IMDB MOVIE Analysis

BY: RUDRA AGGARWAL

IMDB Movie Analysis Project - Description

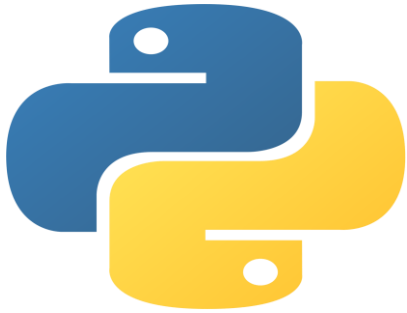
The dataset provided is related to IMDB Movies.
A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?“
Here, success can be defined by high IMDB ratings.
The impact of this problem is significant for movie producers, directors, and investors who want to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

HERE, YOU'LL EXPLORE THE DATA TO UNDERSTAND THE RELATIONSHIPS BETWEEN DIFFERENT VARIABLES. YOU MIGHT LOOK AT THE CORRELATION BETWEEN MOVIE RATINGS AND THE IMPACT OF OTHER FACTORS ON IMDB.

Approach

- Download the dataset and open in Excel for analysis
- Clean the datasets. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.
- Process the data to answer the asked questions
- Use formulas, filters, pivot tables and other functions for finding insights
- Create charts and graphs for easy and meaningful data representation
- Create report and submit the project

Tech Stack Used



Findings & Insights

- Originally the dataset(IMDB Movie Analysis) had 5037 rows.
- Removed empty rows, special characters, and duplicate values
- Columns that were not relevant in finding insights were hidden
- Before cleaning we had 5037 rows (including headings in first row)
- After cleaning we have 3757 rows (including column headings in first row)
- The new dataset was then used for further analysis for correctness of data.

Data cleaning and Preprocessing

We used python in Google colab for cleaning and preprocessing of data.
Colab Link

After we got the dataset we manipulated the genre column and split it into individual columns.

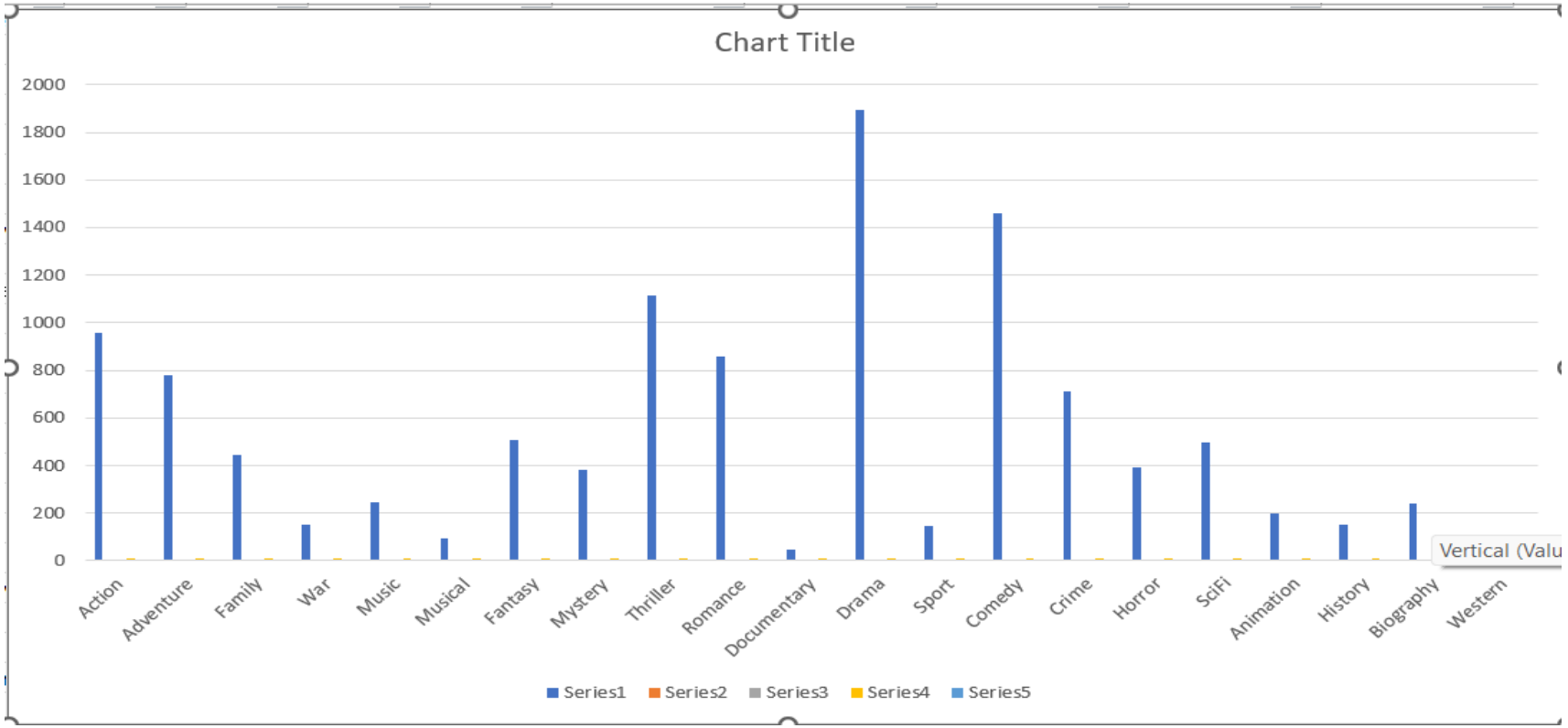
We used various Excel functions like AVERAGEIF, MEAN , MEDIAN, VAR to calculate descriptive statistics and impact of these on IMDB Score for the movies

Cleaned Dataset: Dataset

Here is the final Excel sheet attached Dataset Final

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

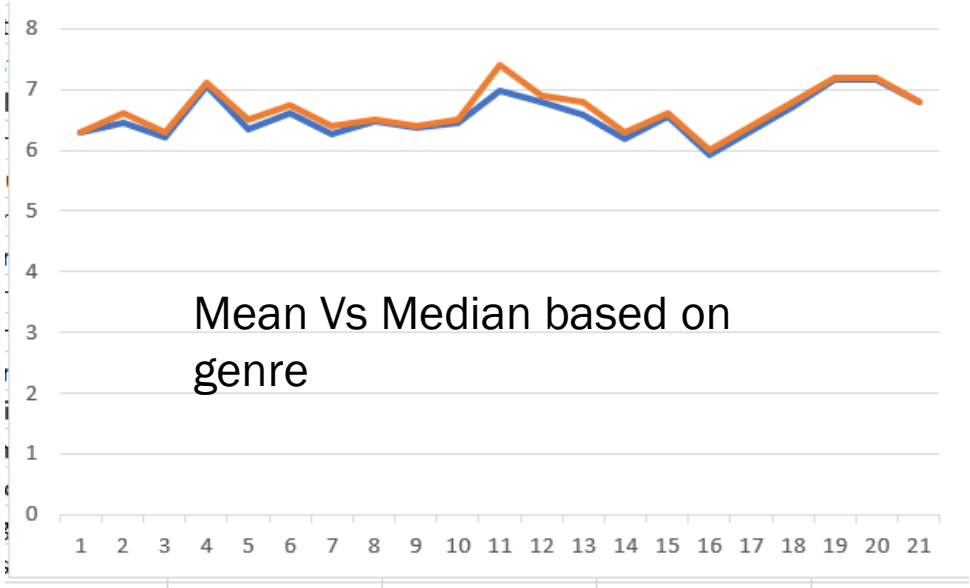
Unique Genre	Count(Genre)	M
Action	959	
Adventure	781	
Family	442	
War	152	
Music	247	
Musical	96	
Fantasy	507	
Mystery	384	
Thriller	1117	
Romance	859	
Documentary	45	
Drama	1893	
Sport	148	
Comedy	1461	
Crime	709	
Horror	392	
SciFi	496	
Animation	196	
History	149	
Biography	239	
Western	59	



Descriptive Statistics Based on genre

Table containing mean median variance based on genre

Unique Genre	Count(Genre)	Mean(Genre)	Median(Genre)	Max(genre)	Min(Genre)
Action	959	6.289781022	6.3	9	2.1
Adventure	781	6.449807939	6.6	8.9	2.3
Family	442	6.213574661	6.3	8.6	1.9
War	152	7.056578947	7.1	8.6	4.3
Music	247	6.343708609	6.5	8.5	1.6
Musical	96	6.596875	6.75	8.5	2.1
Fantasy	507	6.277514793	6.4	8.9	2.2
Mystery	384	6.473958333	6.5	8.6	3.1
Thriller	1117	6.376991943	6.4	9	2.7
Romance	859	6.438300349	6.5	8.5	2.1
Documentary	45	6.988888889	7.4	8.5	1.6
Drama	1893	6.789170629	6.9	9.3	2.1
Sport	148	6.593243243	6.8	8.3	2
Comedy	1461	6.187816564	6.3	8.8	1.9
Crime	709	6.545133992	6.6	9.3	2.4
Horror	392	5.924489796	6	8.6	2.3
SciFi	496	6.327016129	6.4	8.8	1.9
Animation	196	6.70255102	6.8	8.6	2.8
History	149	7.155033557	7.2	8.9	5.5
Biography	239	7.157740586	7.2	8.9	4.5
Western	59	6.793220339	6.8	8.9	4.7



As the mean is almost equal to the median, it indicates that the data is approximately symmetrically distributed. In a symmetric distribution, the mean and median values are close to each other and are often around the center of the data therefore affects the da

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

Pivot table for duration

Row Labels	Average of imdb_score
(blank)	
37-56	7.6
57-76	6.314814815
77-96	5.971092952
97-116	6.394004944
117-136	6.844373402
137-156	7.240888889
157-176	7.468333333
177-196	7.603333333
197-216	7.446666667
217-236	7.685714286
237-256	7.866666667
257-276	7.7
277-296	7.733333333
297-316	6.6
317-336	7.4
Grand Total	**6.465282215**

Plot showing duration vs IMDB score

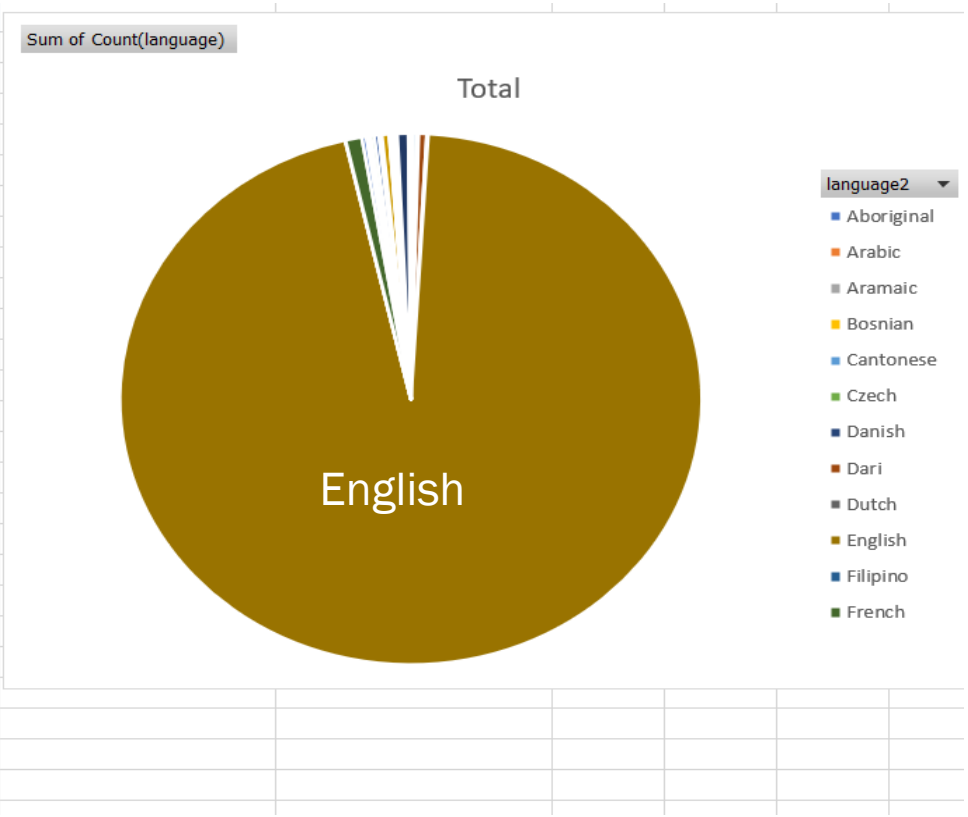
A small value of R indicates that duration of the movie has much less significance on the IMDB score of the movie.

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

Pivot table for language

Row Labels	Sum of Count(language)
Aboriginal	2
Arabic	1
Aramaic	1
Bosnian	1
Cantonese	7
Czech	1
Danish	3
Dari	17
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

Pie chart for distribution



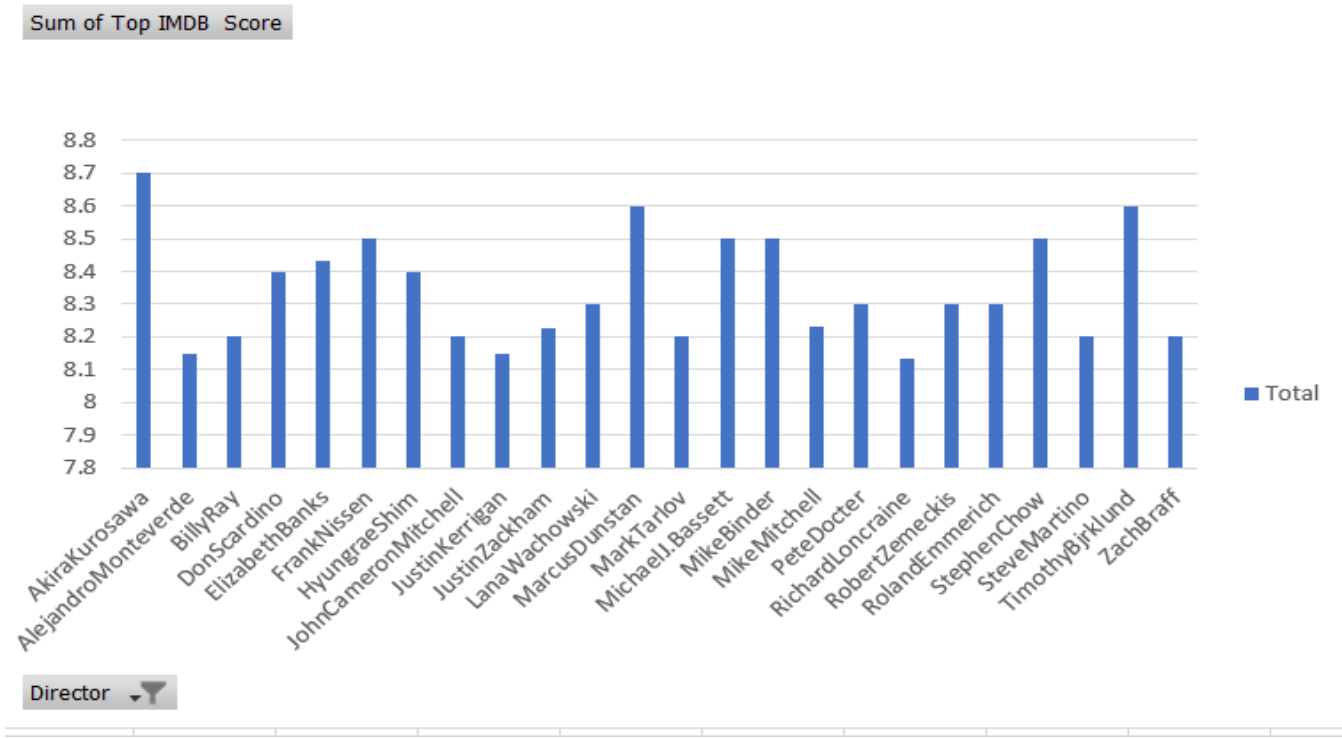
As observing the given pie chart, The **English** language has overwhelming advantage on the IMDB score.
It maybe due to reason that English is a universal language therefore higher the viewer count, more the ratings given.

Director Analysis: Influence of directors on movie ratings.

Pivot Table (Director)

Row Labels	Sum of Top IMDB Score
AkiraKurosawa	8.7
AlejandroMonteverde	8.15
BillyRay	8.2
DonScardino	8.4
ElizabethBanks	8.433333333
FrankNissen	8.5
HyungraeShim	8.4
JohnCameronMitchell	8.2
JustinKerrigan	8.15
JustinZackham	8.225
LanaWachowski	8.3
MarcusDunstan	8.6
MarkTarlov	8.2
MichaelJ.Bassett	8.5
MikeBinder	8.5
MikeMitchell	8.233333333
PeteDocter	8.3
RichardLoncraine	8.133333333
RobertZemeckis	8.3
RolandEmmerich	8.3
StephenChow	8.5
SteveMartino	8.2
TimothyBjrklund	8.6
ZachBraff	8.2
Grand Total	**200.225**

Average Score vs Director(TOP) Chart



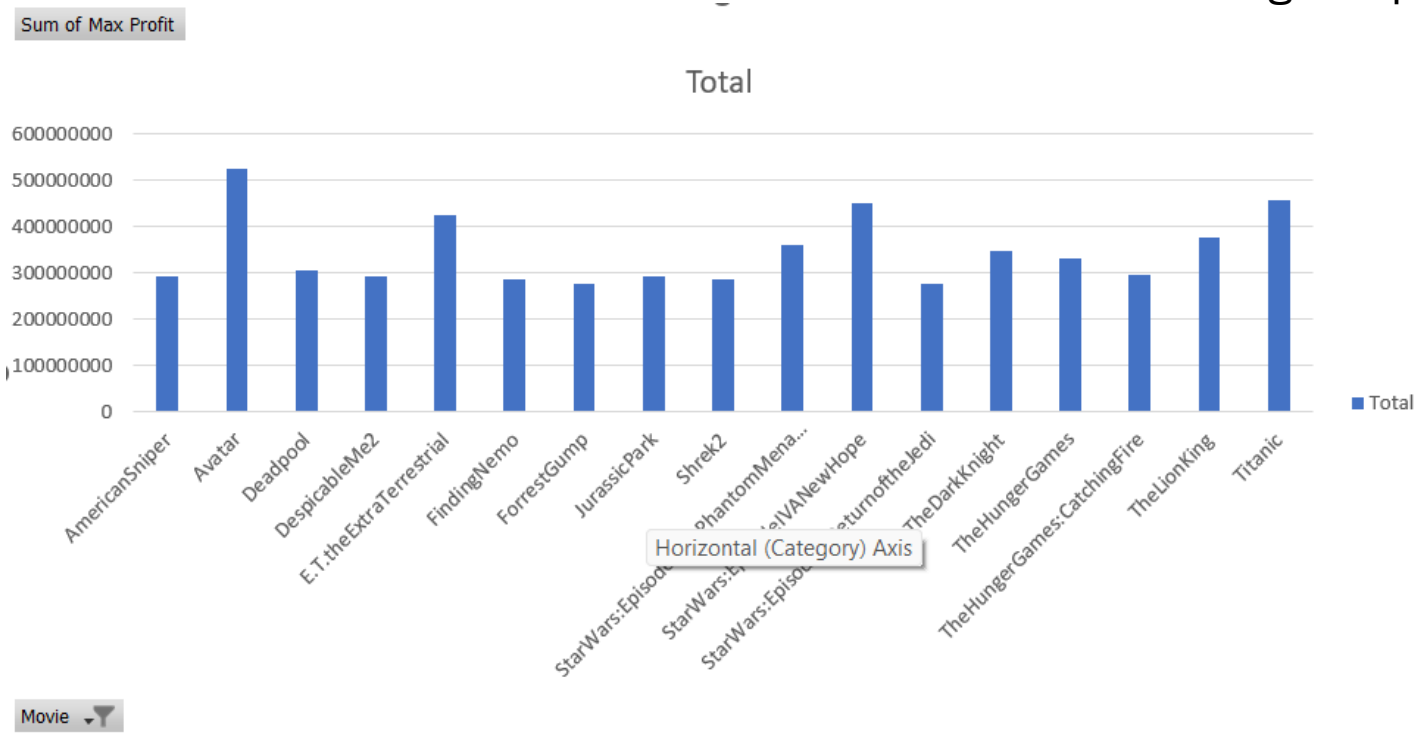
Best director is Akira Kurosawa with mean IMDB score of 8.7

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

Top profit earning companies
Gross - budget

Row Labels	Sum of Max Profit
AmericanSniper	291323553
Avatar	523505847
Deadpool	305024263
DespicableMe2	292049635
E.T.theExtraTerrestrial	424449459
FindingNemo	286838870
ForrestGump	274691196
JurassicPark	293784000
Shrek2	286471036
StarWars:EpisodeIThePhantomMenace	359544677
StarWars:EpisodeIVANewHope	449935665
StarWars:EpisodeVIReturnoftheJedi	276625409
TheDarkKnight	348316061
TheHungerGames	329999255
TheHungerGames:CatchingFire	294645577
TheLionKing	377783777
Titanic	458672302
Grand Total	**5873660582**

Chart showing profit vs movies



Avatar has highest profit.

RESULT

By performing the IMDB Movie Analysis Project

I was able to understand many functions of Excel and able to implement them for calculating descriptive statistics and visualise them using charts and pivot table.

I was able to fulfil all the tasks and the answers posted above are correct to the best of my knowledge.