IMBD MOVIE ANALYSIS

Project description:

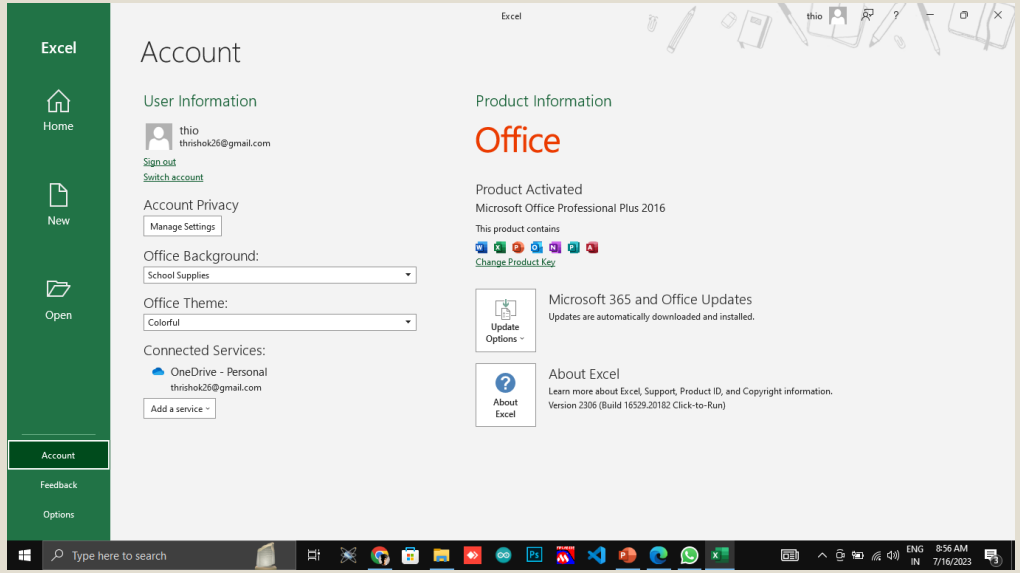
- Providing you with dataset having various columns of different IMDB Movies. You are required to Frame the problem. For this task, you will need to define a problem you want to shed some light on.
- Clean the data as necessary, and use your Data Analysis skills to explore the data set and derive insights.
- Following ‘five why’ approach.

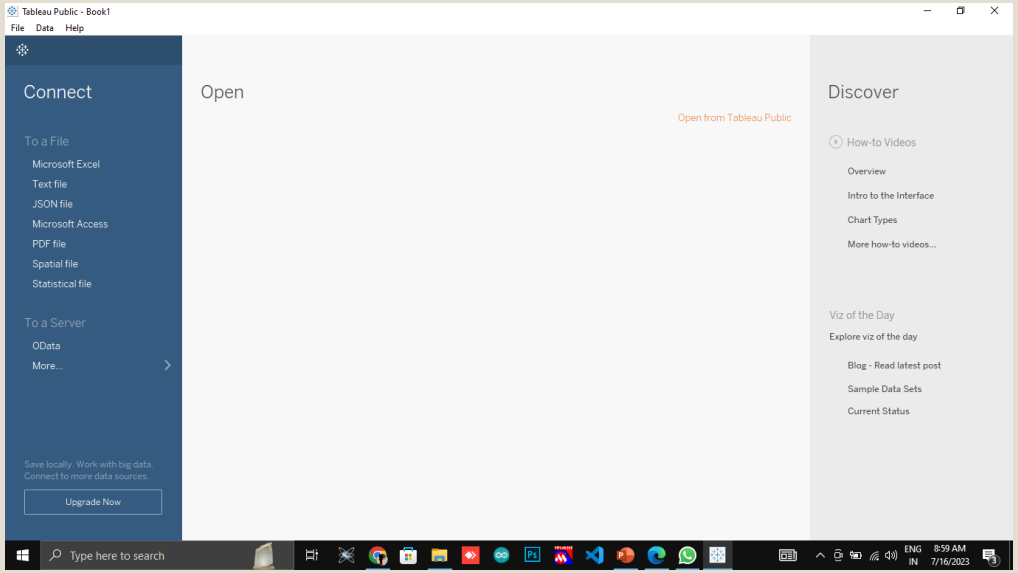
Approach:

- Five 'Whys' approach
- Cleaning data (Dropping unnecessary columns, removing null values, etc.)
- Created a database and then the tables using the structure.
- Joined the data bits and structured the tables to derive business insights, used pivot tables, fetched the required result.

TECH-STACK USED:

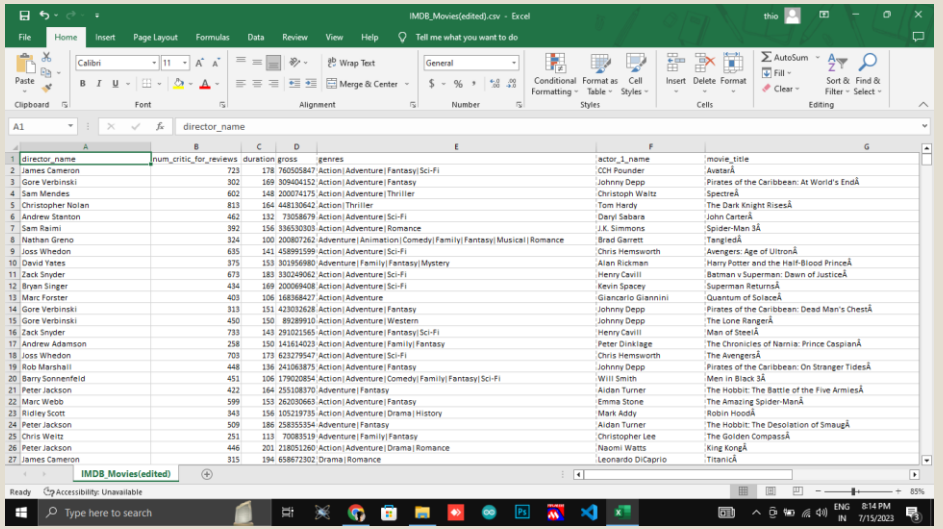
- For this project I have used MS Excel and Tableau.

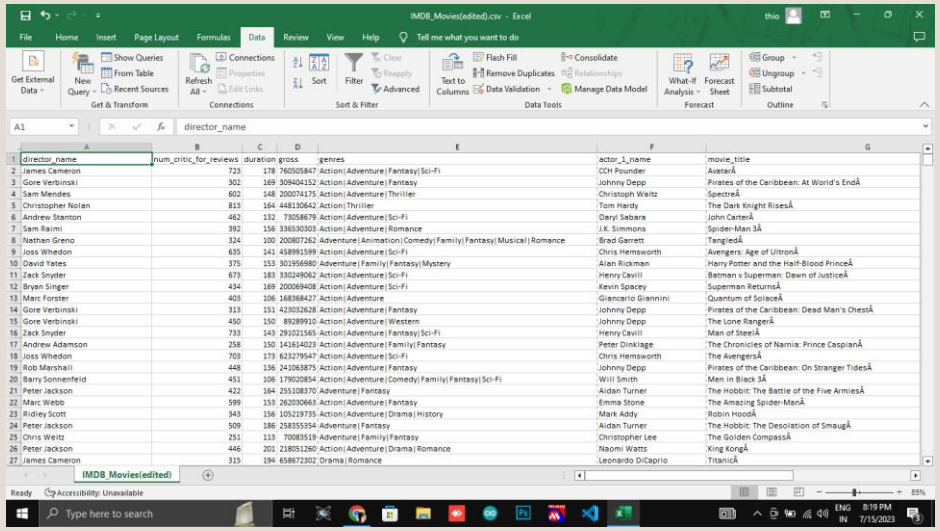


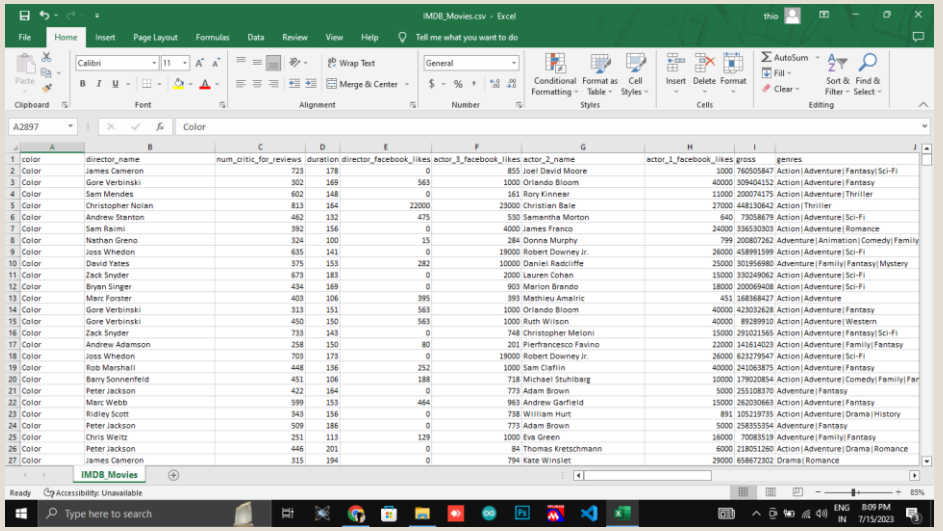


A . Cleaning data:

- Unnecessary columns , rows with null values or blank spaces and duplicate values has been removed.

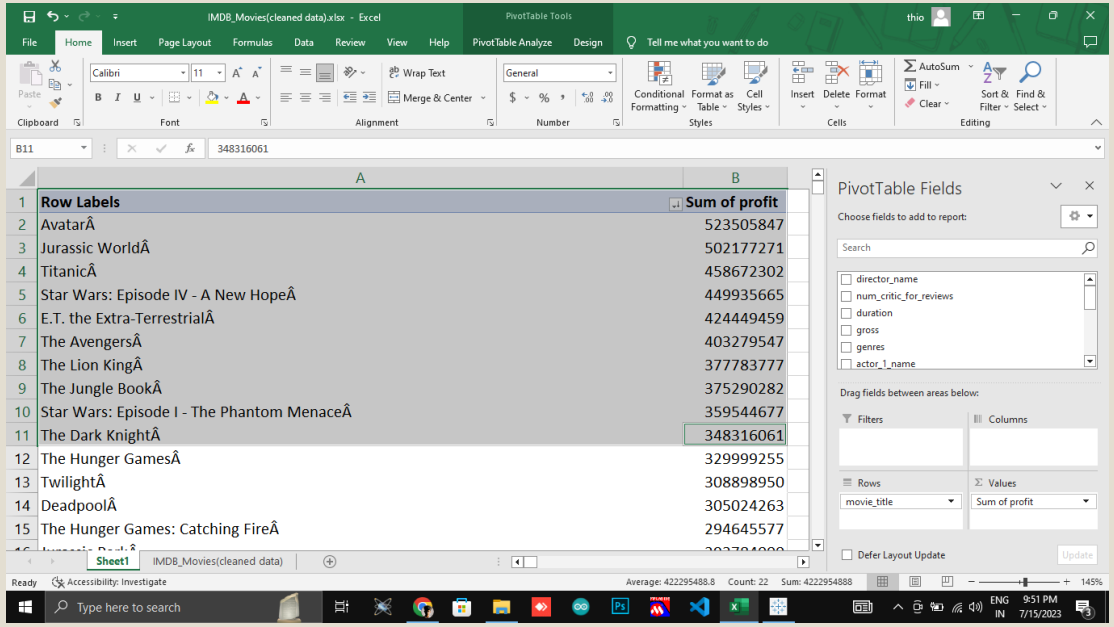




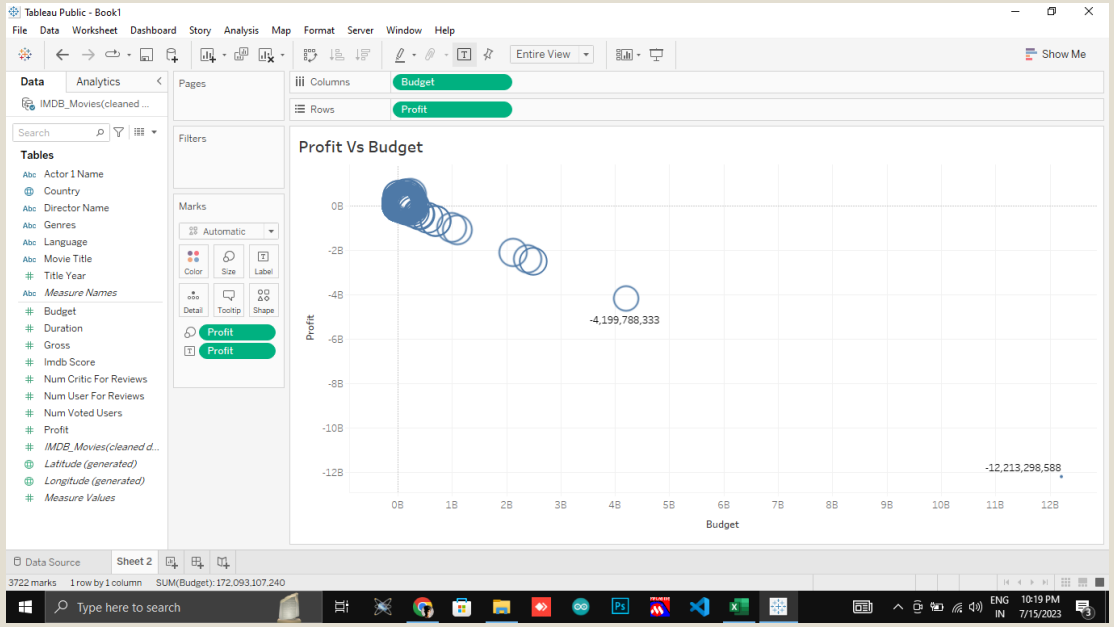


B. Movies with highest profit:

- Movies with highest profit has been calculated by subtracting Budget from Gross.
- Outliers are also found and plotted by Profit in Y-axis and Budget in X-axis.

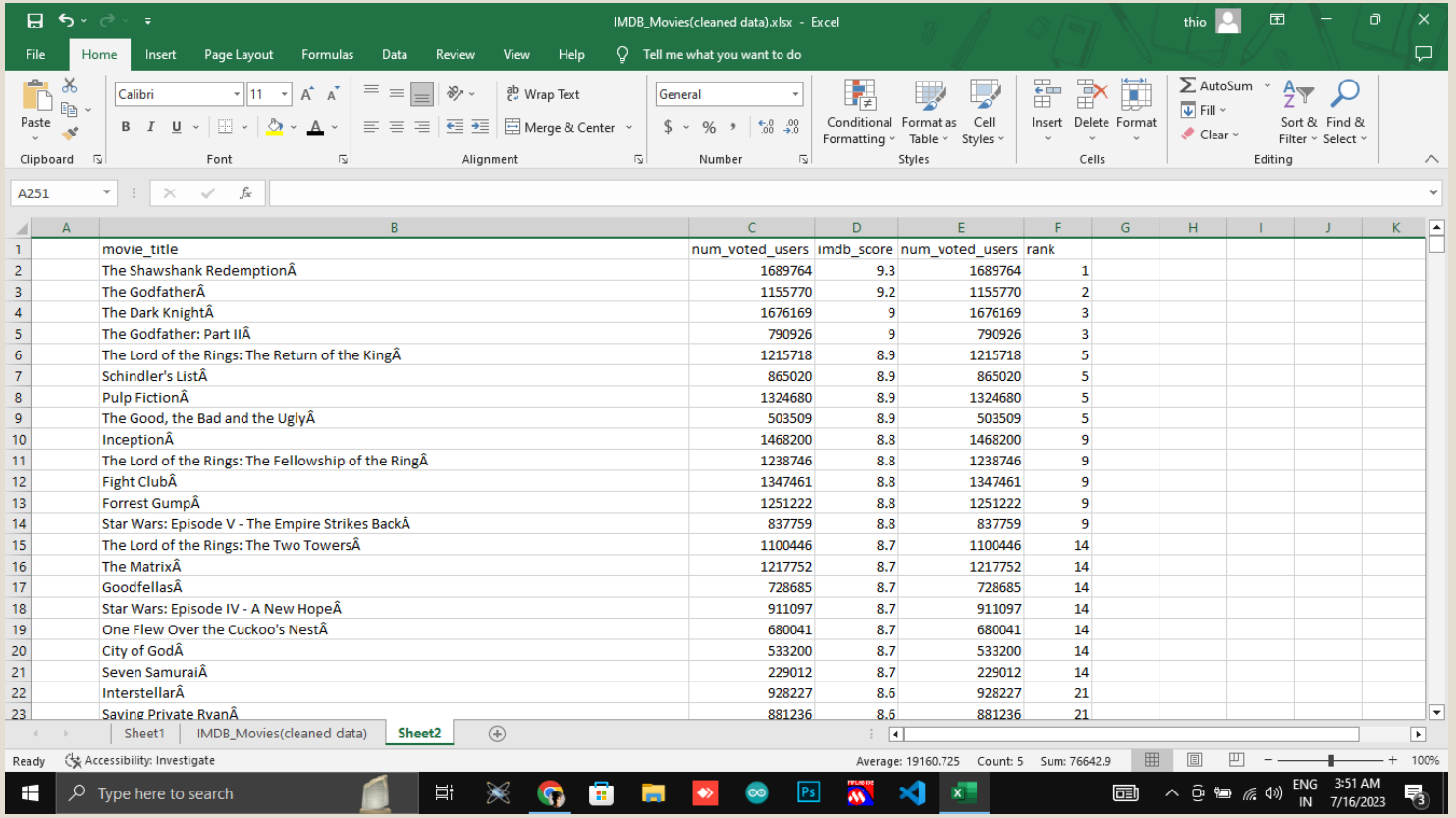


Row Labels	Sum of profit
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A New HopeÂ	449935665
E.T. the Extra-TerrestrialÂ	424449459
The AvengersÂ	403279547
The Lion KingÂ	377783777
The Jungle BookÂ	375290282
Star Wars: Episode I - The Phantom MenaceÂ	359544677
The Dark KnightÂ	348316061
The Hunger GamesÂ	329999255
TwilightÂ	308898950
DeadpoolÂ	305024263
The Hunger Games: Catching FireÂ	294645577



C . IMDB Top 250:

From this table the movie “Shawshank redemption” is the number one movie with IMDB_score of 9.3

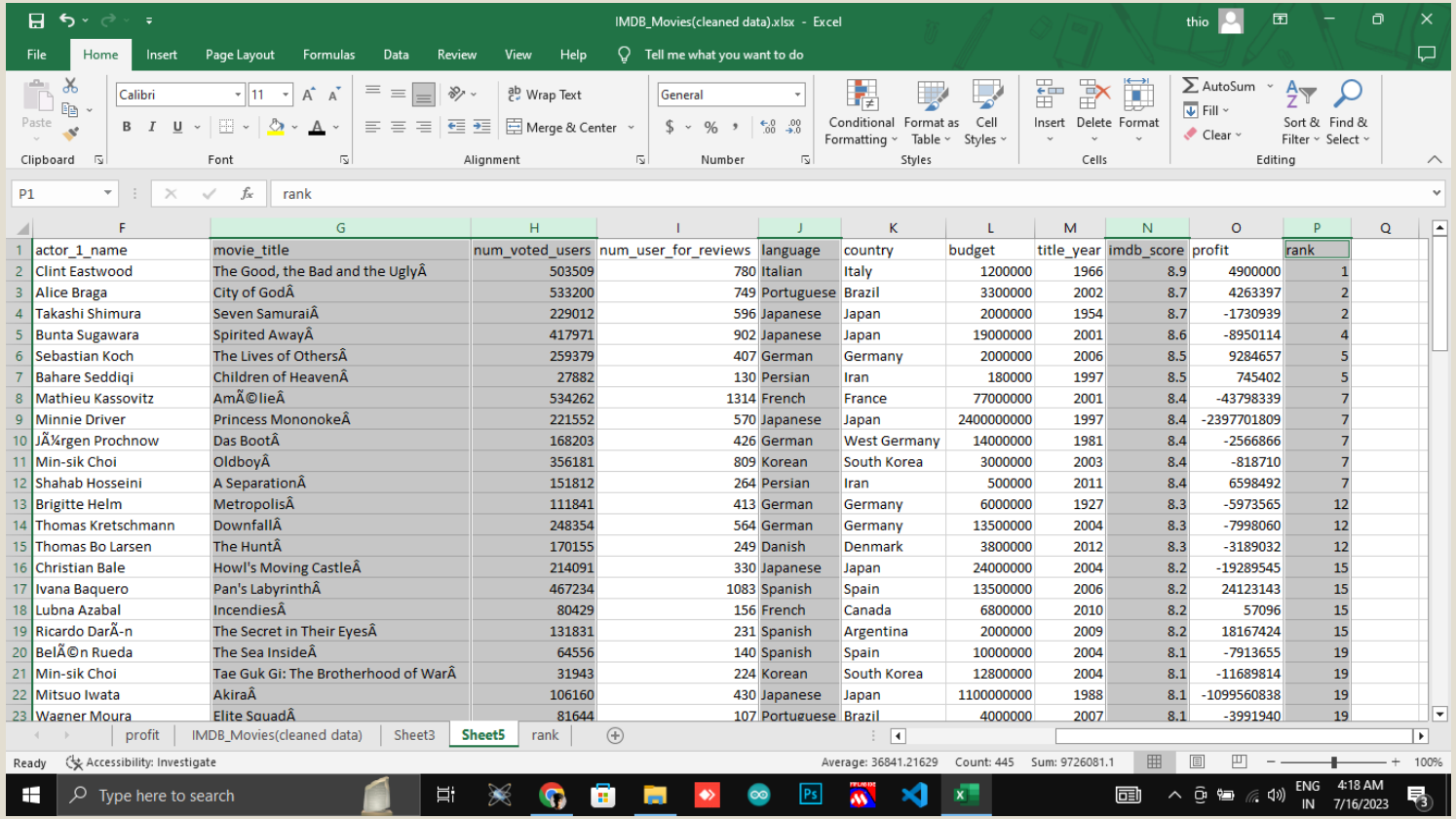


IMDB_Movies(cleaned data).xlsx - Excel

	movie_title	num_voted_users	imdb_score	num_voted_users	rank
2	The Shawshank RedemptionÂ	1689764	9.3	1689764	1
3	The GodfatherÂ	1155770	9.2	1155770	2
4	The Dark KnightÂ	1676169	9	1676169	3
5	The Godfather: Part IIÂ	790926	9	790926	3
6	The Lord of the Rings: The Return of the KingÂ	1215718	8.9	1215718	5
7	Schindler's ListÂ	865020	8.9	865020	5
8	Pulp FictionÂ	1324680	8.9	1324680	5
9	The Good, the Bad and the UglyÂ	503509	8.9	503509	5
10	InceptionÂ	1468200	8.8	1468200	9
11	The Lord of the Rings: The Fellowship of the RingÂ	1238746	8.8	1238746	9
12	Fight ClubÂ	1347461	8.8	1347461	9
13	Forrest GumpÂ	1251222	8.8	1251222	9
14	Star Wars: Episode V - The Empire Strikes BackÂ	837759	8.8	837759	9
15	The Lord of the Rings: The Two TowersÂ	1100446	8.7	1100446	14
16	The MatrixÂ	1217752	8.7	1217752	14
17	GoodfellasÂ	728685	8.7	728685	14
18	Star Wars: Episode IV - A New HopeÂ	911097	8.7	911097	14
19	One Flew Over the Cuckoo's NestÂ	680041	8.7	680041	14
20	City of GodÂ	533200	8.7	533200	14
21	Seven SamuraiÂ	229012	8.7	229012	14
22	InterstellarÂ	928227	8.6	928227	21
23	Saving Private RyanÂ	881236	8.6	881236	21

C . IMDB Top 250(non-English):

From this table the movie
“the good and the bad and the ugly” is the number one foreign language movie with IMDB_score of 8.9

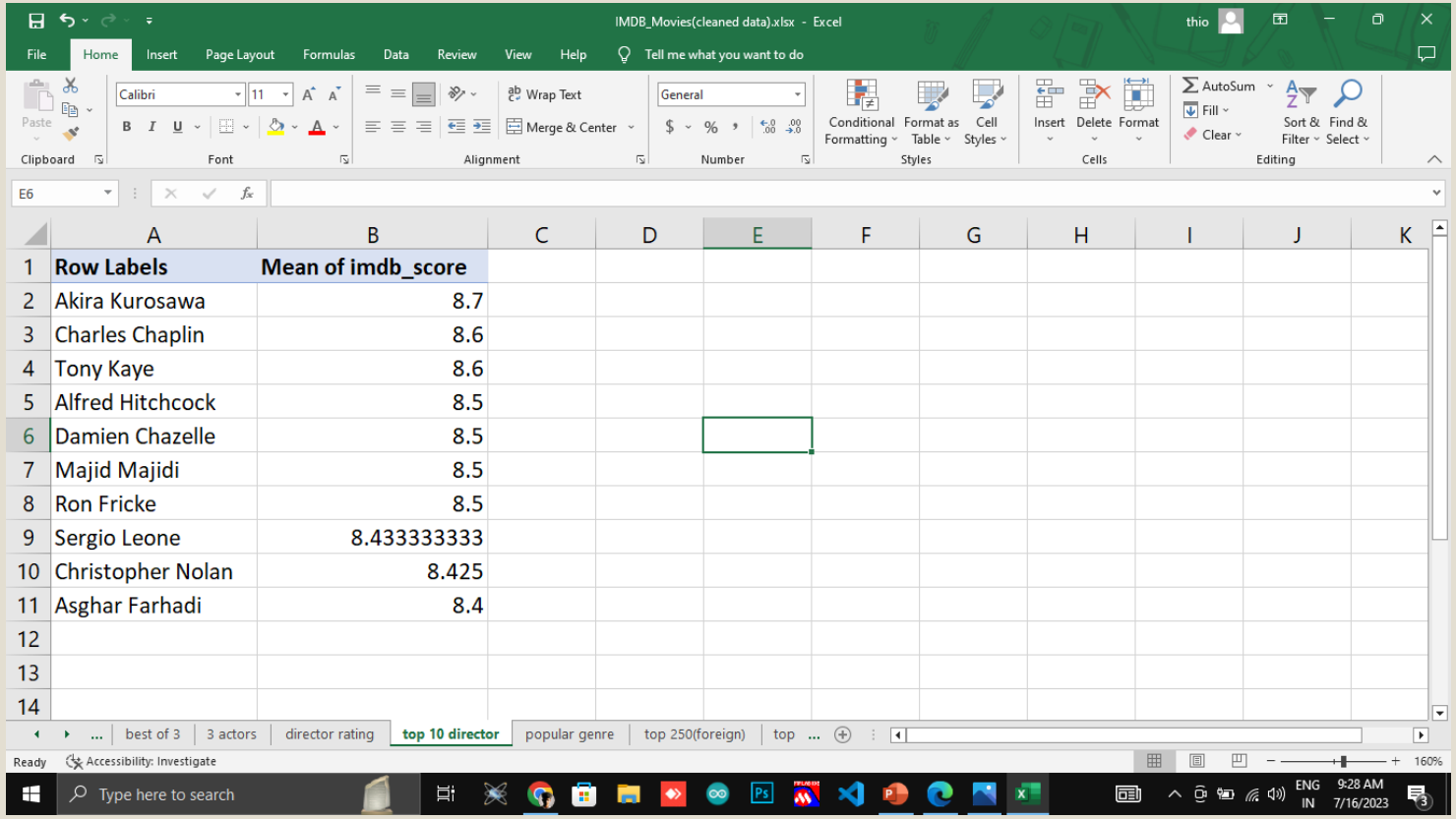


IMDB_Movies(cleaned data).xlsx - Excel

	F	G	H	I	J	K	L	M	N	O	P	Q
1	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	budget	title_year	imdb_score	profit	rank	
2	Clint Eastwood	The Good, the Bad and the UglyÂ	503509	780	Italian	Italy	1200000	1966	8.9	4900000	1	
3	Alice Braga	City of GodÂ	533200	749	Portuguese	Brazil	3300000	2002	8.7	4263397	2	
4	Takashi Shimura	Seven SamuraiÂ	229012	596	Japanese	Japan	2000000	1954	8.7	-1730939	2	
5	Bunta Sugawara	Spirited AwayÂ	417971	902	Japanese	Japan	19000000	2001	8.6	-8950114	4	
6	Sebastian Koch	The Lives of OthersÂ	259379	407	German	Germany	2000000	2006	8.5	9284657	5	
7	Bahare Seddiqi	Children of HeavenÂ	27882	130	Persian	Iran	180000	1997	8.5	745402	5	
8	Mathieu Kassovitz	AmÃ©lieÂ	534262	1314	French	France	77000000	2001	8.4	-43798339	7	
9	Minnie Driver	Princess MononokeÂ	221552	570	Japanese	Japan	2400000000	1997	8.4	-2397701809	7	
10	JÃ¼rgen Prochnow	Das BootÂ	168203	426	German	West Germany	14000000	1981	8.4	-2566866	7	
11	Min-sik Choi	OldboyÂ	356181	809	Korean	South Korea	3000000	2003	8.4	-818710	7	
12	Shahab Hosseini	A SeparationÂ	151812	264	Persian	Iran	500000	2011	8.4	6598492	7	
13	Brigitte Helm	MetropolisÂ	111841	413	German	Germany	6000000	1927	8.3	-5973565	12	
14	Thomas Kretschmann	DownfallÂ	248354	564	German	Germany	13500000	2004	8.3	-7998060	12	
15	Thomas Bo Larsen	The HuntÂ	170155	249	Danish	Denmark	3800000	2012	8.3	-3189032	12	
16	Christian Bale	Howl's Moving CastleÂ	214091	330	Japanese	Japan	24000000	2004	8.2	-19289545	15	
17	Ivana Baquero	Pan's LabyrinthÂ	467234	1083	Spanish	Spain	13500000	2006	8.2	24123143	15	
18	Lubna Azabal	IncendiesÂ	80429	156	French	Canada	6800000	2010	8.2	57096	15	
19	Ricardo DarÃ-n	The Secret in Their EyesÂ	131831	231	Spanish	Argentina	2000000	2009	8.2	18167424	15	
20	BelÃ©n Rueda	The Sea InsideÂ	64556	140	Spanish	Spain	10000000	2004	8.1	-7913655	19	
21	Min-sik Choi	Tae Guk Gi: The Brotherhood of WarÂ	31943	224	Korean	South Korea	12800000	2004	8.1	-11689814	19	
22	Mitsuo Iwata	AkiraÂ	106160	430	Japanese	Japan	1100000000	1988	8.1	-1099560838	19	
23	Wagner Moura	Elite SquadÂ	81644	107	Portuguese	Brazil	4000000	2007	8.1	-3991940	19	

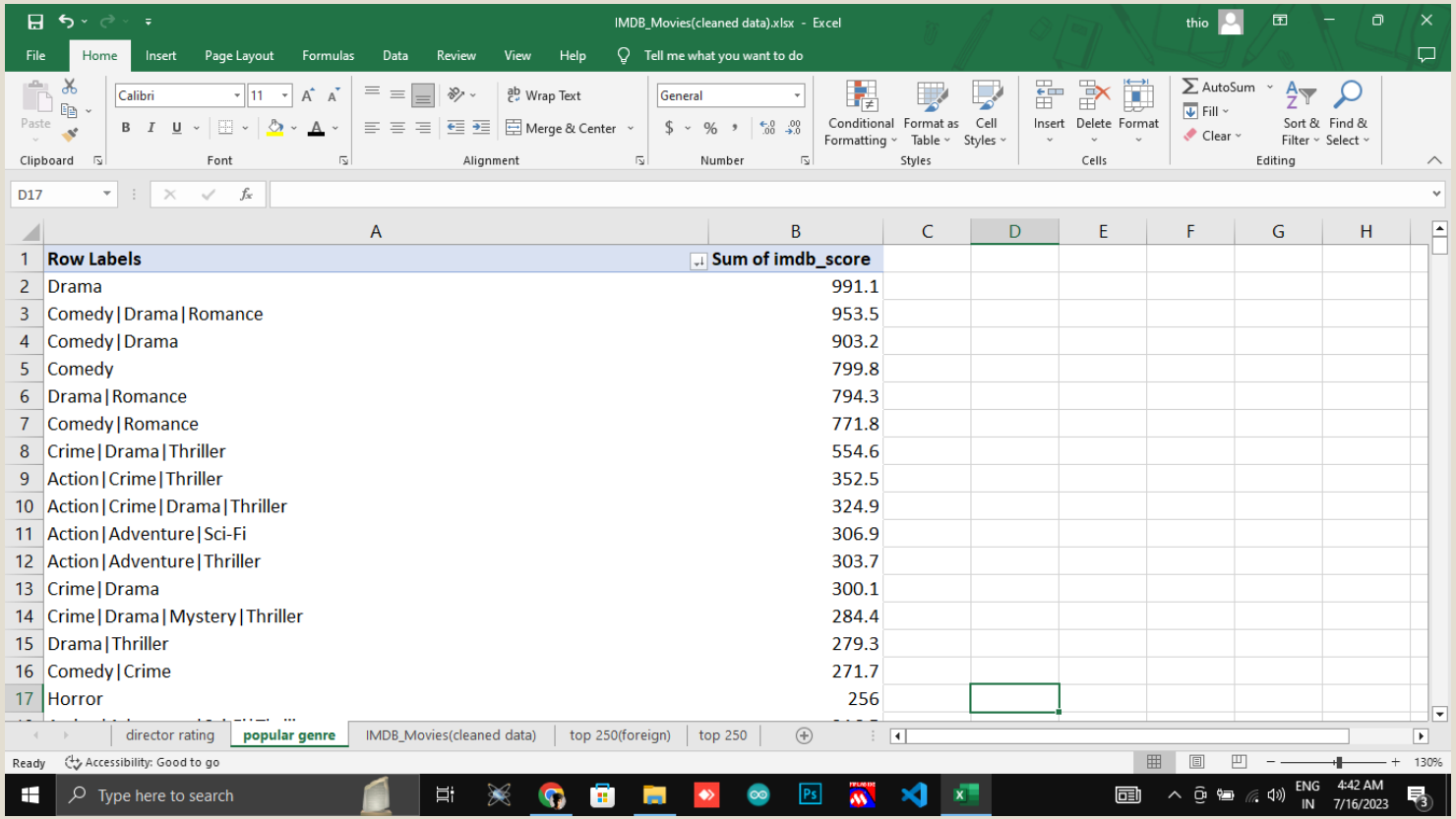
D . Best Directors:

These are the top 10 directors based on IMDB rating. Top director among highest IMDB score directors is “Akira kurosawa”



E . Popular genres:

Based on the analysis Drama , comedy and romance has most number IMDB ratings.

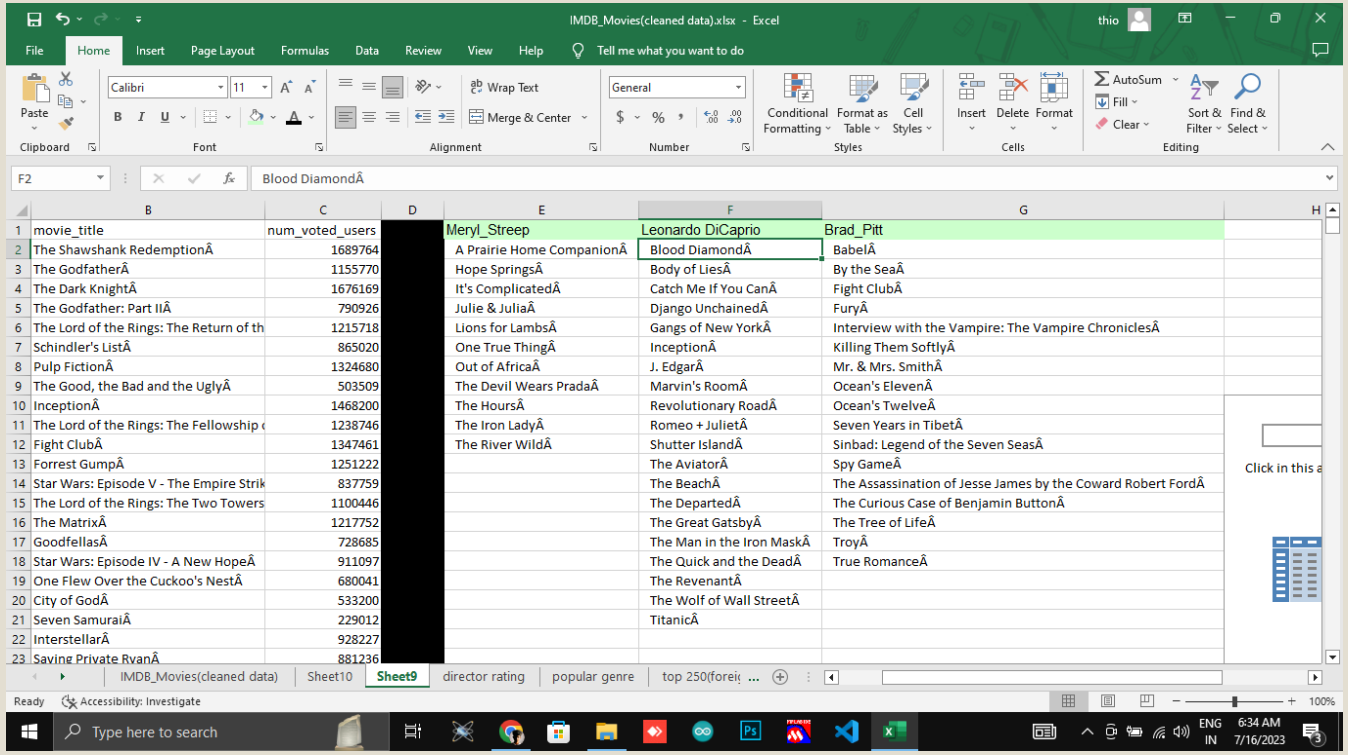


IMDB_Movies(cleaned data).xlsx - Excel

Row Labels	Sum of imdb_score
Drama	991.1
Comedy|Drama|Romance	953.5
Comedy|Drama	903.2
Comedy	799.8
Drama|Romance	794.3
Comedy|Romance	771.8
Crime|Drama|Thriller	554.6
Action|Crime|Thriller	352.5
Action|Crime|Drama|Thriller	324.9
Action|Adventure|Sci-Fi	306.9
Action|Adventure|Thriller	303.7
Crime|Drama	300.1
Crime|Drama|Mystery|Thriller	284.4
Drama|Thriller	279.3
Comedy|Crime	271.7
Horror	256

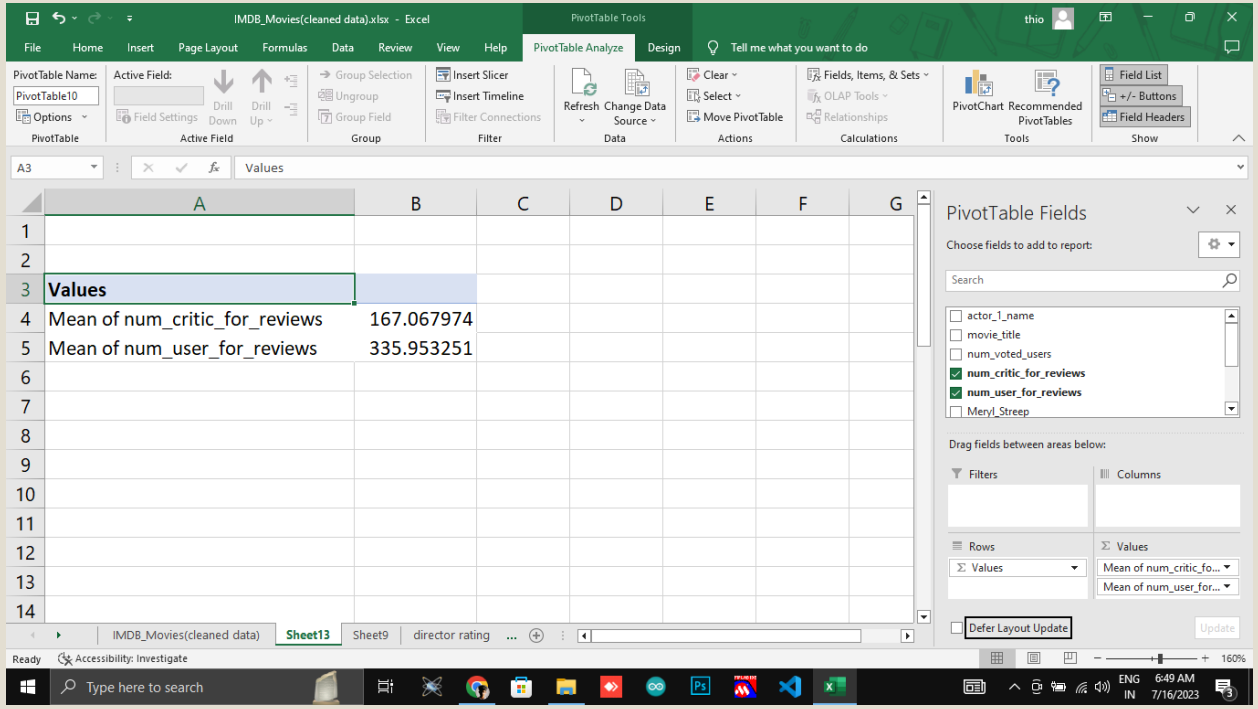
F . Task(1):

- 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.



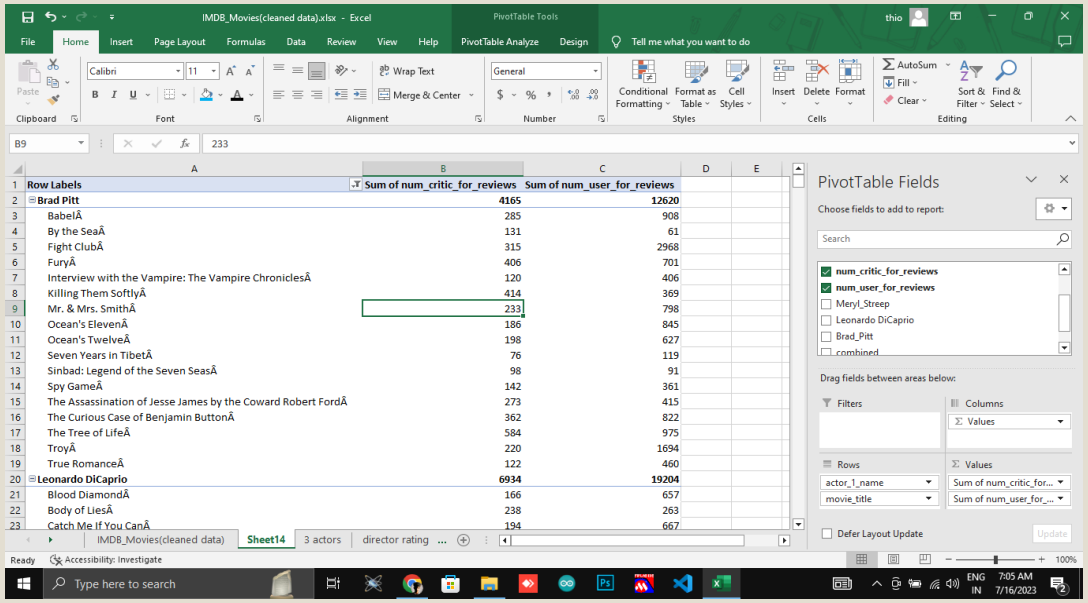
F . Task(2):

- Find the mean of the num_critic_for_reviews and num_users_for_review.

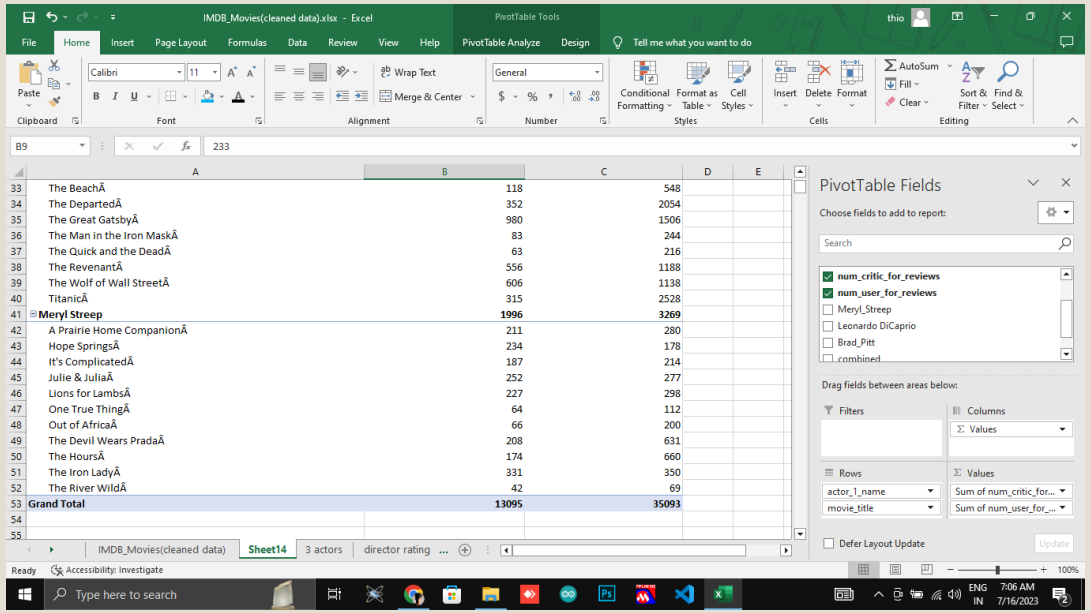


F . Task(3):

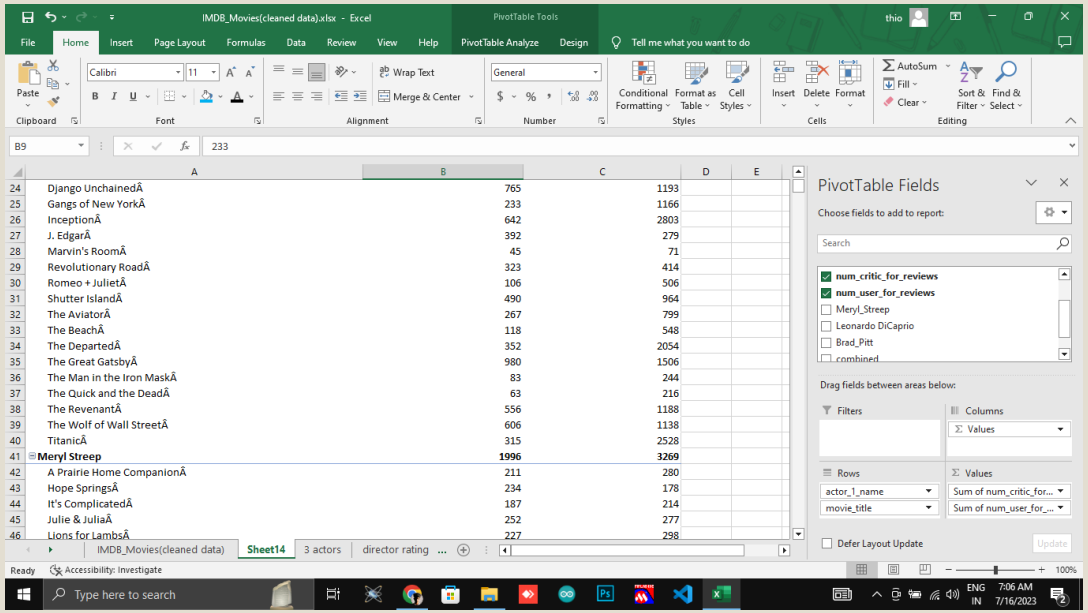
Identify the actors which have the highest mean.



Row Labels	Sum of num_critic_for_reviews	Sum of num_user_for_reviews
Brad Pitt	4165	12620
BabelÂ	285	908
By the SeaÂ	131	61
Fight ClubÂ	315	2968
FuryÂ	406	701
Interview with the Vampire: The Vampire ChroniclesÂ	120	406
Killing Them SoftlyÂ	414	369
Mr. & Mrs. SmithÂ	233	798
Ocean's ElevenÂ	186	845
Ocean's TwelveÂ	198	627
Seven Years in TibetÂ	76	119
Sinbad: Legend of the Seven SeasÂ	98	91
Spy GameÂ	142	361
The Assassination of Jesse James by the Coward Robert FordÂ	273	415
The Curious Case of Benjamin ButtonÂ	362	822
The Tree of LifeÂ	584	975
TroyÂ	220	1694
True RomanceÂ	122	460
Leonardo DiCaprio	6934	19204
Blood DiamondÂ	166	657
Body of LiesÂ	238	263
Catch Me If You CanÂ	194	667



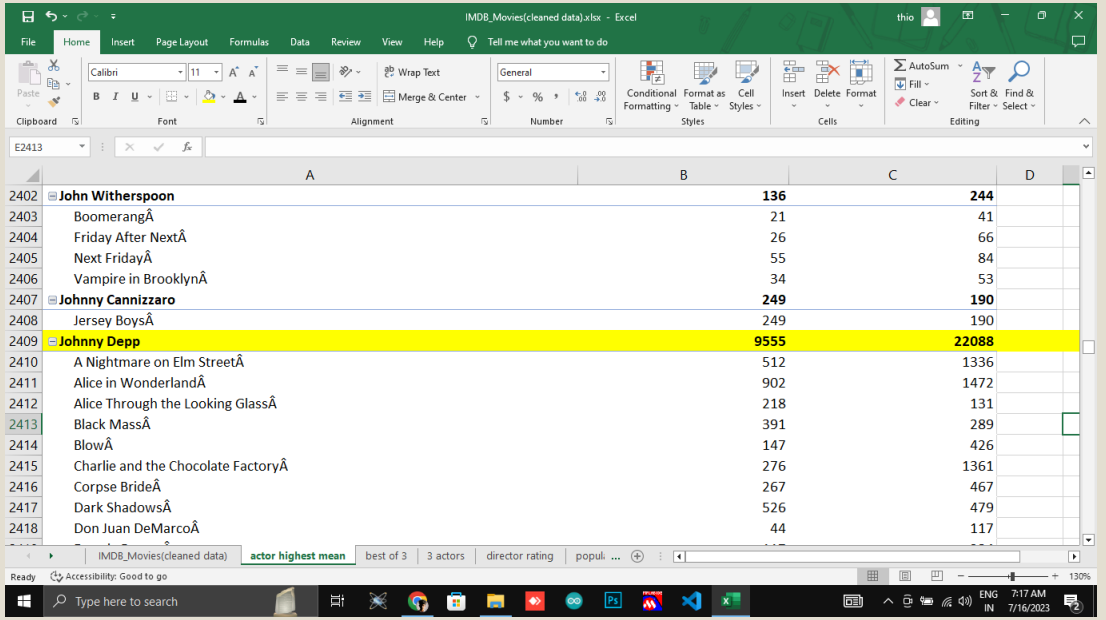
33	The BeachÂ	118	548
34	The DepartedÂ	352	2054
35	The Great GatsbyÂ	980	1506
36	The Man in the Iron MaskÂ	83	244
37	The Quick and the DeadÂ	63	216
38	The RevenantÂ	556	1188
39	The Wolf of Wall StreetÂ	606	1138
40	TitanicÂ	315	2528
41	**Meryl Streep**	**1996**	**3269**
42	A Prairie Home CompanionÂ	211	280
43	Hope SpringsÂ	234	178
44	It's ComplicatedÂ	187	214
45	Julie & JuliaÂ	252	277
46	Lions for LambsÂ	227	298
47	One True ThingÂ	64	112
48	Out of AfricaÂ	66	200
49	The Devil Wears PradaÂ	208	631
50	The HoursÂ	174	660
51	The Iron LadyÂ	331	350
52	The River WildÂ	42	69
53	**Grand Total**	**13095**	**35093**



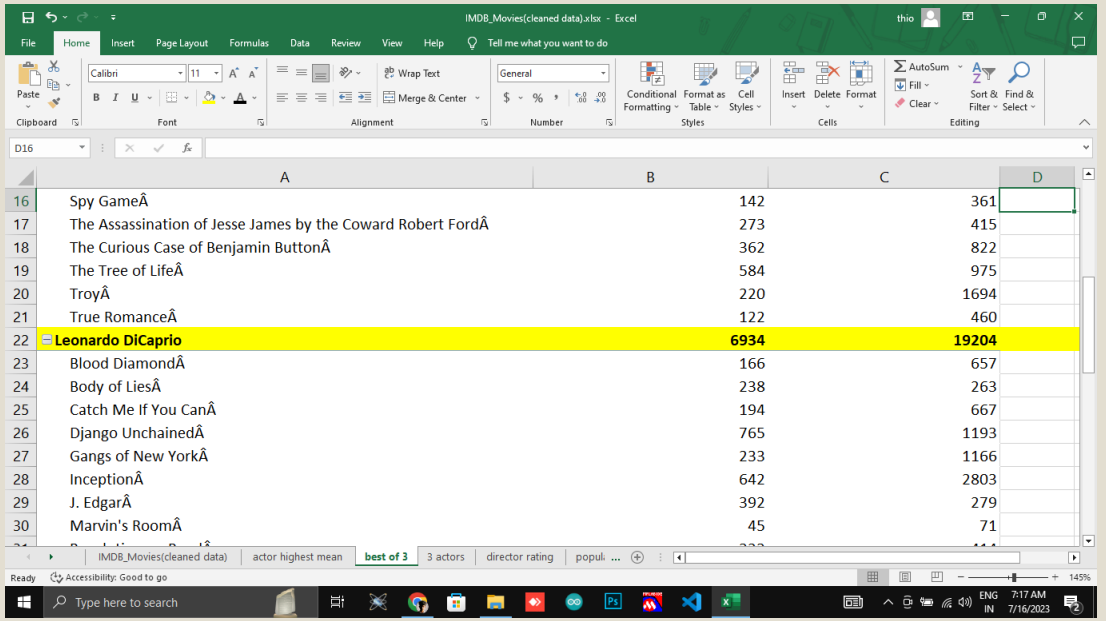
24	Django UnchainedÂ	765	1193
25	Gangs of New YorkÂ	233	1166
26	InceptionÂ	642	2803
27	J. EdgarÂ	392	279
28	Marvin's RoomÂ	45	71
29	Revolutionary RoadÂ	323	414
30	Romeo + JulietÂ	106	506
31	Shutter IslandÂ	490	964
32	The AviatorÂ	267	799
33	The BeachÂ	118	548
34	The DepartedÂ	352	2054
35	The Great GatsbyÂ	980	1506
36	The Man in the Iron MaskÂ	83	244
37	The Quick and the DeadÂ	63	216
38	The RevenantÂ	556	1188
39	The Wolf of Wall StreetÂ	606	1138
40	TitanicÂ	315	2528
41	**Meryl Streep**	**1996**	**3269**
42	A Prairie Home CompanionÂ	211	280
43	Hope SpringsÂ	234	178
44	It's ComplicatedÂ	187	214
45	Julie & JuliaÂ	252	277
46	Lions for LambsÂ	227	298

F . Task(4):

- Favorite actor among these 3 actors is “Leonardo DiCaprio”.
- Favorite actor among all the actors is “Johnny Depp”.



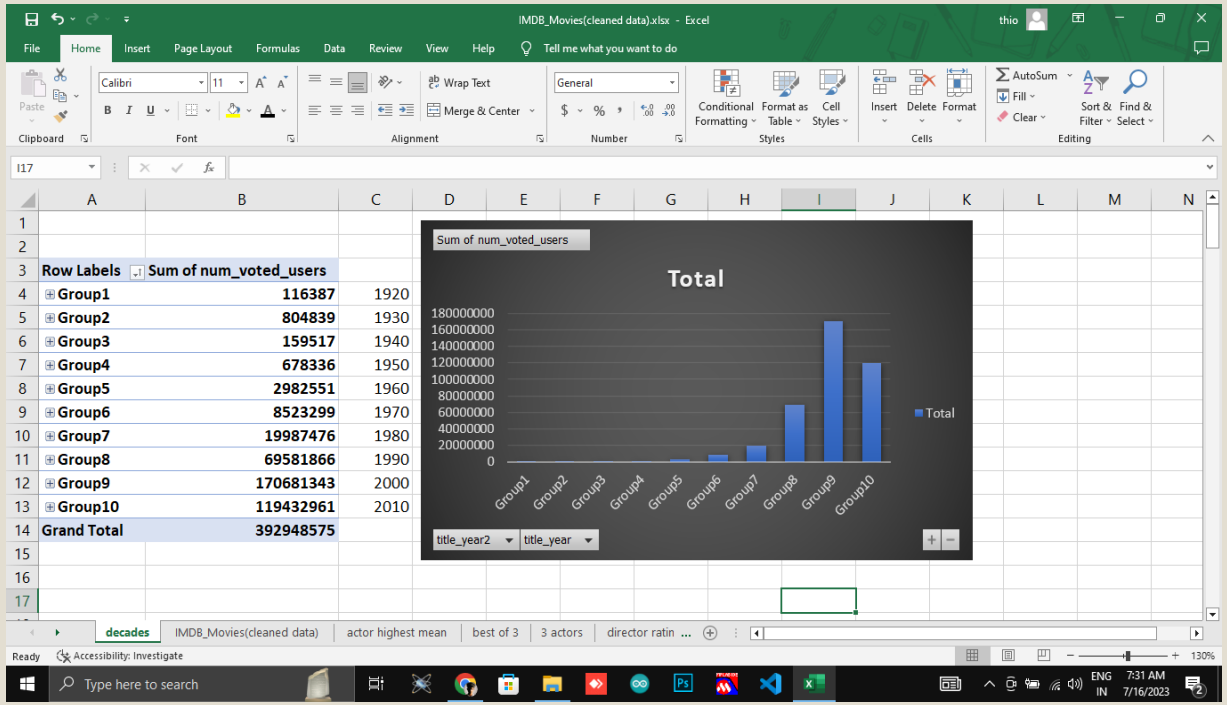
	A	B	C	D
2402	**John Witherspoon**	**136**	**244**	
2403	BoomerangÂ	21	41	
2404	Friday After NextÂ	26	66	
2405	Next FridayÂ	55	84	
2406	Vampire in BrooklynÂ	34	53	
2407	**Johnny Cannizzaro**	**249**	**190**	
2408	Jersey BoysÂ	249	190	
2409	**Johnny Depp**	**9555**	**22088**	
2410	A Nightmare on Elm StreetÂ	512	1336	
2411	Alice in WonderlandÂ	902	1472	
2412	Alice Through the Looking GlassÂ	218	131	
2413	Black MassÂ	391	289	
2414	BlowÂ	147	426	
2415	Charlie and the Chocolate FactoryÂ	276	1361	
2416	Corpse BrideÂ	267	467	
2417	Dark ShadowsÂ	526	479	
2418	Don Juan DeMarcoÂ	44	117	



	A	B	C	D
16	Spy GameÂ	142	361	
17	The Assassination of Jesse James by the Coward Robert FordÂ	273	415	
18	The Curious Case of Benjamin ButtonÂ	362	822	
19	The Tree of LifeÂ	584	975	
20	TroyÂ	220	1694	
21	True RomanceÂ	122	460	
22	**Leonardo DiCaprio**	**6934**	**19204**	
23	Blood DiamondÂ	166	657	
24	Body of LiesÂ	238	263	
25	Catch Me If You CanÂ	194	667	
26	Django UnchainedÂ	765	1193	
27	Gangs of New YorkÂ	233	1166	
28	InceptionÂ	642	2803	
29	J. EdgarÂ	392	279	
30	Marvin's RoomÂ	45	71	

F . Task(5):

- 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. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade.



Row Labels	Sum of num_voted_users	
Group1	116387	1920
Group2	804839	1930
Group3	159517	1940
Group4	678336	1950
Group5	2982551	1960
Group6	8523299	1970
Group7	19987476	1980
Group8	69581866	1990
Group9	170681343	2000
Group10	119432961	2010
Grand Total	392948575	

Insights:

- Cleaning data is a very important in data analysis.
- Avatar is the highest IMDB rating movie.
- Shawshank redemption is the highest IMDB rating movie in English language.
- Highest IMDB score Director is Akira kurosawa.
- Most Popular genre is Drama.
- Each decade the voters increased.
- The critic and audience – favorite actor is Johnny Depp.
- Profit – Budget Analysis is having one outlier.

Result:

- From this project I Understood cleaning and modifying data is very important.
- Removing unnecessary data i.e. columns, blank spaces or rows, etc. makes the data much more easier to work nad readable.
- Outliers identification and removal changes the results of data analysis.

THANK YOU

Excel file link:
https://docs.google.com/spreadsheets/d/17rZcPuOSyUqxho3ROU6qtMh3kLuuz4z-/edit?usp=drive_link&ouid=101181995113560202220&rtpof=true&sd=true