IMDB Movie Analysis

Project Description:

This is detailed overview of complete analysis of imdb_movies. The dataset which used is imdb_movie dataset. I was required to create a problem which I want to shed some light on.

Tech Used:

- **Microsoft excel** : excel is used for complete analysis done in this project.
- **google drive**: drive is used for storing this file in it.
- **canva**: canva is used to make this report.

Approach:

In this project I followed 5 Whys technique to determine its root cause by repeatedly asking the question "Why". It's also called root cause analysis. which is developed by sakichi toyoda founder of toyata industires. This approach is basically asking why in every step.

Insights and reports:

I was required to provide a detailed report for the below data record 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

Insights:

- In this step I removed all the blank cells.
- there are lots of dropping columns present in this dataset which has no use at all, so I removed all the dropping columns.
- I even removed the duplicate value and made the data set look more appealing.

Before cleaning:

After Cleaning:

director_name	num_critic_for_revi	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_review	lang
James Cameron	723	760505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	886204	3054	Engl
Gore Verbinski	302	309404152	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	1238	Engl
Sam Mendes	602	200074175	Action|Adventure|Thriller	Christoph Waltz	Spectre	275868	994	Engl
Christopher Nolan	813	448130642	Action|Thriller	Tom Hardy	The Dark Knight Rises	1144337	2701	Engl
Andrew Stanton	462	73058679	Action|Adventure|Sci-Fi	Daryl Sabara	John Carter	212204	738	Engl
Sam Raimi	392	336530303	Action|Adventure|Romance	J.K. Simmons	Spider-Man 3	383056	1902	Engl
Nathan Greno	324	200807262	Adventure|Animation|Comedy|Family|Fantasy|Musical|Romance	Brad Garrett	Tangled	294810	387	Engl
Joss Whedon	635	458991599	Action|Adventure|Sci-Fi	Chris Hemsworth	Avengers: Age of Ultron	462669	1117	Engl
David Yates	375	301956980	Adventure|Family|Fantasy|Mystery	Alan Rickman	Harry Potter and the Half-Blood Prince	321795	973	Engl
Zack Snyder	673	330249062	Action|Adventure|Sci-Fi	Henry Cavill	Batman v Superman: Dawn of Justice	371639	3018	Engl
Bryan Singer	434	200069408	Action|Adventure|Sci-Fi	Kevin Spacey	Superman Returns	240396	2367	Engl
Marc Forster	403	168368427	Action|Adventure	Giancarlo Giannini	Quantum of Solace	330784	1243	Engl
Gore Verbinski	313	423032628	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: Dead Man's Chest	522040	1832	Engl
Gore Verbinski	450	89289910	Action|Adventure|Western	Johnny Depp	The Lone Ranger	181792	711	Engl
Zack Snyder	733	291021565	Action|Adventure|Fantasy|Sci-Fi	Henry Cavill	Man of Steel	548573	2536	Engl
Andrew Adamson	258	141614023	Action|Adventure|Family|Fantasy	Peter Dinklage	The Chronicles of Narnia: Prince Caspian	149922	438	Engl
Joss Whedon	703	623279547	Action|Adventure|Sci-Fi	Chris Hemsworth	The Avengers	995415	1722	Engl
Rob Marshall	448	241063875	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: On Stranger Tides	370704	484	Engl
Barry Sonnenfeld	451	179020854	Action|Adventure|Comedy|Family|Fantasy|Sci-Fi	Will Smith	Men in Black 3	268154	341	Engl
Peter Jackson	422	255108370	Adventure|Fantasy	Aidan Turner	The Hobbit: The Battle of the Five Armies	354228	802	Engl
Marc Webb	599	262030663	Action|Adventure|Fantasy	Emma Stone	The Amazing Spider-Man	451803	1225	Engl
Ridley Scott	343	105219735	Action|Adventure|Drama|History	Mark Addy	Robin Hood	211765	546	Engl
Peter Jackson	509	258355354	Adventure|Fantasy	Aidan Turner	The Hobbit: The Desolation of Smaug	483540	951	Engl
Chris Weitz	251	70083519	Adventure|Family|Fantasy	Christopher Lee	The Golden Compass	149019	666	Engl
Peter Jackson	446	218051260	Action|Adventure|Drama|Romance	Naomi Watts	King Kong	316018	2618	Engl
James Cameron	315	658672302	Drama|Romance	Leonardo DiCaprio	Titanic	793059	2528	Engl
Anthony Russo	516	407197282	Action|Adventure|Sci-Fi	Robert Downey Jr.	Captain America: Civil War	272670	1022	Engl
Peter Berg	377	65173160	Action|Adventure|Sci-Fi|Thriller	Liam Neeson	Battleship	202382	751	Engl
Colin Trevorrow	644	652177271	Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic World	418214	1290	Engl
Sam Mendes	750	304360277	Action|Adventure|Thriller	Albert Finney	Skyfall	522030	1498	Engl

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?

Insights:

- So basically profit is all about gross - budget. In this step I simply reduce the gross from budget and kept them in separate column.
- on the analysis of this movies, we came to know about the top highest grossed movies ever.
- we can able to see the difference between the budget and profit for these very clearly with the help of charts.
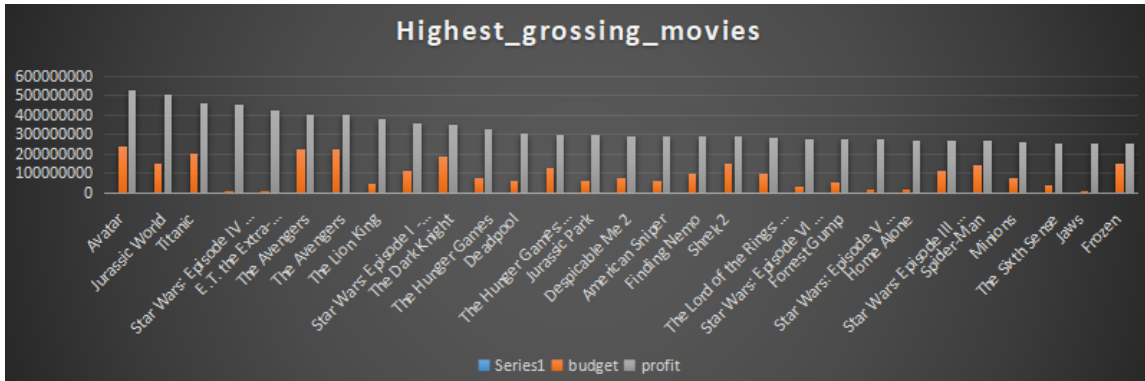
c. Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the 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

Insights:

- In this task we came to know about the imdb top 250 movies. about that movies we only group movies only in English language and rest of them in other column.
- These top movies have filtered based on the num_voted_users. the movies which has num_voted_users is more than 25000 had made into this list.

Top_imdb_movies:

movie_title	imdb_score	num_voted_users	rank
The Shawshank Redemption	9.3	886204	1
The Godfather	9.2	471220	2
The Dark Knight	9	275868	3
The Godfather: Part II	9	1144337	4
The Lord of the Rings: The Return of the King	8.9	212204	5
Schindler's List	8.9	383056	6
Pulp Fiction	8.9	294810	7
The Good, the Bad and the Ugly	8.9	462669	8
Inception	8.8	321795	9
The Lord of the Rings: The Fellowship of the Ring	8.8	371639	10
Fight Club	8.8	240396	11
Forrest Gump	8.8	330784	12
Star Wars: Episode V - The Empire Strikes Back	8.8	522040	13
The Lord of the Rings: The Two Towers	8.7	181792	14
The Matrix	8.7	548573	15
Goodfellas	8.7	149922	16
Star Wars: Episode IV - A New Hope	8.7	995415	17
One Flew Over the Cuckoo's Nest	8.7	370704	18
City of God	8.7	268154	19
Seven Samurai	8.7	354228	20
Interstellar	8.6	451803	21
Saving Private Ryan	8.6	211765	22
Se7en	8.6	483540	23
The Silence of the Lambs	8.6	149019	24
Spirited Away	8.6	316018	25
American History X	8.6	793059	26
The Usual Suspects	8.6	272670	27
Modern Times	8.6	202382	28
The Dark Knight Rises	8.5	418214	29
Gladiator	8.5	522030	30

Top_imdb_movies in foreign languages:

movie_title	imdb_score	num_voted_users	language	rank
The Good, the Bad and the Ugly	8.9	503509	Italian	8
City of God	8.7	533200	Portuguese	19
Seven Samurai	8.7	229012	Japanese	20
Spirited Away	8.6	417971	Japanese	25
The Lives of Others	8.5	259379	German	45
Children of Heaven	8.5	27882	Persian	46
Amélie	8.4	534262	French	49
Princess Mononoke	8.4	221552	Japanese	52
Das Boot	8.4	168203	German	56
Oldboy	8.4	356181	Korean	58
A Separation	8.4	151812	Persian	60
Metropolis	8.3	111841	German	74
Downfall	8.3	248354	German	76
The Hunt	8.3	170155	Danish	83
Howl's Moving Castle	8.2	214091	Japanese	99
Pan's Labyrinth	8.2	467234	Spanish	101
Incendies	8.2	80429	French	103
The Secret in Their Eyes	8.2	131831	Spanish	105
The Act of Killing	8.2	23836	Indonesian	109
The Sea Inside	8.1	64556	Spanish	141
Tae Guk Gi: The Brotherhood of War	8.1	31943	Korean	142
Akira	8.1	106160	Japanese	144
Elite Squad	8.1	81644	Portuguese	145
Amores Perros	8.1	173551	Spanish	154
The Celebration	8.1	65951	Danish	156
My Name Is Khan	8	69759	Hindi	202

Persepolis	8	70194 French	204
Central Station	8	28951 Portuguese	210
Waltz with Bashir	8	46107 Hebrew	213

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 top 10 director.

Your task: Find the best directors

director_name ▾	imdb_score T.
Akira Kurosawa	8.7
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Sergio Leone	8.4
Christopher Nola	8.4
Richard Marquar	8.4
Asghar Farhadi	8.4
Marius A. Marke	8.4

top_10_director
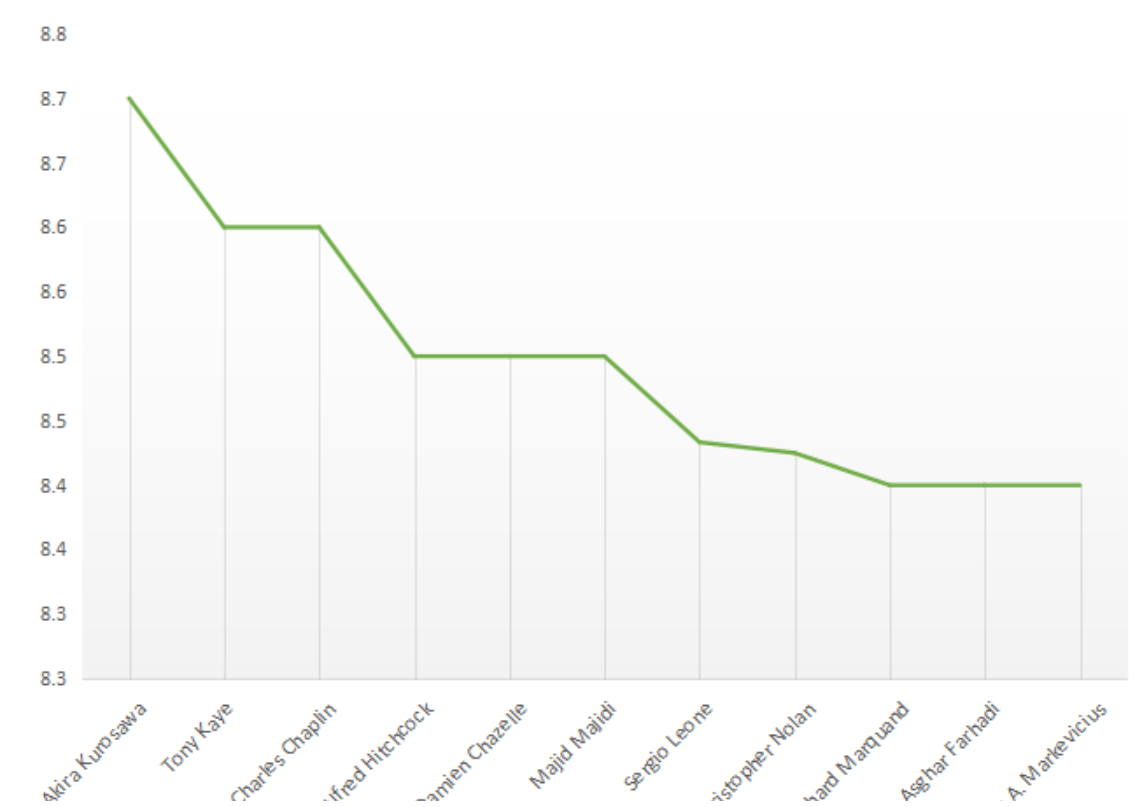
E:Popular Genres: Perform this step using the knowledge gained while performing previous steps.

Your task: Find popular genres

E	F
genres	count
Comedy|Drama|Romance	150
Drama	149
Comedy	145
Comedy|Drama	143
Comedy|Romance	135
Drama|Romance	119
Crime|Drama|Thriller	82
Action|Crime|Thriller	56
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	48

popular_genre
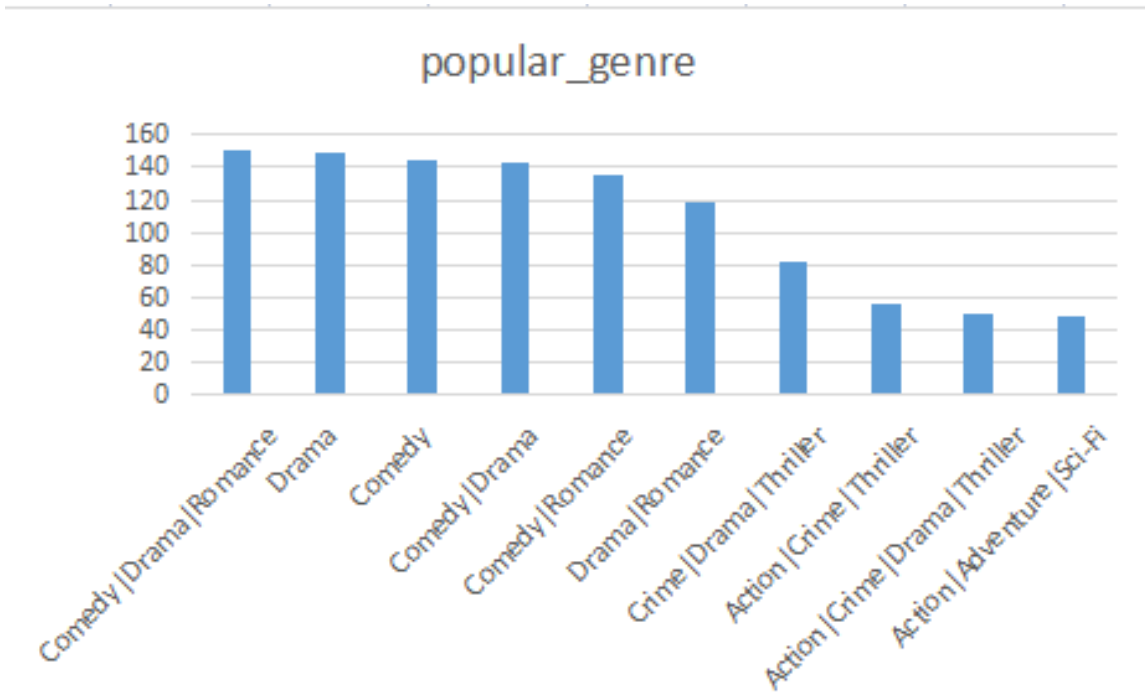
F: Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you

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.

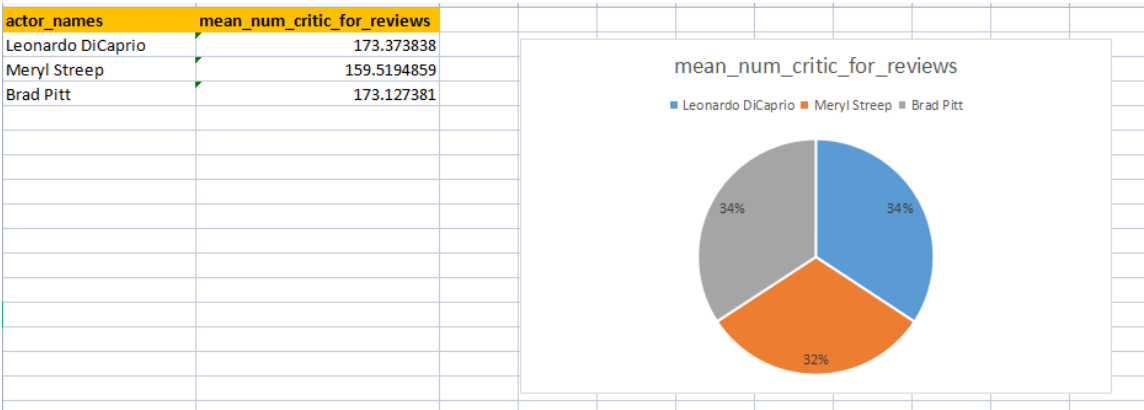Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade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

I	J	K
Leonardo DiCaprio	Meryl Streep	Brad Pitt
Titanic	It's Complicated	The Curious Case of Benjamin Button
The Great Gatsby	The River Wild	Troy
Inception	Julie & Julia	Ocean's Twelve
The Revenant	The Devil Wears Prada	Mr. & Mrs. Smith
The Aviator	Lions for Lambs	Spy Game
Django Unchained	Out of Africa	Ocean's Eleven
Blood Diamond	Hope Springs	Fury
The Wolf of Wall Street	One True Thing	Seven Years in Tibet
Gangs of New York	The Hours	Fight Club
The Departed	The Iron Lady	Sinbad: Legend of the Seven Seas
Shutter Island	A Prairie Home Companion	Interview with the Vampire: The Vampire Chronicles
Body of Lies		The Tree of Life
Catch Me If You Can		The Assassination of Jesse James by the Coward Robert Fo
The Beach		Babel
Revolutionary Road		By the Sea
The Man in the Iron Mask		Killing Them Softly
J. Edgar		True Romance
The Quick and the Dead		
Marvin's Room		
Romeo + Juliet		
The Great Gatsby		

critic favorite actors:

actor_names	mean_num_critic_for_reviews
Leonardo DiCaprio	173.373838
Meryl Streep	159.5194859
Brad Pitt	173.127381

mean_num_critic_for_reviews

■ Leonardo DiCaprio ■ Meryl Streep ■ Brad Pitt

34%   34%
     32%

Audience favorite actors:

actor_names	mean_num_user_for_reviews
Leonardo DiCaprio	354.5391766
Meryl Streep	310.3573798

mean_num_user_for_reviews

Leonardo DiCaprio ■ Meryl Streep ■ Brad Pitt

35% 35%

30%

Num_voted _user by decade:

insights:

- In this graph we can clearly able to see the count of vote by users.
- In over period of time we can able to see the step rise.

years	df_by_decade
1920's	116387
1930's	804839
1940's	159517
1950's	678336
1960's	2982551
1970's	8681156
1980's	20091781
1990's	69934957
2000's	172796030
2010's	100850761

df_by_decade

0

1920's 1930's 1940's 1950's 1960's 1970's 1980's 1990's 2000's 2010's

1920's 1930's 1940's 1950's 1960's 1970's 1980's 1990's 2000's 2010's