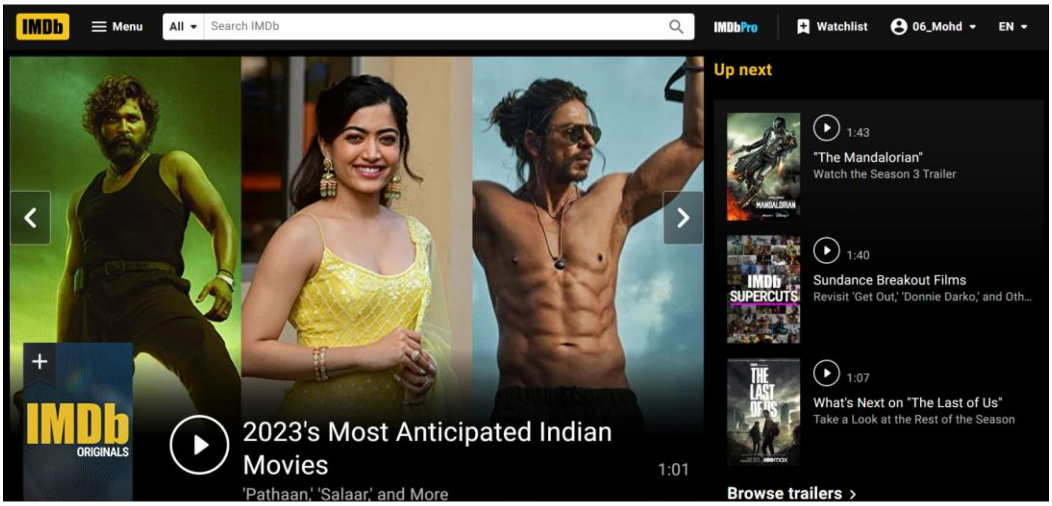
# IMDB Movie Analysis

**Project Description:** In this project, I have performed analysis on the IMDb movies dataset and displayed meaningful information from the dataset through MS Excel. The **Internet Movie Database (IMDb)** is a website that serves as an online database of world cinema containing a large number of public data on films such as the title of the film, the year of release of the film, the genre of the film, the audience, budget, revenue, the rating of critics, the duration of the film, the summary of the film, actors, directors and much more.


IMDb
Menu
All
Search IMDb
IMDbPro
Watchlist
06_Mohd
EN
Up next
1:43
"The Mandalorian"
Watch the Season 3 Trailer
1:40
Sundance Breakout Films
Revisit 'Get Out,' 'Donnie Darko,' and Oth...
1:07
What's Next on "The Last of Us"
Take a Look at the Rest of the Season
IMDb ORIGINALS
2023's Most Anticipated Indian Movies
'Pathaan,' 'Salaar,' and More
1:01
Browse trailers ›

Initially, after studying the columns of the dataset I got to know in the IMDb dataset there are some non-useful columns that are not useful in the analysis work and some rows have null values therefore our analysis work started with cleaning the data.

We are required to provide a detailed report for the below data record mentioning the answers to the questions that follow:

A. **Cleaning the data:** Clean the data

B. **Movies with the highest profit:** Find the movies with the highest profit.

C. **Top 250:** Find IMDB Top 250

D. **Best Directors:** Find the best directors

E. **Popular Genres:** Find popular genres

F. **Charts:** Find the critic-favorite and audience-favorite actors

## Approach:

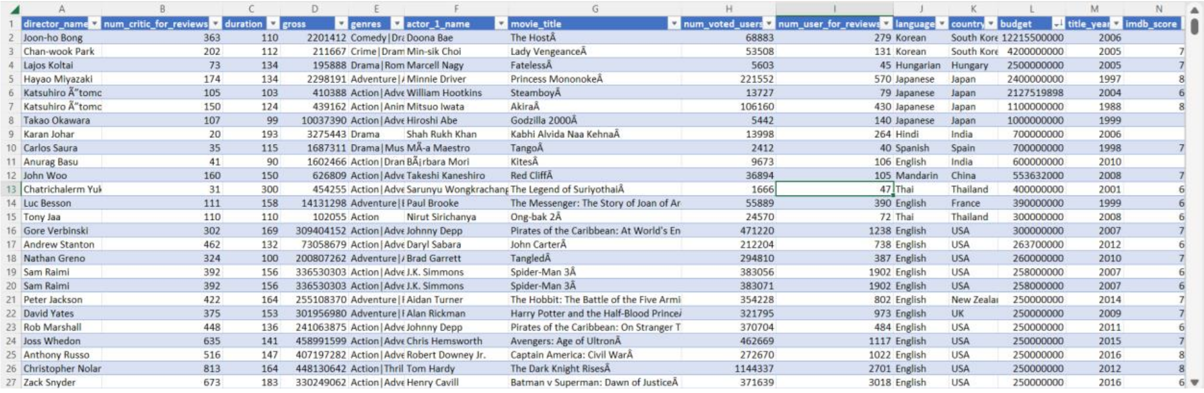
After downloading the provided dataset I used MS Excel, understood each column's data and what exactly they mean, and used mostly pivot tables meanwhile some data can be extracted through statistics and different formulas to lay out required data to the company.

**Note:** The word "data" used in the formulae is the name of the clean table.

We are required to provide a detailed report for the below data record mentioning the answers to the questions that follow:

**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

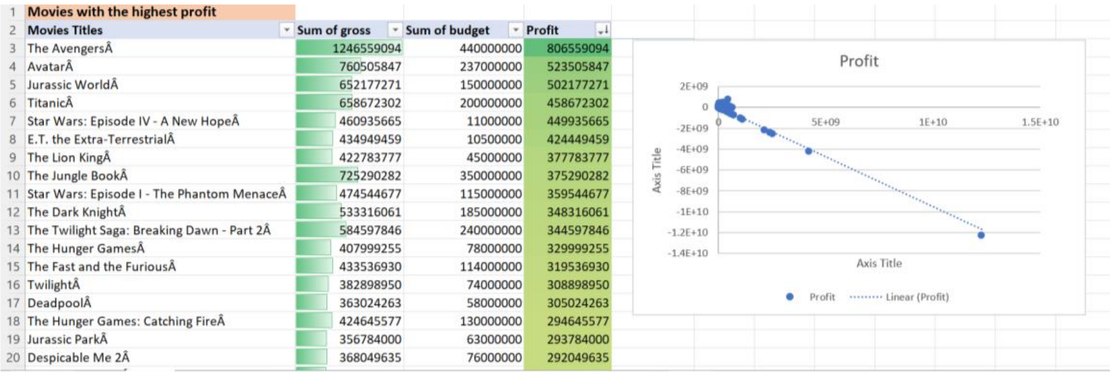
- Initially, I dropped the columns that are not essential for the analysis task and the rows that contain null values.
- It is like purifying the dataset and after purification, the dataset looks like the below containing only useful columns without any null values:



|  | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | director_name | num_critic_for_reviews | duration | gross | genres | actor_1_name | movie_title | num_voted_users | num_user_for_reviews | language | country | budget | title_year | imdb_score |
| 2 | Joon-ho Bong | 363 | 110 | 2201412 | Comedy\|Dr | Doona Bae | The HostÂ | 68883 | 279 | Korean | South Kore | 12215500000 | 2006 |  |
| 3 | Chan-wook Park | 202 | 112 | 211667 | Crime\|Dram | Min-sik Choi | Lady VengeanceÂ | 53508 | 131 | Korean | South Kore | 4200000000 | 2005 | 7 |
| 4 | Lajos Koltai | 73 | 134 | 195888 | Drama\|Rom | Marcell Nagy | FatelessÂ | 5603 | 45 | Hungarian | Hungary | 2500000000 | 2005 | 7 |
| 5 | Hayao Miyazaki | 174 | 134 | 2298191 | Adventure\|/ | Minnie Driver | Princess MononokeÂ | 221552 | 570 | Japanese | Japan | 2400000000 | 1997 | 8 |
| 6 | Katsuhiro Ã”tomc | 105 | 103 | 410388 | Action\|Adve | William Hootkins | SteamboyÂ | 13727 | 79 | Japanese | Japan | 2127519898 | 2004 | 6 |
| 7 | Katsuhiro Ã”tomc | 150 | 124 | 439162 | Action\|Anin | Mitsuo Iwata | AkiraÂ | 106160 | 430 | Japanese | Japan | 1100000000 | 1988 | 8 |
| 8 | Takao Okawara | 107 | 99 | 10037390 | Action\|Adve | Hiroshi Abe | Godzilla 2000Â | 5442 | 140 | Japanese | Japan | 1000000000 | 1999 |  |
| 9 | Karan Johar | 20 | 193 | 3275443 | Drama | Shah Rukh Khan | Kabhi Alvida Naa KehnaÂ | 13998 | 264 | Hindi | India | 700000000 | 2006 |  |
| 10 | Carlos Saura | 35 | 115 | 1687311 | Drama\|Mus | MÃ-a Maestro | TangoÂ | 2412 | 40 | Spanish | Spain | 700000000 | 1998 | 7 |
| 11 | Anurag Basu | 41 | 90 | 1602466 | Action\|Dran | BÃ¡rbara Mori | KitesÂ | 9673 | 106 | English | India | 600000000 | 2010 |  |
| 12 | John Woo | 160 | 150 | 626809 | Action\|Adve | Takeshi Kaneshiro | Red CliffÂ | 36894 | 105 | Mandarin | China | 553632000 | 2008 | 7 |
| 13 | Chatrichalerm Yuk | 31 | 300 | 454255 | Action\|Adve | Sarunyu Wongkrachang | The Legend of SuriyothaiÂ | 1666 | 47 | Thai | Thailand | 400000000 | 2001 | 6 |
| 14 | Luc Besson | 111 | 158 | 14131298 | Adventure\|I | Paul Brooke | The Messenger: The Story of Joan of Ar | 55889 | 390 | English | France | 390000000 | 1999 | 6 |
| 15 | Tony Jaa | 110 | 110 | 102055 | Action | Nirut Sirichanya | Ong-bak 2Â | 24570 | 72 | Thai | Thailand | 300000000 | 2008 | 6 |
| 16 | Gore Verbinski | 302 | 169 | 309404152 | Action\|Adve | Johnny Depp | Pirates of the Caribbean: At World's En | 471220 | 1238 | English | USA | 300000000 | 2007 | 7 |
| 17 | Andrew Stanton | 462 | 132 | 73058679 | Action\|Adve | Daryl Sabara | John CarterÂ | 212204 | 738 | English | USA | 263700000 | 2012 | 6 |
| 18 | Nathan Greno | 324 | 100 | 200807262 | Adventure\|/ | Brad Garrett | TangledÂ | 294810 | 387 | English | USA | 260000000 | 2010 | 7 |
| 19 | Sam Raimi | 392 | 156 | 336530303 | Action\|Adve | J.K. Simmons | Spider-Man 3Â | 383056 | 1902 | English | USA | 258000000 | 2007 | 6 |
| 20 | Sam Raimi | 392 | 156 | 336530303 | Action\|Adve | J.K. Simmons | Spider-Man 3Â | 383071 | 1902 | English | USA | 258000000 | 2007 | 6 |
| 21 | Peter Jackson | 422 | 164 | 255108370 | Adventure\|I | Aidan Turner | The Hobbit: The Battle of the Five Armi | 354228 | 802 | English | New Zealar | 250000000 | 2014 | 7 |
| 22 | David Yates | 375 | 153 | 301956980 | Adventure\|I | Alan Rickman | Harry Potter and the Half-Blood Prince | 321795 | 973 | English | UK | 250000000 | 2009 | 7 |
| 23 | Rob Marshall | 448 | 136 | 241063875 | Action\|Adve | Johnny Depp | Pirates of the Caribbean: On Stranger T | 370704 | 484 | English | USA | 250000000 | 2011 | 6 |
| 24 | Joss Whedon | 635 | 141 | 458991599 | Action\|Adve | Chris Hemsworth | Avengers: Age of UltronÂ | 462669 | 1117 | English | USA | 250000000 | 2015 | 7 |
| 25 | Anthony Russo | 516 | 147 | 407197282 | Action\|Adve | Robert Downey Jr. | Captain America: Civil WarÂ | 272670 | 1022 | English | USA | 250000000 | 2016 | 8 |
| 26 | Christopher Nolar | 813 | 164 | 448130642 | Action\|Thril | Tom Hardy | The Dark Knight RisesÂ | 1144337 | 2701 | English | USA | 250000000 | 2012 | 8 |
| 27 | Zack Snyder | 673 | 183 | 330249062 | Action\|Adve | Henry Cavill | Batman v Superman: Dawn of JusticeÂ | 371639 | 3018 | English | USA | 250000000 | 2016 | 6 |

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?

- For this, a total of four columns were created containing the movie title, movie gross, movie budget, and movie profit.
- This calculation can be done by -
  - A pivot table by creating a measure called "**PROFIT**" and subtracting the sum of **Gross** and sum of **Budget**.
  - And also by manually extracting data regarding each movie and evaluating **PROFIT** by taking the difference between two columns **Gross** and **Budget**.
- And plotted profit (y-axis) vs budget (x-axis) chart using XY Scatter chart.
- Where "**The Avengers**" is the highest profit movie with a gross margin of Rs. 1,24,65,59,094 before **Avatar, Jurassic World,** and others.



Movies with the highest profit

| Movies Titles | Sum of gross | Sum of budget | Profit |
|---|---|---|---|
| The AvengersÂ | 1246559094 | 440000000 | 806559094 |
| AvatarÂ | 760505847 | 237000000 | 523505847 |
| Jurassic WorldÂ | 652177271 | 150000000 | 502177271 |
| TitanicÂ | 658672302 | 200000000 | 458672302 |
| Star Wars: Episode IV - A New HopeÂ | 460935665 | 11000000 | 449935665 |
| E.T. the Extra-TerrestrialÂ | 434949459 | 10500000 | 424449459 |
| The Lion KingÂ | 422783777 | 45000000 | 377783777 |
| The Jungle BookÂ | 725290282 | 350000000 | 375290282 |
| Star Wars: Episode I - The Phantom MenaceÂ | 474544677 | 115000000 | 359544677 |
| The Dark KnightÂ | 533316061 | 185000000 | 348316061 |
| The Twilight Saga: Breaking Dawn - Part 2Â | 584597846 | 240000000 | 344597846 |
| The Hunger GamesÂ | 407999255 | 78000000 | 329999255 |
| The Fast and the FuriousÂ | 433536930 | 114000000 | 319536930 |
| TwilightÂ | 382898950 | 74000000 | 308898950 |
| DeadpoolÂ | 363024263 | 58000000 | 305024263 |
| The Hunger Games: Catching FireÂ | 424645577 | 130000000 | 294645577 |
| Jurassic ParkÂ | 356784000 | 63000000 | 293784000 |
| Despicable Me 2Â | 368049635 | 76000000 | 292049635 |

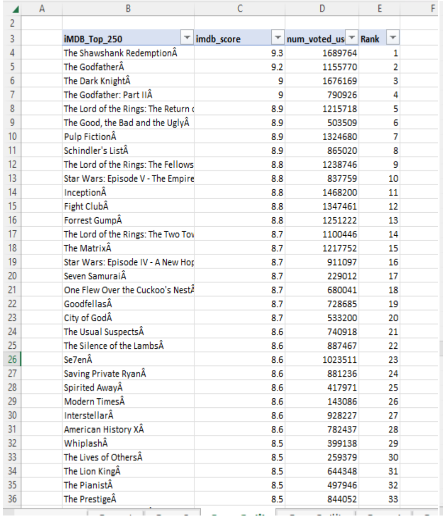
C. **Top 250:** Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add
a Rank column containing the values 1 to 250 indicating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

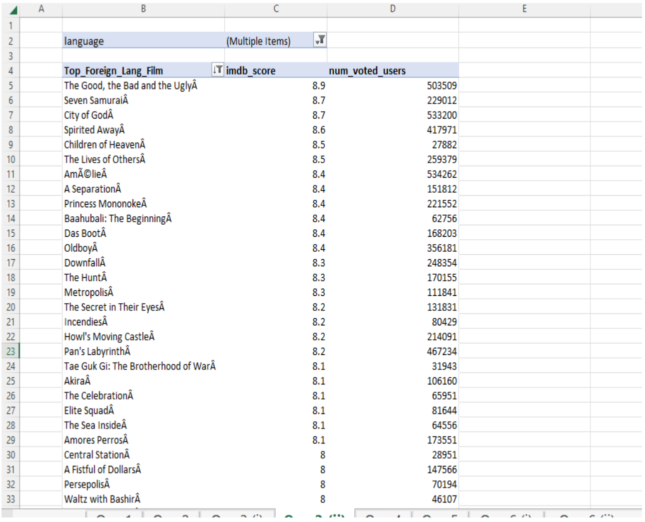
- Here IMDb top 250 movies have been discovered and ranked from 1 to 250 based on the IMDb rating where the number of users who vote for each movie is greater than 25,000.
- For this I have used a pivot table and taken the movie title, IMDb ratings, and the number of user votes for each movie.
- Rank of each movie is calculated using the following formula

  = RANK(C4,$M$2:$M$250,0)+COUNTIF(C$2:$M4,C4)-1

  By the way, the above formula is used to give unique rank values to the specific column.



| iMDB_Top_250 | imdb_score | num_voted_us | Rank |
|---|---|---|---|
| The Shawshank RedemptionÂ | 9.3 | 1689764 | 1 |
| The GodfatherÂ | 9.2 | 1155770 | 2 |
| The Dark KnightÂ | 9 | 1676169 | 3 |
| The Godfather: Part IIÂ | 9 | 790926 | 4 |
| The Lord of the Rings: The Return c | 8.9 | 1215718 | 5 |
| The Good, the Bad and the UglyÂ | 8.9 | 503509 | 6 |
| Pulp FictionÂ | 8.9 | 1324680 | 7 |
| Schindler's ListÂ | 8.9 | 865020 | 8 |
| The Lord of the Rings: The Fellows | 8.8 | 1238746 | 9 |
| Star Wars: Episode V - The Empire | 8.8 | 837759 | 10 |
| InceptionÂ | 8.8 | 1468200 | 11 |
| Fight ClubÂ | 8.8 | 1347461 | 12 |
| Forrest GumpÂ | 8.8 | 1251222 | 13 |
| The Lord of the Rings: The Two Tov | 8.7 | 1100446 | 14 |
| The MatrixÂ | 8.7 | 1217752 | 15 |
| Star Wars: Episode IV - A New Hop | 8.7 | 911097 | 16 |
| Seven SamuraiÂ | 8.7 | 229012 | 17 |
| One Flew Over the Cuckoo's NestÂ | 8.7 | 680041 | 18 |
| GoodfellasÂ | 8.7 | 728685 | 19 |
| City of GodÂ | 8.7 | 533200 | 20 |
| The Usual SuspectsÂ | 8.6 | 740918 | 21 |
| The Silence of the LambsÂ | 8.6 | 887467 | 22 |
| Se7enÂ | 8.6 | 1023511 | 23 |
| Saving Private RyanÂ | 8.6 | 881236 | 24 |
| Spirited AwayÂ | 8.6 | 417971 | 25 |
| Modern TimesÂ | 8.6 | 143086 | 26 |
| InterstellarÂ | 8.6 | 928227 | 27 |
| American History XÂ | 8.6 | 782437 | 28 |
| WhiplashÂ | 8.5 | 399138 | 29 |
| The Lives of OthersÂ | 8.5 | 259379 | 30 |
| The Lion KingÂ | 8.5 | 644348 | 31 |
| The PianistÂ | 8.5 | 497946 | 32 |
| The PrestigeÂ | 8.5 | 844052 | 33 |



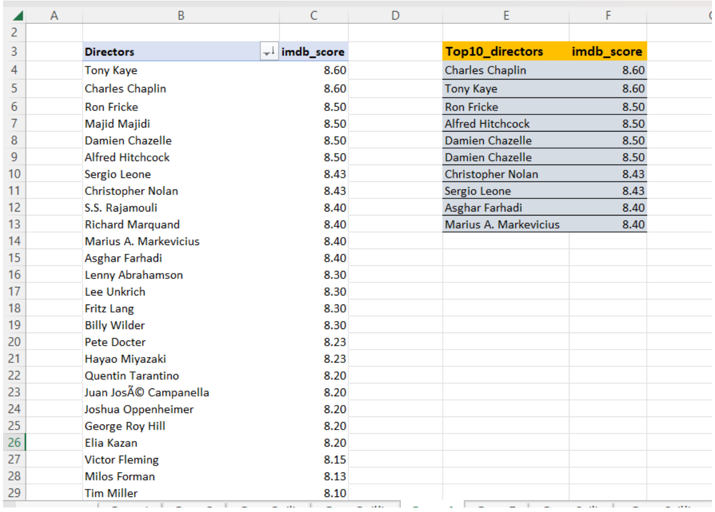
language (Multiple Items)

| Top_Foreign_Lang_Film | imdb_score | num_voted_users |
|---|---|---|
| The Good, the Bad and the UglyÂ | 8.9 | 503509 |
| Seven SamuraiÂ | 8.7 | 229012 |
| City of GodÂ | 8.7 | 533200 |
| Spirited AwayÂ | 8.6 | 417971 |
| Children of HeavenÂ | 8.5 | 27882 |
| The Lives of OthersÂ | 8.5 | 259379 |
| AmÃ©lieÂ | 8.4 | 534262 |
| A SeparationÂ | 8.4 | 151812 |
| Princess MononokeÂ | 8.4 | 221552 |
| Baahubali: The BeginningÂ | 8.4 | 62756 |
| Das BootÂ | 8.4 | 168203 |
| OldboyÂ | 8.4 | 356181 |
| DownfallÂ | 8.3 | 248354 |
| The HuntÂ | 8.3 | 170155 |
| MetropolisÂ | 8.3 | 111841 |
| The Secret in Their EyesÂ | 8.2 | 131831 |
| IncendiesÂ | 8.2 | 80429 |
| Howl's Moving CastleÂ | 8.2 | 214091 |
| Pan's LabyrinthÂ | 8.2 | 467234 |
| Tae Guk Gi: The Brotherhood of WarÂ | 8.1 | 31943 |
| AkiraÂ | 8.1 | 106160 |
| The CelebrationÂ | 8.1 | 65951 |
| Elite SquadÂ | 8.1 | 81644 |
| The Sea InsideÂ | 8.1 | 64556 |
| Amores PerrosÂ | 8.1 | 173551 |
| Central StationÂ | 8 | 28951 |
| A Fistful of DollarsÂ | 8 | 147566 |
| PersepolisÂ | 8 | 70194 |
| Waltz with BashirÂ | 8 | 46107 |

- The next part of the answer contains all the movies from the IMDb_Top_250 column which are not in the **English** language and are stored in the new column named Top_Foreign_Lang_Film.
- Separation of English language film and other language films is done by putting **Language** in the filter field in the pivot table and just unselecting the option **English.**
- Then it will display all the movies which are not In the **English** language.

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

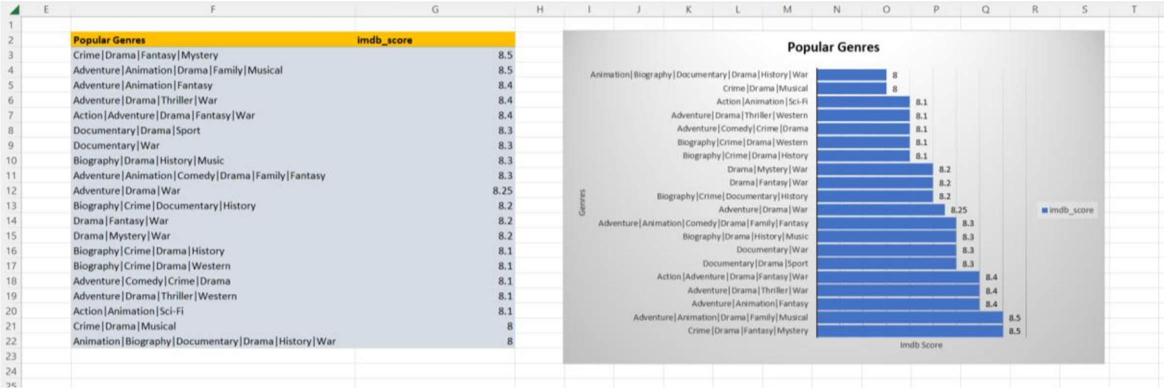
- From the list of directors' names, I have extracted 10 directors with the highest IMDb ratings, and in the case of tie-in ratings, I have manually sorted them in alphabetical order.



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 2 | | | | | | |
| 3 | | Directors | imdb_score | | Top10_directors | imdb_score |
| 4 | | Tony Kaye | 8.60 | | Charles Chaplin | 8.60 |
| 5 | | Charles Chaplin | 8.60 | | Tony Kaye | 8.60 |
| 6 | | Ron Fricke | 8.50 | | Ron Fricke | 8.50 |
| 7 | | Majid Majidi | 8.50 | | Alfred Hitchcock | 8.50 |
| 8 | | Damien Chazelle | 8.50 | | Damien Chazelle | 8.50 |
| 9 | | Alfred Hitchcock | 8.50 | | Damien Chazelle | 8.50 |
| 10 | | Sergio Leone | 8.43 | | Christopher Nolan | 8.43 |
| 11 | | Christopher Nolan | 8.43 | | Sergio Leone | 8.43 |
| 12 | | S.S. Rajamouli | 8.40 | | Asghar Farhadi | 8.40 |
| 13 | | Richard Marquand | 8.40 | | Marius A. Markevicius | 8.40 |
| 14 | | Marius A. Markevicius | 8.40 | | | |
| 15 | | Asghar Farhadi | 8.40 | | | |
| 16 | | Lenny Abrahamson | 8.30 | | | |
| 17 | | Lee Unkrich | 8.30 | | | |
| 18 | | Fritz Lang | 8.30 | | | |
| 19 | | Billy Wilder | 8.30 | | | |
| 20 | | Pete Docter | 8.23 | | | |
| 21 | | Hayao Miyazaki | 8.23 | | | |
| 22 | | Quentin Tarantino | 8.20 | | | |
| 23 | | Juan JosÃ© Campanella | 8.20 | | | |
| 24 | | Joshua Oppenheimer | 8.20 | | | |
| 25 | | George Roy Hill | 8.20 | | | |
| 26 | | Elia Kazan | 8.20 | | | |
| 27 | | Victor Fleming | 8.15 | | | |
| 28 | | Milos Forman | 8.13 | | | |
| 29 | | Tim Miller | 8.10 | | | |

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

- This was simple like the above question where in place of directors just I extracted **GENRES** and their IMDb scores.
- **GENRES** with the highest IMDb scores are named **Popular Genres.**



| Popular Genres | imdb_score |
|---|---|
| Crime\|Drama\|Fantasy\|Mystery | 8.5 |
| Adventure\|Animation\|Drama\|Family\|Musical | 8.5 |
| Adventure\|Animation\|Fantasy | 8.4 |
| Adventure\|Drama\|Thriller\|War | 8.4 |
| Action\|Adventure\|Drama\|Fantasy\|War | 8.4 |
| Documentary\|Drama\|Sport | 8.3 |
| Documentary\|War | 8.3 |
| Biography\|Drama\|History\|Music | 8.3 |
| Adventure\|Animation\|Comedy\|Drama\|Family\|Fantasy | 8.3 |
| Adventure\|Drama\|War | 8.25 |
| Biography\|Crime\|Documentary\|History | 8.2 |
| Drama\|Fantasy\|War | 8.2 |
| Drama\|Mystery\|War | 8.2 |
| Biography\|Crime\|Drama\|History | 8.1 |
| Biography\|Crime\|Drama\|Western | 8.1 |
| Adventure\|Comedy\|Crime\|Drama | 8.1 |
| Adventure\|Drama\|Thriller\|Western | 8.1 |
| Action\|Animation\|Sci-Fi | 8.1 |
| Crime\|Drama\|Musical | 8 |
| Animation\|Biography\|Documentary\|Drama\|History\|War | 8 |

Popular Genres
Genres
Imdb Score
imdb_score
Animation|Biography|Documentary|Drama|History|War 8
Crime|Drama|Musical 8
Action|Animation|Sci-Fi 8.1
Adventure|Drama|Thriller|Western 8.1
Adventure|Comedy|Crime|Drama 8.1
Biography|Crime|Drama|Western 8.1
Biography|Crime|Drama|History 8.1
Drama|Mystery|War 8.2
Drama|Fantasy|War 8.2
Biography|Crime|Documentary|History 8.2
Adventure|Drama|War 8.25
Adventure|Animation|Comedy|Drama|Family|Fantasy 8.3
Biography|Drama|History|Music 8.3
Documentary|War 8.3
Documentary|Drama|Sport 8.3
Action|Adventure|Drama|Fantasy|War 8.4
Adventure|Drama|Thriller|War 8.4
Adventure|Animation|Fantasy 8.4
Adventure|Animation|Drama|Family|Musical 8.5
Crime|Drama|Fantasy|Mystery 8.5

**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.

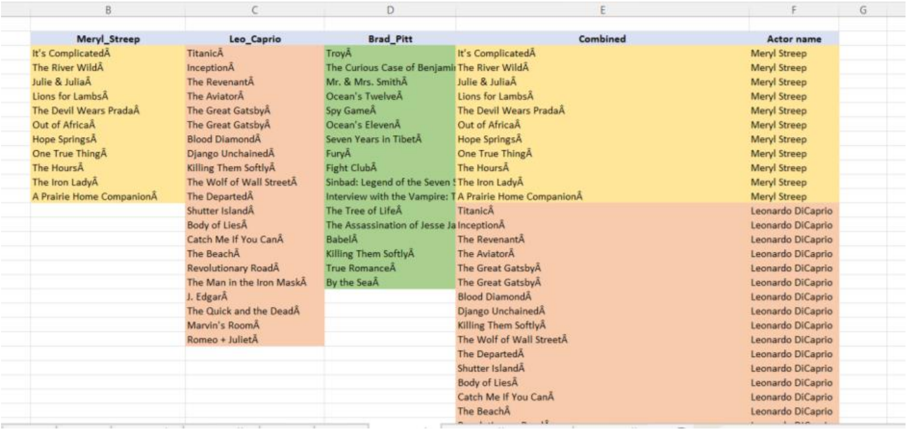
Group the combined column using the actor_1_name column.

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

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

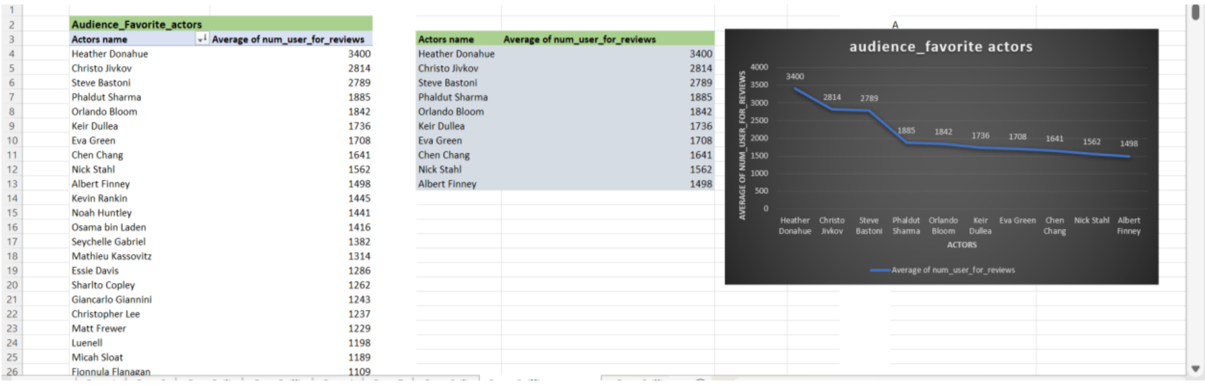
**Your task:** Find the critic-favorite and audience-favorite actors

- Simply listed names of actors in one column and movies in another one then filtered the actors' names and selected only those three actors(**Meryl Streep, Leonardo DiCaprio,** and **Brad Pitt**) then automatically we got movies name where **Meryl Streep, Leonardo DiCaprio,** and **Brad Pitt** are lead actors.
- And made three columns named "Meryl_Streep", "Leo_Caprio", and "Brad_Pitt" and inserted their movies in rows, lastly appended the rows of all three columns in the "Combined" column group by actor name.



| Meryl_Streep | Leo_Caprio | Brad_Pitt | Combined | Actor name |
|---|---|---|---|---|
| It's ComplicatedÂ | TitanicÂ | TroyÂ | It's ComplicatedÂ | Meryl Streep |
| The River WildÂ | InceptionÂ | The Curious Case of Benjami | The River WildÂ | Meryl Streep |
| Julie & JuliaÂ | The RevenantÂ | Mr. & Mrs. SmithÂ | Julie & JuliaÂ | Meryl Streep |
| Lions for LambsÂ | The AviatorÂ | Ocean's TwelveÂ | Lions for LambsÂ | Meryl Streep |
| The Devil Wears PradaÂ | The Great GatsbyÂ | Spy GameÂ | The Devil Wears PradaÂ | Meryl Streep |
| Out of AfricaÂ | The Great GatsbyÂ | Ocean's ElevenÂ | Out of AfricaÂ | Meryl Streep |
| Hope SpringsÂ | Blood DiamondÂ | Seven Years in TibetÂ | Hope SpringsÂ | Meryl Streep |
| One True ThingÂ | Django UnchainedÂ | FuryÂ | One True ThingÂ | Meryl Streep |
| The HoursÂ | Killing Them SoftlyÂ | Fight ClubÂ | The HoursÂ | Meryl Streep |
| The Iron LadyÂ | The Wolf of Wall StreetÂ | Sinbad: Legend of the Seven S | The Iron LadyÂ | Meryl Streep |
| A Prairie Home CompanionÂ | The DepartedÂ | Interview with the Vampire: T | A Prairie Home CompanionÂ | Meryl Streep |
|  | Shutter IslandÂ | The Tree of LifeÂ | TitanicÂ | Leonardo DiCaprio |
|  | Body of LiesÂ | The Assassination of Jesse Ja | InceptionÂ | Leonardo DiCaprio |
|  | Catch Me If You CanÂ | BabelÂ | The RevenantÂ | Leonardo DiCaprio |
|  | The BeachÂ | Killing Them SoftlyÂ | The AviatorÂ | Leonardo DiCaprio |
|  | Revolutionary RoadÂ | True RomanceÂ | The Great GatsbyÂ | Leonardo DiCaprio |
|  | The Man in the Iron MaskÂ | By the SeaÂ | The Great GatsbyÂ | Leonardo DiCaprio |
|  | J. EdgarÂ |  | Blood DiamondÂ | Leonardo DiCaprio |
|  | The Quick and the DeadÂ |  | Django UnchainedÂ | Leonardo DiCaprio |
|  | Marvin's RoomÂ |  | Killing Them SoftlyÂ | Leonardo DiCaprio |
|  | Romeo + JulietÂ |  | The Wolf of Wall StreetÂ | Leonardo DiCaprio |
|  |  |  | The DepartedÂ | Leonardo DiCaprio |
|  |  |  | Shutter IslandÂ | Leonardo DiCaprio |
|  |  |  | Body of LiesÂ | Leonardo DiCaprio |
|  |  |  | Catch Me If You CanÂ | Leonardo DiCaprio |
|  |  |  | The BeachÂ | Leonardo DiCaprio |

- For audience favorite actors, actors who have the highest mean of numbers users' reviews are considered as audience favorite actors whereas **Heather Donahue** holds the 1st position.
- Audience favorite actors are based on the average of number of users' reviews which I calculated using a pivot table.
- Another way of calculating averages with a condition is the **AVERAGEIF** formula and if it has more than one condition **AVERAGEIFS** formula can be used.

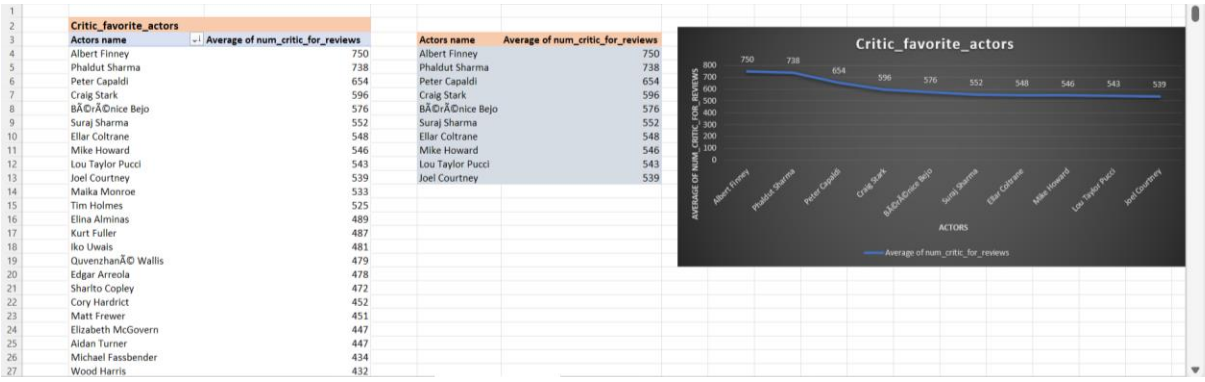


Audience_Favorite_actors

| Actors 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 |
| Eva Green | 1708 |
| Chen Chang | 1641 |
| Nick Stahl | 1562 |
| Albert Finney | 1498 |
| Kevin Rankin | 1445 |
| Noah Huntley | 1441 |
| Osama bin Laden | 1416 |
| Seychelle Gabriel | 1382 |
| Mathieu Kassovitz | 1314 |
| Essie Davis | 1286 |
| Sharlto Copley | 1262 |
| Giancarlo Giannini | 1243 |
| Christopher Lee | 1237 |
| Matt Frewer | 1229 |
| Luenell | 1198 |
| Micah Sloat | 1189 |
| Fionnula Flanagan | 1109 |

| Actors 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 |
| Eva Green | 1708 |
| Chen Chang | 1641 |
| Nick Stahl | 1562 |
| Albert Finney | 1498 |

- ○ Here I used the same approach as for audience favorite actors, just in place of average users' reviews, the average number of critics' reviews has been considered. Albert Finney named as critic favorite actor

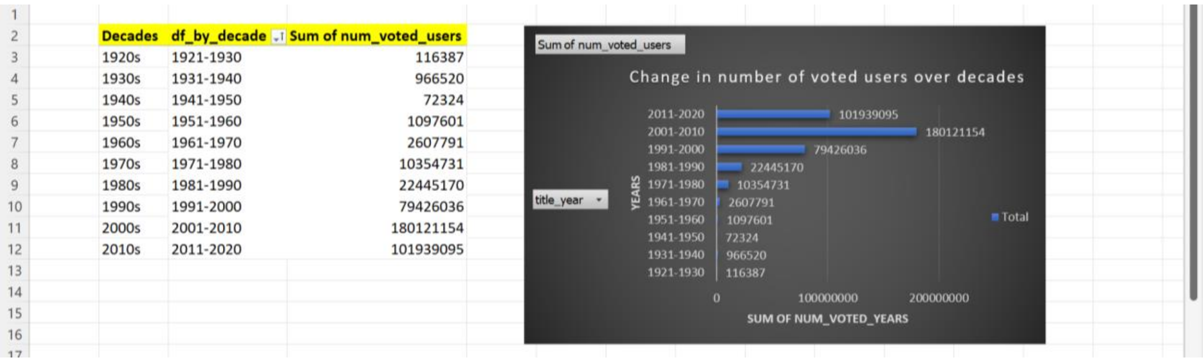


Critic_favorite_actors

| Actors 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 | 479 |
| Edgar Arreola | 478 |
| Sharlto Copley | 472 |
| Cory Hardrict | 452 |
| Matt Frewer | 451 |
| Elizabeth McGovern | 447 |
| Aidan Turner | 447 |
| Michael Fassbender | 434 |
| Wood Harris | 432 |

| Actors 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 |

- Using a pivot table, title_years are taken in a row and the sum of the number of users' votes taken in values.
- And grouped the column "title_years" with the range of 10 years starting from 1921.
- Another column named "Decades" has also been created containing decades.



| Decades | df_by_decade | Sum of num_voted_users |
|---|---|---|
| 1920s | 1921-1930 | 116387 |
| 1930s | 1931-1940 | 966520 |
| 1940s | 1941-1950 | 72324 |
| 1950s | 1951-1960 | 1097601 |
| 1960s | 1961-1970 | 2607791 |
| 1970s | 1971-1980 | 10354731 |
| 1980s | 1981-1990 | 22445170 |
| 1990s | 1991-2000 | 79426036 |
| 2000s | 2001-2010 | 180121154 |
| 2010s | 2011-2020 | 101939095 |

Almost most of the above questions are solved using pivot tables meanwhile for some calculations we have extracted data using statistics and formulae. The PIVOT table technique is used to summarize large amounts of data in a more effective way.

## Tech-Stack Used:

1. **MS Excel:** For answering all the questions I used MS Excel wherein pivot tables played a major role in representing the senseful data in a graphical manner from the dataset for a better understanding.
2. **MS Word:** It is used for making the report.

## Insights:

After solving the above questions, we get some senseful data from the dataset where I discovered:

1. South Korean film "*The Host*" directed by **Joon-ho Bong** in **2006** is the lowest-profit movie.
2. "*The Avengers*" is the highest-grossing film with the highest profit of approx. 806 million whose director is Joss Whedon.
3. The Avatar, Jurassic World, Titanic, Star Wars: Episode IV – A New Hope, The Lion King, etc. like English language films mostly dominated the film industry over the world.
4. According to IMDb ratings, "*The Shawshank Redemption"* movie is the top-rated movie with an IMDb score of 9.3.

5. Other than English films, the French movie “*Amelie*” with 8.4 IMDb ratings has the highest votes by users i.e. 534264 votes. And the Italian movie “*The Good, The Bad and The Ugly*” is the highest IMDb-rated movie with a rating of **8.9.**
6. The director of the “*Modern Times*” movie named **Charles Champlin** and the director of the **“***American History X***”** movie **Tony Clave** have the highest mean IMDb score.
7. **Crime|Drama|Fantasy|Mystery** and **Adventure|Animation|Drama|Family|Musical** types of genres are the most rated genres by users.
8. People love such types of genres such as “*The Lion King”* movie which is an animated movie and “*Bahubali: The Beginning”* movie which is the **Action+Adventure** genre.
9. Over the globe, audiences have shown lots of interest in Science-Fiction (Sci-Fi) movies where “*The Avengers”, “The Avatar”, and “Jurassic World”* are among them
10. Throughout the entire career of Leonardo DiCaprio, Meryl Streep, and Brad Pitt,

| Actors | Total Movies | Total User reviews | Total Critic reviews |
|---|---|---|---|
| **Leonardo DiCaprio** | 21 | 19204 | 6934 |
| **Brad Pitt** | 17 | 12620 | 4165 |
| **Meryl Streep** | 11 | 3269 | 1996 |

11. **Heather Donahue** has been named the audience’s favorite actor and **Albert Finney** is the critic’s favorite actor.
12. From the past 10 decades, **2010**s film has the highest number of users’ votes (approx. 108 million) whereas from 1921 users’ votes seem to be increased over decades.

## Result:

It was a great experience to work on a movie dataset where I got detailed knowledge about how to draw out sense from the raw data. Got very much familiar with pivot tables and charts. Understood better ways of visually presenting data in a graphical manner. Easily performed IMDb movie analysis.

## Excel sheet link

https://docs.google.com/spreadsheets/d/1IyVAf153kUJB82GvMyeQ_hD06uYYHE_J/edit?usp=share_link&ouid=102746205635544467842&rtpof=true&sd=true

Please do open it in MS Excel