IMDB Movie Analysis

Project Description

A large dataset having IMDB scores of movies of different languages, with details of directors, actors, budgets, gross earnings, duration and genres is used to analyse the most favorite genres, best directors, financial success. In addition to it, language and duration analysis was also done based on Descriptive statistics.

Approach

First the given dataset is cleaned. The required data for each analysis is loaded into separate spreadsheets. Statistical Functions, Pivot table, bar charts, scatter plots are adopted in excel for carrying out data analysis and visualization.

Tech-Stack Used

Microsoft Excel 2021 was used because it has many functions to carry out statistical calculations and visualization tools like charts and plots.

Insights

The insights are provided in the document along with each of the five given analysis.

Cleaning the dataset

1. 126 Duplicates are identified based on the movie title and removed.
2. 19 unwanted columns are deleted.
3. 12 rows having no language are removed.
4. 14 rows having no duration are removed.
5. 99 rows with no director name are removed.
6. 744 records without gross earnings are removed.
7. 262 records without budget are removed. Finally, 3785 records are remained after cleaning.
8. Movie titles are edited by removing Â using the function =SUBSTITUTE(D2,” Â ”,” “)

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.
- **In genre analysis, the short and Film-**Noir genres are excluded because they have very less number of movies.

Functions used

Column C of IMDB_movies sheet contains the genres. Column B of Genre analysis sheet contains the unique list of all genres. Column G of IMDB_movies sheet has the IMDB scores

Average of IMDB = =AVERAGEIF(IMDB_Movies!C:C,"*"&'Genre analysis'!B2&"*",IMDB_Movies!G:G)

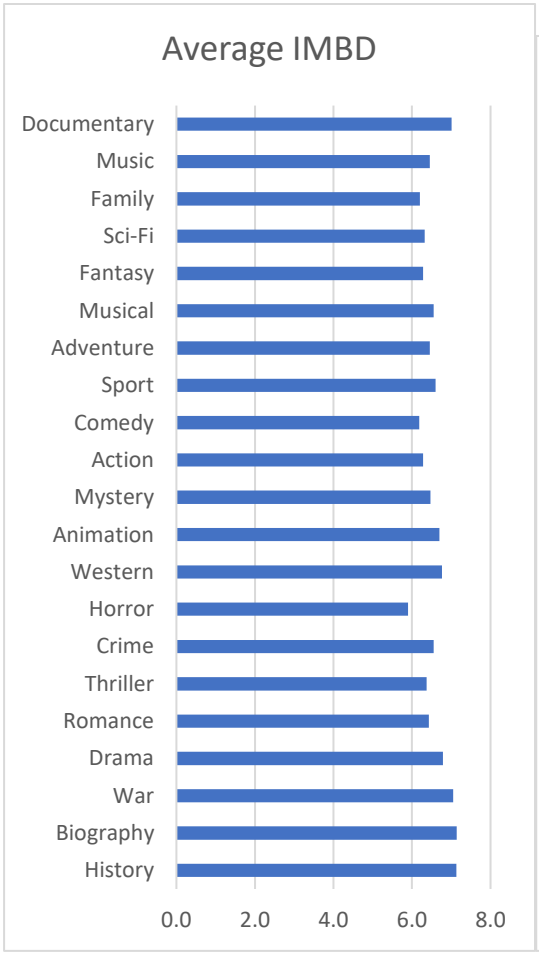
Mode of IMDB score
=MODE.SNGL(IF(ISNUMBER(SEARCH($B2,IMDB_Movies!$C:$C)),IMDB_Movies!$G:$G))

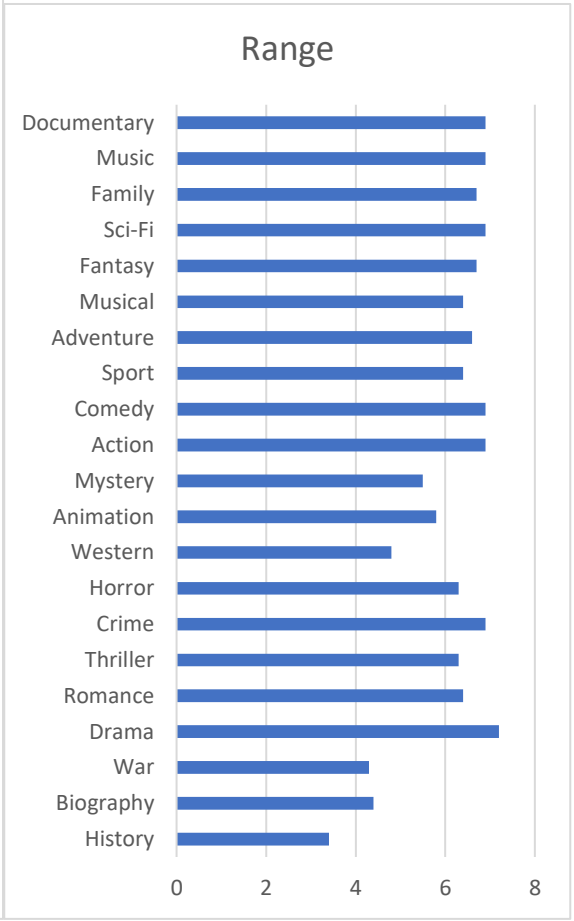
Median of IMDB score =
=MEDIAN(IF(ISNUMBER(SEARCH($B2,IMDB_Movies!$C:$C)),IMDB_Movies!$G:$G))

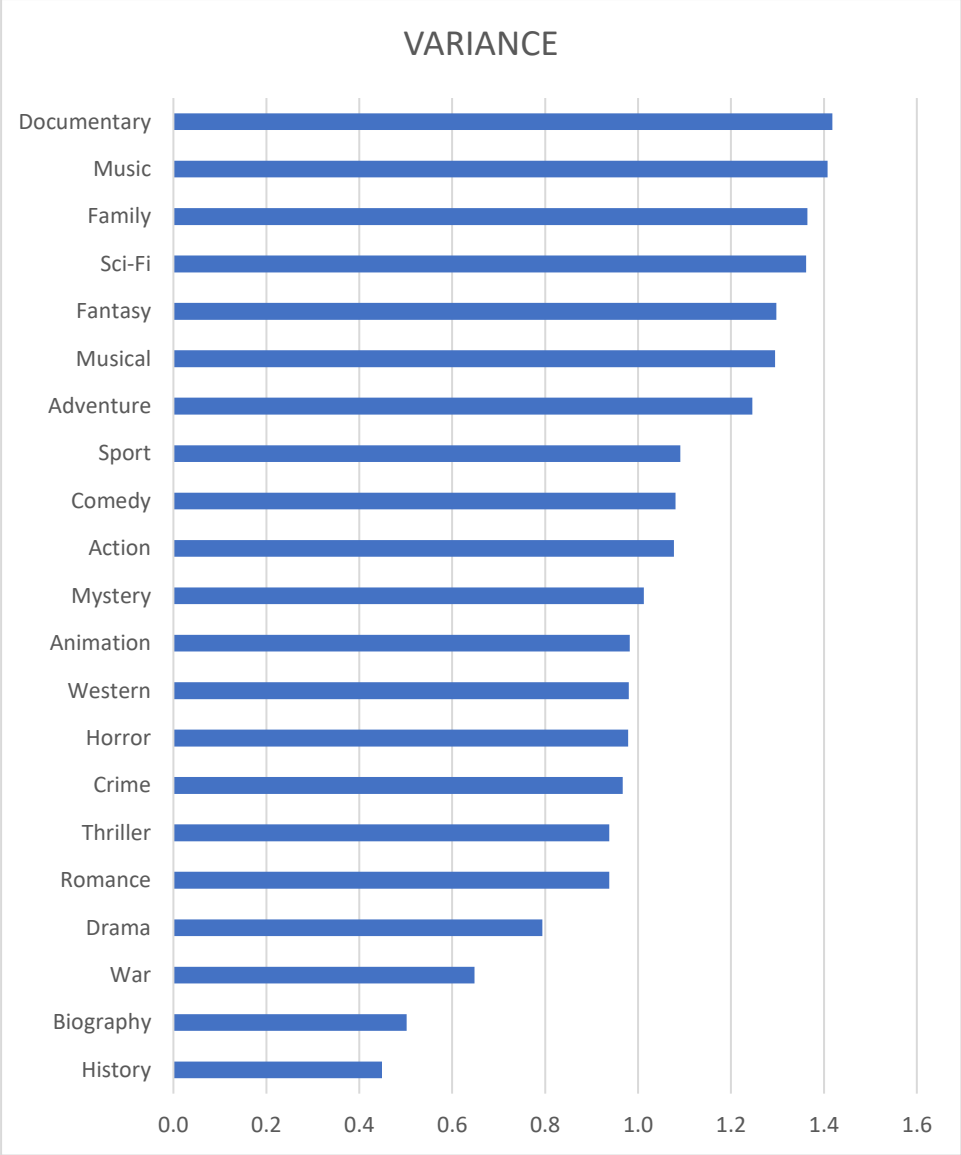
Range of IMDB =
=MAX(IF(ISNUMBER(SEARCH($B2,IMDB_Movies!$C:$C)),IMDB_Movies!$G:$G)) – MIN(IF(ISNUMBER(SEARCH($B2,IMDB_Movies!$C:$C)),IMDB_Movies!$G:$G))

Variance of population
=VAR.P(IF(ISNUMBER(SEARCH($B2,IMDB_Movies!$C:$C)),IMDB_Movies!$G:$G))

Standard deviation of population =
=STDEV.P(IF(ISNUMBER(SEARCH($B2,IMDB_Movies!$C:$C)),IMDB_Movies!$G:$G))







Observations and Insights on Average IMBD scores with respect to genre

Highest Average IMDb Scores:

- History (7.1): History is the genre with the highest average IMDb score in the dataset. This suggests that, on average, movies in the History genre tend to receive higher ratings from viewers.
- Biography (7.1): Similar to History, Biography also has a high average IMDb score, indicating a positive reception among viewers.
- Documentary (7.0): Documentary films have a relatively high average IMDb score, reflecting positive audience opinions.

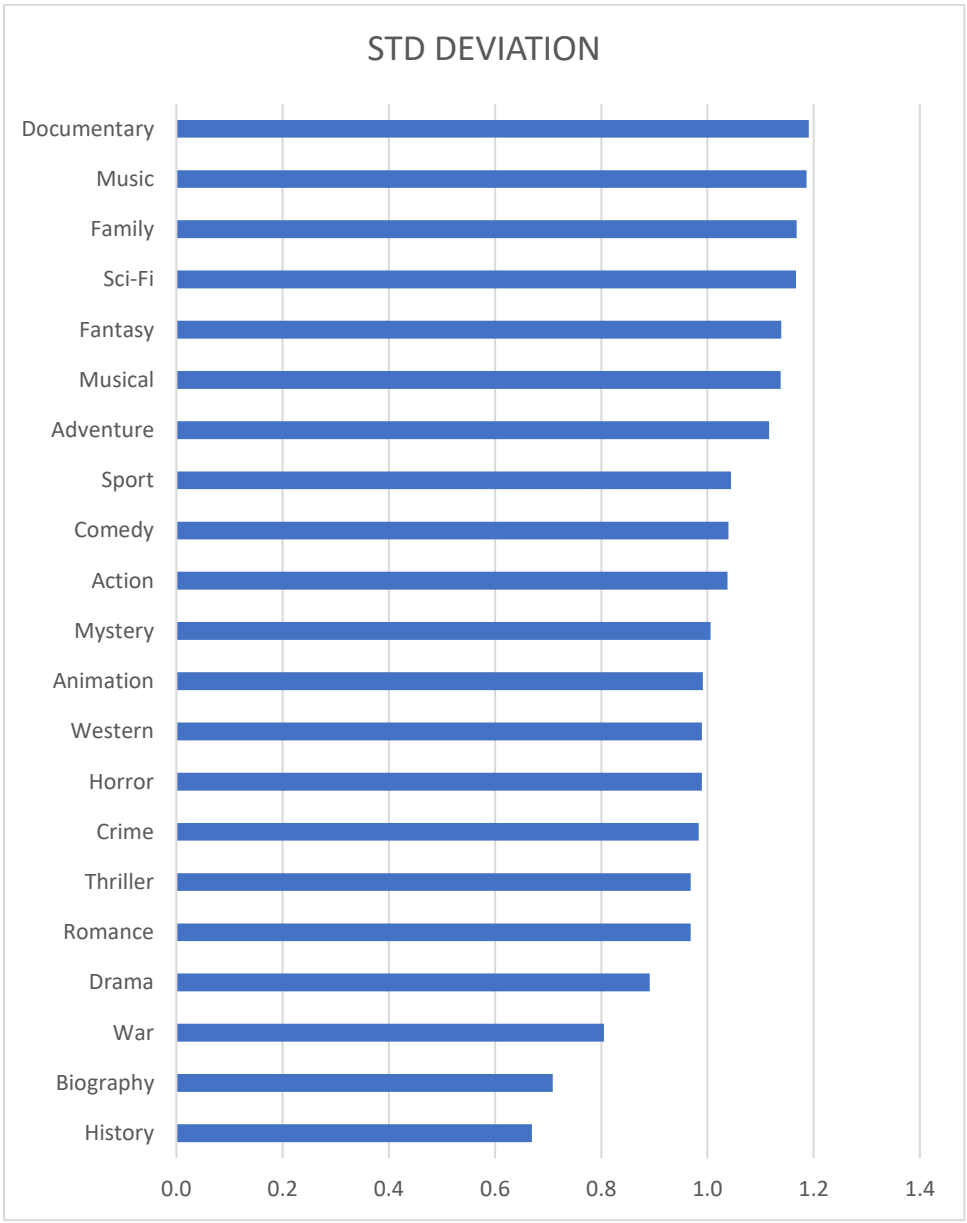
Moderate Average IMDb Scores:

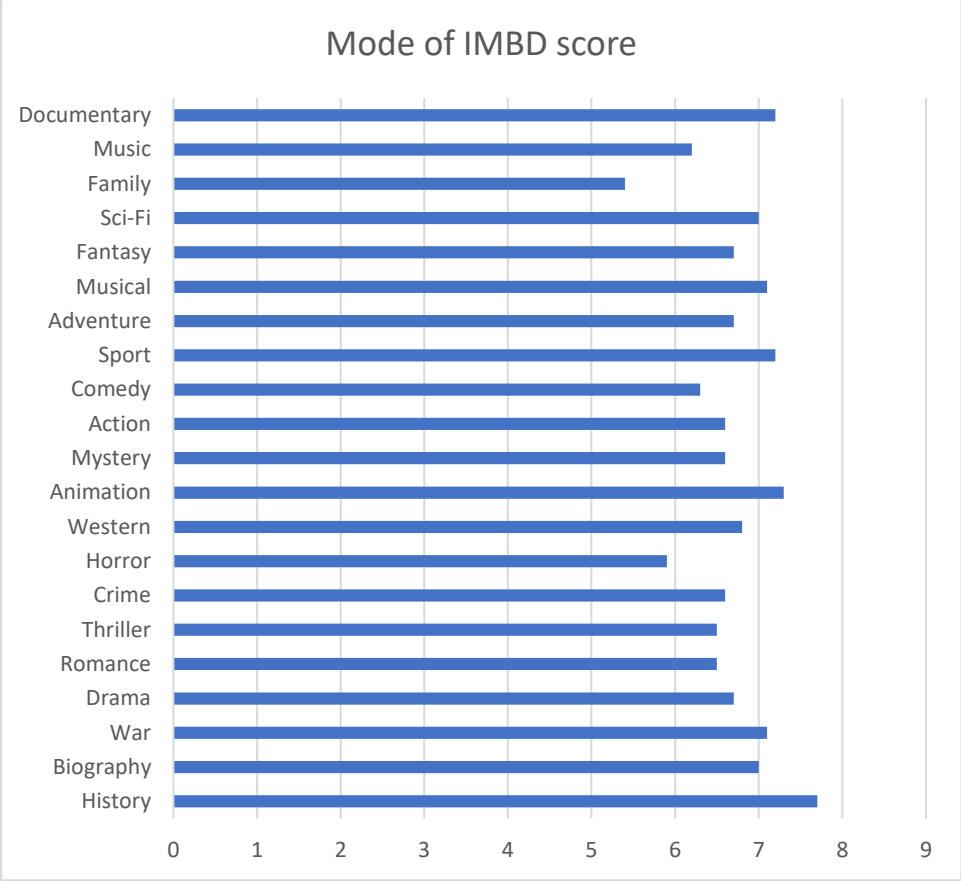
- Drama (6.8): Drama, which has the highest count of movies in the dataset, has a moderate average IMDb score. While Drama films are popular, the ratings vary, resulting in a moderate average.

- Animation (6.7): Animation also falls into the moderate range, indicating a generally positive reception but with some variability in IMDb scores.

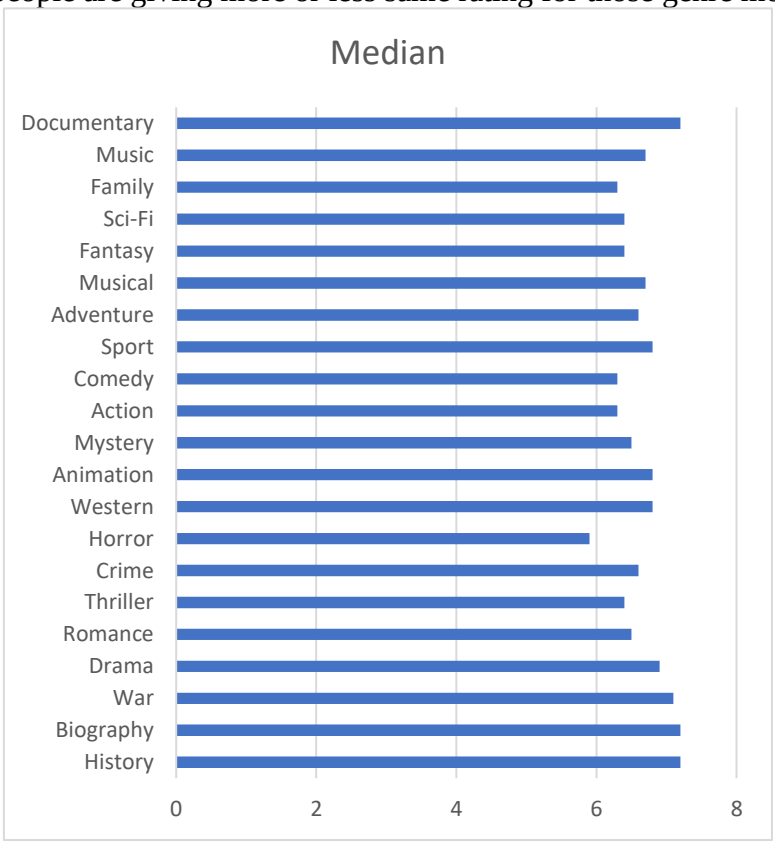
Lower Average IMDb Scores:

- Horror (5.9): Horror has the lowest average IMDb score in the dataset. This suggests that, on average, horror movies may receive lower ratings compared to other genres.





The genres like history, documentary and animation has higher modes representing that people are giving more or less same rating for those genre movies.



Consistency in Scores:

In genres where the mean, mode, and median are close to each other (e.g., Horror, Comedy, Family, and Action), there is a higher level of consistency in IMDb scores. This suggests that the majority of movies in these genres cluster around a central rating.

Variability in Scores:

Genres with larger differences between mean, mode, and median (e.g., Animation, Documentary, Mystery) may have more variability in IMDb scores. This suggests a broader range of audience opinions, with some movies having significantly higher or lower ratings than the central tendency.

Range:

- Low Range (e.g., Horror, Comedy, Family): Genres with a low range have IMDb scores that are relatively close to each other. This suggests a narrower spread of ratings, indicating that the majority of movies in these genres have similar audience opinions.
- High Range (e.g., Animation, Music, Documentary): Genres with a high range have more variability in IMDb scores. This indicates that movies within these genres can have a wider range of ratings, reflecting diverse audience opinions.

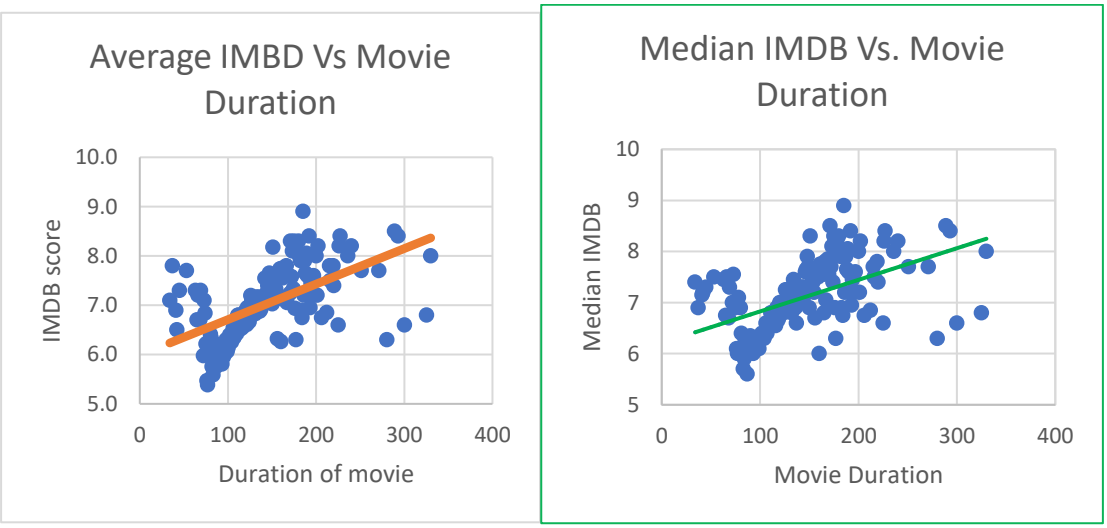
Standard Deviation and Variance of movies wrt Genres:

- Low Standard Deviation and Variance (e.g., Drama, War, History, Biography): Genres with low standard deviation and variance have IMDb scores that are close to the mean. This suggests a more consistent and predictable audience response, with movies generally receiving similar ratings.
- Moderate Standard Deviation and Variance (e.g., Romance, Action, Sci-Fi, Mystery): Genres with moderate standard deviation and variance have IMDb scores that show some variability around the mean. This indicates a moderate level of diversity in audience opinions, with some movies deviating from the average.
- High Standard Deviation and Variance (e.g., Animation, Music, Documentary): Genres with high standard deviation and variance have IMDb scores that are more spread out from the mean. This suggests a higher level of variability in audience responses, with movies eliciting a broader range of ratings.

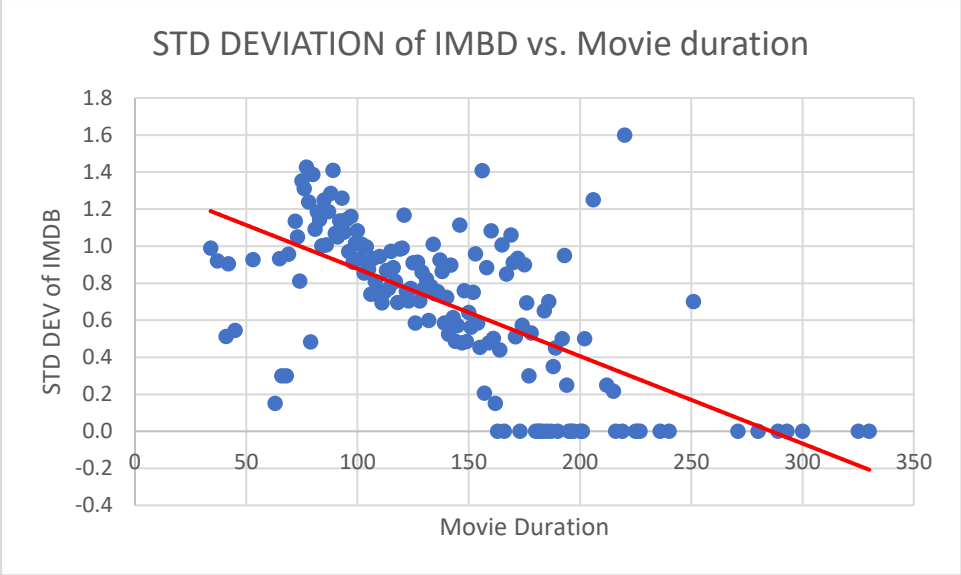
B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

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

- Hint: Calculate descriptive statistics such as mean, median, and standard deviation for movie durations. Use Excel's functions like AVERAGE, MEDIAN, and STDEV. Create a scatter plot to visualize the relationship between movie duration and IMDB score. Add a trendline to assess the direction and strength of the relationship.
- Functions used
- Column B of IMDB_movies has movie duration
- Unique list of movie duration =UNIQUE(IMDB_Movies!B2:B3781)
- Average IMDB =AVERAGEIF(IMDB_Movies!B:B,A2,IMDB_Movies!G:G)
- Median =MEDIAN(IF(ISNUMBER(SEARCH($A2,IMDB_Movies!$B:$B)),IMDB_Movies!$G:$G))
- Standard deviation =STDEV.P(IF(ISNUMBER(SEARCH($A2,IMDB_Movies!$B:$B)),IMDB_Movies!$G:$G))



The trend lines of both graphs – Avg IMDB vs Duration and Median IMDB vs Duration has similar growth following an increasing rate. From both graph, we can conclude that the IMDB score increases with the movie duration. Also, the standard deviation is following a decreasing trend from below graph representing that shorter movies have more variation in scores compared to longer movies.

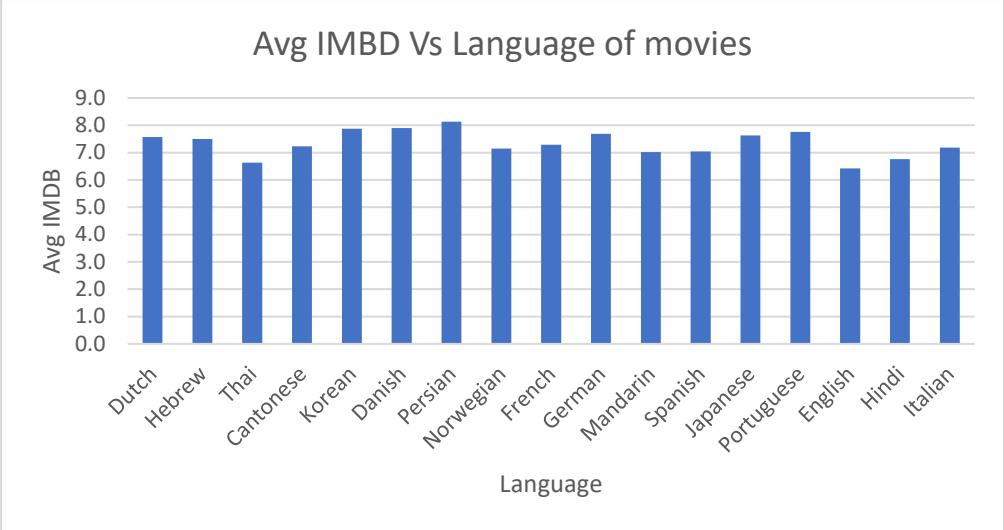


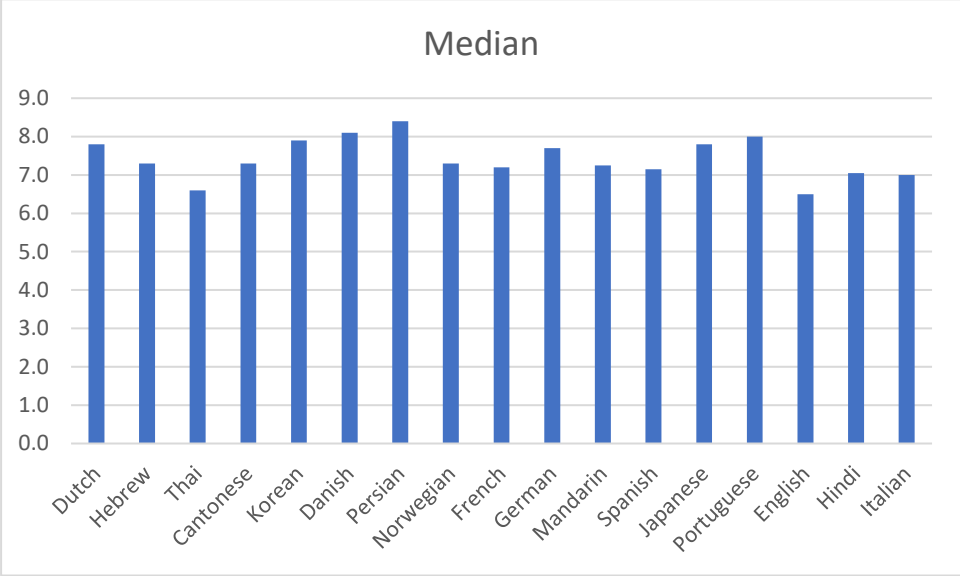
C. Language Analysis: Situation: Examine the distribution of movies based on their language.

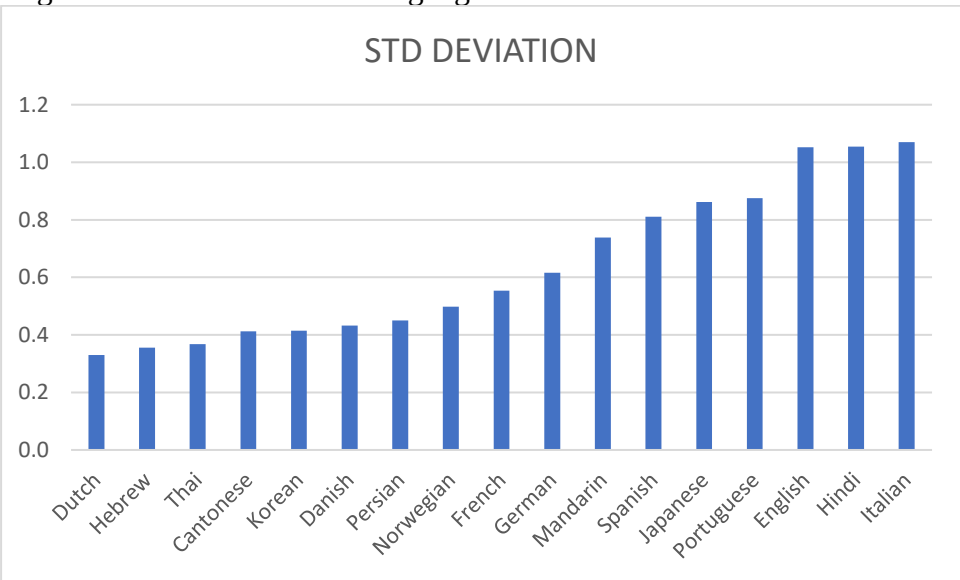
- **Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Functions Used

- Count of movies in each language is listed in column B of "Lang" sheet =COUNTIF(Table1[language],Lang!A2)
- Mean =AVERAGEIF(IMDB_Movies!E:E,A2,IMDB_Movies!$G:$G)
- Median =MEDIAN(IF(ISNUMBER(SEARCH($A2,IMDB_Movies!$E:$E)),IMDB_Movies!$G:$G))
- Standard Deviation =STDEV.P(IF(ISNUMBER(SEARCH($A2,IMDB_Movies!$E:$E)),IMDB_Movies!$G:$G))

- 

- 

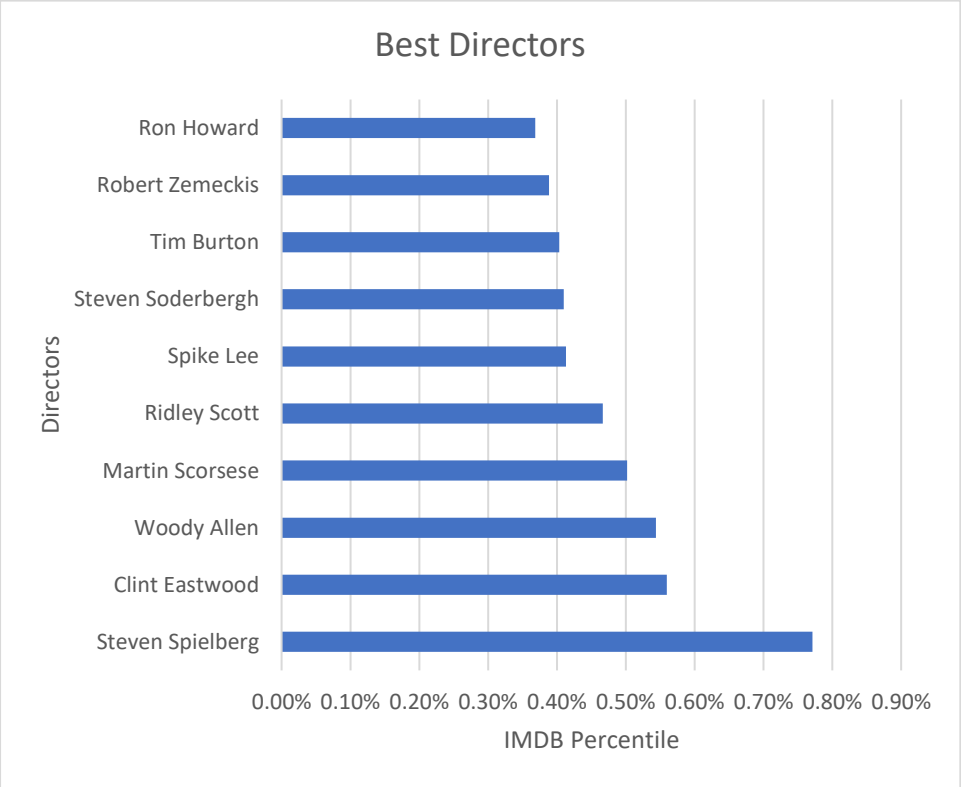
- English movies have lower mean and median IMBD scores because the dataset has huge no. of movies compared to other languages. Due to wide variety of genres and audience across the globe with mixed reactions to the movies, the scores are less for English movies unlike other languages.
- 

English, Hindi and Italian movies got more variation in the score as they have higher standard deviation.

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.

- Hint: Calculate the average IMDB score for each director. Use Excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores.
- Percentile function is used by creating a pivot table with rows as directors and IMDB score as percentile.



All the top directors have directed more than 15 movies in English and their percentile IMDB scores increase with the no. of movies directed. Steven Spielberg scored highest percentile with directing more than 25 movies being greatest number.

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.

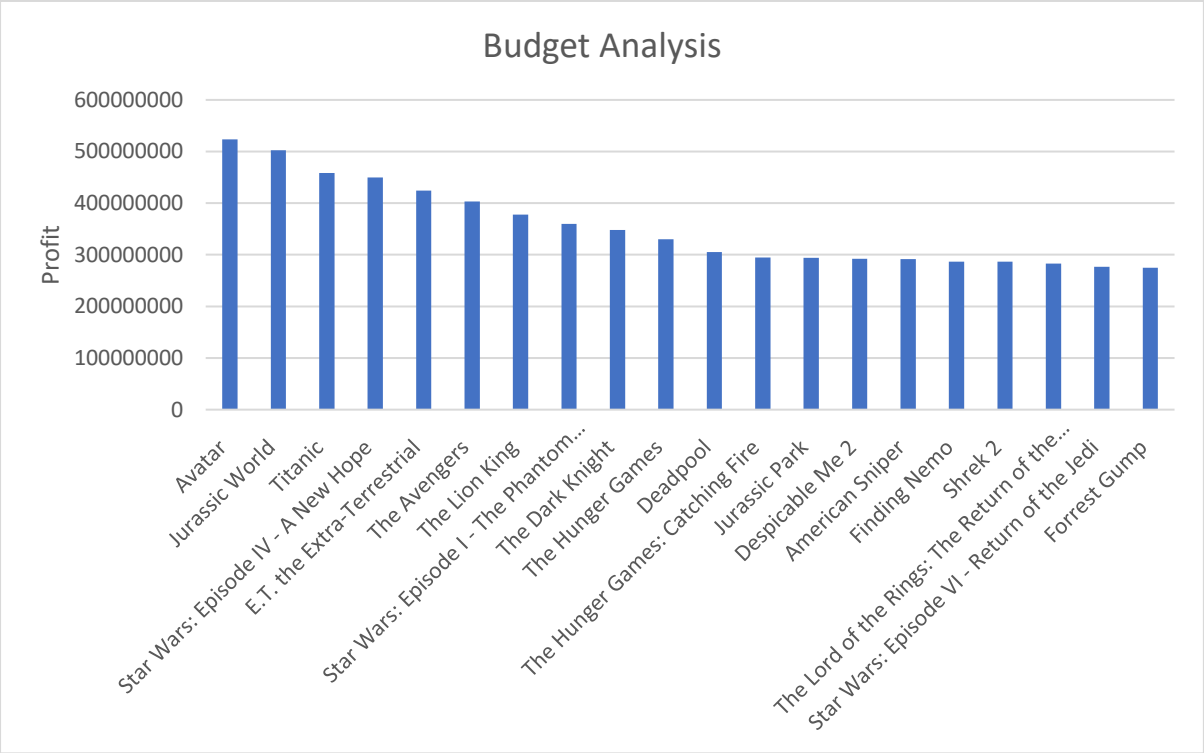
- Hint: Calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function. Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using Excel's MAX function.

Function Used:

CORRELATION COEFFICIENT between budget and profit
=CORREL(IMDB_Movies!H:H,IMDB_Movies!I:I) = 0.223

Profit is calculated by finding the difference in gross earnings and budget for all movies.

The relation between the movie budget and financial success is very weak and positive, i.e., the increase in movie budget is slightly increasing the profit of the movies. It is to be noticed that the Low budget movies can also earn high profits.



All the top movies are made in English language. As English is widely used language in many countries, the profit is enormous for movies made in English. Avatar, Jurassic World and Titanic earned higher profits.

More than 50% of movies have seen profits meanwhile around 47% of total movies have faced loss in their earnings.

Result

Based on the analysis, the production companies identify the genre of the movies and select the directors before investing in the movies. The audience can decide the movies based on the IMBD scores before watching it. Overall, it was a nice experience with the given dataset and I felt interesting in doing the analysis and deriving the insights.

Contact : cmanoj1@gmail.com