IMDB Movie Analysis

(Final Project-1)

Project Description:

The dataset provided by the company contains various columns of different IMDB Movies. We are required to Frame the problem. For this task, we will need to define a problem we want to shed some light on.

We can do this by asking 'What?'. This is where we frame the problem i.e. What is the problem?

We can do this by asking the following 'What?'

• What do we see happening?

• What is our hypothesis for the cause of the problem? (This will be broadly based on intuition initially)

• What is the impact of the problem on stakeholders?

• What is the impact of the problem not being solved? How to handle the things:

• Clean the data.

• Use the Data Analysis skills to explore the data set.

• Derive insights.

The things that we are going to find out through the project are movies with the highest profit, top movies as per IMDB rating, top directors, most popular genres, top foreign language films and more

Approach:

A. Cleaning the data:

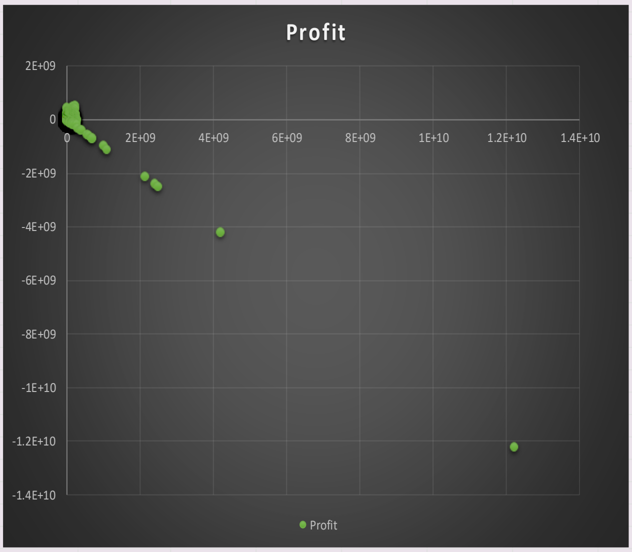
This is the most important step to perform for the better analysis of the data.

- Dropped the unwanted columns as there is no use for the analysis.
- Dropped the rows which are having null/blank.
- Removed the duplicate row values.

B. Movies with highest Profit:

'Avatar' movie is the highest profit in the list of data followed by 'Jurassic World' and 'Titanic'.

movie_title	gross	budget	Profit
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 AvengersÂ	623279547	220000000	403279547
The AvengersÂ	623279547	220000000	403279547
The Lion KingÂ	422783777	45000000	377783777
Star Wars: Episode I - The Phantom Men	474544677	115000000	359544677
The Dark KnightÂ	533316061	185000000	348316061
The Hunger GamesÂ	407999255	78000000	329999255
DeadpoolÂ	363024263	58000000	305024263
The Hunger Games: Catching FireÂ	424645577	130000000	294645577
Jurassic ParkÂ	356784000	63000000	293784000
Despicable Me 2Â	368049635	76000000	292049635
American SniperÂ	350123553	58800000	291323553
Finding NemoÂ	380838870	94000000	286838870
Shrek 2Â	436471036	150000000	286471036
The Lord of the Rings: The Return of the	377019252	94000000	283019252
Star Wars: Episode VI - Return of the Jec	309125409	32500000	276625409
Forrest GumpÂ	329691196	55000000	274691196
Star Wars: Episode V - The Empire Strike	290158751	18000000	272158751
Home AloneÂ	285761243	18000000	267761243
Star Wars: Episode III - Revenge of the S	380262555	113000000	267262555
Spider-ManÂ	403706375	139000000	264706375
MinionsÂ	336029560	74000000	262029560



C. Top 250:

Filtered the 'num_voted_users' column greater than 25,000.

• Created a new column named 'IMDb_Top_250' and stored the top 250 movies with the highest IMDb Rating (sorted the 'imdb_score' column from the largest to the smallest).

• Added a 'Rank' containing the values 1 to 250 using the RANK() function + COUNTIFS() function.

IMDb_Top_250	num_voted_users	language	imdb_score	Rank
The Shawshank RedemptionÂ	1689764	English	9.3	1
The GodfatherÂ	1155770	English	9.2	2
The Dark KnightÂ	1676169	English	9	3
The Godfather: Part IIÂ	790926	English	9	4
Pulp FictionÂ	1324680	English	8.9	5
Schindler's ListÂ	865020	English	8.9	6
The Good, the Bad and the UglyÂ	503509	Italian	8.9	7
The Lord of the Rings: The Return of the Kin	1215718	English	8.9	8
Fight ClubÂ	1347461	English	8.8	9
Forrest GumpÂ	1251222	English	8.8	10
InceptionÂ	1468200	English	8.8	11
Star Wars: Episode V - The Empire Strikes Ba	837759	English	8.8	12
The Lord of the Rings: The Fellowship of the	1238746	English	8.8	13
City of GodÂ	533200	Portuguese	8.7	14
GoodfellasÂ	728685	English	8.7	15
One Flew Over the Cuckoo's NestÂ	680041	English	8.7	16
Seven SamuraiÂ	229012	Japanese	8.7	17
Star Wars: Episode IV - A New HopeÂ	911097	English	8.7	18
The Lord of the Rings: The Two TowersÂ	1100446	English	8.7	19
The MatrixÂ	1217752	English	8.7	20
American History XÂ	782437	English	8.6	21
InterstellarÂ	928227	English	8.6	22
Modern TimesÂ	143086	English	8.6	23
Saving Private RyanÂ	881236	English	8.6	24
Se7enÂ	1023511	English	8.6	25
Spirited AwayÂ	417971	Japanese	8.6	26

• Extracted all the movies in the IMDb_Top_250 column by filtering the ‘language’ column (unselecting English language) and stored them in a new column named ‘Top_Foreign_Lang_Film’.

Top_Foreign_Lang_Film	num_voted_users	imdb_score	language
The Good, the Bad and the UglyÂ	503509	8.9	Italian
City of GodÂ	533200	8.7	Portuguese
Seven SamuraiÂ	229012	8.7	Japanese
Spirited AwayÂ	417971	8.6	Japanese
Children of HeavenÂ	27882	8.5	Persian
The Lives of OthersÂ	259379	8.5	German
A SeparationÂ	151812	8.4	Persian
AmÃ©lieÂ	534262	8.4	French
Baahubali: The BeginningÂ	62756	8.4	Telugu
Das BootÂ	168203	8.4	German
OldboyÂ	356181	8.4	Korean
Princess MononokeÂ	221552	8.4	Japanese
MetropolisÂ	111841	8.3	German
The HuntÂ	170155	8.3	Danish
UnforgivenÂ	248354	8.3	German
Pan's LabyrinthÂ	80429	8.2	French
The Bridge on the River KwaiÂ	131831	8.2	Spanish
The ThingÂ	467234	8.2	Spanish
WarriorÂ	214091	8.2	Japanese
Annie HallÂ	81644	8.1	Portuguese
In the Shadow of the MoonÂ	65951	8.1	Danish
Sling BladeÂ	106160	8.1	Japanese
Tae Guk Gi: The Brotherhood of WarÂ	64556	8.1	Spanish
The Best Years of Our LivesÂ	173551	8.1	Spanish
The Imitation GameÂ	31943	8.1	Korean
Bowling for ColumbineÂ	28951	8	Portuguese
JawsÂ	70194	8	French

D. Best Directors:

• Selected the cleaned dataset done and created a pivot table.

• Put the ‘director_name’ into the Rows and took average of ‘imdb_score’ in the Values section.

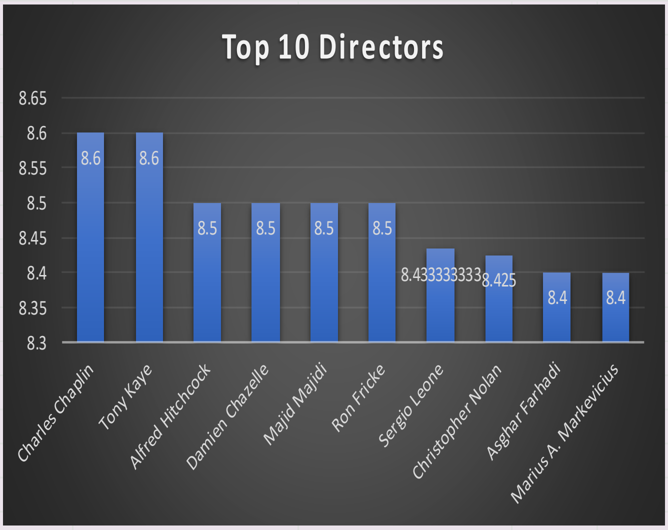
• Sorted the ‘director_name’ in ascending order and then sorted the ‘average of imdb_score’ (largest to smallest).

• Then selected the top 10 directors and their mean of imdb_score in other columns.

• Finally made a Stacked Column chart of the top 10 directors for the better insights.

Row Labels	Average of imdb_score
Tony Kaye	8.6
Charles Chaplin	8.6
Ron Fricke	8.5
Majid Majidi	8.5
Damien Chazelle	8.5
Alfred Hitchcock	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
S.S. Rajamouli	8.4
Richard Marquand	8.4
Marius A. Markevicius	8.4
Asghar Farhadi	8.4
Lenny Abrahamson	8.3
Lee Unkrich	8.3
Fritz Lang	8.3
Billy Wilder	8.3
Pete Docter	8.233333333
Hayao Miyazaki	8.225
Quentin Tarantino	8.2
Joshua Oppenheimer	8.2
Juan JosÃ© Campanella	8.2
Elia Kazan	8.2
George Roy Hill	8.2
Victor Fleming	8.15
Milos Forman	8.133333333
William Wyler	8.1
Tim Miller	8.1

Top 10 Directors	Mean of imdb_score
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
Marius A. Markevicius	8.4

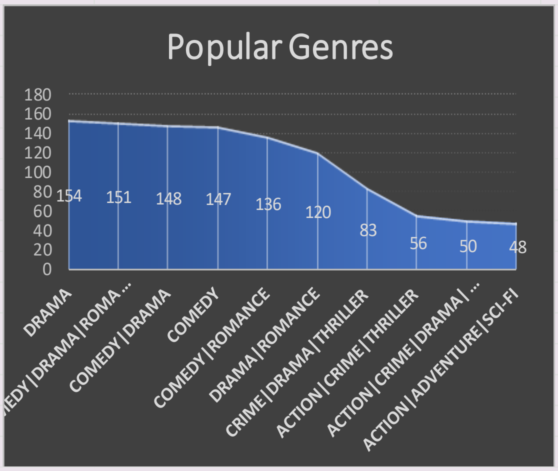


E. Popular Genres:

- Selected the ‘genres’ column from the cleaned dataset and created a pivot table.
- In Pivot, kept the ‘genres’ into the Rows and took count of ‘genres’ in the Values section.
- After Sorted the ‘Count of genres’ in descending order.
- Copied the top 10 genres and their count and pasted it in the other columns.
- Made a Stacked area chart of the top 10 genres for the better insights.

Row Labels	Count of genres
Drama	154
Comedy|Drama|Romance	151
Comedy|Drama	148
Comedy	147
Comedy|Romance	136
Drama|Romance	120
Crime|Drama|Thriller	83
Action|Crime|Thriller	56
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	48
Comedy|Crime	47
Action|Adventure|Thriller	45
Horror	44
Drama|Thriller	42
Crime|Drama	42
Crime|Drama|Mystery|Thriller	42
Horror|Thriller	36
Action|Adventure|Sci-Fi|Thriller	34
Horror|Mystery|Thriller	33
Drama|Mystery|Thriller	30
Biography|Drama	30
Adventure|Animation|Comedy|Family|Fantasy	28
Action|Comedy|Crime	27
Horror|Mystery	25
Action|Adventure|Fantasy	25
Documentary	24

Popular Genres	Count of Genres
Drama	154
Comedy|Drama|Romance	151
Comedy|Drama	148
Comedy	147
Comedy|Romance	136
Drama|Romance	120
Crime|Drama|Thriller	83
Action|Crime|Thriller	56
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	48



F. Charts:

(a) Critic favourite and audience favourite actors

• Created 3 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 using the 'actor_1_name' column.

• Append the rows of all these columns and stored them in a new column named 'Combined'.

• We grouped the column by the actor's name: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt'.

• Then selected the cleaned dataset done in and created a pivot table.

• In Pivot, inserted the 'actor_1_name' into the Rows and took mean/average of 'num_users_for_review' in the Values section.

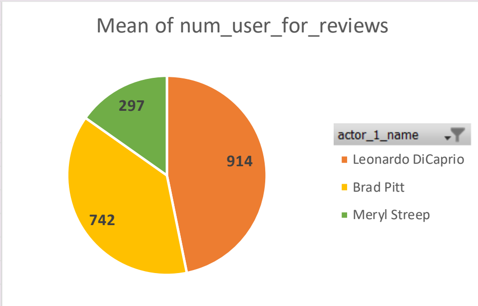
• Sorted the column from largest to smallest by mean of 'num_users_for_review'.

• Made a pivot chart (bar chart) of the mean of 'num_users_for_review'.

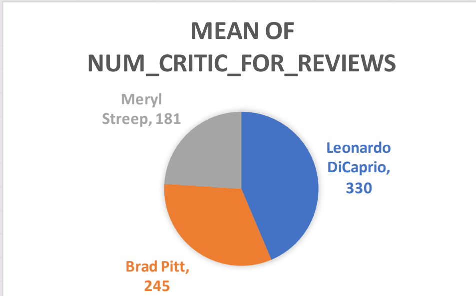
• Same procedure done for the mean of 'num_critic_for_review'.

Meryl_Streep	Leo_Caprio	Brad_Pitt	Combined	Group By
LucyÂ	Django UnchainedÂ	DivergentÂ	LucyÂ	Meryl Streep
MaleficentÂ	InceptionÂ	Shanghai KnightsÂ	MaleficentÂ	Meryl Streep
TranscendenceÂ	How to Train Your DragonÂ	The CircleÂ	TranscendenceÂ	Meryl Streep
Down and Out with the DollsÂ	Pirates of the Caribbean: The Curse of the Black PearlÂ	42nd StreetÂ	Down and Out with the DollsÂ	Meryl Streep
Machine Gun PreacherÂ	The SessionsÂ	Fight ClubÂ	Machine Gun PreacherÂ	Meryl Streep
Like CrazyÂ	Frozen RiverÂ	Scott Pilgrim vs. the WorldÂ	Like CrazyÂ	Meryl Streep
Brown SugarÂ	The MartianÂ	MulanÂ	Brown SugarÂ	Meryl Streep
Bridge of SpiesÂ	Lethal Weapon 4Â	Leap YearÂ	Bridge of SpiesÂ	Meryl Streep
True LiesÂ	The DepartedÂ	The Kids Are All RightÂ	True LiesÂ	Meryl Streep
Hellboy II: The Golden ArmyÂ	The Hunger GamesÂ	Connie and CarlaÂ	Hellboy II: The Golden ArmyÂ	Meryl Streep
GerryÂ	Animal HouseÂ	Star Trek Into DarknessÂ	GerryÂ	Meryl Streep
	The Squid and the WhaleÂ	Sunshine StateÂ	Django UnchainedÂ	Leonardo DiCaprio
	Dream with the FishesÂ	ZookeeperÂ	InceptionÂ	Leonardo DiCaprio
	TransamericaÂ	Black SwanÂ	How to Train Your DragonÂ	Leonardo DiCaprio
	The Iron GiantÂ	Kung Fu PandaÂ	Pirates of the Caribbean: The Curse of the Black PearlÂ	Leonardo DiCaprio
	Casino RoyaleÂ	The Puffy ChairÂ	The SessionsÂ	Leonardo DiCaprio
	Hurricane StreetsÂ	It FollowsÂ	Frozen RiverÂ	Leonardo DiCaprio
	The Devil Wears PradaÂ		The MartianÂ	Leonardo DiCaprio
	Ocean's TwelveÂ		Lethal Weapon 4Â	Leonardo DiCaprio
	The Longest YardÂ		The DepartedÂ	Leonardo DiCaprio
	The Family ManÂ		The Hunger GamesÂ	Leonardo DiCaprio
			Animal HouseÂ	Leonardo DiCaprio
			The Squid and the WhaleÂ	Leonardo DiCaprio

Row Labels	Mean of num_user_for_reviews	Sum of num_user_for_reviews2	Count of num_user_for_reviews3
Leonardo DiCaprio	914	19204	21
Brad Pitt	742	12620	17
Meryl Streep	297	3269	11
Grand Total	**716**	**35093**	**49**



Row Labels	Mean of num_critic_for_reviews	Sum of num_critic_for_reviews2	Count of num_critic_for_reviews3
Leonardo DiCaprio	330	6934	21
Brad Pitt	245	4165	17
Meryl Streep	181	1996	11
Grand Total	**267**	**13095**	**49**



(b) Change in number of voted users over decades

• Selected the cleaned dataset and created a pivot table.

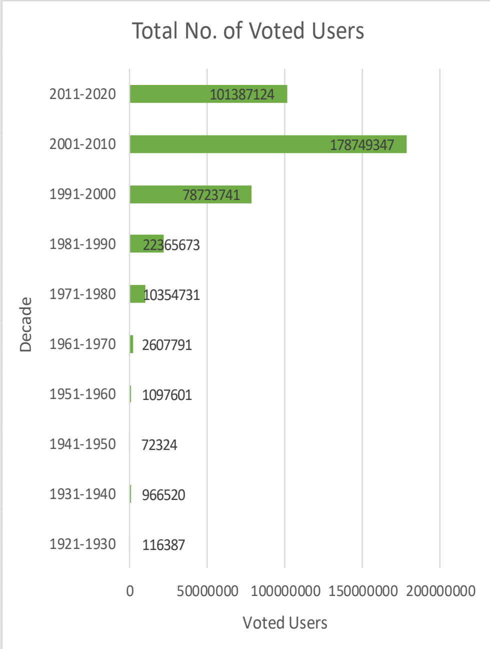
• Insert the ‘title_year’ into the Rows and took the sum of ‘num_voted_users’ in the Values section.

• Then grouped the title_year by decade and stored in df_by_decade column.

• Finally plotted the total no. of voted users against the decade in a bar chart.

Row Labels	Sum of num_voted_users
1927	111841
1929	4546
1933	7921
1935	13269
1936	143086
1937	133348
1939	507215
1940	161681
1946	46663
1947	19236
1948	3258
1950	3167
1952	9456
1953	11171
1954	329902
1957	149444
1959	175196
1960	422432
1961	71919
1962	309417
1963	140280
1964	493685
1965	297839
1966	503509
1967	75280

Title_Year	df_by_decade	Total no. of 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	22365673
1990s	1991-2000	78723741
2000s	2001-2010	178749347
2010s	2011-2020	101387124
	Grand Total	**396441239**



Tech-Stack Used:

Insights:

- There are as many as 5 outliers in the profit columns.
- The movie with the highest profit is ‘Avatar’
- The Shawshank Redemption is the top-most movie with the highest IMDB rating.
- The Good, the Bad and the Ugly (Italian) is the top-most foreign language movie.
- Charles Chaplin is the top-most director followed by Tony Kaye.
- The most popular genres is Drama followed by Comedy.
- 'Leonardo DiCaprio' is the critic-favourite as well as the audience-favourite actor.
- The most users voted in the decade 2000s and the least in the decade 1940s.

Results:

In this Project, I learned especially the “5 Why Analysis” and it helps me to think deeper to analyse and generate valuable insights. I also learned the basic and advanced excel concepts which includes statistics and functions like rank, count etc.,

Finally, It was great learning experience in entire project and very challenging too while executing the solutions.