Movie Dataset Analysis Using Pandas and Matplotlib

Project Overview

This project focuses on analyzing a movie dataset using **Pandas** for data manipulation and **Matplotlib** for data visualization. The dataset contains information about movies, including attributes such as movieID, title, and genres. The project involves cleaning the data through slicing and indexing, and applying various statistical functions to gain insights into the dataset.

Dataset Attributes

The dataset contains the following columns:

- **movieID**: A unique identifier for each movie.
- **title**: The title of the movie.
- **genres**: The genre(s) the movie belongs to (e.g., Action, Comedy, Drama, etc.).

Key Objectives

1. **Data Cleaning**: Perform data slicing and indexing to clean the dataset and ensure that the data is in a usable format.
2. **Statistical Analysis**: Apply statistical functions such as mean, minimum, maximum, standard deviation, mode, and correlation to extract insights from the data.
3. **Visualization**: Use **Matplotlib** to plot graphs and visualizations, helping to better understand trends in the dataset.

Libraries Used

- **Pandas**: For data manipulation, slicing, indexing, and applying statistical functions.
- **Matplotlib**: For plotting graphs and visualizing the statistical analysis.

Features

- **Data Cleaning**:
  - **Slicing and Indexing**: Clean the dataset by slicing unwanted data and indexing relevant columns like movieID, title, and genres.
  - **Handling Duplicates and Missing Values**: Ensure there are no duplicate entries and handle any missing or null values.
- **Statistical Functions**:
  - **Mean**: Calculate the average of numerical values (if applicable) in the dataset, such as movie ratings or counts (optional extension).
  - **Min/Max**: Determine the minimum and maximum values of certain attributes, such as the length of movie titles or the distribution of genres.
  - **Standard Deviation**: Measure the dispersion of the data, particularly around average ratings or movie counts per genre.
  - **Mode**: Identify the most frequently occurring genre or title length.

    - **Correlation**: Analyze the correlation between numerical variables (if any, such as ratings, optional).

- **Visualization**:
    - Plot graphs such as bar charts, histograms, or pie charts to visualize the distribution of movies across genres, movie counts, and other attributes.

Data Analysis Process

1. **Loading and Exploring the Data**:
    - Load the movie dataset into a Pandas DataFrame.
    - Explore the structure of the dataset using functions like head(), info(), and describe().
2. **Data Cleaning**:
    - Slice and index the dataset to isolate relevant columns (movieID, title, genres).
    - Handle any missing values or duplicates, ensuring the dataset is clean for analysis.
3. **Statistical Analysis**:
    - Apply statistical functions like mean(), min(), max(), std(), and mode() to analyze the dataset.
    - If applicable, use correlation analysis to explore relationships between numerical attributes.
4. **Data Visualization**:
    - Use **Matplotlib** to create visualizations such as:
        - Bar charts representing the number of movies per genre.
        - Histograms showing the distribution of movie title lengths.
        - Pie charts to visualize genre distribution.