IMDB Movie Analysis

Project Description:

The project is about IMDB Movie Analysis, where we need to investigates the factors influence the success of a movie on IMDB.

The 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.

Approach:

My first approach is to extract & transform methodology on the dataset provided and follow the below steps.

- The data needs to be cleaned, formatted and to check if there are any missing values in the dataset. If there are, decide on the best strategy to handle them.
- Check and identify the outliers and remove any outliers that may have significant impact on the analysis.
- Apply the best strategy to handle outliers. This could be removing them, replacing them, or leaving them as is, depending on the situation.
- After cleaning Perform relevant descriptive statistic calculations to gain a general understanding of dataset
- This could involve calculating averages, medians, or other statistical measures. It could also involve creating visualizations to better understand the data

Tech-Stack Used:

I used Python – Jupiter Notebook Version 7, As it is open source easy to use programming language platform, Where I can easily extract & transform the data with the help of Pandas- Dataframe libraries. As well as I can define the statistical functions and get the outcomes easily.

Also in terms of visualizations one can visualize the data with the help of Seaborn, Matplotlib libraries to get the meaningful insight from the data.

Insights:

I performed Exploratory Data Analysis on dataset and found below main points.

- The dataset having large amount of missing data, so I checked each and every columns with missing values to check the impact of those missing values columns with other important columns.

- As a result I found gross & budget columns having large missing values and I tried to keep them and replace with 0's but then I saw we might have incorrect data for analysis, so i removed them with other missing attributes.
- In the Language analysis I found English has huge impact on IMDB ratings in both positive & negative way, as the language have most outliers comparing to other languages.
- During Duration analysis can see between 100-150 duration having most IMDB ratings from 5 to 8.
- Christopher Nolan, Alfred Hitchcock, Tony Kaye & Sergio Leone are the best directors on which producers can make decision, as these directors have made 3 or more movies and having good IMDB ratings. Others top directors just have only 1 movie as per the dataset, so we cant make decision based on single success movie.
- Also in the Budget analysis most of the movies which have earned 99% of profit have low IMDB ratings ranging between 2 to 6.
- Based on the above findings we can say IMDB ratings are not rely on single attributes. Most of them are inversely proportional.
- Considering the success of movie depends on many things like Director, Language, Duration, Budget, Actors etc.

Result:

The project helps me to understand how to perform Data Analysis based on all the features of the Movies. As if we stick with one feature then we might have to lose the other important factors. Movie success depends on different factors to get attention of every categories audience, some like Horror movies, others like Drama depending on these categories audience. Also the time duration of movies, their Language if having Multi language might have get more audience, so considering all these factors a Data Analyst have to extract meaningful insights from the data to help producers, make a successful movies.

.

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.

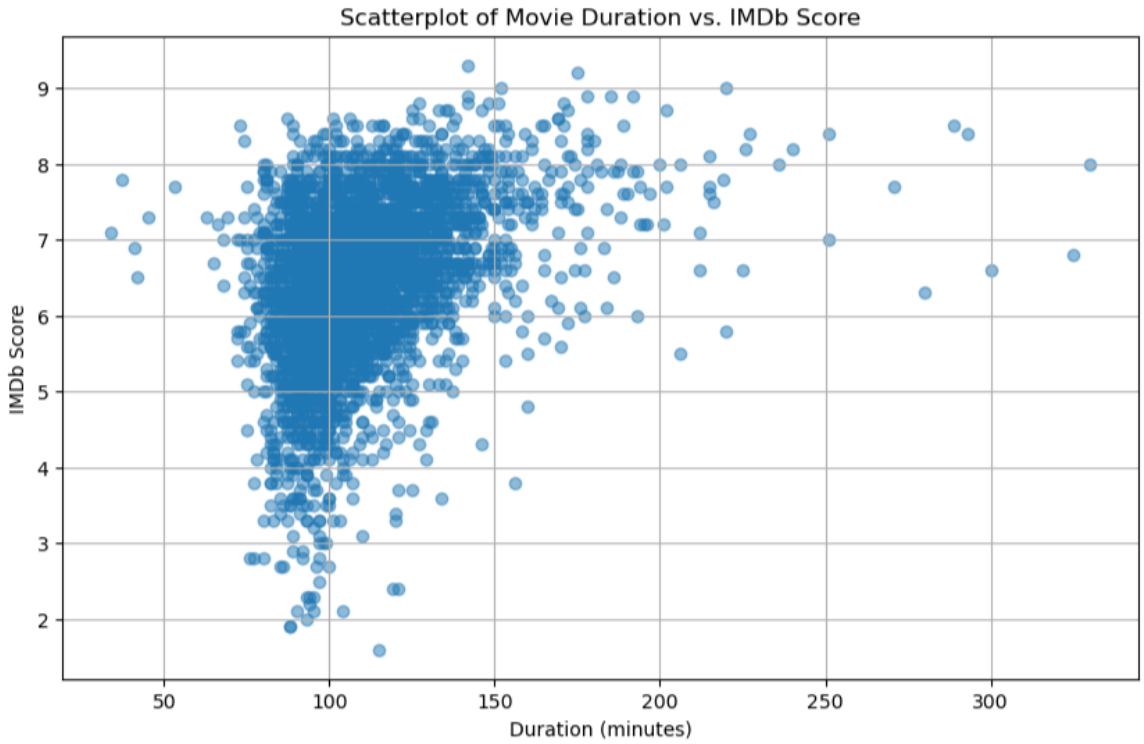
```
Most common genres:
Drama                                  152
Comedy|Drama|Romance                   149
Comedy|Drama                           147
Comedy                                 145
Comedy|Romance                         135
Drama|Romance                          118
Crime|Drama|Thriller                    80
Action|Crime|Thriller                   54
Action|Crime|Drama|Thriller             48
Comedy|Crime                            45
Action|Adventure|Sci-Fi                 45
Action|Adventure|Thriller               43
Horror                                  41
Crime|Drama                             41
Drama|Thriller                          40
Crime|Drama|Mystery|Thriller            40
Action|Adventure|Sci-Fi|Thriller        33
Horror|Thriller                         32
Horror|Mystery|Thriller                 31
Biography|Drama                         30
```

	Mean	Median	Mode	Range	Variance	Std Deviation
Drama	6.789005	6.9	6.7	7.2	0.794389	0.891285
Comedy|Drama|Romance	6.513204	6.6	6.7	7.4	1.066123	1.032532
Comedy|Drama	6.517128	6.6	6.7	7.4	1.062472	1.030763
Comedy	6.182763	6.3	6.3	6.9	1.081709	1.040053
Comedy|Romance	6.301441	6.4	6.7	6.9	1.076757	1.037669
Drama|Romance	6.673146	6.8	7.1	7.2	0.909898	0.953886
Crime|Drama|Thriller	6.616898	6.7	6.7	7.2	0.936684	0.967824
Action|Crime|Thriller	6.409516	6.5	6.6	7.2	1.104428	1.050917
Action|Crime|Drama|Thriller	6.578028	6.7	6.7	7.2	0.998129	0.999064
Comedy|Crime	6.311207	6.4	6.3	7.4	1.085044	1.041655
Action|Adventure|Sci-Fi	6.362007	6.4	6.6	7.1	1.173826	1.083432

Findings : Drama is the most common genres of movies in the dataset followed by Comedy, Romance etc

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

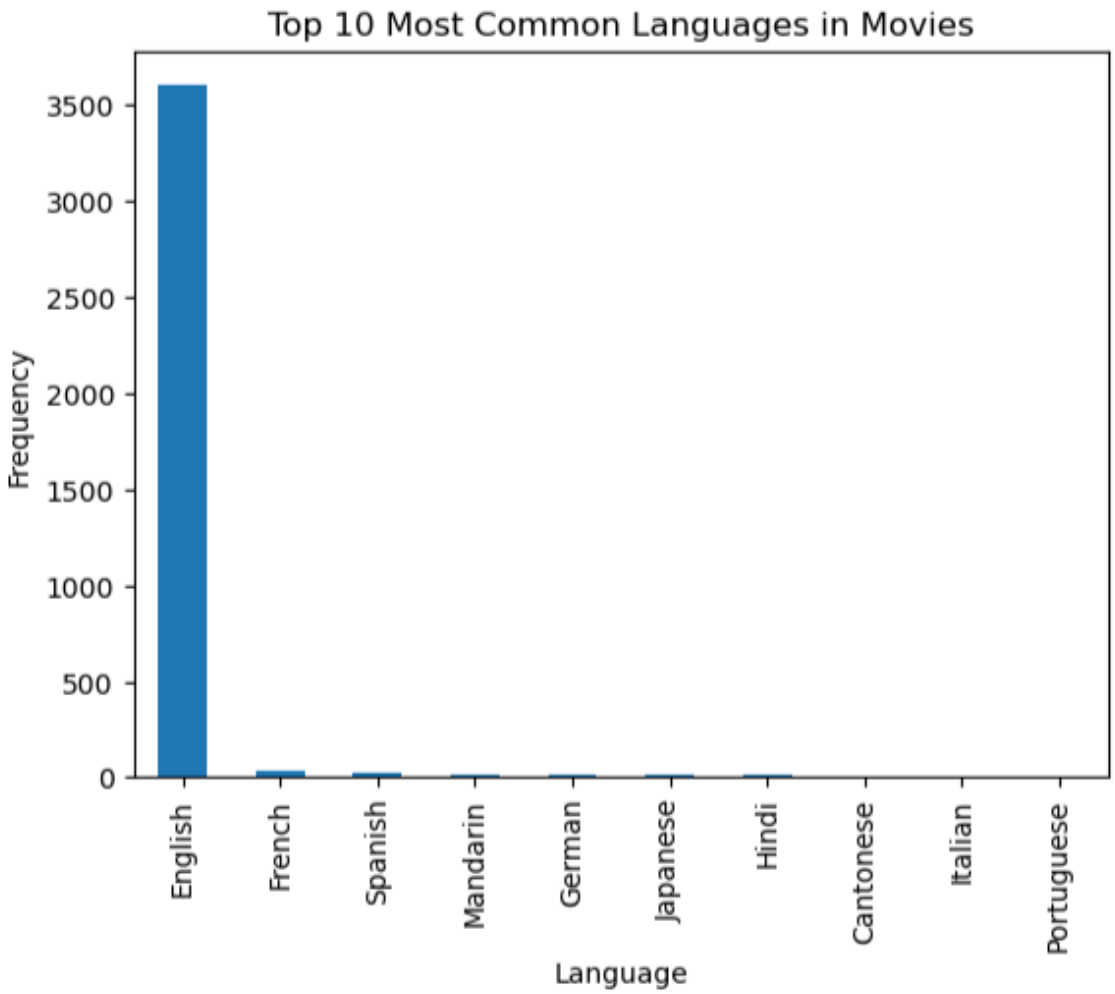
Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.



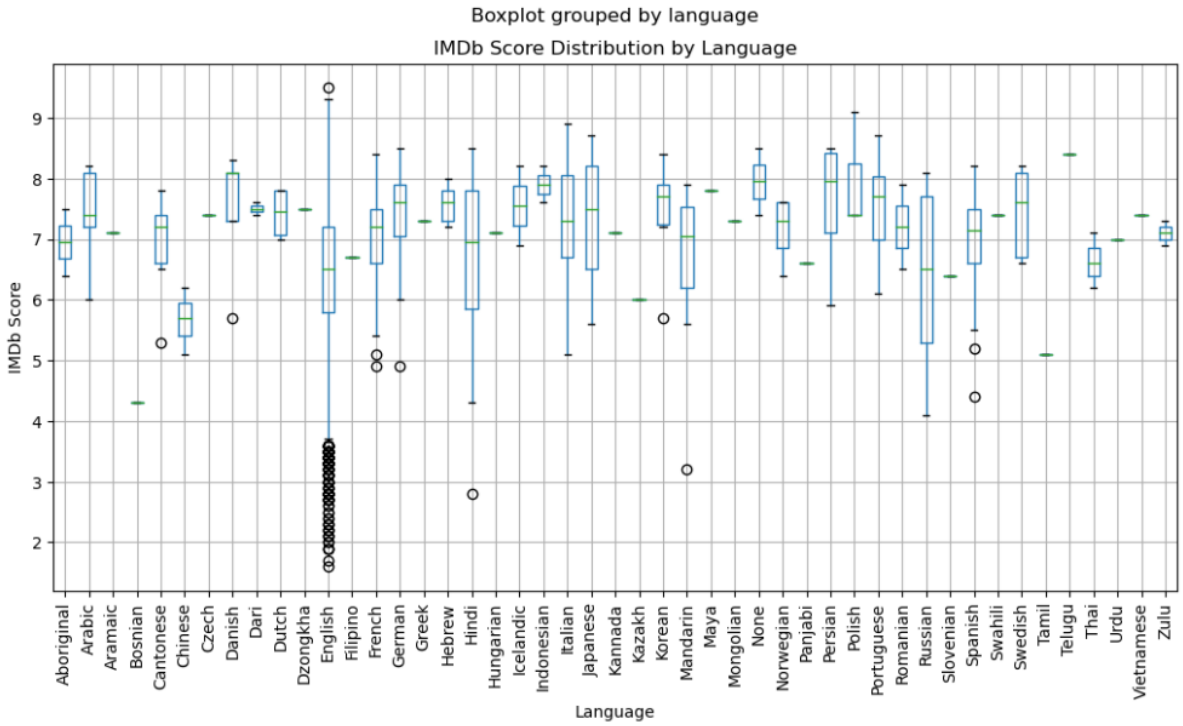
Findings: From the above scatter plot diagram we can most of the mid to high IMDB ratings lies in range 100-150 duration. We can see slight increase in duration have more ratings as well.

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.



language	count	mean	std	min	25%	50%	75%	max
Aboriginal	2.0	6.950000	0.777817	6.4	6.675	6.95	7.225	7.5
Arabic	1.0	7.200000	NaN	7.2	7.200	7.20	7.200	7.2
Aramaic	1.0	7.100000	NaN	7.1	7.100	7.10	7.100	7.1
Bosnian	1.0	4.300000	NaN	4.3	4.300	4.30	4.300	4.3
Cantonese	8.0	7.237500	0.440576	6.5	7.075	7.30	7.525	7.8
Czech	1.0	7.400000	NaN	7.4	7.400	7.40	7.400	7.4
Danish	3.0	7.900000	0.529150	7.3	7.700	8.10	8.200	8.3
Dari	2.0	7.500000	0.141421	7.4	7.450	7.50	7.550	7.6
Dutch	3.0	7.566667	0.404145	7.1	7.450	7.80	7.800	7.8
Dzongkha	1.0	7.500000	NaN	7.5	7.500	7.50	7.500	7.5
English	3602.0	6.420850	1.052605	1.6	5.800	6.50	7.100	9.3
Filipino	1.0	6.700000	NaN	6.7	6.700	6.70	6.700	6.7
French	37.0	7.286486	0.561329	5.8	6.900	7.20	7.700	8.4
German	13.0	7.692308	0.640913	6.1	7.400	7.70	8.300	8.5
Hebrew	3.0	7.500000	0.435890	7.2	7.250	7.30	7.650	8.0
Hindi	10.0	6.760000	1.111755	4.8	6.050	7.05	7.700	8.0
Hungarian	1.0	7.100000	NaN	7.1	7.100	7.10	7.100	7.1



Findings: English is the most common language used in movies. Also it have large amount of outliers shown in above boxplot diagram which are lowest IMDB ratings. On the other hand English has the highest IMDB rating as well.

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.

```
Top Directors based on Average IMDb Score:
              director_name  imdb_score  Percentile
216          Charles Chaplin    8.600000           9
1670               Tony Kaye    8.600000           9
45           Alfred Hitchcock    8.500000           9
1435               Ron Fricke    8.500000           9
1014            Majid Majidi    8.500000           9
302          Damien Chazelle    8.500000           9
1493            Sergio Leone    8.433333           9
260        Christopher Nolan    8.425000           9
1032  Marius A. Markevicius    8.400000           9
1462          S.S. Rajamouli    8.400000           9


Percentile Counts:
0    193
1    190
2    142
3    202
4    147
5    174
6    189
7    160
8    189
9    161
Name: Percentile, dtype: int64
```

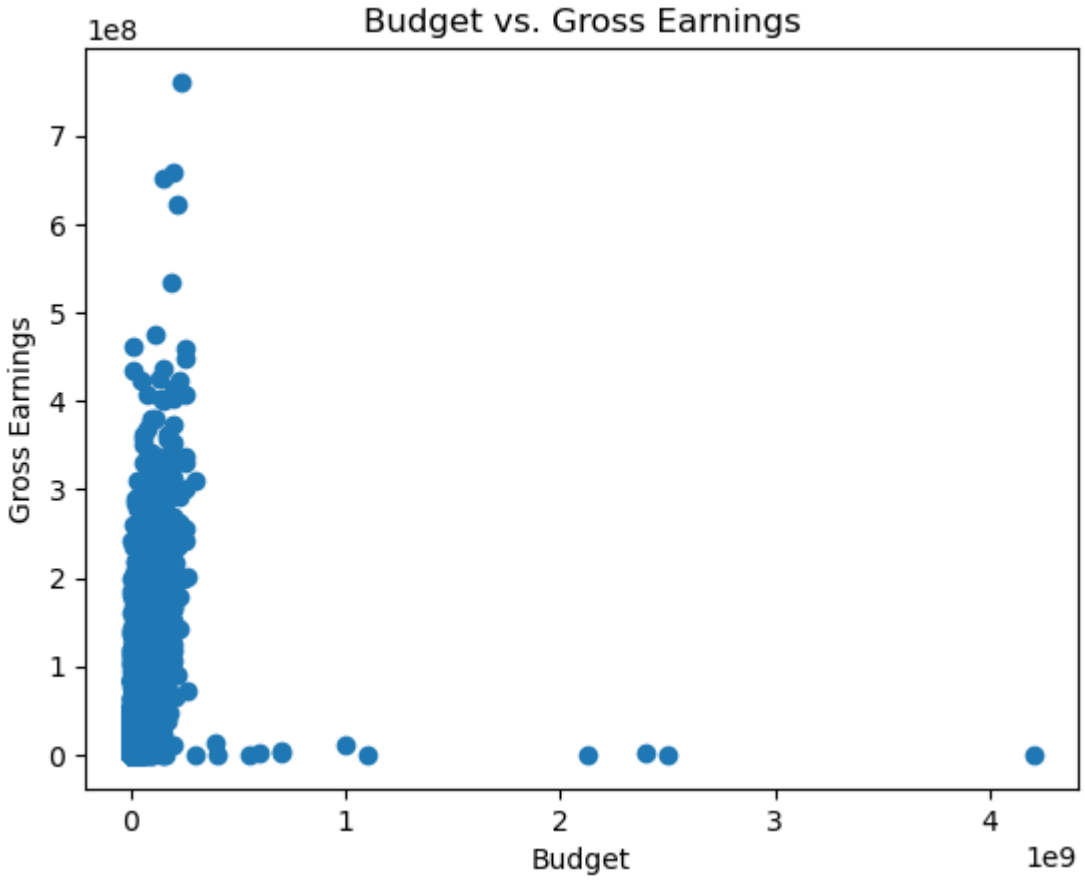
Findings : Christopher Nolan, Alfred Hitchcock, Tony Kaye & Sergio Leone are the directors have made 3 or more movies and having good IMDB ratings. Other top directors just have only 1 movie as per the dataset.

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.

Top 10 Movies with the Highest Profit Margin:

	movie_title	Profit Margin
4793	Paranormal Activity	99.986100
4799	Tarnation	99.963177
4707	The Blair Witch Project	99.957305
4984	The Brothers McMullen	99.756017
3278	The Texas Chain Saw Massacre	99.729311
5035	El Mariachi	99.657017
4956	The Gallows	99.560591
4977	Super Size Me	99.436222
2492	Halloween	99.361702
4674	American Graffiti	99.324348



Findings : As per the above scatter plot diagram we conclude that the movies with low budget have made 99% of profit. So here budget & Gross are inversely proportional.