IMDB Movie Analysis

By Darshan Hedgire

Trainity

IMDB Movie Analysis

Description:

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

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

We can do this by asking the following what questions

- What do you see happening?
- What is your 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 we are going to handle the things

- Cleaning the data
- Data Analysis skills to explore the data set
- Deriving 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:

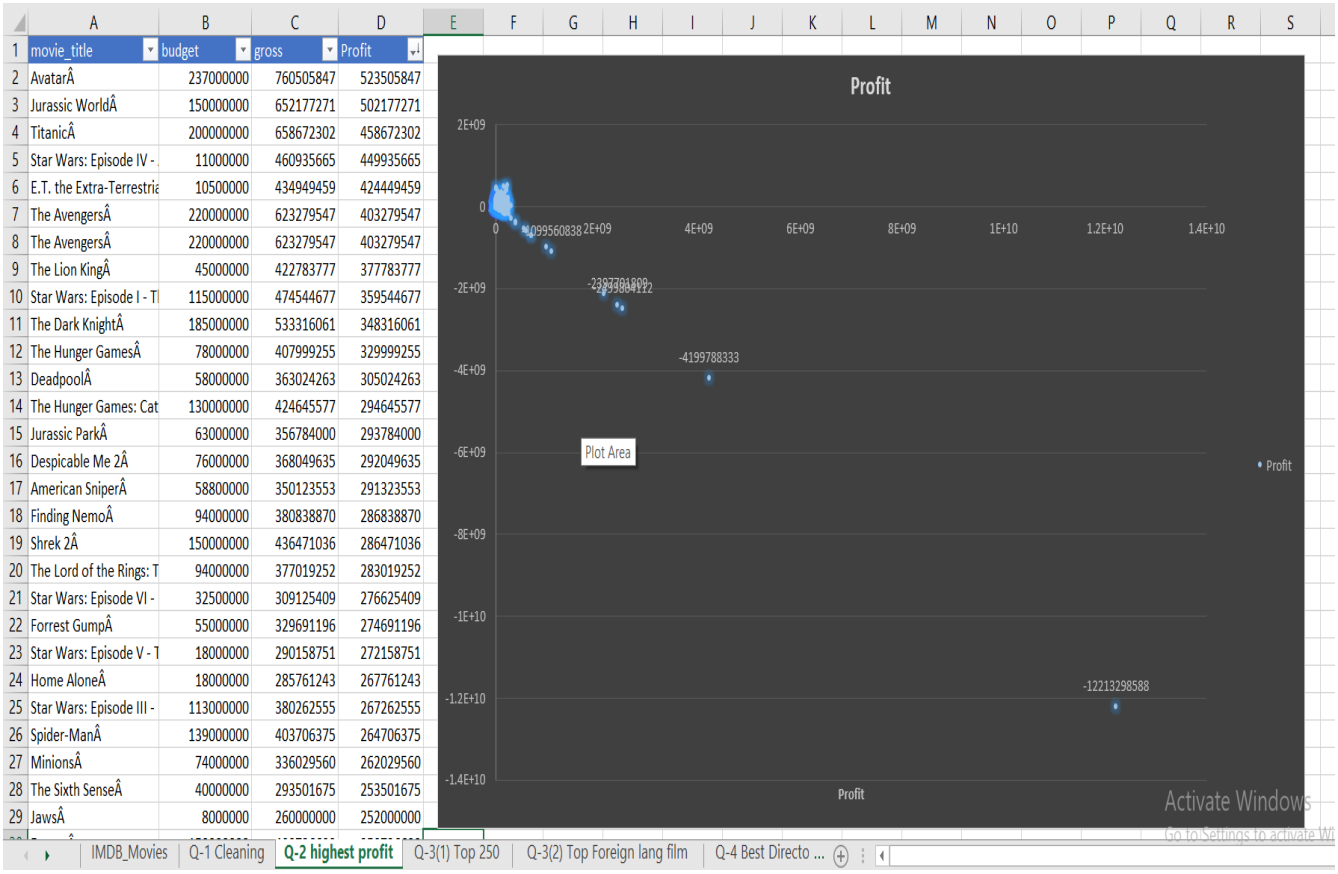
1. Task: Clean the data

   This is one of the most important step to perform before moving forward with the analysis.

- First, we dropped the columns which have no use for the analysis.
- Second, we dropped the rows which are blank/null.
- Third, removed the duplicate row values.

2. Task: Find the movies with the highest profit?

- First, we created a new column ‘Profit’ by subtracting budget column from gross column.
- Second, we sorted the columns using the profit column as reference from the largest to the smallest.
- Third, we plotted the budget and profit in XY Scatter chart to find the outliers.
- There are as many as 5 outliers in the profit columns.
- The movie with the highest profit is ‘Avatar’ followed by ‘Jurassic World’ and ‘Titanic’ and so on.



	movie_title	budget	gross	Profit
2	AvatarÂ	237000000	760505847	523505847
3	Jurassic WorldÂ	150000000	652177271	502177271
4	TitanicÂ	200000000	658672302	458672302
5	Star Wars: Episode IV -	11000000	460935665	449935665
6	E.T. the Extra-Terrestria	10500000	434949459	424449459
7	The AvengersÂ	220000000	623279547	403279547
8	The AvengersÂ	220000000	623279547	403279547
9	The Lion KingÂ	45000000	422783777	377783777
10	Star Wars: Episode I - T	115000000	474544677	359544677
11	The Dark KnightÂ	185000000	533316061	348316061
12	The Hunger GamesÂ	78000000	407999255	329999255
13	DeadpoolÂ	58000000	363024263	305024263
14	The Hunger Games: Cat	130000000	424645577	294645577
15	Jurassic ParkÂ	63000000	356784000	293784000
16	Despicable Me 2Â	76000000	368049635	292049635
17	American SniperÂ	58800000	350123553	291323553
18	Finding NemoÂ	94000000	380838870	286838870
19	Shrek 2Â	150000000	436471036	286471036
20	The Lord of the Rings: T	94000000	377019252	283019252
21	Star Wars: Episode VI -	32500000	309125409	276625409
22	Forrest GumpÂ	55000000	329691196	274691196
23	Star Wars: Episode V - T	18000000	290158751	272158751
24	Home AloneÂ	18000000	285761243	267761243
25	Star Wars: Episode III -	113000000	380262555	267262555
26	Spider-ManÂ	139000000	403706375	264706375
27	MinionsÂ	74000000	336029560	262029560
28	The Sixth SenseÂ	40000000	293501675	253501675
29	JawsÂ	8000000	260000000	252000000

3. Task: Find IMDB Top 250

- First, we filtered the 'num_voted_users' column greater than 25,000.
- Second, we 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).
- Third, added a 'Rank' containing the values 1 to 250 using the RANK() function + COUNTIFS() function.
- Fourth, we 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'.

	A	B	C	D	E
1	Imdb_Top_250	num_voted_users	language	imdb_score	Rank
2	The Shawshank RedemptionÂ	1689764	English	9.3	1
3	The GodfatherÂ	1155770	English	9.2	2
4	The Dark KnightÂ	1676169	English	9	3
5	The Godfather: Part IIÂ	790926	English	9	4
6	Pulp FictionÂ	1324680	English	8.9	5
7	The Lord of the Rings: The Return of	1215718	English	8.9	6
8	Schindler's ListÂ	865020	English	8.9	7
9	The Good, the Bad and the UglyÂ	503509	Italian	8.9	8
10	InceptionÂ	1468200	English	8.8	9
11	Fight ClubÂ	1347461	English	8.8	10
12	Forrest GumpÂ	1251222	English	8.8	11
13	The Lord of the Rings: The Fellowshi	1238746	English	8.8	12
14	Star Wars: Episode V - The Empire St	837759	English	8.8	13
15	The MatrixÂ	1217752	English	8.7	14
16	The Lord of the Rings: The Two Tow	1100446	English	8.7	15
17	Star Wars: Episode IV - A New Hope	911097	English	8.7	16
18	GoodfellasÂ	728685	English	8.7	17
19	One Flew Over the Cuckoo's NestÂ	680041	English	8.7	18
20	City of GodÂ	533200	Portuguese	8.7	19
21	Seven SamuraiÂ	229012	Japanese	8.7	20
22	Se7enÂ	1023511	English	8.6	21
23	InterstellarÂ	928227	English	8.6	22
24	The Silence of the LambsÂ	887467	English	8.6	23
25	Saving Private RyanÂ	881236	English	8.6	24
26	American History XÂ	782437	English	8.6	25
27	The Usual SuspectsÂ	740918	English	8.6	26
28	Spirited AwayÂ	417971	Japanese	8.6	27
29	Modern TimesÂ	143086	English	8.6	28

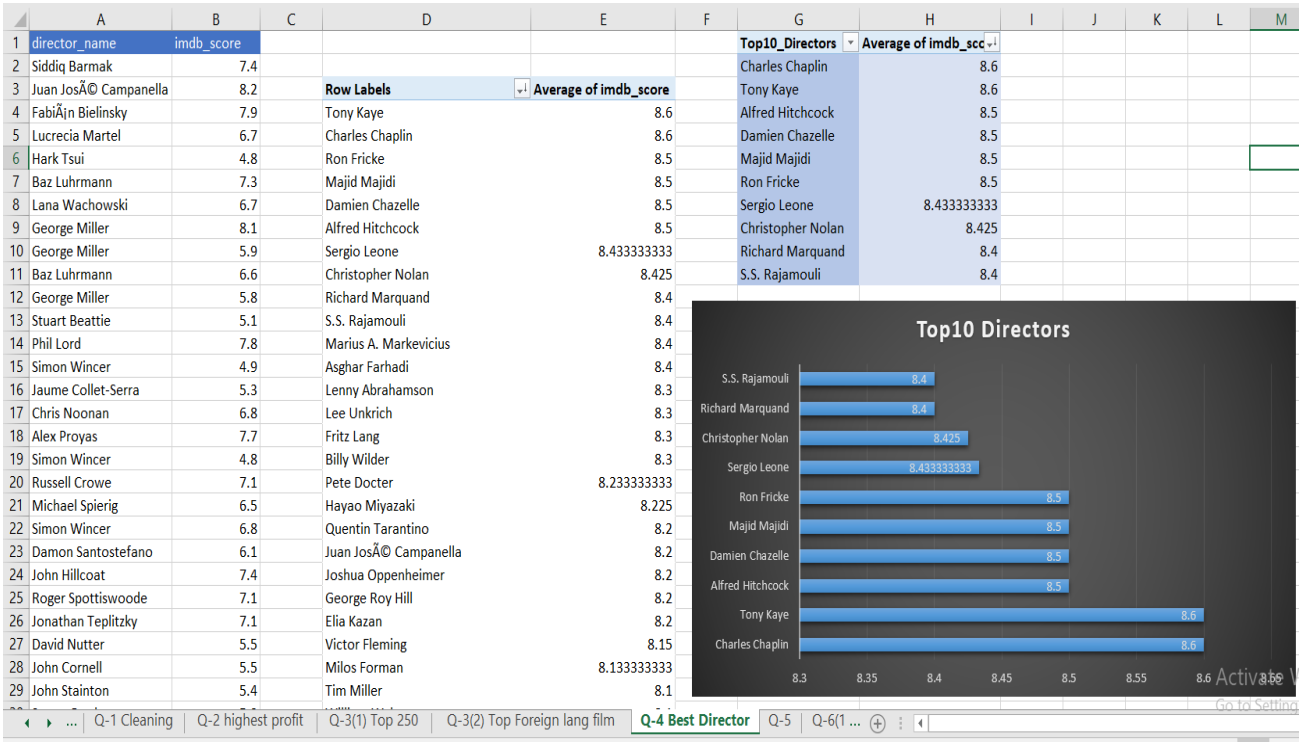
IMDB_Movies | Q-1 Cleaning | Q-2 highest profit | **Q-3(1) Top 250** | Q-3(2) To

	A	B	C	D
1	Top_Foreign_Lang_Film	num_voted_users	language	imdb_score
2	The Good, the Bad and the UglyÂ	503509	Italian	8.9
3	City of GodÂ	533200	Portuguese	8.7
4	Seven SamuraiÂ	229012	Japanese	8.7
5	Spirited AwayÂ	417971	Japanese	8.6
6	The Lives of OthersÂ	259379	German	8.5
7	Children of HeavenÂ	27882	Persian	8.5
8	AmÃ©lieÂ	534262	French	8.4
9	OldboyÂ	356181	Korean	8.4
10	Princess MononokeÂ	221552	Japanese	8.4
11	Das BootÂ	168203	German	8.4
12	A SeparationÂ	151812	Persian	8.4
13	Baahubali: The BeginningÂ	62756	Telugu	8.4
14	DownfallÂ	248354	German	8.3
15	The HuntÂ	170155	Danish	8.3
16	MetropolisÂ	111841	German	8.3
17	Pan's LabyrinthÂ	467234	Spanish	8.2
18	Howl's Moving CastleÂ	214091	Japanese	8.2
19	The Secret in Their EyesÂ	131831	Spanish	8.2
20	IncendiesÂ	80429	French	8.2
21	Amores PerrosÂ	173551	Spanish	8.1
22	AkiraÂ	106160	Japanese	8.1
23	Elite SquadÂ	81644	Portuguese	8.1
24	The CelebrationÂ	65951	Danish	8.1
25	The Sea InsideÂ	64556	Spanish	8.1
26	Tae Guk Gi: The Brotherhood of WarÂ	31943	Korean	8.1
27	A Fistful of DollarsÂ	147566	Italian	8
28	PersepolisÂ	70194	French	8
29	My Name Is KhanÂ	69759	Hindi	8

IMDB_Movies | Q-1 Cleaning | Q-2 highest profit | Q-3(1) Top 250

4. Task: Find the best directors

- First, we selected the cleaned dataset done in Task 1 and created a pivot table.
- Second, we put the ‘director_name’ into the Rows and took average of ‘imdb_score’ in the Values section.
- Third, we sorted the ‘director_name’ in ascending order and then sorted the ‘average of imdb_score’ (largest to smallest).
- Then we selected the top 10 directors and their mean of imdb_score in other columns.
- Next, we made a bar chart of the top 10 directors for the better insights.



director_name	imdb_score
Siddiq Barmak	7.4
Juan JosÃ© Campanella	8.2
FabiÃ¡n Bielinsky	7.9
Lucrecia Martel	6.7
Hark Tsui	4.8
Baz Luhrmann	7.3
Lana Wachowski	6.7
George Miller	8.1
George Miller	5.9
Baz Luhrmann	6.6
George Miller	5.8
Stuart Beattie	5.1
Phil Lord	7.8
Simon Wincer	4.9
Jaume Collet-Serra	5.3
Chris Noonan	6.8
Alex Proyas	7.7
Simon Wincer	4.8
Russell Crowe	7.1
Michael Spierig	6.5
Simon Wincer	6.8
Damon Santostefano	6.1
John Hillcoat	7.4
Roger Spottiswoode	7.1
Jonathan Teplitzky	7.1
David Nutter	5.5
John Cornell	5.5
John Stainton	5.4

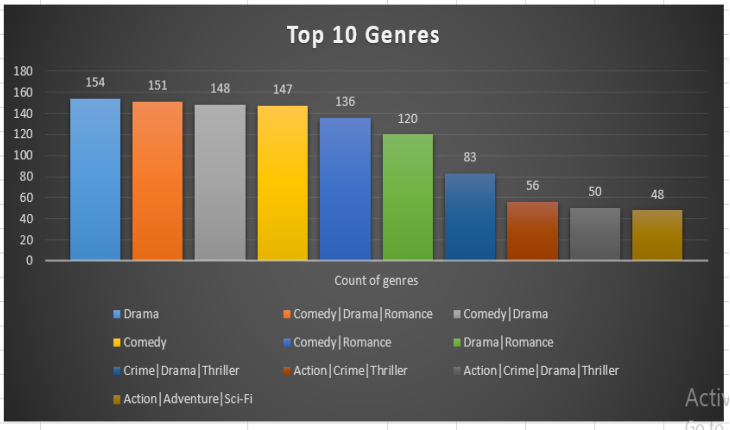
Row Labels	Average of imdb_score
Tony Kaye	8.6
Charles Chaplin	8.6
Ron Fricke	8.5
Majid Majidi	8.5
Damien Chazelle	8.5
Alfred Hitchcock	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Richard Marquand	8.4
S.S. Rajamouli	8.4
Marius A. Markevicius	8.4
Asghar Farhadi	8.4
Lenny Abrahamson	8.3
Lee Unkrich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

Top10_Directors	Average of imdb_sco
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
Richard Marquand	8.4
S.S. Rajamouli	8.4

5. Task: Find popular genres

- First, we selected the ‘genres’ column from the cleaned dataset done in Task 1 and created a pivot table.
- Second, we put the ‘genres’ into the Rows and took count of ‘genres’ in the Values section.
- Third, we sorted the ‘Count of genres’ in descending order.
- Then we copied the top 10 genres and their count and pasted it in the other columns.
- Next, we made a Column chart of the top 10 genres for the better insights.

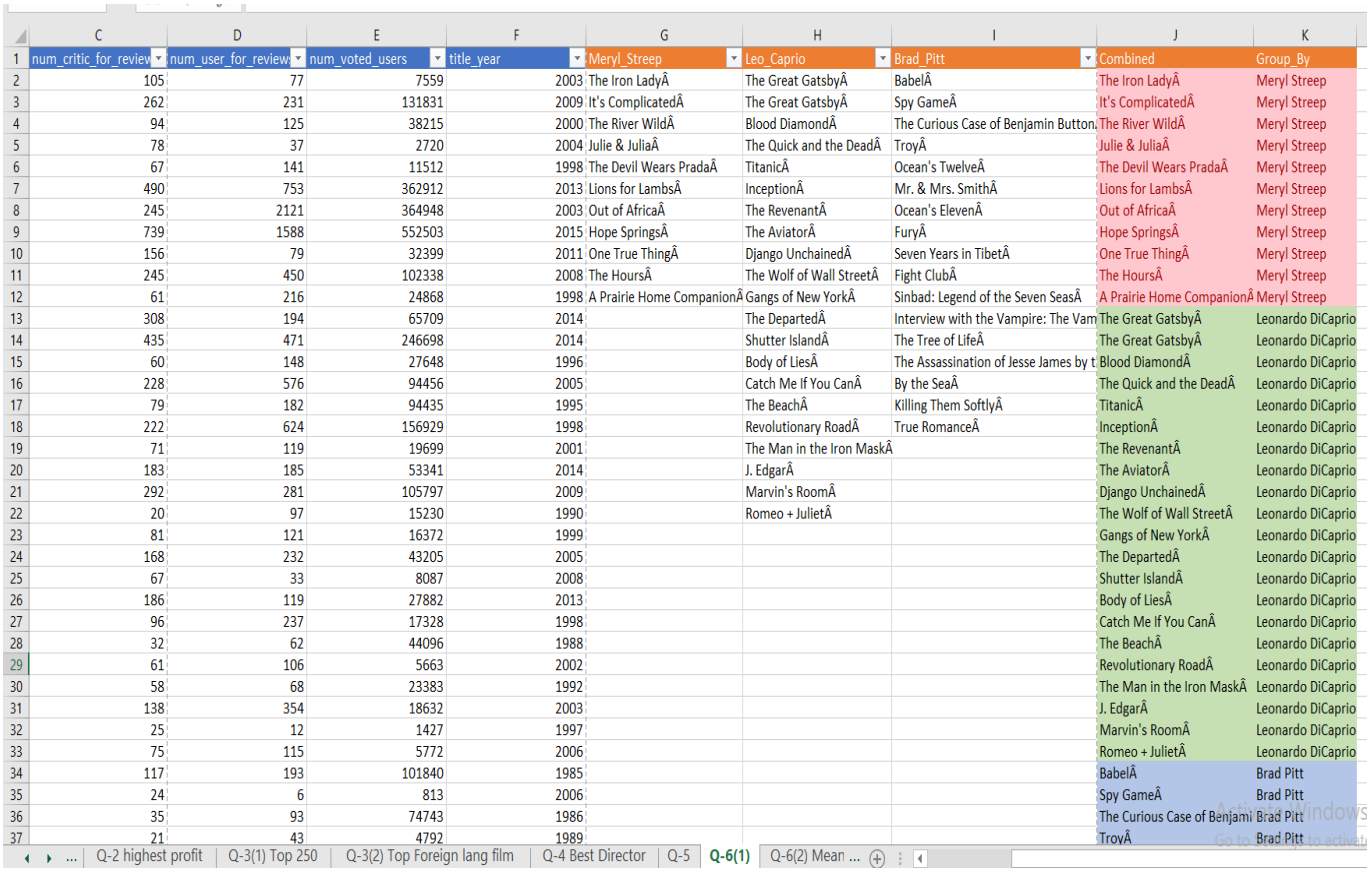
Row Labels	Count of genres		Top10_genres	Count of genres
Drama	154		Drama	154
Comedy|Drama|Romance	151		Comedy|Drama|Romance	151
Comedy|Drama	148		Comedy|Drama	148
Comedy	147		Comedy	147
Comedy|Romance	136		Comedy|Romance	136
Drama|Romance	120		Drama|Romance	120
Crime|Drama|Thriller	83		Crime|Drama|Thriller	83
Action|Crime|Thriller	56		Action|Crime|Thriller	56
Action|Crime|Drama|Thriller	50		Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	48		Action|Adventure|Sci-Fi	48
Comedy|Crime	47			
Action|Adventure|Thriller	45			
Horror	44			
Drama|Thriller	42			
Crime|Drama|Mystery|Thriller	42			
Crime|Drama	41			
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			
Biography|Drama|Sport	23			
Drama|Sport	22			
Drama|Music	22			



... | Q-1 Cleaning | Q-2 highest profit | Q-3(1) Top 250 | Q-3(2) Top Foreign lang film | Q-4 Best Director | Q-5 | Q-6(1 ...

6. (1) Task: Find the critic-favorite and audience-favorite actors

- First, we 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.
- Second, appended 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'.

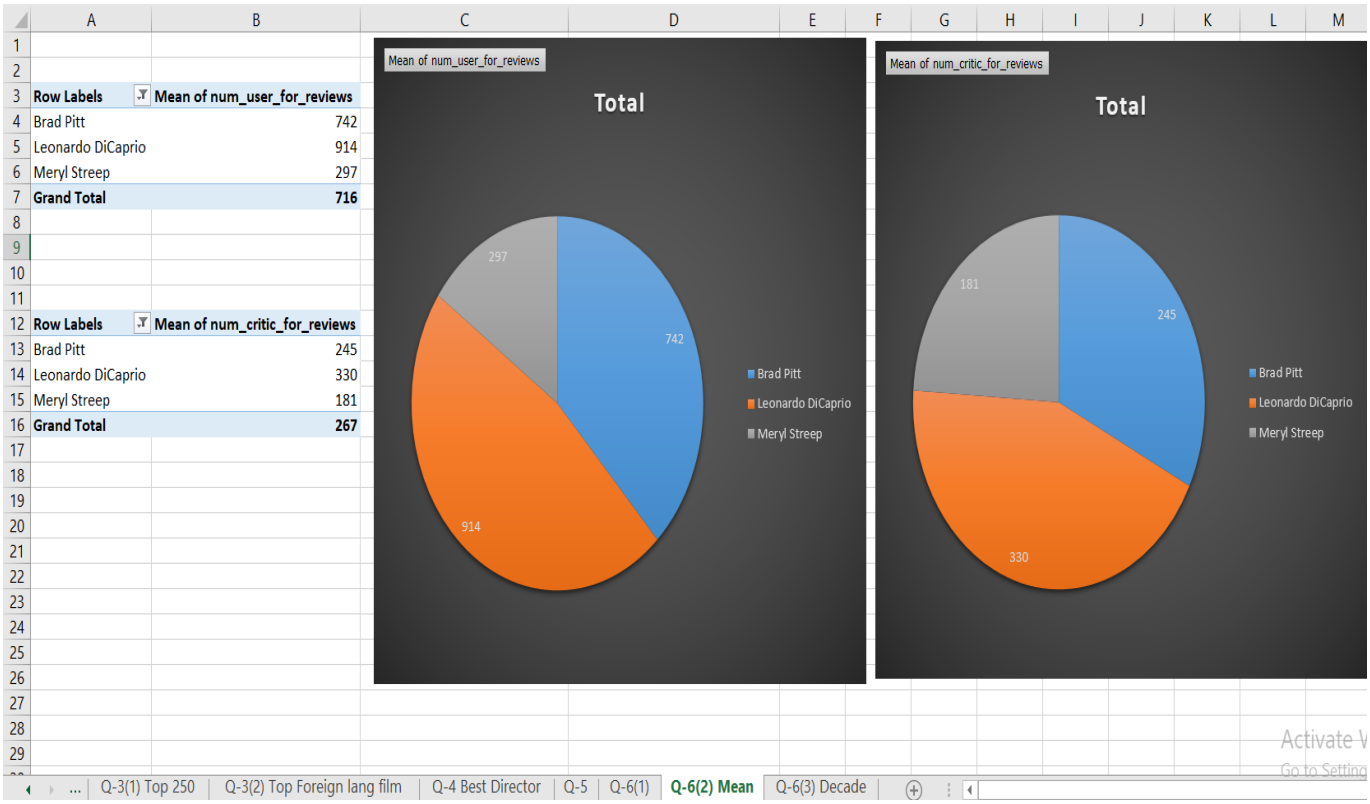


	C	D	E	F	G	H	I	J	K
1	num_critic_for_review	num_user_for_reviews	num_voted_users	title_year	Meryl_Streep	Leo_Caprio	Brad_Pitt	Combined	Group_By
2	105	77	7559	2003	The Iron LadyÂ	The Great GatsbyÂ	BabelÂ	The Iron LadyÂ	Meryl Streep
3	262	231	131831	2009	It's ComplicatedÂ	The Great GatsbyÂ	Spy GameÂ	It's ComplicatedÂ	Meryl Streep
4	94	125	38215	2000	The River WildÂ	Blood DiamondÂ	The Curious Case of Benjamin Button.	The River WildÂ	Meryl Streep
5	78	37	2720	2004	Julie & JuliaÂ	The Quick and the DeadÂ	TroyÂ	Julie & JuliaÂ	Meryl Streep
6	67	141	11512	1998	The Devil Wears PradaÂ	TitanicÂ	Ocean's TwelveÂ	The Devil Wears PradaÂ	Meryl Streep
7	490	753	362912	2013	Lions for LambsÂ	InceptionÂ	Mr. & Mrs. SmithÂ	Lions for LambsÂ	Meryl Streep
8	245	2121	364948	2003	Out of AfricaÂ	The RevenantÂ	Ocean's ElevenÂ	Out of AfricaÂ	Meryl Streep
9	739	1588	552503	2015	Hope SpringsÂ	The AviatorÂ	FuryÂ	Hope SpringsÂ	Meryl Streep
10	156	79	32399	2011	One True ThingÂ	Django UnchainedÂ	Seven Years in TibetÂ	One True ThingÂ	Meryl Streep
11	245	450	102338	2008	The HoursÂ	The Wolf of Wall StreetÂ	Fight ClubÂ	The HoursÂ	Meryl Streep
12	61	216	24868	1998	A Prairie Home CompanionÂ	Gangs of New YorkÂ	Sinbad: Legend of the Seven SeasÂ	A Prairie Home CompanionÂ	Meryl Streep
13	308	194	65709	2014		The DepartedÂ	Interview with the Vampire: The Vam	The Great GatsbyÂ	Leonardo DiCaprio
14	435	471	246698	2014		Shutter IslandÂ	The Tree of LifeÂ	The Great GatsbyÂ	Leonardo DiCaprio
15	60	148	27648	1996		Body of LiesÂ	The Assassination of Jesse James by t	Blood DiamondÂ	Leonardo DiCaprio
16	228	576	94456	2005		Catch Me If You CanÂ	By the SeaÂ	The Quick and the DeadÂ	Leonardo DiCaprio
17	79	182	94435	1995		The BeachÂ	Killing Them SoftlyÂ	TitanicÂ	Leonardo DiCaprio
18	222	624	156929	1998		Revolutionary RoadÂ	True RomanceÂ	InceptionÂ	Leonardo DiCaprio
19	71	119	19699	2001		The Man in the Iron MaskÂ		The RevenantÂ	Leonardo DiCaprio
20	183	185	53341	2014		J. EdgarÂ		The AviatorÂ	Leonardo DiCaprio
21	292	281	105797	2009		Marvin's RoomÂ		Django UnchainedÂ	Leonardo DiCaprio
22	20	97	15230	1990		Romeo + JulietÂ		The Wolf of Wall StreetÂ	Leonardo DiCaprio
23	81	121	16372	1999				Gangs of New YorkÂ	Leonardo DiCaprio
24	168	232	43205	2005				The DepartedÂ	Leonardo DiCaprio
25	67	33	8087	2008				Shutter IslandÂ	Leonardo DiCaprio
26	186	119	27882	2013				Body of LiesÂ	Leonardo DiCaprio
27	96	237	17328	1998				Catch Me If You CanÂ	Leonardo DiCaprio
28	32	62	44096	1988				The BeachÂ	Leonardo DiCaprio
29	61	106	5663	2002				Revolutionary RoadÂ	Leonardo DiCaprio
30	58	68	23383	1992				The Man in the Iron MaskÂ	Leonardo DiCaprio
31	138	354	18632	2003				J. EdgarÂ	Leonardo DiCaprio
32	25	12	1427	1997				Marvin's RoomÂ	Leonardo DiCaprio
33	75	115	5772	2006				Romeo + JulietÂ	Leonardo DiCaprio
34	117	193	101840	1985				BabelÂ	Brad Pitt
35	24	6	813	2006				Spy GameÂ	Brad Pitt
36	35	93	74743	1986				The Curious Case of Benjami	Brad Pitt
37	21	43	4792	1989				TroyÂ	Brad Pitt

... | Q-2 highest profit | Q-3(1) Top 250 | Q-3(2) Top Foreign lang film | Q-4 Best Director | Q-5 | Q-6(1) | Q-6(2) Mean ...

6 (2) Find the mean

- We selected the cleaned dataset done in Task 1 and created a pivot table.
- Next, we put 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'.
- Then, we made a pivot chart (pie chart) of the mean of 'num_users_for_review'.
- Again, we did the same above process for the mean of 'num_critic_for_review'.

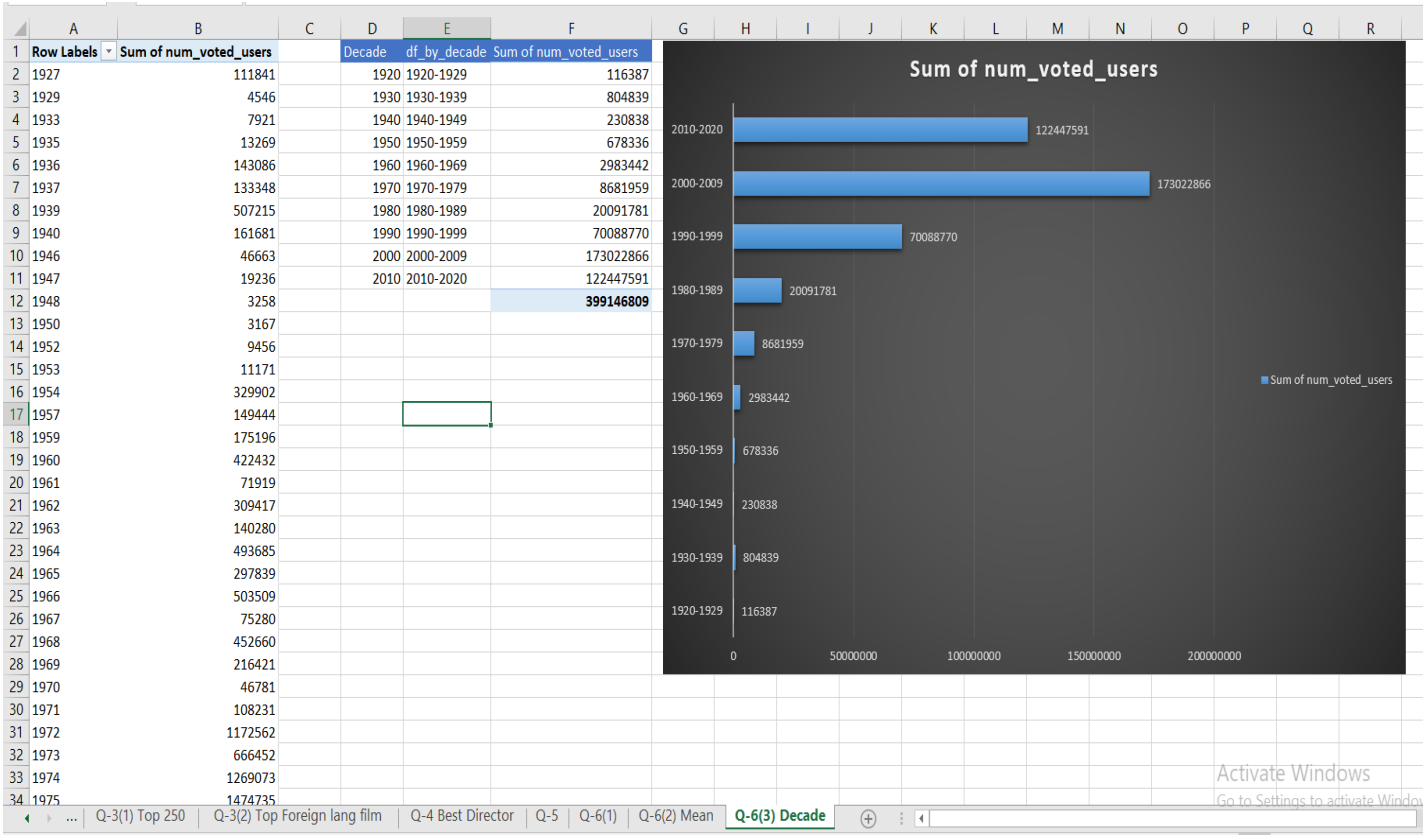


Row Labels	Mean of num_user_for_reviews
Brad Pitt	742
Leonardo DiCaprio	914
Meryl Streep	297
Grand Total	**716**

Row Labels	Mean of num_critic_for_reviews
Brad Pitt	245
Leonardo DiCaprio	330
Meryl Streep	181
Grand Total	**267**

6 (3) Change in number of voted users over decades:

- First, we selected the cleaned dataset done in Task 1 and created a pivot table.
- Second, we put the 'title_year' into the Rows and took the sum of 'num_voted_users' in the Values section.
- Third, we grouped the title_year by decade and stored in df_by_decade column.
- Lastly, we 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
1968	452660
1969	216421
1970	46781
1971	108231
1972	1172562
1973	666452
1974	1269073
1975	1474735

Decade	df_by_decade	Sum of num_voted_users
1920	1920-1929	116387
1930	1930-1939	804839
1940	1940-1949	230838
1950	1950-1959	678336
1960	1960-1969	2983442
1970	1970-1979	8681959
1980	1980-1989	20091781
1990	1990-1999	70088770
2000	2000-2009	173022866
2010	2010-2020	122447591
		399146809

Tech-Stack Used:

• Microsoft Excel 2016: It enables users to format, organize and calculate data in a spreadsheet. It organizes data in an easy-to-navigate way. We need not to perform any complex mathematical functions. And it turns piles of data into helpful graphics and charts.

• Microsoft Word 2016: It is used to make a report (PDF) to be presented to the leadership team.

Insights:

• There are as many as 5 outliers in the profit columns.

• The movie with the highest profit is ‘Avatar’ followed by ‘Jurassic World’ and ‘Titanic’ and so on.

• 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-favorite as well as the audience-favorite actor.

• The most users voted in the decade 2000s and the least in the decade 1940s.

Results:

• In this project, I applied the basic and advance Excel concepts. The concepts related to statistics and EDA have been implemented here by using MS Excel.

• In this task, the concepts regarding the sort, filter, pivot table, charts, different functions like rank, etc have been implemented.

• I learned to implement the learning of Excel in the real-time project.

• I learned how to frame the problem by asking ‘what’ looking at the dataset.

• It helped me in learning the ‘5 Why Analysis’ to determine the root cause of the problem.

• I learned how a data analyst think deeper and deeper to generate the valuable insights.

• It was a great learning experience while doing this project and it was challenging too while asking the different questions and finding their answers.

Excel Sheet Link:

https://drive.google.com/drive/folders/1YDvsRlae9L1LYISLdN5hJ7kAnxFzDN7m?usp=sharing