Name: Nagam Amareswar
Enrollment No: 23324013
Branch: BS MS Physics

Data Cleaning and Exploration Report: Movies Dataset

Introduction

This report details the initial findings from an analysis of a dataset containing information about movies. The data cleaning process is followed by exploratory data analysis (EDA) to summarize key aspects of the dataset.

Data Loading and Inspection

The data was loaded and then converted into Pandas DataFrame. The initial inspection revealed the dataset has the following characteristics:
Columns: The dataset appears to have nine columns: 'MOVIES,' 'YEAR,' 'GENRE,' 'RATING,' 'ONE-LINE,' 'STARS,' 'VOTES,' 'RunTime,' and 'Gross.'
Rows: The dataset contains 9999 rows of data, suggesting a relatively large sample of movies.
Data Types: Based on the output of dtypes, several columns have issues with their data types due to inconsistent or missing values. They would be required to be treated accordingly during the data preparation process. For example:

1) 'YEAR' appears to be mixed with date ranges (e.g., '2010–2022') and needs more cleanup.
2) 'RATING' has some missing values and would also be best handled while cleaning.
3) 'VOTES' and 'RunTime' have some string values and are not automatically read as integers.
4) 'Gross' has currency symbols and needs to be cleaned appropriately

Data Cleaning Steps

Handling Duplicates: All duplicate columns were dropped

Handle Missing Values: Missing values in the dataset were addressed through appropriate methods.

Addressing Invalid characters: and Inconsistent formats:
1) Non-numeric characters were removed from the "YEAR" column to ensure a consistent numeric format.

2)Special characters (except spaces) were stripped from the "ONE-LINE" column (likely a description field).

3)Newline characters were removed from the "GENRE" column to maintain uniformity.

Advanced Data Transformation:

For the YEAR column, a conditional formatting operation is applied to restructure years longer than four digits, introducing a comma separator for improved readability. This transformation converts, for example, "20212022" to "2021, 2022", indicating potentially a range of years for a series or sequel.

One of the more complex transformations involves splitting the STARS column into two separate columns: directors and stars. This operation uses pandas string splitting method with the expand parameter set to True, creating new columns for the split components. The original STARS column is then removed as its information has been redistributed.

Exploratory Data Analysis

Barplots were used to visualize the frequency distribution of categorical variables, such as "Counts (of Movies) vs. Genre." The average rating for each genre was calculated. A bar plot was then created to show the number of movies in each genre that exceeded the average rating.

A heatmap was used to analyze the correlation between numerical variables, such as RATING and RunTime. The heatmap revealed a correlation of -0.15, indicating a weak negative relationship between these variables.

Conclusion

The cleaned movie dataset is now more reliable, with duplicates, missing values, and inconsistencies resolved. Key insights from EDA show genre distribution, movies exceeding average ratings, and a weak negative correlation between RATING and RunTime. This refined dataset sets the stage for deeper analysis, trend exploration, and future predictive modeling.