IMDB Movie Data Analysis Report

Project Description

This project explores the factors that contribute to the success of movies based on a dataset of IMDB movie records. The primary objective was to uncover patterns, correlations, and descriptive insights related to movie genres, languages, directors, budgets, and performance metrics such as IMDB scores and gross earnings. The overall goal was to derive meaningful trends that influence a movie's rating and financial success.

Approach

The project followed a structured and step-wise approach:

1. Data Cleaning
   - Removed duplicates and missing values.
   - Standardized and corrected character encoding issues in director names.
   - Retained only essential columns to streamline analysis.
2. Exploratory Data Analysis (EDA)
   - Used functions like AVERAGE, MEDIAN, MODE, STDEV, VAR, and COUNTIF in Microsoft Excel to perform descriptive statistics.
   - Created charts like bar graphs, pie charts, and scatter plots with trendlines for visualization.
3. Task-wise Analysis
   - Performed genre-wise, language-wise, director-wise, and budget-based analyses.
   - Identified trends using percentile and correlation techniques.

Tech-Stack Used

- Microsoft Excel 2019
  - Data Cleaning: Power Query and manual filtering
  - Statistical Calculations: Built-in Excel formulas
  - Visualization: Excel Charts (bar, pie, scatter plot with trendline)

Formulas Used

Excel Formulas Applied:

- =AVERAGE(range) – to calculate mean (average) values.
- =MEDIAN(range) – to find the central tendency (median).
- =MODE.SNGL(range) – to find the most frequently occurring value (mode).
- =MAX(range) - MIN(range) – to determine the range of scores.
- =VAR.P(range) – to calculate variance for the entire population.
- =STDEV.S(range) – to find standard deviation.
- =COUNTIF(range, criteria) – to count movie occurrences per genre or language.
- =PERCENTILE.EXC(range, 0.9) – to determine top 10% directors based on IMDB ratings.
- =CORREL(range1, range2) – to calculate the correlation coefficient between budget and gross.
- =MAX(range) – to identify the highest profit movie.
- =gross - budget (calculated profit margin for each movie).

Data Cleaning

• Retained only 8 essential columns: movie_title, director_name, duration, imdb_score, genres, language, budget, and gross.

• Removed 122 duplicates using movie_imdb_link, resulting in 4,921 unique rows.

• Handled missing values as follows:

- Replaced 102 missing director_name entries with 'Unknown'
- Replaced all blanks in language with 'English'
- Removed 485 rows with missing budget
- Removed 646 rows with missing gross
- Filled 1 missing duration with the median value

• Corrected encoding errors in director names, replacing special characters with their correct versions.

• Final dataset contains 3,791 rows.

Task A: Movie Genre Analysis

• Drama (1914 movies), Comedy (1492), Thriller (1087) are the top genres.

• Drama movies have the highest mean and median IMDb scores, indicating strong overall audience approval.

• Comedy and Action movies have slightly lower average ratings and higher variance, suggesting more mixed reviews.

• Sci-Fi and Fantasy genres show higher variance and standard deviation, indicating varied audience reception.

• The mode IMDb scores across genres are consistent, mostly around 6.3 to 7.0.

• **Charts Used:**

- Column chart: Total movie count by genre
- 100% stacked column chart: Mean, Median, Mode by genre

Task B: Movie Duration Analysis

• Descriptive statistics show that the average movie duration is approximately [insert mean value], with a median of [insert median value], and standard deviation of [insert SD].

• The scatter plot between duration and IMDB score revealed a moderate positive trend.

• The trendline slope is upward, with an R^2 value of 0.3344.

• **Chart Used:** Scatter plot with trendline (IMDB Score vs Duration)

Task C: Language Analysis

• English dominates the dataset with 3,611 movies but has an average IMDB score of 6.4.

• Non-English languages, such as French (7.3), German (7.7), and Persian (8.1), show higher average ratings.

• Median and standard deviation confirm that some languages yield higher rated movies.

• **Chart Used:** Area chart (Average IMDB Score by Language)

Task D: Director Analysis

• Calculated average IMDB scores for all directors.

• Identified 187 directors in the 90th percentile, with scores above 7.5.

• The overall director score distribution has an average of 6.5 and median of 6.6.

• Top directors significantly elevate the quality and success of their movies.

• **Chart Used:** Top 10 Directors by Average IMDB Score (Column Chart)

Task E: Budget Analysis

• Correlation between budget and gross earnings is 0.097, indicating a very weak positive correlation.

• Budget alone does not strongly predict box office success.

• The highest profit margin movie identified is Avatar, with a budget of $523.5 million and massive gross earnings.

• **Charts Used:**

- Scatter plot with smooth lines for Top 20 Movies (Budget vs Gross)
- Column chart for Top 10 Movies by Profit

Five Whys Approach

1. Why do some movies receive higher IMDb ratings than others?

   - Because they are more critically and publicly appreciated.

2. Why are some movies more appreciated?

   - Because of factors such as better storytelling, direction, acting, or unique content.

3. Why do those factors vary so much between movies?

   - Because different directors and production houses have varying budgets, visions, and creative approaches.

4. Why is budget not a strong predictor of success?

   - Because success depends on execution and audience connection, not just financial investment.

5. Why is audience connection more valuable than budget?

   - Because viewers rate movies based on how the content makes them feel, not how much was spent.

Overall Insights

• Genres: Drama, Comedy, and Thriller are the most common genres. Drama has the highest average IMDB score among the popular genres.

• Languages: English dominates the dataset, but lesser-used languages like Persian, Danish, and Portuguese have higher average ratings.

- Directors: Top 10% of directors have scores significantly above the overall mean of 6.5, with many scoring above 7.5.
- Budgets: There is a weak correlation (r ≈ 0.097) between a movie’s budget and its gross earnings.
- Avatar had the highest profit margin.

Result

Through this project, a deep understanding of movie success factors was achieved. It highlighted the importance of content type, language, and directorial influence on audience reception. The hands-on use of Excel for data analysis and visualization provided practical skills in handling real-world data.

Drive Link

https://docs.google.com/spreadsheets/d/1I4iMqpv9mDbTRceW3k3BwKKDRPHkgD1a/edit?usp=sharing&ouid=115912486417524737249&rtpof=true&sd=true