Movie Correlation Analysis Project

This project analyzes a dataset of movies to explore the relationships and correlations between various features, with a primary focus on predicting a movie's **Gross Earnings**.

Project Overview

The core objective was to clean the movie dataset and use correlation techniques to answer a fundamental question in the film industry: **"What factors correlate most highly with a movie's box office success (Gross Earnings)?"**

Key Steps & Findings

1. **Data Loading & Initial Cleaning:**
    - The dataset was loaded using `pandas`.
    - Missing data was identified, with the `budget` column having the highest percentage of missing values (~28%).
    - Duplicate rows were checked for and found not to exist in the raw dataset.
2. **Exploratory Data Analysis (EDA):**
    - A box plot of `gross` earnings showed the presence of **significant outliers** (high-earning blockbusters).
    - The dataset was sorted by `gross` earnings to identify the top-performing films, which included *Avatar*, *Avengers: Endgame*, and *Titanic*.
3. **Correlation Analysis (Numeric Features):**
    - Scatter plots and regression plots were generated to visualize the relationship between key numeric variables:
        - **`budget` vs. `gross`:** Showed a visually clear **positive correlation**.
        - **`score` vs. `gross`:** Showed a weaker, almost negligible, positive correlation.
    - The **Pearson Correlation Coefficient** for numeric columns confirmed the strong relationship:
        - **`budget` and `gross`**: Strong positive correlation (**~0.74**).
        - **`votes` and `gross`**: Moderate positive correlation (**~0.63**).
4. **Handling Categorical Data:**
    - To analyze the correlation between *all* features, including categorical columns (e.g., `name`, `company`, `star`), the `factorize()` method was used to convert them into a numerical representation.
    - A full correlation matrix (heatmap) was created using this numerized dataset.
5. **Conclusion on Correlation:**
   The analysis strongly suggests that **`budget` and `votes` have the highest positive correlation with a movie's `gross` earnings**.
    - **Budget & Gross:** ≈ 0.74 (Strong Correlation)
    - **Votes & Gross:** ≈ 0.63 (Moderate to Strong Correlation)

6. This supports the hypothesis that the **money spent on a film and its popularity/reach (as indicated by the vote count) are the most significant predictors** of its financial success.