IMDB MOVIE ANALYSIS

TOOL USED

MS- EXCEL

Problem Statement:-

- Based on the massive movie information, it would be interesting to understand, what are the important factors that make a movie more successful than others? So, we would like to analyse what kind of movies are more successful, in other words, get higher IMDB score.
- In this project, we take IMDB scores as response variable and focus on operating predictions by analysing the rest of variables in the IMDB movie dataset. The results can help film companies to understand the secret of generating a commercial success movie.

Please Note:-
Since some of the solutions cannot be shown fully because of size and quality constraint of the screenshot therefore the drive link for the excel sheet is pasted here for full solution and reference.
The drive link:-
https://docs.google.com/spreadsheets/d/1LRMMixIPoC2Ol3CnJmAbiqKGQe6dnRTN/edit?usp=sharing&ouid=107932508938240092754&rtpof=true&sd=true

1.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.)

T**ask:** clean the data

Table_name	Rows
original table	5043
cleaned table	4998
duplicate rows	45

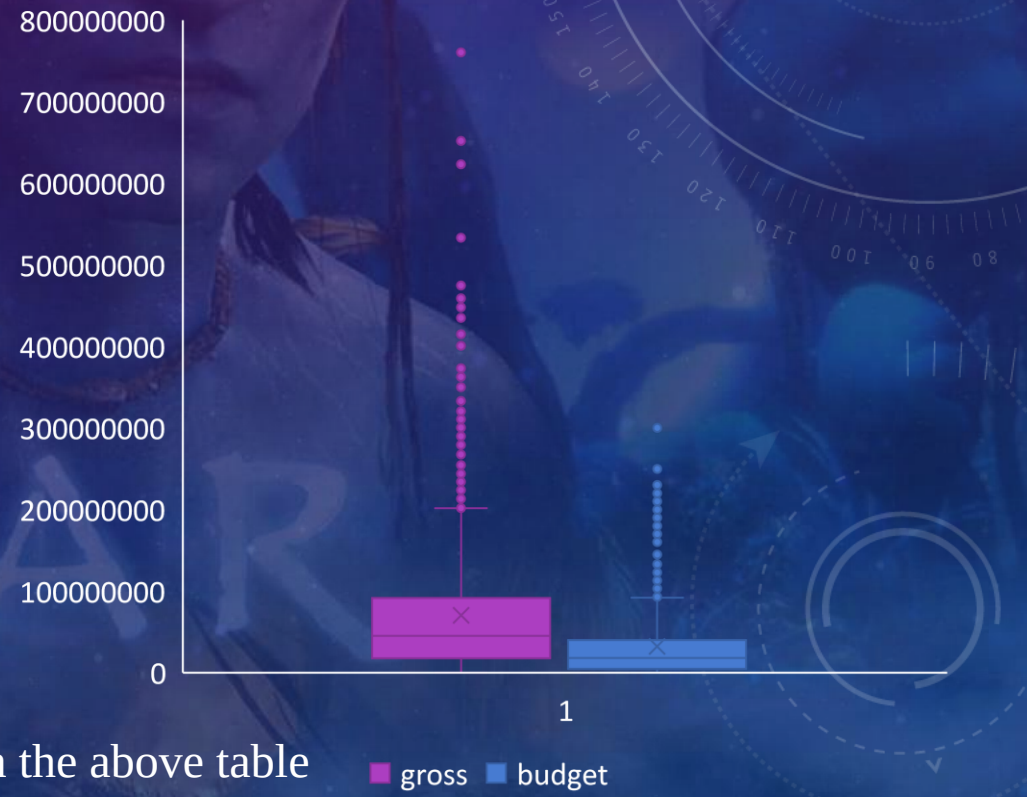
Please Note:-

Under categorical columns the blank cell is replaced by null and under numerical column it is replace by 0 (zero)

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

Movies with highest Profit	profit
Avatar	523505847

Q1	Q3	IQR	UL	LL
5737282	46988319	41251037	108864875	-56139273.5



Box plot shows the outliers which is calculated based on the limits shown in the above table
The points can be referred to the table to get the movie Name which is shown in the next slide where True means outlier for full table follow the sheet.

shows positive outliers

gross	budget	Profit	movie_title	Outliers
760505847	237000000	523505847	Avatar	TRUE
652177271	150000000	502177271	Jurassic World	TRUE
658672302	200000000	458672302	Titanic	TRUE
460935665	11000000	449935665	Star Wars: Episode IV - A New Hope	TRUE
434949459	10500000	424449459	E.T. the Extra-Terrestrial	TRUE
623279547	220000000	403279547	The Avengers	TRUE
422783777	45000000	377783777	The Lion King	TRUE
474544677	115000000	359544677	Star Wars: Episode I - The Phantom Menace	TRUE
533316061	185000000	348316061	The Dark Knight	TRUE
407999255	78000000	329999255	The Hunger Games	TRUE
363024263	58000000	305024263	Deadpool	TRUE
424645577	130000000	294645577	The Hunger Games: Catching Fire	TRUE
356784000	63000000	293784000	Jurassic Park	TRUE
368049635	76000000	292049635	Despicable Me 2	TRUE
350123553	58800000	291323553	American Sniper	TRUE
380838870	94000000	286838870	Finding Nemo	TRUE
436471036	150000000	286471036	Shrek 2	TRUE
377019252	94000000	283019252	The Lord of the Rings: The Return of the King	TRUE
309125409	32500000	276625409	Star Wars: Episode VI - Return of the Jedi	TRUE
329691196	55000000	274691196	Forrest Gump	TRUE
290158751	18000000	272158751	Star Wars: Episode V - The Empire Strikes Back	TRUE

3.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.

Extract all the movies in the imdb_top_250 column which are not in the english language and store them in a new column named top_foreign_lang_film. You can use your own imagination also!

Task: find IMDB top 250

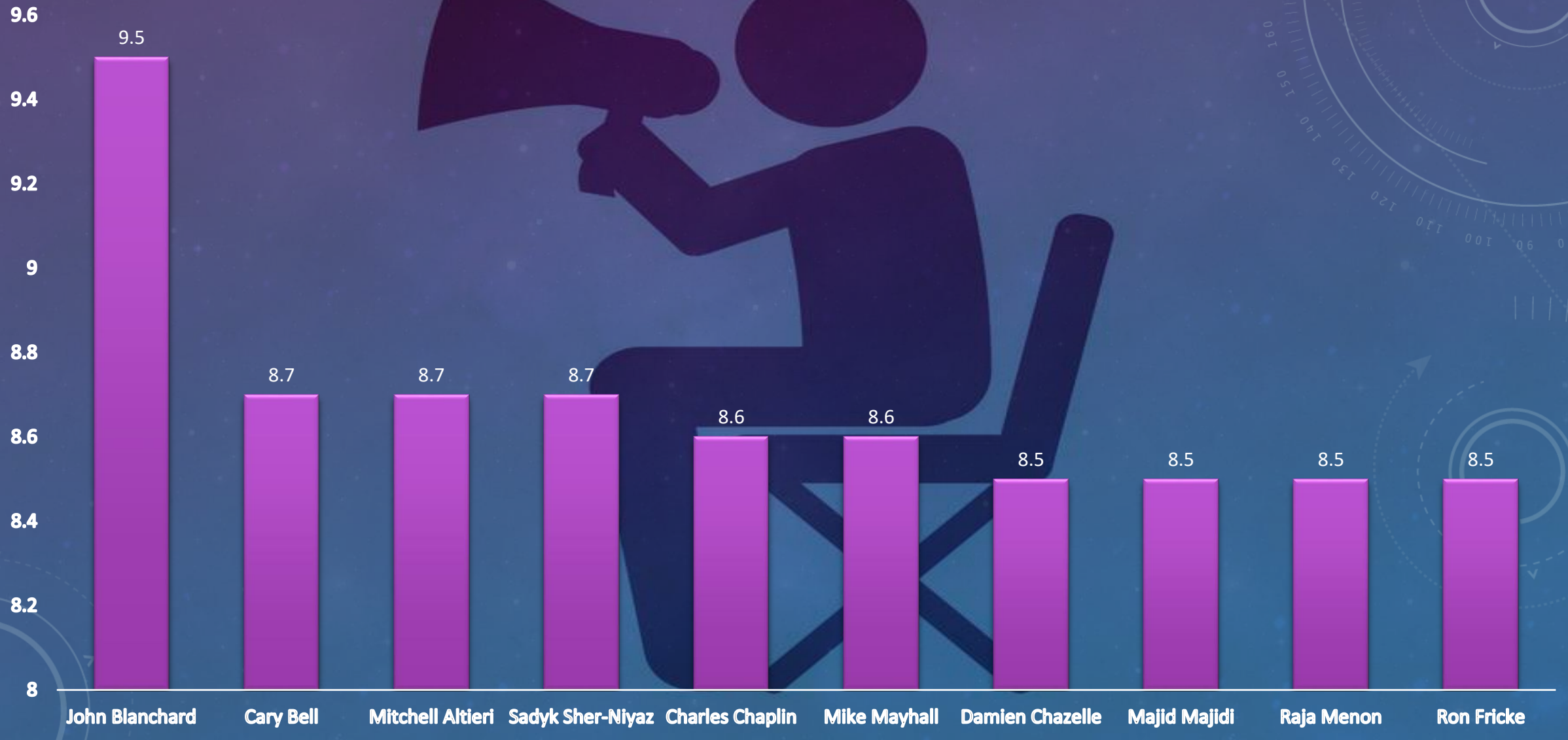
Language→ English

Top_250	language	imdb_score	Rank
The Shawshank Redemption	English	9.3	1
The Godfather	English	9.2	2
Fargo	English	9	3
The Godfather: Part II	English	9	3
The Dark Knight	English	9	3
12 Angry Men	English	8.9	6
The Good, the Bad and the Ugly	Italian	8.9	6
Schindler's List	English	8.9	6
The Lord of the Rings: The Return of the King	English	8.9	6
Pulp Fiction	English	8.9	6
It's Always Sunny in Philadelphia	English	8.8	11
Daredevil	English	8.8	11
Star Wars: Episode V - The Empire Strikes Back	English	8.8	11
The Lord of the Rings: The Fellowship of the Ring	English	8.8	11
Forrest Gump	English	8.8	11

Language→ foreign (other than english)

Top_250_foreign_lang_film	language	imdb_score	rank
Dekalog	Polish	9.1	1
Dekalog	Polish	9.1	1
Seven Samurai	Japanese	8.7	3
Gomorrah	Italian	8.7	3
Spirited Away	Japanese	8.6	5
Children of Heaven	Persian	8.5	6
Airlift	Hindi	8.5	6
The Lives of Others	German	8.5	6
Samsara	None	8.5	6
Baahubali: The Beginning	Telugu	8.4	10
Rang De Basanti	Hindi	8.4	10
A Separation	Persian	8.4	10
Das Boot	German	8.4	10
Princess Mononoke	Japanese	8.4	10
Oldboy	Korean	8.4	10

4.Best directors: group 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.
Task: find the best directors



5.Popular genres: perform this step using the knowledge gained while performing previous steps.
Task: find popular genres

Genres	Mean imdb_score
Action|Adventure|Drama|Fantasy	8.8
Crime|Drama|Horror|Thriller	8.6
Adventure|Drama|Sci-Fi	8.35
Adventure|Animation|Drama|Family|Musical	8.5
Crime|Drama|Fantasy|Mystery	8.5
Drama|Western	8.4
Adventure|Animation|Family|Sci-Fi	8.4
Adventure|Animation|Comedy|Drama|Family|Fantasy	8.3
Biography|Drama|History|Music	8.3
Crime|Documentary	8.3

6.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.

T**ask:** find the critic-favorite and audience-favorite actors

Critic_fav_actor
Phaldut Sharma

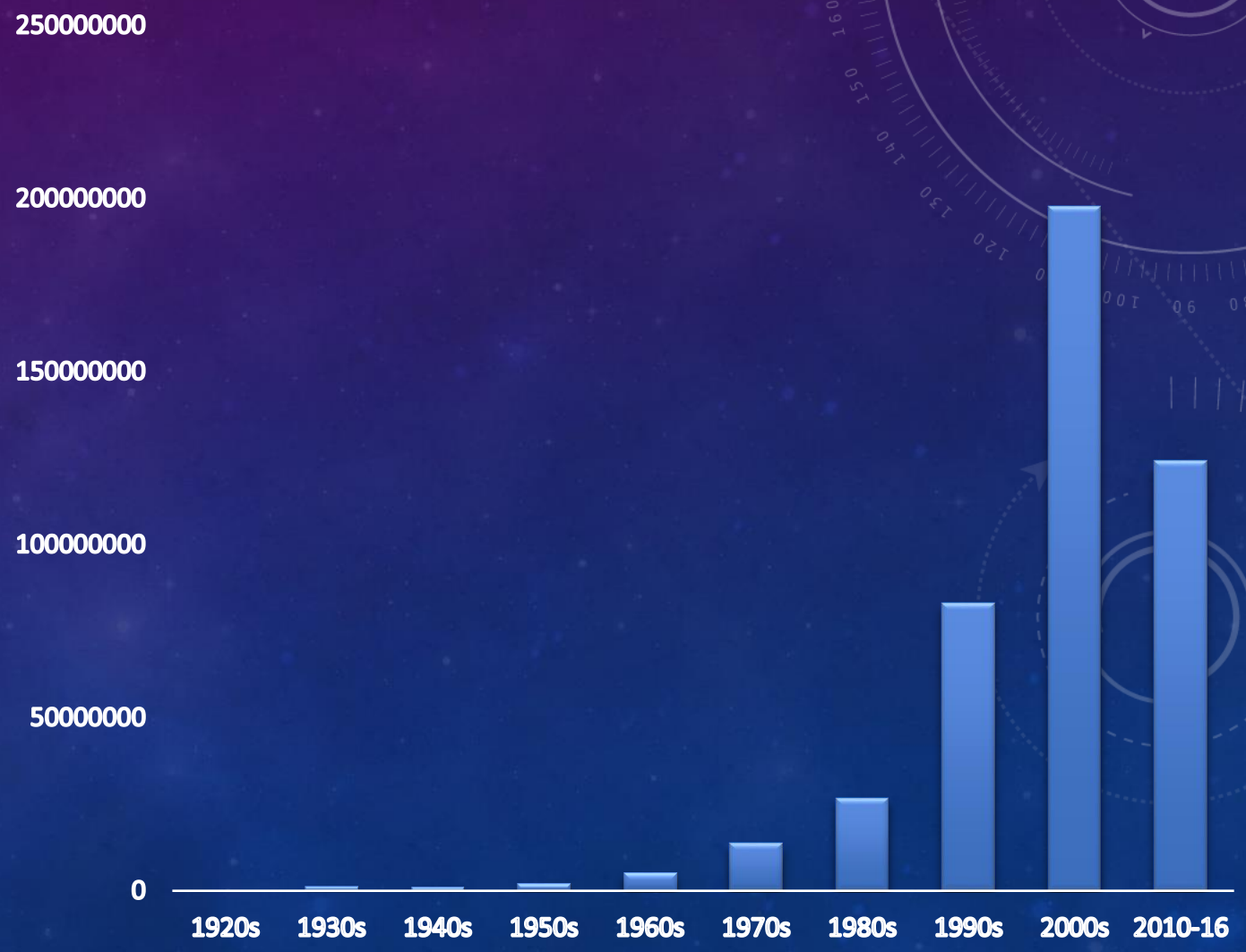
mean_of_num_critic_for_reviews
738

Audience_Fav_actor
Heather Donahue

Mean_num_user_for_reviews
3400

Sum of users voted in each decade→

Decade	sum_of_voted_users
1920s	132425
1930s	1236818
1940s	1215055
1950s	2178269
1960s	5271757
1970s	13707123
1980s	26855844
1990s	83143103
2000s	197683598
2010-16	124384369



LEARNINGS:-

- Statistics approach to find the outliers.
- Application of boxplot
- Replace all the blank cell with a value because if the a blank cell is present it means table ends there and when we use ctrl + shift+ down arrow to select all the rows it will stop at every blank cell and thus we have to keep on pressing ctrl + shift+ down arrow thus to avoid this and have a smooth process we should will blank cell with a appropriate value such as a categorical column with null string and numerical column with 0 so if any numerical calculation is performed error will not be shown.
- How consolidation be use to group by with a categorical data and use a aggregate function with it.

THANK YOU