Data Cleaning in **Python**

Overview

The data cleaning process involves loading the dataset and previewing its structure, then removing prefixes from text columns like "Writtenby" and "Narratedby." Next, we split combined columns, such as ratings and review counts, into separate numerical columns, while handling placeholder values like "Not rated yet." We then adjust data types for optimal analysis, converting prices to float, categorizing columns with limited values, and converting dates to datetime format. For time data, we standardize units, extract hours and minutes, and calculate total time in minutes. Duplicates are identified and removed based on key columns, and missing values are addressed according to analysis needs. Finally, we save the cleaned data, ready for analysis or modeling.

Problem Statement

This script addresses the need to clean and prepare an audiobook dataset downloaded from Audible for further analysis or modeling. The raw data likely contains inconsistencies, missing values, and formatting issues that hinder effective analysis.

Goals

- **Cleanse Data:**
  - Remove unnecessary prefixes from text columns.
  - Split combined columns into separate numerical columns.
  - Standardize data types for each column.
  - Extract and combine time components into a new column.
  - Verify data ranges and address outliers.
  - Identify and remove duplicate entries based on specific criteria.
  - Handle missing values based on analysis needs.

- **Prepare Data:**
  - Transform data into a format suitable for analysis or modeling.

Dataset

The dataset “audible_raw.csv” is provided, containing information about audiobooks downloaded from Audible, covering releases from 1998 through pre-planned launches in 2025. It likely includes details like title, author, narrator, duration, release date, language, rating, price, and potentially more.

- **"name"** - The name of the audiobook.
- **"author"** - The audiobook's author.
- **"narrator"** - The audiobook's narrator.
- **"time"** - The audiobook's duration, in hours and minutes.
- **"releasedate"** - The date the audiobook was published.
- **"language"** - The audiobook's language.
- **"stars"** - The average number of stars (out of 5) and the number of ratings (if available).
- **"price"** - The audiobook's price in INR (Indian Rupee).

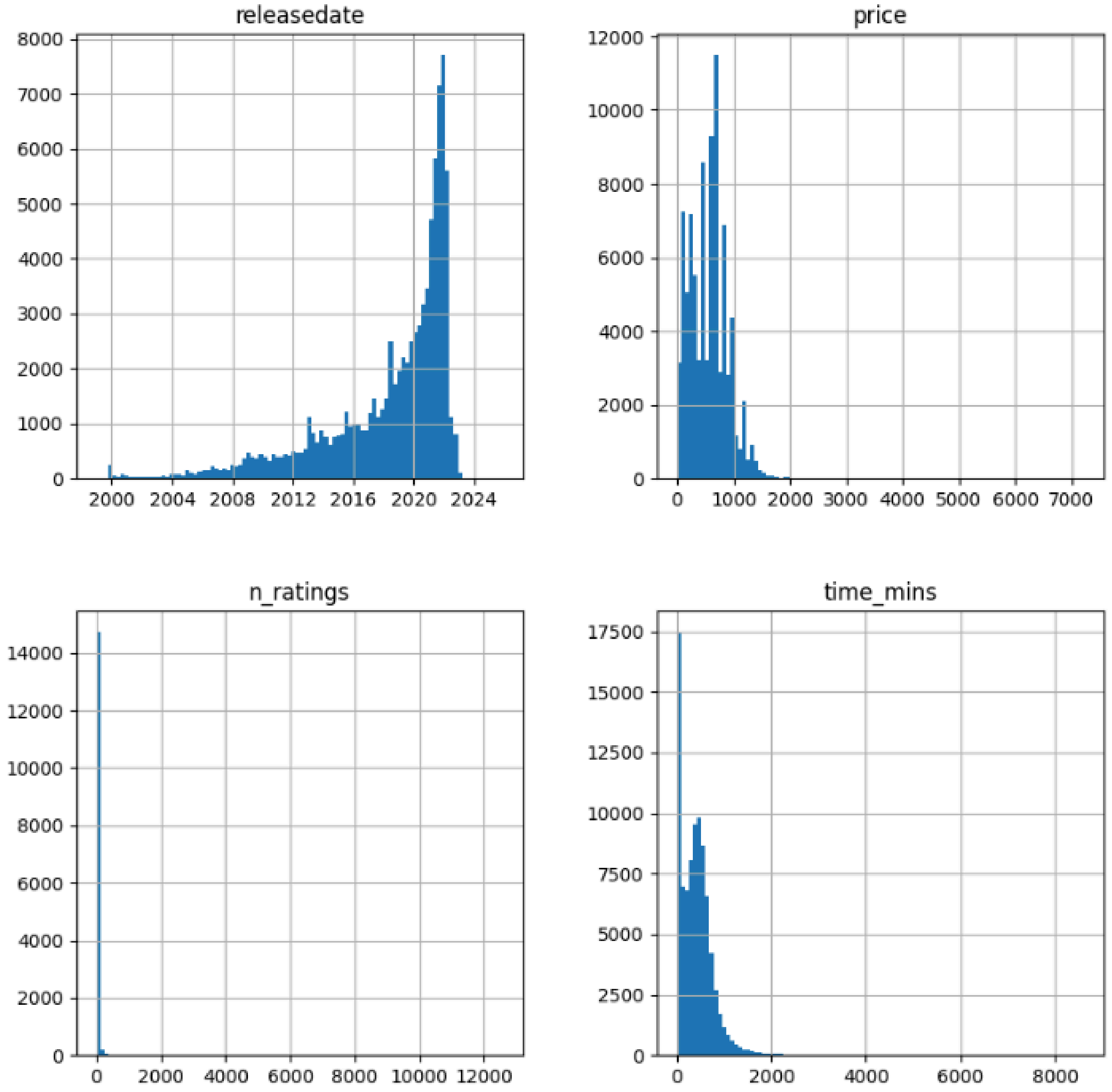
Source: https://www.kaggle.com/datasets/snehangsude/audible-dataset

Outputs

Here are the steps:

1. Loading and Reviewing the Dataset
2. Cleaning Text Data in Author and Narrator Columns
   - Remove prefixes like **"Writtenby:"** and **"Narratedby:"** from author and narrator columns.
3. Splitting the Stars Column into Ratings and Reviews Count
   - Separate the **"stars"** column containing average rating and review count into distinct numerical columns for **"rating_stars"** and **"n_ratings"**.
4. Adjusting Data Types
   - Convert **"price"** to a float data type.
   - Set **"rating_stars"** as a categorical data type.
   - Convert **"releasedate"** to a datetime format.
5. Extracting Hours and Minutes from the Time Column
   - Extract hours and minutes from the "time" column and create a new column **"time_mins"** representing total duration in minutes.
6. Verifying Data Ranges
   - Use histograms and descriptive statistics to check for outliers and ensure values fall within expected ranges.
   - Transform prices from INR to USD for this exercise.
   - Standardize capitalization in the **"language"** column.
7. Identifying and Removing Duplicates
   - Identify and remove duplicate entries based on a subset of columns (excluding release date) to keep the record with the most recent release date.
8. Handling Missing Values
   - Identify missing values using **".isna().sum()"**. Decisions about handling missing values depend on the analysis goals.
9. Saving the Cleaned Data
   - Save the cleaned dataset as **`audible_clean.csv`** without the index.

Verifying Data Ranges



Histogram Analysis

- **Release Date**
  - The histogram shows a clear upward trend, indicating an increasing number of audiobooks released over time.
  - There are a few peaks in the data, which might correspond to periods with increased audiobook production or popularity.
  - The distribution is right-skewed, suggesting that most audiobooks were released relatively recently.
- **Price**
  - The price histogram shows a right-skewed distribution, with most audiobooks priced lower and fewer audiobooks priced higher.
  - There's a peak around the lower price range, suggesting a concentration of audiobooks in the budget-friendly segment.
- **Number of Ratings (n_ratings)**
  - The n_ratings histogram is also right-skewed, with most audiobooks having a relatively low number of ratings.
  - This indicates that a majority of audiobooks receive fewer ratings compared to a few popular titles that garner a large number of reviews.
- **Time in Minutes (time_mins)**
  - The time_mins histogram shows a right-skewed distribution, with most audiobooks having a shorter duration.
  - There's a long tail towards the right, indicating the presence of longer audiobooks.

Insights (Potential)

- **The data cleaning process** helps prepare the audiobook data for analysis, ensuring data consistency and accuracy.
- **Examining data ranges and distributions** helps identify potential outliers or areas requiring further investigation.
- **Handling missing values appropriately** is crucial to avoid introducing bias during analysis.

Conclusion

By applying data cleaning techniques, this script transforms a raw audiobook dataset into a well-structured and prepared format for further analysis or modeling tasks. This allows for the extraction of meaningful insights from the data.