel is a powerful spreadsheet application that can help you store, manage, and analyze data. Here are a few advantages of Microsoft Excel:

- You can perform calculations using formulas and functions that can handle complex mathematical operations and return accurate results.

- You can use data analysis tools such as pivot tables, charts, and filters to summarize, visualize, and explore your data in different ways.

- You can print reports easily by adjusting the page layout, margins, headers, and footers to fit your needs.

- You can use free templates to create professional-looking documents such as invoices, budgets, calendars, and more.

- You can code to automate repetitive tasks using macros and VBA (Visual Basic for Applications), which can save you time and improve your efficiency.

- You can transform and clean data using features such as Power Query, Data Validation, and Text to Columns, which can help you prepare your data for analysis.

- You can store data with millions of rows and columns without compromising the performance of your workbook.

- You can work with Excel online or on a mobile app, which allows you to access and edit your files from anywhere and collaborate with others in real time.

Insights

1. **Top 20 Highest Profit Movies:** Avatar , Jurassic World , Titanic , Star Wars: Episode IV - A New Hope , E.T. the Extra-Terrestrial , The Avengers , The Lion King , Star Wars: Episode I - The Phantom Menace , The Dark Knight , The Hunger Games , Deadpool , The Hunger Games: Catching Fire , Jurassic Park , Despicable Me 2 , American Sniper , Finding Nemo , Shrek 2 , The Lord of the Rings: The Return of the King , Star Wars: Episode VI - Return of the Jedi , Forrest Gump .

2. **Top 20 Highest Number of Voted Users Movies:** The Shawshank Redemption , The Dark Knight , Inception , Fight Club , Pulp Fiction , Forrest Gump , The Lord of the Rings: The Fellowship of the Ring , The Matrix , The Lord of the Rings: The Return of the King , The Godfather , The Dark Knight Rises , The Lord of the Rings: The Two Towers , Se7en , The Avengers , Gladiator , Batman Begins , Django Unchained , Interstellar , Star Wars: Episode IV - A New Hope , The Silence of the Lambs .

3. **Top 20 Best Directors:** Steven Spielberg, Woody Allen, Clint Eastwood, Martin Scorsese, Ridley Scott, Spike Lee, Steven Soderbergh, Tim Burton, Robert Zemeckis, Ron Howard, Oliver Stone, Renny Harlin, Barry Levinson, Tony Scott, Richard Linklater, Michael Bay, David Fincher, Rob Reiner, Joel Schumacher, Robert Rodriguez

4. **Top 5 Genres:**Comedy, Action, Drama, Adventure, Crime

5. **Highest numbers of users voted in which decade?** Highest number of users voted in decade 2000s.

Results

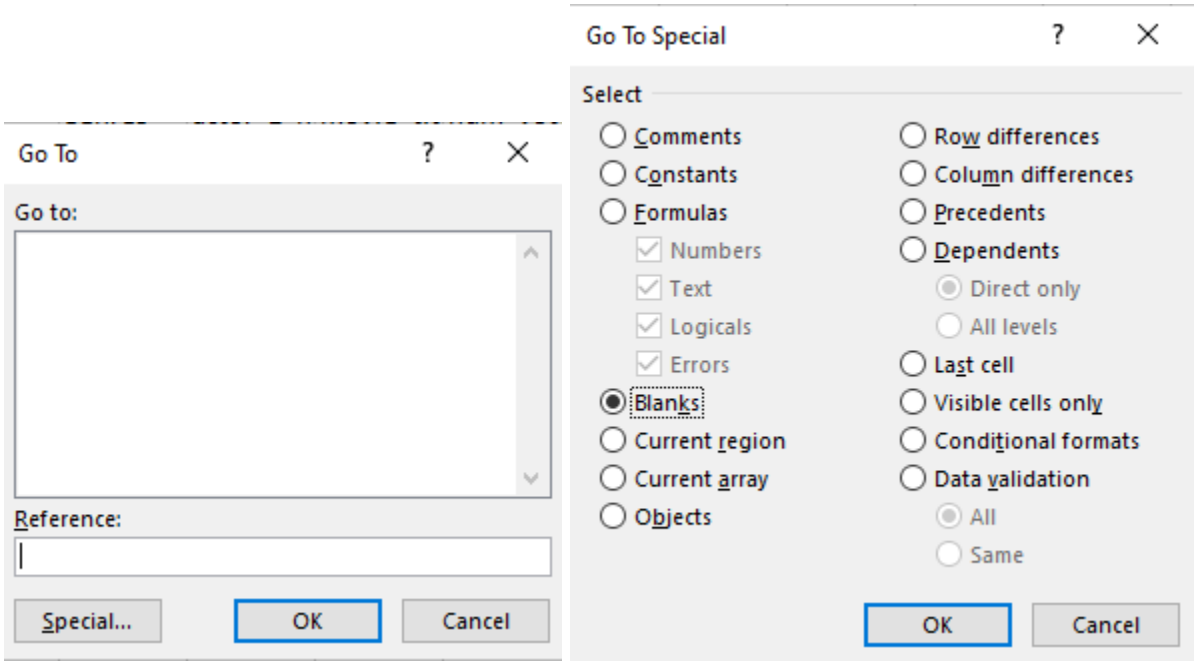
I have learnt EXCEL Formulas and Functions, pivot table, Data Cleaning, Sorting, Filtering and Visualization. This has helped me analyze Dataset using MS Excel.

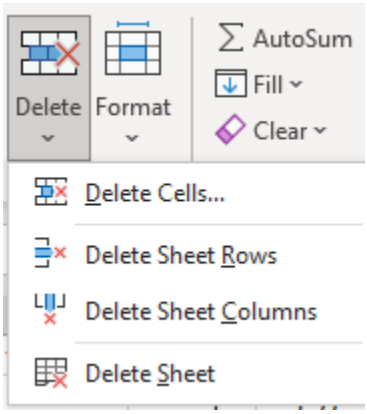
REPORT

Cleaning the data

Deleting unnecessary columns: We need to delete the following column from the dataset color, duration, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, movie_imdb_link, country, content_rating, actor_2_facebook_likes, aspect_ratio and movie_facebook_likes.

Deleting rows with blank cells: For this we need to select all the cells of our table by pressing Ctrl+A then press Ctrl+G. Our go to dialog box will appear. Then click Special. By selecting blanks and pressing the OK button will select all the blank cells. Then press the Delete button on the home tab and click on Delete Sheet Rows**.** This will delete all the rows with blank cells.



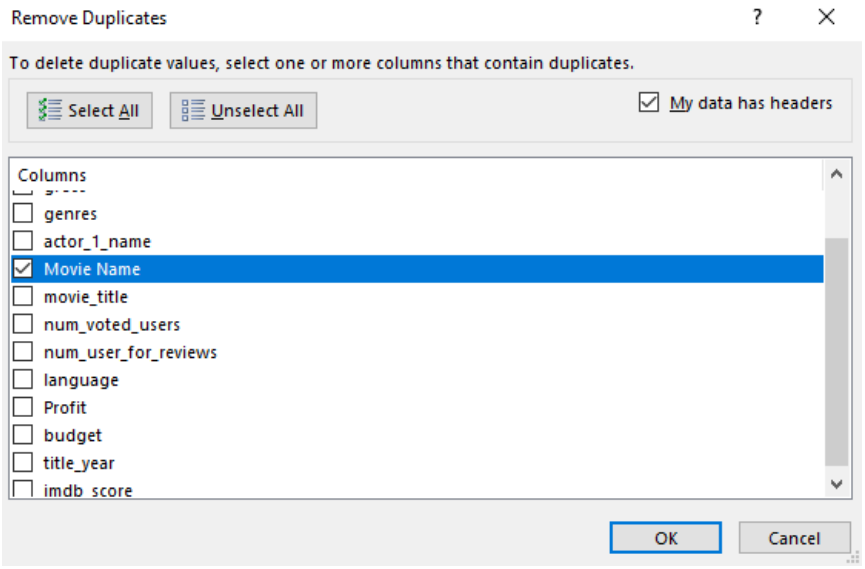


Removing the Â character: For this we need to create a new column Movie Name. Each cell of this column should have values of movie_title column with cleaned **Â** character. For this we need to execute the following formula in our excel sheet-

=SUBSTITUTE(G2,"Â","")

Movie Name	movie_title
Avatar	AvatarÂ
Pirates of the Caribbean: At World's End	Pirates of the Caribbean: At World's End
Spectre	SpectreÂ
The Dark Knight Rises	The Dark Knight RisesÂ
John Carter	John CarterÂ
Spider-Man 3	Spider-Man 3Â
Tangled	TangledÂ
Avengers: Age of Ultron	Avengers: Age of UltronÂ
Harry Potter and the Half-Blood Prince	Harry Potter and the Half-Blood PrinceÂ
Batman v Superman: Dawn of Justice	Batman v Superman: Dawn of JusticeÂ
Superman Returns	Superman ReturnsÂ
Quantum of Solace	Quantum of SolaceÂ

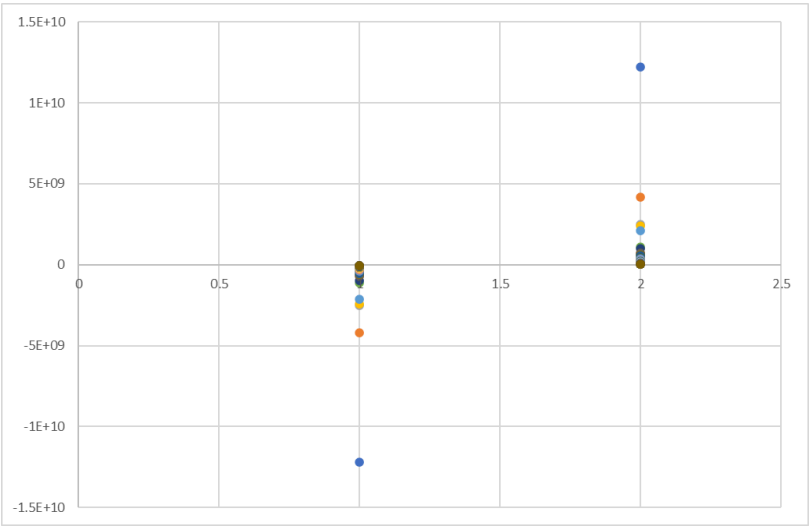
Removing duplicate rows: For this we need to go to the Data tab and click on Remove Duplicates. Now, we check only the Movie Name column to identify rows with the same Movie Name and then click OK. This will remove all the rows with duplicate values.



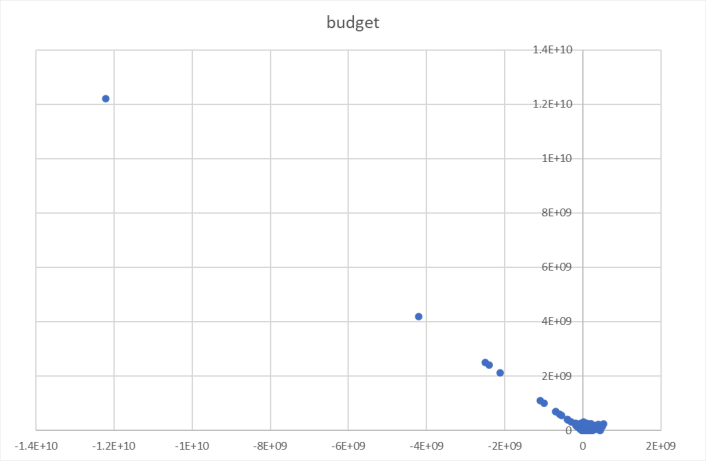
Movies with highest profit

For this, we have created a new column Profit where each cell contains the difference between gross and budget. Now we apply filter on the Profit column and sort using Largest to Smallest. We plot a graph Profit at Y-axis and Budget at X-axis for the first 255 rows (Because Excel allows only 255 data) to find outliers present in our dataset. We can also find the outliers by sorting the Profit column Smallest to Largest or Budget column Largest to smallest.

Profit	budget
-1.2E+10	1E+10
-4.2E+09	4E+09
-2.5E+09	3E+09
-2.4E+09	2E+09
-2.1E+09	2E+09
-1.1E+09	1E+09
-9.9E+08	1E+09
-7E+08	7E+08
-7E+08	7E+08
-6E+08	6E+08
-5.5E+08	6E+08



Profit	budget
-1.2E+10	1E+10
-4.2E+09	4E+09
-2.5E+09	3E+09
-2.4E+09	2E+09
-2.1E+09	2E+09
-1.1E+09	1E+09
-9.9E+08	1E+09



Now, we can plot a graph containing the top 20 movies with the highest profit.

Top 250

For this we have created two new columns IMDb Top 250 and Rank to store top 250 movies with highest IMDb ratings in a new Sheet. Now we use filter function to filter out our required 250 movies and sort it by IMDb score-

=FILTER(SORT(FILTER(F2:H3783,J2:J3783>25000),3,-1,FALSE),ROW(Q2:Q2543)<252)

The top 20 result is given below-

Rank	Top 250 Movies
1	The Shawshank Redemption
2	The Godfather
3	The Dark Knight
4	The Godfather: Part II
5	The Lord of the Rings: The Return of the King
6	Schindler's List
7	Pulp Fiction
8	The Good, the Bad and the Ugly
9	Inception
10	The Lord of the Rings: The Fellowship of the Ring
11	Fight Club
12	Forrest Gump
13	Star Wars: Episode V - The Empire Strikes Back

14	The Lord of the Rings: The Two Towers
15	The Matrix
16	Goodfellas
17	Star Wars: Episode IV - A New Hope
18	One Flew Over the Cuckoo's Nest
19	City of God
20	Seven Samurai

Now out of these 250 movies we need to figure out all non english movies-

First we sort our dataset by IMDb score, Largest to smallest. Then we apply following formula-

=FILTER(F2:F3783,(ROW(F2:F3783)<263)*(G2:G3783<>"English"))

Now we have Top foreign language films. Given below-

Rank	Top Foreign Language Movies
1	The Good, the Bad and the Ugly
2	City of God
3	Seven Samurai
4	Spirited Away
5	The Lives of Others
6	Children of Heaven
7	AmÃ©lie

8	Baahubali: The Beginning
9	Princess Mononoke
10	Das Boot
11	Oldboy
12	A Separation
13	Metropolis
14	Downfall
15	The Hunt
16	Howl's Moving Castle
17	Pan's Labyrinth
18	Incendies
19	The Secret in Their Eyes
20	The Sea Inside
21	Tae Guk Gi: The Brotherhood of War
22	Akira
23	Elite Squad
24	Amores Perros
25	The Celebration
26	My Name Is Khan
27	Persepolis
28	Central Station

29	Waltz with Bashir
30	A Fistful of Dollars
31	Hero
32	Crouching Tiger, Hidden Dragon
33	Letters from Iwo Jima
34	Amour
35	Veer-Zaara
36	The Chorus

Best Directors

To find the top 10 directors with highest value of mean IMDb scores we need to use pivot table and move the director field into rows and IMDb scores into values. We change the field value setting of IMDb scores to mean.

Now we copy the pivot table, paste it in a new sheet and apply filter on it. Now we sort table Largest to Smallest on the mean IMDb score. Here, the top 10 best directors from the table

top10director	Mean IMDb
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	8.43333333
Christopher Nolan	8.425
Asghar Farhadi	8.4
Marius A. Markevicius	8.4

Popular Genres

To find the count of all the genres from movies those have num_voted_users count greater than 25000 i.e popular movies, we need to apply following formula in a new sheet-

=TEXTSPLIT(IMDB_Movies!D2:D3828,,"|")

Then using pivot table we can count the number, each genre appears. Here is the result-

Row Labels	Count of Popular Genre
Comedy	1035
Action	940
Drama	687
Adventure	368
Crime	253
Biography	208
Horror	161
Animation	46
Documentary	44

Fantasy	36
Mystery	22
Sci-Fi	9
Family	5
Thriller	4
Western	4
Romance	3
Musical	2

Charts

The following function will extract movies name to their corresponding actors-

=FILTER('IMDB_Movies (1)'!F2:F3783,'IMDB_Movies (1)'!E2:E3783="Meryl Streep")

Movies with corresponding lead actors are given below

Meryl_Streep	Leo_Caprio	Brad_Pitt
A Prairie Home Companion	Blood Diamond	Babel
Hope Springs	Body of Lies	By the Sea
It's Complicated	Catch Me If You Can	Fight Club
Julie & Julia	Django Unchained	Fury
Lions for Lambs	Gangs of New York	Interview with the Vampire: The Vampire Chronicles
One True Thing	Inception	Killing Them Softly
Out of Africa	J. Edgar	Mr. & Mrs. Smith

The Devil Wears Prada	Marvin's Room	Ocean's Eleven
The Hours	Revolutionary Road	Ocean's Twelve
The Iron Lady	Romeo + Juliet	Seven Years in Tibet
The River Wild	Shutter Island	Sinbad: Legend of the Seven Seas
	The Aviator	Spy Game
	The Beach	The Assassination of Jesse James by the Coward Robert Ford
	The Departed	The Curious Case of Benjamin Button
	The Great Gatsby	The Tree of Life
	The Man in the Iron Mask	Troy
	The Quick and the Dead	True Romance
	The Revenant	
	The Wolf of Wall Street	
	Titanic	

Now, we created a new column Combined and stored all the movies belonging to these actors. Now, we create a new column Actor Name and store all the names of actors who belong to their corresponding movies in the Combined column. The below formula will extract the actor names from our above table-

```
=IFS(ISNUMBER(MATCH(F2,$A$1:$A$12,0)),"Meryl_Streep",ISNUMBER(MATCH(F2,$B$1:$B$21,0)),"Leo_Caprio",ISNUMBER(MATCH(F2,$C$1:$C$18,0)),"Brad_Pitt")
```

Here F2 represents the corresponding Movie Name from the Combined column.

The result is given below-

Combined	Actor Name
A Prairie Home Companion	Meryl_Streep
Hope Springs	Meryl_Streep
It's Complicated	Meryl_Streep
Julie & Julia	Meryl_Streep
Lions for Lambs	Meryl_Streep
One True Thing	Meryl_Streep
Out of Africa	Meryl_Streep
The Devil Wears Prada	Meryl_Streep
The Hours	Meryl_Streep
The Iron Lady	Meryl_Streep
The River Wild	Meryl_Streep
Blood Diamond	Leo_Caprio
Body of Lies	Leo_Caprio
Catch Me If You Can	Leo_Caprio
Django Unchained	Leo_Caprio
Gangs of New York	Leo_Caprio
Inception	Leo_Caprio
J. Edgar	Leo_Caprio
Marvin's Room	Leo_Caprio
Revolutionary Road	Leo_Caprio
Romeo + Juliet	Leo_Caprio
Shutter Island	Leo_Caprio
The Aviator	Leo_Caprio

The Beach	Leo_Caprio
The Departed	Leo_Caprio
The Great Gatsby	Leo_Caprio
The Man in the Iron Mask	Leo_Caprio
The Quick and the Dead	Leo_Caprio
The Revenant	Leo_Caprio
The Wolf of Wall Street	Leo_Caprio
Titanic	Leo_Caprio
Babel	Brad_Pitt
By the Sea	Brad_Pitt
Fight Club	Brad_Pitt
Fury	Brad_Pitt
Interview with the Vampire: The Vampire Chronicles	Brad_Pitt
Killing Them Softly	Brad_Pitt
Mr. & Mrs. Smith	Brad_Pitt
Ocean's Eleven	Brad_Pitt
Ocean's Twelve	Brad_Pitt
Seven Years in Tibet	Brad_Pitt
Sinbad: Legend of the Seven Seas	Brad_Pitt
Spy Game	Brad_Pitt
The Assassination of Jesse James by the Coward Robert Ford	Brad_Pitt
The Curious Case of Benjamin Button	Brad_Pitt
The Tree of Life	Brad_Pitt

Troy	Brad_Pitt
True Romance	Brad_Pitt

These are the top 20 actors with highest mean of num_critic_for_reviews

actor_1_name	Average of num_critic_for_reviews
Albert Finney	750
Phaldut Sharma	738
Peter Capaldi	654
Craig Stark	596
BÃ©rÃ©nice 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
QuvenzhanÃ© Wallis	478.6666667
Edgar Arreola	478
Sharlto Copley	472

Cory Hardrict	452
Matt Frewer	451

These are the top 20 actors with highest mean of num_users _for_reviews

actor_1_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

To find the sum of the number of users voted over decades we created a column Decade and another column df_by_decade to store the number of users voted over decades.

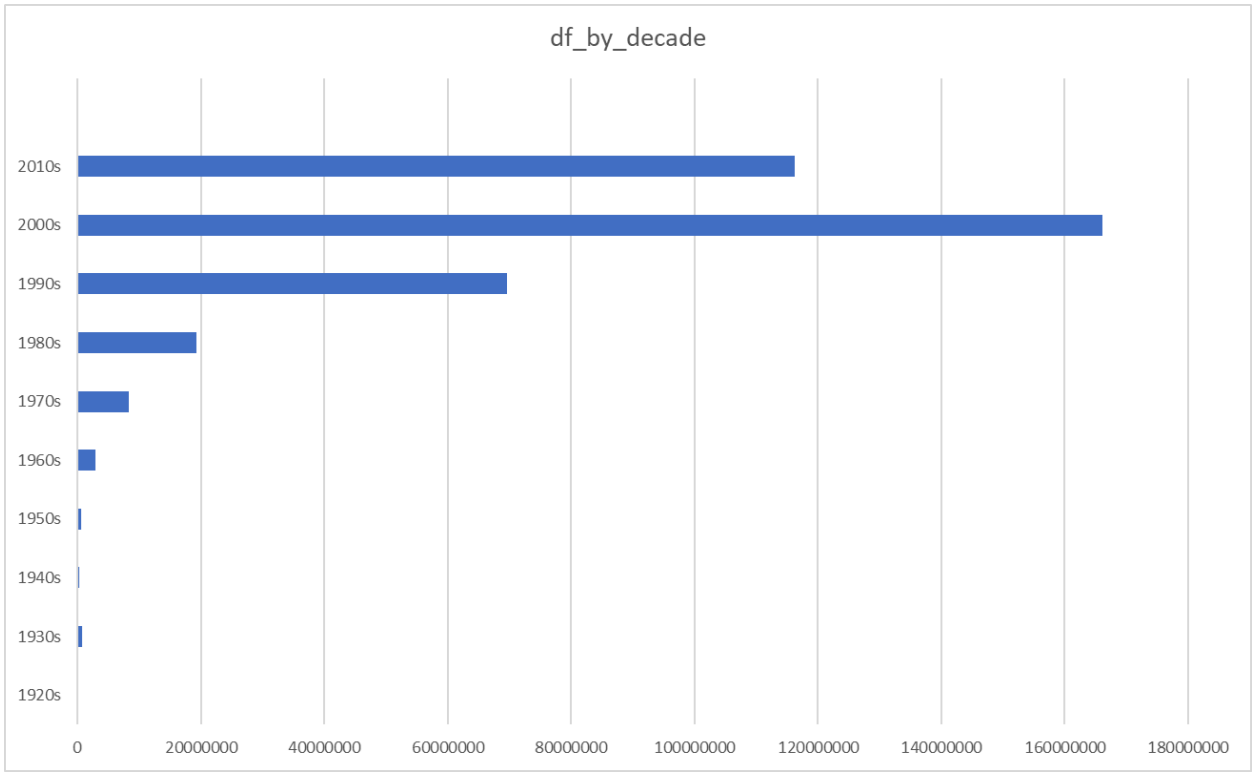
We used following formula to find the values of df_by_decade in reference to Decade column-

=SUMIFS(IMDB_Movies!$H$2:$H$3828,IMDB_Movies!$M$2:$M$3828,(">="&LEFT(A2,4)),IMDB_Movies!$M$2:$M$3828,("<"&LEFT(A2,4)+10))

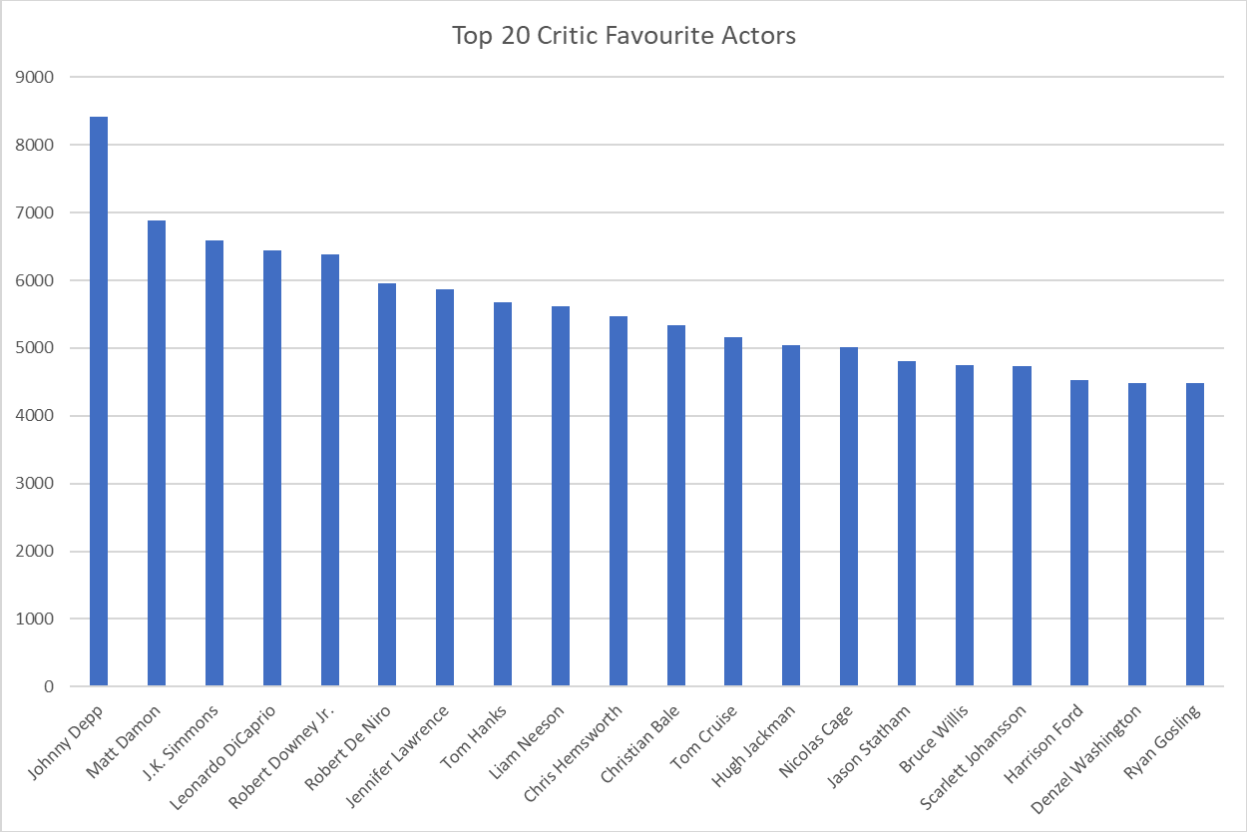
Here is the result-

Decade	df_by_decade
1920s	116387
1930s	804839
1940s	230838
1950s	678336
1960s	2983442
1970s	8318152
1980s	19344369
1990s	69635863
2000s	166058580
2010s	116259722

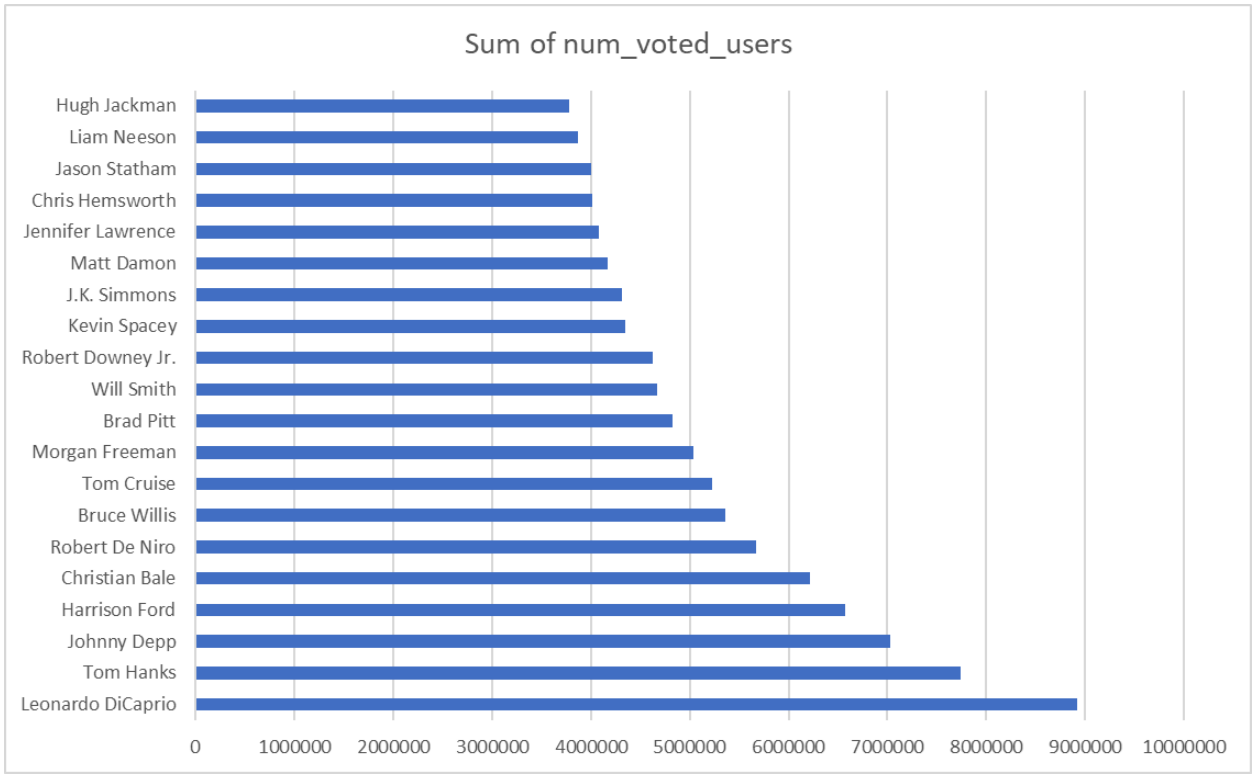
The bar graph plotted from above data-



The critic's favorite actors are those who have the highest number of sum of num_critic_for_reviews value. Here is the graph of top 20 critic favorite actors with sum of num_critic_for_reviews.



The audience favorite actors are those who have the highest sum of num_voted_users. Here is the graph of audience favorite actors with number of voted users-


Sum of num_voted_users
Hugh Jackman
Liam Neeson
Jason Statham
Chris Hemsworth
Jennifer Lawrence
Matt Damon
J.K. Simmons
Kevin Spacey
Robert Downey Jr.
Will Smith
Brad Pitt
Morgan Freeman
Tom Cruise
Bruce Willis
Robert De Niro
Christian Bale
Harrison Ford
Johnny Depp
Tom Hanks
Leonardo DiCaprio
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
9000000
10000000

END