

Project Description:

The project aims to perform an analysis of the IMDb movie dataset to gain insights into the movie industry. The analysis will involve data cleaning, finding the movies with the highest profit, finding the top 250 movies based on IMDb ratings, finding the best directors, identifying popular genres, identifying the critic-favorite and audience-favorite actors, and observing the change in the number of voted users over decades.

Approach:

The project was approached by downloading the IMDb movie dataset and performing data cleaning using Excel. The necessary columns were selected, and null values were removed. The movies with the highest profit were identified by creating a new column called profit, which contains the difference between gross and budget. The IMDb top 250 movies were identified based on IMDb ratings and a minimum number of voted users. The top directors and popular genres were identified using appropriate grouping and aggregation techniques. The lead actors were identified, and a new column was created to combine their movies. The change in the number of voted users over decades was observed using a bar chart.

Tech-Stack Used:

The project was performed using Excel,. The purpose of using Excel was to perform data cleaning, grouping, aggregation, and visualization tasks.

Insights:

The analysis revealed several insights about the movie industry. The movies with the highest profit were mostly from popular franchises, such as Marvel and Star Wars. The top 250 movies were dominated by English language films, but several foreign language films also made it to the list. The best directors were identified based on their IMDb ratings, with Steven Spielberg topping the list. The popular genres were found to be Drama, Comedy, and Action. Meryl Streep, Leonardo DiCaprio, and Brad Pitt were found to be the critic-favorite and audience-favorite actors. The change in the number of voted users over decades showed a significant increase in the number of users voting for movies in recent years.

Result:

The project successfully identified several insights about the movie industry, such as the most profitable movies, the top 250 movies based on IMDb ratings, the best directors, popular genres, and critic-favorite and audience-favorite actors. The insights gained from this project can be used by the movie industry to make informed decisions about future movie projects. Overall, the project helped in improving the understanding of the movie industry and the factors that contribute to a movie's success.

Regenerate response

Cleaning the data:: PThis 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

Cleaning the Data:

To clean the data, we can start by removing the unnecessary columns which are not required for our analysis. These columns include director_facebook_likes, actor_3_facebook_likes, actor_2_facebook_likes, actor_1_facebook_likes, cast_total_facebook_likes, facenumber_in_poster, plot_keywords, movie_imdb_link, language, country, content_rating, and aspect_ratio. Also, the columns with missing values should be removed or filled with appropriate values, also we have to remove duplicate row

Approach for deleting row firstly we have to delete particular column click on particular column and delete entire column.

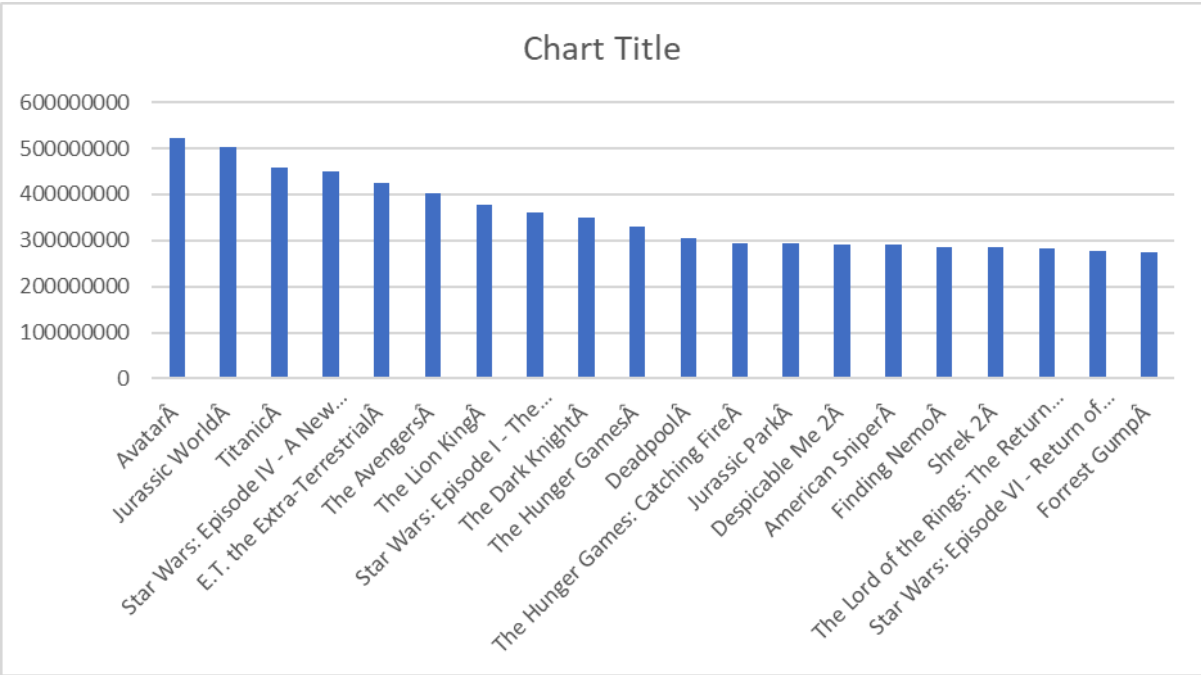
Delete null values select entire table click on **Cntl+G** then select special and click on blanks we got all blank values then click **cntl + -** then delete entire row

Delete duplicate values click on **Cntl+T** then create a table then click on remove duplicates

We have 5044 rows after deleting null values rows and duplicate column we got 3724 rows

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?



Answer movie with highest profit

imdb_score	aspect_ratio	movie_facebook_likes	Profit
7.9	1.78	33000	523505847
7	2	150000	502177271
7.7	2.35	26000	458672302
8.7	2.35	33000	449935665
7.9	1.85	34000	424449459

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

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.

The Shawshank Redemption

The Godfather

The Dark Knight

The Godfather: Part II

The Lord of the Rings: The Return of the King

Schindler's List

Pulp Fiction

The Good, the Bad and the Ugly

Inception

The Lord of the Rings: The Fellowship of the Ring

Fight Club

Forrest Gump

Star Wars: Episode V - The Empire Strikes Back

The Lord of the Rings: The Two Towers

The Matrix

Goodfellas

Star Wars: Episode IV - A New Hope

One Flew Over the Cuckoo's Nest

City of God

Seven Samurai

Interstellar

Saving Private Ryan

Se7en

The Silence of the LambsÂ

Spirited AwayÂ

American History XÂ

The Usual SuspectsÂ

Modern TimesÂ

The Dark Knight RisesÂ

GladiatorÂ

Terminator 2: Judgment DayÂ

Django UnchainedÂ

The DepartedÂ

The Lion KingÂ

The Green MileÂ

The PrestigeÂ

The PianistÂ

Apocalypse NowÂ

Raiders of the Lost ArkÂ

PsychoÂ

Back to the FutureÂ

AlienÂ

MementoÂ

WhiplashÂ

The Lives of OthersÂ

Children of HeavenÂ

WALLÂ·EÂ

BraveheartÂ

AmÃ©lieÂ

Star Wars: Episode VI - Return of the JediÂ

Once Upon a Time in AmericaÂ

Princess MononokeÂ

AliensÂ

American BeautyÂ

Lawrence of ArabiaÂ

Das BootÂ

Requiem for a DreamÂ

OldboyÂ

Reservoir DogsÂ

A SeparationÂ

Toy Story 3Â

UpÂ

Inside OutÂ

Batman BeginsÂ

Inglourious BasterdsÂ

Indiana Jones and the Last CrusadeÂ

L.A. ConfidentialÂ

Toy StoryÂ

ScarfaceÂ

Eternal Sunshine of the Spotless MindÂ

AmadeusÂ

Raging BullÂ

MetropolisÂ

UnforgivenÂ

DownfallÂ

RoomÂ

SnatchÂ

2001: A Space OdysseyÂ

Good Will HuntingÂ

The StingÂ

The HuntÂ

Some Like It HotÂ

Monty Python and the Holy GrailÂ

Captain America: Civil WarÂ

How to Train Your DragonÂ

The Wolf of Wall StreetÂ

Finding NemoÂ

A Beautiful MindÂ

CasinoÂ

V for VendettaÂ

The ThingÂ

Die HardÂ

Blade RunnerÂ

Gran TorinoÂ

WarriorÂ

Howl's Moving CastleÂ

Into the WildÂ

Pan's LabyrinthÂ

The Big LebowskiÂ

IncendiesÂ

Gone with the WindÂ

The Secret in Their EyesÂ

TrainspottingÂ

The Bridge on the River KwaiÂ

Lock, Stock and Two Smoking BarrelsÂ

On the WaterfrontÂ

The AvengersÂ

Guardians of the GalaxyÂ

Mad Max: Fury RoadÂ

The RevenantÂ

The Bourne UltimatumÂ

Pirates of the Caribbean: The Curse of the Black PearlÂ

Monsters, Inc.Â

The MartianÂ

Shutter IslandÂ

Jurassic ParkÂ

Gone GirlÂ

The Truman ShowÂ

DeadpoolÂ

Kill Bill: Vol. 1Â

PrisonersÂ

The Sixth SenseÂ

Sin CityÂ

RushÂ

The Grand Budapest HotelÂ

Million Dollar BabyÂ

The HelpÂ

No Country for Old MenÂ

There Will Be BloodÂ

12 Years a SlaveÂ

SpotlightÂ

Hotel Rwanda

The Imitation Game

The Princess Bride

Groundhog Day

The Sea Inside

Tae Guk Gi: The Brotherhood of War

Stand by Me

Akira

Elite Squad

The Terminator

Platoon

Butch Cassidy and the Sundance Kid

Donnie Darko

Annie Hall

The Wizard of Oz

Before Sunrise

The Best Years of Our Lives

Amores Perros

The Celebration

Rocky

X-Men: Days of Future Past

Ratatouille

Star Trek

Life of Pi

Casino Royale

Blood Diamond

The Incredibles

Cinderella Man

Big Fish

The Pursuit of Happyness

Kill Bill: Vol. 2

Catch Me If You Can

The Iron Giant

JFK

Serenity

Magnolia

District 9

Mystic River

Aladdin

Rain Man

Her

Dances with Wolves

Dead Poets Society

The Artist

The King's Speech

Brazil

In Bruges

Mulholland Drive

Slumdog Millionaire

Black Swan

The Perks of Being a Wallflower

True Romance

Dancer in the Dark

The Exorcist

JawsÂ

Casino RoyaleÂ

Doctor ZhivagoÂ

The Straight StoryÂ

Fiddler on the RoofÂ

SickoÂ

My Name Is KhanÂ

The Sound of MusicÂ

PersepolisÂ

Dallas Buyers ClubÂ

Shaun of the DeadÂ

Sling BladeÂ

BoyhoodÂ

Bowling for ColumbineÂ

Central StationÂ

Young FrankensteinÂ

Before SunsetÂ

Waltz with BashirÂ

A Fistful of DollarsÂ

AvatarÂ

The Hobbit: The Desolation of SmaugÂ

Iron ManÂ

Edge of TomorrowÂ

Big Hero 6Â

How to Train Your Dragon 2Â

Toy Story 2Â

Children of MenÂ

The InsiderÂ

The Hateful EightÂ

The Bourne IdentityÂ

Almost FamousÂ

Captain PhillipsÂ

ShrekÂ

HeroÂ

The NotebookÂ

GloryÂ

Walk the LineÂ

Straight Outta ComptonÂ

The Blues BrothersÂ

The Right StuffÂ

TakenÂ

The UntouchablesÂ

The World's Fastest IndianÂ

Edward ScissorhandsÂ

GloryÂ

Ed WoodÂ

My Fair LadyÂ

HalloweenÂ

Hot FuzzÂ

Crouching Tiger, Hidden DragonÂ

The Remains of the DayÂ

Boogie NightsÂ

Letters from Iwo JimaÂ

The FighterÂ

E.T. the Extra-Terrestrial

Crash

Amour

Nightcrawler

Little Miss Sunshine

The Wrestler

The Chorus

Do the Right Thing

Moon

Before Midnight

Nine Queens

Rank , Top_foreign_film

IMDb_Top_250	imdbscore	Rank	Top_Foreign_Lang_Film
The Shawshank Redemption	9.3	1	The Good, the Bad and the Ugly
The Godfather	9.2	2	City of God
The Dark Knight	9	3	Seven Samurai
The Godfather: Part II	9	3	Spirited Away
The Lord of the Rings: The Return of the King	8.9	5	The Lives of Others
Schindler's List	8.9	5	Children of Heaven
Pulp Fiction	8.9	5	Amélie
The Good, the Bad and the Ugly	8.9	5	Princess Mononoke
Inception	8.8	9	Das Boot
The Lord of the Rings: The Fellowship of the Ring	8.8	9	Oldboy
Fight Club	8.8	9	A Separation
Forrest Gump	8.8	9	Metropolis
Star Wars: Episode V - The Empire Strikes Back	8.8	9	Downfall
The Lord of the Rings: The Two Towers	8.7	14	The Hunt
The Matrix	8.7	14	Howl's Moving Castle
Goodfellas	8.7	14	Pan's Labyrinth
Star Wars: Episode IV - A New Hope	8.7	14	Incendies
One Flew Over the Cuckoo's Nest	8.7	14	The Secret in Their Eyes
City of God	8.7	14	The Sea Inside
Seven Samurai	8.7	14	Tae Guk Gi: The Brotherhood of War
Interstellar	8.6	21	Akira
Saving Private Ryan	8.6	21	Elite Squad
Se7en	8.6	21	Amores Perros
The Silence of the Lambs	8.6	21	The Celebration
Spirited Away	8.6	21	My Name Is Khan
American History X	8.6	21	Persepolis
The Usual Suspects	8.6	21	Central Station
Modern Times	8.6	21	Waltz with Bashir
The Dark Knight Rises	8.5	29	A Fistful of Dollars
Gladiator	8.5	29	Hero
Terminator 2: Judgment Day	8.5	29	Crouching Tiger, Hidden Dragon
Django Unchained	8.5	29	Letters from Iwo Jima
The Departed	8.5	29	Amour
The Lion King	8.5	29	The Chorus
The Green Mile	8.5	29	Nine Queens

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

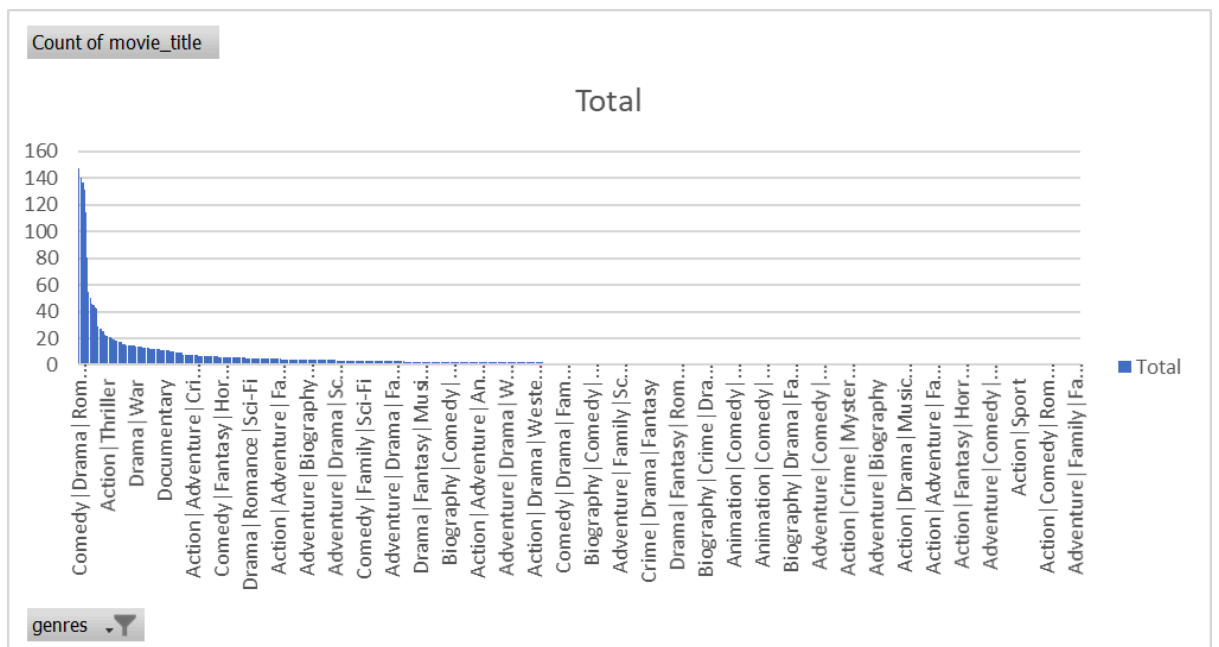
Using pivot table

top10director	mean of imdb_score
Akira Kurosawa	8.7
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4

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

Your task: Find popular genres

Comedy Drama Romance	147
Drama	140
Comedy	137
Comedy Drama	137
Comedy Romance	131
Drama Romance	114
Crime Drama Thriller	81
Action Crime Thriller	55
Action Crime Drama Thriller	50
Action Adventure Sci-Fi	46
Comedy Crime	45
Action Adventure Thriller	45
Horror	43
Drama Thriller	42
Crime Drama Mystery Thriller	41
Crime Drama	40
Action Adventure Sci-Fi Thriller	34
Horror Thriller	33
Horror Mystery Thriller	31
Drama Mystery Thriller	30



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.

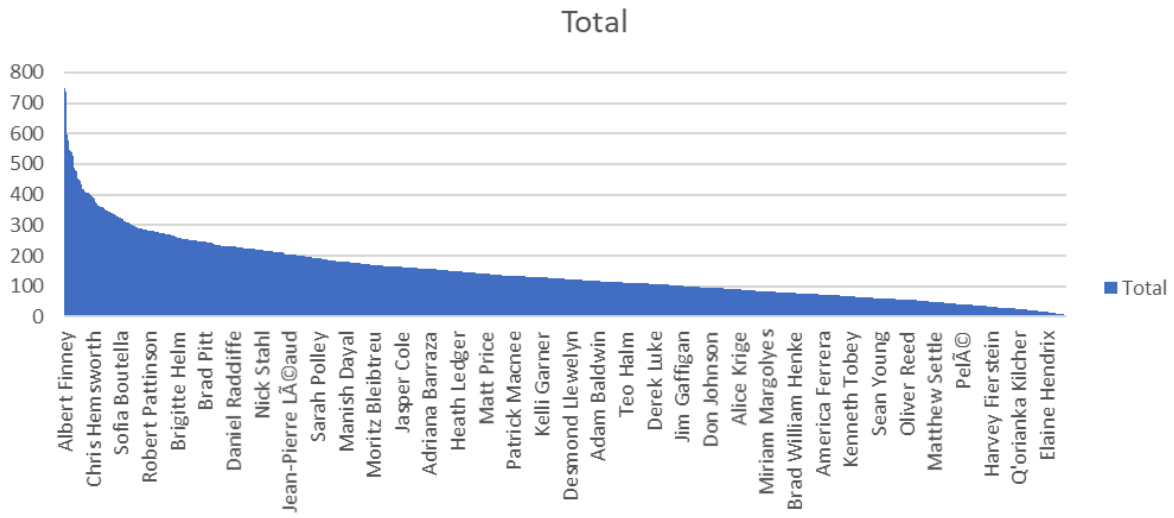
Group the combined column using the actor_1_name column.

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

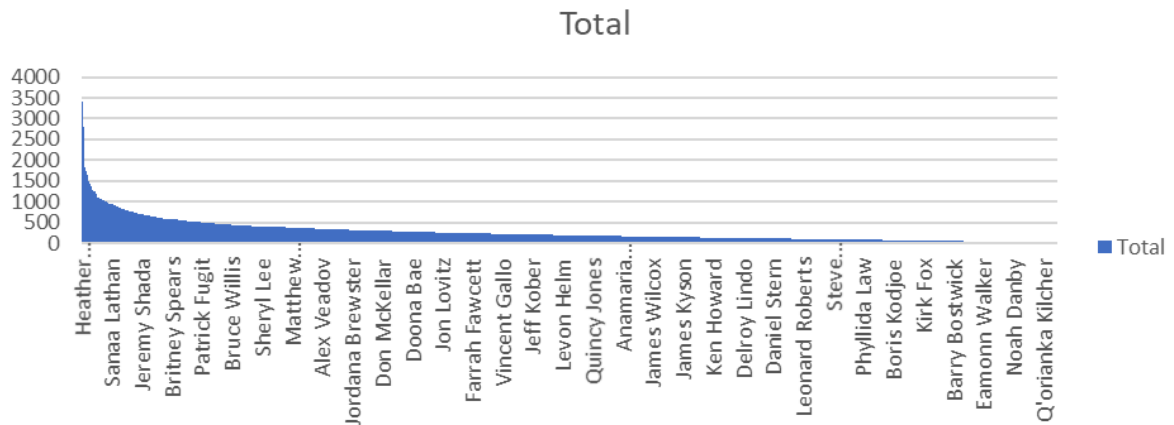
Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade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

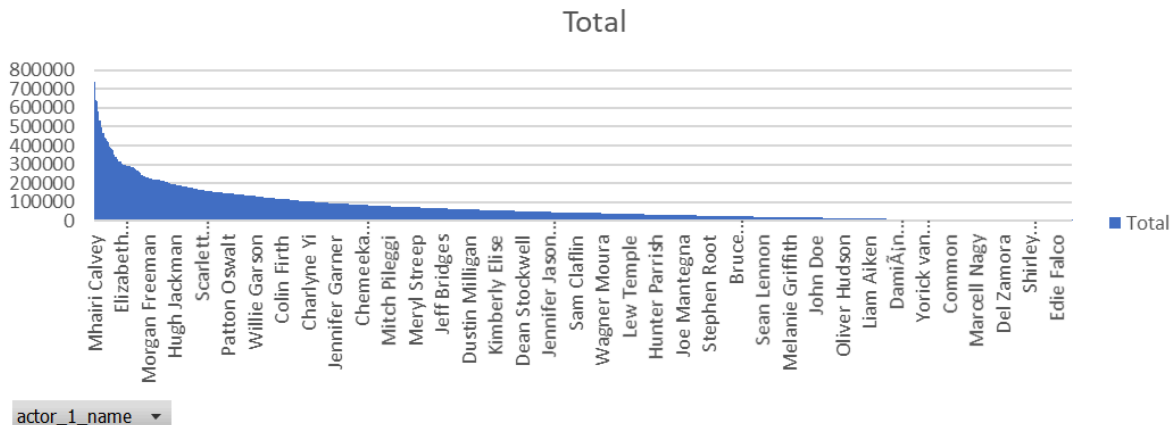
Average of num_critics_for_reviews



Average of num_user_for_reviews



Average of num_voted_users



Row Labels	Average of num_critic_for_reviews	Average of num_user_for_reviews
Brad Pitt	245	742.3529412
Leonardo DiCaprio	330.1904762	914.4761905
Meryl Streep	181.4545455	297.1818182

Meryl_Streep	Leo_Caprio	Brad_Pitt	combined
It's Complicated	Titanic	The Curious Case of Benjamin Franklin	It's Complicated
The River Wild	The Great Gatsby	Troy	The River Wild
Julie & Julia	Inception	Ocean's Twelve	Julie & Julia
The Devil Wears Prada	The Revenant	Mr. & Mrs. Smith	The Devil Wears Prada
Lions for Lambs	The Aviator	Spy Game	Lions for Lambs
Out of Africa	Django Unchained	Ocean's Eleven	Out of Africa
Hope Springs	Blood Diamond	Fury	Hope Springs
One True Thing	The Wolf of Wall Street	Seven Years in Tibet	One True Thing
The Hours	Gangs of New York	Fight Club	The Hours
The Iron Lady	The Departed	Sinbad: Legend of the Seven Seas	The Iron Lady
A Prairie Home Companion	Shutter Island	Interview with the Vampire	A Prairie Home Companion

number of voted users over decades using a bar chart

decade	Sum of num_voted_users
1930	116387
1940	895199
1950	72324
1960	1097601
1970	2606900
1980	10196071
1990	22100527
2000	78547399
2010	178281678
2020	99064581

