Exploratory Data Analysis Report

Introduction

The following report summarizes the exploratory data analysis (EDA) performed on a dataset comprising two sets - a training set and a testing set. The dataset includes fields such as ENTRYTYPE, title, editor, year, publisher, author, and abstract.

Dataset Overview

- **Training Set**: Contains 65,914 entries.
- **Testing Set**: Contains 21,972 entries.
- **Columns**:
    - ENTRYTYPE: Type of the entry
    - Title: Title of the entry.
    - Editor: Editor's name
    - Year: Publication year.
    - Publisher: Publisher's name.
    - Author: Author(s) of the entry.
    - Abstract: Abstract of the entry.

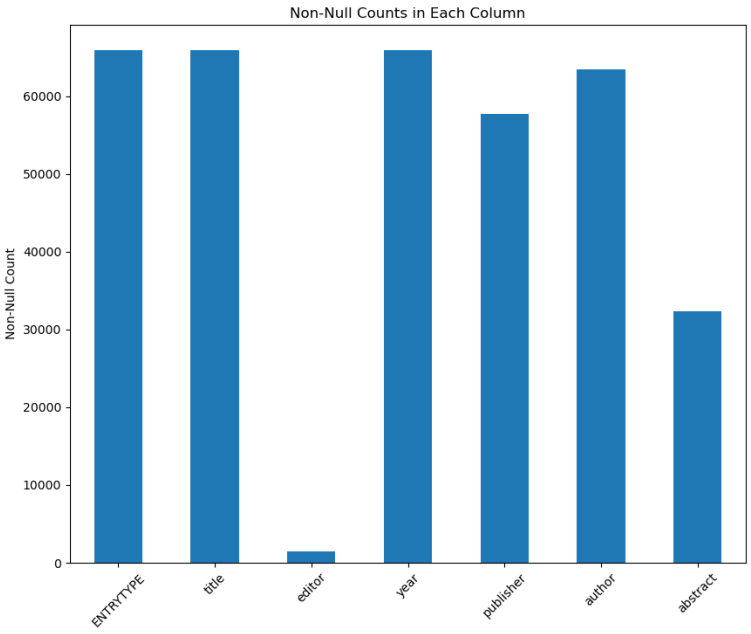
Data Cleaning and Preprocessing

- **Missing Values**:
    - Publisher, author, and abstract columns have significant missing values in the training set (up to 50.46% missing in the abstract column).
    - Similar pattern observed in the testing set.
- **Data Transformation**:
    - The 'year' column converted to numeric, handling non-numeric entries.
    - Concatenated author names into a single string.
    - Dropped the 'editor' column due to high sparsity.
    - Removed duplicate entries based on the title.
    - Standardized the 'year' column.
    - Missing value was treated by filling with empty strings

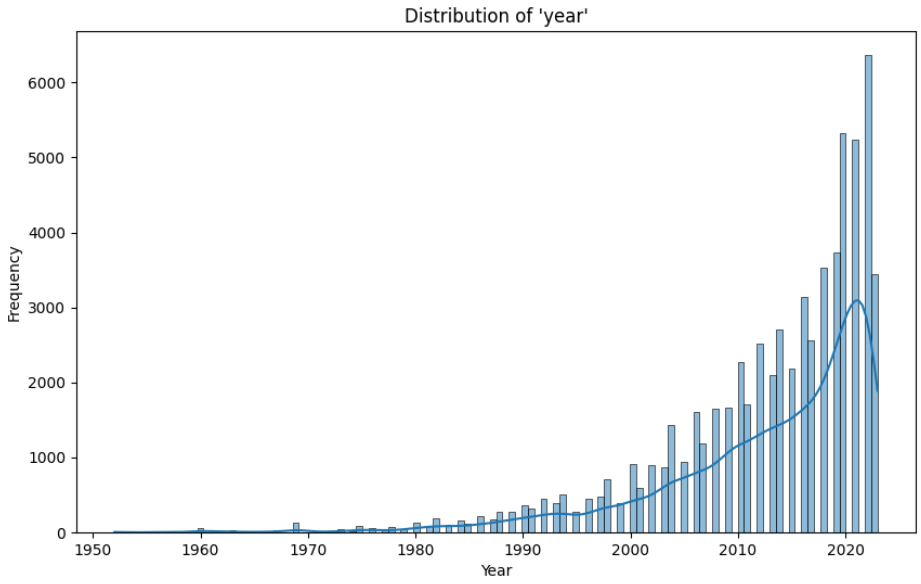
- **Categorization**: ENTRYTYPE converted to a categorical variable for both training and testing sets.

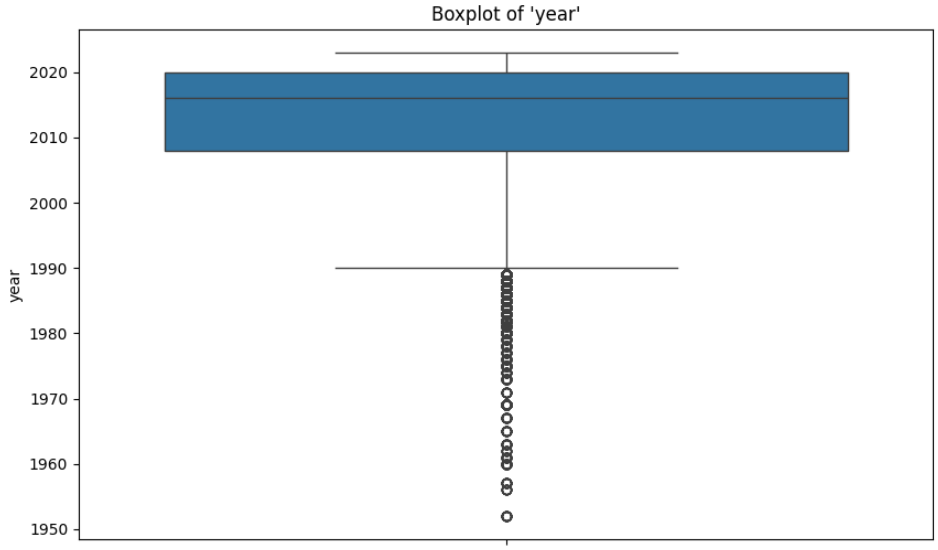
Data Analysis

- **Non-Null Count Analysis**:
    - A bar graph created shows non-null counts across different columns.
    - Publisher, author, and abstract have lower counts indicating missing data.



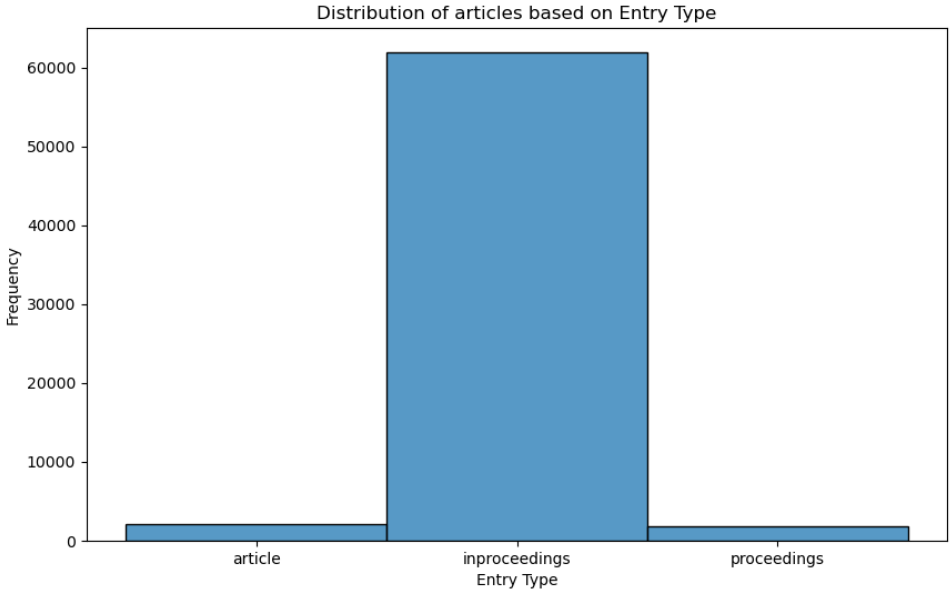
- **Year Analysis**:
    - Distribution of publication year shown using a histogram.
    - Skewness calculated at -1.65, indicating a left-skewed distribution.
    - Boxplot highlights the presence of outliers in the year data.
    - Outliers identified using IQR, ranging from 1952 to 1989.
    - Post-standardization, the distribution of the 'year' column was re-evaluated.

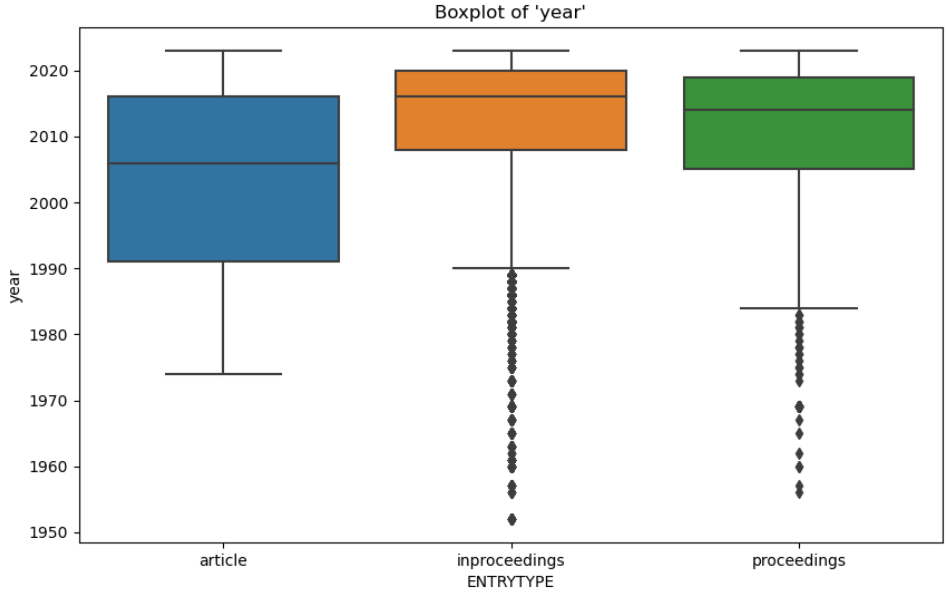




- **Group Analysis**:
  - The data grouped by ENTRYTYPE and descriptive statistics provided for the 'year' column in each group.
  - Distribution of articles based on ENTRYTYPE visualized.
  - The median publication year for all ENTRYTYPEs is relatively recent (post-2000), which may reflect an increasing trend in publications or availability of digital records.

- Inproceedings have the most outliers, which could be due to the nature of conference proceedings often being published in volumes with varied publication dates.
- The presence of outliers in all categories indicates that while most publications occurred in recent years, there is still a significant number of publications dating back several decades.





Observations

- **Year Distribution**: The 'year' feature is left-skewed with most publications being more recent.
- **Outliers**: A significant number of outliers in the earlier years, suggesting fewer publications in the early periods covered by the dataset.
- **Entry Type**: Most entries are of the type 'inproceedings', followed by 'article', and 'proceedings'.
- **Missing Data**: The editor column was mostly empty and thus removed. Abstracts are missing in about half of the cases.

This EDA provides a foundational understanding of the dataset's structure, missing values, and distributions. The insights gathered are crucial for further data modeling and analysis. The cleaned and transformed dataset, particularly with standardized years and categorized entry types, is better suited for advanced analytical techniques.