IMDB Movie Analysis

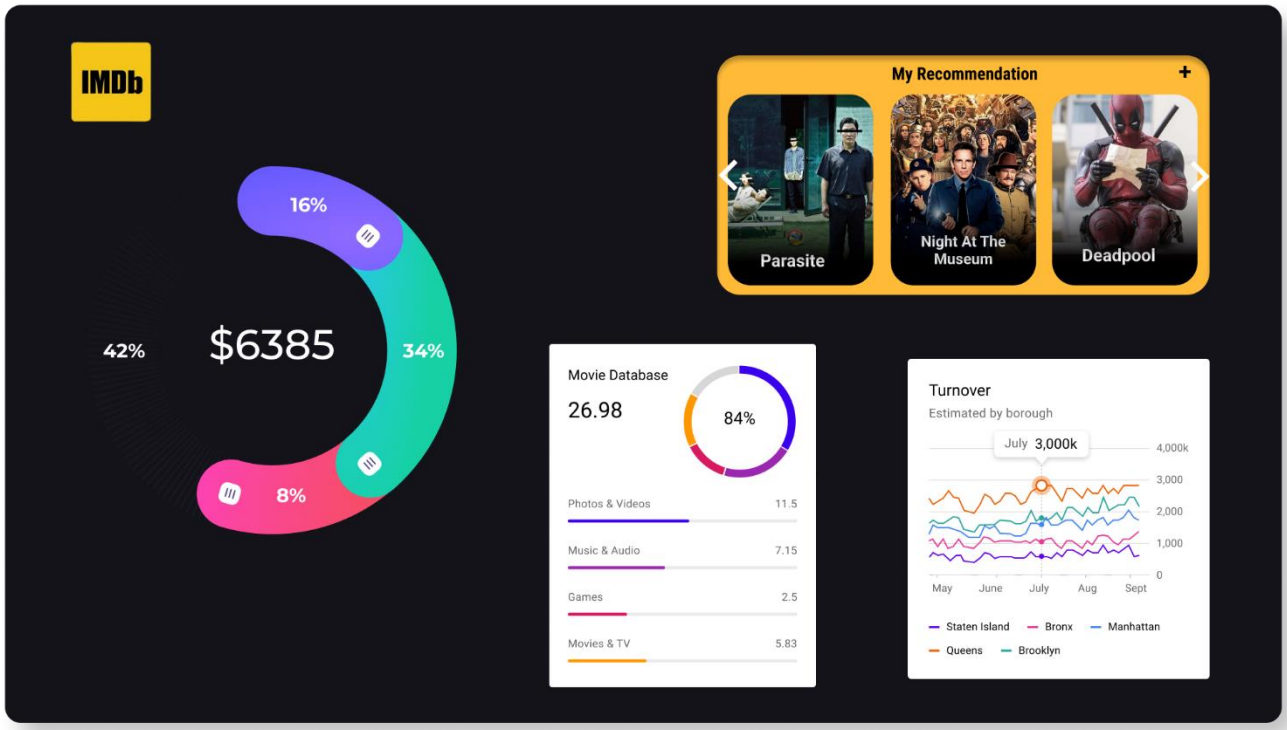
AUGUST 11

By Aastha Kumar

Description

I have been provided with a dataset having various columns of different IMDB Movies and am required to Frame the problem. For this task, I will need to define a problem I want to shed some light on.

I can do this by asking 'What?' This is where I frame the problem i.e. What is the problem? Once I have defined a problem I will use my Data Analysis skills to explore the data set and derive insights.



Approach

First, I downloaded the dataset from Google Sheets onto my personal device for making modifications. I then used my knowledge in statistics and used different formulas in excel to draw necessary conclusions about the company.

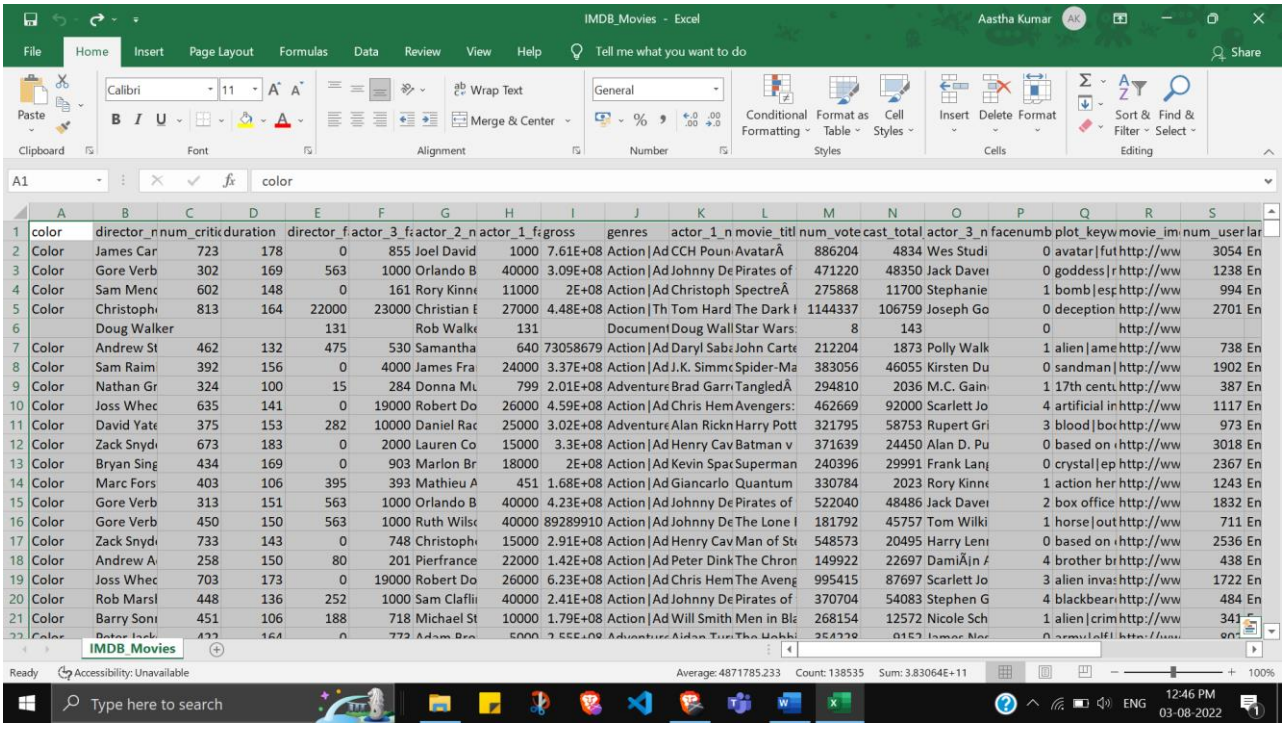
Tech-Stack Used

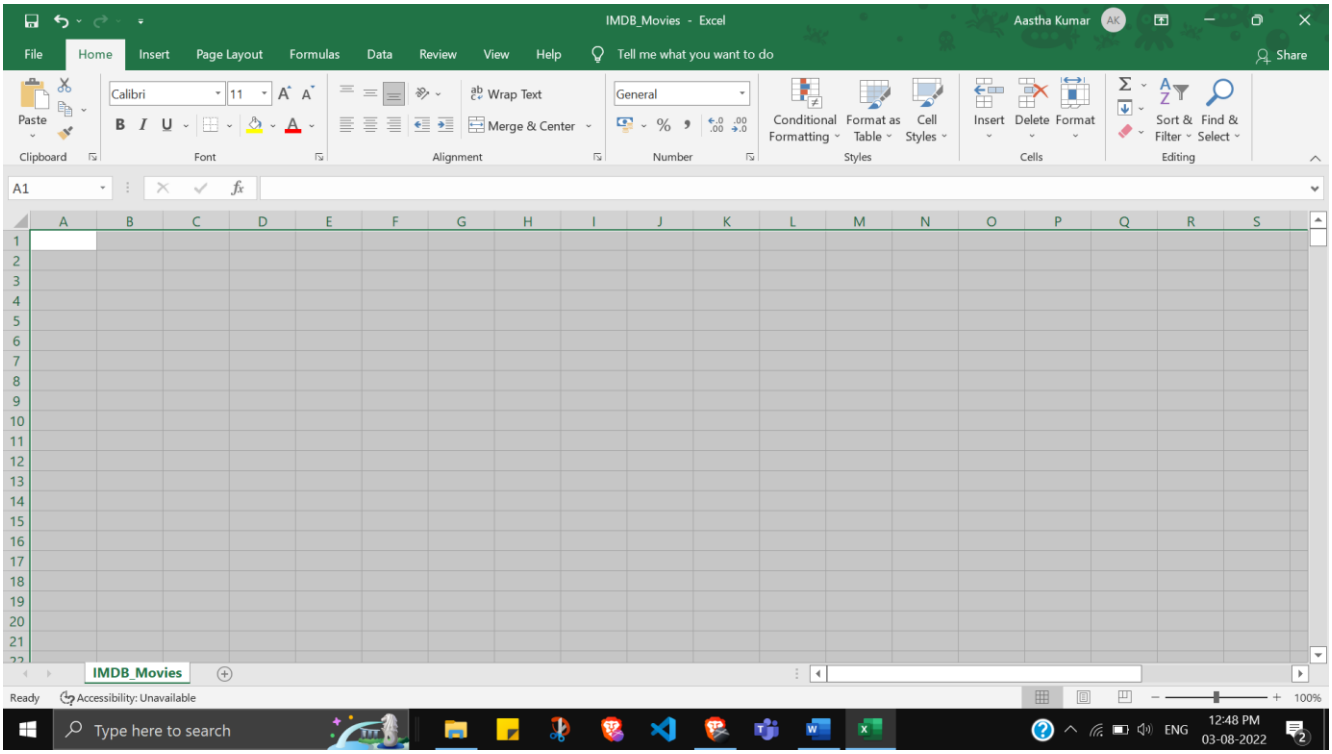
- Google Sheet

Insights

A. Cleaning the data: This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
My Task: Clean the data

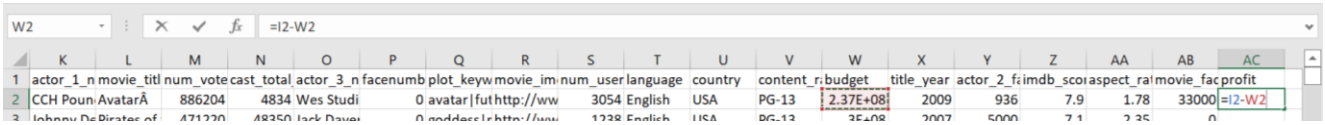
To delete all rows and columns of the spreadsheet, I used the keyboard shortcut ctrl+A to select all the values at once and then hit the delete button.

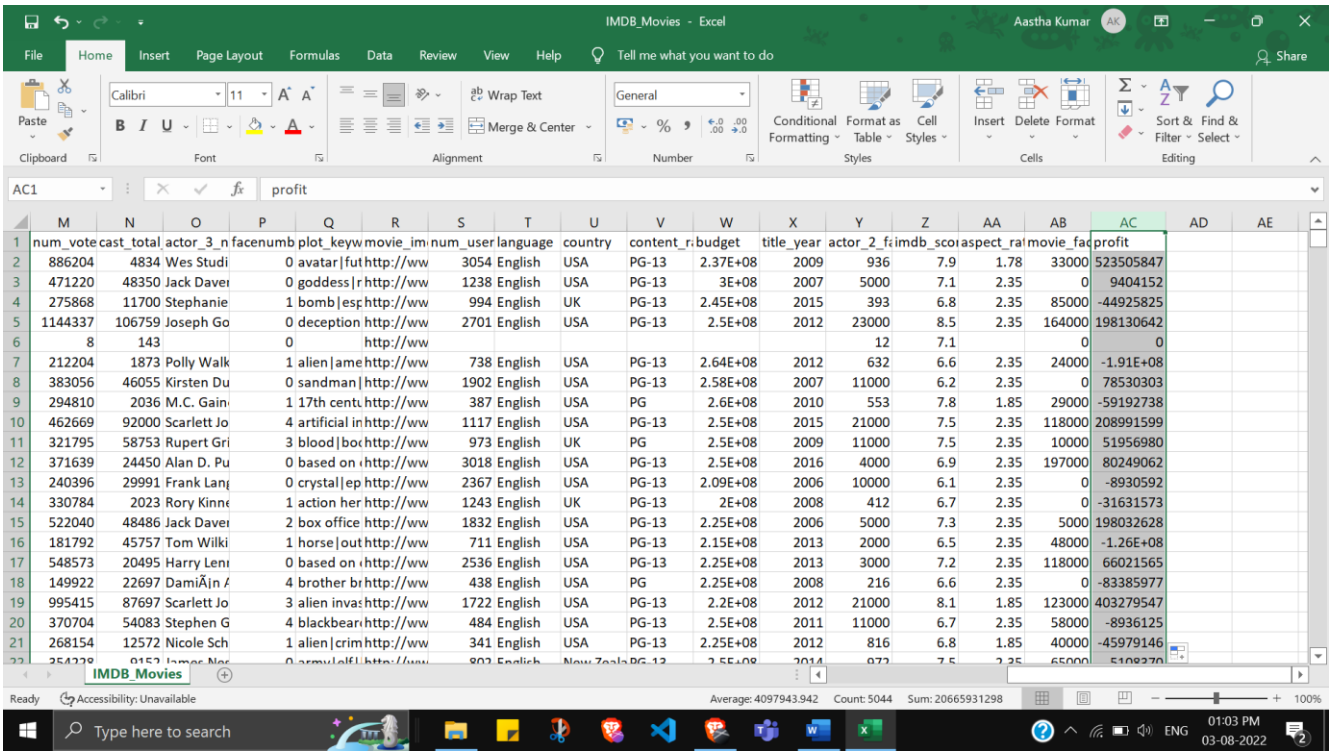



IMDB_Movies - Excel
Aastha Kumar
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Tell me what you want to do
Share
Paste
Calibri
11
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
IMDB_Movies
Ready
Accessibility: Unavailable
100%
Type here to search
ENG
12:48 PM
03-08-2022

B. Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

My Task: Find the movies with the highest profit.



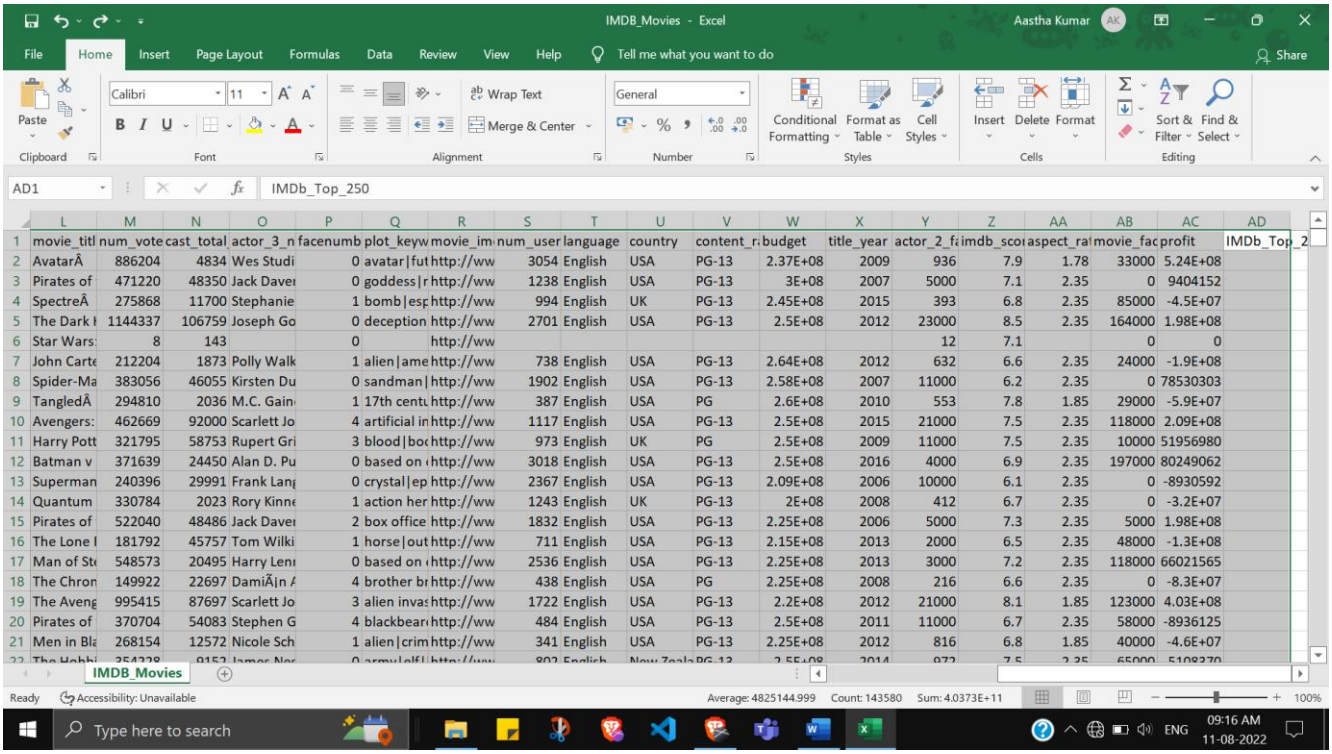


C. Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than

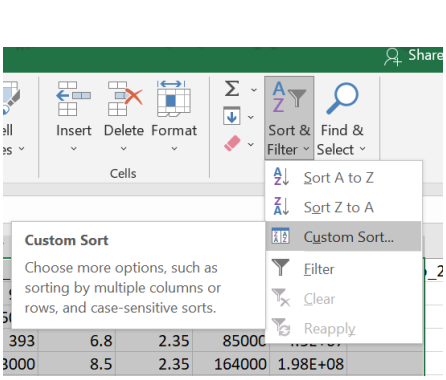
25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

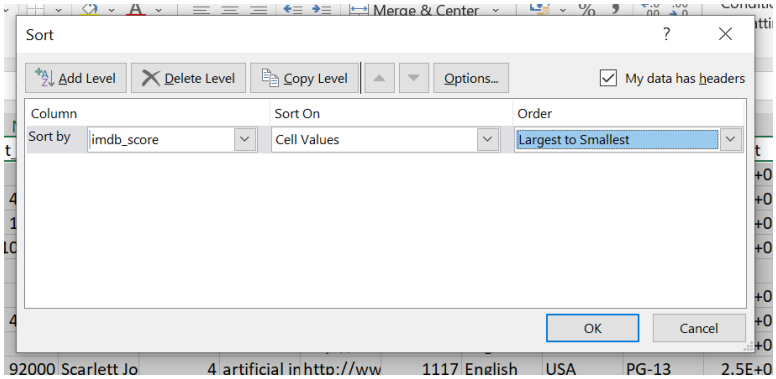
My Task: Find IMDB Top 250

Select the entire spreadsheet by hitting ctrl+A

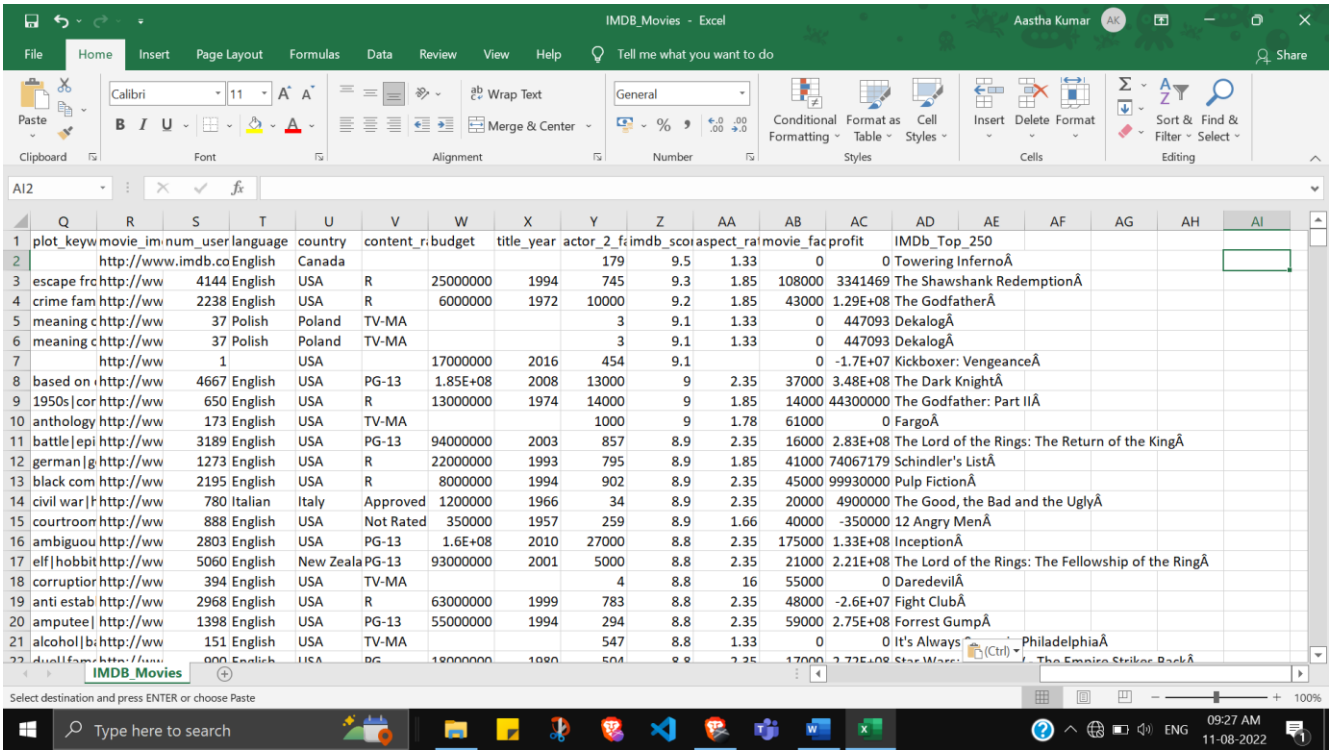


Under Sort and Filter menu, click the custom sort option





Click OK. All the contents of the spreadsheet will be sort according to the highest to lowest imdb_score. Select the top 250 movies corresponding to the highest imdb_scores and paste the values in the new column created (IMDb_Top_250)



	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	plot_keyw	movie_im	num_user	language	country	content_r	budget	title_year	actor_2_f	imdb_sco	aspect_ra	movie_fac	profit	IMDb_Top_250					
2		http://www.imdb.co		English	Canada				179	9.5	1.33	0	0	Towering InfernoÂ					
3	escape fro	http://ww	4144	English	USA	R	25000000	1994	745	9.3	1.85	108000	3341469	The Shawshank RedemptionÂ					
4	crime fam	http://ww	2238	English	USA	R	6000000	1972	10000	9.2	1.85	43000	1.29E+08	The GodfatherÂ					
5	meaning c	http://ww	37	Polish	Poland	TV-MA			3	9.1	1.33	0	447093	DekalogÂ					
6	meaning c	http://ww	37	Polish	Poland	TV-MA			3	9.1	1.33	0	447093	DekalogÂ					
7		http://ww	1		USA		17000000	2016	454	9.1		0	-1.7E+07	Kickboxer: VengeanceÂ					
8	based on	http://ww	4667	English	USA	PG-13	1.85E+08	2008	13000	9	2.35	37000	3.48E+08	The Dark KnightÂ					
9	1950s|cor	http://ww	650	English	USA	R	13000000	1974	14000	9	1.85	14000	44300000	The Godfather: Part IIÂ					
10	anthology	http://ww	173	English	USA	TV-MA			1000	9	1.78	61000	0	FargoÂ					
11	battle|epi	http://ww	3189	English	USA	PG-13	94000000	2003	857	8.9	2.35	16000	2.83E+08	The Lord of the Rings: The Return of the KingÂ					
12	german|g	http://ww	1273	English	USA	R	22000000	1993	795	8.9	1.85	41000	74067179	Schindler's ListÂ					
13	black com	http://ww	2195	English	USA	R	8000000	1994	902	8.9	2.35	45000	99930000	Pulp FictionÂ					
14	civil war|h	http://ww	780	Italian	Italy	Approved	1200000	1966	34	8.9	2.35	20000	4900000	The Good, the Bad and the UglyÂ					
15	courtroon	http://ww	888	English	USA	Not Rated	350000	1957	259	8.9	1.66	40000	-350000	12 Angry MenÂ					
16	ambiguou	http://ww	2803	English	USA	PG-13	1.6E+08	2010	27000	8.8	2.35	175000	1.33E+08	InceptionÂ					
17	elf|hobbit	http://ww	5060	English	New Zeala	PG-13	93000000	2001	5000	8.8	2.35	21000	2.21E+08	The Lord of the Rings: The Fellowship of the RingÂ					
18	corruptior	http://ww	394	English	USA	TV-MA			4	8.8	16	55000	0	DaredevilÂ					
19	anti estab	http://ww	2968	English	USA	R	63000000	1999	783	8.8	2.35	48000	-2.6E+07	Fight ClubÂ					
20	amputee|	http://ww	1398	English	USA	PG-13	55000000	1994	294	8.8	2.35	59000	2.75E+08	Forrest GumpÂ					
21	alcohol|b	http://ww	151	English	USA	TV-MA			547	8.8	1.33	0	0	It's Always ... PhiladelphiaÂ					
22	duel|fam	http://ww	900	English	USA	PG	18000000	1980	504	8.8	2.35	17000	2.72E+08	Star Wars: ... The Empire Strikes BackÂ					

D. Best Directors: TGroup the column using the director_name column.

   Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.

   My Task: Find the best directors

E. Popular Genres: Perform this step using the knowledge gained while performing previous steps.

   My Task: Find popular genres

F. Charts: 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. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

   Append the rows of all these columns and store them in a new column named Combined. Group the combined column using the actor_1_name column. Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean. 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. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on

the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

My Task: Find the critic-favorite and audience-favorite actors

Resources

IMDB_Movies

Result

It would have been impossible or very time consuming for an ordinary human to be able to process such large bits of information. A computer can do it within seconds with the right commands. That's why companies like Instagram hire data analysts to control the waves of data they collect every day, makes sense of it, and then draw conclusions or make predictions. This is the process of turning data into insights, and it's how analysts help businesses put all their data to good use.

The more detailed definition you learned earlier is that data analysis is the collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision-making.

Data analytics can help organizations completely rethink something they do or point them in a totally new direction. For example, maybe data leads them to a new product or unique service, or maybe it helps them find a new way to deliver an incredible customer experience.