**IMDB Movie Analysis**

**DESCRIPTION**

A potential question for film makers 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 question 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.

This project is about giving insights about success of a movie based on IMDB data provided which is helpful for filmmakers and other stakeholders during production of the movie.

**TEXT STACK**

* Microsoft Excel: a spreadsheet editor software used mainly by professionals to enter data in table format.
* Python: Programming language for data processing.

**DATA OVERVIEW**

The dataset provided is related to IMDB Movies and contains records of movies from a number of years and geographical locations.

The Dataset details are:

1. **color**: Movie is Colored or Black and White.

2. **director\_name**: Name of the movie’s director.

3. **num\_critic\_for\_reviews**: Number of reviews by film critics.

4. **duration**: Duration of the movie.

5. **director\_facebook\_likes**: Facebook Likes of the director.

6. **actor\_3\_facebook\_likes**: Facebook Likes of one of the actors.

7. **actor\_2\_name**: Name of one of the actors.

8. **actor\_1\_facebook\_likes**: Facebook Likes of one of the actors.

9. **gross**: Gross collection of the movie.

10. **genres**: Genres of the movie.

11. **actor\_1\_name**: Name of one of the actors.

12. **movie\_title**: Name of the movie.

13. **num\_voted\_users**: Number of users voted for the movie.

14. **cast\_total\_facebook\_likes**: Movie cast’s total facebook likes.

15. **actor\_3\_name**: Name of one of the actors.

16. **facenumber\_in\_poster**: Number of faces in the movie’s poster.

17. **plot\_keywords**: Some keywords from plot of the movie.

18. **movie\_imdb\_link**: IMDB link of the movie

19. **num\_user\_for\_reviews**: Number of users who reviewed the movie.

20. **language**: Original language of the movie.

21. **country**: Country of origin of the movie.

22. **content\_rating**: Content rating of the movie.

23. **budget**: Budget of the movie.

24. **title\_year**: Year in which the movie was released.

25. **actor\_2\_facebook\_likes**: Facebook Likes of one of the actors.

26. **imdb\_score**: IMDB Score of the movie.

27. **aspect\_ratio**: Aspect ratio in which the movie was made.

28. **movie\_facebook\_likes**: Facebook likes of the movie.

**DATA PRE-PROCESSING**

**Handling duplicate values:** found some rows where all column values were duplicate. Keeping the first occurrence of each duplicate, dropped rest of the duplicates.

On checking, of rows with duplicate value of movie title, we observe that except for movie out of ‘Blue’ and ‘The Host’ for almost all movies, the difference between column values for rows with the same movie title are in columns related to Facebook Likes and num\_voted\_user, so except the two movies mentioned above we drop the rest of duplicate rows, leaving just one copy of the rows without much efforts on the overall analysis.

**Handling missing values**: Checked frequency of row wise null values and dropped all the rows where the number of null values were greater than 9.

**Handling outliers:** Per duration column, replaced the values above and below the upper and lower whisker mark respectively, with the median line.

Replaced values of budget less than zero with median value.

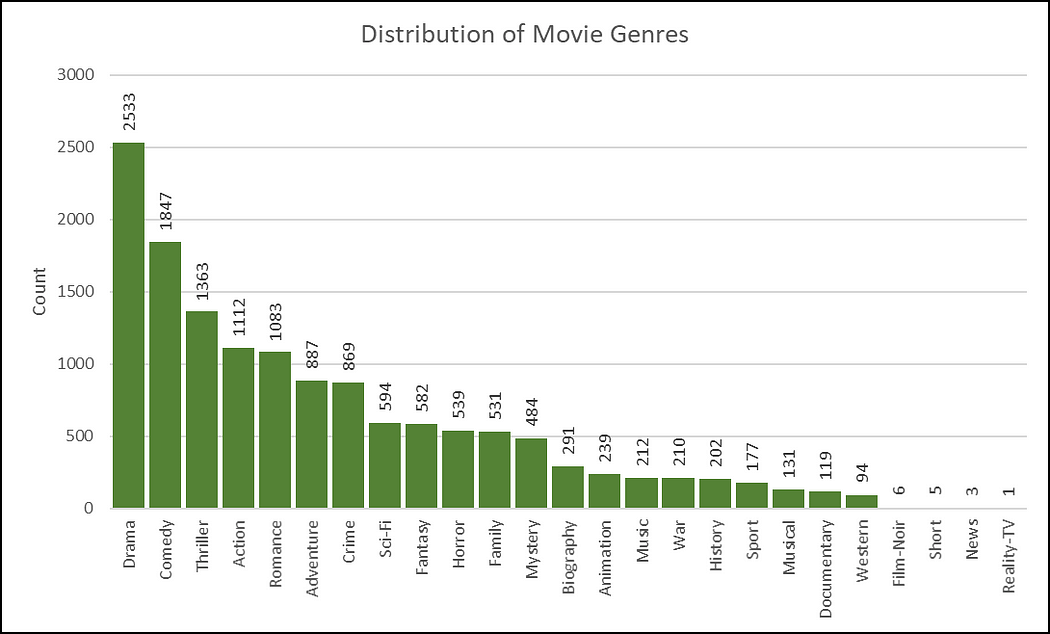
Replace values of gross less than zero with median value.

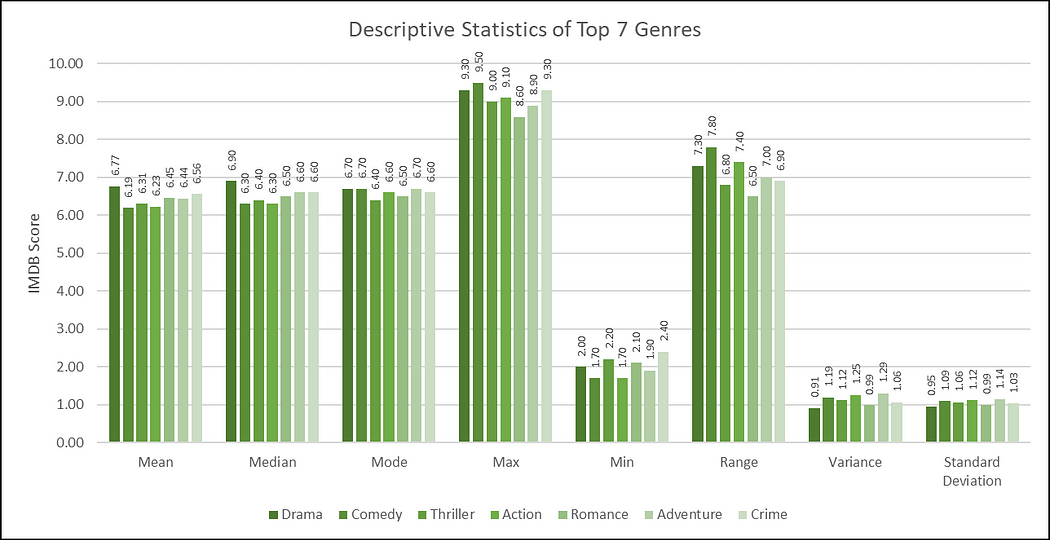
**INSIGHTS**

**Movie Genre Analysis:** A movie genre analysis. Analyse the distribution of movie genre and their impact on 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.

**Result:** The top seven most common genres are drama, comedy, thriller, action, romance, adventure and crime.

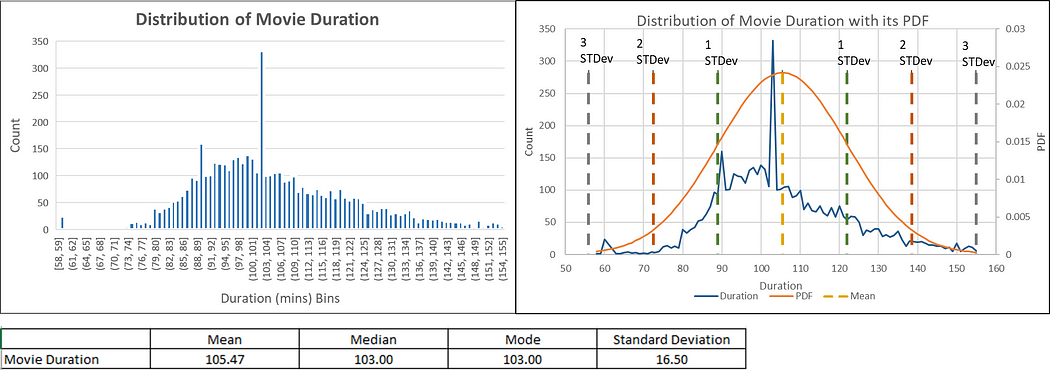


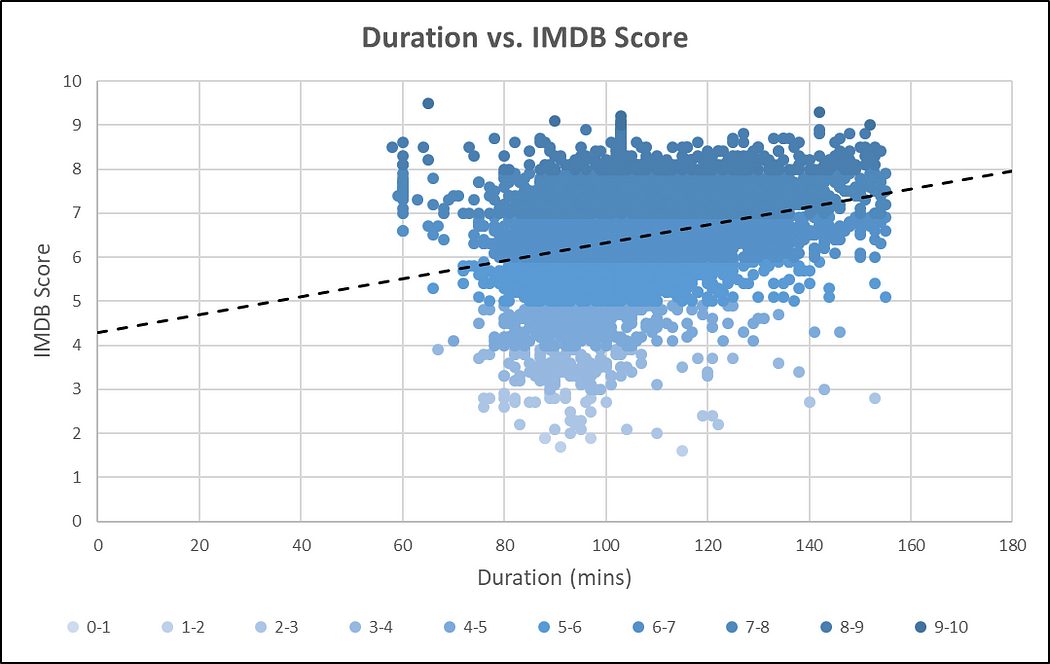


**Movie Duration Analysis:**Analyze the distribution of movie durations and its impact on the IMDB score.

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

**Result:** The distribution of movie duration shows that it closely follows a Normal Distribution. Also the scatter plot shows that duration and IMDB scores have a positive relationship.

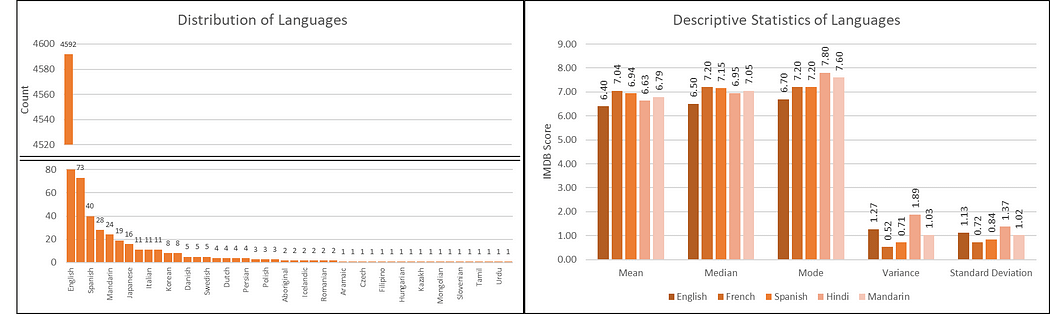




**Language Analysis:** Examine the distribution of movies based on their language.

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

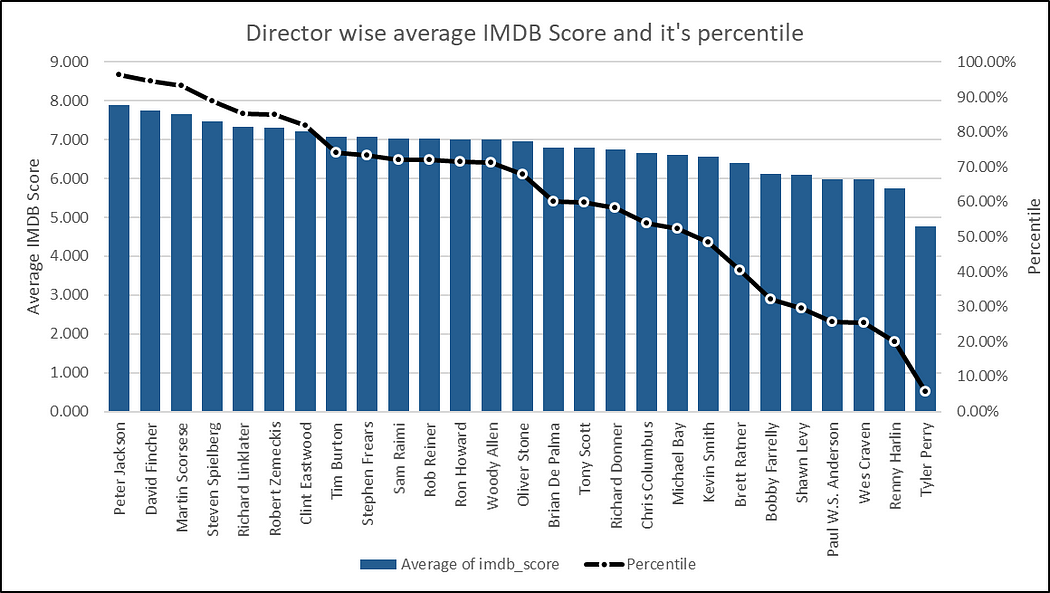
**Result:** The left plot shows that English is the most common language used in movies, followed by French, Spanish, Hindi and Mandarin. The right plot shows that French language has comparatively higher mean and median but the lower variance and standard deviation implying that most the French language movies have their IMDb score on the higher scale.



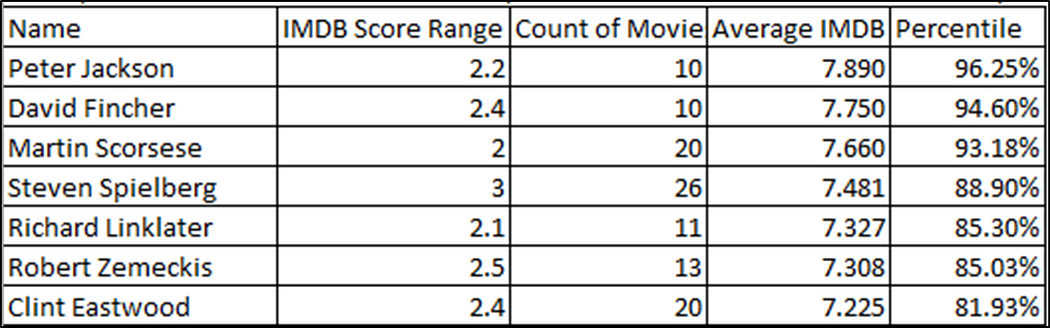
**Director Analysis:**Influence of directors on movie ratings.

**Task:** Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

**Result:** The plot considers only those directors whose movie counts are more than 9 and the range of IMDb scores is less than qual to 3 as otherwise it would be unfair for those who has maintained consistently high scores for large numbers of movies to be compared for top directors to those who has performed well in few movies.



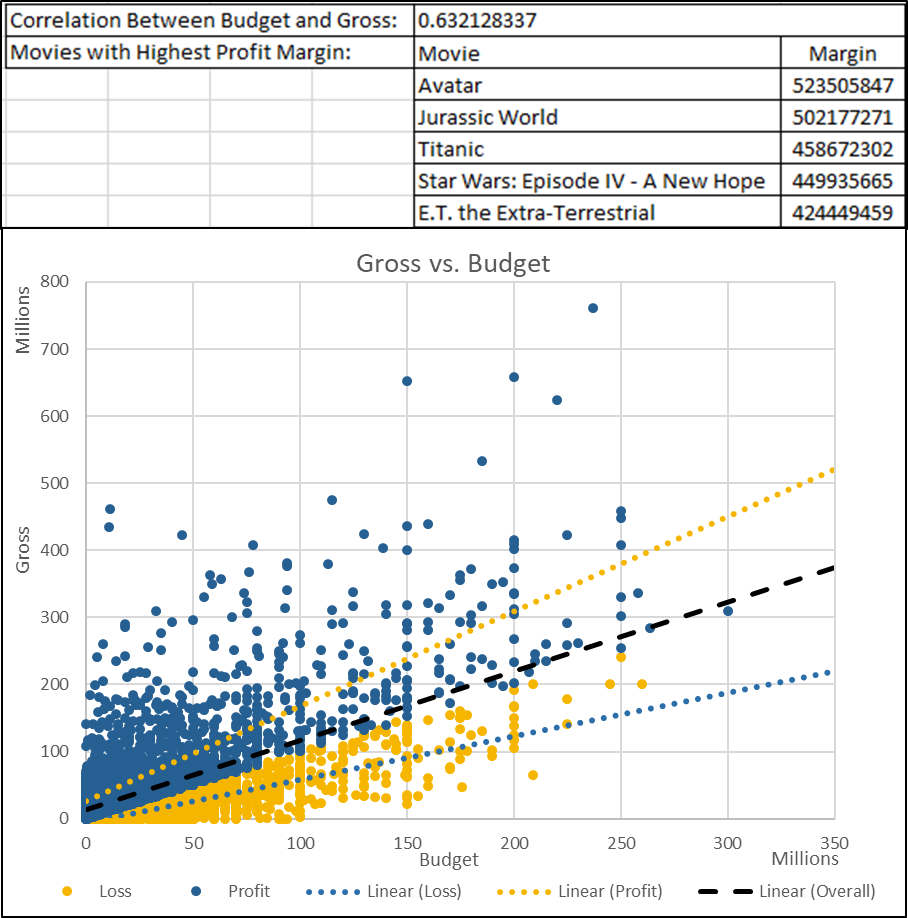
For top directors, only the top 7 directors are considered as there is a drop in percentile after the first 7 directors in the previous plot. The average **IMDB**Scores are between **7**and **8**for the top directors with the above condition. Also their percentile score is above **80%.**



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

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

**Result:** The table shows that the correlation between gross and budget is positive and more than 0.5, That is the relationship shows that as budget of movie increases, there is a very high probability that the gross collection of the movies will also increase.



**CONCLUSION**

Through this project, I was able to understand the importance of **Data Analytics**in **Movies analysis**as it provides valuable insights such as director’s relationship with IMDB Score, genre’s relationship with IMDB Score, budget’s relationship with IMDB Score etc. which helps in making **Data-Driven Decisions**.