IMDB Movie Analysis

Excel Sheet:
https://docs.google.com/spreadsheets/d/1UmGIcJnDVGAI83PuBVEEAl0U4nfdSgSw/edit?usp=drive_link&ouid=111667124296828215097&rtpof=true&sd=true

Video Link:
https://drive.google.com/file/d/1yYDkeioQTEiojK5gEOOeqJCseFlV9zEH/view?usp=drive_link

Victor Shah

Project Description

- As a data analyst in a multinational company, my primary responsibility is about investigating “What factors influence the success of a movie on IMDB?”.
- The aspect of my role is to extract meaningful insights from the dataset. These insights will offer actionable information for various teams within the business.
- The impact of this problem is a significant turn for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

Approach

- Downloading the dataset: The first step is downloading the excel file (.csv) into the local device. Make sure the downloaded file is having the extension (.xlsx)
- Understanding the worksheet: The next step is to examine the structure of the table holding the data in the Excel Sheet. (IMDB_movies.csv)
- Identifying the key tables: Identification of the primary key from the dataset of IMDB_movies.
- Data Cleaning: This is the preprocessing step that makes the data suitable for analysis. It includes handling missing values, removing duplicates.
- Using the Five “Whys” Approach: This technique helps uncover the root cause of the problem dealing during analysis.

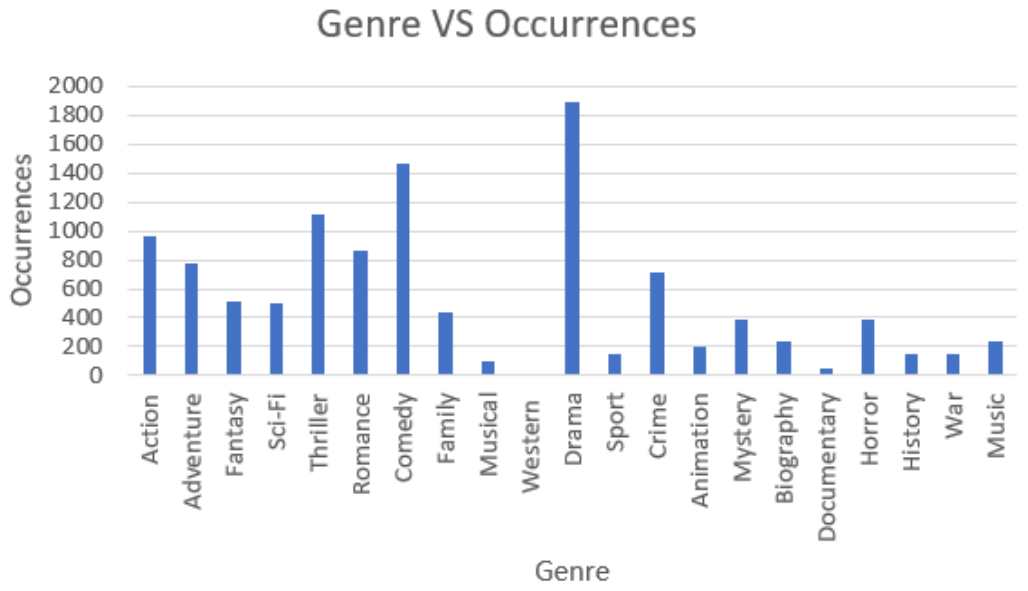
DATA ANALYTICS TASKS

A. Movie Genre Analysis

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.

A) Movie Genre Analysis

Genre	Occurrences	Mean	Median	Mode	Max	Min	Standard Deviation	Variance
Action	959	6.29	6.3	6.6	9	2.1	1.039	1.079
Adventure	781	6.45	6.6	6.6	8.9	2.3	1.114	1.241
Fantasy	507	6.28	6.4	6.7	8.9	2.2	1.134	1.286
Sci-Fi	496	6.33	6.4	6.7	8.8	1.9	1.161	1.348
Thriller	1117	6.38	6.4	6.5	9	2.7	0.972	0.944
Romance	859	6.44	6.5	6.5	8.5	2.1	0.954	0.911
Comedy	1461	6.19	6.3	6.7	8.8	1.9	1.036	1.073
Family	442	6.21	6.3	6.7	8.6	1.9	1.162	1.350
Musical	96	6.60	6.75	7.1	8.5	2.1	1.102	1.214
Western	0	6.79	6.8	6	8.9	4.7	0.932	0.868
Drama	1893	6.79	6.9	6.7	9.3	2.1	0.897	0.804
Sport	148	6.59	6.8	7.2	8.3	2	1.042	1.085
Crime	709	6.55	6.6	6.6	9.3	2.4	0.984	0.968
Animation	196	6.70	6.8	6.7	8.6	2.8	0.990	0.979
Mystery	384	6.47	6.5	6.6	8.6	3.1	1.016	1.032
Biography	239	7.16	7.2	7	8.9	4.5	0.691	0.478
Documentary	45	6.99	7.4	7.6	8.5	1.6	1.385	1.917
Horror	392	5.92	6	5.9	8.6	2.3	0.998	0.997
History	149	7.16	7.2	7.7	8.9	5.5	0.672	0.452
War	152	7.06	7.1	7.1	8.6	4.3	0.802	0.643
Music	233	6.46	6.6	6.2	8.5	1.6	1.184	1.402



The red highlighted cells show the maximum values in the table whereas the yellow highlighted cells show the minimum values in the table.

The most occurred Genre is “DRAMA”

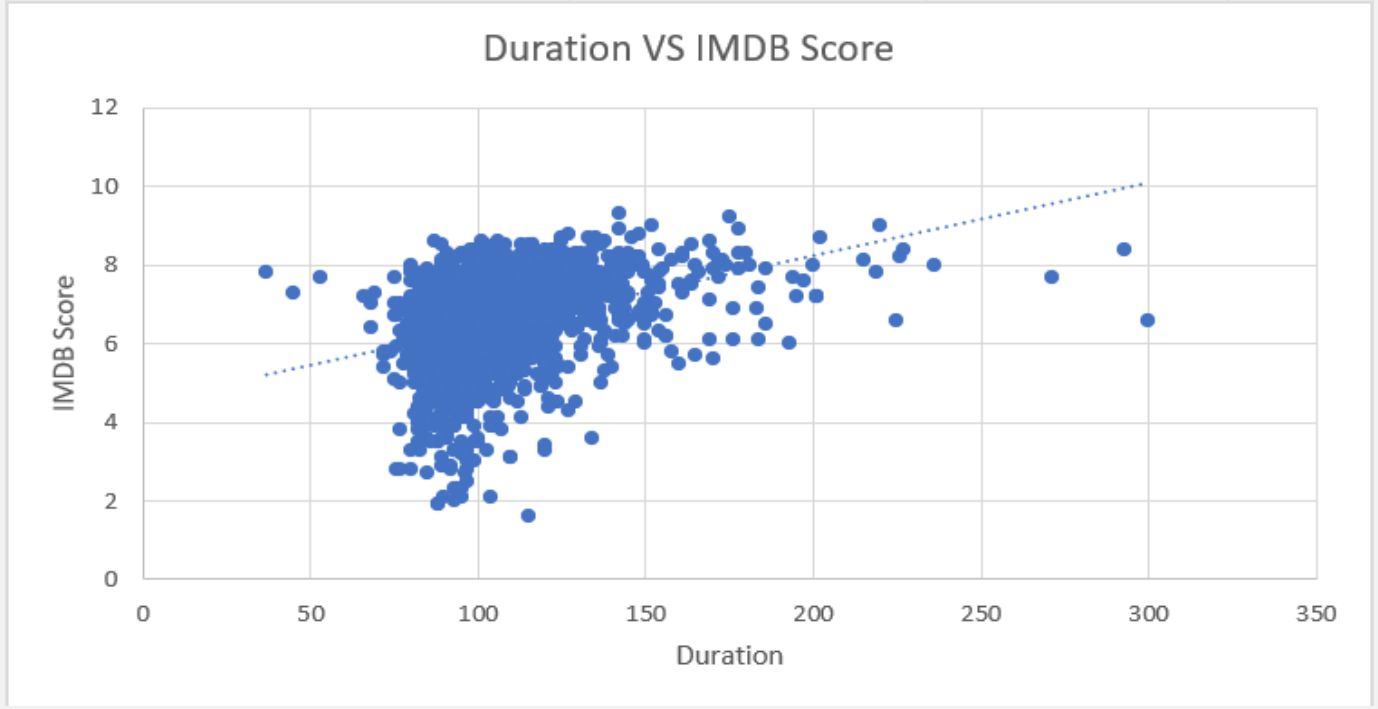
B. Movie Duration Analysis

Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Most of the movies have a mean duration of 110 minutes (around 2 hours long).

B) Movie Duration Analysis

Mean	110.26
Median	106
Standard Deviation	22.65

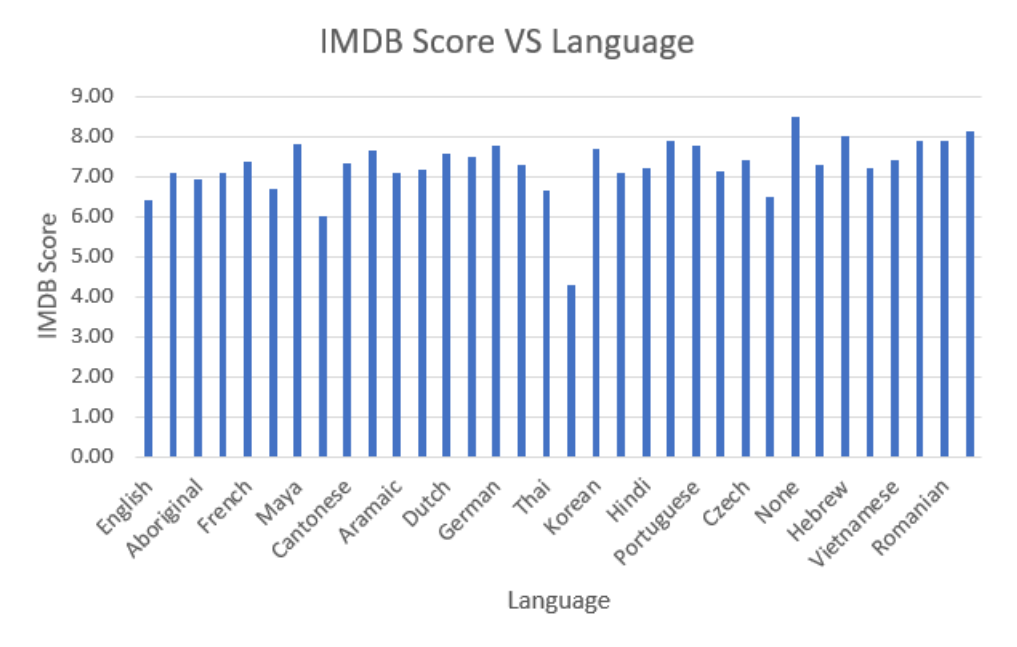


C. Language Analysis

Determine the most common languages used in movies and analyse their impact on the IMDB score using descriptive statistics.

The most occurred language in movies is English.

#DIV/0! Error occurs due to STDEV function. This function can only perform with more than two available values.



C) Language Analysis

Language	Occurrences	Mean	Median	Standard Deviation
English	3598	6.43	6.5	1.051
Mandarin	15	7.08	7.4	0.772
Aboriginal	2	6.95	6.95	0.778
Spanish	23	7.08	7.2	0.861
French	34	7.36	7.3	0.519
Filipino	1	6.70	6.7	#DIV/0!
Maya	1	7.80	7.8	#DIV/0!
Kazakh	1	6.00	6	#DIV/0!
Cantonese	7	7.34	7.3	0.351
Japanese	10	7.66	8	0.990
Aramaic	1	7.10	7.1	#DIV/0!
Italian	7	7.19	7	1.155
Dutch	3	7.57	7.8	0.404
Dari	2	7.50	7.5	0.141
German	10	7.77	7.8	0.712
Mongolian	1	7.30	7.3	#DIV/0!
Thai	3	6.63	6.6	0.451
Bosnian	1	4.30	4.3	#DIV/0!
Korean	5	7.70	7.7	0.570
Hungarian	1	7.10	7.1	#DIV/0!
Hindi	5	7.22	7.4	0.801
Danish	3	7.90	8.1	0.529
Portuguese	5	7.76	8	0.979
Norwegian	4	7.15	7.3	0.574
Czech	1	7.40	7.4	#DIV/0!
Russian	1	6.50	6.5	#DIV/0!
None	1	8.50	8.5	#DIV/0!
Zulu	1	7.30	7.3	#DIV/0!
Hebrew	1	8.00	8	#DIV/0!
Arabic	1	7.20	7.2	#DIV/0!
Vietnamese	1	7.40	7.4	#DIV/0!
Indonesian	2	7.90	7.9	0.424
Romanian	1	7.90	7.9	#DIV/0!
Persian	3	8.13	8.4	0.551

D. Director Analysis

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

The most popular director based on their average movie rating is Akira Kurosawa.

100 percentile score among the average movie ratings came out to be 8.7

D) Director Analysis

director_name	Average_Movie Rating	Number of Movies
Akira Kurosawa	8.7	1
Tony Kaye	8.6	1
Charles Chaplin	8.6	1
Alfred Hitchcock	8.5	1
Ron Fricke	8.5	1
Damien Chazelle	8.5	1
Majid Majidi	8.5	1
Sergio Leone	8.4	3
Christopher Nolan	8.4	8
Richard Marquand	8.4	1
Asghar Farhadi	8.4	1

Percentile	
100%	8.7
90%	7.5
80%	7.1
70%	6.9
60%	6.7
50%	6.5
40%	6.2
30%	5.9
20%	5.6
10%	5.1
0%	2.1

E. Budget Analysis

Analyse the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

The movie with the highest profit is

"Avatar" with a profit made of

523505847 dollars.

The correlation coefficient between

the gross and budget came out to be 0.099

E) Budget Analysis	
Movie With Highest Profit Margin is	Avatar
Profit made is	523505847
Correlation Coefficient Is	0.099

Tech-Stack Used

- Microsoft Excel: It is a spreadsheet program from Microsoft and a component of its Office product for business applications. This enables users to format, calculate and organize data in a spreadsheet.
- MS Excel Functions: They are predefined formulas that perform calculations by using specific values, called arguments, in a particular order or structure. Some of the functions are:
  1. Text functions: clean(), substitute(), replace(), concatenate(), trim(), search(), find(), etc.
  2. Mathematical and Statistical functions: sum(), sumif(), count(), countif(), round(), avg(), min(), max(), subtotal(), etc.
- Data Visualization in Excel: Bar, Column, Line, Histogram, Pie, Scatter, Boxpot, Treemap, etc.

Insights

- We were able to determine the most common genres of movies dataset. We also added the descriptive statistics for each genre.
- To identify the relationship between the movie duration and the IMDB score. The most mean movie duration came out to be 110 minutes.
- Determining the most common languages used in movies and how language impacts the IMDB score.
- Checking the influence of the movie directors on the IMDB ratings. The most popular director came out to be Akira Kurosawa.
- Analyzing the correlation between the movie budgets and their financial success. (gross income)

Results

- Remembering to adapt excel functions on specific datasets.
- These learned insights helped me understand specific business questions which were addressed by MS Excel.
- Learning about Excel Text and Statistical functions. The importance of max(), average(), min(), text() functions.
- We were able to build different charts for visualization for answering the business questions. Some of the charts used were scatter plot and line graph.
- Achieving the ability to learn and write MS Excel functions to execute different business questions.
- Solving Company related problems using different visualization charts offered by MS Excel.

THANK YOU