IMDB Movie Analysis Project

Project Description:

IMDb (Internet Movie Database) is an online database of information related to films, television series, podcasts, home videos, video games, and streaming content online including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews. IMDb began as a fan-operated movie database in 1990, and moved to the Web in 1993. Since 1998, it has been owned and operated by IMDb.com Inc., a subsidiary of Amazon.

Approach:

I went through the Excel data provided by the Trainity IMDB Movie Analysis project and understood all the columns related to the movies in the dataset. Further, I understood the columns and their respective constraints to do the analysis. I was given a set of questions to solve as part of the analysis. By using the Microsoft Excel, I did solve the queries and provided the result as expected.

Tech-Stack Used:

Microsoft Excel 2021 – To answer the queries with the help of Excel Tools and formulas.

Approach:

Started with the data cleaning like:

- Removing null values.
- Removed the columns which we don't use for the analysis.
- Removing the Duplicate rows.
- Autofit the row height and column width

Have used the in-built formulas in excel for the descriptive analysis such as:
Mean – average()

Median – median()

Mode – mode()
Max – max()
Min – min()

Variance – VAR.P()

Standard Deviation - STDEV.P()

With the help of the Excel formulas and pivot tables, I completed the following given tasks.

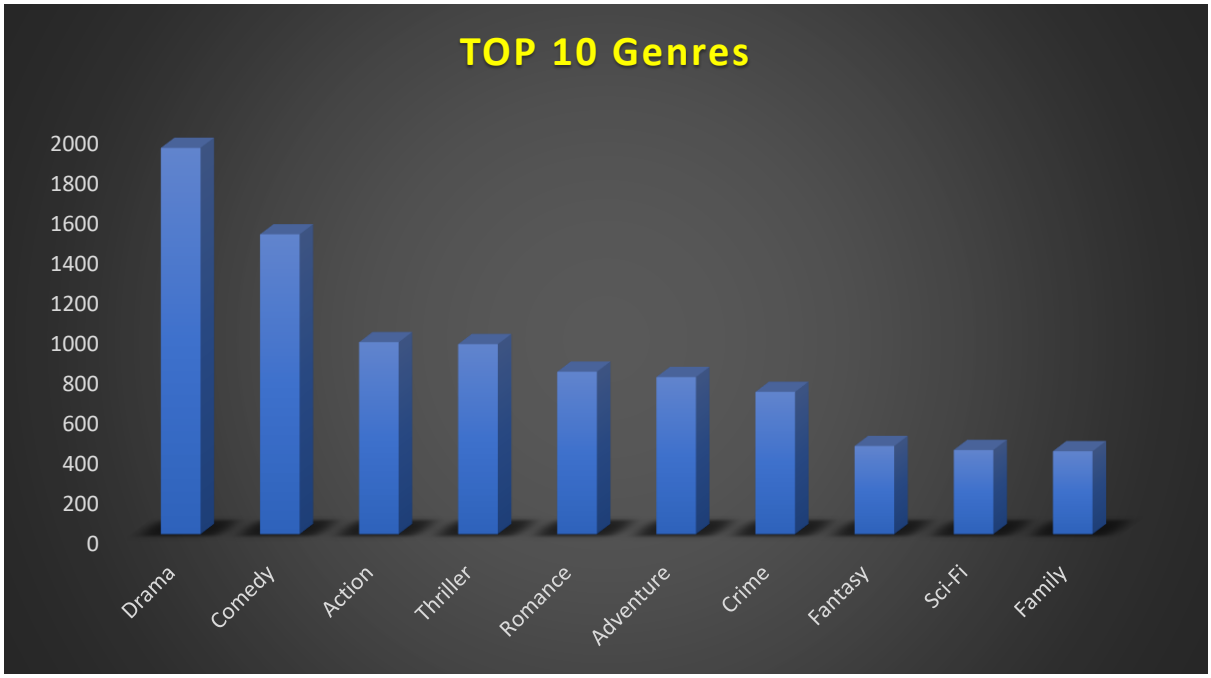
Task A - Genre Analysis:

Determine the most common genres of movies in the dataset.

I used text to columns in the Genre column to separate each genre for further analysis.

After that I found out the count of each genre using the pivot table in Excel.

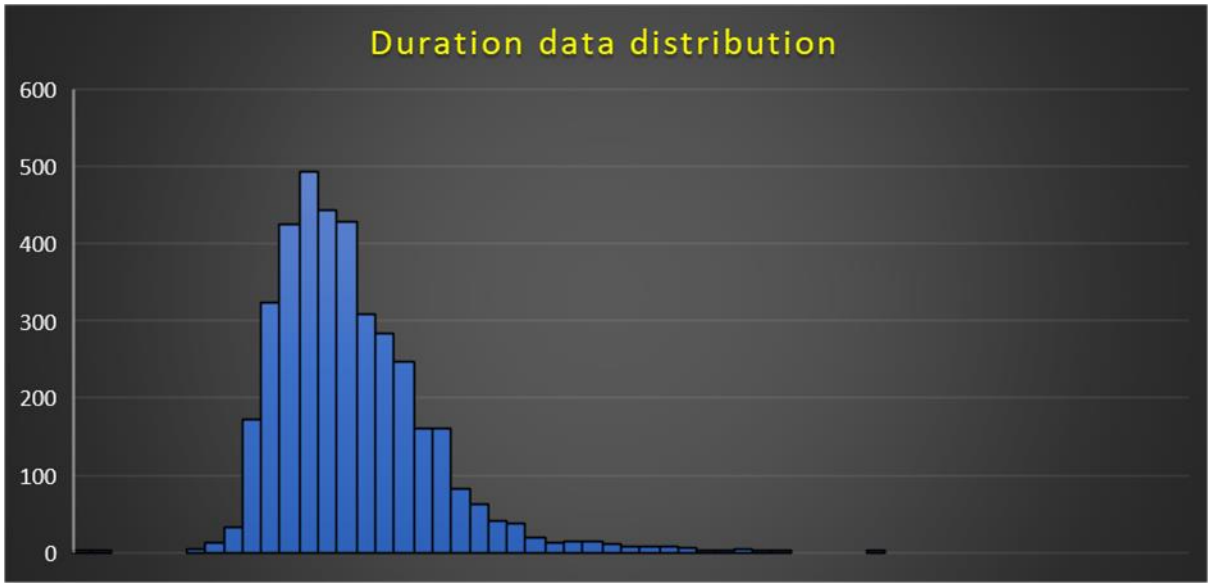
And sorted the pivot table to find out the Top 10 most common genres and created the Column chart for it.



We can see that the most popular genre is "**Drama"** followed by **"Comedy"** and then other genres.

Task B – Movie Duration Analysis:

Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score. Create the Scatter plot as instructed in the task.

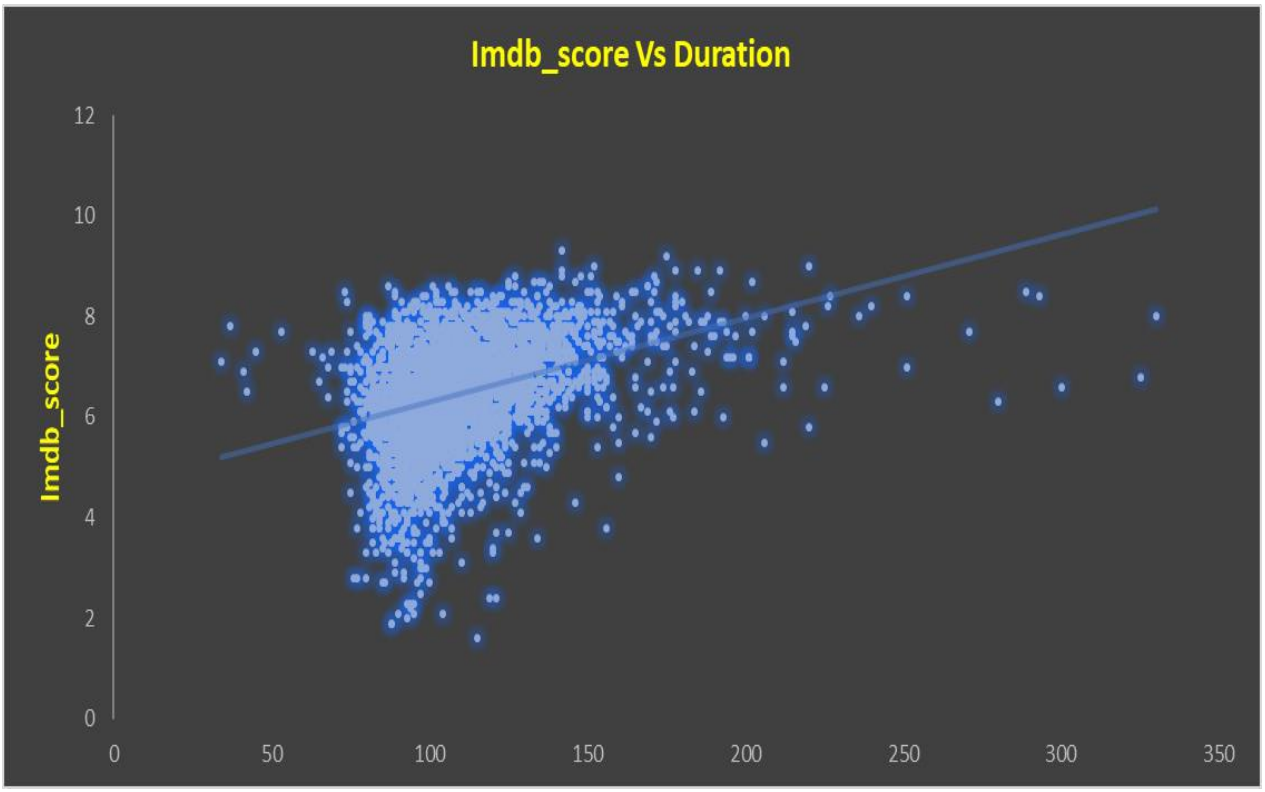


The histogram shows the data distribution in the Duration column. It is positively skewed and has outliers present in the column.

Following is statistical analysis of Duration column:

MEAN	109.91
MEDIAN	106.00
MODE	101.00
MAX	330.00
MIN	34.00
VARIANCE	517.48
STDEV	22.75

This is the scatter plot between Imdb score and Duration.



We could see that the trendline is increasing with the duration increase.

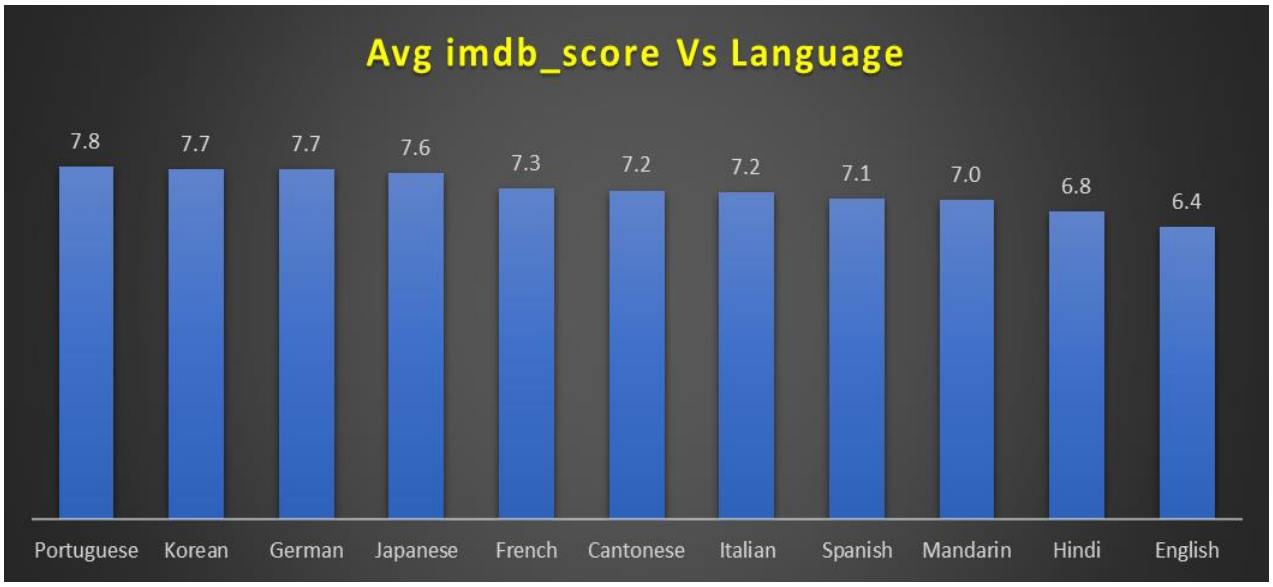
But mostly, the IMDB score is more when the duration is between 80 mins to 150mins.

Task C – Language Analysis:

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

The descriptive analysis of the Top 10 common movie languages is shown below:

	Imdb_score					
Languages	**Count**	**Avg**	**Max**	**Min**	**Var**	**StdDev**
English	3671	6.4	9.3	1.6	1.10	1.05
French	37	7.3	8.4	5.8	0.32	0.56
Spanish	26	7.1	8.2	5.2	0.68	0.83
Mandarin	14	7.0	7.9	5.6	0.59	0.77
German	13	7.7	8.5	6.1	0.41	0.64
Japanese	12	7.6	8.7	6	0.81	0.90
Hindi	10	6.8	8	4.8	1.24	1.11
Cantonese	8	7.2	7.8	6.5	0.19	0.44
Italian	7	7.2	8.9	5.3	1.33	1.16
Korean	5	7.7	8.4	7	0.33	0.57
Portuguese	5	7.8	8.7	6.1	0.96	0.98



The above column chart shows the average imdb score of the Top 10 common movie languages.

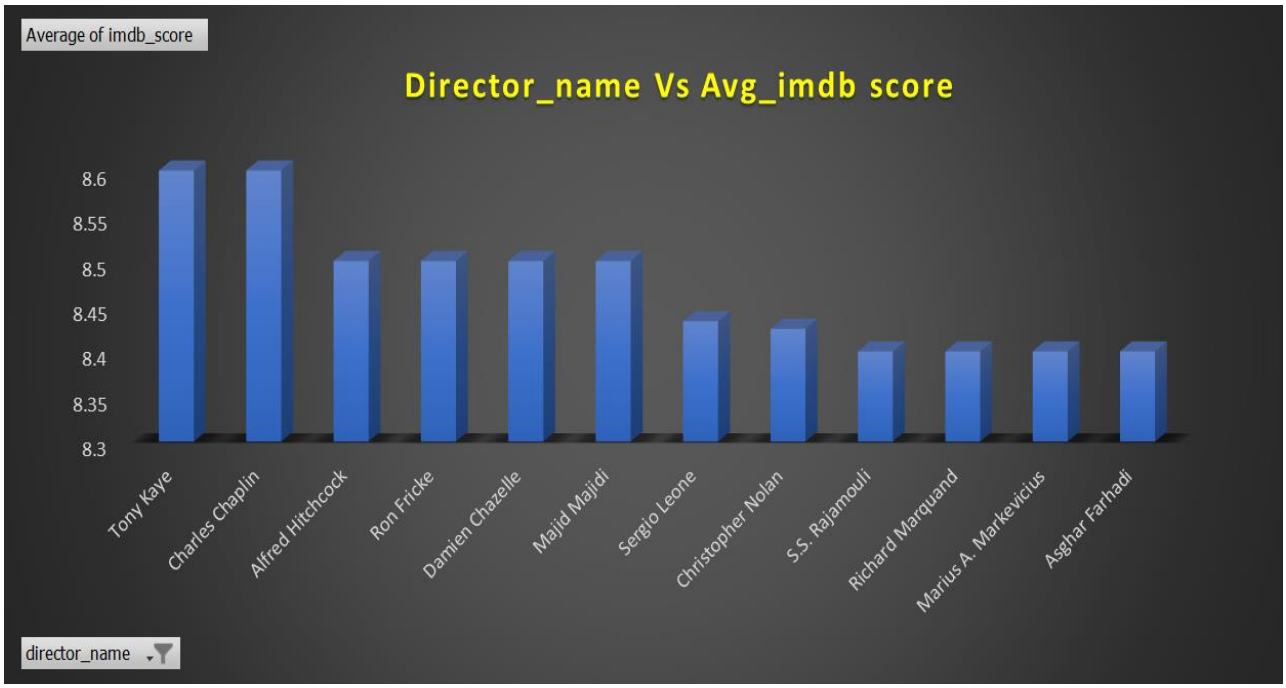
The "**English**" language is the most common language used in the movies with **3671** times present in the dataset and and its average imdb_score is **6.4**. This was found by creating pivot table in excel.

"Portuguese" language movies have the highest imdb score (i.e **7.8**) in the list of most common movie languages.

Task D – Director Analysis:

Identify the top directors based on their average IMDB score.

I have plotted the Bar graph for the Directors with the highest Average Imdb score using Excel pivot tables and pivot chart.

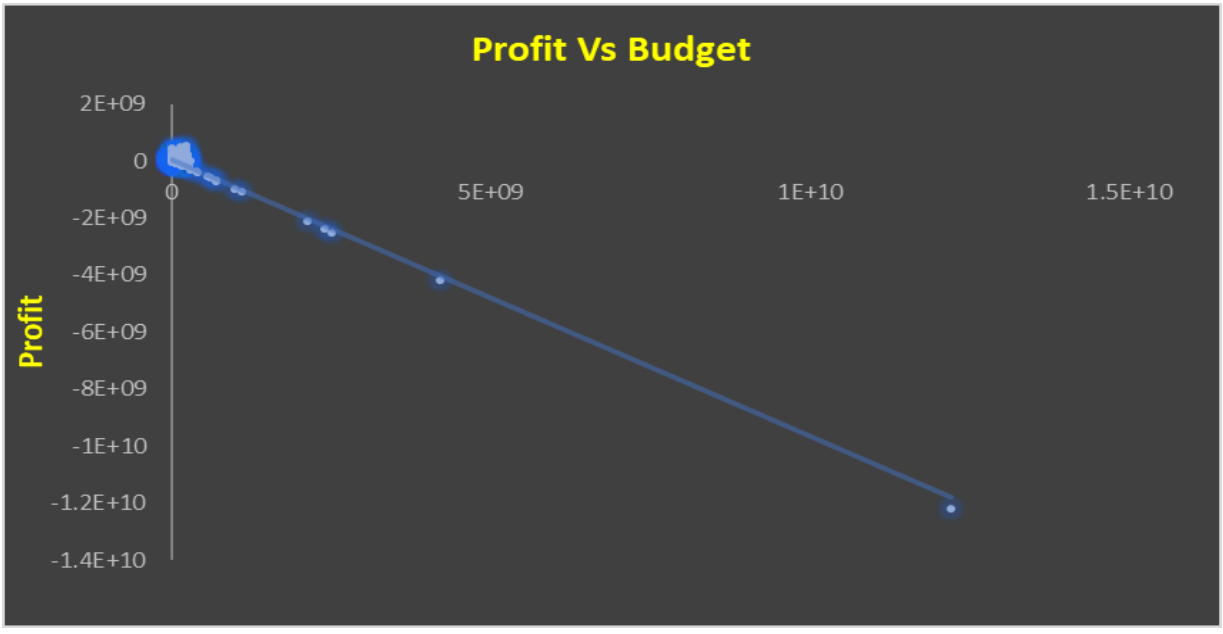


Task E – Budget Analysis:

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

The correlation between Gross earnings and movie Budget is 0.1009, which is very low.

I found the Profit of each movie by the difference of the gross with budget used for the movie.



These are the top 10 Profitable movies in the database:

Movies	Profit_
Avatar	523505847
Jurassic World	502177271
Titanic	458672302
Star Wars: Episode IV - A New Hope	449935665
E.T. the Extra-Terrestrial	424449459
The Avengers	403279547
The Lion King	377783777
The Jungle Book	375290282
Star Wars: Episode I - The Phantom Menace	359544677
The Dark Knight	348316061

The most profitable movie is **Avatar** with a profit of **523,505,847 Dollars (5.2 billion Approx.)**.

Result:

Through this project I was able to understand the formulas being used in the Excel which can be used to find the Statistical measures such as Mean, Median, Mode, Max, Min, Variance and Standard Deviation. I got used to the Excel formulas and how to convert the Raw Data into meaningful insights. And the steps which I used are – cleaning the data, using formulas and pivot tables to find the desired outcome and also learnt how to convert the data into a visualized chart so that the insights can be drawn within seconds by seeing the graphs instead of searching the whole data.

I have achieved the end result and I think I have contributed my full support into the analysis. I hope this project achieve what it was tend to achieve.

Hyperlink for the Excel sheet:

Imdb_movie_analysis