IMDB MOVIE RATING

DATASET LINK: https://drive.google.com/drive/folders/1Md7ZeZda-RhzsAQy3XLIEibbfbrRKr5M?usp=drive_link

Project Description:

The project aimed to analyze IMDb movie ratings to uncover trends and insights within the film industry. Objectives included identifying factors influencing movie ratings, exploring genre preferences, and potentially predicting future ratings based on historical data.

Approach:

The approach involved data collection from IMDb's dataset, preprocessing to clean and organize the data, exploratory data analysis to uncover patterns and correlations, and possibly machine learning techniques for predictive modeling. Techniques such as data visualization, statistical analysis, and regression modeling were employed.

Tech-Stack Used:

Microsoft Excel 2010

Purpose: Data Cleaning, Data Analysis

Data Analytics Tasks:

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.**
- **Hint:** Use Excel's COUNTIF function to count the number of movies for each genre. You might need to manipulate the 'genres' column to separate multiple genres for a single movie. Use Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics. Compare the statistics to understand the impact of genre on movie ratings.

Insights:

Row Labels	Count of movie_title
Action	959
Adventure	369
Animation	45
Biography	205
Comedy	989
Crime	255
Documentary	26
Drama	668
Family	3
Fantasy	37
Horror	164
Musical	2
Mystery	23
Romance	1
Sci-Fi	7
Thriller	1
Western	2
(blank)	
Grand Total	3756

Row Labels	Average of imdb_score
Action	6.289781022
Adventure	6.550406504
Animation	6.74
Biography	7.15902439
Comedy	6.169464105
Crime	6.940392157
Documentary	6.796153846
Drama	6.831586826
Family	6.5
Fantasy	6.281081081
Horror	5.848780488
Musical	6.75
Mystery	6.652173913
Romance	7.1
Sci-Fi	6.628571429
Thriller	4.8
Western	8.1
(blank)	
Grand Total	6.465282215

Row Labels	StdDev of imdb_score
Action	1.038868751
Adventure	1.122621204
Animation	0.970800981
Biography	0.695065961
Comedy	1.032752021
Crime	0.86914285
Documentary	1.716620114
Drama	0.905193675
Family	1.216552506
Fantasy	0.894066191
Horror	1.034881683
Musical	0.636396103
Mystery	1.092482396
Sci-Fi	1.107119815
Western	1.13137085
(blank)	
Grand Total	1.056127552

Row Labels	Var of imdb_score
Action	1.079248282
Adventure	1.260278367
Animation	0.942454545
Biography	0.483116691
Comedy	1.066576737
Crime	0.755409294
Documentary	2.946784615
Drama	0.819375589
Family	1.48
Fantasy	0.799354354
Horror	1.070980099
Musical	0.405
Mystery	1.193517787
Sci-Fi	1.225714286
Western	1.28
Grand Total	1.115405405

AVREAGE	220.9412
MEDIAN	37
MODE	2
MAX	989
MIN	1
VARIANCE	111809.1
STANDARD DEVIATION	334.3786

Descriptive Statistics of the IMDb Scores of each genre are

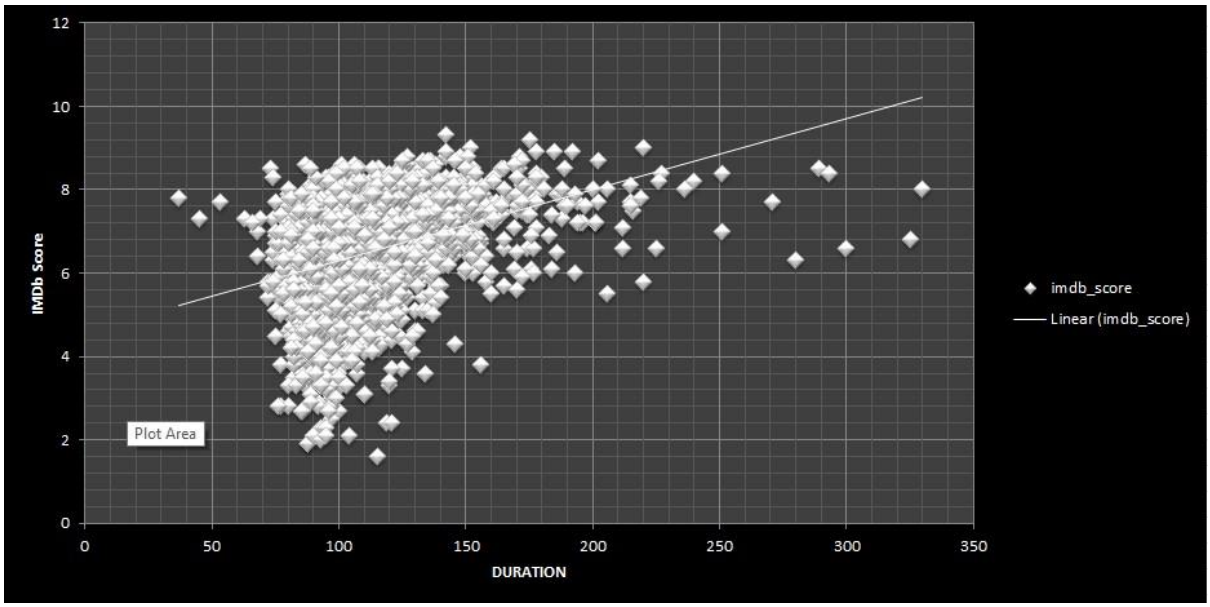
Maximum movies are of "**COMEDY"** genre (989) and minimum movies are of **"Romance"** and **"Thriller"** genre(1).

Total number of Movies are 3756.

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.
- Hint: Calculate descriptive statistics such as mean, median, and standard deviation for movie durations. Use Excel's functions like AVERAGE, MEDIAN, and STDEV. Create a scatter plot to visualize the relationship between movie duration and IMDB score. Add a trendline to assess the direction and strength of the relationship.

Insights:



Average	110.258
Median	106
STD DEV	22.64672

Mean of Duartion is 110.258.

Median of Duration is 106.

Standard Deviation of Duration is 22.64672.

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.
- **Hint:** Use Excel's COUNTIF function to count the number of movies for each language. Calculate the mean, median, and standard deviation of the IMDB scores for each language. Compare the statistics to understand the impact of language on movie ratings.

Insights:

Row Labels	Count of imdb_score
Aboriginal	2
Arabic	1
Aramaic	1
Bosnian	1
Cantonese	7
Czech	1
Danish	3
Dari	2
Dutch	3
English	3598
Filipino	1
French	34
German	10
Hebrew	1
Hindi	5
Hungarian	1
Indonesian	2
Italian	7
Japanese	10
Kazakh	1
Korean	5
Mandarin	15
Maya	1
Mongolian	1
None	1
Norwegian	4
Persian	3
Portuguese	5
Romanian	1
Russian	1
Spanish	23
Thai	3
Vietnamese	1
Zulu	1
(blank)	
Grand Total	**3756**

Q	R	S	T
Aboriginal	mean	median	standard deviation
Arabic	7.9	6.6	1.056246753
Aramaic	7.1	6.6	1.056246753
Bosnian	6.8	6.6	1.056246753
Cantonese	8.5	6.6	1.056246753
Czech	6.6	6.6	1.056246753
Danish	6.2	6.6	1.056246753
Dari	7.8	6.6	1.056246753
Dutch	7.5	6.6	1.056246753
English	7.5	6.6	1.056246753
Filipino	6.9	6.6	1.056246753
French	6.1	6.6	1.056246753
German	6.7	6.6	1.056246753
Hebrew	7.3	6.6	1.056246753
Hindi	6.5	6.6	1.056246753
Hungarian	7.2	6.6	1.056246753
Indonesian	6.6	6.6	1.056246753
Italian	8.1	6.6	1.056246753
Japanese	6.7	6.6	1.056246753
Kazakh	6.8	6.6	1.056246753
Korean	7.1	6.6	1.056246753
Mandarin	7	6.6	1.056246753
Maya	6.7	6.6	1.056246753
Mongolian	7.9	6.6	1.056246753
None	6.1	6.6	1.056246753
Norwegian	7.2	6.6	1.056246753
Persian	7.7	6.6	1.056246753
Portuguese	8.2	6.6	1.056246753
Romanian	5.9	6.6	1.056246753
Russian	7	6.6	1.056246753
Spanish	7.8	6.6	1.056246753
Thai	7.3	6.6	1.056246753
Vietnamese	7.2	6.6	1.056246753
Zulu	6.5	6.6	1.056246753

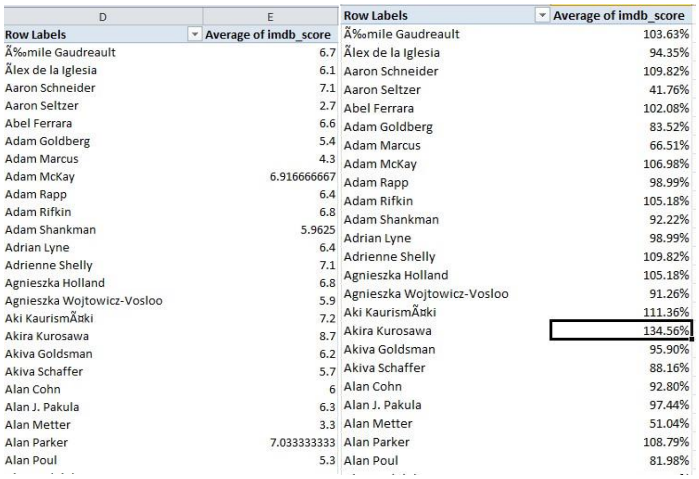
Max language: English

Median and Standard Deviation are same for all languages.

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.
- Hint: Calculate the average IMDB score for each director. Use Excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores.

Insights



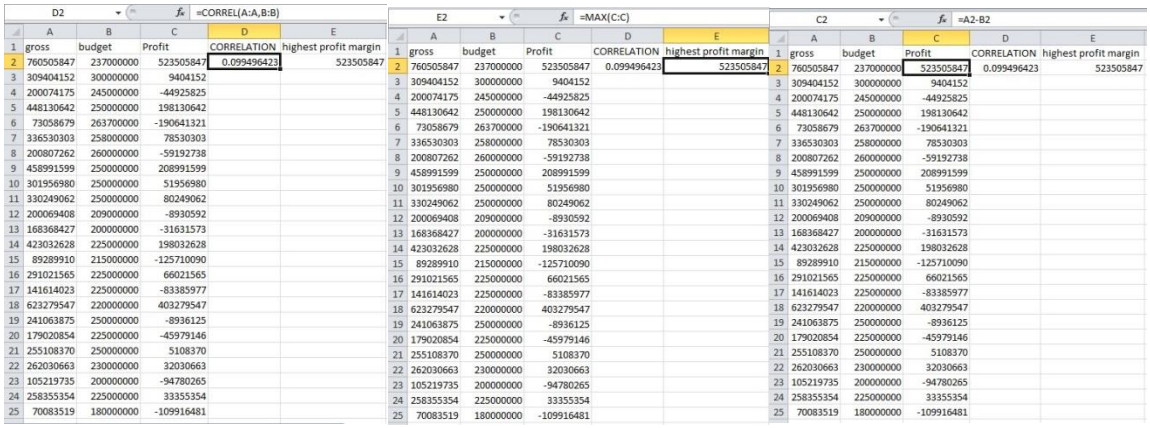
D	E
Row Labels	Average of imdb_score
Ã‰mile Gaudreault	6.7
Ãlex de la Iglesia	6.1
Aaron Schneider	7.1
Aaron Seltzer	2.7
Abel Ferrara	6.6
Adam Goldberg	5.4
Adam Marcus	4.3
Adam McKay	6.916666667
Adam Rapp	6.4
Adam Rifkin	6.8
Adam Shankman	5.9625
Adrian Lyne	6.4
Adrienne Shelly	7.1
Agnieszka Holland	6.8
Agnieszka Wojtowicz-Vosloo	5.9
Aki KaurismÃ¤ki	7.2
Akira Kurosawa	8.7
Akiva Goldsman	6.2
Akiva Schaffer	5.7
Alan Cohn	6
Alan J. Pakula	6.3
Alan Metter	3.3
Alan Parker	7.033333333
Alan Poul	5.3

Row Labels	Average of imdb_score
Ã‰mile Gaudreault	103.63%
Ãlex de la Iglesia	94.35%
Aaron Schneider	109.82%
Aaron Seltzer	41.76%
Abel Ferrara	102.08%
Adam Goldberg	83.52%
Adam Marcus	66.51%
Adam McKay	106.98%
Adam Rapp	98.99%
Adam Rifkin	105.18%
Adam Shankman	92.22%
Adrian Lyne	98.99%
Adrienne Shelly	109.82%
Agnieszka Holland	105.18%
Agnieszka Wojtowicz-Vosloo	91.26%
Aki KaurismÃ¤ki	111.36%
Akira Kurosawa	134.56%
Akiva Goldsman	95.90%
Akiva Schaffer	88.16%
Alan Cohn	92.80%
Alan J. Pakula	97.44%
Alan Metter	51.04%
Alan Parker	108.79%
Alan Poul	81.98%

Director **Akira Kurosawa** has highest scores.

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.
- Hint: Calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function. Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using Excel's MAX function.



gross	budget	Profit	CORRELATION	highest profit margin
760505847	237000000	523505847	0.099496423	523505847
309404152	300000000	9404152		
200074175	245000000	-44925825		
448130642	250000000	198130642		
73058679	263700000	-190641321		
336530303	258000000	78530303		
200807262	260000000	-59192738		
458991599	250000000	208991599		
301956980	250000000	51956980		
330249062	250000000	80249062		
200069408	209000000	-8930592		
168368427	200000000	-31631573		
423032628	225000000	198032628		
89289910	215000000	-125710090		
291021565	225000000	66021565		
141614023	225000000	-83385977		
623279547	220000000	403279547		
241063875	250000000	-8936125		
179020854	225000000	-45979146		
255108370	250000000	5108370		
262030663	230000000	32030663		
105219735	200000000	-94780265		
258355354	225000000	33355354		
70083519	180000000	-109916481		