**Movie Dataset Analysis and Visualization**

**Project Description**

This Python-based project focuses on analyzing a movie dataset (mymoviedb.csv). The dataset is cleaned, processed, and explored for meaningful insights, followed by visualizations to understand the distribution of movie genres, vote averages, and release years. The tool also categorizes movies based on their vote average and explores the relationship between movie popularity and release year.

**Tools/Technologies:**

* Python
* pandas
* seaborn
* matplotlib

**Project Goals or Objectives**

The primary objectives of the project were:

1. Load and clean a movie dataset (mymoviedb.csv).
2. Convert and categorize data into meaningful columns like Release\_year and categorized Vote\_Average.
3. Visualize the distribution of genres, vote averages, and the relationship between popularity and release year.
4. Handle missing data, duplicates, and perform general data wrangling.

**What I Learned or Achieved**

* Enhanced data cleaning skills, including handling missing values and removing unwanted columns.
* Gained experience in feature engineering by creating new columns such as Release\_year.
* Improved understanding of data categorization and using pd.cut() to categorize continuous variables.
* Developed a deeper proficiency in data visualization using seaborn and matplotlib.
* Learned how to analyze data distributions, detect extremes (e.g., max and min popularity), and explore key trends.

**Detailed Steps or Methodology**

1. **Data Loading and Inspection:**
   * Loaded the dataset from a CSV file using pd.read\_csv() with a custom line terminator.
   * Inspected the data with methods such as count(), info(), duplicated().sum(), and isna().sum() to check for missing or duplicate values.
2. **Data Cleaning:**
   * Converted the Release\_Date column into datetime64[ns] format for easier date manipulation.
   * Created a new Release\_year column by extracting the year from the Release\_Date.
   * Dropped rows with missing values (dropna()) and removed irrelevant columns such as Release\_Date, Overview, Original\_Language, and Poster\_Url.
3. **Data Categorization:**
   * Developed a function avg\_cat() to categorize the Vote\_Average column into four levels: "Not Popular", "below\_Average", "Average", and "Popular".
   * Used pd.cut() to bin the values into these categories based on statistical quantiles.
4. **Exploratory Data Analysis (EDA):**
   * Visualized the distribution of genres using a count plot (sns.catplot()), showing how genres are distributed across the dataset.
   * Analyzed the Vote\_Average distribution using another count plot to examine how movies are distributed across vote categories.
   * Identified the highest and lowest popularity values in the dataset to examine the extremes.
   * Plotted a histogram of the Release\_year to visualize the distribution of movies over time.
5. **User Insights and Exploration:**
   * Investigated the most and least popular movies using max() and min() functions on the Popularity column.
   * Provided visualizations that allowed insights into genre distribution and the relationship between popularity and release year.