Lab1 explanation

The lab focuses on data analysis using Pandas and introduces concepts relevant to text mining and natural language processing (NLP). Below is a summary and explanation of the main sections:

Student Name	ID	Section
Lujain Bukassim Almarri	S20106753	1

Learning Goals:

- Systematically handle missing values.
- Parse data columns to create new columns.
- Use groupby to aggregate and analyze data by specific features.

Dataset:

- A collection of approximately 6,000 "best books" from Goodreads, fetched and saved as a CSV file.

Workflow:

1. **Loading and Cleaning Data:**
    - Read and clean data to address missing values.
    - Ensure consistent formatting and structure.
2. **Parsing Data Columns:**

- Extract detailed information from composite columns (e.g., splitting author names or genres into separate attributes).

3. **Grouping Data:**

   - Group data by specific columns (e.g., year or author) to analyze subsets and calculate aggregates.

Steps:

1. **Loading Data:**

   - Import and inspect the dataset for missing or invalid values.

   - Add appropriate column names and ensure the data types align with expected formats.

2. **Cleaning Missing Data:**

   - Identify and handle missing values in columns like year, rating_count, and review_count.

   - Replace or remove rows with NaN values to maintain dataset integrity.

3. **Parsing Data:**

   - Extract authors' names from URLs.

   - Parse genre URLs into a readable format and join multiple genres into a single string.

4. **Grouping:**

   - Use groupby to aggregate data by year or author.

- Explore patterns such as the number of books per author or the best-rated books for each year.

Results:

- The cleaned and parsed data is saved as a new CSV file.
- Analyses like identifying top authors or trends in book ratings are performed.

Explanation of Lab Concepts

1. **Data Cleaning:**
   - Missing data (NaN) can disrupt analysis. Cleaning involves removing or imputing these values to ensure consistent and accurate processing.
2. **Parsing:**
   - Parsing involves breaking down complex data into simpler parts. For instance:
     - Splitting an author's URL to extract their name.
     - Separating genre URLs into individual genres.
3. **Grouping and Aggregating:**
   - Grouping organizes data into subsets based on a shared attribute (e.g., year, author).
   - Aggregation functions (like mean or count) summarize grouped data, enabling insights like:

- The most prolific authors.
- Trends in book ratings over time.

4. **Exploratory Data Analysis (EDA):**
    - Histograms, scatter plots, and aggregation provide insights into global and group-specific patterns, forming the foundation for further modeling or decision-making.
5. **Pandas Functions in Action:**
    - pd.read_csv(): Load datasets.
    - .isnull(), .dropna(): Handle missing values.
    - .astype(): Convert data types.
    - .groupby(): Group and analyze subsets of data.
    - .map(): Apply functions across a DataFrame column.