**Project Report: Analyzing Movie Genres for Microsoft’s new movie studio**

**By: Ahmed Haji**

**Step 1: Problem Statement**

This project aims to identify the movie genres that have been performing exceptionally well at the box office over the past decade. By analyzing metrics such as average ratings, number of votes, runtime, and sales, we provide insights that will assist Microsoft's new movie studio in making data-driven decisions.

**Step 2: Data Collection**

I gathered data from the internet movie database, specifically the 'title.basics.csv' and 'title.ratings.csv' files. My analysis focused on the top ten genres and their performance in terms of ratings, votes, runtime, and sales.

**Step 3: Data Cleaning**

***Handling missing data***: I removed incomplete records, as the volume of missing data was negligible.

***Data formatting:*** Ensured consistent data types and formats.

***Removing duplicates:*** Removed duplicate entries.

***Outlier detection:*** Identified and handled outliers in our datasets.

**Step 4: Data Integration**

I integrated data from ***'title.basics.csv'*** and ***'title.ratings.csv'*** into a comprehensive dataset named ***'merged\_df'***. Further cleaning and preprocessing led to a clean dataset named ***'filtered\_merged\_df'***.

**Step 5: Exploratory Data Analysis (EDA)**

Explored summary statistics to gain initial insights.

Calculated basic statistics (mean, median, standard deviation).

Focused on genre-specific statistics for runtime, ratings, and votes.

Created visualizations to explore relationships between variables and features.

Assessed the relationships between the top ten genres in terms of runtime, ratings, and votes over the last decade.

**Step 6: Feature Engineering**

I created a 'year' column to analyze genre trends over the past ten years.

**Step 7: Data Modeling**

I categorized movies into the top ten genres by sales volume and analyzed their trends over the past decade.

**Step 8: Findings, Results Interpretation, and Recommendations**

**Findings Summary:**

Scatter plot indicates a strong correlation between variables.

Runtime between 75 and 110 minutes is ideal for movie production.

Average ratings correlate with the number of votes, particularly for top genres.

Shifts in audience preferences were observed in genre trends over the years.

Genre Trends over the Years:

Drama dominated from 2010 to 2015, but Documentary surged from 2014 to 2017.

Comedy maintained average popularity.

The decline in 2019 sales was due to streaming service competition and studio mergers.

**Average Ratings:**

Documentary had consistently high ratings (7.2 to 7.5), followed by Drama, Comedy/Drama, and Drama/Romance.

Drama's success was influenced by factors beyond average ratings.

**Number of Votes:**

Drama and Documentary had lower vote counts, while Comedy/Drama/Romance performed well.

**Conclusions:**

Drama is the best-selling genre, followed by Documentary and Comedy.

Ideal runtime is 75 to 110 minutes.

Ratings influence sales more than the number of votes.

Streaming services affect theater sales; online streaming is a significant industry shift.

**Limitations:**

Analysis does not explain audience genre preferences.

Limited datasets restricted deeper exploration of profitability, release dates, and other factors.

**Way Forward:**

Incorporate additional datasets for deeper insights.

Formulate analyses that uncover reasons behind audience preferences.

**Recommendations:**

Prioritize Drama and Documentaries for production.

Maintain a 75 to 110-minute runtime.

Focus on average ratings over vote counts.

Invest in online streaming services for broader audience reach.

***By implementing these findings and recommendations, Microsoft's new movie studio can optimize its strategies for creating successful movies***.