IMDB Movie Analysis

1.1 Project Description

The primary goal of the project is to analyze the factors that help in the success of a movie. Various features related to the making, release, and production of movies are considered to understand the underlying patterns between them. Multiple valuable and data-driven insights are generated and visualized to further understand the dynamics of this field.

1.2 Approach

The dataset is available in .csv and .xlsx format. It is first cleaned, rows with missing values are deleted, and duplicates are removed. All unnecessary features that would not be required for analysis are deleted for simplification of the database. Out of **28 features**, **19 features** are **deleted**. The analysis is done based on a cleaned dataset with **9 features** and **3786 rows**.

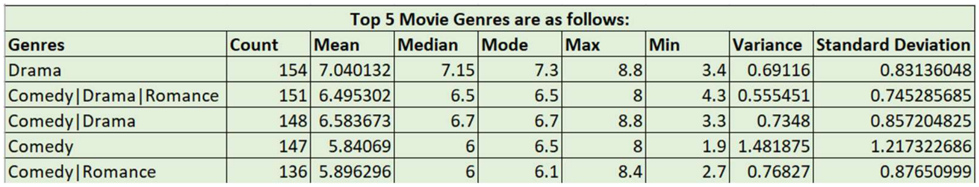
1.3 Tech Stack Used

MS Excel is used for data storage, data processing, manipulation, and visualization. MS Word is used to display actionable insights in an easy-to-understand format, that is, the following report.

1.4 INSIGHTS

1.4.1 Movie Genre Analysis

- 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.
- **Approach:**
  - A column called **Genres** is created using the following formula:
    - =UNIQUE('Cleaned Db'!D1:D3786)
  - A column called **Count** is created using the following:
    - =COUNTIF('Cleaned Db'!I2:I3786,'Cleaned Db'!I2)
  - Both columns are sorted based on **Count** from Largest to Smallest value.
  - Create columns called **Mean**, **Medians**, **Mode**, **Max**, **Min**, **Variance**, **Standard Deviation** using the following formulae:
    - =AVERAGE(IF('Cleaned Db'!$D$2:$D$3786='Movie Genre Analysis'!$K2,'Cleaned Db'!$I$2:$I$3786))
    - =MEDIAN(IF('Cleaned Db'!$D$2:$D$3786='Movie Genre Analysis'!$K2,'Cleaned Db'!$I$2:$I$3786))
    - =MODE(IF('Cleaned Db'!$D$2:$D$3786='Movie Genre Analysis'!$K2,'Cleaned Db'!$I$2:$I$3786))
    - =MAX(IF('Cleaned Db'!$D$2:$D$3786='Movie Genre Analysis'!$K2,'Cleaned Db'!$I$2:$I$3786))
    - =MIN(IF('Cleaned Db'!$D$2:$D$3786='Movie Genre Analysis'!$K2,'Cleaned Db'!$I$2:$I$3786))
    - =VAR.S(IF('Cleaned Db'!$D$2:$D$3786='Movie Genre Analysis'!$K2,'Cleaned Db'!$I$2:$I$3786))
    - =STDEV.S(IF('Cleaned Db'!$D$2:$D$3786='Movie Genre Analysis'!$K2,'Cleaned Db'!$I$2:$I$3786))
- **Output:**



Top 5 Movie Genres are as follows:

Genres	Count	Mean	Median	Mode	Max	Min	Variance	Standard Deviation
Drama	154	7.040132	7.15	7.3	8.8	3.4	0.69116	0.83136048
Comedy|Drama|Romance	151	6.495302	6.5	6.5	8	4.3	0.555451	0.745285685
Comedy|Drama	148	6.583673	6.7	6.7	8.8	3.3	0.7348	0.857204825
Comedy	147	5.84069	6	6.5	8	1.9	1.481875	1.217322686
Comedy|Romance	136	5.896296	6	6.1	8.4	2.7	0.76827	0.87650999

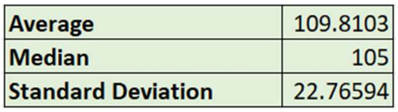
- **Inference:**
  - This analysis gives an understanding of the top movie genres and it's statistical analysis based on their **imdb_score**.

1.4.2 Movie Duration Analysis

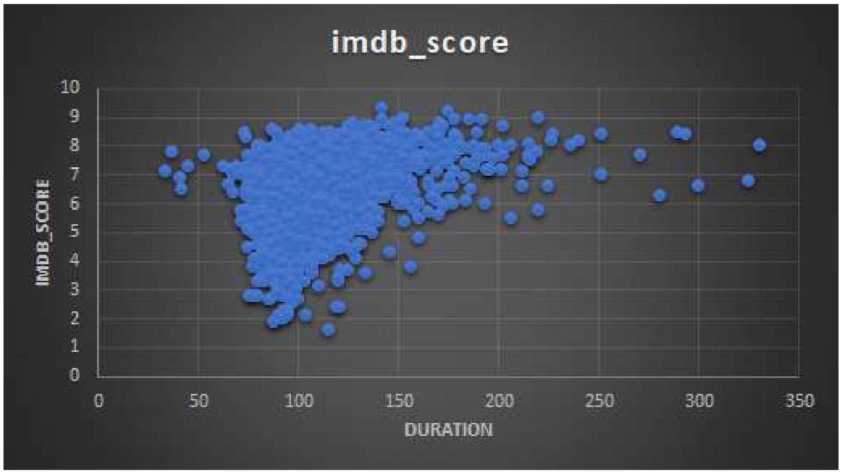
⇨ Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

- **Approach:**
  - The columns of **duration** and **imdb_score** are considered.
  - **Average**, **Median**, **Standard Deviation** are calculated for **duration** based on **imdb_score** using the following formulae:

        - =AVERAGE(A2:A3786)
        - =MEDIAN(A2:A3786)
        - =STDEV.S(A2:A3786)
- **Output:**



Average	109.8103
Median	105
Standard Deviation	22.76594

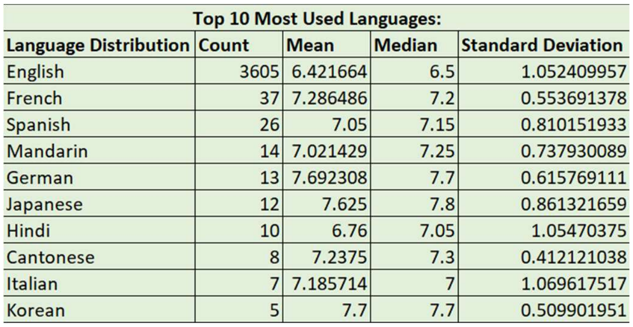


- **Inference:**
    - A scatter plot is used to visualize the impact of **duration** on **imdb_score**.
    - It is observed that the **duration** between **50** to **150** seems to be the sweet spot. A dense cluster in that area indicates that this range of **duration** has a wide variety of movies covering almost the entire range of available **imdb_scores.**

1.4.3 Language Analysis

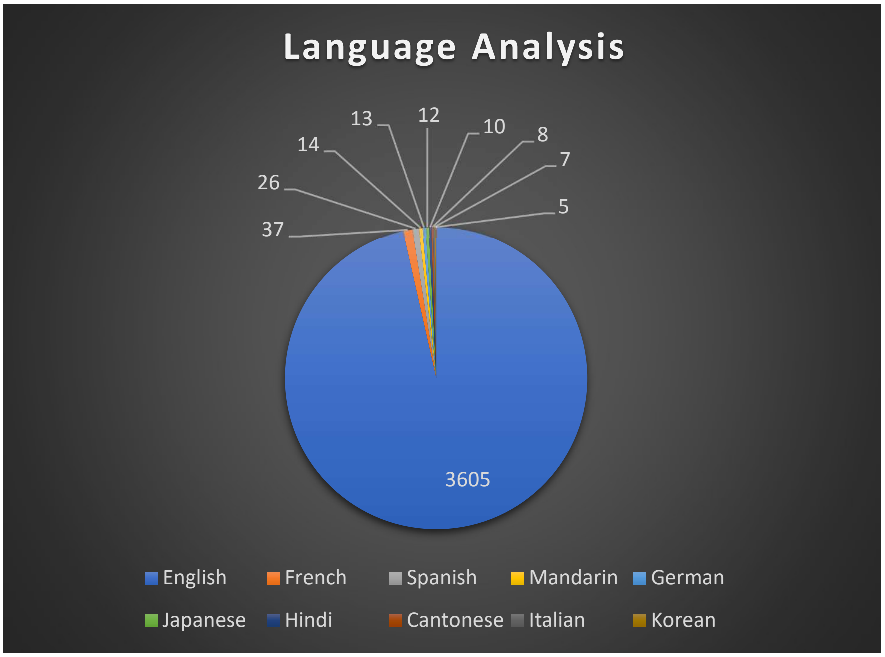
- Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.
- **Approach:**
    - A column called **Language Distribution** is created using the following formula:
        - =UNIQUE('Cleaned Db'!F2:F3786)
    - A column called **Count** is created that has the count of each Language in the dataset using the following formula:
        - =COUNTIF('Cleaned Db'!F2:F3786,'Language Analysis'!I2)
    - Both columns are sorted based on **Count** from Largest to Smallest value.
    - The columns **Mean**, **Median** and **Mode** are created using the following formulae:
        - =AVERAGE(IF('Cleaned Db'!$F$2:$F$3786='Language Analysis'!$I2,'Cleaned Db'!$I$2:$I$3786))
        - =MEDIAN(IF('Cleaned Db'!$F$2:$F$3786='Language Analysis'!$I2,'Cleaned Db'!$I$2:$I$3786))
        - =STDEV.P(IF('Cleaned Db'!$F$2:$F$3786='Language Analysis'!$I2,'Cleaned Db'!$I$2:$I$3786))

- **Output:**



Top 10 Most Used Languages:

Language Distribution	Count	Mean	Median	Standard Deviation
English	3605	6.421664	6.5	1.052409957
French	37	7.286486	7.2	0.553691378
Spanish	26	7.05	7.15	0.810151933
Mandarin	14	7.021429	7.25	0.737930089
German	13	7.692308	7.7	0.615769111
Japanese	12	7.625	7.8	0.861321659
Hindi	10	6.76	7.05	1.05470375
Cantonese	8	7.2375	7.3	0.412121038
Italian	7	7.185714	7	1.069617517
Korean	5	7.7	7.7	0.509901951

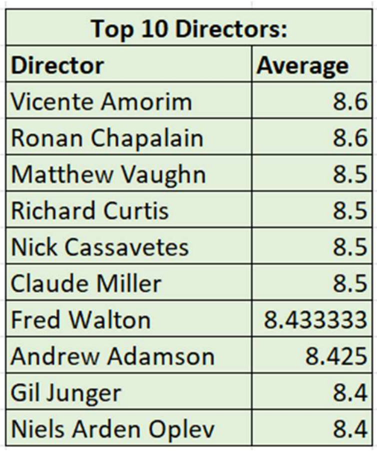


- **Inference:**
  - A pie chart is created to visualize the analysis generated.
  - English is the most common language and Korean is the least common language when top 10 languages are considered.

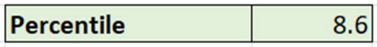
1.4.4 Director Analysis

- Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.
- **Approach:**
  - A column called **Director** is created using the following formula:
    - =UNIQUE('Cleaned Db'!A2:A3786)
  - Average imdb_score is calculated and put into a column called **Average** using the following formula:
    - =AVERAGE(IF('Cleaned Db'!$A$2:$A$3786=A2,'Cleaned Db'!$I$2:$I$3786))

        - Calculate **Percentile** using the following formula:
            - =PERCENTILE.INC(F4:F13,90%)
    - **Output:**



Top 10 Directors:	
Director	**Average**
Vicente Amorim	8.6
Ronan Chapalain	8.6
Matthew Vaughn	8.5
Richard Curtis	8.5
Nick Cassavetes	8.5
Claude Miller	8.5
Fred Walton	8.433333
Andrew Adamson	8.425
Gil Junger	8.4
Niels Arden Oplev	8.4

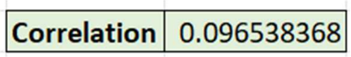


Percentile	8.6

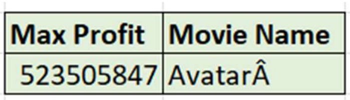
- **Inference:**
    - The **Percentile** gives the average score which is in the top 10% from the entire range of average **imdb_score**. Here, **8.6** is the score with percentile over 90% when compared to the overall distribution in the dataset.

1.4.5 Budget Analysis

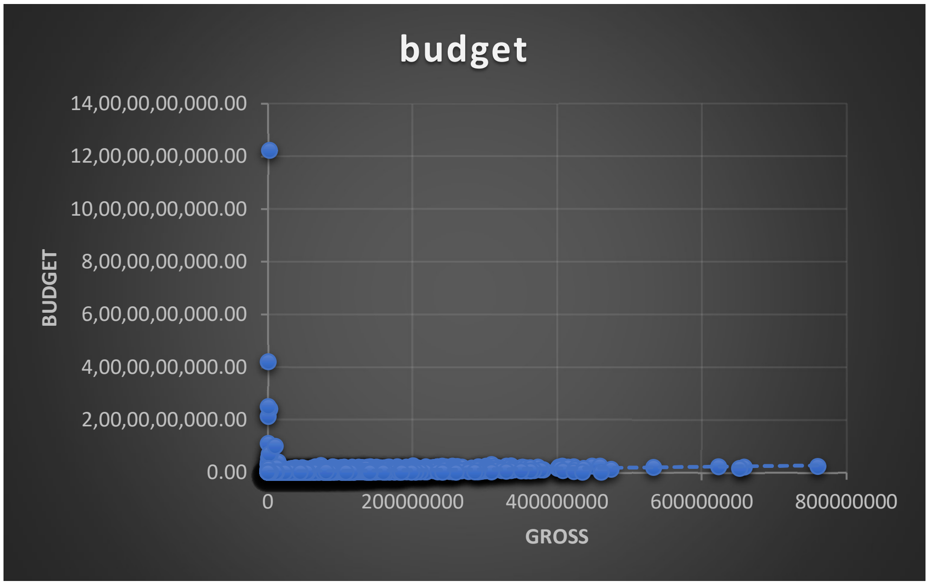
- Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.
- **Approach:**
    - The columns **gross**, **movie_title**, **budget** are considered.
    - **Profit Margin** is calculated using the following formulae:
        - =$A2-$C2
    - **Correlation** is calculated using the following formula:
        - =CORREL(A2:A3786,C2:C3786)
    - Maximum profit and movie that has that profit are calculated using the following formulae:
        - =MAX(D2:D3786)
        - =INDEX(B2:B3786, MATCH(1,IF(D2:D3786=F7, 1),0))
    - A **Scatter Plot** is used to visualize **gross** and **budget** as well as **gross** and **profit margin.**
- **Output:**

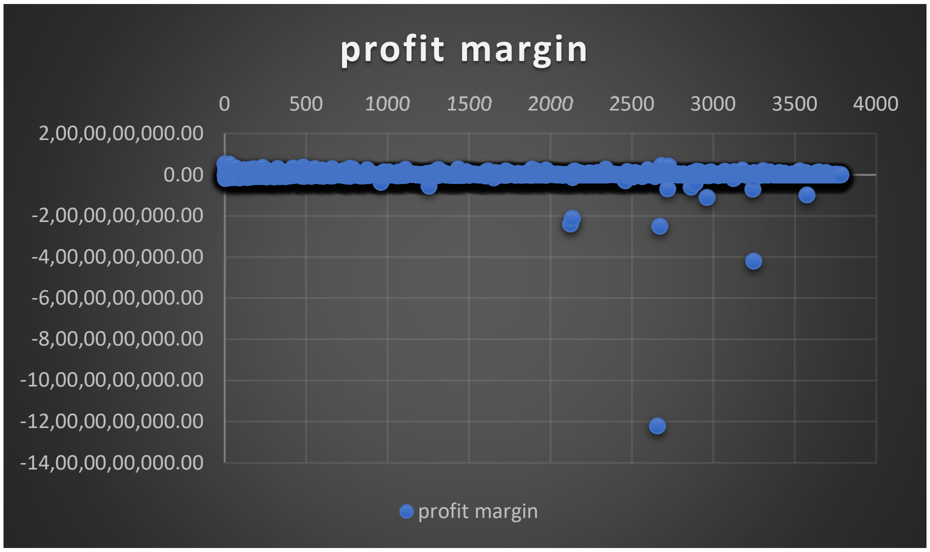


Correlation	0.096538368



Max Profit	**Movie Name**
523505847	AvatarÂ





- **Inference:**
  - Since the correlation constant when calculated for **gross** and **budget** is closer to 1, therefore the relationship is considered to be strong and positively linear. The **budget** graph compares **budget** with **gross.** It can be concluded that **budget** linearly and positively affects the **gross income** generated. In the next graph, **profit margin** is divided in bins for visualization.

1.5 Results

The analysis of the current dataset helps in understanding the movie industry and how various factors affect how a movie does in the market. The results are as follows:

- Drama, comedy, romance and combinations of these genres are very popular among the audience.
- Movies with the duration of 110 minutes on an average are preferred.
- Maximum movies are in English, French and Spanish.
- The score is considered to be more than that of 90 percentage of movies or in other words imdb score is above 90 percentile when it is 8.6.
- Budget linearly affects the gross income in the positive direction.

1.6 Hyperlink of the Excel Sheet

https://docs.google.com/spreadsheets/d/19HRj4SkBIDOs9IjDdU5jWASNumdj0Yl0/edit?usp=sharing&ouid=105545149670713438068&rtpof=true&sd=true