# IMDB MOVIE ANALYSIS

FINAL PROJECT-1
BY NAZIREEN SANIA

# INTRODUCTION

## ABOUT THE PROJECT

The dataset provided is related to IMDB Movies. There are 28 columns and 5043 rows in the dataset.

## Removing Duplicates

Duplicate rows can be r3emoved by using the Data Tab and selecting the “Remove Duplicates” option. 121 rows are deleted which were detected to be duplicates.

## Data Cleaning

This step involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.

We have few columns that are irrelevant to our analysis such as actor name,actor likes, color, director_facebook_likes, facenumber_in_poster, movie_imdb_link, aspect_ratio, content_rating, num_voted_users, movie_facebook_likes hence we delete those rows. We place necesssary columns next to each other to make the analysis easy.

We can find empty cells using the Find&Select option in excel and delete their respective rows as part of data cleaning but deleting all rows together would lead to deleting a huge part of the data. To avoid data loss, we can delete rows for each section of the analysis. For instance, for analysis different genres and their IMDB score, we can filter out rows of genre and score in a separate sheet and analyse them to prevent deleting data from important columns like budget, language, year that might not be empty.

## Feature Engineering

Feature engineering is the process of creating new variables (features) or modifying existing ones to improve analysis and model performance. It helps in uncovering hidden patterns in the data. In our dataset, we can use feature engineering on columns such as splittling Genre, Year (to classify years into decades). creating a profit column where Profit = Gross - Budget and converting IMDB score into categories like "Low", "Average", "High". We can do this during indicidual analysis like for removing null values.

## A. MOVIE GENRE ANALYSIS:
## ANALYZE THE DISTRIBUTION OF MOVIE GENRES AND THEIR IMPACT ON THE IMDB SCORE

On a new Excel sheet, we take the columns - Movie Name, Genre and IMDB Score and remove null values and duplicate rows. We perform feature engineering to extract the different genres.

**1. Split Genres into Separate Columns (Text to Columns Method)**

- Select the "Genres" column.
- Go to Data → Click Text to Columns.
- Choose Delimited → Click Next.
- Select Other and enter | (pipe symbol) → Click Finish.

**2. Create Binary Columns for Each Genre (One-Hot Encoding)**

- Create new column headers for each genre (e.g., Action, Comedy, Drama).
- =IF(COUNTIF($D2:$L2,"Action")>0,1,0) , repeat for each genre

**3. Data Analysis**

- Find mean, median IMDB Score for each genre

| GENRE | AVERAGE IMDB SCORE | MEDIAN IMDB SCORE |
|---|---|---|
| Action | 6.2 | 6.3 |
| Adventure | 6.4 | 6.6 |
| Fantasy | 6.3 | 6.4 |
| Sci-Fi | 6.3 | 6.4 |
| Thriller | 6.3 | 6.4 |
| Romance | 6.4 | 6.5 |
| Animation | 6.6 | 6.7 |
| Family | 6.2 | 6.4 |
| Musical | 6.5 | 6.7 |
| Drama | 6.8 | 6.9 |
| Crime | 6.6 | 6.6 |
| Western | 6.7 | 6.8 |
| Mystery | 6.5 | 6.6 |
| Horror | 5.8 | 6.6 |
| Biography | 7.1 | 5.9 |
| War | 7.1 | 7.2 |
| History | 7.1 | 7.2 |
| Sport | 6.6 | 6.8 |
| Documentary | 7.2 | 7.4 |

| | Action | Adventure | Fantasy | Sci-Fi | Thriller | Romance | Animation | Family | Musical | Drama | Crime | Western | Mystery | Horror | Biography |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SCORE | 6.6 | 6.7 | 6.7 | 6.7 | 6.4 | 6.5 | 6.7 | 6.7 | 7 | 6.7 | 6.6 | 6.5 | 6.6 | 6.2 | 7 |
| IMDB Sc | 1.3 | 1.3 | 1.4 | 1.5 | 1.1 | 1.0 | 1.3 | 1.5 | 1.5 | 0.9 | 1.1 | 1.1 | 1.2 | [illegible] | 0.5 |
| iation | 1.1 | 1.1 | 1.2 | 1.2 | 1.1 | 1.0 | 1.1 | 1.2 | 1.2 | 1.0 | 1.0 | 1.1 | 1.1 | 1.1 | 0.7 |

## B. MOVIE DURATION ANALYSIS

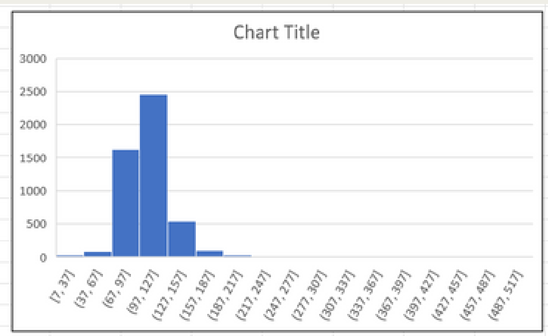
## ANALYZE THE DISTRIBUTION OF MOVIE DURATIONS AND ITS IMPACT ON THE IMDB SCORE.

On a new Excel sheet, we take the columns - Movie Name, Duration and IMDB Score and remove null values and duplicate rows.

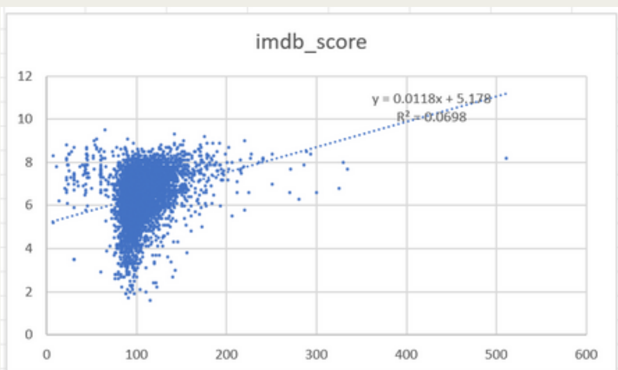
- We create a histogram for the movie durations and increase the bin size to see the distribution of durations. Most movies have a duration between 80-150 minutes

- We calculate the mean, median, mode, and range of the duration for all movies. We also calculate the correlation between duration and IMDB Score. It is 0.2642.
- Since the value is greater than 0, it suggests that longer movies tend to have slightly higher IMDB scores.
- However, because 0.2642 is closer to 0 than to 1, the relationship is weak, meaning movie duration alone is not a strong predictor of IMDB ratings.
- Hence, longer movies might get higher ratings, but the effect is small.

- We then create a scatter plot between movie duration and IMDB scores andinsert a trendline. The $R^2$ value = 0.0698, meaning only 6.98% of the variation in IMDB scores can be explained by movie duration. This is a very weak relationship, meaning that duration alone is not a strong predictor of IMDB score.




| | |
|---|---|
| AVERAGE | 107.099 |
| MEDIAN | 103 |
| MODE | 90 |
| MAXIMUM | 511 |
| MINIMUM | 7 |
| RANGE | 504 |
| STANDARD DEVIATION | 25.2798 |
| CORRELATION | 0.26424 |



**C. LANGUAGE ANALYSIS SITUATION**
**EXAMINE THE DISTRIBUTION OF MOVIES BASED ON THEIR LANGUAGE.**

---

On a new Excel sheet, we take the columns - Movie Name, Language and IMDB Score and remove null values and duplicate rows.

- By using a pivot column, we find the top 5 languages by count of moviesper each language.

- After we find the top 5 languages. we find the average (mean), median and stnadard deviation of IMDB score for each language using the condition:

=AVERAGEIF(B:B, "English", C:C),
=MEDIAN(IF(B:B="French", C:C)),
=STDEV.P(IF(B:B="English", C:C))

We repeat this for the other 4 languages to find their impact on the IMDB Scores.

| Row Labels | Count of movie_title |
|---|---|
| En | |
| Fr | |
| H | |
| M | |
| Sp | |
| **Grand Total** | **4749** |

| MEAN IMDB SCORE FOR TOP 5 LANGUAGES | |
|---|---|
| English | 6.39365 |
| French | 7.03836 |
| Spanish | 6.9375 |
| Hindi | 6.63214 |
| Mandarin | 6.7875 |

| | COUNT | MEAN | MEDIAN | STANDARD DEVIATION |
|---|---|---|---|---|
| English | 4584 | 6.39365 | 7.2 | 1.13 |
| French | 73 | 7.03836 | 7.15 | 0.72 |
| Spanish | 40 | 6.9375 | 6.95 | 0.84 |
| Hindi | 28 | 6.63214 | 7.05 | 1.37 |
| Mandarin | 24 | 6.7875 | 6.5 | 1.02 |

- French has  a higher average score,  it may indicate that movies in that language tend to be better received.

- Higher standard deviation means more variation in movie ratings. Hence, english movie ratings have more variation.

- Why do movies in English have more variation in ratings?

 - English being a widespread international language is well recieved all over the world with Hollywood attracting viewers from all over the world. However, not all movies in English can live upto the same expectations like big franchise movies like Marvel, hence, the variation in the ratings.

## D. DIRECTOR ANALYSIS
## INFLUENCE OF DIRECTORS ON MOVIE RATINGS.

---

On a new Excel sheet, we take the columns - Movie Name, Director and IMDB Score, remove null values, and duplicate rows.

- By using a pivot column, we find the average IMDB score for each director by placing the Director column in rows and Average of IMDB Score as Values. We obtain the top 10 directors by average of IMDB Score by using Sort and Filter.

| Row Labels | Average of imdb_score |
|---|---|
| John Blanchard | 9.5 |
| Sadyk Sher-Niyaz | 8.7 |
| Mitchell Altieri | 8.7 |
| Cary Bell | 8.7 |
| Mike Mayhall | 8.6 |
| Charles Chaplin | 8.6 |
| Damien Chazelle | 8.5 |
| Ron Fricke | 8.5 |
| Raja Menon | 8.5 |
| Majid Majidi | 8.5 |
| **Grand Total** | **8.68** |

- Since we have to use Percentile Function to identify directors with highest scores, we use the function PERCENTILE.INC(C:C, 0.9) where C is the IMDB Score column. The result is 7.7 which means that 90% of movies have an IMDB score below 8.2, and only 10% of movies have a score above this.This helps us identify top directors — those whose average IMDB score exceeds the 90th percentile are in the top 10% of directors.
- To futher filter out the top 10%, we give a conditional statement to classify the director as Top 10% or below 90th Percentile, i.e, IF(AVERAGEIF(B:B, B2, C:C) >= 7.7, "Top 10%", "Below 90th percentile"). Incase we want the top 10 directors out of this Top 10%, we already have the result above that we obtained using a pivot table.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | movie_title | director_name | imdb_score | Percentile |
| 2 | AvatarÂ | James Cameron | 7.9 | Top 10% |
| 5 | The Dark Knight RisesÂ | Christopher Nolan | 8.5 | Top 10% |
| 6 | John CarterÂ | Andrew Stanton | 6.6 | Top 10% |
| 8 | TangledÂ | Nathan Greno | 7.8 | Top 10% |
| 9 | Avengers: Age of UltronÂ | Joss Whedon | 7.5 | Top 10% |
| 18 | The AvengersÂ | Joss Whedon | 8.1 | Top 10% |
| 21 | The Hobbit: The Battle of the Five ArmiesÂ | Peter Jackson | 7.5 | Top 10% |
| 24 | The Hobbit: The Desolation of SmaugÂ | Peter Jackson | 7.9 | Top 10% |
| 26 | King KongÂ | Peter Jackson | 7.2 | Top 10% |
| 27 | TitanicÂ | James Cameron | 7.7 | Top 10% |
| 44 | Toy Story 3Â | Lee Unkrich | 8.3 | Top 10% |
| 58 | WALLÂ·EÂ | Andrew Stanton | 8.4 | Top 10% |
| 66 | The Dark KnightÂ | Christopher Nolan | 9 | Top 10% |
| 67 | UpÂ | Pete Docter | 8.3 | Top 10% |
| 78 | Inside OutÂ | Pete Docter | 8.3 | Top 10% |
| 89 | Big Hero 6Â | Don Hall | 7.9 | Top 10% |
| 90 | Wreck-It RalphÂ | Rich Moore | 7.8 | Top 10% |
| 93 | How to Train Your DragonÂ | Dean DeBlois | 8.2 | Top 10% |
| 96 | InterstellarÂ | Christopher Nolan | 8.6 | Top 10% |
| 97 | InceptionÂ | Christopher Nolan | 8.8 | Top 10% |

## E. BUDGET ANALYSIS

### EXPLORE THE RELATIONSHIP BETWEEN MOVIE BUDGETS AND THEIR FINANCIAL SUCCESS.

---

On a new Excel sheet, we take the columns - Movie Name, Budget, Gross and IMDB Score- remove null values, and duplicate rows.

- We calculate the Profit for each movie by calculating the difference between budget and gross, i.e., Profit = Gross - Budget
- After calculating profit for each movie, we find the movie eith the highest profit by caluclating the maximum profit and indexing it to its respective movie.

Highest Profit = Max(E:E)
Movie with highest profit  = INDEX (A:A, MATCH(MAX(E:E), E:E,0))
where A - Movie Name and E - Profit

The movie with the highest profit is AvatarÂ with a profit of 523505847

| | |
|---|---|
| **correlation between budget and gross** | 0.096619736 |
| **highest profit** | 523505847 |
| **movie with highest profit** | AvatarÂ |

- To analyze the correlation between movie budgets and gross earnings, we use the CORREL function in EXCEL.

- The correlation coefficient 0.0966 is very close to 0, indicating a weak positive relationship between movie budget and gross earnings.

- This means that :

 - Budget does not strongly determine financial success.
 - Some low-budget movies may earn a lot, while some high-budget movies might underperform.
 - Other factors (e.g., genre, director, marketing, reviews) likely have a bigger impact on earnings.