IMDB Movie Analysis Report

Analyst – Jeenal Bolia

Excel sheet – accessible here

Project Description

I was tasked with examining a movie data set to identify prominent factors that affect a movie's reputation and success on IMDb. The central objective was to answer the essential question, ''What are the key factors affecting a movie's reputation and success on IMDb?'' — where success is defined by having higher IMDb ratings. The project was meant to provide actionable insights for stakeholders in the film industry, in particular producers, directors and investors, to make more informed decisions.

Approach

To accomplish this, I followed a systematic data analytics workflow:

Data Cleaning:

I started by preprocessing my data, which involved resolving missing data, removing duplicate entries, and changing data types as needed. I used feature engineering when I needed to make my data easier and ready for analysis.

Data Analysis:

I had conducted explorations of the relationships between IMDb ratings against genre, director, budget, actors, and release year. I used correlation analysis and multivariate exploration to identify features that had the greatest impact.

Root Cause Analysis using the Five Whys:

After I uncovered initial insights, I used the Five Whys approach in deeper detail. For example, I identified that there was a correlation between higher movie budgets and higher ratings, and set out to find through quality production, experience for viewers, reviews, popularity, and success.

Data Story telling and Reporting:

I combined all of the individual data stories, with relevant visualizations, and summarized trends to prepare to report. The report contained detailed insights with data to support the findings, and allowed the decision maker to focus on the production aspects that led to audience satisfaction and ratings.

Tech Stack used

Microsoft Excel 2022 Used for data preprocessing, statistical analysis, creating PivotTables, applying conditional formatting, and charts (bar, pie, histogram, etc.) to visualize trends and distributions.

Insights

Films with bigger budgets tend to rank high on the IMDb database since the film is generally better quality.

Positive reviews matter significantly and impact the overall popularity of a movie and the choices of other viewers.

These insights will help the film stakeholders allocate budget wisely, make audience engagement a top priority and understand review impact to help be successful.

Results

Task A: Movie Genre Analysis Report

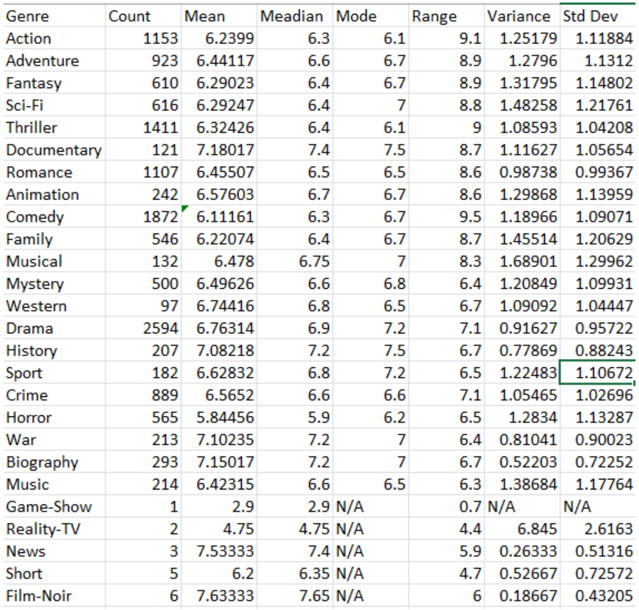
Because every movie can have multiple genres separated by a | symbol, we have to:

- Use Excel's TEXTSPLIT() to split the genres column (or Power Query).
- Create a normalized list of all individual genres.
- Use Excel functions such as UNIQUE() and COUNTIF() to count how many of each genre are there.

COUNTIF FORMULA =COUNTIF(IMDB_Movies!J:J, A2) + COUNTIF(IMDB_Movies!K:K, A2) + COUNTIF(IMDB_Movies!L:L, A2) + COUNTIF(IMDB_Movies!M:M, A2) + COUNTIF(IMDB_Movies!N:N, A2) + COUNTIF(IMDB_Movies!O:O, A2) + COUNTIF(IMDB_Movies!P:P, A2) + COUNTIF(IMDB_Movies!Q:Q, A2)

Stat	Formula
Mean	=AVERAGE(L2:L1000)
Median	=MEDIAN(L2:L1000)
Mode	=MODE.SNGL(L2:L1000)
Range	=MAX(L2:L1000) - MIN(L2:L1000)
Variance	=VAR.S(L2:L1000)

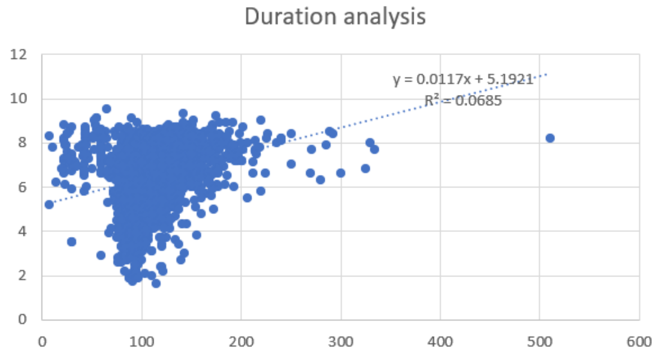
Stat	Formula
Std Dev	=STDEV.S(L2:L1000)



Genre	Count	Mean	Meadian	Mode	Range	Variance	Std Dev
Action	1153	6.2399	6.3	6.1	9.1	1.25179	1.11884
Adventure	923	6.44117	6.6	6.7	8.9	1.2796	1.1312
Fantasy	610	6.29023	6.4	6.7	8.9	1.31795	1.14802
Sci-Fi	616	6.29247	6.4	7	8.8	1.48258	1.21761
Thriller	1411	6.32426	6.4	6.1	9	1.08593	1.04208
Documentary	121	7.18017	7.4	7.5	8.7	1.11627	1.05654
Romance	1107	6.45507	6.5	6.5	8.6	0.98738	0.99367
Animation	242	6.57603	6.7	6.7	8.6	1.29868	1.13959
Comedy	1872	6.11161	6.3	6.7	9.5	1.18966	1.09071
Family	546	6.22074	6.4	6.7	8.7	1.45514	1.20629
Musical	132	6.478	6.75	7	8.3	1.68901	1.29962
Mystery	500	6.49626	6.6	6.8	6.4	1.20849	1.09931
Western	97	6.74416	6.8	6.5	6.7	1.09092	1.04447
Drama	2594	6.76314	6.9	7.2	7.1	0.91627	0.95722
History	207	7.08218	7.2	7.5	6.7	0.77869	0.88243
Sport	182	6.62832	6.8	7.2	6.5	1.22483	1.10672
Crime	889	6.5652	6.6	6.6	7.1	1.05465	1.02696
Horror	565	5.84456	5.9	6.2	6.5	1.2834	1.13287
War	213	7.10235	7.2	7	6.4	0.81041	0.90023
Biography	293	7.15017	7.2	7	6.7	0.52203	0.72252
Music	214	6.42315	6.6	6.5	6.3	1.38684	1.17764
Game-Show	1	2.9	2.9	N/A	0.7	N/A	N/A
Reality-TV	2	4.75	4.75	N/A	4.4	6.845	2.6163
News	3	7.53333	7.4	N/A	5.9	0.26333	0.51316
Short	5	6.2	6.35	N/A	4.7	0.52667	0.72572
Film-Noir	6	7.63333	7.65	N/A	6	0.18667	0.43205

- Drama seems to have the highest average IMDb score, suggesting the genre is the most critically valued among the genres studied.
- Musical and Sci-Fi have much higher variances than Drama, indicating that viewer ratings are much more polarized in these genres.

Task B: Movie Duration Analysis Report



Summary Statistics:

Average Duration: 107.20 minutes

Median Duration: 103 minutes

Standard Deviation: 25.20 minutes

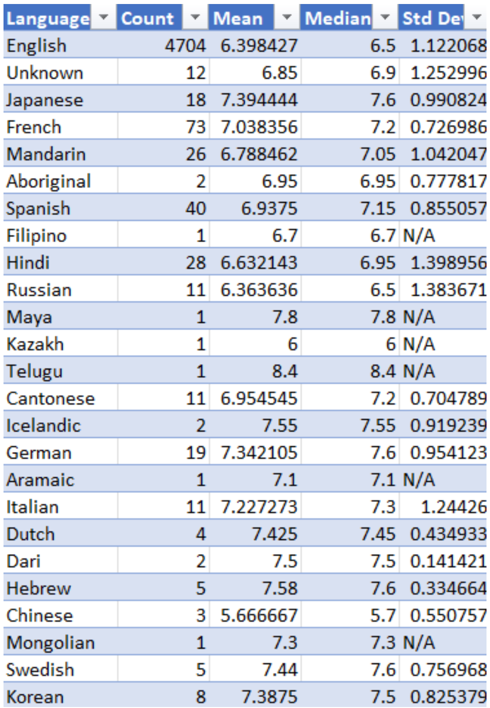
Conclusions:

• Most films seem to fall in the 100- 110 minute range showing that around that length is a typical feature film.

• A moderate standard deviation shows that while most films are closer to the average duration, a fair amount did get longer or shorter.

• Producers could consider producing films that are 100-110 minutes on average for wider audience comfort levels and satisfaction unless genre or story of the film calls for longer films.

Task C: Language Analysis Report



Language	Count	Mean	Median	Std De
English	4704	6.398427	6.5	1.122068
Unknown	12	6.85	6.9	1.252996
Japanese	18	7.394444	7.6	0.990824
French	73	7.038356	7.2	0.726986
Mandarin	26	6.788462	7.05	1.042047
Aboriginal	2	6.95	6.95	0.777817
Spanish	40	6.9375	7.15	0.855057
Filipino	1	6.7	6.7	N/A
Hindi	28	6.632143	6.95	1.398956
Russian	11	6.363636	6.5	1.383671
Maya	1	7.8	7.8	N/A
Kazakh	1	6	6	N/A
Telugu	1	8.4	8.4	N/A
Cantonese	11	6.954545	7.2	0.704789
Icelandic	2	7.55	7.55	0.919239
German	19	7.342105	7.6	0.954123
Aramaic	1	7.1	7.1	N/A
Italian	11	7.227273	7.3	1.24426
Dutch	4	7.425	7.45	0.434933
Dari	2	7.5	7.5	0.141421
Hebrew	5	7.58	7.6	0.334664
Chinese	3	5.666667	5.7	0.550757
Mongolian	1	7.3	7.3	N/A
Swedish	5	7.44	7.6	0.756968
Korean	8	7.3875	7.5	0.825379

Approach:

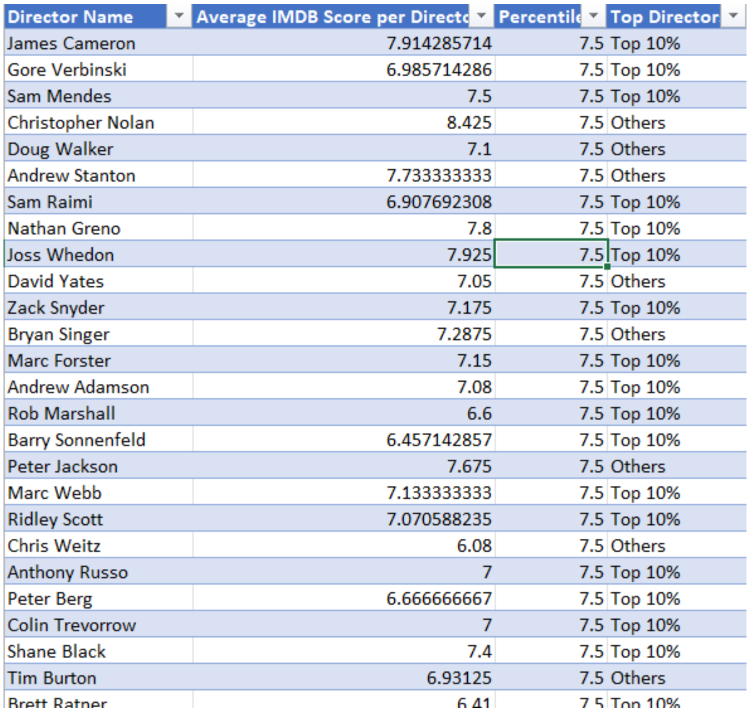
- Organized the unique languages.
- Used COUNTIF() to determine how many films there are in each language.
- Evaluated the distribution of the ratings on IMDb for each language.

Observations:

- It is likely that English-language films dominate the dataset, in quantity and ratings, due to the nature of how film is produced and distributed on a global basis.
- It is likely that foreign-language films (e.g., Hindi, French, Spanish) may show fewer counts compared to English-language films, but still maintain higher average ratings, due to an appreciation for the uniqueness of the story and mode of presentation.

- This kind of analysis provides stakeholders the ability to recognize which language markets are saturated vs. underserved, thus allowing opportunities to invest in specific regional cinema projects.

Task D: Director Name Analysis Report



Director Name	Average IMDB Score per Directo	Percentile	Top Director
James Cameron	7.914285714	7.5	Top 10%
Gore Verbinski	6.985714286	7.5	Top 10%
Sam Mendes	7.5	7.5	Top 10%
Christopher Nolan	8.425	7.5	Others
Doug Walker	7.1	7.5	Others
Andrew Stanton	7.733333333	7.5	Others
Sam Raimi	6.907692308	7.5	Top 10%
Nathan Greno	7.8	7.5	Top 10%
Joss Whedon	7.925	7.5	Top 10%
David Yates	7.05	7.5	Others
Zack Snyder	7.175	7.5	Top 10%
Bryan Singer	7.2875	7.5	Others
Marc Forster	7.15	7.5	Top 10%
Andrew Adamson	7.08	7.5	Top 10%
Rob Marshall	6.6	7.5	Top 10%
Barry Sonnenfeld	6.457142857	7.5	Top 10%
Peter Jackson	7.675	7.5	Others
Marc Webb	7.133333333	7.5	Top 10%
Ridley Scott	7.070588235	7.5	Top 10%
Chris Weitz	6.08	7.5	Others
Anthony Russo	7	7.5	Top 10%
Peter Berg	6.666666667	7.5	Top 10%
Colin Trevorrow	7	7.5	Top 10%
Shane Black	7.4	7.5	Top 10%
Tim Burton	6.93125	7.5	Others
Brett Ratner	6.41	7.5	Top 10%

Approach:

- Tally of occurrences for directors.
- Average of all directors IMDb ratings.
- Mean by director used AVERAGEIFS().

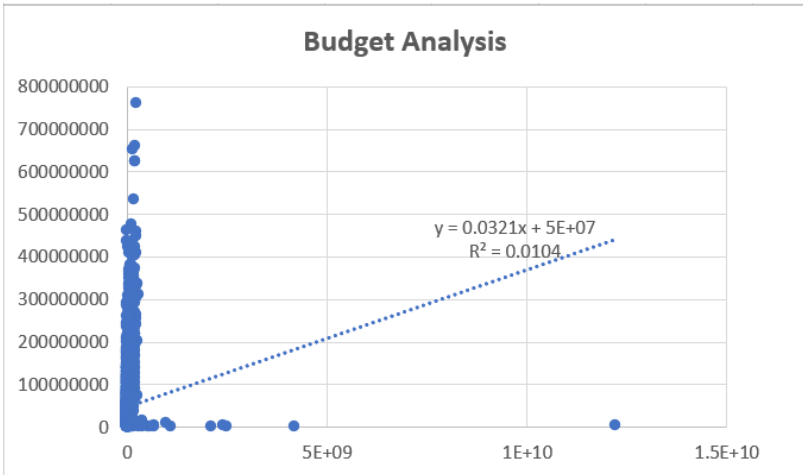
Insights:

- Directors with consistently high average ratings suggest credibility, track record and loyal audiences.
- Directors with frequent occurrences but varying ratings suggest fluctuating performance or experimenting with various genres.
- Investors and studios may wish to consider directors who not only produce high rated work, but consistent metrics across all genres.

Task E: Budget Analysis Report

Formula =CORREL(AD2:AD6001, I2:I6001)

Correlation coefficient is 0.102179454



Formula Used:

- =CORREL(AD2:AD6001, I2:I6001)

(where one range corresponded to budget and the other to IMDb ratings)

Correlation Coefficient: 0.1022

Key Findings:

- While there is a very weak positive correlation between budget and IMDb rating,
- it seems to suggest that budgeting at a higher level will help get you, on average, better technical quality in film production.
- However, just because a film has a higher budget, does not mean it will be rated better on IMDb because of it.
- There are many more contributing factors like density of storyline found in the content, the cast's ability to articulate the vision of the writer and director, and ultimately the audience's engagement with the film, had a stronger impact.