IMDB MOVIE ANALYSIS

Project description: This project is based upon the analysis of imdb movies,

Where a dataset has been provided which contains 29 columns and 5044 rows.Well it’s a big dataset so,in the dataset we may get noisey data or unclean data or missing value or any errors present so first step is to we should clean the data then we should find patterns and analyse them.

Cleaning data: In the data I got so many missing values and noisy data,where I got rid of them by using some cleaning tasks. After then using statistics I analysed which column is useful which column is relatable by using correlations and I removed some of the columns which are not useful for analysing and which are not interrelated to each other. After cleaning the data my data got reduced to 15 columns and 3837 rows which made me felt more easy to analyse the data. Now we can say my data is clean data or good data.

Tech-stack used: Ms-Execl

Approach:

After completion of cleaning I started performing the tasks which are:

A. Cleaning the data

B. Movies with the highest profit

C. Top 250

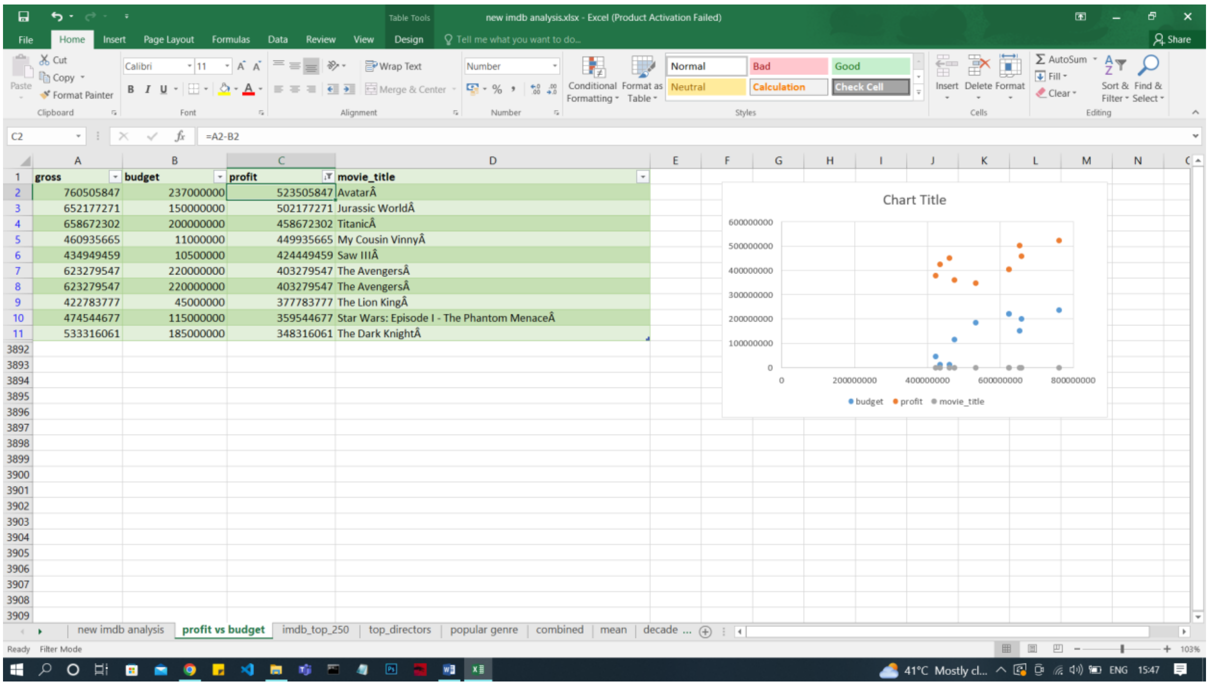
D. Best Directors

E. Popular Genres

F. Charts

As above I said I have completed the cleaning task. Then the next step is to find the movies with the highest profit.

High profit: first I created a new column called profit. Which was difference between the two columns of gross and budget. Then after I plotted a scatter plot chart of profit vs budget and observed which movie got the highest profit. The movie with the highest profit was Avatar followed by Jurassic park and titanic.

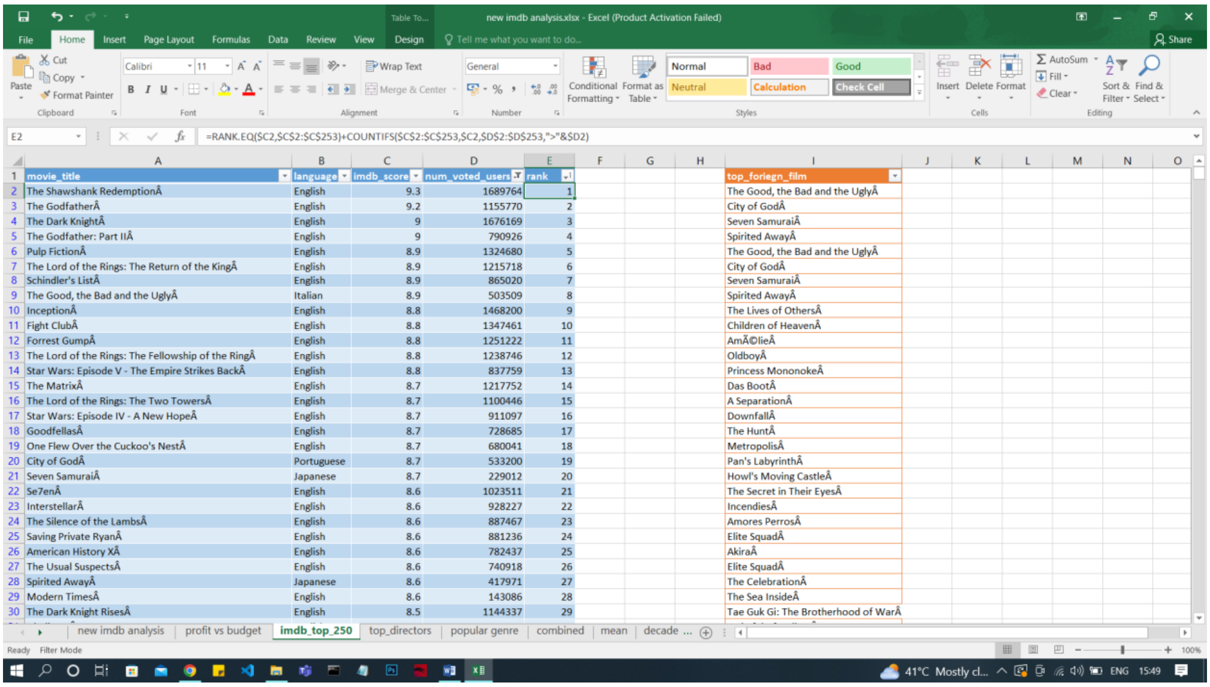


new imdb analysis.xlsx - Excel (Product Activation Failed)

C2 =A2-B2

	gross	budget	profit	movie_title
2	760505847	237000000	523505847	AvatarÂ
3	652177271	150000000	502177271	Jurassic WorldÂ
4	658672302	200000000	458672302	TitanicÂ
5	460935665	11000000	449935665	My Cousin VinnyÂ
6	434949459	10500000	424449459	Saw IIIÂ
7	623279547	220000000	403279547	The AvengersÂ
8	623279547	220000000	403279547	The AvengersÂ
9	422783777	45000000	377783777	The Lion KingÂ
10	474544677	115000000	359544677	Star Wars: Episode I - The Phantom MenaceÂ
11	533316061	185000000	348316061	The Dark KnightÂ

Top 250: In this task I have created a table consisting of movie name, language, imdb_score, num_voted_users, and created a column called rank where I ranked all the movies from 1 to 250. And I made sure that the num_voted_users are greater than 25000. After extracting all the 250 movies I also extracted the movies from top 250 movies which are not English language and stored them in a table called top_foreign_language.



new imdb analysis.xlsx - Excel (Product Activation Failed)

E2 =RANK.EQ($C2,$C$2:$C$253)+COUNTIFS($C$2:$C$253,$C2,$D$2:$D$253,">"&$D2)

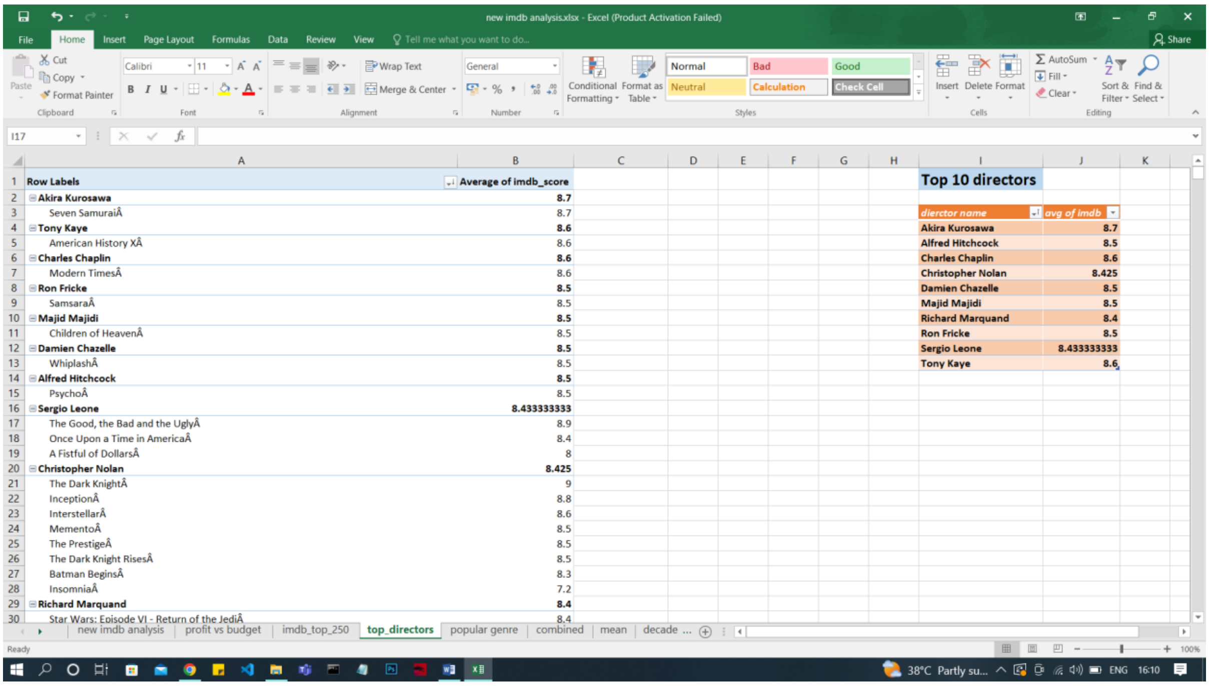
	movie_title	language	imdb_score	num_voted_users	rank
2	The Shawshank RedemptionÂ	English	9.3	1689764	1
3	The GodfatherÂ	English	9.2	1155770	2
4	The Dark KnightÂ	English	9	1676169	3
5	The Godfather: Part IIÂ	English	9	790926	4
6	Pulp FictionÂ	English	8.9	1324680	5
7	The Lord of the Rings: The Return of the KingÂ	English	8.9	1215718	6
8	Schindler's ListÂ	English	8.9	865020	7
9	The Good, the Bad and the UglyÂ	Italian	8.9	503509	8
10	InceptionÂ	English	8.8	1468200	9
11	Fight ClubÂ	English	8.8	1347461	10
12	Forrest GumpÂ	English	8.8	1251222	11
13	The Lord of the Rings: The Fellowship of the RingÂ	English	8.8	1238746	12
14	Star Wars: Episode V - The Empire Strikes BackÂ	English	8.8	837759	13
15	The MatrixÂ	English	8.7	1217752	14
16	The Lord of the Rings: The Two TowersÂ	English	8.7	1100446	15
17	Star Wars: Episode IV - A New HopeÂ	English	8.7	911097	16
18	GoodfellasÂ	English	8.7	728685	17
19	One Flew Over the Cuckoo's NestÂ	English	8.7	680041	18
20	City of GodÂ	Portuguese	8.7	533200	19
21	Seven SamuraiÂ	Japanese	8.7	229012	20
22	Se7enÂ	English	8.6	1023511	21
23	InterstellarÂ	English	8.6	928227	22
24	The Silence of the LambsÂ	English	8.6	887467	23
25	Saving Private RyanÂ	English	8.6	881236	24
26	American History XÂ	English	8.6	782437	25
27	The Usual SuspectsÂ	English	8.6	740918	26
28	Spirited AwayÂ	Japanese	8.6	417971	27
29	Modern TimesÂ	English	8.6	143086	28
30	The Dark Knight RisesÂ	English	8.5	1144337	29

top_foriegn_film
The Good, the Bad and the UglyÂ
City of GodÂ
Seven SamuraiÂ
Spirited AwayÂ
The Good, the Bad and the UglyÂ
City of GodÂ
Seven SamuraiÂ
Spirited AwayÂ
The Lives of OthersÂ
Children of HeavenÂ
AmÃ©lieÂ
OldboyÂ
Princess MononokeÂ
Das BootÂ
A SeparationÂ
DownfallÂ
The HuntÂ
MetropolisÂ
Pan's LabyrinthÂ
Howl's Moving CastleÂ
The Secret in Their EyesÂ
IncendiesÂ
Amores PerrosÂ
Elite SquadÂ
AkiraÂ
Elite SquadÂ
The CelebrationÂ
The Sea InsideÂ
Tae Guk Gi: The Brotherhood of WarÂ

To calculate rank I have used the formula:
=RANK.EQ($C2,$C$2:$C$253)+COUNTIFS($C$2:$C$253,$C2,$D$2:$D$253,">"&

$D2). The top 1 movie is The Shawshank Redemption with the imdb score of 9.3.

Best Directors: In this task I have retrived the director name column,their movies and imdb_score and after I created a pivot table and then created a pivot table and then after I calculated mean of the imdb_score and then sorted them in alphabetical order.



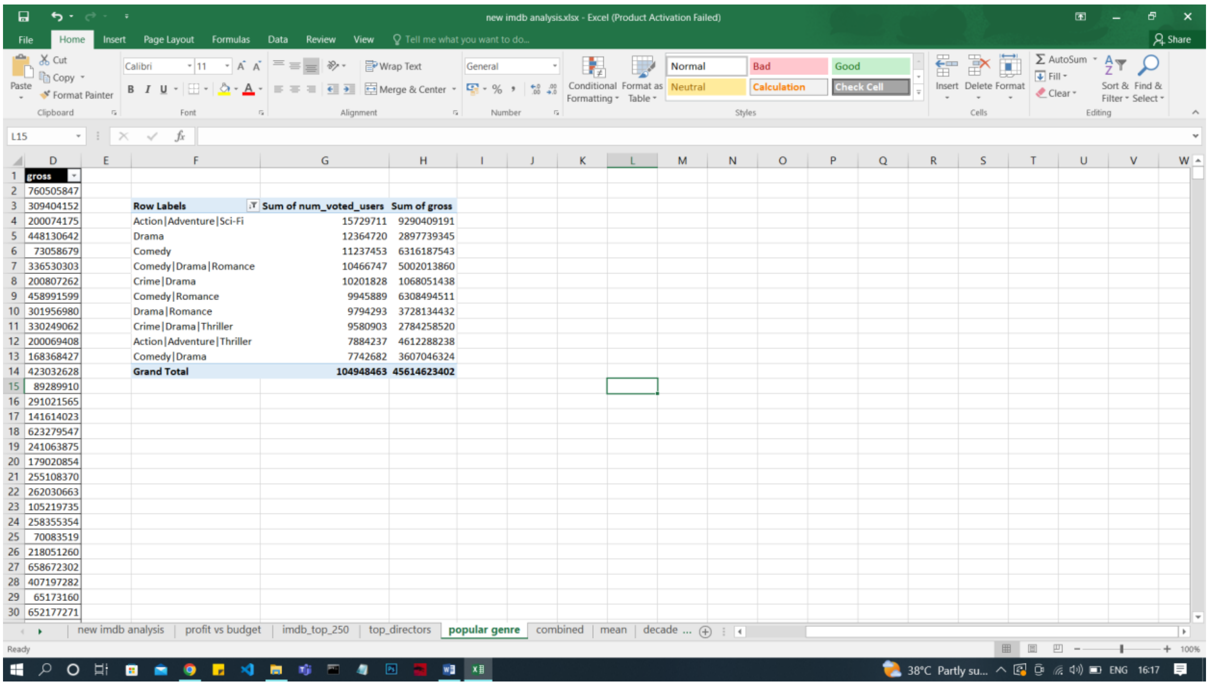
Row Labels	Average of imdb_score
Akira Kurosawa	**8.7**
Seven SamuraiÂ	8.7
Tony Kaye	**8.6**
American History XÂ	8.6
Charles Chaplin	**8.6**
Modern TimesÂ	8.6
Ron Fricke	**8.5**
SamsaraÂ	8.5
Majid Majidi	**8.5**
Children of HeavenÂ	8.5
Damien Chazelle	**8.5**
WhiplashÂ	8.5
Alfred Hitchcock	**8.5**
PsychoÂ	8.5
Sergio Leone	**8.433333333**
The Good, the Bad and the UglyÂ	8.9
Once Upon a Time in AmericaÂ	8.4
A Fistful of DollarsÂ	8
Christopher Nolan	**8.425**
The Dark KnightÂ	9
InceptionÂ	8.8
InterstellarÂ	8.6
MementoÂ	8.5
The PrestigeÂ	8.5
The Dark Knight RisesÂ	8.5
Batman BeginsÂ	8.3
InsomniaÂ	7.2
Richard Marquand	**8.4**
Star Wars: Episode VI - Return of the JediÂ	8.4

Top 10 directors

dierctor name	avg of imdb
Akira Kurosawa	8.7
Alfred Hitchcock	8.5
Charles Chaplin	8.6
Christopher Nolan	8.425
Damien Chazelle	8.5
Majid Majidi	8.5
Richard Marquand	8.4
Ron Fricke	8.5
Sergio Leone	8.433333333
Tony Kaye	8.6

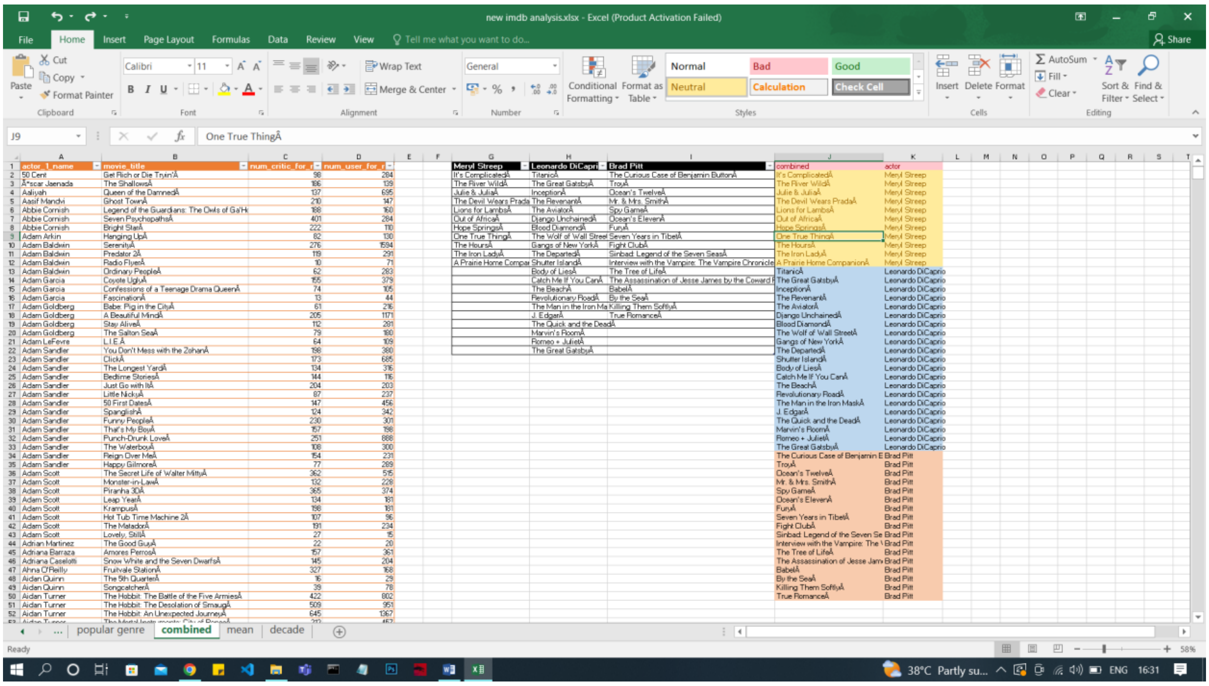
The top director is Akira Kurosawa of 8.7 score of imdb.

Popular Genres: In this task I created a new table which consists director name,genres,gross,num_voted_users and then created a pivot table in the existing worksheet of Genres and sum_of_voted_users and sum_of_gross. According to the data we can see that the most watched genres is action adventure and scifi.



	D	E	F	G	H
1	gross				
2	760505847				
3	309404152		Row Labels	Sum of num_voted_users	Sum of gross
4	200074175		Action|Adventure|Sci-Fi	15729711	9290409191
5	448130642		Drama	12364720	2897739345
6	73058679		Comedy	11237453	6316187543
7	336530303		Comedy|Drama|Romance	10466747	5002013860
8	200807262		Crime|Drama	10201828	1068051438
9	458991599		Comedy|Romance	9945889	6308494511
10	301956980		Drama|Romance	9794293	3728134432
11	330249062		Crime|Drama|Thriller	9580903	2784258520
12	200069408		Action|Adventure|Thriller	7884237	4612288238
13	168368427		Comedy|Drama	7742682	3607046324
14	423032628		Grand Total	104948463	45614623402
15	89289910				
16	291021565				
17	141614023				
18	623279547				
19	241063875				
20	179020854				
21	255108370				
22	262030663				
23	105219735				
24	258355354				
25	70083519				
26	218051260				
27	658672302				
28	407197282				
29	65173160				
30	652177271				

Charts: In this task I have created three new columns called Meryl_streep,Leonardo_Di_Caprio and brad_pitt by retrieving from the table I created with the actor name,movie title ,num_critic_for_review and num_user_from_review and then I sorted the particular movies to particular actors and after that I created another column called combined where I combined the actors and their movies by group by and then calculated mean of num_user_for_reviews and num_critic_for_review by using pivot table and then analysed who is the critic favourite and audience favourite and created a pie chart. The favourite actor is Leonardo di caprio with 47% and also has the highest mean .

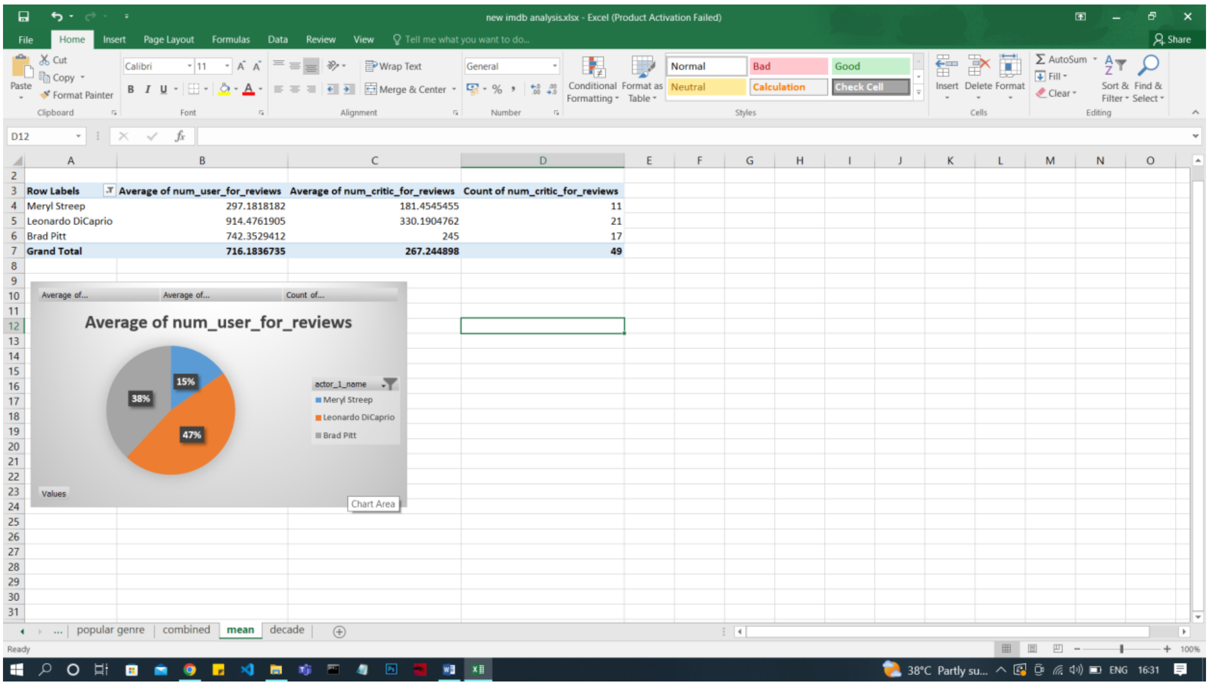


new imdb analysis.xlsx - Excel (Product Activation Failed)

J9 | One True ThingÂ

	A	B	C	D	G	H	I	J	K
1	actor_1_name	movie_title	num_critic_for_r	num_user_for_r	Meryl Streep	Leonardo DiCapri	Brad Pitt	combined	actor
2	50 Cent	Get Rich or Die Tryin'Â	98	284	It's ComplicatedÂ	TitanicÂ	The Curious Case of Benjamin ButtonÂ	It's ComplicatedÂ	Meryl Streep
3	Â*scar Jaenada	The ShallowsÂ	186	139	The River WildÂ	The Great GatsbyÂ	TroyÂ	The River WildÂ	Meryl Streep
4	Aaliyah	Queen of the DamnedÂ	137	695	Julie & JuliaÂ	InceptionÂ	Ocean's TwelveÂ	Julie & JuliaÂ	Meryl Streep
5	Aasif Mandvi	Ghost TownÂ	210	147	The Devil Wears Prada	The RevenantÂ	Mr. & Mrs. SmithÂ	The Devil Wears PradaÂ	Meryl Streep
6	Abbie Cornish	Legend of the Guardians: The Owls of Ga'Hc	188	160	Lions for LambsÂ	The AviatorÂ	Spy GameÂ	Lions for LambsÂ	Meryl Streep
7	Abbie Cornish	Seven PsychopathsÂ	401	284	Out of AfricaÂ	Django UnchainedÂ	Ocean's ElevenÂ	Out of AfricaÂ	Meryl Streep
8	Abbie Cornish	Bright StarÂ	222	110	Hope SpringsÂ	Blood DiamondÂ	FuryÂ	Hope SpringsÂ	Meryl Streep
9	Adam Arkin	Hanging UpÂ	82	130	One True ThingÂ	The Wolf of Wall Street	Seven Years in TibetÂ	One True ThingÂ	Meryl Streep
10	Adam Baldwin	SerenityÂ	276	1594	The HoursÂ	Gangs of New YorkÂ	Fight ClubÂ	The HoursÂ	Meryl Streep
11	Adam Baldwin	Predator 2Â	119	291	The Iron LadyÂ	The DepartedÂ	Sinbad: Legend of the Seven SeasÂ	The Iron LadyÂ	Meryl Streep
12	Adam Baldwin	Radio FlyerÂ	10	71	A Prairie Home Compar	Shutter IslandÂ	Interview with the Vampire: The Vampire Chronicle	A Prairie Home CompanionÂ	Meryl Streep
13	Adam Baldwin	Ordinary PeopleÂ	62	283		Body of LiesÂ	The Tree of LifeÂ	TitanicÂ	Leonardo DiCaprio
14	Adam Garcia	Coyote UglyÂ	155	379		Catch Me If You CanÂ	The Assassination of Jesse James by the Coward F	The Great GatsbyÂ	Leonardo DiCaprio
15	Adam Garcia	Confessions of a Teenage Drama QueenÂ	74	105		The BeachÂ	BabelÂ	InceptionÂ	Leonardo DiCaprio
16	Adam Garcia	FascinationÂ	13	44		Revolutionary RoadÂ	By the SeaÂ	The RevenantÂ	Leonardo DiCaprio
17	Adam Goldberg	Babe: Pig in the CityÂ	61	216		The Man in the Iron Ma	Killing Them SoftlyÂ	The AviatorÂ	Leonardo DiCaprio
18	Adam Goldberg	A Beautiful MindÂ	205	1171		J. EdgarÂ	True RomanceÂ	Django UnchainedÂ	Leonardo DiCaprio
19	Adam Goldberg	Stay AliveÂ	112	281		The Quick and the DeadÂ		Blood DiamondÂ	Leonardo DiCaprio
20	Adam Goldberg	The Salton SeaÂ	79	180		Marvin's RoomÂ		The Wolf of Wall StreetÂ	Leonardo DiCaprio
21	Adam LeFevre	L.I.E.Â	64	109		Romeo + JulietÂ		Gangs of New YorkÂ	Leonardo DiCaprio
22	Adam Sandler	You Don't Mess with the ZohanÂ	198	380		The Great GatsbyÂ		The DepartedÂ	Leonardo DiCaprio
23	Adam Sandler	ClickÂ	173	685				Shutter IslandÂ	Leonardo DiCaprio
24	Adam Sandler	The Longest YardÂ	134	316				Body of LiesÂ	Leonardo DiCaprio
25	Adam Sandler	Bedtime StoriesÂ	144	116				Catch Me If You CanÂ	Leonardo DiCaprio
26	Adam Sandler	Just Go with ItÂ	204	203				The BeachÂ	Leonardo DiCaprio
27	Adam Sandler	Little NickyÂ	87	237				Revolutionary RoadÂ	Leonardo DiCaprio
28	Adam Sandler	50 First DatesÂ	147	456				The Man in the Iron MaskÂ	Leonardo DiCaprio
29	Adam Sandler	SpanglishÂ	124	342				J. EdgarÂ	Leonardo DiCaprio
30	Adam Sandler	Funny PeopleÂ	230	301				The Quick and the DeadÂ	Leonardo DiCaprio
31	Adam Sandler	That's My BoyÂ	157	198				Marvin's RoomÂ	Leonardo DiCaprio
32	Adam Sandler	Punch-Drunk LoveÂ	251	888				Romeo + JulietÂ	Leonardo DiCaprio
33	Adam Sandler	The WaterboyÂ	108	300				The Great GatsbyÂ	Leonardo DiCaprio
34	Adam Sandler	Reign Over MeÂ	154	231				The Curious Case of Benjamin E	Brad Pitt
35	Adam Sandler	Happy GilmoreÂ	77	289				TroyÂ	Brad Pitt
36	Adam Scott	The Secret Life of Walter MittyÂ	362	515				Ocean's TwelveÂ	Brad Pitt
37	Adam Scott	Monster-in-LawÂ	132	228				Mr. & Mrs. SmithÂ	Brad Pitt
38	Adam Scott	Piranha 3DÂ	365	374				Spy GameÂ	Brad Pitt
39	Adam Scott	Leap YearÂ	134	181				Ocean's ElevenÂ	Brad Pitt
40	Adam Scott	KrampusÂ	198	181				FuryÂ	Brad Pitt
41	Adam Scott	Hot Tub Time Machine 2Â	107	96				Seven Years in TibetÂ	Brad Pitt
42	Adam Scott	The MatadorÂ	191	234				Fight ClubÂ	Brad Pitt
43	Adam Scott	Lovely, StillÂ	27	15				Sinbad: Legend of the Seven Se	Brad Pitt
44	Adrian Martinez	The Good GuyÂ	22	20				Interview with the Vampire: The \	Brad Pitt
45	Adriana Barraza	Amores PerrosÂ	157	361				The Tree of LifeÂ	Brad Pitt
46	Adriana Caselotti	Snow White and the Seven DwarfsÂ	145	204				The Assassination of Jesse Jam	Brad Pitt
47	Ahna O'Reilly	Fruitvale StationÂ	327	168				BabelÂ	Brad Pitt
48	Aidan Quinn	The 5th QuarterÂ	16	29				By the SeaÂ	Brad Pitt
49	Aidan Quinn	SongcatcherÂ	39	78				Killing Them SoftlyÂ	Brad Pitt
50	Aidan Turner	The Hobbit: The Battle of the Five ArmiesÂ	422	802				True RomanceÂ	Brad Pitt
51	Aidan Turner	The Hobbit: The Desolation of SmaugÂ	509	951					
52	Aidan Turner	The Hobbit: An Unexpected JourneyÂ	645	1367					

popular genre | combined | mean | decade



new imdb analysis.xlsx - Excel (Product Activation Failed)

D12

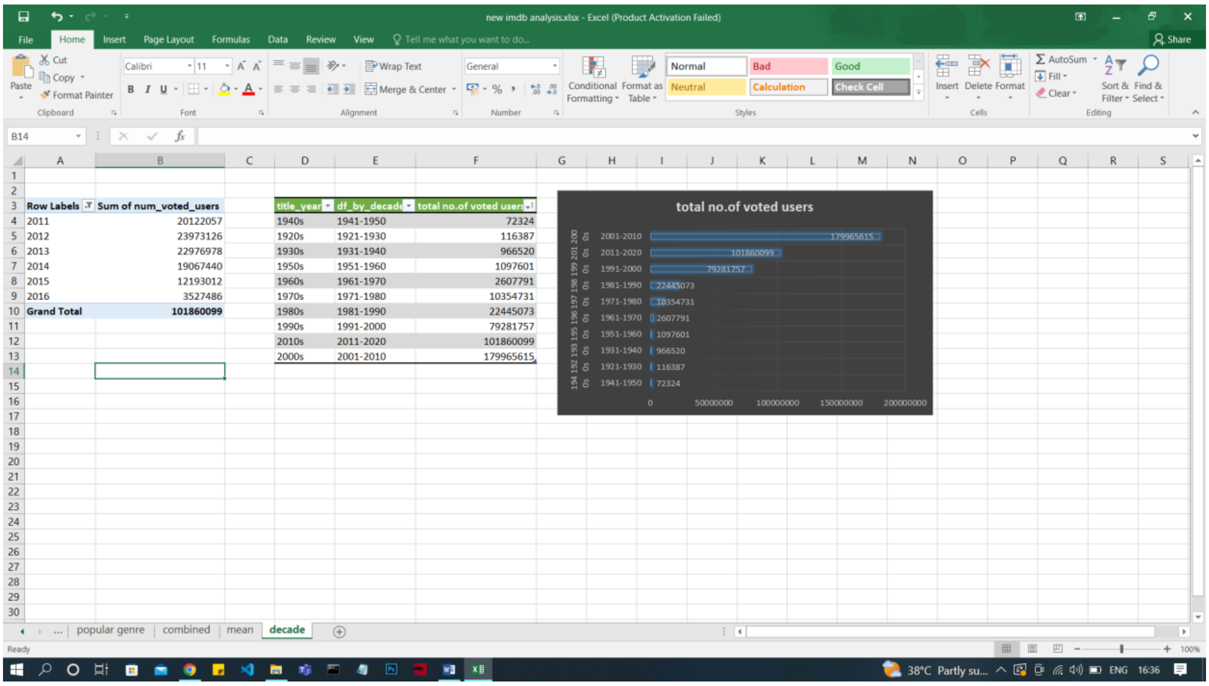
Row Labels	Average of num_user_for_reviews	Average of num_critic_for_reviews	Count of num_critic_for_reviews
Meryl Streep	297.1818182	181.4545455	11
Leonardo DiCaprio	914.4761905	330.1904762	21
Brad Pitt	742.3529412	245	17
Grand Total	716.1836735	267.244898	49

popular genre | combined | mean | decade

After that I have created pivot table for years and sum of num_voted_users

And then I created a timeline For example, the title_year year 1923, 1925 should be stored as 1920s. which I named the column called decade and sorted it and calculated sum of users voted in each decade and created a bar chart.

From 2001-2010 has the highest total no.of voted users of 179965615.



Row Labels	Sum of num_voted_users
2011	20122057
2012	23973126
2013	22976978
2014	19067440
2015	12193012
2016	3527486
Grand Total	**101860099**

title_year	df_by_decade	total no.of voted users
1940s	1941-1950	72324
1920s	1921-1930	116387
1930s	1931-1940	966520
1950s	1951-1960	1097601
1960s	1961-1970	2607791
1970s	1971-1980	10354731
1980s	1981-1990	22445073
1990s	1991-2000	79281757
2010s	2011-2020	101860099
2000s	2001-2010	179965615