Data Cleaning, Schema Matching, and Data Matching Report

1. Introduction

This report details the process of aligning and analyzing two datasets (Table A and Table B) that originally contained different attributes. Our primary objectives were:

1. To unify the schemas of the two tables.
2. To reduce or finalize the schema to a specific set of attributes.
3. To conduct a basic data quality assessment of the resulting table (focusing on missing values, attribute types, textual analyses, outliers, etc.).

We also generated histograms to visualize data distributions and identify potential anomalies.

2. Schema Matching

Original Schemas

- **Table A** initially had the attributes:
  ['ID', 'Title', Tomatometer, 'Audience Score', 'Latest Episode', 'Series URL', 'Image URL']
- **Table B** initially had the attributes:
  ['ID', 'Title', 'Year', 'Rating', 'Genre', 'Description', 'IMDb URL', 'Image URL']

Changes

I changed Tomatometer to Rating since they both are a kind of rating. I'm not sure if I should have different columns for each rating, so if you think I should, please let me know.

Common Columns

After comparing both tables, we identified the following common columns:

['Rating', 'Title', 'Image URL', 'ID']

Final Schema (S)

The final set of attributes is:

['Rating', 'Title', 'Image URL', 'ID']

This leaves us with four attributes in common. All subsequent data-quality analyses were performed on these attributes.

3. Data Quality Analysis (Table A)

Below are the results of the attribute-by-attribute analysis on Table A. (In this case, both tables shared the same attributes, so the analysis can be considered representative.)

1. **Attribute: Rating**
    - **Type:** Numeric
    - **Missing:** 0.00 fraction (0.0%)
    - **Observations / Potential Imputation:**
        - No missing values.
        - Ratings range from 0 to 100, with no apparent data-entry issues based on the histogram.
        - If there were missing values, common strategies could include using mean/median or domain knowledge to fill them in.
2. **Attribute: Title**
    - **Type:** Textual
    - **Missing:** 0.00 fraction (0.0%)
    - **Average Length:** 12.3 characters
    - **Min Length:** 3 characters
    - **Max Length:** 41 characters
    - **Observations / Potential Imputation:**
        - No missing values.
        - Titles are fairly short (typical for TV shows or short movie names).
        - Potential data-quality issues could include inconsistent capitalization or additional metadata in the title.
3. **Attribute: Image URL**
    - **Type:** Textual
    - **Missing:** 0.00 fraction (0.0%)
    - **Average Length:** 206.2 characters

- **Min Length:** 162 characters
- **Max Length:** 221 characters
- **Observations / Potential Imputation:**
  - No missing values.
  - URLs appear to be valid, but no deeper validation (e.g., broken links) was performed.
  - If URLs were missing or malformed, we might try to reconstruct them from other metadata or leave them as null.

4. **Attribute: ID**
   - **Type:** Textual
   - **Missing:** 0.00 fraction (0.0%)
   - **Average Length:** 4.4 characters
   - **Min Length:** 2 characters
   - **Max Length:** 5 characters
   - **Observations / Potential Imputation:**
     - No missing values.
     - The ID appears to be a short string. If meant to be numeric, it could be converted.
     - Verify uniqueness if the ID is intended to be a primary key.

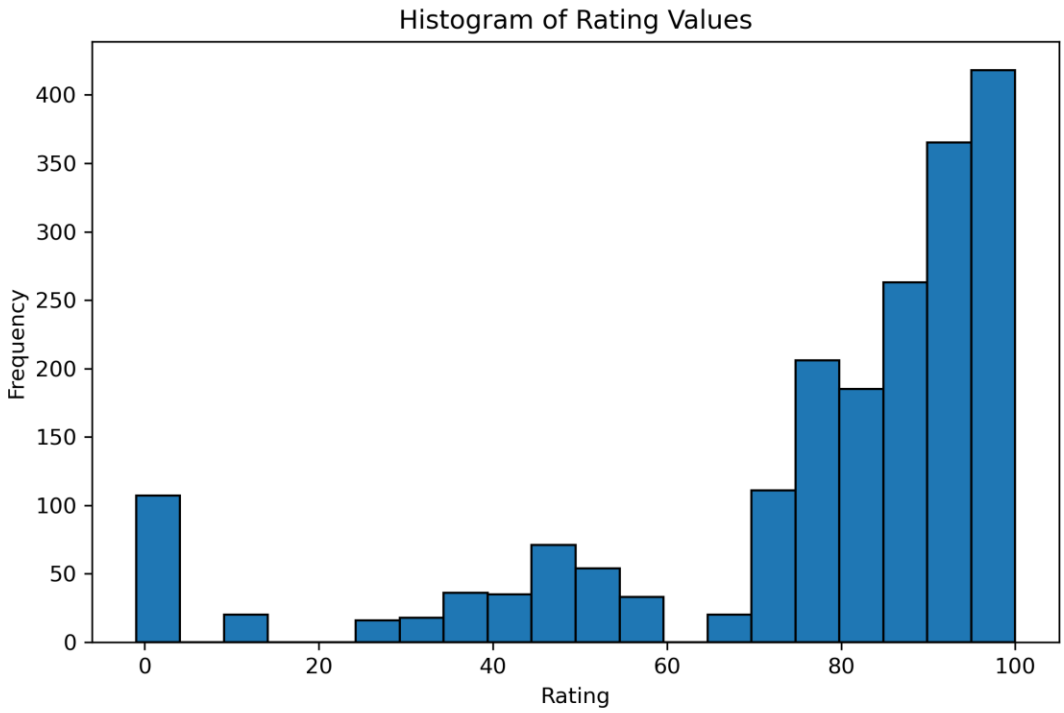
---

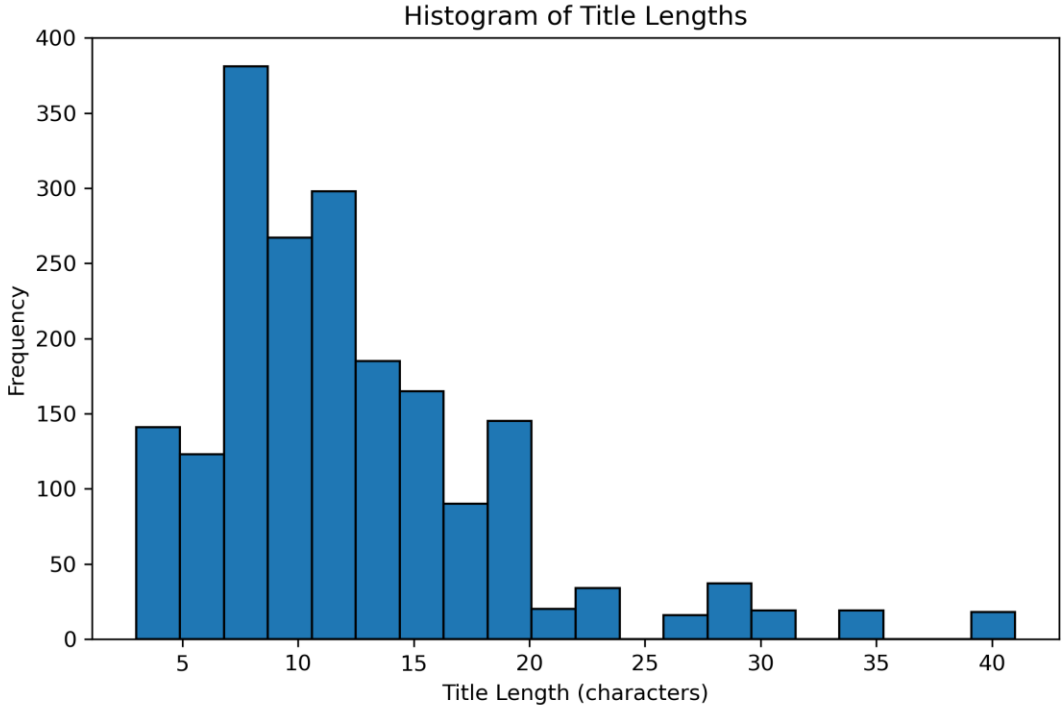
4. Histograms and Visual Analysis

Two histograms were generated to assess potential outliers or anomalies:

1. **Histogram of Rating Values**
   - Shows the distribution of Rating from 0 to 100.
   - The distribution appears to be skewed toward higher ratings (60–100 range).
   - No clear outliers, though the small cluster around 0–10 might warrant closer inspection.

2. **Histogram of Title Lengths**
    - Displays the number of characters in the Title attribute.
    - Most titles range from about 5 to 15 characters in length.
    - A few longer titles (up to 41 characters) may require validation to ensure they aren't including extraneous text or metadata.


Histogram of Rating Values
Frequency
Rating
0
20
40
60
80
100
0
50
100
150
200
250
300
350
400


Histogram of Title Lengths
Frequency
Title Length (characters)
5
10
15
20
25
30
35
40
0
50
100
150
200
250
300
350
400

5. Tools Used

- **Python 3**: Primary programming language.
- **Pandas**: For data manipulation, schema alignment, and missing-value detection.
- **Matplotlib**: For histogram generation and other visualizations.
- **NumPy**: For numeric operations and array handling.
- **Jupyter Notebook / IDE**: For interactive exploration and scripting.

6. Conclusions and Next Steps

- **Schema Alignment**: We successfully unified Table A and Table B to a common schema of four attributes.
- **Data Quality**: The attributes in Table A show no missing values, and the data types (numeric vs. textual) are consistent. Some potential next steps include:
    - Confirming the **ID** uniqueness if it is intended as a primary key.
    - Checking **Image URLs** for validity and removing or correcting broken links.
    - Standardizing **Title** formats (e.g., removing trailing spaces, consistent capitalization).
    - Possibly removing the outliers in the rating and title columns.