Movie Datasetj- Analysis

Milestone 1: Data Collection, Preprocessing, and Exploratory Data Analysis

Project Objective:
The main objective of this script is to **merge, clean, and analyze multiple movie datasets** to create a **comprehensive and high-quality dataset** for further insights, research, or machine learning applications.

Technical Stack:

Programming Language: Python 3.x

Key Libraries:

Data Manipulation: pandas, numpy

Visualization: matplotlib, seaborn

Statistical Analysis: scipy

Machine Learning (future): scikit-learn

Dashboard Development (future): Streamlit

Development Environment:

IDE: Jupyter Notebook

Version Control: Git/GitHub

Documentation: Word/pdf

Dataset Description

This dataset brings together three different sources—**movies.csv, IMDB-TMDB Movie Metadata, and TMDB_movie_dataset_v11.csv**—to create a detailed collection of movie information. It includes key details like **title, release date, budget, revenue, runtime, and popularity**, along with IMDb and TMDb ratings. You'll also find **genres, directors, actors, production companies, and box office earnings**. The data is cleaned by handling missing values, removing outliers, and normalizing numerical features. This makes it perfect for **analyzing movie trends, predicting box office success, or building machine learning**

models to explore what makes a movie successful.

Data Preprocessing Steps

1. Missing Data Handling

Numerical values: Median imputation

Categorical values: Mode imputation

2. Outlier Detection and Treatment

Z-score method to check standard deviation

Removed extreme values that exceeded the threshold to prevent them from affecting the analysis

3. Feature Engineering

Created derived metrics for performance analysis

Standardized time-based features

Key Insights from EDA

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.

Project Timeline

1. Milestone 1(Data Collection & EDA) – February 5-February 21, 2025

2. Milestone 2 (Advanced feature engineering, Model development for performance prediction & Initial dashboard structure) - February 21- March 15, 2025

3. Milestone 3 (Dashboard development, Performance optimization & Final documentation and presentation) - March 15-April 15, 2025

Next Steps

- Development of advanced performance metrics
- Creation of predictive models
- Visualization and dashboard

Dataset Links:

https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies
https://www.kaggle.com/datasets/akshaypawar7/millions-of-movies

https://www.kaggle.com/datasets/ggtejas/tmdb-imdb-merged-movies-dataset