Interactive Dashboard-Movie DataSet MileStone-1 EDA

This code written in python is designed to merge, clean, and analyze movie datasets efficiently. It brings together three different datasets—**movies.csv**, **IMDB-TMDB Movie Metadata Big Dataset**, and **TMDB_movie_dataset_v11.csv**—to create a comprehensive collection of movie information. By handling missing values, identifying outliers, and performing exploratory data analysis (EDA), the script ensures the data is well-prepared for further insights or machine learning applications.

The process begins with loading essential libraries such as pandas, numpy, matplotlib, and seaborn, which help with data handling and visualization. After reading the datasets, the script prints their shape and column names to understand their structure. Some key numerical columns—budget, revenue, runtime, and vote_count—are converted into the correct format (float64) to ensure smooth data processing.

Merging the datasets is done in two phases. First, the movies and imdb_tmdb datasets are combined using an outer join on shared columns like id and title. This method ensures that no data is lost in the process. Missing values in common fields such as language, popularity, release date, budget, and revenue are filled in from either dataset wherever possible. The second phase merges the resulting dataset with the tmdb dataset using the same approach, further enriching the data with additional details.

Once the data is merged, the script moves on to cleaning and preprocessing. Missing values in numerical columns are filled with their average (mean), while categorical values are filled with the most frequent value (mode). Any remaining missing data is dropped. To deal with outliers, the script applies Z-score thresholding, removing extreme values that could distort the analysis. Additionally, numerical data is normalized using Min-Max scaling, making it easier to compare values across different ranges.

With a clean dataset, the script performs exploratory data analysis (EDA) to uncover patterns and trends. It calculates basic statistics such as mean, median, and standard deviation to understand movie attributes. Then, it generates histograms, box plots, and a correlation matrix to visually explore relationships between factors like budget, revenue, popularity, and vote count. This step helps identify key influences on a movie's success.

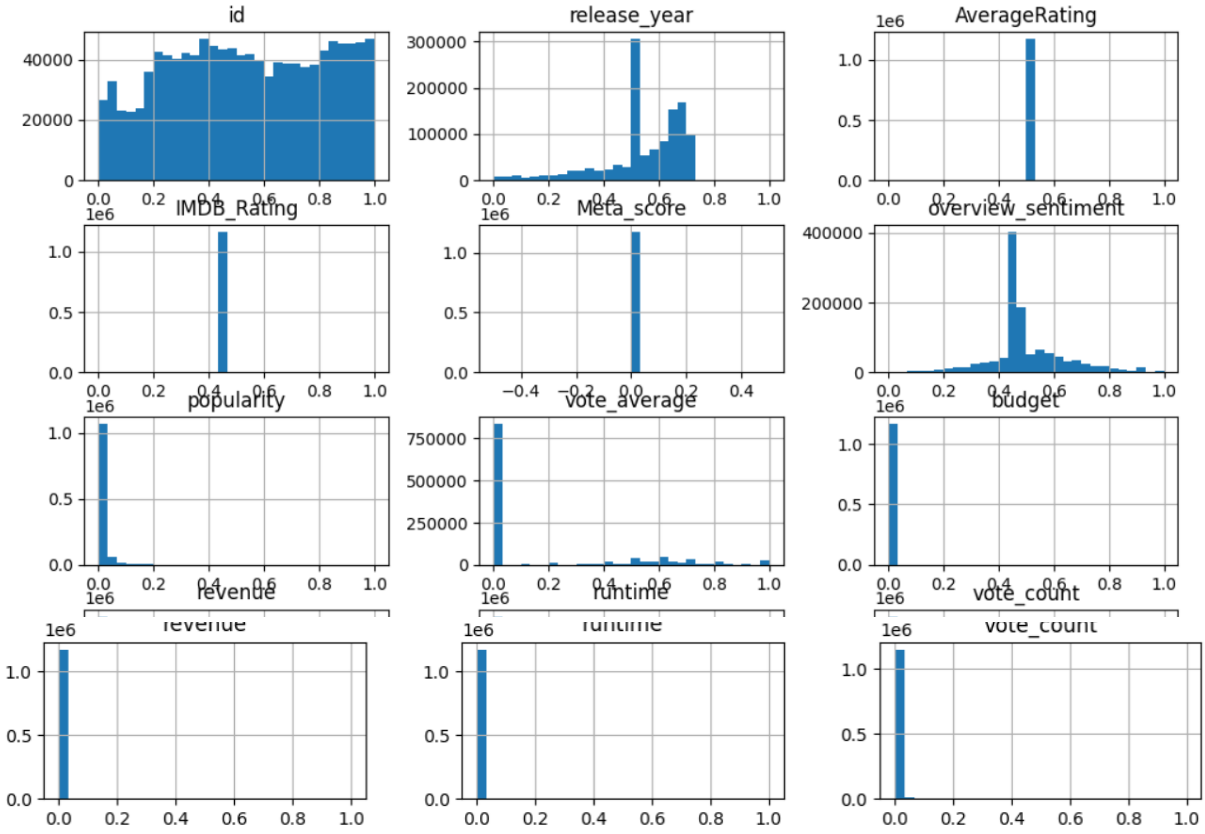
Overall, this script provides a well-structured way to merge, clean, and analyze movie data. It prepares the dataset for further insights, making it valuable for research, business intelligence, and machine learning applications in the film industry.

Key Insights

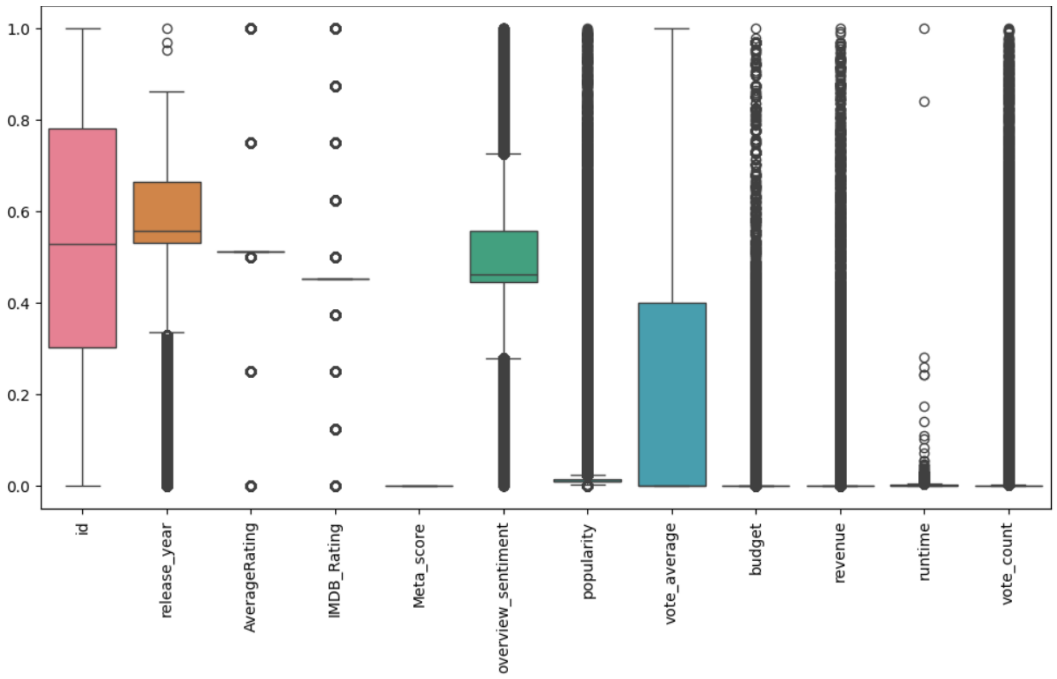
During the EDA process we find that the dataset consists of **1,174,789** movie records
The **mean budget is 0.0019**, and the **mean revenue is 0.0008**, both relatively low.

The minimum for both is 0, which suggests missing or unavailable financial data for many movies.

Most movies have a budget and revenue of nearly 0, indicating that only a few movies have significant financial success.



Below is the box plot



This boxplot visualization provides insights into the distribution and presence of outliers for several key features in the merged movie dataset.

- The budget, revenue, vote count, popularity, and runtime values have a lot of extreme outliers, shown by the many dots above the whiskers in the boxplot.
- Budget and revenue are heavily skewed—most movies have a low budget and revenue, while a few blockbuster films have extremely high numbers.
- Vote count and popularity follow a long-tail pattern, meaning that while most movies receive only a small number of votes and have low popularity, a few movies are widely recognized and get massive attention.