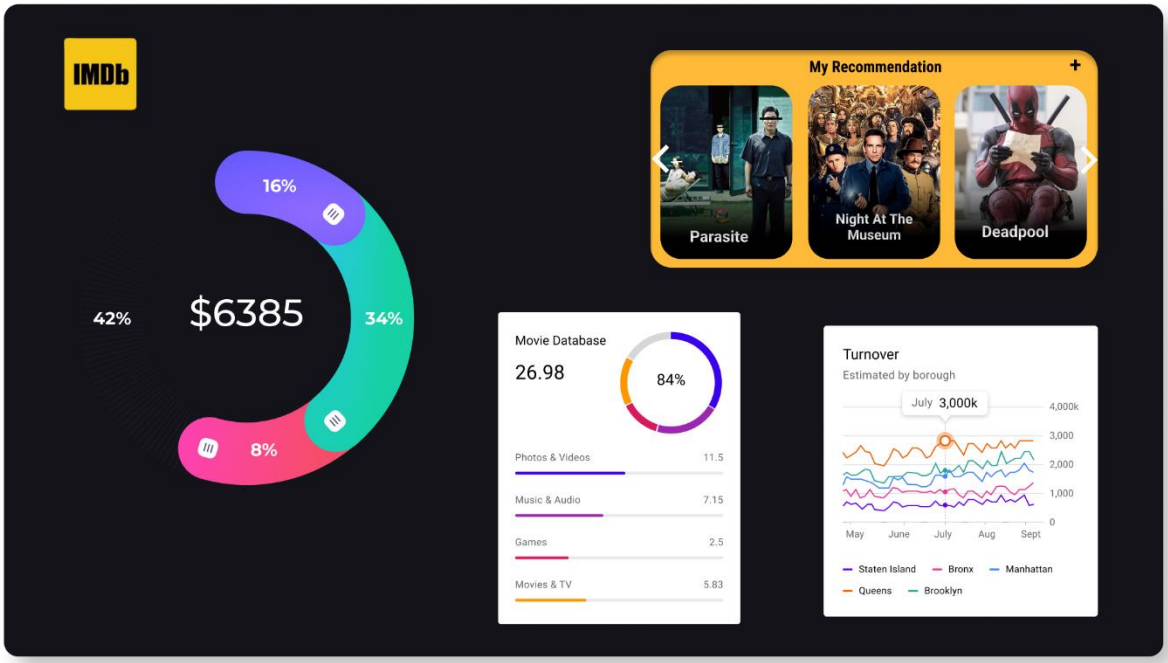
IMDB Movie Analysis

(Final Project-1)

By Shahequa Modabbera

LinkedIn

GitHub



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.

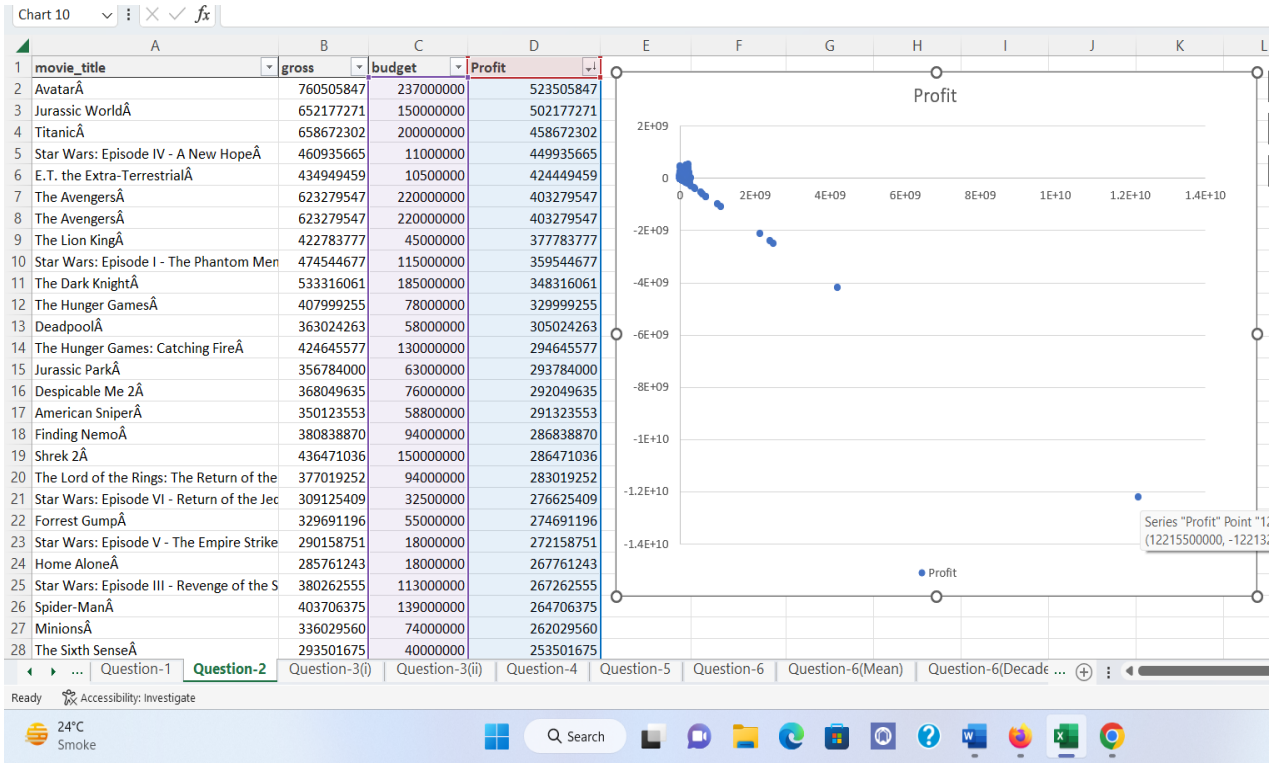
Approach:

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?

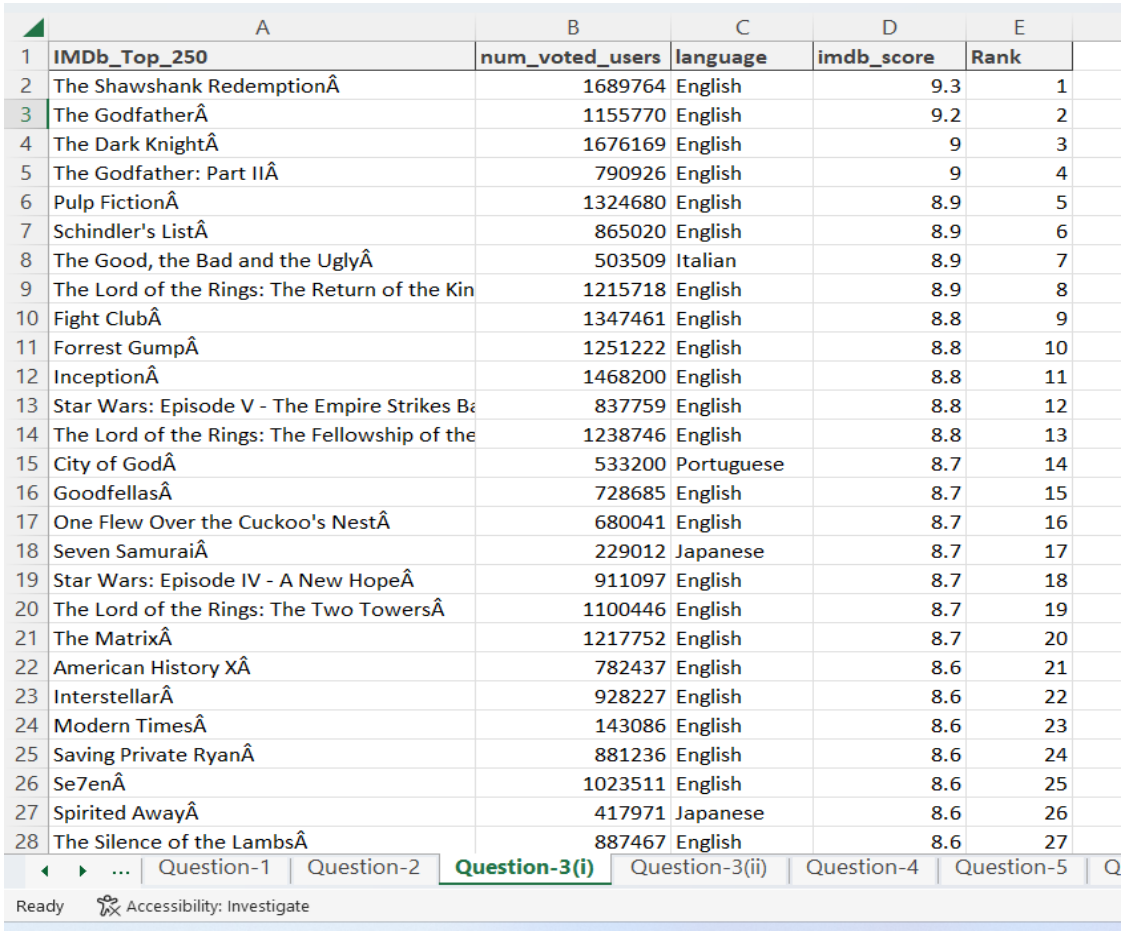
   - First, we created a new column ‘Profit’ by subtracting budget column from gross column.
   - Second, we sorted the columns using the profit column as reference from the largest to the smallest.
   - Third, we plotted the budget and profit in XY Scatter chart to find the outliers.
   - There are as 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	gross	budget	Profit
2	AvatarÂ	760505847	237000000	523505847
3	Jurassic WorldÂ	652177271	150000000	502177271
4	TitanicÂ	658672302	200000000	458672302
5	Star Wars: Episode IV - A New HopeÂ	460935665	11000000	449935665
6	E.T. the Extra-TerrestrialÂ	434949459	10500000	424449459
7	The AvengersÂ	623279547	220000000	403279547
8	The AvengersÂ	623279547	220000000	403279547
9	The Lion KingÂ	422783777	45000000	377783777
10	Star Wars: Episode I - The Phantom Men	474544677	115000000	359544677
11	The Dark KnightÂ	533316061	185000000	348316061
12	The Hunger GamesÂ	407999255	78000000	329999255
13	DeadpoolÂ	363024263	58000000	305024263
14	The Hunger Games: Catching FireÂ	424645577	130000000	294645577
15	Jurassic ParkÂ	356784000	63000000	293784000
16	Despicable Me 2Â	368049635	76000000	292049635
17	American SniperÂ	350123553	58800000	291323553
18	Finding NemoÂ	380838870	94000000	286838870
19	Shrek 2Â	436471036	150000000	286471036
20	The Lord of the Rings: The Return of the	377019252	94000000	283019252
21	Star Wars: Episode VI - Return of the Jed	309125409	32500000	276625409
22	Forrest GumpÂ	329691196	55000000	274691196
23	Star Wars: Episode V - The Empire Strike	290158751	18000000	272158751
24	Home AloneÂ	285761243	18000000	267761243
25	Star Wars: Episode III - Revenge of the S	380262555	113000000	267262555
26	Spider-ManÂ	403706375	139000000	264706375
27	MinionsÂ	336029560	74000000	262029560
28	The Sixth SenseÂ	293501675	40000000	253501675

3. **Task:** Find IMDB Top 250

- First, we filtered the ‘num_voted_users’ column greater than 25,000.
- Second, 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, 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	Schindler's ListÂ	865020	English	8.9	6
8	The Good, the Bad and the UglyÂ	503509	Italian	8.9	7
9	The Lord of the Rings: The Return of the Kin	1215718	English	8.9	8
10	Fight ClubÂ	1347461	English	8.8	9
11	Forrest GumpÂ	1251222	English	8.8	10
12	InceptionÂ	1468200	English	8.8	11
13	Star Wars: Episode V - The Empire Strikes Ba	837759	English	8.8	12
14	The Lord of the Rings: The Fellowship of the	1238746	English	8.8	13
15	City of GodÂ	533200	Portuguese	8.7	14
16	GoodfellasÂ	728685	English	8.7	15
17	One Flew Over the Cuckoo's NestÂ	680041	English	8.7	16
18	Seven SamuraiÂ	229012	Japanese	8.7	17
19	Star Wars: Episode IV - A New HopeÂ	911097	English	8.7	18
20	The Lord of the Rings: The Two TowersÂ	1100446	English	8.7	19
21	The MatrixÂ	1217752	English	8.7	20
22	American History XÂ	782437	English	8.6	21
23	InterstellarÂ	928227	English	8.6	22
24	Modern TimesÂ	143086	English	8.6	23
25	Saving Private RyanÂ	881236	English	8.6	24
26	Se7enÂ	1023511	English	8.6	25
27	Spirited AwayÂ	417971	Japanese	8.6	26
28	The Silence of the LambsÂ	887467	English	8.6	27

	A	B	C	D
1	Top_Foreign_Lang_Film	num_voted_users	imdb_score	language
2	The Good, the Bad and the UglyÂ	503509	8.9	Italian
3	City of GodÂ	533200	8.7	Portuguese
4	Seven SamuraiÂ	229012	8.7	Japanese
5	Spirited AwayÂ	417971	8.6	Japanese
6	Children of HeavenÂ	27882	8.5	Persian
7	The Lives of OthersÂ	259379	8.5	German
8	A SeparationÂ	151812	8.4	Persian
9	AmÃ©lieÂ	534262	8.4	French
10	Baahubali: The BeginningÂ	62756	8.4	Telugu
11	Das BootÂ	168203	8.4	German
12	OldboyÂ	356181	8.4	Korean
13	Princess MononokeÂ	221552	8.4	Japanese
14	MetropolisÂ	111841	8.3	German
15	The HuntÂ	170155	8.3	Danish
16	UnforgivenÂ	248354	8.3	German
17	Pan's LabyrinthÂ	80429	8.2	French
18	The Bridge on the River KwaiÂ	131831	8.2	Spanish
19	The ThingÂ	467234	8.2	Spanish
20	WarriorÂ	214091	8.2	Japanese
21	Annie HallÂ	81644	8.1	Portuguese
22	In the Shadow of the MoonÂ	65951	8.1	Danish
23	Sling BladeÂ	106160	8.1	Japanese
24	Tae Guk Gi: The Brotherhood of WarÂ	64556	8.1	Spanish
25	The Best Years of Our LivesÂ	173551	8.1	Spanish
26	The Imitation GameÂ	31943	8.1	Korean
27	Bowling for ColumbineÂ	28951	8	Portuguese
28	JawsÂ	70194	8	French

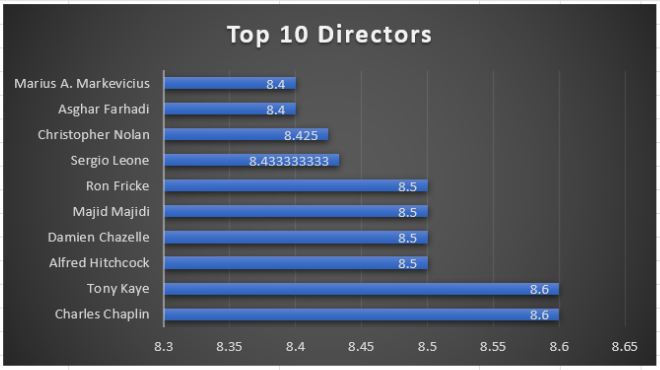
... | Question-1 | Question-2 | Question-3(i) | **Question-3(ii)** | Question-4

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.

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



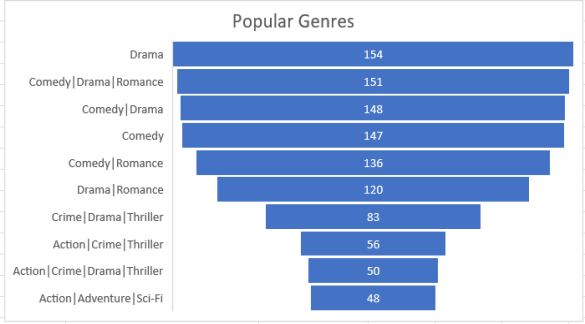
Question-1 | Question-2 | Question-3(i) | Question-3(ii) | **Question-4** | Question-5 | Question-6 | Question-6(Mean) | Question-6(De

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 funnel chart of the top 10 genres for the better insights.

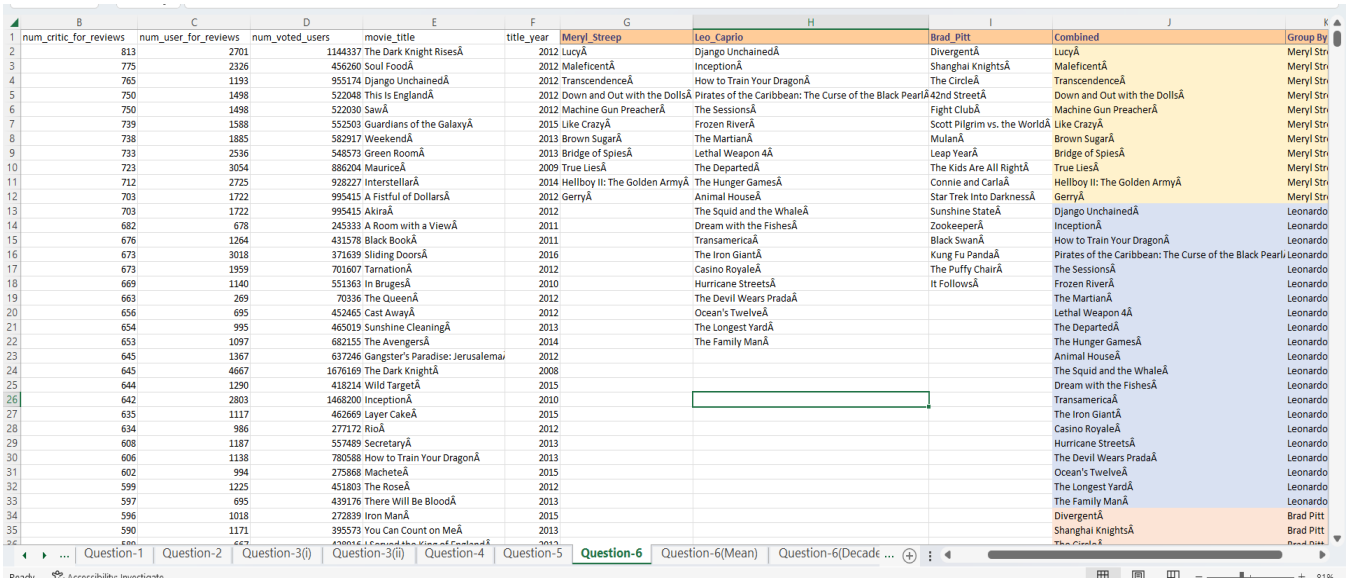
Row Labels	Count of genres
Drama	154
Comedy|Drama|Romance	151
Comedy|Drama	148
Comedy	147
Comedy|Romance	136
Drama|Romance	120
Crime|Drama|Thriller	83
Action|Crime|Thriller	56
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	48
Comedy|Crime	47
Action|Adventure|Thriller	45
Horror	44
Drama|Thriller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

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

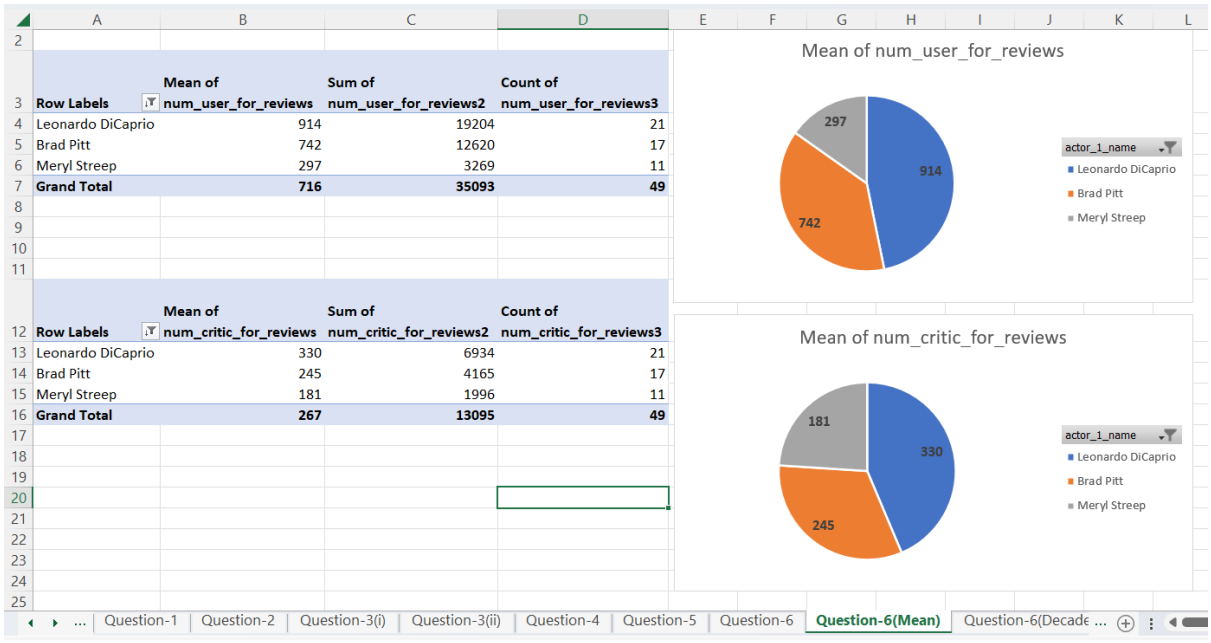


Question-1 | Question-2 | Question-3(i) | Question-3(ii) | Question-4 | **Question-5** | Question-6 | Question-6(Mean) | Question-6(Decade ...

6. **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'.
    - Then, 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 (bar chart) of the mean of ‘num_users_for_review’.
    - Again, we did the same above process for the mean of ‘num_critic_for_review’.



num_critic_for_reviews	num_user_for_reviews	num_voted_users	movie_title	title_year	Meryl_Streep	Leo_Caprio	Brad_Pitt	Combined	Group By
813	2701	1144337	The Dark Knight RisesÂ	2012	LucyÂ	Django UnchainedÂ	DivergentÂ	LucyÂ	Meryl Str
775	2326	456260	Soul FoodÂ	2012	MaleficentÂ	InceptionÂ	Shanghai KnightsÂ	MaleficentÂ	Meryl Str
765	1193	955174	Django UnchainedÂ	2012	TranscendenceÂ	How to Train Your DragonÂ	The CircleÂ	TranscendenceÂ	Meryl Str
750	1498	522048	This Is EnglandÂ	2012	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 Str
750	1498	522030	SawÂ	2012	Machine Gun PreacherÂ	The SessionsÂ	Fight ClubÂ	Machine Gun PreacherÂ	Meryl Str
739	1588	552503	Guardians of the GalaxyÂ	2015	Like CrazyÂ	Frozen RiverÂ	Scott Pilgrim vs. the WorldÂ	Like CrazyÂ	Meryl Str
738	1885	582917	WeekendÂ	2013	Brown SugarÂ	The MartianÂ	MulanÂ	Brown SugarÂ	Meryl Str
733	2536	548573	Green RoomÂ	2013	Bridge of SpiesÂ	Lethal Weapon 4Â	Leap YearÂ	Bridge of SpiesÂ	Meryl Str
723	3054	886204	MauriceÂ	2009	True LiesÂ	The DepartedÂ	The Kids Are All RightÂ	True LiesÂ	Meryl Str
712	2725	928227	InterstellarÂ	2014	Hellboy II: The Golden ArmyÂ	The Hunger GamesÂ	Connie and CarlaÂ	Hellboy II: The Golden ArmyÂ	Meryl Str
703	1722	995415	A Fistful of DollarsÂ	2012	GerryÂ	Animal HouseÂ	Star Trek Into DarknessÂ	GerryÂ	Meryl Str
703	1722	995415	AkiraÂ	2012		The Squid and the WhaleÂ	Sunshine StateÂ	Django UnchainedÂ	Leonardo
682	678	245333	A Room with a ViewÂ	2011		Dream with the FishesÂ	ZookeeperÂ	InceptionÂ	Leonardo
676	1264	431578	Black BookÂ	2011		TransamericaÂ	Black SwanÂ	How to Train Your DragonÂ	Leonardo
673	3018	371639	Sliding DoorsÂ	2016		The Iron GiantÂ	Kung Fu PandaÂ	Pirates of the Caribbean: The Curse of the Black PearlÂ	Leonardo
673	1959	701607	TarnationÂ	2012		Casino RoyaleÂ	The Puffy ChairÂ	The SessionsÂ	Leonardo
669	1140	551363	In BrugesÂ	2010		Hurricane StreetsÂ	It FollowsÂ	Frozen RiverÂ	Leonardo
663	269	70336	The QueenÂ	2012		The Devil Wears PradaÂ		The MartianÂ	Leonardo
656	695	452465	Cast AwayÂ	2012		Ocean's TwelveÂ		Lethal Weapon 4Â	Leonardo
654	995	465019	Sunshine CleaningÂ	2013		The Longest YardÂ		The DepartedÂ	Leonardo
653	1097	682155	The AvengersÂ	2014		The Family ManÂ		The Hunger GamesÂ	Leonardo
645	1367	637246	Gangster's Paradise: Jerusalema	2012				Animal HouseÂ	Leonardo
645	4667	1676169	The Dark KnightÂ	2008				The Squid and the WhaleÂ	Leonardo
644	1290	418214	Wild TargetÂ	2015				Dream with the FishesÂ	Leonardo
642	2803	1468200	InceptionÂ	2010				TransamericaÂ	Leonardo
635	1117	462669	Layer CakeÂ	2015				The Iron GiantÂ	Leonardo
634	986	277172	RioÂ	2012				Casino RoyaleÂ	Leonardo
608	1187	557489	SecretaryÂ	2013				Hurricane StreetsÂ	Leonardo
606	1138	780588	How to Train Your DragonÂ	2013				The Devil Wears PradaÂ	Leonardo
602	994	275868	MacheteÂ	2015				Ocean's TwelveÂ	Leonardo
599	1225	451803	The RoseÂ	2012				The Longest YardÂ	Leonardo
597	695	439176	There Will Be BloodÂ	2013				The Family ManÂ	Leonardo
596	1018	272839	Iron ManÂ	2015				DivergentÂ	Brad Pitt
590	1171	395573	You Can Count on MeÂ	2013				Shanghai KnightsÂ	Brad Pitt

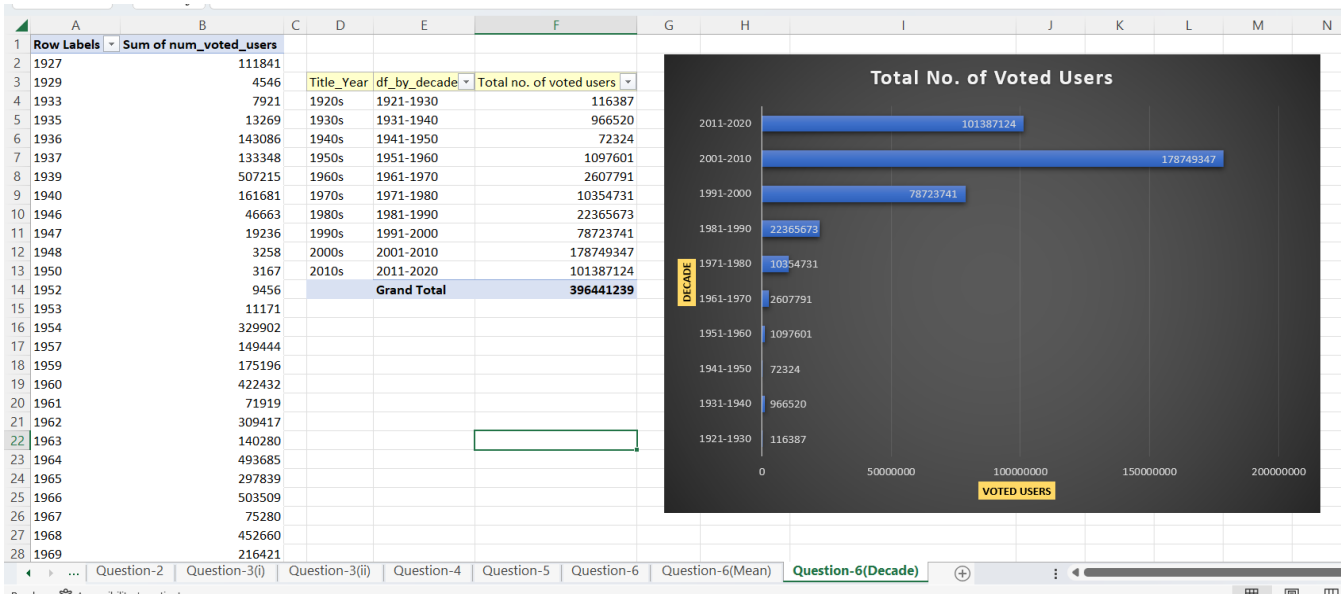


Row Labels	Mean of num_user_for_reviews	Sum of num_user_for_reviews2	Count of num_user_for_reviews3
Leonardo DiCaprio	914	19204	21
Brad Pitt	742	12620	17
Meryl Streep	297	3269	11
Grand Total	**716**	**35093**	**49**

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**

Second part(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

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:

- **Microsoft Excel 365:** It enables users to format, organize and calculate data in a spreadsheet. It organize data in an easy-to-navigate way. We need not to perform any complex mathematical functions. And it turn piles of data into helpful graphics and charts.

- **Microsoft Word 2021:** 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:

- IMDB Movie Analysis.xlsx
- Make sure to open it in the MS Excel for the right visuals.