IMDB Movie Analysis

Problem Statement: The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

Data Cleaning:

- This step involves preprocessing the data to make it suitable for analysis.row
- Here we removed rows with blanks and checked the text has proper cases i.,e upper and lower case and made it proper by using proper function.
- Columns like colour, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, movie_imdb_link, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes are irrelevant data needed to be dropped is dropped.
- Then we removed duplicate rows and separated generes by using text to column.

Data Analysis: Here, you'll explore the data to understand the relationships between different variables. You might look at the correlation between movie ratings and other factors like genre, director, budget, etc. You might also want to consider the year of release, the actors involved, and other relevant factors.

A. **Movie Genre Analysis**: Analyze the distribution of movie genres and their impact on the IMDB score.

   Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

   - Here the separated genres are counted using countif functon i.,e =COUNTIF($D:$D,Q12) and copied to rest of columns.
   - Finding mean using average function as mean is sum of observations divided by total number of observations. i.,e =COUNTIF($D:$D,Q12).
   - Finding median by built-in function =MEDIAN(R12:Y12)
   - Finding mode by built-in function =MODE(R12:Y12)
   - Finding varaince by built-in function as we have both sample and population so here we are calculating variance for both =VAR.P(R12:Y12) and =VAR.S(R12:Y12).

- Finding step deviation by built-in function as we have both sample and population so here we are calculating standard deviation for both =STDEV.P(R12:Y12) and =STDEV.S(R12:Y12).

After splitting the generes we got 8 rows i.,e atleast one movie from the dataset has 8 generes

Genere	G1	G2	G3	G4	G5	G6	G7	G8	Total sum	Mean	Median	Mode	varp	vars	stdevp	stdevs
Action	962	0	0	0	0	0	0	0	962	120.25	0	0	101220.4	115680.5	318.1516	340.1184
Adventure	375	412	0	0	0	0	0	0	787	98.375	0	0	29118.48	33278.27	170.6414	182.4233
Animation	46	125	28	0	0	0	0	0	199	24.875	0	0	1696.859	1939.268	41.19295	44.03712
Drama	691	913	293	42	2	0	0	0	1941	242.625	22	0	115966.5	132533.1	340.5385	364.051
Family	3	137	152	125	30	3	0	0	450	56.25	16.5	3	4137.938	4729.071	64.3268	68.76824
Musical	2	18	29	25	18	9	2	0	103	12.875	13.5	2	109.6094	125.2679	10.46945	11.19231
Mystery	23	177	113	55	10	4	1	0	383	47.875	16.5	#N/A	3679.109	4204.696	60.65566	64.84363
Romance	3	307	372	147	35	8	5	1	878	109.75	21.5	#N/A	19900.69	22743.64	141.0698	150.81
Thriller	3	141	449	387	121	12	2	2	1117	139.625	66.5	2	28761.48	32870.27	169.5921	181.3016
Comedy	1029	291	164	19	0	0	0	0	1503	187.875	9.5	0	111050.4	126914.7	333.2422	356.2509
War	0	21	53	49	29	7	1	0	160	20	14	0	417.75	477.4286	20.43893	21.85014
Horror	160	143	73	11	4	0	0	0	391	48.875	7.5	0	4050.609	4629.268	63.6444	68.03872

https://docs.google.com/spreadsheets/d/1TtZq7ilMNUpgnMK56OvLLbouX7OpqbYe/edit?usp=sharing&ouid=108880336182281145657&rtpof=true&sd=true

B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

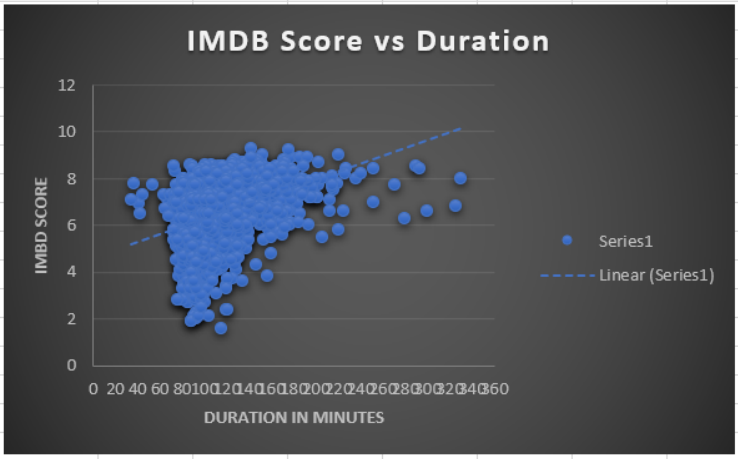
- Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

➢ Here we are considering only two columns in new worksheet i.,e duration of movie and IMDB score.

- Then respective mean, median and standard deviation is calculated for movie duration using built in functions in excel.
- For mean :- =AVERAGE(A2:A3849)
- For mode:- =MEDIAN(A2:A3849)
- For standard deviation:- =STDEV.S(A2:A3849)

➢ And a scatter plot is plotted to visualize the relationship between movie duration and IMDB score.
➢ The trendline is drawn and it indicates that the IMDB increases with increase in duration.

Mean	109.9241164
Median	106
Standard Deviation	22.75364979



https://docs.google.com/spreadsheets/d/18_f3ewLJ07eJeoY2o211mSnZ1UNiF_zB/edit?usp=sharing&ouid=108803361822811456 57&rtpof=true&sd=true

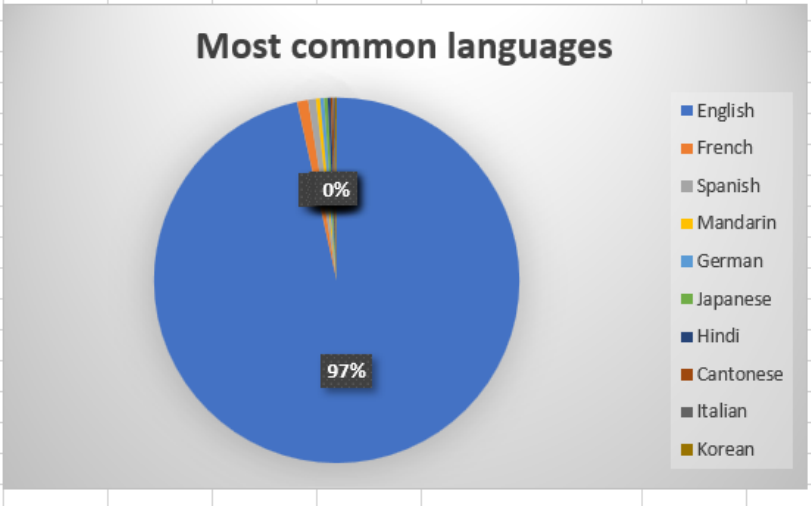
C. Language Analysis: Situation: Examine the distribution of movies based on their language.

- **Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Here we are considering only two columns i.,e language and IMBD score

- By using counif function we will count number of movies for each language i.,e =COUNTIF('cleaned data'!$J$2:$J$3849, J2)
- For mean we will use average of IMDB score using =AVERAGE(IF('cleaned data'!$J$2:$J$3849=J2, 'cleaned data'!$N$2:$N$3849))
- For median we use =MEDIAN(IF('cleaned data'!$J$2:$J$3849=J2, 'cleaned data'!$N$2:$N$3849)
- For Standard deviation we use =STDEV.S(IF('cleaned data'!$J$2:$J$3849=J2, 'cleaned data'!$N$2:$N$3849))

Most common Languages are:-				
Language	Count	Mean	Median	Standard Deviation
English	3668	6.42391	6.5	1.048750752
French	37	7.28649	7.2	0.561328861
Spanish	26	7.05	7.15	0.826196103
Mandarin	14	7.02143	7.25	0.765786244
German	13	7.69231	7.7	0.640912811
Japanese	12	7.625	7.8	0.899621132
Hindi	10	6.76	7.05	1.111755369
Cantonese	8	7.2375	7.3	0.440575922
Italian	7	7.18571	7	1.155318962
Korean	5	7.7	7.7	0.570087713



Here the mean, median and standard deviation is done for IMDB score's of respective languages.

https://docs.google.com/spreadsheets/d/1QaGTHKN_XXhfi1wa9grccNXPx8WwD1cl/edit?usp=sharing&ouid=108803361822811456 57&rtpof=true&sd=true

D. Director Analysis: Influence of directors on movie ratings.

- Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

  - Here we will group by directors and we will find average of their IMDB scores using Average function.
  - And we will sort them decending order using sort and filter.

  - Here we are finding largest value using max function:- =MAX(B2:B1749)
  - And finding percentile using function:- =PERCENTILE(B2:B1748,1)

    Here in the place of 1 we can use a value between 0 and 1 for better output or convenience I used 1.

director_name	Average
Charles Chaplin	8.60
Tony Kaye	8.60
Alfred Hitchcock	8.50
Damien Chazelle	8.50
Majid Majidi	8.50
Ron Fricke	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Asghar Farhadi	8.40
Marius A. Markevicius	8.40
Highest Score	8.60
percentile	8.60

The top directors based on average IMDB score is given in the above table.

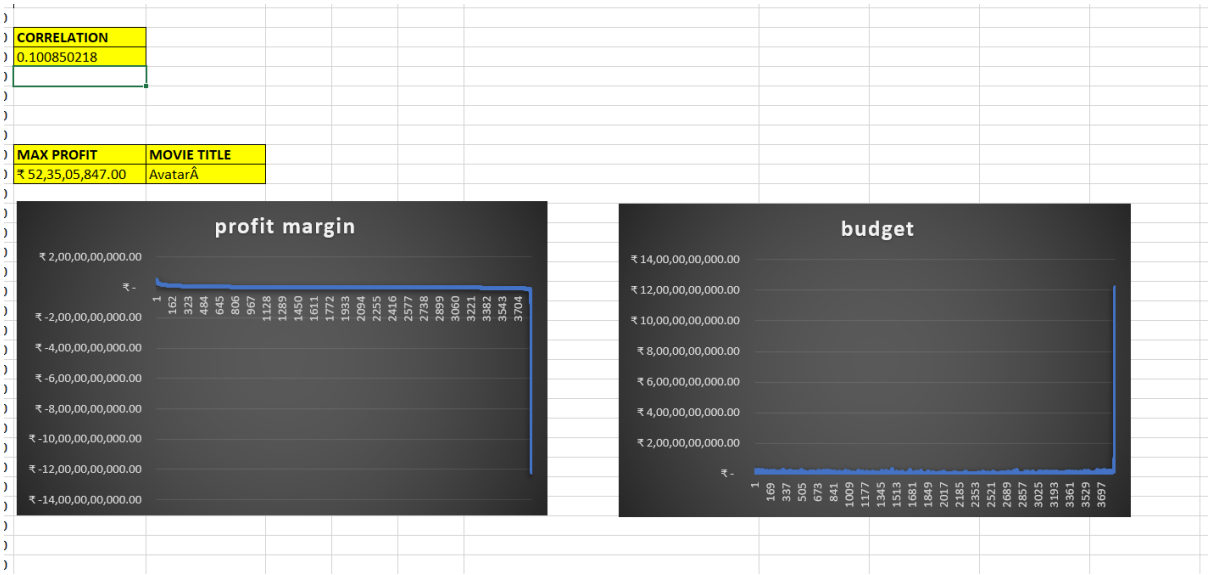
https://docs.google.com/spreadsheets/d/1Su6R5MC0j7Njc2exzJ6Ro4K5fOest0Yn/edit?usp=sharing&ouid=108803361822811456 57&rtpof=true&sd=true

E. Budget Analysis: Explore the relationship between movie budgets and their financial success.

- Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Here we are considering the columns gross and budget which are related to finance.

- Here the profit is given by substracting the gross value with budget.
- And arranging the profit in decending order we get maximum profit margin on top.
- Finding correlation between gross and budget using CORREL function:- =CORREL(A:A,C:C).



https://docs.google.com/spreadsheets/d/1wPUaJ-fk-mQIjeeq23kOQOVapO91fE9H/edit?usp=sharing&ouid=108803361822811456 57&rtpof=true&sd=true

Presentation video

https://drive.google.com/file/d/1Ve-R5FoD3x86wyhOk5Tc4__zJWdtJ0ZU/view?usp=sharing